

```

49     * with the edges of the squares */
50     Point p1 = extend(this.middle(), other.middle(), this.size);
51     Point p2 = extend(this.middle(), other.middle(),
52                         -1 * this.size);
53     Point p3 = extend(other.middle(), this.middle(), other.size);
54     Point p4 = extend(other.middle(), this.middle(),
55                         -1 * other.size);
56
57     /* Of above points, find start and end of lines. Start is
58      * farthest left (with top most as a tie breaker) and end is
59      * farthest right (with bottom most as a tie breaker. */
60     Point start = p1;
61     Point end = p1;
62     Point[] points = {p2, p3, p4};
63     for (int i = 0; i < points.length; i++) {
64         if (points[i].x < start.x ||
65             (points[i].x == start.x && points[i].y < start.y)) {
66             start = points[i];
67         } else if (points[i].x > end.x ||
68                 (points[i].x == end.x && points[i].y > end.y)) {
69             end = points[i];
70         }
71     }
72
73     return new Line(start, end);
74 }

```

The main goal of this problem is to see how careful you are about coding. It's easy to glance over the special cases (e.g., the two squares having the same middle). You should make a list of these special cases before you start the problem and make sure to handle them appropriately. This is a question that requires careful and thorough testing.

- 7.6** Given a two-dimensional graph with points on it, find a line which passes the most number of points.

pg 102

## SOLUTION

This solution seems quite straightforward at first. And it is -- sort of.

We just "draw" an infinite line (that is, not a line segment) between every two points and, using a hashtable, track which line is the most common. This will take  $O(N^2)$  time, since there are  $N^2$  line segments.

We will represent a line as a slope and y-intercept (as opposed to a pair of points), which allows us to easily check to see if the line from  $(x_1, y_1)$  to  $(x_2, y_2)$  is equivalent to the line from  $(x_3, y_3)$  to  $(x_4, y_4)$ .

To find the most common line then, we just iterate through all lines segments, using a

hashtable to count the number of times we've seen each line. Easy enough!

However, there's one little complication. We're defining two lines to be equal if the lines have the same slope and y-intercept. We are then, furthermore, hashing the lines based on these values (specifically, based on the slope). The problem is that floating point numbers cannot always be represented accurately in binary. We resolve this by checking if two floating point numbers are within an epsilon value of each other.

What does this mean for our hashtable? It means that two lines with "equal" slopes may not be hashed to the same value. To solve this, we will round the slope down to the next epsilon and use this flooredSlope as the hash key. Then, to retrieve all lines that are *potentially* equal, we will search the hashtable at three spots: flooredSlope, flooredSlope - epsilon, and flooredSlope + epsilon. This will ensure that we've checked out all lines that might be equal.

```
1 Line findBestLine(GraphPoint[] points) {
2     Line bestLine = null;
3     int bestCount = 0;
4     HashMap<Double, ArrayList<Line>> linesBySlope =
5         new HashMap<Double, ArrayList<Line>>();
6
7     for (int i = 0; i < points.length; i++) {
8         for (int j = i + 1; j < points.length; j++) {
9             Line line = new Line(points[i], points[j]);
10            insertLine(linesBySlope, line);
11            int count = countEquivalentLines(linesBySlope, line);
12            if (count > bestCount) {
13                bestLine = line;
14                bestCount = count;
15            }
16        }
17    }
18    return bestLine;
19 }
20
21 int countEquivalentLines(ArrayList<Line> lines, Line line) {
22     if (lines == null) return 0;
23     int count = 0;
24     for (Line parallelLine : lines) {
25         if (parallelLine.isEquivalent(line)) count++;
26     }
27     return count;
28 }
29
30 int countEquivLines(
31     HashMap<Double, ArrayList<Line>> linesBySlope, Line line) {
32     double key = Line.floorToNearestEpsilon(line.slope);
33     double eps = Line.epsilon;
34     int count = countEquivalentLines(linesBySlope.get(key), line) +
35         countEquivalentLines(linesBySlope.get(key - eps), line) +
36         countEquivalentLines(linesBySlope.get(key + eps), line) +
```

```
36     countEquivalentLines(linesBySlope.get(key + eps), line);
37     return count;
38 }
39
40 void insertLine(HashMap<Double, ArrayList<Line>> linesBySlope,
41     Line line) {
42     ArrayList<Line> lines = null;
43     double key = Line.floorToNearestEpsilon(line.slope);
44     if (!linesBySlope.containsKey(key)) {
45         lines = new ArrayList<Line>();
46         linesBySlope.put(key, lines);
47     } else {
48         lines = linesBySlope.get(key);
49     }
50     lines.add(line);
51 }
52
53 public class Line {
54     public static double epsilon = .0001;
55     public double slope, intercept;
56     private boolean infinite_slope = false;
57
58     public Line(GraphPoint p, GraphPoint q) {
59         if (Math.abs(p.x - q.x) > epsilon) { // if x's are different
60             slope = (p.y - q.y) / (p.x - q.x); // compute slope
61             intercept = p.y - slope * p.x; // y intercept from y=mx+b
62         } else {
63             infinite_slope = true;
64             intercept = p.x; // x-intercept, since slope is infinite
65         }
66     }
67
68     public static double floorToNearestEpsilon(double d) {
69         int r = (int) (d / epsilon);
70         return ((double) r) * epsilon;
71     }
72
73     public boolean isEquivalent(double a, double b) {
74         return (Math.abs(a - b) < epsilon);
75     }
76
77     public boolean isEquivalent(Object o) {
78         Line l = (Line) o;
79         if (isEquivalent(l.slope, slope) &&
80             isEquivalent(l.intercept, intercept) &&
81             (infinite_slope == l.infinite_slope)) {
82             return true;
83         }
84         return false;
85     }
```

86 }

We need to be careful about the calculation of the slope of a line. The line might be completely vertical, which means that it doesn't have a y-intercept and its slope is infinite. We can keep track of this in a separate flag (`infinite_slope`). We need to check this condition in the equals method.

- 7.7** Design an algorithm to find the  $k$ th number such that the only prime factors are 3, 5, and 7.

pg 102

**SOLUTION**

We know that, by definition, any such number will look like  $3^a * 5^b * 7^c$ .

Let's picture a list of all numbers in that form. This problem asks us to find the  $k$ th such number.

1	-	$3^0 * 5^0 * 7^0$
3	3	$3^1 * 5^0 * 7^0$
5	5	$3^0 * 5^1 * 7^0$
7	7	$3^0 * 5^0 * 7^1$
9	$3*3$	$3^2 * 5^0 * 7^0$
15	$3*5$	$3^1 * 5^1 * 7^0$
21	$3*7$	$3^1 * 5^0 * 7^1$
25	$5*5$	$3^0 * 5^2 * 7^0$
27	$3*9$	$3^3 * 5^0 * 7^0$
35	$5*7$	$3^0 * 5^1 * 7^1$
45	$5*9$	$3^2 * 5^1 * 7^0$
49	$7*7$	$3^0 * 5^0 * 7^2$
63	$3*21$	$3^2 * 5^0 * 7^1$

We know that, since  $3^{a-1} * 5^b * 7^c < 3^a * 5^b * 7^c$ ,  $3^{a-1} * 5^b * 7^c$  must have already appeared in our list. In fact, all of the following values have already appeared in our list:

- $3^{a-1} * 5^b * 7^c$
- $3^a * 5^{b-1} * 7^c$
- $3^a * 5^b * 7^{c-1}$

Another way to think about this is that every number can be expressed as one of the following:

- $3 * (\text{some previous number in list})$
- $5 * (\text{some previous number in list})$

- $7 * (\text{some previous number in list})$

So, we know  $A_k$  can be expressed as  $(3, 5 \text{ or } 7) * (\text{some value in } \{A_1, \dots, A_{k-1}\})$ . We also know that  $A_k$  is, by definition, the next number in the list. Therefore,  $A_k$  will be the smallest "new" number (a number that it's already in  $\{A_1, \dots, A_{k-1}\}$ ) that can be formed by multiplying each value in the list by 3, 5 or 7.

How would we find  $A_k$ ? Well, we could actually multiply each number in the list by 3, 5, and 7 and find the smallest element that has not yet been added to our list. This solution is  $O(k^2)$ . Not bad, but I think we can do better.

Rather than  $A_k$  trying to "pull" from a previous element in the list (by multiplying all of them by 3, 5 and 7), we can think about each previous value in the list as "pushing" out three subsequent values in the list. That is, each number  $A_1$  will eventually be used later in the list in the following forms:

- $3 * A_1$
- $5 * A_1$
- $7 * A_1$

We can use this thought to plan in advance. Each time we add a number  $A_1$  to the list, we hold on to the values  $3A_1$ ,  $5A_1$ , and  $7A_1$  in some sort of temporary list. To generate  $A_{i+1}$ , we search through this temporary list to find the smallest value.

Our code looks like this:

```

1  public static int removeMin(Queue<Integer> q) {
2      int min = q.peek();
3      for (Integer v : q) {
4          if (min > v) {
5              min = v;
6          }
7      }
8      while (q.contains(min)) {
9          q.remove(min);
10     }
11     return min;
12 }
13
14 public static void addProducts(Queue<Integer> q, int v) {
15     q.add(v * 3);
16     q.add(v * 5);
17     q.add(v * 7);
18 }
19
20 public static int getKthMagicNumber(int k) {
21     if (k < 0) return 0;
22
23     int val = 1;
24     Queue<Integer> q = new LinkedList<Integer>();
```

```

25     addProducts(q, 1);
26     for (int i = 0; i < k; i++) {
27         val = removeMin(q);
28         addProducts(q, val);
29     }
30     return val;
31 }
```

This algorithm is certainly much, much better than our first algorithm, but it's still not quite perfect.

To generate a new element  $A_1$ , we are searching through a linked list where each element looks like one of:

- $3 * \text{previous element}$
- $5 * \text{previous element}$
- $7 * \text{previous element}$

Where is there unnecessary work that we might be able to optimize out?

Let's imagine our list looks like

$$q_6 = \{7A_1, 5A_2, 7A_2, 7A_3, 3A_4, 5A_4, 7A_4, 5A_5, 7A_5\}$$

When we search this list for the min, we check if  $7A_1 < \text{min}$ , and then later we check if  $7A_5 < \text{min}$ . That seems sort of silly, doesn't it? Since we know that  $A_1 < A_5$ , we should only need to check  $7A_1$ .

If we separated the list from the beginning by the constant factors, then we'd only need to check the first of the multiples of 3, 5 and 7. All subsequent elements would be bigger.

That is, our list above would look like:

$$Q36 = \{3A_4\}$$

$$Q56 = \{5A_2, 5A_4, 5A_5\}$$

$$Q76 = \{7A_1, 7A_2, 7A_3, 7A_4, 7A_5\}$$

To get the min, we only need to look at the fronts of each queue:

$$y = \min(Q3.\text{head}(), Q5.\text{head}(), Q7.\text{head}())$$

Once we compute  $y$ , we need to insert  $3y$  into  $Q3$ ,  $5y$  into  $Q5$ , and  $7y$  into  $Q7$ . But, we only want to insert these elements if they aren't already in another list.

Why might, for example,  $3y$  already be somewhere in the holding queues? Well, if  $y$  was pulled from  $Q7$ , then that means that  $y = 7x$ , for some smaller  $x$ . If  $7x$  is the smallest value, we must have already seen  $3x$ . And what did we do when we saw  $3x$ ? We inserted  $7 * 3x$  into  $Q7$ . Note that  $7 * 3x = 3 * 7x = 3y$ .

To put this another way, if we pull an element from  $Q7$ , it will look like  $7 * \text{suffix}$ , and we know we have already handled  $3 * \text{suffix}$  and  $5 * \text{suffix}$ . In handling  $3 * \text{suffix}$ , we inserted  $7 * 3 * \text{suffix}$  into a  $Q7$ . And in handling  $5 * \text{suffix}$ , we know we inserted  $7 * 5 * \text{suffix}$  in  $Q7$ . The only value we haven't seen yet is  $7 * 7 * \text{suffix}$ .

suffix, so we just insert  $7 * 7 * \text{suffix}$  into Q7.

Let's walk through this with an example to make it really clear.

initialize:

```
Q3 = 3
Q5 = 5
Q7 = 7
```

remove min = 3. insert  $3*3$  in Q3,  $5*3$  into Q5,  $7*3$  into Q7.

```
Q3 = 3*3
Q5 = 5, 5*3
Q7 = 7, 7*3
```

remove min = 5.  $3*5$  is a dup, since we already did  $5*3$ . insert  $5*5$  into Q5,  $7*5$  into Q7.

```
Q3 = 3*3
Q5 = 5*3, 5*5
Q7 = 7, 7*3, 7*5.
```

remove min = 7.  $3*7$  and  $5*7$  are dups, since we already did  $7*3$  and  $7*5$ . insert  $7*7$  into Q7.

```
Q3 = 3*3
Q5 = 5*3, 5*5
Q7 = 7*3, 7*5, 7*7
```

remove min =  $3*3 = 9$ . insert  $3*3*3$  in Q3,  $3*3*5$  into Q5,  $3*3*7$  into Q7.

```
Q3 = 3*3*3
Q5 = 5*3, 5*5, 5*3*3
Q7 = 7*3, 7*5, 7*7, 7*3*3
```

remove min =  $5*3 = 15$ .  $3*(5*3)$  is a dup, since we already did  $5*(3*3)$ . insert  $5*5*3$  in Q5,  $7*5*3$  into Q7.

```
Q3 = 3*3*3
Q5 = 5*5, 5*3*3, 5*5*3
Q7 = 7*3, 7*5, 7*7, 7*3*3, 7*5*3
```

remove min =  $7*3 = 21$ .  $3*(7*3)$  and  $5*(7*3)$  are dups, since we already did  $7*(3*3)$  and  $7*(5*3)$ . insert  $7*7*3$  into Q7.

```
Q3 = 3*3*3
Q5 = 5*5, 5*3*3, 5*5*3
Q7 = 7*5, 7*7, 7*3*3, 7*5*3, 7*7*3
```

Our pseudocode for this problem is as follows:

1. Initialize array and queues Q3, Q5, and Q7
2. Insert 1 into array.
3. Insert  $1*3$ ,  $1*5$  and  $1*7$  into Q3, Q5, and Q7 respectively.
4. Let  $x$  be the minimum element in Q3, Q5, and Q7. Append  $x$  to magic.
5. If  $x$  was found in:

$Q3 \rightarrow$  append  $x*3$ ,  $x*5$  and  $x*7$  to Q3, Q5, and Q7. Remove  $x$  from Q3.

$Q5 \rightarrow$  append  $x*5$  and  $x*7$  to Q5 and Q7. Remove  $x$  from Q5.

Q7 -> only append  $x^7$  to Q7. Remove x from Q7.

6. Repeat steps 4 - 6 until we've found k elements.

The code below implements this algorithm.

```
1  public static int getKthMagicNumber(int k) {  
2      if (k < 0) {  
3          return 0;  
4      }  
5      int val = 0;  
6      Queue<Integer> queue3 = new LinkedList<Integer>();  
7      Queue<Integer> queue5 = new LinkedList<Integer>();  
8      Queue<Integer> queue7 = new LinkedList<Integer>();  
9      queue3.add(1);  
10     /* Include 0th through kth iteration */  
11     for (int i = 0; i <= k; i++) {  
12         int v3 = queue3.size() > 0 ? queue3.peek() :  
13                                         Integer.MAX_VALUE;  
14         int v5 = queue5.size() > 0 ? queue5.peek() :  
15                                         Integer.MAX_VALUE;  
16         int v7 = queue7.size() > 0 ? queue7.peek() :  
17                                         Integer.MAX_VALUE;  
18         val = Math.min(v3, Math.min(v5, v7));  
19         if (val == v3) { // enqueue into queue 3, 5 and 7  
20             queue3.remove();  
21             queue3.add(3 * val);  
22             queue5.add(5 * val);  
23             queue7.add(7 * val);  
24         } else if (val == v5) { // enqueue into queue 5 and 7  
25             queue5.remove();  
26             queue5.add(5 * val);  
27         } else if (val == v7) { // enqueue into Q7  
28             queue7.remove();  
29         }  
30         queue7.add(7 * val); // Always enqueue into Q7  
31     }  
32     return val;  
33 }
```

When you get this question, do your best to solve it—even though it's really difficult. You can start with a brute force approach (challenging, but not quite as tricky), and then you can start trying to optimize it. Or, try to find a pattern in the numbers.

Chances are that your interviewer will help you along when you get stuck. Whatever you do, don't give up! Think out loud, wonder out loud, and explain your thought process. Your interviewer will probably jump in to guide you.

Remember, perfection on this problem is not expected. Your performance is evaluated in comparison to other candidates. Everyone struggles on a tricky problem.

# **Object-Oriented Design**

*Concepts and Algorithms: Solutions*

**Chapter 8**

- 8.1 Design the data structures for a generic deck of cards. Explain how you would subclass the data structures to implement blackjack.

pg 105

### SOLUTION

First, we need to recognize that a “generic” deck of cards can mean many things. Generic could mean a standard deck of cards that can play a poker-like game, or it could even stretch to Uno or Baseball cards. It is important to ask your interviewer what she means by generic.

Let’s assume that your interviewer clarifies that the deck is a standard 52-card set, like you might see used in a blackjack or poker game. If so, the design might look like this:

```
1  public enum Suit {
2      Club (0), Diamond (1), Heart (2), Spade (3);
3      private int value;
4      private Suit(int v) { value = v; }
5      public int getValue() { return value; }
6      public static Suit getSuitFromValue(int value) { ... }
7  }
8
9  public class Deck <T extends Card> {
10     private ArrayList<T> cards; // all cards, dealt or not
11     private int dealtIndex = 0; // marks first undealt card
12
13    public void setDeckOfCards(ArrayList<T> deckOfCards) { ... }
14
15    public void shuffle() { ... }
16    public int remainingCards() {
17        return cards.size() - dealtIndex;
18    }
19    public T[] dealHand(int number) { ... }
20    public T dealCard() { ... }
21 }
22
23 public abstract class Card {
24     private boolean available = true;
25
26     /* number or face that's on card - a number 2 through 10, or 11
27      * for Jack, 12 for Queen, 13 for King, or 1 for Ace */
28     protected int faceValue;
29     protected Suit suit;
30
31     public Card(int c, Suit s) {
32         faceValue = c;
33         suit = s;
34     }
35 }
```

```

36     public abstract int value();
37
38     public Suit suit() { return suit; }
39
40     /* Checks if the card is available to be given out to someone */
41     public boolean isAvailable() { return available; }
42     public void markUnavailable() { available = false; }
43
44     public void markAvailable() { available = true; }
45 }
46
47 public class Hand <T extends Card> {
48     protected ArrayList<T> cards = new ArrayList<T>();
49
50     public int score() {
51         int score = 0;
52         for (T card : cards) {
53             score += card.value();
54         }
55         return score;
56     }
57
58     public void addCard(T card) {
59         cards.add(card);
60     }
61 }
```

In the above code, we have implemented Deck with generics but restricted the type of T to Card. We have also implemented Card as an abstract class, since methods like value() don't make much sense without a specific game attached to them. (You could make a compelling argument that they should be implemented anyway, by defaulting to standard poker rules.)

Now, let's say we're building a blackjack game, so we need to know the value of the cards. Face cards are 10 and an ace is 11 (most of the time, but that's the job of the Hand class, not the following class).

```

1  public class BlackJackHand extends Hand<BlackJackCard> {
2     /* There are multiple possible scores for a blackjack hand,
3      * since aces have multiple values. Return the highest possible
4      * score that's under 21, or the lowest score that's over. */
5     public int score() {
6         ArrayList<Integer> scores = possibleScores();
7         int maxUnder = Integer.MIN_VALUE;
8         int minOver = Integer.MAX_VALUE;
9         for (int score : scores) {
10            if (score > 21 && score < minOver) {
11                minOver = score;
12            } else if (score <= 21 && score > maxUnder) {
13                maxUnder = score;
```

```
14         }
15     }
16     return maxUnder == Integer.MIN_VALUE ? minOver : maxUnder;
17 }
18
19 /* return a list of all possible scores this hand could have
20 * (evaluating each ace as both 1 and 11 */
21 private ArrayList<Integer> possibleScores() { ... }
22
23 public boolean busted() { return score() > 21; }
24 public boolean is21() { return score() == 21; }
25 public boolean isBlackJack() { ... }
26 }
27
28 public class BlackJackCard extends Card {
29     public BlackJackCard(int c, Suit s) { super(c, s); }
30     public int value() {
31         if (isAce()) return 1;
32         else if (faceValue >= 11 && faceValue <= 13) return 10;
33         else return faceValue;
34     }
35
36     public int minValue() {
37         if (isAce()) return 1;
38         else return value();
39     }
40
41     public int maxValue() {
42         if (isAce()) return 11;
43         else return value();
44     }
45
46     public boolean isAce() {
47         return faceValue == 1;
48     }
49
50     public boolean isFaceCard() {
51         return faceValue >= 11 && faceValue <= 13;
52     }
53 }
```

This is just one way of handling aces. We could, alternatively, create a class of type Ace that extends BlackJackCard.

An executable, fully automated version of blackjack is provided in the downloadable code attachment.

- 8.2** Imagine you have a call center with three levels of employees: respondent, manager, and director. An incoming telephone call must be first allocated to a respondent who is free. If the respondent can't handle the call, he or she must escalate the call to a manager. If the manager is not free or not able to handle it, then the call should be escalated to a director. Design the classes and data structures for this problem. Implement a method `dispatchCall()` which assigns a call to the first available employee.

pg 106

## SOLUTION

All three ranks of employees have different work to be done, so those specific functions are profile specific. We should keep these things within their respective class.

There are a few things which are common to them, like address, name, job title, and age. These things can be kept in one class and can be extended or inherited by others.

Finally, there should be one `CallHandler` class which would route the calls to the correct person.

Note that on any object-oriented design question, there are many ways to design the objects. Discuss the trade-offs of different solutions with your interviewer. You should usually design for long-term code flexibility and maintenance.

We'll go through each of the classes below in detail.

`CallHandler` is implemented as a singleton class. It represents the body of the program, and all calls are funneled first through it.

```
1 public class CallHandler {  
2     private static CallHandler instance;  
3  
4     /* 3 levels of employees: respondents, managers, directors. */  
5     private final int LEVELS = 3;  
6  
7     /* Initialize 10 respondents, 4 managers, and 2 directors. */  
8     private final int NUM_RESPONDENTS = 10;  
9     private final int NUM_MANAGERS = 4;  
10    private final int NUM_DIRECTORS = 2;  
11  
12    /* List of employees, by level.  
13     * employeeLevels[0] = respondents  
14     * employeeLevels[1] = managers  
15     * employeeLevels[2] = directors  
16     */  
17    List<List<Employee>> employeeLevels;  
18  
19    /* queues for each call's rank */  
20    List<List<Call>> callQueues;  
21}
```

```

22     protected CallHandler() { ... }
23
24     /* Get instance of singleton class. */
25     public static CallHandler getInstance() {
26         if (instance == null) instance = new CallHandler();
27         return instance;
28     }
29
30     /* Gets the first available employee who can handle this call.
*/
31     public Employee getHandlerForCall(Call call) { ... }
32
33     /* Routes the call to an available employee, or saves in a queue
 * if no employee available. */
34     public void dispatchCall(Caller caller) {
35         Call call = new Call(caller);
36         dispatchCall(call);
37     }
38
39     /* Routes the call to an available employee, or saves in a queue
 * if no employee available. */
40     public void dispatchCall(Call call) {
41         /* Try to route the call to an employee with minimal rank. */
42         Employee emp = getHandlerForCall(call);
43         if (emp != null) {
44             emp.receiveCall(call);
45             call.setHandler(emp);
46         } else {
47             /* Place the call into corresponding call queue according
 * to its rank. */
48             call.reply("Please wait for free employee to reply");
49             callQueues.get(call.getRank().getValue()).add(call);
50         }
51     }
52
53     /* An employee got free. Look for a waiting call that emp. can
 * serve. Return true if we assigned a call, false otherwise. */
54     public boolean assignCall(Employee emp) { ... }
55 }
56
57 */
58
59 }
```

Call represents a call from a user. A call has a minimum rank and is assigned to the first employee who can handle it.

```

1  public class Call {
2      /* Minimal rank of employee who can handle this call. */
3      private Rank rank;
4
5      /* Person who is calling. */
6      private Caller caller;
7
8      /* Employee who is handling call. */
9 }
```

```

9  private Employee handler;
10
11 public Call(Call c) {
12     rank = Rank.Responder;
13     caller = c;
14 }
15
16 /* Set employee who is handling call. */
17 public void setHandler(Employee e) { handler = e; }
18
19 public void reply(String message) { ... }
20 public Rank getRank() { return rank; }
21 public void setRank(Rank r) { rank = r; }
22 public Rank incrementRank() { ... }
23 public void disconnect() { ... }
24 }
```

Employee is a super class for the Director, Manager, and Respondent classes. It is implemented as an abstract class since there should be no reason to instantiate an Employee type directly.

```

1 abstract class Employee {
2     private Call currentCall = null;
3     protected Rank rank;
4
5     public Employee() { } .
6
7     /* Start the conversation */
8     public void receiveCall(Call call) { ... }
9
10    /* the issue is resolved, finish the call */
11    public void callCompleted() { ... }
12
13    /* The issue has not been resolved. Escalate the call, and
14       * assign a new call to the employee. */
15    public void escalateAndReassign() { ... }
16 }
17
18    /* Assign a new call to an employee, if the employee is free. */
19    public boolean assignNewCall() { ... }
20
21    /* Returns whether or not the employee is free. */
22    public boolean isFree() { return currentCall == null; }
23
24    public Rank getRank() { return rank; }
25 }
26
```

The Respondent, Director, and Manager classes are now just simple extensions of the Employee class.

```
1 class Director extends Employee {
```

```
2     public Director() {
3         rank = Rank.Director;
4     }
5 }
6
7 class Manager extends Employee {
8     public Manager() {
9         rank = Rank.Manager;
10    }
11 }
12
13 class Respondent extends Employee {
14     public Respondent() {
15         rank = Rank.Responder;
16    }
17 }
```

This is just one way of designing this problem. Note that there are many other ways that are equally good.

This may seem like an awful lot of code to write in an interview, and it is. We've been much more thorough here than you would need. In a real interview, you would likely be much lighter on some of the details until you have time to fill them in.

### 8.3 Design a musical jukebox using object-oriented principles.

pg 106

#### SOLUTION

In any object-oriented design question, you first want to start off with asking your interviewer some questions to clarify design constraints. Is this jukebox playing CDs? Records? MP3s? Is it a simulation on a computer, or is it supposed to represent a physical jukebox? Does it take money, or is it free? And if it takes money, which currency? And does it deliver change?

Unfortunately, we don't have an interviewer here that we can have this dialogue with. Instead, we'll make some assumptions. We'll assume that the jukebox is a computer simulation that closely mirrors physical jukeboxes, and we'll assume that it's free.

Now that we have that out of the way, we'll outline the basic system components:

- Jukebox
- CD
- Song
- Artist
- Playlist

- Display (displays details on the screen)

Now, let's break this down further and think about the possible actions.

- Playlist creation (includes add, delete, and shuffle)
- CD selector
- Song selector
- Queuing up a song
- Get next song from playlist

A user also can be introduced:

- Adding
- Deleting
- Credit information

Each of the main system components translates roughly to an object, and each action translates to a method. Let's walk through one potential design.

The Jukebox class represents the body of the problem. Many of the interactions between the components of the system, or between the system and the user, are channeled through here.

```

1  public class Jukebox {
2      private CDPlayer cdPlayer;
3      private User user;
4      private Set<CD> cdCollection;
5      private SongSelector ts;
6
7      public Jukebox(CDPlayer cdPlayer, User user,
8                      Set<CD> cdCollection, SongSelector ts) {
9          ...
10     }
11
12     public Song getCurrentSong() {
13         return ts.getCurrentSong();
14     }
15
16     public void setUser(User u) {
17         this.user = u;
18     }
19 }
```

Like a real CD player, the CDPlayer class supports storing just one CD at a time. The CDs that are not in play are stored in the jukebox.

```

1  public class CDPlayer {
2      private Playlist p;
3      private CD c;
4
```

```

5   /* Constructors. */
6   public CDPlayer(CD c, Playlist p) { ... }
7   public CDPlayer(Playlist p) { this.p = p; }
8   public CDPlayer(CD c) { this.c = c; }
9
10  /* Play song */
11  public void playSong(Song s) { ... }
12
13  /* Getters and setters */
14  public Playlist getPlaylist() { return p; }
15  public void setPlaylist(Playlist p) { this.p = p; }
16
17  public CD getCD() { return c; }
18  public void setCD(CD c) { this.c = c; }
19 }

```

The Playlist manages the current and next songs to play. It is essentially a wrapper class for a queue and offers some additional methods for convenience.

```

1  public class Playlist {
2      private Song song;
3      private Queue<Song> queue;
4      public Playlist(Song song, Queue<Song> queue) {
5          ...
6      }
7      public Song getNextSToPlay() {
8          return queue.peek();
9      }
10     public void queueUpSong(Song s) {
11         queue.add(s);
12     }
13 }

```

The classes for CD, Song, and User are all fairly straightforward. They consist mainly of member variables and getters and setters.

```

1  public class CD {
2      /* data for id, artist, songs, etc */
3  }
4
5  public class Song {
6      /* data for id, CD (could be null), title, length, etc */
7  }
8
9  public class User {
10     private String name;
11     public String getName() { return name; }
12     public void setName(String name) { this.name = name; }
13     public long getID() { return ID; }
14     public void setID(long id) { ID = id; }
15     private long ID;
16     public User(String name, long id) { ... }

```

```

17     public User getUser() { return this; }
18     public static User addUser(String name, long id) { ... }
19 }
```

This is by no means the only “correct” implementation. The interviewer’s responses to initial questions, as well as other constraints, will shape the design of the jukebox classes.

#### 8.4 Design a parking lot using object-oriented principles.

pg 106

#### SOLUTION

The wording of this question is vague, just as it would be in an actual interview. This requires you to have a conversation with your interviewer about what types of vehicles it can support, whether the parking lot has multiple levels, and so on.

For our purposes right now, we’ll make the following assumptions. We made these specific assumptions to add a bit of complexity to the problem without adding too much. If you made different assumptions, that’s totally fine.

- The parking lot has multiple levels. Each level has multiple rows of spots.
- The parking lot can park motorcycles, cars, and buses.
- The parking lot has motorcycle spots, compact spots, and large spots.
- A motorcycle can park in any spot.
- A car can park in either a single compact spot or a single large spot.
- A bus can park in five large spots that are consecutive and within the same row. It cannot park in small spots.

In the below implementation, we have created an abstract class `Vehicle`, from which `Car`, `Bus`, and `Motorcycle` inherit. To handle the different parking spot sizes, we have just one class `ParkingSpot` which has a member variable indicating the size.

```

1  public enum VehicleSize { Motorcycle, Compact, Large }
2
3  public abstract class Vehicle {
4      protected ArrayList<ParkingSpot> parkingSpots =
5          new ArrayList<ParkingSpot>();
6      protected String licensePlate;
7      protected int spotsNeeded;
8      protected VehicleSize size;
9
10     public int getSpotsNeeded() { return spotsNeeded; }
11     public VehicleSize getSize() { return size; }
12
13     /* Park vehicle in this spot (among others, potentially) */
```

```
14     public void parkInSpot(ParkingSpot s) { parkingSpots.add(s); }
15
16     /* Remove car from spot, and notify spot that it's gone */
17     public void clearSpots() { ... }
18
19     /* Checks if the spot is big enough for the vehicle (and is
20      * available). This compares the SIZE only. It does not check if
21      * it
22      * has enough spots. */
23     public abstract boolean canFitInSpot(ParkingSpot spot);
24 }
25
26 public class Bus extends Vehicle {
27     public Bus() {
28         spotsNeeded = 5;
29         size = VehicleSize.Large;
30     }
31
32     /* Checks if the spot is a Large. Doesn't check num of spots */
33     public boolean canFitInSpot(ParkingSpot spot) { ... }
34 }
35
36 public class Car extends Vehicle {
37     public Car() {
38         spotsNeeded = 1;
39         size = VehicleSize.Compact;
40     }
41
42     /* Checks if the spot is a Compact or a Large. */
43     public boolean canFitInSpot(ParkingSpot spot) { ... }
44 }
45
46 public class Motorcycle extends Vehicle {
47     public Motorcycle() {
48         spotsNeeded = 1;
49         size = VehicleSize.Motorcycle;
50     }
51
52     public boolean canFitInSpot(ParkingSpot spot) { ... }
53 }
```

The `ParkingLot` class is essentially a wrapper class for an array of `Levels`. By implementing it this way, we are able to separate out logic that deals with actually finding free spots and parking cars out from the broader actions of the `ParkingLot`. If we didn't do it this way, we would need to holding parking spots in some sort of double array (or hash table which maps from a level number to the list of spots). It's cleaner to just separate `ParkingLot` from `Level`.

```
1  public class ParkingLot {
2      private Level[] levels;
```

```

3  private final int NUM_LEVELS = 5;
4
5  public ParkingLot() { ... }
6
7  /* Park the vehicle in a spot (or multiple spots).
8   * Return false if failed. */
9  public boolean parkVehicle(Vehicle vehicle) { ... }
10 }
11
12 /* Represents a level in a parking garage */
13 public class Level {
14  private int floor;
15  private ParkingSpot[] spots;
16  private int availableSpots = 0; // number of free spots
17  private static final int SPOTS_PER_ROW = 10;
18
19  public Level(int flr, int numberSpots) { ... }
20
21  public int availableSpots() { return availableSpots; }
22
23  /* Find a place to park this vehicle. Return false if failed. */
24  public boolean parkVehicle(Vehicle vehicle) { ... }
25
26  /* Park a vehicle starting at the spot spotNumber, and
27   * continuing until vehicle.spotsNeeded. */
28  private boolean parkStartingAtSpot(int num, Vehicle v) { ... }
29
30  /* Find a spot to park this vehicle. Return index of spot, or -1
31   * on failure. */
32  private int findAvailableSpots(Vehicle vehicle) { ... }
33
34  /* When a car was removed from the spot, increment
35   * availableSpots */
36  public void spotFreed() { availableSpots++; }
37 }

```

The `ParkingSpot` is implemented by having just a variable which represents the size of the spot. We could have implemented this by having classes for `LargeSpot`, `CompactSpot`, and `MotorcycleSpot` which inherit from `ParkingSpot`, but this is probably overkill. The spots probably do not have different behaviors, other than their sizes.

```

1  public class ParkingSpot {
2    private Vehicle vehicle;
3    private VehicleSize spotSize;
4    private int row;
5    private int spotNumber;
6    private Level level;
7
8    public ParkingSpot(Level lvl, int r, int n, VehicleSize s) {...}

```

```
9     public boolean isAvailable() { return vehicle == null; }
10    /* Check if the spot is big enough and is available */
11    public boolean canFitVehicle(Vehicle vehicle) { ... }
12
13    /* Park vehicle in this spot. */
14    public boolean park(Vehicle v) { ... }
15
16    public int getRow() { return row; }
17    public int getSpotNumber() { return spotNumber; }
18
19    /* Remove vehicle from spot, and notify level that a new spot is
20     * available */
21    public void removeVehicle() { ... }
22
23 }
```

A full implementation of this code, including executable test code, is provided in the downloadable code attachment.

### 8.5 Design the data structures for an online book reader system.

pg 106

#### SOLUTION

Since the problem doesn't describe much about the functionality, let's assume we want to design a basic online reading system which provides the following functionality:

- User membership creation and extension.
- Searching the database of books.
- Reading a book.
- Only one active user at a time
- Only one active book by this user.

To implement these operations we may require many other functions, like get, set, update, and so on. The objects required would likely include User, Book, and Library.

The class `OnlineReaderSystem` represents the body of our program. We could implement the class such that it stores information about all the books, deals with user management, and refreshes the display, but that would make this class rather hefty. Instead, we've chosen to tear off these components into Library, UserManager, and Display classes.

```
1  public class OnlineReaderSystem {
2      private Library library;
3      private UserManager userManager;
4      private Display display;
```

```

5
6     private Book activeBook;
7     private User activeUser;
8
9     public OnlineReaderSystem() {
10         userManager = new UserManager();
11         library = new Library();
12         display = new Display();
13     }
14
15     public Library getLibrary() { return library; }
16     public UserManager getUserManager() { return userManager; }
17     public Display getDisplay() { return display; }
18
19     public Book getActiveBook() { return activeBook; }
20     public void setActiveBook(Book book) {
21         activeBook = book;
22         display.displayBook(book);
23     }
24
25     public User getActiveUser() { return activeUser; }
26     public void setActiveUser(User user) {
27         activeUser = user;
28         display.displayUser(user);
29     }
30 }
```

We then implement separate classes to handle the user manager, the library, and the display components.

```

1  public class Library {
2     private Hashtable<Integer, Book> books;
3
4     public Book addBook(int id, String details) {
5         if (books.containsKey(id)) {
6             return null;
7         }
8         Book book = new Book(id, details);
9         books.put(id, book);
10        return book;
11    }
12
13    public boolean remove(Book b) { return remove(b.getID()); }
14    public boolean remove(int id) {
15        if (!books.containsKey(id)) {
16            return false;
17        }
18        books.remove(id);
19        return true;
20    }
21 }
```

```
22     public Book find(int id) {
23         return books.get(id);
24     }
25 }
26
27 public class UserManager {
28     private Hashtable<Integer, User> users;
29
30     public User addUser(int id, String details, int accountType) {
31         if (users.containsKey(id)) {
32             return null;
33         }
34         User user = new User(id, details, accountType);
35         users.put(id, user);
36         return user;
37     }
38
39     public boolean remove(User u) {
40         return remove(u.getID());
41     }
42
43     public boolean remove(int id) {
44         if (!users.containsKey(id)) {
45             return false;
46         }
47         users.remove(id);
48         return true;
49     }
50
51     public User find(int id) {
52         return users.get(id);
53     }
54 }
55
56 public class Display {
57     private Book activeBook;
58     private User activeUser;
59     private int pageNumber = 0;
60
61     public void displayUser(User user) {
62         activeUser = user;
63         refreshUsername();
64     }
65
66     public void displayBook(Book book) {
67         pageNumber = 0;
68         activeBook = book;
69
70         refreshTitle();
71         refreshDetails();
```

```

72     refreshPage();
73 }
74
75 public void turnPageForward() {
76     pageNumber++;
77     refreshPage();
78 }
79
80 public void turnPageBackward() {
81     pageNumber--;
82     refreshPage();
83 }
84
85 public void refreshUsername() { /* updates username display */ }
86 public void refreshTitle() { /* updates title display */ }
87 public void refreshDetails() { /* updates details display */ }
88 public void refreshPage() { /* updated page display */ }
89 }
```

The classes for User and Book simply hold data and provide little true functionality.

```

1  public class Book {
2      private int bookId;
3      private String details;
4
5      public Book(int id, String det) {
6          bookId = id;
7          details = det;
8      }
9
10     public int getID() { return bookId; }
11     public void setID(int id) { bookId = id; }
12     public String getDetails() { return details; }
13     public void setDetails(String d) { details = d; }
14 }
15
16 public class User {
17     private int userId;
18     private String details;
19     private int accountType;
20
21     public void renewMembership() { }
22
23     public User(int id, String details, int accountType) {
24         userId = id;
25         this.details = details;
26         this.accountType = accountType;
27     }
28
29     /* Getters and setters */
30     public int getID() { return userId; }
```

```
31     public void setID(int id) { userId = id; }
32     public String getDetails() {
33         return details;
34     }
35
36     public void setDetails(String details) {
37         this.details = details;
38     }
39     public int getAccountType() { return accountType; }
40     public void setAccountType(int t) { accountType = t; }
41 }
```

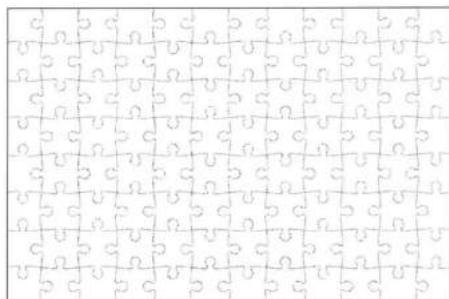
The decision to tear off user management, library, and display into their own classes, when this functionality could have been in the general `OnlineReaderSystem` class, is an interesting one. On a very small system, making this decision could make the system overly complex. However, as the system grows, and more and more functionality gets added to `OnlineReaderSystem`, breaking off such components prevents this main class from getting overwhelmingly lengthy.

- 8.6** *Implement a jigsaw puzzle. Design the data structures and explain an algorithm to solve the puzzle. You can assume that you have a `fitsWith` method which, when passed two puzzle pieces, returns true if the two pieces belong together.*

pg 106

### SOLUTION

We will assume that we have a traditional, simple jigsaw puzzle. The puzzle is grid-like, with rows and columns. Each piece is located in a single row and column and has four edges. Each edge comes in one of three types: inner, outer, and flat. A corner piece, for example, will have two flat edges and two other edges, which could be inner or outer.



As we solve the jigsaw puzzle (manually or algorithmically), we'll need to store the position of each piece. We could think about the position as absolute or relative:

- *Absolute Position:* "This piece is located at position (12, 23)." Absolute position would belong to the `Piece` class itself and would include an orientation as well.

- *Relative Position:* “I don’t know where this piece is actually located, but I know that it is next to this other piece.” The relative position would belong to the Edge class.

For our solution, we will use only the relative position, by adjoining edges to neighboring edges.

A potential object-oriented design looks like the following:

```

1  class Edge {
2      enum Type { inner, outer, flat }
3      Piece parent;
4      Type type;
5      int index; // Index into Piece.edges
6      Edge attached_to; // Relative position.
7
8      /* See Algorithm section. Returns true if the two pieces
9       * should be attached to each other. */
10     boolean fitsWith(Edge edge) { ... };
11 }
12
13 class Piece {
14     Edge[] edges;
15     boolean isCorner() { ... }
16 }
17
18 class Puzzle {
19     Piece[] pieces; /* Remaining pieces left to put away. */
20     Piece[][] solution;
21
22     /* See algorithm section. */
23     Edge[] inners, outers, flats;
24     Piece[] corners;
25
26     /* See Algorithm section. */
27     void sort() { ... }
28     void solve() { ... }
29 }
```

### Algorithm to Solve the Puzzle

We will sketch this algorithm using a mix of pseudocode and real code.

Just as a kid might in solving a puzzle, we’ll start with the easiest pieces first: the corners and edges. We can easily search through all the pieces to find just the edges. While we’re at it though, it probably makes sense to group all the pieces by their edge types.

```

1  void sort() {
2      for each Piece p in pieces {
3          if (p has two flat edges) then add p to corners
```

```
4     for each edge in p.edges {
5         if edge is inner then add to inners
6         if edge is outer then add toouters
7     }
8 }
9 }
```

We now have a quicker way to zero in on potential matches for any given edge. We then go through the puzzle, line by line, to match pieces.

The `solve` method, implemented below, operates by picking an arbitrary corner to start with. It then finds an open edge on the corner and tries to match it to an open piece. When it finds a match, it does the following:

1. Attaches the edge.
2. Removes the edge from the list of open edges.
3. Finds the next open edge.

The next open edge is defined to be the one directly opposite the current edge, if it is available. If it is not available, then the next edge can be any other edge. This will cause the puzzle to be solved in a spiral-like fashion, from the outside to the inside.

The spiral comes from the fact that the algorithm always moves in a straight line, whenever possible. When we reach the end of the first edge, the algorithm moves to the only available edge on that corner piece—a 90-degree turn. It continues to take 90-degree turns at the end of each side until the entire outer edge of the puzzle is completed. When that last edge piece is in place, that piece only has one exposed edge remaining, which is again a 90-degree turn. The algorithm repeats itself for subsequent rings around the puzzle, until finally all the pieces are in place.

This algorithm is implemented below with Java-like pseudocode.

```
1 public void solve() {
2     /* Pick any corner to start with */
3     Edge currentEdge = getExposedEdge(corner[0]);
4
5     /* Loop will iterate in a spiral like fashion until the puzzle
6      * is full. */
7     while (currentEdge != null) {
8         /* Match with opposite edges. Inners withouters, etc. */
9         Edge[] opposites = currentEdge.type == inner ?
10             outers : inners;
11         for each Edge fittingEdge in opposites {
12             if (currentEdge.fitsWith(fittingEdge)) {
13                 attachEdges(currentEdge, fittingEdge); //attach edge
14                 removeFromList(currentEdge);
15                 removeFromList(fittingEdge);
16
17                 /* get next edge */
18                 currentEdge = nextExposedEdge(fittingEdge);
19             }
20         }
21     }
22 }
```

```
19         break; // Break out of inner loop. Continue in outer.
20     }
21 }
22 }
23 }
24
25 public void removeFromList(Edge edge) {
26     if (edge.type == flat) return;
27     Edge[] array = currentEdge.type == inner ? inners :outers;
28     array.remove(edge);
29 }
30
31 /* Return the opposite edge if possible. Else, return any exposed
32 * edge. */
33 public Edge nextExposedEdge(Edge edge) {
34     int next_index = (edge.index + 2) % 4; // Opposite edge
35     Edge next_edge = edge.parent.edges[next_index];
36     if isExposed(next_edge) {
37         return next_edge;
38     }
39     return getExposedEdge(edge.parent);
40 }
41
42 public Edge attachEdges(Edge e1, Edge e2) {
43     e1.attached_to = e2;
44     e2.attached_to = e1;
45 }
46
47 public Edge isExposed(Edge e1) {
48     return edge.type != flat && edge.attached_to == null;
49 }
50
51 public Edge getExposedEdge(Piece p) {
52     for each Edge edge in p.edges {
53         if (isExposed(edge)) {
54             return edge;
55         }
56     }
57     return null;
58 }
```

For simplicity, we've represented `inners` and `outers` as an `Edge` array. This is actually not a great choice, since we need to add and remove elements from it frequently. If we were writing real code, we would probably want to implement these variables as linked lists.

Writing the full code for this problem in an interview would be far, far too much work. More likely, you would be asked to just sketch out the code.

- 8.7 Explain how you would design a chat server. In particular, provide details about the various backend components, classes, and methods. What would be the hardest problems to solve?

pg 106

### SOLUTION

Designing a chat server is a huge project, and it is certainly far beyond the scope of what could be completed in an interview. After all, teams of many people spend months or years creating a chat server. Part of your job, as a candidate, is to focus on an aspect of the problem that is reasonably broad, but focused enough that you could accomplish it during an interview. It need not match real life exactly, but it should be a fair representation of an actual implementation.

For our purposes, we'll focus on the core user management and conversation aspects: adding a user, creating a conversation, updating one's status, and so on. In the interest of time and space, we will not go into the networking aspects of the problem, or how the data actually gets pushed out to the clients.

We will assume that "friending" is mutual; I am only your contact if you are mine. Our chat system will support both group chat and one-on-one (private) chats. We will not worry about voice chat, video chat, or file transfer.

#### What specific actions does it need to support?

This is also something to discuss with your interviewer, but here are some ideas:

- Signing online and offline.
- Add requests (sending, accepting, and rejecting).
- Updating a status message.
- Creating private and group chats.
- Adding new messages to private and group chats.

This is just a partial list. If you have more time, you can add more actions.

#### What can we learn about these requirements?

We must have a concept of users, add request status, online status, and messages.

#### What are the core components of the system?

The system would likely consist of a database, a set of clients, and a set of servers. We won't include these parts in our object-oriented design, but we can discuss the overall view of the system.

The database will be used for more permanent storage, such as the user list or chat

archives. A SQL database is a good bet, or, if we need more scalability, we could potentially use BigTable or a similar system.

For communication between the client and servers, using XML will work well. Although it's not the most compressed format (and you should point this out to your interviewer), it's nice because it's easy for both computers and humans to read. Using XML will make your debugging efforts easier—and that matters a lot.

The server will consist of a set of machines. Data will be divided up across machines, requiring us to potentially hop from machine to machine. When possible, we will try to replicate some data across machines to minimize the lookups. One major design constraint here is to prevent having a single point of failure. For instance, if one machine controlled all the user sign-ins, then we'd cut off millions of users potentially if a single machine lost network connectivity.

### What are the key objects and methods?

The key objects of the system will be a concept of users, conversations, and status messages. We've implemented a `UserManagement` class. If we were looking more at the networking aspects of the problem, or a different component, we might have instead dived into those objects.

```

1  /* UserManager serves as a central place for core user actions. */
2  public class UserManager {
3      private static UserManager instance;
4      /* maps from a user id to a user */
5      private HashMap<Integer, User> usersById;
6
7      /* maps from an account name to a user */
8      private HashMap<String, User> usersByAccountName;
9
10     /* maps from the user id to an online user */
11     private HashMap<Integer, User> onlineUsers;
12
13     public static UserManager getInstance() {
14         if (instance == null) instance = new UserManager();
15         return instance;
16     }
17
18     public void addUser(User fromUser, String toAccountName) { ... }
19     public void approveAddRequest(AddRequest req) { ... }
20     public void rejectAddRequest(AddRequest req) { ... }
21     public void userSignedOn(String accountName) { ... }
22     public void userSignedOff(String accountName) { ... }
23 }
```

The method `receivedAddRequest`, in the `User` class, notifies User B that User A has requested to add him. User B approves or rejects the request (via `UserManager.approveAddRequest` or `rejectAddRequest`), and the `UserManager` takes care of

adding the users to each other's contact lists.

The method `sentAddRequest` in the `User` class is called by `UserManager` to add an `AddRequest` to User A's list of requests. So the flow is:

1. User A clicks "add user" on the client, and it gets sent to the server.
2. User A calls `requestAddUser(User B)`.
3. This method calls `UserManager.addUser`.
4. `UserManager` calls both `User A.sentAddRequest` and `User B.receivedAddRequest`.

Again, this is just *one* way of designing these interactions. It is not the only way, or even the only "good" way.

```
1  public class User {  
2      private int id;  
3      private UserStatus status = null;  
4  
5      /* maps from the other participant's user id to the chat */  
6      private HashMap<Integer, PrivateChat> privateChats;  
7  
8      /* maps from the group chat id to the group chat */  
9      private ArrayList<GroupChat> groupChats;  
10  
11     /* maps from the other person's user id to the add request */  
12     private HashMap<Integer, AddRequest> receivedAddRequests;  
13  
14     /* maps from the other person's user id to the add request */  
15     private HashMap<Integer, AddRequest> sentAddRequests;  
16  
17     /* maps from the user id to the add request */  
18     private HashMap<Integer, User> contacts;  
19  
20     private String accountName;  
21     private String fullName;  
22  
23     public User(int id, String accountName, String fullName) { ... }  
24     public boolean sendMessageToUser(User to, String content){ ... }  
25     public boolean sendMessageToGroupChat(int id, String cnt){...}  
26     public void setStatus(UserStatus status) { ... }  
27     public UserStatus getStatus() { ... }  
28     public boolean addContact(User user) { ... }  
29     public void receivedAddRequest(AddRequest req) { ... }  
30     public void sentAddRequest(AddRequest req) { ... }  
31     public void removeAddRequest(AddRequest req) { ... }  
32     public void requestAddUser(String accountName) { ... }  
33     public void addConversation(PrivateChat conversation) { ... }  
34     public void addConversation(GroupChat conversation) { ... }  
35     public int getId() { ... }
```

```

36     public String getAccountName() { ... }
37     public String getFullName() { ... }
38 }
```

The Conversation class is implemented as an abstract class, since all Conversations must be either a GroupChat or a PrivateChat, and since these two classes each have their own functionality.

```

1  public abstract class Conversation {
2      protected ArrayList<User> participants;
3      protected int id;
4      protected ArrayList<Message> messages;
5
6      public ArrayList<Message> getMessages() { ... }
7      public boolean addMessage(Message m) { ... }
8      public int getId() { ... }
9  }
10
11 public class GroupChat extends Conversation {
12     public void removeParticipant(User user) { ... }
13     public void addParticipant(User user) { ... }
14 }
15
16 public class PrivateChat extends Conversation {
17     public PrivateChat(User user1, User user2) { ... }
18     public User getOtherParticipant(User primary) { ... }
19 }
20
21 public class Message {
22     private String content;
23     private Date date;
24     public Message(String content, Date date) { ... }
25     public String getContent() { ... }
26     public Date getDate() { ... }
27 }
```

AddRequest and UserStatus are simple classes with little functionality. Their main purpose is to group data that other classes will act upon.

```

1  public class AddRequest {
2      private User fromUser;
3      private User toUser;
4      private Date date;
5      RequestStatus status;
6
7      public AddRequest(User from, User to, Date date) { ... }
8      public RequestStatus getStatus() { ... }
9      public User getFromUser() { ... }
10     public User getToUser() { ... }
11     public Date getDate() { ... }
12 }
13
```

```
14 public class UserStatus {  
15     private String message;  
16     private UserStatusType type;  
17     public UserStatus(UserStatusType type, String message) { ... }  
18     public UserStatusType getStatusType() { ... }  
19     public String getMessage() { ... }  
20 }  
21  
22 public enum UserStatusType {  
23     Offline, Away, Idle, Available, Busy  
24 }  
25  
26 public enum RequestStatus {  
27     Unread, Read, Accepted, Rejected  
28 }
```

The downloadable code attachment provides a more detailed look at these methods, including implementations for the methods shown above.

### **What problems would be the hardest to solve (or the most interesting)?**

The following questions may be interesting to discuss with your interviewer further.

*Q1: How do we know if someone is online—I mean, really, really know?*

While we would like users to tell us when they sign off, we can't know for sure. A user's connection might have died, for example. To make sure that we know when a user has signed off, we might try regularly pinging the client to make sure it's still there.

*Q2: How do we deal with conflicting information?*

We have some information stored in the computer's memory and some in the database. What happens if they get out of sync? Which one is "right"?

*Q3: How do we make our server scale?*

While we designed our chat server without worrying—too much—about scalability, in real life this would be a concern. We'd need to split our data across many servers, which would increase our concern about out-of-sync data.

*Q4: How do we prevent denial of service attacks?*

Clients can push data to us—what if they try to DOS (denial of service) us? How do we prevent that?

- 8.8** *Othello is played as follows: Each Othello piece is white on one side and black on the other. When a piece is surrounded by its opponents on both the left and right sides, or both the top and bottom, it is said to be captured and its color is flipped. On your turn, you must capture at least one of your opponent's pieces. The game ends when either user has no more valid moves. The win is assigned to the person with the most pieces. Implement the object-oriented design for Othello.*

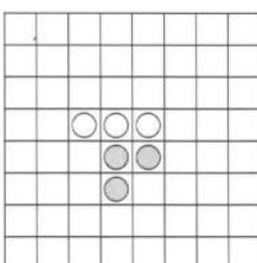
pg 106

## SOLUTION

Let's start with an example. Suppose we have the following moves in an Othello game:

1. Initialize the board with two black and two white pieces in the center. The black pieces are placed at the upper left hand and lower right hand corners.
2. Play a black piece at (row 6, column 4). This flips the piece at (row 5, column 4) from white to black.
3. Play a white piece at (row 4, column 3). This flips the piece at (row 4, column 4) from black to white.

This sequence of moves leads to the board below.



The core objects in Othello are probably the game, the board, the pieces (black or white), and the players. How do we represent these with elegant object-oriented design?

### Should BlackPiece and WhitePiece be classes?

At first, we might think we want to have a `BlackPiece` class and a `WhitePiece` class, which inherit from an abstract `Piece`. However, this is probably not a great idea. Each piece may flip back and forth between colors frequently, so continuously destroying and creating what is really the same object is probably not wise. It may be better to just have a `Piece` class, with a flag in it representing the current color.

### Do we need separate Board and Game classes?

Strictly speaking, it may not be necessary to have both a `Game` object and a `Board` object. Keeping the objects separate allows us to have a logical separation between the

board (which contains just logic involving placing pieces) and the game (which involves times, game flow, etc.). However, the drawback is that we are adding extra layers to our program. A function may call out to a method in Game, only to have it immediately call Board. We have made the choice below to keep Game and Board separate, but you should discuss this with your interviewer.

### Who keeps score?

We know we should probably have some sort of score keeping for the number of black and white pieces. But who should maintain this information? One could make a strong argument for either Game or Board maintaining this information, and possibly even for Piece (in static methods). We have implemented this with Board holding this information, since it can be logically grouped with the board. It is updated by Piece or Board calling colorChanged and colorAdded methods within Board.

### Should Game be a Singleton class?

Implementing Game as a singleton class has the advantage of making it easy for anyone to call a method within Game, without having to pass around references to the Game object.

Making Game a singleton though means that it can only be instantiated once. Can we make this assumption? You should discuss this with your interviewer.

One possible design for Othello is below.

```
1  public enum Direction {  
2      left, right, up, down  
3  }  
4  
5  public enum Color {  
6      White, Black  
7  }  
8  
9  public class Game {  
10     private Player[] players;  
11     private static Game instance;  
12     private Board board;  
13     private final int ROWS = 10;  
14     private final int COLUMNS = 10;  
15  
16     private Game() {  
17         board = new Board(ROWS, COLUMNS);  
18         players = new Player[2];  
19         players[0] = new Player(Color.Black);  
20         players[1] = new Player(Color.White);  
21     }  
22  
23     public static Game getInstance() {
```

```

24     if (instance == null) instance = new Game();
25     return instance;
26   }
27
28   public Board getBoard() {
29     return board;
30   }
31 }
```

The Board class manages the actual pieces themselves. It does not handle much of the game play, leaving that up to the Game class.

```

1  public class Board {
2    private int blackCount = 0;
3    private int whiteCount = 0;
4    private Piece[][][] board;
5
6    public Board(int rows, int columns) {
7      board = new Piece[rows][columns];
8    }
9
10   public void initialize() {
11     /* initialize center black and white pieces */
12   }
13
14   /* Attempt to place a piece of color color at (row, column).
15    * Return true if we were successful. */
16   public boolean placeColor(int row, int column, Color color) {
17     ...
18   }
19
20   /* flips pieces starting at (row, column) and proceeding in
21    * direction d. */
22   private int flipSection(int row, int column, Color color,
23                         Direction d) { ... }
24
25   public int getScoreForColor(Color c) {
26     if (c == Color.Black) return blackCount;
27     else return whiteCount;
28   }
29
30   /* Update board with additional newPieces pieces of color
31    * newColor. Decrease score of opposite color. */
32   public void updateScore(Color newColor, int newPieces) { ... }
33 }
```

As described earlier, we implement the black and white pieces with the Piece class, which has a simple Color variable representing whether it is a black or white piece.

```

1  public class Piece {
2    private Color color;
3    public Piece(Color c) { color = c; }
```

```
4
5     public void flip() {
6         if (color == Color.Black) color = Color.White;
7         else color = Color.Black;
8     }
9
10    public Color getColor() { return color; }
11 }
```

The Player holds only a very limited amount of information. It does not even hold its own score, but it does have a method one can call to get the score. `Player.getScore()` will call out to the `GameManager` to retrieve this value.

```
12 public class Player {
13     private Color color;
14     public Player(Color c) { color = c; }
15
16     public int getScore() { ... }
17
18     public boolean playPiece(int r, int c) {
19         return Game.getInstance().getBoard().placeColor(r, c, color);
20     }
21
22     public Color getColor() { return color; }
23 }
```

A fully functioning (automated) version of this code can be found in the downloadable code attachment.

Remember that in many problems, what you did is less important than *why* you did it. Your interviewer probably doesn't care much whether you chose implement `Game` as a singleton or not, but she probably does care that you took the time to think about it and discuss the trade-offs.

- 8.9** Explain the data structures and algorithms that you would use to design an in-memory file system. Illustrate with an example in code where possible.

pg 106

### SOLUTION

Many candidates may see this problem and instantly panic. A file system seems so low level!

However, there's no need to panic. If we think through the components of a file system, we can tackle this problem just like any other object-oriented design question.

A file system, in its most simplistic version, consists of **Files** and **Directories**. Each **Directory** contains a set of **Files** and **Directories**. Since **Files** and **Directories** share so many characteristics, we've implemented them such that they inherit

from the same class, Entry.

```
1  public abstract class Entry {  
2      protected Directory parent;  
3      protected long created;  
4      protected long lastUpdated;  
5      protected long lastAccessed;  
6      protected String name;  
7  
8      public Entry(String n, Directory p) {  
9          name = n;  
10         parent = p;  
11         created = System.currentTimeMillis();  
12         lastUpdated = System.currentTimeMillis();  
13         lastAccessed = System.currentTimeMillis();  
14     }  
15  
16     public boolean delete() {  
17         if (parent == null) return false;  
18         return parent.deleteEntry(this);  
19     }  
20  
21     public abstract int size();  
22  
23     public String getFullPath() {  
24         if (parent == null) return name;  
25         else return parent.getFullPath() + "/" + name;  
26     }  
27  
28     /* Getters and setters. */  
29     public long getCreationTime() { return created; }  
30     public long getLastUpdatedTime() { return lastUpdated; }  
31     public long getLastAccessedTime() { return lastAccessed; }  
32     public void changeName(String n) { name = n; }  
33     public String getName() { return name; }  
34 }  
35  
36     public class File extends Entry {  
37         private String content;  
38         private int size;  
39  
40         public File(String n, Directory p, int sz) {  
41             super(n, p);  
42             size = sz;  
43         }  
44  
45         public int size() { return size; }  
46         public String getContents() { return content; }  
47         public void setContents(String c) { content = c; }  
48 }
```

```
49
50 public class Directory extends Entry {
51     protected ArrayList<Entry> contents;
52
53     public Directory(String n, Directory p) {
54         super(n, p);
55         contents = new ArrayList<Entry>();
56     }
57
58     public int size() {
59         int size = 0;
60         for (Entry e : contents) {
61             size += e.size();
62         }
63         return size;
64     }
65
66     public int numberOfFiles() {
67         int count = 0;
68         for (Entry e : contents) {
69             if (e instanceof Directory) {
70                 count++; // Directory counts as a file
71                 Directory d = (Directory) e;
72                 count += d.numberOfFiles();
73             } else if (e instanceof File) {
74                 count++;
75             }
76         }
77         return count;
78     }
79
80     public boolean deleteEntry(Entry entry) {
81         return contents.remove(entry);
82     }
83
84     public void addEntry(Entry entry) {
85         contents.add(entry);
86     }
87
88     protected ArrayList<Entry> getContents() { return contents; }
89 }
```

Alternatively, we could have implemented `Directory` such that it contains separate lists for files and subdirectories. This makes the `numberOfFiles()` method a bit cleaner, since it doesn't need to use the `instanceof` operator, but it does prohibit us from cleanly sorting files and directories by dates or names.

- 8.10 Design and implement a hash table which uses chaining (linked lists) to handle collisions.

pg 106

## SOLUTION

Suppose we are implementing a hash table that looks like `Hash<K, V>`. That is, the hash table maps from objects of type K to objects of type V.

At first, we might think our data structure would look something like this:

```
1 public class Hash<K, V> {  
2     LinkedList<V>[] items;  
3     public void put(K key, V value) { ... }  
4     public V get(K key) { ... }  
5 }
```

Note that `items` is an array of linked lists, where `items[i]` is a linked list of all objects with keys that map to index `i` (that is, all the objects that collided at `i`).

This would seem to work until we think more deeply about collisions.

Suppose we have a very simple hash function that uses the string length.

```
1 public int hashCodeOfKey(K key) {  
2     return key.toString().length() % items.length;  
3 }
```

The keys `jim` and `bob` will map to the same index in the array, even though they are different keys. We need to search through the linked list to find the actual object that corresponds to these keys. But how would we do that? All we've stored in the linked list is the value, not the original key.

This is why we need to store both the value and the original key.

One way to do that is to create another object called `Cell` which pairs keys and values. With this implementation, our linked list is of type `Cell`.

The code below uses this implementation.

```
1 public class Hash<K, V> {  
2     private final int MAX_SIZE = 10;  
3     LinkedList<Cell<K, V>>[] items;  
4  
5     public Hash() {  
6         items = (LinkedList<Cell<K, V>>[]) new LinkedList[MAX_SIZE];  
7     }  
8  
9     /* Really, really stupid hash. */  
10    public int hashCodeOfKey(K key) {  
11        return key.toString().length() % items.length;  
12    }  
13}
```

```

14     public void put(K key, V value) {
15         int x = hashCodeOfKey(key);
16         if (items[x] == null) {
17             items[x] = new LinkedList<Cell<K, V>>();
18         }
19
20         LinkedList<Cell<K, V>> collided = items[x];
21
22         /* Look for items with same key and replace if found */
23         for (Cell<K, V> c : collided) {
24             if (c.equivalent(key)) {
25                 collided.remove(c);
26                 break;
27             }
28         }
29
30         Cell<K, V> cell = new Cell<K, V>(key, value);
31         collided.add(cell);
32     }
33
34     public V get(K key) {
35         int x = hashCodeOfKey(key);
36         if (items[x] == null) {
37             return null;
38         }
39         LinkedList<Cell<K, V>> collided = items[x];
40         for (Cell<K, V> c : collided) {
41             if (c.equivalent(key)) {
42                 return c.getValue();
43             }
44         }
45
46         return null;
47     }
48 }

```

The Cell class pairs the data value and its key. This will allow us to search through the linked list (created by "colliding," but different, keys) and find the object with the exact key value.

```

1  public class Cell<K, V> {
2      private K key;
3      private V value;
4      public Cell(K k, V v) {
5          key = k;
6          value = v;
7      }
8
9      public boolean equivalent(Cell<K, V> c) {
10         return equivalent(c.getKey());
11     }

```

```
12
13     public boolean equivalent(K k) {
14         return key.equals(k);
15     }
16
17     public K getKey() { return key; }
18     public V getValue() { return value; }
19 }
```

Another common implementation for a hash table is to use a binary search tree as the underlying data structure. Retrieving an element will no longer be  $O(1)$  (although, technically, it's not  $O(1)$  if there are many collisions), but it prevents us from creating an unnecessarily large array to hold items.



# **Recursion and Dynamic Programming**

*Concepts and Algorithms: Solutions*

**Chapter 9**

- 9.1 A child is running up a staircase with  $n$  steps, and can hop either 1 step, 2 steps, or 3 steps at a time. Implement a method to count how many possible ways the child can run up the stairs.

pg 109

### SOLUTION

We can approach this problem from the top down. On the very last hop, up to the  $n$ th step, the child could have done either a single, double, or triple step hop. That is, the last move might have been a single step hop from step  $n-1$ , a double step hop from step  $n-2$ , or a triple step hop from  $n-3$ . The total number of ways of reaching the last step is therefore the sum of the number of ways of reaching each of the last three steps.

A simple implementation of this code is below.

```
1 public int countWays(int n) {  
2     if (n < 0) {  
3         return 0;  
4     } else if (n == 0) {  
5         return 1;  
6     } else {  
7         return countWays(n - 1) + countWays(n - 2) +  
8                 countWays(n - 3);  
9     }  
10 }
```

Like the Fibonacci problem, the runtime of this algorithm is exponential (specifically,  $O(3^n)$ ), since each call branches out to three more calls. This means that `countWays` is called many times for the same values, which is unnecessary. We can fix this through dynamic programming.

```
11 public static int countWaysDP(int n, int[] map) {  
12     if (n < 0) {  
13         return 0;  
14     } else if (n == 0) {  
15         return 1;  
16     } else if (map[n] > -1) {  
17         return map[n];  
18     } else {  
19         map[n] = countWaysDP(n - 1, map) +  
20                 countWaysDP(n - 2, map) +  
21                 countWaysDP(n - 3, map);  
22     }  
23 }  
24 }
```

Regardless of whether or not you use dynamic programming, note that the number of ways will quickly overflow the bounds of an integer. By the time you get to just  $n = 37$ , the result has already overflowed. Using a `long` will delay, but not completely solve, this issue.

- 9.2** Imagine a robot sitting on the upper left corner of an  $X$  by  $Y$  grid. The robot can only move in two directions: right and down. How many possible paths are there for the robot to go from  $(0, 0)$  to  $(X, Y)$ ?

**FOLLOW UP**

Imagine certain spots are “off limits,” such that the robot cannot step on them. Design an algorithm to find a path for the robot from the top left to the bottom right.

pg 109

**SOLUTION**

We need to count the number of ways of making a path with  $X$  right steps and  $Y$  down steps. This path will have  $X+Y$  steps total.

To build a path, we are essentially selecting  $X$  times to move right out of a total of  $X+Y$  moves. Thus, the number of total paths must be the number of ways of selecting  $X$  items out of  $X+Y$  items. This is given by the binomial expression (a.k.a., “ $n$  choose  $r$ ”):

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

In terms of this problem, the expression is:

$$\binom{X+Y}{X} = \frac{(X+Y)!}{X!Y!}$$

If you didn’t know the binomial expression, you could still deduce how to solve this problem.

Think about each path as a string of length  $X+Y$  consisting  $X$  ‘R’s and  $Y$  ‘D’s. We know that the number of strings we can make with  $X+Y$  unique characters is  $(X+Y)!$ . However, in this case,  $X$  of the characters are ‘R’s and  $Y$  are ‘D’s. Since the ‘R’s can be rearranged in  $X!$  ways, each of which is identical, and we can do an equivalent thing with the ‘D’s, we need to divide out by  $X!$  and  $Y!$ . We then get the same expression as we had before:

$$\frac{(X+Y)!}{X!Y!}$$

**Follow Up: Find a path (with off limit spots)**

If we picture our grid, the only way to move to spot  $(X, Y)$  is by moving to one of the adjacent spots:  $(X-1, Y)$  or  $(X, Y-1)$ . So, we need to find a path to either  $(X-1, Y)$  or  $(X, Y-1)$ .

How do we find a path to those spots? To find a path to  $(X-1, Y)$  or  $(X, Y-1)$ , we need to move to one of its adjacent cells. So, we need to find a path to a spot adjacent to  $(X-1, Y)$ , which are coordinates  $(X-2, Y)$  and  $(X-1, Y-1)$ , or a spot adjacent to

(X,Y-1), which are spots (X-1,Y-1) and (X,Y-2). Observe that we list the point (X-1,Y-1) twice; we'll discuss that issue later.

So then, to find a path from the origin, we just work backwards like this. Starting from the last cell, we try to find a path to each of its adjacent cells. The recursive code below implements this algorithm.

```
1  public boolean getPath(int x, int y, ArrayList<Point> path) {  
2      Point p = new Point(x, y);  
3      path.add(p);  
4      if (x == 0 && y == 0) {  
5          return true; // found a path  
6      }  
7      boolean success = false;  
8      if (x >= 1 && isFree(x - 1, y)) { // Try left  
9          success = getPath(x - 1, y, path); // Free! Go left  
10     }  
11     if (!success && y >= 1 && isFree(x, y - 1)) { // Try up  
12         success = getPath(x, y - 1, path); // Free! Go up  
13     }  
14     if (success) {  
15         path.add(p); // Right way! Add to path.  
16     }  
17     return success;  
18 }
```

Earlier, we'd mentioned an issue with duplicate paths. To find a path to (X,Y), we look for a path to an adjacent coordinate: (X-1,Y) or (X,Y-1). Of course, if one of those squares is off limits, we ignore it. Then, we look at their adjacent coordinates: (X-2,Y), (X-1,Y-1), (X-1,Y-1), and (X,Y-2). The spot (X-1,Y-1) appears twice, which means that we're duplicating effort. Ideally, we should remember that we already visited (X-1,Y-1) so that we don't waste our time.

This is what the dynamic programming algorithm below does.

```
1  public boolean getPath(int x, int y, ArrayList<Point> path,  
2                         Hashtable<Point, Boolean> cache) {  
3      Point p = new Point(x, y);  
4      if (cache.containsKey(p)) { // Already visited this cell  
5          return cache.get(p);  
6      }  
7      if (x == 0 && y == 0) {  
8          return true; // found a path  
9      }  
10     boolean success = false;  
11     if (x >= 1 && isFree(x - 1, y)) { // Try right  
12         success = getPath(x - 1, y, path, cache); // Free! Go right  
13     }  
14     if (!success && y >= 1 && isFree(x, y - 1)) { // Try down  
15         success = getPath(x, y - 1, path, cache); // Free! Go down  
16     }  
17     if (success) {
```

```

18     path.add(p); // Right way! Add to path
19 }
20 cache.put(p, success); // Cache result
21 return success;
22 }
```

This simple change will make our code run substantially faster.

- 9.3** A *magic index* in an array  $A[1 \dots n-1]$  is defined to be an index such that  $A[i] = i$ . Given a sorted array of distinct integers, write a method to find a magic index, if one exists, in array  $A$ .

#### FOLLOW UP

What if the values are not distinct?

pg 109

#### SOLUTION

Immediately, the brute force solution should jump to mind—and there's no shame in mentioning it. We simply iterate through the array, looking for an element which matches this condition.

```

1 public static int magicSlow(int[] array) {
2     for (int i = 0; i < array.length; i++) {
3         if (array[i] == i) {
4             return i;
5         }
6     }
7     return -1;
8 }
```

Given that the array is sorted though, it's very likely that we're supposed to use this condition.

We may recognize that this problem sounds a lot like the classic binary search problem. Leveraging the Pattern Matching approach for generating algorithms, how might we apply binary search here?

In binary search, we find an element  $k$  by comparing it to the middle element,  $x$ , and determining if  $k$  would land on the left or the right side of  $x$ .

Building off this approach, is there a way that we can look at the middle element to determine where a magic index might be? Let's look at a sample array:

-40	-20	-1	1	2	<u>3</u>	5	7	9	12	13
0	1	2	3	4	<u>5</u>	6	7	8	9	10

When we look at the middle element  $A[5] = 3$ , we know that the magic index must be on the right side, since  $A[mid] < mid$ .

Why couldn't the magic index be on the left side? Observe that when we move from  $i$  to  $i-1$ , the value at this index must decrease by at least 1, if not more (since the array is sorted and all the elements are distinct). So, if the middle element is already too small to be a magic index, then when we move to the left, subtracting  $k$  indexes and (at least)  $k$  values, all subsequent elements will also be too small.

We continue to apply this recursive algorithm, developing code that looks very much like binary search.

```
1 public static int magicFast(int[] array, int start, int end) {  
2     if (end < start || start < 0 || end >= array.length) {  
3         return -1;  
4     }  
5     int mid = (start + end) / 2;  
6     if (array[mid] == mid) {  
7         return mid;  
8     } else if (array[mid] > mid){  
9         return magicFast(array, start, mid - 1);  
10    } else {  
11        return magicFast(array, mid + 1, end);  
12    }  
13 }  
14  
15 public static int magicFast(int[] array) {  
16     return magicFast(array, 0, array.length - 1);  
17 }
```

### Follow Up: What if the elements are not distinct?

If the elements are not distinct, then this algorithm fails. Consider the following array:

-10	-5	2	2	2	3	4	7	9	12	13
0	1	2	3	4	5	6	7	8	9	10

When we see that  $A[\text{mid}] < \text{mid}$ , we cannot conclude which side the magic index is on. It could be on the right side, as before. Or, it could be on the left side (as it, in fact, is).

Could it be *anywhere* on the left side? Not exactly. Since  $A[5] = 3$ , we know that  $A[4]$  couldn't be a magic index.  $A[4]$  would need to be 4 to be the magic index, but  $A[4]$  must be less than  $A[5]$ .

In fact, when we see that  $A[5] = 3$ , we'll need to recursively search the right side as before. But, to search the left side, we can skip a bunch of elements and only recursively search elements  $A[0]$  through  $A[3]$ .  $A[3]$  is the first element that could be a magic index.

The general pattern is that we compare `midIndex` and `midValue` for equality first. Then, if they are not equal, we recursively search the left and right sides as follows:

- Left side: search indices `start` through `Math.min(midIndex - 1, midValue)`.

- Right side: search indices `Math.max(midIndex + 1, midValue)` through `end`.

The code below implements this algorithm.

```

1  public static int magicFast(int[] array, int start, int end) {
2      if (end < start || start < 0 || end >= array.length) {
3          return -1;
4      }
5      int midIndex = (start + end) / 2;
6      int midValue = array[midIndex];
7      if (midValue == midIndex) {
8          return midIndex;
9      }
10     /* Search left */
11     int leftIndex = Math.min(midIndex - 1, midValue);
12     int left = magicFast(array, start, leftIndex);
13     if (left >= 0) {
14         return left;
15     }
16     /* Search right */
17     int rightIndex = Math.max(midIndex + 1, midValue);
18     int right = magicFast(array, rightIndex, end);
19
20     return right;
21 }
22
23 public static int magicFast(int[] array) {
24     return magicFast(array, 0, array.length - 1);
25 }
```

Note that in the above code, if the elements are all distinct, the method operates almost identically to the first solution.

#### 9.4 Write a method to return all subsets of a set.

pg 109

#### SOLUTION

We should first have some reasonable expectations of our time and space complexity. How many subsets of a set are there? We can compute this by realizing that when we generate a subset, each element has the “choice” of either being in there or not. That is, for the first element, there are two choices: it is either in the set, or it is not. For the second, there are two, etc. So, doing  $\{2 * 2 * \dots\}$   $n$  times gives us  $2^n$  subsets. We will therefore not be able to do better than  $O(2^n)$  in time or space complexity.

The subsets of  $\{a_1, a_2, \dots, a_n\}$  are also called the powerset,  $P(\{a_1, a_2, \dots, a_n\})$ , or just  $P(n)$ .

### Solution #1: Recursion

This problem is a good candidate for the Base Case and Build approach. Imagine that we are trying to find all subsets of a set like  $S = \{a_1, a_2, \dots, a_n\}$ . We can start with the Base Case.

*Base Case:*  $n = 0$ .

There is just one subset of the empty set: {}.

*Case:*  $n = 1$ .

There are two subsets of the set  $\{a_1\}$ : {},  $\{a_1\}$ .

*Case:*  $n = 2$ .

There are four subsets of the set  $\{a_1, a_2\}$ : {},  $\{a_1\}$ ,  $\{a_2\}$ ,  $\{a_1, a_2\}$ .

*Case:*  $n = 3$ .

Now here's where things get interesting. We want to find a way of generating the solution for  $n = 3$  based on the prior solutions.

What is the difference between the solution for  $n = 3$  and the solution for  $n = 2$ ? Let's look at this more deeply:

$$\begin{aligned} P(2) &= \{\}, \{a_1\}, \{a_2\}, \{a_1, a_2\} \\ P(3) &= \{\}, \{a_1\}, \{a_2\}, \{a_3\}, \{a_1, a_2\}, \{a_1, a_3\}, \{a_2, a_3\}, \\ &\quad \{a_1, a_2, a_3\} \end{aligned}$$

The difference between these solutions is that  $P(2)$  is missing all the subsets containing  $a_3$ .

$$P(3) - P(2) = \{a_3\}, \{a_1, a_3\}, \{a_2, a_3\}, \{a_1, a_2, a_3\}$$

How can we use  $P(2)$  to create  $P(3)$ ? We can simply clone the subsets in  $P(2)$  and add  $a_3$  to them:

$$\begin{aligned} P(2) &= \{\}, \{a_1\}, \{a_2\}, \{a_1, a_2\} \\ P(2) + a_3 &= \{a_3\}, \{a_1, a_3\}, \{a_2, a_3\}, \{a_1, a_2, a_3\} \end{aligned}$$

When merged together, the lines above make  $P(3)$ .

*Case:*  $n > 0$

Generating  $P(n)$  for the general case is just a simple generalization of the above steps. We compute  $P(n-1)$ , clone the results, and then add  $a_n$  to each of these cloned sets.

The following code implements this algorithm:

```
1 ArrayList<ArrayList<Integer>> getSubsets(ArrayList<Integer> set,
2                                         int index) {
3     ArrayList<ArrayList<Integer>> allsubsets;
4     if (set.size() == index) { // Base case - add empty set
5         allsubsets = new ArrayList<ArrayList<Integer>>();
6         allsubsets.add(new ArrayList<Integer>()); // Empty set
```

```

7     } else {
8         allsubsets = getSubsets(set, index + 1);
9         int item = set.get(index);
10        ArrayList<ArrayList<Integer>> moresubsets =
11            new ArrayList<ArrayList<Integer>>();
12        for (ArrayList<Integer> subset : allsubsets) {
13            ArrayList<Integer> newsubset = new ArrayList<Integer>();
14            newsubset.addAll(subset); //
15            newsubset.add(item);
16            moresubsets.add(newsubset);
17        }
18        allsubsets.addAll(moresubsets);
19    }
20    return allsubsets;
21 }

```

This solution will be  $O(2^n)$  in time and space, which is the best we can do. For a slight optimization, we could also implement this algorithm iteratively.

### Solution #2: Combinatorics

While there's nothing wrong with the above solution, there's another way to approach it.

Recall that when we're generating a set, we have two choices for each element: (1) the element is in the set (the "yes" state) or (2) the element is not in the set (the "no" state). This means that each subset is a sequence of yeses / nos—e.g., "yes, yes, no, no, yes, no"

This gives us  $2^n$  possible subsets. How can we iterate through all possible sequences of "yes" / "no" states for all elements? If each "yes" can be treated as a 1 and each "no" can be treated as a 0, then each subset can be represented as a binary string.

Generating all subsets, then, really just comes down to generating all binary numbers (that is, all integers). We iterate through all numbers from 1 to  $2^n$  and translate the binary representation of the numbers into a set. Easy!

```

1  ArrayList<ArrayList<Integer>> getSubsets2(ArrayList<Integer> set) {
2      ArrayList<ArrayList<Integer>> allsubsets =
3          new ArrayList<ArrayList<Integer>>();
4      int max = 1 << set.size(); /* Compute 2^n */
5      for (int k = 0; k < max; k++) {
6          ArrayList<Integer> subset = convertIntToSet(k, set);
7          allsubsets.add(subset);
8      }
9      return allsubsets;
10 }
11
12 ArrayList<Integer> convertIntToSet(int x, ArrayList<Integer> set) {
13     ArrayList<Integer> subset = new ArrayList<Integer>();
14     int index = 0;

```

```
15    for (int k = x; k > 0; k >>= 1) {  
16        if ((k & 1) == 1) {  
17            subset.add(set.get(index));  
18        }  
19        index++;  
20    }  
21    return subset;  
22 }
```

There's nothing substantially better or worse about this solution compared to the first one.

### 9.5 Write a method to compute all permutations of a string

pg 109

#### SOLUTION

Like in many recursive problems, the Base Case and Build approach will be useful. Assume we have a string  $S$  represented by the characters  $a_1 a_2 \dots a_n$ .

*Base Case:*  $n = 1$

The only permutation of  $S = a_1$  is the string  $a_1$ .

*Case:*  $n = 2$

The permutations of  $S = a_1 a_2$  are the strings  $a_1 a_2$  and  $a_2 a_1$ .

*Case:*  $n = 3$

Here is where the cases get more interesting. How can we generate all permutations of  $a_1 a_2 a_3$  given the permutations of  $a_1 a_2$ ? That is, we need to generate

$a_1 a_2 a_3, a_1 a_3 a_2, a_2 a_1 a_3, a_2 a_3 a_1, a_3 a_1 a_2, a_3 a_2 a_1$

given

$a_1 a_2, a_2 a_1.$

The difference between these lists is that the first one contains  $a_3$  while the second list does not. So, how can we generate  $f(3)$  from  $f(2)$ ? By pushing  $a_3$  into every possible spot in the strings in  $f(2)$ .

*Case:*  $n > 0$

For the general case, we just repeat this process. We solve for  $f(n-1)$ , and then push  $a_n$  into every spot in each of these strings.

The code below does just this.

```
1  public static ArrayList<String> getPerms(String str) {  
2      if (str == null) {
```

```

3     return null;
4 }
5 ArrayList<String> permutations = new ArrayList<String>();
6 if (str.length() == 0) { // base case
7     permutations.add("");
8     return permutations;
9 }
10
11 char first = str.charAt(0); // get the first character
12 String remainder = str.substring(1); // remove the 1st character
13 ArrayList<String> words = getPerms(remainder);
14 for (String word : words) {
15     for (int j = 0; j <= word.length(); j++) {
16         String s = insertCharAt(word, first, j);
17         permutations.add(s);
18     }
19 }
20 return permutations;
21 }
22
23 public static String insertCharAt(String word, char c, int i) {
24     String start = word.substring(0, i);
25     String end = word.substring(i);
26     return start + c + end;
27 }

```

This solution takes  $O(n!)$  time, since there are  $n!$  permutations. We cannot do better than this.

- 9.6** Implement an algorithm to print all valid (i.e., properly opened and closed) combinations of  $n$ -pairs of parentheses.

pg 110

## SOLUTION

Our first thought here might be to apply a recursive approach where we build the solution for  $f(n)$  by adding pairs of parentheses to  $f(n-1)$ . That's certainly a good instinct.

Let's consider the solution for  $n = 3$ :

((())) ((()) ()()) (())() ()()()

How might we build this from  $n = 2$ ?

(()) ()()

We can do this by inserting a pair of parentheses inside every existing pair of parentheses, as well as one at the beginning of the string. Any other places that we could insert parentheses, such as at the end of the string, would reduce to the earlier cases.

So, we have the following:

```

((() -> ((()) /* inserted pair after 1st left paren */
-> (((()) /* inserted pair after 2nd left paren */
-> (((() /* inserted pair at beginning of string */
()() -> ((()) /* inserted pair after 1st left paren */
-> ((() /* inserted pair after 2nd left paren */
-> ((()) /* inserted pair at beginning of string */

```

But wait—we have some duplicate pairs listed. The string `()()` is listed twice.

If we're going to apply this approach, we'll need to check for duplicate values before adding a string to our list.

```

1  public static Set<String> generateParens(int remaining) {
2      Set<String> set = new HashSet<String>();
3      if (remaining == 0) {
4          set.add("");
5      } else {
6          Set<String> prev = generateParens(remaining - 1);
7          for (String str : prev) {
8              for (int i = 0; i < str.length(); i++) {
9                  if (str.charAt(i) == '(') {
10                      String s = insertInside(str, i);
11                      /* Add s to set if it's not already in there. Note:
12                         * HashSet automatically checks for duplicates before
13                         * adding, so an explicit check is not necessary. */
14                      set.add(s);
15                  }
16              }
17              if (!set.contains("()" + str)) {
18                  set.add("()" + str);
19              }
20          }
21      }
22      return set;
23  }
24
25  public String insertInside(String str, int leftIndex) {
26      String left = str.substring(0, leftIndex + 1);
27      String right = str.substring(leftIndex + 1, str.length());
28      return left + "(" + right;
29  }

```

This works, but it's not very efficient. We waste a lot of time coming up with the duplicate strings.

We can avoid this duplicate string issue by building the string from scratch. Under this approach, we add left and right parens, as long as our expression stays valid.

On each recursive call, we have the index for a particular character in the string. We need to select either a left or a right paren. When can we use a left paren, and when can we use a right paren?

- Left Paren:* As long as we haven't used up all the left parentheses, we can always insert a left paren.
- Right Paren:* We can insert a right paren as long as it won't lead to a syntax error. When will we get a syntax error? We will get a syntax error if there are more right parentheses than left.

So, we simply keep track of the number of left and right parentheses allowed. If there are left parens remaining, we'll insert a left paren and recurse. If there are more right parens remaining than left (i.e., if there are more left parens in use than right parens), then we'll insert a right paren and recurse.

```

1  public void addParen(ArrayList<String> list, int leftRem,
2                      int rightRem, char[] str, int count) {
3      if (leftRem < 0 || rightRem < leftRem) return; // invalid state
4
5      if (leftRem == 0 && rightRem == 0) { /* no more parens left */
6          String s = String.valueOf(str);
7          list.add(s);
8      } else {
9          /* Add left paren, if there are any left parens remaining. */
10         if (leftRem > 0) {
11             str[count] = '(';
12             addParen(list, leftRem - 1, rightRem, str, count + 1);
13         }
14
15         /* Add right paren, if expression is valid */
16         if (rightRem > leftRem) {
17             str[count] = ')';
18             addParen(list, leftRem, rightRem - 1, str, count + 1);
19         }
20     }
21 }
22
23 public ArrayList<String> generateParens(int count) {
24     char[] str = new char[count*2];
25     ArrayList<String> list = new ArrayList<String>();
26     addParen(list, count, count, str, 0);
27     return list;
28 }
```

Because we insert left and right parentheses at each index in the string, and we never repeat an index, each string is guaranteed to be unique.

- 9.7 *Implement the "paint fill" function that one might see on many image editing programs. That is, given a screen (represented by a two-dimensional array of colors), a point, and a new color, fill in the surrounding area until the color changes from the original color.*

**SOLUTION**

First, let's visualize how this method works. When we call `paintFill` (i.e., "click" paint fill in the image editing application) on, say, a green pixel, we want to "bleed" outwards. Pixel by pixel, we expand outwards by calling `paintFill` on the surrounding pixel. When we hit a pixel that is not green, we stop.

We can implement this algorithm recursively:

```

1  enum Color {
2      Black, White, Red, Yellow, Green
3  }
4
5  boolean paintFill(Color[][] screen, int x, int y, Color ocolor,
6                      Color ncolor) {
7      if (x < 0 || x >= screen[0].length ||
8          y < 0 || y >= screen.length) {
9          return false;
10     }
11     if (screen[y][x] == ocolor) {
12         screen[y][x] = ncolor;
13         paintFill(screen, x - 1, y, ocolor, ncolor); // left
14         paintFill(screen, x + 1, y, ocolor, ncolor); // right
15         paintFill(screen, x, y - 1, ocolor, ncolor); // top
16         paintFill(screen, x, y + 1, ocolor, ncolor); // bottom
17     }
18     return true;
19 }
20
21 boolean paintFill(Color[][] screen, int x, int y, Color ncolor) {
22     if (screen[y][x] == ncolor) return false;
23     return paintFill(screen, x, y, screen[y][x], ncolor);
24 }
```

Note the ordering of the `x` and `y` in `screen[y][x]`, and remember this when you hit graphical problem. Because `x` represents the *horizontal* axis (that is, it's left to right), it actually corresponds to the number of a column, not the number of rows. The value of `y` equals the number of rows. This is a very easy place to make a mistake in an interview, as well as in your daily coding.

- 9.8** Given an infinite number of quarters (25 cents), dimes (10 cents), nickels (5 cents) and pennies (1 cent), write code to calculate the number of ways of representing  $n$  cents.

pg 110

**SOLUTION**

This is a recursive problem, so let's figure out how to compute `makeChange(n)` using

prior solutions (i.e., sub-problems).

Let's say  $n = 100$ . We want to compute the number of ways of making change for 100 cents. What is the relationship between this problem and its sub-problems?

We know that making change for 100 cents will involve either 0, 1, 2, 3, or 4 quarters. So:

`makeChange(100) =`

```
makeChange(100 using 0 quarters) +
makeChange(100 using 1 quarter) +
makeChange(100 using 2 quarters) +
makeChange(100 using 3 quarters) +
makeChange(100 using 4 quarters)
```

Inspecting this further, we can see that some of these problems reduce. For example, `makeChange(100 using 1 quarter)` will equal `makeChange(75 using 0 quarters)`. This is because, if we must use exactly one quarter to make change for 100 cents, then our only remaining choices involve making change for the remaining 75 cents.

We can apply the same logic to `makeChange(100 using 2 quarters)`, `makeChange(100 using 3 quarters)` and `makeChange(100 using 4 quarters)`. We have thus reduced the above statement to the following.

```
makeChange(100) =
makeChange(100 using 0 quarters) +
makeChange(75 using 0 quarters) +
makeChange(50 using 0 quarters) +
makeChange(25 using 0 quarters) +
1
```

Note that the final statement from above, `makeChange(100 using 4 quarters)`, equals 1. We call this "fully reduced."

Now what? We've used up all our quarters, so now we can start applying our next biggest denomination: dimes.

Our approach for quarters applies to dimes as well, but we apply this for *each* of the four of five parts of the above statement. So, for the first part, we get the following statements:

```
makeChange(100 using 0 quarters) =
makeChange(100 using 0 quarters, 0 dimes) +
makeChange(100 using 0 quarters, 1 dime) +
makeChange(100 using 0 quarters, 2 dimes) +
...
makeChange(100 using 0 quarters, 10 dimes)

makeChange(75 using 0 quarters) =
makeChange(75 using 0 quarters, 0 dimes) +
makeChange(75 using 0 quarters, 1 dime) +
makeChange(75 using 0 quarters, 2 dimes) +
```

```

...
makeChange(75 using 0 quarters, 7 dimes)

makeChange(50 using 0 quarters) =
    makeChange(50 using 0 quarters, 0 dimes) +
    makeChange(50 using 0 quarters, 1 dime) +
    makeChange(50 using 0 quarters, 2 dimes) +
    ...
makeChange(50 using 0 quarters, 5 dimes)

makeChange(25 using 0 quarters) =
    makeChange(25 using 0 quarters, 0 dimes) +
    makeChange(25 using 0 quarters, 1 dime) +
    makeChange(25 using 0 quarters, 2 dimes)

```

Each one of these, in turn, expands out once we start applying nickels. We end up with a tree-like recursive structure where each call expands out to four or more calls.

The base case of our recursion is the fully reduced statement. For example, `makeChange(50 using 0 quarters, 5 dimes)` is fully reduced to 1, since 5 dimes equals 50 cents.

This leads to a recursive algorithm that looks like this:

```

1 public int makeChange(int n, int denom) {
2     int next_denom = 0;
3     switch (denom) {
4         case 25:
5             next_denom = 10;
6             break;
7         case 10:
8             next_denom = 5;
9             break;
10        case 5:
11            next_denom = 1;
12            break;
13        case 1:
14            return 1;
15    }
16
17    int ways = 0;
18    for (int i = 0; i * denom <= n; i++) {
19        ways += makeChange(n - i * denom, next_denom);
20    }
21    return ways;
22 }
23
24 System.out.writeln(makeChange(100, 25));

```

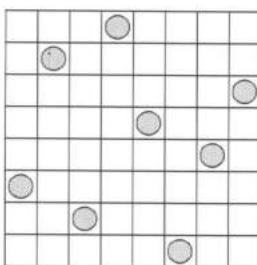
Although we've implemented this to work for US currency, it can be easily extended to work for any other set of denominations.

- 9.9** Write an algorithm to print all ways of arranging eight queens on an 8x8 chess board so that none of them share the same row, column or diagonal. In this case, "diagonal" means all diagonals, not just the two that bisect the board.

pg 110

**SOLUTION**

We have eight queens which must be lined up on an 8x8 chess board such that none share the same row, column or diagonal. So, we know that each row and column (and diagonal) must be used exactly once.



A "Solved" Board with 8 Queens

Picture the queen that is placed last, which we'll assume is on row 8. (This is an okay assumption to make since the ordering of placing the queens is irrelevant.) On which cell in row 8 is this queen? There are eight possibilities, one for each column.

So if we want to know all the valid ways of arranging 8 queens on an 8x8 chess board, it would be:

```

ways to arrange 8 queens on an 8x8 board =
  ways to arrange 8 queens on an 8x8 board with queen at (7, 0) +
  ways to arrange 8 queens on an 8x8 board with queen at (7, 1) +
  ways to arrange 8 queens on an 8x8 board with queen at (7, 2) +
  ways to arrange 8 queens on an 8x8 board with queen at (7, 3) +
  ways to arrange 8 queens on an 8x8 board with queen at (7, 4) +
  ways to arrange 8 queens on an 8x8 board with queen at (7, 5) +
  ways to arrange 8 queens on an 8x8 board with queen at (7, 6) +
  ways to arrange 8 queens on an 8x8 board with queen at (7, 7)

```

We can compute each one of these using a very similar approach:

```

ways to arrange 8 queens on an 8x8 board with queen at (7, 3) =
  ways to ... with queens at (7, 3) and (6, 0) +
  ways to ... with queens at (7, 3) and (6, 1) +
  ways to ... with queens at (7, 3) and (6, 2) +
  ways to ... with queens at (7, 3) and (6, 4) +
  ways to ... with queens at (7, 3) and (6, 5) +
  ways to ... with queens at (7, 3) and (6, 6) +
  ways to ... with queens at (7, 3) and (6, 7)

```

Note that we don't need to consider combinations with queens at (7, 3) and (6, 3),

since this is a violation of the requirement that every queen is in its own row, column and diagonal.

Implementing this is now reasonably straightforward.

```

1 int GRID_SIZE = 8;
2
3 void placeQueens(int row, Integer[] columns,
4                   ArrayList<Integer[]> results) {
5     if (row == GRID_SIZE) { // Found valid placement
6         results.add(columns.clone());
7     } else {
8         for (int col = 0; col < GRID_SIZE; col++) {
9             if (checkValid(columns, row, col)) {
10                 columns[row] = col; // Place queen
11                 placeQueens(row + 1, columns, results);
12             }
13         }
14     }
15 }
16
17 /* Check if (row1, column1) is a valid spot for a queen by checking
18 * if there is a queen in the same column or diagonal. We don't
19 * need to check it for queens in the same row because the calling
20 * placeQueen only attempts to place one queen at a time. We know
21 * this row is empty. */
22 boolean checkValid(Integer[] columns, int row1, int column1) {
23     for (int row2 = 0; row2 < row1; row2++) {
24         int column2 = columns[row2];
25         /* Check if (row2, column2) invalidates (row1, column1) as a
26          * queen spot. */
27
28         /* Check if rows have a queen in the same column */
29         if (column1 == column2) {
30             return false;
31         }
32
33         /* Check diagonals: if the distance between the columns
34          * equals the distance between the rows, then they're in the
35          * same diagonal. */
36         int columnDistance = Math.abs(column2 - column1);
37
38         /* row1 > row2, so no need for abs */
39         int rowDistance = row1 - row2;
40         if (columnDistance == rowDistance) {
41             return false;
42         }
43     }
44     return true;
45 }
```

Observe that since each row can only have one queen, we don't need to store our board as a full 8x8 matrix. We only need a single array where  $\text{column}[row] = c$  indicates that row  $r$  has a queen at column  $c$ .

- 9.10** You have a stack of  $n$  boxes, with widths  $w_i$ , heights  $h_i$ , and depths  $d_i$ . The boxes cannot be rotated and can only be stacked on top of one another if each box in the stack is strictly larger than the box above it in width, height, and depth. Implement a method to build the tallest stack possible, where the height of a stack is the sum of the heights of each box.

pg 110

### SOLUTION

To tackle this problem, we need to recognize the relationship between the different sub-problems.

Imagine we had the following boxes:  $b_1, b_2, \dots, b_n$ . The biggest stack that we can build with all the boxes equals the max of (biggest stack with bottom  $b_1$ , biggest stack with bottom  $b_2, \dots$ , biggest stack with bottom  $b_n$ ). That is, if we experimented with each box as a bottom and built the biggest stack possible with each, we would find the biggest stack possible.

But, how would we find the biggest stack with a particular bottom? Essentially the same way. We experiment with different boxes for the second level, and so on for each level

Of course, we only experiment with valid boxes. If  $b_s$  is bigger than  $b_1$ , then there's no point in trying to build a stack that looks like  $\{b_1, b_s, \dots\}$ . We already know  $b_1$  can't be below  $b_s$ .

The code below implements this algorithm recursively.

```

1  public ArrayList<Box> createStackR(Box[] boxes, Box bottom) {
2      int max_height = 0;
3      ArrayList<Box> max_stack = null;
4      for (int i = 0; i < boxes.length; i++) {
5          if (boxes[i].canBeAbove(bottom)) {
6              ArrayList<Box> new_stack = createStackR(boxes, boxes[i]);
7              int new_height = stackHeight(new_stack);
8              if (new_height > max_height) {
9                  max_stack = new_stack;
10                 max_height = new_height;
11             }
12         }
13     }
14
15    if (max_stack == null) {
16        max_stack = new ArrayList<Box>();
17    }

```

```

18     if (bottom != null) {
19         max_stack.add(0, bottom); // Insert in bottom of stack
20     }
21
22     return max_stack;
23 }

```

The problem in this code is that it gets very inefficient. We try to find the best solution that looks like { $b_3$ ,  $b_4$ , ...} even though we may have already found the best solution with  $b_4$  at the bottom. Instead of generating these solutions from scratch, we can cache these results using dynamic programming.

```

1  public ArrayList<Box> createStackDP(Box[] boxes, Box bottom,
2      HashMap<Box, ArrayList<Box>> stack_map) {
3      if (bottom != null && stack_map.containsKey(bottom)) {
4          return stack_map.get(bottom);
5      }
6
7      int max_height = 0;
8      ArrayList<Box> max_stack = null;
9      for (int i = 0; i < boxes.length; i++) {
10         if (boxes[i].canBeAbove(bottom)) {
11             ArrayList<Box> new_stack =
12                 createStackDP(boxes, boxes[i], stack_map);
13             int new_height = stackHeight(new_stack);
14             if (new_height > max_height) {
15                 max_stack = new_stack;
16                 max_height = new_height;
17             }
18         }
19     }
20
21     if (max_stack == null) max_stack = new ArrayList<Box>();
22     if (bottom != null) max_stack.add(0, bottom);
23     stack_map.put(bottom, max_stack);
24
25     return (ArrayList<Box>)max_stack.clone();
26 }

```

You might ask why, in line 25, we have to cast `max_stack.clone()`. Isn't `max_stack` already of the correct data type? Yes, but we still need to cast.

The `clone()` method originally comes from the `Object` class and has the signature:

```
1  protected Object clone() { ... }
```

When we override a method, we can change the parameters, but we cannot change the return type. Therefore, if `Foo` were to inherit from `Object` and override `clone()`, its `clone()` method will still return an instance of type `Object`.

This is precisely what happens with the statement `(ArrayList<Box>)max_stack.clone()`. The `stack` class overrides `clone()`, but the method still must return an

Object. So, we must cast its return value.

- 9.11** Given a boolean expression consisting of the symbols 0, 1, &, |, and ^, and a desired boolean result value `result`, implement a function to count the number of ways of parenthesizing the expression such that it evaluates to `result`.

pg 110

### SOLUTION

As in other recursive problems, the key to this problem is to figure out the relationship between a problem and its sub-problems.

Suppose `int f(expression, result)` is a function which returns the count of all valid expressions which evaluate to `result`. We want to compute `f(1^0|0|1, true)` (that is, all ways of parenthesizing the expression `1^0|0|1` such that the expression evaluates to `true`). Each parenthesized expression must have an outermost pair of parentheses. So, we can say that:

$$\begin{aligned}f(1^0|0|1, \text{true}) = & f(1 \wedge (0|0|1), \text{true}) + \\& f((1^0) \mid (0|1), \text{true}) + \\& f((1^0|0) \mid 1, \text{true})\end{aligned}$$

That is, we can iterate through the expression, treating each operator as the first operator to be parenthesized.

Now, how do we compute one of these inner expressions, like `f((1^0) \mid (0|1), true)`? Well, in order for that expression to evaluate to `true`, either the left half or the right half must evaluate to `true`. So, the expression breaks down as:

$$\begin{aligned}f((1^0) \mid (0|1), \text{true}) = & f(1^0, \text{true}) * f(0|1, \text{true}) + \\& f(1^0, \text{false}) * f(0|1, \text{true}) + \\& f(1^0, \text{true}) * f(0|1, \text{false})\end{aligned}$$

We can implement a similar break down for each of the boolean operators:

$$\begin{aligned}f(\text{exp1} \mid \text{exp2}, \text{true}) &= f(\text{exp1}, \text{true}) * f(\text{exp2}, \text{true}) + \\&\quad f(\text{exp1}, \text{true}) * f(\text{exp2}, \text{false}) + \\&\quad f(\text{exp1}, \text{false}) * f(\text{exp2}, \text{true}) \\f(\text{exp1} \& \text{exp2}, \text{true}) &= f(\text{exp1}, \text{true}) * f(\text{exp2}, \text{true}) \\f(\text{exp1} \wedge \text{exp2}, \text{true}) &= f(\text{exp1}, \text{true}) * f(\text{exp2}, \text{false}) + \\&\quad f(\text{exp1}, \text{false}) * f(\text{exp2}, \text{true})\end{aligned}$$

For the `false` results, we can perform a very similar operation:

$$\begin{aligned}f(\text{exp1} \mid \text{exp2}, \text{false}) &= f(\text{exp1}, \text{false}) * f(\text{exp2}, \text{false}) \\f(\text{exp1} \& \text{exp2}, \text{false}) &= f(\text{exp1}, \text{false}) * f(\text{exp2}, \text{false}) + \\&\quad f(\text{exp1}, \text{true}) * f(\text{exp2}, \text{false}) + \\&\quad f(\text{exp1}, \text{false}) * f(\text{exp2}, \text{true}) \\f(\text{exp1} \wedge \text{exp2}, \text{false}) &= f(\text{exp1}, \text{true}) * f(\text{exp2}, \text{true}) + \\&\quad f(\text{exp1}, \text{false}) * f(\text{exp2}, \text{false})\end{aligned}$$

Implementing this is now just a matter of applying this recurrence relation. (Note: to

keep the lines from wrapping unnecessarily and making the code very confusing, we've implemented this with extra short variable names.)

```

1  public int f(String exp, boolean result, int s, int e) {
2      if (s == e) {
3          if (exp.charAt(s) == '1' && result) {
4              return 1;
5          } else if (exp.charAt(s) == '0' && !result) {
6              return 1;
7          }
8      }
9      return 0;
10 }
11 int c = 0;
12 if (result) {
13     for (int i = s + 1; i <= e; i += 2) {
14         char op = exp.charAt(i);
15         if (op == '&') {
16             c += f(exp, true, s, i - 1) * f(exp, true, i + 1, e);
17         } else if (op == '|') {
18             c += f(exp, true, s, i - 1) * f(exp, false, i + 1, e);
19             c += f(exp, false, s, i - 1) * f(exp, true, i + 1, e);
20             c += f(exp, true, s, i - 1) * f(exp, true, i + 1, e);
21         } else if (op == '^') {
22             c += f(exp, true, s, i - 1) * f(exp, false, i + 1, e);
23             c += f(exp, false, s, i - 1) * f(exp, true, i + 1, e);
24         }
25     }
26 } else {
27     for (int i = s + 1; i <= e; i += 2) {
28         char op = exp.charAt(i);
29         if (op == '&') {
30             c += f(exp, false, s, i - 1) * f(exp, true, i + 1, e);
31             c += f(exp, true, s, i - 1) * f(exp, false, i + 1, e);
32             c += f(exp, false, s, i - 1) * f(exp, false, i + 1, e);
33         } else if (op == '|') {
34             c += f(exp, false, s, i - 1) * f(exp, false, i + 1, e);
35         } else if (op == '^') {
36             c += f(exp, true, s, i - 1) * f(exp, true, i + 1, e);
37             c += f(exp, false, s, i - 1) * f(exp, false, i + 1, e);
38         }
39     }
40 }
41 return c;
42 }
```

Although this works, it's not very efficient. This method will recompute  $f(\text{exp})$  many times, for the same value of  $\text{exp}$ .

To solve this issue, we can use dynamic programming and cache the results of the different expressions. Note that we need to cache based on the expression and the result.

```

1  public int f(String exp, boolean result, int s, int e,
2              HashMap<String, Integer> q) {
3      String key = "" + result + s + e;
4      if (q.containsKey(key)) {
5          return q.get(key);
6      }
7
8      if (s == e) {
9          if (exp.charAt(s) == '1' && result == true) {
10             return 1;
11         } else if (exp.charAt(s) == '0' && result == false) {
12             return 1;
13         }
14         return 0;
15     }
16     int c = 0;
17     if (result) {
18         for (int i = s + 1; i <= e; i += 2) {
19             char op = exp.charAt(i);
20             if (op == '&') {
21                 c += f(exp,true,s,i-1,q) * f(exp,true,i+1,e,q);
22             } else if (op == '|') {
23                 c += f(exp,true,s,i-1,q) * f(exp,false,i+1,e,q);
24                 c += f(exp,false,s,i-1,q) * f(exp,true,i+1,e,q);
25                 c += f(exp,true,s,i-1,q) * f(exp,true,i+1,e,q);
26             } else if (op == '^') {
27                 c += f(exp,true,s,i-1,q) * f(exp,false,i+1,e,q);
28                 c += f(exp,false,s,i-1,q) * f(exp,true,i+1,e,q);
29             }
30         }
31     } else {
32         for (int i = s + 1; i <= e; i += 2) {
33             char op = exp.charAt(i);
34             if (op == '&') {
35                 c += f(exp,false,s,i-1,q) * f(exp,true,i+1,e,q);
36                 c += f(exp,true,s,i-1,q) * f(exp,false,i+1,e,q);
37                 c += f(exp,false,s,i-1,q) * f(exp,false,i+1,e,q);
38             } else if (op == '|') {
39                 c += f(exp,false,s,i-1,q) * f(exp,false,i+1,e,q);
40             } else if (op == '^') {
41                 c += f(exp,true,s,i-1,q) * f(exp,true,i+1,e,q);
42                 c += f(exp,false,s,i-1,q) * f(exp,false,i+1,e,q);
43             }
44         }
45     }
46     q.put(key, c);
47     return c;
48 }

```

While this is nicely optimized with dynamic programming, it's still not as optimized as it

could be. If we knew how many ways there were of parenthesizing an expression, then we could compute  $f(\text{exp} = \text{false})$  by doing  $\text{total}(\text{exp}) - f(\text{exp} = \text{true})$ .

There is a closed form expression for the number of ways of parenthesizing an expression, but you wouldn't be expected to know it. It is given by the Catalan numbers, where  $n$  is the number of operators:

$$C_n = \frac{(2n)!}{(n+1)!n!}$$

With this adjustment, the code looks like the following.

```

1  public int f(String exp, boolean result, int s, int e,
2      HashMap<String, Integer> q) {
3      String key = "" + s + e;
4      int c = 0;
5      if (!q.containsKey(key)) {
6          if (s == e) {
7              if (exp.charAt(s) == '1') c = 1;
8              else c = 0;
9          }
10
11         for (int i = s + 1; i <= e; i += 2) {
12             char op = exp.charAt(i);
13             if (op == '&') {
14                 c += f(exp,true,s,i-1,q) * f(exp,true,i+1,e,q);
15             } else if (op == '|') {
16                 int left_ops = (i-1-s)/2; // parens on left
17                 int right_ops = (e - i - 1) / 2; // parens on right
18                 int total_ways = total(left_ops) * total(right_ops);
19                 int total_false = f(exp,false,s,i-1,q) *
20                     f(exp,false,i+1,e,q);
21                 c += total_ways - total_false;
22             } else if (op == '^') {
23                 c += f(exp,true,s,i-1,q) * f(exp,false,i+1,e,q);
24                 c += f(exp,false,s,i-1,q) * f(exp,true,i+1,e,q);
25             }
26         }
27         q.put(key, c);
28     } else {
29         c = q.get(key);
30     }
31     if (result) {
32         return c;
33     } else {
34         int num_ops = (e - s) / 2;
35         return total(num_ops) - c;
36     }
37 }
```





## Scalability and Memory Limits

*Concepts and Algorithms: Solutions*

**Chapter 10**

- 10.1** Imagine you are building some sort of service that will be called by up to 1000 client applications to get simple end-of-day stock price information (open, close, high, low). You may assume that you already have the data, and you can store it in any format you wish. How would you design the client-facing service which provides the information to client applications? You are responsible for the development, rollout, and ongoing monitoring and maintenance of the feed. Describe the different methods you considered and why you would recommend your approach. Your service can use any technologies you wish, and can distribute the information to the client applications in any mechanism you choose.

pg 115

### SOLUTION

From the statement of the problem, we want to focus on how we actually distribute the information to clients. We can assume that we have some scripts that magically collect the information.

We want to start off by thinking about what the different aspects we should consider in a given proposal are:

- *Client Ease of Use:* We want the service to be easy for the clients to implement and useful for them.
- *Ease for Ourselves:* This service should be as easy as possible for us to implement, as we shouldn't impose unnecessary work on us. We need to consider in this not only the cost of implementing, but also the cost of maintenance.
- *Flexibility for Future Demands:* This problem is stated in a "what would you do in the real world" way, so we should think like we would in a real-world problem. Ideally, we do not want to overly constrain ourselves in the implementation, such that we can't be flexible if the requirements or demands change.
- *Scalability and Efficiency:* We should be mindful of the efficiency of our solution, so as not to overly burden our service.

With this framework in mind, we can consider various proposals.

#### Proposal #1

One option is that we could keep the data in simple text files and let clients download the data through some sort of FTP server. This would be easy to maintain in some sense, since files can be easily viewed and backed up, but it would require more complex parsing to do any sort of query. And, if additional data were added to our text file, it might break the clients' parsing mechanism.

#### Proposal #2

We could use a standard SQL database, and let the clients plug directly into that. This

would provide the following benefits:

- Facilitates an easy way for the clients to do query processing over the data, in case there are additional features we need to support. For example, we could easily and efficiently perform a query such as “return all stocks having an open price greater than N and a closing price less than M.”
- Rolling back, backing up data, and security could be provided using standard database features. We don’t have to “reinvent the wheel,” so it’s easy for us to implement.
- Reasonably easy for the clients to integrate into existing applications. SQL integration is a standard feature in software development environments.

What are the disadvantages of using a SQL database?

- It’s much heavier weight than we really need. We don’t necessarily need all the complexity of a SQL backend to support a feed of a few bits of information.
- It’s difficult for humans to be able to read it, so we’ll likely need to implement an additional layer to view and maintain the data. This increases our implementation costs.
- Security: While a SQL database offers pretty well-defined security levels, we would still need to be very careful to not give clients access that they shouldn’t have. Additionally, even if clients aren’t doing anything “malicious,” they might perform expensive and inefficient queries, and our servers would bear the costs of that.

These disadvantages don’t mean that we shouldn’t provide SQL access. Rather, they mean that we should be aware of the disadvantages.

### Proposal #3

XML is another great option for distributing the information. Our data has fixed format and fixed size: company\_name, open, high, low, closing price. The XML could look like this:

```
1  <root>
2    <date value="2008-10-12">
3      <company name="foo">
4        <open>126.23</open>
5        <high>130.27</high>
6        <low>122.83</low>
7        <closingPrice>127.30</closingPrice>
8      </company>
9      <company name="bar">
10     <open>52.73</open>
11     <high>60.27</high>
12     <low>50.29</low>
13     <closingPrice>54.91</closingPrice>
14   </company>
15 </date>
```

```
16    <date value="2008-10-11"> . . . </date>
17 </root>
```

The advantages of this approach include the following:

- It's very easy to distribute, and it can also be easily read by both machines and humans. This is one reason that XML is a standard data model to share and distribute data.
- Most languages have a library to perform XML parsing, so it's reasonably easy for clients to implement.
- We can add new data to the XML file by adding additional nodes. This would not break the client's parser (provided they have implemented their parser in a reasonable way).
- Since the data is being stored as XML files, we can use existing tools for backing up the data. We don't need to implement our own backup tool.

The disadvantages may include:

- This solution sends the clients all the information, even if they only want part of it. It is inefficient in that way.
- Performing any queries on the data requires parsing the entire file.

Regardless of which solution we use for data storage, we could provide a web service (e.g., SOAP) for client data access. This adds a layer to our work, but it can provide additional security, and it may even make it easier for clients to integrate the system.

However—and this is a pro and a con—clients will be limited to grabbing the data only how we expect or want them to. By contrast, in a pure SQL implementation, clients could query for the highest stock price, even if this wasn't a procedure we "expected" them to need.

So which one of these would we use? There's no clear answer. The pure text file solution is probably a bad choice, but you can make a compelling argument for the SQL or XML solution, with or without a web service.

The goal of a question like this is not to see if you get the "correct" answer (there is no single correct answer). Rather, it's to see how you design a system, and how you evaluate trade-offs.

**10.2** *How would you design the data structures for a very large social network like Facebook or LinkedIn? Describe how you would design an algorithm to show the connection, or path, between two people (e.g., Me -> Bob -> Susan -> Jason -> You).*

pg 115

### SOLUTION

A good way to approach this problem is to remove some of the constraints and solve it for that situation first.

### Step 1: Simplify the Problem—Forget About the Millions of Users

First, let's forget that we're dealing with millions of users. Design this for the simple case.

We can construct a graph by treating each person as a node and letting an edge between two nodes indicate that the two users are friends.

```
1 class Person {
2     Person[] friends;
3     // Other info
4 }
```

If I wanted to find the connection between two people, I would start with one person and do a simple breadth first search.

Why wouldn't a depth first search work well? Because it would be very inefficient. Two users might be only one degree of separation apart, but I could search millions of nodes in their "subtrees" before finding this relatively immediate connection.

### Step 2: Handle the Millions of Users

When we deal with a service the size of LinkedIn or Facebook, we cannot possibly keep all of our data on one machine. That means that our simple Person data structure from above doesn't quite work—our friends may not live on the same machine as we do. Instead, we can replace our list of friends with a list of their IDs, and traverse as follows:

1. For each friend ID: int machine\_index = getMachineIDForUser(personID);
2. Go to machine #machine\_index
3. On that machine, do: Person friend = getPersonWithID(person\_id);

The code below outlines this process. We've defined a class Server, which holds a list of all the machines, and a class Machine which represents a single machine. Both classes have hash tables to efficiently lookup data.

```
1 public class Server {
2     HashMap<Integer, Machine> machines =
3         new HashMap<Integer, Machine>();
4     HashMap<Integer, Integer> personToMachineMap =
5         new HashMap<Integer, Integer>();
6
7     public Machine getMachineWithId(int machineID) {
8         return machines.get(machineID);
9     }
10
11    public int getMachineIDForUser(int personID) {
12        Integer machineID = personToMachineMap.get(personID);
13        return machineID == null ? -1 : machineID;
14    }
15}
```

```
14    }
15
16    public Person getPersonWithID(int personID) {
17        Integer machineID = personToMachineMap.get(personID);
18        if (machineID == null) return null;
19
20        Machine machine = getMachineWithId(machineID);
21        if (machine == null) return null;
22
23        return machine.getPersonWithID(personID);
24    }
25 }
26
27 public class Person {
28     private ArrayList<Integer> friendIDs;
29     private int personID;
30
31     public Person(int id) { this.personID = id; }
32
33     public int getID() { return personID; }
34     public void addFriend(int id) { friends.add(id); }
35 }
36
37 public class Machine {
38     public HashMap<Integer, Person> persons =
39         new HashMap<Integer, Person>();
40     public int machineID;
41
42     public Person getPersonWithID(int personID) {
43         return persons.get(personID);
44     }
45 }
```

There are more optimizations and follow up questions here than we could possibly discuss, but here are just a few thoughts.

### Optimization: Reduce Machine Jumps

Jumping from one machine to another is expensive. Instead of randomly jumping from machine to machine with each friend, try to batch these jumps—e.g., if five of my friends live on one machine, I should look them up all at once.

### Optimization: Smart Division of People and Machines

People are much more likely to be friends with people who live in the same country as they do. Rather than randomly dividing people up across machines, try to divide them up by country, city, state, and so on. This will reduce the number of jumps.

**Question: Breadth First Search usually requires “marking” a node as visited. How**

**do you do that in this case?**

Usually, in BFS, we mark a node as visited by setting a flag `visited` in its node class. Here, we don't want to do that. There could be multiple searches going on at the same time, so it's bad to just edit our data.

Instead, we could mimic the marking of nodes with a hash table to look up a node id and whether or not it's been visited.

**Other Follow-Up Questions:**

- In the real world, servers fail. How does this affect you?
- How could you take advantage of caching?
- Do you search until the end of the graph (infinite)? How do you decide when to give up?
- In real life, some people have more friends of friends than others, and are therefore more likely to make a path between you and someone else. How could you use this data to pick where you start traversing?

These are just a few of the follow-up questions you or the interviewer could raise. There are many others.

**10.3** Given an input file with four billion non-negative integers, provide an algorithm to generate an integer which is not contained in the file. Assume you have 1 GB of memory available for this task.

**FOLLOW UP**

What if you have only 10 MB of memory? Assume that all the values are distinct and we now have no more than one billion non-negative integers.

pg 115

**SOLUTION**

There are a total of  $2^{32}$ , or 4 billion, distinct integers possible (and  $2^{31}$  non-negative integers). We have 1 GB of memory, or 8 billion bits.

Thus, with 8 billion bits, we can map all possible integers to a distinct bit with the available memory. The logic is as follows:

1. Create a bit vector (BV) with 4 billion bits. Recall that a bit vector is an array that compactly stores boolean values by using an array of ints (or another data type). Each int stores a sequence of 32 bits, or boolean values.
2. Initialize BV with all 0's.
3. Scan all numbers (num) from the file and call `BV.set(num, 1)`.

4. Now scan again BV from the 0th index.
5. Return the first index which has a value of 0.

The following code demonstrates our algorithm.

```
1 long numberOInts = ((long) Integer.MAX_VALUE) + 1;
2 byte[] bitfield = new byte [(int) (numberOInts / 8)];
3 void findOpenNumber() throws FileNotFoundException {
4     Scanner in = new Scanner(new FileReader("file.txt"));
5     while (in.hasNextInt()) {
6         int n = in.nextInt ();
7         /* Finds the corresponding number in the bitfield by using
8          * the OR operator to set the nth bit of a byte
9          * (e.g., 10 would correspond to the 2nd bit of index 2 in
10         * the byte array). */
11         bitfield [n / 8] |= 1 << (n % 8);
12     }
13
14     for (int i = 0; i < bitfield.length; i++) {
15         for (int j = 0; j < 8; j++) {
16             /* Retrieves the individual bits of each byte. When 0 bit
17              * is found, finds the corresponding value. */
18             if ((bitfield[i] & (1 << j)) == 0) {
19                 System.out.println (i * 8 + j);
20                 return;
21             }
22         }
23     }
24 }
```

### Follow Up: What if we have only 10 MB memory?

It's possible to find a missing integer with two passes of the data set. We can divide up the integers into blocks of some size (we'll discuss how to decide on a size later). Let's just assume that we divide up the integers into blocks of 1000. So, block 0 represents the numbers 0 through 999, block 1 represents numbers 1000 - 1999, and so on.

Since all the values are distinct, we know how many values we *should* find in each block. So, we search through the file and count how many values are between 0 and 999, how many are between 1000 and 1999, and so on. If we count only 999 values in a particular range, then we know that a missing int must be in that range.

In the second pass, we'll actually look for which number in that range is missing. We use the bit vector approach from the first part of this problem. We can ignore any number outside of this specific range.

The question, now, is what is the appropriate block size? Let's define some variables as follows:

- Let `rangeSize` be the size of the ranges that each block in the first pass represents.

- Let `arraySize` represent the number of blocks in the first pass. Note that  $\text{arraySize} = 2^{31} / \text{rangeSize}$ , since there are  $2^{31}$  non-negative integers.

We need to select a value for `rangeSize` such that the memory from the first pass (the array) and the second pass (the bit vector) fit.

#### *First Pass: The Array*

The array in the first pass can fit in 10 megabytes, or roughly  $2^{23}$  bytes, of memory. Since each element in the array is an int, and an int is 4 bytes, we can hold an array of at most about  $2^{21}$  elements. So, we can deduce the following:

$$\begin{aligned}\text{arraySize} &= \frac{2^{31}}{\text{rangeSize}} \leq 2^{21} \\ \text{rangeSize} &\geq \frac{2^{31}}{2^{21}} \\ \text{rangeSize} &\geq 2^{10}\end{aligned}$$

#### *Second Pass: The Bit Vector*

We need to have enough space to store `rangeSize` bits. Since we can fit  $2^{23}$  bytes in memory, we can fit  $2^{26}$  bits in memory. Therefore, we can conclude the following:

$$2^{10} \leq \text{rangeSize} \leq 2^{26}$$

These conditions give us a good amount of “wiggle room,” but the nearer to the middle that we pick, the less memory will be used at any given time.

The below code provides one implementation for this algorithm.

```

1  int bitsize = 1048576; // 2^20 bits (2^17 bytes)
2  int blockNum = 4096; // 2^12
3  byte[] bitfield = new byte[bitsize/8];
4  int[] blocks = new int[blockNum];
5
6  void findOpenNumber() throws FileNotFoundException {
7      int starting = -1;
8      Scanner in = new Scanner (new FileReader ("file.txt"));
9      while (in.hasNextInt()) {
10          int n = in.nextInt();
11          blocks[n / (bitfield.length * 8)]++;
12      }
13
14     for (int i = 0; i < blocks.length; i++) {
15         if (blocks[i] < bitfield.length * 8){
16             /* if value < 2^20, then at least 1 number is missing in
17              * that section. */
18             starting = i * bitfield.length * 8;
19             break;
20         }
21     }
22 }
```

```
23     in = new Scanner(new FileReader("file.txt"));
24     while (in.hasNextInt()) {
25         int n = in.nextInt();
26         /* If the number is inside the block that's missing
27          * numbers, we record it */
28         if (n >= starting && n < starting + bitfield.length * 8) {
29             bitfield [(n-starting) / 8] |= 1 << ((n - starting) % 8);
30         }
31     }
32
33     for (int i = 0 ; i < bitfield.length; i++) {
34         for (int j = 0; j < 8; j++) {
35             /* Retrieves the individual bits of each byte. When 0 bit
36              * is found, finds the corresponding value. */
37             if ((bitfield[i] & (1 << j)) == 0) {
38                 System.out.println(i * 8 + j + starting);
39                 return;
40             }
41         }
42     }
43 }
```

What if, as a potential follow up question, the interviewer asked you to solve the problem with even less memory? In this case we would do repeated passes using the approach from the first step. We'd first check to see how many integers are found within each sequence of a million elements. Then, in the second pass, we'd check how many integers are found in each sequence of a thousand elements. Finally, in the third pass, we'd apply the bit vector.

- 10.4** You have an array with all the numbers from 1 to N, where N is at most 32,000. The array may have duplicate entries and you do not know what N is. With only 4 kilobytes of memory available, how would you print all duplicate elements in the array?

pg 116

### SOLUTION

We have 4 kilobytes of memory which means we can address up to  $8 * 4 * 2^{10}$  bits. Note that  $32 * 2^{10}$  bits is greater than 32000. We can create a bit vector with 32000 bits, where each bit represents one integer.

Using this bit vector, we can then iterate through the array, flagging each element v by setting bit v to 1. When we come across a duplicate element, we print it.

```
1  public static void checkDuplicates(int[] array) {
2      BitSet bs = new BitSet(32000);
3      for (int i = 0; i < array.length; i++) {
4          int num = array[i];
5          int num0 = num - 1; // bitset starts at 0, numbers start at 1
6          if (bs.get(num0)) {
```

```

7         System.out.println(num);
8     } else {
9         bs.set(num0);
10    }
11 }
12 }
13
14 class BitSet {
15     int[] bitset;
16
17     public BitSet(int size) {
18         bitset = new int[size >> 5]; // divide by 32
19     }
20
21     boolean get(int pos) {
22         int wordNumber = (pos >> 5); // divide by 32
23         int bitNumber = (pos & 0x1F); // mod 32
24         return (bitset[wordNumber] & (1 << bitNumber)) != 0;
25     }
26
27     void set(int pos) {
28         int wordNumber = (pos >> 5); // divide by 32
29         int bitNumber = (pos & 0x1F); // mod 32
30         bitset[wordNumber] |= 1 << bitNumber;
31     }
32 }
```

Note that while this isn't an especially difficult problem, it's important to implement this cleanly. This is why we defined our own bit vector class to hold a large bit vector. If our interviewer lets us (she may or may not), we could have of course used Java's built in `BitSet` class.

#### 10.5 If you were designing a web crawler, how would you avoid getting into infinite loops?

pg 116

#### SOLUTION

The first thing to ask ourselves in this problem is how an infinite loop might occur. The simplest answer is that, if we picture the web as a graph of links, an infinite loop will occur when a cycle occurs.

To prevent infinite loops, we just need to detect cycles. One way to do this is to create a hash table where we set `hash[v]` to true after we visit page v.

This solution would mean crawling the web using breadth first search. Each time we visit a page, we gather all its links and insert them at the end of a queue. If we've already visited a page, we ignore it.

This is great—but what does it mean to visit page v? Is page v defined based on its

content or its URL?

If it's defined based on its URL, we must recognize that URL parameters might indicate a completely different page. For example, the page `www.careercup.com/page?id=microsoft-interview-questions` is totally different from the page `www.careercup.com/page?id=google-interview-questions`. But, we can also append URL parameters arbitrarily to any URL without truly changing the page, provided it's not a parameter that the web application recognizes and handles. The page `www.careercup.com?foobar=hello` is the same as `www.careercup.com`.

"Okay, then," you might say, "let's define it based on its content." That sounds good too, at first, but it also doesn't quite work. Suppose I have some randomly generated content on the `careercup.com` home page. Is it a different page each time you visit it? Not really.

The reality is that there is probably no perfect way to define a "different" page, and this is where this problem gets tricky.

One way to tackle this is to have some sort of estimation for degree of similarity. If, based on the content and the URL, a page is deemed to be sufficiently similar to other pages, we *deprioritize* crawling its children. For each page, we would come up with some sort of signature based on snippets of the content and the page's URL.

Let's see how this would work.

We have a database which stores a list of items we need to crawl. On each iteration, we select the highest priority page to crawl. We then do the following:

1. Open up the page and create a signature of the page based on specific subsections of the page and its URL.
2. Query the database to see whether anything with this signature has been crawled recently.
3. If something with this signature has been recently crawled, insert this page back into the database at a low priority.
4. If not, crawl the page and insert its links into the database.

Under the above implementation, we never "complete" crawling the web, but we will avoid getting stuck in a loop of pages. If we want to allow for the possibility of "finishing" crawling the web (which would clearly happen only if the "web" were actually a smaller system, like an intranet), then we can set a minimum priority that a page must have to be crawled.

This is just one, simplistic solution, and there are many others that are equally valid. A problem like this will more likely resemble a conversation with your interviewer which could take any number of paths. In fact, the discussion of this problem could have taken the path of the very next problem.

- 10.6** You have 10 billion URLs. How do you detect the duplicate documents? In this case, assume that “duplicate” means that the URLs are identical.

pg 116

### SOLUTION

Just how much space do 10 billion URLs take up? If each URL is an average of 100 characters, and each character is 4 bytes, then this list of 10 billion URLs will take up about 4 terabytes. We are probably not going to hold that much data in memory.

But, let's just pretend for a moment that we were miraculously holding this data in memory, since it's useful to first construct a solution for the simple version. Under this version of the problem, we would just create a hash table where each URL maps to true if it's already been found elsewhere in the list. (As an alternative solution, we could sort the list and look for the duplicate values that way. That will take a bunch of extra time and offers few advantages.)

Now that we have a solution for the simple version, what happens when we have all 4000 gigabytes of data and we can't store it all in memory? We could solve this either by storing some of the data on disk or by splitting up the data across machines.

#### Solution #1: Disk Storage

If we stored all the data on one machine, we would do two passes of the document. The first pass would split the list of URLs into 4000 chunks of 1 GB each. An easy way to do that might be to store each URL  $u$  in a file named  $\langle x \rangle.txt$  where  $x = \text{hash}(u) \% 4000$ . That is, we divide up the URLs based on their hash value (modulo the number of chunks). This way, all URLs with the same hash value would be in the same file.

In the second pass, we would essentially implement the simple solution we came up with earlier: load each file into memory, create a hash table of the URLs, and look for duplicates.

#### Solution #2: Multiple Machines

The other solution is to perform essentially the same procedure, but to use multiple machines. In this solution, rather than storing the data in file  $\langle x \rangle.txt$ , we would send the URL to machine  $x$ .

Using multiple machines has pros and cons.

The main pro is that we can parallelize the operation, such that all 4000 chunks are processed simultaneously. For large amounts of data, this might result in a faster solution.

The disadvantage though is that we are now relying on 4000 different machines to operate perfectly. That may not be realistic (particularly with more data and more

machines), and we'll need to start considering how to handle failure. Additionally, we have increased the complexity of the system simply by involving so many machines.

Both are good solutions though, and both should be discussed with your interviewer.

**10.7** *Imagine a web server for a simplified search engine. This system has 100 machines to respond to search queries, which may then call out using processSearch(string query) to another cluster of machines to actually get the result. The machine which responds to a given query is chosen at random, so you can not guarantee that the same machine will always respond to the same request. The method processSearch is very expensive. Design a caching mechanism to cache the results of the most recent queries. Be sure to explain how you would update the cache when data changes.*

pg 116

### SOLUTION

Before getting into the design of this system, we first have to understand what the question means. Many of the details are somewhat ambiguous, as is expected in questions like this. We will make reasonable assumptions for the purposes of this solution, but you should discuss these details—in depth—with your interviewer.

#### Assumptions

Here are a few of the assumptions we make for this solution. Depending on the design of your system and how you approach the problem, you may make other assumptions. Remember that while some approaches are better than others, there is no one "correct" approach.

- Other than calling out to processSearch as necessary, all query processing happens on the initial machine that was called.
- The number of queries we wish to cache is large (millions).
- Calling between machines is relatively quick.
- The result for a given query is an ordered list of URLs, each of which has an associated 50 character title and 200 character summary.
- The most popular queries are extremely popular, such that they would always appear in the cache.

Again, these aren't the *only* valid assumptions. This is just one reasonable set of assumptions.

#### System Requirements

When designing the cache, we know we'll need to support two primary functions:

- Efficient lookups given a key.
- Expiration of old data so that it can be replaced with new data.

In addition, we must also handle updating or clearing the cache when the results for a query change. Because some queries are very common and may permanently reside in the cache, we cannot just wait for the cache to naturally expire.

### Step 1: Design a Cache for a Single System

A good way to approach this problem is to start by designing it for a single machine. So, how would you create a data structure that enables you to easily purge old data and also efficiently look up a value based on a key?

- A linked list would allow easy purging of old data, by moving “fresh” items to the front. We could implement it to remove the last element of the linked list when the list exceeds a certain size.
- A hash table allows efficient lookups of data, but it wouldn’t ordinarily allow easy data purging.

How can we get the best of both worlds? By merging the two data structures. Here’s how this works:

Just as before, we create a linked list where a node is moved to the front every time it’s accessed. This way, the end of the linked list will always contain the stalest information.

In addition, we have a hash table that maps from a query to the corresponding node in the linked list. This allows us to not only efficiently return the cached results, but also to move the appropriate node to the front of the list, thereby updating its “freshness.”

For illustrative purposes, abbreviated code for the cache is below. The code attachment provides the full code for this part. Note that in your interview, it is unlikely that you would be asked to write the full code for this as well as perform the design for the larger system.

```
1  public class Cache {  
2      public static int MAX_SIZE = 10;  
3      public Node head, tail;  
4      public HashMap<String, Node> map;  
5      public int size = 0;  
6  
7      public Cache() {  
8          map = new HashMap<String, Node>();  
9      }  
10     ,  
11     /* Moves node to front of linked list */  
12     public void moveToFront(Node node) { ... }  
13     public void moveToFront(String query) { ... }  
14  
15     /* Removes node from linked list */
```

```
16     public void removeFromLinkedList(Node node) { ... }
17
18     /* Gets results from cache, and updates linked list */
19     public String[] getResults(String query) {
20         if (!map.containsKey(query)) return null;
21
22         Node node = map.get(query);
23         moveToFront(node); // update freshness
24         return node.results;
25     }
26
27     /* Inserts results into linked list and hash */
28     public void insertResults(String query, String[] results) {
29         if (map.containsKey(query)) { // update values
30             Node node = map.get(query);
31             node.results = results;
32             moveToFront(node); // update freshness
33             return;
34         }
35
36         Node node = new Node(query, results);
37         moveToFront(node);
38         map.put(query, node);
39
40         if (size > MAX_SIZE) {
41             map.remove(tail.query);
42             removeFromLinkedList(tail);
43         }
44     }
45 }
```

### Step 2: Expand to Many Machines

Now that we understand how to design this for a single machine, we need to understand how we would design this when queries could be sent to many different machines. Recall from the problem statement that there's no guarantee that a particular query will be consistently sent to the same machine.

The first thing we need to decide is to what extent the cache is shared across machines. We have several options to consider.

#### *Option 1: Each machine has its own cache.*

A simple option is to give each machine its own cache. This means that if "foo" is sent to machine 1 twice in a short amount of time, the result would be recalled from the cache on the second time. But, if "foo" is sent first to machine 1 and then to machine 2, it would be treated as a totally fresh query both times.

This has the advantage of being relatively quick, since no machine-to-machine calls are used. The cache, unfortunately, is somewhat less effective as an optimization tool as

many repeat queries would be treated as fresh queries.

*Option 2: Each machine has a copy of the cache.*

On the other extreme, we could give each machine a complete copy of the cache. When new items are added to the cache, they are sent to all machines. The entire data structure—linked list and hash table—would be duplicated.

This design means that common queries would nearly always be in the cache, as the cache is the same everywhere. The major drawback however is that updating the cache means firing off data to  $N$  different machines, where  $N$  is the size of the response cluster. Additionally, because each item effectively takes up  $N$  times as much space, our cache would hold much less data.

*Option 3: Each machine stores a segment of the cache.*

A third option is to divide up the cache, such that each machine holds a different part of it. Then, when machine  $i$  needs to look up the results for a query, machine  $i$  would figure out which machine holds this value, and then ask this other machine (machine  $j$ ) to look up the query in  $j$ 's cache.

But how would machine  $i$  know which machine holds this part of the hash table?

One option is to assign queries based on the formula  $\text{hash}(\text{query}) \% N$ . Then, machine  $i$  only needs to apply this formula to know that machine  $j$  should store the results for this query.

So, when a new query comes in to machine  $i$ , this machine would apply the formula and call out to machine  $j$ . Machine  $j$  would then return the value from its cache or call `processSearch(query)` to get the results. Machine  $j$  would update its cache and return the results back to  $i$ .

Alternatively, you could design the system such that machine  $j$  just returns null if it doesn't have the query in its current cache. This would require machine  $i$  to call `processSearch` and then forward the results to machine  $j$  for storage. This implementation actually increases the number of machine-to-machine calls, with few advantages.

### Step 3: Updating results when contents change

Recall that some queries may be so popular that, with a sufficiently large cache, they would permanently be cached. We need some sort of mechanism to allow cached results to be refreshed, either periodically or "on-demand" when certain content changes.

To answer this question, we need to consider when results would change (and you need to discuss this with your interviewer). The primary times would be when:

1. The content at a URL changes (or the page at that URL is removed).
2. The ordering of results change in response to the rank of a page changing.

### 3. New pages appear related to a particular query.

To handle situations #1 and #2, we could create a separate hash table that would tell us which cached queries are tied to a specific URL. This could be handled completely separately from the other caches, and reside on different machines. However, this solution may require a lot of data.

Alternatively, if the data doesn't require instant refreshing (which it probably doesn't), we could periodically crawl through the cache stored on each machine to purge queries tied to the updated URLs.

Situation #3 is substantially more difficult to handle. We could update single word queries by parsing the content at the new URL and purging these one-word queries from the caches. But, this will only handle the one word queries.

A good way to handle Situation #3 (and likely something we'd want to do anyway) is to implement an "automatic time-out" on the cache. That is, we'd impose a time out where *no* query, regardless of how popular it is, can sit in the cache for more than  $x$  minutes. This will ensure that all data is periodically refreshed.

### Step 4: Further Enhancements

There are a number of improvement and tweaks you could make to this design depending on the assumptions you make and the situations you optimize for.

One such optimization is to better support the situation where some queries are very popular. For example, suppose (as an extreme example) a particular string constitutes 1% of all queries. Rather than machine  $i$  forwarding the request to machine  $j$  every time, machine  $i$  could forward the request just once to  $j$ , and then  $i$  could store the results in its own cache as well.

Alternatively, there may also be some possibility of doing some sort of re-architecture of the system to assign queries to machines based on their hash value (and therefore the location of the cache), rather than randomly. However, this decision may come with its own set of trade-offs.

Another optimization we could make is to the "automatic time out" mechanism. As initially described, this mechanism purges any data after  $X$  minutes. However, we may want to update some data (like current news) much more frequently than other data (like historical stock prices). We could implement time outs based on topic or based on URLs. In the latter situation, each URL would have a time out value based on how frequently the page has been updated in the past. The time out for the query would be the minimum of the time outs for each URL.

These are just a few of the enhancements we can make. Remember that in questions like this, there is no single correct way to solve the problem. These questions are about having a discussion with your interviewer about design criteria and demonstrating your general approach and methodology.

# **Sorting and Searching**

*Concepts and Algorithms: Solutions*

**Chapter 11**

- 11.1 You are given two sorted arrays, A and B, where A has a large enough buffer at the end to hold B. Write a method to merge B into A in sorted order.

pg 121

### SOLUTION

Since we know that A has enough buffer at the end, we won't need to allocate additional space. Our logic should involve simply comparing elements of A and B and inserting them in order, until we've exhausted all elements in A and in B.

The only issue with this is that if we insert an element into the front of A, then we'll have to shift the existing elements backwards to make room for it. It's better to insert elements into the back of the array, where there's empty space.

The code below does just that. It works from the back of A and B, moving the largest elements to the back of A.

```
1  public static void merge(int[] a, int[] b, int lastA, int lastB) {  
2      int indexA = lastA - 1; /* Index of last element in array a */  
3      int indexB = lastB - 1; /* Index of last element in array b */  
4      int indexMerged = lastB + lastA - 1; /* end of merged array */  
5  
6      /* Merge a and b, starting from the last element in each */  
7      while (indexA >= 0 && indexB >= 0) {  
8          /* end of a is > than end of b */  
9          if (a[indexA] > b[indexB]) {  
10              a[indexMerged] = a[indexA]; // copy element  
11              indexMerged--; // move indices  
12              indexA--;  
13          } else {  
14              a[indexMerged] = b[indexB]; // copy element  
15              indexMerged--; // move indices  
16              indexB--;  
17          }  
18      }  
19  
20      /* Copy remaining elements from b into place */  
21      while (indexB >= 0) {  
22          a[indexMerged] = b[indexB];  
23          indexMerged--;  
24          indexB--;  
25      }  
26  }
```

Note that you don't need to copy the contents of A after running out of elements in B. They are already in place.

- 11.2** Write a method to sort an array of strings so that all the anagrams are next to each other.

pg 121

## SOLUTION

---

This problem asks us to group the strings in an array such that the anagrams appear next to each other. Note that no specific ordering of the words is required, other than this.

One way to do this is to just apply any standard sorting algorithm, like merge sort or quick sort, and modify the comparator. This comparator will be used to indicate that two strings which are anagrams of each other are equivalent.

What's the easiest way of checking if two words are anagrams? We could count the occurrences of the distinct characters in each string and return true if they match. Or, we could just sort the string. After all, two words which are anagrams will look the same once they're sorted.

The code below implements the comparator.

```

1  public class AnagramComparator implements Comparator<String> {
2      public String sortChars(String s) {
3          char[] content = s.toCharArray();
4          Arrays.sort(content);
5          return new String(content);
6      }
7
8      public int compare(String s1, String s2) {
9          return sortChars(s1).compareTo(sortChars(s2));
10     }
11 }
```

Now, just sort the arrays using this `compareTo` method instead of the usual one.

```
12 Arrays.sort(array, new AnagramComparator());
```

This algorithm will take  $O(n \log(n))$  time.

This may be the best we can do for a general sorting algorithm, but we don't actually need to fully sort the array. We only need to *group* the strings in the array by anagram.

We can do this by using a hash table which maps from the sorted version of a word to a list of its anagrams. So, for example, `acre` will map to the list `{acre, race, care}`. Once we've grouped all the words into these lists by anagram, we can then put them back into the array.

The code below implements this algorithm.

```

1  public void sort(String[] array) {
2      Hashtable<String, LinkedList<String>> hash =
3          new Hashtable<String, LinkedList<String>>();
```

```
4      /* Group words by anagram */
5      for (String s : array) {
6          String key = sortChars(s);
7          if (!hash.containsKey(key)) {
8              hash.put(key, new LinkedList<String>());
9          }
10         LinkedList<String> anagrams = hash.get(key);
11         anagrams.push(s);
12     }
13 }
14
15 /* Convert hash table to array */
16 int index = 0;
17 for (String key : hash.keySet()) {
18     LinkedList<String> list = hash.get(key);
19     for (String t : list) {
20         array[index] = t;
21         index++;
22     }
23 }
24 }
```

You may notice that the algorithm above is a modification of bucket sort.

- 11.3** Given a sorted array of  $n$  integers that has been rotated an unknown number of times, write code to find an element in the array. You may assume that the array was originally sorted in increasing order.

pg 121

### SOLUTION

If this problem smells like binary search to you, you're right!

In classic binary search, we compare  $x$  with the midpoint to figure out if  $x$  belongs on the left or the right side. The complication here is that the array is rotated and may have an inflection point. Consider, for example, the following two arrays:

```
Array1: {10, 15, 20, 0, 5}
Array2: {50, 5, 20, 30, 40}
```

Note that both arrays have a midpoint of 20, but 5 appears on the left side of one and on the right side of the other. Therefore, comparing  $x$  with the midpoint is insufficient.

However, if we look a bit deeper, we can see that one half of the array must be ordered normally (in increasing order). We can therefore look at the normally ordered half to determine whether we should search the left or right half.

For example, if we are searching for 5 in Array1, we can look at the left element (10) and middle element (20). Since  $10 < 20$ , the left half must be ordered normally. And, since 5 is not between those, we know that we must search the right half.

In Array2, we can see that since  $50 > 20$ , the right half must be ordered normally. We turn to the middle (20) and right (40) element to check if 5 would fall between them. The value 5 would not; therefore, we search the left half.

The tricky condition is if the left and the middle are identical, as in the example array {2, 2, 2, 3, 4, 2}. In this case, we can check if the rightmost element is different. If it is, we can search just the right side. Otherwise, we have no choice but to search both halves.

```

1  public int search(int a[], int left, int right, int x) {
2      int mid = (left + right) / 2;
3      if (x == a[mid]) { // Found element
4          return mid;
5      }
6      if (right < left) {
7          return -1;
8      }
9
10     /* Either the left or right half must be normally ordered. Find
11      * out which side is normally ordered, and then use the normally
12      * ordered half to figure out which side to search to find x. */
13     if (a[left] < a[mid]) { // Left is normally ordered.
14         if (x >= a[left] && x <= a[mid]) {
15             return search(a, left, mid - 1, x); // Search left
16         } else {
17             return search(a, mid + 1, right, x); // Search right
18         }
19     } else if (a[mid] < a[left]) { // Right is normally ordered.
20         if (x >= a[mid] && x <= a[right]) {
21             return search(a, mid + 1, right, x); // Search right
22         } else {
23             return search(a, left, mid - 1, x); // Search left
24         }
25     } else if (a[left] == a[mid]) { // Left half is all repeats
26         if (a[mid] != a[right]) { // If right is diff., search it
27             return search(a, mid + 1, right, x); // search right
28         } else { // Else, we have to search both halves
29             int result = search(a, left, mid - 1, x); // Search left
30             if (result == -1) {
31                 return search(a, mid + 1, right, x); // Search right
32             } else {
33                 return result;
34             }
35         }
36     }
37     return -1;
38 }
```

This code will run in  $O(\log n)$  if all the elements are unique. However, with many duplicates, the algorithm is actually  $O(n)$ . This is because with many duplicates, we will often

have to search both the left and right sides of the array (or subarrays).

Note that while this problem is not conceptually very complex, it is actually very difficult to implement flawlessly. Don't feel bad if you had trouble implementing it without a few bugs. Because of the ease of making off-by-one and other minor errors, you should make sure to test your code very thoroughly.

- 11.4** Imagine you have a 20 GB file with one string per line. Explain how you would sort the file.

pg 122

### SOLUTION

When an interviewer gives a size limit of 20 gigabytes, it should tell you something. In this case, it suggests that they don't want you to bring all the data into memory.

So what do we do? We only bring part of the data into memory.

We'll divide the file into chunks which are  $x$  megabytes each, where  $x$  is the amount of memory we have available. Each chunk is sorted separately and then saved back to the file system.

Once all the chunks are sorted, we then merge the chunks, one by one. At the end, we have a fully sorted file.

This algorithm is known as external sort.

- 11.5** Given a sorted array of strings which is interspersed with empty strings, write a method to find the location of a given string.

pg 122

### SOLUTION

If it weren't for the empty strings, we could simply use binary search. We would compare the string to be found, `str`, with the midpoint of the array, and go from there.

With empty strings interspersed, we can implement a simple modification of binary search. All we need to do is fix the comparison against `mid`, in case `mid` is an empty string. We simply move `mid` to the closest non-empty string.

The recursive code below to solve this problem can easily be modified to be iterative. We provide such an implementation in the code attachment.

```
1 public int searchR(String[] strings, String str, int first,
2                     int last) {
3     if (first > last) return -1;
4     /* Move mid to the middle */
5     int mid = (last + first) / 2;
```

```

6
7     /* If mid is empty, find closest non-empty string. */
8     if (strings[mid].isEmpty()) {
9         int left = mid - 1;
10        int right = mid + 1;
11        while (true) {
12            if (left < first && right > last) {
13                return -1;
14            } else if (right <= last && !strings[right].isEmpty()) {
15                mid = right;
16                break;
17            } else if (left >= first && !strings[left].isEmpty()) {
18                mid = left;
19                break;
20            }
21            right++;
22            left--;
23        }
24    }
25
26    /* Check for string, and recurse if necessary */
27    if (str.equals(strings[mid])) { // Found it!
28        return mid;
29    } else if (strings[mid].compareTo(str) < 0) { // Search right
30        return searchR(strings, str, mid + 1, last);
31    } else { // Search left
32        return searchR(strings, str, first, mid - 1);
33    }
34 }
35
36 public int search(String[] strings, String str) {
37     if (strings == null || str == null || str == "") {
38         return -1;
39     }
40     return searchR(strings, str, 0, strings.length - 1);
41 }

```

Careful consideration should be given to the situation when someone searches for the empty string. Should we find the location (which is an  $O(n)$  operation)? Or should we handle this as an error?

There's no correct answer here. This is an issue you should raise with your interviewer. Simply asking this question will demonstrate that you are a careful coder.

- 11.6** Given an  $M \times N$  matrix in which each row and each column is sorted in ascending order, write a method to find an element.

pg 122

## SOLUTION

As a first approach, we can do binary search on every row to find the element. This algorithm will be  $O(M \log(N))$ , since there are  $M$  rows and it takes  $O(\log(N))$  time to search each one. This is a good approach to mention to your interviewer before you proceed with generating a better algorithm.

To develop an algorithm, let's start with a simple example.

15	20	40	85
20	35	80	95
30	55	95	105
40	80	100	120

Suppose we are searching for the element 55. How can we identify where it is?

If we look at the start of a row or the start of a column, we can start to deduce the location. If the start of a column is greater than 55, we know that 55 can't be in that column, since the start of the column is always the minimum element. Additionally, we know that 55 can't be in any columns on the right, since the first element of each column must increase in size from left to right. Therefore, if the start of the column is greater than the element  $x$  that we are searching for, we know that we need to move further to the left.

For rows, we can follow identical logic. If the start of a row is bigger than  $x$ , then we know we need to move upwards.

Observe that we can also make a similar conclusion by looking at the ends of columns or rows. If the end of a column or row is less than  $x$ , then we know that we must move down (for rows) or to the right (for columns) to find  $x$ . This is because the end is always the maximum element.

We can bring these observations together into a solution. The observations are the following:

- If the start of a column is greater than  $x$ , then  $x$  is to the left of the column.
- If the end of a column is less than  $x$ , then  $x$  is to the right of the column.
- If the start of a row is greater than  $x$ , then  $x$  is above that row.
- If the end of a row is less than  $x$ , then  $x$  is below that row.

We can begin in any number of places, but let's begin with looking at the starts of columns.

We need to start with the greatest column and work our way to the left. This means that our first element for comparison is  $\text{array}[0][c-1]$ , where  $c$  is the number of columns. By comparing the start of columns to  $x$  (which is 55), we'll find that  $x$  must be in columns 0, 1, or 2. We will have stopped at  $\text{array}[0][2]$ .

This element may not be the end of a row in the full matrix, but it is an end of a row of a submatrix. The same conditions apply. The value at  $\text{array}[0][2]$ , which is 40, is less

than 55, so we know we can move downwards.

We now have a submatrix to consider that looks like the following (the gray squares have been eliminated).

15	20	40	85
20	35	80	95
30	55	95	105
40	80	100	120

We can repeatedly apply these conditions to search for 55. Note that the only conditions we actually use are conditions 1 and 4.

The code below implements this elimination algorithm.

```

1  public static boolean findElement(int[][] matrix, int elem) {
2      int row = 0;
3      int col = matrix[0].length - 1;
4      while (row < matrix.length && col >= 0) {
5          if (matrix[row][col] == elem) {
6              return true;
7          } else if (matrix[row][col] > elem) {
8              col--;
9          } else {
10              row++;
11          }
12      }
13      return false;
14 }
```

Alternatively, we can apply a solution that more directly looks like binary search. The code is considerably more complicated, but it applies many of the same learnings.

## Solution #2: Binary Search

Let's again look at a simple example.

15	20	70	85
29	35	80	95
30	55	95	105
40	80	100	120

We want to be able to leverage the sorting property to more efficiently find an element. So, we might ask ourselves, what does the unique ordering property of this matrix imply about where an element might be located?

We are told that every row and column is sorted. This means that element  $a[i][j]$  will be greater than the elements in row  $i$  between columns 0 and  $j - 1$  and the elements

in column  $j$  between rows 0 and  $i - 1$ .

Or, in other words:

$$\begin{aligned} a[i][0] &\leq a[i][1] \leq \dots \leq a[i][j-1] \leq a[i][j] \\ a[0][j] &\leq a[1][j] \leq \dots \leq a[i-1][j] \leq a[i][j] \end{aligned}$$

Looking at this visually, the dark gray element below is bigger than all the light gray elements.

15	20	70	85
20	35	80	95
30	55	95	105
40	80	100	120

The light gray elements also have an ordering to them: each is bigger than the elements to the left of it, as well as the elements above it. So, by transitivity, the dark gray element is bigger than the entire square.

15	20	70	85
20	35	80	95
30	55	95	105
40	80	100	120

This means that for any rectangle we draw in the matrix, the bottom right hand corner will always be the biggest.

Likewise, the top left hand corner will always be the smallest. The colors below indicate what we know about the ordering of elements (light gray < dark gray < black):

15	20	70	85
20	35	80	95
30	55	95	105
40	80	120	120

Let's return to the original problem: suppose we were searching for the value 85. If we look along the diagonal, we'll find the elements 35 and 95. What does this tell us about the location of 85?

15	20	70	85
25	35	80	95
30	55	95	105
40	80	120	120

85 can't be in the black area, since 95 is in the upper left hand corner and is therefore the smallest element in that square.

85 can't be in the light gray area either, since 35 is in the lower right hand corner of that square.

85 must be in one of the two white areas.

So, we partition our grid into four quadrants and recursively search the lower left quadrant and the upper right quadrant. These, too, will get divided into quadrants and searched.

Observe that since the diagonal is sorted, we can efficiently search it using binary search.

The code below implements this algorithm.

```

1  public Coordinate findElement(int[][] matrix, Coordinate origin,
2                                Coordinate dest, int x) {
3      if (!origin.inbounds(matrix) || !dest.inbounds(matrix)) {
4          return null;
5      }
6      if (matrix[origin.row][origin.column] == x) {
7          return origin;
8      } else if (!origin.isBefore(dest)) {
9          return null;
10     }
11
12    /* Set start to start of diagonal and end to the end of the
13     * diagonal. Since the grid may not be square, the end of the
14     * diagonal may not equal dest. */
15    Coordinate start = (Coordinate) origin.clone();
16    int diagDist = Math.min(dest.row - origin.row,
17                           dest.column - origin.column);
18    Coordinate end = new Coordinate(start.row + diagDist,
19                                    start.column + diagDist);
20    Coordinate p = new Coordinate(0, 0);
21
22    /* Do binary search on the diagonal, looking for the first
23     * element greater than x */
24    while (start.isBefore(end)) {
25        p.setToAverage(start, end);
26        if (x > matrix[p.row][p.column]) {
27            start.row = p.row + 1;
28            start.column = p.column + 1;
29        } else {
30            end.row = p.row - 1;
31            end.column = p.column - 1;
32        }
33    }
34
35    /* Split the grid into quadrants. Search the bottom left and the
36     * top right. */
37    return partitionAndSearch(matrix, origin, dest, start, x);
38 }
```

```
39
40 public Coordinate partitionAndSearch(int[][] matrix,
41     Coordinate origin, Coordinate dest, Coordinate pivot,
42     int elem) {
43     Coordinate lowerLeftOrigin =
44         new Coordinate(pivot.row, origin.column);
45     Coordinate lowerLeftDest =
46         new Coordinate(dest.row, pivot.column - 1);
47     Coordinate upperRightOrigin =
48         new Coordinate(origin.row, pivot.column);
49     Coordinate upperRightDest =
50         new Coordinate(pivot.row - 1, dest.column);
51
52     Coordinate lowerLeft =
53         findElement(matrix, lowerLeftOrigin, lowerLeftDest, elem);
54     if (lowerLeft == null) {
55         return findElement(matrix, upperRightOrigin,
56                             upperRightDest, elem);
56     }
57     return lowerLeft;
58 }
59
60
61 public static Coordinate findElement(int[][] matrix, int x) {
62     Coordinate origin = new Coordinate(0, 0);
63     Coordinate dest = new Coordinate(matrix.length - 1,
64                                     matrix[0].length - 1);
65     return findElement(matrix, origin, dest, x);
66 }
67
68 public class Coordinate implements Cloneable {
69     public int row;
70     public int column;
71     public Coordinate(int r, int c) {
72         row = r;
73         column = c;
74     }
75
76     public boolean inbounds(int[][] matrix) {
77         return row >= 0 && column >= 0 &&
78             row < matrix.length && column < matrix[0].length;
79     }
80
81     public boolean isBefore(Coordinate p) {
82         return row <= p.row && column <= p.column;
83     }
84
85     public Object clone() {
86         return new Coordinate(row, column);
87     }
88 }
```

```

91     public void setToAverage(Coordinate min, Coordinate max) {
92         row = (min.row + max.row) / 2;
93         column = (min.column + max.column) / 2;
94     }

```

If you read all this code and thought, “there’s no way I could do all this in an interview!” you’re probably right. You couldn’t. But, your performance on any problem is evaluated compared to other candidates on the same problem. So while you couldn’t implement all this, neither can they. You are at no disadvantage when you get a tricky problem like this.

You help yourself out a bit by separating code out into other methods. For example, by pulling `partitionAndSearch` out into its own method, you will have an easier time outlining key aspects of the code. You can then come back to fill in the body for `partitionAndSearch` if you have time.

**11.7** *A circus is designing a tower routine consisting of people standing atop one another’s shoulders. For practical and aesthetic reasons, each person must be both shorter and lighter than the person below him or her. Given the heights and weights of each person in the circus, write a method to compute the largest possible number of people in such a tower.*

pg 122

## SOLUTION

---

When we cut out all the “fluff” to this problem, we can understand that the problem is really the following.

*We have a list of pairs of items. Find the longest sequence such that both the first and second items are in non-decreasing order.*

If we apply the Simplify and Generalize approach (or the Pattern Matching approach), we can relate this problem to finding the longest increasing sequence in an array.

### Sub-Problem: Longest Increasing Subsequence

If the elements do not need to stay in the same (relative) order, then we would simply sort the array. This makes the problem too trivial, so let’s assume that the elements need to stay in the same relative order.

We can attempt to derive a recursive algorithm by looking at the array, element by element. First, we need to observe that knowing the longest increasing subsequence for  $A[0]$  through  $A[i]$  will not give us the answer for  $A[i + 1]$  and  $A[i + 2]$ . We can see this through a simple example:

Array: 13, 14, 10, 11, 12  
Longest(0 through 0): 13

```
Longest(0 through 1): 13, 14
Longest(0 through 2): 13, 14
Longest(0 through 3): 13, 14 OR 10, 11
Longest(0 through 4): 10, 11, 12
```

If we only tried to build `Longest(0 through 4)` and `Longest(0 through 3)` off of the most recent solutions, we would not find the optimal solution.

However, we can use a different recursive approach. Rather than trying to find the longest increasing subsequence across elements 0 through *i*, we can find the longest subsequence which *ends* with element *i*. Using the above example, we do the following:

```
Array: 13, 14, 10, 11, 12
Longest(ending with A[0]): 13
Longest(ending with A[1]): 13, 14
Longest(ending with A[2]): 10
Longest(ending with A[3]): 10, 11
Longest(ending with A[4]): 10, 11, 12
```

Note that the longest sequence ending with `A[i]` can be found by looking at all the prior solutions. We simply append `A[i]` to the longest “valid” one, where valid means any list where `A[i] > list.tail`.

### The Real Problem: Longest Increasing Subsequence of Pairs

Now that we know how to find the longest increasing subsequence of an array of integers, we can solve the real problem fairly easily. We simply sort the list of people by their heights, and then apply the `longestIncreasingSubsequence` algorithm on just their weights.

The code below implements this algorithm.

```
1 ArrayList<HtWt> getIncreasingSequence(ArrayList<HtWt> items) {
2     Collections.sort(items);
3     return longestIncreasingSubsequence(items);
4 }
5
6 void longestIncreasingSubsequence(ArrayList<HtWt> array,
7     ArrayList<HtWt>[] solutions, int current_index) {
8     if (current_index >= array.size() || current_index < 0) return;
9     HtWt current_element = array.get(current_index);
10
11    /* Find longest sequence we can append current_element to */
12    ArrayList<HtWt> best_sequence = null;
13    for (int i = 0; i < current_index; i++) {
14        if (array.get(i).isBefore(current_element)) {
15            best_sequence = seqWithMaxLength(best_sequence,
16                                         solutions[i]);
17        }
18    }
19}
```

```

20  /* Append current_element */
21  ArrayList<HtWt> new_solution = new ArrayList<HtWt>();
22  if (best_sequence != null) {
23      new_solution.addAll(best_sequence);
24  }
25  new_solution.add(current_element);
26
27  /* Add to list and recurse */
28  solutions[current_index] = new_solution;
29  longestIncreasingSubsequence(array, solutions, current_index+1);
30 }
31
32 ArrayList<HtWt> longestIncreasingSubsequence(
33     ArrayList<HtWt> array) {
34     ArrayList<HtWt>[] solutions = new ArrayList[array.size()];
35     longestIncreasingSubsequence(array, solutions, 0);
36
37     ArrayList<HtWt> best_sequence = null;
38     for (int i = 0; i < array.size(); i++) {
39         best_sequence = seqWithMaxLength(best_sequence, solutions[i]);
40     }
41
42     return best_sequence;
43 }
44
45 /* Returns longer sequence */
46 ArrayList<HtWt> seqWithMaxLength(ArrayList<HtWt> seq1,
47     ArrayList<HtWt> seq2) {
48     if (seq1 == null) return seq2;
49     if (seq2 == null) return seq1;
50     return seq1.size() > seq2.size() ? seq1 : seq2;
51 }
52
53 public class HtWt implements Comparable {
54     /* declarations, etc */
55
56     /* used for sort method */
57     public int compareTo( Object s ) {
58         HtWt second = (HtWt) s;
59         if (this.Ht != second.Ht) {
60             return ((Integer)this.Ht).compareTo(second.Ht);
61         } else {
62             return ((Integer)this.Wt).compareTo(second.Wt);
63         }
64     }
65
66     /* Returns true if "this" should be lined up before "other."
67      * Note that it's possible that this.isBefore(other) and
68      * other.isBefore(this) are both false. This is different from
69      * the compareTo method, where if a < b then b > a. */

```

```
70     public boolean isBefore(HtWt other) {  
71         if (this.Ht < other.Ht && this.Wt < other.Wt) return true;  
72         else return false;  
73     }  
74 }
```

This algorithm operates in  $O(n^2)$  time. An  $O(n \log(n))$  algorithm does exist, but it is considerably more complicated and it is highly unlikely that you would derive this in an interview—even with some help. However, if you are interested in exploring this solution, a quick internet search will turn up a number of explanations of this solution.

- 11.8** *Imagine you are reading in a stream of integers. Periodically, you wish to be able to look up the rank of a number  $x$  (the number of values less than or equal to  $x$ ). Implement the data structures and algorithms to support these operations. That is, implement the method `track(int x)`, which is called when each number is generated, and the method `getRankOfNumber(int x)`, which returns the number of values less than or equal to  $x$  (not including  $x$  itself).*

pg 122

### SOLUTION

A relatively easy way to implement this would be to have an array that holds all the elements in sorted order. When a new element comes in, we would need to shift the other elements to make room. Implementing `getRankOfNumber` would be quite efficient though. We would simply perform a binary search for  $n$ , and return the index.

However, this is very inefficient for inserting elements (that is, the `track(int x)` function). We need a data structure which is good at keeping relative ordering, as well as updating when we insert new elements. A binary search tree can do just that.

Instead of inserting elements into an array, we insert elements into a binary search tree. The method `track(int x)` will run in  $O(\log n)$  time, where  $n$  is the size of the tree (provided, of course, that the tree is balanced).

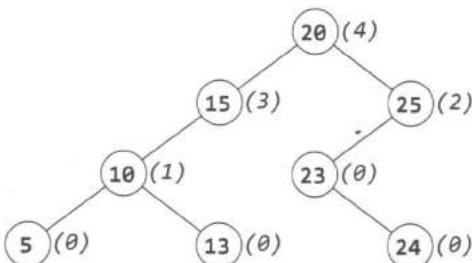
To find the rank of a number, we could do an in-order traversal, keeping a counter as we traverse. The goal is that, by the time we find  $x$ , counter will equal the number of elements less than  $x$ .

As long as we're moving left during searching for  $x$ , the counter won't change. Why? Because all the values we're skipping on the right side are greater than  $x$ . After all, the very smallest element (with rank of 1) is the leftmost node.

When we move to the right though, we skip over a bunch of elements on the left. All of these elements are less than  $x$ , so we'll need to increment counter by the number of elements in the left subtree.

Rather than counting the size of the left subtree (which would be inefficient), we can track this information as we add new elements to the tree.

Let's walk through an example on the following tree. In the below example, the value in parentheses indicates the number of nodes in the left subtree (or, in other words, the rank of the node *relative* to its subtree).



Suppose we want to find the rank of 24 in the tree above. We would compare 24 with the root, 20, and find that 24 must reside on the right. The root has 4 nodes in its left subtree, and when we include the root itself, this gives us five total nodes smaller than 24. We set counter to 5.

Then, we compare 24 with node 25 and find that 24 must be on the left. The value of counter does not update, since we're not "passing over" any smaller nodes. The value of counter is still 5.

Next, we compare 24 with node 23, and find that 24 must be on the right. Counter gets incremented by just 1 (to 6), since 23 has no left nodes.

Finally, we find 24 and we return counter: 6.

Recursively, the algorithm is the following:

```

1 int getRank(Node node, int x) {
2     if x is node.data
3         return node.leftSize()
4     if x is on left of node
5         return getRank(node.left, x)
6     if x is on right of node
7         return node.leftSize() + 1 + getRank(node.right, x)
8 }
  
```

The full code for this is below.

```

1 public class Question {
2     private static RankNode root = null;
3
4     public static void track(int number) {
5         if (root == null) {
6             root = new RankNode(number);
7         } else {
8             root.insert(number);
9         }
10    }
  
```

```
12     public static int getRankOfNumber(int number) {
13         return root.getRank(number);
14     }
15
16     ...
17 }
18
19 public class RankNode {
20     public int left_size = 0;
21     public RankNode left, right;
22     public int data = 0;
23     public RankNode(int d) {
24         data = d;
25     }
26
27     public void insert(int d) {
28         if (d <= data) {
29             if (left != null) left.insert(d);
30             else left = new RankNode(d);
31             left_size++;
32         } else {
33             if (right != null) right.insert(d);
34             else right = new RankNode(d);
35         }
36     }
37
38     public int getRank(int d) {
39         if (d == data) {
40             return left_size;
41         } else if (d < data) {
42             if (left == null) return -1;
43             else return left.getRank(d);
44         } else {
45             int right_rank = right == null ? -1 : right.getRank(d);
46             if (right_rank == -1) return -1;
47             else return left_size + 1 + right_rank;
48         }
49     }
50 }
```

Note how we've handled the case in which  $d$  is not found in the tree. We check for the -1 return value, and, when we find it, return -1 up the tree. It is important that you handle cases like this.

# **Testing**

*Concepts and Algorithms: Solutions*

**Chapter 12**

**12.1** Find the mistake(s) in the following code:

```
1 unsigned int i;
2 for (i = 100; i >= 0; --i)
3     printf("%d\n", i);
```

pg 129

### SOLUTION

There are two mistakes in this code.

First, note that an `unsigned int` is, by definition, always greater than or equal to zero. The for loop condition will therefore always be true, and it will loop infinitely.

The correct code to print all numbers from 100 to 1, is `i > 0`. If we truly wanted to print zero, we could add an additional `printf` statement after the for loop.

```
1 unsigned int i;
2 for (i = 100; i > 0; --i)
3     printf("%d\n", i);
```

One additional correction is to use `%u` in place of `%d`, as we are printing `unsigned int`.

```
1 unsigned int i;
2 for (i = 100; i > 0; --i)
3     printf("%u\n", i);
```

This code will now correctly print the list of all numbers from 100 to 1, in descending order.

**12.2** You are given the source to an application which crashes when it is run. After running it ten times in a debugger, you find it never crashes in the same place. The application is single threaded, and uses only the C standard library. What programming errors could be causing this crash? How would you test each one?

pg 129

### SOLUTION

The question largely depends on the type of application being diagnosed. However, we can give some general causes of random crashes.

1. *“Random Variable”*: The application may use some random number or variable component that may not be fixed for every execution of the program. Examples include user input, a random number generated by the program, or the time of day.
2. *Uninitialized Variable*: The application could have an uninitialized variable which, in some languages, may cause it to take on an arbitrary value. The values of this variable could result in the code taking a slightly different path each time.
3. *Memory Leak*: The program may have run out of memory. Other culprits are totally random for each run since it depends on the number of processes running at that

particular time. This also includes heap overflow or corruption of data on the stack.

4. *External Dependencies:* The program may depend on another application, machine, or resource. If there are multiple dependencies, the program could crash at any point.

To track down the issue, we should start with learning as much as possible about the application. Who is running it? What are they doing with it? What kind of application is it?

Additionally, although the application doesn't crash in exactly the same place, it's possible that it is linked to specific components or scenarios. For example, it could be that the application never crashes if it's simply launched and left untouched, and that crashes only appear at some point after loading a file. Or, it may be that all the crashes take place within the lower level components, such as file I/O.

It may be useful to approach this by elimination. Close down all other applications on the system. Track resource use very careful. If there are parts of the program we can disable, do so. Run it on a different machine and see if we experience the same issue. The more we can eliminate (or change), the easier we can track down the issue.

Additionally, we may be able to use tools to check for specific situations. For example, to investigate issue #2, we can utilize runtime tools which check for uninitialized variables.

These problems are as much about your brainstorming ability as they are about your approach. Do you jump all over the place, shouting out random suggestions? Or do you approach it in a logical, structured manner? Hopefully, it's the latter.

- 12.3** We have the following method used in a chess game: boolean `canMoveTo(int x, int y)`. This method is part of the `Piece` class and returns whether or not the piece can move to position  $(x, y)$ . Explain how you would test this method.

pg 129

## SOLUTION

In this problem, there are two primary types of testing: extreme case validation (ensuring that the program doesn't crash on bad input), and general case testing. We'll start with the first type.

### Testing Type #1: Extreme Case Validation

We need to ensure that the program handles bad or unusual input gracefully. This means checking the following conditions:

- Test with negative numbers for  $x$  and  $y$
- Test with  $x$  larger than the width
- Test with  $y$  larger than the height

- Test with a completely full board
- Test with an empty or nearly empty board
- Test with far more white pieces than black
- Test with far more black pieces than white

For the error cases above, we should ask our interviewer whether we want to return false or throw an exception, and we should test accordingly.

### Testing Type #2: General Testing:

General testing is much more expansive. Ideally, we would test every possible board, but there are far too many boards. We can, however, perform a reasonable coverage of different boards.

There are 6 pieces in chess, so we can test each piece against every other piece, in every possible direction. This would look something like the below code:

```
1   foreach piece a:  
2       for each other type of piece b (6 types + empty space)  
3           foreach direction d  
4               Create a board with piece a.  
5               Place piece b in direction d.  
6               Try to move - check return value.
```

The key to this problem is recognizing that we can't test every possible scenario, even if we would like to. So, instead, we must focus on the essential areas.

### 12.4 How would you load test a webpage without using any test tools?

pg 129

#### SOLUTION

Load testing helps to identify a web application's maximum operating capacity, as well as any bottlenecks that may interfere with its performance. Similarly, it can check how an application responds to variations in load.

To perform load testing, we must first identify the performance critical scenarios and the metrics which fulfill our performance objectives. Typical criteria include:

- Response time
- Throughput
- Resource utilization
- Maximum load that the system can bear.

Then, we design tests to simulate the load, taking care to measure each of these criteria.

In the absence of formal testing tools, we can basically create our own. For example,

we could simulate concurrent users by creating thousands of virtual users. We would write a multi-threaded program with thousands of threads, where each thread acts as a real-world user loading the page. For each user, we would programmatically measure response time, data I/O, etc.

We would then analyze the results based on the data gathered during the tests and compare it with the accepted values.

### 12.5 How would you test a pen?

pg 129

#### SOLUTION

This problem is largely about understanding the constraints and approaching the problem in a structured manner.

To understand the constraints, you should ask a lot of questions to understand the "who, what, where, when, how and why" of a problem (or as many of those as apply to the problem). Remember that a good tester understands exactly what he is testing before starting the work.

To illustrate the technique in this problem, let us guide you through a mock conversation.

**Interviewer:** How would you test a pen?

**Candidate:** Let me find out a bit about the pen. Who is going to use the pen?

**Interviewer:** Probably children.

**Candidate:** Okay, that's interesting. What will they be doing with it? Will they be writing, drawing, or doing something else with it?

**Interviewer:** Drawing.

**Candidate:** Ok, great. On what? Paper? Clothing? Walls?

**Interviewer:** On clothing.

**Candidate:** Great. What kind of tip does the pen have? Felt? Ball point? Is it intended to wash off, or is it intended to be permanent?

**Interviewer:** It's intended to wash off.

Many questions later, you may get to this:

**Candidate:** Okay, so as I understand it, we have a pen that is being targeted at 5 to 10 year olds. The pen has a felt tip and comes in red, green, blue and black. It's intended to wash off when clothing is washed. Is that correct?

The candidate now has a problem that is significantly different from what it initially

seemed to be. This is not uncommon. In fact, many interviewers intentionally give a problem that seems clear (everyone knows what a pen is!), only to let you discover that it's quite a different problem from what it seemed. Their belief is that users do the same thing, though users do so accidentally.

Now that you understand what you're testing, it's time to come up with a plan of attack. The key here is *structure*.

Consider what the different components of the object or problem, and go from there. In this case, the components might be:

- *Fact check:* Verify that the pen is felt tip and that the ink is one of the allowed colors.
- *Intended use:* Drawing. Does the pen write properly on clothing?
- *Intended use:* Washing. Does it wash off of clothing (even if it's been there for an extended period of time)? Does it wash off in hot, warm and cold water?
- *Safety:* Is the pen safe (non-toxic) for children?
- *Unintended uses:* How else might children use the pen? They might write on other surfaces, so you need to check whether the behavior there is correct. They might also stomp on the pen, throw it, and so on. You'll need to make sure that the pen holds up under these conditions.

Remember that in any testing question, you need to test both the intended and unintended scenarios. People don't always use the product the way you want them to.

### 12.6 How would you test an ATM in a distributed banking system?

pg 129

#### SOLUTION

The first thing to do on this question is to clarify assumptions. Ask the following questions:

- Who is going to use the ATM? Answers might be "anyone," or it might be "blind people," or any number of other answers.
- What are they going to use it for? Answers might be "withdrawing money," "transferring money," "checking their balance," or many other answers.
- What tools do we have to test? Do we have access to the code, or just to the ATM?

Remember: a good tester makes sure she knows what she's testing!

Once we understand what the system looks like, we'll want to break down the problem into different testable components. These components include:

- Logging in
- Withdrawing money

- Depositing money
- Checking balance
- Transferring money

We would probably want to use a mix of manual and automated testing.

Manual testing would involve going through the steps above, making sure to check for all the error cases (low balance, new account, nonexistent account, and so on).

Automated testing is a bit more complex. We'll want to automate all the standard scenarios, as shown above, and we also want to look for some very specific issues, such as race conditions. Ideally, we would be able to set up a closed system with fake accounts and ensure that, even if someone withdraws and deposits money rapidly from different locations, he never gets money or loses money that he shouldn't.

Above all, we need to prioritize security and reliability. People's accounts must always be protected, and we must make sure that money is always properly accounted for. No one wants to unexpectedly lose money! A good tester understands the system priorities.



## C and C++

*Knowledge Based: Solutions*

**Chapter 13**

### 13.1 Write a method to print the last K lines of an input file using C++.

pg 139

#### SOLUTION

One brute force way could be to count the number of lines (N) and then print from N-K to Nth line. But, this requires two reads of the file, which is unnecessarily costly. We need a solution which allows us to read just once and be able to print the last K lines.

We can allocate an array for all K lines and the last K lines we've read in the array. So, initially, our array has lines 0 through K, then 1 through K+1, then 2 through K+2, and so on. Each time that we read a new line, we purge the oldest line from the array.

But, you might ask, wouldn't this require shifting elements in the array, which is also very expensive? No, not if we do it correctly. Instead of shifting the array each time, we will use a circular array.

With a circular array, we always replace the oldest item when we read a new line. The oldest item is tracked in a separate variable, which adjusts as we add new items.

The following is an example of a circular array:

```
step 1 (initially): array = {a, b, c, d, e, f}. p = 0
step 2 (insert g):   array = {g, b, c, d, e, f}. p = 1
step 3 (insert h):   array = {g, h, c, d, e, f}. p = 2
step 4 (insert i):   array = {g, h, i, d, e, f}. p = 3
```

The code below implements this algorithm.

```
1 void printLast10Lines(char* fileName) {
2     const int K = 10;
3     ifstream file (fileName);
4     string L[K];
5     int size = 0;
6
7     /* read file line by line into circular array */
8     while (file.good()) {
9         getline(file, L[size % K]);
10        size++;
11    }
12
13    /* compute start of circular array, and the size of it */
14    int start = size > K ? (size % K) : 0;
15    int count = min(K, size);
16
17    /* print elements in the order they were read */
18    for (int i = 0; i < count; i++) {
19        cout << L[(start + i) % K] << endl;
20    }
21 }
```

This solution will require reading in the whole file, but only ten lines will be in memory

at any given point.

- 13.2** Compare and contrast a hash table and an STL map. How is a hash table implemented? If the number of inputs is small, which data structure options can be used instead of a hash table?

pg 139

## SOLUTION

In a hash table, a value is stored by calling a hash function on a key. Values are not stored in sorted order. Additionally, since hash tables use the key to find the index that will store the value, an insert or lookup can be done in amortized  $O(1)$  time (assuming few collisions in the hash table). In a hash table, one must also handle potential collisions. This is often done by chaining, which means to create a linked list of all the values whose keys map to a particular index.

An STL map inserts the key/value pairs into a binary search tree based on the keys. There is no need to handle collisions, and, since the tree is balanced, the insert and lookup time is guaranteed to be  $O(\log N)$ .

### How is a hash table implemented?

A hash table is traditionally implemented with an array of linked lists. When we want to insert a key/value pair, we map the key to an index in the array using a hash function. The value is then inserted into the linked list at that position.

Note that the elements in a linked list at a particular index of the array do not have the same key. Rather, `hashFunction(key)` is the same for these values. Therefore, in order to retrieve the value for a specific key, we need to store in each node both the exact key and the value.

To summarize, the hash table will be implemented with an array of linked lists, where each node in the linked list holds two pieces of data: the value and the original key. In addition, we will want to note the following design criteria:

1. We want to use a good hash function to ensure that the keys are well distributed. If they are not well distributed, then we would get a lot of collisions and the speed to find an element would decline.
2. No matter how good our hash function is, we will still have collisions, so we need a method for handling them. This often means chaining via a linked list, but it's not the only way.
3. We may also wish to implement methods to dynamically increase or decrease the hash table size depending on capacity. For example, when the ratio of the number of elements to the table size exceeds a certain threshold, we may wish to increase the hash table size. This would mean creating a new hash table and transferring the

entries from the old table to the new table. Because this is an expensive operation, we want to be careful to not do it too often.

### What can be used instead of a hash table, if the number of inputs is small?

You can use an STL map or a binary tree. Although this takes  $O(\log(n))$  time, the number of inputs may be small enough to make this time negligible.

#### 13.3 How do virtual functions work in C++?

pg 139

### SOLUTION

A virtual function depends on a “vtable” or “Virtual Table.” If any function of a class is declared to be virtual, a vtable is constructed which stores addresses of the virtual functions of this class. The compiler also adds a hidden `vptr` variable in all such classes which points to the vtable of that class. If a virtual function is not overridden in the derived class, the vtable of the derived class stores the address of the function in its parent class. The vtable is used to resolve the address of the function when the virtual function is called. Dynamic binding in C++ is performed through the vtable mechanism.

Thus, when we assign the derived class object to the base class pointer, the `vptr` variable points to the vtable of the derived class. This assignment ensures that the most derived virtual function gets called.

Consider the following code.

```
1 class Shape {
2     public:
3         int edge_length;
4         virtual int circumference () {
5             cout << "Circumference of Base Class\n";
6             return 0;
7         }
8     };
9 class Triangle: public Shape {
10    public:
11        int circumference () {
12            cout << "Circumference of Triangle Class\n";
13            return 3 * edge_length;
14        }
15    };
16 void main() {
17     Shape * x = new Shape();
18     x->circumference(); // "Circumference of Base Class"
19     Shape *y = new Triangle();
20     y->circumference(); // "Circumference of Triangle Class"
21 }
```

In the previous example, `circumference` is a virtual function in the `Shape` class, so it becomes virtual in each of the derived classes (`Triangle`, etc). C++ non-virtual function calls are resolved at compile time with static binding, while virtual function calls are resolved at runtime with dynamic binding.

- 13.4** What is the difference between deep copy and shallow copy? Explain how you would use each.

pg 139

### SOLUTION

A shallow copy copies all the member values from one object to another. A deep copy does all this and also deep copies any pointer objects.

An example of shallow and deep copy is below.

```

1 struct Test {
2     char * ptr;
3 };
4
5 void shallow_copy(Test & src, Test & dest) {
6     dest.ptr = src.ptr;
7 }
8
9 void deep_copy(Test & src, Test & dest) {
10    dest.ptr = (char*)malloc(strlen(src.ptr) + 1);
11    strcpy(dest.ptr, src.ptr);
12 }
```

Note that `shallow_copy` may cause a lot of programming runtime errors, especially with the creation and deletion of objects. Shallow copy should be used very carefully and only when a programmer really understands what he wants to do. In most cases, shallow copy is used when there is a need to pass information about a complex structure without actual duplication of data. One must also be careful with destruction of objects in a shallow copy.

In real life, shallow copy is rarely used. Deep copy should be used in most cases, especially when the size of the copied structure is small.

- 13.5** What is the significance of the keyword "volatile" in C?

pg 140

### SOLUTION

The keyword `volatile` informs the compiler that the value of variable it is applied to can change from the outside, without any update done by the code. This may be done by the operating system, the hardware, or another thread. Because the value can change

unexpectedly, the compiler will therefore reload the value each time from memory.

A volatile integer can be declared by either of the following statements:

```
int volatile x;  
volatile int x;
```

To declare a pointer to a volatile integer, we do the following:

```
volatile int * x;  
int volatile * x;
```

A volatile pointer to non-volatile data is rare, but can be done.

```
int * volatile x;
```

If you wanted to declare a volatile variable pointer for volatile memory (both pointer address and memory contained are volatile), you would do the following:

```
int volatile * volatile x;
```

Volatile variables are not optimized, which can be very useful. Imagine this function:

```
1 int opt = 1;  
2 void Fn(void) {  
3     start:  
4         if (opt == 1) goto start;  
5         else break;  
6 }
```

At first glance, our code appears to loop infinitely. The compiler may try to optimize it to:

```
1 void Fn(void) {  
2     start:  
3         int opt = 1;  
4         if (true)  
5             goto start;  
6 }
```

This becomes an infinite loop. However, an external operation might write '0' to the location of variable opt, thus breaking the loop.

To prevent the compiler from performing such optimization, we want to signal that another element of the system could change the variable. We do this using the volatile keyword, as shown below.

```
1 volatile int opt = 1;  
2 void Fn(void) {  
3     start:  
4         if (opt == 1) goto start;  
5         else break;  
6 }
```

Volatile variables are also useful when multi-threaded programs have global variables and any thread can modify these shared variables. We may not want optimization on these variables.

**13.6 Why does a destructor in base class need to be declared virtual?**

pg 140

**SOLUTION**

Let's think about why we have virtual methods to start with. Suppose we have the following code:

```
1 class Foo {  
2     public:  
3         void f();  
4     };  
5  
6 class Bar : public Foo {  
7     public:  
8         void f();  
9     };  
10  
11 Foo * p = new Bar();  
12 p->f();
```

Calling `p->f()` will result in a call to `Foo::f()`. This is because `p` is a pointer to `Foo`, and `f()` is not virtual.

To ensure that `p->f()` will invoke the most derived implementation of `f()`, we need to declare `f()` to be a virtual function.

Now, let's go back to our destructor. Destructors are used to clean up memory and resources. If `Foo`'s destructor were not virtual, then `Foo`'s destructor would be called, even when `p` is *really* of type `Bar`.

This is why we declare destructors to be virtual; we want to ensure that the destructor for the most derived class is called.

**13.7 Write a method that takes a pointer to a Node structure as a parameter and returns a complete copy of the passed in data structure. The Node data structure contains two pointers to other Nodes.**

pg 140

**SOLUTION**

The algorithm will maintain a mapping from a node address in the original structure to the corresponding node in the new structure. This mapping will allow us to discover previously copied nodes during a traditional depth first traversal of the structure. Traversals often mark visited nodes—the mark can take many forms and does not necessarily need to be stored in the node.

Thus, we have a simple recursive algorithm:

```
1  typedef map<Node*, Node*> NodeMap;
2
3  Node * copy_recursive(Node * cur, NodeMap & nodeMap) {
4      if(cur == NULL) {
5          return NULL;
6      }
7
8      NodeMap::iterator i = nodeMap.find(cur);
9      if (i != nodeMap.end()) {
10          // we've been here before, return the copy
11          return i->second;
12      }
13
14     Node * node = new Node;
15     nodeMap[cur] = node; // map current before traversing links
16     node->ptr1 = copy_recursive(cur->ptr1, nodeMap);
17     node->ptr2 = copy_recursive(cur->ptr2, nodeMap);
18     return node;
19 }
20
21 Node * copy_structure(Node * root) {
22     NodeMap nodeMap; // we will need an empty map
23     return copy_recursive(root, nodeMap);
24 }
```

- 13.8** Write a smart pointer class. A smart pointer is a data type, usually implemented with templates, that simulates a pointer while also providing automatic garbage collection. It automatically counts the number of references to a `SmartPointer<T*>` object and frees the object of type `T` when the reference count hits zero.

pg 140

### SOLUTION

A smart pointer is the same as a normal pointer, but it provides safety via automatic memory management. It avoids issues like dangling pointers, memory leaks and allocation failures. The smart pointer must maintain a single reference count for all references to a given object.

This is one of those problems that seems, at first glance, pretty overwhelming, especially if you're not a C++ expert. One useful way to approach the problem is to divide the problem into two parts: (1) outline the pseudocode and approach and then (2) implement the detailed code.

In terms of the approach, we need a reference count variable that is incremented when we add a new reference to the object and decremented when we remove a reference. The code should look something like the below pseudocode:

```

1 template <class T> class SmartPointer {
2     /* The smart pointer class needs pointers to both the object
3      * itself and to the ref count. These must be pointers, rather
4      * than the actual object or ref count value, since the goal of
5      * a smart pointer is that the reference count is tracked across
6      * multiple smart pointers to one object. */
7     T * obj;
8     unsigned * ref_count;
9 }
```

We know we need constructors and a single destructor for this class, so let's add those first.

```

1 SmartPointer(T * object) {
2     /* We want to set the value of T * obj, and set the reference
3      * counter to 1. */
4 }
5
6 SmartPointer(SmartPointer<T>& sptr) {
7     /* This constructor creates a new smart pointer that points to
8      * an existing object. We will need to first set obj and
9      * ref_count to pointer to sptr's obj and ref_count. Then,
10     * because we created a new reference to obj, we need to
11     * increment ref_count. */
12 }
13
14 ~SmartPointer(SmartPointer<T> sptr) {
15     /* We are destroying a reference to the object. Decrement
16     * ref_count. If ref_count is 0, then free the memory created by
17     * the integer and destroy the object. */
18 }
```

There's one additional way that references can be created: by setting one SmartPointer equal to another. We'll want to override the equal operator to handle this, but for now, let's sketch the code like this.

```

19 onSetEquals(SmartPoint<T> ptr1, SmartPoint<T> ptr2) {
20     /* If ptr1 has an existing value, decrement its reference count.
21      * Then, copy the pointers to obj and ref_count over. Finally,
22      * since we created a new reference, we need to increment
23      * ref_count. */
24 }
```

Getting just the approach, even without filling in the complicated C++ syntax, would count for a lot. Finishing out the code is now just a matter of filling the details.

```

1 template <class T> class SmartPointer {
2 public:
3     SmartPointer(T * ptr) {
4         ref = ptr;
5         ref_count = (unsigned*)malloc(sizeof(unsigned));
6         *ref_count = 1;
7     }
```

```
8     SmartPointer(SmartPointer<T> & sptr) {
9         ref = sptr.ref;
10        ref_count = sptr.ref_count;
11        ++(*ref_count);
12    }
13
14
15    /* Override the equal operator, so that when you set one smart
16       * pointer equal to another the old smart pointer has its
17       * reference count decremented and the new smart pointer has its
18       * reference count incremented. */
19    SmartPointer<T> & operator=(SmartPointer<T> & sptr) {
20        if (this == &sptr) return *this;
21
22        /* If already assigned to an object, remove one reference. */
23        if (*ref_count > 0) {
24            remove();
25        }
26
27        ref = sptr.ref;
28        ref_count = sptr.ref_count;
29        ++(*ref_count);
30        return *this;
31    }
32
33    ~SmartPointer() {
34        remove(); // Remove one reference to object.
35    }
36
37    T getValue() {
38        return *ref;
39    }
40
41 protected:
42    void remove() {
43        --(*ref_count);
44        if (*ref_count == 0) {
45            delete ref;
46            free(ref_count);
47            ref = NULL;
48            ref_count = NULL;
49        }
50    }
51
52    T * ref;
53    unsigned * ref_count;
54};
```

The code for this problem is complicated, and you probably wouldn't be expected to complete it flawlessly.

- 13.9** Write an aligned malloc and free function that supports allocating memory such that the memory address returned is divisible by a specific power of two.

pg 140

### SOLUTION

Typically, with `malloc`, we do not have control over where the memory is allocated within the heap. We just get a pointer to a block of memory which could start at any memory address within the heap.

We need to work with these constraints by requesting enough memory that we can return a memory address which is divisible by the desired value.

Suppose we are requesting a 100-byte chunk of memory, and we want it to start at a memory address that is a multiple of 16. How much extra memory would we need to allocate to ensure that we can do so? We would need to allocate an extra 15 bytes. With these 15 bytes, plus another 100 bytes right after that sequence, we know that we would have a memory address divisible by 16 with space for 100 bytes.

We could then do something like:

```

1 void* aligned_malloc(size_t required_bytes, size_t alignment) {
2     int offset = alignment - 1;
3     void* p = (void*) malloc(required_bytes + offset);
4     void* q = (void*) (((size_t)(p) + offset) & ~(alignment - 1));
5     return q;
6 }
```

Line 4 is a bit tricky, so let's discuss it. Suppose `alignment` is 16. We know that somewhere in the first 16 bytes there is a memory address divisible by 16. By doing `(p1 + 16) & 11..10000`, we are moving `q` back to a memory address divisible by 16. Doing an AND of the last three bits of the memory address with `0000` guarantees us that this new value will be divisible by 16.

This solution is *almost* perfect, except for one big issue: how do we free the memory?

We've allocated an extra 15 bytes, in the above example, and we need to free them when we free the "real" memory.

We can do this by storing, in this "extra" memory, the address of where the full memory block begins. We will store this immediately before the aligned memory block. Of course, this means that we now need to allocate even *more* extra memory to ensure that we have enough space to store this pointer.

Specifically, to bytes aligned with `alignment`, we will need to allocate an additional `alignment - 1 + sizeof(void*)` bytes.

The code below implements this approach.

```

1 void* aligned_malloc(size_t required_bytes, size_t alignment) {
2     void* p1; // original block
3     void** p2; // aligned block
```

```
4     int offset = alignment - 1 + sizeof(void*);
5     if ((p1 = (void*)malloc(required_bytes + offset)) == NULL) {
6         return NULL;
7     }
8     p2 = (void**)((size_t)(p1) + offset) & ~(alignment - 1));
9     p2[-1] = p1;
10    return p2;
11 }
12
13 void aligned_free(void *p2) {
14     /* for consistency, we use the same names as aligned_malloc*/.
15     void* p1 = ((void**)p2)[-1];
16     free(p1);
17 }
```

Let's look at how `aligned_free` works. The `aligned_free` method is passed in `p2` (the same `p2` as we had in `aligned_malloc`). We know, however, that the value of `p1` (which points to the beginning of the full memory block) was stored just before `p2`.

If we treat `p2` as a `void**` (or an array of `void*`'s), we can just look at the index `- 1` to retrieve `p1`. Then, by freeing `p1`, we have deallocated the whole memory block.

- 13.10** Write a function in C called `my2DAlloc` which allocates a two-dimensional array. Minimize the number of calls to `malloc` and make sure that the memory is accessible by the notation `arr[i][j]`.

pg 140

### SOLUTION

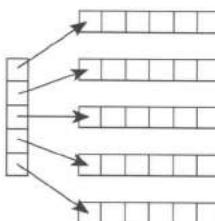
As you may know, a two-dimensional array is essentially an array of arrays. Since we use pointers with arrays, we can use double pointers to create a double array.

The basic idea is to create a one-dimensional array of pointers. Then, for each array index, we create a new one-dimensional array. This gives us a two-dimensional array that can be accessed via array indices.

The code below implements this.

```
1 int** my2DAlloc(int rows, int cols) {
2     int** rowptr;
3     int i;
4     rowptr = (int**) malloc(rows * sizeof(int*));
5     for (i = 0; i < rows; i++) {
6         rowptr[i] = (int*) malloc(cols * sizeof(int));
7     }
8     return rowptr;
9 }
```

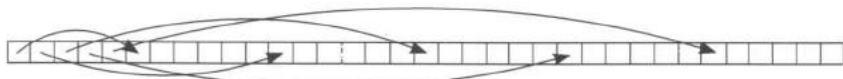
Observe how, in the above code, we've told `rowptr` where exactly each index should point. The below diagram represents how this memory is allocated.



To free this memory, we cannot simply call `free` on `rowptr`. We need to make sure to free not only the memory from the first `malloc` call, but also each subsequent call.

```
1 void my2DDealloc(int** rowptr, int rows) {
2     for (i = 0; i < rows; i++) {
3         free(rowptr[i]);
4     }
5     free(rowptr);
6 }
```

Rather than allocating the memory in many different blocks (one block for each row, plus one block to specify *where* each row is located), we can allocate this in a consecutive block of memory. Conceptually, for a two dimensional array with five rows and six columns, this would look like the following.



If it seems strange to view the 2D array like this (and it probably does), remember that this is fundamentally no different than the first diagram. The only difference is that the memory is in a contiguous block, so our first five (in this example) elements point elsewhere in the same block of memory.

To implement this solution, we do the following.

```
1 int** my2DAlloc(int rows, int cols) {
2     int i;
3     int header = rows * sizeof(int*);
4     int data = rows * cols * sizeof(int);
5     int** rowptr = (int**)malloc(header + data);
6     if (rowptr == NULL) {
7         return NULL;
8     }
9
10    int* buf = (int*) (rowptr + rows);
11    for (i = 0; i < rows; i++) {
12        rowptr[i] = buf + i * cols;
13    }
14    return rowptr;
15 }
```

You should carefully observe what is happening on lines 11 through 13. If there are five rows of six columns each, `array[0]` will point to `array[5]`, `array[1]` will point to `array[11]`, and so on.

Then, when we actually call `array[1][3]`, the computer looks up `array[1]`, which is a pointer to another spot in memory—specifically, a pointer to `array[5]`. This element is treated as its own array, and we then get the third (zero-indexed) element from it.

Constructing the array in a single call to `malloc` has the added benefit of allowing disposal of the array with a single `free` call rather than using a special function to free the remaining data blocks.

## **Java**

---

*Knowledge Based: Solutions*

**Chapter 14**

**14.1** In terms of inheritance, what is the effect of keeping a constructor private?

pg 145

### SOLUTION

Declaring the constructor private will ensure that no one outside of the class can directly instantiate the class. In this case, the only way to create an instance of the class is to provide a static public method, as is done when using the Factory Method Pattern.

Additionally, because the constructor is private, the class also cannot be inherited.

**14.2** In Java, does the finally block get executed if we insert a return statement inside the try block of a try-catch-finally?

pg 145

### SOLUTION

Yes, it will get executed. The finally block gets executed when the try block exits. Even when we attempt to exit within the try block (via a return statement, a continue statement, a break statement or any exception), the finally block will still be executed.

Note that there are some cases in which the finally block will not get executed, such as the following:

- If the virtual machine exits during try/catch block execution.
- If the thread which is executing the try/catch block gets killed.

**14.3** What is the difference between final, finally, and finalize?

pg 145

### SOLUTIONS

Despite their similar sounding names, final, finally and finalize have very different purposes. To speak in very general terms, final is used to control whether a variable, method, or class is “changeable.” The finally keyword is used in a try/catch block to ensure that a segment of code is always executed. The finalize() method is called by the garbage collector once it determines that no more references exist.

Further detail on these keywords and methods is provided below.

#### final

The final statement has a different meaning depending on its context.

- When applied to a variable (primitive): The value of the variable cannot change.
- When applied to a variable (reference): The reference variable cannot point to any other object on the heap.
- When applied to a method: The method cannot be overridden.
- When applied to a class: The class cannot be subclassed.

### **finally**

There is an optional **finally** block after the **try** block or after the **catch** block. Statements in the **finally** block will always be executed (except if Java Virtual Machine exits from the **try** block). The **finally** block is used to write the clean-up code.

### **finalize()**

The **finalize()** method is called by the garbage collector when it determines that no more references exist. It is typically used to clean up resources, such as closing a file.

## **14.4 Explain the difference between templates in C++ and generics in Java.**

pg 145

### **SOLUTION**

Many programmers consider the concepts of templates and generics to be equivalent simply because both allow you to do something along the lines of `List<String>`. But, how each language does this, and why, varies significantly.

The implementation of Java generics is rooted in an idea of “type erasure.” This technique eliminates the parameterized types when source code is translated to the Java Virtual Machine (JVM) byte code.

For example, suppose you have the Java code below:

```
1 Vector<String> vector = new Vector<String>();
2 vector.add(new String("hello"));
3 String str = vector.get(0);
```

During compilation, this code is re-written into:

```
1 Vector vector = new Vector();
2 vector.add(new String("hello"));
3 String str = (String) vector.get(0);
```

The use of Java generics didn’t really change much about our capabilities; it just made things a bit prettier. For this reason, Java generics are sometimes called “syntactic sugar.”

This is quite different from C++. In C++, templates are essentially a glorified macro set, with the compiler creating a new copy of the template code for each type. Proof of this is in the fact that an instance of `MyClass<Foo>` will not share a static variable with

`MyClass<Bar>`. Two instances of `MyClass<Foo>`, however, will share a static variable.

To illustrate this, consider the code below:

```
1  /*** MyClass.h ***/
2  template<class T> class MyClass {
3     public:
4         static int val;
5         MyClass(int v) { val = v; }
6     };
7
8  /*** MyClass.cpp ***/
9  template<typename T>
10 int MyClass<T>::bar;
11
12 template class MyClass<Foo>;
13 template class MyClass<Bar>;
14
15 /*** main.cpp ***/
16 MyClass<Foo> * foo1 = new MyClass<Foo>(10);
17 MyClass<Foo> * foo2 = new MyClass<Foo>(15);
18 MyClass<Bar> * bar1 = new MyClass<Bar>(20);
19 MyClass<Bar> * bar2 = new MyClass<Bar>(35);
20
21 int f1 = foo1->val; // will equal 15
22 int f2 = foo2->val; // will equal 15
23 int b1 = bar1->val; // will equal 35
24 int b2 = bar2->val; // will equal 35
```

In Java, static variables would be shared across instances of `MyClass`, regardless of the different type parameters.

Because of their architectural differences, Java generics and C++ templates have a number of other differences. These include:

- C++ templates can use primitive types, like `int`. Java cannot and must instead use `Integer`.
- In Java, you can restrict the template's type parameters to be of a certain type. For instance, you might use generics to implement a `CardDeck` and specify that the type parameter must extend from `CardGame`.
- In C++, the type parameter can be instantiated, whereas Java does not support this.
- In Java, the type parameter (i.e., the `Foo` in `MyClass<Foo>`) cannot be used for static methods and variables, since these would be shared between `MyClass<Foo>` and `MyClass<Bar>`. In C++, these classes are different, so the type parameter can be used for static methods and variables.
- In Java, all instances of `MyClass`, regardless of their type parameters, are the same type. The type parameters are erased at runtime. In C++, instances with different type parameters are different types.

Remember that although Java generics and C++ templates look the same in many ways, they are very different.

#### 14.5 Explain what object reflection is in Java and why it is useful.

pg 145

#### SOLUTION

Object Reflection is a feature in Java which provides a way to get reflective information about Java classes and objects, and perform operations such as:

1. Getting information about the methods and fields present inside the class at runtime.
2. Creating a new instance of a class.
3. Getting and setting the object fields directly by getting field reference, regardless of what the access modifier is.

The code below offers an example of object reflection.

```
1  /* Parameters */
2  Object[] doubleArgs = new Object[] { 4.2, 3.9 };
3
4  /* Get class */
5  Class rectangleDefinition = Class.forName("MyProj.Rectangle");
6
7  /* Equivalent: Rectangle rectangle = new Rectangle(4.2, 3.9); */
8  Class[] doubleArgsClass = new Class[] {double.class, double.class};
9  Constructor doubleArgsConstructor =
10    rectangleDefinition.getConstructor(doubleArgsClass);
11 Rectangle rectangle =
12   (Rectangle) doubleArgsConstructor.newInstance(doubleArgs);
13
14 /* Equivalent: Double area = rectangle.area(); */
15 Method m = rectangleDefinition.getDeclaredMethod("area");
16 Double area = (Double) m.invoke(rectangle);
```

This code does the equivalent of:

```
1 Rectangle rectangle = new Rectangle(4.2, 3.9);
2 Double area = rectangle.area();
```

#### Why Is Object Reflection Useful?

Of course, it doesn't seem very useful in the above example, but reflection can be very useful in particular cases.

Object reflection is useful for three main reasons:

1. It helps in observing or manipulating the runtime behavior of applications.
2. It can help in debugging or testing programs, as we have direct access to methods, constructors, and fields.
3. We can call methods by name when we don't know the method in advance. For example, we may let the user pass in a class name, parameters for the constructor, and a method name. We can then use this information to create an object and call a method. Doing these operations without reflection would require a complex series of if-statements, if it's possible at all.

**14.6** Implement a `CircularArray` class that supports an array-like data structure which can be efficiently rotated. The class should use a generic type, and should support iteration via the standard `for (Obj o : circularArray)` notation.

pg 145

### SOLUTION

This problem really has two parts to it. First, we need to implement the `CircularArray` class. Second, we need to support iteration. We will address these parts separately.

#### Implementing the `CircularArray` class

One way to implement the `CircularArray` class is to actually shift the elements each time we call `rotate(int shiftRight)`. Doing this is, of course, not very efficient.

Instead, we can just create a member variable `head` which points to what should be conceptually viewed as the start of the circular array. Rather than shifting around the elements in the array, we just increment `head` by `shiftRight`.

The code below implements this approach.

```
1  public class CircularArray<T> {  
2      private T[] items;  
3      private int head = 0;  
4  
5      public CircularArray(int size) {  
6          items = (T[]) new Object[size];  
7      }  
8  
9      private int convert(int index) {  
10         if (index < 0) {  
11             index += items.length;  
12         }  
13         return (head + index) % items.length;  
14     }  
15  
16     public void rotate(int shiftRight) {  
17         head = convert(shiftRight);  
18     }  
19 }
```

```

18    }
19
20    public T get(int i) {
21        if (i < 0 || i >= items.length) {
22            throw new java.lang.IndexOutOfBoundsException("...");
23        }
24        return items[convert(i)];
25    }
26
27    public void set(int i, T item) {
28        items[convert(i)] = item;
29    }
30 }

```

There are a number of things here which are easy to make mistakes on, such as:

- We cannot create an array of the generic type. Instead, we must either cast the array or define `items` to be of type `List<T>`. For simplicity, we have done the former.
- The `%` operator will return a negative value when we do `negValue % posVal`. For example, `-8 % 3` is `-2`. This is different from how mathematicians would define the modulus function. We must add `items.length` to a negative index to get the correct positive result.
- We need to be sure to consistently convert the raw index to the rotated index. For this reason, we have implemented a `convert` function that is used by other methods. Even the `rotate` function uses `convert`. This is a good example of code reuse.

Now that we have the basic code for `CircularArray` out of the way, we can focus on implementing an iterator.

### Implementing the Iterator Interface

The second part of this question asks us to implement the `CircularArray` class such that we can do the following:

```

1 CircularArray<String> array = ...
2 for (String s : array) { ... }

```

Implementing this requires implementing the `Iterator` interface.

To implement the `Iterator` interface, we need to do the following:

- Modify the `CircularArray<T>` definition to add `implements Iterable<T>`. This will also require us to add an `iterator()` method to `CircularArray<T>`.
- Create a `CircularArrayIterator<T>` which implements `Iterator<T>`. This will also require us to implement, in the `CircularArrayIterator`, the methods `hasNext()`, `next()`, and `remove()`.

Once we've done the above items, the `for` loop will "magically" work.

In the code below, we have removed the aspects of `CircularArray` which were iden-

tical to the earlier implementation.

```
1  public class CircularArray<T> implements Iterable<T> {
2      ...
3      public Iterator<T> iterator() {
4          return new CircularArrayIterator<T>(this);
5      }
6
7      private class CircularArrayIterator<TI> implements Iterator<TI>{
8          /* current reflects the offset from the rotated head, not
9             * from the actual start of the raw array. */
10         private int _current = -1;
11         private TI[] _items;
12
13         public CircularArrayIterator(CircularArray<TI> array){
14             _items = array.items;
15         }
16
17         @Override
18         public boolean hasNext() {
19             return _current < items.length - 1;
20         }
21
22         @Override
23         public TI next() {
24             _current++;
25             TI item = (TI) _items[convert(_current)];
26             return item;
27         }
28
29         @Override
30         public void remove() {
31             throw new UnsupportedOperationException("...");
32         }
33     }
34 }
```

In the above code, note that the first iteration of the for loop will call `hasNext()` and then `next()`. Be very sure that your implementation will return the correct values here.

When you get a problem like this one in an interview, there's a good chance you don't remember exactly what the various methods and interfaces are called. In this case, work through the problem as well as you can. If you can reason out what sorts of methods one might need, that alone will show a good degree of competency.

## Databases

*Knowledge Based: Solutions*

**Chapter 15**

Questions 1 through 3 refer to the below database schema:

Apartments		Buildings		Tenants	
AptID	int	BuildingID	int	TenantID	int
UnitNumber	varchar	ComplexID	int	TenantName	varchar
BuildingID	int	BuildingName	varchar		
		Address	varchar		

Complexes		AptTenants		Requests	
ComplexID	int	TenantID	int	RequestID	int
ComplexName	varchar	AptID	int	Status	varchar
				AptID	int
				Description	varchar

Note that each apartment can have multiple tenants, and each tenant can have multiple apartments. Each apartment belongs to one building, and each building belongs to one complex.

- 15.1** Write a SQL query to get a list of tenants who are renting more than one apartment.

pg 152

### SOLUTION

To implement this, we can use the HAVING and GROUP BY clauses and then perform an INNER JOIN with Tenants.

```
1  SELECT TenantName
2  FROM Tenants
3  INNER JOIN
4      (SELECT TenantID
5       FROM AptTenants
6       GROUP BY TenantID
7       HAVING count(*) > 1) C
8  ON Tenants.TenantID = C.TenantID
```

Whenever you write a GROUP BY clause in an interview (or in real life), make sure that anything in the SELECT clause is either an aggregate function or contained within the GROUP BY clause.

- 15.2** Write a SQL query to get a list of all buildings and the number of open requests (Requests in which status equals 'Open').

pg 152

### SOLUTION

This problem uses a straightforward join of Requests and Apartments to get a list of building IDs and the number of open requests. Once we have this list, we join it again with the Buildings table.

```

1  SELECT BuildingName, ISNULL(Count, 0) as 'Count'
2  FROM Buildings
3  LEFT JOIN
4      (SELECT Apartments.BuildingID, count(*) as 'Count'
5       FROM Requests INNER JOIN Apartments
6       ON Requests.AptID = Apartments.AptID
7       WHERE Requests.Status = 'Open'
8       GROUP BY Apartments.BuildingID) ReqCounts
9  ON ReqCounts.BuildingID = Buildings.BuildingID

```

Queries like this that utilize sub-queries should be thoroughly tested, even when coding by hand. It may be useful to test the inner part of the query first, and then test the outer part.

- 15.3** *Building #11 is undergoing a major renovation. Implement a query to close all requests from apartments in this building.*

pg 152

### SOLUTION

UPDATE queries, like SELECT queries, can have WHERE clauses. To implement this query, we get a list of all apartment IDs within building #11 and the list of update requests from those apartments.

```

1  UPDATE Requests
2  SET Status = 'Closed'
3  WHERE AptID IN
4      (SELECT AptID
5       FROM Apartments
6       WHERE BuildingID = 11)

```

- 15.4** *What are the different types of joins? Please explain how they differ and why certain types are better in certain situations.*

pg 152

### SOLUTION

JOIN is used to combine the results of two tables. To perform a JOIN, each of the tables must have at least one field that will be used to find matching records from the other table. The join type defines which records will go into the result set.

Let's take for example two tables: one table lists the "regular" beverages, and another lists the calorie-free beverages. Each table has two fields: the beverage name and its product code. The "code" field will be used to perform the record matching.

Regular Beverages:

Name	Code
Budweiser	BUDWEISER
Coca-Cola	COCACOLA
Pepsi	PEPSI

Calorie-Free Beverages:

Name	Code
Diet Coca-Cola	COCACOLA
Fresca	FRESCA
Diet Pepsi	PEPSI
Pepsi Light	PEPSI
Purified Water	Water

If we wanted to join Beverage with Calorie-Free Beverages, we would have many options. These are discussed below.

- **INNER JOIN:** The result set would contain only the data where the criteria match. In our example, we would get three records: one with a COCACOLA code and two with PEPSI codes.
- **OUTER JOIN:** An OUTER JOIN will always contain the results of INNER JOIN, but it may also contain some records that have no matching record in the other table. OUTER JOINS are divided into the following subtypes:
  - » **LEFT OUTER JOIN**, or simply **LEFT JOIN**: The result will contain all records from the left table. If no matching records were found in the right table, then its fields will contain the NULL values. In our example, we would get four records. In addition to INNER JOIN results, BUDWEISER would be listed, because it was in the left table.
  - » **RIGHT OUTER JOIN**, or simply **RIGHT JOIN**: This type of join is the opposite of LEFT JOIN. It will contain every record from the right table; the missing fields from the left table will be NULL. Note that if we have two tables, A and B, then we can say that the statement A LEFT JOIN B is equivalent to the statement B RIGHT JOIN A. In our example above, we will get five records. In addition to INNER JOIN results, FRESCA and WATER records will be listed.
  - » **FULL OUTER JOIN**: This type of join combines the results of the LEFT and RIGHT JOINS. All records from both tables will be included in the result set, regardless of whether or not a matching record exists in the other table. If no matching record was found, then the corresponding result fields will have a NULL value. In our example, we will get six records.

**15.5** What is denormalization? Explain the pros and cons.

pg 152

**SOLUTION**

Denormalization is a database optimization technique in which we add redundant data to one or more tables. This can help us avoid costly joins in a relational database.

By contrast, in a traditional normalized database, we store data in separate logical tables and attempt to minimize redundant data. We may strive to have only one copy of each piece of data in the database.

For example, in a normalized database, we might have a Courses table and a Teachers table. Each entry in Courses would store the teacherID for a Course but not the teacherName. When we need to retrieve a list of all Courses with the Teacher name, we would do a join between these two tables.

In some ways, this is great; if a teacher changes his or her name, we only have to update the name in one place.

The drawback, however, is that if the tables are large, we may spend an unnecessarily long time doing joins on tables.

Denormalization, then, strikes a different compromise. Under denormalization, we decide that we're okay with some redundancy and some extra effort to update the database in order to get the efficiency advantages of fewer joins.

Cons of Denormalization	Pros of Denormalization
Updates and inserts are more expensive.	Retrieving data is faster since we do fewer joins.
Denormalization can make update and insert code harder to write.	Queries to retrieve can be simpler (and therefore less likely to have bugs), since we need to look at fewer tables.
Data may be inconsistent. Which is the "correct" value for a piece of data?	
Data redundancy necessitates more storage.	

In a system that demands scalability, like that of any major tech companies, we almost always use elements of both normalized and denormalized databases.

- 15.6 Draw an entity-relationship diagram for a database with companies, people, and professionals (people who work for companies).

pg 152

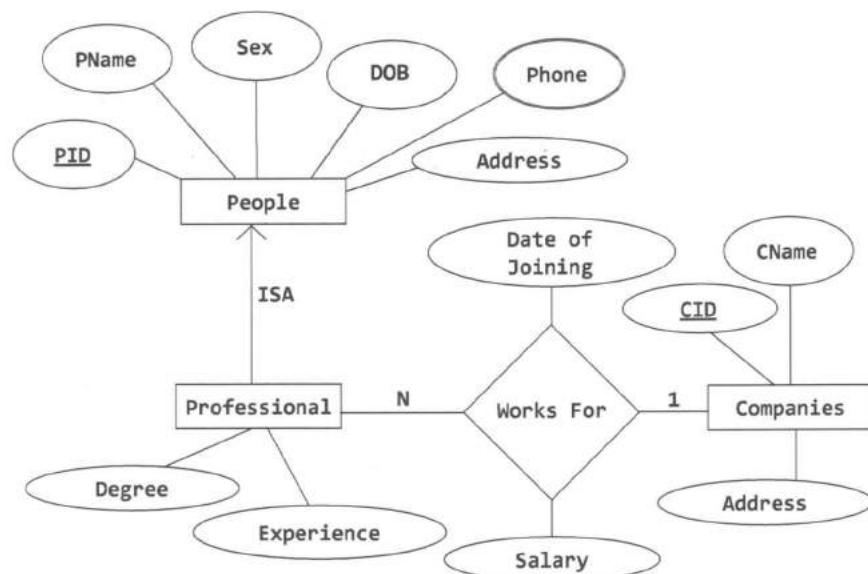
### SOLUTION

People who work for Companies are Professionals. So, there is an ISA ("is a") relationship between People and Professionals (or we could say that a Professional is derived from People).

Each Professional has additional information such as degree and work experiences in addition to the properties derived from People.

A Professional works for one company at a time, but Companies can hire many Professionals. So, there is a many-to-one relationship between Professionals and Companies. This "Works For" relationship can store attributes such as an employee's start date and salary. These attributes are defined only when we relate a Professional with a Company.

A Person can have multiple phone numbers, which is why Phone is a multi-valued attribute.



- 15.7 Imagine a simple database storing information for students' grades. Design what this database might look like and provide a SQL query to return a list of the honor roll students (top 10%), sorted by their grade point average.

pg 152

**SOLUTION**

In a simplistic database, we'll have at least three objects: Students, Courses, and CourseEnrollment. Students will have at least a student name and ID and will likely have other personal information. Courses will contain the course name and ID and will likely contain the course description, professor, and other information. CourseEnrollment will pair Students and Courses and will also contain a field for CourseGrade.

Students	
StudentID	int
StudentName	varchar(100)
Address	varchar(500)

Courses	
CourseID	int
CourseName	varchar(100)
ProfessorID	int

CourseEnrollment	
CourseID	int
StudentID	int
Grade	float
Term	int

This database could get arbitrarily more complicated if we wanted to add in professor information, billing information, and other data.

Using the Microsoft SQL Server TOP . . . PERCENT function, we might (incorrectly) first try a query like this:

```

1 /* Incorrect Code */
2 SELECT TOP 10 PERCENT AVG(CourseEnrollment.Grade) AS GPA,
3     CourseEnrollment.StudentID
4 FROM CourseEnrollment
5 GROUP BY CourseEnrollment.StudentID
6 ORDER BY AVG(CourseEnrollment.Grade)

```

The problem with the above code is that it will return literally the top 10% of rows, when sorted by GPA. Imagine a scenario in which there are 100 students, and the top 15 students all have 4.0 GPAs. The above function will only return 10 of those students, which is not really what we want. In case of a tie, we want to include the students who tied for the top 10% -- even if this means that our honor roll includes more than 10% of the class.

To correct this issue, we can build something similar to this query, but instead first get

the GPA cut off.

```
1  DECLARE @GPACutOff float;
2  SET @GPACutOff = (SELECT min(GPA) as 'GPAMin'
3      FROM (
4          SELECT TOP 10 PERCENT AVG(CourseEnrollment.Grade) AS GPA
5              FROM CourseEnrollment
6                  GROUP BY CourseEnrollment.StudentID
7                      ORDER BY GPA desc)
8      Grades);
```

Then, once we have @GPACutOff defined, selecting the students with at least this GPA is reasonably straightforward.

```
1  SELECT StudentName, GPA
2  FROM (
3      SELECT AVG(CourseEnrollment.Grade) AS GPA,
4          CourseEnrollment.StudentID
5      FROM CourseEnrollment
6          GROUP BY CourseEnrollment.StudentID
7          HAVING AVG(CourseEnrollment.Grade) >= @GPACutOff) Honors
8  INNER JOIN Students ON Honors.StudentID = Student.StudentID
```

Be very careful about what implicit assumptions you make. If you look at the above database description, what potentially incorrect assumption do you see? One is that each course can only be taught by one professor. At some schools, courses may be taught by multiple professors.

However, you *will* need to make some assumptions, or you'd drive yourself crazy. Which assumptions you make is less important than just recognizing *that* you made assumptions. Incorrect assumptions, both in the real world and in an interview, can be dealt with *as long as they are acknowledged*.

Remember, additionally, that there's a trade-off between flexibility and complexity. Creating a system in which a course can have multiple professors does increase the database's flexibility, but it also increases its complexity. If we tried to make our database flexible to every possible situation, we'd wind up with something hopelessly complex.

Make your design reasonably flexible, and state any other assumptions or constraints. This goes for not just database design, but object-oriented design and programming in general.

## **Threads and Locks**

---

*Knowledge Based: Solutions*

**Chapter 16**

### 16.1 What's the difference between a thread and a process?

pg 159

#### SOLUTION

Processes and threads are related to each other but are fundamentally different.

A process can be thought of as an instance of a program in execution. A process is an independent entity to which system resources (e.g., CPU time and memory) are allocated. Each process is executed in a separate address space, and one process cannot access the variables and data structures of another process. If a process wishes to access another process' resources, inter-process communications have to be used. These include pipes, files, sockets, and other forms.

A thread exists within a process and shares the process' resources (including its heap space). Multiple threads within the same process will share the same heap space. This is very different from processes, which cannot directly access the memory of another process. Each thread still has its own registers and its own stack, but other threads can read and write the heap memory.

A thread is a particular execution path of a process. When one thread modifies a process resource, the change is immediately visible to sibling threads.

### 16.2 How would you measure the time spent in a context switch?

pg 159

#### SOLUTION

This is a tricky question, but let's start with a possible solution.

A context switch is the time spent switching between two processes (i.e., bringing a waiting process into execution and sending an executing process into waiting/terminated state). This happens in multitasking. The operating system must bring the state information of waiting processes into memory and save the state information of the currently running process.

In order to solve this problem, we would like to record the timestamps of the last and first instruction of the swapping processes. The context switch time is the difference in the timestamps between the two processes.

Let's take an easy example: Assume there are only two processes,  $P_1$  and  $P_2$ .

$P_1$  is executing and  $P_2$  is waiting for execution. At some point, the operating system must swap  $P_1$  and  $P_2$ —let's assume it happens at the Nth instruction of  $P_1$ . If  $t_{x,k}$  indicates the timestamp in microseconds of the kth instruction of process x, then the context switch would take  $t_{2,1} - t_{1,n}$  microseconds.

The tricky part is this: how do we know when this swapping occurs? We cannot, of

course, record the timestamp of every instruction in the process.

Another issue is that swapping is governed by the scheduling algorithm of the operating system and there may be many kernel level threads which are also doing context switches. Other processes could be contending for the CPU or the kernel handling interrupts. The user does not have any control over these extraneous context switches. For instance, if at time  $t_{1,n}$ , the kernel decides to handle an interrupt, then the context switch time would be overstated.

In order to overcome these obstacles, we must first construct an environment such that after  $P_1$  executes, the task scheduler immediately selects  $P_2$  to run. This may be accomplished by constructing a data channel, such as a pipe, between  $P_1$  and  $P_2$  and having the two processes play a game of ping-pong with a data token.

That is, let's allow  $P_1$  to be the initial sender and  $P_2$  to be the receiver. Initially,  $P_2$  is blocked (sleeping) as it awaits the data token. When  $P_1$  executes, it delivers the token over the data channel to  $P_2$  and immediately attempts to read a response token. However, since  $P_2$  has not yet had a chance to run, no such token is available for  $P_1$  and the process is blocked. This relinquishes the CPU.

A context switch results and the task scheduler must select another process to run. Since  $P_2$  is now in a ready-to-run state, it is a desirable candidate to be selected by the task scheduler for execution. When  $P_2$  runs, the roles of  $P_1$  and  $P_2$  are swapped.  $P_2$  is now acting as the sender and  $P_1$  as the blocked receiver. The game ends when  $P_2$  returns the token to  $P_1$ .

To summarize, an iteration of the game is played with the following steps:

- $P_2$  blocks awaiting data from  $P_1$ .
- $P_1$  marks the start time.
- $P_1$  sends token to  $P_2$ .
- $P_1$  attempts to read a response token from  $P_2$ . This induces a context switch.
- $P_2$  is scheduled and receives the token.
- $P_2$  sends a response token to  $P_1$ .
- $P_2$  attempts read a response token from  $P_1$ . This induces a context switch.
- $P_1$  is scheduled and receives the token.
- $P_1$  marks the end time.

The key is that the delivery of a data token induces a context switch. Let  $T_d$  and  $T_r$  be the time it takes to deliver and receive a data token, respectively, and let  $T_c$  be the amount of time spent in a context switch. At step 2,  $P_1$  records the timestamp of the delivery of the token, and at step 9, it records the timestamp of the response. The amount of time elapsed,  $T$ , between these events may be expressed by:

$$T = 2 * (T_d + T_c + T_r)$$

This formula arises because of the following events:  $P_1$  sends a token (3), the CPU context switches (4),  $P_2$  receives it (5).  $P_2$  then sends the response token (6), the CPU context switches (7), and finally  $P_1$  receives it (8).

$P_1$  will be able to easily compute  $T$ , since this is just the time between events 3 and 8. So, to solve for  $T_c$ , we must first determine the value of  $T_d + T_r$ .

How can we do this? We can do this by measuring the length of time it takes  $P_1$  to send and receive a token to itself. This will not induce a context switch since  $P_1$  is running on the CPU at the time it sent the token and will not block to receive it.

The game is played a number of iterations to average out any variability in the elapsed time between steps 2 and 9 that may result from unexpected kernel interrupts and additional kernel threads contending for the CPU. We select the smallest observed context switch time as our final answer.

However, all we can ultimately say that this is an approximation which depends on the underlying system. For example, we make the assumption that  $P_2$  is selected to run once a data token becomes available. However, this is dependent on the implementation of the task scheduler and we cannot make any guarantees.

That's okay; it's important in an interview to recognize when your solution might not be perfect.

- 16.3** In the famous dining philosophers problem, a bunch of philosophers are sitting around a circular table with one chopstick between each of them. A philosopher needs both chopsticks to eat, and always picks up the left chopstick before the right one. A deadlock could potentially occur if all the philosophers reached for the left chopstick at the same time. Using threads and locks, implement a simulation of the dining philosophers problem that prevents deadlocks.

pg 160

### SOLUTION

First, let's implement a simple simulation of the dining philosophers problem in which we don't concern ourselves with deadlocks. We can implement this solution by having Philosopher extend Thread, and Chopstick call lock.lock() when it is picked up and lock.unlock() when it is put down.

```
1 public class Chopstick {
2     private Lock lock;
3
4     public Chopstick() {
5         lock = new ReentrantLock();
6     }
7
8     public void pickUp() {
9         lock.lock();
```

```

10    }
11
12    public void putDown() {
13        lock.unlock();
14    }
15 }
16
17 public class Philosopher extends Thread {
18     private int bites = 10;
19     private Chopstick left;
20     private Chopstick right;
21
22     public Philosopher(Chopstick left, Chopstick right) {
23         this.left = left;
24         this.right = right;
25     }
26
27     public void eat() {
28         pickUp();
29         chew();
30         putDown();
31     }
32
33     public void pickUp() {
34         left.pickUp();
35         right.pickUp();
36     }
37
38     public void chew() { }
39
40     public void putDown() {
41         left.putDown();
42         right.putDown();
43     }
44
45     public void run() {
46         for (int i = 0; i < bites; i++) {
47             eat();
48         }
49     }
50 }
```

Running the above code may lead to a deadlock if all the philosophers have a left chopstick and are waiting for the right one.

To prevent deadlocks, we can implement a strategy where a philosopher will put down his left chopstick if he is unable to obtain the right one.

```

1  public class Chopstick {
2      /* same as before */
3 }
```

```
4     public boolean pickUp() {
5         return lock.tryLock();
6     }
7 }
8
9 public class Philosopher extends Thread {
10    /* same as before */
11
12    public void eat() {
13        if (pickUp()) {
14            chew();
15            putDown();
16        }
17    }
18
19    public boolean pickUp() {
20        /* attempt to pick up */
21        if (!left.pickUp()) {
22            return false;
23        }
24        if (!right.pickUp()) {
25            left.putDown();
26            return false;
27        }
28        return true;
29    }
30 }
```

In the above code, we need to be sure to release the left chopstick if we can't pick up the right one—and to not call `putDown()` on the chopsticks if we never had them in the first place.

### 16.4 Design a class which provides a lock only if there are no possible deadlocks.

pg 160

#### SOLUTION

There are several common ways to prevent deadlocks. One of the popular ways is to require a process to declare upfront what locks it will need. We can then verify if a deadlock would be created by issuing these locks, and we can fail if so.

With these constraints in mind, let's investigate how we can detect deadlocks. Suppose this was the order of locks requested:

```
A = {1, 2, 3, 4}
B = {1, 3, 5}
C = {7, 5, 9, 2}
```

This may create a deadlock because we could have the following scenario:

- A locks 2, waits on 3
- B locks 3, waits on 5
- C locks 5, waits on 2

We can think about this as a graph, where 2 is connected to 3, 3 is connected to 5, and 5 is connected to 2. A deadlock is represented by a cycle. An edge  $(w, v)$  exists in the graph if a process declares that it will request lock  $v$  immediately after lock  $w$ . For the earlier example, the following edges would exist in the graph:  $(1, 2)$ ,  $(2, 3)$ ,  $(3, 4)$ ,  $(1, 3)$ ,  $(3, 5)$ ,  $(7, 5)$ ,  $(5, 9)$ ,  $(9, 2)$ . The “owner” of the edge does not matter.

This class will need a `declare` method, which threads and processes will use to declare what order they will request resources in. This `declare` method will iterate through the `declareOrder`, adding each contiguous pair of elements  $(v, w)$  to the graph. Afterwards, it will check to see if any cycles have been created. If any cycles have been created, it will backtrack, removing these edges from the graph, and then exit.

We just have one final component to discuss: how do we detect a cycle? We can detect a cycle by doing a depth first search through each connected component (i.e., each connected part of the graph). Complex algorithms exist to select all the connected components of a graph, but our work in this problem does not require this degree of complexity.

We know that if a cycle was created, one of our new edges must be to blame. Thus, as long as our depth first search touches all of these edges at some point, then we know that we have fully searched for a cycle.

The pseudocode for this special case cycle detection looks like this:

```

1  boolean checkForCycle(locks[] locks) {
2      touchedNodes = hash table(lock -> boolean)
3      initialize touchedNodes to false for each lock in locks
4      for each (lock x in process.locks) {
5          if (touchedNodes[x] == false) {
6              if (hasCycle(x, touchedNodes)) {
7                  return true;
8              }
9          }
10     }
11     return false;
12 }

13
14 boolean hasCycle(node x, touchedNodes) {
15     touchedNodes[r] = true;
16     if (x.state == VISITING) {
17         return true;
18     } else if (x.state == FRESH) {
19         ... (see full code below)
20     }
21 }
```

In the above code, note that we may do several depth first searches, but `touchedNodes` is only initialized once. We iterate until all the values in `touchedNodes` are false.

The code below provides further details. For simplicity, we assume that all locks and processes (owners) are ordered sequentially.

```
1  public class LockFactory {
2      private static LockFactory instance;
3
4      private int numberOfLocks = 5; /* default */
5      private LockNode[] locks;
6
7      /* Maps from a process or owner to the order that the owner
8       * claimed it would call the locks in */
9      private Hashtable<Integer, LinkedList<LockNode>> lockOrder;
10
11     private LockFactory(int count) { ... }
12     public static LockFactory getInstance() { return instance; }
13
14     public static synchronized LockFactory initialize(int count) {
15         if (instance == null) instance = new LockFactory(count);
16         return instance;
17     }
18
19     public boolean hasCycle(
20         Hashtable<Integer, Boolean> touchedNodes,
21         int[] resourcesInOrder) {
22         /* check for a cycle */
23         for (int resource : resourcesInOrder) {
24             if (touchedNodes.get(resource) == false) {
25                 LockNode n = locks[resource];
26                 if (n.hasCycle(touchedNodes)) {
27                     return true;
28                 }
29             }
30         }
31         return false;
32     }
33
34     /* To prevent deadlocks, force the processes to declare upfront
35      * what order they will need the locks in. Verify that this
36      * order does not create a deadlock (a cycle in a directed
37      * graph) */
38     public boolean declare(int ownerId, int[] resourcesInOrder) {
39         Hashtable<Integer, Boolean> touchedNodes =
40             new Hashtable<Integer, Boolean>();
41
42         /* add nodes to graph */
43         int index = 1;
44         touchedNodes.put(resourcesInOrder[0], false);
```

```
45     for (index = 1; index < resourcesInOrder.length; index++) {
46         LockNode prev = locks[resourcesInOrder[index - 1]];
47         LockNode curr = locks[resourcesInOrder[index]];
48         prev.joinTo(curr);
49         touchedNodes.put(resourcesInOrder[index], false);
50     }
51
52     /* if we created a cycle, destroy this resource list and
53      * return false */
54     if (hasCycle(touchedNodes, resourcesInOrder)) {
55         for (int j = 1; j < resourcesInOrder.length; j++) {
56             LockNode p = locks[resourcesInOrder[j - 1]];
57             LockNode c = locks[resourcesInOrder[j]];
58             p.remove(c);
59         }
60         return false;
61     }
62
63     /* No cycles detected. Save the order that was declared, so
64      * that we can verify that the process is really calling the
65      * locks in the order it said it would. */
66     LinkedList<LockNode> list = new LinkedList<LockNode>();
67     for (int i = 0; i < resourcesInOrder.length; i++) {
68         LockNode resource = locks[resourcesInOrder[i]];
69         list.add(resource);
70     }
71     lockOrder.put(ownerId, list);
72
73     return true;
74 }
75
76     /* Get the lock, verifying first that the process is really
77      * calling the locks in the order it said it would. */
78     public Lock getLock(int ownerId, int resourceId) {
79         LinkedList<LockNode> list = lockOrder.get(ownerId);
80         if (list == null) return null;
81
82         LockNode head = list.getFirst();
83         if (head.getId() == resourceId) {
84             list.removeFirst();
85             return head.getLock();
86         }
87         return null;
88     }
89 }
90
91     public class LockNode {
92         public enum VisitState { FRESH, VISITING, VISITED };
93
94         private ArrayList<LockNode> children;
```

```
95     private int lockId;
96     private Lock lock;
97     private int maxLocks;
98
99     public LockNode(int id, int max) { ... }
100
101    /* Join "this" to "node", checking to make sure that it doesn't
102     * create a cycle */
103    public void joinTo(LockNode node) { children.add(node); }
104    public void remove(LockNode node) { children.remove(node); }
105
106    /* Check for a cycle by doing a depth-first-search. */
107    public boolean hasCycle(
108        Hashtable<Integer, Boolean> touchedNodes) {
109        VisitState[] visited = new VisitState[maxLocks];
110        for (int i = 0; i < maxLocks; i++) {
111            visited[i] = VisitState.FRESH;
112        }
113        return hasCycle(visited, touchedNodes);
114    }
115
116    private boolean hasCycle(VisitState[] visited,
117        Hashtable<Integer, Boolean> touchedNodes) {
118        if (touchedNodes.containsKey(lockId)) {
119            touchedNodes.put(lockId, true);
120        }
121
122        if (visited[lockId] == VisitState.VISITING) {
123            /* We looped back to this node while still visiting it, so
124             * we know there's a cycle. */
125            return true;
126        } else if (visited[lockId] == VisitState.FRESH) {
127            visited[lockId] = VisitState.VISITING;
128            for (LockNode n : children) {
129                if (n.hasCycle(visited, touchedNodes)) {
130                    return true;
131                }
132            }
133            visited[lockId] = VisitState.VISITED;
134        }
135        return false;
136    }
137
138    public Lock getLock() {
139        if (lock == null) lock = new ReentrantLock();
140        return lock;
141    }
142
143    public int getId() { return lockId; }
144 }
```

As always, when you see code this complicated and lengthy, you wouldn't be expected to write all of it. More likely, you would be asked to sketch out pseudocode and possibly implement one of these methods.

**16.5** Suppose we have the following code:

```
public class Foo {
    public Foo() { ... }
    public void first() { ... }
    public void second() { ... }
    public void third() { ... }
}
```

The same instance of *Foo* will be passed to three different threads. *ThreadA* will call *first*, *threadB* will call *second*, and *threadC* will call *third*. Design a mechanism to ensure that *first* is called before *second* and *second* is called before *third*.

pg 160

### SOLUTION

The general logic is to check if *first()* has completed before executing *second()*, and if *second()* has completed before calling *third()*. Because we need to be very careful about thread safety, simple boolean flags won't do the job.

What about using a lock to do something like the below code?

```
1  public class FooBad {
2      public int pauseTime = 1000;
3      public ReentrantLock lock1, lock2, lock3;
4
5      public FooBad() {
6          try {
7              lock1 = new ReentrantLock();
8              lock2 = new ReentrantLock();
9              lock3 = new ReentrantLock();
10
11             lock1.lock();
12             lock2.lock();
13             lock3.lock();
14         } catch (...) { ... }
15     }
16
17     public void first() {
18         try {
19             ...
20             lock1.unlock(); // mark finished with first()
21         } catch (...) { ... }
```

```
22     }
23
24     public void second() {
25         try {
26             lock1.lock(); // wait until finished with first()
27             lock1.unlock();
28             ...
29
30             lock2.unlock(); // mark finished with second()
31         } catch (...) { ... }
32     }
33
34     public void third() {
35         try {
36             lock2.lock(); // wait until finished with third()
37             lock2.unlock();
38             ...
39         } catch (...) { ... }
40     }
41 }
```

This code won't actually quite work due to the concept of *lock ownership*. One thread is actually performing the lock (in the `FooBad` constructor), but different threads attempt to unlock the locks. This is not allowed, and your code will raise an exception. A lock in Java is owned by the same thread which locked it.

Instead, we can replicate this behavior with semaphores. The logic is identical.

```
1  public class Foo {
2      public Semaphore sem1, sem2, sem3;
3
4      public Foo() {
5          try {
6              sem1 = new Semaphore(1);
7              sem2 = new Semaphore(1);
8              sem3 = new Semaphore(1);
9
10             sem1.acquire();
11             sem2.acquire();
12             sem3.acquire();
13         } catch (...) { ... }
14     }
15
16     public void first() {
17         try {
18             ...
19             sem1.release();
20         } catch (...) { ... }
21     }
22
23     public void second() {
```

```
24     try {
25         sem1.acquire();
26         sem1.release();
27         ...
28         sem2.release();
29     } catch (...) { ... }
30 }
31
32 public void third() {
33     try {
34         sem2.acquire();
35         sem2.release();
36         ...
37     } catch (...) { ... }
38 }
39 }
```

- 16.6** You are given a class with synchronized method A and a normal method B. If you have two threads in one instance of a program, can they both execute A at the same time? Can they execute A and B at the same time?

pg 160

### SOLUTION

By applying the word synchronized to a method, we ensure that two threads cannot execute synchronized methods *on the same object instance* at the same time.

So, the answer to the first part really depends. If the two threads have the same instance of the object, then no, they cannot simultaneously execute method A. However, if they have different instances of the object, then they can.

Conceptually, you can see this by considering locks. A synchronized method applies a “lock” on *all* synchronized methods in that instance of the object. This blocks other threads from executing synchronized methods within that instance.

In the second part, we’re asked if thread1 can execute synchronized method A while thread2 is executing non-synchronized method B. Since B is not synchronized, there is nothing to block thread1 from executing A while thread2 is executing B. This is true regardless of whether thread1 and thread2 have the same instance of the object or not.

Ultimately, the key concept to remember here is that only one synchronized method can be in execution per instance of that object. Other threads can execute non-synchronized methods on that instance, or they can execute any method on a different instance of the object.



## **Moderate**

---

*Additional Review Problems: Solutions*

**Chapter 17**

17.1 Write a function to swap a number in place (that is, without temporary variables).

pg 163

### SOLUTION

This is a classic interview problem, and it's a reasonably straightforward one. We'll walk through this using  $a_0$  to indicate the original value of  $a$  and  $b_0$  to indicate the original value of  $b$ . We'll also use  $\text{diff}$  to indicate the value of  $a_0 - b_0$ .

Let's picture these on a number line for the case where  $a > b$ .



First, we briefly set  $a$  to  $\text{diff}$ , which is the right side of the above number line. Then, when we add  $b$  and  $\text{diff}$  (and store that value in  $b$ ), we get  $a_0$ . We now have  $b = a_0$  and  $a = \text{diff}$ . All that's left to do is to set  $a$  equal to  $a_0 - \text{diff}$ , which is just  $b - a$ .

The code below implements this.

```
1 public static void swap(int a, int b) {  
2     // Example for a = 9, b = 4  
3     a = a - b; // a = 9 - 4 = 5  
4     b = a + b; // b = 5 + 4 = 9  
5     a = b - a; // a = 9 - 5  
6  
7     System.out.println("a: " + a);  
8     System.out.println("b: " + b);  
9 }
```

We can implement a similar solution with bit manipulation. The benefit of this solution is that it works for more data types than just integers.

```
1 public static void swap_opt(int a, int b) {  
2     // Example for a = 101 (in binary) and b = 110  
3     a = a^b; // a = 101^110 = 011  
4     b = a^b; // b = 011^110 = 101  
5     a = a^b; // a = 011^101 = 110  
6  
7     System.out.println("a: " + a);  
8     System.out.println("b: " + b);  
9 }
```

This code works by using XORs. The easiest way to see how this works is by looking at it for a two bits  $p$  and  $q$ . We will again use the notation  $p_0$  and  $q_0$  to indicate the original values.

If we can correctly swap two bits, then we know the entire operation works correctly. Let's walk through this line-by-line.

```

1 p = pe ^ qe /* 0 if pe = qe, 1 if pe != qe */
2 q = p ^ qe /* equals value of pe */
3 p = p ^ q /* equals value of qe */

```

In line 1, doing the operation  $p = p_e \wedge q_e$  will result in a 0 if  $p_e = q_e$  and a 1 if  $p_e \neq q_e$ .

In line 2, we do  $q = p \wedge q_e$ . We can examine this for the two possible values of  $p$ . Since we are trying to eventually swap the original values of  $p$  and  $q$ , we want this operation to return the value of  $p_e$ .

- Case  $p = 0$ : In this case,  $p_e = q_e$ , so we need this operation to return either  $p_e$  or  $q_e$ . XORing any value with 0 will always return the original value, so we know that this operation will correctly return  $q_e$  (or  $p_e$ ).
- Case  $p = 1$ : In this case,  $p_e \neq q_e$ . We need this operation to return a 1 if  $q_e$  is 0 and a 0 if  $p_e$  is 1. This is exactly what an XOR with 1 will do.

In line 3, we do  $p = p \wedge q$ . Let's again examine this for the two possible values of  $p$ . Our goal is to return  $q_e$ . Observe that  $q$  now equals  $p_e$ , so we are really doing  $p \wedge p_e$ .

- Case  $p = 0$ : Since  $p_e = q_e$ , we want to return either  $p_e$  or  $q_e$ —either will do. By doing  $0 \wedge p_e$ , we return  $p_e$ , which equals  $q_e$ .
- Case  $p = 1$ : Here, we are doing  $1 \wedge p_e$ . This will flip the value of  $p_e$ , which is exactly what we want, since  $p_e \neq q_e$ .

We have now set  $p$  equal to  $q_e$  and  $q$  equal to  $p_e$ . Thus, since our operation correctly swaps each bit, we know it will correctly swap the entire int.

## 17.2 Design an algorithm to figure out if someone has won a game of tic-tac-toe.

pg 164

### SOLUTION

At first glance, this problem seems really straightforward. We're just checking a tic-tac-toe board; how hard could it be? It turns out that the problem is a bit more complex, and there is no single "perfect" answer. The optimal solution very much depends on your preferences.

There are a few major design decisions to consider:

1. Will `hasWon` be called just once, or many times (for instance, as part of a tic-tac-toe website)? If the latter is the case, we may want to add pre-processing time in order to optimize the runtime of `hasWon`.
2. Tic-tac-toe is usually on a 3x3 board. Do we want to design for just that, or do we want to implement it as an NxN solution?
3. In general, how much do we prioritize compactness of code vs. speed of execution vs. clarity of code? Remember that the most efficient code may not always be the best. Your ability to understand and maintain the code matters too.

### Solution #1: If hasWon is called many times

There are only  $3^9$ , or about 20,000 tic-tac-toe boards (assuming a 3x3 board). We can therefore represent our tic-tac-toe board as an int, with each digit representing a piece (0 means Empty, 1 means Red, 2 means Blue). We set up a hash table or array in advance with all possible boards as keys and the value indicating who has won. Our function then is simply this:

```
1 public int hasWon(int board) {  
2     return winnerHashtable[board];  
3 }
```

To convert a board (represented by a char array) to an int, we can use what is essentially a “base 3” representation. Each board is represented as  $3^0 v_0 + 3^1 v_1 + 3^2 v_2 + \dots + 3^8 v_8$ , where  $v_i$  is a 0 if the space is empty, a 1 if it’s a “blue spot” and a 2 if it’s a “red spot.”

```
1 public static int convertBoardToInt(char[][] board) {  
2     int factor = 1;  
3     int sum = 0;  
4     for (int i = 0; i < board.length; i++) {  
5         for (int j = 0; j < board[i].length; j++) {  
6             int v = 0;  
7             if (board[i][j] == 'x') {  
8                 v = 1;  
9             } else if (board[i][j] == 'o') {  
10                 v = 2;  
11             }  
12             sum += v * factor;  
13             factor *= 3;  
14         }  
15     }  
16     return sum;  
17 }
```

Now, looking up the winner of a board is just a matter of looking it up in a hash table.

Of course, if we need to convert a board into this format every time we want to check for a winner, we haven’t saved ourselves any time compared with the other solutions. But, if we can store the board this way from the very beginning, then the lookup process will be very efficient.

### Solution #2: Designing for just a 3x3 board

If we really only want to implement a solution for a 3x3 board, the code is relatively short and simple. The only complex part is trying to be clean and organized, without writing too much duplicated code.

```
1 Piece hasWon1(Piece[][] board) {  
2     for (int i = 0; i < board.length; i++) {  
3         /* Check Rows */  
4         if (board[i][0] != Piece.Empty &&  
5             board[i][0] == board[i][1] &&
```

```

6         board[i][0] == board[i][2]) {
7     return board[i][0];
8 }
9
10    /* Check Columns */
11    if (board[0][i] != Piece.Empty &&
12        board[0][i] == board[1][i] &&
13        board[0][i] == board[2][i]) {
14    return board[0][i];
15 }
16 }
17
18    /* Check Diagonal */
19    if (board[0][0] != Piece.Empty &&
20        board[0][0] == board[1][1] &&
21        board[0][0] == board[2][2]) {
22    return board[0][0];
23 }
24
25    /* Check Reverse Diagonal */
26    if (board[2][0] != Piece.Empty &&
27        board[2][0] == board[1][1] &&
28        board[2][0] == board[0][2]) {
29    return board[2][0];
30 }
31
32 } 
```

**Solution #3: Designing for an NxN board**

This is a straightforward extension of the code for a 3x3 board. The code attachment provides several other ways, but we have listed one below.

```

1 Piece hasWon3(Piece[][] board) {
2     int N = board.length;
3     int row = 0;
4     int col = 0;
5
6     /* Check rows */
7     for (row = 0; row < N; row++) {
8         if (board[row][0] != Piece.Empty) {
9             for (col = 1; col < N; col++) {
10                 if (board[row][col] != board[row][col-1]) break;
11             }
12             if (col == N) return board[row][0];
13         }
14     }
15
16     /* Check columns */
17     for (col = 0; col < N; col++) { 
```

```
18     if (board[0][col] != Piece.Empty) {
19         for (row = 1; row < N; row++) {
20             if (board[row][col] != board[row-1][col]) break;
21         }
22         if (row == N) return board[0][col];
23     }
24 }
25
26 /* Check diagonal (top left to bottom right) */
27 if (board[0][0] != Piece.Empty) {
28     for (row = 1; row < N; row++) {
29         if (board[row][row] != board[row-1][row-1]) break;
30     }
31     if (row == N) return board[0][0];
32 }
33
34 /* Check diagonal (bottom left to top right) */
35 if (board[N-1][0] != Piece.Empty) {
36     for (row = 1; row < N; row++) {
37         if (board[N-row-1][row] != board[N-row][row-1]) break;
38     }
39     if (row == N) return board[N-1][0];
40 }
41
42 return Piece.Empty;
43 }
```

Regardless of how you solve the problem, the algorithms for the problem are not especially challenging. The tricky part is understanding how to code in a clean and maintainable way, and this is exactly what your interviewer is attempting to assess.

### 17.3 Write an algorithm which computes the number of trailing zeros in $n$ factorial.

pg 163

#### SOLUTION

A simple approach is to compute the factorial, and then count the number of trailing zeros by continuously dividing by ten. The problem with this though is that the bounds of an `int` would be exceeded very quickly. To avoid this issue, we can look at this problem mathematically.

Consider a factorial like  $19!$ :

$$19! = 1 * 2 * 3 * 4 * 5 * 6 * 7 * 8 * 9 * 10 * 11 * 12 * 13 * 14 * 15 * 16 * 17 * 18 * 19$$

A trailing zero is created with multiples of 10, and multiples of 10 are created with pairs of 5-multiples and 2-multiples.

For example, in  $19!$ , the following terms create the trailing zeros:

$$19! = 2 * \dots * 5 * \dots * 10 * \dots * 15 * 16 * \dots$$

To count the number of zeros, therefore, we only need to count the pairs of multiples of 5 and 2. There will always be more multiples of 2 than 5 though, so simply counting the number of multiples of 5 is sufficient.

One “gotcha” here is 15 contributes a multiple of 5 (and therefore one trailing zero), while 25 contributes two (because  $25 = 5 * 5$ ).

There are two different ways of writing this code.

The first way is to iterate through all the numbers from 2 through n, counting the number of times that 5 goes into each number.

```

1  /* If the number is a 5 or five, return which power of 5.
2   *      5 -> 1,
3   *      25-> 2, etc.
4 */
5  public int factorsOf5(int i) {
6      int count = 0;
7      while (i % 5 == 0) {
8          count++;
9          i /= 5;
10     }
11     return count;
12 }
13
14 public int countFactZeros(int num) {
15     int count = 0;
16     for (int i = 2; i <= num; i++) {
17         count += factorsOf5(i);
18     }
19     return count;
20 }
```

This isn't bad, but we can make it a little more efficient by directly counting the factors of 5. Under this approach, we would first count the number of multiples of 5 that are between 1 and n (which is  $n/5$ ), then the number of multiples of 25 ( $n/25$ ), then 125, and so on.

To count how many multiples of m are in n, we can just divide n by m.

```

1  public int countFactZeros(int num) {
2      int count = 0;
3      if (num < 0) {
4          return -1;
5      }
6      for (int i = 5; num / i > 0; i *= 5) {
7          count += num / i;
8      }
9      return count;
10 }
```

This problem is a bit of a brain teaser, but it can be approached logically (as shown above). By thinking through what exactly will contribute a zero, you can come up with a solution. You should be very clear in your rules upfront so that you can implement this correctly.

- 17.4** Write a method which finds the maximum of two numbers. You should not use if-else or any other comparison operator.

pg 163

### SOLUTION

A common way of implementing a max function is to look at the sign of  $a - b$ . In this case, we can't use a comparison operator on this sign, but we *can* use multiplication.

Let  $k$  equal the sign of  $a - b$  such that if  $a - b \geq 0$ , then  $k$  is 1. Else,  $k = 0$ . Let  $q$  be the inverse of  $k$ .

We can then implement the code as follows:

```
1  /* Flips a 1 to a 0 and a 0 to a 1 */
2  public static int flip(int bit) {
3      return 1^bit;
4  }
5
6  /* Returns 1 if a is positive, and 0 if a is negative */
7  public static int sign(int a) {
8      return flip((a >> 31) & 0x1);
9  }
10
11 public static int getMaxNaive(int a, int b) {
12     int k = sign(a - b);
13     int q = flip(k);
14     return a * k + b * q;
15 }
```

This code almost works. It fails, unfortunately, when  $a - b$  overflows. Suppose, for example, that  $a$  is `INT_MAX - 2` and  $b$  is `-15`. In this case,  $a - b$  will be greater than `INT_MAX` and will overflow, resulting in a negative value.

We can implement a solution to this problem by using the same approach. Our goal is to maintain the condition where  $k$  is 1 when  $a > b$ . We will need to use more complex logic to accomplish this.

When does  $a - b$  overflow? It will overflow only when  $a$  is positive and  $b$  is negative, or the other way around. It may be difficult to specially detect the overflow condition, but we *can* detect when  $a$  and  $b$  have different signs. Note that if  $a$  and  $b$  have different signs, then we want  $k$  to equal `sign(a)`.

The logic looks like:

```

1 if a and b have different signs:
2   // if a > 0, then b < 0, and k = 1.
3   // if a < 0, then b > 0, and k = 0.
4   // so either way, k = sign(a)
5   let k = sign(a)
6 else
7   let k = sign(a - b) // overflow is impossible

```

The code below implements this, using multiplication instead of if-statements.

```

1 public static int getMax(int a, int b) {
2     int c = a - b;
3
4     int sa = sign(a); // if a >= 0, then 1 else 0
5     int sb = sign(b); // if b >= 1, then 1 else 0
6     int sc = sign(c); // depends on whether or not a - b overflows
7
8     /* Goal: define a value k which is 1 if a > b and 0 if a < b.
9      * (if a = b, it doesn't matter what value k is) */
10
11    // If a and b have different signs, then k = sign(a)
12    int use_sign_of_a = sa ^ sb;
13
14    // If a and b have the same sign, then k = sign(a - b)
15    int use_sign_of_c = flip(sa ^ sb);
16
17    int k = use_sign_of_a * sa + use_sign_of_c * sc;
18    int q = flip(k); // opposite of k
19
20    return a * k + b * q;
21 }

```

Note that for clarity, we split up the code into many different methods and variables. This is certainly not the most compact or efficient way to write it, but it does make what we're doing much cleaner.

### 17.5 The Game of Master Mind is played as follows:

The computer has four slots, and each slot will contain a ball that is red (R), yellow (Y), green (G) or blue (B). For example, the computer might have RGGB (Slot #1 is red, Slots #2 and #3 are green, Slot #4 is blue).

You, the user, are trying to guess the solution. You might, for example, guess YRGB.

When you guess the correct color for the correct slot, you get a "hit." If you guess a color that exists but is in the wrong slot, you get a "pseudo-hit." Note that a slot that is a hit can never count as a pseudo-hit.

For example, if the actual solution is RGBY and you guess GGRR, you have one hit and one pseudo-hit.

Write a method that, given a guess and a solution, returns the number of hits and pseudo-hits.

pg 163

### SOLUTION

This problem is straightforward, but it's surprisingly easy to make little mistakes. You should check your code extremely thoroughly, on a variety of test cases.

We'll implement this code by first creating a frequency array which stores how many times each character occurs in solution, excluding times when the slot is a "hit." Then, we iterate through guess to count the number of pseudo-hits.

The code below implements this algorithm.

```
1  public class Result {  
2      public int hits = 0;  
3      public int pseudoHits = 0;  
4  
5      public String toString() {  
6          return "(" + hits + ", " + pseudoHits + ")";  
7      }  
8  }  
9  
10 public int code(char c) {  
11     switch (c) {  
12         case 'B':  
13             return 0;  
14         case 'G':  
15             return 1;  
16         case 'R':  
17             return 2;  
18         case 'Y':  
19             return 3;  
20         default:  
21             return -1;
```

```

22     }
23 }
24
25 public static int MAX_COLORS = 4;
26
27 public Result estimate(String guess, String solution) {
28     if (guess.length() != solution.length()) return null;
29
30     Result res = new Result();
31     int[] frequencies = new int[MAX_COLORS];
32
33     /* Compute hits and build frequency table */
34     for (int i = 0; i < guess.length(); i++) {
35         if (guess.charAt(i) == solution.charAt(i)) {
36             res.hits++;
37         } else {
38             /* Only increment the frequency table (which will be used
39              * for pseudo-hits) if it's not a hit. If it's a hit, the
40              * slot has already been "used." */
41             int code = code(solution.charAt(i));
42             frequencies[code]++;
43         }
44     }
45
46     /* Compute pseudo-hits */
47     for (int i = 0; i < guess.length(); i++) {
48         int code = code(guess.charAt(i));
49         if (code >= 0 && frequencies[code] > 0 &&
50             guess.charAt(i) != solution.charAt(i)) {
51             res.pseudoHits++;
52             frequencies[code]--;
53         }
54     }
55     return res;
56 }
```

Note that the easier the algorithm-for a problem is, the more important it is to write clean and correct code. In this case, we've pulled `code(char c)` into its own method, and we've created a `Result` class to hold the result, rather than just printing it.

- 17.6** Given an array of integers, write a method to find indices  $m$  and  $n$  such that if you sorted elements  $m$  through  $n$ , the entire array would be sorted. Minimize  $n - m$  (that is, find the smallest such sequence).

pg 164

## SOLUTION

Before we begin, let's make sure we understand what our answer will look like. If we're

looking for just two indices, this indicates that some middle section of the array will be sorted, with the start and end of the array already being in order.

Now, let's approach this problem by looking at an example.

1, 2, 4, 7, 10, 11, 7, 12, 6, 7, 16, 18, 19

Our first thought might be to just find the longest increasing subsequence at the beginning and the longest increasing subsequence at the end.

```
left: 1, 2, 4, 7, 10, 11  
middle: 7, 12  
right: 6, 7, 16, 18, 19
```

These subsequences are easy to generate. We just start from the left and the right sides, and work our way inward. When an element is out of order, then we have found the end of our increasing / decreasing subsequence.

In order to solve our problem though, we would need to be able to sort the middle part of the array and, by doing just that, get all the elements in the array in order. Specifically, the following would have to be true:

```
/* all items on left are smaller than all items in middle */  
min(middle) > end(left)  
  
/* all items in middle are smaller than all items in right */  
max(middle) < start(right)
```

Or, in other words, for all elements:

```
left < middle < right
```

In fact, this condition will *never* be met. The middle section is, by definition, the elements that were out of order. That is, it is *always* the case that `left.end > middle.start` and `middle.end > right.start`. Thus, you cannot sort the middle to make the entire array sorted.

But, what we can do is *shrink* the left and right subsequences until the earlier conditions are met.

Let `min` equal `min(middle)` and `max` equal `max(middle)`.

On the left side, we start with the end of the subsequence (value 11, at element 5) and move to the left. Once we find an element `i` such that `array[i] < min`, we know that we could sort the middle and have that part of the array appear in order.

Then, we do a similar thing on the right side. The value `max` equals 12. So, we begin with the start of the right subsequence (value 6) and move to the right. We compare the max of 12 to 6, then 7, then 16. When reach 16, we know that no elements smaller than 12 could be after it (since it's an increasing subsequence). Thus, the middle of the array could now be sorted to make the entire array sorted.

The following code implements this algorithm.

```

1 int findEndOfLeftSubsequence(int[] array) {
2     for (int i = 1; i < array.length; i++) {
3         if (array[i] < array[i - 1]) return i - 1;
4     }
5     return array.length - 1;
6 }
7
8 int findStartOfRightSubsequence(int[] array) {
9     for (int i = array.length - 2; i >= 0; i--) {
10        if (array[i] > array[i + 1]) return i + 1;
11    }
12    return 0;
13 }
14
15 int shrinkLeft(int[] array, int min_index, int start) {
16     int comp = array[min_index];
17     for (int i = start - 1; i >= 0; i--) {
18         if (array[i] <= comp) return i + 1;
19     }
20     return 0;
21 }
22
23 int shrinkRight(int[] array, int max_index, int start) {
24     int comp = array[max_index];
25     for (int i = start; i < array.length; i++) {
26         if (array[i] >= comp) return i - 1;
27     }
28     return array.length - 1;
29 }
30
31 void findUnsortedSequence(int[] array) {
32     /* find left subsequence */
33     int end_left = findEndOfLeftSubsequence(array);
34
35     /* find right subsequence */
36     int start_right = findStartOfRightSubsequence(array);
37
38     /* find min and max element of middle */
39     int min_index = end_left + 1;
40     if (min_index >= array.length) return; // Already sorted
41
42     int max_index = start_right - 1;
43     for (int i = end_left; i <= start_right; i++) {
44         if (array[i] < array[min_index]) min_index = i;
45         if (array[i] > array[max_index]) max_index = i;
46     }
47
48     /* slide left until less than array[min_index] */
49     int left_index = shrinkLeft(array, min_index, end_left);
50

```

```
51     /* slide right until greater than array[max_index] */  
52     int right_index = shrinkRight(array, max_index, start_right);  
53  
54     System.out.println(left_index + " " + right_index);  
55 }
```

Note the use of other methods in this solution. Although we could have jammed it all into one method, it would have made the code a lot harder to understand, maintain, and test. In your interview coding, you should prioritize these aspects.

- 17.7** Given any integer, print an English phrase that describes the integer (e.g., “One Thousand, Two Hundred Thirty Four”).

pg 164

### SOLUTION

This is not an especially challenging problem, but it is a somewhat tedious one. The key is to be organized in how you approach the problem—and to make sure you have good test cases.

We can think about converting a number like 19,323,984 as converting each of three 3-digit segments of the number, and inserting “thousands” and “millions” in between as appropriate. That is,

$$\text{convert}(19,323,984) = \text{convert}(19) + \text{“ million ”} + \\ \text{convert}(323) + \text{“ thousand ”} + \\ \text{convert}(984)$$

The code below implements this algorithm.

```
1  public String[] digits = {"One", "Two", "Three", "Four", "Five",  
2      "Six", "Seven", "Eight", "Nine"};  
3  public String[] teens = {"Eleven", "Twelve", "Thirteen",  
4      "Fourteen", "Fifteen", "Sixteen", "Seventeen", "Eighteen",  
5      "Nineteen"};  
6  public static String[] tens = {"Ten", "Twenty", "Thirty", "Forty",  
7      "Fifty", "Sixty", "Seventy", "Eighty", "Ninety"};  
8  public static String[] bigs = {"", "Thousand", "Million"};  
9  
10 public static String numToString(int number) {  
11     if (number == 0) {  
12         return "Zero";  
13     } else if (number < 0) {  
14         return "Negative " + numToString(-1 * number);  
15     }  
16  
17     int count = 0;  
18     String str = "";  
19  
20     while (number > 0) {
```

```

21     if (number % 1000 != 0) {
22         str = numToString100(number % 1000) + bigs[count] +
23             " " + str;
24     }
25     number /= 1000;
26     count++;
27 }
28
29     return str;
30 }
31
32 public static String numToString100(int number) {
33     String str = "";
34
35     /* Convert hundreds place */
36     if (number >= 100) {
37         str += digits[number / 100 - 1] + " Hundred ";
38         number %= 100;
39     }
40
41     /* Convert tens place */
42     if (number >= 11 && number <= 19) {
43         return str + teens[number - 11] + " ";
44     } else if (number == 10 || number >= 20) {
45         str += tens[number / 10 - 1] + " ";
46         number %= 10;
47     }
48
49     /* Convert ones place */
50     if (number >= 1 && number <= 9) {
51         str += digits[number - 1] + " ";
52     }
53
54     return str;
55 }

```

The key in a problem like this is to make sure you consider all the special cases. There are a lot of them.

- 17.8** You are given an array of integers (both positive and negative). Find the contiguous sequence with the largest sum. Return the sum.

pg 164

### SOLUTION

This is a challenging problem, but an extremely common one. Let's approach this by looking at an example:

2    3    -8    -1    2    4    -2    3

If we think about our array as having alternating sequences of positive and negative numbers, we can observe that we would never include only part of a negative subsequence or part of a positive sequence. Why would we? Including part of a negative subsequence would make things unnecessarily negative, and we should just instead not include that negative sequence at all. Likewise, including only part of a positive subsequence would be strange, since the sum would be even bigger if we included the whole thing.

For the purposes of coming up with our algorithm, we can think about our array as being a sequence of alternating negative and positive *numbers*. Each number corresponds to the sum of a subsequence of positive numbers of a subsequence of negative numbers. For the array above, our new reduced array would be:

5   -9   6   -2   3

This doesn't give away a great algorithm immediately, but it does help us to better understand what we're working with.

Consider the array above. Would it ever make sense to have {5, -9} in a subsequence? No. These numbers sum to -4, so we're better off not including either number, or possibly just having the sequence be just {5}).

When would we want negative numbers included in a subsequence? Only if it allows us to join two positive subsequences, each of which have a sum greater than the negative value.

We can approach this in a step-wise manner, starting with the first element in the array.

When we look at 5, this is the biggest sum we've seen so far. We set `maxSum` to 5, and `sum` to 5. Then, we consider -9. If we added it to `sum`, we'd get a negative value. There's no sense in extending the subsequence from 5 to -9 (which "reduces" to a sequence of just -4), so we just reset the value of `sum`.

Now, we consider 6. This subsequence is greater than 5, so we update both `maxSum` and `sum`.

Next, we look at -2. Adding this to 6 will set `sum` to 4. Since this is still a "value add" (when adjoined to another, bigger sequence), we *might* want {6, -2} in our max subsequence. We'll update `sum`, but not `maxSum`.

Finally, we look at 3. Adding 3 to `sum` (4) gives us 7, so we update `maxSum`. The max subsequence is therefore the sequence {6, -2, 3}.

When we look at this in the full expanded array, our logic is identical. The code below implements this algorithm.

```
1  public static int getMaxSum(int[] a) {
2      int maxsum = 0;
3      int sum = 0;
4      for (int i = 0; i < a.length; i++) {
5          sum += a[i];
```

```

6     if (maxsum < sum) {
7         maxsum = sum;
8     } else if (sum < 0) {
9         sum = 0;
10    }
11 }
12 return maxsum;
13 }

```

If the array is all negative numbers, what is the correct behavior? Consider this simple array `{-3, -10, -5}`. You could make a good argument that the maximum sum is either:

1. -3 (if you assume the subsequence can't be empty)
2. 0 (the subsequence has length 0)
3. MINIMUM\_INT (essentially the error case).

We went with option #2 (`maxSum = 0`), but there's no "correct" answer. This is a great thing to discuss with your interviewer; it will show how detail oriented you are.

### **17.9 Design a method to find the frequency of occurrences of any given word in a book.**

pg 164

#### **SOLUTION**

The first question that you should ask your interviewer is if you'll be doing this operation once or repeatedly. That is, are you simply asking for the frequency of "dog," or might you ask for "dog," and then "cat," "mouse," and so on?

##### **Solution: Single Query**

In this case, we simply go through the book, word by word, and count the number of times that a word appears. This will take  $O(n)$  time. We know we can't do better than that since we must look at every word in the book.

##### **Solution: Repetitive Queries**

If we're doing the operation repeatedly, then we can probably afford to take some time and extra memory to do pre-processing on the book. We can create a hash table which maps from a word to its frequency. The frequency of any word can be easily looked up in  $O(1)$  time. The code for this is below.

```

1 Hashtable<String, Integer> setupDictionary(String[] book) {
2     Hashtable<String, Integer> table =
3         new Hashtable<String, Integer>();
4     for (String word : book) {
5         word = word.toLowerCase();

```

```
6     if (word.trim() != "") {
7         if (!table.containsKey(word)) {
8             table.put(word, 0);
9         }
10        table.put(word, table.get(word) + 1);
11    }
12 }
13 return table;
14 }
15
16 int getFrequency(Hashtable<String, Integer> table, String word) {
17     if (table == null || word == null) return -1;
18     word = word.toLowerCase();
19     if (table.containsKey(word)) {
20         return table.get(word);
21     }
22     return 0;
23 }
```

Note that a problem like this is actually relatively easy. Thus, the interviewer is going to be looking heavily at how careful you are. Did you check for error conditions?

- 17.10** Since XML is very verbose, you are given a way of encoding it where each tag gets mapped to a pre-defined integer value. The language/grammar is as follows:

Element	-->	Tag Attributes END Children END
Attribute	-->	Tag Value
END	-->	0
Tag	-->	some predefined mapping to int
Value	-->	string value END

For example, the following XML might be converted into the compressed string below (assuming a mapping of family -> 1, person -> 2, firstName -> 3, LastName -> 4, state -> 5).

```
<family lastName="McDowell" state="CA">
    <person firstName="Gayle">Some Message</person>
</family>
```

Becomes:

```
1 4 McDowell 5 CA 0 2 3 Gayle 0 Some Message 0 0
```

Write code to print the encoded version of an XML element (passed in Element and Attribute objects).

pg 164

### SOLUTION

Since we know the element will be passed in as an Element and Attribute, our code is reasonably simple. We can implement this by applying a tree-like approach.

We repeatedly call `encode()` on parts of the XML structure, handling the code in slightly different ways depending on the type of the XML element.

```

1  public static void encode(Element root, StringBuffer sb) {
2      encode(root.getNameCode(), sb);
3      for (Attribute a : root.attributes) {
4          encode(a, sb);
5      }
6      encode("0", sb);
7      if (root.value != null && root.value != "") {
8          encode(root.value, sb);
9      } else {
10         for (Element e : root.children) {
11             encode(e, sb);
12         }
13     }
14     encode("0", sb);
15 }
16
17 public static void encode(String v, StringBuffer sb) {
18     sb.append(v);
19     sb.append(" ");
20 }
21
22 public static void encode(Attribute attr, StringBuffer sb) {
23     encode(attr.getTagCode(), sb);
24     encode(attr.value, sb);
25 }
26
27 public static String encodeToString(Element root) {
28     StringBuffer sb = new StringBuffer();
29     encode(root, sb);
30     return sb.toString();
31 }
```

Observe in line 17, the use of the very simple `encode` method for a string. This is somewhat unnecessary; all it does is insert the string and a space following it. However, using this method is a nice touch as it ensures that every element will be inserted with a space surrounding it. Otherwise, it might be easy to break the encoding by forgetting to append the empty string.

- 17.11** Implement a method `rand7()` given `rand5()`. That is, given a method that generates a random number between 0 and 4 (inclusive), write a method that generates a random number between 0 and 6 (inclusive).

pg 165

## SOLUTION

To implement this function correctly, we must have each of the values between 0 and 6

returned with 1/7th probability.

### First Attempt (Fixed Number of Calls)

As a first attempt, we might try to generate all numbers between 0 and 9, and then mod the resulting value by 7. Our code for it might look something like this:

```
1 int rand7() {
2     int v = rand5() + rand5();
3     return v % 7;
4 }
```

Unfortunately, the above code will not generate the values with equal probability. We can see this by looking at how the results of each call to `rand5()` will correspond to the return result of the `rand7()` function.

1st Call	2nd Call	Result
0	0	0
0	1	1
0	2	2
0	3	3
0	4	4
1	0	1
1	1	2
1	2	3
1	3	4
1	4	5
2	0	2
2	1	3
2	2	4

1st Call	2nd Call	Result
2	3	5
2	4	6
3	0	3
3	1	4
3	2	5
3	3	6
3	4	0
4	0	4
4	1	5
4	2	6
4	3	0
4	4	1

Each individual row has a 1 in 25 chance of occurring, since there are two calls to `rand5()` and each distributes its results with 1/5th probability. If you count up the number of times each number occurs, you'll note that this `rand7()` function will return 4 with 5/25th probability but return 0 with just 3/25th probability. This means that our function has failed; the results do not have probability 1/7th.

Now, imagine we modify our function to add an if-statement, to change the constant multiplier, or to insert a new call to `rand5()`. We will still wind up with a similar looking table, and the probability of getting any one of those rows will be  $1/5^k$ , where  $k$  is the number of calls to `rand5()` in that row. Different rows may have different number of calls.

The probability of winding up with the result of the `rand7()` function being, say, 6

would be the sum of the probabilities of all rows that result in 6. That is:

$$P(\text{rand7}() = 6) = 1/5^1 + 1/5^3 + \dots + 1/5^6$$

We know that, in order for our function to be correct, this probability must equal 1/7.

This is impossible though. Because 5 and 7 are relatively prime, no series of reciprocal powers of 5 will result in 1/7.

Does this mean the problem is impossible? Not exactly. Strictly speaking, it means that, as long as we can list out the combinations of `rand5()` results that will result in a particular value of `rand7()`, the function will not give well distributed results.

We can still solve this problem. We just have to use a while loop, and realize that there's no telling just how many turns will be required to return a result.

### Second Attempt (Nondeterministic Number of Calls)

As soon as we've allowed for a while loop, our work gets much easier. We just need to generate a range of values where each value is equally likely (and where the range has at least seven elements). If we can do this, then we can discard the elements greater than the previous multiple of 7, and mod the rest of them by 7. This will get us a value within the range of 0 to 6, with each value being equally likely.

In the below code, we generate the range 0 through 24 by doing `5 * rand5() + rand5()`. Then, we discard the values between 21 and 24, since they would otherwise make `rand7()` unfairly weighted towards 0 through 3. Finally, we mod by 7 to give us the values in the range 0 to 6 with equal probability.

Note that because we discard values in this approach, we have no guarantee on the number of `rand5()` calls it may take to return a value. This is what is meant by a *nondeterministic* number of calls.

```

1 public static int rand7() {
2     while (true) {
3         int num = 5 * rand5() + rand5();
4         if (num < 21) {
5             return num % 7;
6         }
7     }
8 }
```

Observe that doing `5 * rand5() + rand5()` gives us exactly one way of getting each number in its range (0 to 24). This ensures that each value is equally probable.

Could we instead do `2 * rand5() + rand5()`? No, because the values wouldn't be equally distributed. For example, there would be two ways of getting a 6 ( $6=2*1+4$  and  $6=2*2+2$ ) but only one way of getting a 0 ( $0=2*0+0$ ). The values in the range are not equally probable.

There is a way that we can use `2 * rand5()` and still get an identically distributed

range, but it's much more complicated. See below.

```
1 public int rand7() {  
2     while (true) {  
3         int r1 = 2 * rand5(); /* evens between 0 and 9 */  
4         int r2 = rand5(); /* used later to generate a 0 or 1 */  
5         if (r2 != 4) { /* r2 has extra even num-discard the extra */  
6             int rand1 = r2 % 2; /* Generate 0 or 1 */  
7             int num = r1 + rand1; /* will be in the range 0 to 9 */  
8             if (num < 7) {  
9                 return num;  
10            }  
11        }  
12    }  
13 }
```

In fact, there is an infinite number of ranges we can use. The key is to make sure that the range is big enough and that all values are equally likely.

- 17.12** Design an algorithm to find all pairs of integers within an array which sum to a specified value.

pg 165

### SOLUTION

We can approach this problem in two ways. The “preferred” solution depends on your preferences between time efficiency, space efficiency, and code complexity.

#### Simple Solution

One easy and (time) efficient solution involves a hash map from integers to integers. This algorithm works by iterating through the array. On each element  $x$ , look up  $\text{sum} - x$  in the hash table and, if it exists, print  $(x, \text{sum} - x)$ . Add  $x$  to the hash table, and go to the next element.

#### Alternate Solution

First, let's start with a definition. If we're trying to find a pair of numbers that sums to  $z$ , the *complement* of  $x$  will be  $z - x$  (that is, the number that can be added to  $x$  to make  $z$ ). For example, if we're trying to find a pair of numbers that sums to 12, the complement of -5 would be 17.

Now, imagine we have the following sorted array:  $\{-2 \ -1 \ 0 \ 3 \ 5 \ 6 \ 7 \ 9 \ 13 \ 14\}$ . Let `first` point to the head of the array and `last` point to the end of the array. To find the complement of `first`, we just move `last` backwards until we find it. If `first + last < sum`, then there is no complement for `first`. We can therefore move `first` forward. We stop when `first` is greater than `last`.

Why must this find all complements for `first`? Because the array is sorted and we're trying progressively smaller numbers. When the sum of `first` and `last` is less than the sum, we know that trying even smaller numbers (as `last`) won't help us find a complement.

Why must this find all complements for `last`? Because all pairs must be made up of a `first` and a `last`. We've found all complements for `first`, therefore we've found all complements of `last`.

```

1 void printPairSums(int[] array, int sum) {
2     Arrays.sort(array);
3     int first = 0;
4     int last = array.length - 1;
5     while (first < last) {
6         int s = array[first] + array[last];
7         if (s == sum) {
8             System.out.println(array[first] + " " + array[last]);
9             first++;
10            last--;
11        } else {
12            if (s < sum) first++;
13            else last--;
14        }
15    }
16 }
```

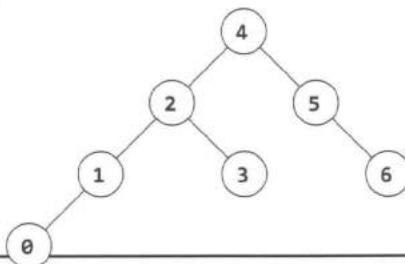
- 17.13** Consider a simple node-like data structure called `BiNode`, which has pointers to two other nodes. The data structure `BiNode` could be used to represent both a binary tree (where `node1` is the left node and `node2` is the right node) or a doubly linked list (where `node1` is the previous node and `node2` is the next node). Implement a method to convert a binary search tree (implemented with `BiNode`) into a doubly linked list. The values should be kept in order and the operation should be performed in place (that is, on the original data structure).

pg 165

## SOLUTION

This seemingly complex problem can be implemented quite elegantly using recursion. You will need to understand recursion very well to solve it.

Picture a simple binary search tree:



The convert method should transform it into the below doubly linked list:

```
0 <-> 1 <-> 2 <-> 3 <-> 4 <-> 5 <-> 6
```

Let's approach this recursively, starting with the root (node 4).

We know that the left and right halves of the tree form their own "sub-parts" of the linked list (that is, they appear consecutively in the linked list). So, if we recursively converted the left and right subtrees to a doubly linked list, could we build the final linked list from those parts?

Yes! We would simply merge the different parts.

The pseudocode looks something like:

```
1 BiNode convert(BiNode node) {  
2     BiNode left = convert(node.left);  
3     BiNode right = convert(node.right);  
4     mergeLists(left, node, right);  
5     return left; // front of left  
6 }
```

To actually implement the nitty-gritty details of this, we'll need to get the head and tail of each linked list. We can do this several different ways.

### Solution #1: Additional Data Structure

The first, and easier, approach is to create a new data structure called `NodePair` which holds just the head and tail of a linked list. The convert method can then return something of type `NodePair`.

The code below implements this approach.

```
1 private class NodePair {  
2     BiNode head;  
3     BiNode tail;  
4  
5     public NodePair(BiNode head, BiNode tail) {  
6         this.head = head;  
7         this.tail = tail;  
8     }  
9 }  
10  
11 public NodePair convert(BiNode root) {  
12     if (root == null) {  
13         return null;  
14     }  
15  
16     NodePair part1 = convert(root.node1);  
17     NodePair part2 = convert(root.node2);  
18  
19     if (part1 != null) {
```

```

20     concat(part1.tail, root);
21 }
22
23 if (part2 != null) {
24     concat(root, part2.head);
25 }
26
27 return new NodePair(part1 == null ? root : part1.head,
28                     part2 == null ? root : part2.tail);
29 }
30
31 public static void concat(BiNode x, BiNode y) {
32     x.node2 = y;
33     y.node1 = x;
34 }
```

The above code still converts the BiNode data structure in place. We're just using NodePair as a way to return additional data. We could have alternatively used a two-element BiNode array to fulfill the same purposes, but it looks a bit messier (and we like clean code, especially in an interview).

It'd be nice, though, if we could do this without these extra data structures—and we can.

### Solution #2: Retrieving the Tail

Instead of returning the head and tail of the linked list with NodePair, we can return just the head, and then we can use the head to find the tail of the linked list.

```

1  public static BiNode convert(BiNode root) {
2      if (root == null) {
3          return null;
4      }
5
6      BiNode part1 = convert(root.node1);
7      BiNode part2 = convert(root.node2);
8
9      if (part1 != null) {
10          concat(getTail(part1), root);
11      }
12
13      if (part2 != null) {
14          concat(root, part2);
15      }
16
17      return part1 == null ? root : part1;
18 }
19
20 public static BiNode getTail(BiNode node) {
21     if (node == null) return null;
```

```
22     while (node.node2 != null) {  
23         node = node.node2;  
24     }  
25     return node;  
26 }
```

Other than a call to `getTail`, this code is almost identical to the first solution. It is not, however, very efficient. A leaf node at depth  $d$  will be “touched” by the `getTail` method  $d$  times (one for each node above it), leading to an  $O(N^2)$  overall runtime, where  $N$  is the number of nodes in the tree.

### Solution #3: Building a Circular Linked List

We can build our third and final approach off of the second one.

This approach requires returning the head and tail of the linked list with `BiNode`. We can do this by returning each list as the head of a *circular* linked list. To get the tail, then, we simply call `head.node1`.

```
1  public static BiNode convertToCircular(BiNode root) {  
2      if (root == null) {  
3          return null;  
4      }  
5  
6      BiNode part1 = convertToCircular(root.node1);  
7      BiNode part3 = convertToCircular(root.node2);  
8  
9      if (part1 == null && part3 == null) {  
10         root.node1 = root;  
11         root.node2 = root;  
12         return root;  
13     }  
14     BiNode tail3 = (part3 == null) ? null : part3.node1;  
15  
16     /* join left to root */  
17     if (part1 == null) {  
18         concat(part3.node1, root);  
19     } else {  
20         concat(part1.node1, root);  
21     }  
22  
23     /* join right to root */  
24     if (part3 == null) {  
25         concat(root, part1);  
26     } else {  
27         concat(root, part3);  
28     }  
29  
30     /* join right to left */  
31     if (part1 != null && part3 != null) {
```

```

32     concat(tail3, part1);
33 }
34
35 return part1 == null ? root : part1;
36 }
37
38 /* Convert list to a circular linked list, and then break the
39 * circular connection. */
40 public static BiNode convert(BiNode root) {
41     BiNode head = convertToCircular(root);
42     head.node1.node2 = null;
43     head.node1 = null;
44     return head;
45 }

```

Observe that we have moved the main parts of the code into `convertToCircular`. The `convert` method calls this method to get the head of the circular linked list, and then breaks the circular connection.

The approach takes  $O(N)$  time, since each node is only touched an average of once (or, more accurately,  $O(1)$  times).

**17.14** *Oh, no! You have just completed a lengthy document when you have an unfortunate Find/Replace mishap. You have accidentally removed all spaces, punctuation, and capitalization in the document. A sentence like "I reset the computer. It still didn't boot!" would become "iresetthecomputeritstilldidntboot". You figure that you can add back in the punctuation and capitalization later, once you get the individual words properly separated. Most of the words will be in a dictionary, but some strings, like proper names, will not.*

*Given a dictionary (a list of words), design an algorithm to find the optimal way of "unconcatenating" a sequence of words. In this case, "optimal" is defined to be the parsing which minimizes the number of unrecognized sequences of characters.*

*For example, the string "jesslookedjustliketimherbrother" would be optimally parsed as "JESS looked just like TIM her brother". This parsing has seven unrecognized characters, which we have capitalized for clarity.*

pg 165

## SOLUTION

---

Some interviewers like to cut to the chase and give you the specific problems. Others, though, like to give you a lot of unnecessary context, like this problem has. It's useful in such cases to boil down the problem to what it's really all about.

In this case, the problem is really about finding a way to break up a string into separate words such that as few characters as possible are "left out" of the parsing.

Note that we do not attempt to “understand” the string. We could just as well parse “thisisawesome” to be “this is a we some” as we could “this is awesome.”

The key to this problem is to find a way to define the solution (that is, parsed string) in terms of its subproblems. One way to do this is to recurse through the string. At each point, the optimal parsing is the better of two possible decisions:

1. Inserting a space after this character.
2. Not inserting a space after this character.

We'll walk through this problem for the string **thit** as shown below. For clarity, we will use the following notation:

- **CAPITALIZE** the invalid (non-dictionary) words
- Underline valid words
- **Bold** characters that are adjoined to each other (since they have no space between them).

These **bold** characters are within strings that are still being “processed,” and we have therefore not yet made a decision as to whether they are valid or invalid (in the dictionary or not).

```
1 p(thit)
2     = min(T + p(hit), p(thit)) --> 1 inv.
3     T + p(hit) = min(T + H + p(it), T + p(hit)) --> 1 inv.
4     T + H + p(it) = min(T + H + i + p(t), T + H + p(it)) --> 2
5     T + H + i + p(t) = T + H + i + T = 3 invalid
6     T + H + p(it) = T + H + it = 2 invalid
7     T + p(hit) = min(T + hi + p(t), T + p(hit)) --> 1 inv.
8     T + hi + p(t) = T + hi + T = 2 invalid
9     T + p(hit) = T + hit = 1 invalid
10    p(thit) = min(TH + p(it), p(thit)) --> 2 inv.
11    TH + p(it) = min(TH + i + p(t), TH + p(it)) --> 2 inv.
12    TH + i + p(t) = TH + i + T = 3 invalid
13    TH + p(it) = TH + it = 2 invalid
14    p(thit) = min(THI + p(t), p(thit)) --> 4 inv.
15    THI + p(t) = THI + T = 4 invalid
16    p(thit) = THIT = 4 invalid
```

In the above steps, observe that each level divides into two parts. The first splits the string and the second adjoins it.

For example, when we first call **p(thit)**, the current character being processed is just the first **t**. We recurse in two different directions. The first one (line 3) inserts a space after **t** and tries to find the most optimal way of parsing **hit**. The second one (line 10) tries to find the most optimal way of parsing such that there is no space between **t** and **h**. As we do this repeatedly, we will eventually look at every possible way of parsing the string.

The code below implements this solution. For simplicity, we have implemented this

algorithm to return only the number of invalid characters.

```

1  public int parseSimple(int wordStart, int wordEnd) {
2      if (wordEnd >= sentence.length()) {
3          return wordEnd - wordStart;
4      }
5
6      String word = sentence.substring(wordStart, wordEnd + 1);
7
8      /* break current word */
9      int bestExact = parseSimple(wordEnd + 1, wordEnd + 1);
10     if (!dictionary.contains(word)) {
11         bestExact += word.length();
12     }
13
14     /* extend current word */
15     int bestExtend = parseSimple(wordStart, wordEnd + 1);
16
17     /* find best */
18     return Math.min(bestExact, bestExtend);
19 }
```

There are two major optimizations we can make to this code:

- Some of the recursive cases overlap. For example, in the earlier walkthrough, we repeatedly compute the optimal parsing for it. Instead, we should just cache this result the first time we compute it, and reuse it later. We can use dynamic programming to do this.
- In some cases, we may be able to predict that a particular parsing will produce invalid strings. For example, if we were trying to parse the string xten, there are no words that begin with xt. Our current solution, however, will attempt to parse the string as xt + p(en), xte + p(n), and xten. Each time we will find that such words do not exist in the dictionary. Instead, we should just put a space after the x and do the best parsing we can from there. How can we quickly know that there are no strings beginning with xt? By using a trie.

The code below implements both of these optimizations.

```

1  public int parseOptimized(int wordStart, int wordEnd,
2                             Hashtable<Integer, Integer> cache) {
3      if (wordEnd >= sentence.length()) {
4          return wordEnd - wordStart;
5      }
6      if (cache.containsKey(wordStart)) {
7          return cache.get(wordStart);
8      }
9
10     String currentWord = sentence.substring(wordStart, wordEnd + 1);
11
12     /* check if prefix is in dictionary (false --> partial match) */
```

```
13     boolean validPartial = dictionary.contains(currentWord, false);
14
15     /* break current word */
16     int bestExact = parseOptimized(wordEnd + 1, wordEnd + 1, cache);
17
18     /* if full string is not in dictionary, add to invalid count. */
19     if (!validPartial || !dictionary.contains(currentWord, true)) {
20         bestExact += currentWord.length();
21     }
22
23     /* extend current word */
24     int bestExtend = Integer.MAX_VALUE;
25     if (validPartial) {
26         bestExtend = parseOptimized(wordStart, wordEnd + 1, cache);
27     }
28
29     /* find best */
30     int min = Math.min(bestExact, bestExtend);
31     cache.put(wordStart, min); // Cache result
32     return min;
33 }
```

Observe that we use a hash table to cache results. The key is the index of the *start* of the word. That is, we're caching the best way to parse the rest of the string.

We can make this code return the fully parsed string, but it's a bit more tedious. We will need to use a wrapper class called *Result* so that we can return both the number of invalid characters and the optimal string. Had we been implementing this with C++, we could have instead just passed a value by reference.

```
1  public class Result {
2      public int invalid = Integer.MAX_VALUE;
3      public String parsed = "";
4      public Result(int inv, String p) {
5          invalid = inv;
6          parsed = p;
7      }
8
9      public Result clone() {
10         return new Result(this.invalid, this.parsed);
11     }
12
13     public static Result min(Result r1, Result r2) {
14         if (r1 == null) {
15             return r2;
16         } else if (r2 == null) {
17             return r1;
18         }
19         return r2.invalid < r1.invalid ? r2 : r1;
20     }
21 }
```

```
22
23 public Result parse(int wordStart, int wordEnd,
24                     Hashtable<Integer, Result> cache) {
25     if (wordEnd >= sentence.length()) {
26         return new Result(wordEnd - wordStart,
27                            sentence.substring(wordStart).toUpperCase());
28     }
29     if (cache.containsKey(wordStart)) {
30         return cache.get(wordStart).clone();
31     }
32     String currentWord = sentence.substring(wordStart, wordEnd + 1);
33     boolean validPartial = dictionary.contains(currentWord, false);
34     boolean validExact = validPartial &&
35                         dictionary.contains(currentWord, true);
36
37     /* break current word */
38     Result bestExact = parse(wordEnd + 1, wordEnd + 1, cache);
39     if (validExact) {
40         bestExact.parsed = currentWord + " " + bestExact.parsed;
41     } else {
42         bestExact.invalid += currentWord.length();
43         bestExact.parsed = currentWord.toUpperCase() + " " +
44                             bestExact.parsed;
45     }
46
47     /* extend current word */
48     Result bestExtend = null;
49     if (validPartial) {
50         bestExtend = parse(wordStart, wordEnd + 1, cache);
51     }
52
53     /* find best */
54     Result best = Result.min(bestExact, bestExtend);
55     cache.put(wordStart, best.clone());
56     return best;
57 }
```

Be very careful in dynamic programming problems of how you cache objects. If the value you are caching is an object and not a primitive data type, it is very likely that you need to clone the object. You will see this in lines 30 and 55 above. If we don't clone it, future calls to parse will unintentionally modify the values in the cache.



## **Hard**

---

*Additional Review Problems: Solutions*

**Chapter 18**

- 18.1 Write a function that adds two numbers. You should not use + or any arithmetic operators.

pg 167

### SOLUTION

Our first instinct in problems like these should be that we're going to have to work with bits. Why? Because when you take away the + sign, what other choice do we have? Plus, that's how computers do it!

Our next thought should be to really, really understand how addition works. We can walk through an addition problem to see if we can understand something new—some pattern—and then see if we can replicate that with code.

So let's do just that—let's walk through an addition problem. We'll work in base 10 so that it's easier to see.

To add 759 + 674, I would usually add `digit[0]` from each number, carry the one, add `digit[1]` from each number, carry the one, and so on. You could take the same approach in binary: add each digit, and carry the one as necessary.

Can we make this a little easier? Yes! Imagine I decided to split apart the "addition" and "carry" steps. That is, I do the following:

1. Add 759 + 674, but "forget" to carry. I then get 323.
2. Add 759 + 674 but only do the carrying, rather than the addition of each digit. I then get 1110.
3. Add the result of the first two operations (recursively, using the same process described in step 1 and 2):  $1110 + 323 = 1433$ .

Now, how would we do this in binary?

1. If I add two binary numbers together, but forget to carry, the  $i$ th bit in the sum will be 0 only if  $a$  and  $b$  have the same  $i$ th bit (both 0 or both 1). This is essentially an XOR.
2. If I add two numbers together but *only* carry, I will have a 1 in the  $i$ th bit of the sum only if bits  $i - 1$  of  $a$  and  $b$  are both 1s. This is an AND, shifted.
3. Now, recurse until there's nothing to carry.

The following code implements this algorithm.

```
1 public static int add(int a, int b) {  
2     if (b == 0) return a;  
3     int sum = a ^ b; // add without carrying  
4     int carry = (a & b) << 1; // carry, but don't add  
5     return add(sum, carry); // recurse  
6 }
```

Problems requiring us to implement core operations like addition and subtraction are relatively common. The key in all of these problems is to dig into how these operations

are usually implemented, so that we can re-implement them with the constraints of the given problem.

- 18.2** Write a method to shuffle a deck of cards. It must be a perfect shuffle—in other words, each of the  $52!$  permutations of the deck has to be equally likely. Assume that you are given a random number generator which is perfect.

pg 167

## SOLUTION

---

This is a very well-known interview question, and a well-known algorithm. If you aren't one of the lucky few to already know this algorithm, read on.

Let's imagine our  $n$ -element array. Suppose it looks like this:

[1] [2] [3] [4] [5]

Using our Base Case and Build approach, we can ask this question: suppose we had a method `shuffle(...)` that worked on  $n - 1$  elements. Could we use this to shuffle  $n$  elements?

Sure. In fact, that's quite easy. We would first shuffle the first  $n - 1$  elements. Then, we would take the  $n$ th element and randomly swap it with an element in the array. That's it!

Recursively, that algorithm looks like this:

```

1  /* Random number between lower and higher, inclusive */
2  int rand(int lower, int higher) {
3      return lower + (int)(Math.random() * (higher - lower + 1));
4  }
5
6  int[] shuffleArrayRecursively(int[] cards, int i) {
7      if (i == 0) return cards;
8
9      shuffleArrayRecursively(cards, i - 1); // Shuffle earlier part
10     int k = rand(0, i); // Pick random index to swap with
11
12     /* Swap element k and i */
13     int temp = cards[k];
14     cards[k] = cards[i];
15     cards[i] = temp;
16
17     /* Return shuffled array */
18     return cards;
19 }
```

What would this algorithm look like iteratively? Let's think about it. All it does is moving through the array and, for each element  $i$ , swapping `array[i]` with a random element between 0 and  $i$ , inclusive.

This is actually a very clean algorithm to implement iteratively:

```
1 void shuffleArrayIteratively(int[] cards) {  
2     for (int i = 0; i < cards.length; i++) {  
3         int k = rand(0, i);  
4         int temp = cards[k];  
5         cards[k] = cards[i];  
6         cards[i] = temp;  
7     }  
8 }
```

The iterative approach is usually how we see this algorithm written.

- 18.3** Write a method to randomly generate a set of  $m$  integers from an array of size  $n$ . Each element must have equal probability of being chosen.

pg 167

### SOLUTION

Like the prior problem which was similar, (problem 18.2 on page 463), we can look at this problem recursively using the Base Case and Build approach.

Suppose we have an algorithm that can pull a random set of  $m$  elements from an array of size  $n - 1$ . How can we use this algorithm to pull a random set of  $m$  elements from an array of size  $n$ ?

We can first pull a random set of size  $m$  from the first  $n - 1$  elements. Then, we just need to decide if  $\text{array}[n]$  should be inserted into our subset (which would require pulling out a random element from it). An easy way to do this is to pick a random number  $k$  from  $0$  through  $n$ . If  $k < m$ , then insert  $\text{array}[n]$  into  $\text{subset}[k]$ . This will both "fairly" (i.e., with proportional probability) insert  $\text{array}[n]$  into the subset and "fairly" remove a random element from the subset.

The pseudocode for this recursive algorithm would look like this:

```
1 int[] pickMRecursively(int[] original, int m, int i) {  
2     if (i + 1 == m) { // Base case  
3         /* return first m elements of original */  
4     } else if (i + m > m) {  
5         int[] subset = pickMRecursively(original, m, i - 1);  
6         int k = random value between 0 and i, inclusive  
7         if (k < m) {  
8             subset[k] = original[i];  
9         }  
10        return subset;  
11    }  
12    return null;  
13 }
```

This is even cleaner to write iteratively. In this approach, we initialize an array  $\text{subset}$  to be the first  $m$  elements in  $\text{original}$ . Then, we iterate through the array, starting at element  $m$ , inserting  $\text{array}[i]$  into the subset at (random) position  $k$  whenever  $k < m$ .

```

1 int[] pickMIteratively(int[] original, int m) {
2     int[] subset = new int[m];
3
4     /* Fill in subset array with first part of original array */
5     for (int i = 0; i < m ; i++) {
6         subset[i] = original[i];
7     }
8
9     /* Go through rest of original array. */
10    for (int i = m; i < original.length; i++) {
11        int k = rand(0, i); // Random # between 0 and i, inclusive
12        if (k < m) {
13            subset[k] = original[i];
14        }
15    }
16
17    return subset;
18 }
```

Both solutions are, not surprisingly, very similar to the algorithm to shuffle an array.

#### 18.4 Write a method to count the number of 2s between 0 and n.

pg 167

#### SOLUTION

Our first approach to this problem can be—and probably should be—a brute force solution. Remember that interviewers want to see how you’re approaching a problem. Offering a brute force solution is a great way to start.

```

1 /* Counts the number of '2' digits between 0 and n */
2 int numberOf2sInRange(int n) {
3     int count = 0;
4     for (int i = 2; i <= n; i++) { // Might as well start at 2
5         count += numberOf2s(i);
6     }
7     return count;
8 }
9
10 /* Counts the number of '2' digits in a single number */
11 int numberOf2s(int n) {
12     int count = 0;
13     while (n > 0) {
14         if (n % 10 == 2) {
15             count++;
16         }
17         n = n / 10;
18     }
19     return count;
```

20 }

The only interesting part is that it's probably cleaner to separate out `numberOf2s` into a separate method. This demonstrates an eye for code cleanliness.

### Improved Solution

Rather than looking at the problem by ranges of numbers, we can look at the problem digit by digit. Picture a sequence of numbers:

0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
...									
110	111	112	113	114	115	116	117	118	119

We know that roughly one tenth of the time, the last digit will be a 2 since it happens once in any sequence of ten numbers. In fact, any digit is a 2 roughly one tenth of the time.

We say "roughly" because there are (very common) boundary conditions. For example, between 1 and 100, the 10s digit is a 2 exactly  $1/10^{\text{th}}$  of the time. However, between 1 and 37, the 10s digit is a 2 much more than  $1/10^{\text{th}}$  of the time.

We can work out what exactly the ratio is by looking at the three cases individually: `digit < 2`, `digit = 2`, and `digit > 2`.

#### Case $\text{digit} < 2$

Consider the value  $x = 61523$  and  $d = 3$ , and observe that  $x[d] = 1$  (that is, the dth digit of  $x$  is 1). There are 2s at the 3rd digit in the ranges 2000 - 2999, 12000 - 12999, 22000 - 22999, 32000 - 32999, 42000 - 42999, and 52000 - 52999. We will not yet have hit the range 62000 - 62999, so there are 6000 2s total in the 3rd digit. This is the same amount as if we were just counting all the 2s in the 3rd digit between 1 and 60000.

In other words, we can round *down* to the nearest  $10^{d+1}$ , and then divide by 10, to compute the number of 2s in the dth digit.

```
if x[d] < 2: count2sInRangeAtDigit(x, d) =
    let y = round down to nearest  $10^{d+1}$ 
    return y / 10
```

#### Case $\text{digit} > 2$

Now, let's look at the case where dth digit of  $x$  is greater than 2 ( $x[d] > 2$ ). We can apply almost the exact same logic to see that there are the same number of 2s in the 3rd digit in the range 0 - 63525 as there are in the range 0 - 70000. So, rather than rounding down, we round up.

```
if x[d] > 2: count2sInRangeAtDigit(x, d) =
    let y = round up to nearest  $10^{d+1}$ 
```

```
return y / 10
```

*Case digit = 2*

The final case may be the trickiest, but it follows from the earlier logic. Consider  $x = 62523$  and  $d = 3$ . We know that there are the same ranges of 2s from before (that is, the ranges  $2000 - 2999$ ,  $12000 - 12999$ , ...,  $52000 - 52999$ ). How many appear in the 3rd digit in the final, partial range from  $62000 - 62523$ ? Well, that should be pretty easy. It's just 524 ( $62000, 62001, \dots, 62523$ ).

```
if x[d] = 2: count2sInRangeAtDigit(x, d) =
    let y = round down to nearest  $10^{d-1}$ 
    let z = right side of x (i.e.,  $x \% 10^d$ )
    return y / 10 + z + 1
```

Now, all you need is to iterate through each digit in the number. Implementing this code is reasonably straightforward.

```
1  public static int count2sInRangeAtDigit(int number, int d) {
2      int powerOf10 = (int) Math.pow(10, d);
3      int nextPowerOf10 = powerOf10 * 10;
4      int right = number % powerOf10;
5
6      int roundDown = number - number % nextPowerOf10;
7      int roundUp = roundDown + nextPowerOf10;
8
9      int digit = (number / powerOf10) % 10;
10     if (digit < 2) { // if the digit in spot digit is
11         return roundDown / 10;
12     } else if (digit == 2) {
13         return roundDown / 10 + right + 1;
14     } else {
15         return roundUp / 10;
16     }
17 }
18
19 public static int count2sInRange(int number) {
20     int count = 0;
21     int len = String.valueOf(number).length();
22     for (int digit = 0; digit < len; digit++) {
23         count += count2sInRangeAtDigit(number, digit);
24     }
25     return count;
26 }
```

This question requires very careful testing. Make sure to generate a list of test cases, and to work through each of them.

- 18.5 You have a large text file containing words. Given any two words, find the shortest distance (in terms of number of words) between them in the file. If the operation will be repeated many times for the same file (but different pairs of words), can you optimize your solution?

pg 167

### SOLUTION

We will assume for this question that it doesn't matter whether word1 or word2 appears first. This is a question you should ask your interviewer. If the word order does matter, we can make the small modification shown in the code below.

To solve this problem, we can traverse the file just once. We remember throughout our traversal where we've last seen word1 and word2, storing the locations in lastPosWord1 and lastPosWord2. When we come across word1, we compare it to lastPosWord2 and update min as necessary, and then update lastPosWord1. We do the equivalent operation on word2. At the end of the traversal, we will have the minimum distance.

The code below implements this algorithm.

```
1  public int shortest(String[] words, String word1, String word2) {  
2      int min = Integer.MAX_VALUE;  
3      int lastPosWord1 = -1;  
4      int lastPosWord2 = -1;  
5      for (int i = 0; i < words.length; i++) {  
6          String currentWord = words[i];  
7          if (currentWord.equals(word1)) {  
8              lastPosWord1 = i;  
9              // Comment following 3 lines if word order matters  
10             int distance = lastPosWord1 - lastPosWord2;  
11             if (lastPosWord2 >= 0 && min > distance) {  
12                 min = distance;  
13             }  
14         } else if (currentWord.equals(word2)) {  
15             lastPosWord2 = i;  
16             int distance = lastPosWord2 - lastPosWord1;  
17             if (lastPosWord1 >= 0 && min > distance) {  
18                 min = distance;  
19             }  
20         }  
21     }  
22     return min;  
23 }
```

If we need to repeat the operation for other pairs of words, we can create a hash table with each word and the locations where it occurs. We then just need to find the minimum (arithmetic) difference between a value in listA and a value in listB.

There are several ways to compute the minimum difference between a value in listA

and a value in `listB`. Consider the following lists:

```
listA: {1, 2, 9, 15, 25}
listB: {4, 10, 19}
```

We can merge these lists into one sorted list, but “tag” each number with the original list. Tagging can be done by wrapping each value in a class that has two member variables: `data` (to store the actual value) and `listNumber`.

```
list: {1a, 2a, 4b, 9a, 10b, 15a, 19b, 25a}
```

Finding the minimum distance is now just a matter of traversing this merged list to find the minimum distance between two consecutive numbers which have different list tags. In this case, the solution would be a distance of 1 (between 9a and 10b). After the initial indexing of the file, this takes  $O(p + k)$  time, where  $p$  and  $k$  are the number of occurrences of each word.

- 18.6** *Describe an algorithm to find the smallest one million numbers in one billion numbers. Assume that the computer memory can hold all one billion numbers.*

pg 167

## SOLUTION

There are a number of ways to approach this problem. We will go through three of them: sorting, min heap, and selection rank.

### Solution 1: Sorting

We can sort the elements in ascending order and then take the first million numbers from that. The time complexity is  $O(n \log(n))$ .

### Solution 2: Min Heap

We can use a min heap to solve this problem. We first create a max heap (largest element at the top) for the first million numbers.

Then, we traverse through the list. On each element, we insert it into the list and delete the largest element.

At the end of the traversal, we will have a heap containing the smallest one million numbers. This algorithm is  $O(n \log(m))$ , where  $m$  is the number of values we are looking for.

### Approach 3: Selection Rank Algorithm (if you can modify the original array)

Selection Rank is a well-known algorithm in computer science to find the  $i$ th smallest (or largest) element in an array in linear time.

If the elements are unique, you can find the  $i$ th smallest element in expected  $O(n)$

time. The basic algorithm operates like this:

1. Pick a random element in the array and use it as a “pivot.” Partition elements around the pivot, keeping track of the number of elements on the left side of the partition.
2. If there are exactly  $i$  elements on the left, then you just return the biggest element on the left.
3. If the left side is bigger than  $i$ , repeat the algorithm on just the left part of the array.
4. If the left side is smaller than  $i$ , repeat the algorithm on the right, but look for the element with rank  $i - \text{leftSize}$ .

The code below implements this algorithm.

```
1  public int partition(int[] array, int left, int right, int pivot) {  
2      while (true) {  
3          while (left <= right && array[left] <= pivot) {  
4              left++;  
5          }  
6  
7          while (left <= right && array[right] > pivot) {  
8              right--;  
9          }  
10         if (left > right) {  
11             return left - 1;  
12         }  
13         swap(array, left, right);  
14     }  
15 }  
16 }  
17  
18 public int rank(int[] array, int left, int right, int rank) {  
19     int pivot = array[randomIntInRange(left, right)];  
20  
21     /* Partition and return end of left partition */  
22     int leftEnd = partition(array, left, right, pivot);  
23  
24     int leftSize = leftEnd - left + 1;  
25     if (leftSize == rank + 1) {  
26         return max(array, left, leftEnd);  
27     } else if (rank < leftSize) {  
28         return rank(array, left, leftEnd, rank);  
29     } else {  
30         return rank(array, leftEnd + 1, right, rank - leftSize);  
31     }  
32 }
```

Once you have found the  $i$ th smallest element, you can run through the array to find all values less than or equal to this element.

If the elements are not unique (which they are unlikely to be), we can tweak this algo-

rithm slightly to accomodate this change. However, we no longer have as tight of a guarantee on the runtime.

There is an algorithm that can *guarantee* that we can find the  $i$ th smallest element in linear time, regardless of the uniqueness of the elements. However, the complexity of this algorithm is far beyond the scope of this interview. If you are interested though, you can look up the algorithm in CLRS' Introduction to Algorithms textbook.

- 18.7** Given a list of words, write a program to find the longest word made of other words in the list.

pg 168

### SOLUTION

This problem seems complex, so let's simplify it. What if we just wanted to know the longest word made of two other words in the list?

We could solve this by iterating through the list, from the longest word to the shortest word. For each word, we would split it into all possible pairs and check if both the left and right side are contained in the list.

The pseudocode for this would look like the following:

```

1  String getLongestWord(String[] list) {
2      String[] array = list.SortByLength();
3      /* Create map for easy lookup */
4      HashMap<String, Boolean> map = new HashMap<String, Boolean>;
5
6      for (String str : array) {
7          map.put(str, true);
8      }
9
10     for (String s : array) {
11         // Divide into every possible pair
12         for (int i = 1; i < s.length(); i++) {
13             String left = s.substring(0, i);
14             String right = s.substring(i);
15             // Check if both sides are in the array
16             if (map[left] == true && map[right] == true) {
17                 return s;
18             }
19         }
20     }
21     return str;
22 }
```

This works greater for when we just want to know composites of two words. But what if a word could be formed by any number of other words?

In this case, we could apply a very similar approach, with one modification: rather than simply looking up if the right side is in the array, we would recursively see if we can build the right side from the other elements in the array.

The code below implements this algorithm:

```
1 String printLongestWord(String[] arr) {
2     HashMap<String, Boolean> map = new HashMap<String, Boolean>();
3     for (String str : arr) {
4         map.put(str, true);
5     }
6     Arrays.sort(arr, new LengthComparator()); // Sort by length
7     for (String s : arr) {
8         if (canBuildWord(s, true, map)) {
9             System.out.println(s);
10            return s;
11        }
12    }
13    return "";
14 }
15
16 boolean canBuildWord(String str, boolean isOriginalWord,
17                       HashMap<String, Boolean> map) {
18     if (map.containsKey(str) && !isOriginalWord) {
19         return map.get(str);
20     }
21     for (int i = 1; i < str.length(); i++) {
22         String left = str.substring(0, i);
23         String right = str.substring(i);
24         if (map.containsKey(left) && map.get(left) == true &&
25             canBuildWord(right, false, map)) {
26             return true;
27         }
28     }
29     map.put(str, false);
30     return false;
31 }
```

Note that in this solution we have performed a small optimization. We use a dynamic programming approach to cache the results between calls. This way, if we repeatedly need to check if there's any way to build "testingtester," we'll only have to compute it once.

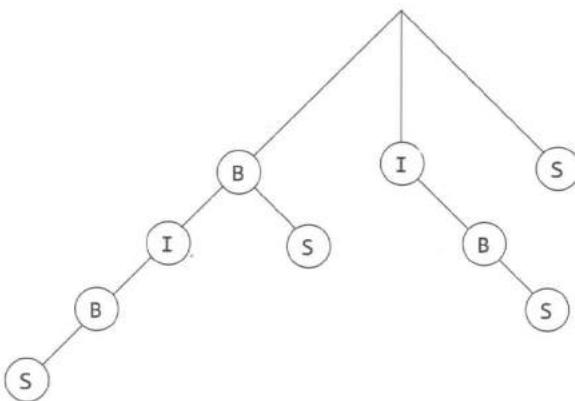
A boolean flag `isOriginalWord` is used to complete the above optimization. The method `canBuildWord` is called for the original word and for each substring, and its first step is to check the cache for a previously calculated result. However, for the original words, we have a problem: `map` is initialized to `true` for them, but we don't want to return `true` (since a word cannot be composed solely of itself). Therefore, for the original word, we simply bypass this check using the `isOriginalWord` flag.

- 18.8** Given a string  $s$  and an array of smaller strings  $T$ , design a method to search  $s$  for each small string in  $T$ .

pg 168

**SOLUTION**

First, create a suffix tree for  $s$ . For example, if your word were `bibs`, you would create the following tree:



Then, all you need to do is search in the suffix tree for each string in  $T$ . Note that if "B" were a word, you would come up with two locations.

```

1  public class SuffixTree {
2      SuffixTreeNode root = new SuffixTreeNode();
3      public SuffixTree(String s) {
4          for (int i = 0; i < s.length(); i++) {
5              String suffix = s.substring(i);
6              root.insertString(suffix, i);
7          }
8      }
9
10     public ArrayList<Integer> search(String s) {
11         return root.search(s);
12     }
13 }
14
15 public class SuffixTreeNode {
16     HashMap<Character, SuffixTreeNode> children = new
17             HashMap<Character, SuffixTreeNode>();
18     char value;
19     ArrayList<Integer> indexes = new ArrayList<Integer>();
20     public SuffixTreeNode() { }
21
22     public void insertString(String s, int index) {
  
```

```
23     indexes.add(index);
24     if (s != null && s.length() > 0) {
25         value = s.charAt(0);
26         SuffixTreeNode child = null;
27         if (children.containsKey(value)) {
28             child = children.get(value);
29         } else {
30             child = new SuffixTreeNode();
31             children.put(value, child);
32         }
33         String remainder = s.substring(1);
34         child.insertString(remainder, index);
35     }
36 }
37
38 public ArrayList<Integer> search(String s) {
39     if (s == null || s.length() == 0) {
40         return indexes;
41     } else {
42         char first = s.charAt(0);
43         if (children.containsKey(first)) {
44             String remainder = s.substring(1);
45             return children.get(first).search(remainder);
46         }
47     }
48     return null;
49 }
50 }
```

- 18.9** Numbers are randomly generated and passed to a method. Write a program to find and maintain the median value as new values are generated.

pg 168

### SOLUTIONS

One solution is to use two priority heaps: a max heap for the values below the median, and a min heap for the values above the median. This will divide the elements roughly in half, with the middle two elements as the top of the two heaps. This makes it trivial to find the median.

What do we mean by “roughly in half,” though? “Roughly” means that, if we have an odd number of values, one heap will have an extra value. Observe that the following is true:

- If `maxHeap.size() > minHeap.size()`, then `heap1.top()` will be the median.
- If `maxHeap.size() == minHeap.size()`, then the average of `maxHeap.top()` and `minHeap.top()` will be the median.

By the way in which we rebalance the heaps, we will ensure that it is always `maxHeap`

with extra element.

The algorithm works as follows. When a new value arrives, it is placed in the maxHeap if the value is less than or equal to the median, otherwise it is placed into the minHeap. The heap sizes can be equal, or the maxHeap may have one extra element. This constraint can easily be restored by shifting an element from one heap to the other. The median is available in constant time, by looking at the top element(s). Updates take  $O(\log(n))$  time.

```

1  private Comparator<Integer> maxHeapComparator;
2  private Comparator<Integer> ,minHeapComparator;
3  private PriorityQueue<Integer> maxHeap, minHeap;
4
5  public void addNewNumber(int randomNumber) {
6      /* Note: addNewNumber maintains a condition that
7         * maxHeap.size() >= minHeap.size() */
8      if (maxHeap.size() == minHeap.size()) {
9          if ((minHeap.peek() != null) &&
10              randomNumber > minHeap.peek()) {
11                  maxHeap.offer(minHeap.poll());
12                  minHeap.offer(randomNumber);
13              } else {
14                  maxHeap.offer(randomNumber);
15              }
16      } else {
17          if (randomNumber < maxHeap.peek()) {
18              minHeap.offer(maxHeap.poll());
19              maxHeap.offer(randomNumber);
20          }
21          else {
22              minHeap.offer(randomNumber);
23          }
24      }
25  }
26
27  public static double getMedian() {
28      /* maxHeap is always at least as big as minHeap. So if maxHeap
29         * is empty, then minHeap is also. */
30      if (maxHeap.isEmpty()) {
31          return 0;
32      }
33      if (maxHeap.size() == minHeap.size()) {
34          return ((double)minHeap.peek()+(double)maxHeap.peek()) / 2;
35      } else {
36          /* If maxHeap and minHeap are of different sizes, then
37             * maxHeap must have one extra element. Return maxHeap's
38             * top element.*/
39          return maxHeap.peek();
40      }
41  }

```

- 18.10 Given two words of equal length that are in a dictionary, write a method to transform one word into another word by changing only one letter at a time. The new word you get in each step must be in the dictionary.

pg 168

### SOLUTION

Although this problem seems tough, it's actually a straightforward modification of breadth first search. Each word in our "graph" branches to all words in the dictionary that are one edit away. The interesting part is the implementation. Specifically, should we build a graph as we go?

We could, but there's an easier way. We can instead use a "backtrack map." In this backtrack map, if  $B[v] = w$ , then you know that you edited v to get w. When we reach our end word, we can use this backtrack map repeatedly to reverse our path. See the code below:

```
1  LinkedList<String> transform(String startWord, String stopWord,
2      Set<String> dictionary) {
3      startWord = startWord.toUpperCase();
4      stopWord = stopWord.toUpperCase();
5      Queue<String> actionQueue = new LinkedList<String>();
6      Set<String> visitedSet = new HashSet<String>();
7      Map<String, String> backtrackMap =
8          new TreeMap<String, String>();
9
10     actionQueue.add(startWord);
11     visitedSet.add(startWord);
12
13     while (!actionQueue.isEmpty()) {
14         String w = actionQueue.poll();
15         /* For each possible word v from w with one edit operation */
16         for (String v : getOneEditWords(w)) {
17             if (v.equals(stopWord)) {
18                 // Found our word! Now, back track.
19                 LinkedList<String> list = new LinkedList<String>();
20                 // Append v to list
21                 list.add(v);
22                 while (w != null) {
23                     list.add(0, w);
24                     w = backtrackMap.get(w);
25                 }
26                 return list;
27             }
28             /* If v is a dictionary word */
29             if (dictionary.contains(v)) {
30                 if (!visitedSet.contains(v)) {
31                     actionQueue.add(v);
32                     visitedSet.add(v); // mark visited
```

```

33         backtrackMap.put(v, w);
34     }
35 }
36 }
37 }
38 return null;
39 }
40
41 Set<String> getOneEditWords(String word) {
42     Set<String> words = new TreeSet<String>();
43     for (int i = 0; i < word.length(); i++) {
44         char[] wordArray = word.toCharArray();
45         // change that letter to something else
46         for (char c = 'A'; c <= 'Z'; c++) {
47             if (c != word.charAt(i)) {
48                 wordArray[i] = c;
49                 words.add(new String(wordArray));
50             }
51         }
52     }
53     return words;
54 }

```

Let  $n$  be the length of the start word and  $m$  be the number of like-sized words in the dictionary. The runtime of this algorithm is  $O(nm)$  since the while loop will dequeue at most  $m$  unique words. The for loop is  $O(n)$  since it walks down the string applying a fixed number of replacements for each character.

- 18.11** *Imagine you have a square matrix, where each cell (pixel) is either black or white. Design an algorithm to find the maximum subsquare such that all four borders are filled with black pixels.*

pg 168

## SOLUTION

Like many problems, there's an easy way and a hard way to solve this. We'll go through both solutions.

### The “Simple” Solution: $O(N^4)$

We know that the biggest possible square has a length of size  $N$ , and there is only one possible square of size  $N \times N$ . We can easily check for that square and return if we find it.

If we do not find a square of size  $N \times N$ , we can try the next best thing:  $(N-1) \times (N-1)$ . We iterate through all squares of this size and return the first one we find. We then do the same for  $N-2$ ,  $N-3$ , and so on. Since we are searching progressively smaller squares, we know that the first square we find is the biggest.

Our code works as follows:

```
1 Subsquare findSquare(int[][] matrix) {
2     for (int i = matrix.length; i >= 1; i--) {
3         Subsquare square = findSquareWithSize(matrix, i);
4         if (square != null) return square;
5     }
6     return null;
7 }
8
9 Subsquare findSquareWithSize(int[][] matrix, int squareSize) {
10    /* On an edge of length N, there are (N - sz + 1) squares of
11       * length sz. */
12    int count = matrix.length - squareSize + 1;
13
14    /* Iterate through all squares with side length squareSize. */
15    for (int row = 0; row < count; row++) {
16        for (int col = 0; col < count; col++) {
17            if (isSquare(matrix, row, col, squareSize)) {
18                return new Subsquare(row, col, squareSize);
19            }
20        }
21    }
22    return null;
23 }
24
25 boolean isSquare(int[][] matrix, int row, int col, int size) {
26    // Check top and bottom border.
27    for (int j = 0; j < size; j++){
28        if (matrix[row][col+j] == 1) {
29            return false;
30        }
31        if (matrix[row+size-1][col+j] == 1){
32            return false;
33        }
34    }
35
36    // Check left and right border.
37    for (int i = 1; i < size - 1; i++){
38        if (matrix[row+i][col] == 1){
39            return false;
40        }
41        if (matrix[row+i][col+size-1] == 1){
42            return false;
43        }
44    }
45    return true;
46 }
```

**Pre-Processing Solution: O(N<sup>3</sup>)**

A large part of the slowness of the “simple” solution above is due to the fact we have to do  $O(N)$  work each time we want to check a potential square. By doing some pre-processing, we can cut down the time of `isSquare` to  $O(1)$ . The time of the whole algorithm is reduced to  $O(N^3)$ .

If we analyze what `isSquare` does, we realize that all it ever needs to know is if the next `squareSize` items, on the right of as well as below particular cells, are zeros. We can pre-compute this data in a straightforward, iterative fashion.

We iterate from right to left, bottom to top. At each cell, we do the following computation:

```
if A[r][c] is white, zeros right and zeros below are 0
else A[r][c].zerosRight = A[r][c + 1].zerosRight + 1
      A[r][c].zerosBelow = A[r + 1][c].zerosBelow + 1
```

Below is an example of these values for a potential matrix.

(0s right, 0s below)			Original Matrix		
0,0	1,3	0,0		W	B
2,2	1,2	0,0		B	B
2,1	1,1	0,0		B	W

Now, instead of iterating through  $O(N)$  elements, the `isSquare` method just needs to check `zerosRight` and `zerosBelow` for the corners.

Our code for this algorithm is below. Note that `findSquare` and `findSquareWithSize` is equivalent, other than a call to `processMatrix` and working with a new data type thereafter.

```
1 public class SquareCell {
2     public int zerosRight = 0;
3     public int zerosBelow = 0;
4     /* declaration, getters, setters */
5 }
6
7 Subsquare findSquare(int[][][] matrix) {
8     SquareCell[][] processed = processSquare(matrix);
9     for (int i = matrix.length; i >= 1; i--) {
10         Subsquare square = findSquareWithSize(processed, i);
11         if (square != null) return square;
12     }
13     return null;
14 }
```

## Solutions to Chapter 18 | Hard

```
16 Subsquare findSquareWithSize(SquareCell[][][] processed,
17                               int squareSize) {
18     /* equivalent to first algorithm */
19 }
20
21
22 boolean isSquare(SquareCell[][][] matrix, int row, int col,
23                   int size) {
24     SquareCell topLeft = matrix[row][col];
25     SquareCell topRight = matrix[row][col + size - 1];
26     SquareCell bottomLeft = matrix[row + size - 1][col];
27     if (topLeft.zerosRight < size) { // Check top edge
28         return false;
29     }
30     if (topLeft.zerosBelow < size) { // Check left edge
31         return false;
32     }
33     if (topRight.zerosBelow < size) { // Check right edge
34         return false;
35     }
36     if (bottomLeft.zerosRight < size) { // Check bottom edge
37         return false;
38     }
39     return true;
40 }
41
42 SquareCell[][][] processSquare(int[][][] matrix) {
43     SquareCell[][][] processed =
44         new SquareCell[matrix.length][matrix.length];
45
46     for (int r = matrix.length - 1; r >= 0; r--) {
47         for (int c = matrix.length - 1; c >= 0; c--) {
48             int rightZeros = 0;
49             int belowZeros = 0;
50             // only need to process if it's a black cell
51             if (matrix[r][c] == 0) {
52                 rightZeros++;
53                 belowZeros++;
54                 // next column over is on same row
55                 if (c + 1 < matrix.length) {
56                     SquareCell previous = processed[r][c + 1];
57                     rightZeros += previous.zerosRight;
58                 }
59                 if (r + 1 < matrix.length) {
60                     SquareCell previous = processed[r + 1][c];
61                     belowZeros += previous.zerosBelow;
62                 }
63             }
64             processed[r][c] = new SquareCell(rightZeros, belowZeros);
65         }
66     }
67 }
```

```

66     }
67     return processed;
68 }
```

- 18.12** Given an  $N \times N$  matrix of positive and negative integers, write code to find the submatrix with the largest possible sum.

pg 168

### SOLUTION

This problem can be approached in a variety of ways. We'll start with the brute force solution and then optimize the solution from there.

#### Brute Force Solution: $O(N^6)$

Like many "maximizing" problems, this problem has a straightforward brute force solution. This solution simply iterates through all possible submatrices, computes the sum, and finds the biggest.

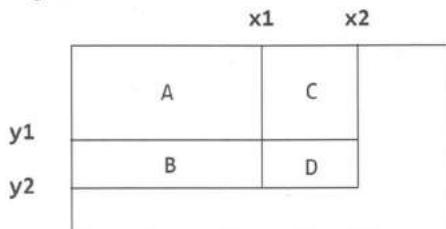
To iterate through all possible submatrices (with no duplicates), we simply need to iterate through all order pairs of rows, and then all ordered pairs of columns.

This solution is  $O(N^6)$ , since we iterate through  $O(N^4)$  submatrices and it takes  $O(N^2)$  time to compute the area of each.

#### Dynamic Programming Solution: $O(N^4)$

Notice that the earlier solution is made slower by a factor of  $O(N^2)$  simply because computing the sum of a matrix is so slow. Can we reduce the time to compute the area? Yes! In fact, we can reduce the time of `computeSum` to  $O(1)$ .

Consider the following rectangle:



Suppose we knew the following values:

```

ValD = area(point(0, 0) -> point(x2, y2))
ValC = area(point(0, 0) -> point(x2, y1))
ValB = area(point(0, 0) -> point(x1, y2))
ValA = area(point(0, 0) -> point(x1, y1))
```

Each `Val*` starts at the origin and ends at the bottom right corner of a subrectangle.

With these values, we know the following:

$$\text{area}(D) = \text{Val}_D - \text{area}(A \cup C) - \text{area}(A \cup B) + \text{area}(A).$$

Or, written another way:

$$\text{area}(D) = \text{Val}_D - \text{Val}_B - \text{Val}_C + \text{Val}_A$$

We can efficiently compute these values for all points in the matrix by using similar logic:

$$\text{Val}(x, y) = \text{Val}(x - 1, y) + \text{Val}(y - 1, x) - \text{Val}(x - 1, y - 1) + M[x][y]$$

We can precompute all such values, and then efficiently find the maximum submatrix.

The following code implements this algorithm.

```
1  int getMaxMatrix(int[][] original) {
2      int maxArea = Integer.MIN_VALUE; // Important! Max may be < 0
3      int rowCount = original.length;
4      int columnCount = original[0].length;
5      int[][] matrix = precomputeMatrix(original);
6      for (int row1 = 0; row1 < rowCount; row1++) {
7          for (int row2 = row1; row2 < rowCount; row2++) {
8              for (int col1 = 0; col1 < columnCount; col1++) {
9                  for (int col2 = col1; col2 < columnCount; col2++) {
10                      maxArea = Math.max(maxArea, computeSum(matrix,
11                                         row1, row2, col1, col2));
12                  }
13              }
14          }
15      }
16      return maxArea;
17  }
18
19  int[][] precomputeMatrix(int[][][] matrix) {
20      int[][] sumMatrix = new int[matrix.length][matrix[0].length];
21      for (int i = 0; i < matrix.length; i++) {
22          for (int j = 0; j < matrix.length; j++) {
23              if (i == 0 && j == 0) { // first cell
24                  sumMatrix[i][j] = matrix[i][j];
25              } else if (j == 0) { // cell in first column
26                  sumMatrix[i][j] = sumMatrix[i - 1][j] + matrix[i][j];
27              } else if (i == 0) { // cell in first row
28                  sumMatrix[i][j] = sumMatrix[i][j - 1] + matrix[i][j];
29              } else {
30                  sumMatrix[i][j] = sumMatrix[i - 1][j] +
31                      sumMatrix[i][j - 1] - sumMatrix[i - 1][j - 1] +
32                      matrix[i][j];
33              }
34          }
35      }
36      return sumMatrix;
37  }
```

```

38
39 int computeSum(int[][] sumMatrix, int i1, int i2, int j1, int j2) {
40     if (i1 == 0 && j1 == 0) { // starts at row 0, column 0
41         return sumMatrix[i2][j2];
42     } else if (i1 == 0) { // start at row 0
43         return sumMatrix[i2][j2] - sumMatrix[i2][j1 - 1];
44     } else if (j1 == 0) { // start at column 0
45         return sumMatrix[i2][j2] - sumMatrix[i1 - 1][j2];
46     } else {
47         return sumMatrix[i2][j2] - sumMatrix[i2][j1 - 1]
48             - sumMatrix[i1 - 1][j2] + sumMatrix[i1 - 1][j1 - 1];
49     }
50 }

```

**Optimized Solution: O(N<sup>3</sup>)**

Believe it or not, an even more optimal solution exists. If we have R rows and C columns, we can solve it in O(R<sup>2</sup>C) time.

Recall the solution to the maximum subarray problem: Given an array of integers, find the subarray with the largest sum."We can find the maximum subarray in O(N) time. We will leverage this solution for this problem.

Every submatrix can be represented by a contiguous sequence of rows and a contiguous sequence of columns. If we were to iterate through every contiguous sequence of rows, we would then just need to find, for each of those, the set of columns that give us the highest sum. That is:

```

1 maxSum = 0
2 foreach rowStart in rows
3     foreach rowEnd in rows
4         /* We have a number of possible submatrices with rowStart
5            * and rowEnd as the top and bottom edge of the matrix.
6            * Find the colStart and colEnd edges that give the
7            * highest sum. */
8         maxSum = max(runningMaxSum, maxSum)
9     return maxSum

```

Now, the question is, how do we efficiently find the "best" colStart and colEnd? Here's where things get really fun.

Picture a submatrix:

rowStart				
9	-8	1	3	-2
-3	7	6	-2	4
6	-4	-4	8	-7
12	-5	3	9	-5
rowEnd				

We want to find the `colStart` and `colEnd` that give us the highest possible sum of all submatrices with `rowStart` as their top edge and `rowEnd` as their bottom edge. To do this, we can sum up each column and then apply the `maximumSubArray` function explained at the beginning of this problem.

For the earlier example, the maximum subarray is the first through fourth column. This means that the maximum submatrix is (`rowStart, first column`) through (`rowEnd, fourth column`).

We now have pseudocode that looks like the following.

```
1 maxSum = 0
2 foreach rowStart in rows
3     foreach rowEnd in rows
4         foreach col in columns
5             partialSum[col] = sum of matrix[rowStart, col] through
6                             matrix[rowEnd, col]
7             runningMaxSum = maxSubArray(partialSum)
8             maxSum = max(runningMaxSum, maxSum)
9 return maxSum
```

The sum in lines 5 and 6 takes  $R \times C$  time to compute (since it iterates through `rowStart` through `rowEnd`), so this gives us a runtime of  $O(R^3C)$ . We're not quite done yet.

In lines 5 and 6, we're basically adding up `a[0]...a[i]` from scratch, even though in the previous iteration of the outer for loop, we already added up `a[0]...a[i-1]`. Let's cut out this duplicated effort.

```
1 maxSum = 0
2 foreach rowStart in rows
3     clear array partialSum
4     foreach rowEnd in rows
5         foreach col in columns
6             partialSum[col] += matrix[rowEnd, col]
7             runningMaxSum = maxSubArray(partialSum)
8             maxSum = max(runningMaxSum, maxSum)
9 return maxSum
```

Our full code looks like this:

```
1 public void clearArray(int[] array) {
2     for (int i = 0; i < array.length; i++) {
3         array[i] = 0;
4     }
5 }
6
7 public static int maxSubMatrix(int[][] matrix) {
8     int rowCount = matrix.length;
9     int colCount = matrix[0].length;
10    int[] partialSum = new int[colCount];
11    int maxSum = 0; // Max sum is an empty matrix
12
13}
```

```

14    for (int rowStart = 0; rowStart < rowCount; rowStart++) {
15        clearArray(partialSum);
16
17        for (int rowEnd = rowStart; rowEnd < rowCount; rowEnd++) {
18            for (int i = 0; i < colCount; i++) {
19                partialSum[i] += matrix[rowEnd][i];
20            }
21
22            int tempMaxSum = maxSubArray(partialSum, colCount);
23
24            /* If you want to track the coordinates, add code here to
25             * do that. */
26            maxSum = Math.max(maxSum, tempMaxSum);
27        }
28    }
29    return maxSum;
30 }
31
32 public static int maxSubArray(int array[], int N) {
33     int maxSum = 0;
34     int runningSum = 0;
35
36     for (int i = 0; i < N; i++) {
37         runningSum += array[i];
38         maxSum = Math.max(maxSum, runningSum);
39
40         /* If runningSum is < 0, no point in trying to continue the
41          * series. Reset. */
42         if (runningSum < 0) {
43             runningSum = 0;
44         }
45     }
46     return maxSum;
47 }

```

This was an extremely complex problem. You would not be expected to figure out this entire problem in an interview without a lot of help from your interviewer.

- 18.13** Given a list of millions of words, design an algorithm to create the largest possible rectangle of letters such that every row forms a word (reading left to right) and every column forms a word (reading top to bottom). The words need not be chosen consecutively from the list, but all rows must be the same length and all columns must be the same height.

pg 168

## SOLUTION

Many problems involving a dictionary can be solved by doing some pre-processing.

Where can we do pre-processing?

Well, if we're going to create a rectangle of words, we know that each row must be the same length and each column must be the same length. So, let's group the words of the dictionary based on their sizes. Let's call this grouping D, where  $D[i]$  contains the list of words of length  $i$ .

Next, observe that we're looking for the largest rectangle. What is the absolute largest rectangle that could be formed? It's  $\text{length}(\text{largest word})^2$ .

```
1 int maxRectangle = longestWord * longestWord;
2 for z = maxRectangle to 1 {
3     for each pair of numbers (i, j) where i*j = z {
4         /* attempt to make rectangle. return if successful. */
5     }
6 }
```

By iterating from the biggest possible rectangle to the smallest, we ensure that the first valid rectangle we find will be the largest possible one.

Now, for the hard part: `makeRectangle(int l, int h)`. This method attempts to build a rectangle of words which has length  $l$  and height  $h$ .

One way to do this is to iterate through all (ordered) sets of  $h$  words and then check if the columns are also valid words. This will work, but it's rather inefficient.

Imagine that we are trying to build a  $6 \times 5$  rectangle and the first couple rows are:

```
there
queen
pizza
....
```

At this point, we know that the first column starts with `tqp`. We know—or *should* know—that no dictionary word starts with `tqp`. Why do we bother continuing to build a rectangle when we know we'll fail to create a valid one in the end?

This leads us to a more optimal solution. We can build a trie to easily look up if a substring is a prefix of a word in the dictionary. Then, when we build our rectangle, row by row, we check to see if the columns are all valid prefixes. If not, we fail immediately, rather than continue to try to build this rectangle.

The code below implements this algorithm. It is long and complex, so we will go through it step-by-step.

First, we do some pre-processing to group words by their lengths. We create an array of tries (one for each word length), but hold off on building the tries until we need them.

```
1 WordGroup[] groupList = WordGroup.createWordGroups(list);
2 int maxWordLength = groupList.length;
3 Trie trieList[] = new Trie[maxWordLength];
```

The `maxRectangle` method is the “main” part of our code. It starts with the biggest

possible rectangle area (which is  $\text{maxWordLength}^2$ ) and tries to build a rectangle of that size. If it fails, it subtracts one from the area and attempts this new, smaller size. The first rectangle that can be successfully built is guaranteed to be the biggest.

```

1  Rectangle maxRectangle() {
2      int maxSize = maxWordLength * maxWordLength;
3      for (int z = maxSize; z > 0; z--) { // start from biggest area
4          for (int i = 1; i <= maxWordLength; i++) {
5              if (z % i == 0) {
6                  int j = z / i;
7                  if (j <= maxWordLength) {
8                      /* Create a rectangle of length i and height j. Note
9                         * that i * j = z. */
10                     Rectangle rectangle = makeRectangle(i, j);
11                     if (rectangle != null) {
12                         return rectangle;
13                     }
14                 }
15             }
16         }
17     }
18     return null;
19 }
```

The `makeRectangle` method is called by `maxRectangle` and tries to build a rectangle of a specific length and height.

```

1  Rectangle makeRectangle(int length, int height) {
2      if (groupList[length - 1] == null ||
3          groupList[height - 1] == null) {
4          return null;
5      }
6
7      /* Create trie for word length if we haven't yet */
8      if (trieList[height - 1] == null) {
9          LinkedList<String> words = groupList[height - 1].getWords();
10         trieList[height - 1] = new Trie(words);
11     }
12
13     return makePartialRectangle(length, height,
14                                 new Rectangle(length));
15 }
```

The `makePartialRectangle` method is where the action happens. It is passed in the intended, final length and height, and a partially formed rectangle. If the rectangle is already of the final height, then we just check to see if the columns form valid, complete words, and return.

Otherwise, we check to see if the columns form valid prefixes. If they do not, then we immediately break since there is no way to build a valid rectangle off of this partial one.

But, if everything is okay so far, and all the columns are valid prefixes of words, then we

search through all the words of the right length, append each to the current rectangle, and recursively try to build a rectangle off of {current rectangle with new word appended}.

```
1  Rectangle makePartialRectangle(int l, int h, Rectangle rectangle) {
2      if (rectangle.height == h) { // Check if complete rectangle
3          if (rectangle.isComplete(l, h, groupList[h - 1])) {
4              return rectangle;
5          } else {
6              return null;
7          }
8      }
9
10     /* Compare columns to trie to see if potentially valid rect */
11     if (!rectangle.isPartialOK(l, trieList[h - 1])) {
12         return null;
13     }
14
15     /* Go through all words of the right length. Add each one to the
16      * current partial rectangle, and attempt to build a rectangle
17      * recursively. */
18     for (int i = 0; i < groupList[l-1].length(); i++) {
19         /* Create a new rectangle which is this rect + new word. */
20         Rectangle orgPlus =
21             rectangle.append(groupList[l-1].getWord(i));
22
23         /* Try to build a rectangle with this new, partial rect */
24         Rectangle rect = makePartialRectangle(l, h, orgPlus);
25         if (rect != null) {
26             return rect;
27         }
28     }
29     return null;
30 }
```

The Rectangle class represents a partially or fully formed rectangle of words. The method `isPartialOk` can be called to check if the rectangle is, thus far, a valid one (that is, all the columns are prefixes of words). The method `isComplete` serves a similar function, but checks if each of the columns makes a full word.

```
1  public class Rectangle {
2      public int height, length;
3      public char [][] matrix;
4
5      /* Construct an "empty" rectangular. Length is fixed, but height
6       * vary as we add words. */
7      public Rectangle(int l) {
8          height = 0;
9          length = l;
10     }
11 }
```

```

12  /* Construct a rectangular array of letters of the specified
13   * length and height, and backed by the specified matrix of
14   * letters. (It is assumed that the length and height specified
15   * as arguments are consistent with the array argument's
16   * dimensions.) */
17  public Rectangle(int length, int height, char[][] letters) {
18      this.height = letters.length;
19      this.length = letters[0].length;
20      matrix = letters;
21  }
22
23  public char getLetter (int i, int j) { return matrix[i][j]; }
24  public String getColumn(int i) { ... }
25
26  /* Check if all columns are valid. All rows are already known to
27   * be valid since they were added directly from dictionary. */
28  public boolean isComplete(int l, int h, WordGroup groupList) {
29      if (height == h) {
30          /* Check if each column is a word in the dictionary. */
31          for (int i = 0; i < l; i++) {
32              String col = getColumn(i);
33              if (!groupList.containsWord(col)) {
34                  return false;
35              }
36          }
37          return true;
38      }
39      return false;
40  }
41
42  public boolean isPartialOK(int l, Trie trie) {
43      if (height == 0) return true;
44      for (int i = 0; i < l; i++ ) {
45          String col = getColumn(i);
46          if (!trie.contains(col)) {
47              return false;
48          }
49      }
50      return true;
51  }
52
53  /* Create a new Rectangle by taking the rows of the current
54   * rectangle and appending s. */
55  public Rectangle append(String s) { ... }
56 }
```

The WordGroup class is a simple container for all words of a specific length. For easy lookup, we store the words in a hash table as well as in an ArrayList.

The lists in WordGroup are created through a static method called `createWordGroups`.

```
1  public class WordGroup {
2      private Hashtable<String, Boolean> lookup =
3          new Hashtable<String, Boolean>();
4      private ArrayList<String> group = new ArrayList<String>();
5
6      public boolean containsWord(String s) {
7          return lookup.containsKey(s);
8      }
9
10     public void addWord (String s) {
11         group.add(s);
12         lookup.put(s, true);
13     }
14
15     public int length() { return group.size(); }
16     public String getWord(int i) { return group.get(i); }
17     public ArrayList<String> getWords() { return group; }
18
19     public static WordGroup[] createWordGroups(String[] list) {
20         WordGroup[] groupList;
21         int maxWordLength = 0;
22         /* Find the length of the longest word */
23         for (int i = 0; i < list.length; i++) {
24             if (list[i].length() > maxWordLength) {
25                 maxWordLength = list[i].length();
26             }
27         }
28
29         /* Group the words in the dictionary into lists of words of
30          * same length. groupList[i] will contain a list of words,
31          * each of length (i+1). */
32         groupList = new WordGroup[maxWordLength];
33         for (int i = 0; i < list.length; i++) {
34             /* We do wordLength - 1 instead of just wordLength since
35              * this is used as an index and no words are of length 0 */
36             int wordLength = list[i].length() - 1;
37             if (groupList[wordLength] == null) {
38                 groupList[wordLength] = new WordGroup();
39             }
40             groupList[wordLength].addWord(list[i]);
41         }
42         return groupList;
43     }
44 }
```

The full code for this problem, including the code for `Trie` and `TrieNode`, can be found in the code attachment. Note that in a problem as complex as this, you'd most likely only need to write the pseudocode. Writing the entire code would be nearly impossible in such a short amount of time.