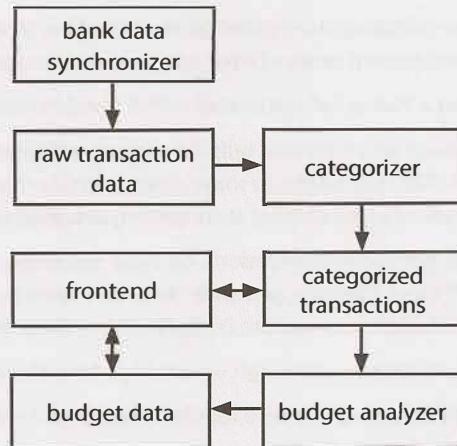


Step 3: Draw the Major Components

The most naive system would be one that pulls bank data on each login, categorizes all the data, and then analyzes the user's budget. This wouldn't quite fit the requirements, though, as we want email notifications on particular events.

We can do a bit better.



With this basic architecture, the bank data is pulled at periodic times (hourly or daily). The frequency may depend on the behavior of the users. Less active users may have their accounts checked less frequently.

Once new data arrives, it is stored in some list of raw, unprocessed transactions. This data is then pushed to the categorizer, which assigns each transaction to a category and stores these categorized transactions in another datastore.

The budget analyzer pulls in the categorized transactions, updates each user's budget per category, and stores the user's budget.

The frontend pulls data from both the categorized transactions datastore as well as from the budget datastore. Additionally, a user could also interact with the frontend by changing the budget or the categorization of their transactions.

Step 4: Identify the Key Issues

We should now reflect on what the major issues here might be.

This will be a very data-heavy system. We want it to feel snappy and responsive, though, so we'll want as much processing as possible to be asynchronous.

We will almost certainly want at least one task queue, where we can queue up work that needs to be done. This work will include tasks such as pulling in new bank data, re-analyzing budgets, and categorizing new bank data. It would also include re-trying tasks that failed.

These tasks will likely have some sort of priority associated with them, as some need to be performed more often than others. We want to build a task queue system that can prioritize some task types over others, while still ensuring that all tasks will be performed eventually. That is, we wouldn't want a low priority task to essentially "starve" because there are always higher priority tasks.

One important part of the system that we haven't yet addressed will be the email system. We could use a task to regularly crawl user's data to check if they're exceeding their budget, but that means checking every

single user daily. Instead, we'll want to queue a task whenever a transaction occurs that potentially exceeds a budget. We can store the current budget totals by category to make it easy to understand if a new transaction exceeds the budget.

We should also consider incorporating the knowledge (or assumption) that a system like this will probably have a large number of inactive users—users who signed up once and then haven't touched the system since. We may want to either remove them from the system entirely or deprioritize their accounts. We'll want some system to track their account activity and associate priority with their accounts.

The biggest bottleneck in our system will likely be the massive amount of data that needs to be pulled and analyzed. We should be able to fetch the bank data asynchronously and run these tasks across many servers. We should drill a bit deeper into how the categorizer and budget analyzer work.

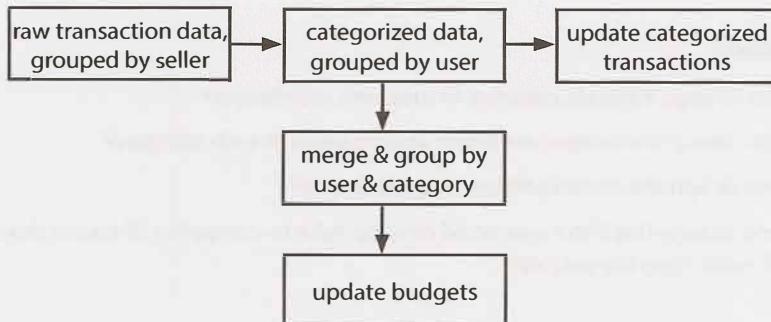
Categorizer and Budget Analyzer

One thing to note is that transactions are not dependent on each other. As soon as we get a transaction for a user, we can categorize it and integrate this data. It might be inefficient to do so, but it won't cause any inaccuracies.

Should we use a standard database for this? With lots of transactions coming in at once, that might not be very efficient. We certainly don't want to do a bunch of joins.

It may be better instead to just store the transactions to a set of flat text files. We assumed earlier that the categorizations are based on the seller's name alone. If we're assuming a lot of users, then there will be a lot of duplicates across the sellers. If we group the transaction files by seller's name, we can take advantage of these duplicates.

The categorizer can do something like this:



It first gets the raw transaction data, grouped by seller. It picks the appropriate category for the seller (which might be stored in a cache for the most common sellers), and then applies that category to all those transactions.

After applying the category, it re-groups all the transactions by user. Then, those transactions are inserted into the datastore for this user.

before categorizer	after categorizer
amazon/ user121,\$5.43,Aug 13 user922,\$15.39,Aug 27 ... comcast/ user922,\$9.29,Aug 24 user248,\$40.13,Aug 18 ...	user121/ amazon,shopping,\$5.43,Aug 13 ... user922/ amazon,shopping,\$15.39,Aug 27 comcast,utilities,\$9.29,Aug 24 ... user248/ comcast,utilities,\$40.13,Aug 18 ...

Then, the budget analyzer comes in. It takes the data grouped by user, merges it across categories (so all Shopping tasks for this user in this timespan are merged), and then updates the budget.

Most of these tasks will be handled in simple log files. Only the final data (the categorized transactions and the budget analysis) will be stored in a database. This minimizes writing and reading from the database.

User Changing Categories

The user might selectively override particular transactions to assign them to a different category. In this case, we would update the datastore for the categorized transactions. It would also signal a quick recomputation of the budget to decrement the item from the old category and increment the item in the other category.

We could also just recompute the budget from scratch. The budget analyzer is fairly quick as it just needs to look over the past few weeks of transactions for a single user.

Follow Up Questions

- How would this change if you also needed to support a mobile app?
- How would you design the component which assigns items to each category?
- How would you design the recommended budgets feature?
- How would you change this if the user could develop rules to categorize all transactions from a particular seller differently than the default?

9.8 Pastebin: Design a system like Pastebin, where a user can enter a piece of text and get a randomly generated URL for public access.

pg 145

SOLUTION

We can start with clarifying the specifics of this system.

Step 1: Scope the Problem

- The system does not support user accounts or editing documents.
- The system tracks analytics of how many times each page is accessed.
- Old documents get deleted after not being accessed for a sufficiently long period of time.
- While there isn't true authentication on accessing documents, users should not be able to "guess" docu-

ment URLs easily.

- The system has a frontend as well as an API.
- The analytics for each URL can be accessed through a “stats” link on each page. It is not shown by default, though.

Step 2: Make Reasonable Assumptions

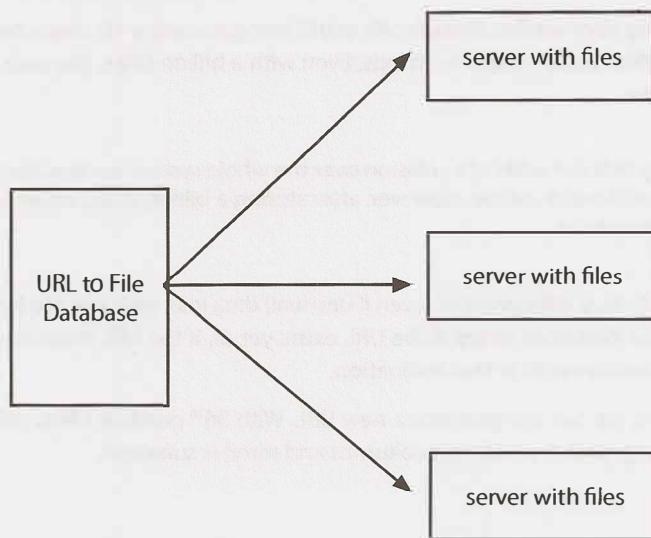
- The system gets heavy traffic and contains many millions of documents.
- Traffic is not equally distributed across documents. Some documents get much more access than others.

Step 3: Draw the Major Components

We can sketch out a simple design. We'll need to keep track of URLs and the files associated with them, as well as analytics for how often the files have been accessed.

How should we store the documents? We have two options: we can store them in a database or we can store them on a file. Since the documents can be large and it's unlikely we need searching capabilities, storing them on a file is probably the better choice.

A simple design like this might work well:



Here, we have a simple database that looks up the location (server and path) of each file. When we have a request for a URL, we look up the location of the URL within the datastore and then access the file.

Additionally, we will need a database that tracks analytics. We can do this with a simple datastore that adds each visit (including timestamp, IP address, and location) as a row in a database. When we need to access the stats of each visit, we pull the relevant data in from this database.

Step 4: Identify the Key Issues

The first issue that comes to mind is that some documents will be accessed much more frequently than others. Reading data from the filesystem is relatively slow compared with reading from data in memory. Therefore, we probably want to use a cache to store the most recently accessed documents. This will ensure

that items accessed very frequently (or very recently) will be quickly accessible. Since documents cannot be edited, we will not need to worry about invalidating this cache.

We should also potentially consider sharding the database. We can shard it using some mapping from the URL (for example, the URL's hash code modulo some integer), which will allow us to quickly locate the database which contains this file.

In fact, we could even take this a step further. We could skip the database entirely and just let a hash of the URL indicate which server contains the document. The URL itself could reflect the location of the document. One potential issue from this is that if we need to add servers, it could be difficult to redistribute the documents.

Generating URLs

We have not yet discussed how to actually generate the URLs. We probably do not want a monotonically increasing integer value, as this would be easy for a user to "guess." We want URLs to be difficult to access without being provided the link.

One simple path is to generate a random GUID (e.g., 5d50e8ac-57cb-4a0d-8661-bcdee2548979). This is a 128-bit value that, while not strictly guaranteed to be unique, has low enough odds of a collision that we can treat it as unique. The drawback of this plan is that such a URL is not very "pretty" to the user. We could hash it to a smaller value, but then that increases the odds of collision.

We could do something very similar, though. We could just generate a 10-character sequence of letters and numbers, which gives us 36^{10} possible strings. Even with a billion URLs, the odds of a collision on any specific URL are very low.

This is not to say that the odds of a collision over the whole system are low. They are not. Any one specific URL is unlikely to collide. However, after storing a billion URLs, we are very likely to have a collision at some point.

Assuming that we aren't okay with periodic (even if unusual) data loss, we'll need to handle these collisions. We can either check the datastore to see if the URL exists yet or, if the URL maps to a specific server, just detect whether a file already exists at the destination.

When a collision occurs, we can just generate a new URL. With 36^{10} possible URLs, collisions would be rare enough that the lazy approach here (detect collisions and retry) is sufficient.

Analytics

The final component to discuss is the analytics piece. We probably want to display the number of visits, and possibly break this down by location or time.

We have two options here:

- Store the raw data from each visit.
- Store just the data we know we'll use (number of visits, etc.).

You can discuss this with your interviewer, but it probably makes sense to store the raw data. We never know what features we'll add to the analytics down the road. The raw data allows us flexibility.

This does not mean that the raw data needs to be easily searchable or even accessible. We can just store a log of each visit in a file, and back this up to other servers.

One issue here is that this amount of data could be substantial. We could potentially reduce the space usage considerably by storing data only probabilistically. Each URL would have a `storage_probability` associated with it. As the popularity of a site goes up, the `storage_probability` goes down. For example, a popular document might have data logged only one out of every ten times, at random. When we look up the number of visits for the site, we'll need to adjust the value based on the probability (for example, by multiplying it by 10). This will of course lead to a small inaccuracy, but that may be acceptable.

The log files are not designed to be used frequently. We will want to also store this precomputed data in a datastore. If the analytics just displays the number of visits plus a graph over time, this could be kept in a separate database.

URL	Month and Year	Visits
12ab31b92p	December 2013	242119
12ab31b92p	January 2014	429918
...

Every time a URL is visited, we can increment the appropriate row and column. This datastore can also be sharded by the URL.

As the stats are not listed on the regular pages and would generally be of less interest, it should not face as heavy of a load. We could still cache the generated HTML on the frontend servers, so that we don't continuously reaccess the data for the most popular URLs.

Follow-Up Questions

- How would you support user accounts?
- How would you add a new piece of analytics (e.g., referral source) to the stats page?
- How would your design change if the stats were shown with each document?

10

Solutions to Sorting and Searching

- 10.1 Sorted Merge:** 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 149

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 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 (indexB >= 0) {  
8         /* end of a is > than end of b */  
9         if (indexA >= 0 && a[indexA] > b[indexB]) {  
10             a[indexMerged] = a[indexA]; // copy element  
11             indexA--;  
12         } else {  
13             a[indexMerged] = b[indexB]; // copy element  
14             indexB--;  
15         }  
16         indexMerged--; // move indices  
17     }  
18 }
```

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

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

pg 150

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.

We need a quick and easy way of determining if two strings are anagrams of each other. What defines if two words are anagrams of each other? Well, anagrams are words that have the same characters but in different orders. It follows then that if we can put the characters in the same order, we can easily check if the new words are identical.

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 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 void sort(String[] array) {
2     HashMapList<String, String> mapList = new HashMapList<String, String>();
3
4     /* Group words by anagram */
5     for (String s : array) {
6         String key = sortChars(s);
7         mapList.put(key, s);
8     }
9 }
```

```
9
10 / *Convert hash table to array */
11 int index = 0;
12 for (String key : mapList.keySet()) {
13     ArrayList<String> list = mapList.get(key);
14     for (String t : list) {
15         array[index] = t;
16         index++;
17     }
18 }
19 }
20
21 String sortChars(String s) {
22     char[] content = s.toCharArray();
23     Arrays.sort(content);
24     return new String(content);
25 }
26
27 / *HashMapList<String, Integer> is a HashMap that maps from Strings to
28 * ArrayList<Integer>. See appendix for implementation. */
```

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

10.3 Search in Rotated Array: 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.

EXAMPLE

Input: find 5 in {15, 16, 19, 20, 25, 1, 3, 4, 5, 7, 10, 14}

Output: 8 (the index of 5 in the array)

pg 150

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 Array 2, 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}. 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  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 out which side
11      * is normally ordered, and then use the normally ordered half to figure out
12      * 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 or right half is all repeats
26         if (a[mid] != a[right]) { // If right is different, 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.

10.4 Sorted Search, No Size: You are given an array-like data structure `Listy` which lacks a size method. It does, however, have an `elementAt(i)` method that returns the element at index `i` in $O(1)$ time. If `i` is beyond the bounds of the data structure, it returns `-1`. (For this reason, the data structure only supports positive integers.) Given a `Listy` which contains sorted, positive integers, find the index at which an element `x` occurs. If `x` occurs multiple times, you may return any index.

pg 150

SOLUTION

Our first thought here should be binary search. The problem is that binary search requires us knowing the length of the list, so that we can compare it to the midpoint. We don't have that here.

Could we compute the length? Yes!

We know that `elementAt` will return `-1` when `i` is too large. We can therefore just try bigger and bigger values until we exceed the size of the list.

But how much bigger? If we just went through the list linearly—1, then 2, then 3, then 4, and so on—we'd wind up with a linear time algorithm. We probably want something faster than this. Otherwise, why would the interviewer have specified the list is sorted?

It's better to back off exponentially. Try 1, then 2, then 4, then 8, then 16, and so on. This ensures that, if the list has length `n`, we'll find the length in at most $O(\log n)$ time.

Why $O(\log n)$? Imagine we start with pointer `q` at `q = 1`. At each iteration, this pointer `q` doubles, until `q` is bigger than the length `n`. How many times can `q` double in size before it's bigger than `n`? Or, in other words, for what value of `k` does $2^k = n$? This expression is equal when `k = \log n`, as this is precisely what `\log` means. Therefore, it will take $O(\log n)$ steps to find the length.

Once we find the length, we just perform a (mostly) normal binary search. I say "mostly" because we need to make one small tweak. If the mid point is `-1`, we need to treat this as a "too big" value and search left. This is on line 16 below.

There's one more little tweak. Recall that the way we figure out the length is by calling `elementAt` and comparing it to `-1`. If, in the process, the element is bigger than the value `x` (the one we're searching for), we'll jump over to the binary search part early.

```
1 int search(Listy list, int value) {
2     int index = 1;
3     while (list.elementAt(index) != -1 && list.elementAt(index) < value) {
4         index *= 2;
5     }
6     return binarySearch(list, value, index / 2, index);
7 }
8
9 int binarySearch(Listy list, int value, int low, int high) {
10    int mid;
11
12    while (low <= high) {
13        mid = (low + high) / 2;
14        int middle = list.elementAt(mid);
15        if (middle > value || middle == -1) {
16            high = mid - 1;
17        } else if (middle < value) {
```

```

18     low = mid + 1;
19 } else {
20     return mid;
21 }
22 }
23 return -1;
24 }

```

It turns out that not knowing the length didn't impact the runtime of the search algorithm. We find the length in $O(\log n)$ time and then do the search in $O(\log n)$ time. Our overall runtime is $O(\log n)$, just as it would be in a normal array.

- 10.5 Sparse Search:** Given a sorted array of strings that is interspersed with empty strings, write a method to find the location of a given string.

EXAMPLE

Input: ball, {"at", "", "", "", "ball", "", "", "car", "", "", "dad", "", ""}
Output: 4

pg 150

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 int search(String[] strings, String str, int first, int last) {
2     if (first > last) return -1;
3     /* Move mid to the middle */
4     int mid = (last + first) / 2;
5
6     /* If mid is empty, find closest non-empty string. */
7     if (strings[mid].isEmpty()) {
8         int left = mid - 1;
9         int right = mid + 1;
10        while (true) {
11            if (left < first && right > last) {
12                return -1;
13            } else if (right <= last && !strings[right].isEmpty()) {
14                mid = right;
15                break;
16            } else if (left >= first && !strings[left].isEmpty()) {
17                mid = left;
18                break;
19            }
20            right++;
21            left--;
22        }
23    }

```

```
24
25     /* Check for string, and recurse if necessary */
26     if (str.equals(strings[mid])) { // Found it!
27         return mid;
28     } else if (strings[mid].compareTo(str) < 0) { // Search right
29         return search(strings, str, mid + 1, last);
30     } else { // Search left
31         return search(strings, str, first, mid - 1);
32     }
33 }
34
35 int search(String[] strings, String str) {
36     if (strings == null || str == null || str == "") {
37         return -1;
38     }
39     return search(strings, str, 0, strings.length - 1);
40 }
```

The worst-case runtime for this algorithm is $O(n)$. In fact, it's impossible to have an algorithm for this problem that is better than $O(n)$ in the worst case. After all, you could have an array of all empty strings except for one non-empty string. There is no "smart" way to find this non-empty string. In the worst case, you will need to look at every element in the array.

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.

10.6 Sort Big File: Imagine you have a 20 GB file with one string per line. Explain how you would sort the file.

pg 150

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 merge the chunks, one by one. At the end, we have a fully sorted file.

This algorithm is known as external sort.

- 10.7 Missing Int:** Given an input file with four billion non-negative integers, provide an algorithm to generate an integer that 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 150

SOLUTION

There are a total of 2^{32} , or 4 billion, distinct integers possible and 2^{31} non-negative integers. Therefore, we know the input file (assuming it is `ints` rather than `longs`) contains some duplicates.

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 represents 32 boolean values.
2. Initialize BV with all 0s.
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 numberOfInts = ((long) Integer.MAX_VALUE) + 1;
2 byte[] bitfield = new byte [(int) (numberOfInts / 8)];
3 String filename = ...
4
5 void findOpenNumber() throws FileNotFoundException {
6     Scanner in = new Scanner(new FileReader(filename));
7     while (in.hasNextInt()) {
8         int n = in.nextInt ();
9         /* Finds the corresponding number in the bitfield by using the OR operator to
10          * set the nth bit of a byte (e.g., 10 would correspond to the 2nd bit of
11          * index 2 in the byte array). */
12         bitfield [n / 8] |= 1 << (n % 8);
13     }
14
15     for (int i = 0; i < bitfield.length; i++) {
16         for (int j = 0; j < 8; j++) {
17             /* Retrieves the individual bits of each byte. When 0 bit is found, print
18              * the corresponding value. */
19             if ((bitfield[i] & (1 << j)) == 0) {
20                 System.out.println (i * 8 + j);
21                 return;
22             }
23         }
24     }
25 }
```

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 $\text{arraySize} = \frac{2^{31}}{\text{rangeSize}}$ 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:

$$2^{11} \leq \text{rangeSize} \leq 2^{26}$$

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 findOpenNumber(String filename) throws FileNotFoundException {
2      int rangeSize = (1 << 20); // 2^20 bits (2^17 bytes)
3
4      /* Get count of number of values within each block. */
5      int[] blocks = getCountPerBlock(filename, rangeSize);
6
7      /* Find a block with a missing value. */
8      int blockIndex = findBlockWithMissing(blocks, rangeSize);
9      if (blockIndex < 0) return -1;
```

```
10  /* Create bit vector for items within this range. */
11 byte[] bitVector = getBitVectorForRange(filename, blockIndex, rangeSize);
12
13
14 /* Find a zero in the bit vector */
15 int offset = findZero(bitVector);
16 if (offset < 0) return -1;
17
18 /* Compute missing value. */
19 return blockIndex * rangeSize + offset;
20 }
21
22 /* Get count of items within each range. */
23 int[] getCountPerBlock(String filename, int rangeSize)
24     throws FileNotFoundException {
25     int arraySize = Integer.MAX_VALUE / rangeSize + 1;
26     int[] blocks = new int[arraySize];
27
28     Scanner in = new Scanner (new FileReader(filename));
29     while (in.hasNextInt()) {
30         int value = in.nextInt();
31         blocks[value / rangeSize]++;
32     }
33     in.close();
34     return blocks;
35 }
36
37 /* Find a block whose count is low. */
38 int findBlockWithMissing(int[] blocks, int rangeSize) {
39     for (int i = 0; i < blocks.length; i++) {
40         if (blocks[i] < rangeSize){
41             return i;
42         }
43     }
44     return -1;
45 }
46
47 /* Create a bit vector for the values within a specific range. */
48 byte[] getBitVectorForRange(String filename, int blockIndex, int rangeSize)
49     throws FileNotFoundException {
50     int startRange = blockIndex * rangeSize;
51     int endRange = startRange + rangeSize;
52     byte[] bitVector = new byte[rangeSize/Byte.SIZE];
53
54     Scanner in = new Scanner(new FileReader(filename));
55     while (in.hasNextInt()) {
56         int value = in.nextInt();
57         /* If the number is inside the block that's missing numbers, we record it */
58         if (startRange <= value && value < endRange) {
59             int offset = value - startRange;
60             int mask = (1 << (offset % Byte.SIZE));
61             bitVector[offset / Byte.SIZE] |= mask;
62         }
63     }
64     in.close();
65     return bitVector;
```

```
66 }
67
68 /* Find bit index that is 0 within byte. */
69 int findZero(byte b) {
70     for (int i = 0; i < Byte.SIZE; i++) {
71         int mask = 1 << i;
72         if ((b & mask) == 0) {
73             return i;
74         }
75     }
76     return -1;
77 }
78
79 /* Find a zero within the bit vector and return the index. */
80 int findZero(byte[] bitVector) {
81     for (int i = 0; i < bitVector.length; i++) {
82         if (bitVector[i] != ~0) { // If not all 1s
83             int bitIndex = findZero(bitVector[i]);
84             return i * Byte.SIZE + bitIndex;
85         }
86     }
87     return -1;
88 }
```

What if, as a follow up question, you are asked to solve the problem with even less memory? In this case, we can 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.

10.8 Find Duplicates: 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 151

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 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) + 1]; // 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.

- 10.9 Sorted Matrix Search:** 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 151

SOLUTION

We can approach this in two ways: a more naive solution that only takes advantage of part of the sorting, and a more optimal way that takes advantage of both parts of the sorting.

Solution #1: Naive 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 use identical logic. If the start of a row is bigger than x , 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  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
20	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  Coordinate findElement(int[][] matrix, Coordinate origin, Coordinate dest, int x){  
2      if (!origin.inbounds(matrix) || !dest.inbounds(matrix)) {  
3          return null;  
4      }  
5      if (matrix[origin.row][origin.column] == x) {  
6          return origin;  
7      } else if (!origin.isBefore(dest)) {  
8          return null;  
9      }  
10     /* Set start to start of diagonal and end to the end of the diagonal. Since the  
11        * grid may not be square, the end of the diagonal may not equal dest. */  
12     Coordinate start = (Coordinate) origin.clone();  
13     int diagDist = Math.min(dest.row - origin.row, dest.column - origin.column);  
14     Coordinate end = new Coordinate(start.row + diagDist, start.column + diagDist);  
15     Coordinate p = new Coordinate(0, 0);  
16  
17     /* Do binary search on the diagonal, looking for the first element > x */  
18     while (start.isBefore(end)) {  
19         p.setToAverage(start, end);  
20         if (x > matrix[p.row][p.column]) {  
21             start.row = p.row + 1;  
22             start.column = p.column + 1;  
23         } else {  
24             end.row = p.row - 1;  
25             end.column = p.column - 1;  
26         }  
27     }  
28  
29     /* Split the grid into quadrants. Search the bottom left and the top right. */  
30     return partitionAndSearch(matrix, origin, dest, start, x);  
31 }  
32  
33  
34 Coordinate partitionAndSearch(int[][] matrix, Coordinate origin, Coordinate dest,  
35                               Coordinate pivot, int x) {  
36     Coordinate lowerLeftOrigin = new Coordinate(pivot.row, origin.column);  
37     Coordinate lowerLeftDest = new Coordinate(dest.row, pivot.column - 1);  
38     Coordinate upperRightOrigin = new Coordinate(origin.row, pivot.column);  
39     Coordinate upperRightDest = new Coordinate(pivot.row - 1, dest.column);  
40  
41     Coordinate lowerLeft = findElement(matrix, lowerLeftOrigin, lowerLeftDest, x);  
42     if (lowerLeft == null) {  
43         return findElement(matrix, upperRightOrigin, upperRightDest, x);  
44     }  
45 }
```

```
45     return lowerLeft;
46 }
47
48 Coordinate findElement(int[][] matrix, int x) {
49     Coordinate origin = new Coordinate(0, 0);
50     Coordinate dest = new Coordinate(matrix.length - 1, matrix[0].length - 1);
51     return findElement(matrix, origin, dest, x);
52 }
53
54 public class Coordinate implements Cloneable {
55     public int row, column;
56     public Coordinate(int r, int c) {
57         row = r;
58         column = c;
59     }
60
61     public boolean inbounds(int[][] matrix) {
62         return row >= 0 && column >= 0 &&
63             row < matrix.length && column < matrix[0].length;
64     }
65
66     public boolean isBefore(Coordinate p) {
67         return row <= p.row && column <= p.column;
68     }
69
70     public Object clone() {
71         return new Coordinate(row, column);
72     }
73
74     public void setToAverage(Coordinate min, Coordinate max) {
75         row = (min.row + max.row) / 2;
76         column = (min.column + max.column) / 2;
77     }
78 }
```

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 could 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.

10.10 Rank from Stream: 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).

EXAMPLE

Stream (in order of appearance): 5, 1, 4, 4, 5, 9, 7, 13, 3

```
getRankOfNumber(1) = 0  
getRankOfNumber(3) = 1  
getRankOfNumber(4) = 3
```

pg 151

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).

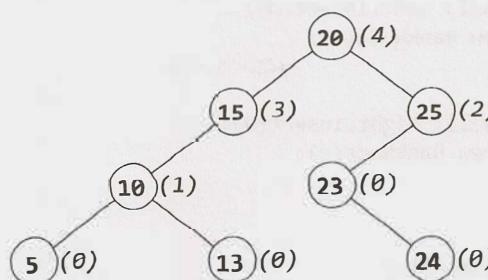
To find the rank of a number, we could do an in-order traversal, keeping a counter as we traverse. The goal is that, by the time we find x , counter will equal the number of elements less than x .

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, return node.leftSize()
3     if x is on left of node, return getRank(node.left, x)
4     if x is on right of node, return node.leftSize() + 1 + getRank(node.right, x)
5 }
```

The full code for this is below.

```

1 RankNode root = null;
2
3 void track(int number) {
4     if (root == null) {
5         root = new RankNode(number);
6     } else {
7         root.insert(number);
8     }
9 }
10
11 int getRankOfNumber(int number) {
12     return root.getRank(number);
13 }
14
15
16 public class RankNode {
17     public int left_size = 0;
18     public RankNode left, right;
19     public int data = 0;
20     public RankNode(int d) {
21         data = d;
22     }
23
24     public void insert(int d) {
25         if (d <= data) {
```

```
26     if (left != null) left.insert(d);
27     else left = new RankNode(d);
28     left_size++;
29 } else {
30     if (right != null) right.insert(d);
31     else right = new RankNode(d);
32 }
33 }
34
35 public int getRank(int d) {
36     if (d == data) {
37         return left_size;
38     } else if (d < data) {
39         if (left == null) return -1;
40         else return left.getRank(d);
41     } else {
42         int right_rank = right == null ? -1 : right.getRank(d);
43         if (right_rank == -1) return -1;
44         else return left_size + 1 + right_rank;
45     }
46 }
47 }
```

The `track` method and the `getRankOfNumber` method will both operate in $O(\log N)$ on a balanced tree and $O(N)$ on an unbalanced tree.

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.

10.11 Peaks and Valleys: In an array of integers, a “peak” is an element which is greater than or equal to the adjacent integers and a “valley” is an element which is less than or equal to the adjacent integers. For example, in the array $\{5, 8, 6, 2, 3, 4, 6\}$, $\{8, 6\}$ are peaks and $\{5, 2\}$ are valleys. Given an array of integers, sort the array into an alternating sequence of peaks and valleys.

EXAMPLE

Input: $\{5, 3, 1, 2, 3\}$

Output: $\{5, 1, 3, 2, 3\}$

pg 151

SOLUTION

Since this problem asks us to sort the array in a particular way, one thing we can try is doing a normal sort and then “fixing” the array into an alternating sequence of peaks and valleys.

Suboptimal Solution

Imagine we were given an unsorted array and then sort it to become the following:

0 1 4 7 8 9

We now have an ascending list of integers.

How can we rearrange this into a proper alternating sequence of peaks and valleys? Let's walk through it and try to do that.

- The 0 is okay.

- The 1 is in the wrong place. We can swap it with either the 0 or 4. Let's swap it with the 0.
1 0 4 7 8 9
- The 4 is okay.
- The 7 is in the wrong place. We can swap it with either the 4 or the 8. Let's swap it with the 4.
1 0 7 4 8 9
- The 9 is in the wrong place. Let's swap it with the 8.
1 0 7 4 9 8

Observe that there's nothing special about the array having these values. The relative order of the elements matters, but all sorted arrays will have the same relative order. Therefore, we can take this same approach on any sorted array.

Before coding, we should clarify the exact algorithm, though.

- Sort the array in ascending order.
- Iterate through the elements, starting from index 1 (not 0) and jumping two elements at a time.
- At each element, swap it with the previous element. Since every three elements appear in the order `small <= medium <= large`, swapping these elements will always put `medium` as a peak: `medium <= small <= large`.

This approach will ensure that the peaks are in the right place: indexes 1, 3, 5, and so on. As long as the odd-numbered elements (the peaks) are bigger than the adjacent elements, then the even-numbered elements (the valleys) must be smaller than the adjacent elements.

The code to implement this is below.

```
1 void sortValleyPeak(int[] array) {  
2     Arrays.sort(array);  
3     for (int i = 1; i < array.length; i += 2) {  
4         swap(array, i - 1, i);  
5     }  
6 }  
7  
8 void swap(int[] array, int left, int right) {  
9     int temp = array[left];  
10    array[left] = array[right];  
11    array[right] = temp;  
12 }
```

This algorithm runs in $O(n \log n)$ time.

Optimal Solution

To optimize past the prior solution, we need to cut out the sorting step. The algorithm must operate on an unsorted array.

Let's revisit an example.

9 1 0 4 8 7

For each element, we'll look at the adjacent elements. Let's imagine some sequences. We'll just use the numbers 0, 1 and 2. The specific values don't matter.

```
0 1 2  
0 2 1      // peak  
1 0 2  
1 2 0      // peak  
2 1 0
```

2 0 1

If the center element needs to be a peak, then two of those sequences work. Can we fix the other ones to make the center element a peak?

Yes. We can fix this sequence by swapping the center element with the largest adjacent element.

```
0 1 2 -> 0 2 1  
0 2 1 // peak  
1 0 2 -> 1 2 0  
1 2 0 // peak  
2 1 0 -> 1 2 0  
2 0 1 -> 0 2 1
```

As we noted before, if we make sure the peaks are in the right place then we know the valleys are in the right place.

We should be a little cautious here. Is it possible that one of these swaps could “break” an earlier part of the sequence that we’d already processed? This is a good thing to worry about, but it’s not an issue here. If we’re swapping `middle` with `left`, then `left` is currently a valley. `Middle` is smaller than `left`, so we’re putting an even smaller element as a valley. Nothing will break. All is good!

The code to implement this is below.

```
1 void sortValleyPeak(int[] array) {  
2     for (int i = 1; i < array.length; i += 2) {  
3         int biggestIndex = maxIndex(array, i - 1, i, i + 1);  
4         if (i != biggestIndex) {  
5             swap(array, i, biggestIndex);  
6         }  
7     }  
8 }  
9  
10 int maxIndex(int[] array, int a, int b, int c) {  
11     int len = array.length;  
12     int aValue = a >= 0 && a < len ? array[a] : Integer.MIN_VALUE;  
13     int bValue = b >= 0 && b < len ? array[b] : Integer.MIN_VALUE;  
14     int cValue = c >= 0 && c < len ? array[c] : Integer.MIN_VALUE;  
15  
16     int max = Math.max(aValue, Math.max(bValue, cValue));  
17     if (aValue == max) return a;  
18     else if (bValue == max) return b;  
19     else return c;  
20 }
```

This algorithm takes $O(n)$ time.

11

Solutions to Testing

11.1 Mistake:

Find the mistake(s) in the following code:

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

pg 157

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.

11.2 Random Crashes:

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 157

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 carefully. 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.

11.3 Chess Test: 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 157

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.

11.4 No Test Tools: How would you load test a webpage without using any test tools?

pg 157

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.

11.5 Test a Pen: How would you test a pen?

pg 157

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:** Okay, great. On what? Paper? Clothing? Walls?
- **Interviewer:** On clothing.
- **Candidate:** Great. What kind of tip does the pen have? Felt? Ballpoint? 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.

11.6 Test an ATM: How would you test an ATM in a distributed banking system?

pg 157

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.

12

Solutions to C and C++

12.1 Last K Lines: Write a method to print the last K lines of an input file using C++.

pg 163

SOLUTION

One brute force way could be to count the number of lines (N) and then print from $N-K$ to N th 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., 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     /* peek() so an EOF following a line ending is not considered a separate line */
9     while (file.peek() != EOF) {
10         getline(file, L[size % K]);
11         size++;
12     }
13
14     /* compute start of circular array, and the size of it */
15     int start = size > K ? (size % K) : 0;
16     int count = min(K, size);
17 }
```

```

18  /* print elements in the order they were read */
19  for (int i = 0; i < count; i++) {
20      cout << L[(start + i) % K] << endl;
21  }
22 }
```

This solution will require reading in the whole file, but only ten lines will be in memory at any given point.

- 12.2 Reverse String:** Implement a function void reverse(char* str) in C or C++ which reverses a null-terminated string.

pg 163

SOLUTION

This is a classic interview question. The only “gotcha” is to try to do it in place, and to be careful for the null character.

We will implement this in C.

```

1 void reverse(char *str) {
2     char* end = str;
3     char tmp;
4     if (str) {
5         while (*end) { /* find end of the string */
6             ++end;
7         }
8         --end; /* set one char back, since last char is null */
9
10        /* swap characters from start of string with the end of the string, until the
11           * pointers meet in middle. */
12        while (str < end) {
13            tmp = *str;
14            *str++ = *end;
15            *end-- = tmp;
16        }
17    }
18 }
```

This is just one of many ways to implement this solution. We could even implement this code recursively (but we wouldn't recommend it).

- 12.3 Hash Table vs STL Map:** 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 163

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.

12.4 Virtual Functions: How do virtual functions work in C++?

pg 164

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
10 class Triangle: public Shape {
11 public:
12     int circumference () {
13         cout<< "Circumference of Triangle Class\n";
14         return 3 * edge_length;
15     }
16 };
17
18 void main() {
19     Shape * x = new Shape();
20     x->circumference(); // "Circumference of Base Class"
21     Shape *y = new Triangle();
22     y->circumference(); // "Circumference of Triangle Class"
23 }

```

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.

12.5 Shallow vs Deep Copy: What is the difference between deep copy and shallow copy? Explain how you would use each.

pg 164

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.

12.6 Volatile: What is the significance of the keyword "volatile" in C?

pg 164

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.

12.7 Virtual Base Class: Why does a destructor in base class need to be declared virtual?

pg 164

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.

12.8 Copy Node: 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 164

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     Node * node = new Node;
14     nodeMap[cur] = node; // map current before traversing links
15     node->ptr1 = copy_recursive(cur->ptr1, nodeMap);
16     node->ptr2 = copy_recursive(cur->ptr2, nodeMap);
17     return node;
18 }
19
20
21 Node * copy_structure(Node * root) {
22     NodeMap nodeMap; // we will need an empty map
23     return copy_recursive(root, nodeMap);
24 }
```

- 12.9 Smart Pointer:** 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 164

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 itself and to the
3      * ref count. These must be pointers, rather than the actual object or ref count
4      * value, since the goal of a smart pointer is that the reference count is
5      * tracked across multiple smart pointers to one object. */
6     T * obj;
7     unsigned * ref_count;
8 }
```

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 counter to 1. */
3 }
4
5 SmartPointer(SmartPointer<T>& sptr) {
6     /* This constructor creates a new smart pointer that points to an existing
7      * object. We will need to first set obj and ref_count to pointer to sptr's obj
8      * and ref_count. Then, because we created a new reference to obj, we need to
9      * increment ref_count. */
10 }
11
12 ~SmartPointer(SmartPointer<T> sptr) {
13     /* We are destroying a reference to the object. Decrement ref_count. If
```

```

14     * ref_count is 0, then free the memory created by the integer and destroy the
15     * object. */
16 }

```

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.

```

1 onSetEquals(SmartPointer<T> ptr1, SmartPointer<T> ptr2) {
2     /* If ptr1 has an existing value, decrement its reference count. Then, copy the
3      * pointers to obj and ref_count over. Finally, since we created a new
4      * reference, we need to increment ref_count. */
5 }

```

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
9     SmartPointer(SmartPointer<T> & sptr) {
10        ref = sptr.ref;
11        ref_count = sptr.ref_count;
12        ++(*ref_count);
13    }
14
15    /* Override the equal operator, so that when you set one smart pointer equal to
16     * another the old smart pointer has its reference count decremented and the new
17     * smart pointer has its reference count incremented. */
18    SmartPointer<T> & operator=(SmartPointer<T> & sptr) {
19        if (this == &sptr) return *this;
20
21        /* If already assigned to an object, remove one reference. */
22        if (*ref_count > 0) {
23            remove();
24        }
25
26        ref = sptr.ref;
27        ref_count = sptr.ref_count;
28        ++(*ref_count);
29        return *this;
30    }
31
32    ~SmartPointer() {
33        remove(); // Remove one reference to object.
34    }
35
36    T getValue() {
37        return *ref;
38    }
39
40 protected:
41     void remove() {
42         --(*ref_count);
43         if (*ref_count == 0) {

```

```
44     delete ref;
45     free(ref_count);
46     ref = NULL;
47     ref_count = NULL;
48 }
49 }
50
51 T * ref;
52 unsigned * ref_count;
53 };
```

The code for this problem is complicated, and you probably wouldn't be expected to complete it flawlessly.

12.10 Malloc: 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.

EXAMPLE

`align_malloc(1000, 128)` will return a memory address that is a multiple of 128 and that points to memory of size 1000 bytes.

`aligned_free()` will free memory allocated by `align_malloc`.

pg 164

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 one of the first 16 memory address in the block at `p` must be divisible by 16. With `(p + 15) & 11...10000` we advance as need to this address. ANDing the last four bits of `p + 15` with `0000` guarantees that this new value will be divisible by 16 (either at the original `p` or in one of the following 15 addresses).

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.

Therefore, to guarantee both an aligned address and space for this pointer, we will need to allocate an additional $\text{alignment} - 1 + \text{sizeof}(\text{void}^*)$ bytes.

The code below implements this approach.

```

1 void* aligned_malloc(size_t required_bytes, size_t alignment) {
2     void* p1; // initial block
3     void* p2; // aligned block inside initial 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     ((void **)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 the pointer arithmetic in lines 9 and 15. If we treat p2 as a `void**` (or an array of `void*`'s), we can just look at the index -1 to retrieve p1.

In `aligned_free`, we take p2 as the same p2 returned from `aligned_malloc`. As before, we know that the value of p1 (which points to the beginning of the full memory block) was stored just before p2. By freeing p1, we deallocate the whole memory block.

12.11 2D Alloc: 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 164

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.

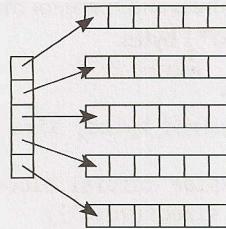
The basic idea is to create a one-dimensional array of pointers. Then, for each array index, we create a new one-dimensional array. This gives us a two-dimensional array that can be accessed via array indices.

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 following diagram represents how this memory is allocated.

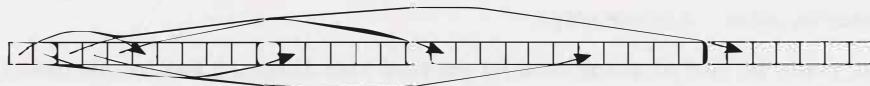


To free this memory, we cannot simply call `free` on `rowptr`. We need to make sure to free not only the memory from the first `malloc` call, but also each subsequent call.

```

1 void my2DDAlloc(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) return NULL;
7
8     int* buf = (int*) (rowptr + rows);
9     for (i = 0; i < rows; i++) {
10         rowptr[i] = buf + i * cols;
11     }
12     return rowptr;
13 }
```

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.

13

Solutions to Java

13.1 Private Constructor: In terms of inheritance, what is the effect of keeping a constructor private?

pg 167

SOLUTION

Declaring a constructor `private` on class A means that you can only access the (private) constructor if you could also access A's private methods. Who, other than A, can access A's private methods and constructor? A's inner classes can. Additionally, if A is an inner class of Q, then Q's other inner classes can.

This has direct implications for inheritance, since a subclass calls its parent's constructor. The class A can be inherited, but only by its own or its parent's inner classes.

13.2 Return from Finally: In Java, does the finally block get executed if we insert a return statement inside the try block of a try-catch-finally?

pg 167

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 during the `try/catch` block gets killed.

13.3 Final, etc.: What is the difference between `final`, `finally`, and `finalize`?

pg 167

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 keyword

There is an optional `finally` block after the `try` block or after the `catch` block. Statements in the `finally` block will always be executed, even if an exception is thrown (except if Java Virtual Machine exits from the `try` block). The `finally` block is often used to write the clean-up code. It will be executed after the `try` and `catch` blocks, but before control transfers back to its origin.

Watch how this plays out in the example below.

```
1  public static String lem() {  
2      System.out.println("lem");  
3      return "return from lem";  
4  }  
5  
6  public static String foo() {  
7      int x = 0;  
8      int y = 5;  
9      try {  
10          System.out.println("start try");  
11          int b = y / x;  
12          System.out.println("end try");  
13          return "returned from try";  
14      } catch (Exception ex) {  
15          System.out.println("catch");  
16          return lem() + " | returned from catch";  
17      } finally {  
18          System.out.println("finally");  
19      }  
20  }  
21  
22 public static void bar() {  
23     System.out.println("start bar");  
24     String v = foo();  
25     System.out.println(v);  
26     System.out.println("end bar");  
27  }  
28  
29 public static void main(String[] args) {  
30     bar();  
31 }
```

The output for this code is the following:

```
1  start bar
```

```

2 start try
3 catch
4 lem
5 finally
6 return from lem | returned from catch
7 end bar

```

Look carefully at lines 3 to 5 in the output. The `catch` block is fully executed (including the function call in the `return` statement), then the `finally` block, and then the function actually returns.

finalize()

The automatic garbage collector calls the `finalize()` method just before actually destroying the object. A class can therefore override the `finalize()` method from the `Object` class in order to define custom behavior during garbage collection.

```

1 protected void finalize() throws Throwable {
2     /* Close open files, release resources, etc */
3 }

```

13.4 Generics vs. Templates: Explain the difference between templates in C++ and generics in Java.

pg 167

SOLUTION

Many programmers consider templates and generics to be essentially equivalent because both allow you to do something like `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 are shared across instances of `MyClass`, regardless of the different type parameters.

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: Although Java generics and C++ templates look the same in many ways, they are very different.

13.5 TreeMap, HashMap, LinkedHashMap: Explain the differences between TreeMap, HashMap, and LinkedHashMap. Provide an example of when each one would be best.

pg 167

SOLUTION

All offer a key->value map and a way to iterate through the keys. The most important distinction between these classes is the time guarantees and the ordering of the keys.

- `HashMap` offers $O(1)$ lookup and insertion. If you iterate through the keys, though, the ordering of the keys is essentially arbitrary. It is implemented by an array of linked lists.
- `TreeMap` offers $O(\log N)$ lookup and insertion. Keys are ordered, so if you need to iterate through the keys in sorted order, you can. This means that keys must implement the `Comparable` interface. `TreeMap` is implemented by a Red-Black Tree.
- `LinkedHashMap` offers $O(1)$ lookup and insertion. Keys are ordered by their insertion order. It is implemented by doubly-linked buckets.

Imagine you passed an empty `TreeMap`, `HashMap`, and `LinkedHashMap` into the following function:

```

1 void insertAndPrint(AbstractMap<Integer, String> map) {
2     int[] array = {1, -1, 0};
3     for (int x : array) {
4         map.put(x, Integer.toString(x));
5     }
6
7     for (int k : map.keySet()) {
8         System.out.print(k + ", ");
9     }
10 }
```

The output for each will look like the results below.

HashMap	LinkedHashMap	TreeMap
(any ordering)	{1, -1, 0}	{-1, 0, 1}

Very important: The output of LinkedHashMap and TreeMap must look like the above. For HashMap, the output was, in my own tests, {0, 1, -1}, but it could be any ordering. There is no *guarantee* on the ordering.

When might you need ordering in real life?

- Suppose you were creating a mapping of names to Person objects. You might want to periodically output the people in alphabetical order by name. A TreeMap lets you do this.
- A TreeMap also offers a way to, given a name, output the next 10 people. This could be useful for a "More" function in many applications.
- A LinkedHashMap is useful whenever you need the ordering of keys to match the ordering of insertion. This might be useful in a caching situation, when you want to delete the oldest item.

Generally, unless there is a reason not to, you would use HashMap. That is, if you need to get the keys back in insertion order, then use LinkedHashMap. If you need to get the keys back in their true/natural order, then use TreeMap. Otherwise, HashMap is probably best. It is typically faster and requires less overhead.

13.6 Object Reflection:

Explain what object reflection is in Java and why it is useful.

pg 168

SOLUTION

Object Reflection is a feature in Java that provides a way to get reflective information about Java classes and objects, and perform operations such as:

- Getting information about the methods and fields present inside the class at runtime.
- Creating a new instance of a class.
- 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 = (Rectangle) doubleArgsConstructor.newInstance(doubleArgs);
12
13 /* Equivalent: Double area = rectangle.area(); */
14 Method m = rectangleDefinition.getDeclaredMethod("area");
15 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 some cases. Three main reasons are:

1. It can help you observe or manipulate the runtime behavior of applications.
2. It can help you debug or test programs, as you have direct access to methods, constructors, and fields.
3. You can call methods by name when you 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.

13.7 Lambda Expressions: There is a class `Country` that has methods `getContinent()` and `getPopulation()`. Write a function `int getPopulation(List<Country> countries, String continent)` that computes the total population of a given continent, given a list of all countries and the name of a continent.

pg 168

SOLUTION

This question really comes in two parts. First, we need to generate a list of the countries in North America. Then, we need to compute their total population.

Without lambda expressions, this is fairly straightforward to do.

```
1 int getPopulation(List<Country> countries, String continent) {
2     int sum = 0;
3     for (Country c : countries) {
4         if (c.getContinent().equals(continent)) {
5             sum += c.getPopulation();
6         }
7     }
8     return sum;
9 }
```

To implement this with lambda expressions, let's break this up into multiple parts.

First, we use `filter` to get a list of the countries in the specified continent.

```
1 Stream<Country> northAmerica = countries.stream().filter(
2     country -> { return country.getContinent().equals(continent); })
```

```
3 );
```

Second, we convert this into a list of populations using `map`.

```
1 Stream<Integer> populations = northAmerica.map(
2   c -> c.getPopulation()
3 );
```

Third and finally, we compute the sum using `reduce`.

```
1 int population = populations.reduce(0, (a, b) -> a + b);
```

This function puts it all together.

```
1 int getPopulation(List<Country> countries, String continent) {
2   /* Filter countries. */
3   Stream<Country> sublist = countries.stream().filter(
4     country -> { return country.getContinent().equals(continent); }
5   );
6
7   /* Convert to list of populations. */
8   Stream<Integer> populations = sublist.map(
9     c -> c.getPopulation()
10  );
11
12  /* Sum list. */
13  int population = populations.reduce(0, (a, b) -> a + b);
14  return population;
15 }
```

Alternatively, because of the nature of this specific problem, we can actually remove the `filter` entirely. The `reduce` operation can have logic that maps the population of countries not in the right continent to zero. The sum will effectively disregard countries not within `continent`.

```
1 int getPopulation(List<Country> countries, String continent) {
2   Stream<Integer> populations = countries.stream().map(
3     c -> c.getContinent().equals(continent) ? c.getPopulation() : 0);
4   return populations.reduce(0, (a, b) -> a + b);
5 }
```

Lambda functions were new to Java 8, so if you don't recognize them, that's probably why. Now is a great time to learn about them, though!

13.8 Lambda Random: Using Lambda expressions, write a function `List<Integer> getRandomSubset(List<Integer> list)` that returns a random subset of arbitrary size. All subsets (including the empty set) should be equally likely to be chosen.

pg 439

SOLUTION

It's tempting to approach this problem by picking a subset size from 0 to N and then generating a random subset of that size.

That creates two issues:

1. We'd have to weight those probabilities. If $N > 1$, there are more subsets of size $N/2$ than there are of subsets of size N (of which there is always only one).
2. It's actually more difficult to generate a subset of a restricted size (e.g., specifically 10) than it is to generate a subset of any size.

Instead, rather than generating a subset based on sizes, let's think about it based on elements. (The fact that we're told to use lambda expressions is also a hint that we should think about some sort of iteration or processing through the elements.)

Imagine we were iterating through {1, 2, 3} to generate a subset. Should 1 be in this subset?

We've got two choices: yes or no. We need to weight the probability of "yes" vs. "no" based on the percent of subsets that contain 1. So, what percent of elements contain 1?

For any specific element, there are as many subsets that contain the element as do not contain it. Consider the following:

{}	{1}
{2}	{1, 2}
{3}	{1, 3}
{2, 3}	{1, 2, 3}

Note how the difference between the subsets on the left and the subsets on the right is the existence of 1. The left and right sides must have the same number of subsets because we can convert from one to the other by just adding an element.

This means that we can generate a random subset by iterating through the list and flipping a coin (i.e., deciding on a 50/50 chance) to pick whether or not each element will be in it.

Without lambda expressions, we can write something like this:

```
1 List<Integer> getRandomSubset(List<Integer> list) {
2     List<Integer> subset = new ArrayList<Integer>();
3     Random random = new Random();
4     for (int item : list) {
5         /* Flip coin. */
6         if (random.nextBoolean()) {
7             subset.add(item);
8         }
9     }
10    return subset;
11 }
```

To implement this approach using lambda expressions, we can do the following:

```
1 List<Integer> getRandomSubset(List<Integer> list) {
2     Random random = new Random();
3     List<Integer> subset = list.stream().filter(
4         k -> { return random.nextBoolean(); /* Flip coin. */ }
5     ).collect(Collectors.toList());
6     return subset;
7 }
```

Or, we can use a predicate (defined within the class or within the function):

```
1 Random random = new Random();
2 Predicate<Object> flipCoin = o -> {
3     return random.nextBoolean();
4 };
5
6 List<Integer> getRandomSubset(List<Integer> list) {
7     List<Integer> subset = list.stream().filter(flipCoin).
8         collect(Collectors.toList());
9     return subset;
10 }
```

The nice thing about this implementation is that now we can apply the `flipCoin` predicate in other places.

14

Solutions to Databases

Questions 1 through 3 refer to the following database schema:

Apartments	
AptID	int
UnitNumber	varchar(10)
BuildingID	int

Buildings	
BuildingID	int
ComplexID	int
BuildingName	varchar(100)
Address	varchar(500)

Requests	
RequestID	int
Status	varchar(100)
AptID	int
Description	varchar(500)

Complexes	
ComplexID	int
ComplexName	varchar(100)

AptTenants	
TenantID	int
AptID	int

Tenants	
TenantID	int
TenantName	varchar(100)

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.

14.1 Multiple Apartments: Write a SQL query to get a list of tenants who are renting more than one apartment.

pg 172

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 FROM AptTenants GROUP BY TenantID HAVING count(*) > 1) C
5 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.

- 14.2 Open Requests:** Write a SQL query to get a list of all buildings and the number of open requests (Requests in which status equals 'Open').

pg 173

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.

- 14.3 Close All Requests:** Building #11 is undergoing a major renovation. Implement a query to close all requests from apartments in this building.

pg 173

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 (SELECT AptID FROM Apartments WHERE BuildingID = 11)
```

- 14.4 Joins:** What are the different types of joins? Please explain how they differ and why certain types are better in certain situations.

pg 173

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

Name	Code
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.

14.5 Denormalization: What is denormalization? Explain the pros and cons.

pg 173

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.

14.6 Entity-Relationship Diagram: Draw an entity-relationship diagram for a database with companies, people, and professionals (people who work for companies).

pg 173

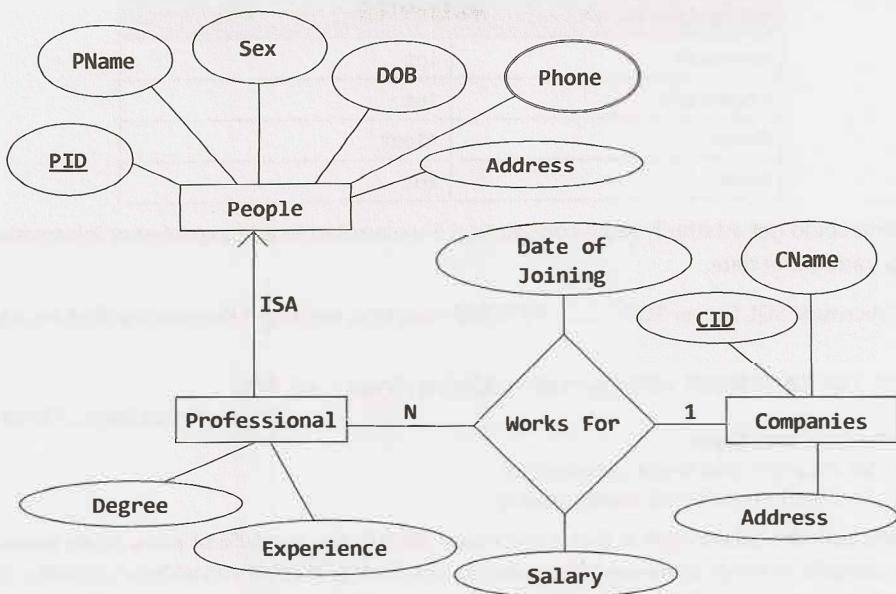
SOLUTION

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 (probably—you might want to validate this assumption), 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.



- 14.7 Design Grade Database:** 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 173

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  SELECT TOP 10 PERCENT AVG(CourseEnrollment.Grade) AS GPA,
2                               CourseEnrollment.StudentID
3  FROM CourseEnrollment
4  GROUP BY CourseEnrollment.StudentID
5  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' FROM (
3      SELECT TOP 10 PERCENT AVG(CourseEnrollment.Grade) AS GPA
4      FROM CourseEnrollment
5      GROUP BY CourseEnrollment.StudentID
6      ORDER BY GPA desc) Grades);

```

Then, once we have @GPACutOff defined, selecting the students with at least this GPA is reasonably straightforward.

```

1  SELECT StudentName, GPA
2  FROM (SELECT AVG(CourseEnrollment.Grade) AS GPA, CourseEnrollment.StudentID
3         FROM CourseEnrollment
4         GROUP BY CourseEnrollment.StudentID
5         HAVING AVG(CourseEnrollment.Grade) >= @GPACutOff) Honors
6  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.

15

Solutions to Threads and Locks

15.1 Thread vs. Process: What's the difference between a thread and a process?

pg 179

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.

15.2 Context Switch: How would you measure the time spent in a context switch?

pg 179

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 N th instruction of P_1 . If $t_{x,k}$ indicates the timestamp in microseconds of the k th 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:

1. P_2 blocks awaiting data from P_1 .
2. P_1 marks the start time.
3. P_1 sends token to P_2 .
4. P_1 attempts to read a response token from P_2 . This induces a context switch.
5. P_2 is scheduled and receives the token.
6. P_2 sends a response token to P_1 .
7. P_2 attempts read a response token from P_1 . This induces a context switch.
8. P_1 is scheduled and receives the token.
9. 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_c , 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.

15.3 Dining Philosophers: 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 180

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  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 class Philosopher extends Thread {
18     private int bites = 10;
19     private Chopstick left, right;
20
21     public Philosopher(Chopstick left, Chopstick right) {
22         this.left = left;
23         this.right = right;
24     }

```

```
25     public void eat() {
26         pickUp();
27         chew();
28         putDown();
29     }
30
31
32     public void pickUp() {
33         left.pickUp();
34         right.pickUp();
35     }
36
37     public void chew() { }
38
39     public void putDown() {
40         right.putDown();
41         left.putDown();
42     }
43
44     public void run() {
45         for (int i = 0; i < bites; i++) {
46             eat();
47         }
48     }
49 }
```

Running the above code may lead to a deadlock if all the philosophers have a left chopstick and are waiting for the right one.

Solution #1: All or Nothing

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.

One issue with this is that if all the philosophers were perfectly synchronized, they could simultaneously pick up their left chopstick, be unable to pick up the right one, and then put back down the left one—only to have the process repeated again.

Solution #2: Prioritized Chopsticks

Alternatively, we can label the chopsticks with a number from 0 to $N - 1$. Each philosopher attempts to pick up the lower numbered chopstick first. This essentially means that each philosopher goes for the left chopstick before right one (assuming that's the way you labeled it), except for the last philosopher who does this in reverse. This will break the cycle.

```

1  public class Philosopher extends Thread {
2      private int bites = 10;
3      private Chopstick lower, higher;
4      private int index;
5      public Philosopher(int i, Chopstick left, Chopstick right) {
6          index = i;
7          if (left.getNumber() < right.getNumber()) {
8              this.lower = left;
9              this.higher = right;
10         } else {
11             this.lower = right;
12             this.higher = left;
13         }
14     }
15
16    public void eat() {
17        pickUp();
18        chew();
19        putDown();
20    }
21
22    public void pickUp() {
23        lower.pickUp();
24        higher.pickUp();
25    }
26
27    public void chew() { ... }
28
29    public void putDown() {
30        higher.putDown();
31        lower.putDown();
32    }
33
34    public void run() {
35        for (int i = 0; i < bites; i++) {
36            eat();
```

```
37     }
38 }
39 }
40
41 public class Chopstick {
42     private Lock lock;
43     private int number;
44
45     public Chopstick(int n) {
46         lock = new ReentrantLock();
47         this.number = n;
48     }
49
50     public void pickUp() {
51         lock.lock();
52     }
53
54     public void putDown() {
55         lock.unlock();
56     }
57
58     public int getNumber() {
59         return number;
60     }
61 }
```

With this solution, a philosopher can never hold the larger chopstick without holding the smaller one. This prevents the ability to have a cycle, since a cycle means that a higher chopstick would “point” to a lower one.

15.4 Deadlock-Free Class:

Design a class which provides a lock only if there are no possible deadlocks.

pg 180

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 `declare` order, 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 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 find 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 }

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  class LockFactory {
2      private static LockFactory instance;
3
4      private int numberLocks = 5; /* default */
5      private LockNode[] locks;
6
7      /* Maps from a process or owner to the order that the owner claimed it would
8       * call the locks in */
9      private HashMap<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(HashMap<Integer, Boolean> touchedNodes,
20                         int[] resourcesInOrder) {
21     /*check for a cycle */
22     for (int resource : resourcesInOrder) {
23         if (touchedNodes.get(resource) == false) {
24             LockNode n = locks[resource];
25             if (n.hasCycle(touchedNodes)) {
26                 return true;
27             }
28         }
29     }
30     return false;
31 }
32
33 /*To prevent deadlocks, force the processes to declare upfront what order they
34 * will need the locks in. Verify that this order does not create a deadlock (a
35 * cycle in a directed graph) */
36 public boolean declare(int ownerId, int[] resourcesInOrder) {
37     HashMap<Integer, Boolean> touchedNodes = new HashMap<Integer, Boolean>();
38
39     /*add nodes to graph */
40     int index = 1;
41     touchedNodes.put(resourcesInOrder[0], false);
42     for (index = 1; index < resourcesInOrder.length; index++) {
43         LockNode prev = locks[resourcesInOrder[index - 1]];
44         LockNode curr = locks[resourcesInOrder[index]];
45         prev.joinTo(curr);
46         touchedNodes.put(resourcesInOrder[index], false);
47     }
48
49     /*if we created a cycle, destroy this resource list and return false */
50     if (hasCycle(touchedNodes, resourcesInOrder)) {
51         for (int j = 1; j < resourcesInOrder.length; j++) {
52             LockNode p = locks[resourcesInOrder[j - 1]];
53             LockNode c = locks[resourcesInOrder[j]];
54             p.remove(c);
55         }
56         return false;
57     }
58
59     /*No cycles detected. Save the order that was declared, so that we can
60     * verify that the process is really calling the locks in the order it said
61     * it would. */
62     LinkedList<LockNode> list = new LinkedList<LockNode>();
63     for (int i = 0; i < resourcesInOrder.length; i++) {
64         LockNode resource = locks[resourcesInOrder[i]];
65         list.add(resource);
66     }
67     lockOrder.put(ownerId, list);
68
69     return true;
70 }
71 }
```

```

72  /* Get the lock, verifying first that the process is really calling the locks in
73   * the order it said it would. */
74  public Lock getLock(int ownerId, int resourceId) {
75      LinkedList<LockNode> list = lockOrder.get(ownerId);
76      if (list == null) return null;
77
78      LockNode head = list.getFirst();
79      if (head.getId() == resourceId) {
80          list.removeFirst();
81          return head.getLock();
82      }
83      return null;
84  }
85 }
86
87 public class LockNode {
88     public enum VisitState { FRESH, VISITING, VISITED };
89
90     private ArrayList<LockNode> children;
91     private int lockId;
92     private Lock lock;
93     private int maxLocks;
94
95     public LockNode(int id, int max) { ... }
96
97     /* Join "this" to "node", checking that it doesn't create a cycle */
98     public void joinTo(LockNode node) { children.add(node); }
99     public void remove(LockNode node) { children.remove(node); }
100
101    /* Check for a cycle by doing a depth-first-search. */
102    public boolean hasCycle(HashMap<Integer, Boolean> touchedNodes) {
103        VisitState[] visited = new VisitState[maxLocks];
104        for (int i = 0; i < maxLocks; i++) {
105            visited[i] = VisitState.FRESH;
106        }
107        return hasCycle(visited, touchedNodes);
108    }
109
110    private boolean hasCycle(VisitState[] visited,
111                             HashMap<Integer, Boolean> touchedNodes) {
112        if (touchedNodes.containsKey(lockId)) {
113            touchedNodes.put(lockId, true);
114        }
115
116        if (visited[lockId] == VisitState.VISITING) {
117            /* We looped back to this node while still visiting it, so we know there's
118             * a cycle. */
119            return true;
120        } else if (visited[lockId] == VisitState.FRESH) {
121            visited[lockId] = VisitState.VISITING;
122            for (LockNode n : children) {
123                if (n.hasCycle(visited, touchedNodes)) {
124                    return true;
125                }
126            }
127            visited[lockId] = VisitState.VISITED;

```

```

128     }
129     return false;
130 }
131
132 public Lock getLock() {
133     if (lock == null) lock = new ReentrantLock();
134     return lock;
135 }
136
137 public int getId() { return lockId; }
138 }
```

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.

15.5 Call In Order:

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 180

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;
4
5      public FooBad() {
6          try {
7              lock1 = new ReentrantLock();
8              lock2 = new ReentrantLock();
9
10             lock1.lock();
11             lock2.lock();
12         } catch (...) { ... }
13     }
14
15     public void first() {
16         try {
17             ...
18             lock1.unlock(); // mark finished with first()
19         } catch (...) { ... }
20     }
```

```

21
22     public void second() {
23         try {
24             lock1.lock(); // wait until finished with first()
25             lock1.unlock();
26             ...
27
28             lock2.unlock(); // mark finished with second()
29         } catch (...) { ... }
30     }
31
32     public void third() {
33         try {
34             lock2.lock(); // wait until finished with third()
35             lock2.unlock();
36             ...
37         } catch (...) { ... }
38     }
39 }
```

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;
3
4     public Foo() {
5         try {
6             sem1 = new Semaphore(1);
7             sem2 = new Semaphore(1);
8
9             sem1.acquire();
10            sem2.acquire();
11        } catch (...) { ... }
12    }
13
14    public void first() {
15        try {
16            ...
17            sem1.release();
18        } catch (...) { ... }
19    }
20
21    public void second() {
22        try {
23            sem1.acquire();
24            sem1.release();
25            ...
26            sem2.release();
27        } catch (...) { ... }
28    }
29
30    public void third() {
31        try {
32            sem2.acquire();
```

```
33         sem2.release();  
34         ...  
35     } catch (...) { ... }  
36 }  
37 }
```

- 15.6 Synchronized Methods:** 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 180

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.

Ultimately, the key concept to remember 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.

- 15.7 FizzBuzz:** In the classic problem FizzBuzz, you are told to print the numbers from 1 to n. However, when the number is divisible by 3, print “Fizz”. When it is divisible by 5, print “Buzz”. When it is divisible by 3 and 5, print “FizzBuzz”. In this problem, you are asked to do this in a multithreaded way. Implement a multithreaded version of FizzBuzz with four threads. One thread checks for divisibility of 3 and prints “Fizz”. Another thread is responsible for divisibility of 5 and prints “Buzz”. A third thread is responsible for divisibility of 3 and 5 and prints “FizzBuzz”. A fourth thread does the numbers.

pg 180

SOLUTION

Let’s start off with implementing a single threaded version of FizzBuzz.

Single Threaded

Although this problem (in the single threaded version) shouldn’t be hard, a lot of candidates overcomplicate it. They look for something “beautiful” that reuses the fact that the divisible by 3 and 5 case (“FizzBuzz”) seems to resemble the individual cases (“Fizz” and “Buzz”).

In actuality, the best way to do it, considering readability and efficiency, is just the straightforward way.

```
1 void fizzbuzz(int n) {
```

```

2   for (int i = 1; i <= n; i++) {
3     if (i % 3 == 0 && i % 5 == 0) {
4       System.out.println("FizzBuzz");
5     } else if (i % 3 == 0) {
6       System.out.println("Fizz");
7     } else if (i % 5 == 0) {
8       System.out.println("Buzz");
9     } else {
10      System.out.println(i);
11    }
12  }
13 }
```

The primary thing to be careful of here is the order of the statements. If you put the check for divisibility by 3 before the check for divisibility by 5, it won't print the right thing.

Multithreaded

To do this multithreaded, we want a structure that looks something like this:

FizzBuzz Thread	Fizz Thread
if i div by 3 && 5 print FizzBuzz increment i repeat until i > n	if i div by only 3 print Fizz increment i repeat until i > n
Buzz Thread	Number Thread
if i div by only 5 print Buzz increment i repeat until i > n	if i not div by 3 or 5 print i increment i repeat until i > n

The code for this will look something like:

```

1 while (true) {
2   if (current > max) {
3     return;
4   }
5   if /* divisibility test */ {
6     System.out.println/* print something */;
7     current++;
8   }
9 }
```

We'll need to add some synchronization in the loop. Otherwise, the value of `current` could change between lines 2 - 4 and lines 5 - 8, and we can inadvertently exceed the intended bounds of the loop. Additionally, incrementing is not thread-safe.

To actually implement this concept, there are many possibilities. One possibility is to have four entirely separate thread classes that share a reference to the `current` variable (which can be wrapped in an object).

The loop for each thread is substantially similar. They just have different target values for the divisibility checks, and different print values.

	FizzBuzz	Fizz	Buzz	Number
current % 3 == 0	true	true	false	false
current % 5 == 0	true	false	true	false
to print	FizzBuzz	Fizz	Buzz	current

For the most part, this can be handled by taking in “target” parameters and the value to print. The output for the Number thread needs to be overwritten, though, as it’s not a simple, fixed string.

We can implement a FizzBuzzThread class which handles most of this. A NumberThread class can extend FizzBuzzThread and override the print method.

```

1  Thread[] threads = {new FizzBuzzThread(true, true, n, "FizzBuzz"),
2                      new FizzBuzzThread(true, false, n, "Fizz"),
3                      new FizzBuzzThread(false, true, n, "Buzz"),
4                      new NumberThread(false, false, n)};
5  for (Thread thread : threads) {
6      thread.start();
7  }
8
9  public class FizzBuzzThread extends Thread {
10     private static Object lock = new Object();
11     protected static int current = 1;
12     private int max;
13     private boolean div3, div5;
14     private String toPrint;
15
16     public FizzBuzzThread(boolean div3, boolean div5, int max, String toPrint) {
17         this.div3 = div3;
18         this.div5 = div5;
19         this.max = max;
20         this.toPrint = toPrint;
21     }
22
23     public void print() {
24         System.out.println(toPrint);
25     }
26
27     public void run() {
28         while (true) {
29             synchronized (lock) {
30                 if (current > max) {
31                     return;
32                 }
33
34                 if ((current % 3 == 0) == div3 &&
35                     (current % 5 == 0) == div5) {
36                     print();
37                     current++;
38                 }
39             }
40         }
41     }
42 }
43
44 public class NumberThread extends FizzBuzzThread {

```

```

45     public NumberThread(boolean div3, boolean div5, int max) {
46         super(div3, div5, max, null);
47     }
48
49     public void print() {
50         System.out.println(current);
51     }
52 }
```

Observe that we need to put the comparison of `current` and `max` before the if statement, to ensure the value will only get printed when `current` is less than or equal to `max`.

Alternatively, if we're working in a language which supports this (Java 8 and many other languages do), we can pass in a `validate` method and a `print` method as parameters.

```

1  int n = 100;
2  Thread[] threads = {
3      new FBThread(i -> i % 3 == 0 && i % 5 == 0, i -> "FizzBuzz", n),
4      new FBThread(i -> i % 3 == 0 && i % 5 != 0, i -> "Fizz", n),
5      new FBThread(i -> i % 3 != 0 && i % 5 == 0, i -> "Buzz", n),
6      new FBThread(i -> i % 3 != 0 && i % 5 != 0, i -> Integer.toString(i), n)};
7  for (Thread thread : threads) {
8      thread.start();
9  }
10
11 public class FBThread extends Thread {
12     private static Object lock = new Object();
13     protected static int current = 1;
14     private int max;
15     private Predicate<Integer> validate;
16     private Function<Integer, String> printer;
17     int x = 1;
18
19     public FBThread(Predicate<Integer> validate,
20                     Function<Integer, String> printer, int max) {
21         this.validate = validate;
22         this.printer = printer;
23         this.max = max;
24     }
25
26     public void run() {
27         while (true) {
28             synchronized (lock) {
29                 if (current > max) {
30                     return;
31                 }
32                 if (validate.test(current)) {
33                     System.out.println(printer.apply(current));
34                     current++;
35                 }
36             }
37         }
38     }
39 }
```

There are of course many other ways of implementing this as well.

16

Solutions to Moderate

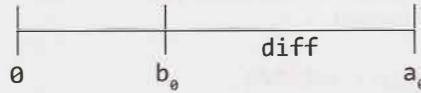
16.1 Number Swapper: Write a function to swap a number in place (that is, without temporary variables).

pg 181

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 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 diff , which is the right side of the above number line. Then, when we add b and 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 // Example for a = 9, b = 4
2 a = a - b; // a = 9 - 4 = 5
3 b = a + b; // b = 5 + 4 = 9
4 a = b - a; // a = 9 - 5
```

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 // Example for a = 101 (in binary) and b = 110
2 a = a^b; // a = 101^110 = 011
3 b = a^b; // b = 011^110 = 101
4 a = a^b; // a = 011^101 = 110
```

This code works by using XORs. The easiest way to see how this works is by focusing on a specific bit. If we can correctly swap two bits, then we know the entire operation works correctly.

Let's take two bits, x and y , and walk through this line by line.

1. $x = x \wedge y$

This line essentially checks if x and y have different values. It will result in 1 if and only if $x \neq y$.

2. $y = x \wedge y$

Or: $y = \{0 \text{ if originally same, 1 if different}\} \wedge \{\text{original } y\}$

Observe that XORing a bit with 1 always flips the bit, whereas XORing with 0 will never change it.

Therefore, if we do $y = 1 \wedge \{\text{original } y\}$ when $x \neq y$, then y will be flipped and therefore have x 's original value.

Otherwise, if $x == y$, then we do $y = 0 \wedge \{\text{original } y\}$ and the value of y does not change.

Either way, y will be equal to the original value of x .

3. $x = x \wedge y$

Or: $x = \{0 \text{ if originally same, 1 if different}\} \wedge \{\text{original } x\}$

At this point, y is equal to the original value of x . This line is essentially equivalent to the line above it, but for different variables.

If we do $x = 1 \wedge \{\text{original } x\}$ when the values are different, x will be flipped.

If we do $x = 0 \wedge \{\text{original } x\}$ when the values are the same, x will not be changed.

This operation happens for each bit. Since it correctly swaps each bit, it will correctly swap the entire number.

16.2 Word Frequencies: Design a method to find the frequency of occurrences of any given word in a book. What if we were running this algorithm multiple times?

pg 181

SOLUTION

Let's start with the simple case.

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.

```

1 int getFrequency(String[] book, String word) {
2     word = word.trim().toLowerCase();
3     int count = 0;
4     for (String w : book) {
5         if (w.trim().toLowerCase().equals(word)) {
6             count++;
7         }
8     }
9     return count;
10 }
```

We have also converted the string to lowercase and trimmed it. You can discuss with your interviewer if this is necessary (or even desired).

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 HashMap<String, Integer> setupDictionary(String[] book) {
2     HashMap<String, Integer> table =
```

```
3     new HashMap<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(HashMap<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?

16.3 Intersection: Given two straight line segments (represented as a start point and an end point), compute the point of intersection, if any.

pg 181

SOLUTION

We first need to think about what it means for two line segments to intersect.

For two infinite lines to intersect, they only have to have different slopes. If they have the same slope, then they must be the exact same line (same y-intercept). That is:

slope 1 != slope 2
OR
slope 1 == slope 2 AND intersect 1 == intersect 2

For two straight lines to intersect, the condition above must be true, *plus* the point of intersection must be within the ranges of each line segment.

extended infinite segments intersect
AND
intersection is within line segment 1 (x and y coordinates)
AND
intersection is within line segment 2 (x and y coordinates)

What if the two segments represent the same infinite line? In this case, we have to ensure that some portion of their segments overlap. If we order the line segments by their x locations (start is before end, point 1 is before point 2), then an intersection occurs only if:

Assume:
start1.x < start2.x && start1.x < end1.x && start2.x < end2.x
Then intersection occurs if:
start2 is between start1 and end1

We can now go ahead and implement this algorithm.

```

1 Point intersection(Point start1, Point end1, Point start2, Point end2) {
2     /* Rearranging these so that, in order of x values: start is before end and
3      * point 1 is before point 2. This will make some of the later logic simpler. */
4     if (start1.x > end1.x) swap(start1, end1);
5     if (start2.x > end2.x) swap(start2, end2);
6     if (start1.x > start2.x) {
7         swap(start1, start2);
8         swap(end1, end2);
9     }
10
11    /* Compute lines (including slope and y-intercept). */
12    Line line1 = new Line(start1, end1);
13    Line line2 = new Line(start2, end2);
14
15    /* If the lines are parallel, they intercept only if they have the same y
16     * intercept and start 2 is on line 1. */
17    if (line1.slope == line2.slope) {
18        if (line1.yintercept == line2.yintercept &&
19            isBetween(start1, start2, end1)) {
20            return start2;
21        }
22        return null;
23    }
24
25    /* Get intersection coordinate. */
26    double x = (line2.yintercept - line1.yintercept) / (line1.slope - line2.slope);
27    double y = x * line1.slope + line1.yintercept;
28    Point intersection = new Point(x, y);
29
30    /* Check if within line segment range. */
31    if (isBetween(start1, intersection, end1) &&
32        isBetween(start2, intersection, end2)) {
33        return intersection;
34    }
35    return null;
36}
37
38 /* Checks if middle is between start and end. */
39 boolean isBetween(double start, double middle, double end) {
40     if (start > end) {
41         return end <= middle && middle <= start;
42     } else {
43         return start <= middle && middle <= end;
44     }
45}
46
47 /* Checks if middle is between start and end. */
48 boolean isBetween(Point start, Point middle, Point end) {
49     return isBetween(start.x, middle.x, end.x) &&
50            isBetween(start.y, middle.y, end.y);
51}
52
53 /* Swap coordinates of point one and two. */
54 void swap(Point one, Point two) {
55     double x = one.x;
56     double y = one.y;

```

```
57     one.setLocation(two.x, two.y);
58     two.setLocation(x, y);
59 }
60
61 public class Line {
62     public double slope, yintercept;
63
64     public Line(Point start, Point end) {
65         double deltaY = end.y - start.y;
66         double deltaX = end.x - start.x;
67         slope = deltaY / deltaX; // Will be Infinity (not exception) when deltaX = 0
68         yintercept = end.y - slope * end.x;
69     }
70
71     public class Point {
72         public double x, y;
73         public Point(double x, double y) {
74             this.x = x;
75             this.y = y;
76         }
77
78         public void setLocation(double x, double y) {
79             this.x = x;
80             this.y = y;
81         }
82     }
83 }
```

For simplicity and compactness (it really makes the code easier to read), we've chosen to make the variables within `Point` and `Line` `public`. You can discuss with your interviewer the advantages and disadvantages of this choice.

16.4 Tic Tac Win: Design an algorithm to figure out if someone has won a game of tic-tac-toe.

pg 181

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 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 to optimize the runtime of `hasWon`.
2. Do we know the last move that was made?
3. 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 $N \times N$ solution?
4. In general, how much do we prioritize compactness of code versus speed of execution vs. clarity of code? Remember: 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). Therefore, we can 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 Piece 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 $3^0v_0 + 3^1v_1 + 3^2v_2 + \dots + 3^8v_8$, 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 enum Piece { Empty, Red, Blue };
2
3 int convertBoardToInt(Piece[][] board) {
4     int sum = 0;
5     for (int i = 0; i < board.length; i++) {
6         for (int j = 0; j < board[i].length; j++) {
7             /* Each value in enum has an integer associated with it. We
8              * can just use that. */
9             int value = board[i][j].ordinal();
10            sum = sum * 3 + value;
11        }
12    }
13    return sum;
14 }
```

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: If we know the last move

If we know the very last move that was made (and we’ve been checking for a winner up until now), then we only need to check the row, column, and diagonal that overlaps with this position.

```
1 Piece hasWon(Piece[][] board, int row, int column) {
2     if (board.length != board[0].length) return Piece.Empty;
3
4     Piece piece = board[row][column];
5
6     if (piece == Piece.Empty) return Piece.Empty;
7
8     if (hasWonRow(board, row) || hasWonColumn(board, column)) {
9         return piece;
10    }
11
12    if (row == column && hasWonDiagonal(board, 1)) {
13        return piece;
14    }
15
16    if (row == (board.length - column - 1) && hasWonDiagonal(board, -1)) {
17        return piece;
18    }
```

```
19     return Piece.Empty;
20 }
21 }
22
23 boolean hasWonRow(Piece[][] board, int row) {
24     for (int c = 1; c < board[row].length; c++) {
25         if (board[row][c] != board[row][0]) {
26             return false;
27         }
28     }
29     return true;
30 }
31
32 boolean hasWonColumn(Piece[][] board, int column) {
33     for (int r = 1; r < board.length; r++) {
34         if (board[r][column] != board[0][column]) {
35             return false;
36         }
37     }
38     return true;
39 }
40
41 boolean hasWonDiagonal(Piece[][] board, int direction) {
42     int row = 0;
43     int column = direction == 1 ? 0 : board.length - 1;
44     Piece first = board[0][column];
45     for (int i = 0; i < board.length; i++) {
46         if (board[row][column] != first) {
47             return false;
48         }
49         row += 1;
50         column += direction;
51     }
52     return true;
53 }
```

There is actually a way to clean up this code to remove some of the duplicated code. We'll see this approach in a later function.

Solution #3: 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.

The code below checks each row, column, and diagonal to see if there is a winner.

```
1 Piece hasWon(Piece[][] board) {
2     for (int i = 0; i < board.length; i++) {
3         /* Check Rows */
4         if (hasWinner(board[i][0], board[i][1], board[i][2])) {
5             return board[i][0];
6         }
7
8         /* Check Columns */
9         if (hasWinner(board[0][i], board[1][i], board[2][i])) {
10            return board[0][i];
11        }
12    }
```

```

12 }
13
14 /* Check Diagonal */
15 if (hasWinner(board[0][0], board[1][1], board[2][2])) {
16     return board[0][0];
17 }
18
19 if (hasWinner(board[0][2], board[1][1], board[2][0])) {
20     return board[0][2];
21 }
22
23 return Piece.Empty;
24 }
25
26 boolean hasWinner(Piece p1, Piece p2, Piece p3) {
27     if (p1 == Piece.Empty) {
28         return false;
29     }
30     return p1 == p2 && p2 == p3;
31 }
```

This is an okay solution in that it's relatively easy to understand what is going on. The problem is that the values are hard coded. It's easy to accidentally type the wrong indices.

Additionally, it won't be easy to scale this to an NxN board.

Solution #4: Designing for an NxN board

There are a number of ways to implement this on an NxN board.

Nested For-Loops

The most obvious way is through a series of nested for-loops.

```

1 Piece hasWon(Piece[][] board) {
2     int size = board.length;
3     if (board[0].length != size) return Piece.Empty;
4     Piece first;
5
6     /* Check rows. */
7     for (int i = 0; i < size; i++) {
8         first = board[i][0];
9         if (first == Piece.Empty) continue;
10        for (int j = 1; j < size; j++) {
11            if (board[i][j] != first) {
12                break;
13            } else if (j == size - 1) { // Last element
14                return first;
15            }
16        }
17    }
18
19    /* Check columns. */
20    for (int i = 0; i < size; i++) {
21        first = board[0][i];
22        if (first == Piece.Empty) continue;
23        for (int j = 1; j < size; j++) {
24            if (board[j][i] != first) {
```

```

25         break;
26     } else if (j == size - 1) { // Last element
27         return first;
28     }
29 }
30 }
31 /* Check diagonals. */
32 first = board[0][0];
33 if (first != Piece.Empty) {
34     for (int i = 1; i < size; i++) {
35         if (board[i][i] != first) {
36             break;
37         } else if (i == size - 1) { // Last element
38             return first;
39         }
40     }
41 }
42 }
43
44 first = board[0][size - 1];
45 if (first != Piece.Empty) {
46     for (int i = 1; i < size; i++) {
47         if (board[i][size - i - 1] != first) {
48             break;
49         } else if (i == size - 1) { // Last element
50             return first;
51         }
52     }
53 }
54
55 return Piece.Empty;
56 }

```

This is, to say the least, pretty ugly. We're doing nearly the same work each time. We should look for a way of reusing the code.

Increment and Decrement Function

One way that we can reuse the code better is to just pass in the values to another function that increments/decrements the rows and columns. The hasWon function now just needs the starting position and the amount to increment the row and column by.

```

1  class Check {
2      public int row, column;
3      private int rowIncrement, columnIncrement;
4      public Check(int row, int column, int rowI, int colI) {
5          this.row = row;
6          this.column = column;
7          this.rowIncrement = rowI;
8          this.columnIncrement = colI;
9      }
10
11     public void increment() {
12         row += rowIncrement;
13         column += columnIncrement;
14     }
15
16     public boolean inBounds(int size) {

```

```

17     return row >= 0 && column >= 0 && row < size && column < size;
18 }
19 }
20
21 Piece hasWon(Piece[][] board) {
22     if (board.length != board[0].length) return Piece.Empty;
23     int size = board.length;
24
25     /* Create list of things to check. */
26     ArrayList<Check> instructions = new ArrayList<Check>();
27     for (int i = 0; i < board.length; i++) {
28         instructions.add(new Check(0, i, 1, 0));
29         instructions.add(new Check(i, 0, 0, 1));
30     }
31     instructions.add(new Check(0, 0, 1, 1));
32     instructions.add(new Check(0, size - 1, 1, -1));
33
34     /* Check them. */
35     for (Check instr : instructions) {
36         Piece winner = hasWon(board, instr);
37         if (winner != Piece.Empty) {
38             return winner;
39         }
40     }
41     return Piece.Empty;
42 }
43
44 Piece hasWon(Piece[][] board, Check instr) {
45     Piece first = board[instr.row][instr.column];
46     while (instr.inBounds(board.length)) {
47         if (board[instr.row][instr.column] != first) {
48             return Piece.Empty;
49         }
50         instr.increment();
51     }
52     return first;
53 }

```

The Check function is essentially operating as an iterator.

Iterator

Another way of doing it is, of course, to actually build an iterator.

```

1  Piece hasWon(Piece[][] board) {
2     if (board.length != board[0].length) return Piece.Empty;
3     int size = board.length;
4
5     ArrayList<PositionIterator> instructions = new ArrayList<PositionIterator>();
6     for (int i = 0; i < board.length; i++) {
7         instructions.add(new PositionIterator(new Position(0, i), 1, 0, size));
8         instructions.add(new PositionIterator(new Position(i, 0), 0, 1, size));
9     }
10    instructions.add(new PositionIterator(new Position(0, 0), 1, 1, size));
11    instructions.add(new PositionIterator(new Position(0, size - 1), 1, -1, size));
12
13    for (PositionIterator iterator : instructions) {
14        Piece winner = hasWon(board, iterator);

```

```
15     if (winner != Piece.Empty) {
16         return winner;
17     }
18 }
19 return Piece.Empty;
20 }
21
22 Piece hasWon(Piece[][] board, PositionIterator iterator) {
23     Position firstPosition = iterator.next();
24     Piece first = board[firstPosition.row][firstPosition.column];
25     while (iterator.hasNext()) {
26         Position position = iterator.next();
27         if (board[position.row][position.column] != first) {
28             return Piece.Empty;
29         }
30     }
31     return first;
32 }
33
34 class PositionIterator implements Iterator<Position> {
35     private int rowIncrement, colIncrement, size;
36     private Position current;
37
38     public PositionIterator(Position p, int rowIncrement,
39                             int colIncrement, int size) {
40         this.rowIncrement = rowIncrement;
41         this.colIncrement = colIncrement;
42         this.size = size;
43         current = new Position(p.row - rowIncrement, p.column - colIncrement);
44     }
45
46     @Override
47     public boolean hasNext() {
48         return current.row + rowIncrement < size &&
49                current.column + colIncrement < size;
50     }
51
52     @Override
53     public Position next() {
54         current = new Position(current.row + rowIncrement,
55                               current.column + colIncrement);
56         return current;
57     }
58 }
59
60 public class Position {
61     public int row, column;
62     public Position(int row, int column) {
63         this.row = row;
64         this.column = column;
65     }
66 }
```

All of this is potentially overkill, but it's worth discussing the options with your interviewer. The point of this problem is to assess your understanding of how to code in a clean and maintainable way.

16.5 Factorial Zeros: Write an algorithm which computes the number of trailing zeros in n factorial.

pg 181

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:

$$19! = 2 * \dots * 5 * \dots * 10 * \dots * 15 * 16 * \dots$$

Therefore, to count the number of zeros, 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 to write 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. For example: 5 -> 1,
2   * 25-> 2, etc. */
3  int factorsOf5(int i) {
4      int count = 0;
5      while (i % 5 == 0) {
6          count++;
7          i /= 5;
8      }
9      return count;
10 }
11
12 int countFactZeros(int num) {
13     int count = 0;
14     for (int i = 2; i <= num; i++) {
15         count += factorsOf5(i);
16     }
17     return count;
18 }
```

This isn't bad, but we can make it a little more efficient by directly counting the factors of 5. Using this approach, we would first count the number of multiples of 5 between 1 and n (which is $\frac{n}{5}$), then the number of multiples of 25 ($\frac{n}{25}$), then 125, and so on.

To count how many multiples of m are in n, we can just divide n by m.

```

1  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 brainteaser, 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 it correctly.

16.6 Smallest Difference: Given two arrays of integers, compute the pair of values (one value in each array) with the smallest (non-negative) difference. Return the difference.

EXAMPLE

Input: {1, 3, 15, 11, 2}, {23, 127, 235, 19, 8}

Output: 3. That is, the pair (11, 8).

pg 181

SOLUTION

Let's start first with a brute force solution.

Brute Force

The simple brute force way is to just iterate through all pairs, compute the difference, and compare it to the current minimum difference.

```
1  int findSmallestDifference(int[] array1, int[] array2) {  
2      if (array1.length == 0 || array2.length == 0) return -1;  
3  
4      int min = Integer.MAX_VALUE;  
5      for (int i = 0; i < array1.length; i++) {  
6          for (int j = 0; j < array2.length; j++) {  
7              if (Math.abs(array1[i] - array2[j]) < min) {  
8                  min = Math.abs(array1[i] - array2[j]);  
9              }  
10         }  
11     }  
12     return min;  
13 }
```

One minor optimization we could perform from here is to return immediately if we find a difference of zero, since this is the smallest difference possible. However, depending on the input, this might actually be slower.

This will only be faster if there's a pair with difference zero early in the list of pairs. But to add this optimization, we need to execute an additional line of code each time. There's a tradeoff here; it's faster for some inputs and slower for others. Given that it adds complexity in reading the code, it may be best to leave it out.

With or without this "optimization," the algorithm will take $O(AB)$ time.

Optimal

A more optimal approach is to sort the arrays. Once the arrays are sorted, we can find the minimum difference by iterating through the array.

Consider the following two arrays:

A: {1, 2, 11, 15}
 B: {4, 12, 19, 23, 127, 235}

Try the following approach:

- Suppose a pointer *a* points to the beginning of A and a pointer *b* points to the beginning of B. The current difference between *a* and *b* is 3. Store this as the *min*.
- How can we (potentially) make this difference smaller? Well, the value at *b* is bigger than the value at *a*, so moving *b* will only make the difference larger. Therefore, we want to move *a*.
- Now *a* points to 2 and *b* (still) points to 4. This difference is 2, so we should update *min*. Move *a*, since it is smaller.
- Now *a* points to 11 and *b* points to 4. Move *b*.
- Now *a* points to 11 and *b* points to 12. Update *min* to 1. Move *b*.

And so on.

```

1 int findSmallestDifference(int[] array1, int[] array2) {
2     Arrays.sort(array1);
3     Arrays.sort(array2);
4     int a = 0;
5     int b = 0;
6     int difference = Integer.MAX_VALUE;
7     while (a < array1.length && b < array2.length) {
8         if (Math.abs(array1[a] - array2[b]) < difference) {
9             difference = Math.abs(array1[a] - array2[b]);
10        }
11
12        /* Move smaller value. */
13        if (array1[a] < array2[b]) {
14            a++;
15        } else {
16            b++;
17        }
18    }
19    return difference;
20 }
```

This algorithm takes $O(A \log A + B \log B)$ time to sort and $O(A + B)$ time to find the minimum difference. Therefore, the overall runtime is $O(A \log A + B \log B)$.

16.7 Number Max: Write a method that finds the maximum of two numbers. You should not use if-else or any other comparison operator.

pg 181

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 int flip(int bit) {
```

```

3     return 1^bit;
4 }
5
6 /* Returns 1 if a is positive, and 0 if a is negative */
7 int sign(int a) {
8     return flip((a >> 31) & 0x1);
9 }
10
11 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 $\text{INT_MAX} - 2$ 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 $\text{sign}(a)$.

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  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 >= 0, 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.

- 16.8 English Int:** Given any integer, print an English phrase that describes the integer (e.g., "One Thousand, Two Hundred Thirty Four").

pg 182

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,

```
convert(19,323,984) = convert(19) + " million " + convert(323) + " thousand "
                           + convert(984)
```

The code below implements this algorithm.

```

1  String[] smalls = {"Zero", "One", "Two", "Three", "Four", "Five", "Six", "Seven",
2   "Eight", "Nine", "Ten", "Eleven", "Twelve", "Thirteen", "Fourteen", "Fifteen",
3   "Sixteen", "Seventeen", "Eighteen", "Nineteen"};
4  String[] tens = {"", "", "Twenty", "Thirty", "Forty", "Fifty", "Sixty", "Seventy",
5   "Eighty", "Ninety"};
6  String[] bigs = {"", "Thousand", "Million", "Billion"};
7  String hundred = "Hundred";
8  String negative = "Negative";
9
10 String convert(int num) {
11   if (num == 0) {
12     return smalls[0];
13   } else if (num < 0) {
14     return negative + " " + convert(-1 * num);
15   }
16
17   LinkedList<String> parts = new LinkedList<String>();
18   int chunkCount = 0;
19
20   while (num > 0) {
21     if (num % 1000 != 0) {
22       String chunk = convertChunk(num % 1000) + " " + bigs[chunkCount];
23       parts.addFirst(chunk);
24     }
25     num /= 1000; // shift chunk
26     chunkCount++;
27   }
28
29   return listToString(parts);
30 }
31
32 String convertChunk(int number) {
33   LinkedList<String> parts = new LinkedList<String>();
34
35   /* Convert hundreds place */
36   if (number >= 100) {
37     parts.addLast(smalls[number / 100]);
```

```
38     parts.addLast(hundred);
39     number %= 100;
40 }
41
42 /* Convert tens place */
43 if (number >= 10 && number <= 19) {
44     parts.addLast(smalls[number]);
45 } else if (number >= 20) {
46     parts.addLast(tens[number / 10]);
47     number %= 10;
48 }
49
50 /* Convert ones place */
51 if (number >= 1 && number <= 9) {
52     parts.addLast(smalls[number]);
53 }
54
55 return listToString(parts);
56 }
57 /* Convert a linked list of strings to a string, dividing it up with spaces. */
58 String listToString(LinkedList<String> parts) {
59     StringBuilder sb = new StringBuilder();
60     while (parts.size() > 1) {
61         sb.append(parts.pop());
62         sb.append(" ");
63     }
64     sb.append(parts.pop());
65     return sb.toString();
66 }
```

The key in a problem like this is to make sure you consider all the special cases. There are a lot of them.

- 16.9 Operations:** Write methods to implement the multiply, subtract, and divide operations for integers. The results of all of these are integers. Use only the add operator.

pg 182

SOLUTION

The only operation we have to work with is the add operator. In each of these problems, it's useful to think in depth about what these operations really do or how to phrase them in terms of other operations (either add or operations we've already completed).

Subtraction

How can we phrase subtraction in terms of addition? This one is pretty straightforward. The operation $a - b$ is the same thing as $a + (-1) * b$. However, because we are not allowed to use the $*$ (multiply) operator, we must implement a negate function.

```
1  /* Flip a positive sign to negative or negative sign to pos. */
2  int negate(int a) {
3      int neg = 0;
4      int newSign = a < 0 ? 1 : -1;
5      while (a != 0) {
6          neg += newSign;
7          a += newSign;
8      }
```

```

9     return neg;
10 }
11
12 /* Subtract two numbers by negating b and adding them */
13 int minus(int a, int b) {
14     return a + negate(b);
15 }

```

The negation of the value k is implemented by adding -1 k times. Observe that this will take $O(k)$ time.

If optimizing is something we value here, we can try to get a to zero faster. (For this explanation, we'll assume that a is positive.) To do this, we can first reduce a by 1, then 2, then 4, then 8, and so on. We'll call this value δ . We want a to reach exactly zero. When reducing a by the next δ it would change the sign of a , we reset δ back to 1 and repeat the process.

For example:

a:	29	28	26	22	14	13	11	7	6	4	0
delta:	-1	-2	-4	-8	-1	-2	-4	-1	-2	-4	

The code below implements this algorithm.

```

1 int negate(int a) {
2     int neg = 0;
3     int newSign = a < 0 ? 1 : -1;
4     int delta = newSign;
5     while (a != 0) {
6         boolean differentSigns = (a + delta > 0) != (a > 0);
7         if (a + delta != 0 && differentSigns) { // If delta is too big, reset it.
8             delta = newSign;
9         }
10        neg += delta;
11        a += delta;
12        delta *= 2; // Double the delta
13    }
14    return neg;
15 }

```

Figuring out the runtime here takes a bit of calculation.

Observe that reducing a by half takes $O(\log a)$ work. Why? For each round of “reduce a by half”, the absolute values of a and δ always add up to the same number. The values of δ and a will converge at $\frac{a}{2}$. Since δ is being doubled each time, it will take $O(\log a)$ steps to reach half of a .

We do $O(\log a)$ rounds.

- Reducing a to $\frac{a}{2}$ takes $O(\log a)$ time.
- Reducing $\frac{a}{2}$ to $\frac{a}{4}$ takes $O(\log \frac{a}{2})$ time.
- Reducing $\frac{a}{4}$ to $\frac{a}{8}$ takes $O(\log \frac{a}{4})$ time.

... As so on, for $O(\log a)$ rounds.

The runtime therefore is $O(\log a + \log(\frac{a}{2}) + \log(\frac{a}{4}) + \dots)$, with $O(\log a)$ terms in the expression.

Recall two rules of logs:

- $\log(xy) = \log x + \log y$
- $\log(\frac{x}{y}) = \log x - \log y$.

If we apply this to the above expression, we get:

1. $O(\log a + \log(\frac{a}{2}) + \log(\frac{a}{4}) + \dots)$
2. $O(\log a + (\log a - \log 2) + (\log a - \log 4) + (\log a - \log 8) + \dots)$
3. $O((\log a)^2) // O(\log a)$ terms
4. $O((\log a)^2) - (1 + 2 + 3 + \dots + \log a) // \text{computing the values of logs}$
5. $O((\log a)^2) - \frac{(\log a)(1 + \log a)}{2} // \text{apply equation for sum of 1 through } k$
6. $O((\log a)^2) // \text{drop second term from step 5}$

Therefore, the runtime is $O((\log a)^2)$.

This math is considerably more complicated than most people would be able to do (or expected to do) in an interview. You could make a simplification: You do $O(\log a)$ rounds and the longest round takes $O(\log a)$ work. Therefore, as an upper bound, negate takes $O((\log a)^2)$ time. In this case, the upper bound happens to be the true time.

There are some faster solutions too. For example, rather than resetting `delta` to 1 at each round, we could change `delta` to its previous value. This would have the effect of `delta` “counting up” by multiples of two, and then “counting down” by multiples of two. The runtime of this approach would be $O(\log a)$. However, this implementation would require a stack, division, or bit shifting—any of which might violate the spirit of the problem. You could certainly discuss those implementations with your interviewer though.

Multiplication

The connection between addition and multiplication is equally straightforward. To multiply `a` by `b`, we just add `a` to itself `b` times.

```
1  /* Multiply a by b by adding a to itself b times */
2  int multiply(int a, int b) {
3      if (a < b) {
4          return multiply(b, a); // algorithm is faster if b < a
5      }
6      int sum = 0;
7      for (int i = abs(b); i > 0; i = minus(i, 1)) {
8          sum += a;
9      }
10     if (b < 0) {
11         sum = negate(sum);
12     }
13     return sum;
14 }
15
16 /* Return absolute value */
17 int abs(int a) {
18     if (a < 0) {
19         return negate(a);
20     } else {
21         return a;
22     }
23 }
```

The one thing we need to be careful of in the above code is to properly handle multiplication of negative numbers. If `b` is negative, we need to flip the value of `sum`. So, what this code really does is:

```
multiply(a, b) <- abs(b) * a * (-1 if b < 0).
```

We also implemented a simple `abs` function to help.

Division

Of the three operations, division is certainly the hardest. The good thing is that we can use the `multiply`, `subtract`, and `negate` methods now to implement `divide`.

We are trying to compute x where $X = \frac{a}{b}$. Or, to put this another way, find x where $a = bx$. We've now changed the problem into one that can be stated with something we know how to do: multiplication.

We could implement this by multiplying b by progressively higher values, until we reach a . That would be fairly inefficient, particularly given that our implementation of `multiply` involves a lot of adding.

Alternatively, we can look at the equation $a = xb$ to see that we can compute x by adding b to itself repeatedly until we reach a . The number of times we need to do that will equal x .

Of course, a might not be evenly divisible by b , and that's okay. Integer division, which is what we've been asked to implement, is supposed to truncate the result.

The code below implements this algorithm.

```

1  int divide(int a, int b) throws java.lang.ArithmetricException {
2      if (b == 0) {
3          throw new java.lang.ArithmetricException("ERROR");
4      }
5      int absa = abs(a);
6      int absb = abs(b);
7
8      int product = 0;
9      int x = 0;
10     while (product + absb <= absa) { /* don't go past a */
11         product += absb;
12         x++;
13     }
14
15     if ((a < 0 && b < 0) || (a > 0 && b > 0)) {
16         return x;
17     } else {
18         return negate(x);
19     }
20 }
```

In tackling this problem, you should be aware of the following:

- A logical approach of going back to what exactly multiplication and division do comes in handy. Remember that. All (good) interview problems can be approached in a logical, methodical way!
- The interviewer is looking for this sort of logical work-your-way-through-it approach.
- This is a great problem to demonstrate your ability to write clean code—specifically, to show your ability to reuse code. For example, if you were writing this solution and didn't put `negate` in its own method, you should move it into its own method once you see that you'll use it multiple times.
- Be careful about making assumptions while coding. Don't assume that the numbers are all positive or that a is bigger than b .

16.10 Living People: Given a list of people with their birth and death years, implement a method to compute the year with the most number of people alive. You may assume that all people were born between 1900 and 2000 (inclusive). If a person was alive during any portion of that year, they should be included in that year's count. For example, Person (birth = 1908, death = 1909) is included in the counts for both 1908 and 1909.

pg 182

SOLUTION

The first thing we should do is outline what this solution will look like. The interview question hasn't specified the exact form of input. In a real interview, we could ask the interviewer how the input is structured. Alternatively, you can explicitly state your (reasonable) assumptions.

Here, we'll need to make our own assumptions. We will assume that we have an array of simple Person objects:

```
1  public class Person {  
2      public int birth;  
3      public int death;  
4      public Person(int birthYear, int deathYear) {  
5          birth = birthYear;  
6          death = deathYear;  
7      }  
8  }
```

We could have also given Person a `getBirthYear()` and `getDeathYear()` objects. Some would argue that's better style, but for compactness and clarity, we'll just keep the variables public.

The important thing here is to actually use a Person object. This shows better style than, say, having an integer array for birth years and an integer array for death years (with an implicit association of `births[i]` and `deaths[i]` being associated with the same person). You don't get a lot of chances to demonstrate great coding style, so it's valuable to take the ones you get.

With that in mind, let's start with a brute force algorithm.

Brute Force

The brute force algorithm falls directly out from the wording of the problem. We need to find the year with the most number of people alive. Therefore, we go through each year and check how many people are alive in that year.

```
1  int maxAliveYear(Person[] people, int min, int max) {  
2      int maxAlive = 0;  
3      int maxAliveYear = min;  
4  
5      for (int year = min; year <= max; year++) {  
6          int alive = 0;  
7          for (Person person : people) {  
8              if (person.birth <= year && year <= person.death) {  
9                  alive++;  
10             }  
11         }  
12         if (alive > maxAlive) {  
13             maxAlive = alive;  
14             maxAliveYear = year;  
15         }  
16     }
```

```

17
18     return maxAliveYear;
19 }

```

Note that we have passed in the values for the min year (1900) and max year (2000). We shouldn't hard code these values.

The runtime of this is $O(RP)$, where R is the range of years (100 in this case) and P is the number of people.

Slightly Better Brute Force

A slightly better way of doing this is to create an array where we track the number of people born in each year. Then, we iterate through the list of people and increment the array for each year they are alive.

```

1  int maxAliveYear(Person[] people, int min, int max) {
2     int[] years = createYearMap(people, min, max);
3     int best = getMaxIndex(years);
4     return best + min;
5 }
6
7 /* Add each person's years to a year map. */
8 int[] createYearMap(Person[] people, int min, int max) {
9     int[] years = new int[max - min + 1];
10    for (Person person : people) {
11        incrementRange(years, person.birth - min, person.death - min);
12    }
13    return years;
14 }
15
16 /* Increment array for each value between left and right. */
17 void incrementRange(int[] values, int left, int right) {
18    for (int i = left; i <= right; i++) {
19        values[i]++;
20    }
21 }
22
23 /* Get index of largest element in array. */
24 int getMaxIndex(int[] values) {
25    int max = 0;
26    for (int i = 1; i < values.length; i++) {
27        if (values[i] > values[max]) {
28            max = i;
29        }
30    }
31    return max;
32 }

```

Be careful on the size of the array in line 9. If the range of years is 1900 to 2000 inclusive, then that's 101 years, not 100. That is why the array has size `max - min + 1`.

Let's think about the runtime by breaking this into parts.

- We create an R-sized array, where R is the min and max years.
- Then, for P people, we iterate through the years (Y) that the person is alive.
- Then, we iterate through the R-sized array again.

The total runtime is $O(PY + R)$. In the worst case, Y is R and we have done no better than we did in the first algorithm.

More Optimal

Let's create an example. (In fact, an example is really helpful in almost all problems. Ideally, you've already done this.) Each column below is matched, so that the items correspond to the same person. For compactness, we'll just write the last two digits of the year.

birth:	12	20	10	01	10	23	13	90	83	75
death:	15	90	98	72	98	82	98	98	99	94

It's worth noting that it doesn't really matter whether these years are matched up. Every birth adds a person and every death removes a person.

Since we don't actually need to match up the births and deaths, let's sort both. A sorted version of the years might help us solve the problem.

birth:	01	10	10	12	13	20	23	75	83	90
death:	15	72	82	90	94	98	98	98	99	99

We can try walking through the years.

- At year 0, no one is alive.
- At year 1, we see one birth.
- At years 2 through 9, nothing happens.
- Let's skip ahead until year 10, when we have two births. We now have three people alive.
- At year 15, one person dies. We are now down to two people alive.
- And so on.

If we walk through the two arrays like this, we can track the number of people alive at each point.

```
1  int maxAliveYear(Person[] people, int min, int max) {  
2      int[] births = getSortedYears(people, true);  
3      int[] deaths = getSortedYears(people, false);  
4  
5      int birthIndex = 0;  
6      int deathIndex = 0;  
7      int currentlyAlive = 0;  
8      int maxAlive = 0;  
9      int maxAliveYear = min;  
10  
11     /* Walk through arrays. */  
12     while (birthIndex < births.length) {  
13         if (births[birthIndex] <= deaths[deathIndex]) {  
14             currentlyAlive++; // include birth  
15             if (currentlyAlive > maxAlive) {  
16                 maxAlive = currentlyAlive;  
17                 maxAliveYear = births[birthIndex];  
18             }  
19             birthIndex++; // move birth index  
20         } else if (births[birthIndex] > deaths[deathIndex]) {  
21             currentlyAlive--; // include death  
22             deathIndex++; // move death index  
23         }  
24     }  
25  
26     return maxAliveYear;  
27 }  
28  
29 /* Copy birth years or death years (depending on the value of copyBirthYear into
```

```

30 * integer array, then sort array. */
31 int[] getSortedYears(Person[] people, boolean copyBirthYear) {
32     int[] years = new int[people.length];
33     for (int i = 0; i < people.length; i++) {
34         years[i] = copyBirthYear ? people[i].birth : people[i].death;
35     }
36     Arrays.sort(years);
37     return years;
38 }

```

There are some very easy things to mess up here.

On line 13, we need to think carefully about whether this should be a less than ($<$) or a less than or equals (\leq). The scenario we need to worry about is that you see a birth and death in the same year. (It doesn't matter whether the birth and death is from the same person.)

When we see a birth and death from the same year, we want to include the birth *before* we include the death, so that we count this person as alive for that year. That is why we use a \leq on line 13.

We also need to be careful about where we put the updating of `maxAlive` and `maxAliveYear`. It needs to be after the `currentAlive++`, so that it takes into account the updated total. But it needs to be before `birthIndex++`, or we won't have the right year.

This algorithm will take $O(P \log P)$ time, where P is the number of people.

More Optimal (Maybe)

Can we optimize this further? To optimize this, we'd need to get rid of the sorting step. We're back to dealing with unsorted values:

birth: 12	20	10	01	10	23	13	90	83	75
death: 15	90	98	72	98	82	98	98	99	94

Earlier, we had logic that said that a birth is just adding a person and a death is just subtracting a person. Therefore, let's represent the data using the logic:

01: +1	10: +1	10: +1	12: +1	13: +1
15: -1	20: +1	23: +1	72: -1	75: +1
82: -1	83: +1	90: +1	90: -1	94: -1
98: -1	98: -1	98: -1	98: -1	99: -1

We can create an array of the years, where the value at `array[year]` indicates how the population changed in that year. To create this array, we walk through the list of people and increment when they're born and decrement when they die.

Once we have this array, we can walk through each of the years, tracking the current population as we go (adding the value at `array[year]` each time).

This logic is reasonably good, but we should think about it more. Does it really work?

One edge case we should consider is when a person dies the same year that they're born. The increment and decrement operations will cancel out to give 0 population change. According to the wording of the problem, this person should be counted as living in that year.

In fact, the "bug" in our algorithm is broader than that. This same issue applies to all people. People who die in 1908 shouldn't be removed from the population count until 1909.

There's a simple fix: instead of decrementing `array[deathYear]`, we should decrement `array[deathYear + 1]`.

```
1 int maxAliveYear(Person[] people, int min, int max) {
```

```
2  /* Build population delta array. */
3  int[] populationDeltas = getPopulationDeltas(people, min, max);
4  int maxAliveYear = getMaxAliveYear(populationDeltas);
5  return maxAliveYear + min;
6 }
7
8 /* Add birth and death years to deltas array.*/
9 int[] getPopulationDeltas(Person[] people, int min, int max) {
10    int[] populationDeltas = new int[max - min + 2];
11    for (Person person : people) {
12        int birth = person.birth - min;
13        populationDeltas[birth]++;
14
15        int death = person.death - min;
16        populationDeltas[death + 1]--;
17    }
18    return populationDeltas;
19 }
20
21 /* Compute running sums and return index with max.*/
22 int getMaxAliveYear(int[] deltas) {
23    int maxAliveYear = 0;
24    int maxAlive = 0;
25    int currentlyAlive = 0;
26    for (int year = 0; year < deltas.length; year++) {
27        currentlyAlive += deltas[year];
28        if (currentlyAlive > maxAlive) {
29            maxAliveYear = year;
30            maxAlive = currentlyAlive;
31        }
32    }
33
34    return maxAliveYear;
35 }
```

This algorithm takes $O(R + P)$ time, where R is the range of years and P is the number of people. Although $O(R + P)$ might be faster than $O(P \log P)$ for many expected inputs, you cannot directly compare the speeds to say that one is faster than the other.

16.11 Diving Board: You are building a diving board by placing a bunch of planks of wood end-to-end. There are two types of planks, one of length `shorter` and one of length `longer`. You must use exactly K planks of wood. Write a method to generate all possible lengths for the diving board.

pg 182

SOLUTION

One way to approach this is to think about the choices we make as we're building a diving board. This leads us to a recursive algorithm.

Recursive Solution

For a recursive solution, we can imagine ourselves building a diving board. We make K decisions, each time choosing which plank we will put on next. Once we've put on K planks, we have a complete diving board and we can add this to the list (assuming we haven't seen this length before).

We can follow this logic to write recursive code. Note that we don't need to track the sequence of planks. All we need to know is the current length and the number of planks remaining.

```

1  HashSet<Integer> allLengths(int k, int shorter, int longer) {
2      HashSet<Integer> lengths = new HashSet<Integer>();
3      getAllLengths(k, 0, shorter, longer, lengths);
4      return lengths;
5  }
6
7  void getAllLengths(int k, int total, int shorter, int longer,
8                      HashSet<Integer> lengths) {
9      if (k == 0) {
10          lengths.add(total);
11          return;
12      }
13      getAllLengths(k - 1, total + shorter, shorter, longer, lengths);
14      getAllLengths(k - 1, total + longer, shorter, longer, lengths);
15  }

```

We've added each length to a hash set. This will automatically prevent adding duplicates.

This algorithm takes $O(2^k)$ time, since there are two choices at each recursive call and we recurse to a depth of K .

Memoization Solution

As in many recursive algorithms (especially those with exponential runtimes), we can optimize this through memorization (a form of dynamic programming).

Observe that some of the recursive calls will be essentially equivalent. For example, picking plank 1 and then plank 2 is equivalent to picking plank 2 and then plank 1.

Therefore, if we've seen this (`total`, `plank count`) pair before then we stop this recursive path. We can do this using a `HashSet` with a key of (`total`, `plank count`).

Many candidates will make a mistake here. Rather than stopping only when they've seen (`total`, `plank count`), they'll stop whenever they've seen just `total` before. This is incorrect. Seeing two planks of length 1 is not the same thing as one plank of length 2, because there are different numbers of planks remaining. In memoization problems, be very careful about what you choose for your key.

The code for this approach is very similar to the earlier approach.

```

1  HashSet<Integer> allLengths(int k, int shorter, int longer) {
2      HashSet<Integer> lengths = new HashSet<Integer>();
3      HashSet<String> visited = new HashSet<String>();
4      getAllLengths(k, 0, shorter, longer, lengths, visited);
5      return lengths;
6  }
7
8  void getAllLengths(int k, int total, int shorter, int longer,
9                      HashSet<Integer> lengths, HashSet<String> visited) {
10     if (k == 0) {
11         lengths.add(total);
12         return;
13     }
14     String key = k + " " + total;

```

```
15     if (visited.contains(key)) {
16         return;
17     }
18     getAllLengths(k - 1, total + shorter, shorter, longer, lengths, visited);
19     getAllLengths(k - 1, total + longer, shorter, longer, lengths, visited);
20     visited.add(key);
21 }
```

For simplicity, we've set the key to be a string representation of `total` and the current plank count. Some people may argue it's better to use a data structure to represent this pair. There are benefits to this, but there are drawbacks as well. It's worth discussing this tradeoff with your interviewer.

The runtime of this algorithm is a bit tricky to figure out.

One way we can think about the runtime is by understanding that we're basically filling in a table of SUMS \times PLANK COUNTS. The biggest possible sum is $K^2 \times \text{LONGER}$ and the biggest possible plank count is K . Therefore, the runtime will be no worse than $O(K^2 \times \text{LONGER})$.

Of course, a bunch of those sums will never actually be reached. How many unique sums can we get? Observe that any path with the same number of each type of planks will have the same sum. Since we can have at most K planks of each type, there are only K different sums we can make. Therefore, the table is really $K \times K$, and the runtime is $O(K^2)$.

Optimal Solution

If you re-read the prior paragraph, you might notice something interesting. There are only K distinct sums we can get. Isn't that the whole point of the problem—to find all possible sums?

We don't actually need to go through all arrangements of planks. We just need to go through all unique sets of K planks (sets, not orders!). There are only K ways of picking K planks if we only have two possible types: {0 of type A, K of type B}, {1 of type A, $K-1$ of type B}, {2 of type A, $K-2$ of type B}, ...

This can be done in just a simple for loop. At each "sequence", we just compute the sum.

```
1  HashSet<Integer> allLengths(int k, int shorter, int longer) {
2      HashSet<Integer> lengths = new HashSet<Integer>();
3      for (int nShorter = 0; nShorter <= k; nShorter++) {
4          int nLonger = k - nShorter;
5          int length = nShorter * shorter + nLonger * longer;
6          lengths.add(length);
7      }
8      return lengths;
9  }
```

We've used a `HashSet` here for consistency with the prior solutions. This isn't really necessary though, since we shouldn't get any duplicates. We could instead use an `ArrayList`. If we do this, though, we just need to handle an edge case where the two types of planks are the same length. In this case, we would just return an `ArrayList` of size 1.

16.12 XML Encoding: 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
  
```

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 182

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 void encode(Element root, StringBuilder 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 void encode(String v, StringBuilder sb) {
18     sb.append(v);
19     sb.append(" ");
20 }
21
22 void encode(Attribute attr, StringBuilder sb) {
23     encode(attr.getTagCode(), sb);
24     encode(attr.value, sb);
25 }
26
  
```

```
27 String encodeToString(Element root) {  
28     StringBuilder sb = new StringBuilder();  
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.

16.13 Bisect Squares: Given two squares on a two-dimensional plane, find a line that would cut these two squares in half. Assume that the top and the bottom sides of the square run parallel to the x-axis.

pg 182

SOLUTION

Before we start, we should think about what exactly this problem means by a “line.” Is a line defined by a slope and a y-intercept? Or by any two points on the line? Or, should the line be really a line segment, which starts and ends at the edges of the squares?

We will assume, since it makes the problem a bit more interesting, that we mean the third option: that the line should end at the edges of the squares. In an interview situation, you should discuss this with your interviewer.

This line that cuts two squares in half must connect the two middles. We can easily calculate the slope, knowing that $\text{slope} = \frac{y_1 - y_2}{x_1 - x_2}$. Once we calculate the slope using the two middles, we can use the same equation to calculate the start and end points of the line segment.

In the below code, we will assume the origin $(0, 0)$ is in the upper left-hand corner.

```
1  public class Square {  
2      ...  
3      public Point middle() {  
4          return new Point((this.left + this.right) / 2.0,  
5                             (this.top + this.bottom) / 2.0);  
6      }  
7  
8      /* Return the point where the line segment connecting mid1 and mid2 intercepts  
9       * the edge of square 1. That is, draw a line from mid2 to mid1, and continue it  
10      * out until the edge of the square. */  
11      public Point extend(Point mid1, Point mid2, double size) {  
12          /* Find what direction the line mid2 -> mid1 goes. */  
13          double xdir = mid1.x < mid2.x ? -1 : 1;  
14          double ydir = mid1.y < mid2.y ? -1 : 1;  
15  
16          /* If mid1 and mid2 have the same x value, then the slope calculation will  
17             * throw a divide by 0 exception. So, we compute this specially. */  
18          if (mid1.x == mid2.x) {  
19              return new Point(mid1.x, mid1.y + ydir * size / 2.0);  
20          }  
21  
22          double slope = (mid1.y - mid2.y) / (mid1.x - mid2.x);  
23          double x1 = 0;  
24          double y1 = 0;  
25      }
```

```

26     /* Calculate slope using the equation (y1 - y2) / (x1 - x2).
27      * Note: if the slope is "steep" (>1) then the end of the line segment will
28      * hit size / 2 units away from the middle on the y axis. If the slope is
29      * "shallow" (<1) the end of the line segment will hit size / 2 units away
30      * from the middle on the x axis. */
31     if (Math.abs(slope) == 1) {
32         x1 = mid1.x + xdir * size / 2.0;
33         y1 = mid1.y + ydir * size / 2.0;
34     } else if (Math.abs(slope) < 1) { // shallow slope
35         x1 = mid1.x + xdir * size / 2.0;
36         y1 = slope * (x1 - mid1.x) + mid1.y;
37     } else { // steep slope
38         y1 = mid1.y + ydir * size / 2.0;
39         x1 = (y1 - mid1.y) / slope + mid1.x;
40     }
41     return new Point(x1, y1);
42 }
43
44 public Line cut(Square other) {
45     /* Calculate where a line between each middle would collide with the edges of
46      * the squares */
47     Point p1 = extend(this.middle(), other.middle(), this.size);
48     Point p2 = extend(this.middle(), other.middle(), -1 * this.size);
49     Point p3 = extend(other.middle(), this.middle(), other.size);
50     Point p4 = extend(other.middle(), this.middle(), -1 * other.size);
51
52     /* Of above points, find start and end of lines. Start is farthest left (with
53      * top most as a tie breaker) and end is farthest right (with bottom most as
54      * a tie breaker. */
55     Point start = p1;
56     Point end = p1;
57     Point[] points = {p2, p3, p4};
58     for (int i = 0; i < points.length; i++) {
59         if (points[i].x < start.x ||
60             (points[i].x == start.x && points[i].y < start.y)) {
61             start = points[i];
62         } else if (points[i].x > end.x ||
63             (points[i].x == end.x && points[i].y > end.y)) {
64             end = points[i];
65         }
66     }
67
68     return new Line(start, end);
69 }

```

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.

16.14 Best Line: Given a two-dimensional graph with points on it, find a line which passes the most number of points.

pg 183

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 hash table, 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 hash table 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 hash table? 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 hash table at three spots: flooredSlope, flooredSlope - epsilon, and flooredSlope + epsilon. This will ensure that we’ve checked out all lines that might be equal.

```
1 /* Find line that goes through most number of points. */
2 Line findBestLine(GraphPoint[] points) {
3     HashMapList<Double, Line> linesBySlope = getListOfLines(points);
4     return getBestLine(linesBySlope);
5 }
6
7 /* Add each pair of points as a line to the list. */
8 HashMapList<Double, Line> getListOfLines(GraphPoint[] points) {
9     HashMapList<Double, Line> linesBySlope = new HashMapList<Double, Line>();
10    for (int i = 0; i < points.length; i++) {
11        for (int j = i + 1; j < points.length; j++) {
12            Line line = new Line(points[i], points[j]);
13            double key = Line.floorToNearestEpsilon(line.slope);
14            linesBySlope.put(key, line);
15        }
16    }
17    return linesBySlope;
18 }
19
20 /* Return the line with the most equivalent other lines. */
21 Line getBestLine(HashMapList<Double, Line> linesBySlope) {
22     Line bestLine = null;
23     int bestCount = 0;
24
25     Set<Double> slopes = linesBySlope.keySet();
26     for (double slope : slopes) {
```

```

28     ArrayList<Line> lines = linesBySlope.get(slope);
29     for (Line line : lines) {
30         /* count lines that are equivalent to current line */
31         int count = countEquivalentLines(linesBySlope, line);
32
33         /* if better than current line, replace it */
34         if (count > bestCount) {
35             bestLine = line;
36             bestCount = count;
37             bestLine.Print();
38             System.out.println(bestCount);
39         }
40     }
41 }
42 return bestLine;
43 }
44
45 /* Check hashmap for lines that are equivalent. Note that we need to check one
46 * epsilon above and below the actual slope since we're defining two lines as
47 * equivalent if they're within an epsilon of each other. */
48 int countEquivalentLines(HashMapList<Double, Line> linesBySlope, Line line) {
49     double key = Line.floorToNearestEpsilon(line.slope);
50     int count = countEquivalentLines(linesBySlope.get(key), line);
51     count += countEquivalentLines(linesBySlope.get(key - Line.epsilon), line);
52     count += countEquivalentLines(linesBySlope.get(key + Line.epsilon), line);
53     return count;
54 }
55
56 /* Count lines within an array of lines which are "equivalent" (slope and
57 * y-intercept are within an epsilon value) to a given line */
58 int countEquivalentLines(ArrayList<Line> lines, Line line) {
59     if (lines == null) return 0;
60
61     int count = 0;
62     for (Line parallelLine : lines) {
63         if (parallelLine.isEquivalent(line)) {
64             count++;
65         }
66     }
67     return count;
68 }
69
70 public class Line {
71     public static double epsilon = .0001;
72     public double slope, intercept;
73     private boolean infinite_slope = false;
74
75     public Line(GraphPoint p, GraphPoint q) {
76         if (Math.abs(p.x - q.x) > epsilon) { // if x's are different
77             slope = (p.y - q.y) / (p.x - q.x); // compute slope
78             intercept = p.y - slope * p.x; // y intercept from y=mx+b
79         } else {
80             infinite_slope = true;
81             intercept = p.x; // x-intercept, since slope is infinite
82         }
83     }

```

```
84
85     public static double floorToNearestEpsilon(double d) {
86         int r = (int) (d / epsilon);
87         return ((double) r) * epsilon;
88     }
89
90     public boolean isEquivalent(double a, double b) {
91         return (Math.abs(a - b) < epsilon);
92     }
93
94     public boolean isEquivalent(Object o) {
95         Line l = (Line) o;
96         if (isEquivalent(l.slope, slope) && isEquivalent(l.intercept, intercept) &&
97             (infinite_slope == l.infinite_slope)) {
98             return true;
99         }
100        return false;
101    }
102 }
103
104 /* HashMapList<String, Integer> is a HashMap that maps from Strings to
105 * ArrayList<Integer>. See appendix for implementation. */
```

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.

16.15 Master Mind:

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 183

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 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 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 int MAX_COLORS = 4;
26
27 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 for pseudo-hits)
39              * if it's not a hit. If it's a hit, the slot has already been "used." */
40             int code = code(solution.charAt(i));
41             frequencies[code]++;
42         }
43     }
44
45     /* Compute pseudo-hits */
46     for (int i = 0; i < guess.length(); i++) {
47         int code = code(guess.charAt(i));
48         if (code >= 0 && frequencies[code] > 0 &&
49             guess.charAt(i) != solution.charAt(i)) {
50             res.pseudoHits++;
51             frequencies[code]--;
52         }
53     }
54     return res;
55 }
```

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.

16.16 Sub Sort: 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).

EXAMPLE

Input: 1, 2, 4, 7, 10, 11, 7, 12, 6, 7, 16, 18, 19

Output: (3, 9)

pg 183

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, 8, 12, 5, 6, 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: 8, 12  
right: 5, 6, 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. We need the left part to be smaller than all the elements in the middle and right side, and the right part to be bigger than all the elements on the left and right side.

Let `min` equal `min(middle and right side)` and `max` equal `max(middle and left side)`. Observe that since the right and left sides are already in sorted order, we only actually need to check their start or end point.

On the left side, we start with the end of the subsequence (value 11, at element 5) and move to the left. The value `min` equals 5. 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 we 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 void findUnsortedSequence(int[] array) {
2     // find left subsequence
3     int end_left = findEndOfLeftSubsequence(array);
4     if (end_left >= array.length - 1) return; // Already sorted
5
6     // find right subsequence
7     int start_right = findStartOfRightSubsequence(array);
8
9     // get min and max
10    int max_index = end_left; // max of left side
11    int min_index = start_right; // min of right side
12    for (int i = end_left + 1; i < start_right; i++) {
13        if (array[i] < array[min_index]) min_index = i;
14        if (array[i] > array[max_index]) max_index = i;
15    }
16
17    // slide left until less than array[min_index]
18    int left_index = shrinkLeft(array, min_index, end_left);
19
20    // slide right until greater than array[max_index]
21    int right_index = shrinkRight(array, max_index, start_right);
22
23    System.out.println(left_index + " " + right_index);
24 }
25
26 int findEndOfLeftSubsequence(int[] array) {
27     for (int i = 1; i < array.length; i++) {
28         if (array[i] < array[i - 1]) return i - 1;
29     }
30     return array.length - 1;
31 }
32
33 int findStartOfRightSubsequence(int[] array) {
34     for (int i = array.length - 2; i >= 0; i--) {
35         if (array[i] > array[i + 1]) return i + 1;
36     }
37     return 0;
38 }
39
40 int shrinkLeft(int[] array, int min_index, int start) {
41     int comp = array[min_index];
42     for (int i = start - 1; i >= 0; i--) {
43         if (array[i] <= comp) return i + 1;
44     }
45     return 0;
46 }
47
48 int shrinkRight(int[] array, int max_index, int start) {
49     int comp = array[max_index];
50     for (int i = start; i < array.length; i++) {

```

```
51     if (array[i] >= comp) return i - 1;
52 }
53 return array.length - 1;
54 }
```

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.

16.17 Contiguous Sequence: You are given an array of integers (both positive and negative). Find the contiguous sequence with the largest sum. Return the sum.

EXAMPLE

Input: 2, -8, 3, -2, 4, -10

Output: 5 (i.e., {3, -2, 4})

pg 183

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 fully expanded array, our logic is identical. The code below implements this algorithm.

```

1  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.

16.18 Pattern Matching: You are given two strings, `pattern` and `value`. The pattern string consists of just the letters a and b, describing a pattern within a string. For example, the string `catcatgocatgo` matches the pattern `aabab` (where `cat` is a and `go` is b). It also matches patterns like `a`, `ab`, and `b`. Write a method to determine if `value` matches `pattern`.

pg 183

SOLUTION

As always, we can start with a simple brute force approach.

Brute Force

A brute force algorithm is to just try all possible values for a and b and then check if this works.

We could do this by iterating through all substrings for a and all possible substrings for b. There are $O(n^2)$ substrings in a string of length n, so this will actually take $O(n^4)$ time. But then, for each value of a and b, we need to build the new string of this length and compare it for equality. This building/comparison step takes $O(n)$ time, giving an overall runtime of $O(n^5)$.

```

1  for each possible substring a
2      for each possible substring b
3          candidate = buildFromPattern(pattern, a, b)
4          if candidate equals value
5              return true
```

Ouch.

One easy optimization is to notice that if the pattern starts with 'a', then the string must start at the beginning of value. (Otherwise, the b string must start at the beginning of value.) Therefore, there aren't $O(n^2)$ possible values for a; there are $O(n)$.

The algorithm then is to check if the pattern starts with a or b. If it starts with b, we can "invert" it (flipping each 'a' to a 'b' and each 'b' to an 'a') so that it starts with 'a'. Then, iterate through all possible substrings for a (each of which must begin at index 0) and all possible substrings for b (each of which must begin at some character after the end of a). As before, we then compare the string for this pattern with the original string.

This algorithm now takes $O(n^4)$ time.

There's one more minor (optional) optimization we can make. We don't actually need to do this "inversion" if the string starts with 'b' instead of 'a'. The `buildFromPattern` method can take care of this. We can think about the first character in the pattern as the "main" item and the other character as the alternate character. The `buildFromPattern` method can build the appropriate string based on whether 'a' is the main character or alternate character.

```
1  boolean doesMatch(String pattern, String value) {
2      if (pattern.length() == 0) return value.length() == 0;
3
4      int size = value.length();
5      for (int mainSize = 0; mainSize < size; mainSize++) {
6          String main = value.substring(0, mainSize);
7          for (int altStart = mainSize; altStart <= size; altStart++) {
8              for (int altEnd = altStart; altEnd <= size; altEnd++) {
9                  String alt = value.substring(altStart, altEnd);
10                 String cand = buildFromPattern(pattern, main, alt);
11                 if (cand.equals(value)) {
12                     return true;
13                 }
14             }
15         }
16     }
17     return false;
18 }
19
20 String buildFromPattern(String pattern, String main, String alt) {
21     StringBuffer sb = new StringBuffer();
22     char first = pattern.charAt(0);
23     for (char c : pattern.toCharArray()) {
24         if (c == first) {
25             sb.append(main);
26         } else {
27             sb.append(alt);
28         }
29     }
30     return sb.toString();
31 }
```

We should look for a more optimal algorithm.

Optimized

Let's think through our current algorithm. Searching through all values for the main string is fairly fast (it takes $O(n)$ time). It's the alternate string that is so slow: $O(n^2)$ time. We should study how to optimize that.

Suppose we have a pattern like aabab and we're comparing it to the string catcatgocatgo. Once we've picked "cat" as the value for a to try, then the a strings are going to take up nine characters (three a strings with length three each). Therefore, the b strings must take up the remaining four characters, with each having length two. Moreover, we actually know exactly where they must occur, too. If a is cat, and the pattern is aabab, then b must be go.

In other words, once we've picked a, we've picked b too. There's no need to iterate. Gathering some basic stats on pattern (number of as, number of bs, first occurrence of each) and iterating through values for a (or whichever the main string is) will be sufficient.

```

1  boolean doesMatch(String pattern, String value) {
2      if (pattern.length() == 0) return value.length() == 0;
3
4      char mainChar = pattern.charAt(0);
5      char altChar = mainChar == 'a' ? 'b' : 'a';
6      int size = value.length();
7
8      int countOfMain = countOf(pattern, mainChar);
9      int countOfAlt = pattern.length() - countOfMain;
10     int firstAlt = pattern.indexOf(altChar);
11     int maxMainSize = size / countOfMain;
12
13    for (int mainSize = 0; mainSize <= maxMainSize; mainSize++) {
14        int remainingLength = size - mainSize * countOfMain;
15        String first = value.substring(0, mainSize);
16        if (countOfAlt == 0 || remainingLength % countOfAlt == 0) {
17            int altIndex = firstAlt * mainSize;
18            int altSize = countOfAlt == 0 ? 0 : remainingLength / countOfAlt;
19            String second = countOfAlt == 0 ? "" :
20                            value.substring(altIndex, altSize + altIndex);
21
22            String cand = buildFromPattern(pattern, first, second);
23            if (cand.equals(value)) {
24                return true;
25            }
26        }
27    }
28    return false;
29 }
30
31 int countOf(String pattern, char c) {
32     int count = 0;
33     for (int i = 0; i < pattern.length(); i++) {
34         if (pattern.charAt(i) == c) {
35             count++;
36         }
37     }
38     return count;
39 }
40
41 String buildFromPattern(...) { /* same as before */ }
```

This algorithm takes $O(n^2)$, since we iterate through $O(n)$ possibilities for the main string and do $O(n)$ work to build and compare the strings.

Observe that we've also cut down the possibilities for the main string that we try. If there are three instances of the main string, then its length cannot be any more than one third of value.

Optimized (Alternate)

If you don't like the work of building a string only to compare it (and then destroy it), we can eliminate this.

Instead, we can iterate through the values for a and b as before. But this time, to check if the string matches the pattern (given those values for a and b), we walk through value, comparing each substring to the first instance of the a and b strings.

```

1  boolean doesMatch(String pattern, String value) {
2      if (pattern.length() == 0) return value.length() == 0;
3
4      char mainChar = pattern.charAt(0);
5      char altChar = mainChar == 'a' ? 'b' : 'a';
6      int size = value.length();
7
8      int countOfMain = countOf(pattern, mainChar);
9      int countOfAlt = pattern.length() - countOfMain;
10     int firstAlt = pattern.indexOf(altChar);
11     int maxMainSize = size / countOfMain;
12
13    for (int mainSize = 0; mainSize <= maxMainSize; mainSize++) {
14        int remainingLength = size - mainSize * countOfMain;
15        if (countOfAlt == 0 || remainingLength % countOfAlt == 0) {
16            int altIndex = firstAlt * mainSize;
17            int altSize = countOfAlt == 0 ? 0 : remainingLength / countOfAlt;
18            if (matches(pattern, value, mainSize, altSize, altIndex)) {
19                return true;
20            }
21        }
22    }
23    return false;
24 }
25
26 /* Iterates through pattern and value. At each character within pattern, checks if
27 * this is the main string or the alternate string. Then checks if the next set of
28 * characters in value match the original set of those characters (either the main
29 * or the alternate. */
30 boolean matches(String pattern, String value, int mainSize, int altSize,
31                 int firstAlt) {
32     int stringIndex = mainSize;
33     for (int i = 1; i < pattern.length(); i++) {
34         int size = pattern.charAt(i) == pattern.charAt(0) ? mainSize : altSize;
35         int offset = pattern.charAt(i) == pattern.charAt(0) ? 0 : firstAlt;
36         if (!isEqual(value, offset, stringIndex, size)) {
37             return false;
38         }
39         stringIndex += size;
40     }
41     return true;
42 }
43
44 /* Checks if two substrings are equal, starting at given offsets and continuing to
45 * size. */
46 boolean isEqual(String s1, int offset1, int offset2, int size) {
47     for (int i = 0; i < size; i++) {
48         if (s1.charAt(offset1 + i) != s1.charAt(offset2 + i)) {
49             return false;

```

```

50     }
51 }
52 return true;
53 }

```

This algorithm will still take $O(n^2)$ time, but the benefit is that it can short circuit when matches fail early (which they usually will). The previous algorithm must go through all the work to build the string before it can learn that it has failed.

- 16.19 Pond Sizes:** You have an integer matrix representing a plot of land, where the value at that location represents the height above sea level. A value of zero indicates water. A pond is a region of water connected vertically, horizontally, or diagonally. The size of the pond is the total number of connected water cells. Write a method to compute the sizes of all ponds in the matrix.

EXAMPLE

Input:

```

0 2 1 0
0 1 0 1
1 1 0 1
0 1 0 1

```

Output: 2, 4, 1 (in any order)

pg 184

SOLUTION

The first thing we can try is just walking through the array. It's easy enough to find water: when it's a zero, that's water.

Given a water cell, how can we compute the amount of water nearby? If the cell is not adjacent to any zero cells, then the size of this pond is 1. If it is, then we need to add in the adjacent cells, plus any water cells adjacent to those cells. We need to, of course, be careful to not recount any cells. We can do this with a modified breadth-first or depth-first search. Once we visit a cell, we permanently mark it as visited.

For each cell, we need to check eight adjacent cells. We could do this by writing in lines to check up, down, left, right, and each of the four diagonal cells. It's even easier, though, to do this with a loop.

```

1  ArrayList<Integer> computePondSizes(int[][] land) {
2      ArrayList<Integer> pondSizes = new ArrayList<Integer>();
3      for (int r = 0; r < land.length; r++) {
4          for (int c = 0; c < land[r].length; c++) {
5              if (land[r][c] == 0) { // Optional. Would return anyway.
6                  int size = computeSize(land, r, c);
7                  pondSizes.add(size);
8              }
9          }
10     }
11     return pondSizes;
12 }
13
14 int computeSize(int[][] land, int row, int col) {
15     /* If out of bounds or already visited. */
16     if (row < 0 || col < 0 || row >= land.length || col >= land[row].length ||
17         land[row][col] != 0) { // visited or not water
18     return 0;
19 }

```

```

20 int size = 1;
21 land[row][col] = -1; // Mark visited
22 for (int dr = -1; dr <= 1; dr++) {
23     for (int dc = -1; dc <= 1; dc++) {
24         size += computeSize(land, row + dr, col + dc);
25     }
26 }
27 return size;
28 }
```

In this case, we marked a cell as visited by setting its value to `-1`. This allows us to check, in one line (`land[row][col] != 0`), if the value is valid dry land or visited. In either case, the value will be zero.

You might also notice that the for loop iterates through nine cells, not eight. It includes the current cell. We could add a line in there to not recurse if `dr == 0` and `dc == 0`. This really doesn't save us much. We'll execute this if-statement in eight cells unnecessarily, just to avoid one recursive call. The recursive call returns immediately since the cell is marked as visited.

If you don't like modifying the input matrix, you can create a secondary `visited` matrix.

```

1 ArrayList<Integer> computePondSizes(int[][] land) {
2     boolean[][] visited = new boolean[land.length][land[0].length];
3     ArrayList<Integer> pondSizes = new ArrayList<Integer>();
4     for (int r = 0; r < land.length; r++) {
5         for (int c = 0; c < land[r].length; c++) {
6             int size = computeSize(land, visited, r, c);
7             if (size > 0) {
8                 pondSizes.add(size);
9             }
10        }
11    }
12    return pondSizes;
13 }
14
15 int computeSize(int[][] land, boolean[][] visited, int row, int col) {
16     /* If out of bounds or already visited. */
17     if (row < 0 || col < 0 || row >= land.length || col >= land[row].length ||
18         visited[row][col] || land[row][col] != 0) {
19         return 0;
20     }
21     int size = 1;
22     visited[row][col] = true;
23     for (int dr = -1; dr <= 1; dr++) {
24         for (int dc = -1; dc <= 1; dc++) {
25             size += computeSize(land, visited, row + dr, col + dc);
26         }
27     }
28     return size;
29 }
```

Both implementations are $O(WH)$, where W is the width of the matrix and H is the height.

Note: Many people say " $O(N)$ " or " $O(N^2)$ ", as though N has some inherent meaning. It doesn't. Suppose this were a square matrix. You could describe the runtime as $O(N)$ or $O(N^2)$. Both are correct, depending on what you mean by N . The runtime is $O(N^2)$, where N is the length of one side. Or, if N is the number of cells, it is $O(N)$. Be careful by what you mean by N . In fact, it might be safer to just not use N at all when there's any ambiguity as to what it could mean.

Some people will miscompute the runtime to be $O(N^4)$, reasoning that the `computeSize` method could take as long as $O(N^2)$ time and you might call it as much as $O(N^2)$ times (and apparently assuming an $N \times N$ matrix, too). While those are both basically correct statements, you can't just multiply them together. That's because as a single call to `computeSize` gets more expensive, the number of times it is called goes down.

For example, suppose the very first call to `computeSize` goes through the entire matrix. That might take $O(N^2)$ time, but then we never call `computeSize` again.

Another way to compute this is to think about how many times each cell is "touched" by either call. Each cell will be touched once by the `computePondSizes` function. Additionally, a cell might be touched once by each of its adjacent cells. This is still a constant number of touches per cell. Therefore, the overall runtime is $O(N^2)$ on an $N \times N$ matrix or, more generally, $O(WH)$.

- 16.20 T9:** On old cell phones, users typed on a numeric keypad and the phone would provide a list of words that matched these numbers. Each digit mapped to a set of 0 - 4 letters. Implement an algorithm to return a list of matching words, given a sequence of digits. You are provided a list of valid words (provided in whatever data structure you'd like). The mapping is shown in the diagram below:

1	2 abc	3 def
4 ghi	5 jkl	6 mno
7 pqrs	8 tuv	9 wxyz
		0

EXAMPLE

Input: 8733
Output: tree, used

pg 184

SOLUTION

We could approach this in a couple of ways. Let's start with a brute force algorithm.

Brute Force

Imagine how you would solve the problem if you had to do it by hand. You'd probably try every possible value for each digit with all other possible values.

This is exactly what we do algorithmically. We take the first digit and run through all the characters that map to that digit. For each character, we add it to a `prefix` variable and recurse, passing the `prefix` downward. Once we run out of characters, we print `prefix` (which now contains the full word) if the string is a valid word.

We will assume the list of words is passed in as a `HashSet`. A `HashSet` operates similarly to a hash table, but rather than offering key->value lookups, it can tell us if a word is contained in the set in $O(1)$ time.

```

1  ArrayList<String> getValidT9Words(String number, HashSet<String> wordList) {
2      ArrayList<String> results = new ArrayList<String>();
3      getValidWords(number, 0, "", wordList, results);
4      return results;
5  }
6

```

```

7 void getValidWords(String number, int index, String prefix,
8                     HashSet<String> wordSet, ArrayList<String> results) {
9     /* If it's a complete word, print it. */
10    if (index == number.length() && wordSet.contains(prefix)) {
11        results.add(prefix);
12        return;
13    }
14
15    /* Get characters that match this digit. */
16    char digit = number.charAt(index);
17    char[] letters = getT9Chars(digit);
18
19    /* Go through all remaining options. */
20    if (letters != null) {
21        for (char letter : letters) {
22            getValidWords(number, index + 1, prefix + letter, wordSet, results);
23        }
24    }
25 }
26
27 /* Return array of characters that map to this digit. */
28 char[] getT9Chars(char digit) {
29     if (!Character.isDigit(digit)) {
30         return null;
31     }
32     int dig = Character.getNumericValue(digit) - Character.getNumericValue('0');
33     return t9Letters[dig];
34 }
35
36 /* Mapping of digits to letters. */
37 char[][] t9Letters = {null, null, {'a', 'b', 'c'}, {'d', 'e', 'f'},
38                      {'g', 'h', 'i'}, {'j', 'k', 'l'}, {'m', 'n', 'o'}, {'p', 'q', 'r', 's'},
39                      {'t', 'u', 'v'}, {'w', 'x', 'y', 'z'}};
40 };

```

This algorithm runs in $O(4^N)$ time, where N is the length of the string. This is because we recursively branch four times for each call to `getValidWords`, and we recurse until a call stack depth of N .

This is very, very slow on large strings.

Optimized

Let's return to thinking about how you would do this, if you were doing it by hand. Imagine the example of 33835676368 (which corresponds to development). If you were doing this by hand, I bet you'd skip over solutions that start with fftf [3383], as no valid words start with those characters.

Ideally, we'd like our program to make the same sort of optimization: stop recursing down paths which will obviously fail. Specifically, if there are no words in the dictionary that start with `prefix`, stop recursing.

The Trie data structure (see "Tries (Prefix Trees)" on page 105) can do this for us. Whenever we reach a string which is not a valid prefix, we exit.

```

1 ArrayList<String> getValidT9Words(String number, Trie trie) {
2     ArrayList<String> results = new ArrayList<String>();
3     getValidWords(number, 0, "", trie.getRoot(), results);
4     return results;
5 }
6

```

```

7 void getValidWords(String number, int index, String prefix, TrieNode trieNode,
8                     ArrayList<String> results) {
9     /* If it's a complete word, print it. */
10    if (index == number.length()) {
11        if (trieNode.terminates()) { // Is complete word
12            results.add(prefix);
13        }
14    }
15    return;
16 }
17 /* Get characters that match this digit */
18 char digit = number.charAt(index);
19 char[] letters = getT9Chars(digit);
20
21 /* Go through all remaining options. */
22 if (letters != null) {
23     for (char letter : letters) {
24         TrieNode child = trieNode.getChild(letter);
25         /* If there are words that start with prefix + letter,
26          * then continue recursing. */
27         if (child != null) {
28             getValidWords(number, index + 1, prefix + letter, child, results);
29         }
30     }
31 }
32 }
```

It's difficult to describe the runtime of this algorithm since it depends on what the language looks like. However, this "short-circuiting" will make it run much, much faster in practice.

Most Optimal

Believe or not, we can actually make it run even faster. We just need to do a little bit of preprocessing. That's not a big deal though. We were doing that to build the trie anyway.

This problem is asking us to list all the words represented by a particular number in T9. Instead of trying to do this "on the fly" (and going through a lot of possibilities, many of which won't actually work), we can just do this in advance.

Our algorithm now has a few steps:

Pre-Computation:

1. Create a hash table that maps from a sequence of digits to a list of strings.
2. Go through each word in the dictionary and convert it to its T9 representation (e.g., APPLE -> 27753). Store each of these in the above hash table. For example, 8733 would map to {used, tree}.

Word Lookup:

1. Just look up the entry in the hash table and return the list.

That's it!

```

1 /* WORD LOOKUP */
2 ArrayList<String> getValidT9Words(String numbers,
3                                     HashMapList<String, String> dictionary) {
4     return dictionary.get(numbers);
5 }
```

```

7  /* PRECOMPUTATION */
8
9  /* Create a hash table that maps from a number to all words that have this
10 * numerical representation. */
11 HashMapList<String, String> initializeDictionary(String[] words) {
12     /* Create a hash table that maps from a letter to the digit */
13     HashMap<Character, Character> letterToNumberMap = createLetterToNumberMap();
14
15     /* Create word -> number map. */
16     HashMapList<String, String> wordsToNumbers = new HashMapList<String, String>();
17     for (String word : words) {
18         String numbers = convertToT9(word, letterToNumberMap);
19         wordsToNumbers.put(numbers, word);
20     }
21     return wordsToNumbers;
22 }
23
24 /* Convert mapping of number->letters into letter->number. */
25 HashMap<Character, Character> createLetterToNumberMap() {
26     HashMap<Character, Character> letterToNumberMap =
27         new HashMap<Character, Character>();
28     for (int i = 0; i < t9Letters.length; i++) {
29         char[] letters = t9Letters[i];
30         if (letters != null) {
31             for (char letter : letters) {
32                 char c = Character.forDigit(i, 10);
33                 letterToNumberMap.put(letter, c);
34             }
35         }
36     }
37     return letterToNumberMap;
38 }
39
40 /* Convert from a string to its T9 representation. */
41 String convertToT9(String word, HashMap<Character, Character> letterToNumberMap) {
42     StringBuilder sb = new StringBuilder();
43     for (char c : word.toCharArray()) {
44         if (letterToNumberMap.containsKey(c)) {
45             char digit = letterToNumberMap.get(c);
46             sb.append(digit);
47         }
48     }
49     return sb.toString();
50 }
51
52 char[][] t9Letters = /* Same as before */
53
54 /* HashMapList<String, Integer> is a HashMap that maps from Strings to
55 * ArrayList<Integer>. See appendix for implementation. */

```

Getting the words that map to this number will run in $O(N)$ time, where N is the number of digits. The $O(N)$ comes in during the hash table look up (we need to convert the number to a hash table). If you know the words are never longer than a certain max size, then you could also describe the runtime as $O(1)$.

Note that it's easy to think, "Oh, linear—that's not that fast." But it depends what it's linear on. Linear on the length of the word is extremely fast. Linear on the length of the dictionary is not so fast.

16.21 Sum Swap: Given two arrays of integers, find a pair of values (one value from each array) that you can swap to give the two arrays the same sum.

EXAMPLE

Input: {4, 1, 2, 1, 1, 2} and {3, 6, 3, 3}

Output: {1, 3}

pg 184

SOLUTION

We should start by trying to understand what exactly we're looking for.

We have two arrays and their sums. Although we likely aren't given their sums upfront, we can just act like we are for now. After all, computing the sum is an O(N) operation and we know we can't beat O(N) anyway. Computing the sum, therefore, won't impact the runtime.

When we move a (positive) value a from array A to array B, then the sum of A drops by a and the sum of B increases by a .

We are looking for two values, a and b , such that:

$$\text{sumA} - a + b = \text{sumB} - b + a$$

Doing some quick math:

$$2a - 2b = \text{sumA} - \text{sumB}$$

$$a - b = (\text{sumA} - \text{sumB}) / 2$$

Therefore, we're looking for two values that have a specific target difference: $(\text{sumA} - \text{sumB}) / 2$.

Observe that because that the target must be an integer (after all, you can't swap two integers to get a non-integer difference), we can conclude that the difference between the sums must be even to have a valid pair.

Brute Force

A brute force algorithm is simple enough. We just iterate through the arrays and check all pairs of values.

We can either do this the "naive" way (compare the new sums) or by looking for a pair with that difference.

Naive approach:

```

1 int[] findSwapValues(int[] array1, int[] array2) {
2     int sum1 = sum(array1);
3     int sum2 = sum(array2);
4
5     for (int one : array1) {
6         for (int two : array2) {
7             int newSum1 = sum1 - one + two;
8             int newSum2 = sum2 - two + one;
9             if (newSum1 == newSum2) {
10                 int[] values = {one, two};
11                 return values;
12             }
13         }
14     }
15
16     return null;
17 }
```

Target approach:

```
1 int[] findSwapValues(int[] array1, int[] array2) {
```

```
2     Integer target = getTarget(array1, array2);
3     if (target == null) return null;
4
5     for (int one : array1) {
6         for (int two : array2) {
7             if (one - two == target) {
8                 int[] values = {one, two};
9                 return values;
10            }
11        }
12    }
13
14    return null;
15 }
```

```
17 Integer getTarget(int[] array1, int[] array2) {
18     int sum1 = sum(array1);
19     int sum2 = sum(array2);
20
21     if ((sum1 - sum2) % 2 != 0) return null;
22     return (sum1 - sum2) / 2;
23 }
```

We've used an `Integer` (a boxed data type) as the return value for `getTarget`. This allows us to distinguish an "error" case.

This algorithm takes $O(AB)$ time.

Optimal Solution

This problem reduces to finding a pair of values that have a particular difference. With that in mind, let's revisit what the brute force does.

In the brute force, we're looping through A and then, for each element, looking for an element in B which gives us the "right" difference. If the value in A is 5 and the target is 3, then we must be looking for the value 2. That's the only value that could fulfill the goal.

That is, rather than writing `one - two == target`, we could have written `two == one - target`. How can we more quickly find an element in B that equals `one - target`?

We can do this very quickly with a hash table. We just throw all the elements in B into a hash table. Then, iterate through A and look for the appropriate element in B.

```
1  int[] findSwapValues(int[] array1, int[] array2) {
2      Integer target = getTarget(array1, array2);
3      if (target == null) return null;
4      return findDifference(array1, array2, target);
5  }
6
7  /* Find a pair of values with a specific difference. */
8  int[] findDifference(int[] array1, int[] array2, int target) {
9      HashSet<Integer> contents2 = getContents(array2);
10     for (int one : array1) {
11         int two = one - target;
12         if (contents2.contains(two)) {
13             int[] values = {one, two};
14             return values;
15         }
16     }
17 }
```

```

16     }
17
18     return null;
19 }
20
21 /* Put contents of array into hash set. */
22 HashSet<Integer> getContents(int[] array) {
23     HashSet<Integer> set = new HashSet<Integer>();
24     for (int a : array) {
25         set.add(a);
26     }
27     return set;
28 }
```

This solution will take $O(A+B)$ time. This is the Best Conceivable Runtime (BCR), since we have to at least touch every element in the two arrays.

Alternate Solution

If the arrays are sorted, we can iterate through them to find an appropriate pair. This will require less space.

```

1 int[] findSwapValues(int[] array1, int[] array2) {
2     Integer target = getTarget(array1, array2);
3     if (target == null) return null;
4     return findDifference(array1, array2, target);
5 }
6
7 int[] findDifference(int[] array1, int[] array2, int target) {
8     int a = 0;
9     int b = 0;
10
11    while (a < array1.length && b < array2.length) {
12        int difference = array1[a] - array2[b];
13        /* Compare difference to target. If difference is too small, then make it
14         * bigger by moving a to a bigger value. If it is too big, then make it
15         * smaller by moving b to a bigger value. If it's just right, return this
16         * pair. */
17        if (difference == target) {
18            int[] values = {array1[a], array2[b]};
19            return values;
20        } else if (difference < target) {
21            a++;
22        } else {
23            b++;
24        }
25    }
26
27    return null;
28 }
```

This algorithm takes $O(A + B)$ time but requires the arrays to be sorted. If the arrays aren't sorted, we can still apply this algorithm but we'd have to sort the arrays first. The overall runtime would be $O(A \log A + B \log B)$.

16.22 Langton's Ant: An ant is sitting on an infinite grid of white and black squares. It initially faces right. At each step, it does the following:

(1) At a white square, flip the color of the square, turn 90 degrees right (clockwise), and move forward one unit.

(2) At a black square, flip the color of the square, turn 90 degrees left (counter-clockwise), and move forward one unit.

Write a program to simulate the first K moves that the ant makes and print the final board as a grid. Note that you are not provided with the data structure to represent the grid. This is something you must design yourself. The only input to your method is K. You should print the final grid and return nothing. The method signature might be something like void printKMoves(int K).

pg 185

SOLUTION

At first glance, this problem seems very straightforward: create a grid, remember the ant's position and orientation, flip the cells, turn, and move. The interesting part comes in how to handle an infinite grid.

Solution #1: Fixed Array

Technically, since we're only running the first K moves, we do have a max size for the grid. The ant cannot move more than K moves in either direction. If we create a grid that has width 2K and height 2K (and place the ant at the center), we know it will be big enough.

The problem with this is that it's not very extensible. If you run K moves and then want to run another K moves, you might be out of luck.

Additionally, this solution wastes a good amount of space. The max might be K moves in a particular dimension, but the ant is probably going in circles a bit. You probably won't need all this space.

Solution #2: Resizable Array

One thought is to use a resizable array, such as Java's `ArrayList` class. This allows us to grow an array as necessary, while still offering $O(1)$ amortized insertion.

The problem is that our grid needs to grow in two dimensions, but the `ArrayList` is only a single array. Additionally, we need to grow "backward" into negative values. The `ArrayList` class doesn't support this.

However, we take a similar approach by building our own resizable grid. Each time the ant hits an edge, we double the size of the grid in that dimension.

What about the negative expansions? While conceptually we can talk about something being at negative positions, we cannot actually access array indices with negative values.

One way we can handle this is to create "fake indices." Let us treat the ant as being at coordinates (-3, -10), but track some sort of offset or delta to translate these coordinates into array indices.

This is actually unnecessary, though. The ant's location does not need to be publicly exposed or consistent (unless, of course, indicated by the interviewer). When the ant travels into negative coordinates, we can double the size of the array and just move the ant and all cells into the positive coordinates. Essentially, we are relabeling all the indices.

This relabeling will not impact the big O time since we have to create a new matrix anyway.

```
1 public class Grid {  
2     private boolean[][] grid;
```

```

3  private Ant ant = new Ant();
4
5  public Grid() {
6      grid = new boolean[1][1];
7  }
8
9  /* Copy old values into new array, with an offset/shift applied to the row and
10   * columns. */
11 private void copyWithShift(boolean[][] oldGrid, boolean[][] newGrid,
12                           int shiftRow, int shiftColumn) {
13     for (int r = 0; r < oldGrid.length; r++) {
14         for (int c = 0; c < oldGrid[0].length; c++) {
15             newGrid[r + shiftRow][c + shiftColumn] = oldGrid[r][c];
16         }
17     }
18 }
19
20 /* Ensure that the given position will fit on the array. If necessary, double
21  * the size of the matrix, copy the old values over, and adjust the ant's
22  * position so that it's in a positive range. */
23 private void ensureFit(Position position) {
24     int shiftRow = 0;
25     int shiftColumn = 0;
26
27     /* Calculate new number of rows. */
28     int numRows = grid.length;
29     if (position.row < 0) {
30         shiftRow = numRows;
31         numRows *= 2;
32     } else if (position.row >= numRows) {
33         numRows *= 2;
34     }
35
36     /* Calculate new number of columns. */
37     int numColumns = grid[0].length;
38     if (position.column < 0) {
39         shiftColumn = numColumns;
40         numColumns *= 2;
41     } else if (position.column >= numColumns) {
42         numColumns *= 2;
43     }
44
45     /* Grow array, if necessary. Shift ant's position too. */
46     if (numRows != grid.length || numColumns != grid[0].length) {
47         boolean[][] newGrid = new boolean[numRows][numColumns];
48         copyWithShift(grid, newGrid, shiftRow, shiftColumn);
49         ant.adjustPosition(shiftRow, shiftColumn);
50         grid = newGrid;
51     }
52 }
53
54 /* Flip color of cells. */
55 private void flip(Position position) {
56     int row = position.row;
57     int column = position.column;
58     grid[row][column] = grid[row][column] ? false : true;

```

```

59     }
60
61     /* Move ant. */
62     public void move() {
63         ant.turn(grid[ant.position.row][ant.position.column]);
64         flip(ant.position);
65         ant.move();
66         ensureFit(ant.position); // grow
67     }
68
69     /* Print board. */
70     public String toString() {
71         StringBuilder sb = new StringBuilder();
72         for (int r = 0; r < grid.length; r++) {
73             for (int c = 0; c < grid[0].length; c++) {
74                 if (r == ant.position.row && c == ant.position.column) {
75                     sb.append(ant.orientation);
76                 } else if (grid[r][c]) {
77                     sb.append("X");
78                 } else {
79                     sb.append("_");
80                 }
81             }
82             sb.append("\n");
83         }
84         sb.append("Ant: " + ant.orientation + ". \n");
85         return sb.toString();
86     }
87 }

```

We pulled the Ant code into a separate class. The nice thing about this is that if we need to have multiple ants for some reason, we can easily extend the code to support this.

```

1  public class Ant {
2      public Position position = new Position(0, 0);
3      public Orientation orientation = Orientation.right;
4
5      public void turn(boolean clockwise) {
6          orientation = orientation.getTurn(clockwise);
7      }
8
9      public void move() {
10         if (orientation == Orientation.left) {
11             position.column--;
12         } else if (orientation == Orientation.right) {
13             position.column++;
14         } else if (orientation == Orientation.up) {
15             position.row--;
16         } else if (orientation == Orientation.down) {
17             position.row++;
18         }
19     }
20
21     public void adjustPosition(int shiftRow, int shiftColumn) {
22         position.row += shiftRow;
23         position.column += shiftColumn;
24     }
25 }

```

Orientation is also its own enum, with a few useful functions.

```

1  public enum Orientation {
2      left, up, right, down;
3
4      public Orientation getTurn(boolean clockwise) {
5          if (this == left) {
6              return clockwise ? up : down;
7          } else if (this == up) {
8              return clockwise ? right : left;
9          } else if (this == right) {
10             return clockwise ? down : up;
11         } else { // down
12             return clockwise ? left : right;
13         }
14     }
15
16    @Override
17    public String toString() {
18        if (this == left) {
19            return "\u2190";
20        } else if (this == up) {
21            return "\u2191";
22        } else if (this == right) {
23            return "\u2192";
24        } else { // down
25            return "\u2193";
26        }
27    }
28 }
```

We've also put Position into its own simple class. We could just as easily track the row and column separately.

```

1  public class Position {
2      public int row;
3      public int column;
4
5      public Position(int row, int column) {
6          this.row = row;
7          this.column = column;
8      }
9  }
```

This works, but it's actually more complicated than is necessary.

Solution #3: HashSet

Although it may seem "obvious" that we would use a matrix to represent a grid, it's actually easier not to do that. All we actually need is a list of the white squares (as well as the ant's location and orientation).

We can do this by using a HashSet of the white squares. If a position is in the hash set, then the square is white. Otherwise, it is black.

The one tricky bit is how to print the board. Where do we start printing? Where do we end?

Since we will need to print a grid, we can track what should be top-left and bottom-right corner of the grid. Each time the ant moves, we compare the ant's position to the most top-left position and most bottom-right position, updating them if necessary.

```
1  public class Board {
2      private HashSet<Position> whites = new HashSet<Position>();
3      private Ant ant = new Ant();
4      private Position topLeftCorner = new Position(0, 0);
5      private Position bottomRightCorner = new Position(0, 0);
6
7      public Board() { }
8
9      /* Move ant. */
10     public void move() {
11         ant.turn(isWhite(ant.position)); // Turn
12         flip(ant.position); // flip
13         ant.move(); // move
14         ensureFit(ant.position);
15     }
16
17     /* Flip color of cells. */
18     private void flip(Position position) {
19         if (whites.contains(position)) {
20             whites.remove(position);
21         } else {
22             whites.add(position.clone());
23         }
24     }
25
26     /* Grow grid by tracking the most top-left and bottom-right positions.*/
27     private void ensureFit(Position position) {
28         int row = position.row;
29         int column = position.column;
30
31         topLeftCorner.row = Math.min(topLeftCorner.row, row);
32         topLeftCorner.column = Math.min(topLeftCorner.column, column);
33
34         bottomRightCorner.row = Math.max(bottomRightCorner.row, row);
35         bottomRightCorner.column = Math.max(bottomRightCorner.column, column);
36     }
37
38     /* Check if cell is white. */
39     public boolean isWhite(Position p) {
40         return whites.contains(p);
41     }
42
43     /* Check if cell is white. */
44     public boolean isWhite(int row, int column) {
45         return whites.contains(new Position(row, column));
46     }
47
48     /* Print board. */
49     public String toString() {
50         StringBuilder sb = new StringBuilder();
51         int rowMin = topLeftCorner.row;
52         int rowMax = bottomRightCorner.row;
53         int colMin = topLeftCorner.column;
54         int colMax = bottomRightCorner.column;
55         for (int r = rowMin; r <= rowMax; r++) {
56             for (int c = colMin; c <= colMax; c++) {
```

```

57         if (r == ant.position.row && c == ant.position.column) {
58             sb.append(ant.orientation);
59         } else if (isWhite(r, c)) {
60             sb.append("X");
61         } else {
62             sb.append("_");
63         }
64     }
65     sb.append("\n");
66 }
67 sb.append("Ant: " + ant.orientation + ", \n");
68 return sb.toString();
69 }
```

The implementation of Ant and Orientation is the same.

The implementation of Position gets updated slightly, in order to support the HashSet functionality. The position will be the key, so we need to implement a hashCode() function.

```

1  public class Position {
2      public int row;
3      public int column;
4
5      public Position(int row, int column) {
6          this.row = row;
7          this.column = column;
8      }
9
10     @Override
11     public boolean equals(Object o) {
12         if (o instanceof Position) {
13             Position p = (Position) o;
14             return p.row == row && p.column == column;
15         }
16         return false;
17     }
18
19     @Override
20     public int hashCode() {
21         /* There are many options for hash functions. This is one. */
22         return (row * 31) ^ column;
23     }
24
25     public Position clone() {
26         return new Position(row, column);
27     }
28 }
```

The nice thing about this implementation is that if we do need to access a particular cell elsewhere, we have consistent row and column labeling.

- 16.23 Rand7 from Rand5:** 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 186

SOLUTION

To implement this function correctly, we must have each of the values between 0 and 6 returned with $\frac{1}{7}$ th probability.

First Attempt (Fixed Number of Calls)

As a first attempt, we might try generating 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 the results of each call to `rand5()` and the return result of the `rand7()` function.

1st Call	2nd Call	Result	1st Call	2nd Call	Result
0	0	0	2	3	5
0	1	1	2	4	6
0	2	2	3	0	3
0	3	3	3	1	4
0	4	4	3	2	5
1	0	1	3	3	6
1	1	2	3	4	0
1	2	3	4	0	4
1	3	4	4	1	5
1	4	5	4	2	6
2	0	2	4	3	0
2	1	3	4	4	1
2	2	4			

Each individual row has a $\frac{1}{25}$ chance of occurring, since there are two calls to `rand5()` and each distributes its results with $\frac{1}{5}$ th probability. If you count up the number of times each number occurs, you'll note that this `rand7()` function will return 4 with $\frac{5}{25}$ th probability but return 0 with just $\frac{3}{25}$ th probability. This means that our function has failed; the results do not have probability $\frac{1}{7}$ th.

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 $\frac{1}{5^k}$, 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:

$$P(\text{rand7}() = 6) = \frac{1}{5^1} + \frac{1}{5^1} + \dots + \frac{1}{5^m}$$

We know that, in order for our function to be correct, this probability must equal $\frac{1}{7}$. This is impossible though. Because 5 and 7 are relatively prime, no series of reciprocal powers of 5 will result in $\frac{1}{7}$.

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 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 three ways of getting a 6 ($6 = 2 * 1 + 4$, $6 = 2 * 2 + 2$, and $6 = 2 * 3 + 0$) 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 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.

16.24 Pairs with Sum: Design an algorithm to find all pairs of integers within an array which sum to a specified value.

pg 185

SOLUTION

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.

Brute Force

A brute force solution is to just iterate through all pairs and print the pair if its sum matches the target sum.

```
1 ArrayList<Pair> printPairSums(int[] array, int sum) {  
2     ArrayList<Pair> result = new ArrayList<Pair>();  
3     for (int i = 0 ; i < array.length; i++) {  
4         for (int j = i + 1; j < array.length; j++) {  
5             if (array[i] + array[j] == sum) {  
6                 result.add(new Pair(array[i], array[j]));  
7             }  
8         }  
9     }  
10    return result;  
11 }
```

If there are duplicates in the array (e.g., $\{5, 6, 5\}$), it might print the same sum twice. You should discuss this with your interviewer.

Optimized Solution

We can optimize this with a hash map, where the value in the hash map reflects the number of "unpaired" instances of a key. We walk through the array. At each element x , check how many unpaired instances of x 's complement preceded it in the array. If the count is at least one, then there is an unpaired instance of x 's complement. We add this pair and decrement x 's complement to signify that this element has been paired. If the count is zero, then increment the value of x in the hash table to signify that x is unpaired.

```
1 ArrayList<Pair> printPairSums(int[] array, int sum) {  
2     ArrayList<Pair> result = new ArrayList<Pair>();  
3     HashMap<Integer, Integer> unpairedCount = new HashMap<Integer, Integer>();  
4     for (int x : array) {  
5         int complement = sum - x;  
6         if (unpairedCount.getOrDefault(complement, 0) > 0) {  
7             result.add(new Pair(x, complement));  
8             adjustCounterBy(unpairedCount, complement, -1); // decrement complement  
9         } else {  
10             adjustCounterBy(unpairedCount, x, 1); // increment count  
11         }  
12     }  
13     return result;  
14 }  
15 }
```

```

16 void adjustCounterBy(HashMap<Integer, Integer> counter, int key, int delta) {
17     counter.put(key, counter.getOrDefault(key, 0) + delta);
18 }

```

This solution will print duplicate pairs, but will not reuse the same instance of an element. It will take $O(N)$ time and $O(N)$ space.

Alternate Solution

Alternatively, we can sort the array and then find the pairs in a single pass. Consider this 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 }

```

This algorithm takes $O(N \log N)$ time to sort and $O(N)$ time to find the pairs.

Note that since the array is presumably unsorted, it would be equally fast in terms of big O to just do a binary search at each element for its complement. This would give us a two-step algorithm, where each step is $O(N \log N)$.

16.25 LRU Cache: Design and build a “least recently used” cache, which evicts the least recently used item. The cache should map from keys to values (allowing you to insert and retrieve a value associated with a particular key) and be initialized with a max size. When it is full, it should evict the least recently used item. You can assume the keys are integers and the values are strings.

pg 185

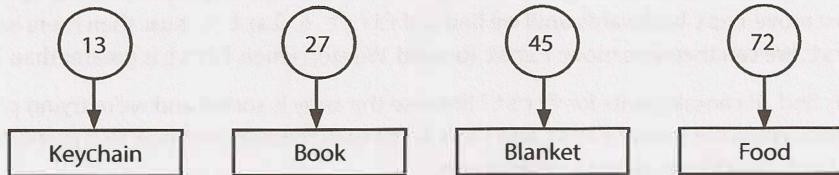
SOLUTION

We should start off by defining the scope of the problem. What exactly do we need to achieve?

- **Inserting Key, Value Pair:** We need to be able to insert a (key, value) pair.

- Retrieving Value by Key:** We need to be able to retrieve the value using the key.
- Finding Least Recently Used:** We need to know the least recently used item (and, likely, the usage ordering of all items).
- Updating Most Recently Used:** When we retrieve a value by key, we need to update the order to be the most recently used item.
- Eviction:** The cache should have a max capacity and should remove the least recently used item when it hits capacity.

The (key, value) mapping suggests a hash table. This would make it easy to look up the value associated with a particular key.



Unfortunately, a hash table usually would not offer a quick way to remove the most recently used item. We could mark each item with a timestamp and iterate through the hash table to remove the item with the lowest timestamp, but that can get quite slow ($O(N)$ for insertions).

Instead, we could use a linked list, ordered by the most recently used. This would make it easy to mark an item as the most recently used (just put it in the front of the list) or to remove the least recently used item (remove the end).

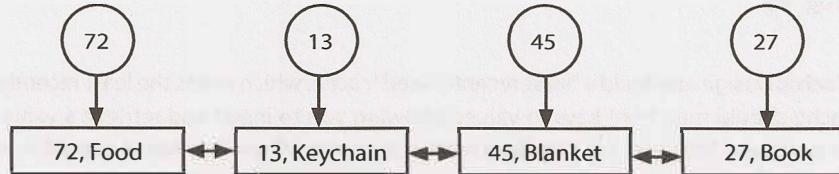


Unfortunately, this does not offer a quick way to look up an item by its key. We could iterate through the linked list and find the item by key. But this could get very slow ($O(N)$ for retrieval).

Each approach does half of the problem (different halves) very well, but neither approach does both parts well.

Can we get the best parts of each? Yes. By using both!

The linked list looks as it did in the earlier example, but now it's a doubly linked list. This allows us to easily remove an element from the middle of the linked list. The hash table now maps to each linked list node rather than the value.



The algorithms now operate as follows:

- Inserting Key, Value Pair:** Create a linked list node with key, value. Insert into head of linked list. Insert key \rightarrow node mapping into hash table.
- Retrieving Value by Key:** Look up node in hash table and return value. Update most recently used item

(see below).

- **Finding Least Recently Used:** Least recently used item will be found at the end of the linked list.
- **Updating Most Recently Used:** Move node to front of linked list. Hash table does not need to be updated.
- **Eviction:** Remove tail of linked list. Get key from linked list node and remove key from hash table.

The code below implements these classes and algorithms.

```

1  public class Cache {
2      private int maxCacheSize;
3      private HashMap<Integer, LinkedListNode> map =
4          new HashMap<Integer, LinkedListNode>();
5      private LinkedListNode listHead = null;
6      public LinkedListNode listTail = null;
7
8      public Cache(int maxSize) {
9          maxCacheSize = maxSize;
10     }
11
12     /* Get value for key and mark as most recently used. */
13     public String getValue(int key) {
14         LinkedListNode item = map.get(key);
15         if (item == null) return null;
16
17         /* Move to front of list to mark as most recently used. */
18         if (item != listHead) {
19             removeFromLinkedList(item);
20             insertAtFrontOfLinkedList(item);
21         }
22         return item.value;
23     }
24
25     /* Remove node from linked list. */
26     private void removeFromLinkedList(LinkedListNode node) {
27         if (node == null) return;
28
29         if (node.prev != null) node.prev.next = node.next;
30         if (node.next != null) node.next.prev = node.prev;
31         if (node == listTail) listTail = node.prev;
32         if (node == listHead) listHead = node.next;
33     }
34
35     /* Insert node at front of linked list. */
36     private void insertAtFrontOfLinkedList(LinkedListNode node) {
37         if (listHead == null) {
38             listHead = node;
39             listTail = node;
40         } else {
41             listHead.prev = node;
42             node.next = listHead;
43             listHead = node;
44         }
45     }
46
47     /* Remove key/value pair from cache, deleting from hashtable and linked list. */
48     public boolean removeKey(int key) {

```

```
49     LinkedListNode node = map.get(key);
50     removeFromLinkedList(node);
51     map.remove(key);
52     return true;
53 }
54
55 /* Put key, value pair in cache. Removes old value for key if necessary. Inserts
56 * pair into linked list and hash table.*/
57 public void setKeyValue(int key, String value) {
58     /* Remove if already there. */
59     removeKey(key);
60
61     /* If full, remove least recently used item from cache. */
62     if (map.size() >= maxCacheSize && listTail != null) {
63         removeKey(listTail.key);
64     }
65
66     /* Insert new node. */
67     LinkedListNode node = new LinkedListNode(key, value);
68     insertAtFrontOfLinkedList(node);
69     map.put(key, node);
70 }
71
72 private static class LinkedListNode {
73     private LinkedListNode next, prev;
74     public int key;
75     public String value;
76     public LinkedListNode(int k, String v) {
77         key = k;
78         value = v;
79     }
80 }
81 }
```

Note that we've chosen to make `LinkedListNode` an inner class of `Cache`, since no other classes should need access to this class and really should only exist within the scope of `Cache`.

16.26 Calculator: Given an arithmetic equation consisting of positive integers, `+, -, *` and `/` (no parentheses), compute the result.

EXAMPLE

Input: $2*3+5/6*3+15$

Output: 23.5

pg 185

SOLUTION

The first thing we should realize is that the dumb thing—just applying each operator left to right—won't work. Multiplication and division are considered "higher priority" operations, which means that they have to happen before addition.

For example, if you have the simple expression $3+6*2$, the multiplication must be performed first, and then the addition. If you just processed the equation left to right, you would end up with the incorrect result, 18, rather than the correct one, 15. You know all of this, of course, but it's worth really spelling out what it means.

Solution #1

We can still process the equation from left to right; we just have to be a little smarter about how we do it. Multiplication and division need to be grouped together such that whenever we see those operations, we perform them immediately on the surrounding terms.

For example, suppose we have this expression:

$2 - 6 - 7*8/2 + 5$

It's fine to compute $2 - 6$ immediately and store it into a `result` variable. But, when we see $7*$ (something), we know we need to fully process that term before adding it to the result.

We can do this by reading left to right and maintaining two variables.

- The first is `processing`, which maintains the result of the current cluster of terms (both the operator and the value). In the case of addition and subtraction, the cluster will be just the current term. In the case of multiplication and division, it will be the full sequence (until you get to the next addition or subtraction).
- The second is the `result` variable. If the next term is an addition or subtraction (or there is no next term), then `processing` is applied to `result`.

On the above example, we would do the following:

1. Read $+$. Apply it to `processing`. Apply `processing` to `result`. Clear `processing`.

```
processing = {+, 2} --> null
result = 0           --> 2
```

2. Read -6 . Apply it to `processing`. Apply `processing` to `result`. Clear `processing`.

```
processing = {-, 6} --> null
result = 2           --> -4
```

3. Read -7 . Apply it to `processing`. Observe next sign is a $*$. Continue.

```
processing = {-, 7}
result = -4
```

4. Read $*8$. Apply it to `processing`. Observe next sign is a $/$. Continue.

```
processing = {-, 56}
result = -4
```

5. Read $/2$. Apply it to `processing`. Observe next sign is a $+$, which terminates this multiplication and division cluster. Apply `processing` to `result`. Clear `processing`.

```
processing = {-, 28} --> null
result = -4           --> -32
```

6. Read $+5$. Apply it to `processing`. Apply `processing` to `result`. Clear `processing`.

```
processing = {+, 5} --> null
result = -32          --> -27
```

The code below implements this algorithm.

```
1  /* Compute the result of the arithmetic sequence. This works by reading left to
2   * right and applying each term to a result. When we see a multiplication or
3   * division, we instead apply this sequence to a temporary variable. */
4  double compute(String sequence) {
5      ArrayList<Term> terms = Term.parseTermSequence(sequence);
6      if (terms == null) return Integer.MIN_VALUE;
7
8      double result = 0;
9      Term processing = null;
10     for (int i = 0; i < terms.size(); i++) {
```

```

11     Term current = terms.get(i);
12     Term next = i + 1 < terms.size() ? terms.get(i + 1) : null;
13
14     /* Apply the current term to "processing". */
15     processing = collapseTerm(processing, current);
16
17     /* If next term is + or -, then this cluster is done and we should apply
18      * "processing" to "result". */
19     if (next == null || next.getOperator() == Operator.ADD
20         || next.getOperator() == Operator.SUBTRACT) {
21         result = applyOp(result, processing.getOperator(), processing.getNumber());
22         processing = null;
23     }
24 }
25
26 return result;
27 }
28
29 /* Collapse two terms together using the operator in secondary and the numbers
30  * from each. */
31 Term collapseTerm(Term primary, Term secondary) {
32     if (primary == null) return secondary;
33     if (secondary == null) return primary;
34
35     double value = applyOp(primary.getNumber(), secondary.getOperator(),
36                           secondary.getNumber());
37     primary.setNumber(value);
38     return primary;
39 }
40
41 double applyOp(double left, Operator op, double right) {
42     if (op == Operator.ADD) return left + right;
43     else if (op == Operator.SUBTRACT) return left - right;
44     else if (op == Operator.MULTIPLY) return left * right;
45     else if (op == Operator.DIVIDE) return left / right;
46     else return right;
47 }
48
49 public class Term {
50     public enum Operator {
51         ADD, SUBTRACT, MULTIPLY, DIVIDE, BLANK
52     }
53
54     private double value;
55     private Operator operator = Operator.BLANK;
56
57     public Term(double v, Operator op) {
58         value = v;
59         operator = op;
60     }
61
62     public double getNumber() { return value; }
63     public Operator getOperator() { return operator; }
64     public void setNumber(double v) { value = v; }
65
66     /* Parses arithmetic sequence into a list of Terms. For example, 3-5*6 becomes

```

```

67     * something like: [{BLANK,3}, {SUBTRACT, 5}, {MULTIPLY, 6}].  

68     * If improperly formatted, returns null. */  

69     public static ArrayList<Term> parseTermSequence(String sequence) {  

70         /* Code can be found in downloadable solutions. */  

71     }  

72 }

```

This takes $O(N)$ time, where N is the length of the initial string.

Solution #2

Alternatively, we can solve this problem using two stacks: one for numbers and one for operators.

$2 - 6 - 7 * 8 / 2 + 5$

The processing works as follows:

- Each time we see a number, it gets pushed onto `numberStack`.
- Operators get pushed onto `operatorStack`—as long as the operator has higher priority than the current top of the stack. If `priority(currentOperator) <= priority(operatorStack.top())`, then we “collapse” the top of the stacks:
 - » Collapsing: pop two elements off `numberStack`, pop an operator off `operatorStack`, apply the operator, and push the result onto `numberStack`.
 - » Priority: addition and subtraction have equal priority, which is lower than the priority of multiplication and division (also equal priority).

This collapsing continues until the above inequality is broken, at which point `currentOperator` is pushed onto `operatorStack`.

- At the very end, we collapse the stack.

Let's see this with an example: $2 - 6 - 7 * 8 / 2 + 5$

	action	numberStack	operatorStack
2	numberStack.push(2)	2	[empty]
-	operatorStack.push(-)	2	-
6	numberStack.push(6)	6, 2	-
-	collapseStacks [2 - 6] operatorStack.push(-)	-4 -4	[empty] -
7	numberStack.push(7)	7, -4	-
*	operatorStack.push(*)	7, -4	*, -
8	numberStack.push(8)	8, 7, -4	*, -
/	collapseStack [7 * 8] numberStack.push(/)	56, -4 56, -4	- /, -
2	numberStack.push(2)	2, 56, -4	/, -
+	collapseStack [56 / 2] collapseStack [-4 - 28] operatorStack.push(+)	28, -4 -32 -32	- [empty] +
5	numberStack.push(5)	5, -32	+
	collapseStack [-32 + 5]	-27	[empty]
	return -27		

The code below implements this algorithm.

```

1  public enum Operator {
2      ADD, SUBTRACT, MULTIPLY, DIVIDE, BLANK
3  }
4
5  double compute(String sequence) {
6      Stack<Double> numberStack = new Stack<Double>();
7      Stack<Operator> operatorStack = new Stack<Operator>();
8
9      for (int i = 0; i < sequence.length(); i++) {
10         try {
11             /* Get number and push. */
12             int value = parseNextNumber(sequence, i);
13             numberStack.push((double) value);
14
15             /* Move to the operator. */
16             i += Integer.toString(value).length();
17             if (i >= sequence.length()) {
18                 break;
19             }
20
21             /* Get operator, collapse top as needed, push operator. */
22             Operator op = parseNextOperator(sequence, i);
23             collapseTop(op, numberStack, operatorStack);
24             operatorStack.push(op);
25         } catch (NumberFormatException ex) {
26             return Integer.MIN_VALUE;
27         }
28     }
29
30     /* Do final collapse. */
31     collapseTop(Operator.BLANK, numberStack, operatorStack);
32     if (numberStack.size() == 1 && operatorStack.size() == 0) {
33         return numberStack.pop();
34     }
35     return 0;
36 }
37
38 /* Collapse top until priority(futureTop) > priority(top). Collapsing means to pop
39 * the top 2 numbers and apply the operator popped from the top of the operator
40 * stack, and then push that onto the numbers stack.*/
41 void collapseTop(Operator futureTop, Stack<Double> numberStack,
42                  Stack<Operator> operatorStack) {
43     while (operatorStack.size() >= 1 && numberStack.size() >= 2) {
44         if (priorityOfOperator(futureTop) <=
45             priorityOfOperator(operatorStack.peek())) {
46             double second = numberStack.pop();
47             double first = numberStack.pop();
48             Operator op = operatorStack.pop();
49             double collapsed = applyOp(first, op, second);
50             numberStack.push(collapsed);
51         } else {
52             break;
53         }
54     }

```

```

55 }
56
57 /* Return priority of operator. Mapped so that:
58 *      addition == subtraction < multiplication == division. */
59 int priorityOfOperator(Operator op) {
60     switch (op) {
61         case ADD: return 1;
62         case SUBTRACT: return 1;
63         case MULTIPLY: return 2;
64         case DIVIDE: return 2;
65         case BLANK: return 0;
66     }
67     return 0;
68 }
69
70 /* Apply operator: left [op] right. */
71 double applyOp(double left, Operator op, double right) {
72     if (op == Operator.ADD) return left + right;
73     else if (op == Operator.SUBTRACT) return left - right;
74     else if (op == Operator.MULTIPLY) return left * right;
75     else if (op == Operator.DIVIDE) return left / right;
76     else return right;
77 }
78
79 /* Return the number that starts at offset. */
80 int parseNextNumber(String seq, int offset) {
81     StringBuilder sb = new StringBuilder();
82     while (offset < seq.length() && Character.isDigit(seq.charAt(offset))) {
83         sb.append(seq.charAt(offset));
84         offset++;
85     }
86     return Integer.parseInt(sb.toString());
87 }
88
89 /* Return the operator that occurs as offset. */
90 Operator parseNextOperator(String sequence, int offset) {
91     if (offset < sequence.length()) {
92         char op = sequence.charAt(offset);
93         switch(op) {
94             case '+': return Operator.ADD;
95             case '-': return Operator.SUBTRACT;
96             case '*': return Operator.MULTIPLY;
97             case '/': return Operator.DIVIDE;
98         }
99     }
100    return Operator.BLANK;
101 }

```

This code also takes $O(N)$ time, where N is the length of the string.

This solution involves a lot of annoying string parsing code. Remember that getting all these details out is not that important in an interview. In fact, your interviewer might even let you assume the expression is passed in pre-parsed into some sort of data structure.

Focus on modularizing your code from the beginning and “farming out” tedious or less interesting parts of the code to other functions. You want to focus on getting the core `compute` function working. The rest of the details can wait!

17

Solutions to Hard

17.1 Add Without Plus: Write a function that adds two numbers. You should not use + or any arithmetic operators.

pg 186

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 deeply 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 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 with sum + carry
6 }
```

Alternatively, you can implement this iteratively.

```

1 int add(int a, int b) {
2     while (b != 0) {
3         int sum = a ^ b; // add without carrying
4         int carry = (a & b) << 1; // carry, but don't add
5         a = sum;
6         b = carry;
7     }
8     return a;
9 }
```

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.

- 17.2 Shuffle:** 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 186

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 $\text{array}[i]$ 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.

17.3 Random Set: 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 186

SOLUTION

Like the prior problem which was similar, (problem 17.2 on page 531), 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 + 1 > 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 subset to be the first m elements in original . Then, we iterate through the array, starting at element m , inserting $\text{array}[i]$ into the subset at (random) position $k < m$.

```
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.

17.4 Missing Number: An array A contains all the integers from 0 to n, except for one number which is missing. In this problem, we cannot access an entire integer in A with a single operation. The elements of A are represented in binary, and the only operation we can use to access them is “fetch the jth bit of A[i],” which takes constant time. Write code to find the missing integer. Can you do it in O(n) time?

pg 186

SOLUTION

You may have seen a very similar sounding problem: Given a list of numbers from 0 to n, with exactly one number removed, find the missing number. This problem can be solved by simply adding the list of numbers and comparing it to the actual sum of 0 through n, which is $\frac{n(n+1)}{2}$. The difference will be the missing number.

We could solve this by computing the value of each number, based on its binary representation, and calculating the sum.

The runtime of this solution is $n * \text{length}(n)$, when length is the number of bits in n. Note that $\text{length}(n) = \log_2(n)$. So, the runtime is actually $O(n \log(n))$. Not quite good enough!

So how else can we approach it?

We can actually use a similar approach, but leverage the bit values more directly.

Picture a list of binary numbers (the ----- indicates the value that was removed):

00000	00100	01000	01100
00001	00101	01001	01101
00010	00110	01010	
-----	00111	01011	

Removing the number above creates an imbalance of 1s and 0s in the least significant bit, which we'll call LSB. In a list of numbers from 0 to n, we would expect there to be the same number of 0s as 1s (if n is odd), or an additional 0 if n is even. That is:

```

if n % 2 == 1 then count(0s) = count(1s)
if n % 2 == 0 then count(0s) = 1 + count(1s)

```

Note that this means that count(0s) is always greater than or equal to count(1s).

When we remove a value v from the list, we'll know immediately if v is even or odd just by looking at the least significant bits of all the other values in the list.

	$n \% 2 == 0$ $\text{count}(0s) = 1 + \text{count}(1s)$	$n \% 2 == 1$ $\text{count}(0s) = \text{count}(1s)$
$v \% 2 == 0$ $\text{LSB}_1(v) = 0$	a 0 is removed. $\text{count}(0s) = \text{count}(1s)$	a 0 is removed. $\text{count}(0s) < \text{count}(1s)$
$v \% 2 == 1$ $\text{LSB}_1(v) = 1$	a 1 is removed. $\text{count}(0s) > \text{count}(1s)$	a 1 is removed. $\text{count}(0s) > \text{count}(1s)$

So, if $\text{count}(0s) \leq \text{count}(1s)$, then v is even. If $\text{count}(0s) > \text{count}(1s)$, then v is odd.

We can now remove all the evens and focus on the odds, or remove all the odds and focus on the evens.

Okay, but how do we figure out what the next bit in v is? If v were contained in our (now smaller) list, then we should expect to find the following (where count_2 indicates the number of 0s or 1s in the second least significant bit):

$$\text{count}_2(0s) = \text{count}_2(1s) \quad \text{OR} \quad \text{count}_2(0s) = 1 + \text{count}_2(1s)$$

As in the earlier example, we can deduce the value of the second least significant bit (LSB_2) of v .

	$\text{count}_2(0s) = 1 + \text{count}_2(1s)$	$\text{count}_2(0s) = \text{count}_2(1s)$
$\text{LSB}_2(v) == 0$	a 0 is removed. $\text{count}_2(0s) = \text{count}_2(1s)$	a 0 is removed. $\text{count}_2(0s) < \text{count}_2(1s)$
$\text{LSB}_2(v) == 1$	a 1 is removed. $\text{count}_2(0s) > \text{count}_2(1s)$	a 1 is removed. $\text{count}_2(0s) > \text{count}_2(1s)$

Again, we have the same conclusion:

- If $\text{count}_2(0s) \leq \text{count}_2(1s)$, then $\text{LSB}_2(v) = 0$.
- If $\text{count}_2(0s) > \text{count}_2(1s)$, then $\text{LSB}_2(v) = 1$.

We can repeat this process for each bit. On each iteration, we count the number of 0s and 1s in bit i to check if $\text{LSB}_i(v)$ is 0 or 1. Then, we discard the numbers where $\text{LSB}_i(x) \neq \text{LSB}_i(v)$. That is, if v is even, we discard the odd numbers, and so on.

By the end of this process, we will have computed all bits in v . In each successive iteration, we look at n , then $n / 2$, then $n / 4$, and so on, bits. This results in a runtime of $O(N)$.

If it helps, we can also move through this more visually. In the first iteration, we start with all the numbers:

00000	00100	01000	01100
00001	00101	01001	01101
00010	00110	01010	
-----	00111	01011	

Since $\text{count}_1(0s) > \text{count}_1(1s)$, we know that $\text{LSB}_1(v) = 1$. Now, discard all numbers x where $\text{LSB}_1(x) \neq \text{LSB}_1(v)$.

00000	00100	01000	01100
00001	00101	01001	01101
00010	00110	01010	
-----	00111	01011	

Now, $\text{count}_2(0s) > \text{count}_2(1s)$, so we know that $\text{LSB}_2(v) = 1$. Now, discard all numbers x where $\text{LSB}_2(x) \neq \text{LSB}_2(v)$.

00000	00100	01000	01100
00001	00101	01001	01101
00010	00110	01010	
-----	00111	01011	

This time, $\text{count}_3(0s) \leq \text{count}_3(1s)$, we know that $\text{LSB}_3(v) = 0$. Now, discard all numbers x where $\text{LSB}_3(x) \neq \text{LSB}_3(v)$.

00000	00100	01000	01100
00001	00101	01001	01101
00010	00110	01010	
-----	00111	01011	

We're down to just one number. In this case, $\text{count}_4(0s) \leq \text{count}_4(1s)$, so $\text{LSB}_4(v) = 0$.

When we discard all numbers where $\text{LSB}_4(x) \neq 0$, we'll wind up with an empty list. Once the list is empty, then $\text{count}_1(0s) \leq \text{count}_1(1s)$, so $\text{LSB}_1(v) = 0$. In other words, once we have an empty list, we can fill in the rest of the bits of v with 0.

This process will compute that, for the example above, $v = 00011$.

The code below implements this algorithm. We've implemented the discarding aspect by partitioning the array by bit value as we go.

```

1 int findMissing(ArrayList<BitInteger> array) {
2     /* Start from the least significant bit, and work our way up */
3     return findMissing(array, 0);
4 }
5
6 int findMissing(ArrayList<BitInteger> input, int column) {
7     if (column >= BitInteger.INTEGER_SIZE) { // We're done!
8         return 0;
9     }
10    ArrayList<BitInteger> oneBits = new ArrayList<BitInteger>(input.size()/2);
11    ArrayList<BitInteger> zeroBits = new ArrayList<BitInteger>(input.size()/2);
12
13    for (BitInteger t : input) {
14        if (t.fetch(column) == 0) {
15            zeroBits.add(t);
16        } else {
17            oneBits.add(t);
18        }
19    }
20    if (zeroBits.size() <= oneBits.size()) {
21        int v = findMissing(zeroBits, column + 1);
22        return (v << 1) | 0;
23    } else {
24        int v = findMissing(oneBits, column + 1);
25        return (v << 1) | 1;
26    }
27 }
```

In lines 24 and 27, we recursively calculate the other bits of v . Then, we insert either a 0 or 1, depending on whether or not $\text{count}_1(0s) \leq \text{count}_1(1s)$.

17.5 Letters and Numbers: Given an array filled with letters and numbers, find the longest subarray with an equal number of letters and numbers.

pg 186

SOLUTION

In the introduction, we discussed the importance of creating a really good, general-purpose example. That's absolutely true. It's also important, though, to understand what matters.

In this case, we just want an equal number of letters and numbers. All letters are treated identically and all numbers are treated identically. Therefore, we can use an example with a single letter and a single number—or, for that matter, As and Bs, 0s and 1s, or Thing1s and Thing2s.

With that said, let's start with an example:

```
[A, B, A, A, A, B, B, A, B, A, A, B, B, A, A, A, A, A, A]
```

We're looking for the smallest subarray where $\text{count}(A, \text{ subarray }) = \text{count}(B, \text{ subarray })$.

Brute Force

Let's start with the obvious solution. Just go through all subarrays, count the number of As and Bs (or letters and numbers), and find the longest one that is equal.

We can make one small optimization to this. We can start with the longest subarray and, as soon as we find one which fits this equality condition, return it.

```
1 /* Return the largest subarray with equal number of 0s and 1s. Look at each
2  * subarray, starting from the longest. As soon as we find one that's equal, we
3  * return.
4 char[] findLongestSubarray(char[] array) {
5     for (int len = array.length; len > 1; len--) {
6         for (int i = 0; i <= array.length - len; i++) {
7             if (hasEqualLettersNumbers(array, i, i + len - 1)) {
8                 return extractSubarray(array, i, i + len - 1);
9             }
10        }
11    }
12    return null;
13 }
14
15 /* Check if subarray has equal number of letters and numbers. */
16 boolean hasEqualLettersNumbers(char[] array, int start, int end) {
17     int counter = 0;
18     for (int i = start; i <= end; i++) {
19         if (Character.isLetter(array[i])) {
20             counter++;
21         } else if (Character.isDigit(array[i])) {
22             counter--;
23         }
24     }
25     return counter == 0;
26 }
27
28 /* Return subarray of array between start and end (inclusive). */
29 char[] extractSubarray(char[] array, int start, int end) {
30     char[] subarray = new char[end - start + 1];
31     for (int i = start; i <= end; i++) {
32         subarray[i - start] = array[i];
```

```

33     }
34     return subarray;
35 }
```

Despite the one optimization we made, this algorithm is still $O(N^2)$, where N is the length of the array.

Optimal Solution

What we're trying to do is find a subarray where the count of letters equals the count of numbers. What if we just started from the beginning, counting the number of letters and numbers?

	a	a	a	a	1	1	a	1	1	a	a	1	a	a	1	a	a	a	a	
#a	1	2	3	4	4	4	5	5	5	6	7	7	8	9	9	10	11	12	13	14
#1	0	0	0	0	1	2	2	3	4	4	4	5	5	5	6	6	6	6	6	6

Certainly, whenever the number of letters equals the number of numbers, we can say that from index 0 to that index is an "equal" subarray.

That will only tell us equal subarrays that start at index 0. How can we identify all equal subarrays?

Let's picture this. Suppose we inserted an equal subarray (like a11a1a) after an array like a1aaa1. How would that impact the counts?

	a	1	a	a	a	1		a	1	1	a	1	a	1				
#a	1	1	2	3	4	4		5	5	5	6	6	7					
#1	0	1	1	1	1	2		2	3	4	4	5	5					

Study the numbers before the subarray (4, 2) and the end (7, 5). You might notice that, while the values aren't the same, the differences are: $4 - 2 = 7 - 5$. This makes sense. Since they've added the same number of letters and numbers, they should maintain the same difference.

Observe that when the difference is the same, the subarray starts one after the initial matching index and continues through the final matching index. This explains line 10 in the code below.

Let's update the earlier array with the differences.

	a	a	a	a	1	1	a	1	1	a	a	1	a	a	1	a	a	a	a	
#a	1	2	3	4	4	4	5	5	5	6	7	7	8	9	9	10	11	12	13	14
#1	0	0	0	0	1	2	2	3	4	4	4	5	5	5	6	6	6	6	6	6
-	1	2	3	4	3	2	3	2	1	2	3	2	3	4	3	4	5	6	7	8

Whenever we return the same difference, then we know we have found an equal subarray. To find the biggest subarray, we just have to find the two indices farthest apart with the same value.

To do so, we use a hash table to store the first time we see a particular difference. Then, each time we see the same difference, we see if this subarray (from first occurrence of this index to current index) is bigger than the current max. If so, we update the max.

```

1 char[] findLongestSubarray(char[] array) {
2     /* Compute deltas between count of numbers and count of letters. */
3     int[] deltas = computeDeltaArray(array);
4
5     /* Find pair in deltas with matching values and largest span. */
6     int[] match = findLongestMatch(deltas);
7
8     /* Return the subarray. Note that it starts one *after* the initial occurrence of
9      * this delta. */
10    return extract(array, match[0] + 1, match[1]);
11 }
12 }
```

```
13 /* Compute the difference between the number of letters and numbers between the
14 * beginning of the array and each index. */
15 int[] computeDeltaArray(char[] array) {
16     int[] deltas = new int[array.length];
17     int delta = 0;
18     for (int i = 0; i < array.length; i++) {
19         if (Character.isLetter(array[i])) {
20             delta++;
21         } else if (Character.isDigit(array[i])) {
22             delta--;
23         }
24         deltas[i] = delta;
25     }
26     return deltas;
27 }
28
29 /* Find the matching pair of values in the deltas array with the largest
30 * difference in indices. */
31 int[] findLongestMatch(int[] deltas) {
32     HashMap<Integer, Integer> map = new HashMap<Integer, Integer>();
33     map.put(0, -1);
34     int[] max = new int[2];
35     for (int i = 0; i < deltas.length; i++) {
36         if (!map.containsKey(deltas[i])) {
37             map.put(deltas[i], i);
38         } else {
39             int match = map.get(deltas[i]);
40             int distance = i - match;
41             int longest = max[1] - max[0];
42             if (distance > longest) {
43                 max[1] = i;
44                 max[0] = match;
45             }
46         }
47     }
48     return max;
49 }
50
51 char[] extract(char[] array, int start, int end) { /* same */ }
```

This solution takes $O(N)$ time, where N is size of the array.

17.6 Count of 2s: Write a method to count the number of 2s between 0 and n.

pg 186

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 10's digit is a 2 exactly $\frac{1}{10}$ th of the time. However, between 1 and 37, the 10's digit is a 2 much more than $1/10$ 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 `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 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, ..., 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 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 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.

- 17.7 Baby Names:** Each year, the government releases a list of the 10,000 most common baby names and their frequencies (the number of babies with that name). The only problem with this is that some names have multiple spellings. For example, "John" and "Jon" are essentially the same name but would be listed separately in the list. Given two lists, one of names/frequencies and the other of pairs of equivalent names, write an algorithm to print a new list of the true frequency of each name. Note that if John and Jon are synonyms, and Jon and Johnny are synonyms, then John and Johnny are synonyms. (It is both transitive and symmetric.) In the final list, any name can be used as the "real" name.

EXAMPLE

Input:

Names: John (15), Jon (12), Chris (13), Kris (4), Christopher (19)

Synonyms: (Jon, John), (John, Johnny), (Chris, Kris), (Chris, Christopher)

Output: John (27), Kris (36)

pg 187

SOLUTION

Let's start off with a good example. We want an example with some names with multiple synonyms and some with none. Additionally, we want the synonym list to be diverse in which name is on the left side and which is on the right. For example, we wouldn't want Johnny to always be the name on the left side as we're creating the group of (John, Jonathan, Jon, and Johnny).

This list should work fairly well.

Name	Count
John	10
Jon	3
Davis	2
Kari	3
Johnny	11
Carlton	8
Carleton	2
Jonathan	9
Carrie	5

Name	Alternate
Jonathan	John
Jon	Johnny
Johnny	John
Kari	Carrie
Carleton	Carlton

The final list should be something like: John (33), Kari (8), Davis(2), Carleton (10).

Solution #1

Let's assume our baby names list is given to us as a hash table. (If not, it's easy enough to build one.)

We can start reading pairs in from the synonyms list. As we read the pair (Jonathan, John), we can merge the counts for Jonathan and John together. We'll need to remember, though, that we saw this pair, because, in the future, we could discover that Jonathan is equivalent to something else.

We can use a hash table (L1) that maps from a name to its "true" name. We'll also need to know, given a "true" name, all the names equivalent to it. This will be stored in a hash table L2. Note that L2 acts as a reverse lookup of L1.

READ (Jonathan, John)

```
L1.ADD Jonathan -> John
L2.ADD John -> Jonathan
READ (Jon, Johnny)
    L1.ADD Jon -> Johnny
    L2.ADD Johnny -> Jon
READ (Johnny, John)
    L1.ADD Johnny -> John
    L1.UPDATE Jon -> John
    L2.UPDATE John -> Jonathan, Johnny, Jon
```

If we later find that John is equivalent to, say, Jonny, we'll need to look up the names in L1 and L2 and merge together all the names that are equivalent to them.

This will work, but it's unnecessarily complicated to keep track of these two lists.

Instead, we can think of these names as "equivalence classes." When we find a pair (Jonathan, John), we put these in the same set (or equivalence classes). Each name maps to its equivalence class. All items in the set map to the same instance of the set.

If we need to merge two sets, then we copy one set into the other and update the hash table to point to the new set.

```
READ (Jonathan, John)
CREATE Set1 = Jonathan, John
    L1.ADD Jonathan -> Set1
    L1.ADD John -> Set1
READ (Jon, Johnny)
    CREATE Set2 = Jon, Johnny
    L1.ADD Jon -> Set2
    L1.ADD Johnny -> Set2
READ (Johnny, John)
    COPY Set2 into Set1.
        Set1 = Jonathan, John, Jon, Johnny
    L1.UPDATE Jon -> Set1
    L1.UPDATE Johnny -> Set1
```

In the last step above, we iterated through all items in Set2 and updated the reference to point to Set1. As we do this, we keep track of the total frequency of names.

```
1  HashMap<String, Integer> trulyMostPopular(HashMap<String, Integer> names,
2                                              String[][] synonyms) {
3      /* Parse list and initialize equivalence classes.*/
4      HashMap<String, NameSet> groups = constructGroups(names);
5
6      /* Merge equivalence classes together. */
7      mergeClasses(groups, synonyms);
8
9      /* Convert back to hash map. */
10     return convertToMap(groups);
11 }
12
13 /* This is the core of the algorithm. Read through each pair. Merge their
14 * equivalence classes and update the mapping of the secondary class to point to
15 * the first set.*/
16 void mergeClasses(HashMap<String, NameSet> groups, String[][] synonyms) {
17     for (String[] entry : synonyms) {
18         String name1 = entry[0];
19         String name2 = entry[1];
20         NameSet set1 = groups.get(name1);
```

```

21     NameSet set2 = groups.get(name2);
22     if (set1 != set2) {
23         /* Always merge the smaller set into the bigger one. */
24         NameSet smaller = set2.size() < set1.size() ? set2 : set1;
25         NameSet bigger = set2.size() < set1.size() ? set1 : set2;
26
27         /* Merge lists */
28         Set<String> otherNames = smaller.getNames();
29         int frequency = smaller.getFrequency();
30         bigger.copyNamesWithFrequency(otherNames, frequency);
31
32         /* Update mapping */
33         for (String name : otherNames) {
34             groups.put(name, bigger);
35         }
36     }
37 }
38 }
39
40 /* Read through (name, frequency) pairs and initialize a mapping of names to
41 * NameSets (equivalence classes).*/
42 HashMap<String, NameSet> constructGroups(HashMap<String, Integer> names) {
43     HashMap<String, NameSet> groups = new HashMap<String, NameSet>();
44     for (Entry<String, Integer> entry : names.entrySet()) {
45         String name = entry.getKey();
46         int frequency = entry.getValue();
47         NameSet group = new NameSet(name, frequency);
48         groups.put(name, group);
49     }
50     return groups;
51 }
52
53 HashMap<String, Integer> convertToMap(HashMap<String, NameSet> groups) {
54     HashMap<String, Integer> list = new HashMap<String, Integer>();
55     for (NameSet group : groups.values()) {
56         list.put(group.getRootName(), group.getFrequency());
57     }
58     return list;
59 }
60
61 public class NameSet {
62     private Set<String> names = new HashSet<String>();
63     private int frequency = 0;
64     private String rootName;
65
66     public NameSet(String name, int freq) {
67         names.add(name);
68         frequency = freq;
69         rootName = name;
70     }
71
72     public void copyNamesWithFrequency(Set<String> more, int freq) {
73         names.addAll(more);
74         frequency += freq;
75     }
76 }
```

```
77     public Set<String> getNames() { return names; }
78     public String getRootName() { return rootName; }
79     public int getFrequency() { return frequency; }
80     public int size() { return names.size(); }
81 }
```

The runtime of the algorithm is a bit tricky to figure out. One way to think about it is to think about what the worst case is.

For this algorithm, the worst case is where all names are equivalent—and we have to constantly merge sets together. Also, for the worst case, the merging should come in the worst possible way: repeated pairwise merging of sets. Each merging requires copying the set's elements into an existing set and updating the pointers from those items. It's slowest when the sets are larger.

If you notice the parallel with merge sort (where you have to merge single-element arrays into two-element arrays, and then two-element arrays into four-element arrays, until finally having a full array), you might guess it's $O(N \log N)$. That is correct.

If you don't notice that parallel, here's another way to think about it.

Imagine we had the names (a, b, c, d, . . . , z). In our worst case, we'd first pair up the items into equivalence classes: (a, b), (c, d), (e, f), . . . , (y, z). Then, we'd merge pairs of those: (a, b, c, d), (e, f, g, h), . . . , (w, x, y, z). We'd continue doing this until we wind up with just one class.

At each "sweep" through the list where we merge sets together, half of the items get moved into a new set. This takes $O(N)$ work per sweep. (There are fewer sets to merge, but each set has grown larger.)

How many sweeps do we do? At each sweep, we have half as many sets as we did before. Therefore, we do $O(\log N)$ sweeps.

Since we're doing $O(\log N)$ sweeps and $O(N)$ work per sweep, the total runtime is $O(N \log N)$.

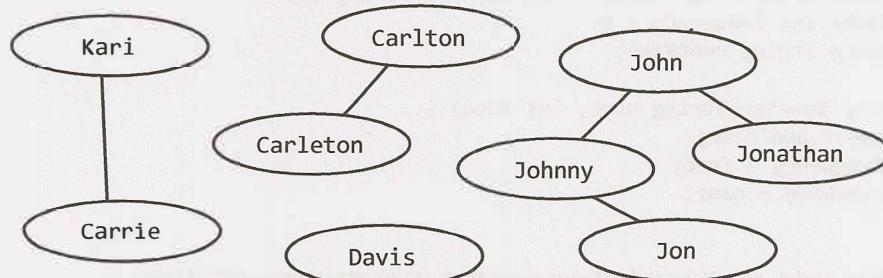
This is pretty good, but let's see if we can make it even faster.

Optimized Solution

To optimize the old solution, we should think about what exactly makes it slow. Essentially, it's the merging and updating of pointers.

So what if we just didn't do that? What if we marked that there was an equivalence relationship between two names, but didn't actually do anything with the information yet?

In this case, we'd be building essentially a graph.



Now what? Visually, it seems easy enough. Each component is an equivalent set of names. We just need to group the names by their component, sum up their frequencies, and return a list with one arbitrarily chosen name from each group.

In practice, how does this work? We could pick a name and do a depth-first (or breadth-first) search to sum the frequencies of all the names in one component. We would have to make sure that we hit each component exactly once. That's easy enough to achieve: mark a node as `visited` after it's discovered in the graph search, and only start the search for nodes where `visited` is false.

```

1  HashMap<String, Integer> trulyMostPopular(HashMap<String, Integer> names,
2                                              String[][] synonyms) {
3      /* Create data. */
4      Graph graph = constructGraph(names);
5      connectEdges(graph, synonyms);
6
7      /* Find components. */
8      HashMap<String, Integer> rootNames = getTrueFrequencies(graph);
9      return rootNames;
10 }
11
12 /* Add all names to graph as nodes. */
13 Graph constructGraph(HashMap<String, Integer> names) {
14     Graph graph = new Graph();
15     for (Entry<String, Integer> entry : names.entrySet()) {
16         String name = entry.getKey();
17         int frequency = entry.getValue();
18         graph.createNode(name, frequency);
19     }
20     return graph;
21 }
22
23 /* Connect synonymous spellings. */
24 void connectEdges(Graph graph, String[][] synonyms) {
25     for (String[] entry : synonyms) {
26         String name1 = entry[0];
27         String name2 = entry[1];
28         graph.addEdge(name1, name2);
29     }
30 }
31
32 /* Do DFS of each component. If a node has been visited before, then its component
33 * has already been computed. */
34 HashMap<String, Integer> getTrueFrequencies(Graph graph) {
35     HashMap<String, Integer> rootNames = new HashMap<String, Integer>();
36     for (GraphNode node : graph.getNodes()) {
37         if (!node.isVisited()) { // Already visited this component
38             int frequency = getComponentFrequency(node);
39             String name = node.getName();
40             rootNames.put(name, frequency);
41         }
42     }
43     return rootNames;
44 }
45
46 /* Do depth-first search to find the total frequency of this component, and mark
47 * each node as visited.*/
48 int getComponentFrequency(GraphNode node) {
49     if (node.isVisited()) return 0; // Already visited
50
51     node.setIsVisited(true);
52     int sum = node.getFrequency();

```

```
53     for (GraphNode child : node.getNeighbors()) {  
54         sum += getComponentFrequency(child);  
55     }  
56     return sum;  
57 }  
58  
59 /* Code for GraphNode and Graph is fairly self-explanatory, but can be found in  
60 * the downloadable code solutions.*/
```

To analyze the efficiency, we can think about the efficiency of each part of the algorithm.

- Reading in the data is linear with respect to the size of the data, so it takes $O(B + P)$ time, where B is the number of baby names and P is the number of pairs of synonyms. This is because we only do a constant amount of work per piece of input data.
- To compute the frequencies, each edge gets “touched” exactly once across all of the graph searches and each node gets touched exactly once to check if it’s been visited. The time of this part is $O(B + P)$.

Therefore, the total time of the algorithm is $O(B + P)$. We know we cannot do better than this since we must at least read in the $B + P$ pieces of data.

- 17.8 Circus Tower:** 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 187

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.

One thing we might first try is sorting the items on an attribute. This is useful actually, but it won’t get us all the way there.

By sorting the items by height, we have a relative order the items must appear in. We still need to find the longest increasing subsequence of weight though.

Solution 1: Recursive

One approach is to essentially try all possibilities. After sorting by height, we iterate through the array. At each element, we branch into two choices: add this element to the subsequence (if it’s valid) or do not.

```
1  ArrayList<HtWt> longestIncreasingSeq(ArrayList<HtWt> items) {  
2      Collections.sort(items);  
3      return bestSeqAtIndex(items, new ArrayList<HtWt>(), 0);  
4  }  
5  
6  ArrayList<HtWt> bestSeqAtIndex(ArrayList<HtWt> array, ArrayList<HtWt> sequence,  
7                                     int index) {  
8      if (index >= array.size()) return sequence;  
9  
10     HtWt value = array.get(index);  
11 }
```

```

12 ArrayList<HtWt> bestWith = null;
13 if (canAppend(sequence, value)) {
14     ArrayList<HtWt> sequenceWith = (ArrayList<HtWt>) sequence.clone();
15     sequenceWith.add(value);
16     bestWith = bestSeqAtIndex(array, sequenceWith, index + 1);
17 }
18
19 ArrayList<HtWt> bestWithout = bestSeqAtIndex(array, sequence, index + 1);
20
21 if (bestWith == null || bestWithout.size() > bestWith.size()) {
22     return bestWithout;
23 } else {
24     return bestWith;
25 }
26 }
27
28 boolean canAppend(ArrayList<HtWt> solution, HtWt value) {
29     if (solution == null) return false;
30     if (solution.size() == 0) return true;
31
32     HtWt last = solution.get(solution.size() - 1);
33     return last.isBefore(value);
34 }
35
36 ArrayList<HtWt> max(ArrayList<HtWt> seq1, ArrayList<HtWt> seq2) {
37     if (seq1 == null) {
38         return seq2;
39     } else if (seq2 == null) {
40         return seq1;
41     }
42     return seq1.size() > seq2.size() ? seq1 : seq2;
43 }
44
45 public class HtWt implements Comparable<HtWt> {
46     private int height;
47     private int weight;
48     public HtWt(int h, int w) { height = h; weight = w; }
49
50     public int compareTo(HtWt second) {
51         if (this.height != second.height) {
52             return ((Integer)this.height).compareTo(second.height);
53         } else {
54             return ((Integer)this.weight).compareTo(second.weight);
55         }
56     }
57
58     /* Returns true if "this" should be lined up before "other". Note that it's
59      * possible that this.isBefore(other) and other.isBefore(this) are both false.
60      * This is different from the compareTo method, where if a < b then b > a. */
61     public boolean isBefore(HtWt other) {
62         if (height < other.height && weight < other.weight) {
63             return true;
64         } else {
65             return false;
66         }
67     }

```

68 }

This algorithm will take $O(2^n)$ time. We can optimize it using memoization (that is, caching the best sequences).

There's a cleaner way to do this though.

Solution #2: Iterative

Imagine we had the longest subsequence that terminates with each element, $A[0]$ through $A[3]$. Could we use this to find the longest subsequence that terminates with $A[4]$?

```
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
```

Sure. We just append $A[4]$ on to the longest subsequence that it can be appended to.

This is now fairly straightforward to implement.

```
1  ArrayList<HtWt> longestIncreasingSeq(ArrayList<HtWt> array) {
2      Collections.sort(array);
3
4      ArrayList<ArrayList<HtWt>> solutions = new ArrayList<ArrayList<HtWt>>();
5      ArrayList<HtWt> bestSequence = null;
6
7      /* Find the longest subsequence that terminates with each element. Track the
8         * longest overall subsequence as we go. */
9      for (int i = 0; i < array.size(); i++) {
10          ArrayList<HtWt> longestAtIndex = bestSeqAtIndex(array, solutions, i);
11          solutions.add(i, longestAtIndex);
12          bestSequence = max(bestSequence, longestAtIndex);
13      }
14
15      return bestSequence;
16  }
17
18  /* Find the longest subsequence which terminates with this element. */
19  ArrayList<HtWt> bestSeqAtIndex(ArrayList<HtWt> array,
20      ArrayList<ArrayList<HtWt>> solutions, int index) {
21      HtWt value = array.get(index);
22
23      ArrayList<HtWt> bestSequence = new ArrayList<HtWt>();
24
25      /* Find the longest subsequence that we can append this element to. */
26      for (int i = 0; i < index; i++) {
27          ArrayList<HtWt> solution = solutions.get(i);
28          if (canAppend(solution, value)) {
29              bestSequence = max(solution, bestSequence);
30          }
31      }
32
33      /* Append element. */
34      ArrayList<HtWt> best = (ArrayList<HtWt>) bestSequence.clone();
35      best.add(value);
36  }
```

```

37     return best;
38 }

```

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.

- 17.9 Kth Multiple:** Design an algorithm to find the kth number such that the only prime factors are 3, 5, and 7. Note that 3, 5, and 7 do not have to be factors, but it should not have any other prime factors. For example, the first several multiples would be (in order) 1, 3, 5, 7, 9, 15, 21.

pg 187

SOLUTION

Let's first understand what this problem is asking for. It's asking for the kth smallest number that is in the form $3^a * 5^b * 7^c$. Let's start with a brute force way of finding this.

Brute Force

We know that biggest this kth number could be is $3^k * 5^k * 7^k$. So, the “stupid” way of doing this is to compute $3^a * 5^b * 7^c$ for all values of a, b, and c between 0 and k. We can throw them all into a list, sort the list, and then pick the kth smallest value.

```

1 int getKthMagicNumber(int k) {
2     ArrayList<Integer> possibilities = allPossibleKFactors(k);
3     Collections.sort(possibilities);
4     return possibilities.get(k);
5 }
6
7 ArrayList<Integer> allPossibleKFactors(int k) {
8     ArrayList<Integer> values = new ArrayList<Integer>();
9     for (int a = 0; a <= k; a++) { // loop 3
10         int powA = (int) Math.pow(3, a);
11         for (int b = 0; b <= k; b++) { // loop 5
12             int powB = (int) Math.pow(5, b);
13             for (int c = 0; c <= k; c++) { // loop 7
14                 int powC = (int) Math.pow(7, c);
15                 int value = powA * powB * powC;
16
17                 /* Check for overflow. */
18                 if (value < 0 || powA == Integer.MAX_VALUE ||
19                     powB == Integer.MAX_VALUE ||
20                     powC == Integer.MAX_VALUE) {
21                     value = Integer.MAX_VALUE;
22                 }
23                 values.add(value);
24             }
25         }
26     }
27     return values;
28 }

```

What is the runtime of this approach? We have nested for loops, each of which runs for k iterations. The runtime of the `allPossibleKFactors` is $O(k^3)$. Then, we sort the k^3 results in $O(k^3 \log(k^3))$ time (which is equivalent to $O(k^3 \log k)$). This gives us a runtime of $O(k^3 \log k)$.

There are a number of optimizations you could make to this (and better ways of handling the integer overflow), but honestly this algorithm is fairly slow. We should instead focus on reworking the algorithm.

Improved

Let's picture what our results will look like.

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	$3*21$	$3^2 * 5^0 * 7^1$

The question is: what is the next value in the list? The next value will be one of these:

- $3 * (\text{some previous number in list})$
- $5 * (\text{some previous number in list})$
- $7 * (\text{some previous number in list})$

If this doesn't immediately jump out at you, think about it this way: whatever the next value (let's call it nv) is, divide it by 3. Will that number have already appeared? As long as nv has factors of 3 in it, yes. The same can be said for dividing it by 5 and 7.

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_i will eventually be used later in the list in the following forms:

- $3 * A_i$
- $5 * A_i$
- $7 * A_i$

We can use this thought to plan in advance. Each time we add a number A_i to the list, we hold on to the values $3A_i$, $5A_i$, and $7A_i$ 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 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 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 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.

Optimal Algorithm

To generate a new element A_i , 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:

$$\begin{aligned} Q3_6 &= \{3A_4\} \\ Q5_6 &= \{5A_2, 5A_4, 5A_5\} \\ Q7_6 &= \{7A_1, 7A_2, 7A_3, 7A_4, 7A_5\} \end{aligned}$$

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 $7 * 7 * \text{suffix}$, 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 $x \times 3$, $x \times 5$ and $x \times 7$ to Q3, Q5, and Q7. Remove x from Q3.

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

$Q7 \rightarrow$ only append $x \times 7$ to Q7. Remove x from Q7.

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

The code below implements this algorithm.

```

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

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.

17.10 Majority Element: A majority element is an element that makes up more than half of the items in an array. Given a positive integers array, find the majority element. If there is no majority element, return -1. Do this in $O(N)$ time and $O(1)$ space.

Input: 1 2 5 9 5 9 5 5 5

Output: 5

pg 187

SOLUTION

Let's start off with an example:

3 1 7 1 3 7 3 7 1 7 7

One thing we can notice here is that if the majority element (in this case 7) appears less often in the beginning, it must appear much more often toward the end. That's a good observation to make.

This interview question specifically requires us to do this in $O(N)$ time and $O(1)$ space. Nonetheless, sometimes it can be useful to relax one of those requirements and develop an algorithm. Let's try relaxing the time requirement but staying firm on the $O(1)$ space requirement.

Solution #1 (Slow)

One simple way to do this is to just iterate through the array and check each element for whether it's the majority element. This takes $O(N^2)$ time and $O(1)$ space.

```
1 int findMajorityElement(int[] array) {
2     for (int x : array) {
3         if (validate(array, x)) {
4             return x;
5         }
6     }
7     return -1;
8 }
```



```
10 boolean validate(int[] array, int majority) {
11     int count = 0;
12     for (int n : array) {
13         if (n == majority) {
14             count++;
15         }
16     }
17
18     return count > array.length / 2;
19 }
```

This does not fit the time requirements of the problem, but it is potentially a starting point. We can think about optimizing this.

Solution #2 (Optimal)

Let's think about what that algorithm did on a particular example. Is there anything we can get rid of?

3	1	7	1	1	7	7	3	7	7	7
0	1	2	3	4	5	6	7	8	9	10

In the very first validation pass, we select 3 and validate it as the majority element. Several elements later, we've still counted just one 3 and several non-3 elements. Do we need to continue checking for 3?

On one hand, yes. 3 could redeem itself and be the majority element, if there are a bunch of 3s later in the array.

On the other hand, not really. If 3 does redeem itself, then we'll encounter those 3s later on, in a subsequent validation step. We could terminate this `validate(3)` step.

That logic is fine for the first element, but what about the next one? We would immediately terminate `validate(1)`, `validate(7)`, and so on.

Since the logic was okay for the first element, what if we treated all subsequent elements like they're the first element of some new subarray? This would mean that we start `validate(array[1])` at index 1, `validate(array[2])` at index 2, and so on.

What would this look like?

```

validate(3)
    sees 3 -> countYes = 1, countNo = 0
    sees 1 -> countYes = 1, countNo = 1
    TERMINATE. 3 is not majority thus far.

validate(1)
    sees 1 -> countYes = 0, countNo = 0
    sees 7 -> countYes = 1, countNo = 1
    TERMINATE. 1 is not majority thus far.

validate(7)
    sees 7 -> countYes = 1, countNo = 0
    sees 1 -> countYes = 1, countNo = 1
    TERMINATE. 7 is not majority thus far.

validate(1)
    sees 1 -> countYes = 1, countNo = 0
    sees 1 -> countYes = 2, countNo = 0
    sees 7 -> countYes = 2, countNo = 1
    sees 7 -> countYes = 2, countNo = 1
    TERMINATE. 1 is not majority thus far.

validate(1)
    sees 1 -> countYes = 1, countNo = 0
    sees 7 -> countYes = 1, countNo = 1
    TERMINATE. 1 is not majority thus far.

validate(7)
    sees 7 -> countYes = 1, countNo = 0
    sees 7 -> countYes = 2, countNo = 0
    sees 3 -> countYes = 2, countNo = 1
    sees 7 -> countYes = 3, countNo = 1
    sees 7 -> countYes = 4, countNo = 1
    sees 7 -> countYes = 5, countNo = 1
  
```

Do we know at this point that 7 is the majority element? Not necessarily. We have eliminated everything before that 7, and everything after it. But there could be no majority element. A quick `validate(7)` pass that starts from the beginning can confirm if 7 is actually the majority element. This validate step will be $O(N)$ time, which is also our Best Conceivable Runtime. Therefore, this final validate step won't impact our total runtime.

This is pretty good, but let's see if we can make this a bit faster. We should notice that some elements are being "inspected" repeatedly. Can we get rid of this?

Look at the first `validate(3)`. This fails after the subarray [3, 1], because 3 was not the majority element. But because `validate` fails the instant an element is not the majority element, it also means nothing else in that subarray was the majority element. By our earlier logic, we don't need to call `validate(1)`. We know that 1 did not appear more than half the time. If it is the majority element, it'll pop up later.

Let's try this again and see if it works out.

```
validate(3)
    sees 3 -> countYes = 1, countNo = 0
    sees 1 -> countYes = 1, countNo = 1
    TERMINATE. 3 is not majority thus far.

skip 1
validate(7)
    sees 7 -> countYes = 1, countNo = 0
    sees 1 -> countYes = 1, countNo = 1
    TERMINATE. 7 is not majority thus far.

skip 1
validate(1)
    sees 1 -> countYes = 1, countNo = 0
    sees 7 -> countYes = 1, countNo = 1
    TERMINATE. 1 is not majority thus far.

skip 7
validate(7)
    sees 7 -> countYes = 1, countNo = 0
    sees 3 -> countYes = 1, countNo = 1
    TERMINATE. 7 is not majority thus far.

skip 3
validate(7)
    sees 7 -> countYes = 1, countNo = 0
    sees 7 -> countYes = 2, countNo = 0
    sees 7 -> countYes = 3, countNo = 0
```

Good! We got the right answer. But did we just get lucky?

We should pause for a moment to think what this algorithm is doing.

1. We start off with [3] and we expand the subarray until 3 is no longer the majority element. We fail at [3, 1]. At the moment we fail, the subarray can have no majority element.
2. Then we go to [7] and expand until [7, 1]. Again, we terminate and nothing could be the majority element in that subarray.
3. We move to [1] and expand to [1, 7]. We terminate. Nothing there could be the majority element.
4. We go to [7] and expand to [7, 3]. We terminate. Nothing there could be the majority element.
5. We go to [7] and expand until the end of the array: [7, 7, 7]. We have found the majority element (and now we must validate that).

Each time we terminate the `validate` step, the subarray has no majority element. This means that there are at least as many non-7s as there are 7s. Although we're essentially removing this subarray from the original array, the majority element will still be found in the rest of the array—and will still have majority status. Therefore, at some point, we will discover the majority element.

Our algorithm can now be run in two passes: one to find the possible majority element and another to validate it. Rather than using two variables to count (`countYes` and `countNo`), we'll just use a single `count` variable that increments and decrements.

```
1 int findMajorityElement(int[] array) {
2     int candidate = getCandidate(array);
3     return validate(array, candidate) ? candidate : -1;
4 }
5
6 int getCandidate(int[] array) {
7     int majority = 0;
```

```

8     int count = 0;
9     for (int n : array) {
10         if (count == 0) { // No majority element in previous set.
11             majority = n;
12         }
13         if (n == majority) {
14             count++;
15         } else {
16             count--;
17         }
18     }
19     return majority;
20 }
21
22 boolean validate(int[] array, int majority) {
23     int count = 0;
24     for (int n : array) {
25         if (n == majority) {
26             count++;
27         }
28     }
29
30     return count > array.length / 2;
31 }
```

This algorithm runs in $O(N)$ time and $O(1)$ space.

- 17.11 Word Distance:** 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 187

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.

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 `location1` and `location2`. If the current locations are better than our best known location, we update the best locations.

The code below implements this algorithm.

```

1 LocationPair findClosest(String[] words, String word1, String word2) {
2     LocationPair best = new LocationPair(-1, -1);
3     LocationPair current = new LocationPair(-1, -1);
4     for (int i = 0; i < words.length; i++) {
5         String word = words[i];
6         if (word.equals(word1)) {
7             current.location1 = i;
8             best.updateWithMin(current);
9         } else if (word.equals(word2)) {
10            current.location2 = i;
11            best.updateWithMin(current); // If shorter, update values
12        }
13    }
```

```
14     return best;
15 }
16
17 public class LocationPair {
18     public int location1, location2;
19     public LocationPair(int first, int second) {
20         setLocations(first, second);
21     }
22
23     public void setLocations(int first, int second) {
24         this.location1 = first;
25         this.location2 = second;
26     }
27
28     public void setLocations(LocationPair loc) {
29         setLocations(loc.location1, loc.location2);
30     }
31
32     public int distance() {
33         return Math.abs(location1 - location2);
34     }
35
36     public boolean isValid() {
37         return location1 >= 0 && location2 >= 0;
38     }
39
40     public void updateWithMin(LocationPair loc) {
41         if (!isValid() || loc.distance() < distance()) {
42             setLocations(loc);
43         }
44     }
45 }
```

If we need to repeat the operation for other pairs of words, we can create a hash table that maps from each word to the locations where it occurs. We'll only need to read through the list of words once. After that point, we can do a very similar algorithm but just iterate through the locations directly.

Consider the following lists of locations.

```
listA: {1, 2, 9, 15, 25}
listB: {4, 10, 19}
```

Picture pointers pA and pB that point to the beginning of each list. Our goal is to make pA and pB point to values as close together as possible.

The first potential pair is (1, 4).

What is the next pair we can find? If we moved pB, then the distance would definitely get larger. If we moved pA, though, we might get a better pair. Let's do that.

The second potential pair is (2, 4). This is better than the previous pair, so let's record this as the best pair.

We move pA again and get (9, 4). This is worse than we had before.

Now, since the value at pA is bigger than the one at pB, we move pB. We get (9, 10).

Next we get (15, 10), then (15, 19), then (25, 19).

We can implement this algorithm as shown below.

```
1 LocationPair findClosest(String word1, String word2,
```

```

2                     HashMapList<String, Integer> locations) {
3     ArrayList<Integer> locations1 = locations.get(word1);
4     ArrayList<Integer> locations2 = locations.get(word2);
5     return findMinDistancePair(locations1, locations2);
6 }
7
8 LocationPair findMinDistancePair(ArrayList<Integer> array1,
9                                 ArrayList<Integer> array2) {
10    if (array1 == null || array2 == null || array1.size() == 0 ||
11        array2.size() == 0) {
12        return null;
13    }
14
15    int index1 = 0;
16    int index2 = 0;
17    LocationPair best = new LocationPair(array1.get(0), array2.get(0));
18    LocationPair current = new LocationPair(array1.get(0), array2.get(0));
19
20    while (index1 < array1.size() && index2 < array2.size()) {
21        current.setLocations(array1.get(index1), array2.get(index2));
22        best.updateWithMin(current); // If shorter, update values
23        if (current.location1 < current.location2) {
24            index1++;
25        } else {
26            index2++;
27        }
28    }
29
30    return best;
31 }
32
33 /* Precomputation. */
34 HashMapList<String, Integer> getWordLocations(String[] words) {
35     HashMapList<String, Integer> locations = new HashMapList<String, Integer>();
36     for (int i = 0; i < words.length; i++) {
37         locations.put(words[i], i);
38     }
39     return locations;
40 }
41
42 /* HashMapList<String, Integer> is a HashMap that maps from Strings to
43 * ArrayList<Integer>. See appendix for implementation. */

```

The precomputation step of this algorithm will take $O(N)$ time, where N is the number of words in the string.

Finding the closest pair of locations will take $O(A + B)$ time, where A is the number of occurrences of the first word and B is the number of occurrences of the second word.

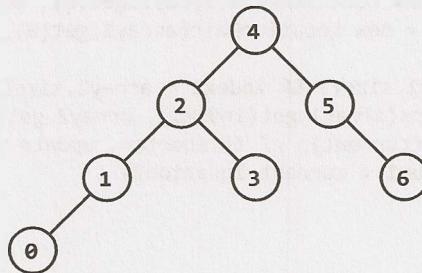
17.12 BiNode: Consider a simple data structure called BiNode, which has pointers to two other nodes. The data structure BiNode could be used to represent both a binary tree (where node1 is the left node and node2 is the right node) or a doubly linked list (where node1 is the previous node and node2 is 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 188

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, tail;
3
4     public NodePair(BiNode head, BiNode tail) {
5         this.head = head;
6         this.tail = tail;
7     }
8 }
9
10    public NodePair convert(BiNode root) {
11        if (root == null) return null;
12
13        NodePair part1 = convert(root.node1);
14        NodePair part2 = convert(root.node2);
15
16        if (part1 != null) {
17            concat(part1.tail, root);
18        }
19
20        if (part2 != null) {
21            concat(root, part2.head);
22        }
23
24        return new NodePair(part1 == null ? root : part1.head,
25                            part2 == null ? root : part2.tail);
26    }
27
28    public static void concat(BiNode x, BiNode y) {
29        x.node2 = y;
30        y.node1 = x;
31    }

```

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     BiNode convert(BiNode root) {
2         if (root == null) return null;
3
4         BiNode part1 = convert(root.node1);
5         BiNode part2 = convert(root.node2);
6
7         if (part1 != null) {
8             concat(getTail(part1), root);
9         }
10
11        if (part2 != null) {
12            concat(root, part2);
13        }
14
15        return part1 == null ? root : part1;

```

```
16 }
17
18 public static BiNode getTail(BiNode node) {
19     if (node == null) return null;
20     while (node.node2 != null) {
21         node = node.node2;
22     }
23     return node;
24 }
```

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 BiNode convertToCircular(BiNode root) {
2     if (root == null) return null;
3
4     BiNode part1 = convertToCircular(root.node1);
5     BiNode part3 = convertToCircular(root.node2);
6
7     if (part1 == null && part3 == null) {
8         root.node1 = root;
9         root.node2 = root;
10        return root;
11    }
12    BiNode tail3 = (part3 == null) ? null : part3.node1;
13
14    /* join left to root */
15    if (part1 == null) {
16        concat(part3.node1, root);
17    } else {
18        concat(part1.node1, root);
19    }
20
21    /*join right to root */
22    if (part3 == null) {
23        concat(root, part1);
24    } else {
25        concat(root, part3);
26    }
27
28    /*join right to left */
29    if (part1 != null && part3 != null) {
30        concat(tail3, part1);
31    }
32
33    return part1 == null ? root : part1;
34 }
35
36 /*Convert list to a circular linked list, then break the circular connection. */
```

```

37 BiNode convert(BiNode root) {
38     BiNode head = convertToCircular(root);
39     head.node1.node2 = null;
40     head.node1 = null;
41     return head;
42 }

```

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).

17.13 Re-Space: Oh, no! You have accidentally removed all spaces, punctuation, and capitalization in a lengthy document. A sentence like "I reset the computer. It still didn't boot!" became "iresetthecomputeritstilldidntboot". You'll deal with the punctuation and capitalization later; right now you need to re-insert the spaces. Most of the words are in a dictionary but a few are not. Given a dictionary (a list of strings) and the document (a string), design an algorithm to unconcatenate the document in a way that minimizes the number of unrecognized characters.

EXAMPLE

Input: jesslookedjustliketimherbrother

Output: jess looked just like tim her brother (7 unrecognized characters)

pg 188

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."

Brute Force

The key to this problem is finding a way to define the solution (that is, parsed string) in terms of its subproblems. One way to do this is recursing through the string.

The very first choice we make is where to insert the first space. After the first character? Second character? Third character?

Let's imagine this in terms of a string like `thisismikesfavoritefood`. What is the first space we insert?

- If we insert a space after `t`, this gives us one invalid character.
- After `th` is two invalid characters.
- After `thi` is three invalid characters.
- At `this` we have a complete word. This is zero invalid characters.
- At `thisi` is five invalid characters.
- ... and so on.

After we choose the first space, we can recursively pick the second space, then the third space, and so on, until we are done with the string.

We take the best (fewest invalid characters) out of all these choices and return.

What should the function return? We need both the number of invalid characters in the recursive path as well as the actual parsing. Therefore, we just return both by using a custom-built `ParseResult` class.

```
1  String bestSplit(HashSet<String> dictionary, String sentence) {
2      ParseResult r = split(dictionary, sentence, 0);
3      return r == null ? null : r.parsed;
4  }
5
6  ParseResult split(HashSet<String> dictionary, String sentence, int start) {
7      if (start >= sentence.length()) {
8          return new ParseResult(0, "");
9      }
10
11     int bestInvalid = Integer.MAX_VALUE;
12     String bestParsing = null;
13     String partial = "";
14     int index = start;
15     while (index < sentence.length()) {
16         char c = sentence.charAt(index);
17         partial += c;
18         int invalid = dictionary.contains(partial) ? 0 : partial.length();
19         if (invalid < bestInvalid) { // Short circuit
20             /* Recurse, putting a space after this character. If this is better than
21              * the current best option, replace the best option. */
22             ParseResult result = split(dictionary, sentence, index + 1);
23             if (invalid + result.invalid < bestInvalid) {
24                 bestInvalid = invalid + result.invalid;
25                 bestParsing = partial + " " + result.parsed;
26                 if (bestInvalid == 0) break; // Short circuit
27             }
28         }
29         index++;
30     }
31     return new ParseResult(bestInvalid, bestParsing);
32 }
33
34
35 public class ParseResult {
36     public int invalid = Integer.MAX_VALUE;
37     public String parsed = " ";
38     public ParseResult(int inv, String p) {
39         invalid = inv;
40         parsed = p;
41     }
42 }
```

We've applied two short circuits here.

- Line 22: If the number of current invalid characters exceeds the best known one, then we know this recursive path will not be ideal. There's no point in even taking it.
- Line 30: If we have a path with zero invalid characters, then we know we can't do better than this. We might as well accept this path.

What's the runtime of this? It's difficult to truly describe in practice as it depends on the (English) language. One way of looking at it is to imagine a bizarre language where essentially all paths in the recursion are taken. In this case, we are making both choices at each character. If there are n characters, this is an $O(2^n)$ runtime.

Optimized

Commonly, when we have exponential runtimes for a recursive algorithm, we optimize them through memoization (that is, caching results). To do so, we need to find the common subproblems.

Where do recursive paths overlap? That is, where are the common subproblems?

Let's again imagine the string `thisismikesfavoritefood`. Again, imagine that everything is a valid word.

In this case, we attempt to insert the first space after `t` as well as after `th` (and many other choices). Think about what the next choice is.

```

split(thisismikesfavoritefood) ->
    t + split(hisismikesfavoritefood)
OR th + split(isismikesfavoritefood)
OR ...

split(hisismikesfavoritefood) ->
    h + split(isismikesfavoritefood)
OR ...

...

```

Adding a space after `t` and `h` leads to the same recursive path as inserting a space after `th`. There's no sense in computing `split(isismikesfavoritefood)` twice when it will lead to the same result.

We should instead cache the result. We do this using a hash table which maps from the current substring to the `ParseResult` object.

We don't actually need to make the current substring a key. The `start` index in the string sufficiently represents the substring. After all, if we were to use the substring, we'd really be using `sentence.substring(start, sentence.length())`. This hash table will map from a start index to the best parsing from that index to the end of the string.

And, since the start index is the key, we don't need a true hash table at all. We can just use an array of `ParseResult` objects. This will also serve the purpose of mapping from an index to an object.

The code is essentially identical to the earlier function, but now takes in a memo table (a cache). We look up when we first call the function and set it when we return.

```

1 String bestSplit(HashSet<String> dictionary, String sentence) {
2     ParseResult[] memo = new ParseResult[sentence.length()];
3     ParseResult r = split(dictionary, sentence, 0, memo);
4     return r == null ? null : r.parsed;
5 }
6
7 ParseResult split(HashSet<String> dictionary, String sentence, int start,
8                     ParseResult[] memo) {
9     if (start >= sentence.length()) {
10         return new ParseResult(0, "");
11     } if (memo[start] != null) {
12         return memo[start];
13     }

```

```

14
15     int bestInvalid = Integer.MAX_VALUE;
16     String bestParsing = null;
17     String partial = "";
18     int index = start;
19     while (index < sentence.length()) {
20         char c = sentence.charAt(index);
21         partial += c;
22         int invalid = dictionary.contains(partial) ? 0 : partial.length();
23         if (invalid < bestInvalid) { // Short circuit
24             /* Recurse, putting a space after this character. If this is better than
25              * the current best option, replace the best option. */
26             ParseResult result = split(dictionary, sentence, index + 1, memo);
27             if (invalid + result.invalid < bestInvalid) {
28                 bestInvalid = invalid + result.invalid;
29                 bestParsing = partial + " " + result.parsed;
30                 if (bestInvalid == 0) break; // Short circuit
31             }
32         }
33     }
34     index++;
35 }
36 memo[start] = new ParseResult(bestInvalid, bestParsing);
37 return memo[start];
38 }
```

Understanding the runtime of this is even trickier than in the prior solution. Again, let's imagine the truly bizarre case, where essentially everything looks like a valid word.

One way we can approach it is to realize that `split(i)` will only be computed once for each value of `i`. What happens when we call `split(i)`, assuming we've already called `split(i+1)` through `split(n - 1)`?

```

split(i) -> calls:
    split(i + 1)
    split(i + 2)
    split(i + 3)
    split(i + 4)
    ...
    split(n - 1)
```

Each of the recursive calls has already been computed, so they just return immediately. Doing $n - i$ calls at $O(1)$ time each takes $O(n - i)$ time. This means that `split(i)` takes $O(i)$ time at most.

We can now apply the same logic to `split(i - 1)`, `split(i - 2)`, and so on. If we make 1 call to compute `split(n - 1)`, 2 calls to compute `split(n - 2)`, 3 calls to compute `split(n - 3)`, ..., n calls to compute `split(0)`, how many calls total do we do? This is basically the sum of the numbers from 1 through n , which is $O(n^2)$.

Therefore, the runtime of this function is $O(n^2)$.

17.14 Smallest K: Design an algorithm to find the smallest K numbers in an array.

pg 188

SOLUTION

There are a number of ways to approach this problem. We will go through three of them: sorting, max heap, and selection rank.

Some of these algorithms require modifying the array. This is something you should discuss with your interviewer. Note, though, that even if modifying the original array is not acceptable, you can always clone the array and modify the clone instead. This will not impact the overall big O time of any algorithm.

Solution 1: Sorting

We can sort the elements in ascending order and then take the first million numbers from that.

```

1 int[] smallestK(int[] array, int k) {
2     if (k <= 0 || k > array.length) {
3         throw new IllegalArgumentException();
4     }
5
6     /* Sort array. */
7     Arrays.sort(array);
8
9     /* Copy first k elements. */
10    int[] smallest = new int[k];
11    for (int i = 0; i < k; i++) {
12        smallest[i] = array[i];
13    }
14    return smallest;
15 }
```

The time complexity is $O(n \log(n))$.

Solution 2: Max Heap

We can use a max 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, if it's smaller than the root, we insert it into the heap and delete the largest element (which will be the root).

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.

```

1 int[] smallestK(int[] array, int k) {
2     if (k <= 0 || k > array.length) {
3         throw new IllegalArgumentException();
4     }
5
6     PriorityQueue<Integer> heap = getKMaxHeap(array, k);
7     return heapToIntArray(heap);
8 }
9
10 /* Create max heap of smallest k elements. */
11 PriorityQueue<Integer> getKMaxHeap(int[] array, int k) {
12     PriorityQueue<Integer> heap =
```

```

13     new PriorityQueue<Integer>(k, new MaxHeapComparator());
14     for (int a : array) {
15         if (heap.size() < k) { // If space remaining
16             heap.add(a);
17         } else if (a < heap.peek()) { // If full and top is small
18             heap.poll(); // remove highest
19             heap.add(a); // insert new element
20         }
21     }
22     return heap;
23 }
24
25 /* Convert heap to int array. */
26 int[] heapToIntArray(PriorityQueue<Integer> heap) {
27     int[] array = new int[heap.size()];
28     while (!heap.isEmpty()) {
29         array[heap.size() - 1] = heap.poll();
30     }
31     return array;
32 }
33
34 class MaxHeapComparator implements Comparator<Integer> {
35     public int compare(Integer x, Integer y) {
36         return y - x;
37     }
38 }

```

Java's uses the `PriorityQueue` class to offer heap-like functionality. By default, it operates as a min heap, with the smallest element on the top. To switch it to the biggest element on the top, we can pass in a different comparator.

Approach 3: Selection Rank Algorithm (if elements are unique)

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 $i - \text{leftSize}$.

Once you have found the i th smallest element, you know that all elements smaller than this will be to the left of this (since you've partitioned the array accordingly). You can now just return the first i elements.

The code below implements this algorithm.

```

1 int[] smallestK(int[] array, int k) {
2     if (k <= 0 || k > array.length) {
3         throw new IllegalArgumentException();
4     }
5

```

```

6  int threshold = rank(array, k - 1);
7  int[] smallest = new int[k];
8  int count = 0;
9  for (int a : array) {
10     if (a <= threshold) {
11         smallest[count] = a;
12         count++;
13     }
14 }
15 return smallest;
16 }
17
18 /* Get element with rank. */
19 int rank(int[] array, int rank) {
20     return rank(array, 0, array.length - 1, rank);
21 }
22
23 /* Get element with rank between left and right indices. */
24 int rank(int[] array, int left, int right, int rank) {
25     int pivot = array[randomIntInRange(left, right)];
26     int leftEnd = partition(array, left, right, pivot);
27     int leftSize = leftEnd - left + 1;
28     if (rank == leftSize - 1) {
29         return max(array, left, leftEnd);
30     } else if (rank < leftSize) {
31         return rank(array, left, leftEnd, rank);
32     } else {
33         return rank(array, leftEnd + 1, right, rank - leftSize);
34     }
35 }
36
37 /* Partition array around pivot such that all elements <= pivot come before all
38 * elements > pivot. */
39 int partition(int[] array, int left, int right, int pivot) {
40     while (left <= right) {
41         if (array[left] > pivot) {
42             /* Left is bigger than pivot. Swap it to the right side, where we know it
43             * should be. */
44             swap(array, left, right);
45             right--;
46         } else if (array[right] <= pivot) {
47             /* Right is smaller than the pivot. Swap it to the left side, where we know
48             * it should be. */
49             swap(array, left, right);
50             left++;
51         } else {
52             /* Left and right are in correct places. Expand both sides. */
53             left++;
54             right--;
55         }
56     }
57     return left - 1;
58 }
59
60 /* Get random integer within range, inclusive. */
61 int randomIntInRange(int min, int max) {

```

```

62     Random rand = new Random();
63     return rand.nextInt(max + 1 - min) + min;
64 }
65
66 /* Swap values at index i and j. */
67 void swap(int[] array, int i, int j) {
68     int t = array[i];
69     array[i] = array[j];
70     array[j] = t;
71 }
72
73 /* Get largest element in array between left and right indices. */
74 int max(int[] array, int left, int right) {
75     int max = Integer.MIN_VALUE;
76     for (int i = left; i <= right; i++) {
77         max = Math.max(array[i], max);
78     }
79     return max;
80 }

```

If the elements are not unique, we can tweak this algorithm slightly to accommodate this.

Approach 4: Selection Rank Algorithm (if elements are not unique)

The major change that needs to be made is to the `partition` function. When we partition the array around a pivot element, we now partition it into three chunks: less than pivot, equal to pivot, and greater than pivot.

This requires minor tweaks to `rank` as well. We now compare the size of left and middle partitions to rank.

```

1  class PartitionResult {
2      int leftSize, middleSize;
3      public PartitionResult(int left, int middle) {
4          this.leftSize = left;
5          this.middleSize = middle;
6      }
7  }
8
9  int[] smallestK(int[] array, int k) {
10    if (k <= 0 || k > array.length) {
11        throw new IllegalArgumentException();
12    }
13
14    /* Get item with rank k - 1. */
15    int threshold = rank(array, k - 1);
16
17    /* Copy elements smaller than the threshold element. */
18    int[] smallest = new int[k];
19    int count = 0;
20    for (int a : array) {
21        if (a < threshold) {
22            smallest[count] = a;
23            count++;
24        }
25    }
26
27    /* If there's still room left, this must be for elements equal to the threshold

```

```

28     * element. Copy those in. */
29     while (count < k) {
30         smallest[count] = threshold;
31         count++;
32     }
33
34     return smallest;
35 }
36
37 /* Find value with rank k in array. */
38 int rank(int[] array, int k) {
39     if (k >= array.length) {
40         throw new IllegalArgumentException();
41     }
42     return rank(array, k, 0, array.length - 1);
43 }
44
45 /* Find value with rank k in sub array between start and end. */
46 int rank(int[] array, int k, int start, int end) {
47     /* Partition array around an arbitrary pivot. */
48     int pivot = array[randomIntInRange(start, end)];
49     PartitionResult partition = partition(array, start, end, pivot);
50     int leftSize = partition.leftSize;
51     int middleSize = partition.middleSize;
52
53     /* Search portion of array. */
54     if (k < leftSize) { // Rank k is on left half
55         return rank(array, k, start, start + leftSize - 1);
56     } else if (k < leftSize + middleSize) { // Rank k is in middle
57         return pivot; // middle is all pivot values
58     } else { // Rank k is on right
59         return rank(array, k - leftSize - middleSize, start + leftSize + middleSize,
60                     end);
61     }
62 }
63
64 /* Partition result into < pivot, equal to pivot -> bigger than pivot. */
65 PartitionResult partition(int[] array, int start, int end, int pivot) {
66     int left = start; /* Stays at (right) edge of left side. */
67     int right = end; /* Stays at (left) edge of right side. */
68     int middle = start; /* Stays at (right) edge of middle. */
69     while (middle <= right) {
70         if (array[middle] < pivot) {
71             /* Middle is smaller than the pivot. Left is either smaller or equal to
72                 * the pivot. Either way, swap them. Then middle and left should move by
73                 * one. */
74             swap(array, middle, left);
75             middle++;
76             left++;
77         } else if (array[middle] > pivot) {
78             /* Middle is bigger than the pivot. Right could have any value. Swap them,
79                 * then we know that the new right is bigger than the pivot. Move right by
80                 * one. */
81             swap(array, middle, right);
82             right--;
83         } else if (array[middle] == pivot) {

```

```
84     /* Middle is equal to the pivot. Move by one. */
85     middle++;
86 }
87 }
88
89 /* Return sizes of left and middle. */
90 return new PartitionResult(left - start, right - left + 1);
91 }
```

Notice the change made to `smallestK` too. We can't simply copy all elements less than or equal to `threshold` into the array. Since we have duplicates, there could be many more than `k` elements that are less than or equal to `threshold`. (We also can't just say "okay, only copy `k` elements over." We could inadvertently fill up the array early on with "equal" elements, and not leave enough space for the smaller ones.)

The solution for this is fairly simple: only copy over the smaller elements first, then fill up the array with equal elements at the end.

17.15 Longest Word: Given a list of words, write a program to find the longest word made of other words in the list.

pg 188

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 great 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 }

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/memoization 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.

17.16 The Masseuse: A popular masseuse receives a sequence of back-to-back appointment requests and is debating which ones to accept. She needs a 15-minute break between appointments and therefore she cannot accept any adjacent requests. Given a sequence of back-to-back appointment requests (all multiples of 15 minutes, none overlap, and none can be moved), find the optimal (highest total booked minutes) set the masseuse can honor. Return the number of minutes.

EXAMPLE

Input: {30, 15, 60, 75, 45, 15, 15, 45}

Output: 180 minutes ({30, 60, 45, 45}).

pg 188

SOLUTION

Let's start with an example. We'll draw it visually to get a better feel for the problem. Each number indicates the number of minutes in the appointment.

$r_0 = 75$	$r_1 = 105$	$r_2 = 120$	$r_3 = 75$	$r_4 = 90$	$r_5 = 135$
------------	-------------	-------------	------------	------------	-------------

Alternatively, we could have also divided all the values (including the break) by 15 minutes, to give us the array {5, 7, 8, 5, 6, 9}. This would be equivalent, but now we would want a 1-minute break.

The best set of appointments for this problem has 330 minutes total, formed with $\{r_0 = 75, r_2 = 120, r_5 = 135\}$. Note that we've intentionally chosen an example in which the best sequence of appointments was not formed through a strictly alternating sequence.

We should also recognize that choosing the longest appointment first (the "greedy" strategy) would not necessarily be optimal. For example, a sequence like {45, 60, 45, 15} would not have 60 in the optimal set.

Solution #1: Recursion

The first thing that may come to mind is a recursive solution. We have essentially a sequence of choices as we walk down the list of appointments: Do we use this appointment or do we not? If we use appointment i , we must skip appointment $i + 1$ as we can't take back-to-back appointments. Appointment $i + 2$ is a possibility (but not necessarily the best choice).

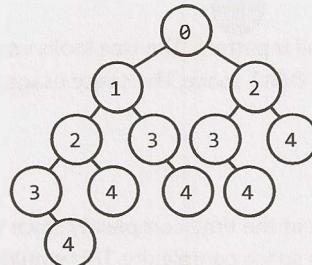
```

1  int maxMinutes(int[] massages) {
2      return maxMinutes(massages, 0);
3  }
4
5  int maxMinutes(int[] massages, int index) {
6      if (index >= massages.length) { // Out of bounds
7          return 0;
8      }
9
10     /* Best with this reservation. */
11     int bestWith = massages[index] + maxMinutes(massages, index + 2);
12
13     /* Best without this reservation. */
14     int bestWithout = maxMinutes(massages, index + 1);
15
16     /* Return best of this subarray, starting from index. */
17     return Math.max(bestWith, bestWithout);
18 }
```

The runtime of this solution is $O(2^n)$ because at each element we're making two choices and we do this n times (where n is the number of massages).

The space complexity is $O(n)$ due to the recursive call stack.

We can also depict this through a recursive call tree on an array of length 5. The number in each node represents the index value in a call to `maxMinutes`. Observe that, for example, `maxMinutes(massages, 0)` calls `maxMinutes(massages, 1)` and `maxMinutes(massages, 2)`.



As with many recursive problems, we should evaluate if there's a possibility to memoize repeated subproblems. Indeed, there is.

Solution #2: Recursion + Memoization

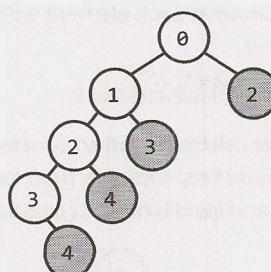
We will repeatedly call `maxMinutes` on the same inputs. For example, we'll call it on index 2 when we're deciding whether to take appointment 0. We'll also call it on index 2 when we're deciding whether to take appointment 1. We should memoize this.

Our memo table is just a mapping from index to the max minutes. Therefore, a simple array will suffice.

```

1  int maxMinutes(int[] massages) {
2      int[] memo = new int[massages.length];
3      return maxMinutes(massages, 0, memo);
4  }
5
6  int maxMinutes(int[] massages, int index, int[] memo) {
7      if (index >= massages.length) {
8          return 0;
9      }
10     if (memo[index] == 0) {
11         int bestWith = massages[index] + maxMinutes(massages, index + 2, memo);
12         int bestWithout = maxMinutes(massages, index + 1, memo);
13         memo[index] = Math.max(bestWith, bestWithout);
14     }
15
16     return memo[index];
17 }
18 }
```

To determine the runtime, we'll draw the same recursive call tree as before but gray-out the calls that will return immediately. The calls that will never happen will be deleted entirely.



If we drew a bigger tree, we'd see a similar pattern. The tree looks very linear, with one branch down to the left. This gives us an $O(n)$ runtime and $O(n)$ space. The space usage comes from the recursive call stack as well as from the memo table.

Solution #3: Iterative

Can we do better? We certainly can't beat the time complexity since we have to look at each appointment. However, we might be able to beat the space complexity. This would mean not solving the problem recursively.

Let's look at our first example again.

$r_0 = 30$	$r_1 = 15$	$r_2 = 60$	$r_3 = 75$	$r_4 = 45$	$r_5 = 15$	$r_6 = 15$	$r_7 = 45$
------------	------------	------------	------------	------------	------------	------------	------------

As we noted in the problem statement, we cannot take adjacent appointments.

There's another observation, though, that we can make: We should never *skip* three consecutive appointments. That is, we might skip r_1 and r_2 if we wanted to take r_0 and r_3 . But we would never skip r_1 , r_2 , and r_3 . This would be suboptimal since we could always improve our set by grabbing that middle element.

This means that if we take r_0 , we know we'll definitely skip r_1 and definitely take either r_2 or r_3 . This substantially limits the options we need to evaluate and opens the door to an iterative solution.

Let's think about our recursive + memoization solution and try to reverse the logic; that is, let's try to approach it iteratively.

A useful way to do this is to approach it from the back and move toward the start of the array. At each point, we find the solution for the subarray.

- **best(7):** What's the best option for $\{r_7 = 45\}$? We can get 45 min. if we take r_7 , so $\text{best}(7) = 45$.
- **best(6):** What's the best option for $\{r_6 = 15, \dots\}$? Still 45 min., so $\text{best}(6) = 45$.
- **best(5):** What's the best option for $\{r_5 = 15, \dots\}$? We can either:
 - » take $r_5 = 15$ and merge it with $\text{best}(7) = 45$, or:
 - » take $\text{best}(6) = 45$.

The first gives us 60 minutes, $\text{best}(5) = 60$.

- **best(4):** What's the best option for $\{r_4 = 45, \dots\}$? We can either:
 - » take $r_4 = 45$ and merge it with $\text{best}(6) = 45$, or:
 - » take $\text{best}(5) = 60$.

The first gives us 90 minutes, $\text{best}(4) = 90$.

- **best(3):** What's the best option for $\{r_3 = 75, \dots\}$? We can either:
 - » take $r_3 = 75$ and merge it with $\text{best}(5) = 60$, or:

» take best(4) = 90.

The first gives us 135 minutes, best(3) = 135.

- best(2): What's the best option for $\{r_2 = 60, \dots\}$? We can either:

» take $r_2 = 60$ and merge it with best(4) = 90, or:

» take best(3) = 135.

The first gives us 150 minutes, best(2) = 150.

- best(1): What's the best option for $\{r_1 = 15, \dots\}$? We can either:

» take $r_1 = 15$ and merge it with best(3) = 135, or:

» take best(2) = 150.

Either way, best(1) = 150.

- best(0): What's the best option for $\{r_0 = 30, \dots\}$? We can either:

» take $r_0 = 30$ and merge it with best(2) = 150, or:

» take best(1) = 150.

The first gives us 180 minutes, best(0) = 180.

Therefore, we return 180 minutes.

The code below implements this algorithm.

```

1 int maxMinutes(int[] massages) {
2     /* Allocating two extra slots in the array so we don't have to do bounds
3      * checking on lines 7 and 8. */
4     int[] memo = new int[massages.length + 2];
5     memo[massages.length] = 0;
6     memo[massages.length + 1] = 0;
7     for (int i = massages.length - 1; i >= 0; i--) {
8         int bestWith = massages[i] + memo[i + 2];
9         int bestWithout = memo[i + 1];
10        memo[i] = Math.max(bestWith, bestWithout);
11    }
12    return memo[0];
13 }
```

The runtime of this solution is $O(n)$ and the space complexity is also $O(n)$.

It's nice in some ways that it's iterative, but we haven't actually "won" anything here. The recursive solution had the same time and space complexity.

Solution #4: Iterative with Optimal Time and Space

In reviewing the last solution, we can recognize that we only use the values in the memo table for a short amount of time. Once we are several elements past an index, we never use that element's index again.

In fact, at any given index i , we only need to know the best value from $i + 1$ and $i + 2$. Therefore, we can get rid of the memo table and just use two integers.

```

1 int maxMinutes(int[] massages) {
2     int oneAway = 0;
3     int twoAway = 0;
4     for (int i = massages.length - 1; i >= 0; i--) {
5         int bestWith = massages[i] + twoAway;
6         int bestWithout = oneAway;
```

```
7     int current = Math.max(bestWith, bestWithout);
8     twoAway = oneAway;
9     oneAway = current;
10    }
11    return oneAway;
12 }
```

This gives us the most optimal time and space possible: $O(n)$ time and $O(1)$ space.

Why did we look backward? It's a common technique in many problems to walk backward through an array.

However, we can walk forward if we want. This is easier for some people to think about, and harder for others. In this case, rather than asking "What's the best set that starts with $a[i]$?", we would ask "What's the best set that ends with $a[i]$?"

17.17 Multi Search: Given a string b and an array of smaller strings T , design a method to search b for each small string in T .

pg 189

SOLUTION

Let's start with an example:

```
T = {"is", "ppi", "hi", "sis", "i", "ssippi"}
b = "mississippi"
```

Note that in our example, we made sure to have some strings (like "is") that appear multiple times in b .

Solution #1

The naive solution is reasonably straightforward. Just search through the bigger string for each instance of the smaller string.

```
1  HashMapList<String, Integer> searchAll(String big, String[] smalls) {
2      HashMapList<String, Integer> lookup =
3          new HashMapList<String, Integer>();
4      for (String small : smalls) {
5          ArrayList<Integer> locations = search(big, small);
6          lookup.put(small, locations);
7      }
8      return lookup;
9  }
10
11 /* Find all locations of the smaller string within the bigger string. */
12 ArrayList<Integer> search(String big, String small) {
13     ArrayList<Integer> locations = new ArrayList<Integer>();
14     for (int i = 0; i < big.length() - small.length() + 1; i++) {
15         if (isSubstringAtLocation(big, small, i)) {
16             locations.add(i);
17         }
18     }
19     return locations;
20 }
21
22 /* Check if small appears at index offset within big. */
23 boolean isSubstringAtLocation(String big, String small, int offset) {
24     for (int i = 0; i < small.length(); i++) {
25         if (big.charAt(offset + i) != small.charAt(i)) {
```

```

26         return false;
27     }
28 }
29 return true;
30 }
31
32 /* HashMapList<String, Integer> is a HashMap that maps from Strings to
33 * ArrayList<Integer>. See appendix for implementation. */

```

We could have also used a `substring` and `equals` function, instead of writing `isAtLocation`. This is slightly faster (though not in terms of big O) because it doesn't require creating a bunch of substrings.

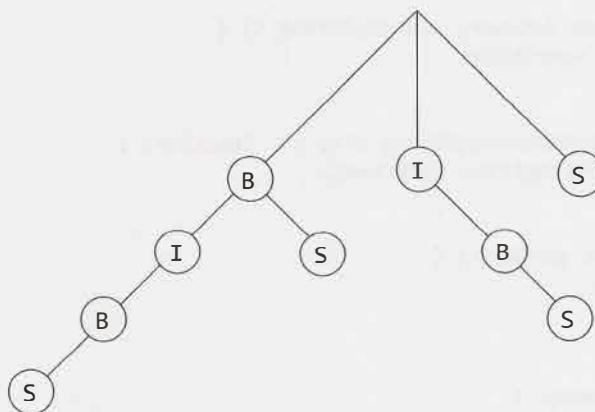
This will take $O(kbt)$ time, where k is the length of the longest string in T , b is the length of the bigger string, and t is the number of smaller strings within T .

Solution #2

To optimize this, we should think about how we can tackle all the elements in T at once, or somehow re-use work.

One way is to create a trie-like data structure using each suffix in the bigger string. For the string `bibs`, the suffix list would be: `bibs`, `ibs`, `bs`, `s`.

The tree for this is below.



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  HashMapList<String, Integer> searchAll(String big, String[] smalls) {
2      HashMapList<String, Integer> lookup = new HashMapList<String, Integer>();
3      Trie tree = createTrieFromString(big);
4      for (String s : smalls) {
5          /* Get terminating location of each occurrence.*/
6          ArrayList<Integer> locations = tree.search(s);
7
8          /* Adjust to starting location. */
9          subtractValue(locations, s.length());
10
11         /* Insert. */
12         lookup.put(s, locations);
13     }
14

```

```
15 }
16
17 Trie createTrieFromString(String s) {
18     Trie trie = new Trie();
19     for (int i = 0; i < s.length(); i++) {
20         String suffix = s.substring(i);
21         trie.insertString(suffix, i);
22     }
23     return trie;
24 }
25
26 void subtractValue(ArrayList<Integer> locations, int delta) {
27     if (locations == null) return;
28     for (int i = 0; i < locations.size(); i++) {
29         locations.set(i, locations.get(i) - delta);
30     }
31 }
32
33 public class Trie {
34     private TrieNode root = new TrieNode();
35
36     public Trie(String s) { insertString(s, 0); }
37     public Trie() {}
38
39     public ArrayList<Integer> search(String s) {
40         return root.search(s);
41     }
42
43     public void insertString(String str, int location) {
44         root.insertString(str, location);
45     }
46
47     public TrieNode getRoot() {
48         return root;
49     }
50 }
51
52 public class TrieNode {
53     private HashMap<Character, TrieNode> children;
54     private ArrayList<Integer> indexes;
55     private char value;
56
57     public TrieNode() {
58         children = new HashMap<Character, TrieNode>();
59         indexes = new ArrayList<Integer>();
60     }
61
62     public void insertString(String s, int index) {
63         indexes.add(index);
64         if (s != null && s.length() > 0) {
65             value = s.charAt(0);
66             TrieNode child = null;
67             if (children.containsKey(value)) {
68                 child = children.get(value);
69             } else {
70                 child = new TrieNode();
```

```

71         children.put(value, child);
72     }
73     String remainder = s.substring(1);
74     child.insertString(remainder, index + 1);
75 } else {
76     children.put('\0', null); // Terminating character
77 }
78 }
79
80 public ArrayList<Integer> search(String s) {
81     if (s == null || s.length() == 0) {
82         return indexes;
83     } else {
84         char first = s.charAt(0);
85         if (children.containsKey(first)) {
86             String remainder = s.substring(1);
87             return children.get(first).search(remainder);
88         }
89     }
90     return null;
91 }
92
93 public boolean terminates() {
94     return children.containsKey('\0');
95 }
96
97 public TrieNode getChild(char c) {
98     return children.get(c);
99 }
100 }
101
102 /* HashMapList<String, Integer> is a HashMap that maps from Strings to
103 * ArrayList<Integer>. See appendix for implementation. */

```

It takes $O(b^2)$ time to create the tree and $O(kt)$ time to search for the locations.

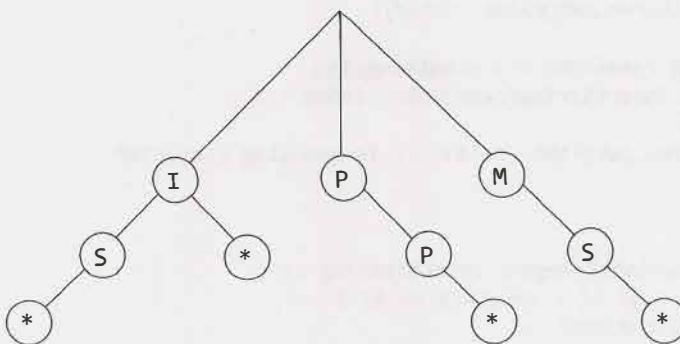
Reminder: k is the length of the longest string in T , b is the length of the bigger string, and t is the number of smaller strings within T .

The total runtime is $O(b^2 + kt)$.

Without some additional knowledge of the expected input, you cannot directly compare $O(bkt)$, which was the runtime of the prior solution, to $O(b^2 + kt)$. If b is very large, then $O(bkt)$ is preferable. But if you have a lot of smaller strings, then $O(b^2 + kt)$ might be better.

Solution #3

Alternatively, we can add all the smaller strings into a trie. For example, the strings {i, is, pp, ms} would look like the trie below. The asterisk (*) hanging from a node indicates that this node completes a word.



Now, when we want to find all words in mississippi, we search through this trie starting with each word.

- m: We would first look up in the trie starting with m, the first letter in mississippi. As soon as we go to mi, we terminate.
- i: Then, we go to i, the second character in mississippi. We see that i is a complete word, so we add it to the list. We also keep going with i over to is. The string is is also a complete word, so we add that to the list. This node has no more children, so we move onto the next character in mississippi.
- s: We now go to s. There is no upper-level node for s, so we go onto the next character.
- s: Another s. Go on to the next character.
- i: We see another i. We go to the i node in the trie. We see that i is a complete word, so we add it to the list. We also keep going with i over to is. The string is is also a complete word, so we add that to the list. This node has no more children, so we move onto the next character in mississippi.
- s: We go to s. There is no upper-level node for s.
- s: Another s. Go on to the next character.
- i: We go to the i node. We see that i is a complete word, so we add it to the trie. The next character in mississippi is a p. There is no node p, so we break here.
- p: We see a p. There is no node p.
- p: Another p.
- i: We go to the i node. We see that i is a complete word, so we add it to the trie. There are no more characters left in mississippi, so we are done.

Each time we find a complete "small" word, we add it to a list along with the location in the bigger word (mississippi) where we found the small word.

The code below implements this algorithm.

```

1  HashMapList<String, Integer> searchAll(String big, String[] smalls) {
2      HashMapList<String, Integer> lookup = new HashMapList<String, Integer>();
3      int maxLen = big.length();
4      TrieNode root = createTreeFromStrings(smallss, maxLen).getRoot();
5
6      for (int i = 0; i < big.length(); i++) {
7          ArrayList<String> strings = findStringsAtLoc(root, big, i);
8          insertIntoHashMap(strings, lookup, i);
9      }
10
11     return lookup;
  
```

```

12 }
13
14 /* Insert each string into trie (provided string is not longer than maxLen). */
15 Trie createTreeFromStrings(String[] smalls, int maxLen) {
16     Trie tree = new Trie("");
17     for (String s : smalls) {
18         if (s.length() <= maxLen) {
19             tree.insertString(s, 0);
20         }
21     }
22     return tree;
23 }
24
25 /* Find strings in trie that start at index "start" within big. */
26 ArrayList<String> findStringsAtLoc(TrieNode root, String big, int start) {
27     ArrayList<String> strings = new ArrayList<String>();
28     int index = start;
29     while (index < big.length()) {
30         root = root.getChild(big.charAt(index));
31         if (root == null) break;
32         if (root.terminates()) { // Is complete string, add to list
33             strings.add(big.substring(start, index + 1));
34         }
35         index++;
36     }
37     return strings;
38 }
39
40 /* HashMapList<String, Integer> is a HashMap that maps from Strings to
41 * ArrayList<Integer>. See appendix for implementation. */

```

This algorithm takes $O(kt)$ time to create the trie and $O(bk)$ time to search for all the strings.

Reminder: k is the length of the longest string in T , b is the length of the bigger string, and t is the number of smaller strings within T .

The total time to solve the question is $O(kt + bk)$.

Solution #1 was $O(kbt)$. We know that $O(kt + bk)$ will be faster than $O(kbt)$.

Solution #2 was $O(b^2 + kt)$. Since b will always be bigger than k (or if it's not, then we know this really long string k cannot be found in b), we know Solution #3 is also faster than Solution #2.

17.18 Shortest Supersequence: You are given two arrays, one shorter (with all distinct elements) and one longer. Find the shortest subarray in the longer array that contains all the elements in the shorter array. The items can appear in any order.

EXAMPLE

Input:

```
{1, 5, 9}
{7, 5, 9, 0, 2, 1, 3, 5, 7, 9, 1, 1, 5, 8, 8, 9, 7}
Output: [7, 10] (the underlined portion above)
```

pg 189

SOLUTIONS

As usual, a brute force approach is a good way to start. Try thinking about it as if you were doing it by hand. How would you do it?

Let's use the example from the problem to walk through this. We'll call the smaller array `smallArray` and the bigger array `bigArray`.

Brute Force

The slow, "easy" way to do this is to iterate through `bigArray` and do repeated small passes through it.

At each index in `bigArray`, scan forward to find the next occurrence of each element in `smallArray`. The largest of these next occurrences will tell us the shortest subarray that starts at that index. (We'll call this concept "closure." That is, the closure is the element that "closes" a complete subarray starting at that index. For example, the closure of index 3—which has value 0—in the example is index 9.)

By finding the closures for each index in the array, we can find the shortest subarray overall.

```
1 Range shortestSupersequence(int[] bigArray, int[] smallArray) {
2     int bestStart = -1;
3     int bestEnd = -1;
4     for (int i = 0; i < bigArray.length; i++) {
5         int end = findClosure(bigArray, smallArray, i);
6         if (end == -1) break;
7         if (bestStart == -1 || end - i < bestEnd - bestStart) {
8             bestStart = i;
9             bestEnd = end;
10        }
11    }
12    return new Range(bestStart, bestEnd);
13 }
14
15 /* Given an index, find the closure (i.e., the element which terminates a complete
16 * subarray containing all elements in smallArray). This will be the max of the
17 * next locations of each element in smallArray. */
18 int findClosure(int[] bigArray, int[] smallArray, int index) {
19     int max = -1;
20     for (int i = 0; i < smallArray.length; i++) {
21         int next = findNextInstance(bigArray, smallArray[i], index);
22         if (next == -1) {
23             return -1;
24         }
25         max = Math.max(next, max);
26     }
}
```

```

27     return max;
28 }
29
30 /* Find next instance of element starting from index. */
31 int findNextInstance(int[] array, int element, int index) {
32     for (int i = index; i < array.length; i++) {
33         if (array[i] == element) {
34             return i;
35         }
36     }
37     return -1;
38 }
39
40 public class Range {
41     private int start;
42     private int end;
43     public Range(int s, int e) {
44         start = s;
45         end = e;
46     }
47
48     public int length() { return end - start + 1; }
49     public int getStart() { return start; }
50     public int getEnd() { return end; }
51
52     public boolean shorterThan(Range other) {
53         return length() < other.length();
54     }
55 }

```

This algorithm will potentially take $O(SB^2)$ time, where B is the length of `bigString` and S is the length of `smallString`. This is because at each of the B characters, we potentially do $O(SB)$ work: S scans of the rest of the string, which has potentially B characters.

Optimized

Let's think about how we can optimize this. The core reason why it's slow is the repeated searches. Is there a faster way that we can find, given an index, the next occurrence of a particular character?

Let's think about it with an example. Given the array below, is there a way we could quickly find the next 5 from each location?

7, 5, 9, 0, 2, 1, 3, 5, 7, 9, 1, 1, 5, 8, 8, 9, 7

Yes. Because we're going to have to do this repeatedly, we can precompute this information in just a single (backwards) sweep. Iterate through the array backwards, tracking the last (most recent) occurrence of 5.

value	7	5	9	0	2	1	3	5	7	9	1	1	5	8	8	9	7
index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
next 5	1	1	7	7	7	7	7	12	12	12	12	12	x	x	x	x	

Doing this for each of {1, 5, 9} takes just 3 backwards sweeps.

Some people want to merge this into one backwards sweep that handles all three values. It feels faster—but it's not really. Doing it in one backwards sweep means doing three comparisons at each iteration. N moves through the list with three comparisons at each move is no better than $3N$ moves and one comparison at each move. You might as well keep the code clean by doing it in separate sweeps.

value	7	5	9	0	2	1	3	5	7	9	1	1	5	8	8	9	7
index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
next 1	5	5	5	5	5	5	10	10	10	10	10	11	x	x	x	x	x
next 5	1	1	7	7	7	7	7	12	12	12	12	12	x	x	x	x	x
next 9	2	2	2	9	9	9	9	9	9	9	15	15	15	15	15	15	x

The `findNextInstance` function can now just use this table to find the next occurrence, rather than doing a search.

But, actually, we can make it a bit simpler. Using the table above, we can quickly compute the closure of each index. It's just the max of the column. If a column has an x in it, then there is no closure, at this indicates that there's no next occurrence of that character.

The difference between the index and the closure is the smallest subarray starting at that index.

value	7	5	9	0	2	1	3	5	7	9	1	1	5	8	8	9	7
index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
next 1	5	5	5	5	5	5	10	10	10	10	10	11	x	x	x	x	x
next 5	1	1	7	7	7	7	7	7	12	12	12	12	12	x	x	x	x
next 9	2	2	2	9	9	9	9	9	9	9	15	15	15	15	15	15	x
closure	5	5	7	9	9	9	10	10	12	12	15	15	x	x	x	x	x
diff.	5	4	5	6	5	4	4	3	4	3	5	4	x	x	x	x	x

Now, all we have to do is to find the minimum distance in this table.

```

1 Range shortestSupersequence(int[] big, int[] small) {
2     int[][] nextElements = getNextElementsMulti(big, small);
3     int[] closures = getClosures(nextElements);
4     return getShortestClosure(closures);
5 }
6
7 /* Create table of next occurrences. */
8 int[][] getNextElementsMulti(int[] big, int[] small) {
9     int[][] nextElements = new int[small.length][big.length];
10    for (int i = 0; i < small.length; i++) {
11        nextElements[i] = getNextElement(big, small[i]);
12    }
13    return nextElements;
14 }
15
16 /* Do backwards sweep to get a list of the next occurrence of value from each
17 * index. */
18 int[] getNextElement(int[] bigArray, int value) {
19    int next = -1;
20    int[] nexts = new int[bigArray.length];
21    for (int i = bigArray.length - 1; i >= 0; i--) {
22        if (bigArray[i] == value) {
23            next = i;
24        }
25        nexts[i] = next;
26    }
27    return nexts;
28 }
29
30 /* Get closure for each index. */

```

```

31 int[] getClosures(int[][] nextElements) {
32     int[] maxNextElement = new int[nextElements[0].length];
33     for (int i = 0; i < nextElements[0].length; i++) {
34         maxNextElement[i] = getClosureForIndex(nextElements, i);
35     }
36     return maxNextElement;
37 }
38
39 /* Given an index and the table of next elements, find the closure for this index
40 * (which will be the min of this column). */
41 int getClosureForIndex(int[][] nextElements, int index) {
42     int max = -1;
43     for (int i = 0; i < nextElements.length; i++) {
44         if (nextElements[i][index] == -1) {
45             return -1;
46         }
47         max = Math.max(max, nextElements[i][index]);
48     }
49     return max;
50 }
51
52 /* Get shortest closure. */
53 Range getShortestClosure(int[] closures) {
54     int bestStart = -1;
55     int bestEnd = -1;
56     for (int i = 0; i < closures.length; i++) {
57         if (closures[i] == -1) {
58             break;
59         }
60         int current = closures[i] - i;
61         if (bestStart == -1 || current < bestEnd - bestStart) {
62             bestStart = i;
63             bestEnd = closures[i];
64         }
65     }
66     return new Range(bestStart, bestEnd);
67 }

```

This algorithm will potentially take $O(SB)$ time, where B is the length of `bigString` and S is the length of `smallString`. This is because we do S sweeps through the array to build up the next occurrences table and each sweep takes $O(B)$ time.

It uses $O(SB)$ space.

More Optimized

While our solution is fairly optimal, we can reduce the space usage. Remember the table we created:

value	7	5	9	0	2	1	3	5	7	9	1	1	5	8	8	9	7
index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
next 1	5	5	5	5	5	5	10	10	10	10	10	11	x	x	x	x	x
next 5	1	1	7	7	7	7	7	7	12	12	12	12	12	x	x	x	x
next 9	2	2	2	9	9	9	9	9	9	9	15	15	15	15	15	15	x
closure	5	5	7	9	9	9	10	10	12	12	15	15	x	x	x	x	x

In actuality, all we need is the closure row, which is the minimum of all the other rows. We don't need to store all the other next occurrence information the entire time.

Instead, as we do each sweep, we just update the closure row with the minimums. The rest of the algorithm works essentially the same way.

```
1 Range shortestSupersequence(int[] big, int[] small) {
2     int[] closures = getClosures(big, small);
3     return getShortestClosure(closures);
4 }
5
6 /* Get closure for each index. */
7 int[] getClosures(int[] big, int[] small) {
8     int[] closure = new int[big.length];
9     for (int i = 0; i < small.length; i++) {
10         sweepForClosure(big, closure, small[i]);
11     }
12     return closure;
13 }
14
15 /* Do backwards sweep and update the closures list with the next occurrence of
16 * value, if it's later than the current closure. */
17 void sweepForClosure(int[] big, int[] closures, int value) {
18     int next = -1;
19     for (int i = big.length - 1; i >= 0; i--) {
20         if (big[i] == value) {
21             next = i;
22         }
23         if ((next == -1 || closures[i] < next) &&
24             (closures[i] != -1)) {
25             closures[i] = next;
26         }
27     }
28 }
29
30 /* Get shortest closure. */
31 Range getShortestClosure(int[] closures) {
32     Range shortest = new Range(0, closures[0]);
33     for (int i = 1; i < closures.length; i++) {
34         if (closures[i] == -1) {
35             break;
36         }
37         Range range = new Range(i, closures[i]);
38         if (!shortest.shorterThan(range)) {
39             shortest = range;
40         }
41     }
42     return shortest;
43 }
```

This still runs in $O(SB)$ time, but it now only takes $O(B)$ additional memory.

Alternative & More Optimal Solution

There's a totally different way to approach it. Let's suppose we had a list of the occurrences of each element in `smallArray`.

value	7	5	9	9	2	1	3	5	7	9	1	1	5	8	8	9	7
index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

1 -> {5, 10, 11}
 5 -> {1, 7, 12}
 9 -> {2, 3, 9, 15}

What is the very first valid subsequence (which contains 1, 5, and 9)? We can just look at the heads of each list to tell us this. The minimum of the heads is the start of the range and the max of the heads is the end of the range. In this case, the first range is [1, 5]. This is currently our “best” subsequence.

How can we find the next one? Well, the next one will not include index 1, so let’s remove that from the list.

1 -> {5, 10, 11}
 5 -> {7, 12}
 9 -> {2, 3, 9, 15}

The next subsequence is [2, 7]. This is worse than the earlier best, so we can toss it.

Now, what’s the next subsequence? We can remove the min from earlier (2) and find out.

1 -> {5, 10, 11}
 5 -> {7, 12}
 9 -> {3, 9, 15}

The next subsequence is [3, 7], which is no better or worse than our current best.

We can continue down this path each time, repeating this process. We will end up iterating through all “minimal” subsequences that start from a given point.

1. Current subsequence is [min of heads, max of heads]. Compare to best subsequence and update if necessary.
2. Remove the minimum head.
3. Repeat.

This will give us an $O(SB)$ time complexity. This is because for each of B elements, we are doing a comparison to the S other list heads to find the minimum.

This is pretty good, but let’s see if we can make that minimum computation faster.

What we’re doing in these repeated minimum calls is taking a bunch of elements, finding and removing the minimum, adding in one more element, and then finding the minimum again.

We can make this faster by using a min-heap. First, put each of the heads in a min-heap. Remove the minimum. Look up the list that this minimum came from and add back the new head. Repeat.

To get the list that the minimum element came from, we’ll need to use a `HeapNode` class that stores both the `locationWithinList` (the index) and the `listId`. This way, when we remove the minimum, we can jump back to the correct list and add its new head to the heap.

```

1 Range shortestSupersequence(int[] array, int[] elements) {
2     ArrayList<Queue<Integer>> locations = getLocationsForElements(array, elements);
3     if (locations == null) return null;
4     return getShortestClosure(locations);
5 }
6
7 /* Get list of queues (linked lists) storing the indices at which each element in
8 * smallArray appears in bigArray. */
9 ArrayList<Queue<Integer>> getLocationsForElements(int[] big, int[] small) {
10    /* Initialize hash map from item value to locations. */
11    HashMap<Integer, Queue<Integer>> itemLocations =

```

```
12     new HashMap<Integer, Queue<Integer>>();
13     for (int s : small) {
14         Queue<Integer> queue = new LinkedList<Integer>();
15         itemLocations.put(s, queue);
16     }
17
18     /*Walk through big array, adding the item locations to hash map */
19     for (int i = 0; i < big.length; i++) {
20         Queue<Integer> queue = itemLocations.get(big[i]);
21         if (queue != null) {
22             queue.add(i);
23         }
24     }
25
26     ArrayList<Queue<Integer>> allLocations = new ArrayList<Queue<Integer>>();
27     allLocations.addAll(itemLocations.values());
28     return allLocations;
29 }
30
31 Range getShortestClosure(ArrayList<Queue<Integer>> lists) {
32     PriorityQueue<HeapNode> minHeap = new PriorityQueue<HeapNode>();
33     int max = Integer.MIN_VALUE;
34
35     /*Insert min element from each list. */
36     for (int i = 0; i < lists.size(); i++) {
37         int head = lists.get(i).remove();
38         minHeap.add(new HeapNode(head, i));
39         max = Math.max(max, head);
40     }
41
42     int min = minHeap.peek().locationWithinList;
43     int bestRangeMin = min;
44     int bestRangeMax = max;
45
46     while (true) {
47         /*Remove min node. */
48         HeapNode n = minHeap.poll();
49         Queue<Integer> list = lists.get(n.listId);
50
51         /*Compare range to best range. */
52         min = n.locationWithinList;
53         if (max - min < bestRangeMax - bestRangeMin) {
54             bestRangeMax = max;
55             bestRangeMin = min;
56         }
57
58         /*If there are no more elements, then there's no more subsequences and we
59          * can break. */
60         if (list.size() == 0) {
61             break;
62         }
63
64         /*Add new head of list to heap. */
65         n.locationWithinList = list.remove();
66         minHeap.add(n);
67         max = Math.max(max, n.locationWithinList);
```

```

68     }
69
70     return new Range(bestRangeMin, bestRangeMax);
71 }
```

We're going through B elements in `getShortestClosure`, and each time pass in the for loop will take $O(\log S)$ time (the time to insert/remove from the heap). This algorithm will therefore take $O(B \log S)$ time in the worst case.

17.19 Missing Two: You are given an array with all the numbers from 1 to N appearing exactly once, except for one number that is missing. How can you find the missing number in $O(N)$ time and $O(1)$ space? What if there were two numbers missing?

pg 189

SOLUTIONS

Let's start with the first part: find a missing number in $O(N)$ time and $O(1)$ space.

Part 1: Find One Missing Number

We have a very constrained problem here. We can't store all the values (that would take $O(N)$ space) and yet, somehow, we need to have a "record" of them such that we can identify the missing number.

This suggests that we need to do some sort of computation with the values. What characteristics does this computation need to have?

- **Unique.** If this computation gives the same result on two arrays (which fit the description in the problem), then those arrays must be equivalent (same missing number). That is, the result of the computation must uniquely correspond to the specific array and missing number.
- **Reversible.** We need some way of getting from the result of the calculation to the missing number.
- **Constant Time:** The calculation can be slow, but it must be constant time per element in the array.
- **Constant Space:** The calculation can require additional memory, but it must be $O(1)$ memory.

The "unique" requirement is the most interesting—and the most challenging. What calculations can be performed on a set of numbers such that the missing number will be discoverable?

There are actually a number of possibilities.

We could do something with prime numbers. For example, for each value x in the array, we multiply `result` by the x th prime. We would then get some value that is indeed unique (since two different sets of primes can't have the same product).

Is this reversible? Yes. We could take `result` and divide it by each prime number: 2, 3, 5, 7, and so on. When we get a non-integer for the i th prime, then we know i was missing from our array.

Is it constant time and space, though? Only if we had a way of getting the i th prime number in $O(1)$ time and $O(1)$ space. We don't have that.

What other calculations could we do? We don't even need to do all this prime number stuff. Why not just multiply all the numbers together?

- **Unique?** Yes. Picture $1 * 2 * 3 * \dots * n$. Now, imagine crossing off one number. This will give us a different result than if we crossed off any other number.
- **Constant time and space?** Yes.

- **Reversible?** Let's think about this. If we compare what our product is to what it would have been without a number removed, can we find the missing number? Sure. We just divide `full_product` by `actual_product`. This will tell us which number was missing from `actual_product`.

There's just one issue: this product is really, really, really big. If n is 20, the product will be somewhere around 2,000,000,000,000,000.

We can still approach it this way, but we'll need to use the `BigInteger` class.

```

1  int missingOne(int[] array) {
2      BigInteger fullProduct = productToN(array.length + 1);
3
4      BigInteger actualProduct = new BigInteger("1");
5      for (int i = 0; i < array.length; i++) {
6          BigInteger value = new BigInteger(array[i] + "");
7          actualProduct = actualProduct.multiply(value);
8      }
9
10     BigInteger missingNumber = fullProduct.divide(actualProduct);
11     return Integer.parseInt(missingNumber.toString());
12 }
13
14 BigInteger productToN(int n) {
15     BigInteger fullProduct = new BigInteger("1");
16     for (int i = 2; i <= n; i++) {
17         fullProduct = fullProduct.multiply(new BigInteger(i + ""));
18     }
19     return fullProduct;
20 }
```

There's no need for all of this, though. We can use the sum instead. It too will be unique.

Doing the sum has another benefit: there is already a closed form expression to compute the sum of numbers between 1 and n . This is $\frac{n(n+1)}{2}$.

Most candidates probably won't remember the expression for the sum of numbers between 1 and n , and that's okay. Your interviewer might, however, ask you to derive it. Here's how to think about that: you can pair up the low and high values in the sequence of $0 + 1 + 2 + 3 + \dots + n$ to get: $(0, n) + (1, n-1) + (2, n-3)$, and so on. Each of those pairs has a sum of n and there are $\frac{n+1}{2}$ pairs. But what if n is even, such that $\frac{n+1}{2}$ is not an integer? In this case, pair up low and high values to get $\frac{n}{2}$ pairs with sum $n+1$. Either way, the math works out to $\frac{n(n+1)}{2}$.

Switching to a sum will delay the overflow issue substantially, but it won't wholly prevent it. You should discuss the issue with your interviewer to see how he/she would like you to handle it. Just mentioning it is plenty sufficient for many interviewers.

Part 2: Find Two Missing Numbers

This is substantially more difficult. Let's start with what our earlier approaches will tell us when we have two missing numbers.

- Sum: Using this approach will give us the sum of the two values that are missing.
- Product: Using this approach will give us the product of the two values that are missing.

Unfortunately, knowing the sum isn't enough. If, for example, the sum is 10, that could correspond to (1, 9), (2, 8), and a handful of other pairs. The same could be said for the product.

We're again at the same point we were in the first part of the problem. We need a calculation that can be applied such that the result is unique across all potential pairs of missing numbers.

Perhaps there is such a calculation (the prime one would work, but it's not constant time), but your interviewer probably doesn't expect you to know such math.

What else can we do? Let's go back to what we can do. We can get $x + y$ and we can also get $x * y$. Each result leaves us with a number of possibilities. But using both of them narrows it down to the specific numbers.

$$\begin{aligned}x + y &= \text{sum} \quad \rightarrow y = \text{sum} - x \\x * y &= \text{product} \quad \rightarrow x(\text{sum} - x) = \text{product} \\&\quad x * \text{sum} - x^2 = \text{product} \\&\quad x * \text{sum} - x^2 - \text{product} = 0 \\&\quad -x^2 + x * \text{sum} - \text{product} = 0\end{aligned}$$

At this point, we can apply the quadratic formula to solve for x . Once we have x , we can then compute y .

There are actually a number of other calculations you can perform. In fact, almost any other calculation (other than "linear" calculations) will give us values for x and y .

For this part, let's use a different calculation. Instead of using the product of $1 * 2 * \dots * n$, we can use the sum of the squares: $1^2 + 2^2 + \dots + n^2$. This will make the `BigInteger` usage a little less critical, as the code will at least run on small values of n . We can discuss with our interviewer whether or not this is important.

$$\begin{aligned}x + y &= s \quad \rightarrow y = s - x \\x^2 + y^2 &= t \quad \rightarrow x^2 + (s-x)^2 = t \\&\quad 2x^2 - 2sx + s^2 - t = 0\end{aligned}$$

Recall the quadratic formula:

$$x = [-b \pm \sqrt{b^2 - 4ac}] / 2a$$

where, in this case:

$$\begin{aligned}a &= 2 \\b &= -2s \\c &= s^2 - t\end{aligned}$$

Implementing this is now somewhat straightforward.

```

1 int[] missingTwo(int[] array) {
2     int max_value = array.length + 2;
3     int rem_square = squareSumToN(max_value, 2);
4     int rem_one = max_value * (max_value + 1) / 2;
5
6     for (int i = 0; i < array.length; i++) {
7         rem_square -= array[i] * array[i];
8         rem_one -= array[i];
9     }
10
11    return solveEquation(rem_one, rem_square);
12 }
13
14 int squareSumToN(int n, int power) {
15     int sum = 0;
16     for (int i = 1; i <= n; i++) {
17         sum += (int) Math.pow(i, power);
18     }
19     return sum;
20 }
```

```

21
22 int[] solveEquation(int r1, int r2) {
23     /* ax^2 + bx + c
24     * -->
25     * x = [-b +- sqrt(b^2 - 4ac)] / 2a
26     * In this case, it has to be a + not a -
27     int a = 2;
28     int b = -2 * r1;
29     int c = r1 * r1 - r2;
30
31     double part1 = -1 * b;
32     double part2 = Math.sqrt(b*b - 4 * a * c);
33     double part3 = 2 * a;
34
35     int solutionX = (int) ((part1 + part2) / part3);
36     int solutionY = r1 - solutionX;
37
38     int[] solution = {solutionX, solutionY};
39     return solution;
40 }

```

You might notice that the quadratic formula usually gives us two answers (see the + or - part), yet in our code, we only use the (+) result. We never checked the (-) answer. Why is that?

The existence of the “alternate” solution doesn’t mean that one is the correct solution and one is “fake.” It means that there are exactly two values for x which will correctly fulfill our equation: $2x^2 - 2sx + (s^2-t) = 0$.

That’s true. There are. What’s the other one? The other value is y !

If this doesn’t immediately make sense to you, remember that x and y are interchangeable. Had we solved for y earlier instead of x , we would have wound up with an identical equation: $2y^2 - 2sy + (s^2-t) = 0$. So of course y could fulfill x ’s equation and x could fulfill y ’s equation. They have the exact same equation. Since x and y are both solutions to equations that look like $2[\text{something}]^2 - 2s[\text{something}] + s^2-t = 0$, then the other something that fulfills that equation must be y .

Still not convinced? Okay, we can do some math. Let’s say we took the alternate value for x : $[-b - \sqrt{b^2 - 4ac}] / 2a$. What’s y ?

$$\begin{aligned}
 x + y &= r_1 \\
 y &= r_1 - x \\
 &= r_1 - [-b - \sqrt{b^2 - 4ac}] / 2a \\
 &= [2a * r_1 + b + \sqrt{b^2 - 4ac}] / 2a
 \end{aligned}$$

Partially plug in values for a and b , but keep the rest of the equation as-is:

$$\begin{aligned}
 &= [2(2)*r_1 + (-2r_1) + \sqrt{b^2 - 4ac}] / 2a \\
 &= [2r_1 + \sqrt{b^2 - 4ac}] / 2a
 \end{aligned}$$

Recall that $b = -2r_1$. Now, we wind up with this equation:

$$= [-b + \sqrt{b^2 - 4ac}] / 2a$$

Therefore, if we use $x = (\text{part1} + \text{part2}) / \text{part3}$, then we’ll get $(\text{part1} - \text{part2}) / \text{part3}$ for the value for y .

We don’t care which one we call x and which one we call y , so we can use either one. It’ll work out the same in the end.

- 17.20 Continuous Median:** 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 189

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()`, `maxHeap.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 Comparator<Integer> maxHeapComparator, minHeapComparator;
2 PriorityQueue<Integer> maxHeap, minHeap;
3
4 void addNewNumber(int randomNumber) {
5     /* Note: addNewNumber maintains a condition that
6      * maxHeap.size() >= minHeap.size() */
7     if (maxHeap.size() == minHeap.size()) {
8         if ((minHeap.peek() != null) &&
9             randomNumber > minHeap.peek()) {
10            maxHeap.offer(minHeap.poll());
11            minHeap.offer(randomNumber);
12        } else {
13            maxHeap.offer(randomNumber);
14        }
15    } else {
16        if (randomNumber < maxHeap.peek()) {
17            minHeap.offer(maxHeap.poll());
18            maxHeap.offer(randomNumber);
19        }
20        else {
21            minHeap.offer(randomNumber);
22        }
23    }
24 }
25
26 double getMedian() {
27     /* maxHeap is always at least as big as minHeap. So if maxHeap is empty, then
28      * minHeap is also. */
29     if (maxHeap.isEmpty()) {
30         return 0;
31     }

```

```

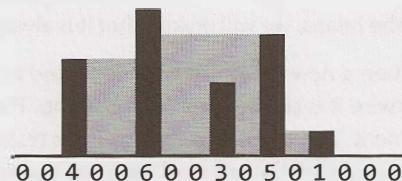
32     if (maxHeap.size() == minHeap.size()) {
33         return ((double)minHeap.peek()+(double)maxHeap.peek()) / 2;
34     } else {
35         /* If maxHeap and minHeap are of different sizes, then maxHeap must have one
36          * extra element. Return maxHeap's top element.*/
37         return maxHeap.peek();
38     }
39 }
```

17.21 Volume of Histogram: Imagine a histogram (bar graph). Design an algorithm to compute the volume of water it could hold if someone poured water across the top. You can assume that each histogram bar has width 1.

EXAMPLE

Input: {0, 0, 4, 0, 0, 6, 0, 0, 3, 0, 5, 0, 1, 0, 0, 0}

(Black bars are the histogram. Gray is water.)

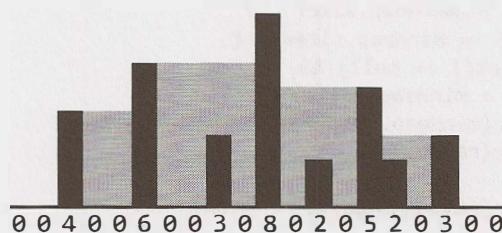


Output: 26

pg 189

SOLUTION

This is a difficult problem, so let's come up with a good example to help us solve it.



We should study this example to see what we can learn from it. What exactly dictates how big those gray areas are?

Solution #1

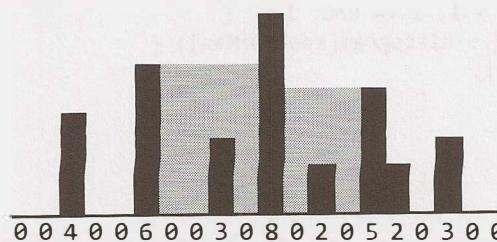
Let's look at the tallest bar, which has size 8. What role does that bar play? It plays an important role for being the highest, but it actually wouldn't matter if that bar instead had height 100. It wouldn't affect the volume.

The tallest bar forms a barrier for water on its left and right. But the volume of water is actually controlled by the next highest bar on the left and right.

- **Water on immediate left of tallest bar:** The next tallest bar on the left has height 6. We can fill up the area in between with water, but we have to deduct the height of each histogram between the tallest and next tallest. This gives a volume on the immediate left of: $(6-0) + (6-0) + (6-3) + (6-0) = 21$.
- **Water on immediate right of tallest bar:** The next tallest bar on the right has height 5. We can now

compute the volume: $(5-0) + (5-2) + (5-0) = 13$.

This just tells us part of the volume.



What about the rest?

We have essentially two subgraphs, one on the left and one on the right. To find the volume there, we repeat a very similar process.

1. Find the max. (Actually, this is given to us. The highest on the left subgraph is the right border (6) and the highest on the right subgraph is the left border (5).)
2. Find the second tallest in each subgraph. In the left subgraph, this is 4. In the right subgraph, this is 3.
3. Compute the volume between the tallest and the second tallest.
4. Recurse on the edge of the graph.

The code below implements this algorithm.

```

1  int computeHistogramVolume(int[] histogram) {
2      int start = 0;
3      int end = histogram.length - 1;
4
5      int max = findIndexOfMax(histogram, start, end);
6      int leftVolume = subgraphVolume(histogram, start, max, true);
7      int rightVolume = subgraphVolume(histogram, max, end, false);
8
9      return leftVolume + rightVolume;
10 }
11
12 /* Compute the volume of a subgraph of the histogram. One max is at either start
13 * or end (depending on isLeft). Find second tallest, then compute volume between
14 * tallest and second tallest. Then compute volume of subgraph. */
15 int subgraphVolume(int[] histogram, int start, int end, boolean isLeft) {
16     if (start >= end) return 0;
17     int sum = 0;
18     if (isLeft) {
19         int max = findIndexOfMax(histogram, start, end - 1);
20         sum += borderedVolume(histogram, max, end);
21         sum += subgraphVolume(histogram, start, max, isLeft);
22     } else {
23         int max = findIndexOfMax(histogram, start + 1, end);
24         sum += borderedVolume(histogram, start, max);
25         sum += subgraphVolume(histogram, max, end, isLeft);
26     }
27
28     return sum;
29 }
30

```

```
31 /* Find tallest bar in histogram between start and end. */
32 int findIndexOfMax(int[] histogram, int start, int end) {
33     int indexOfMax = start;
34     for (int i = start + 1; i <= end; i++) {
35         if (histogram[i] > histogram[indexOfMax]) {
36             indexOfMax = i;
37         }
38     }
39     return indexOfMax;
40 }
41
42 /* Compute volume between start and end. Assumes that tallest bar is at start and
43 * second tallest is at end. */
44 int borderedVolume(int[] histogram, int start, int end) {
45     if (start >= end) return 0;
46
47     int min = Math.min(histogram[start], histogram[end]);
48     int sum = 0;
49     for (int i = start + 1; i < end; i++) {
50         sum += min - histogram[i];
51     }
52     return sum;
53 }
```

This algorithm takes $O(N^2)$ time in the worst case, where N is the number of bars in the histogram. This is because we have to repeatedly scan the histogram to find the max height.

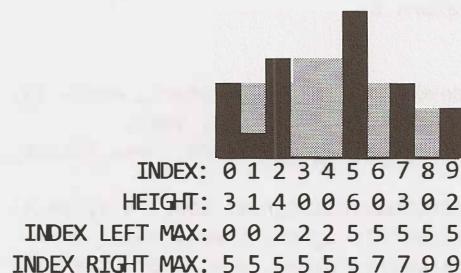
Solution #2 (Optimized)

To optimize the previous algorithm, let's think about the exact cause of the inefficiency of the prior algorithm. The root cause is the perpetual calls to `findIndexOfMax`. This suggests that it should be our focus for optimizing.

One thing we should notice is that we don't pass in arbitrary ranges into the `findIndexOfMax` function. It's actually always finding the max from one point to an edge (either the right edge or the left edge). Is there a quicker way we could know what the max height is from a given point to each edge?

Yes. We could precompute this information in $O(N)$ time.

In two sweeps through the histogram (one moving right to left and the other moving left to right), we can create a table that tells us, from any index i , the location of the max index on the right and the max index on the left.



The rest of the algorithm proceeds essentially the same way.

We've chosen to use a `HistogramData` object to store this extra information, but we could also use a two-dimensional array.

```

1 int computeHistogramVolume(int[] histogram) {
2     int start = 0;
3     int end = histogram.length - 1;
4
5     HistogramData[] data = createHistogramData(histogram);
6
7     int max = data[0].getRightMaxIndex(); // Get overall max
8     int leftVolume = subgraphVolume(data, start, max, true);
9     int rightVolume = subgraphVolume(data, max, end, false);
10
11    return leftVolume + rightVolume;
12 }
13
14 HistogramData[] createHistogramData(int[] histo) {
15     HistogramData[] histogram = new HistogramData[histo.length];
16     for (int i = 0; i < histo.length; i++) {
17         histogram[i] = new HistogramData(histo[i]);
18     }
19
20     /* Set left max index. */
21     int maxIndex = 0;
22     for (int i = 0; i < histo.length; i++) {
23         if (histo[maxIndex] < histo[i]) {
24             maxIndex = i;
25         }
26         histogram[i].setLeftMaxIndex(maxIndex);
27     }
28
29     /* Set right max index. */
30     maxIndex = histogram.length - 1;
31     for (int i = histogram.length - 1; i >= 0; i--) {
32         if (histo[maxIndex] < histo[i]) {
33             maxIndex = i;
34         }
35         histogram[i].setRightMaxIndex(maxIndex);
36     }
37
38     return histogram;
39 }
40
41 /* Compute the volume of a subgraph of the histogram. One max is at either start
42 * or end (depending on isLeft). Find second tallest, then compute volume between
43 * tallest and second tallest. Then compute volume of subgraph. */
44 int subgraphVolume(HistogramData[] histogram, int start, int end,
45                     boolean isLeft) {
46     if (start >= end) return 0;
47     int sum = 0;
48     if (isLeft) {
49         int max = histogram[end - 1].getLeftMaxIndex();
50         sum += borderedVolume(histogram, max, end);
51         sum += subgraphVolume(histogram, start, max, isLeft);
52     } else {
53         int max = histogram[start + 1].getRightMaxIndex();
54         sum += borderedVolume(histogram, start, max);

```

```

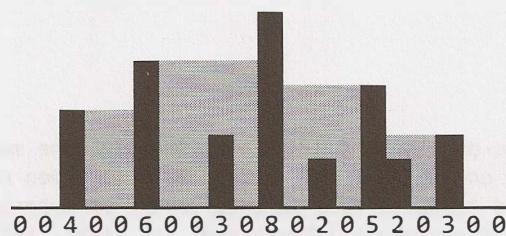
55     sum += subgraphVolume(histogram, max, end, isLeft);
56 }
57
58 return sum;
59 }
60
61 /* Compute volume between start and end. Assumes that tallest bar is at start and
62 * second tallest is at end. */
63 int borderedVolume(HistogramData[] data, int start, int end) {
64     if (start >= end) return 0;
65
66     int min = Math.min(data[start].getHeight(), data[end].getHeight());
67     int sum = 0;
68     for (int i = start + 1; i < end; i++) {
69         sum += min - data[i].getHeight();
70     }
71     return sum;
72 }
73
74 public class HistogramData {
75     private int height;
76     private int leftMaxIndex = -1;
77     private int rightMaxIndex = -1;
78
79     public HistogramData(int v) { height = v; }
80     public int getHeight() { return height; }
81     public int getLeftMaxIndex() { return leftMaxIndex; }
82     public void setLeftMaxIndex(int idx) { leftMaxIndex = idx; }
83     public int getRightMaxIndex() { return rightMaxIndex; }
84     public void setRightMaxIndex(int idx) { rightMaxIndex = idx; }
85 }

```

This algorithm takes $O(N)$ time. Since we have to look at every bar, we cannot do better than this.

Solution #3 (Optimized & Simplified)

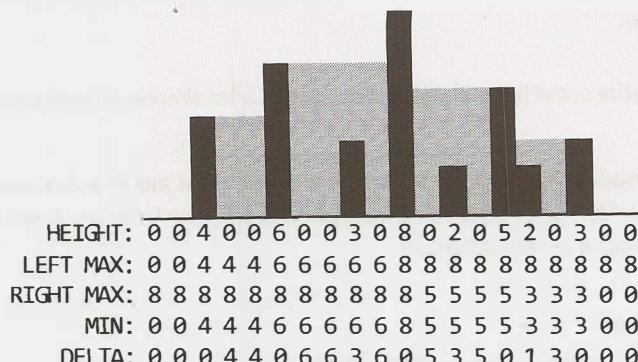
While we can't make the solution faster in terms of big O, we can make it much, much simpler. Let's look at an example again in light of what we've just learned about potential algorithms.



As we've seen, the volume of water in a particular area is determined by the tallest bar to the left and to the right (specifically, by the shorter of the two tallest bars on the left and the tallest bar on the right). For example, water fills in the area between the bar with height 6 and the bar with height 8, up to a height of 6. It's the second tallest, therefore, that determines the height.

The total volume of water is the volume of water above each histogram bar. Can we efficiently compute how much water is above each histogram bar?

Yes. In Solution #2, we were able to precompute the height of the tallest bar on the left and right of each index. The minimums of these will indicate the “water level” at a bar. The difference between the water level and the height of this bar will be the volume of water.



Our algorithm now runs in a few simple steps:

1. Sweep left to right, tracking the max height you've seen and setting left max.
2. Sweep right to left, tracking the max height you've seen and setting right max.
3. Sweep across the histogram, computing the minimum of the left max and right max for each index.
4. Sweep across the histogram, computing the delta between each minimum and the bar. Sum these deltas.

In the actual implementation, we don't need to keep so much data around. Steps 2, 3, and 4 can be merged into the same sweep. First, compute the left maxes in one sweep. Then sweep through in reverse, tracking the right max as you go. At each element, calculate the min of the left and right max and then the delta between that (the “min of maxes”) and the bar height. Add this to the sum.

```

1  /* Go through each bar and compute the volume of water above it.
2   * Volume of water at a bar =
3   *   height - min(tallest bar on left, tallest bar on right)
4   *   [where above equation is positive]
5   * Compute the left max in the first sweep, then sweep again to compute the right
6   * max, minimum of the bar heights, and the delta. */
7 int computeHistogramVolume(int[] histo) {
8     /* Get left max */
9     int[] leftMaxes = new int[histo.length];
10    int leftMax = histo[0];
11    for (int i = 0; i < histo.length; i++) {
12        leftMax = Math.max(leftMax, histo[i]);
13        leftMaxes[i] = leftMax;
14    }
15
16    int sum = 0;
17
18    /* Get right max */
19    int rightMax = histo[histo.length - 1];
20    for (int i = histo.length - 1; i >= 0; i--) {
21        rightMax = Math.max(rightMax, histo[i]);
22        int secondTallest = Math.min(rightMax, leftMaxes[i]);
23
24        /* If there are taller things on the left and right side, then there is water
25         * above this bar. Compute the volume and add to the sum. */
26        if (secondTallest > histo[i]) {

```

```
27         sum += secondTallest - histo[i];
28     }
29 }
30
31     return sum;
32 }
```

Yes, this really is the entire code! It is still $O(N)$ time, but it's a lot simpler to read and write.

17.22 Word Transformer: 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.

EXAMPLE

Input: DAMP, LIKE

Output: DAMP -> LAMP -> LIMP -> LIME -> LIKE

pg 189

SOLUTION

Let's start with a naive solution and then work our way to a more optimal solution.

Brute Force

One way of solving this problem is to just transform the words in every possible way (of course checking at each step to ensure each is a valid word), and then see if we can reach the final word.

So, for example, the word bold would be transformed into:

- aold, bold, ..., zold
- bald, bbld, ..., bzld
- bod, bobd, ..., bozd
- bola, bolb, ..., bolz

We will terminate (not pursue this path) if the string is not a valid word or if we've already visited this word.

This is essentially a depth-first search where there is an "edge" between two words if they are only one edit apart. This means that this algorithm will not find the shortest path. It will only find a path.

If we wanted to find the shortest path, we would want to use breadth-first search.

```
1  LinkedList<String> transform(String start, String stop, String[] words) {
2      HashSet<String> dict = setupDictionary(words);
3      HashSet<String> visited = new HashSet<String>();
4      return transform(visited, start, stop, dict);
5  }
6
7  HashSet<String> setupDictionary(String[] words) {
8      HashSet<String> hash = new HashSet<String>();
9      for (String word : words) {
10          hash.add(word.toLowerCase());
11      }
12      return hash;
13  }
14
15  LinkedList<String> transform(HashSet<String> visited, String startWord,
```

```

16                     String stopWord, Set<String> dictionary) {
17     if (startWord.equals(stopWord)) {
18         LinkedList<String> path = new LinkedList<String>();
19         path.add(startWord);
20         return path;
21     } else if (visited.contains(startWord) || !dictionary.contains(startWord)) {
22         return null;
23     }
24
25     visited.add(startWord);
26     ArrayList<String> words = wordsOneAway(startWord);
27
28     for (String word : words) {
29         LinkedList<String> path = transform(visited, word, stopWord, dictionary);
30         if (path != null) {
31             path.addFirst(startWord);
32             return path;
33         }
34     }
35
36     return null;
37 }
38
39 ArrayList<String> wordsOneAway(String word) {
40     ArrayList<String> words = new ArrayList<String>();
41     for (int i = 0; i < word.length(); i++) {
42         for (char c = 'a'; c <= 'z'; c++) {
43             String w = word.substring(0, i) + c + word.substring(i + 1);
44             words.add(w);
45         }
46     }
47     return words;
48 }
```

One major inefficiency in this algorithm is finding all strings that are one edit away. Right now, we're finding the strings that are one edit away and then eliminating the invalid ones.

Ideally, we want to only go to the ones that are valid.

Optimized Solution

To travel to only valid words, we clearly need a way of going from each word to a list of all the valid related words.

What makes two words "related" (one edit away)? They are one edit away if all but one character is the same. For example, ball and bill are one edit away, because they are both in the form b_ll. Therefore, one approach is to group all words that look like b_ll together.

We can do this for the whole dictionary by creating a mapping from a "wildcard word" (like b_ll) to a list of all words in this form. For example, for a very small dictionary like {all, ill, ail, ape, ale} the mapping might look like this:

```

_il -> ail
_le -> ale
_ll -> all, ill
_pe -> ape
_ae -> ape, ale
_a_l -> all, ail
```

```
i_l -> ill
ai_ -> ail
al_ -> all, ale
ap_ -> ape
il_ -> ill
```

Now, when we want to know the words that are one edit away from a word like ale, we look up _le, a_e, and al_ in the hash table.

The algorithm is otherwise essentially the same.

```
1  LinkedList<String> transform(String start, String stop, String[] words) {
2      HashMapList<String, String> wildcardToWordList = createWildcardToWordMap(words);
3      HashSet<String> visited = new HashSet<String>();
4      return transform(visited, start, stop, wildcardToWordList);
5  }
6
7  /* Do a depth-first search from startWord to stopWord, traveling through each word
8   * that is one edit away. */
9  LinkedList<String> transform(HashSet<String> visited, String start, String stop,
10                           HashMapList<String, String> wildcardToWordList) {
11     if (start.equals(stop)) {
12         LinkedList<String> path = new LinkedList<String>();
13         path.add(start);
14         return path;
15     } else if (visited.contains(start)) {
16         return null;
17     }
18
19     visited.add(start);
20     ArrayList<String> words = getValidLinkedWords(start, wildcardToWordList);
21
22     for (String word : words) {
23         LinkedList<String> path = transform(visited, word, stop, wildcardToWordList);
24         if (path != null) {
25             path.addFirst(start);
26             return path;
27         }
28     }
29
30     return null;
31 }
32
33 /* Insert words in dictionary into mapping from wildcard form -> word. */
34 HashMapList<String, String> createWildcardToWordMap(String[] words) {
35     HashMapList<String, String> wildcardToWords = new HashMapList<String, String>();
36     for (String word : words) {
37         ArrayList<String> linked = getWildcardRoots(word);
38         for (String linkedWord : linked) {
39             wildcardToWords.put(linkedWord, word);
40         }
41     }
42     return wildcardToWords;
43 }
44
45 /* Get list of wildcards associated with word. */
46 ArrayList<String> getWildcardRoots(String w) {
47     ArrayList<String> words = new ArrayList<String>();
```

```

48     for (int i = 0; i < w.length(); i++) {
49         String word = w.substring(0, i) + "_" + w.substring(i + 1);
50         words.add(word);
51     }
52     return words;
53 }
54
55 /* Return words that are one edit away. */
56 ArrayList<String> getValidLinkedWords(String word,
57     HashMapList<String, String> wildcardToWords) {
58     ArrayList<String> wildcards = getWildcardRoots(word);
59     ArrayList<String> linkedWords = new ArrayList<String>();
60     for (String wildcard : wildcards) {
61         ArrayList<String> words = wildcardToWords.get(wildcard);
62         for (String linkedWord : words) {
63             if (!linkedWord.equals(word)) {
64                 linkedWords.add(linkedWord);
65             }
66         }
67     }
68     return linkedWords;
69 }
70
71 /* HashMapList<String, String> is a HashMap that maps from Strings to
72 * ArrayList<String>. See appendix for implementation. */

```

This will work, but we can still make it faster.

One optimization is to switch from depth-first search to breadth-first search. If there are zero paths or one path, the algorithms are equivalent speeds. However, if there are multiple paths, breadth-first search may run faster.

Breadth-first search finds the shortest path between two nodes, whereas depth-first search finds any path. This means that depth-first search might take a very long, windy path in order to find a connection when, in fact, the nodes were quite close.

Optimal Solution

As noted earlier, we can optimize this using breadth-first search. Is this as fast as we can make it? Not quite.

Imagine that the path between two nodes has length 4. With breadth-first search, we will visit about 15^4 nodes to find them.

Breadth-first search spans out very quickly.

Instead, what if we searched out from the source and destination nodes simultaneously? In this case, the breadth-first searches would collide after each had done about two levels each.

- Nodes travelled to from source: 15^2
- Nodes travelled to from destination: 15^2
- Total nodes: $15^2 + 15^2$

This is much better than the traditional breadth-first search.

We will need to track the path that we've travelled at each node.

To implement this approach, we've used an additional class `BFSData`. `BFSData` helps us keep things a bit clearer, and allows us to keep a similar framework for the two simultaneous breadth-first searches. The alternative is to keep passing around a bunch of separate variables.

```
1  LinkedList<String> transform(String startWord, String stopWord, String[] words) {
2      HashMapList<String, String> wildcardToWordList = getWildcardToWordList(words);
3
4      BFSData sourceData = new BFSData(startWord);
5      BFSData destData = new BFSData(stopWord);
6
7      while (!sourceData.isFinished() && !destData.isFinished()) {
8          /* Search out from source. */
9          String collision = searchLevel(wildcardToWordList, sourceData, destData);
10         if (collision != null) {
11             return mergePaths(sourceData, destData, collision);
12         }
13
14         /* Search out from destination. */
15         collision = searchLevel(wildcardToWordList, destData, sourceData);
16         if (collision != null) {
17             return mergePaths(sourceData, destData, collision);
18         }
19     }
20
21     return null;
22 }
23
24 /* Search one level and return collision, if any. */
25 String searchLevel(HashMapList<String, String> wildcardToWordList,
26                     BFSData primary, BFSData secondary) {
27     /* We only want to search one level at a time. Count how many nodes are
28     * currently in the primary's level and only do that many nodes. We'll continue
29     * to add nodes to the end. */
30     int count = primary.toVisit.size();
31     for (int i = 0; i < count; i++) {
32         /* Pull out first node. */
33         PathNode pathNode = primary.toVisit.poll();
34         String word = pathNode.getWord();
35
36         /* Check if it's already been visited. */
37         if (secondary.visited.containsKey(word)) {
38             return pathNode.getWord();
39         }
40
41         /* Add friends to queue. */
42         ArrayList<String> words = getValidLinkedWords(word, wildcardToWordList);
43         for (String w : words) {
44             if (!primary.visited.containsKey(w)) {
45                 PathNode next = new PathNode(w, pathNode);
46                 primary.visited.put(w, next);
47                 primary.toVisit.add(next);
48             }
49         }
50     }
51     return null;
52 }
```

```

54 LinkedList<String> mergePaths(BFSData bfs1, BFSData bfs2, String connection) {
55     PathNode end1 = bfs1.visited.get(connection); // end1 -> source
56     PathNode end2 = bfs2.visited.get(connection); // end2 -> dest
57     LinkedList<String> pathOne = end1.collapse(false); // forward
58     LinkedList<String> pathTwo = end2.collapse(true); // reverse
59     pathTwo.removeFirst(); // remove connection
60     pathOne.addAll(pathTwo); // add second path
61     return pathOne;
62 }
63
64 /* Methods getWildcardRoots, getWildcardToWordList, and getValidLinkedWords are
65 * the same as in the earlier solution. */
66
67 public class BFSData {
68     public Queue<PathNode> toVisit = new LinkedList<PathNode>();
69     public HashMap<String, PathNode> visited = new HashMap<String, PathNode>();
70
71     public BFSData(String root) {
72         PathNode sourcePath = new PathNode(root, null);
73         toVisit.add(sourcePath);
74         visited.put(root, sourcePath);
75     }
76
77     public boolean isFinished() {
78         return toVisit.isEmpty();
79     }
80 }
81
82 public class PathNode {
83     private String word = null;
84     private PathNode previousNode = null;
85     public PathNode(String word, PathNode previous) {
86         this.word = word;
87         previousNode = previous;
88     }
89
90     public String getWord() {
91         return word;
92     }
93
94     /* Traverse path and return linked list of nodes. */
95     public LinkedList<String> collapse(boolean startsWithRoot) {
96         LinkedList<String> path = new LinkedList<String>();
97         PathNode node = this;
98         while (node != null) {
99             if (startsWithRoot) {
100                 path.addLast(node.word);
101             } else {
102                 path.addFirst(node.word);
103             }
104             node = node.previousNode;
105         }
106         return path;
107     }
108 }
109

```

```
110 /* HashMapList<String, Integer> is a HashMap that maps from Strings to  
111 * ArrayList<Integer>. See appendix for implementation. */
```

This algorithm's runtime is a bit harder to describe since it depends on what the language looks like, as well as the actual source and destination words. One way of expressing it is that if each word has E words that are one edit away and the source and destination are distance D , the runtime is $O(E^{D/2})$. This is how much work each breadth-first search does.

Of course, this is a lot of code to implement in an interview. It just wouldn't be possible. More realistically, you'd leave out a lot of the details. You might write just the skeleton code of `transform` and `searchLevel`, but leave out the rest.

17.23 Max Square Matrix: 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 190

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 length sz. */  
11    int count = matrix.length - squareSize + 1;  
12  
13    /* Iterate through all squares with side length squareSize. */  
14    for (int row = 0; row < count; row++) {  
15        for (int col = 0; col < count; col++) {  
16            if (isSquare(matrix, row, col, squareSize)) {  
17                return new Subsquare(row, col, squareSize);  
18            }  
19        }  
20    }  
21    return null;  
22 }  
23  
24 boolean isSquare(int[][] matrix, int row, int col, int size) {
```

```

25    // Check top and bottom border.
26    for (int j = 0; j < size; j++){
27        if (matrix[row][col+j] == 1) {
28            return false;
29        }
30        if (matrix[row+size-1][col+j] == 1){
31            return false;
32        }
33    }
34
35    // Check left and right border.
36    for (int i = 1; i < size - 1; i++){
37        if (matrix[row+i][col] == 1){
38            return false;
39        }
40        if (matrix[row+i][col+size-1] == 1) {
41            return false;
42        }
43    }
44    return true;
45 }

```

Pre-Processing Solution: $O(N^3)$

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																		
<table border="1"> <tr> <td>0,0</td><td>1,3</td><td>0,0</td></tr> <tr> <td>2,2</td><td>1,2</td><td>0,0</td></tr> <tr> <td>2,1</td><td>1,1</td><td>0,0</td></tr> </table>	0,0	1,3	0,0	2,2	1,2	0,0	2,1	1,1	0,0	<table border="1"> <tr> <td>W</td><td>B</td><td>W</td></tr> <tr> <td>B</td><td>B</td><td>W</td></tr> <tr> <td>B</td><td>B</td><td>W</td></tr> </table>	W	B	W	B	B	W	B	B	W
0,0	1,3	0,0																	
2,2	1,2	0,0																	
2,1	1,1	0,0																	
W	B	W																	
B	B	W																	
B	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 }
15
16 Subsquare findSquareWithSize(SquareCell[][] processed, int size) {
17     /* equivalent to first algorithm */
18 }
19
20 boolean isSquare(SquareCell[][] matrix, int row, int col, int sz) {
21     SquareCell topLeft = matrix[row][col];
22     SquareCell topRight = matrix[row][col + sz - 1];
23     SquareCell bottomLeft = matrix[row + sz - 1][col];
24
25     /* Check top, left, right, and bottom edges, respectively. */
26     if (topLeft.zerosRight < sz || topLeft.zerosBelow < sz ||
27         topRight.zerosBelow < sz || bottomLeft.zerosRight < sz) {
28         return false;
29     }
30     return true;
31 }
32
33 SquareCell[][] processSquare(int[][] matrix) {
34     SquareCell[][] processed =
35         new SquareCell[matrix.length][matrix.length];
36
37     for (int r = matrix.length - 1; r >= 0; r--) {
38         for (int c = matrix.length - 1; c >= 0; c--) {
39             int rightZeros = 0;
40             int belowZeros = 0;
41             // only need to process if it's a black cell
42             if (matrix[r][c] == 0) {
43                 rightZeros++;
44                 belowZeros++;
45                 // next column over is on same row
46                 if (c + 1 < matrix.length) {
47                     SquareCell previous = processed[r][c + 1];
48                     rightZeros += previous.zerosRight;
49                 }
50                 if (r + 1 < matrix.length) {
51                     SquareCell previous = processed[r + 1][c];
52                     belowZeros += previous.zerosBelow;
53                 }
54             }
55             processed[r][c] = new SquareCell(rightZeros, belowZeros);
56         }
57     }
58     return processed;
}
```

59 }

- 17.24 Max Submatrix:** Given an $N \times N$ matrix of positive and negative integers, write code to find the submatrix with the largest possible sum.

pg 190

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 largest.

To iterate through all possible submatrices (with no duplicates), we simply need to iterate through all ordered 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.

```

1  SubMatrix getMaxMatrix(int[][] matrix) {
2      int rowCount = matrix.length;
3      int columnCount = matrix[0].length;
4      SubMatrix best = null;
5      for (int row1 = 0; row1 < rowCount; row1++) {
6          for (int row2 = row1; row2 < rowCount; row2++) {
7              for (int col1 = 0; col1 < columnCount; col1++) {
8                  for (int col2 = col1; col2 < columnCount; col2++) {
9                      int sum = sum(matrix, row1, col1, row2, col2);
10                     if (best == null || best.getSum() < sum) {
11                         best = new SubMatrix(row1, col1, row2, col2, sum);
12                     }
13                 }
14             }
15         }
16     }
17     return best;
18 }
19
20 int sum(int[][] matrix, int row1, int col1, int row2, int col2) {
21     int sum = 0;
22     for (int r = row1; r <= row2; r++) {
23         for (int c = col1; c <= col2; c++) {
24             sum += matrix[r][c];
25         }
26     }
27     return sum;
28 }
29
30 public class SubMatrix {
31     private int row1, row2, col1, col2, sum;
32     public SubMatrix(int r1, int c1, int r2, int c2, int sm) {
33         row1 = r1;
34         col1 = c1;

```

```

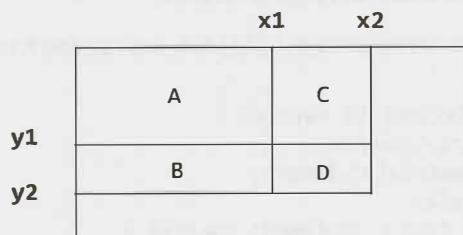
35     row2 = r2;
36     col2 = c2;
37     sum = sm;
38 }
39
40 public int getSum() {
41     return sum;
42 }
43 }
```

It is good practice to pull the sum code into its own function since it's a fairly distinct set of code.

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:

$$\text{area}(D) = \text{ValD} - \text{area}(A \cup C) - \text{area}(A \cup B) + \text{area}(A).$$

Or, written another way:

$$\text{area}(D) = \text{ValD} - \text{ValB} - \text{ValC} + \text{ValA}$$

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 SubMatrix getMaxMatrix(int[][] matrix) {
2     SubMatrix best = null;
3     int rowCount = matrix.length;
4     int columnCount = matrix[0].length;
5     int[][] sumThrough = precomputeSums(matrix);
6
7     for (int row1 = 0; row1 < rowCount; row1++) {
8         for (int row2 = row1; row2 < rowCount; row2++) {
9             for (int col1 = 0; col1 < columnCount; col1++) {
10                for (int col2 = col1; col2 < columnCount; col2++) {
```

```

11         int sum = sum(sumThrough, row1, col1, row2, col2);
12         if (best == null || best.getSum() < sum) {
13             best = new SubMatrix(row1, col1, row2, col2, sum);
14         }
15     }
16 }
17 }
18 }
19 return best;
20 }
21
22 int[][] precomputeSums(int[][] matrix) {
23     int[][] sumThrough = new int[matrix.length][matrix[0].length];
24     for (int r = 0; r < matrix.length; r++) {
25         for (int c = 0; c < matrix[0].length; c++) {
26             int left = c > 0 ? sumThrough[r][c - 1] : 0;
27             int top = r > 0 ? sumThrough[r - 1][c] : 0;
28             int overlap = r > 0 && c > 0 ? sumThrough[r-1][c-1] : 0;
29             sumThrough[r][c] = left + top - overlap + matrix[r][c];
30         }
31     }
32     return sumThrough;
33 }
34
35 int sum(int[][] sumThrough, int r1, int c1, int r2, int c2) {
36     int topAndLeft = r1 > 0 && c1 > 0 ? sumThrough[r1-1][c1-1] : 0;
37     int left = c1 > 0 ? sumThrough[r2][c1 - 1] : 0;
38     int top = r1 > 0 ? sumThrough[r1 - 1][c2] : 0;
39     int full = sumThrough[r2][c2];
40     return full - left - top + topAndLeft;
41 }

```

This algorithm takes $O(N^4)$ time, since it goes through each pair of rows and each pair of columns.

Optimized Solution: $O(N^3)$

Believe it or not, an even more optimal solution exists. If we have R rows and C columns, we can solve it in $O(R^2C)$ 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 gives us the highest sum. That is:

```

1 maxSum = 0
2 foreach rowStart in rows
3     foreach rowEnd in rows
4         /* We have many possible submatrices with rowStart and rowEnd as the top and
5            * bottom edges of the matrix. Find the colStart and colEnd edges that give
6            * the highest sum. */
7         maxSum = max(runningMaxSum, maxSum)
8     return maxSum

```

Now the question is, how do we efficiently find the "best" colStart and colEnd?

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				

Given a `rowStart` and `rowEnd`, we want to find the `colStart` and `colEnd` that give us the highest possible sum. 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 columns. 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 matrix[rowEnd, col]
6              runningMaxSum = maxSubArray(partialSum)
7              maxSum = max(runningMaxSum, maxSum)
8  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 $a[0] \dots a[i]$ from scratch, even though in the previous iteration of the outer for loop, we already added up $a[0] \dots a[i-1]$. 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  SubMatrix getMaxMatrix(int[][] matrix) {
2      int rowCount = matrix.length;
3      int colCount = matrix[0].length;
4      SubMatrix best = null;
5
6      for (int rowStart = 0; rowStart < rowCount; rowStart++) {
7          int[] partialSum = new int[colCount];
8
9          for (int rowEnd = rowStart; rowEnd < rowCount; rowEnd++) {
10              /* Add values at row rowEnd. */
11              for (int i = 0; i < colCount; i++) {

```

```

12         partialSum[i] += matrix[rowEnd][i];
13     }
14
15     Range bestRange = maxSubArray(partialSum, colCount);
16     if (best == null || best.getSum() < bestRange.sum) {
17         best = new SubMatrix(rowStart, bestRange.start, rowEnd,
18                               bestRange.end, bestRange.sum);
19     }
20 }
21 }
22 return best;
23 }
24
25 Range maxSubArray(int[] array, int N) {
26     Range best = null;
27     int start = 0;
28     int sum = 0;
29
30     for (int i = 0; i < N; i++) {
31         sum += array[i];
32         if (best == null || sum > best.sum) {
33             best = new Range(start, i, sum);
34         }
35
36         /* If running_sum is < 0 no point in trying to continue the series. Reset. */
37         if (sum < 0) {
38             start = i + 1;
39             sum = 0;
40         }
41     }
42     return best;
43 }
44
45 public class Range {
46     public int start, end, sum;
47     public Range(int start, int end, int sum) {
48         this.start = start;
49         this.end = end;
50         this.sum = sum;
51     }
52 }

```

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.

17.25 Word Rectangle: 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 190

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 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 6x5 rectangle and the first few 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 `maxWordLength2`) 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 rectangle of length i and height j. Note that i * j = z. */
9                     Rectangle rectangle = makeRectangle(i, j);
```

```

10         if (rectangle != null) return rectangle;
11     }
12   }
13 }
14 }
15 return null;
16 }
```

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 || groupList[height-1] == null) {
3     return null;
4   }
5
6   /* Create trie for word length if we haven't yet */
7   if (trieList[height - 1] == null) {
8     LinkedList<String> words = groupList[height - 1].getWords();
9     trieList[height - 1] = new Trie(words);
10  }
11
12  return makePartialRectangle(length, height, new Rectangle(length));
13 }
```

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     }
6     return null;
7   }
8
9   /* Compare columns to trie to see if potentially valid rect */
10  if (!rectangle.isPartialOK(l, trieList[h - 1])) {
11    return null;
12  }
13
14  /* Go through all words of the right length. Add each one to the current partial
15   * rectangle, and attempt to build a rectangle recursively. */
16  for (int i = 0; i < groupList[l-1].length(); i++) {
17    /* Create a new rectangle which is this rect + new word. */
18    Rectangle orgPlus = rectangle.append(groupList[l-1].getWord(i));
19
20    /* Try to build a rectangle with this new, partial rect */
21    Rectangle rect = makePartialRectangle(l, h, orgPlus);
22    if (rect != null) {
23      return rect;
24    }
25  }
26}
```

```

24     }
25 }
26 return null;
27 }

```

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” rectangle. Length is fixed, but height varies as we add
6      * words. */
7      public Rectangle(int l) {
8          height = 0;
9          length = l;
10     }
11
12     / *Construct a rectangular array of letters of the specified length and height,
13     * and backed by the specified matrix of letters. (It is assumed that the length
14     * and height specified as arguments are consistent with the array argument’s
15     * dimensions.) */
16     public Rectangle(int length, int height, char[][] letters) {
17         this.height = letters.length;
18         this.length = letters[0].length;
19         matrix = letters;
20     }
21
22     public char getLetter (int i, int j) { return matrix[i][j]; }
23     public String getColumn(int i) { ... }
24
25     / *Check if all columns are valid. All rows are already known to be valid since
26     * they were added directly from dictionary. */
27     public boolean isComplete(int l, int h, WordGroup groupList) {
28         if (height == h) {
29             / *Check if each column is a word in the dictionary. */
30             for (int i = 0; i < l; i++) {
31                 String col = getColumn(i);
32                 if (!groupList.containsWord(col)) {
33                     return false;
34                 }
35             }
36             return true;
37         }
38         return false;
39     }
40
41     public boolean isPartialOK(int l, Trie trie) {
42         if (height == 0) return true;
43         for (int i = 0; i < l; i++ ) {
44             String col = getColumn(i);
45             if (!trie.contains(col)) {
46                 return false;
47             }
48         }

```

```

49     return true;
50 }
51
52 / *Create a new Rectangle by taking the rows of the current rectangle and
53 * appending s. */
54 public Rectangle append(String s) { ... }
55 }
```

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 HashMap<String, Boolean> lookup = new HashMap<String, Boolean>();
3      private ArrayList<String> group = new ArrayList<String>();
4      public boolean containsWord(String s) { return lookup.containsKey(s); }
5      public int length() { return group.size(); }
6      public String getWord(int i) { return group.get(i); }
7      public ArrayList<String> getWords() { return group; }
8
9      public void addWord (String s) {
10         group.add(s);
11         lookup.put(s, true);
12     }
13
14     public static WordGroup[] createWordGroups(String[] list) {
15         WordGroup[] groupList;
16         int maxWordLength = 0;
17         / *Find the length of the longest word */
18         for (int i = 0; i < list.length; i++) {
19             if (list[i].length() > maxWordLength) {
20                 maxWordLength = list[i].length();
21             }
22         }
23
24         / *Group the words in the dictionary into lists of words of same length.
25          * groupList[i] will contain a list of words, each of length (i+1). */
26         groupList = new WordGroup[maxWordLength];
27         for (int i = 0; i < list.length; i++) {
28             / *We do wordLength - 1 instead of just wordLength since this is used as
29              * an index and no words are of length 0 */
30             int wordLength = list[i].length() - 1;
31             if (groupList[wordLength] == null) {
32                 groupList[wordLength] = new WordGroup();
33             }
34             groupList[wordLength].addWord(list[i]);
35         }
36         return groupList;
37     }
38 }
```

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.

17.26 Sparse Similarity: The similarity of two documents (each with distinct words) is defined to be the size of the intersection divided by the size of the union. For example, if the documents consist of integers, the similarity of {1, 5, 3} and {1, 7, 2, 3} is 0.4, because the intersection has size 2 and the union has size 5.

We have a long list of documents (with distinct values and each with an associated ID) where the similarity is believed to be “sparse.” That is, any two arbitrarily selected documents are very likely to have similarity 0. Design an algorithm that returns a list of pairs of document IDs and the associated similarity.

Print only the pairs with similarity greater than 0. Empty documents should not be printed at all. For simplicity, you may assume each document is represented as an array of distinct integers.

EXAMPLE

Input:

```
13: {14, 15, 100, 9, 3}  
16: {32, 1, 9, 3, 5}  
19: {15, 29, 2, 6, 8, 7}  
24: {7, 10}
```

Output:

```
ID1, ID2 : SIMILARITY  
13, 19   : 0.1  
13, 16   : 0.25  
19, 24   : 0.14285714285714285
```

pg 190

SOLUTION

This sounds like quite a tricky problem, so let's start off with a brute force algorithm. If nothing else, it will help wrap our heads around the problem.

Remember that each document is an array of distinct “words”, and each is just an integer.

Brute Force

A brute force algorithm is as simple as just comparing all arrays to all other arrays. At each comparison, we compute the size of the intersection and size of the union of the two arrays.

Note that we only want to print this pair if the similarity is greater than 0. The union of two arrays can never be zero (unless both arrays are empty, in which case we don't want them printed anyway). Therefore, we are really just printing the similarity if the intersection is greater than 0.

How do we compute the size of the intersection and the union?

The intersection means the number of elements in common. Therefore, we can just iterate through the first array (A) and check if each element is in the second array (B). If it is, increment an `intersection` variable.

To compute the union, we need to be sure that we don't double count elements that are in both. One way to do this is to count up all the elements in A that are *not* in B. Then, add in all the elements in B. This will avoid double counting as the duplicate elements are only counted with B.

Alternatively, we can think about it this way. If we *did* double count elements, it would mean that elements in the intersection (in both A and B) were counted twice. Therefore, the easy fix is to just remove these duplicate elements.

```
union(A, B) = A + B - intersection(A, B)
```

This means that all we really need to do is compute the intersection. We can derive the union, and therefore similarity, from that immediately.

This gives us an $O(AB)$ algorithm, just to compare two arrays (or documents).

However, we need to do this for all pairs of D documents. If we assume each document has at most W words then the runtime is $O(D^2 W^2)$.

Slightly Better Brute Force

As a quick win, we can optimize the computation for the similarity of two arrays. Specifically, we need to optimize the intersection computation.

We need to know the number of elements in common between the two arrays. We can throw all of A's elements into a hash table. Then we iterate through B, incrementing intersection every time we find an element in A.

This takes $O(A + B)$ time. If each array has size W and we do this for D arrays, then this takes $O(D^2 W)$.

Before implementing this, let's first think about the classes we'll need.

We'll need to return a list of document pairs and their similarities. We'll use a `DocPair` class for this. The exact return type will be a hash table that maps from `DocPair` to a double representing the similarity.

```
1  public class DocPair {
2      public int doc1, doc2;
3
4      public DocPair(int d1, int d2) {
5          doc1 = d1;
6          doc2 = d2;
7      }
8
9      @Override
10     public boolean equals(Object o) {
11         if (o instanceof DocPair) {
12             DocPair p = (DocPair) o;
13             return p.doc1 == doc1 && p.doc2 == doc2;
14         }
15         return false;
16     }
17
18     @Override
19     public int hashCode() { return (doc1 * 31) ^ doc2; }
20 }
```

It will also be useful to have a class that represents the documents.

```
1  public class Document {
2      private ArrayList<Integer> words;
3      private int docId;
4
5      public Document(int id, ArrayList<Integer> w) {
6          docId = id;
7          words = w;
8      }
9
10     public ArrayList<Integer> getWords() { return words; }
11     public int getId() { return docId; }
```

```
12     public int size() { return words == null ? 0 : words.size(); }
13 }
```

Strictly speaking, we don't need any of this. However, readability is important, and it's a lot easier to read `ArrayList<Document>` than `ArrayList<ArrayList<Integer>>`.

Doing this sort of thing not only shows good coding style, it also makes your life in an interview a lot easier. You have to write a lot less. (You probably would not define the entire `Document` class, unless you had extra time or your interviewer asked you to.)

```
1  HashMap<DocPair, Double> computeSimilarities(ArrayList<Document> documents) {
2      HashMap<DocPair, Double> similarities = new HashMap<DocPair, Double>();
3      for (int i = 0; i < documents.size(); i++) {
4          for (int j = i + 1; j < documents.size(); j++) {
5              Document doc1 = documents.get(i);
6              Document doc2 = documents.get(j);
7              double sim = computeSimilarity(doc1, doc2);
8              if (sim > 0) {
9                  DocPair pair = new DocPair(doc1.getId(), doc2.getId());
10                 similarities.put(pair, sim);
11             }
12         }
13     }
14     return similarities;
15 }
16
17 double computeSimilarity(Document doc1, Document doc2) {
18     int intersection = 0;
19     HashSet<Integer> set1 = new HashSet<Integer>();
20     set1.addAll(doc1.getWords());
21
22     for (int word : doc2.getWords()) {
23         if (set1.contains(word)) {
24             intersection++;
25         }
26     }
27
28     double union = doc1.size() + doc2.size() - intersection;
29     return intersection / union;
30 }
```

Observe what's happening on line 28. Why did we make `union` a `double`, when it's obviously an `integer`?

We did this to avoid an integer division bug. If we didn't do this, the division would "round" down to an integer. This would mean that the similarity would almost always return 0. Oops!

Slightly Better Brute Force (Alternate)

If the documents were sorted, you could compute the intersection between two documents by walking through them in sorted order, much like you would when doing a sorted merge of two arrays.

This would take $O(A + B)$ time. This is the same time as our current algorithm, but less space. Doing this on D documents with W words each would take $O(D^2 W)$ time.

Since we don't know that the arrays are sorted, we could first sort them. This would take $O(D * W \log W)$ time. The full runtime then is $O(D * W \log W + D^2 W)$.

We cannot necessarily assume that the second part “dominates” the first one, because it doesn’t necessarily. It depends on the relative size of D and $\log W$. Therefore, we need to keep both terms in our runtime expression.

Optimized (Somewhat)

It is useful to create a larger example to really understand the problem.

```
13: {14, 15, 100, 9, 3}
16: {32, 1, 9, 3, 5}
19: {15, 29, 2, 6, 8, 7}
24: {7, 10, 3}
```

At first, we might try various techniques that allow us to more quickly eliminate potential comparisons. For example, could we compute the min and max values in each array? If we did that, then we’d know that arrays with no overlap in ranges don’t need to be compared.

The problem is that this doesn’t really fix our runtime issue. Our best runtime thus far is $O(D^2 W)$. With this change, we’re still going to be comparing all $O(D^2)$ pairs, but the $O(W)$ part might go to $O(1)$ sometimes. That $O(D^2)$ part is going to be a really big problem when D gets large.

Therefore, let’s focus on reducing that $O(D^2)$ factor. That is the “bottleneck” in our solution. Specifically, this means that, given a document $docA$, we want to find all documents with some similarity—and we want to do this without “talking” to each document.

What would make a document similar to $docA$? That is, what characteristics define the documents with similarity > 0 ?

Suppose $docA$ is $\{14, 15, 100, 9, 3\}$. For a document to have similarity > 0 , it needs to have a 14, a 15, a 100, a 9, or a 3. How can we quickly gather a list of all documents with one of those elements?

The slow (and, really, only way) is to read every single word from every single document to find the documents that contain a 14, a 15, a 100, a 9, or a 3. That will take $O(DW)$ time. Not good.

However, note that we’re doing this repeatedly. We can reuse the work from one call to the next.

If we build a hash table that maps from a word to all documents that contain that word, we can very quickly know the documents that overlap with $docA$.

```
1 -> 16
2 -> 19
3 -> 13, 16, 24
5 -> 16
6 -> 19
7 -> 19, 24
8 -> 19
9 -> 13, 16
...

```

When we want to know all the documents that overlap with $docA$, we just look up each of $docA$ ’s items in this hash table. We’ll then get a list of all documents with some overlap. Now, all we have to do is compare $docA$ to each of those documents.

If there are P pairs with similarity > 0 , and each document has W words, then this will take $O(PW)$ time (plus $O(DW)$ time to create and read this hash table). Since we expect P to be much less than D^2 , this is much better than before.

Optimized (Better)

Let's think about our previous algorithm. Is there any way we can make it more optimal?

If we consider the runtime— $O(PW + DW)$ —we probably can't get rid of the $O(DW)$ factor. We have to touch each word at least once, and there are $O(DW)$ words. Therefore, if there's an optimization to be made, it's probably in the $O(PW)$ term.

It would be difficult to eliminate the P part in $O(PW)$ because we have to at least print all P pairs (which takes $O(P)$ time). The best place to focus, then, is on the W part. Is there some way we can do less than $O(W)$ work for each pair of similar documents?

One way to tackle this is to analyze what information the hash table gives us. Consider this list of documents:

```
12: {1, 5, 9}  
13: {5, 3, 1, 8}  
14: {4, 3, 2}  
15: {1, 5, 9, 8}  
17: {1, 6}
```

If we look up document 12's elements in a hash table for this document, we'll get:

```
1 -> {12, 13, 15, 17}  
5 -> {12, 13, 15}  
9 -> {12, 15}
```

This tells us that documents 13, 15, and 17 have some similarity. Under our current algorithm, we would now need to compare document 12 to documents 13, 15, and 17 to see the number of elements document 12 has in common with each (that is, the size of the intersection). The union can be computed from the document sizes and the intersection, as we did before.

Observe, though, that document 13 appeared twice in the hash table, document 15 appeared three times, and document 17 appeared once. We discarded that information. But can we use it instead? What does it indicate that some documents appeared multiple times and others didn't?

Document 13 appeared twice because it has two elements (1 and 5) in common. Document 17 appeared once because it has only one element (1) in common. Document 15 appeared three times because it has three elements (1, 5, and 9) in common. This information can actually directly give us the size of the intersection.

We could go through each document, look up the items in the hash table, and then count how many times each document appears in each item's lists. There's a more direct way to do it.

1. As before, build a hash table for a list of documents.
2. Create a new hash table that maps from a document pair to an integer (which will indicate the size of the intersection).
3. Read the first hash table by iterating through each list of documents.
4. For each list of documents, iterate through the pairs in that list. Increment the intersection count for each pair.

Comparing this runtime to the previous one is a bit tricky. One way we can look at it is to realize that before we were doing $O(W)$ work for each similar pair. That's because once we noticed that two documents were similar, we touched every single word in each document. With this algorithm, we're only touching the words that actually overlap. The worst cases are still the same, but for many inputs this algorithm will be faster.

```
1  HashMap<DocPair, Double>  
2  computeSimilarities(HashMap<Integer, Document> documents) {
```

```

3  HashMapList<Integer, Integer> wordToDocs = groupWords(documents);
4  HashMap<DocPair, Double> similarities = computeIntersections(wordToDocs);
5  adjustToSimilarities(documents, similarities);
6  return similarities;
7 }
8
9 /* Create hash table from each word to where it appears. */
10 HashMapList<Integer, Integer> groupWords(HashMap<Integer, Document> documents) {
11  HashMapList<Integer, Integer> wordToDocs = new HashMapList<Integer, Integer>();
12
13  for (Document doc : documents.values()) {
14      ArrayList<Integer> words = doc.getWords();
15      for (int word : words) {
16          wordToDocs.put(word, doc.getId());
17      }
18  }
19
20  return wordToDocs;
21 }
22
23 /* Compute intersections of documents. Iterate through each list of documents and
24 * then each pair within that list, incrementing the intersection of each page. */
25 HashMap<DocPair, Double> computeIntersections(
26     HashMapList<Integer, Integer> wordToDocs {
27     HashMap<DocPair, Double> similarities = new HashMap<DocPair, Double>();
28     Set<Integer> words = wordToDocs.keySet();
29     for (int word : words) {
30         ArrayList<Integer> docs = wordToDocs.get(word);
31         Collections.sort(docs);
32         for (int i = 0; i < docs.size(); i++) {
33             for (int j = i + 1; j < docs.size(); j++) {
34                 increment(similarities, docs.get(i), docs.get(j));
35             }
36         }
37     }
38
39     return similarities;
40 }
41
42 /* Increment the intersection size of each document pair. */
43 void increment(HashMap<DocPair, Double> similarities, int doc1, int doc2) {
44     DocPair pair = new DocPair(doc1, doc2);
45     if (!similarities.containsKey(pair)) {
46         similarities.put(pair, 1.0);
47     } else {
48         similarities.put(pair, similarities.get(pair) + 1);
49     }
50 }
51
52 /* Adjust the intersection value to become the similarity. */
53 void adjustToSimilarities(HashMap<Integer, Document> documents,
54                         HashMap<DocPair, Double> similarities) {
55     for (Entry<DocPair, Double> entry : similarities.entrySet()) {
56         DocPair pair = entry.getKey();
57         Double intersection = entry.getValue();
58         Document doc1 = documents.get(pair.doc1);

```

```
59     Document doc2 = documents.get(pair.doc2);
60     double union = (double) doc1.size() + doc2.size() - intersection;
61     entry.setValue(intersection / union);
62 }
63 }
64
65 /* HashMapList<Integer, Integer> is a HashMap that maps from Integer to
66 * ArrayList<Integer>. See appendix for implementation. */
```

For a set of documents with sparse similarity, this will run much faster than the original naive algorithm, which compares all pairs of documents directly.

Optimized (Alternative)

There's an alternative algorithm that some candidates might come up with. It's slightly slower, but still quite good.

Recall our earlier algorithm that computed the similarity between two documents by sorting them. We can extend this approach to multiple documents.

Imagine we took all of the words, tagged them by their original document, and then sorted them. The prior list of documents would look like this:

$1_{12}, 1_{13}, 1_{15}, 1_{16}, 2_{14}, 3_{13}, 3_{14}, 4_{14}, 5_{12}, 5_{13}, 5_{15}, 6_{16}, 8_{13}, 8_{15}, 9_{12}, 9_{15}$

Now we have essentially the same approach as before. We iterate through this list of elements. For each sequence of identical elements, we increment the intersection counts for the corresponding pair of documents.

We will use an Element class to group together documents and words. When we sort the list, we will sort first on the word but break ties on the document ID.

```
1 class Element implements Comparable<Element> {
2     public int word, document;
3     public Element(int w, int d) {
4         word = w;
5         document = d;
6     }
7
8     /* When we sort the words, this function will be used to compare the words. */
9     public int compareTo(Element e) {
10         if (word == e.word) {
11             return document - e.document;
12         }
13         return word - e.word;
14     }
15 }
16
17 HashMap<DocPair, Double> computeSimilarities(
18     HashMap<Integer, Document> documents) {
19     ArrayList<Element> elements = sortWords(documents);
20     HashMap<DocPair, Double> similarities = computeIntersections(elements);
21     adjustToSimilarities(documents, similarities);
22     return similarities;
23 }
24
25 /* Throw all words into one list, sorting by the word and then the document. */
26 ArrayList<Element> sortWords(HashMap<Integer, Document> docs) {
27     ArrayList<Element> elements = new ArrayList<Element>();
```

```

28     for (Document doc : docs.values()) {
29         ArrayList<Integer> words = doc.getWords();
30         for (int word : words) {
31             elements.add(new Element(word, doc.getId()));
32         }
33     }
34     Collections.sort(elements);
35     return elements;
36 }
37
38 /* Increment the intersection size of each document pair. */
39 void increment(HashMap<DocPair, Double> similarities, int doc1, int doc2) {
40     DocPair pair = new DocPair(doc1, doc2);
41     if (!similarities.containsKey(pair)) {
42         similarities.put(pair, 1.0);
43     } else {
44         similarities.put(pair, similarities.get(pair) + 1);
45     }
46 }
47
48 /* Adjust the intersection value to become the similarity. */
49 HashMap<DocPair, Double> computeIntersections(ArrayList<Element> elements) {
50     HashMap<DocPair, Double> similarities = new HashMap<DocPair, Double>();
51
52     for (int i = 0; i < elements.size(); i++) {
53         Element left = elements.get(i);
54         for (int j = i + 1; j < elements.size(); j++) {
55             Element right = elements.get(j);
56             if (left.word != right.word) {
57                 break;
58             }
59             increment(similarities, left.document, right.document);
60         }
61     }
62     return similarities;
63 }
64
65 /* Adjust the intersection value to become the similarity. */
66 void adjustToSimilarities(HashMap<Integer, Document> documents,
67                         HashMap<DocPair, Double> similarities) {
68     for (Entry<DocPair, Double> entry : similarities.entrySet()) {
69         DocPair pair = entry.getKey();
70         Double intersection = entry.getValue();
71         Document doc1 = documents.get(pair.doc1);
72         Document doc2 = documents.get(pair.doc2);
73         double union = (double) doc1.size() + doc2.size() - intersection;
74         entry.setValue(intersection / union);
75     }
76 }

```

The first step of this algorithm is slower than that of the prior algorithm, since it has to sort rather than just add to a list. The second step is essentially equivalent.

Both will run much faster than the original naive algorithm.

Advanced Topics

XI

This section includes topics that are mostly beyond the scope of interviews but can come up on occasion. Interviewers shouldn't be surprised if you don't know these topics well. Feel free to dive into these topics if you want to. If you're pressed for time, they're low priority.

XI

Advanced Topics

When writing the 6th edition, I had a number of debates about what should and shouldn't be included. Red-black trees? Dijkstra's algorithm? Topological sort?

On one hand, I'd had a number of requests to include these topics. Some people insisted that these topics are asked "all the time" (in which case, they have a very different idea of what this phrase means!). There was clearly a desire—at least from some people—to include them. And learning more can't hurt, right?

On the other hand, I know these topics to be rarely asked. It happens, of course. Interviewers are individuals and might have their own ideas of what is "fair game" or "relevant" for an interview. But it's rare. When it does come up, if you don't know the topic, it's unlikely to be a big red flag.

Admittedly, as an interviewer, I *have* asked candidates questions where the solution was essentially an application of one of these algorithms. On the rare occasions that a candidate already knew the algorithm, they did not benefit from this knowledge (nor were they hurt by it). I want to evaluate your ability to solve a problem you haven't seen before. So, I'll take into account whether you know the underlying algorithm in advance.

I believe in giving people a fair expectation of the interview, not scaring people into excess studying. I also have no interest in making the book more "advanced" so as to help book sales, at the expense of your time and energy. That's not fair or right to do to you.

(Additionally, I didn't want to give interviewers—who I know to be reading this—the impression that they can or should be covering these more advanced topics. Interviewers: If you ask about these topics, you're testing knowledge of algorithms. You're just going to wind up eliminating a lot of perfectly smart people.)

But there are many borderline "important" topics. They're not often asked, but sometimes they are.

Ultimately, I decided to leave the decision in your hands. After all, you know better than I do how thorough you want to be in your preparation. If you want to do an extra thorough job, read this. If you just love learning data structures and algorithms, read this. If you want to see new ways of approaching problems, read this.

But if you're pressed for time, this studying isn't a super high priority.

► Useful Math

Here's some math that can be useful in some questions. There are more formal proofs that you can look up online, but we'll focus here on giving you the intuition behind them. You can think of these as informal proofs.

XI. Advanced Topics

Sum of Integers 1 through N

What is $1 + 2 + \dots + n$? Let's figure it out by pairing up low values with high values.

If n is even, we pair 1 with n , 2 with $n - 1$, and so on. We will have $\frac{n}{2}$ pairs each with sum $n + 1$.

If n is odd, we pair 0 with n , 1 with $n - 1$, and so on. We will have $\frac{n+1}{2}$ pairs with sum n .

n is even			
pair #	a	b	a + b
1	1	n	$n + 1$
2	2	$n - 1$	$n + 1$
3	3	$n - 2$	$n + 1$
4	4	$n - 3$	$n + 1$
...
$\frac{n}{2}$	$\frac{n}{2}$	$\frac{n}{2} + 1$	$n + 1$
total:	$\frac{n}{2} * (n+1)$		

n is odd			
pair #	a	b	a + b
1	0	n	n
2	1	$n - 1$	n
3	2	$n - 2$	n
4	3	$n - 3$	n
...
$\frac{n+1}{2}$	$\frac{n-1}{2}$	$\frac{n+1}{2}$	n
total:	$\frac{n+1}{2} * n$		

In either case, the sum is $\frac{n(n+1)}{2}$.

This reasoning comes up a lot in nested loops. For example, consider the following code:

```
1  for (int i = 0; i < n; i++) {  
2      for (int j = i + 1; j < n; j++) {  
3          System.out.println(i + j);  
4      }  
5  }
```

On the first iteration of the outer for loop, the inner for loop iterates $n - 1$ times. On the second iteration of the outer for loop, the inner for loop iterates $n - 2$ times. Next, $n - 3$, then $n - 4$, and so on. There are $\frac{n(n-1)}{2}$ total iterations of the inner for loop. Therefore, this code takes $O(n^2)$ time.

Sum of Powers of 2

Consider this sequence: $2^0 + 2^1 + 2^2 + \dots + 2^n$. What is its result?

A nice way to see this is by looking at these values in binary.

	Power	Binary	Decimal
	2^0	00001	1
	2^1	00010	2
	2^2	00100	4
	2^3	01000	8
	2^4	10000	16
sum:	$2^5 - 1$	11111	$32 - 1 = 31$

Therefore, the sum of $2^0 + 2^1 + 2^2 + \dots + 2^n$ would, in base 2, be a sequence of $(n + 1)$ 1s. This is $2^{n+1} - 1$.

Takeaway: The sum of a sequence of powers of two is roughly equal to the next value in the sequence.

Bases of Logs

Suppose we have something in \log_2 (log base 2). How do we convert that to \log_{10} ? That is, what's the relationship between $\log_b k$ and $\log_x k$?

Let's do some math. Assume $c = \log_b k$ and $y = \log_x k$.

$$\begin{aligned} \log_b k = c \rightarrow b^c = k & \quad // \text{ This is the definition of log.} \\ \log_x(b^c) = \log_x k & \quad // \text{ Take log of both sides of } b^c = k. \\ c \log_x b = \log_x k & \quad // \text{ Rules of logs. You can move out the exponents.} \\ c = \log_b k = \frac{\log_x k}{\log_x b} & \quad // \text{ Dividing above expression and substituting } c. \end{aligned}$$

Therefore, if we want to convert $\log_2 p$ to $\log_{10} p$, we just do this:

$$\log_{10} p = \frac{\log_2 p}{\log_2 10}$$

Takeaway: Logs of different bases are only off by a constant factor. For this reason, we largely ignore what the base of a log within a big O expression. It doesn't matter since we drop constants anyway.

Permutations

How many ways are there of rearranging a string of n unique characters? Well, you have n options for what to put in the first characters, then $n - 1$ options for what to put in the second slot (one option is taken), then $n - 2$ options for what to put in the third slot, and so on. Therefore, the total number of strings is $n!$.

$$n! = n * n - 1 * n - 2 * n - 3 * \dots * 1$$

What if you were forming a k -length string (with all unique characters) from n total unique characters? You can follow similar logic, but you'd just stop your selection/multiplication earlier.

$$\frac{n!}{(n-k)!} = n * n - 1 * n - 2 * n - 3 * \dots * n - k + 1$$

Combinations

Suppose you have a set of n distinct characters. How many ways are there of selecting k characters into a new set (where order doesn't matter)? That is, how many k -sized subsets are there out of n distinct elements? This is what the expression n -choose- k means, which is often written $\binom{n}{k}$.

Imagine we made a list of all the sets by first writing all k -length substrings and then taking out the duplicates.

From the above *Permutations* section, we'd have $\frac{n!}{(n-k)!}$ k -length substrings.

Since each k -sized subset can be rearranged $k!$ unique ways into a string, each subset will be duplicated $k!$ times in this list of substrings. Therefore, we need to divide by $k!$ to take out these duplicates.

$$\binom{n}{k} = \frac{1}{k!} * \frac{n!}{(n-k)!} = \frac{n!}{k!(n-k)!}$$

Proof by Induction

Induction is a way of proving something to be true. It is closely related to recursion. It takes the following form.

Task: Prove statement $P(k)$ is true for all $k \geq b$.

- Base Case: Prove the statement is true for $P(b)$. This is usually just a matter of plugging in numbers.
- Assumption: Assume the statement is true for $P(n)$.
- Inductive Step: Prove that if the statement is true for $P(n)$, then it's true for $P(n+1)$.

This is like dominoes. If the first domino falls, and one domino always knocks over the next one, then all the dominoes must fall.

Let's use this to prove that there are 2^n subsets of an n -element set.

- Definitions: let $S = \{a_1, a_2, a_3, \dots, a_n\}$ be the n -element set.

- Base case: Prove there are 2^0 subsets of $\{\}$. This is true, since the only subset of $\{\}$ is $\{\}$.

- Assume that there are 2^n subsets of $\{a_1, a_2, a_3, \dots, a_n\}$.

- Prove that there are 2^{n+1} subsets of $\{a_1, a_2, a_3, \dots, a_{n+1}\}$.

Consider the subsets of $\{a_1, a_2, a_3, \dots, a_{n+1}\}$. Exactly half will contain a_{n+1} and half will not.

The subsets that do not contain a_{n+1} are just the subsets of $\{a_1, a_2, a_3, \dots, a_n\}$. We assumed there are 2^n of those.

Since we have the same number of subsets with x as without x , there are 2^n subsets with a_{n+1} .

Therefore, we have $2^n + 2^n$ subsets, which is 2^{n+1} .

Many recursive algorithms can be proved valid with induction.

► Topological Sort

A topological sort of a directed graph is a way of ordering the list of nodes such that if (a, b) is an edge in the graph then a will appear before b in the list. If a graph has cycles or is not directed, then there is no topological sort.

There are a number of applications for this. For example, suppose the graph represents parts on an assembly line. The edge (Handle, Door) indicates that you need to assemble the handle before the door. The topological sort would offer a valid ordering for the assembly line.

We can construct a topological sort with the following approach.

1. Identify all nodes with no incoming edges and add those nodes to our topological sort.

- » We know those nodes are safe to add first since they have nothing that needs to come before them. Might as well get them over with!
- » We know that such a node must exist if there's no cycle. After all, if we picked an arbitrary node we could just walk edges backwards arbitrarily. We'll either stop at some point (in which case we've found a node with no incoming edges) or we'll return to a prior node (in which case there is a cycle).

2. When we do the above, remove each node's outbound edges from the graph.

- » Those nodes have already been added to the topological sort, so they're basically irrelevant. We can't violate those edges anymore.

3. Repeat the above, adding nodes with no incoming edges and removing their outbound edges. When all the nodes have been added to the topological sort, then we are done.

More formally, the algorithm is this:

1. Create a queue `order`, which will eventually store the valid topological sort. It is currently empty.
2. Create a queue `processNext`. This queue will store the next nodes to process.
3. Count the number of incoming edges of each node and set a class variable `Node.inbound`. Nodes typically only store their outgoing edges. However, you can count the inbound edges by walking through each node n and, for each of its outgoing edges (n, x) , incrementing $x.inbound$.
4. Walk through the nodes again and add to `processNext` any node where $x.inbound == 0$.
5. While `processNext` is not empty, do the following:
 - » Remove first node n from `processNext`.

- » For each edge (n, x) , decrement $x.inbound$. If $x.inbound == 0$, append x to $processNext$.
 - » Append n to $order$.
6. If $order$ contains all the nodes, then it has succeeded. Otherwise, the topological sort has failed due to a cycle.

This algorithm does sometimes come up in interview questions. Your interviewer probably wouldn't expect you to know it offhand. However, it would be reasonable to have you derive it even if you've never seen it before.

► Dijkstra's Algorithm

In some graphs, we might want to have edges with weights. If the graph represented cities, each edge might represent a road and its weight might represent the travel time. In this case, we might want to ask, just as your GPS mapping system does, what's the shortest path from your current location to another point p ? This is where Dijkstra's algorithm comes in.

Dijkstra's algorithm is a way to find the shortest path between two points in a weighted directed graph (which might have cycles). All edges must have positive values.

Rather than just stating what Dijkstra's algorithm is, let's try to derive it. Consider the earlier described graph. We could find the shortest path from s to t by literally taking all possible routes using actual time. (Oh, and we'll need a machine to clone ourselves.)

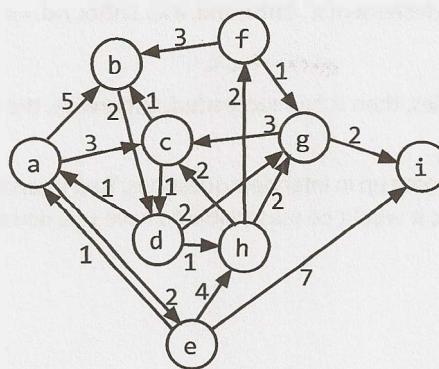
1. Start off at s .
2. For each of s 's outbound edges, clone ourselves and start walking. If the edge (s, x) has weight 5, we should actually take 5 minutes to get there.
3. Each time we get to a node, check if anyone's been there before. If so, then just stop. We're automatically not as fast as another path since someone beat us here from s . If no one has been here before, then clone ourselves and head out in all possible directions.
4. The first one to get to t wins.

This works just fine. But, of course, in the real algorithm we don't want to literally use a timer to find the shortest path.

Imagine that each clone could jump immediately from one node to its adjacent nodes (regardless of the edge weight), but it kept a `time_so_far` log of how long its path would have taken if it did walk at the "true" speed. Additionally, only one person moves at a time, and it's always the one with the lowest `time_so_far`. This is sort of how Dijkstra's algorithm works.

Dijkstra's algorithm finds the minimum weight path from a start node s to every node on the graph.

Consider the following graph.



Assume we are trying to find the shortest path from a to i . We'll use Dijkstra's algorithm to find the shortest path from a to all other nodes, from which we will clearly have the shortest path from a to i .

We first initialize several variables:

- `path_weight[node]`: maps from each node to the total weight of the shortest path. All values are initialized to infinity, except for `path_weight[a]` which is initialized to 0.
- `previous[node]`: maps from each node to the previous node in the (current) shortest path.
- `remaining`: a priority queue of all nodes in the graph, where each node's priority is defined by its `path_weight`.

Once we've initialized these values, we can start adjusting the values of `path_weight`.

A (min) **priority queue** is an abstract data type that—at least in this case—supports insertion of an object and key, removing the object with the minimum key, and decreasing a key. (Think of it like a typical queue, except that, instead of removing the oldest item, it removes the item with the lowest or highest priority.) It is an abstract data type because it is defined by its behavior (its operations). Its underlying implementation can vary. You could implement a priority queue with an array or a min (or max) heap (or many other data structures).

We iterate through the nodes in `remaining` (until `remaining` is empty), doing the following:

1. Select the node in `remaining` with the lowest value in `path_weight`. Call this node n .
2. For each adjacent node, compare `path_weight[x]` (which is the weight of the current shortest path from a to x) to `path_weight[n] + edge_weight[(n, x)]`. That is, could we get a path from a to x with lower weight by going through n instead of our current path? If so, update `path_weight` and `previous`.
3. Remove n from `remaining`.

When `remaining` is empty, then `path_weight` stores the weight of the current shortest path from a to each node. We can reconstruct this path by tracing through `previous`.

Let's walk through this on the above graph.

1. The first value of n is a . We look at its adjacent nodes (b , c , and e), update the values of `path_weight` (to 5, 3, and 2) and `previous` (to a) and then remove a from `remaining`.
2. Then, we go to the next smallest node, which is e . We previously updated `path_weight[e]` to be 2. Its adjacent nodes are h and i , so we update `path_weight` (to 6 and 9) and `previous` for both of those.

Observe that 6 is `path_weight[e]` (which is 2) + the weight of the edge (`e, h`) (which is 4).

- The next smallest node is `c`, which has `path_weight` 3. Its adjacent nodes are `b` and `d`. The value of `path_weight[d]` is infinity, so we update it to 4 (which is `path_weight[c]` + weight(edge `c, d`)). The value of `path_weight[b]` has been previously set to 5. However, since `path_weight[c] + weight(edge c, b)` (which is $3 + 1 = 4$) is less than 5, we update `path_weight[b]` to 4 and previous to `c`. This indicates that we would improve the path from `a` to `b` by going through `c`.

We continue doing this until `remaining` is empty. The following diagram shows the changes to the `path_weight` (left) and `previous` (right) at each step. The topmost row shows the current value for `n` (the node we are removing from `remaining`). We black out a row after it has been removed from `remaining`.

	INITIAL		<code>n=a</code>		<code>n=e</code>		<code>n=c</code>		<code>n=b</code>		<code>n=d</code>		<code>n=h</code>		<code>n=g</code>		<code>n=f</code>		FINAL		
	wt	pr	wt	pr	wt	pr	wt	pr	wt	pr	wt	pr	wt	pr	wt	pr	wt	pr	wt	pr	
a	0	-																	0	-	
b	∞	-	5	a			4	c											4	c	
c	∞	-	3	a															3	a	
d	∞	-					4	c											4	c	
e	∞	-	2	a															2	a	
f	∞	-												7	h				7	h	
g	∞	-									6	d							6	d	
h	∞	-			6	e				5	d								5	d	
i	∞	-	∞	-	9	e										8	g			8	g

Once we're done, we can follow this chart backwards, starting at `i` to find the actual path. In this case, the smallest weight path has weight 8 and is `a -> c -> d -> g -> i`.

Priority Queue and Runtime

As mentioned earlier, our algorithm used a priority queue, but this data structure can be implemented in different ways.

The runtime of this algorithm depends heavily on the implementation of the priority queue. Assume you have v vertices and e nodes.

- If you implemented the priority queue with an array, then you would call `remove_min` up to v times. Each operation would take $O(v)$ time, so you'd spend $O(v^2)$ time in the `remove_min` calls. Additionally, you would update the values of `path_weight` and `previous` at most once per edge, so that's $O(e)$ time doing those updates. Observe that e must be less than or equal to v^2 since you can't have more edges than there are pairs of vertices. Therefore, the total runtime is $O(v^2)$.
- If you implemented the priority queue with a min heap, then the `remove_min` calls will each take $O(\log v)$ time (as will inserting and updating a key). We will do one `remove_min` call for each vertex, so that's $O(v \log v)$ (v vertices at $O(\log v)$ time each). Additionally, on each edge, we might call one update key or insert operation, so that's $O(e \log v)$. The total runtime is $O((v + e) \log v)$.

Which one is better? Well, that depends. If the graph has a lot of edges, then v^2 will be close to e . In this case, you might be better off with the array implementation, as $O(v^2)$ is better than $O((v + v^2) \log v)$. However, if the graph is sparse, then e is much less than v^2 . In this case, the min heap implementation may be better.

► Hash Table Collision Resolution

Essentially any hash table can have collisions. There are a number of ways of handling this.

Chaining with Linked Lists

With this approach (which is the most common), the hash table's array maps to a linked list of items. We just add items to this linked list. As long as the number of collisions is fairly small, this will be quite efficient.

In the worst case, lookup is $O(n)$, where n is the number of elements in the hash table. This would only happen with either some very strange data or a very poor hash function (or both).

Chaining with Binary Search Trees

Rather than storing collisions in a linked list, we could store collisions in a binary search tree. This will bring the worst-case runtime to $O(\log n)$.

In practice, we would rarely take this approach unless we expected an extremely nonuniform distribution.

Open Addressing with Linear Probing

In this approach, when a collision occurs (there is already an item stored at the designated index), we just move on to the next index in the array until we find an open spot. (Or, sometimes, some other fixed distance, like the `index + 5`.)

If the number of collisions is low, this is a very fast and space-efficient solution.

One obvious drawback of this is that the total number of entries in the hash table is limited by the size of the array. This is not the case with chaining.

There's another issue here. Consider a hash table with an underlying array of size 100 where indexes 20 through 29 are filled (and nothing else). What are the odds of the next insertion going to index 30? The odds are 10% because an item mapped to any index between 20 and 30 will wind up at index 30. This causes an issue called *clustering*.

Quadratic Probing and Double Hashing

The distance between probes does not need to be linear. You could, for example, increase the probe distance quadratically. Or, you could use a second hash function to determine the probe distance.

► Rabin-Karp Substring Search

The brute force way to search for a substring S in a larger string B takes $O(s(b-s))$ time, where s is the length of S and b is the length of B . We do this by searching through the first $b - s + 1$ characters in B and, for each, checking if the next s characters match S .

The Rabin-Karp algorithm optimizes this with a little trick: if two strings are the same, they must have the same hash value. (The converse, however, is not true. Two different strings can have the same hash value.)

Therefore, if we efficiently precompute a hash value for each sequence of s characters within B , we can find the locations of S in $O(b)$ time. We then just need to validate that those locations really do match S .

For example, imagine our hash function was simply the sum of each character (where space = 0, a = 1, b = 2, and so on). If S is ear and B = doe are hearing me, we'd then just be looking for sequences where the sum is 24 (e + a + r). This happens three times. For each of those locations, we'd check if the string really is ear.

char:	d	o	e		a	r	e		h	e	a	r	i	n	g		m	e
code:	4	15	5	0	1	18	5	0	8	5	1	18	9	14	7	0	13	5
sum of next 3:	24	20	6	19	24	23	13	13	14	24	28	41	30	21	20	18		

If we computed these sums by doing $\text{hash}(\text{'doe'})$, then $\text{hash}(\text{'oe'})$, then $\text{hash}(\text{'e a'})$, and so on, we would still be at $O(s(b-s))$ time.

Instead, we compute the hash values by recognizing that $\text{hash}(\text{'oe'}) = \text{hash}(\text{'doe'}) - \text{code}(\text{'d'}) + \text{code}(\text{' '})$. This takes $O(b)$ time to compute all the hashes.

You might argue that, still, in the worst case this will take $O(s(b-s))$ time since many of the hash values could match. That's absolutely true—for this hash function.

In practice, we would use a better *rolling hash function*, such as the Rabin fingerprint. This essentially treats a string like *doe* as a base 128 (or however many characters are in our alphabet) number.

$$\text{hash}(\text{'doe'}) = \text{code}(\text{'d'}) * 128^2 + \text{code}(\text{'o'}) * 128^1 + \text{code}(\text{'e'}) * 128^0$$

This hash function will allow us to remove the *d*, shift the *o* and *e*, and then add in the space.

$$\text{hash}(\text{'oe'}) = (\text{hash}(\text{'doe'}) - \text{code}(\text{'d'}) * 128^2) * 128 + \text{code}(\text{' '})$$

This will considerably cut down on the number of false matches. Using a good hash function like this will give us expected time complexity of $O(s + b)$, although the worst case is $O(sb)$.

Usage of this algorithm comes up fairly frequently in interviews, so it's useful to know that you can identify substrings in linear time.

► AVL Trees

An AVL tree is one of two common ways to implement tree balancing. We will only discuss insertions here, but you can look up deletions separately if you're interested.

Properties

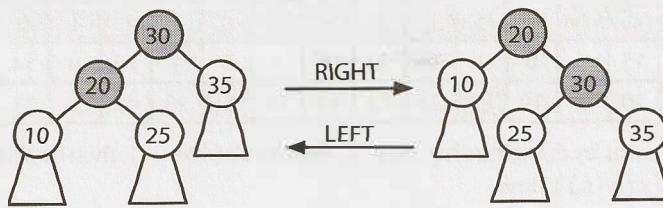
An AVL tree stores in each node the height of the subtrees rooted at this node. Then, for any node, we can check if it is height balanced: that the height of the left subtree and the height of the right subtree differ by no more than one. This prevents situations where the tree gets too lopsided.

$$\begin{aligned}\text{balance}(n) &= n.\text{left.height} - n.\text{right.height} \\ -1 &\leq \text{balance}(n) \leq 1\end{aligned}$$

Inserts

When you insert a node, the balance of some nodes might change to -2 or 2. Therefore, when we "unwind" the recursive stack, we check and fix the balance at each node. We do this through a series of rotations.

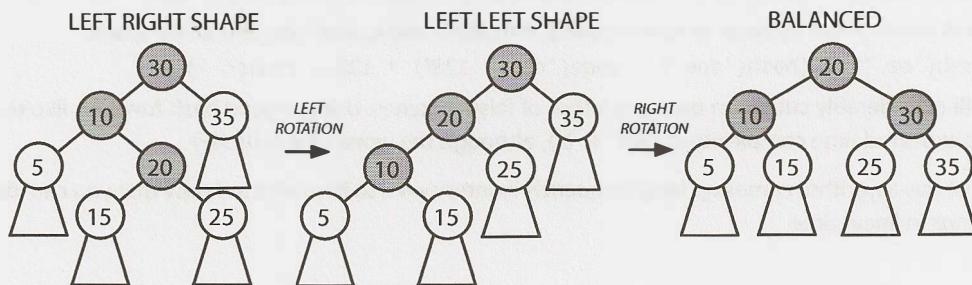
Rotations can be either left or right rotations. The right rotation is an inverse of the left rotation.



Depending on the balance and where the imbalance occurs, we fix it in a different way.

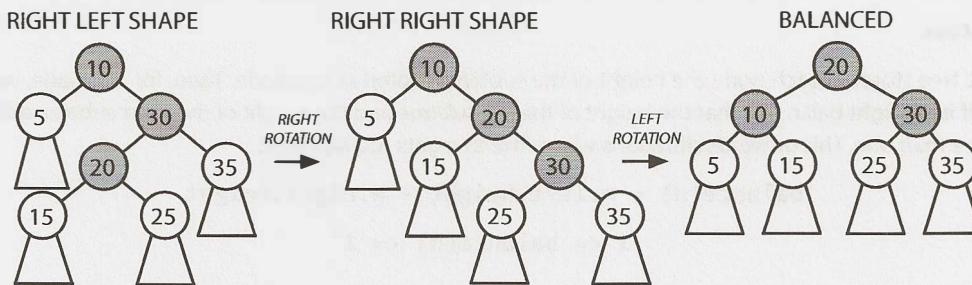
- *Case 1: Balance is 2.*

In this case, the left's height is two bigger than the right's height. If the left side is larger, the left subtree's extra nodes must be hanging to the left (as in LEFT LEFT SHAPE) or hanging to the right (as in LEFT RIGHT SHAPE). If it looks like the LEFT RIGHT SHAPE, transform it with the rotations below into the LEFT LEFT SHAPE then into BALANCED. If it looks like the LEFT LEFT SHAPE already, just transform it into BALANCED.



- *Case 2: Balance is -2.*

This case is the mirror image of the prior case. The tree will look like either the RIGHT LEFT SHAPE or the RIGHT RIGHT SHAPE. Perform the rotations below to transform it into BALANCED.



In both cases, “balanced” just means that the balance of the tree is between -1 and 1. It does not mean that the balance is 0.

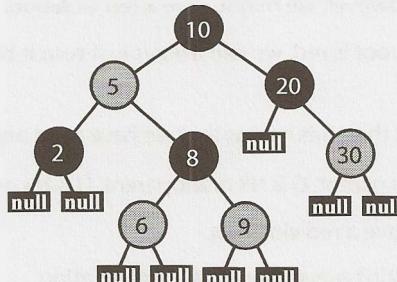
We recurse up the tree, fixing any imbalances. If we ever achieve a balance of 0 on a subtree, then we know that we have completed all the balances. This portion of the tree will not cause another, higher subtree to have a balance of -2 or 2. If we were doing this non-recursively, then we could break from the loop.

► Red-Black Trees

Red-black trees (a type of self-balancing binary search tree) do not ensure quite as strict balancing, but the balancing is still good enough to ensure $O(\log N)$ insertions, deletions, and retrievals. They require a bit less memory and can rebalance faster (which means faster insertions and removals), so they are often used in situations where the tree will be modified frequently.

Red-black trees operate by enforcing a quasi-alternating red and black coloring (under certain rules, described below) and then requiring every path from a node to its leaves to have the same number of black nodes. Doing so leads to a reasonably balanced tree.

The tree below is a red-black tree (where the red nodes are indicated with gray):



Properties

1. Every node is either red or black.
2. The root is black.
3. The leaves, which are NULL nodes, are considered black.
4. Every red node must have two black children. That is, a red node cannot have red children (although a black node can have black children).
5. Every path from a node to its leaves must have the same number of black children.

Why It Balances

Property #4 means that two red nodes cannot be adjacent in a path (e.g., parent and child). Therefore, no more than half the nodes in a path can be red.

Consider two paths from a node (say, the root) to its leaves. The paths must have the same number of black nodes (property #5), so let's assume that their red node counts are as different as possible: one path contains the minimum number of red nodes and the other one contains the maximum number.

- Path 1 (Min Red): The minimum number of red nodes is zero. Therefore, path 1 has b nodes total.
- Path 2 (Max Red): The maximum number of red nodes is b , since red nodes must have black children and there are b black nodes. Therefore, path 2 has $2b$ nodes total.

Therefore, even in the most extreme case, the lengths of paths cannot differ by more than a factor of two. That's good enough to ensure an $O(\log N)$ find and insert runtime.

If we can maintain these properties, we'll have a (sufficiently) balanced tree—good enough to ensure $O(\log N)$ insert and find, anyway. The question then is how to maintain these properties efficiently. We'll only discuss insertion here, but you can look up deletion on your own.