Section 8.5

Introduction to Threads

Like people, computers can multitask. That is, they can be working on several different tasks at the same time. A computer that has just a single central processing unit can't literally do two things at the same time, any more than a person can, but it can still switch its attention back and forth among several tasks. Furthermore, it is increasingly common for computers to have more than one processing unit, and such computers can literally work on several tasks simultaneously. It is likely that from now on, most of the increase in computing power will come from adding additional processors to computers rather than from increasing the speed of individual processors. To use the full power of these multiprocessing computers, a programmer must do parallel programming, which means writing a program as a set of several tasks that can be executed simultaneously. Even on a single-processor computer, parallel programming techniques can be useful, since some problems can be tackled most naturally by breaking the solution into a set of simultaneous tasks that cooperate to solve the problem.

In Java, a single task is called a thread. The term "thread" refers to a "thread of control" or "thread of execution," meaning a sequence of instructions that are executed one after another -- the thread extends through time, connecting each instruction to the next. In a multithreaded program, there can be many threads of control, weaving through time in parallel and forming the complete fabric of the program. (Ok, enough with the metaphor, already!) Every Java program has at least one thread; when the Java virtual machine runs your program, it creates a thread that is responsible for executing the main routine of the program. This main thread can in turn create other threads that can continue even after the main thread has terminated. In a GUI program, there is at least one additional thread, which is responsible for handling events and drawing components on the screen. This GUI thread is created when the first window is opened. So in fact, you have already done parallel programming! When a main routine opens a window, both the main thread and the GUI thread can continue to run in parallel. Of course, parallel programming can be used in much more interesting ways.

Unfortunately, parallel programming is even more difficult than ordinary, single-threaded programming. When several threads are working together on a problem, a whole new category of errors is possible. This just means that techniques for writing correct and robust programs are even more important for parallel programming than they are for normal programming. (That's one excuse for having this section in this chapter -- another is that we will need threads at several points in future chapters, and I didn't have another place in the book where the topic fits more naturally.) Since threads are a difficult topic, you will probably not fully understand everything in this section the first time through the material. Your understanding should improve as you encounter more examples of threads in future sections.

8.5.1 Creating and Running Threads

In Java, a thread is represented by an object belonging to the class java.lang.Thread (or to a subclass of this class). The purpose of a Thread object is to execute a single method. The method is executed in its own thread of control, which can run in parallel with other threads. When the execution of the method is finished, either because the method terminates normally or because of an uncaught exception, the thread stops running. Once this happens, there is no way to restart the thread or to use the same Thread object to start another thread.

There are two ways to program a thread. One is to create a subclass of Thread and to define the method public void run() in the subclass. This run() method defines the task that will be performed by the thread; that is, when the thread is started, it is the run() method that will be executed in the thread. For example, here is a simple, and rather useless, class that defines a thread that does nothing but print a message on standard output:

```
public class NamedThread extends Thread {
   private String name;  // The name of this thread.
   public NamedThread(String name) {  // Constructor gives name to
thread.
      this.name = name;
```

```
    }
    public void run() {  // The run method prints a message to standard output.
        System.out.println("Greetings from thread '" + name + "'!");
    }
}
```

To use a NamedThread, you must of course create an object belonging to this class. For example,

```
NamedThread greetings = new NamedThread("Fred");
```

However, creating the object does not automatically start the thread running. To do that, you must call the start() method in the thread object. For the example, this would be done with the statement

```
greetings.start();
```

The purpose of the start() method is to create a new thread of control that will execute the Thread object's run() method. The new thread runs in parallel with the thread in which the start() method was called, along with any other threads that already existed. This means that the code in the run() method will execute at the same time as the statements that follow the call to greetings.start(). Consider this code segment:

```
NamedThread greetings = new NamedThread("Fred");
greetings.start();
System.out.println("Thread has been started.");
```

After greetings.start() is executed, there are two threads. One of them will print "Thread has been started." while the other one wants to print "Greetings from thread 'Fred'!". It is important to note that *these messages can be printed in either order*. The two threads run simultaneously and will compete for access to standard output, so that they can print their messages. Whichever thread happens to be the first to get access will be the first to print its message. In a normal, single-threaded program, things happen in a definite, predictable order from beginning to end. In a multi-threaded program, there is a fundamental indeterminacy. You can't be sure what order things will happen in. This indeterminacy is what makes parallel programming so difficult!

Note that calling greetings.start() is **very** different from calling greetings.run(). Calling greetings.run() will execute the run() method in the same thread, rather than creating a new thread. This means that all the work of the run() will be done before the computer moves on to the statement that follows the call to greetings.run() in the program. There is no parallelism and no indeterminacy.

I mentioned that there are two ways to program a thread. The first way was to define a subclass of Thread. The second is to define a class that implements the interface java.lang.Runnable. The Runnable interface defines a single method, public void run(). An object that implements the Runnable interface can be passed as a parameter to the constructor of an object of type Thread. When that thread's start method is called, the thread will execute the run() method in the Runnable object. For example, as an alternative to the NamedThread class, we could define the class:

```
public class NamedRunnable implements Runnable {
   private String name;  // The name of this thread.
   public NamedRunnable(String name) {  // Constructor gives name to object.
      this.name = name;
   }
   public void run() {  // The run method prints a message to standard output.
      System.out.println("Greetings from thread '" + name +"'!");
   }
}
```

To use this version of the class, we would create a NamedRunnable object and use that object to create an object of type Thread:

```
NamedRunnable greetings = new NamedRunnable("Fred");
Thread greetingsThread = new Thread(greetings);
greetingsThread.start();
```

Finally, I'll note that it is sometimes convenient to define a thread using an anonymous inner class (Subsection 5.7.3). For example:

```
Thread greetingsFromFred = new Thread() {
   public void run() {
      System.out.println("Greetings from Fred!");
   }
};
greetingsFromFred.start();
```

To help you understand how multiple threads are executed in parallel, we consider the sample program ThreadTest1.java. This program creates several threads. Each thread performs exactly the same task. The task is to count the number of integers less than 1000000 that are prime. (The particular task that is done is not important for our purposes here.) On my computer, this task takes a little more than one second of processing time. The threads that perform this task are defined by the following static nested class:

```
/**
 * When a thread belonging to this class is run it will count the
 * number of primes between 2 and 1000000.  It will print the result
 * to standard output, along with its ID number and the elapsed
 * time between the start and the end of the computation.
 */
private static class CountPrimesThread extends Thread {
   int id;  // An id number for this thread; specified in the
constructor.
   public CountPrimesThread(int id) {
      this.id = id;
   }
   public void run() {
      long startTime = System.currentTimeMillis();
      int count = countPrimes(2,1000000); // Counts the primes.
      long elapsedTime = System.currentTimeMillis() - startTime;
      System.out.println("Thread " + id + " counted " +
            count + " primes in " + (elapsedTime/1000.0) + " seconds.");
   }
}
```

The main program asks the user how many threads to run, and then creates and starts the specified number of threads:

```
public static void main(String[] args) {
   int numberOfThreads = 0;
   while (numberOfThreads < 1 || numberOfThreads > 25) {
      System.out.print("How many threads do you want to use  (1 to 25) ?  ");
       numberOfThreads = TextIO.getlnInt();
       if (numberOfThreads < 1 || numberOfThreads > 25)
          System.out.println("Please enter a number between 1 and 25 !");
   }
   System.out.println("\nCreating " + numberOfThreads
                                          + " prime counting
threads...");
   CountPrimesThread[] worker = new CountPrimesThread[numberOfThreads];
   for (int i = 0; i < numberOfThreads; i++)
      worker[i] = new CountPrimesThread( i );
   for (int i = 0; i < numberOfThreads; i++)
      worker[i].start();
   System.out.println("Threads have been created and started.");
}
```

Here is an applet that simulates the program. Try running the program for various numbers of threads. In particular, you should at least try it with one thread and with two threads:

When I ran the program with one thread, it took 1.18 seconds for my computer to do the computation. When I ran it using six threads, the output was:

```
Creating 6 prime counting threads...
Threads have been created and started.
Thread 1 counted 78498 primes in 6.706 seconds.
Thread 4 counted 78498 primes in 6.693 seconds.
Thread 0 counted 78498 primes in 6.838 seconds.
Thread 2 counted 78498 primes in 6.825 seconds.
Thread 3 counted 78498 primes in 6.893 seconds.
Thread 5 counted 78498 primes in 6.859 seconds.
```

The second line was printed immediately after the first. At this point, the main program has ended but the six threads continue to run. After a pause of about seven seconds, all six threads completed at about the same time. The order in which the threads complete is not the same as the order in which they were started, and the order is indeterminate. That is, if the program is run again, the order in which the threads complete will probably be different.

On my computer, six threads take about six times longer than one thread. This is because my computer has only one processor. Six threads, all doing the same task, take six times as much processing as one thread. With only one processor to do the work, the total elapsed time for six threads is about six times longer than the time for one thread. On a computer with two processors, the computer can work on two tasks at the same time, and six threads might complete in as little as three times the time it takes for one thread. On a computer with six or more processors, six threads might take no more time than a single thread. Because of overhead and other reasons, the actual speedup will probably be smaller than this analysis indicates, but on a multiprocessor machine, you should see a definite speedup. What happens when you run the program on your own computer? How many processors do you have?

Whenever there are more threads to be run than there are processors to run them, the computer divides its attention among all the runnable threads by switching rapidly from one thread to another. That is, each processor runs one thread for a while then switches to another thread and runs that one for a while, and so on. Typically, these "context switches" occur about 100 times or more per second. The result is that the computer makes progress on all the tasks, and it looks to the user as if all the tasks are being executed simultaneously. This is why in the sample program, in which each thread has the same amount of work to do, all the threads complete at about the same time: Over any time period longer than a fraction of a second, the computer's time is divided approximately equally among all the threads.

When you do parallel programming in order to spread the work among several processors, you might want to take into account the number of available processors. You might, for example, want to create one thread for each processor. In Java, you can find out the number of processors by calling the function

```
Runtime.getRuntime().availableProcessors()
```

which returns an int giving the number of processors that are available to the Java Virtual Machine. In some cases, this might be less than the actual number of processors in the computer.

8.5.2 Operations on Threads

The Thread class includes several useful methods in addition to the start() method that was discussed above. I will mention just a few of them.

If thrd is an object of type Thread, then the boolean-valued function thrd.isAlive() can be used to test whether or not the thread is alive. A thread is "alive" between the time it is started and the time when it terminates. After the thread has terminated it is said to be "dead". (The rather gruesome metaphor is also used when we refer to "killing" or "aborting" a thread.)

The static method Thread.sleep(milliseconds) causes the thread that executes this method to "sleep" for the specified number of milliseconds. A sleeping thread is still alive, but it is not running. While a thread is sleeping, the computer will work on any other runnable threads (or on other programs). Thread.sleep() can be used to insert a pause in the execution of a thread. The sleep method can throw an exception of type InterruptedException, which is an exception class that requires mandatory exception handling (see

Subsection 8.3.4). In practice, this means that the sleep method is usually used in a try..catch statement that catches the potential InterruptedException:

```
try {
   Thread.sleep(lengthOfPause);
}
catch (InterruptedException e) {
}
```

One thread can interrupt another thread to wake it up when it is sleeping or paused for some other reason. A Thread, thrd, can be interrupted by calling its method thrd.interrupt(), but you are not likely to do this until you start writing rather advanced applications, and you are not likely to need to do anything in response to an InterruptedException (except to catch it). It's unfortunate that you have to worry about it at all, but that's the way that mandatory exception handling works.

Sometimes, it's necessary for one thread to wait for anther thread to die. This is done with the join() method from the Thread class. Suppose that thrd is a Thread. Then, if another thread calls thrd.join(), that other thread will go to sleep until thrd terminates. If thrd is already dead when thrd.join() is called, then it simply has no effect -- the thread that called thrd.join() proceeds immediately. The method join() can throw an InterruptedException, which must be handled. As an example, the following code starts several threads, waits for them all to terminate, and then outputs the elapsed time:

```
CountPrimesThread[] worker = new CountPrimesThread[numberOfThreads];
long startTime = System.currentTimeMillis();
for (int i = 0; i < numberOfThreads; i++) {
   worker[i] = new CountPrimesThread();
   worker[i].start();
}
for (int i = 0; i < numberOfThreads; i++) {
   try {
      worker[i].join();  // Sleep until worker[i] has terminated.
   }
   catch (InterruptedException e) {
   }
}
// At this point, all the worker threads have terminated.
long elapsedTime = System.currentTimeMillis() - startTime;
System.out.println("Elapsed time:  " + (elapsedTime/1000.0) + "
seconds.");
```

An observant reader will note that this code assumes that no InterruptedException will occur. To be absolutely sure that the thread worker[i] has terminated in an environment where InterruptedExceptions are possible, you would have to do something like:

```
while (worker[i].isAlive()) {
   try {
      worker[i].join();
   }
   catch (InterruptedException e) {
   }
}
```

8.5.3 Mutual Exclusion with "synchronized"

Programming several threads to carry out independent tasks is easy. The real difficulty arises when threads have to interact in some way. One way that threads interact is by sharing resources. When two threads need access to the same resource, such as a variable or a window on the screen, some care must be taken that they don't try to use the same resource at the same time. Otherwise, the situation could be something like this: Imagine several cooks sharing the use of just one measuring cup, and imagine that Cook A fills the measuring cup with milk, only to have Cook B grab the cup before Cook A has a chance to empty the milk into his bowl. There has to be some way for Cook A to claim exclusive rights to the cup while he performs the two operations: Add-Milk-To-Cup and Empty-Cup-Into-Bowl.

Something similar happens with threads, even with something as simple as adding one to a counter. The statement

```
count = count + 1;
```

is actually a sequence of three operations:

```
Step 1.  Get the value of count
Step 2.  Add 1 to the value.
Step 3.  Store the new value in count
```

Suppose that several threads perform these three steps. Remember that it's possible for two threads to run at the same time, and even if there is only one processor, it's possible for that processor to switch from one thread to another at any point. Suppose that while one thread is between Step 2 and Step 3, another thread starts executing the same sequence of steps. Since the first thread has not yet stored the new value in count, the second thread reads the **old** value of count and adds one to that old value. After both threads have executed Step 3, the value of count has gone up only by 1 instead of by 2! This type of problem is called a race condition. This occurs when one thread is in the middle of a multi-step operation, and another thread changes some value or condition that the first thread is depending upon. (The first thread is "in a race" to complete all the steps before it is interrupted by another thread.) Another example of a race condition can occur in an if statement. Suppose the following statement, which is meant to avoid a division-by-zero error is executed by a thread:

```
if ( A != 0 )
   B = C / A;
```

If the variable A is shared by several threads, and if nothing is done to guard against the race condition, then it is possible that a second thread will change the value of A to zero between the time that the first thread checks the condition A != 0 and the time that it does the division. This means that the thread ends up dividing by zero, even though it just checked that A was not zero!

To fix the problem of race conditions, there has to be some way for a thread to get exclusive access to a shared resource. This is not a trivial thing to implement, but Java provides a high level and relatively easy-to-use approach to exclusive access. It's done with synchronized methods and with the synchronized statement. These are used to protect shared resources by making sure that only one thread at a time will try to access the resource. Synchronization in Java actually provides only mutual exclusion, which means that exclusive access to a resource is only guaranteed if **every** thread that needs access to that resource uses synchronization. Synchronization is like a cook leaving a note that says, "I'm using the measuring cup." This will get the cook exclusive access to the cup -- but only if all the cooks agree to check the note before trying to grab the cup.

Because this is a difficult topic, I will start with a simple example. Suppose that we want to avoid the race condition that occurs when several threads all want to add 1 to a counter. We can do this by defining a class to represent the counter and by using synchronized methods in that class:

```
public class ThreadSafeCounter {

   private int count = 0;  // The value of the counter.
```

```
   synchronized public void increment() {
      count = count + 1;
   }

   synchronized public int getValue() {
      return count;
   }

}
```

If tsc is of type ThreadSafeCounter, then any thread can call tsc.increment() to add 1 to the counter in a completely safe way. The fact that tsc.increment() is synchronized means that only one thread can be in this method at a time; once a thread starts executing this method, it is guaranteed that it will finish executing it without having another thread change the value of tsc.count in the meantime. There is no possibility of a race condition. Note that the guarantee depends on the fact that count is a private variable. This forces all access to tsc.count to occur in the synchronized methods that are provided by the class. If count were public, it would be possible for a thread to bypass the synchronization by, for example, saying tsc.count++. This could change the value of count while another thread is in the middle of the tsc.increment(). Synchronization does **not** guarantee exclusive access; it only guarantees **mutual exclusion** among all the threads that are properly synchronized.

The ThreadSafeCounter class does not prevent all possible race conditions that might arise when using a counter. Consider the if statement:

```
if ( tsc.getValue() == 0 )
   doSomething();
```

where doSomething() is some method that requires the value of the counter to be zero. There is still a race condition here, which occurs if a second thread increments the counter between the time the first thread tests tsc.getValue() == 0 and the time it executes doSomething(). The first thread needs exclusive access to the counter during the execution of the whole if statement. (The synchronization in the ThreadSafeCounter class only gives it exclusive access during the time it is evaluating tsc.getValue().) We can solve the race condition by putting the if statement in a synchronized statement:

```
synchronized(tsc) {
   if ( tsc.getValue() == 0 )
      doSomething();
}
```

Note that the synchronized statement takes an object -- tsc in this case -- as a kind of parameter. The syntax of the synchronized statement is:

```
synchronized( object ) {
   statements
```

```
}
```

In Java, mutual exclusion is always associated with an object; we say that the synchronization is "on" that object. For example, the if statement above is "synchronized on tsc." A synchronized instance method, such as those in the class ThreadSafeCounter, is synchronized on the object that contains the instance method. In fact, adding the synchronized modifier to the definition of an instance method is pretty much equivalent to putting the body of the method in a synchronized statement, synchronized(this) {...}. It is also possible to have synchronized static methods; a synchronized static method is synchronized on a special class object that represents the class that contains the static method.

The real rule of synchronization in Java is: **Two threads cannot be synchronized on the same object at the same time**; that is, they cannot simultaneously be executing code segments that are synchronized on that object. If one thread is synchronized on an object, and a second thread tries to synchronize on the same object, the second thread is forced to wait until the first thread has finished with the object. This is implemented using something called a lock. Every object has a lock, and that lock can be "held" by only one thread at a time. To enter a synchronized statement or synchronized method, a thread must obtain the associated object's lock. If the lock is available, then the thread obtains the lock and immediately begins executing the synchronized code. It releases the lock after it finishes executing the synchronized code. If Thread A tries to obtain a lock that is already held by Thread B, then Thread A has to wait until Thread B releases the lock. In fact, Thread A will go to sleep, and will not be awoken until the lock becomes available.

As a simple example of shared resources, we return to the prime-counting problem. Suppose that we want to count all the primes in a given range of integers, and suppose that we want to divide the work up among several threads. Each thread will be assigned part of the range of integers and will count the primes in its assigned range. At the end of its computation, the thread has to add its count to the overall total number of primes found. The variable that represents the total is shared by all the threads. If each thread just says

```
total = total + count;
```

then there is a (small) chance that two threads will try to do this at the same time and that the final total will be wrong. To prevent this race condition, access to total has to be synchronized. My program uses a synchronized method to add the counts to the total:

```
synchronized private static void addToTotal(int x) {
   total = total + x;
   System.out.println(total + " primes found so far.");
}
```

The source code for the program can be found in ThreadTest2.java. This program counts the primes in the range 3000001 to 6000000. (The numbers are rather arbitrary.) The main() routine in this program creates between 1 and 5 threads and assigns part of the job to each thread. It then waits for all the threads to finish, using the join() method as described above, and reports the total elapsed time. If you run the program on a

multiprocessor computer, it should take less time for the program to run when you use more than one thread. Here is an applet that simulates the program:

Synchronization can help to prevent race conditions, but it introduces the possibility of another type of error, deadlock. A deadlock occurs when a thread waits forever for a resource that it will never get. In the kitchen, a deadlock might occur if two very simple-minded cooks both want to measure a cup of milk at the same time. The first cook grabs the measuring cup, while the second cook grabs the milk. The first cook needs the milk, but can't find it because the second cook has it. The second cook needs the measuring cup, but can't find it because the first cook has it. Neither cook can continue and nothing more gets done. This is deadlock. Exactly the same thing can happen in a program, for example if there are two threads (like the two cooks) both of which need to obtain locks on the same two objects (like the milk and the measuring cup) before they can proceed. Deadlocks can easily occur, unless great care is taken to avoid them. Fortunately, we won't be looking at any examples that require locks on more than one object, so we will avoid that source of deadlock.

8.5.4 Wait and Notify

Threads can interact with each other in other ways besides sharing resources. For example, one thread might produce some sort of result that is needed by another thread. This imposes some restriction on the order in which the threads can do their computations. If the second thread gets to the point where it needs the result from the first thread, it might have to stop and wait for the result to be produced. Since the second thread can't continue, it might as well go to sleep. But then there has to be some way to notify the second thread when the result is ready, so that it can wake up and continue its computation. Java, of course, has a way to do this kind of waiting and notification: It has wait() and notify() methods that are defined as instance methods in class Object and so can be used with any object. The reason why wait() and notify() should be associated with objects is not obvious, so don't worry about it at this point. It does, at least, make it possible to direct different notifications to a different recipients, depending on which object's notify() method is called.

The general idea is that when a thread calls a wait() method in some object, that thread goes to sleep until the notify() method in the same object is called. It will have to be called, obviously, by another thread, since the thread that called wait() is sleeping. A typical pattern is that Thread A calls wait() when it needs a result from Thread B, but that result is not yet available. When Thread B has the result ready, it calls notify(), which will wake Thread A up so that it can use the result. It is not an error to call notify() when no one is

waiting; it just has no effect. To implement this, Thread A will execute code similar to the following, where obj is some object:

```
if ( resultIsAvailable() == false )
   obj.wait();  // wait for noification that the result is available
useTheResult();
```

while Thread B does something like:

```
generateTheResult();
obj.notify();  // send out a notification that the result is available
```

Now, there is a really nasty race condition in this code. The two threads might execute their code in the following order:

```
1.  Thread A checks resultIsAvailable() and finds that the result is not
ready,
        so it decides to execute the obj.wait() statement, but before it
does,
2.  Thread B finishes generating the result and calls obj.notify()
3.  Thread A calls obj.wait() to wait for notification that the result
is ready.
```

In Step 3, Thread A is waiting for a notification that will never come, because notify() has already been called. This is a kind of deadlock that can leave Thread A waiting forever. Obviously, we need some kind of synchronization. The solution is to enclose both Thread A's code and Thread B's code in synchronized statements, and it is very natural to synchronize on the same object, obj, that is used for the calls to wait() and notify(). In fact, since synchronization is almost always needed when wait() and notify() are used, Java makes it an absolute requirement. In Java, a thread can legally call obj.wait() or obj.notify() only if that thread holds the synchronization lock associated with the object obj. If it does not hold that lock, then an exception is thrown. (The exception is of type IllegalMonitorStateException, which does not require mandatory handling and which is typically not caught.) One further complication is that the wait() method can throw an InterruptedException and so should be called in a try statement that handles the exception.

To make things more definite, lets consider a producer/consumer problem where one thread produces a result that is consumed by another thread. Assume that there is a shared variable named sharedResult that is used to transfer the result from the producer to the consumer. When the result is ready, the producer sets the variable to a non-null value. The producer can check whether the result is ready by testing whether the value of sharedResult is null. We will use a variable named lock for synchronization. The code for the producer thread could have the form:

```
makeResult = generateTheResult();  // Not synchronized!
synchronized(lock) {
   sharedResult = makeResult;
   lock.notify();
```

```
}
```

while the consumer would execute code such as:

```
synchronized(lock) {
   while ( sharedResult == null ) {
      try {
         lock.wait();
      }
      catch (InterruptedException e) {
      }
   }
   useResult = sharedResult;
}
useTheResult(useResult);  // Not synchronized!
```

The calls to generateTheResult() and useTheResult() are not synchronized, which allows them to run in parallel with other threads that might also synchronize on lock. Since sharedResult is a shared variable, all references to sharedResult should be synchronized, so the references to sharedResult must be inside the synchronized statements. The goal is to do as little as possible (but not less) in synchronized code segments.

If you are uncommonly alert, you might notice something funny: lock.wait() does not finish until lock.notify() is executed, but since both of these methods are called in synchronized statements that synchronize on the same object, shouldn't it be impossible for both methods to be running at the same time? In fact, lock.wait() is a special case: When the consumer thread calls lock.wait(), it gives up the lock that it holds on the synchronization object, lock. This gives the producer thread a chance to execute the synchronized(lock) block that contains the lock.notify() statement. After the producer thread exits from this block, the lock is returned to the consumer thread so that it can continue.

The producer/consumer pattern can be generalized and made more useful without making it any more complex. In the general case, multiple results are produced by one or more producer threads and are consumed by one or more consumer threads. Instead of having just one sharedResult object, we keep a list of objects that have been produced but not yet consumed. Producer threads add objects to this list. Consumer threads remove objects from this list. The only time when a thread is blocked from running is when a consumer thread tries to get a result from the list, and no results are available. It is easy to encapsulate the whole producer/consumer pattern in a class (where I assume that there is a class ResultType that represents the result objects):

```
/**
 * An object of type ProducerConsumer represents a list of results
 * that are available for processing.  Results are added to the list
 * by calling the produce method and are removed by calling consume.
 * If no result is available when consume is called, the method will
```

```
 * not return until a result becomes available.
 */
private static class ProducerConsumer {

   private ArrayList<ResultType> items = new ArrayList<ResultType>();
       // This ArrayList holds results that have been produced and are waiting
       // to be consumed.  See Subsection 7.3.3 for information on ArrayList.

   public void produce(ResultType item) {
      synchronized(items) {
         items.add(item); // Add item to the list of results.
         items.notify();  // Notify any thread waiting in consume()
method.
      }
   }

   public ResultType consume() {
      ResultType item;
      synchronized(items) {
        // If no results are available, wait for notification from produce().
         while (items.size() == 0) {
            try {
```

```
                items.wait();
            }
            catch (InterruptedException e) {
            }
        }
        // At this point, we know that at least one result is
available.
        item = items.remove(0);
      }
      return item;
   }

}
```

For an example of a program that uses a ProducerConsumer class, see ThreadTest3.java. This program performs the same task as ThreadTest2.java, but the threads communicate using the producer/consumer pattern instead of with a shared variable.

Going back to our kitchen analogy for a moment, consider a restaurant with several waiters and several cooks. If we look at the flow of customer orders into the kitchen, the waiters "produce" the orders and leave them in a pile. The orders are "consumed" by the cooks; whenever a cook needs a new order to work on, she picks one up from the pile. The pile of orders, or course, plays the role of the list of result objects in the

producer/consumer pattern. Note that the only time that a cook has to wait is when she needs a new order to work on, and there are no orders in the pile. The cook must wait until one of the waiters places an order in the pile. We can complete the analogy by imagining that the waiter rings a bell when he places the order in the pile -- ringing the bell is like calling the notify() method to notify the cooks that an order is available.

A final note on notify: It is possible for several threads to be waiting for notification. A call to obj.notify() will wake only one of the threads that is waiting on obj. If you want to wake all threads that are waiting on obj, you can call obj.notifyAll(). And a final note on wait: There is another version of wait() that takes a number of milliseconds as a parameter. A thread that calls obj.wait(milliseconds) will wait only up to the specified number of milliseconds for a notification. If a notification doesn't occur during that period, the thread will wake up and continue without the notification. In practice, this feature is most often used to let a waiting thread wake periodically while it is waiting in order to perform some periodic task, such as causing a message "Waiting for computation to finish" to blink.

8.5.5 Volatile Variables

And a final note on communication among threads: In general, threads communicate by sharing variables and accessing those variables in synchronized methods or synchronized statements. However, synchronization is fairly expensive computationally, and excessive use of it should be avoided. So in some cases, it can make sense for threads to refer to shared variables without synchronizing their access to those variables.

However, a subtle problem arises when the value of a shared variable is set is one thread and used in another. Because of the way that threads are implemented in Java, the second thread might not see the changed value of the variable immediately. That is, it is possible that a thread will continue to see the **old** value of the shared variable for some time after the value of the variable has been changed by another thread. This is because threads are allowed to cache shared data. That is, each thread can keep its own local copy of the shared data. When one thread changes the value of a shared variable, the local copies in the caches of other threads are not immediately changed, so the other threads continue to see the old value.

When a synchronized method or statement is entered, threads are forced to update their caches to the most current values of the variables in the cache. So, using shared variables in synchronized code is always safe.

It is still possible to use a shared variable **outside** of synchronized code, but in that case, the variable must be declared to be volatile. The volatile keyword is a modifier that can be added to a variable declaration, as in

```
private volatile int count;
```

If a variable is declared to be volatile, no thread will keep a local copy of that variable in its cache. Instead, the thread will always use the official, main copy of the variable. This means that any change made to the variable will immediately be available to all threads. This makes it safe for threads to refer to volatile shared variables even outside of synchronized code. (Remember, though, that synchronization is still the only way to prevent race conditions.)

When the volatile modifier is applied to an object variable, only the variable itself is declared to be volatile, not the contents of the object that the variable points to. For this reason, volatile is generally only used for variables of simple types such as primitive types and enumerated types.

A typical example of using volatile variables is to send a signal from one thread to another that tells the second thread to terminate. The two threads would share a variable

```
volatile boolean terminate = false;
```

The run method of the second thread would check the value of terminate frequently and end when the value of terminate becomes true:

```
public void run() {
   while (true) {
      if (terminate)
```

```
            return;
        .
        .   // Do some work
        .
    }
}
```

This thread will run until some other thread sets the value of terminate to true. Something like this is really the only clean way for one thread to cause another thread to die.

(By the way, you might be wondering why threads should use local data caches in the first place, since it seems to complicate things unnecessarily. Caching is allowed because of the structure of multiprocessing computers. In many multiprocessing computers, each processor has some local memory that is directly connected to the processor. A thread's cache is stored in the local memory of the processor on which the thread is running. Access to this local memory is much faster than access to other memory, so it is more efficient for a thread to use a local copy of a shared variable rather than some "master copy" that is stored in non-local memory.)

Section 11.4

Networking

As far as a program is concerned, a network is just another possible source of input data, and another place where data can be output. That does oversimplify things, because networks are not as easy to work with as files are. But in Java, you can do network communication using input streams and output streams, just as you can use such streams to communicate with the user or to work with files. Nevertheless, opening a network connection between two computers is a bit tricky, since there are two computers involved and they have to somehow agree to open a connection. And when each computer can send data to the other, synchronizing communication can be a problem. But the fundamentals are the same as for other forms of I/O.

One of the standard Java packages is called java.net. This package includes several classes that can be used for networking. Two different styles of network I/O are supported. One of these, which is fairly high-level, is based on the World-Wide Web, and provides the sort of network communication capability that is used by

a Web browser when it downloads pages for you to view. The main classes for this style of networking are java.net.URL and java.net.URLConnection. An object of type URL is an abstract representation of a Universal Resource Locator, which is an address for an HTML document or other resource on the Web. A URLConnection represents a network connection to such a resource.

The second style of I/O, which is more general and much more important, views the network at a lower level. It is based on the idea of a socket. A socket is used by a program to establish a connection with another program on a network. Communication over a network involves two sockets, one on each of the computers involved in the communication. Java uses a class called java.net.Socket to represent sockets that are used for network communication. The term "socket" presumably comes from an image of physically plugging a wire into a computer to establish a connection to a network, but it is important to understand that a socket, as the term is used here, is simply an object belonging to the class Socket. In particular, a program can have several sockets at the same time, each connecting it to another program running on some other computer on the network. All these connections use the same physical network connection.

This section gives a brief introduction to these basic networking classes, and shows how they relate to input and output streams.

11.4.1 URLs and URLConnections

The URL class is used to represent resources on the World-Wide Web. Every resource has an address, which identifies it uniquely and contains enough information for a Web browser to find the resource on the network and retrieve it. The address is called a "url" or "universal resource locator."

An object belonging to the URL class represents such an address. Once you have a URL object, you can use it to open a URLConnection to the resource at that address. A url is ordinarily specified as a string, such as "http://math.hws.edu/eck/index.html". There are also relative url's. A relative url specifies the location of a resource relative to the location of another url, which is called the base or context for the relative url. For example, if the context is given by the url http://math.hws.edu/eck/, then the incomplete, relative url "index.html" would really refer to http://math.hws.edu/eck/index.html.

An object of the class URL is not simply a string, but it can be constructed from a string representation of a url. A URL object can also be constructed from another URL object, representing a context, and a string that specifies a url relative to that context. These constructors have prototypes

```
public URL(String urlName) throws MalformedURLException
```

and

```
public URL(URL context, String relativeName) throws
MalformedURLException
```

Note that these constructors will throw an exception of type MalformedURLException if the specified strings don't represent legal url's. The MalformedURLException class is a subclass of IOException, and it requires mandatory exception handling. That is, you must call the constructor inside a try..catch statement that handles the exception or in a subroutine that is declared to throw the exception.

The second constructor is especially convenient when writing applets. In an applet, two methods are available that provide useful URL contexts. The method getDocumentBase(), defined in the Applet and JApplet classes, returns an object of type URL. This URL represents the location from which the HTML page that contains the applet was downloaded. This allows the applet to go back and retrieve other files that are stored in the same location as that document. For example,

```
URL url = new URL(getDocumentBase(), "data.txt");
```

constructs a URL that refers to a file named data.txt on the same computer and in the same directory as the source file for the web page on which the applet is running. Another method, getCodeBase(), returns a URL that gives the location of the applet class file (which is not necessarily the same as the location of the document).

Once you have a valid URL object, you can call its openConnection() method to set up a connection. This method returns a URLConnection. The URLConnection object can, in turn, be used to create an InputStream for reading data from the resource represented by the URL. This is done by calling its getInputStream() method. For example:

```
URL url = new URL(urlAddressString);
URLConnection connection = url.openConnection();
InputStream in = connection.getInputStream();
```

The openConnection() and getInputStream() methods can both throw exceptions of type IOException. Once the InputStream has been created, you can read from it in the usual way, including wrapping it in another input stream type, such as TextReader, or using a Scanner. Reading from the stream can, of course, generate exceptions.

One of the other useful instance methods in the URLConnection class is getContentType(), which returns a String that describes the type of information available from the URL. The return value can be null if the type of information is not yet known or if it is not possible to determine the type. The type might not be available until after the input stream has been created, so you should generally call getContentType() after getInputStream(). The string returned by getContentType() is in a format called a mime type. Mime types include "text/plain", "text/html", "image/jpeg", "image/gif", and many others. All mime types contain two parts: a general type, such as "text" or "image", and a more specific type within that general category, such as "html" or "gif". If you are only interested in text data, for example, you can check whether the string returned by getContentType() starts with "text". (Mime types were first introduced to describe the content of email messages. The name stands for "Multipurpose Internet Mail Extensions." They are now used almost universally to specify the type of information in a file or other resource.)

Let's look at a short example that uses all this to read the data from a URL. This subroutine opens a connection to a specified URL, checks that the type of data at the URL is text, and then copies the text onto the screen. Many of the operations in this subroutine can throw exceptions. They are handled by declaring that the subroutine "throws IOException" and leaving it up to the main program to decide what to do when an error occurs.

```
static void readTextFromURL( String urlString ) throws IOException {

   /* Open a connection to the URL, and get an input stream
      for reading data from the URL. */

   URL url = new URL(urlString);
   URLConnection connection = url.openConnection();
   InputStream urlData = connection.getInputStream();

   /* Check that the content is some type of text. */

   String contentType = connection.getContentType();
   if (contentType == null || contentType.startsWith("text") == false)
     throw new IOException("URL does not seem to refer to a text file.");

   /* Copy lines of text from the input stream to the screen, until
```

```
            end-of-file is encountered  (or an error occurs). */

         BufferedReader in;  // For reading from the connection's input stream.
         in = new BufferedReader( new InputStreamReader(urlData) );

         while (true) {
            String line = in.readLine();
            if (line == null)
               break;
            System.out.println(line);
         }

      } // end readTextFromURL()
```

A complete program that uses this subroutine can be found in the file ReadURL.java. When using the program, note that you have to specify a complete url, including the "http://" at the beginning. Here is an applet that does much the same thing. The applet lets you enter a URL, which can be either a complete URL or a relative URL. A relative URL will be interpreted relative to the document base of the applet. Error messages or text loaded from the URL will be displayed in the text area of the applet. (The amount of text is limited to 10000 characters.) When the applet starts up, it is configured to load the file ReadURL.java from this book's source code directory; just click the "Load" button:

You can also try to use this applet to look at the HTML source code for this very page. Just type s4.html into the input box at the bottom of the applet and then click on the Load button. You might want to experiment with other urls to see what types of errors can occur. For example, entering "bogus.html" is likely to generate a FileNotFoundException, since no document of that name exists in the directory that contains this page. As another example, you can probably generate a SecurityException by trying to connect to http://www.whitehouse.gov. (Not because it's an official secret -- any url that does not lead back to the same computer from which the applet was loaded will generate a security exception. To protect you from malicious applets, an applet is allowed to open network connections only back to the computer from which it came.) The source code for the applet is in the file ReadURLApplet.java.

11.4.2 TCP/IP and Client/Server

Communication over the Internet is based on a pair of protocols called the Transmission Control Protocol and the Internet Protocol, which are collectively referred to as TCP/IP. (In fact, there is a more basic communication protocol called UDP that can be used instead of TCP in certain applications. UDP is supported in Java, but for this discussion, I'll stick to the full TCP/IP, which provides reliable two-way communication between networked computers.)

For two programs to communicate using TCP/IP, each program must create a socket, as discussed earlier in this section, and those sockets must be connected. Once such a connection is made, communication takes place using input streams and output streams. Each program has its own input stream and its own output stream. Data written by one program to its output stream is transmitted to the other computer. There, it enters the input stream of the program at the other end of the network connection. When that program reads data from its input stream, it is receiving the data that was transmitted to it over the network.

The hard part, then, is making a network connection in the first place. Two sockets are involved. To get things started, one program must create a socket that will wait passively until a connection request comes in from another socket. The waiting socket is said to be listening for a connection. On the other side of the connection-to-be, another program creates a socket that sends out a connection request to the listening socket. When the listening socket receives the connection request, it responds, and the connection is established. Once that is done, each program can obtain an input stream and an output stream for sending data over the connection. Communication takes place through these streams until one program or the other closes the connection.

A program that creates a listening socket is sometimes said to be a server, and the socket is called a server socket. A program that connects to a server is called a client, and the socket that it uses to make a connection is called a client socket. The idea is that the server is out there somewhere on the network, waiting for a connection request from some client. The server can be thought of as offering some kind of service, and the client gets access to that service by connecting to the server. This is called the client/server model of network communication. In many actual applications, a server program can provide connections to several

clients at the same time. When a client connects to a server's listening socket, that socket does not stop listening. Instead, it continues listening for additional client connections at the same time that the first client is being serviced. To do this, it is necessary to use threads (Section 8.5). We'll look at how it works in the next section.

The URL class that was discussed at the beginning of this section uses a client socket behind the scenes to do any necessary network communication. On the other side of that connection is a server program that accepts a connection request from the URL object, reads a request from that object for some particular file on the server computer, and responds by transmitting the contents of that file over the network back to the URL object. After transmitting the data, the server closes the connection.

A client program has to have some way to specify which computer, among all those on the network, it wants to communicate with. Every computer on the Internet has an IP address which identifies it uniquely among all the computers on the net. Many computers can also be referred to by domain names such as math.hws.edu or www.whitehouse.gov. (See Section 1.7.) Traditional (or IPv4) IP addresses are 32-bit integers. They are usually written in the so-called "dotted decimal" form, such as 69.9.161.200, where each of the four numbers in the address represents an 8-bit integer in the range 0 through 255. A new version of the Internet Protocol, IPv6, is currently being introduced. IPv6 addresses are 128-bit integers and are usually written in hexadecimal form (with some colons and maybe some extra information thrown in). In actual use, IPv6 addresses are still fairly rare.

A computer can have several IP addresses, and can have both IPv4 and IPv6 addresses. Usually, one of these is the loopback address, which can be used when a program wants to communicate with another program *on the same computer*. The loopback address has IPv4 address 127.0.0.1 and can also, in general, be referred to using the domain name localhost. In addition, there can be one or more IP addresses associated with physical network connections. Your computer probably has some utility for displaying your computer's IP addresses. I have written a small Java program, ShowMyNetwork.java, that does the same thing. When I run ShowMyNetwork on my computer, the output is:

```
en1 :  /192.168.1.47  /fe80:0:0:0:211:24ff:fe9c:5271%5
lo0 :  /127.0.0.1  /fe80:0:0:0:0:0:0:1%1  /0:0:0:0:0:0:0:1%0
```

The first thing on each line is a network interface name, which is really meaningful only to the computer's operating system. The output also contains the IP addresses for that interface. In this example, lo0 refers to the loopback address, which has IPv4 address 127.0.0.1 as usual. The most important number here is 192.168.1.47, which is the IPv4 address that can be used for communication over the network.

Now, a single computer might have several programs doing network communication at the same time, or one program communicating with several other computers. To allow for this possibility, a network connection is actually identified by a port number in combination with an IP address. A port number is just a 16-bit integer. A server does not simply listen for connections -- it listens for connections *on a particular port*. A potential client must know both the Internet address (or domain name) of the computer on which the server is running and the port number on which the server is listening. A Web server, for example, generally

listens for connections on port 80; other standard Internet services also have standard port numbers. (The standard port numbers are all less than 1024, and are reserved for particular services. If you create your own server programs, you should use port numbers greater than 1024.)

11.4.3 Sockets

To implement TCP/IP connections, the java.net package provides two classes, ServerSocket and Socket. A ServerSocket represents a listening socket that waits for connection requests from clients. A Socket represents one endpoint of an actual network connection. A Socket can be a client socket that sends a connection request to a server. But a Socket can also be created by a server to handle a connection request from a client. This allows the server to create multiple sockets and handle multiple connections. A ServerSocket does not itself participate in connections; it just listens for connection requests and creates Sockets to handle the actual connections.

When you construct a ServerSocket object, you have to specify the port number on which the server will listen. The specification for the constructor is

```
public ServerSocket(int port) throws IOException
```

The port number must be in the range 0 through 65535, and should generally be greater than 1024. (A value of 0 tells the server socket to listen on any available port.) The constructor might throw a SecurityException

if a smaller port number is specified. An IOException can occur if, for example, the specified port number is already in use.

As soon as a ServerSocket is created, it starts listening for connection requests. The accept() method in the ServerSocket class accepts such a request, establishes a connection with the client, and returns a Socket that can be used for communication with the client. The accept() method has the form

```
public Socket accept() throws IOException
```

When you call the accept() method, it will not return until a connection request is received (or until some error occurs). The method is said to block while waiting for the connection. (While the method is blocked, the thread that called the method can't do anything else. However, other threads in the same program can proceed.) You can call accept() repeatedly to accept multiple connection requests. The ServerSocket will continue listening for connections until it is closed, using its close() method, or until some error occurs, or until the program is terminated in some way.

Suppose that you want a server to listen on port 1728, and suppose that you've written a method provideService(Socket) to handle the communication with one client. Then the basic form of the server program would be:

```
try {
   ServerSocket server = new ServerSocket(1728);
   while (true) {
```

```
            Socket connection = server.accept();
            provideService(connection);
        }
    }
    catch (IOException e) {
        System.out.println("Server shut down with error: " + e);
    }
```

On the client side, a client socket is created using a constructor in the Socket class. To connect to a server on a known computer and port, you would use the constructor

```
public Socket(String computer, int port) throws IOException
```

The first parameter can be either an IP number or a domain name. This constructor will block until the connection is established or until an error occurs.

Once you have a connected socket, no matter how it was created, you can use the Socket methods getInputStream() and getOutputStream() to obtain streams that can be used for communication over the connection. These methods return objects of type InputStream and OutputStream, respectively. Keeping all this in mind, here is the outline of a method for working with a client connection:

```
/**
 * Open a client connection to a specified server computer and
```

```
 * port number on the server, and then do communication through
 * the connection.
 */
void doClientConnection(String computerName, int serverPort) {
   Socket connection;
   InputStream in;
   OutputStream out;
   try {
      connection = new Socket(computerName,serverPort);
      in = connection.getInputStream();
      out = connection.getOutputStream();
   }
   catch (IOException e) {
      System.out.println(
          "Attempt to create connection failed with error: " + e);
      return;
   }
    .
    .  // Use the streams, in and out, to communicate with server.
    .
   try {
```

```
            connection.close();
                // (Alternatively, you might depend on the server
                //  to close the connection.)
        }
        catch (IOException e) {
        }
    }  // end doClientConnection()
```

All this makes network communication sound easier than it really is. (And if you think it sounded hard, then it's even harder.) If networks were completely reliable, things would be almost as easy as I've described. The problem, though, is to write robust programs that can deal with network and human error. I won't go into detail here. However, what I've covered here should give you the basic ideas of network programming, and it is enough to write some simple network applications. Let's look at a few working examples of client/server programming.

11.4.4 A Trivial Client/Server

The first example consists of two programs. The source code files for the programs are DateClient.java and DateServer.java. One is a simple network client and the other is a matching server. The client makes a connection to the server, reads one line of text from the server, and displays that text on the screen. The text sent by the server consists of the current date and time on the computer where the server is running. In order to open a connection, the client must know the computer on which the server is running and the port on which it is listening. The server listens on port number 32007. The port number could be anything between 1025 and 65535, as long the server and the client use the same port. Port numbers between 1 and 1024 are reserved for standard services and should not be used for other servers. The name or IP number of the computer on which the server is running must be specified as a command-line argument For example, if the server is running on a computer named math.hws.edu, then you would typically run the client with the command "java DateClient math.hws.edu". Here is the complete client program:

```
import java.net.*;
import java.io.*;

/**
 * This program opens a connection to a computer specified
 * as the first command-line argument.  The connection is made to
```

```
 * the port specified by LISTENING_PORT.  The program reads one
 * line of text from the connection and then closes the
 * connection.  It displays the text that it read on
 * standard output.  This program is meant to be used with
 * the server program, DateServer, which sends the current
 * date and time on the computer where the server is running.
 */
public class DateClient {

   public static final int LISTENING_PORT = 32007;

   public static void main(String[] args) {

      String hostName;         // Name of the server computer to connect to.
      Socket connection;       // A socket for communicating with server.
       BufferedReader incoming; // For reading data from the connection.

       /* Get computer name from command line. */

       if (args.length > 0)
          hostName = args[0];
       else {
```

```
      // No computer name was given.  Print a message and exit.
   System.out.println("Usage:  java DateClient
                                  <server_host_name>");
   return;
}

/* Make the connection, then read and display a line of text. */

try {
   connection = new Socket( hostName, LISTENING_PORT );
   incoming = new BufferedReader(
                     new
        InputStreamReader(connection.getInputStream()) );
   String lineFromServer = incoming.readLine();
   if (lineFromServer == null) {
         // A null from incoming.readLine() indicates that
         // end-of-stream was encountered.
      throw new IOException("Connection was opened, " +
            "but server did not send any data.");
   }
   System.out.println();
```

```
            System.out.println(lineFromServer);
            System.out.println();
            incoming.close();
        }
        catch (Exception e) {
            System.out.println("Error:  " + e);
        }

    }  // end main()


} //end class DateClient
```

Note that all the communication with the server is done in a try..catch statement. This will catch the IOExceptions that can be generated when the connection is opened or closed and when data is read from the input stream. The connection's input stream is wrapped in a BufferedReader, which has a readLine() method that makes it easy to read one line of text. (See Subsection 11.1.4.)

In order for this program to run without error, the server program must be running on the computer to which the client tries to connect. By the way, it's possible to run the client and the server program on the same computer. For example, you can open two command windows, start the server in one window and then run the client in the other window. To make things like this easier, most computers will recognize the domain

name localhost and the IP number 127.0.0.1 as referring to "this computer." This means that the command "java DateClient localhost" will tell the DateClient program to connect to a server running on the same computer. If that command doesn't work, try "java DateClient 127.0.0.1".

The server program that corresponds to the DateClient client program is called DateServer. The DateServer program creates a ServerSocket to listen for connection requests on port 32007. After the listening socket is created, the server will enter an infinite loop in which it accepts and processes connections. This will continue until the program is killed in some way -- for example by typing a CONTROL-C in the command window where the server is running. When a connection is received from a client, the server calls a subroutine to handle the connection. In the subroutine, any Exception that occurs is caught, so that it will not crash the server. Just because a connection to one client has failed for some reason, it does not mean that the server should be shut down; the error might have been the fault of the client. The connection-handling subroutine creates a PrintWriter for sending data over the connection. It writes the current date and time to this stream and then closes the connection. (The standard class java.util.Date is used to obtain the current time. An object of type Date represents a particular date and time. The default constructor, "new Date()", creates an object that represents the time when the object is created.) The complete server program is as follows:

```
import java.net.*;
import java.io.*;
import java.util.Date;
```

```
/**
 * This program is a server that takes connection requests on
 * the port specified by the constant LISTENING_PORT.  When a
 * connection is opened, the program sends the current time to
 * the connected socket.  The program will continue to receive
 * and process connections until it is killed (by a CONTROL-C,
 * for example).  Note that this server processes each connection
 * as it is received, rather than creating a separate thread
 * to process the connection.
 */
public class DateServer {

   public static final int LISTENING_PORT = 32007;

   public static void main(String[] args) {

      ServerSocket listener;  // Listens for incoming connections.
     Socket connection;      // For communication with the connecting program.

     /* Accept and process connections forever, or until some error occurs.
        (Note that errors that occur while communicating with a connected
         program are caught and handled in the sendDate() routine, so
```

```
        they will not crash the server.) */

    try {
        listener = new ServerSocket(LISTENING_PORT);
        System.out.println("Listening on port " + LISTENING_PORT);
        while (true) {
                // Accept next connection request and handle it.
            connection = listener.accept();
            sendDate(connection);
        }
    }
    catch (Exception e) {
        System.out.println("Sorry, the server has shut down.");
        System.out.println("Error:  " + e);
        return;
    }

}  // end main()


/**
```

```
    * The parameter, client, is a socket that is already connected to another
    * program.  Get an output stream for the connection, send the current time,
    * and close the connection.
    */
   private static void sendDate(Socket client) {
      try {
         System.out.println("Connection from " +
                                       client.getInetAddress().toString()
);
         Date now = new Date();  // The current date and time.
         PrintWriter outgoing;   // Stream for sending data.
         outgoing = new PrintWriter( client.getOutputStream() );
         outgoing.println( now.toString() );
         outgoing.flush();  // Make sure the data is actually sent!
         client.close();
      }
      catch (Exception e){
         System.out.println("Error: " + e);
      }
   } // end sendDate()
```

```
} //end class DateServer
```

When you run DateServer in a command-line interface, it will sit and wait for connection requests and report them as they are received. To make the DateServer service permanently available on a computer, the program really should be run as a daemon. A daemon is a program that runs continually on a computer, independently of any user. The computer can be configured to start the daemon automatically as soon as the computer boots up. It then runs in the background, even while the computer is being used for other purposes. For example, a computer that makes pages available on the World Wide Web runs a daemon that listens for requests for pages and responds by transmitting the pages. It's just a souped-up analog of the DateServer program! However, the question of how to set up a program as a daemon is not one I want to go into here. For testing purposes, it's easy enough to start the program by hand, and, in any case, my examples are not really robust enough or full-featured enough to be run as serious servers. (By the way, the word "daemon" is just an alternative spelling of "demon" and is usually pronounced the same way.)

Note that after calling out.println() to send a line of data to the client, the server program calls out.flush(). The flush() method is available in every output stream class. Calling it ensures that data that has been written to the stream is actually sent to its destination. You should generally call this function every time you use an output stream to send data over a network connection. If you don't do so, it's possible that the stream will collect data until it has a large batch of data to send. This is done for efficiency, but it can impose unacceptable delays when the client is waiting for the transmission. It is even possible that some of the data might remain untransmitted when the socket is closed, so it is especially important to call flush() before closing the connection. This is one of those unfortunate cases where different implementations of

Java can behave differently. If you fail to flush your output streams, it is possible that your network application will work on some types of computers but not on others.

11.4.5 A Simple Network Chat

In the DateServer example, the server transmits information and the client reads it. It's also possible to have two-way communication between client and server. As a first example, we'll look at a client and server that allow a user on each end of the connection to send messages to the other user. The program works in a command-line interface where the users type in their messages. In this example, the server waits for a connection from a single client and then closes down its listener so that no other clients can connect. After the client and server are connected, both ends of the connection work in much the same way. The user on the client end types a message, and it is transmitted to the server, which displays it to the user on that end. Then the user of the server types a message that is transmitted to the client. Then the client user types another message, and so on. This continues until one user or the other enters "quit" when prompted for a message. When that happens, the connection is closed and both programs terminate. The client program and the server program are very similar. The techniques for opening the connections differ, and the client is programmed to send the first message while the server is programmed to receive the first message. The client and server programs can be found in the files CLChatClient.java and CLChatServer.java. (The name "CLChat" stands for "command-line chat.") Here is the source code for the server:

```
import java.net.*;
import java.io.*;

/**
 * This program is one end of a simple command-line interface chat program.
 * It acts as a server which waits for a connection from the CLChatClient
 * program.  The port on which the server listens can be specified as a
 * command-line argument.  If it is not, then the port specified by the
 * constant DEFAULT_PORT is used.  Note that if a port number of zero is
 * specified, then the server will listen on any available port.
 * This program only supports one connection.  As soon as a connection is
 * opened, the listening socket is closed down.  The two ends of the connection
 * each send a HANDSHAKE string to the other, so that both ends can verify
 * that the program on the other end is of the right type.  Then the connected
 * programs alternate sending messages to each other.  The client always sends
 * the first message.  The user on either end can close the connection by
 * entering the string "quit" when prompted for a message.  Note that the first
 * character of any string sent over the connection must be 0 or 1; this
 * character is interpreted as a command.
 */
public class CLChatServer {

   /**
```

```
 * Port to listen on, if none is specified on the command line.
 */
static final int DEFAULT_PORT = 1728;

/**
 * Handshake string. Each end of the connection sends this  string to the
 * other just after the connection is opened.  This is done to confirm that
 * the program on the other side of the connection is a CLChat program.
 */
static final String HANDSHAKE = "CLChat";

/**
 * This character is prepended to every message that is sent.
 */
static final char MESSAGE = '0';

/**
 * This character is sent to the connected program when the user quits.
 */
static final char CLOSE = '1';
```

```
public static void main(String[] args) {

   int port;   // The port on which the server listens.

   ServerSocket listener;  // Listens for a connection request.
   Socket connection;      // For communication with the client.

  BufferedReader incoming;  // Stream for receiving data from client.
   PrintWriter outgoing;     // Stream for sending data to client.
   String messageOut;        // A message to be sent to the client.
   String messageIn;         // A message received from the client.

   BufferedReader userInput; // A wrapper for System.in, for reading
                             // lines of input from the user.

   /* First, get the port number from the command line,
      or use the default port if none is specified. */

   if (args.length == 0)
      port = DEFAULT_PORT;
   else {
```

```
      try {
         port= Integer.parseInt(args[0]);
         if (port < 0 || port > 65535)
            throw new NumberFormatException();
      }
      catch (NumberFormatException e) {
         System.out.println("Illegal port number, " + args[0]);
         return;
      }
   }

   /* Wait for a connection request.  When it arrives, close
      down the listener.  Create streams for communication
      and exchange the handshake. */

   try {
      listener = new ServerSocket(port);
      System.out.println("Listening on port " +
                                       listener.getLocalPort());
      connection = listener.accept();
      listener.close();
```

```
            incoming = new BufferedReader(
                               new InputStreamReader
                                  (connection.getInputStream()) );
            outgoing = new PrintWriter(connection.getOutputStream());
            outgoing.println(HANDSHAKE);  // Send handshake to client.
            outgoing.flush();
          messageIn = incoming.readLine();  // Receive handshake from client.
            if (! HANDSHAKE.equals(messageIn) ) {
               throw new Exception("Connected program is not a CLChat!");
            }
          System.out.println("Connected.  Waiting for the first message.");
        }
        catch (Exception e) {
          System.out.println("An error occurred while opening connection.");
            System.out.println(e.toString());
            return;
        }

      /* Exchange messages with the other end of the connection until one side
         or the other closes the connection.  This server program waits for
         the first message from the client.  After that, messages alternate
            strictly back and forth. */
```

```
      try {
         userInput = new BufferedReader(new
                     InputStreamReader(System.in));
         System.out.println("NOTE: Enter 'quit' to end the program.\n");
         while (true) {
            System.out.println("WAITING...");
            messageIn = incoming.readLine();
            if (messageIn.length() > 0) {
                  // The first character of the message is a command.
If
                 // the command is CLOSE, then the connection is closed.
                  // Otherwise, remove the command character from the
                  // message and proceed.
               if (messageIn.charAt(0) == CLOSE) {
                  System.out.println("Connection closed at other end.");
                  connection.close();
                  break;
               }
               messageIn = messageIn.substring(1);
            }
```

```
            System.out.println("RECEIVED:  " + messageIn);
            System.out.print("SEND:      ");
            messageOut = userInput.readLine();
            if (messageOut.equalsIgnoreCase("quit"))  {
                  // User wants to quit.  Inform the other side
                  // of the connection, then close the connection.
               outgoing.println(CLOSE);
               outgoing.flush();  // Make sure the data is sent!
               connection.close();
               System.out.println("Connection closed.");
               break;
            }
            outgoing.println(MESSAGE + messageOut);
            outgoing.flush(); // Make sure the data is sent!
            if (outgoing.checkError()) {
             throw new IOException("Error occurred while transmitting message.");
            }
         }
      }
      catch (Exception e) {
        System.out.println("Sorry, an error has occurred.  Connection lost.");
         System.out.println("Error:  " + e);
```

```
            System.exit(1);
        }

    }  // end main()


} //end class CLChatServer
```

This program is a little more robust than DateServer. For one thing, it uses a handshake to make sure that a client who is trying to connect is really a CLChatClient program. A handshake is simply information sent between client and server as part of setting up the connection, before any actual data is sent. In this case, each side of the connection sends a string to the other side to identify itself. The handshake is part of the protocol that I made up for communication between CLChatClient and CLChatServer. A protocol is a detailed specification of what data and messages can be exchanged over a connection, how they must be represented, and what order they can be sent in. When you design a client/server application, the design of the protocol is an important consideration. Another aspect of the CLChat protocol is that after the handshake, every line of text that is sent over the connection begins with a character that acts as a command. If the character is 0, the rest of the line is a message from one user to the other. If the character is 1, the line indicates that a user has entered the "quit" command, and the connection is to be shut down.

Remember that if you want to try out this program on a single computer, you can use two command-line windows. In one, give the command "java CLChatServer" to start the server. Then, in the other, use the command "java CLChatClient localhost" to connect to the server that is running on the same machine.

Section 11.5

Network Programming and Threads

In the previous section, we looked at several examples of network programming. Those examples showed how to create network connections and communicate through them, but they didn't deal with one of the fundamental characteristics of network programming, the fact that network communication is fundamentally asynchronous. From the point of view of a program on one end of a network connection, messages can arrive from the other side of the connection at any time; the arrival of a message is an *event* that is not under the control of the program that is receiving the message. Certainly, it is possible to design a network communication protocol that proceeds in a synchronous, step-by-step process from beginning to end -- but whenever the process gets to a point in the protocol where it needs to read a message from the other side of the connection, it has to *wait* for that message to arrive. Essentially, the process has to wait for a message-arrival event to occur before it can proceed. While it is waiting for the message, we say that the process is blocked.

Perhaps an event-oriented networking API would be a good approach to dealing with the asynchronous nature of network communication, but that is not the approach that is taken in Java (or, typically, in other languages). Instead, a serious network program in Java uses **threads**. Threads were introduced in Section 8.5. A thread is a separate computational process that can run in parallel with other threads. When a program uses threads to do network communication, it is possible that some threads will be blocked, waiting for incoming messages, but other threads will still be able to continue performing useful work.

11.5.1 A Threaded GUI Chat Program.

The command-line chat programs, CLChatClient.java and CLChatServer.java, from the previous section use a straight-through, step-by-step protocol for communication. After a user on one side of a connection enters a message, the user must wait for a reply from the other side of the connection. An asynchronous chat program would be much nicer. In such a program, a user could just keep typing lines and sending messages without waiting for any response. Messages that arrive -- asynchronously -- from the other side would be displayed as soon as they arrive. It's not easy to do this in a command-line interface, but it's a natural application for a graphical user interface. The basic idea for a GUI chat program is to create a thread whose job is to read messages that arrive from the other side of the connection. As soon as the message arrives, it is displayed to the user; then, the message-reading thread blocks until the next incoming message arrives. While it is blocked, however, other threads can continue to run. In particular, the GUI event-handling thread

that responds to user actions keeps running; that thread can send outgoing messages as soon as the user generates them.

In case this is not clear to you, here is an applet that simulates such a program. Enter a message in the input box at the bottom of the applet, and press return (or, equivalently, click the "Send" button):

Both incoming messages and messages that you send are posted to the JTextArea that occupies most of the applet. This is not a real network connection. When you send your first message, a separate thread is started by the applet. This thread **simulates** incoming messages from the other side of a network connection. In fact, it just chooses the messages at random from a pre-set list. At the same time, you can continue to enter and send messages. Of course, this applet doesn't really do any network communication, but the same idea can be used to write a GUI network chat program. The program GUIChat.java allows two-way network chatting that works similarly to this simulation, except that the incoming messages really do come from the other side of a network connection

The GUIChat program can act as both the client end and the server end of a connection. When GUIChat is started, a window appears on the screen. This window has a "Listen" button that the user can click to create a server socket that will listen for an incoming connection request; this makes the program act as a server. It also has a "Connect" button that the user can click to send a connection request; this makes the program act as a client. As usual, the server listens on a specified port number. The client needs to know the computer on

which the server is running and the port on which the server is listening. There are input boxes in the GUIChat window where the user can enter this information. Once a connection has been established between two GUIChat windows, each user can send messages to the other. The window has an input box where the user types the message. Pressing return while typing in this box sends the message. This means that the sending of the message is handled by the usual event-handling thread, in response to an event generated by a user action. Messages are received by a separate thread that just sits around waiting for incoming messages. This thread blocks while waiting for a message to arrive; when a message does arrive, it displays that message to the user. The window contains a large transcript area that displays both incoming and outgoing messages, along with other information about the network connection.

I urge you to compile the source code, GUIChat.java, and try the program. To make it easy to try it on a single computer, you can make a connection between one window and another window on the same computer, using "localhost" or "127.0.0.1" as the name of the computer. (Once you have one GUIChat window open, you can open a second one by clicking the "New" button.) I also urge you to read the source code. I will discuss only parts of it here.

The program uses a nested class, ConnectionHandler, to handle most network-related tasks. ConnectionHandler is a subclass of Thread. The ConnectionHandler thread is responsible for opening the network connection and then for reading incoming messages once the connection has been opened. (By putting the connection-opening code in a separate thread, we make sure that the GUI is not blocked while the connection is being opened. Like reading incoming messages, opening a connection is a blocking operation that can take some time to complete.) A ConnectionHandler is created when the user clicks the

"Listen" or "Connect" button. The "Listen" button should make the thread act as a server, while "Connect" should make it act as a client. To distinguish these two cases, the ConnectionHandler class has two constructors:

```
/**
 * Listen for a connection on a specified port.  The constructor
 * does not perform any network operations; it just sets some
 * instance variables and starts the thread.  Note that the
 * thread will only listen for one connection, and then will
 * close its server socket.
 */
ConnectionHandler(int port) {
   state = ConnectionState.LISTENING;
   this.port = port;
   postMessage("\nLISTENING ON PORT " + port + "\n");
   start();
}

/**
 * Open a connection to specified computer and port.  The constructor
 * does not perform any network operations; it just sets some
 * instance variables and starts the thread.
```

```
 */
ConnectionHandler(String remoteHost, int port) {
   state = ConnectionState.CONNECTING;
   this.remoteHost = remoteHost;
   this.port = port;
   postMessage("\nCONNECTING TO " + remoteHost + " ON PORT " + port +
"\n");
   start();
}
```

Here, state is an instance variable whose type is defined by an enumerated type

```
enum ConnectionState { LISTENING, CONNECTING, CONNECTED, CLOSED };
```

The values of this enum represent different possible states of the network connection. It is often useful to treat a network connection as a state machine (see Subsection 6.5.4), since the response to various events can depend on the state of the connection when the event occurs. Setting the state variable to LISTENING or CONNECTING tells the thread whether it should act as a server or as a client. Note that the postMessage() method posts a message to the transcript area of the window, where it will be visible to the user.

Once the thread has been started, it executes the following run() method:

```
/**
 * The run() method that is executed by the thread.  It opens a
 * connection as a client or as a server (depending on which
 * constructor was used).
 */
public void run() {
   try {
      if (state == ConnectionState.LISTENING) {
            // Open a connection as a server.
         listener = new ServerSocket(port);
         socket = listener.accept();
         listener.close();
      }
      else if (state == ConnectionState.CONNECTING) {
            // Open a connection as a client.
         socket = new Socket(remoteHost,port);
      }
      connectionOpened();  // Sets up to use the connection (including
                           // creating a BufferedReader, in, for reading
                           // incoming messages).
      while (state == ConnectionState.CONNECTED) {
```

```
            // Read one line of text from the other side of
            // the connection, and report it to the user.
         String input = in.readLine();
         if (input == null)
            connectionClosedFromOtherSide();
         else
            received(input);  // Report message to user.
      }
   }
   catch (Exception e) {
         // An error occurred.  Report it to the user, but not
         // if the connection has been closed (since the error
         // might be the expected error that is generated when
         // a socket is closed).
      if (state != ConnectionState.CLOSED)
         postMessage("\n\n ERROR:  " + e);
   }
   finally {  // Clean up before terminating the thread.
      cleanUp();
   }
}
```

This method calls several other methods to do some of its work, but you can see the general outline of how it works. After opening the connection as either a server or client, the run() method enters a while loop in which it receives and processes messages from the other side of the connection until the connection is closed. It is important to understand how the connection can be closed. The GUIChat window has a "Disconnect" button that the user can click to close the connection. The program responds to this event by closing the socket that represents the connection. It is likely that when this happens, the connection-handling thread is blocked in the in.readLine() method, waiting for an incoming message. When the socket is closed by another thread, this method will fail and will throw an exception; this exception causes the thread to terminate. (If the connection-handling thread happens to be between calls to in.readLine() when the socket is closed, the while loop will terminate because the connection state changes from CONNECTED to CLOSED.) Note that closing the window will also close the connection in the same way.

It is also possible for the user on the other side of the connection to close the connection. When that happens, the stream of incoming messages ends, and the in.readLine() on this side of the connection returns the value null, which indicates end-of-stream and acts as a signal that the connection has been closed by the remote user.

For a final look into the GUIChat code, consider the methods that send and receive messages. These methods are called from different threads. The send() method is called by the event-handling thread in response to a user action. Its purpose is to transmit a message to the remote user. It uses a PrintWriter, out, that writes to the socket's output stream. Synchronization of this method prevents the connection state from changing in the middle of the send operation:

```
/**
 * Send a message to the other side of the connection, and post the
 * message to the transcript.  This should only be called when the
 * connection state is ConnectionState.CONNECTED; if it is called at
 * other times, it is ignored.
 */
synchronized void send(String message) {
   if (state == ConnectionState.CONNECTED) {
      postMessage("SEND:  " + message);
      out.println(message);
      out.flush();
      if (out.checkError()) {
         postMessage("\nERROR OCCURRED WHILE TRYING TO SEND DATA.");
         close();  // Closes the connection.
      }
   }
}
```

The received() method is called by the connection-handling thread **after** a message has been read from the remote user. Its only job is to display the message to the user, but again it is synchronized to avoid the race condition that could occur if the connection state were changed by another thread while this method is being executed:

```
/**
 * This is called by the run() method when a message is received from
 * the other side of the connection.  The message is posted to the
 * transcript, but only if the connection state is CONNECTED.  (This
 * is because a message might be received after the user has clicked
 * the "Disconnect" button; that message should not be seen by the
 * user.)
 */
synchronized private void received(String message) {
   if (state == ConnectionState.CONNECTED)
      postMessage("RECEIVE:  " + message);
}
```

11.5.2 A Multithreaded Server

There is still one big problem with the GUIChat program. In order to open a connection to another computer, a user must know that there is a GUIChat program listening on some particular port on some particular computer. Except in rather contrived situations, there is no way for a user to know that. It would be nice if it were possible to discover, somehow, who's out there on the Internet waiting for a connection. Unfortunately, this is not possible. And yet, applications such as AOL Instant Messenger seem to do just that -- they can show you a list of users who are available to receive messages. How can they do that?

I don't know the details of instant messenger protocols, but it has to work something like this: When you start the client program, that program contacts a server program that runs constantly on some particular computer and on some particular port. Since the server is always available at the same computer and port, the information needed to contact it can be built into the client program or otherwise made available to the users of the program. The purpose of the server is to keep a list of available users. When your client program contacts the server, it gets a list of available users, along with whatever information is necessary to send messages to those users. At the same time, your client program registers you with the server, so that the server can tell other users that you are on-line. When you shut down the client program, you are removed from the server's list of users, and other users can be informed that you have gone off-line.

Of course, in an application like AOL server, you only get to see a list of available users from your "buddy list," a list of your friends who are also AOL users. To implement this, you need to have an account on the AOL server. The server needs to keep a database of information about all user accounts, including the buddy list for each user. This makes the server program rather complicated, and I won't consider that aspect of its functionality here. However, it is not very difficult to write a scaled-down application that uses the network in a similar way. I call my scaled-down version "BuddyChat." It doesn't keep separate buddy lists for each user; it assumes that you're willing to be buddies with anyone who happens to connect to the server. In this application, the server keeps a list of connected users and makes that list available to each connected user. A user can connect to another user and chat with that user, using a window that is very similar to the chat window in GUIChat. BuddyChat is still just a toy, compared to serious network applications, but it does illustrate some core ideas.

The BuddyChat application comes in several pieces. BuddyChatServer.java is the server program, which keeps the list of available users and makes that list available to clients. Ideally, the server program would run constantly (as a daemon) on a computer and port that are known to all the possible client users. For testing, of course, it can simply be stated like any other program. The client program is BuddyChat.java. This program is to be run by any user who wants to use the BuddyChat service. When a user starts the client program, it connects to the server, and it gets from the server a list of other users who are currently connected. The list is displayed to the user of the client program, who can send a request for a chat connection with any user on the list. The client can also receive incoming chat connection requests from other users. The window that is used for chatting is defined by BuddyChatWindow.java, which is not itself a program but just a subclass of JFrame that defines the chat window. (There is also a fourth piece, BuddyChatServerShutdown.java. This is a program that can be run to shut down the BuddyChatServer gracefully. I will not discuss it further here. See the source code for more information, if you are interested.)

I urge you to compile the programs and try them out. For testing, you can try them on a single computer (although all the windows can get a little confusing). First, start BuddyChatServer. The server has no GUI interface, but it does print some information to standard output as it runs. Then start the BuddyChat client program. When BuddyChat starts up, it presents a window where you can enter the name and port number for the server and your "handle," which is just a name that will identify you in the server's list of users. The server info is already set up to connect to a server on the same machine. When you hit the "Connect" button, a new window will open with a list, currently empty, of other users connected to the server. Now, start another copy of the BuddyChat client program. When you click "Connect", you'll have two client list

windows, one for each copy of the client program that you've started. (One of these windows will be exactly on top of the other, so you'll have to move it to see the second window.) Each client window will display the other client in its list of users. You can run additional copies of the client program, if you want, and you might want to try connecting from another computer if one is available.

At this point, there is a network connection in place between the server and each client. Whenever a client connects to or disconnects from the server, the server sends a notification of the event to each connected client, so that the client can modify its own list of connected users. The server also maintains a listening socket that listens for connection requests from new clients. In order to manage all this, the server is running several threads. One thread waits for connection requests on the listening socket. In addition to this, there are two threads for each connected client -- one thread for sending messages to the client and one thread for reading messages sent by the client to the server.

Back to trying out the program. Remember that the whole point was to provide each user with a list of potential chat partners. Click on a user in one of the client user lists, and then click the "Connect to Selected Buddy" button. When you do this, your BuddyChat program sends a connection request to the BuddyChat program that is being run by the selected user. Each BuddyChat program, one on each side of the connection, opens a chat window (of type BuddyChatWindow). A network connection between these two windows is set up without any further action on the part of the two users, and the users can use the windows to send messages back and forth to each other. The BuddyChatServer program has nothing to do with opening, closing, or using the connection between its two clients (although a different design might have had the messages go through the server).

In order to open the chat connection from one program to another, the second program must be listening for connection requests and the first program must know the computer and port on which the first user is listening. In the BuddyChat system, the BuddyChatServer knows this information and provides it to each BuddyChat client program. The **users** of the client programs never have to be aware of this information.

How does the server know about the clients' computers and port numbers? When a BuddyChat client program is run, in addition to opening a connection to the BuddyChatServer, the client also creates a listening socket to accept connection requests from other users. When the client registers with the server, it tells the server the port number of the client's listening socket. The server also knows the IP address of the computer on which the client is running, since it has a network connection to that computer. This means that the BuddyChatServer knows the IP address and listening socket port number of every BuddyChat client. A copy of this information is provided (along with the users' handles) to each connected client program. The net result is that every BuddyChat client program has the information that it needs to contact all the other clients.

The basic techniques used in the BuddyChat system are the same as those used in previous networking examples: server sockets, client sockets, input and output streams for sending messages over the network, and threads to handle the communication. The important difference is how these basic building blocks are combined to build a more complex application. I have tried to explain the logic of that application here. I will not discuss the BuddyChat source code here, since it is locally similar to examples that we have already looked at, but I encourage you to study the source code if you are interested in network programming.

BuddyChat seems to have a lot of functionality, yet I said it was still a "toy" program. What exactly makes it a toy? There are at least two big problems. First of all, it is not scalable. A network program is scalable if it will work well for a large number of simultaneous users. BuddyChat would have problems with a large number of users because it uses so many threads (two for each user). It takes a certain amount of processing for a computer to switch its attention from one thread to another. On a very busy server, the constant switching between threads would soon start to degrade the performance. One solution to this is to use a more advanced network API. Java has a class SelectableChannel that makes it possible for one thread to manage communication over a large number of network connections. This class is part of the package java.nio that provides a number of advanced I/O capabilities for working with files and networking. However, I will not cover those capabilities in this book.

But the biggest problem is that BuddyChat offers absolutely no defense against denial of service attacks. In a denial of service, a malicious user attacks a network server in some way that prevents other users from accessing the service or severely degrades the performance of the service for those users. It would be simple to launch a denial of service attack on BuddyChat by making a huge number of connections to the server. The server would then spend most of its time servicing those bogus connections. The server could guard against this to some extent by putting a limit on the number of simultaneous connections that it will accept from a given IP address. It would also be helpful to add some security to the server by requiring users to know a password in order to connect. However, neither of these measures would fully solve the problem, and it is very difficult to find a complete defense against denial of service attacks.

11.5.3 Distributed Computing

In Section 8.5, we saw how threads can be used to do parallel processing, where a number of processors work together to complete some task. In that section, it was assumed that all the processors were inside one multi-processor computer. But parallel processing can also be done using processors that are in different computers, as long as those computers are connected to a network over which they can communicate. This type of parallel processing -- in which a number of computers work together on a task and communicate over a network -- is called distributed computing.

In some sense, the whole Internet is an immense distributed computation, but here I am interested in how computers on a network can cooperate to solve some computational problem. There are several approaches to distributed computing that are supported in Java. RMI and CORBA are standards that enable a program running on one computer to call methods in objects that exist on other computers. This makes it possible to design an object-oriented program in which different parts of the program are executed on different computers. RMI (Remote Method Invocation) only supports communication between Java objects. CORBA (Common Object Request Broker Architecture) is a more general standard that allows objects written in various programming languages, including Java, to communicate with each other. As is commonly the case in networking, there is the problem of locating services (where in this case, a "service" means an object that is available to be called over the network). That is, how can one computer know which computer a service is located on and what port it is listening on? RMI and CORBA solve this problem using something like our little BuddyChatServer example -- a server running at a known location keeps a list of services that are

available on other computers. Computers that offer services register those services with the server; computers that need services contact the server to find out where they are located.

RMI and CORBA are complex systems that are not very easy to use. I mention them here because they are part of Java's standard network API, but I will not discuss them further. Instead, we will look at a relatively simple demonstration of distributed computing that uses only basic networking.

The problem that we will look at uses the simplest type of parallel programming, in which the problem can be broken down into tasks that can be performed independently, with no communication between the tasks. To apply distributed computing to this type of problem, we can use one "master" program that divides the problem into tasks and sends those tasks over the network to "worker" programs that do the actual work. The worker programs send their results back to the master program, which combines the results from all the tasks into a solution of the overall problem. In this context, the worker programs are often called "slaves," and the program uses the so-called master/slave approach to distributed computing.

The demonstration program is defined by three source code files: CLMandelbrotMaster.java defines the master program; CLMandelbrotWorker.java defines the worker programs; and CLMandelbrotTask.java defines the class, CLMandelbrotTask, that represents an individual task that is performed by the workers. To run the demonstration, you must start the CLMandelbrotWorker program on several computers (probably by running it on the command line). This program uses CLMandelbrotTask, so both class files, CLMandelbrotWorker.class and CLMandelbrotTask.class, must be present on the worker computers. You can then run CLMandelbrotMaster on the master computer. Note that this program also requires the class

CLMandelbrotTask. You must specify the host name or IP address of each of the worker computers as command line arguments for CLMandelbrotMaster. A worker program listens for connection requests from the master program, and the master program must be told where to send those requests. For example, if the worker program is running on three computers with IP addresses 172.30.217.101, 172.30.217.102, and 172.30.217.103, then you can run CLMandelbrotMaster with the command

```
java CLMandelbrotMaster 172.30.217.101 172.30.217.102 172.30.217.103
```

The master will make a network connection to the worker at each IP address; these connections will be used for communication between the master program and the workers.

It is possible to run several copies of CLMandelbrotWorker on the same computer, but they must listen for network connections on different ports. It is also possible to run CLMandelbrotWorker on the same computer as CLMandelbrotMaster. You might even see some speed-up when you do this, if your computer has several processors. See the comments in the program source code files for more information, but here are some commands that you can use to run the master program and two copies of the worker program on the same computer. Give these commands in separate command windows:

```
java CLMandelbrotWorker                    (Listens on default
port)
```

```
java  CLMandelbrotWorker  1501                          (Listens on port
1501)

java  CLMandelbrotMaster  localhost  localhost:1501
```

Every time CLMandelbrotMaster is run, it solves exactly the same problem. (For this demonstration, the nature of the problem is not important, but the problem is to compute the data needed for a picture of a small piece of the famous "Mandelbrot Set." If you are interested in seeing the picture that is produced, uncomment the call to the saveImage() method at the end of the main() routine in CLMandelbrotMaster.java. We will encounter the Mandelbrot Set again as an example in Chapter 12.)

You can run CLMandelbrotMaster with different numbers of worker programs to see how the time required to solve the problem depends on the number of workers. (Note that the worker programs continue to run after the master program exists, so you can run the master program several times without having to restart the workers.) In addition, if you run CLMandelbrotMaster with no command line arguments, it will solve the entire problem on its own, so you can see how long it takes to do so without using distributed computing. In a trial that I ran, it took 40 seconds for CLMandelbrotMaster to solve the problem on its own. Using just one worker, it took 43 seconds. The extra time represents extra work involved in using the network; it takes time to set up a network connection and to send messages over the network. Using two workers (on different computers), the problem was solved in 22 seconds. In this case, each worker did about half of the work, and their computations were performed in parallel, so that the job was done in about half the time. With larger numbers of workers, the time continued to decrease, but only up to a point. The master program itself has a

certain amount of work to do, no matter how many workers there are, and the total time to solve the problem can never be less than the time it takes for the master program to do its part. In this case, the minimum time seemed to be about five seconds.

Let's take a look at how this distributed application is programmed. The master program divides the overall problem into a set of tasks. Each task is represented by an object of type CLMandelbrotTask. These tasks have to be communicated to the worker programs, and the worker programs must send back their results. Some protocol is needed for this communication. I decided to use character streams. The master encodes a task as a line of text, which is sent to a worker. The worker decodes the text (into an object of type CLMandelbrotTask) to find out what task it is supposed to perform. It performs the assigned task. It encodes the results as another line of text, which it sends back to the master program. Finally, the master decodes the results and combines them with the results from other tasks. After all the tasks have been completed and their results have been combined, the problem has been solved.

The problem is divided into a fairly large number of tasks. A worker receives not just one task, but a sequence of tasks. Each time it finishes a task and sends back the result, it is assigned a new task. After all tasks are complete, the worker receives a "close" command that tells it to close the connection. In CLMandelbrotWorker.java, all this is done in a method named handleConnection() that is called to handle a connection that has already been opened to the master program. It uses a method readTask() to decode a task

that it receives from the master and a method writeResults() to encode the results of the task for transmission back to the master. It must also handle any errors that occur:

```
private static void handleConnection(Socket connection) {
   try {
      BufferedReader in = new BufferedReader( new InputStreamReader(
                                          connection.getInputStream())
);
      PrintWriter out = new PrintWriter(connection.getOutputStream());
      while (true) {
         String line = in.readLine();  // Message from the master.
         if (line == null) {
               // End-of-stream encountered -- should not happen.
            throw new Exception("Connection closed unexpectedly.");
         }
         if (line.startsWith(CLOSE_CONNECTION_COMMAND)) {
               // Represents the normal termination of the connection.
            System.out.println("Received close command.");
            break;
         }
         else if (line.startsWith(TASK_COMMAND)) {
               // Represents a CLMandelbrotTask that this worker is
```

```
                // supposed to perform.
             CLMandelbrotTask task = readTask(line);  // Decode the message.
              task.compute();  // Peform the task.
              out.println(writeResults(task));  //  Send back the results.
              out.flush();
           }
           else {
                // No other messages are part of the protocol.
              throw new Exception("Illegal command received.");
           }
        }
     }
     catch (Exception e) {
        System.out.println("Client connection closed with error " + e);
     }
     finally {
        try {
           connection.close();  // Make sure the socket is closed.
        }
        catch (Exception e) {
        }
```

```
    }
}
```

Note that this method is **not** executed in a separate thread. The worker has only one thing to do at a time and does not need to be multithreaded.

You might wonder why so many tasks are used. Why not just divide the problem into one task for each worker? The reason is that using a larger number of tasks makes it possible to do load balancing. Not all tasks take the same amount of time to execute. This is true for many reasons. Some of the tasks might simply be more computationally complex than others. Some of the worker computers might be slower than others. Or some worker computers might be busy running other programs, so that they can only give part of their processing power to the worker program. If we assigned one task per worker, it is possible that a complex task running on a slow, busy computer would take much longer than the other tasks to complete. This would leave the other workers idle and delay the completion of the job while that worker completes its task. To complete the job as quickly as possible, we want to keep all the workers busy and have them all finish at about the same time. This is called load balancing. If we have a large number of tasks, the load will automatically be approximately balanced: A worker is not assigned a new task until it finishes the task that it is working on. A slow worker, or one that happens to receive more complex tasks, will complete fewer tasks than other workers, but all workers will be kept busy until close to the end of the job. On the other hand, individual tasks shouldn't be too small. Network communication takes some time. If it takes longer to transmit a task and its results than it does to perform the task, then using distributed computing will take

more time than simply doing the whole job on one computer! A problem is a good candidate for distributed computing if it can be divided into a fairly large number of fairly large tasks.

Turning to the master program, CLMandelbrotMaster.java, we encounter a more complex situation. The master program must communicate with several workers over several network connections. To accomplish this, the master program is multi-threaded, with one thread to manage communication with each worker. A pseudocode outline of the main() routine is quite simple:

```
create a list of all tasks that must be performed
if there are no command line arguments {
      // The master program does all the tasks itself.
   Perform each task.
}
else {
      // The tasks will be performed by worker programs.
   for each command line argument:
      Get information about a worker from command line argument.
      Create and start a thread to communicate with the worker.
   Wait for all threads to terminate.
}
// All tasks are now complete (assuming no error occurred).
```

The list of tasks is stored in a variable, tasks, of type ArrayList<CLMandelbrotTask>. The communication threads take tasks from this list and send them to worker programs. The method getNextTask() gets one task from the list. If the list is empty, it returns null as a signal that all tasks have been assigned and the communication thread can terminate. Since tasks is a resource that is shared by several threads, access to it must be controlled; this is accomplished by writing getNextTask() as a synchronized method:

```
synchronized private static CLMandelbrotTask getNextTask() {
   if (tasks.size() == 0)
      return null;
   else
      return tasks.remove(0);
}
```

(The reason for the synchronization is to avoid the race condition that could occur between the time that the value of tasks.size() is tested and the time that tasks.remove() is called. See Subsection 8.5.3 for information about parallel programming, race conditions, and synchronized.)

The job of a thread is to send a sequence of tasks to a worker thread and to receive the results that the worker sends back. The thread is also responsible for opening the connection in the first place. A pseudocode outline for the process executed by the thread might look like:

```
Create a socket connected to the worker program.
```

```
Create input and output streams for communicating with the worker.
while (true) {
   Let task = getNextTask().
   If task == null
      break;  // All tasks have been assigned.
   Encode the task into a message and transmit it to the worker.
   Read the response from the worker.
   Decode and process the response.
}
Send a "close" command to the worker.
Close the socket.
```

This would work OK. However, there are a few subtle points. First of all, the thread must be ready to deal with a network error. For example, a worker might shut down unexpectedly. But if that happens, the master program can continue, provided other workers are still available. (You can try this when you run the program: Stop one of the worker programs, with CONTROL-C, and observe that the master program still completes successfully.) A difficulty arises if an error occurs while the thread is working on a task: If the problem as a whole is going to be completed, that task will have to be reassigned to another worker. I take care of this by putting the uncompleted task back into the task list. (Unfortunately, my program does not handle all possible errors. If a network connection "hangs" indefinitely without actually generating an error,

my program will also hang, waiting for a response from a worker that will never arrive. A more robust program would have some way of detecting the problem and reassigning the task.)

Another defect in the procedure outlined above is that it leaves the worker program idle while the thread is processing the worker's response. It would be nice to get a new task to the worker before processing the response from the previous task. This would keep the worker busy and allow two operations to proceed simultaneously instead of sequentially. (In this example, the time it takes to process a response is so short that keeping the worker waiting while it is done probably makes no significant difference. But as a general principle, it's desirable to have as much parallelism as possible in the algorithm.) We can modify the procedure to take this into account:

```
try {
   Create a socket connected to the worker program.
   Create input and output streams for communicating with the worker.
   Let currentTask = getNextTask().
   Encode currentTask into a message and send it to the worker.
   while (true) {
      Read the response from the worker.
      Let nextTask = getNextTask().
      If nextTask != null {
            // Send nextTask to the worker before processing the
            // response to currentTask.
```

```
            Encode nextTask into a message and send it to the worker.
         }
         Decode and process the response to currentTask.
         currentTask = nextTask.
         if (currentTask == null)
            break; // All tasks have been assigned.
      }
      Send a "close" command to the worker.
      Close the socket.
   }
   catch (Exception e) {
      Put uncompleted task, if any, back into the task list.
   }
   finally {
      Close the connection.
   }
```

Finally, here is how this translates into Java. The pseudocode presented above becomes the run() method in the class that defines the communication threads used by the master program:

```
/**
 * This class represents one worker thread.  The job of a worker thread
```

```
 * is to send out tasks to a CLMandelbrotWorker program over a network
 * connection, and to get back the results computed by that program.
 */
private static class WorkerConnection extends Thread {

   int id;         // Identifies this thread in output statements.
   String host;    // The host to which this thread will connect.
   int port;       // The port number to which this thread will connect.

   /**
    * The constructor just sets the values of the instance
    * variables id, host, and port and starts the thread.
    */
   WorkerConnection(int id, String host, int port) {
      this.id = id;
      this.host = host;
      this.port = port;
      start();
   }

   /**
```

```
     * The run() method of the thread opens a connection to the host and
     * port specified in the constructor, then sends tasks to the
     * CLMandelbrotWorker program on the other side of that connection.
     * If the thread terminates normally, it outputs  the number of tasks
     * that it processed.  If it terminates with an error, it outputs
     * an error message.
     */
    public void run() {

       int tasksCompleted = 0; // How many tasks has this thread handled.
       Socket socket;  // The socket for the connection.

       try {
          socket = new Socket(host,port);  // open the connection.
       }
       catch (Exception e) {
         System.out.println("Thread " + id + " could not open connection to " +
                 host + ":" + port);
          System.out.println("   Error: " + e);
          return;
       }
```

```
CLMandelbrotTask currentTask = null;
CLMandelbrotTask nextTask = null;

try {
   PrintWriter out = new PrintWriter(socket.getOutputStream());
   BufferedReader in = new BufferedReader(
                          new
                     InputStreamReader(socket.getInputStream()) );
   currentTask = getNextTask();
   if (currentTask != null) {
         // Send first task to the worker program.
      String taskString = writeTask(currentTask);
      out.println(taskString);
      out.flush();
   }
   while (currentTask != null) {
     String resultString = in.readLine(); // Get results for currentTask.
      if (resultString == null)
         throw new IOException("Connection closed unexpectedly.");
      if (! resultString.startsWith(RESULT_COMMAND))
       throw new IOException("Illegal string received from worker.");
```

```
        nextTask = getNextTask();  // Get next task and send it to worker.
        if (nextTask != null) {
            // Send nextTask to worker before processing results for
            // currentTask, so that the worker can work on nextTask
              // while the currentTask results are processed.
           String taskString = writeTask(nextTask);
           out.println(taskString);
           out.flush();
        }
        readResults(resultString, currentTask);
        finishTask(currentTask);  // Process results from currentTask.
        tasksCompleted++;
        currentTask = nextTask;   // We are finished with old currentTask.
        nextTask = null;
      }
    out.println(CLOSE_CONNECTION_COMMAND);  // Send close command to worker.
     out.flush();
  }
  catch (Exception e) {
    System.out.println("Thread " + id + " terminated because of an error");
     System.out.println("   Error: " + e);
     e.printStackTrace();
```

```
            // Put uncompleted task, if any, back into the task list.
         if (currentTask != null)
            reassignTask(currentTask);
         if (nextTask != null)
            reassignTask(nextTask);
      }
      finally {
        System.out.println("Thread " + id + " ending after completing " +
               tasksCompleted + " tasks");
         try {
            socket.close();
         }
         catch (Exception e) {
         }
      }

   } //end run()

} // end nested class WorkerConnection
```