

Locks, Semaphores and Monitors

Java's synchronization mechanisms are based upon the notion of a *lock*. A lock is a special value that can be held by at most one thread.

If a thread holds a lock, it has permission to do some "critical" operation like writing a shared variable or restructuring a shared data object.

If a thread wants to do an operation but doesn't hold the necessary lock, it must wait until it gets the lock.

In Java there is a lock associated with each run-time object.

Lock Granularity and Access

Though each Java object has a lock, you often don't want to lock and unlock each object you access.

If you are manipulating a large data structure (like a tree or hash table), acquiring the lock for each object in the tree or table can be costly and error-prone.

Instead, it is common to create a lock corresponding to a group of objects. Hence holding the lock to the root of a tree may give you permission to access the whole tree.

There is a danger though—if all or most of a large structure is held by one thread, then other threads won't be able to access that structure concurrently.

For example, a large shared data base (like one used to record current bookings for an airline) shouldn't be held exclusively by one thread—hundreds of concurrent threads may want to access it at any time. An intermediate lock, like all reservations for a single flight, is more reasonable. There is also an issue of how long you hold a lock. The ideal is to have exclusive access for as short a period as is necessary. Work that is not subject to interference by other threads (like computations using local variables) should be done before a lock is obtained. Hence Java's **synchronized** statement allows a method to get exclusive access to an object for a limited region, enhancing shared access.

Deadlock

A variety of programming problems appear in concurrent programs that don't exist in ordinary sequential programs.

The most serious of these is *deadlock*:

Two or more threads hold locks that other threads require. Each waits for the other thread to release a needed lock, and no thread is able to execute.

As an example of how deadlock may occur, consider two threads, τ_1 and τ_2 . Each requires two files, a master file and a log file. Since these files are shared, each has a lock.

Assume τ_1 gets the lock for the master file while τ_2 (at the same instant) gets the lock for the log file.

Now each is stuck. Each has one file, and will wait forever for the other file to be released.

In Java deadlock avoidance is wholly up to the programmer. There are no language-level guarantees that deadlock can't happen.

Some languages have experimented with ways to help programmers avoid deadlock:

- If all locks must be claimed at once, deadlock can be avoided. You either get all of them or none, but you can't block other threads while making no progress yourself.
- Locks (and the resources they control) can be ordered, with the rule that you must acquire locks in the

proper order. Now two threads can't each hold locks the other needs.

- The language can require that the largest set of locks ever needed be declared in advance. When locks are requested, the operating system can track what's claimed and what may be needed, and refuse to honor unsafe requests.

Fairness & Starvation

When one thread has a lock, other threads who want the lock will be suspended until the lock is released.

It can happen that a waiting thread may be forced to wait indefinitely to acquire a lock, due to an unfair waiting policy. A waiting thread that never gets a lock it needs due to unfair lock allocation faces *starvation*.

As an example, if we place waiting threads on a stack, newly arrived threads will get access to a lock before earlier arrivals. This can lead to starvation. Most thread managers try to be fair and guarantee that all waiting threads have a fair chance to acquire a lock.

How are Locks Implemented?

Internally, Java needs operations to acquire a lock and release a lock. These operations can be implemented using the notion of a *semaphore*.

A semaphore is an integer value (often just a single bit) with two atomic operations: *up* and *down*.

up(s) increments **s** atomically.

down(s) decrements **s** atomically.

But if **s** is already zero, the process doing the **down** operation is put in a wait state until **s** becomes positive (eventually some other process should do an **up** operation).

Now locks are easy to implement.

You do a **down(lock)** to claim a lock. If someone else has it, you are forced

to wait until the lock is released. If the lock value is > 0 you get it and all others are “locked out.”

When you want to release a lock, you do `up(lock)`, which makes `lock` non-zero and eligible for another thread to claim.

In fact, since only one thread will ever have a lock, the lock value needs to be only one bit, with 1 meaning currently free and unlocked and 0 meaning currently claimed and locked.

Monitors

Direct manipulation of semaphores is tedious and error-prone. If you acquire a lock but forget to release it, threads may be blocked forever.

Depending on where **down** and **up** operations are placed, it may be difficult to understand how synchronization is being performed.

Few modern languages allow direct use of semaphores. Instead, semaphores are used in the implementation of higher-level constructs like *monitors*.

A monitor is a language construct that guarantees that a block of code will be executed synchronously (one thread at a time).

The Java **synchronized** statement is a form of monitor.

When

```
synchronized(obj) { ... }
```

is executed, “invisible” **getLock** and **freeLock** operations are added:

```
synchronized(obj) {  
    getLock(obj)  
    ...  
    freeLock(obj);  
}
```

This allows the body of the **synchronized** statement to execute only when it has the lock for **obj**.

Thus two different threads can never simultaneously execute the body of a **synchronized** statement because two threads can't simultaneously hold **obj**'s lock.

In fact, synchronized methods are really just methods whose bodies are enclosed in an invisible **synchronized** statement.

If we execute

```
obj.method( )
```

where **method** is synchronized,

method's body is executed as if it were of the form

```
synchronized(obj) {  
    body of method  
}
```

Operations like **sleep**, **wait**, **notify** and **notifyAll** also implicitly cause threads to release locks, allowing other threads to proceed.