Concurrent Programming with CXL

Robbert van Renesse

July 2, 2020

Contents

1 On Concurrent Programming **3**

2 Introduction to Programming with CXL **6**

3 The Problem of Concurrent Programming **8**

4 The CXL Virtual Machine **13**

5 Critical Sections **20**

6 Peterson's Algorithm **26**

7 CXL Methods and Pointers **32**

8 Spinlock **35**

9 Locks and Blocking **39**

10 Reader/Writer Locks using Busy Waiting **45**

11 Reader/Writer Locks with Blocking **50**

12 Semaphores **54**

13 Bounded Buffer **59**

14 Split Binary Semaphores **65**

15 Starvation **69**

16 Monitors **74**

17 Deadlock **80**

18 Actors and Message Passing **87**

19 Networking: Alternating Bit Protocol **91**

20 Non-Blocking Synchronization **94**

21 Barrier Synchronization **99**

Bibliography **102**

A List of Values **104**

B List of Operators **106**

C List of Statements **108**

D List of Modules **109**
D.1 The `alloc` module 109
D.2 The `bag` module 109
D.3 The `list` module 109
D.4 The `synch` module 110

E List of Machine Instructions **111**

F Contexts and Processes **113**

G The CXL Virtual Machine **115**

Acknowledgments **116**

Index **117**

Glossary **119**

Chapter 1

On Concurrent Programming

Programming with concurrency is hard. On the one hand concurrency can make programs faster than sequential ones, but having multiple processes (aka *threads*[1]) read and update shared variables concurrently and synchronize with one another makes programs more complicated than programs where only one thing happens at a time.

Why are concurrent programs more complicated than sequential ones? There are, at least, two reasons:

- The execution of a sequential program is mostly *deterministic*. If you run it twice with the same input, the same output will be produced. Bugs are typically easily reproducible and easy to track down, for example by instrumenting the program. The output of running concurrent programs depends on how the execution of the various processes are *interleaved*. Some bugs may occur only occasionally and may never occur when the program is instrumented to find them (so-called *Heisenbugs*).
- In a sequential program, each statement and each function can be thought of as happening *atomically* (indivisibly) because there is no other activity interfering with their execution. Even though a statement or function may be compiled into multiple machine instructions, they are executed back-to-back until completion. Not so with a concurrent program, where other processes may update memory locations while a statement or function is being executed.

The lack of determinism and atomicity in concurrent programs make them not only hard to reason about, but also hard to test. Running the same test of concurrent code twice is likely to produce two different results. More problematically, a test may trigger a bug only for certain "lucky" executions. Due to the probabilistic nature of concurrent code, some bugs may be highly unlikely to get triggered even when running a test millions of times. And even if a bug does get triggered, the source of the bug may be hard to find because it is hard to reproduce.

This book is intended to help people with understanding and developing concurrent code. In particular, it uses a new tool called CXL that helps with *testing* concurrent code. The approach is based on *model checking*: instead of relying on luck, CXL will run *all possible executions* of a

[1] We do not make a distinction between processes and threads.

particular test program. So even if a bug is unlikely to occur, if the test *can* expose the bug it *will*. Importantly, if the bug is found, the model checker precisely shows how to trigger it in the smallest number of steps.

Model checking is not a replacement for formal verification. Formal verification proves that a program is correct. Model checking only verifies that a program is correct for some *model*. Think of a model as a test program. Because model checking tries every possible execution, the test program needs to be simple. Otherwise it may take longer than we care to wait for or run out of memory. In particular, the model needs to have a relatively small number of reachable states.

So if model checking does not prove a program correct, why is it useful? To answer that question, let us consider a sorting algorithm. Now suppose we create a test program, a model, that tries sorting *all* lists of up to five numbers chosen from the set { 1, 2, 3, 4, 5 }. Model checking proves that for those particular scenarios the sorting algorithm works: the output is a sorted permutation of the input. In some sense it is an excellent test: it will have considered all *corner cases*, including lists where all numbers are the same, lists that are already sorted or reversely sorted, etc. If there is a bug in the sorting algorithm, most likely it would be triggered and the model checker would produce a scenario that would make it easy to find the source of the bug. However, if the model checker does not find any bugs, we do not know for sure that the algorithm works for lists with more than five numbers or for lists that have values other than the numbers 1 through 5. Still, we would expect that the likelihood that there are bugs remaining in the sorting algorithm is small. Note, however, that it would be easy to write a program that sorts all lists of up to five numbers correctly but fails to do so for a list of 6 numbers. (Hint: simply use an **if** statement.)

While model checking does not in general prove an algorithm correct, it can help with proving an algorithm correct. The reason is that the correctness of many programs is built around *invariants*: predicates that must hold for every state in the execution of a program. A model checker can find violations of proposed invariants when evaluating a model and provide valuable early feedback to somebody who is trying to construct a proof, even an informal one. We will include examples of such invariants as they often provide excellent insight into why a particular algorithm works.

So what is CXL? CXL is a concurrent programming language. It was designed to teach the basics of concurrent programming, but it is also useful for testing new concurrent algorithms or even sequential and distributed algorithms. CXL programs are not intended to be "run" like programs in most other programming languages—instead CXL programs are model checked to test that the program has certain desirable properties and does not suffer from bugs.

The syntax and semantics of CXL is similar to that of Python. Python is familiar to many programmers and is easy to learn and use. We will assume that the reader is familiar with the basics of Python programming. We also will assume that the reader understands some basics of machine architecture and how programs are executed. For example, we assume that the reader is familiar with the concepts of CPU, memory, register, stack, and machine instructions.

CXL is heavily influenced by Leslie Lamport's work on TLA+, TLC, and PlusCal [Lam02, Lam09]. CXL is designed to have a lower learning curve than those systems, but is not as powerful. When you finish this book and want to learn more, we strongly encourage checking those out. Another excellent resource is Fred Schneider's book "On Concurrent Programming" [Sch97]. (This chapter is named after that book.)

The book proceeds as follows:

- Chapter 2 introduces the CXL programming language, as it provides the language for presenting synchronization problems and solutions.

- Chapter 3 illustrates the problem of concurrent programming through a simple example in which two processes are concurrently incrementing a counter.
- Chapter 4 presents the CXL virtual machine to understand the problem underlying concurrency better.
- Chapter 5 introduces the concept of a *critical section* and presents various flawed implementations of critical sections to demonstrate that implementing a critical section is not trivial.
- Chapter 6 introduces Peterson's Algorithm, an elegant solution to implementating a critical section.
- Chapter 7 gives some more details on the CXL language.
- Chapter 8 introduces "hardware" locks for implemented critical sections.
- Chapter 9 presents CXL modules that supports locks and other synchronization primitives.
- Chapters 10 and 11 present the abstraction and implementation of reader/writer locks, which allow multiple processes in a critical section if they are all only executing read-only accesses. Chapter 10 also talks about busy-waiting, which is an undesirable approach to synchronize processes.
- Chapter 12 introduces *semaphores*, a generalization of locks that is good not only for mutual exclusion but also for waiting for certain application-level conditions.
- Chapters 13 and 14 introduce the *bounded buffer problem* (aka *producer/consumer problem*) and various solutions.
- Chapter 15 talks about *starvation*: the problem that in some synchronization approaches processes may not be able to get access to a resource they need.
- Chapter 16 presents *monitors* and *condition variables*, another approach to process synchronication.
- Chapter 17 describes *deadlock* where a set of processes are indefinitely waiting for one another to release a resource.
- Chapter 18 presents the *actor model* and *message passing* as an approach to synchronization.
- Chapter 19 presents a problem and a solution to the distributed systems problem of having two processes communicate reliably over an unreliable network.
- Chapter 20 introduces *non-blocking* or *wait-free* synchronization algorithms, which prevent processes waiting for one another more than a bounded number of steps.
- Chapter 21 describes *barrier synchronization*, useful in high-performance computing applications such as parallel simulations.

Chapter 2

Introduction to Programming with CXL

Like Python, CXL is an imperative, dynamically typed, and garbage collected programming language. There are also some important differences:

- Syntax: Every statement in CXL must be terminated by a semicolon (even **while** statements and **def** statements). In Python semicolons are optional and rarely used. There is no syntactic significance to indentation in CXL. Correct indentation in CXL is encouraged but not enforced. CXL only supports basic operator precedence or associativity. Use parentheses liberally to remove ambiguity.
- CXL does not (currently) support floating point, iterators, or I/O; CXL does support **for** loops and various comprehensions over sets.
- Python is object-oriented, supporting classes with methods and inheritance; CXL has objects but does not support classes. CXL supports pointers to objects.

There are also many unimportant ones that you will discover as you get more familiar with programming in CXL.

Figure 2.1 shows a simple example of a CXL program. (The code for examples in this book can be found in the `code` folder under the name listed in the caption of the example.) The example is sequential and has a method `triangle` that takes an integer number as argument. Each method has a variable called *result* that eventually contains the result of the method (there is no **return** statement in CXL). The method also has a variable called n containing the value of the argument. The $x..y$ notation generates a set containing the numbers from x to y (inclusive). The last two lines in the program are the most interesting. The first assigns to x some unspecified value in the range `0..N` and the second verifies that `triangle`(x) equals $x(x+1)/2$.

"Running" this CXL program will try all possible executions, which includes all possible values for x. Try it out (here `$` represents a shell prompt):

```
const N = 10;

def triangle(n):   # computes the n'th triangle number
    result = 0;
    for i in 1..n:      # for each integer from 1 to n inclusive
        result += i;    # add i to result
    ;
;
x = choose(0..N);       # select an x between 0 and N inclusive
assert triangle(x) == ((x * (x + 1)) / 2);
```

Figure 2.1: [triangle.cxl] Computing triangle numbers.

```
$ cxl triangle.cxl
#states = 13
no issues found
$
```

(For this to work, make sure `cxl` is in your command shell's search path.) Essentially, the `choose`(S) operator provides the input to the program by selecting some value from the set S, while the **assert** statement checks that the output is correct. If the program is correct, the output of CXL is the size of the "state graph" (13 states in this case). If not, CXL also reports what went wrong, typically by displaying a summary of an execution in which something went wrong.

In CXL, constants have a default value, but those can be overridden on the command line using the `-c` option. For example, if you want to test the code for `N = 100`, run:

```
$ cxl -c N=100 triangle.cxl
#states = 103
no issues found
$
```

Exercises

2.1 See what happens if, instead of initializing *result* to 0, you initialize it to 1. (You do not need to understand the error report at this time. They will be explained in more detail in Chapter 4.)

2.2 Write a CXL program that computes squares by repeated adding. So the program should compute the square of x by adding x to an initial value of 0 x times.

Chapter 3

The Problem of Concurrent Programming

Concurrent programming, aka multithreaded programming, involves multiple processes running in parallel while sharing variables. Figure 3.1 presents a simple example. The program initializes two shared variables: an integer *count* and an array *done* with two booleans. Method `incrementer` takes a parameter called *self*. It increments *count* and sets *done*[*self*] to `True`. Method `main` waits until *done* flags are set to `True`. After that, it verifies that the value of *count* equals 2. Like Python, CXL supports `assert` statements with two arguments: the first is a Boolean condition, and the second is a value that is reported if the condition fails, in this case the actual value of *count*. The program spawns three processes. The first runs `incrementer(0)`, the second runs `incrementer(1)`, and the last runs `main()`. Note that although the processes are *spawned* one at a time, they will execute concurrently. It is, for example, quite possible that `incrementer(1)` finishes before `incrementer(0)` even gets going.

- Before you run the program, what do you think will happen? Is the program correct in that *count* will always end up being 2? (You may assume that `load` and `store` instructions of the underlying machine architecture are atomic (indivisible)—in fact they are.)

What is going on is that the CXL program is compiled to machine instructions, and it is the machine instructions that are executed by the underlying CXL machine. The details of this appear in Chapter 4, but suffice it to say that the machine has instructions that load values from memory and store values into memory. Importantly, it does not have instructions to atomically increment or decrement values in memory locations. So to increment a value in memory, the machine must do at least three machine instructions. Conceptually:

1. load the value from the memory location;
2. add 1 to the value;
3. store the value to the memory location.

```
def incrementer(self):
    count = count + 1;
    done[self] = True;
;
def main():
    while not (done[0] and done[1]):
        pass;
    ;
    assert count == 2, count;
;
count = 0;
done = [ False, False ];
spawn incrementer(0);
spawn incrementer(1);
spawn main();
```

Figure 3.1: [Up.cxl] Incrementing twice in parallel.

When running multiple processes, each essentially runs an instantiation of the machine, and they do so in parallel. As they execute, their machine instructions are interleaved in unspecified and often random ways. A program is correct if it works for any interleaving. In fact, CXL will try all possible interleavings of the processes executing machine instructions.

The following is a possible interleaving of incrementers 0 and 1:

1. incrementer 0 loads the value of *count*, which is 0;
2. incrementer 1 loads the value of *count*, which is still 0;
3. incrementer 1 adds one to the value that it loaded (0), and stores 1 into *count*;
4. incrementer 0 adds one to the value that it loaded (0), and stores 1 into *count*;
5. incrementer 0 sets *done*[0] to `True`;
6. incrementer 1 sets *done*[1] to `True`.

The result in this particular interleaving is that *count* ends up being 1. This is known as a *race condition*. When running CXL, it will report violations of assertions. It also provides an example of an interleaving, like the one above, in which an assertion fails.

If one thinks of the assertion as providing the specification of the program, then clearly its implementation does not satisfy its specification. Either the specification or the implementation (or both) must have a bug. We could change the specification by changing the assertion as follows:

```
assert (count == 1) or (count == 2);
```

This would fix the issue, but more likely it is the program that must be fixed.

Exercises

The following exercises are intended to show you that on the one hand it is easy to translate CXL programs into Python, but on the other hand, it is easier to find concurrency bugs using CXL.

3.1 CXL programs can usually be easily translated into Python. For example, Figure 3.2 is a Python version of Figure 3.1.

1. Run Figure 3.2 using Python. Does the assertion fail?
2. Using a script, run Figure 3.2 1000 times. For example, if you are using the bash shell, you can do the following:

   ```
   for i in {1..1000}
   do
       python Up.py
   done
   ```

   How many times does the assertion fail (if any).

3.2 Figure 3.3 is a version of Figure 3.2 that has each incrementer thread increment *count* `N` times. Run Figure 3.2 10 times (using Python). Report how many times the assertion fails and what the value of *count* was for each of the failed runs. Also experiment with lower values of `N`. How large does `N` need to be for assertions to fail?

3.3 Can you think of a fix to Figure 3.1? Try one or two different fixes and run them through CXL. Do not worry about having to come up with a correct fix at this time—the important thing is to develop an understanding of concurrency.

```python
import threading

def incrementer(self):
    global count
    count = count + 1
    done[self] = True

def main():
    while not (done[0] and done[1]):
        pass
    assert count == 2, count
    print("Done")

def spawn(f, a):
    threading.Thread(target=f, args=a).start()

count = 0
done = [ False, False ]
spawn(incrementer, (0,))
spawn(incrementer, (1,))
spawn(main, ())
```

Figure 3.2: [Up.py] Python implementation of Figure 3.1.

```
import threading

N = 10000000

def incrementer(self):
    global count
    for i in range(N):
        count = count + 1
    done[self] = True

def main():
    while not (done[0] and done[1]):
        pass
    assert count == 2*N, count
    print("Done")

def spawn(f, a):
    threading.Thread(target=f, args=a).start()

count = 0
done = [ False, False ]
spawn(incrementer, (0,))
spawn(incrementer, (1,))
spawn(main, ())
```

Figure 3.3: [UpMany.py] Incrementing `N` times.

Chapter 4

The CXL Virtual Machine

CXL programs are compiled to CXL bytecode (machine instructions for a virtual machine), which in turn is executed by the CXL virtual machine (CVM). To understand the problem of concurrent computing, it is important to have a basic understanding of the CVM.

CXL Values

CXL programs, and indeed the CVM, manipulate CXL values. CXL values are recursively defined: they include booleans (`False` and `True`), integers (but not floating point numbers), strings (enclosed by double quotes), sets of CXL values, and dictionaries that map CXL values to other CXL values. Another type of CXL value is the *atom*. It is essentially just a name. An atom is denoted using a period followed by the name. For example, `.main` is an atom.

CXL makes extensive use of dictionaries. A dictionary maps values, known as *keys*, to values. The CXL dictionary syntax and properties are a little different from Python. Unlike Python, any CXL value can be a key, including another dictionary. CXL dictionaries are written as `dict`$\{k_0 : v_0,\ k_1 : v_1,\ ...\}$. If d is a dictionary, and k is a key, then the following expression retrieves the CXL value that k maps to in d:

```
d k
```

This is unfamiliar to Python programers, but in CXL square brackets can be used in the same way as parentheses, so you can express the same thing in the form that is familiar to Python programmers:

```
d[k]
```

However, if d = `dict{ .count: 3 }`, then you can write can write d`.count` (which has value 3) instead of having to write d`[.count]` (although both will work). Thus using atoms, a dictionary can be made to look much like a Python object. The meaning of $d\ a\ b$ ... is $(((d\ a)\ b)$ ...).

Tuples are special forms of dictionaries where the keys are the indexes into the tuple. For example, the tuple `(5, False)` is the same CXL value as `dict{ 0:5, 1:False }`. The empty

tuple `()` is the same value as `dict{}`. Note that this is different from the empty set, which is {}. As in Python, you can create singleton tuples by including a comma. For example, `(1,)`.

Again, square brackets and parentheses work the same in CXL, so `[a, b, c]` (which looks like a Python list) is the same CXL value as `(a, b, c)` (which looks like a Python tuple), which in turn is the same CXL value as `dict{ 0:a, 1:b, 2:c }`. So if x = `[ False, True ]`, then $x[0]$ = `False` and $x[1]$ = `True`, just like in Python. However, when creating a singleton list, make sure you include the comma, as in `[False,]`.

CXL is not an object-oriented language, so objects don't have built-in methods. However, CXL does have some powerful operators to make up for some of that. For example, dictionaries have two handy unary operators. If d is a dictionary, then `keys` d (or equivalently `keys`(d) returns the set of keys and `len` d returns the size of this set.

Appendix 4 provides details on all the values that CXL currently supports.

CXL Bytecode

A CXL program is translated into *bytecode*, which is list of machine instructions that the CXL virtual machine (CVM) executes. The CVM is not an ordinary virtual machine, but its architecture is nonetheless representative of conventional computers and virtual machines such as the Java Virtual Machine.

Instead of bits and bytes, a CVM manipulates CXL values. A CVM has the following components:

- Code: This is an immutable and finite list of CVM instructions, generated from a CXL program. The types of instructions will be described later.
- Shared memory: A CVM has just one memory location containing an CXL value.
- Processes: Any process can spawn an unbounded number of other processes and processes may terminate. Each process has an immutable (but not necessarily unique) *name tag*, a program counter, a stack of CXL values, and a single mutable general purpose *register* that contains a CXL value.

A name tag consists of the name of the main method of the process, along with an optional tag specified in the `spawn` statement. The default tag is the first argument to the method. A name tag is represented as a dictionary with keys `.name` and `.tag`, but we shall often abbreviate it using the notation `name/tag`. For example, in Figure 3.1, the created processes have name tags `incrementer/0`, `incrementer/1`, and `main/()`.

The state of a process is called a *context* (aka *continuation*): it contains the value of its name tag, program counter, stack, and register. The state of a CVM consists of the value of its memory and the multiset (or *bag*) of contexts. It is a multiset of contexts because two different processes can be in the same state. The initial state of the CXL memory is the empty dictionary, `dict{}`. The context bag has an initial context in it with name tag `__main__/()`, pc 0, register `()`, and an empty stack. Each machine instruction updates the state in some way.

It may seem strange that there is only one memory location and that each process has only one register. However, this is not a limitation because CXL values are unbounded trees. Both the memory and the register of a process always contain a dictionary that maps atoms to CXL values.

```
Up.cxl:1 def incrementer(self):
    0 Jump 34
    1 Frame incrementer(self)
Up.cxl:2     count = count + 1;
    2 Load count
    3 Push 1
    4 2-ary +
    5 Store count
Up.cxl:3     done[self] = True;
    6 LoadVar self
    7 PushAddress done
    8 Address 2
    9 Push True
   10 Store
   11 Return
```

Figure 4.1: The first part of the CVM bytecode corresponding to Figure 3.1.

We call this a *directory*. A directory represents the state of a collection of variables named by the atoms. Because directories are CXL values themselves, directories can be organized into a tree. Each node in a directory tree is then identified by a sequence of atoms, like a path name in the file system hierarchy. We call such a sequence the *address* of a CXL value, and it is relative to the "root" of the directory tree.

Compiling the code in Figure 3.1 results in the CVM bytecode listed in Figure 4.1. You can obtain this code by invoking `cxl` with the `-a` flag like so:

```
cxl -a Up.cxl
```

```
#states = 79 diameter = 5
==== Safety violation ====
__init__/()   [0,34-49]     50 { count=0, done=[False, False] }
incrementer/0 [1-4]          5 { count=0, done=[False, False] }
incrementer/1 [1-10]        11 { count=1, done=[False, True]  }
incrementer/0 [5-10]        11 { count=1, done=[True,  True]  }
main/() [13-17,20-24,26-31] 31 { count=1, done=[True,  True]  }
>>> CXL Assertion failed : 1
Open file://.../cxl.html for more information
```

Figure 4.2: The text output of running CXL on Figure 3.1.

Each process in the CVM is predominantly a *stack machine*, but it also has a register. All instructions are atomically executed. Most instructions pop values from the stack or push values onto the stack. At first there is one process, with name tag `__init__/()`, that initializes the state. It starts executing at instruction 0 and keeps executing until the last execution in the program. In this case, the first instruction is a `JUMP` instruction that sets the program counter to 39. At program counter 1 is the code for the `incrementer` method. All methods start with a `Frame` instruction and end with a `Return` instruction.

Appendix E provides a list of all CVM machine instructions, in case you want to read about the details. The `Frame` instruction lists the name of the method and the names of its arguments. The code generated from $count := count + 1$ in line 2 of `Up.cxl` is as follows (see Figure 4.1):

2. The `Load` instruction pushes the value of the *count* variable onto the stack.

3. the `Push` instruction pushes the constant 1 onto the stack of the process.

4. `2-ary` is a `+` operation with 2 arguments. It pops two values from the stack (1 and the value of *count*), adds them and pushes the result back onto the stack.

5. The `Store` instruction pops a CXL value (the sum of the *count* variable and 1) and stores it in the *count* variable.

Once initialization completes, any processes that were spawned can run. You can think of CXL as trying every possible interleaving of processes executing instructions. Figure 4.2 shows the output produced by running CXL on the `Up.cxl` program.

CXL can report a variety of failure types:

- `Safety violation`: this means something went wrong with at least one of the executions of the program that it tried. This can include a failing assertion, divide by zero, using an uninitialized or non-existent variable, dividing a set by an integer, and so on. It also includes certain infinite loops. CXL will print at trace of the shortest bad execution that it found.

- `Non-terminating States`: CXL found one or more states from which there does not exist an execution such that all processes terminate. CXL will not only print the non-terminating state with the shortest trace, but also the list of processes that are running at that state, along with their name tags and program counters.

- `Stopped States`: Similar to non-terminating states, these are states in which some processes are left that cannot terminate. Chapter 9 will explain what it means for a process to be stopped.

In this case, CXL reports a safety violation, in particular it reports that the assertion on line 9 failed and that *count* has value 1 instead of 2. Next it reports an execution that failed this assertion. The program got to the faulty execution in 5 "steps." The output has four columns:

1. The name tag of the process;

2. The sequence of program counters of the CVM instructions that the process executed;

3. The current program counter of the process;

4. The contents of the shared memory.

file://.../cxl.html

Issue: process failure		Shared Variables	
Process	**Steps**	**count**	**done**
__init__/()	0 34-50	0	[False, False]
incrementer/0	1-4	0	[False, False]
incrementer/1	1-11	1	[False, True]
incrementer/0	5-11	1	[True, True]
main/()	13-21 23-24 26-31	1	[True, True]

1	Frame incrementer(self)
Up.cxl:2 count = count + 1;	
2	Load count
3	Push 1
4	2-ary +
5	Store count
Up.cxl:3 done[self] = True;	
6	PushAddress done
7	LoadVar self
8	Address 2

Process	Status	Stack Trace	Variables
main/()	failed	31 main(): assert count == 2, count; CXL Assertion failed : 1	result = ()

Figure 4.3: The HTML output of running CXL on Figure 3.1. There are three sections. The top shows the steps to the problematic state. The bottom left shows the source code and bytecode of the program. The bottom right shows the state of each process after the selected step. Each row in the stack trace of a process shows the current program counter, the method that is being evaluated, and the line of code that is being evaluated.

The first step is initialization. It sets shared variable *count* to 0 and shared variable *done* to `[False, False]`. If we look at the next three steps, we see that:

- Process `implementer/0` executed instructions 1 through 4, loading the value of *count* but stopping just before storing 1 into *count*;
- Process `implementer/1` executed instructions 1 through 10, storing 1 into into *count* and storing `True` into *done*[1];
- Process `implementer/0` continues execution, storing value 1 into *count* and storing `True` into *done*[0].

Finally, process `main()` runs and finds the problem. This makes precise the concurrency issue that we encountered.

CXL also generates an HTML file that allows exploring more details of the execution interactively. Open the suggested HTML file and you should see something like Figure 4.3. If you hover the mouse over a machine instruction, it provides a brief explanation.

Exercises

4.1 Figure 4.4 shows an attempt at trying to fix the code of Figure 3.1. Run it through CXL and see what happens. Based on the error output, describe in English what is wrong with the code by describing, in broad steps, how running the program can get into a bad state.

4.2 What if we moved line 2 of Figure 4.4 to after the **if** statement (between lines 7 and 8)? Do you think that would work? Run it through CXL and describe either why it works or why it does not work.

```
def incrementer(self):
    entered[self] = True;
    if entered[1 − self]:
        while not done[1 − self]:
            pass;
        ;
    ;
    count = count + 1;
    done[self] = True;
;
def main():
    while not (done[0] and done[1]):
        pass;
    ;
    assert count == 2, count;
;
count = 0;
entered = [ False, False ];
done = [ False, False ];
spawn incrementer(0);
spawn incrementer(1);
spawn main();
```

Figure 4.4: [UpEnter.cxl] Broken attempt at fixing the code of Figure 3.1.

Chapter 5

Critical Sections

Hopefully you have started thinking of how to solve the concurrency problem and you may already have prototyped some solutions. In this chapter we will go through a few reasonable but broken attempts. At the heart of the problem is that we would like make sure that, when the *count* variable is being updated, no other process is trying to do the same thing. This is called a *critical section* (aka critical region) [Dij65]: a set of instructions where only one process is allowed to execute at a time.

Critical sections are useful when accessing a shared data structure, particularly when that access requires multiple underlying machine instructions. A counter is a very simple example of a data structure, but as we have seen it too requires multiple instructions. A more involved one would be accessing a binary tree. Adding a node to a binary tree, or re-balancing a tree, often requires multiple operations. Maintaining "consistency" is certainly much easier (although not necessarily impossible) if during this time no other process also tries to access the binary tree. An implementation of a data structure that can be safely accessed by multiple processes (or threads) and free of race conditions is called *thread-safe*.

```
def process(self):
   while True:
      #enter critical section
      @cs: assert atLabel.cs == dict{ nametag(): 1 };
      #exit critical section
   ;
;
spawn process(0);
spawn process(1);
```

Figure 5.1: [csbarebones.cxl] A bare bones critical section with two processes.

```
def process(self):
    while choose({ False, True }):
        #enter critical section
        @cs: assert atLabel.cs == dict{ nametag(): 1 };
        #exit critical section
    ;
;
spawn process(0);
spawn process(1);
```

Figure 5.2: [cx.cxl] CXL model of a critical section.

A critical section is often modeled as processes in an infinite loop entering and exiting the critical section. Figure 5.1 shows the CXL code. Here `@cs` is a *label*, identifying a location in the CVM bytecode. The first thing we need to ensure is that there can never be two processes in the critical section. This property is called *mutual exclusion*. We would like to place an assertion at the `@cs` label that specifies that only the current process can be there.

CXL in fact supports this. It has an operator **atLabel** L, where L is the atom containing the name of the label (in this case, `.cs`). The operator returns a bag (multiset) of name tags of processes executing at that label. The bag is represented by a dictionary that maps each element in the bag to the number of times the element appears in the bag. Method **atLabel** only exists for specification purposes—do not use it in normal code. The assertion also makes use of the **nametag**() operator that returns the name tag of the current process. If you run the code through CXL, the assertion should fail because there is no code yet for safely entering and exiting the critical section.

However, mutual exclusion by itself is easy to ensure. For example, we could insert the following code to enter the critical section:

```
while True:
    pass;
;
```

This code will surely prevent two or more processes from being at label `@cs` at the same time. But it does so by preventing *any* process from reaching the critical section. We clearly need another property besides mutual exclusion.

Mutual exclusion is an example of a *safety property*, a property that ensures that *nothing bad will happen*, in this case two processes being in the critical section. What we need now a *liveness property*: we want to ensure that *eventually something good will happen*. There are various possible liveness properties we could use, but here we will propose the following informally: if (1) there exists a non-empty set S of processes that are trying to enter the critical section and (2) processes in the critical section always leave eventually, then eventually one process in S will enter the critical section. We call this *progress*.

In order to detect violations of progress, and other liveness problems in algorithms in general, CXL requires that every execution must be able to reach a state in which all processes have terminated. Clearly, even if mutual exclusion holds in Figure 5.1, the spawned processes never terminate.

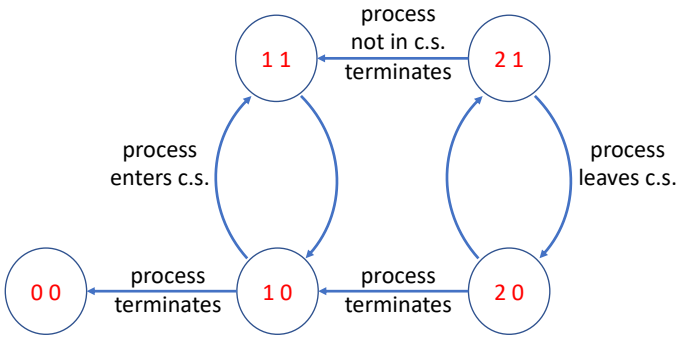

Figure 5.3: High-level state diagram specification of mutual exclusion with up to two processes. The first number in a state gives the number of processes; the second number is the number of processes in the critical section.

We will model processes in critical sections using the framework in Figure 5.2: a process can *choose* to enter a critical section more than once, but it can also choose to terminate, even without entering the critical section ever. (Recall that CXL will try every possible execution, and so it will evaluate both choices.)

Figure 5.3 shows a high-level state diagram of mutual exclusion and progress.The circles represent states, while the arrows represent possible state transitions (high-level steps). The label in each circle summarizes the state; the label on an arrow describes the transition. In this case, the label contains two numbers: the total number of processes and the number of processes in the critical section. A process that is in the critical section cannot terminate until after leaving the critical section. We will now consider various approaches toward implementing this specification.

You may already have heard of the concept of a *lock* and have realized that it could be used to implement a critical section. The idea is that the lock is like a baton that at most one process can own (or hold) at a time. A process that wants to enter the critical section at a time must obtain the lock first and release it upon exiting the critical section. Note that the word "lock" does not describe the concept well—it really is more like a baton or token that only one process can hold.

Using a lock is a good thought, but how does one implement one? Figure 5.4 presents a mutual exclusion attempt based on a naïve (and, as it turns out, broken) implementation of a lock. Initially the lock is not owned, indicated by *lockTaken* being `False`. To enter the critical section, a process waits until *lockTaken* is `False` and then sets it to `True` to indicate that the lock has been taken. The process then executes the critical section. Finally the process releases the lock by setting *lockTaken* back to `False`.

Unfortunately, if we run the program through CXL, we find that the assertion still fails. Diagnosing the problem, we see that the *lockTaken* variable suffers from the same problem as the *count* variable in Figure 3.1: operations on it consist of several instructions. It is thus possible for both processes to believe the lock is available and to obtain the lock at the same time.

Figure 5.5 presents a solution based on each process having a flag indicating that it is trying to enter the critical section. A process can write its own flag and read the flag of its peer. After setting its flag, the process waits until the other process $(1 - self)$ is not trying to enter the critical

```
def process(self):
    while choose({ False, True }):
        # Enter critical section
        while lockTaken:
            pass;
        ;
        lockTaken = True;

        # Critical section
        @cs: assert atLabel.cs == dict{ nametag(): 1 };

        # Leave critical section
        lockTaken = False;
    ;
;
lockTaken = False;
spawn process(0);
spawn process(1);
```

Figure 5.4: [naiveLock.cxl] Naïve implementation of a shared lock.

```
def process(self):
    while choose({ False, True }):
        # Enter critical section
        flags[self] = True;
        while flags[1 − self]:
            pass;
        ;

        # Critical section
        @cs: assert atLabel.cs == dict{ nametag(): 1 };

        # Leave critical section
        flags[self] = False;
    ;
;
flags = [ False, False ];
spawn process(0);
spawn process(1);
```

Figure 5.5: [naiveFlags.cxl] Naïve use of flags to solve mutual exclusion.

```
def process(self):
    while choose({ False, True }):
        # Enter critical section
        while turn == (1 - self):
            pass;
        ;

        # Critical section
        @cs: assert atLabel.cs == dict{ nametag(): 1 };

        # Leave critical section
        turn = 1 - self;
    ;
;
turn = 0;
spawn process(0);
spawn process(1);
```

Figure 5.6: [naiveTurn.cxl] Naïve use of turn variable to solve mutual exclusion.

section. If we run this program, the assertion does not fail. In fact, this solution does prevent both processes being in the critical section at the same time.

To see why, first note the invariant that if process i is in the critical section, then *flags*$[i]$ == `True`. Without loss of generality, suppose that process 0 sets *flags*[0] at time t_0. Process 0 can only reach the critical section if at some time t_1, $t_1 > t_0$, it finds that *flags*[1] == `False`. Because of the invariant, *flags*[1] == `False` implies that process 1 is not in the critical section at time t_1. Let t_2 be the time at which process 0 sets *flags*[0] to `False`. Process 0 is in the critical section sometime between t_1 and t_2. It is easy to see that process 1 cannot enter the critical section between t_1 and t_2, because *flags*[1] == `False` at time t_1. To reach the critical section between t_1 and t_2, it would first have to set *flags*[1] to `True` and then wait until *flags*[0] == `False`. But that does not happen until time t_2.

However, if you run the program through CXL, it turns out the solution does has a problem: if both try to enter the critical section at the same time, they may end up waiting for one another indefinitely. Thus the solution violates *progress*.

The final naïve solution that we propose is based on a variable called *turn* that alternates between 0 and 1. When *turn* $= i$, process i can enter the critical section, while process $1 - i$ has to wait. When done, process i sets *turn* to $1 - i$ to give the other process an opportunity to enter. An invariant of this solution is that while process i is in the critical section, *turn* $== i$. Since *turn* cannot be 0 and 1 at the same time, mutual exclusion is satisfied. The solution also has the nice property that processes 0 and 1 alternate entering the critical section.

Run the program through CXL. It turns out that this solution also violates *progress*, albeit for a different reason: if process i terminates instead of entering the critical section when it is process i's turn, process $1 - i$ ends up waiting indefinitely for its turn.

Exercises

5.1 Run Figure 5.2 using CXL. As there is no protection of the critical section, mutual exclusion is violated, the assertion should fail, and a trace should be reported. Now plug the following code **while** `True`: **pass**;; just before entering the critical section in Figure 5.2 and run CXL again. Mutual exclusion is guaranteed but progress is violated. CXL should print a trace to a state from which a terminating state cannot be reached. Describe in English the difference in the failure reports before and after inserting the code.

5.2 See if you can come up with some different approaches that satisfy both mutual exclusion and progress. Try them with CXL and see if they work or not. If they don't, try to understand why. Do not despair if you can't figure it out—as we will find out, it is possible but not easy.

Chapter 6

Peterson's Algorithm

In 1981, Peterson came up with a beautiful solution to the mutual exclusion problem, now known as "Peterson's Algorithm" [Pet81]. The algorithm is an amalgam of the (incorrect) algorithms in Figures 5.5 and 5.6, and is presented in Figure 6.1. A process first indicates its interest in entering the critical section by setting its flag. It then politely gives way to the other process should it also want to enter the critical section—if both do so at the same time one will win because writes to memory in CXL are atomic. The process continues to be polite, waiting in a **while** loop until either the other process is nowhere near the critical section ($flag[1 - self] == \texttt{False}$) or has given way ($turn == self$). Running the algorithm with CXL shows that it satisfies both mutual exclusion and progress.

Why does it work? We will focus here on how one might go about proving mutual exclusion for an algorithm such as Peterson's. For that, we have to understand a little bit more about how the CXL machine works. In Chapter 4 we talked about the concept of *state*: at any point in time the CVM is in a specific state. A state is comprised of the values of the shared variables as well as the values of the process variables for each process, including its program counter and the contents of its stack. Everytime a process executes a CVM machine instruction, the state changes (if only because the program counter of the process changes). We call that a *step*. Steps in CXL are atomic.

The CVM starts in an initial state in which there is only one process and its program counter is 0. A *trace* is a sequence of steps starting from the initial state. When making a step, there are two sources of non-determinism in CXL. One is when there is more than one process that can make a step. The other is when a process executes a `choose` operation and there is more than one choice. Because there is non-determinism, there are multiple possible traces. We call a state *reachable* if it is either the initial state or it can be reached from the initial state through a trace. A state is final when there are no processes left to make state changes.

It is often useful to classify states into sets. *Initial*, *final*, and *reachable*, and *unreachable* are all examples of classes of states. Figure 6.2 depicts a Venn diagram of various classes of states and a trace. One way to classify states is to define a predicate over states. All states in which $x = 1$, or all states where there are two or more processes executing, are examples of such predicates. For our purposes, it is useful to define a predicate that says that at most one process is in the critical section. We shall call such states *exclusive*.

```
def process(self):
    while choose({ False, True }):
        # Enter critical section
        flags[self] = True;
        turn = 1 - self;
        while flags[1 - self] and (turn == (1 - self)):
            pass;
        ;

        # critical section is here
        @cs: assert atLabel.cs == dict{ nametag(): 1 };

        # Leave critical section
        flags[self] = False;
    ;
;
flags = [ False, False ];
turn = choose({0, 1});
spawn process(0);
spawn process(1);
```

Figure 6.1: [Peterson.cxl] Peterson's Algorithm

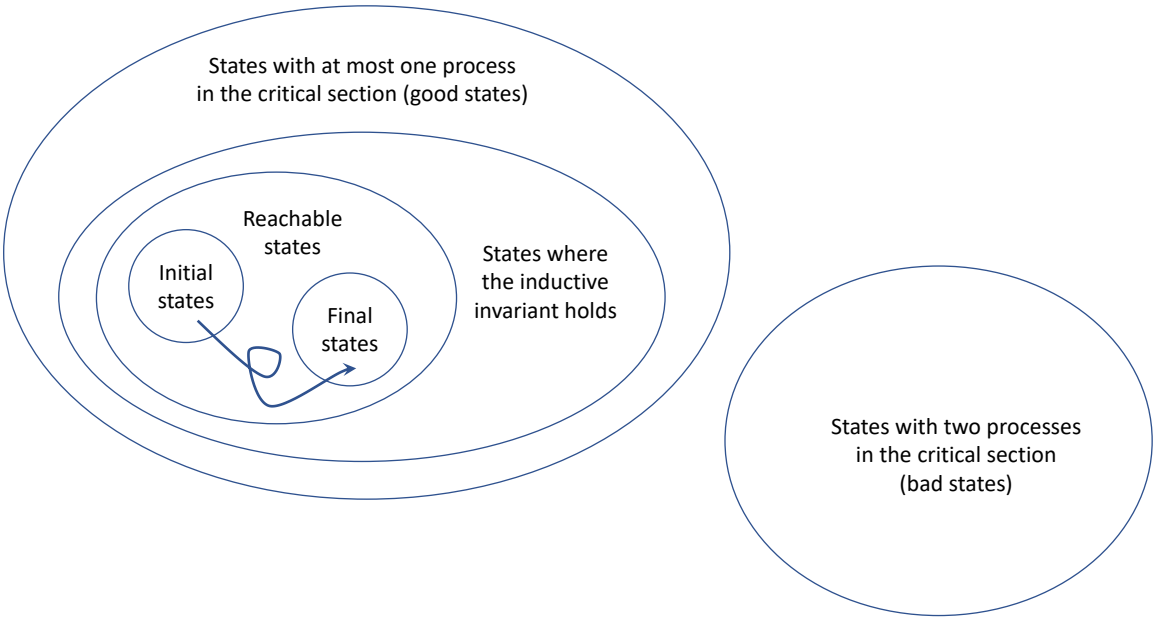

Figure 6.2: Venn diagram classifying all states and a trace.

An *invariant* of a program is a predicate that holds over all states that are reachable by that program. We want to show that exclusivity is an invariant. In other words, we want to show that the set of reachable states of executing the program is a subset of the set of states where there is at most one process in the critical section.

One way to prove that a predicate is an invariant is through induction on the number of steps. First you prove that the predicate holds over the initial state. Then you prove that for every reachable state, and for every step from that reachable state, the predicate also holds over the resulting state. For this you need a predicate that describes exactly which states are reachable. But we do not have such a predicate: we know how to describe the set of reachable states, but given an arbitrary state it is not easy to see whether it is reachable or not.

To solve this problem, we will use what is called an *inductive invariant*. An inductive invariant $\mathcal{I}$ is a predicate over states that satisfies the following:

- $\mathcal{I}$ holds in the initial state.
- For any state in which $\mathcal{I}$ holds and any process in the state takes a step, then $\mathcal{I}$ also holds in the resulting state.

Unlike an invariant, an inductive invariant must also hold over unreachable states.

One candidate for such an invariant is exclusivity itself. After all, it certainly holds over the initial state. And as CXL has already determined, exclusivity is an invariant: it holds over every reachable state. Unfortunately, it is not an *inductive* invariant. To see why, we need to consider an *unreachable* state. It is easy to construct one: let process 0 be at label `@cs` and process 1 at the start of the **while** loop. Also imagine that in this particular state $turn = 1$. Now let process 1 make

```
def process(self):
   while choose({ False, True }):
      # Enter critical section
      flags[self] = True;
      @gate: turn = 1 − self;
      while flags[1 − self] and (turn == (1 − self)):
         pass;
      ;

      # Critical section
      @cs: assert (not (flags[1 − self] and (turn == (1 − self))))
               or (atLabel.gate == dict{nametags[1 − self] : 1})
         ;

      # Leave critical section
      flags[self] = False;
   ;
;
flags = [ False, False ];
turn = choose({0, 1});
nametags = [ dict{ .name: .process, .tag: tag } for tag in 0..1 ];
spawn process(0);
spawn process(1);
```

Figure 6.3: [PetersonInductive.cxl] Peterson's Algorithm with Inductive Invariant

a sequence of steps. Because $turn = 1$, process 1 will break out of the while loop and also enter the critical section, entering a state that is not exclusive. So exclusivity is an invariant, but not an inductive invariant. Doing a inductive proof with an invariant that is not inductive is usually much harder than doing one with an invariant that is.

So we are looking for an inductive invariant that *implies* exclusivity. In other words, the set of states where the inducive invariant holds must be a subset of the set of states where there is at most one process in the critical section. Let $\mathcal{C}(i) = flags[1-i] \wedge turn = 1-i$, that is, the condition on the **while** loop for process i. Let predicate $\mathcal{I}_p(i)$ be the following. if process i is at label `@cs` (i.e., process i is in the critical section), then $\mathcal{C}(i)$ does not hold or process $1-i$ is executing after setting $flags[i]$ but still before setting $turn$ to $1-i$. More formally, $\mathcal{I}_p(i) = \texttt{process}(i)@cs) \Rightarrow (\neg\mathcal{C}(i) \vee \texttt{process}(1-i)@gate)$. For Peterson's Algorithm, an inductive invariant that works well is $\mathcal{I}_p(0) \wedge \mathcal{I}_p(1)$.

Figure 6.3 formalizes $\mathcal{I}_p(i)$ in CXL. The label `@gate` refers to the instruction that sets $turn$ to $1-i$. You can run Figure 6.3 to determine that $\mathcal{I}_p(i)$ is indeed an invariant for $i = 0, 1$.

To see that the inductive invariant implies exclusivity, suppose not. Then when both process 0 and process 1 are in the critical section, the following must hold: $(\neg\mathcal{C}(0) \vee \texttt{process}(1)@gate) \wedge (\neg\mathcal{C}(1) \vee \texttt{process}(0)@gate)$. We know that both flags are set. We also know that neither process 0

nor process 1 is at label `@gate` (because they are both at label `@cs`), so this simplifies to $(\neg\mathcal{C}(0)) \wedge (\neg\mathcal{C}(1))$. So we conclude that $turn = 0 \wedge turn = 1$, a logical impossibility. Thus $\mathcal{I}_p(i)$ implies exclusivity.

To see that $\mathcal{I}_p(0) \wedge \mathcal{I}_p(1)$ is, in fact, an inductive invariant, first note that it certainly holds in the initial state, because in the initial state no process is in the critical section. Without loss of generality, suppose $i = 0$ (a benefit from the fact that the algorithm is symmetric for both processes). We still have to show that if we are in a state in which $\mathcal{I}_p(0)$ holds, then any step will result in a state in which $\mathcal{I}_p(0)$ still holds. If $\mathcal{I}_p(0)$ holds, process 0 is at label `@cs`. If process 0 were to take a step, then in the next state process 0 would be no longer at that label and $\mathcal{I}_p(0)$ would hold trivially over the next state. Therefore we only need to consider a step by process 1.

From $\mathcal{I}_p(0)$ we know that one of the following three cases must hold before process 1 takes a step:

1. $flags[1] =$ `False`;
2. $turn = 0$;
3. process 1 is at label `@gate`.

Let us consider each of these cases. In the first case, if process 1 takes a step, there are two possibilities: either $flags[1]$ will still be `False` (in which case the first case continues to hold), or $flags[1]$ will be `True` and process 1 will be at label `@gate` (in which case the third case will hold). We know that process 1 never sets *turn* to 1, so if the second case holds before the step, it will also hold after the step. Finally, if process 1 is at label `@gate` before the step, then after the step *turn* will equal 0, and therefore the second case will hold after the step. qed.

We have now demonstrated mutual exclusion in Peterson's Algorithm in two different ways: one by letting CXL explore all possible executions, the other using an inductive invariant and proof by induction. The former is certainly easier, but it does not provide intuition for why the algorithm works. The second provides much more insight. We therefore encourage to include inductive invariants in your CXL code.

A cool anecdote is the following. When the author of CXL had to teach Peterson's Algorithm, he refreshed his memory by looking at the Wikipedia page. The page claimed that the following predicate is invariant: if process i is in the critical section, then $\neg\mathcal{C}(i)$ (i.e., $\mathcal{I}_p(i)$ without the disjunct that process $1 - i$ is at label `@gate`). To demonstrate that this predicate is not invariant, you can remove the disjunct from Figure 6.3 and run it to get a counterexample.

This anecdote suggests the following. If you need to do an induction proof of an algorithm, you have to come up with an inductive invariant. Before trying to prove the algorithm, you can check that the predicate is at least invariant by testing it using CXL. Doing so could potentially avoid wasting your time on a proof that will not work because the predicate is not invariant, and therefore not an inductive invariant either. (The author fixed the Wikipedia page.)

Exercises

6.1 Figure 6.4 presents another correct solution to the mutual exclusion problem. It is similar to the one in Figure 5.5, but has a process *back out and try again* if it finds that the other process is either trying to enter the critical section or already has. Compare this algorithm with Peterson's. What are the advantages and disadvantages?

```
def process(self):
    while choose({ False, True }):
        # Enter critical section
        flags[self] = True;
        while flags[1 − self]:
            flags[self] = False;
            flags[self] = True;
        ;

        # Critical section
        @cs: assert atLabel.cs == dict{ nametag(): 1 };

        # Leave critical section
        flags[self] = False;
    ;
;
flags = [ False, False ];
spawn process(0);
spawn process(1);
```

Figure 6.4: [csonebit.cxl] Mutual exclusion using a flag per process.

6.2 Can you find an inductive invariant for the algorithm in Figure 6.4 to prove it correct? Here's a pseudo-code version of the algorithm to help you. Each line is an atomic action:

```
initially: flagX = flagY = False

process X:                          process Y:
    X0: flagX = True                   Y0: flagY = True
    X1: if not flagY goto X4           Y1: if not flagX goto Y4
    X2: flagX = False                  Y2: flagY = False
    X3: goto X0                        Y3: goto Y0
    X4: ...critical section...         Y4: ...critical section...
    X5: flagX = False                  Y5: flagY = False
```

6.3 A colleague of the author asked if the first two assignments in Peterson's algorithm (setting *flags*[*self*]) to `True` and *turn* to 1 - *self*) can be reversed. After all, they are different variables assigned independent values—in a sequential program one could surely swap the two assignments. See if you can figure out for yourself if the two assignments can be reversed. Then run the program in Figure 6.1 after reversing the two assignments and describe in English what happens.

6.4 Bonus question: Can you generalize Peterson's algorithm to more than two processes?

Chapter 7

CXL Methods and Pointers

A method `m` with argument a is invoked in its most basic form as follows (assigning the result to `r`).

```
r = m a;
```

That's right, no parentheses are required. In fact, if you invoke `m`(a), the argument is (a), which is the same as a. If you invoke `m()`, the argument is `()`, which is the empty tuple. If you invoke `m`(a, b), the argument is (a, b), the tuple consisting of values a and a.

You may note that all this looks familiar. Indeed, the syntax is the same as that for dictionaries (see Chapter 4). Both dictionaries and methods map CXL values to CXL values, and their syntax is indistinguishable. If `f` is either a method or a dictionary, and x is an arbitrary CXL value, then `f` x, `f`(x), and `f`[x] are all the same expression in CXL.

CXL is not an object-oriented language like Python is. In Python you can pass a reference to an object to a method, and that method can then update the object. In CXL, it is also sometimes convenient to have a method update a shared variable specified as an argument. For this (as well as some other uses), CXL supports *pointers* to shared variables. If x is a shared variable, then the expression &x is a pointer to x (also known as the *address* of x). Conversely, if p is a pointer to a shared variable, then the expression `^`p is the value of the shared variable. In particular, `^&`x == x. This is similar to how C pointers work.

Figure 7.1 again shows Peterson's algorithm, but this time with methods defined to enter and exit the critical section. The name *mutex* is often used to denote a variable or value that is used for mutual exclusion. `P_mutex` is a method that returns a "mutex," which, in this case, is a dictionary that contains Peterson's Algorithm's state: a turn variable and two flags. Both methods `P_enter` and `P_exit` take two arguments: a pointer to a mutex and the process identifier (0 or 1). `(^`pm`.turn` is the value of the `.turn` key in the dictionary that pm points to. Note that the parentheses in this expression are needed, as `^`pm`.turn` would wrongly evaluate to `^`(pm`.turn`).

You can put the first three methods in its own CXL source file and include it using the CXL `import` statement. This would make the code re-usable by other applications.

For those who are extra adventurous: you can add the `P_enter` and `P_exit` methods to the `P_mutex` dictionary like so:

```
dict{ .turn: 0, .flags: [ False, False ], .enter: P_enter, .exit: P_exit }
```

```
def P_enter(pm, pid):
    (^pm).flags[pid] = True;
    (^pm).turn = 1 - pid;
    while (^pm).flags[1 - pid] and ((^pm).turn == (1 - pid)):
        pass;
    ;
;
def P_exit(pm, pid):
    (^pm).flags[pid] = False;
;
def P_mutex():
    result = dict{ .turn: choose({0, 1}), .flags: [ False, False ] };
;

#### The code above can go into its own CXL module ####

def process(self):
    while choose({ False, True }):
        P_enter(&mutex, self);
        @cs: assert atLabel.cs == dict{ nametag(): 1 };
        P_exit(&mutex, self);
    ;
;
mutex = P_mutex();
spawn process(0);
spawn process(1);
```

Figure 7.1: [PetersonMethod.cxl] Peterson's Algorithm accessed through methods.

That would allow you to simulate object methods.

Chapter 8

Spinlock

Figure 5.4 showed a faulty attempt at solving mutual exclusion using a lock. The problem with the implementation of the lock is that checking the lock and setting it if it is available is not *atomic*. Thus multiple processes contending for the lock can all "grab the lock" at the same time. While Peterson's algorithm gets around the problem, it is not efficient, especially if generalized to multiple processes. Instead, multi-core processors provide so-called *interlock instructions*: special machine instructions that can read memory and then write it in an indivisible step.

While the CVM does not have any specific built-in interlock instructions, it does have support for executing multiple instructions atomically. This feature is available in the CXL language in two ways. First, any CXL statement can be made atomic by placing a label in front of it. Second, a group of CXL statements can be made atomic through its **atomic** statement. We can use **atomic** statement blocks to implement a wide variety of interlock operations. For example, we could fix the program in Figure 3.1 by constructing an atomic increment operation for a counter, like so:

```
1   def atomic_inc(ptr):
2       atomic:
3           ^ptr += 1;
4       ;
5   ;
6   count = 0;
7   atomic_inc(&count);
```

Many CPUs have an atomic "test-and-set" (TAS) operation. Method `tas` in Figure 8.1 shows its specification. Here s points to a shared Boolean variable and p to a private Boolean variable, belonging to some process. The operation copies the value of the shared variable to the private variable (the "test") and then sets the shared variable to `True` ("set").

Figure 8.1 goes on how to implement mutual exclusion for a set of `N` processes. The approach is called *spinlock*, because each process is "spinning" until it can acquire the lock. The program uses `N` + 1 variables. Variable *shared* is initialized to `False` while *private*[i] for each process i is initialized to `True`. An important invariant, $\mathcal{I}_1$, of the program is that at any time at most one of

```
const N = 3;

def tas(s, p):
    atomic:
        ^p = ^s;
        ^s = True;
    ;
;
def process(self):
    while choose({ False, True }):
        # Enter critical section
        while private[self]:
            tas(&shared, &private[self]);
        ;

        # Critical section
        @cs: assert (not private[self]) and
                (atLabel.cs == dict{ nametag(): 1 })
            ;

        # Leave critical section
        private[self] = True;
        shared = False;
    ;
;
shared = False;
private = [ True for i in 0..(N–1) ];
for i in 0..(N–1):
    spawn process(i);
;
```

Figure 8.1: [spinlock.cxl] Mutual Exclusion using a “spinlock” based on test-and-set.

these variables is `False`. Another invariant, $\mathcal{I}_2(i)$, is that if process i is in the critical section, then *private*$[i]$ == `False`. Between the two (i.e., $\mathcal{I}_1 \wedge \forall i : \mathcal{I}_2(i)$), it is clear that only one process can be in the critical section at the same time.

$\mathcal{I}_1$ is an inductive invariant. To see that invariant $\mathcal{I}_1$ is maintained, note that `^`p == `True` upon entry of `tas` (because of the condition on the **while** loop that the `tas` method is invoked in). So there are two cases:

1. `^`s is `False` upon entry to `tas`. Then upon exit `^`p == `False` and `^`s == `True`, maintaining the invariant.

2. `^`s is `True` upon entry to `tas`. Then upon exit nothing has changed, maintaining the invariant.

Invariant $\mathcal{I}_1$ is also easy to verify for exiting the critical section. Invariant $\mathcal{I}_2(i)$ is obvious as (i) process i only proceeds to the critical section if *private*$[i]$ == `False`, and (ii) no other process modifies *private*$[i]$.

CXL can check these invariants as well. Figure 8.1 already has the code to check $\mathcal{I}_2(i)$. But how would one go about checking an invariant like $\mathcal{I}_1$? Invariants must hold for every state. For $\mathcal{I}_2$ we only need an assertion at label `@cs` because the premisse is that there is a process at that label. However, we would like to check $\mathcal{I}_1$ in *every state* (after the variables have been initialized).

We can do this by adding another process that, in a loop, checks the invariant. Figure 8.2 shows the code. Method `checkInvariant()` checks to see if the invariant holds in a state. It introduces a new feature of CXL: the ability to have variables local to a method. In this case, the process variable *sum* is used to compute the number of shared variables that have value `False`. The function is invoked by in a loop by a process that runs alongside the other processes. In CXL, **assert** statements are executed atomically, so the evaluation of the assertion is not interleaved with the execution of other processes. Because CXL tries every possible execution, the process is guaranteed to find violations of the invariant if it does not hold.

Exercises

8.1 Implement an atomic swap operation. It should take two pointer arguments and swap the values.

8.2 Implement a spinlock using the atomic swap operation.

8.3 Write out the invariants that need to hold and check them using CXL.

```
const N = 3;

def checkInvariant():
    let sum = 0:
        if not shared:
            sum = 1;
        ;
        for i in 0..(N–1):
            if not private[i]:
                sum += 1;
            ;
        ;
        result = sum ¡= 1;
    ;
;
def invariantChecker():
    while choose({ False, True }):
        assert checkInvariant();
    ;
;

# tas() and process() code omitted to save space

shared = False;
private = [ True for i in 0..(N–1) ];
spawn invariantChecker();
for i in 0..(N–1):
    spawn process(i);
;
```

Figure 8.2: [spinlockInv.cxl] Checking invariants.

Chapter 9

Locks and Blocking

In Figure 8.1 we have shown a solution based on a shared variable and a private variable for each process. The private variables themselves are actually implemented as shared variables, but they are accessed only by their respective processes. There is no need to keep *private* as a shared variable—we only did so to be able to show and check the invariants. Figure 9.1 shows a more straightforward implementation of spinlock. The solution is similar to the naïve solution of Figure 5.4, but uses test-and-set to check and set the lock variable atomically. This approach is general for any number of processes.

It is important to appreciate the difference between an *atomic section* (the statements executed within an **atomic** statement) and a *critical section* (protected by a lock of some sort). The former ensures that while the atomic statement is executing no other process can execute. The latter allows multiple processes to run concurrently, just not within the critical section. The former is rarely available to a programmer, while the latter is very common.

In CXL, atomic statements allow you to *implement* your own synchronization primitives like test-and-set. Other common examples include compare-and swap and fetch-and-add. Atomic statements are not intended to *replace* locks or other synchonization primitives. When solving synchronization problems you should not directly use atomic statement but use the synchronization primitives that are available to you. But if you want to design a new synchronization primitive, then use **atomic** by all means.

Locks are probably the most prevalent and basic form of synchronization in concurrent programs. For this reason, CXL has a module called `synch` that includes support for locks. Figure 9.2 shows how they are implemented, and Figure 9.3 gives an example of how they may be used, in this case to fix the program of Figure 3.1. Notice that the module completely hides the implementation of the lock. The `synch` module includes a variety of other useful synchronization primitives, which will be discussed in later chapters.

We call a process *blocked* in a state if the process is waiting on a low-level synchronization variable and cannot terminate without the help of another process. A process trying to acquire a test-and-set spinlock held by another process is a good example of a process being blocked. The only way forward is if the other process releases the lock.

In most operating systems, processes are virtual and can be suspended until some condition changes. For example, a process can be suspended until the lock is available to it. In CXL, a

```
def tas(s):
    atomic:
        result = ^s;
        ^s = True;
    ;
;
def process():
    while choose({ False, True }):
        while tas(&lock):
            pass;
        ;
        @cs: assert atLabel.cs == dict{ nametag(): 1 };
        lock = False;
    ;
;
lock = False;
for i in 1..10:
    spawn process();
;
```

Figure 9.1: [csTAS.cxl] Fixed version of Figure 5.4 using test-and-set.

```
def tas(lk):
    atomic:
        result = ^lk;
        ^lk = True;
    ;
;
def Lock():
    result = False;
;
def lock(lk):
    while tas(lk):
        pass;
    ;
;
def unlock(lk):
    ^lk = False;
;
```

Figure 9.2: The `Lock` interface in the `synch` module.

```
import synch;

def process(self):
    lock(&countlock);
    count = count + 1;
    unlock(&countlock);
    done[self] = True;
;
def main(self):
    while not (done[0] and done[1]):
        pass;
    ;
    assert count == 2, count;
;
count = 0;
countlock = Lock();
done = [ False, False ];
spawn process(0);
spawn process(1);
spawn main();
```

Figure 9.3: [UpLock.cxl] Program of Figure 3.1 fixed with a lock.

```
import list;

def Lock():
    result = dict{ .locked: False, .suspended: [] };
;
def lock(lk):
    atomic:
        if (^lk).locked:
            stop (^lk).suspended;
            assert (^lk).locked;
        else:
            (^lk).locked = True;
        ;
    ;
;
def unlock(lk):
    atomic:
        if (^lk).suspended == []:
            (^lk).locked = False;
        else:
            go (head((^lk).suspended)) ();
            (^lk).suspended = tail((^lk).suspended);
        ;
    ;
;
```

Figure 9.4: The `Lock` interface in the `synchS` module uses suspension.

process can suspend itself and save its context (state) in a list. Recall that the context of a process consists of its name tag, its program counter, and the contents of its register and stack. A context is a regular CXL value. The syntax of the expression is as follows:

stop L

Here L is a shared variable containing a list. Another process can revive the process using the **go** statement:

go C R

Here C is a context and R is a CXL value. It causes a process with context C to be added to the state that has just executed the **stop** expression. The **stop** expression returns the value R.

There is a second version of the `synch` module that uses suspension instead of spinlocks. Figure 9.4 shows the same `Lock` interface implemented using suspension. The interface is implemented as follows:

- A `Lock` maintains both a boolean indicating whether the lock is taken and a list of contexts of processes that want to acquire the lock.
- `lock` sets the lock if available and suspends itself if not. Note that **stop** is called within an **atomic** statement—this is the only exception to an atomic statement running to completion. While the process is running no other processes can run, but when it suspends itself it allows other processes to run.
- `unlock` checks to see if any processes are waiting to get the lock. If so, it uses the `head` and `tail` methods from the `list` module (see Appendix D) to resume the first process that got suspended and to remove its context from the list.

Selecting the first process is a design choice. Another implementation could have picked the last one, and yet another implementation could have used `choose` to pick an arbitrary one. Selecting the first is a common choice in lock implementations as it prevents *starvation*: every process gets a chance to obtain the lock (assuming every process eventually releases it). Chapter 15 will talk more about starvation and how to prevent it.

CXL allows you to select which version of the `synch` module you would like to use with the `-m` flag. For example,

```
cxl -m synch=synchS x.cxl
```

runs the file `x.cxl` using the suspension version of the `synch` module. You will find that using the `synchS` module often leads to searching a significantly larger search space than using the `synch` module. Part of the reason is that the `synchS` module keeps track of the order in which processes wait for a lock.

Exercises

9.1 Write a synchronization module using your own implementation of `Lock` (for example, using Peterson's algorithm or using `swap`).

9.2 Run Figure 9.3 using (1) `synch`, (2), `synchS`, and (3) your own module. Report how many states were explored by CXL for each version of the module.

Chapter 10

Reader/Writer Locks using Busy Waiting

Locks are useful when accessing a shared data structure. By preventing more than one process from accessing the data structure at the same time, conflicting accesses are avoided. However, not all concurrent accesses conflict, and opportunities for concurrency may be lost, hurting performance. One important case is when multiple processes are simply reading the data structure. In many applications, reads are the majority of all accesses. Allowing reads to proceed concurrently can significantly improve performance.

What we want is a special kind of lock that allows either one writer or one or more readers to be in the critical section. This is called a *reader/writer lock* [CHP71]. We will explore various ways of implementing reader/writer locks in this chapter and future ones.

Figure 10.1 presents a solution that uses a single (ordinary) lock and two counters: one that maintains the number of readers and one that maintains the number of writers. Figure 10.2 shows how the code may be used. We will call the lock the *mutex* to distinguish it clearly from the reader/writer lock that we are implementing. The mutex is used to protect shared access to the counters. The program shows a process that executes in a loop. Each time, it decides whether to read or write. The critical section is spread between two labels: readers access `@rcs` and writers access `@wcs`. The specification is that if a reader is at label `@rcs`, no writer is allowed to be at label `@wcs`. Vice versa, if a writer is at label `@wcs`, no reader is allowed to be at label `@rcs` *and* there cannot be another writer at label `@wcs`. Figure 10.3 shows a high-level specification for two processes.

A process that wants to read first waits until there are no writers: $nwriters == 0$. Then it increments the number of readers. Similarly, a process that wants to write waits until there are no readers *or* writers. Then it increments the number of writers. The important invariants in this code are:

- n readers at `@rcs` $\Rightarrow nreaders \geq n$,
- n writers at `@wcs` $\Rightarrow nwriters \geq n$,
- $(nreaders \geq 0 \wedge nwriters = 0) \vee (nreaders = 0 \wedge 0 \leq nwriters \leq 1)$.

```
import synch;

def acquire_rlock():
    let blocked = True:
        while blocked:
            lock(&rwlock);
            if nwriters == 0:
                nreaders += 1;
                blocked = False;
            ;
            unlock(&rwlock);
        ;
    ;
;
def release_rlock():
    lock(&rwlock);
    nreaders -= 1;
    unlock(&rwlock);
;
def acquire_wlock():
    let blocked = True:
        while blocked:
            lock(&rwlock);
            if (nreaders == 0) and (nwriters == 0):
                nwriters = 1;
                blocked = False;
            ;
            unlock(&rwlock);
        ;
    ;
;
def release_wlock():
    lock(&rwlock);
    nwriters = 0;
    unlock(&rwlock);
;
rwlock = Lock();
nreaders = 0;
nwriters = 0;
```

Figure 10.1: [RW.cxl] Busy-Waiting Reader/Writer Lock implementation.

```
import RW;

def process():
    while choose({ False, True }):
        if choose({ .read, .write }) == .read:
            acquire_rlock();
            @rcs: assert atLabel.wcs == dict{};
            release_rlock();
        else:                    # .write
            acquire_wlock();
            @wcs: assert (atLabel.wcs == dict{ nametag(): 1 }) and
                    (atLabel.rcs == dict{})
                ;
            release_wlock();
        ;
    ;
;
for i in 1..4:
    spawn process();
;
```

Figure 10.2: [RWtest.cxl] Test code for Figure 10.1.

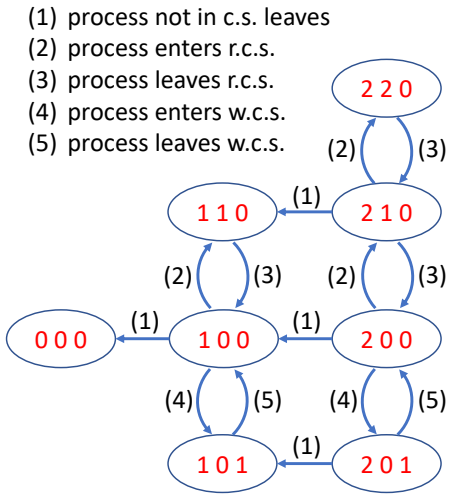

Figure 10.3: High-level state diagram specification of reader/writer locks with up to two processes. The first number in a state gives the number of processes; the second number is the number of processes reading in the critical section; the third is the number of processes writing in the critical section.

```
import RW;

const NREADERS = 3;
const NWRITERS = 3;

def reader(self):
    acquire_rlock();
    @rcs: assert atLabel.wcs == dict{};
    release_rlock();
;
def writer():
    acquire_wlock();
    @wcs: assert (atLabel.wcs == dict{ nametag(): 1 }) and
                (atLabel.rcs == dict{})
        ;
    release_wlock();
;
rwlock = Lock();
rlock = Lock();
nreaders = 0;
for i in 1..NREADERS:
    spawn reader(i);
;
for i in 1..NWRITERS:
    spawn writer();
;
acquire_rlock();
```

Figure 10.4: [RWbusychk.cxl] Checking for Busy Waiting.

It is easy to see that the invariants hold and imply the reader/writer specification. The solution also supports progress: if no process is in the critical section then any process can enter. Better still: if any reader is in the critical section, any other reader is also able to enter.

While correct, it is not considered a good solution. The solution is an example of what is called *busy-waiting* (aka *spin-waiting*): processes spin in a loop until some desirable application-level condition is met. Figure 3.1 also has an example of busy-waiting: the `main` process waits for the other two processes to finish by checking their `done` flags.

The astute reader might wonder if obtaining the mutex itself is an example of busy-waiting. After all, the CXL `synch` implementation of `lock()` spins in a loop until the mutex is available (see Figure 9.2). However, there is a big difference. As we pointed out, the processes that are waiting for a spinlock are *blocked*. In most operating systems and programming language runtimes, when a process acquires a mutex, the process is placed on a scheduling queue and stops using CPU cycles

until the mutex becomes available. CXL supports this with the `synchS` module. Busy waiting disables all this: even when using the `synchS` module, readers and writers only get suspended temporarily in case there is contention for the mutex. A process trying to obtain a read or write lock in Figure 10.1 has to obtain the mutex repeatedly until the read or write lock is available.

Thus, it is considered ok to have processes be blocked while waiting for application-specific conditions, but not for them to busy-wait for application-specific conditions.

CXL does not complain about Figure 10.1, but, using the right test, CXL can be used to check for busy-waiting. Figure 10.4 presents such a test. Note that the initialization code obtains a read lock, preventing all readers and writers from entering the critical section. In that case, those processes should ideally be blocked and CXL can test for that by running it with the `-b` flag. This flag tells CXL to check that all states are non-terminating and lead to a state in which all processes are blocked.

Exercises

10.1 Draw additional states and steps in Figure 10.3 for three processes.

10.2 Using busy waiting, implement a "bound lock" that allows up to `M` processes to acquire it at the same time. A bound lock with `M == 1` is an ordinary lock. You should define a constant `M` and two methods: `acquire_bound_lock()` and `release_bound_lock()`.

10.3 Write a test program for your bound lock that checks that no more than `M` processes can acquire the bound lock.

Chapter 11

Reader/Writer Locks with Blocking

Figure 11.1 presents a reader/writer lock implementation that does not busy-wait. You can test the code by running `cxl -m RW=RWlock RWtest`. It uses two ordinary locks: *rwlock* is used by both readers and writers, while *rlock* is only used by readers. *rwlock* is held either when readers are at label `@rcs` or when a writer is at label `@wcs`. *rlock* is used to protect the *nreaders* variable that counts the number of readers in the critical section.

The invariants that imply the reader/writer specification (which are again easy to verify) are as follows:

- n readers at `@rcs` $\Rightarrow$ $nreaders \geq n$,
- $\exists$ writer at `@wcs` $\Rightarrow$ $nreaders = 0$,

A writer simply acquires *rwlock* to enter the critical section and releases it to exit. The *first* reader to enter the critical section acquires *rwlock* and the *last* reader to exit the critical section releases *rwlock*. The implementation satisfies progress: if no process is in the critical section than any process enter.

It is instructive to see what happens when a writer is in the critical section and two readers try to enter. The first reader successfully acquires *rlock* and but blocks when trying to acquire *rwlock*, which is held by the writer. The second reader blocks trying acquire *rlock* because it is held by the first reader. When the writer leaves the critical section, the first reader acquires *rwlock*, sets *nreaders* to 1, and releases *rlock*. *rwlock* is still held. Then the second reader acquires *rlock* and, assuming the first reader is still in the critical section, increments *nreaders* to 2 and enters the critical section *without* acquiring *rwlock*. *rwlock* is essentially jointly held by both readers. It does not matter in which order they leave: the second will release *rwlock*.

While CXL does not detect any issues, we must remember what it is checking for: it checks to make sure that there are never a reader and a writer in the critical section, and there are never two readers in the critical section. Moreover, it checks progress: all executions eventually terminate. What it does *not* check is if the code allows more than one reader in the critical section. Indeed, we could have implemented the reader/writer lock as in Figure 11.2 using an ordinary lock, never allowing more than one process in the critical section.

```
import synch;

def acquire_rlock():
    lock(&rlock);
    if nreaders == 0:
        lock(&rwlock);
    ;
    nreaders += 1;
    unlock(&rlock);
;
def release_rlock():
    lock(&rlock);
    nreaders -= 1;
    if nreaders == 0:
        unlock(&rwlock);
    ;
    unlock(&rlock);
;
def acquire_wlock():
    lock(&rwlock);
;
def release_wlock():
    unlock(&rwlock);
;
rwlock = Lock();
rlock = Lock();
nreaders = 0;
```

Figure 11.1: [RWlock.cxl] Reader/Writer with Two Locks.

```
import RWlock;

const NREADERS = 3;
const NWRITERS = 2;

def reader(self):
    acquire_rlock();
    @rcs: assert atLabel.wcs == dict{};
    flags[self] = True;
    if choose({ False, True }):
        let blocked = True:
            while blocked:
                if { flags[i] for i in 1..NREADERS } == { True }:
                    blocked = False;
                ;
            ;
        ;
    ;
    release_rlock();
;
def writer():
    acquire_wlock();
    @wcs: assert (atLabel.wcs == dict{ nametag(): 1 }) and
                (atLabel.rcs == dict{})
        ;
    release_wlock();
;
rwlock = Lock();
nreaders = 0;
flags = dict{ False for i in 1..NREADERS };
for i in 1..NREADERS:
    spawn reader(i);
;
for i in 1..NWRITERS:
    spawn writer();
;
```

Figure 11.2: [RWmulti.cxl] Checking that multiple readers can acquire the read lock.

Figure 11.2 contains a new test that ensures that indeed more than one reader can enter the critical section at a time. It does so by having processes sometimes wait in the critical section until all other readers are there. If other readers cannot enter, then that process enters in an infinite loop that CXL will readily detect. Try it out!

An important lesson here is that one should not celebrate too early if CXL does not find any issues. While a test may be successful, a test may not explore all desirable executions. While CXL can help explore all cornercases in a test, it cannot find problems that the test is not looking for.

Exercises

11.1 Why do you suppose the code in Figure 11.2 uses a flag per process rather than a simple counter? Can you write a version of this test that uses a counter instead of a flag per process?

11.2 Write a test program for "bound locks" (as implemented in the exercises of the last chapter) that checks that up to `M` processes can acquire the bound lock at the same time.

11.3 It can be useful for a process to *downgrade* its write lock to a read lock when its done writing but not done reading. Implement `downgrade`. Note, you cannot do this simply by releasing the write lock and then obtaining the read lock.

Chapter 12

Semaphores

So far we have looked at how to protect a single resource. Some types of resources may have multiple units. If there are only n units, no more than n units can be used at a time; if all are in use and a process comes along that needs one of the units, it has to wait until another process releases one of the units. Note that we cannot solve this problem simply using a lock per unit—allocating a unit requires access to the entire collection.

Consider the following synchronization problem that we call *Italian Philosophers* (a simplification of the famous Dining Philosophers problem described in Chapter 17). There are five philosophers sitting around a table, each with a plate of spaghetti in front of them. There is a glass sitting in the middle of the table with only three forks in it. A philosopher needs one of those forks to eat, but clearly, no more than three philosophers can eat at a time. But a lock can only count to one.

Introduced by Dijkstra, a *semaphore* is a synchronization primitive that can solve this kind of problem [Dij62]. A semaphore is essentially a counter that can be incremented and decremented but is not allowed to go below zero. The semaphore counter is typically initialized to the number of units of a resource initially available. When allocating a resource, a process decrements the counter using the `P` operation. You can think of `P` standing for *Procure*, as in procuring the resource associated with the semaphore. The `P` operation blocks the invoking process if the counter is zero. To release the resource, a process increments the resource using the `V` operation. You can think of `V` standing for *vacate*, as in vacating the resource. One can think of a semaphore as a generalization of a lock. In fact, a lock can be implemented by a semaphore initialized to 1. Then `P` procures the lock, and `V` vacates the lock.

Figure 12.1 shows the `synch` module implementation of semaphores. It represents a semaphore as an integer. Figure 12.2 shows a solution to the Italian Philosphers problem using a semaphore.

You may find the free "The Little Book of Semaphores" by Allen Downey a great resource for working with semaphores [Dow09].

Exercices

12.1 Add an assertion that demonstrates that at most three Italian philosophers are eating at the same time?

```
def Semaphore(cnt):
    result = cnt;
;
def P(sema):
    let blocked = True:
        while blocked:
            atomic:
                if (^sema) > 0:
                    ^sema -= 1;
                    blocked = False;
                ;
            ;
        ;
    ;
;
def V(sema):
    atomic:
        ^sema += 1;
    ;
;
```

Figure 12.1: The `Semaphore` interface in the `synch` module.

```
import synch;

const NDINERS = 5;
const NFORKS = 3;

def diner(which):
    while choose({ False, True }):
        P(&forks);
        # dine
        V(&forks);
        # think
    ;
;
forks = Semaphore(NFORKS);
for i in 1..NDINERS:
    spawn diner(i);
;
```

Figure 12.2: [DinersSema.cxl] Italian Philosophers.

12.2 Implement a test program that checks to see that the solution allows more than one Italian philosopher to eat at the same time?

12.3 Figure 12.3 shows an iterative implementation of the Qsort algorithm, and Figure 12.4 an accompanying test program. The array to be sorted is stored in shared variable a. Another shared variable, *todo*, contains the ranges of the array that need to be sorted (initially the entire array). Re-using as much of this code as you can, implement a parallel version of this. You should not have to change the methods `swap`, `partition`, or `sortrange` for this. Create `NWORKERS` "worker processes" that should replace the `qsort` code. Each worker loops until *todo* is empty and sorts the ranges that it finds until then. The `main` process needs to wait until all workers are done. Use a semaphore for this: the `main` process should use `P` to wait for each worker to terminate. After that it should check that the array is sorted and that it is a permutation of the original array.

```
def swap(p, q):                # swap p̂ and q̂
    let t = ^p:
        ^p = ^q;
        ^q = t;
    ;
;
def partition(lo, hi):
    result = lo;
    for i in lo..(hi − 1):
        if a[i] ¡= a[hi]:
            swap(&a[result], &a[i]);
            result += 1;
        ;
    ;
    swap(&a[result], &a[hi]);
;
def sortrange(range):
    let lo, hi = range, pivot = partition(lo, hi):
        if (pivot − 1) > lo:
            todo += { (lo, pivot − 1) };
        ;
        if (pivot + 1) < hi:
            todo += { (pivot + 1, hi) };
        ;
    ;
;
def qsort():
    while todo != {}:
        let range = choose(todo):
            todo −= { range };
            sortrange(range);
        ;
    ;
;
```

Figure 12.3: [qsort.cxl] Iterative qsort() implementation.

```
import list;
import qsort;

const NITEMS = 4;

def sorted():                    # return whether array 'a' is sorted
    result = True;
    for i in 1..(len(a) - 1):
        if a[i - 1] > a[i]:
            result = False;
        ;
    ;
;
def main(copy):
    qsort();                         # sort array 'a'
    assert sorted();                     # is the array sorted?:
    assert list2bag(a) == list2bag(copy); # is it a permutation?
;

const values = 1..NITEMS;
a = [ choose(values) for i in 1..choose(1..NITEMS) ];
todo = { (0, len(a) - 1) };
spawn main(a);
```

Figure 12.4: [qsorttest.cxl] Test program for Figure 12.3.

Chapter 13

Bounded Buffer

A good example of a resource with multiple units is the so-called *bounded buffer* (aka *ring buffer*). Suppose you have a *producer* (some process) wanting to communicate a stream of items with a *consumer* (another process). A bounded buffer is essentially a queue implemented using a circular buffer of a certain length and two pointers: one where the producer inserts items, and one where the consumer extracts items. If the buffer is full, the producer must wait; if the buffer is empty, the consumer must wait. This problem is known as the "Producer/Consumer Problem" and was proposed by Dijkstra [Dij72].

The producer/consumer pattern is common. Processes may be arranged in *pipelines*, where each upstream process is a producer and each downstream process is a consumer. Or processes may be arranged in a manager/worker pattern, with a manager producing jobs and workers executing them in parallel. Or, in the client/server model, some process may act as a *server* that clients can send requests to and receive responses from.

In general, we allow multiple producers and multiple consumers all sharing the same bounded buffer. Figure 13.1 gives a high-level description. The two numbers in a state specify the number of items that the producers still can produce and the number of items that the consumers have consumed. In this case there are three items to produce initially.

Figure 13.2 presents the implementation of a bounded buffer using semaphores. It features a circular bounded buffer *buf* with slots numbered 1 through `NSLOTS`. There are two indexes into the buffer: *b_in* specifies where the next produced item is inserted, while *b_out* specifies from where the next can be consumed. As there may be multiple producers and multiple consumers, updates to the indexes, which are shared variables, must be protected. To this end, we use two semaphores as locks: *l_in* for *b_in* and *l_out* for *b_out*.

Without additional synchronization, the indexes may overflow and point to invalid entries in the circular buffer. For this, there is a semaphore *n_full* that keeps track of the number of filled entries in the buffer and a semaphore *n_empty* that keeps track of the number of empty entries in the buffer.

To add an item to the bounded buffer (`produce`(*item*)), the producing process first has to wait until there is room in the bounded buffer. To this end, it invokes `P`(*n_empty*), which waits until *n_empty* > 0 and atomically decrements *n_empty* once this is the case. In other words, the producer procures an empty slot. Next, it adds the item to the next position in the buffer. Since there may

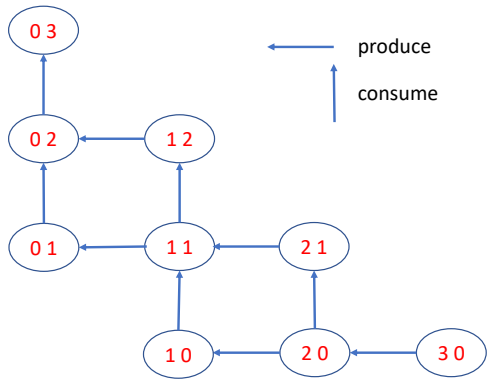

Figure 13.1: High-level state diagram specification of producers/consumers.

```
import synch;

const NSLOTS = 2;     # size of bounded buffer

def produce(item):
    P(&n_empty);
    P(&l_in);
    buf[b_in] = item;
    b_in = (b_in % NSLOTS) + 1;
    V(&l_in);
    V(&n_full);
;
def consume():
    P(&n_full);
    P(&l_out);
    result = buf[b_out];
    b_out = (b_out % NSLOTS) + 1;
    V(&l_out);
    V(&n_empty);
;
buf = dict{ () for x in 1..NSLOTS };
b_in = 1;
b_out = 1;
l_in = Semaphore(1);
l_out = Semaphore(1);
n_full = Semaphore(0);
n_empty = Semaphore(NSLOTS);
```

Figure 13.2: [BBsema.cxl] Bounded Buffer implementation using semaphores.

```
import BBsema;

const NPRODS = 3;     # number of producers
const NCONSS = 3;     # number of consumers

for i in 1..NPRODS:
   spawn produce(i);
;
for i in 1..NCONSS:
   spawn consume();
;
```

Figure 13.3: [BBsematest.cxl] Test program for Figure 13.2.

```
cxl -b -c NSLOTS=0 -c NPRODS=1 -c NCONSS=1 BBsematest.cxl
cxl -b -c NSLOTS=1 -c NPRODS=0 -c NCONSS=1 BBsematest.cxl
cxl    -c NSLOTS=1 -c NPRODS=1 -c NCONSS=0 BBsematest.cxl
cxl    -c NSLOTS=1 -c NPRODS=1 -c NCONSS=1 BBsematest.cxl
cxl -b -c NSLOTS=1 -c NPRODS=1 -c NCONSS=2 BBsematest.cxl
cxl -b -c NSLOTS=1 -c NPRODS=2 -c NCONSS=0 BBsematest.cxl
cxl    -c NSLOTS=1 -c NPRODS=2 -c NCONSS=1 BBsematest.cxl
cxl    -c NSLOTS=1 -c NPRODS=2 -c NCONSS=2 BBsematest.cxl
cxl -b -c NSLOTS=1 -c NPRODS=2 -c NCONSS=3 BBsematest.cxl
cxl    -c NSLOTS=2 -c NPRODS=1 -c NCONSS=0 BBsematest.cxl
cxl    -c NSLOTS=2 -c NPRODS=1 -c NCONSS=1 BBsematest.cxl
cxl -b -c NSLOTS=2 -c NPRODS=1 -c NCONSS=2 BBsematest.cxl
cxl    -c NSLOTS=2 -c NPRODS=2 -c NCONSS=0 BBsematest.cxl
cxl    -c NSLOTS=2 -c NPRODS=2 -c NCONSS=1 BBsematest.cxl
cxl    -c NSLOTS=2 -c NPRODS=2 -c NCONSS=2 BBsematest.cxl
cxl -b -c NSLOTS=2 -c NPRODS=2 -c NCONSS=3 BBsematest.cxl
cxl -b -c NSLOTS=2 -c NPRODS=3 -c NCONSS=0 BBsematest.cxl
cxl    -c NSLOTS=2 -c NPRODS=3 -c NCONSS=1 BBsematest.cxl
cxl    -c NSLOTS=2 -c NPRODS=3 -c NCONSS=2 BBsematest.cxl
cxl    -c NSLOTS=2 -c NPRODS=3 -c NCONSS=3 BBsematest.cxl
```

Figure 13.4: Script for testing producer/consumer synchronization (assuming the code is in file `BBsematest.cxl`).

```
import BBsema;

const NPRODS = 2;        # number of producers
const NCONSS = NPRODS;   # number of consumers

def producer(i):
    produce((i, 0));
    produce((i, 1));
;
def consumer():
    let first = consume():
        let second = consume():
            assert (first[0] != second[0]) or (first[1] < second[1]);
            atomic:
                received += { first, second };
            ;
        ;
    ;
;
def main():
    while cardinality(received) < (2 * NCONSS):
        pass;
    ;
    assert received == ({ (i, 0) for i in 1..NPRODS } +
                        { (i, 1) for i in 1..NPRODS })
    ;
;
received = {};
for i in 1..NPRODS:
    spawn producer(i);
;
for i in 1..NCONSS:
    spawn consumer();
;
spawn main();
```

Figure 13.5: [BBsemadata.cxl] Test whether the bounded buffer delivers the right data in the right order.

be more than one empty slot, multiple producers may attempt to do the same thing concurrently. Hence the use of the *l_in* semaphore used as a lock. Finally, once the item is added, the process increments the *n_full* semaphore to signal to consumers that an item is available. The consumer code is symmetric.

Note the quite different usage of the *l_* semaphores and *n_* semaphores. People often call semaphores that are used as locks *binary semaphores*, as they only take on the values 0 and 1. They protect critical sections. The second type of semaphores are called *counting semaphores*. They can be used to send signals between processes. However, binary and counting semaphores are implemented the same way.

Figure 13.3 invokes the code of Figure 13.2. We can run this through CXL, but what would it test? There are no assert statements in the code. We can split the functionality of this code into two pieces. One is the synchronization part: producers should never run more than `NSLOTS` ahead of the consumers, but they should be able to produce up to `NSLOTS` items even if there are no consumers. The other part is the data: we want to make sure that the items are sent through the buffer in order and without loss. We will first focus on the synchronization.

The first part can be tested with the code given in Figure 13.1, by experimenting with different non-negative values for the three constants. In particular, any run with a positive `NSLOTS` and `NCONSS` $\leq$ `NPRODS` $\leq$ `NCONSS` + `NSLOTS` should terminate, while other values should lead to processes blocking. To this this, one can write a script like the one in Figure 13.4.

Figure 13.5 tests whether the correct data is sent in the correct order. There are the same number of producers and consumers. Each producer i produces two tuples: $(i, 0)$ and $(i, 1)$ in that order. Each consumer consumes two tuples and checks that the tuples are different and that, if the two tuples come from the same producer, then they have to be in the order sent. Moreover, consumers add the tuples they received to the set *received* (atomically—a lock could have been used here as well).) The `main()` process waits for all consumers to have received exactly the set of tuples that the producers produce.

Exercises

13.1 Change the code in Figure 13.2 so that `produce()` fails if the buffer is full rather than block. `produce()` should return `False` if it fails and return `True` if it succeeds.

13.2 Implement a *bounded stack*. Method `push`(*item*) should block if the stack is full. Method `pop()` should block until the stack is non-empty and then return an item.

13.3 Implement a thread-safe GPU allocator by modifying Figure 13.6. There are `N` GPUs identified by the numbers 1 through `N`. Method `gpuAlloc()` returns the identifier of an available GPU, blocking if there is currently no available GPU. Method `gpuRelease`(*gpu*) releases the given GPU. It never needs to block.

```
const N = 10;

def gpuAlloc():
    result = choose(availGPUs);
    availGPUs -= { result };
;
def gpuRelease(gpu):
    availGPUs += { gpu };
;
availGPUs = 1..N;
```

Figure 13.6: [gpu.cxl] A thread-unsafe GPU allocator.

Chapter 14

Split Binary Semaphores

Reader/writer locks and bounded buffers are both examples of what are sometimes called *conditional critical sections* (CCSs) [Hoa73]. All critical sections are, in a way, conditional. Even the basic critical section has the condition that a process can only enter if no other process is in the critical section. Reader/writer locks refined that condition, loosening the restriction for improved performance. In the case of bounded buffers, there were two CCSs. First, a CCS to add an entry to the bounded buffer that can only be entered by one process and only if there is room in the bounded buffer (`produce`). Second, a CCS that can only be entered by one process and only if there is an item in the bounded buffer (`consume`).

CCSs are easy to implement using busy waiting:

```
let blocked = True:
   while blocked:
      lock();
      if application–condition–holds:
         blocked = False;
      else:
         unlock():
@cr: application–code;
unlock():
```

We have seen solutions to the reader/writer problem and the producer/consumer problem that do not busy-wait, but they are quite different. It would be nice to have a general technique for all CCS problems. The "Split Binary Semaphore" (SBS) is such a technique, originally proposed by Tony Hoare and popularized by Edsger Dijkstra [Dij79]. SBS is an example of what one could call a "baton-passing" technique. A process needs the baton to access shared variables. When a process has the baton and does not need it any more, it first checks to see if there are any processes that are waiting for the baton and can make progress. If so, it passes the baton directly to one of those processes. If not, it simply gives up the baton.

```
import synch;

def V_one():
    if (w_entered == 0) and (r_waiting > 0):
        V(&r_sema);
    elif ((r_entered + w_entered) == 0) and (w_waiting > 0):
        V(&w_sema);
    else:
        V(&mutex);
    ;
;
def acquire_rlock():
    P(&mutex);
    if w_entered > 0:
        r_waiting += 1;
        V(&mutex); P(&r_sema);
        r_waiting -= 1;
    ;
    r_entered += 1;
    V_one();
;
def release_rlock():
    P(&mutex);
    r_entered -= 1;
    V_one();
;
def acquire_wlock():
    P(&mutex);
    if (r_entered + w_entered) > 0:
        w_waiting += 1;
        V(&mutex); P(&w_sema);
        w_waiting -= 1;
    ;
    w_entered += 1;
    V_one();
;
def release_wlock():
    P(&mutex);
    w_entered -= 1;
    V_one();
;
mutex = Semaphore(1); r_sema = Semaphore(0); w_sema = Semaphore(0);
r_entered = 0; r_waiting = 0; w_entered = 0; w_waiting = 0;
```

Figure 14.1: [RWsbs.cxl] Reader/Writer Lock implementation using Split Binary Semaphores.

We will illustrate the technique using the reader/writer problem. Figure 14.1 shows the `rlock_acquire`, `rlock_release`, `wlock_acquire`, and `wlock_release` methods implemented using SBS.

Step 1 is to enumerate all waiting conditions. In the case of the reader/writer problem, there are two: a process that wants to read may have to wait for a writer to leave the critical region, while a process that wants to write may have to wait until all readers have left the critical section. If there are N waiting conditions, the SBS technique will use $N + 1$ binary semaphores: one for each waiting condition and an extra one that processes use when they first try to enter the critical section and do not yet know if they need to wait. We shall call this the *mutex*.

An important invariant of the SBS technique is that the sum of the $N + 1$ semaphores must always be 0 or 1, and it should be 0 when a process is accessing the shared variables and 1 when not. As it were, these semaphores taken together are emulating a single binary semaphore. Initially, *mutex* is 1 and the other N semaphores are all 0. To maintain the invariant, each process alternates `P` and `V` operations, starting with `P`(&*mutex*) when it tries to enter the CCS, and ending with a `V` operation on one of the semaphores (yet to be determined).

Another important part of the SBS technique is to keep careful track of the state. In the case of the reader/writer lock, the state consists of the following shared variables:

- *r_entered*: the number of readers in the critical section;
- *w_entered*: the number of writers in the critical section;
- *r_waiting*: the number of readers waiting to enter;
- *w_waiting*: the number of writers waiting to enter.

Each of the `rlock_acquire`, `rlock_release`, `wlock_acquire`, and `wlock_release` methods must maintain this state. To wit, check out `rlock_acquire`. It first checks to see if there are any writers. If so, it increments *r_waiting*. If not, it increments *r_entered*.

Either way, it vacates one of the semaphores next. `V_one()` is the baton passing function that selects which of the binary semaphores to vacate. Method `V_one()` first checks to see if there are any readers or writers waiting. If there are readers waiting and there are no writers in the critical section, it vacates the semaphore associated with readers waiting. If there are writers waiing and there are no readers nor writers in the critical section, it vacates the semaphore associated with writers waiting. If both conditions hold, it could vacate either one—it would not matter. However, it can only vacate one of the semaphores. If none of the conditions hold, `V_one()` vacates *mutex*.

Let us return now to `rlock_acquire`. In the case that the process, say p, found there is a writer in the critical section, it vacates *mutex* and procures *r_sema*, blocking on the semaphore associated with waiting readers. When later some other process passes the baton to p by incrementing *r_sema*, p updates the state by decrementing *r_waiting* and incrementing *r_entered*. p then invokes `V_one()` one more time.

As an example, consider the case where there is a writer in the critical section and there are two readers waiting. Let us see what happens when the writer calls `wlock_release`:

1. After procuring *mutex*, the sum of the three semaphores is 0. The writer then decrements *w_entered* and calls `V_one()`.
2. `V_one()` finds that there are no writers in the critical section and there are two readers waiting. It therefore vacates *r_sema*. The sum of the semaphores is now 1, because *r_sema* is 1. This guarantees that only a waiting reader, procuring *r_sema*, can enter the critical section.

3. When it does, the reader decrements *r_waiting* from 2 to 1, and increments *r_entered* from 0 to 1. The reader finally calls `V_one()`.

4. Again, `V_one()` finds that there are no writers and that there are readers waiting, so again it vacates *r_sema*, passing the baton to the one remaining waiting reader.

5. The remaining reader decrements *r_waiting* from 1 to 0 and increments *r_entered* from 1 to 2.

6. Finally, the remaining reader calls `V_one()`. `V_one()` does not find any processes waiting, and so it vacates *mutex*.

Exercises

14.1 Assume that you have only binary semaphores. Implement counting semaphores using binary semaphores using split binary semaphores.

14.2 Implement a solution to the producer/consumer problem (Figure 13.2) using split binary semaphores.

14.3 Cornell's campus features some one-lane bridges. Cars can only go in one direction at a time. Consider northbound and southbound cars wanting to cross a one-lane bridge. The bridge allows arbitrary many cars, as long as they're going in the same direction. Implement a lock that observes this requirement using SBS. Write methods `nb_enter()` that a car must invoke before going northbound on the bridge and `nb_leave()` that the car must invoke after leaving the bridge. Similarly write `sb_enter()` and `sb_leave()` for southbound cars.

Chapter 15

Starvation

A *property* is a set of traces. If a program has a certain property, that means that the traces that that program allows are a subset of the traces in the property. So far we have pursued two properties: *mutual exclusion* and *progress*. The former is an example of a *safety property*—it prevents something "bad" from happening, like a reader and writer process both entering the critical section. The *progress* property is an example of a *liveness property*—guaranteeing that something good eventually happens. Informally (and inexactly), progress states that if no processes are in the critical section, then some process that wants to enter can.

Progress is a weak form of liveness. It says that *some* process can enter, but it does not prevent a scenario such as the following. There are three processes repeatedly trying to enter a critical section using a spinlock. Two of the processes successfully keep entering, alternating, but the third process never gets a turn. This is an example of `starvation`. With a spinlock, this scenario could even happen with two processes. Initially both processes try to acquire the spinlock. One of the processes is successful and enters. After the process leaves, it immediately tries to re-enter. This state is identical to the initial state, and there is nothing that prevents the same process from acquiring the lock yet again.

It is worth noting that Peterson's Algorithm (Chapter 6) does not suffer from starvation, thanks to the `turn` variable that alternates between 0 and 1 when two processes are contending for the critical section.

While spinlocks suffer from starvation, it is a uniform random process and each process has an equal chance of entering the critical section. Thus the probability of starvation is exponentially vanishing. We shall call such a solution *fair* (although it does not quite match the usual formal nor vernacular concepts of fairness).

Unfortunately, such is not the case for the reader/writer solutions that we presented thus far. Consider this scenario: there are two readers and one writer. One reader is in the critical section while the writer is waiting. Now the second reader tries to enter and is able to. The first reader leaves. We are now in a similar situation as the initial state with one reader in the critical section and the writer waiting, but it is not the same reader. Unfortunately for the writer, this scenario

```
import synch;

def V_one_r():  # prefers reader next over writer next
    if (w_entered == 0) and (w_waiting == 0) and (r_waiting > 0):
        V(&r_sema);
    elif ((r_entered + w_entered) == 0) and (w_waiting > 0):
        V(&w_sema);
    else:
        V(&mutex);
    ;
;
def V_one_w():  # prefers writer next over reader next
    if ((r_entered + w_entered) == 0) and (w_waiting > 0):
        V(&w_sema);
    elif (w_entered == 0) and (w_waiting == 0) and (r_waiting > 0):
        V(&r_sema);
    else:
        V(&mutex);
    ;
;
def acquire_rlock():
    P(&mutex);
    if (w_entered > 0) or (w_waiting > 0):
        r_waiting += 1; V(&mutex);
        P(&r_sema); r_waiting -= 1;
    ;
    r_entered += 1;
    V_one_r();     # only other readers can enter
;
def release_rlock():
    P(&mutex); r_entered -= 1;
    V_one_w();     # other writers have right of way
;
def acquire_wlock():
    P(&mutex);
    if (r_entered + w_entered) > 0:
        w_waiting += 1; V(&mutex);
        P(&w_sema); w_waiting -= 1;
    ;
    w_entered = 1;
    V(&mutex);    # no other process can enter
;
def release_wlock():
    P(&mutex); w_entered = 0;
    V_one_r();     # other readers have right of way
;
mutex = Semaphore(1); r_sema = Semaphore(0); w_sema = Semaphore(0);
r_entered = 0; r_waiting = 0; w_entered = 0; w_waiting = 0;
```

Figure 15.1: [RWfair.cxl] Reader/Writer Lock SBS implementation addressing fairness.

can repeat itself indefinitely. So even if neither reader was in the critical section all of the time, and the second reader arrived well after the writer, the writer never had a chance.

SBSs allow much control over which thread runs next and is therefore a good starting point for developing fair synchronization algorithm In this chapter, we will present two SBS-based fair reader/writer lock implementations. The first, in Figure 15.1, is based on Figure 14.1. There are two differences:

1. When a reader tries to enter the critical section, it yields not only if there are writers in the critical section, but also if there are writers waiting to enter the critical section;

2. There are two versions of `V_one()`. In particular, when the last reader leaves the critical section, it preferentially passes the baton to a waiting writer rather than to a waiting reader. Similarly, when a writer leaves the critical section, it preferentially passes the baton to a waiting reader rather than to a waiting writer.

The net effect of this is that if there is contention between readers and writers, than readers and writers end up alternating entering the critical section. While readers can still starve other readers and writers can still starve other writers, they cannot starve one another. And since contention for a semaphore is resolved probabilistically, each reader and each writer will eventually have a chance to enter the critical section almost surely.

Figures 15.2 and 15.3 present a different approach in which processes are allowed into the critical region in a purely First Come First Served manner. When a reader arrives, it is allowed in only if there is no writer in the critical section or waiting to enter. When a writer arrives, it is allowed in only if there is no reader or writer in the critical section (and therefore also not waiting to enter). To maintain the arrival order, the solution uses a semaphore per process and a queue. The `acquire_` methods take a pointer to the semaphore of the process as argument. Each entry in the queue is a pair consisting of a type of access (`.read` or `.write`) and a pointer to the semaphore of the processs that is waiting. Method `V_one` considers semaphores to vacate in the order in which they appear in the queue.

Exercises

15.1 Write a fair solution to the one-lane bridge problem of Exercise 14.3. If you want to use the queue method, you can change the signature of the methods to enter the bridge, adding an argument that contains a pointer to a semaphore.

```
import synch;
import list;

def V_one():
    if queue == []:
        V(&mutex);
    else:
        let h = head(queue):
            if (h[0] == .read) and (w_entered == 0):
                V(h[1]);
            elif (h[0] == .write) and ((r_entered + w_entered) == 0):
                V(h[1]);
            else:
                V(&mutex);
            ;
        ;
    ;
;
def acquire_rlock(psema):
    P(&mutex);
    if (w_entered > 0) or (queue != []):
        queue = append(queue, (.read, psema)); V(&mutex);
        P(psema); queue = tail(queue);
    ;
    r_entered += 1;
    V_one();
;
def release_rlock():
    P(&mutex); r_entered −= 1; V_one();
;
def acquire_wlock(psema):
    P(&mutex);
    if (r_entered + w_entered) > 0:
        queue = append(queue, (.write, psema)); V(&mutex);
        P(psema); queue = tail(queue);
    ;
    w_entered = 1;
    V_one();
;
def release_wlock():
    P(&mutex); w_entered = 0; V_one();
;
mutex = Semaphore(1);
r_entered = 0; w_entered = 0;
queue = [];
```

Figure 15.2: [RWqueue.cxl] Reader/Writer Lock SBS implementation using a queue.

```
import RWqueue;

const NPROCS = 3;

def process(self):
    while choose({ False, True }):
        if choose({ .read, .write }) == .read:
            acquire_rlock(&semas[self]);
            @rcs: assert atLabel.wcs == dict{};
            release_rlock();
        else:                   # .write
            acquire_wlock(&semas[self]);
            @wcs: assert (atLabel.wcs == dict{ nametag(): 1 }) and
                        (atLabel.rcs == dict{})
                    ;
            release_wlock();
        ;
    ;
;
semas = [ Semaphore(0) for i in 0..(NPROCS–1) ];
for i in 0..(NPROCS–1):
    spawn process(i);
;
```

Figure 15.3: [RWqtest.cxl] Test program for Figure 15.2.

Chapter 16

Monitors

Tony Hoare, who came up with the concept of split binary semaphores, devised an abstraction of the concept in a programming language paradigm called *monitors* [Hoa74]. (A similar construct was independently invented by Per Brinch Hansen [BH73].) A monitor is a special version of an object-oriented *class*, comprising a set of variables and methods that operate on those variables. There is a split binary semaphore associated with each such class. The *mutex* is hidden: it is automatically acquired when invoking a method and released upon exit. The other semaphores are called *condition variables*. There are two operations on condition variables: `wait` and `signal`. If c is a pointer to a condition variable, then the semantics of the operations, in CXL, are as follows:

```
def wait(c):
    V(&mutex);
    P(c);
;
def signal(c):
    V(c);
    P(&mutex);
;
```

`wait` releases the mutex and blocks while trying to procure the condition variable (which is a semaphore that is 0 when `wait` is invoked). `signal` vacates the condition variable, passing the baton to a process trying to procure it, and then procures the mutex. The use of `wait` and `signal` ensures that `P` and `V` operations alternate as prescribed by the SBS technique. Our implementations of reader/writer locks can be easily changed to use `wait` and `signal` instead of using `P` and `V`.

In the late 70s, Xerox PARC developed a new programming language called Mesa [LR80]. Mesa introduced various important concepts to programming languages, including software exceptions and incremental compilation. Mesa also incorporated a version of monitors. However, there are some subtle but important differences with Hoare monitors that make Mesa monitors quite unlike split binary semaphores.

As in Hoare monitors, there is a hidden mutex associated with each monitor, and the mutex is automatically acquired upon entry to a method and released upon exit. Mesa monitors also have

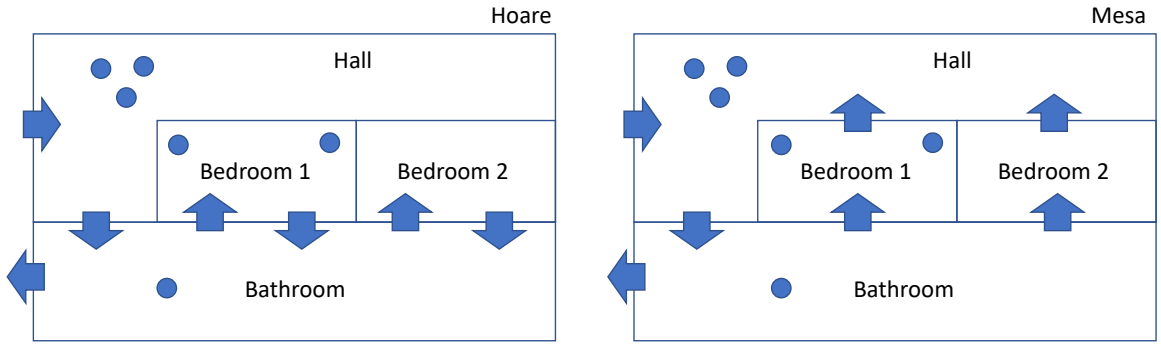

Figure 16.1: High-level depictions of Hoare and Mesa monitors. The hall is where processes wait to enter the bathroom (the critical section). The bedrooms illustrate two condition variables. The circles are processes.

condition variable that a process can wait on. Like in Hoare monitors, the `wait` operation releases the mutex. The most important difference is in what `signal` does. To make the distinction more clear, we shall call the corresponding Mesa operation `notify` rather than `signal`. When a process p invokes `notify`, it does not immediately pass the baton to some process q requiring the baton (assuming there even is such a process). Instead, p holds onto the baton until it leaves the monitor. At that point, any process that was notified will have a chance to obtain the baton.

Figure 16.1 illustrates the difference with a drawing. Here a bathroom represents the critical section, allowing only one process at a time. The bedrooms represent condition variables. There is some condition associated with each bedroom. In the hall are processes waiting to enter the critical section.

On the left is a Hoare monitor. When a process leaves the monitor (let's call that process p), it must be in the bathroom. It can choose to either let in a process in one of the bedrooms or a process in the hall, assuming there are any. To make the choice, p checks for each bedroom if its condition holds and if there are processes in the bedroom. If there are such bedrooms, p selects one and one of its processes into the bathroom. Let's call that process q. Importantly, when q enters the bathroom, the condition that q was waiting for still holds.

On the right is a Mesa monitor. The only way into the bathroom is through the hall. When process p leaves the bathroom, if also checks each of the bedrooms to see if there is a process q that can run. But instead of letting q go straight into the bathroom, q goes into the hall, joining any other processes that are also waiting to enter the bathroom. Finally, when p has left the bathroom, one of the processes in the hall is let into the bathroom. But that may not be q. Assuming every process eventually leaves the bathroom and the system is fair, eventually q will enter the bathroom. However, it is no longer guaranteed that its condition holds. Therefore, the first thing q must do is check, again, to see if the condition holds. If not, it should go back into the bedroom corresponding to the condition.

Mesa monitors provide another important option: when process p leaves the bathroom, it can choose to let any number of processes in the bedrooms into the hall. It can do so by calling `notify` repeatedly, each time letting one process in one of the bedrooms into the hall. Using `notify` on an empty bedroom is considered a no-op in Mesa. (Hoare monitors do not allow using `signal` on

an empty bathroom.) Mesa also provides an operation called `notifyAll` (aka `broadcast`)) that, when applied to a bedroom, lets all processes in that bedroom into the hall. This is not possible with Hoare monitors: when needing to wake up a set of processes, only one can be signaled at a time. Each of those processes then has to signal the next until all processes have been signaled.

While Hoare monitors are conceptually simpler, the so-called "Mesa monitor semantics" or "Mesa condition variable" semantics have become more popular, adopted by all major programming languages. That said, few programming languages provide full support for monitors. In Java, each object has a hidden lock *and* a hidden condition variable associated with it. Methods declared with the `synchronized` keyword automatically obtain the lock. Java objects also support `wait`, `notify`, and `notifyAll`. Java also supports explicit allocations of locks and condition variables. In Python, locks and condition variables must be explicitly declared. The `with` statement makes it easy to acquire and release a lock for a section of code. In C and C++, support for locks and condition variables is through libraries.

The CXL `synch` module supports Mesa condition variables. However, like C, Java, and Python, it does not have built-in language support for them but instead associates condition variables with explicit locks. Figure 16.2 shows the implementation of condition variables in the `synch` module. `Condition`(*lk*) creates a new condition variable associated with the given pointer to a lock *lk*. The condition variable itself is represented by a dictionary containing the pointer to the lock and a bag of name tags of processes waiting on the condition variable. (The `synchS` library instead uses a list of contexts.)

`wait` adds the nametag of the process to the bag. This increments the number of processes in the bag with the same context. `wait` then waits until that count is restored to the value that it had upon entry to `wait`. `notify` removes an arbitrary context from the bag, allowing one of the processes with that context to return from `wait`. `notifyAll` empties out the entire bag, allowing all processes in the bag to resume.

To illustrate how Mesa condition variables are used in practice, we demonstrate using, once again, an implementation of reader/writer locks. Figure 16.3 shows the code. *rwlock* is the global lock or mutex for the reader/writer lock. There are two condition variables: readers wait on *rcond* and writers wait on *wcond*. The implementation also keeps track of the number of readers and writers in the critical section.

When inspecting the code, one important feature to notice is that `wait` is always invoked within a **while** loop that checks for the condition that the process is waiting for. We explained why above: when a process resumes after being notified, it is not guaranteed that the condition it was waiting for holds, even if it did at the time of invoking `notify` or `notifyAll`. It is *imperative* that there is always a **while** loop around any invocation of `wait` containing the negation of the condition that the process is waiting for. Many implementation of condition variables depend on this, and optimized implementations of condition variables often allow so-called "spurious wakeups," where `wait` may sometimes return even if the conditon variable has not been notified.

In `release_rlock`, notice that `notify`(&*wcond*) is invoked when there are no readers left, *without* checking if there are writers waiting to enter. With Mesa monitors this is ok, because calling `notify` is a no-op if no process is waiting.

`release_wlock` executes `notifyAll`(&*wcond*) as well as `notify`(&*wcond)*. Again, because we do not keep track of the number of waiting readers or writers, we have to conservatively assume that all waiting readers can enter, or, alternatively, up to one waiting writer can enter. So `release_wlock` wakes up all potential candidates. There are two things to note here. First, unlike split binary semaphores or Hoare monitors, where multiple waiting readers would have to be signaled one at a

```
import bag;

def Condition(lk):
    result = dict{ .lock: lk, .waiters: bagEmpty() };
;
def wait(c):
    let lk = (^c).lock, blocked = True,
            cnt = bagCount((^c).waiters, nametag()):
        bagAdd(&(^c).waiters, nametag());
        ^lk = False;
        while blocked:
            atomic:
                if (not (^lk)) and
                    (bagCount((^c).waiters, nametag()) ¡= cnt):
                    ^lk = True;
                    blocked = False;
                ;
            ;
        ;
    ;
;
def notify(c):
    let waiters = (^c).waiters:
        if waiters != bagEmpty():
            bagRemove(&(^c).waiters, bagChoose(waiters));
        ;
    ;
;
def notifyAll(c):
    (^c).waiters = bagEmpty();
;
```

Figure 16.2: Implementation of condition variables in the `synch` module.

```
import synch;

def acquire_rlock():
    lock(&rwlock);
    while nwriters > 0:
        wait(&rcond);
    ;
    nreaders += 1;
    unlock(&rwlock);
;
def release_rlock():
    lock(&rwlock);
    nreaders −= 1;
    if nreaders == 0:
        notify(&wcond);
    ;
    unlock(&rwlock);
;
def acquire_wlock():
    lock(&rwlock);
    while (nreaders + nwriters) > 0:
        wait(&wcond);
    ;
    nwriters = 1;
    unlock(&rwlock);
;
def release_wlock():
    lock(&rwlock);
    nwriters = 0;
    notifyAll(&rcond);
    notify(&wcond);
    unlock(&rwlock);
;
rwlock = Lock();
rcond = Condition(&rwlock);
wcond = Condition(&rwlock);
nreaders = 0;
nwriters = 0;
```

Figure 16.3: [RWcv.cxl] Reader/Writer Lock implementation using Mesa-style condition variables.

time in a baton-passing fashion (see Figure 14.1), with Mesa monitors all readers are awaked in one fell swoop using `notifyAll`. Second, both readers and writers are awakened—this is ok because both execute `wait` within a **while** loop, re-checking the condition that they are waiting for. So if both type of processes are waiting, either all the readers get to enter next or one of the writers gets to enter next.

Much of the complexity of programming with Mesa condition variables is in figuring out when to invoke `notify` and when to invoke `notifyAll`. As a rule of thumb: be conservative—it is better to wake up too many processes than too few. Waking up too many processes may lead to some inefficiency, but waking up to few may cause the application to get stuck. CXL can be particularly helpful here, as it examines every corner case. You can safely try to replace each `notifyAll` with `notify` and see if every possible execution of the the application still terminates.

Andrew Birrell's paper on Programming with Threads gives an excellent introduction to working with Mesa-style condition variables [Bir89].

Exercises

16.1 Implement semaphores using Mesa condition variables.

16.2 Write a new version of the GPU allocator in Exercise 13.3 using Mesa condition variables. In this version, a process is allowed to allocate a set of GPUs and release a set of GPUs that it has allocated. Method `gpuAllocSet(n)` should block until n GPUs are available, but it should grant them as soon as they are available. Method `gpuReleaseSet(s)` takes a set of GPU identifiers as argument. A process does not have to return all the GPUs it allocated at once.

16.3 The specification in the previous question makes the solution unfair. Explain why this is so. Then change the specification and the solution so that it is fair.

Chapter 17

Deadlock

When multiple processes are synchronizing access to shared resources, they may end up in a *deadlock* situation where one or more of the processes end up being blocked indefinitely because each is waiting for another to give up a resource. The famous Dutch computer scientist Edsger W. Dijkstra illustrated this using a scenario he called "Dining Philosophers."

Imagine five philosopers sitting around a table, each with a plate of food in front of them and a fork between each two plates. Each philosopher requires two forks to eat. To start eating, a philosopher first picks up the fork on the left, then the fork on the right. Each philospher likes to take breaks from eating to think for a while. To do so, the philosopher puts down both forks. Each philosopher repeat this procedure. Dijkstra had them repeating this for ever, but for the purposes of this book, philosphers can leave the table when they're not eating.

Figure 17.1 implements the dining philosophers in CXL, using a process for each philosopher and a lock for each fork. If you run it, CXL complains that the execution may not terminate, with all five processes being blocked trying to acquire the lock.

- Do you see what the problem is?
- Does it depend on `N`, the number of philosophers?
- Does it matter in what order the philosophers lay down their forks?

There are four conditions that must hold for deadlock to occur [CES71]:

1. *Mutual Exclusion*: each resource can only be used by one process at a time:
2. *Hold and Wait*: each process holds resources it already allocated while it waits for other resources that it needs;
3. *No Preemption*: Resources cannot be forcibly taken back from processes that allocated them;
4. *Circular Wait*: There exists a directed circular chain of processes, each waiting to allocate a resource held by the next.

Preventing deadlock thus means preventing that one of these conditions occurs. However, mutual exclusion is not easily prevented (although, for some resources it is possible, as demonstrated in Chapter 20). Havender proposed the following techniques that avoid the remaining three conditions [Hav68]:

```
import synch;

const N = 5;

def diner(which):
    let left = which, right = (which % N) + 1:
        while choose({ False, True }):
            lock(&forks[left]);
            lock(&forks[right]);
            # dine
            unlock(&forks[left]);
            unlock(&forks[right]);
            # think
        ;
    ;
;
forks = dict{ Lock() for i in 1..N };
for i in 1..N:
    spawn diner(i);
;
```

Figure 17.1: [Diners.cxl] Dining Philosophers.

```
import synch;

const N = 5;

def diner(which):
    let left = which, right = (which % N) + 1:
        while choose({ False, True }):
            lock(&mutex);
            while forks[left] or forks[right]:
                if forks[left]:
                    wait(&conds[left]);
                ;
                if forks[right]:
                    wait(&conds[right]);
                ;
            ;
            assert not (forks[left] or forks[right]);
            forks[left] = True;
            forks[right] = True;
            unlock(&mutex);

            # dine

            lock(&mutex);
            forks[left] = False;
            forks[right] = False;
            notify(&conds[left]);
            notify(&conds[right]);
            unlock(&mutex);

            # think
        ;
    ;
;
mutex = Lock();
forks = dict{ False for i in 1..N };
conds = dict{ Condition(&mutex) for i in 1..N };
for i in 1..N:
    spawn diner(i);
;
```

Figure 17.2: [DinersCV.cxl] Dining Philosophers that grab both forks at the same time.

- *No Hold and Wait*: a process must request all resources it is going to need at the same time;
- *Preemption*: if a process is denied a request for a resource, it must release all resources that it has already acquired and start over;
- *No Circular Wait*: define an ordering on all resources and allocate resources in a particular order.

To implement a *No Hold and Wait* solution, a philosopher would need a way to lock both the left and right forks at the same time. Locks do not have such an ability, and neither do semaphores. so we re-implement the Dining Philosophers using condition variables that allow one to wait for arbitrary application-specific conditions.

Figure 17.2 demonstrates how this might be done. We use a single mutex for the diners, and, for each fork, a boolean and a condition variable. The boolean indicates if the fork has been taken. Each diner waits if either the left or right fork is already taken. But which condition variable to wait on? The code demonstrates an important technique to use when waiting for multiple conditions. The condition in the **while** statement is the negation of the condition that the diner is waiting and consists of two disjuncts. Within the **while** statement, there is an **if** statement for each disjunct. The code waits for either or both forks if necessary. After that, it goes back to the top of the **while** loop.

A common mistake is to write the following code instead:

```
while forks[left]:
    wait(&conds[left]);
;
while forks[right]:
    wait(&conds[right]);
;
```

- Can you see why this does not work? What can go wrong?
- Run it through CXL in case you are not sure!

The *Preemption* approach suggested by Havender is to allow processes to back out. While this could be done, this invariably leads to a busy waiting solution where a process keeps obtaining locks and releasing them again until it finally is able to get all of them.

The *No Circular Waiting* approach is to prevent a cycle from forming, with each process waiting for the next process on the cycle. We can do this by establishing an ordering among the resources (in this case the forks) and, when needing more than one resource, always acquiring them in order. In the case of the philosopers, they could prevent deadlock by always picking up the lower numbered fork before the higher numbered fork, like so:

```
if left < right:
    lock(&forks[left]);
    lock(&forks[right]);
else:
    unlock(&forks[right]);
    unlock(&forks[left]);
;
```

This completes all the Havender methods. There is, however, another approach, which is sometimes called deadlock *avoidance* instead of deadlock *prevention*. In the case of the Dining Philosophers, we want to avoid the situation where each diner picks up a fork. So if we can prevent more than four diners from starting to eat at the same time, then we can avoid the conditions for deadlock from ever happening. Figure 17.3 demonstrates this concept. It uses a semaphore to restrict the number of diners at any time to four.

This concept can be generalized using something called the Banker's Algorithm [Dij64], but it is outside the scope of this book. The problem with these kinds of schemes is that one needs to know ahead of time the set of processes and what the maximum number of resources is that each process wants to allocate, making them generally quite impractical.

Exercises

17.1 Figure 17.4 shows an implementation of a bank with various accounts and transfers between those accounts. Unfortunately, running the test reveals that it sometimes leaves unterminated processes. Can you fix the problem?

17.2 Add a method `total()` to the solution of the previous question that computes the total over all balances. It needs to obtain a lock on all accounts. Make sure that it cannot cause deadlock.

17.3 Implement an extra process that checks, once, if the total of the balances is the same as that at the beginning (using an assertion). Is once enough to determine that no money gets lost or create, or should it do so continually?

```
import synch;

const N = 5;

def diner(which):
    let left = which, right = (which % N) + 1:
        while choose({ False, True }):
            P(&sema);
            lock(&forks[left]);
            lock(&forks[right]);
            # dine
            unlock(&forks[left]);
            unlock(&forks[right]);
            V(&sema);
            # think
        ;
    ;
;
forks = dict{ Lock() for i in 1..N };
sema = Semaphore(N - 1);
for i in 1..N:
    spawn diner(i);
;
```

Figure 17.3: [DinersAvoid.cxl] Dining Philosophers that carefully avoid getting into a deadlock scenario.

```
import synch;

const MAX = 2;
const NACCOUNTS = 2;
const NPROCESSES = 2;

def transfer(a1, a2, amount):
    lock(&accounts[a1].lock);
    if amount ¡= accounts[a1].balance:
        accounts[a1].balance −= amount;
        lock(&accounts[a2].lock);
        accounts[a2].balance += amount;
        unlock(&accounts[a2].lock);
        result = True;
    else:
        result = False;
    ;
    unlock(&accounts[a1].lock);
;
def process():
    let a1 = choose(0..(NACCOUNTS–1)),
                a2 = choose((0..(NACCOUNTS–1)) − { a1 }):
        transfer(a1, a2, choose(0..MAX));
    ;
;
accounts = [ dict{ .lock: Lock(), .balance: choose(0..MAX) }
            for i in 0..(NACCOUNTS–1) ]
;
for i in 1..NPROCESSES:
    spawn process();
;
```

Figure 17.4: [bank.cxl] Bank accounts.

Chapter 18

Actors and Message Passing

Some programming languages favor a different way of implementing synchronization using so-called *actors* [HBS73]. Actors are processes that have only private memory and communicate through *message passing*. See Figure 18.1 for an illustration. Given that there is no shared memory in the actor model (other than the message queues, which have built-in synchronization), there is no need for critical sections. Instead, some sequential process owns a particular piece of data and other processes access it by sending request messages to the process and optionally waiting for response messages. Each process handles one message at a time, serializing all access to the data it owns. As message queues are FCFS, starvation is prevented.

The actor synchronization model is popular in a variety of programming languages, including Erlang and Scala. Actor support is also available through popular libraries such as Akka, which is available for various programming languages. In Python, Java, and C/C++, actors can be easily emulated using threads and *synchronized queues* (aka *blocking queues*) for messaging. Each thread would have one such queue for receiving messages. Dequeuing from an empty synchronized queue blocks the thread until another thread enqueues a message on the queue.

The `synch` library supports a synchronized message queue, similar to the `Queue` object in Python. Its interface is as follows:

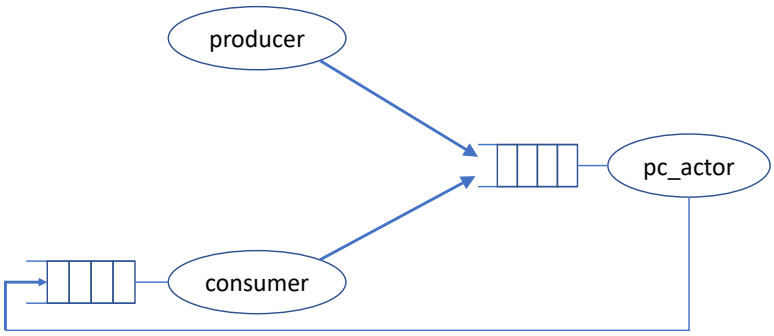

Figure 18.1: Depiction of three actors. The producer does not receive messages.

```
import synch;

def pc_actor(q, nrequests):
   let requests = {}, balance = 0:
      while nrequests > 0:
         let req = dequeue(q):
            if req.type == .produce:
               if balance ¿= 0:
                  requests += { req };
               else:
                  let r = choose(requests):
                     assert r.type == .consume;
                     enqueue(r.queue, req.item);
                     requests -= { r };
                  ;
               ;
               balance += 1;
            else:
               assert req.type == .consume;
               if balance ¡= 0:
                  requests += { req };
               else:
                  let r = choose(requests):
                     assert r.type == .produce;
                     enqueue(req.queue, r.item);
                     requests -= { r };
                  ;
               ;
               balance -= 1;
            ;
         ;
         nrequests -= 1;
      ;
   ;
;
def produce(q, item):
   enqueue(&pc_queue, dict{ .type: .produce, .item: item });
;
def consume(q1, q2):
   enqueue(q1, dict{ .type: .consume, .queue: q2 });
   result = dequeue(q2);
;
```

Figure 18.2: [actor.cxl] A producer/consumer actor.

```
import actor;

const NITEMS = 3;

pc_queue = Queue();
spawn pc_actor(&pc_queue, 2 * NITEMS);
queues = [ Queue() for i in 0..(NITEMS – 1) ];
for i in 0..(NITEMS–1):
    spawn produce(&pc_queue, i);
    spawn consume(&pc_queue, &queues[i]);
;
```

Figure 18.3: [actortest.cxl] Test code for producer/consumer actor.

- `Queue()` returns a new message queue;
- `enqueue`(q, *item*) adds *item* to the queue pointed to by q;
- `dequeue`(q) waits for and returns an item on the queue pointed to by q.

Note that a `Queue` behaves much like a zero-initialized semaphore. `enqueue` is much like `V`, except that it is accompanied by data. `dequeue` is much like `P`, except that it also returns data. Thus, synchronized queues can be considered a generalization of counting semaphores.

Figure 18.2 shows a CXL solution to the "multiple producer / multiple consumer problem" (with an unbounded buffer) using an actor (as illustrated in Figure 18.1. `pc_actor` is essentially a matchmaker: it matches `.produce` requests with `.consume` requests. Sometimes it may have buffered `.consume` requests and is waiting for `.produce` requests, and sometimes it is vice versa. The variable *balance* specifies what the situation is: if it is positive then it has buffered `.produce` requests; if it is negative then it has buffered `.consume` requests. Note that both *balance* and *requests* are variables that are local to the `pc_actor` process.

Each request message is a dictionary with two fields. One of the fields is `.type`, which either is `.produce` or `.consume`. In case of `.produce`, the second field is `.item` and holds the data that is produced. In case of `.consume`, the second field is `.queue` and holds a pointer to the queue where the response message must be enqueued. `.produce` messages do not get a response.

The `pc_actor` process has two arguments. q is a pointer to the queue on which it receives messages. *nrequests* is the total number of requests that it will handle. We need the latter because in CXL all processes are required to terminate. In a more realistic implementation, the `pc_actor` would be in an infinite loop awaiting requests.

The shared memory consists entirely of message queues. The `produce` method takes two arguments: a pointer to the queue of the `pc_actor` and the item. The `consume` method takes two queue pointer arguments. The first is a pointer to the queue of the `pc_actor`. The second is a pointer to the queue of the invoker: it is the queue to which the response must be posted.

Figure 18.3 shows how this code may be used. It spawns `NITEMS` producer and consumer processes. Note that in that case the `pc_actor` is expected to receive `2 * NITEMS` requests.

Exercises

18.1 A popular model is the *client/server* model. Here a single actor implements some service, like computing the square of a number. A client can send a message containing an integer to the server, and the server returns a response message containing the square of that message. Implement this.

18.2 Actors and message queues are good for building pipelines. Develop a pipeline that computes Mersenne primes (primes that are one less than a factor of two). Write four actors:

1. an actor that generates a sequence of integers 1 through `N`;
2. an actor that receives integers and forwards only those that are prime;
3. an actor that receives integers and forwards only those that are one less than a factor of two;
4. an actor that receives integers but otherwise ignores them.

Configure two versions of the pipeline, one that first checks if a number is prime and then if it is one less than a factor of two, the other in the opposite order. Which do you think is better?

Chapter 19

Networking: Alternating Bit Protocol

A distributed system is a concurrent system in which a collection of processes communicate by message passing, much the same as in the actor model. The most important difference between distributed and concurrent systems is that the former takes *failures* into account, including failures of processes and failures of shared memory. In this chapter, we will consider two actors, Alice and Bob. Alice wants to send a sequence of application messages to Bob, but the underlying network may lose messages. The network does not re-order messages: when sending messages m_1 and m_2 in that order, then if both messages are received, m_1 is received before m_2. Also, the network does not create messages out of nothing: if message m is received, then message m was sent.

We need a network protocol that Alice and Bob can run. In particular, it has to be the case that if Alice sends application messages $m_1, ..., m_n$ in that order, then if Bob receives an application message m, then $m = m_i$ for some i and Bob will already have received application messages $m_1, ..., m_i$ (safety). Also, if the network is fair, then if Alice sends application message m, then eventually Bob should deliver m (liveness).

The *Alternating Bit Protocol* is suitable for our purposes. We assume that there are two unreliable network channels: one from Alice to Bob and one from Bob to Alice. Alice and Bob each maintain a zero-initialized sequence number, *s_seq* and *r_seq* resp. Alice sends a network message to Bob containing an application message as *payload* and Alice's sequence number as *header*. When Bob receives such a network message, Bob returns an an *acknowledgment* to Alice, which is a network message containing the sequence number in the message that Bob received.

In the protocol, Alice keeps sending the same network message until she receives an acknowledgment with the same sequence number. When this happens, Alice increments her sequence number. She is now ready to send the next message she wants to send to Bob. Bob, on the other hand, waits until he receives a message matching Bob's sequence number. If so, Bob *delivers* the payload in the message and increments his sequence number. Because of the network properties, a one-bit sequence number suffices.

Because of the network's properties, we can model each channel as a variable that either contains a network message or nothing (we use `()` in the model). Let *s_chan* be the channel from Alice to Bob and *r_chan* the channel from Bob to Alice. `net_send`(*pchan*, *m*, *reliable*) models sending a message *m* to ^*pchan*, where *pchan* is either &*s_chan* or &*r_chan*. The method places either *m*

```
def net_send(pchan, m, reliable):
    ^pchan = m if (reliable or choose({ False, True })) else ();
;
def net_recv(pchan):
    result = ^pchan;
;
def app_send(payload):
    s_seq = 1 - s_seq;
    let m = dict{ .seq: s_seq, .payload: payload }, blocked = True:
        while blocked:
            net_send(&s_chan, m, False);
            let response = net_recv(&r_chan):
                blocked = (response == ()) or (response.ack != s_seq);
            ;
        ;
    ;
;
def app_recv(reliable):
    r_seq = 1 - r_seq;
    let blocked = True:
        while blocked:
            let m = net_recv(&s_chan):
                if m != ():
                    net_send(&r_chan, dict{ .ack: m.seq }, reliable);
                    if m.seq == r_seq:
                        result = m.payload;
                        blocked = False;
                    ;
                ;
            ;
        ;
    ;
;
s_chan = ();
r_chan = ();
s_seq = 0;
r_seq = 0;
```

Figure 19.1: [abp.cxl] Alternating Bit Protocol.

```
import abp;

const NMSGS = 5;

def sender():
    for i in 1..NMSGS:
        app_send(i);
    ;
;
def receiver():
    for i in 1..NMSGS:
        let payload = app_recv(i == NMSGS):
            assert payload == i;
        ;
    ;
;
spawn sender();
spawn receiver();
```

Figure 19.2: [abptest.cxl] Test code for alternating bit protocol.

(to model a successful send) or `()` (to model loss) in `^`*pchan*. The use of the *reliable* flag will be explained later. `net_recv`(*pchan*) models checking `^pchan` for the next message.

Method `app_send`(m) tries *retransmitting* m until an acknowledgment is received. Method `app_recv`(*reliable*) returns the next successfully received message. Figure 19.2 shows how the methods may be used to send and receive a stream of `NMSGS` messages reliably. It has to be bounded, because model checking requires a finite model.

Only the last invocation of `app_recv`(*reliable*) is invoked with *reliable* == `True`. It causes the last acknowledgment to be sent reliably. It allows the receiver (Bob) to stop, as well as the sender (Alice) once the last acknowledgment has been received. Without something like this, either the sender may be left hanging waiting for the last acknowledgment, or the receiver waiting for the last message.

Exercises

19.1 Exercise 18.1 explored the *client/server*. It is popular in distributed systems as well. Develop a protocol for a single client and server using the same network model as for the ABP protocol. Hint: the response to a request can contain the same sequence number as the request.

19.2 Generalize the solution in the previous exercise to multiple clients. Each client is uniquely identified. You may either use separate channel pairs for each client, or solve the problem using a single pair of channels.

Chapter 20

Non-Blocking Synchronization

So far we have concentrated on critical sections to synchronize multiple processes. Certainly, preventing multiple processes from accessing certain code at the same time simplifies how to think about synchronization. However, it can lead to starvation. Even in the absence of starvation, if some process is slow for some reason while being in the critical section, the other processes have to wait for it to finish executing the critical section. In this chapter, we will have a look at how one can develop concurrent code in which processes do not have to wait for other processes.

As a first example, we will revisit the producer/consumer problem. The code in Figure 20.1 is based on code developed by Herlihy and Wing [HW87]. The code is a "proof of existence" for non-blocking synchronization; it is not necessarily practical. There are two variables. *items* is an unbounded array with each entry initialized to `()`. *back* is an index into the array and points to the next slot where a new value is inserted. The code uses two interlock instructions:

- `inc`(p): atomically increments `^`p and returns the old value;
- `swap`(p): sets `^`p to `()` and returns the old value.

Method `produce`(*item*) uses `inc`(&*back*) to allocate the next available slot in the *items* array. It stores the item as a singleton tuple. Method `consume()` repeatedly scans the array, up to the *back* index, trying to find an item to return. To check an entry, it uses `swap()` to atomically remove an item from a slot if there is one. This way, if two or more processes attempt to extract an item from the same slot, at most one will succeed.

There is no critical section. If one process is executing instructions very slowly, this does not negatively impact the other processes, as it would with solutions based on critical sections. On the contrary, it helps them because it creates less contention. Unfortunately, the solution is not practical for the following reasons:

- The *items* array must be of infinite size if an unbounded number of items may be produced;
- Each slot in the array is only used once, which is inefficient;
- the `inc` and `swap` interlock instructions are not universally available on existing processors.

```
const MAX_ITEMS = 3;

def inc(pcnt):
   atomic:
      result = ^pcnt;
      ^pcnt += 1;
   ;
;
def swap(pv):
   atomic:
      result = ^pv;
      ^pv = ();
   ;
;
def produce(item):
   items[inc(&back)] = (item,);
;
def consume():
   result = ();
   while result == ():
      let range = back, i = 0:
         while (i < range) and (result == ()):
            result = swap(&items[i]);
            i += 1;
         ;
      ;
   ;
   result = result[0];     # convert (item,) into item
;
back = 0;
items = [ () for i in 0..MAX_ITEMS ];

for i in 1..MAX_ITEMS:
   spawn produce(i);
;
for i in 1..choose(0..MAX_ITEMS):
   spawn consume();
;
```

Figure 20.1: [hw.cxl] Non-blocking queue.

However, in the literature there are many examples of practical non-blocking (aka *wait-free*) synchronization algorithms [].

There are also various algorithms where read-only access is wait-free but updates do require a lock—these are useful in situations where most accesses are read-only and the performance of updates is less critical []. Figures 20.2 and 20.3 implements an ordered linked list of integers without duplicates. There are two update operations: values can be added using `lst_insert` or deleted using `lst_remove`. There is also a read-only operation: `lst_contains` checks if a particular value is in the list.

The implementation uses the `alloc` module for dynamic allocation of nodes in the list using `recAlloc` and `recFree`. The read-only operation `lst_contains` is wait_free: it scans through the list without having to obtain a lock. Nonetheless, the implementation of the list is *linearizable* [HW90], a strong notion of consistency that makes it appear as if each of the operations executes atomically at some point between their invocation and return.

The list has two "book-end" nodes with values `-inf` and `inf` (like the Python `math.inf` constants). Each node has a lock, a value, and *next*, a pointer to the next node (which is `None` for the final `inf` node). The `lst_find`(*lst*, *v*) method is a helper function for update methods that finds and locks two consecutive nodes *before* and *after* such that *before*.`data.value` $< v \leq$ *after*.`data.value`. It does so by performing something called *hand-over-hand locking*. It first locks the first node, which is the `-inf` node. Then, iteratively, it obtains a lock on the next node and release the lock on the last one, and so on, similar to climbing a tree hand-over-hand.

Having the two adjoining nodes locked, implementing `lst_insert` and `lst_remove` is straightforward. The `lst_contains` method can simply scan through the list because the *next* pointers are updated atomically in such a way that they always point to a legal suffix of the list. An invariant of the algorithm is that at any point in time the list is "complete", starting with a `-inf` node and ending with a `inf` node.

Determining if an implementation of a concurrent data structure is linearizable involves finding what are known as the *linearization points* of the operations in an execution. These are the unique points in time at which an operation appears to execute atomically. The linearization points for the `lst_insert` and `lst_remove` operations coincide exactly with the update of the *before*.`next` pointer. The linearization point of a `lst_contains` method execution depends on whether the value is found or not. If found, it coincides with retrieving the pointer to the node that has the value. If not found, it coincides with retrieving the pointer to the `inf` node.

Exercises

20.1 Add read-only methods to the data structure in Figures 20.2 and 20.3 that report the size of the list, the minimum value in the list, the maximum value in the list, and the sum of the values in the list. Are they linearizable? If so, what are their linearization points?

20.2 A *seqlock* consists of a lock and a version number. An update operation acquires the lock, increments the version number, makes the changes to the data structure, and then releases the lock. A read-only operation does not use the lock. Instead, it retrieves the version number, reads the data structure, and then checks if the version number has not changed. If it has, read-only operation is retried. Use a seqlock to implement a bank much like Exercise 17.1, with one seqlock for the entire

```
import synch;
import alloc;

def lst_node(v, n):
    result = recAlloc();
    (^result).data = dict{ .lock: Lock(), .value: v };
    (^result).next = n;
;
def lst_new():
    result = lst_node(–inf, lst_node(inf, None));
;
def lst_contains(lst, v):
    let n = lst:
        while (^n).data.value < v:
            n = (^n).next;
        ;
        result = (^n).data.value == v;
    ;
;
```

Figure 20.2: [lst1.cxl] List with non-blocking read operations, part 1.

bank (i.e., no locks on individual accounts). Method `transfer` is an update operation; method `total` is a read-only operation. Explain how a seqlock can lead to starvation.

```
import lst1;

def lst_find(lst, v):
    let before = lst, after = (^before).next:
        lock(&(^before).data.lock);
        lock(&(^after).data.lock);
        while (^after).data.value < v:
            unlock(&(^before).data.lock);
            before = after;
            after = (^before).next;
            lock(&(^after).data.lock);
        ;
        result = (before, after);
    ;
;
def lst_insert(lst, v):
    let before, after = lst_find(lst, v):
        if (^after).data.value != v:
            (^before).next = lst_node(v, after);
        ;
        unlock(&(^before).data.lock);
        unlock(&(^after).data.lock);
    ;
;
def lst_remove(lst, v):
    let before, after = lst_find(lst, v):
        if (^after).data.value == v:
            (^before).next = (^after).next;
            unlock(&(^after).data.lock);
            recFree(after);
        else:
            unlock(&(^after).data.lock);
        ;
        unlock(&(^before).data.lock);
    ;
;

mylist = lst_new();
lst_insert(mylist, 3);
assert lst_contains(mylist, 3);
lst_remove(mylist, 3);
assert not lst_contains(mylist, 3);
```

Figure 20.3: [lst2.cxl] List with non-blocking read operations, part 2.

Chapter 21

Barrier Synchronization

Barrier synchronization is a problem that comes up in high-performance parallel computing, used, among others, for scalable simulation. A barrier is almost the opposite of a critical section: the intention is to get a group of processes to run some code at the same time, instead of having them execute it one at a time. More precisely, with barrier synchronization the processes execute in rounds. Between each round there is a so-called *barrier* where processes wait until all processes have completed the round and before they start the next one. For example, in an iterative matrix algorithm, the matrix may be cut up into fragments. During a round, the processes run concurrently one for each fragment. The next round is not allowed to start until all processes have completed processing their fragment.

Figure 21.1 shows a high-level depiction with three processes. Initially all processes are in round 0. Then each process can start and finish the round. However, none of the processes can progress to the next round until all processes have finished the current round.

Counting semaphores work well for implementing barrier synchronization. Figure 21.2 shows an example. There is a 0-initialized semaphore for each of the `N` processes. Before process i enters a round, it first sends a signal to every other process and then waits until it receives a signal from

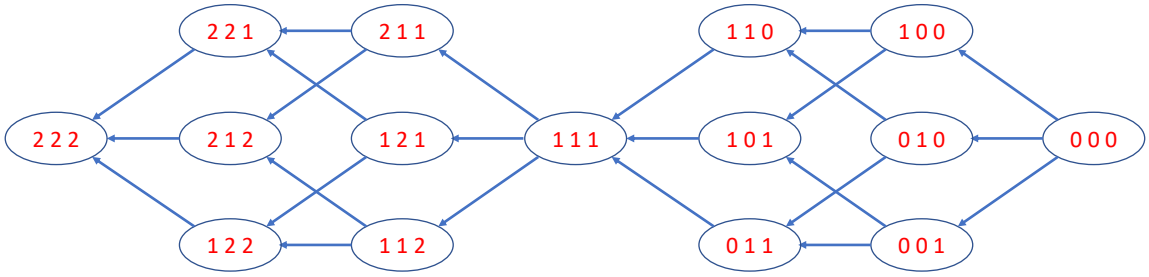

Figure 21.1: High-level state diagram specification of barrier synchronization with three processes and two rounds. The three numbers specify how many rounds each process has completed.

```
import synch;
import list;

const N = 3;        # number of processes
const R = 4;        # number of rounds

def process(self):
   for r in 1..R:
      for i in 0..(N–1):
         if i != self:
            V(&sema[i]);
         ;
      ;
      for i in 0..(N–1):
         if i != self:
            P(&sema[self]);
         ;
      ;
      round[self] = (round[self] + 1);
      assert (listMax(round) – listMin(round)) ¡= 1;
   ;
;
round = [ 0 for i in 0..(N–1) ];
sema = [ Semaphore(0) for i in 0..(N–1) ];
for i in 0..(N–1):
   spawn process(i);
;
```

Figure 21.2: [barrier.cxl] Barrier synchronization with semaphores.

every other process. Process i sends a signal to process j by incrementing `sema[`j`]` (using `V`), while process i waits for a signal be decrementing `sema[`i`]` (using `P`).

The *round* array is kept to check the correctness of this approach. Each process increments its entry every time it enters a round. If the algorithm is correct, it can never be that two processes are more than one round apart.

Exercises

21.1 See if you can implement barrier synchronization for `N` processes with just three semaphores. Busy waiting is not allowed. How about just two? (As always, the Little Book of Semaphores [Dow09] is a good resource for solving semaphore problems. Look for the double turnstile solution.)

21.2 Implement barrier synchronization with Mesa condition variables. (You may want to use a double turnstile approach here as well.)

Bibliography

[BH73] Per Brinch Hansen. *Operating System Principles*. Prentice-Hall, Inc., USA, 1973.

[Bir89] Andrew D. Birrell. An introduction to programming with threads. SRC report 35, Digital Systems Research Center, Palo Alto, CA, USA, January 1989.

[CES71] Edward G. Coffman, Melanie Elphick, and Arie Shoshani. System deadlocks. *ACM Comput. Surv.*, 3(2):67–78, June 1971.

[CHP71] Pierre-Jacques Courtois, Frans Heymans, and David L. Parnas. Concurrent control with "readers" and "writers". *Commun. ACM*, 14(10):667–668, October 1971.

[Dij62] Edsger W. Dijkstra. EWD-35: Over de sequentialiteit van procesbeschrijvingen. circulated privately, approx. 1962.

[Dij64] Edsger W. Dijkstra. EWD-108: Een algorithme ter voorkoming van de dodelijke omarming. circulated privately, approx. 1964.

[Dij65] Edsger W. Dijkstra. EWD-123: Cooperating Sequential Processes. circulated privately, 1965.

[Dij72] Edsger W. Dijkstra. EWD-329 information streams sharing a finite buffer. 1972.

[Dij79] Edsger W. Dijkstra. EWD-703: A tutorial on the split binary semaphore. circulated privately, March 1979.

[Dow09] Allen B. Downey. *The Little Book Of Semaphores*. Green Tea Press, 2009.

[Hav68] James W. Havender. Avoiding deadlock in multitasking systems. *IBM Syst. J.*, 7(2):74–84, June 1968.

[HBS73] Carl Hewitt, Peter Bishop, and Richard Steiger. A universal modular actor formalism for artificial intelligence. In *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, IJCAI'73, page 235–245, San Francisco, CA, USA, 1973. Morgan Kaufmann Publishers Inc.

[Hoa73] C. A. R. Hoare. Towards a theory of parallel programming. In C. A. R. Hoare and R. H. Perrott, editors, *Operating Systems Techniques*. Academic Press, New York, NY, 1973.

[Hoa74] C. A. R. Hoare. Monitors: An operating system structuring concept. *Commun. ACM*, 17(10):549–557, October 1974.

[HW87] Maurice P. Herlihy and Jeannette M. Wing. Axioms for concurrent objects. In *Proceedings of the 14th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*, POPL '87, page 13–26, New York, NY, USA, 1987. Association for Computing Machinery.

[HW90] Maurice P. Herlihy and Jeannette M. Wing. Linearizability: A correctness condition for concurrent objects. *ACM Trans. Program. Lang. Syst.*, 12(3):463–492, July 1990.

[Lam02] Leslie Lamport. *Specifying Systems: The TLA+ Language and Tools for Hardware and Software Engineers*. Addison-Wesley Longman Publishing Co., Inc., USA, 2002.

[Lam09] Leslie Lamport. The PlusCal Algorithm Language. In Martin Leucker and Carroll Morgan, editors, *Theoretical Aspects of Computing - ICTAC 2009*, pages 36–60, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[LR80] Butler W. Lampson and David D. Redell. Experience with processes and monitors in Mesa. *Commun. ACM*, 23(2):105–117, February 1980.

[Pet81] Gary L. Peterson. Myths about the mutual exclusion problem. *Information Processing Letters*, 12(3):115 – 116, 1981.

[Sch97] Fred B. Schneider. *On Concurrent Programming*. Springer-Verlag, Berlin, Heidelberg, 1997.

Appendix A

List of Values

Chapter 4 provides an introduction to CXL values. Below is a complete list of CXL value types with examples:

Boolean	`False`, `True`
Integer	`-inf, ..., -2, -1, 0, 1, 2, ..., inf`
Atom	`.example`, `.test1`
Program Counter	(method names are program counter constants)
Dictionary	`dict{ .account: 12345, .valid: False }`
Set	`{ 1, 2, 3 }, { False, .id, 3 }`
Address	`&lock, &flags[2], None`
Context	(generated by **stop** expression)

Tuples, lists, strings, and bags are all special cases of dictionaries. Both tuples and lists map indexes (starting at 0) to CXL values. Their format is either `(`*e*`, `*e*`, ..., `*e*`,)` or `[`*e*`, `*e*`, ..., `*e*`,]`. If the tuple or list has two or more elements, then the final comma is optional. A string must be enclosed by double quotes and is represented as a tuple of its characters. Characters are one-character atoms. A bag or multiset is a dictionary that maps a value to how many times it occurs in the bag.

All CXL values are ordered with respect to one another. First they are ordered by type according to the table above. So, for example, `True < 0 < .xyz < { 0 }`). Within types, the following rules apply:

- `False < True`;
- integers are ordered in the natural way;
- atoms are lexicographically ordered;
- program counters are ordered by their integer values;
- dictionaries are first converted into a list of ordered (key, value) pairs. Then two dictionaries are lexicographically ordered by this representation;
- a set is first converted into an ordered list, then lexicographically ordered;

- an address is a list of atoms. `None` is the empty list of atoms. Addresses are lexicographically ordered accordingly;
- contexts (Appendix F) are ordered first by name tag, then by program counter, then by hash.

Appendix B

List of Operators

CXL currently supports the following operators:

e `==` *e*, *e* `!=` *e*	two CXL values of the same type
e `<` *e*, *e* `<=` *e*	any two CXL values
e `>` *e*, *e* `>=` *e*	any two CXL values
e `and` *e* `and ...`	two or more booleans
e `or` *e* `or ...`	two or more booleans
`not` *e*	a boolean
`-`*e*	an integer
e `+` *e* `+ ...`	two or more integers, sets, or lists
e `-` *e*	two integers or sets
e `*` *e* `* ...`	two or more integers or sets
e `/` *e*	two integers (integer division)
e `%` *e*	two integers (division remainder)
e `..` *e*	two integers
e `[not] in` *e*	first is any CXL value, second is set
e `if` *e* `else` *e*	middle *e* must be a boolean
`^`*e*	an address
`hash` *e*	any CXL value
`choose` *e*	a set
`min` *e*, `max` *e*	a set
`cardinality` *e*	a set
`keys` *e*, `len` *e*	a dictionary (includes tuple or list)
`bagsize` *e*	a bag
`atLabel` *e*	atom (corresponding to a label)
`nametag` *e*	*e* must be `()`
`processes` *e*	*e* must be `()`
`&`*lv*	*lv* is an lvalue
`stop` *lv*	*lv* is an lvalue

`+` computes the union when applied to sets and the concatenation when applied to lists or tuples. `-` computes set difference when applied to two sets. `*` computes the intersection when applied to sets. x`..`y computes the set consisting of the integers x through y inclusive. It returns the empty set if $x > y$.

An *lvalue* (short for left hand value of an assignment statement) is something that can be assigned. This can be a shared variable, a process variable, or a dereferenced pointer variable. It can also be indexed. Examples include *var*, *var*[*e*], and `^`*p*.

CXL also supports the following comprehensions:

Set comprehension	{ `f(`e`) for` e `in` s }
List comprehension	[`f(`e`) for` e `in` s]
Dict comprehension	`dict`{ `f(`e`) for` e `in` s }

In each case, s must be a set. For list comprehensions, the list will be sorted by e. For dictionary comprehensions, each e will be mapped to `f`(e).

Appendix C

List of Statements

CXL currently supports the following statements:

lv = *e*	*lv* is an lvalue and *e* is an expression
lv `[op]=` *e*	`op` is one of `+`, `-`, `*`, `/`, `and`, and `or`
`pass`	do nothing
`del` *lv*	delete
`assert` *b* `[,` *x*`]`	*b* is a Boolean expression. Optionally report value of expression *e*
`const v =` *e*	`v` is an identifier, *e* is a constant expression
`def m(`*v*`, ...): S`	`m` and `v` are identifiers, `S` a list of statements
`let` *v* `=` *e*`, ...: S`	*v* is an identifier, *e* is an expression, `S` a list of statements
`if` *b*`: S`	*b* is a Boolean expression, `S` a list of statements
`while` *b*`: S`	*b* is a Boolean expression, `S` a list of statements
`for` *v* `in` *e*`: S`	*v* is an identifier, *e* is a set evaluated in order, `S` a list of statements
`atomic: S`	`S` a list of statements
`spawn m` *e* `[,` *t*`]`	`m` is a method, *e* is an expression, *t* is a tag (an expression)
`go` *c* *e*	*c* is a context, *e* is an expression
`import m`	`m` identifies a module

CXL does not support Python-style iterators. To iterate through a dictionary, list, or tuple, use: `for` *v* `in keys(`*e*`): S`.

Appendix D

List of Modules

D.1 The `alloc` module

The `alloc` module support thread-safe dynamic allocation of "records." Each record has two fields: `.data` and `.next` that can be used arbitrarily, although `.next` is intended to be used as a pointer to another record. There are just two methods:

`recAlloc()`	allocate a record
`recFree(`r`)`	free record r

The usage is similar to `malloc` and `free` in C. `recAlloc` returns `None` when running out of memory.

D.2 The `bag` module

The `bag` module has various useful methods that operate on bags or multisets:

`bagEmpty()`	returns an empty bag
`bagFromSet(`s`)`	create a bag from set s
`bagCount(`b`, `e`)`	count how many times e occurs in bag b
`bagChoose(`b`)`	like `choose(s)`, but applied to a bag
`bagAdd(`pb, e`)`	add one copy of e to bag `^`pb
`bagRemove(`pb, e`)`	remove one copy of e from bag `^`pb

D.3 The `list` module

The `list` module has various useful methods that operate on lists or tuples:

`subseq(`t`, `b`, `b`)`	return a *slice* of list t starting at index b and ending just before index f
`append(`t`, `e`)`	append e to list t
`head(`t`)`	return the first element of list t
`tail(`t`)`	return all but the first element of list t
`listQsort(`t`)`	sort list t using quicksort
`list2bag(`t`)`	convert list t into a bag
`list2set(`t`)`	convert list t into a set
`listMin(`t`)`	return the minimum of all elements in t
`listMax(`t`)`	return the maximum of all elements in t
`listSum(`t`)`	return the sum of all elements in t

D.4 The `synch` module

The `synch` module provides the following methods:

tas(*lk*)	test-and-set on `^`*lk*
Lock()	return a lock object
lock(*lk*)	acquire `^`*lk*
unlock(*lk*)	release `^`*lk*
Condition(*lk*)	return a condition variable on `^`*lk*
wait(*c*)	wait on condition variable `^`*c*
notify(*c*)	notify a process waiting on condition variable `^`*c*
notifyAll(*c*)	notify all processes waiting on condition variable `^`*c*
Semaphore(*cnt*)	return a semaphore object initialized to *cnt*
P(*sema*)	procure `^`*sema*
V(*sema*)	vacate `^`*sema*
Queue()	return a synchronized queue object
dequeue(*q*)	return next element of *q*, blocking if empty
enqueue(*q*, *item*)	add *item* to *a*

Appendix E

List of Machine Instructions

Address n	compute address from n components
Apply	pop m and i and apply i to m, pushing a value
Assert	pop b and check that it is `True`
AtomicInc	increment the atomic counter of this context
AtomicDec	decrement the atomic counter of this context
Continue	no-op (but causes a context switch)
Choose	choose an element from the set on top of the stack
Del [v]	delete shared variable v
DelVar [v]	delete process variable v
Dict	pop n and n key/value pairs and push a dictionary
Dup	duplicate the top element of the stack
Frame `m`(a)	start method `m` with arguments a, initializing variables.
Go	pop context and value, push value on context's stack, and add to context bag
Jump p	set program counter to p
JumpCond e p	pop expression and, if equal to e, set program counter to p
Load [v]	push the value of a shared variable onto the stack
LoadVar [v]	push the value of a process variable onto the stack
n-ary op	apply n-ary operator op to the top n elements on the stack
Pop	pop a value of the stack and discard it
Push c	push constant c onto the stack
PushAddress v	push the address of variable v
Return	pop return address, push `result`, and restore program counter
Set	pop n and n elements and push a set of the elements
Spawn	pop tag, argument, and method and spawn a new context
Split	pop set and push minimum and remainder
Stop [v]	save context into shared variable v and remove from context bag
Store [v]	pop a value from the stack and store it in a shared variable
StoreVar [v]	pop a value from the stack and store it in a process variable
Swap	swap the two values on top of the stack

Clarifications:

- The `Address` instruction expects n values $v_1, ..., v_n$ on the stack. The top value must be an address value. The remaining values are keys into dictionaries. The instruction then computes the address of `^`v_1 v_2 ... v_n.
- Even though CXL code does not allow taking addresses of process variables, both shared and process variables can have addresses.
- The `Load`, `LoadVar`, `Del`, `DelVar`, and `Stop` instructions have an optional variable name: if omitted the top of the stack must contain the address of the variable.
- `Store` and `StoreVar` instructions have an optional variable name. In both cases the value to be assigned is on the top of the stack. If the name is omitted, the address is underneath that value on the stack.
- The effect of the `Apply` instructions depends much on m. If m is a dictionary, then `Apply` finds i in the dictionary and pushes the value. If m is a program counter, then `Apply` invokes method m by pushing the current program counter and setting the program counter to m. m is supposed to leave the result on the stack.
- The `Frame` instruction pushes the value of the process register (*i.e.*, the values of the process variables) onto the stack. It initializes the `result` variable to the empty dictionary. The `Return` instruction restores the process register by popping its value of the stack.
- All method calls have exactly one argument, although it sometimes appears otherwise:
  - `m()` invokes method `m` with the empty dictionary `()` as argument;
  - `m(`a`)` invokes method `m` with argument a;
  - `m(`a`, `b`, `c`)` invokes method `m` with tuple `(`a`, `b`, `c`)` as argument.

  The `Frame` instruction unpacks the argument to the method and places them into process variables by the given names.
- Every `Stop` instruction must immediately be followed by a `Continue` instruction.

Appendix F

Contexts and Processes

A context captures the state of a process. Each time the process executes an instruction, it goes from one context to another. All instructions update the program counter, and so no instruction leaves the context the same. There may be multiple processes with the same state at the same time. A context consists of the following:

name tag	a dictionary with atoms `.name` and `.tag`
program counter	an integer value pointing into the code
frame pointer	an integer value pointing into the stack
atomic	if non-zero, the process is in atomic mode
stack	a list of CXL values
register	a CXL dictionary mapping atoms (representing process variables) to CXL values
stopped	a boolean indicating if the context is stopped
failure	if not None, string that describes how the process failed

Details:

- The name in a name tag is the name of the method that the process is executing;
- The tag in a name tag is its argument by default. Optionally `spawn` allows the tag to be specified explicitly;
- The frame pointer points to the current *stack frame*, which consists of the caller's frame and variables, the argument to the method, and the return address.
- A process terminates when it reaches the `Return` instruction of the method or when it hits a fault. Faults include divide by zero, reading a non-existent key in a dictionary, accessing a non-existent variable, as well as when an assertion fails;
- The execution of a process in *atomic mode* does not get interleaved with that of other processes.
- The register of a process always contains a dictionary, mapping atoms to arbitrary values. The atoms correspond to the variable names in a CXL program.

A context is considered *blocked* if there does not exist a sequence instructions that the process can execute by itself that leads to its termination.

Appendix G

The CXL Virtual Machine

The CXL Virtual Machine (CVM) has the following state:

code	a list of CVM machine instructions
labels	a dictionary of atoms to program counters
variables	a dictionnary mapping atoms to values
ctxbag	a bag of non-stopped contexts
stopbag	a bag of stopped contexts
choosing	if not `None`, indicates a context that is choosing
initializing	boolean indicating that the state is not fully initialized.

There is initially a single context with nametag `__init__/()` and program counter 0. It starts executing in atomic mode until it finishes executing the last instruction in the code. All states until then are `initializing` states. Other processes, created through `spawn` statements, do not start executing until then.

A *micro step* is the execution of a single CVM machine instruction by a context. Each micro step generates a new state. When there are multiple contexts, the CVM can interleave them. However, trying to interleave every microstep would be needlessly expensive, as many micro steps involve changes to a context that are invisible to other contexts.

A *macro step* can involve multiple micro steps. The following instructions start a new macro step: `Load`, `Store`, `AtomicInc`, and `Continue`. The CVM interleaves macro steps, not micro steps. Like micro steps, each macro step involves a single context. Unlike a micro step, a macro step can leave the state unchanged.

Executing a CXL program results in a graph where the nodes are CXL states and the edges are macro steps. When a state is `choosing`, the edges from that state are by a single context, one for each choice. If not, the edges from the state are one per context. A terminating state is a state with an empty context bag or a state where all contexts are blocked.

Acknowledgments

I received considerable help and inspiration from various people while writing this book.

First and foremost I would like to think my student Haobin Ni with whom I've had numerous discussions about the design of CXL. Haobin even contributed some code to the CXL compiler.

Most of what I know about concurrent programming I learned from my colleague Fred Schneider. He suggested I write this book after demonstrating CXL to him.

Leslie Lamport introduced me to using model checking to test properties of a concurrent system. My experimentation with using TLC on Peterson's Algorithm became an aha momment for me.

I first demonstrated CXL to the students in my CS6480 class on systems and formal verification and received valuable feedback from them.

I would also like to thank my proofreaders Anneke van Renesse and Zhuoyu Xu.

Finally, I would like to thank my family who had to suffer as I obsessed over writing the code and the book.

Index

actor model, 87
address, 15, 32
alloc module, 109
alternating bit protocol, 91
atLabel operator, 21
atom, 13
atomic statement, 35
atomicity, 3

bag, 14
bag module, 109
barrier synchronization, 99
binary semaphore, 63
blocked process, 39, 114
blocking queue, 87
bounded buffer, 59
broadcast, 76
busy waiting, 48
bytecode, 13

choose operator, 7
circular buffer, 59
client/server model, 59
condition variable, 74
conditional critical section, 65
constant, 7
context, 14
continuation, 14
corner case, 4
counting semaphore, 63
critical region, 20
critical section, 20
CVM, 13
CXL method, 32
CXL Virtual Machine, 13

deadlock, 80
deadlock avoidance, 84
determinism, 3
dictionary, 13
dining philosopher, 80
directory, 15
distributed system, 91
dynamic allocation, 96

failure, 91
fairness, 69
formal verification, 4

go statement, 43

hand-over-hand locking, 96
Heisenbug, 3

import statement, 32
inductive invariant, 28
interleaving, 9
interlock instruction, 35
invariant, 4, 28

label, 21
linearizable, 96
linearization point, 96
list module, 109
liveness property, 21
lock, 22, 39, 43
lvalue, 107

machine instruction, 8
macro step, 115
Mesa, 74
message passing, 87
model checking, 3
module, 32, 43

monitor, 74
multiple conditions, waiting on, 83
multiset, 14
mutual exclusion, 21

name tag, 14
nametag operator, 21
network, 91
non-blocking synchronization, 94
non-determinism, 26
notify, 75
notifyAll, 76

P, 54
Peterson's Algorithm, 26
pipeline, 59
pointer, 32
process, 8
process variable, 26
procure, 54
producer/consumer problem, 59
program counter, 14
progress, 21
property, 69
protocol, 91

race condition, 9
reachable state, 26
reader/writer lock, 45
register, 14
ring buffer, 59

safety property, 21
semaphore, 54
seqlock, 96
sequential, 3
shared variable, 3
signal, 74
spawn statement, 14
spin waiting, 48
spinlock, 35
split binary semaphore, 65
stack machine, 16
starvation, 43, 69
state, 26
state diagram, 22
step, 26
stop expression, 43
synch module, 39, 110
synchronized queue, 87

tag, 14
TAS, 35
test, 3
test-and-set, 35
thread, 3, 20
thread safety, 20
trace, 26

unlock, 43

V, 54
vacate, 54
virtual machine, 13

wait, 74
wait-free synchronization, 96

Glossary

actor model is a concurrency model where there are no shared variables, only processes with private variables that communicate through message passing. 87

atomicity describes that a certain machine instruction or sequence of machine instructions by a process is executed indivisibly and cannot be interleaved with machine instructions of another process. 3

barrier synchronization is when a set of processes execute in rounds, waiting for one another to complete each round. 99

blocked process is a process that is waiting on a low-level synchronization primitive and cannot terminate without the help of another process. For example, a process that is waiting for a spinlock to become available. 39

busy waiting (aka spin-waiting) is when a process waits in a loop for some application-defined condition. 48

concurrent execution (aka parallel execution) is when there are multiple processes executing and their machine instructions are interleaved in an unpredictable manner. 3

condition variable a variable that keeps track of which processes are waiting for a specific application-level condition. The variable can be waited on as well as signaled or notified. 74

conditional critical section is a critical section with, besides mutual exclusion, additional conditions on when a process is allowed to enter the critical section. 65

context (aka continuation) describes the state of a running process, including its program counter, the values of its variables (stored in its register), and the contents of its stack. 14

critical section (aka critical region) is a set of instructions that only one process is allowed to execute at a time. The instructions are, however, not executed atomically, as other processes can continue to execute and access shared variables. 20

deadlock is when there are two or more processes waiting indefinitely for one another to release a resource. 80

determinism is when the outcome of an execution is uniquely determined by the initial state. 3

fairness is when each process eventually can access each resource it needs to access with high probability. 69

interlock instruction a machine instruction that involves multiple memory load and/or store operations, executed atomically. 35

invariant is a binary predicate over states that must hold for every reachable state of a process. 4

linearizable is a consistency condition for concurrent access to an object, requiring that each access must appear to execute atomically sometime between the invocation of the access and its completion. 96

lock an object that can be owned by at most one process at a time. Useful for implementing mutual exclusion. 22

machine instruction is an atomic operation on the CXL virtual machine, executed by a process. 8

model checking is a formal verification method that explores all possible executions of a program, which must have a finite number of states. 3

monitor is a programming language paradigm that supports mutual exclusion as well as waiting for resources to become available. 74

mutual exclusion is a property that two processes never enter the same critical section. 20

non-blocking synchronization (aka wait-free synchronization) is when access to a shared resource can be guaranteed in a bounded number of steps even if other processes are not making progress. 94

process is a method in execution. We do not make the distinction between processes and threads. A process has a program counter, a register, and a stack. 8

process variable is a variable that is private to a single process and stored in its register. 26

producer/consumer problem is a synchronization problem whereby one or more producing processes submit items and one or more consuming process want to receive them. No item can get lost or forged or be delivered to more than one consumer, and producers and consumers should block if resources are exhausted. 59

property describes a set of execution traces or histories that are allowed by a program. Safety properties are properties in which "no bad things happen," such as violating mutual exclusion in a critical section. Liveness properties are properties where "something good eventually happens," like processes being able to enter the critical section if they want to. 69

race condition describes when multiple processes access shared state concurrently, leading to undesirable outcomes. 9

reader/writer lock is a lock on a resource that can be held by multiple processes if they all only read the resource. 45

semaphore is a counter that can be atomically incremented and decremented, but blocks the process until the counter is larger than zero first. 54

sequential execution is when there is just one process executing, as opposed to concurrent execution. 3

shared variable is a variable that is stored in the memory of the CXL virtual machine and shared between multiple processes, as opposed to a process variable. 3

spinlock is an implementation of a lock whereby a process loops until the lock is available, at which point the process atomically obtains the lock. 35

stack machine is a model of computing where the state of a process is kept on a stack. CXL uses a combination of a stack machine and a register-based machine. 15

starvation is when a process cannot make progress because it is continuously losing a competition with other processes to get access to a resource. 69

state an assignment of values to variables. In a CXL virtual machine, this includes the contents of its shared memory and the set of contexts. 26

step is the execution of a machine instruction by a process, updating its state. CXL distinguishes micro steps (machine instructions) and macro step (a sequence of instructions with at most one effect visible by other processes). 26

thread safety is when the implementation of a data structure allows concurrent access with well-defined semantics. 20

trace is a sequence of steps, starting from an initial state. An infinite trace is also called a *behavior*. 26