CSCI 3323 (Principles of Operating Systems), Fall 2014

Homework 2

Credit: 50 points.

1 Reading

Be sure you have read Chapter 2.

2 Problems

Answer the following questions. You may write out your answers by hand or using a word processor or other program, but please submit hard copy, either in class or in one of my mailboxes (outside my office or in the ASO).

1. (5 points) In class we discussed a proposed solution to the mutual-exclusion problem based on disabling interrupts, and rejected it because it doesn't work for systems with more than one CPU. For a system with a single CPU, however, this could be an acceptable solution, especially if the critical region is short. Write pseudocode for an implementation of semaphores for a single-CPU system that might not have a TSL instruction but does have library functions `enable_int()` and `disable_int()` to enable and disable interrupts respectively. (I.e., say what variables you would need for each semaphore, and give pseudocode for `up()` and `down()`.)

2. (5 points) The programming assignment for Homework 1 asked you to write a simple shell program using `fork()` to create a new process for each command executed by the shell. `fork()` essentially creates this new process by duplicating the process that calls it, including the state of any data structures related to open files. What advantages does this have? What are some possible disadvantages? Consider both situations in which the parent process waits for the child to finish (as in the shell program) and situations in which both processes continue concurrently. (*Hint:* Think about the standard input/output/error streams and also about other kinds of open files. Also try to apply what you know about buffering of input/output.)

3. (5 points) Solve the dining philosophers problem with monitors rather than semaphores. (Do this yourself, though, rather than looking for a solution online or in another book!)

4. (5 points) Restrooms are usually designated as men-only or women-only, but this requires having two restrooms if everyone is to be accommodated. A less expensive approach consistent with cultural norms in the U.S. would be to have one restroom with a sign on the door that indicates its current state — empty, in use by at least one woman, or in use by at least one man. If it is empty, either a man or a women may enter; if it is occupied, a person of the same sex may enter, but a person of the opposite sex must wait until it is empty. Write pseudocode for four functions to implement this approach: `woman_enter`, `man_enter`, `woman_leave`, and `man_leave`, to be used by the following pseudocode:

```
/* woman process */
while (TRUE) {
    woman_enter();
```

```
        use_restroom();
        woman_leave();
        do_other_stuff();
    }
    /* man process */
    while (TRUE) {
        man_enter();
        use_restroom();
        man_leave();
        do_other_stuff();
    }
```

You can use any of the synchronization mechanisms we have talked about (shared variables, semaphores, monitors, or even message passing).

5. (5 points) Five batch jobs (call them A through E) arrive at a computer center at almost the same time, in the order shown below. Their estimated running times (in minutes) and priorities are as follows, with 5 indicating the highest priority:

job	*running time*	*priority*
A	10	3
B	6	5
C	2	2
D	4	1
E	8	4

For each of the following scheduling algorithms, determine the turnaround time for each job and the average turnaround time. Assume that all jobs are completely CPU-bound (i.e., they do not block). (Before doing this by hand, decide how much of programming problem 2 you want to do.)

- First-come, first-served (run them in alphabetic order by job name).
- Shortest job first.
- Round robin, using a time quantum of 1 minute.
- Round robin, using a time quantum of 2 minutes.
- Preemptive priority scheduling.

6. (5 points) Suppose that a scheduling algorithm favors processes that have used the least amount of processor time in the recent past. Why will this algorithm favor I/O-bound processes yet not permanently starve CPU-bound processes, even if there is always an I/O-bound process ready to run?

3 Programming Problems

Do the following programming problems. You will end up with at least one code file per problem. Submit your program source (and any other needed files) by sending mail to `bmassing@cs.trinity.edu`, with each file as an attachment. Please use a subject line that mentions the course number and the assignment (e.g., "csci 3323 homework 2"). You can develop your programs on

any system that provides the needed functionality, but I will test them on one of the department's Linux machines, so you should probably make sure they work in that environment before turning them in.

1. (10 points) The starting point for this problem is a simple implementation of the mutual exclusion problem in C with POSIX threads m-e-problem.c[1]. Each thread executes a loop similar to the one presented in class for this problem, except that:

   - Rather than looping forever, each thread makes a finite number of trips through the loop.
   - The critical region is represented by code to print some messages and sleep for a random interval.
   - The non-critical region is represented by code to sleep for a random interval.

   Currently no attempt is made to ensure that only one thread at a time is in its critical region, and if you run it you will see that in fact it frequently happens that all the threads are in their critical region at the same time. Your mission is to correct this.

   Start by compiling the program, running it, and observing its behavior. To compile with `gcc`, you will need the extra flag `-pthread` and also `-std=c99`, e.g.,

   ```
   gcc -Wall -std=c99 -pthread m-e-problem.c
   ```

   (Or download this Makefile[2] and type `make m-e-problem`.) The program requires several command-line arguments, described in comments at the top of the code. (If you have trouble remembering the order, notice that the program prints a meant-to-be-helpful usage message if run with no arguments.)

   You are to produce two corrected versions of this program:

   - The first version should use shared variables only and one of the following algorithms:
     - Strict alternation, extended to work for an arbitrary number of threads. (No, this isn't a perfect solution, but it does enforce the "one at a time" condition.)
     - Peterson's algorithm, for two threads only. For extra credit, research and implement a variation that works for more than two threads. Cite a source for your solution if appropriate — e.g., "I found pseudocode for this solution at the following Web site." Or look up and implement Leslie Lamport's bakery algorithm.
   - The second version should use one of the following sets of library functions:
     - The POSIX threads mutex functions. `man pthread_mutex_init` is a good starting point for finding out about these functions.
     - The POSIX threads semaphore functions. `man sem_init` is a good starting point for finding out about these functions.

   Places in the program that should change are marked with "TODO" comments. You should not need to add much code. Confirm that your two improved versions behave as expected,

[1] `http://www.cs.trinity.edu/~bmassing/Classes/CS3323_2014fall/Homeworks/HW02/Problems/m-e-problem.c`

[2] `http://www.cs.trinity.edu/~bmassing/Classes/CS3323_2014fall/Homeworks/HW02/Problems/Makefile`

i.e., when one thread starts its critical region no other thread can start *its* critical region until the first one finishes.

NOTE about shared variables: Optimizing compilers play a lot of tricks to reduce actual accesses to memory, as do most processors. What this means for multithreaded programs is that it is very difficult to guarantee that changes made to a shared variable in one thread are visible to other threads. Declaring shared variables `volatile` avoids at least some compile-time optimizations but does not provide any guarantees about what will happen at runtime, especially if there are multiple processors. For the latter, what is needed is a "memory fence", i.e., a way of specifying that at a particular point in the program all memory reads and writes have completed. As far as I know there is no portable way to achieve this in C99; one must fall back on compiler- or processor-specific code. The starter code includes a function `memory_fence` that invokes a gcc-specific function providing a memory fence and recommends its use in the functions to begin and end the critical region. (*Disclaimer:* Last year the version of this function present on our classroom/lab machines apparently did nothing! This may be a bug in `gcc`, and whether is has been fixed I do not know. My sample solutions seem to work correctly anyway.) Note that some library functions for synchronization (e.g., the ones included with POSIX threads) incorporate this functionality as well.

2. (10 points) The starting point for this problem is a C++ program scheduler.cpp[3] that simulates execution of a scheduler, i.e., generates solutions to problems such as the one in the written part of this assignment. Comments describe input and desired output. Currently the program simulates only the FCFS algorithm. Your mission is to make it simulate one or more of the other algorithms mentioned in the written problem (FCFS, SJF, round robin using time quantums of 1 minute and 2 minutes, and preemptive priority scheduling). You will get full credit for simulating one algorithm, extra points for simulating additional algorithms.

   - Sample input[4].
   - Output for sample input[5].

   I chose C++ for the starter code because in theory all of you have had at least some exposure to C++, and this might be a good opportunity for you to dust off that skill. The starter code also makes use of some library classes (`string` and `vector`) that you may not have worked with before. `string` is functionally pretty similar to strings in languages such as Java and Scala; `vector` represents a templated expandable array (i.e., one with a type parameter that lets you specify the type of elements in the array). I'm cautiously optimistic that between the starter code, this toy example[6] of using `vector`, and what you can find on the Web about these classes (the Wikipedia articles seem okay), you will be able to use them to implement your choice of scheduling algorithm(s). If you don't remember, or didn't learn, how to compile C++ from the command line in Linux:

[3] `http://www.cs.trinity.edu/~bmassing/Classes/CS3323_2014fall/Homeworks/HW02/Problems/scheduler/scheduler.cpp`

[4] `http://www.cs.trinity.edu/~bmassing/Classes/CS3323_2014fall/Homeworks/HW02/Problems/scheduler/sample-in.txt`

[5] `http://www.cs.trinity.edu/~bmassing/Classes/CS3323_2014fall/Homeworks/HW02/Problems/scheduler/sample-out.txt`

[6] `http://www.cs.trinity.edu/~bmassing/Classes/CS3323_2014fall/Homeworks/HW02/Problems/scheduler/vector-example.cpp`

```
g++ -Wall -pedantic scheduler.cpp
```

However, feel free to rewrite anything about this program, including starting over in a language of your choice. Just remember that the program has to run on one of the department Linux machines, and it needs to accept input from command-line arguments and files — i.e., no GUIs, Web-based programs, etc. The latter requirement is to make it possible for me to automate testing your code. If you make changes to the format of the input — and I prefer that you don't — change the comments so they describe the changed requirements.