CS3331 Concurrent Computing Exam 1 Solutions
Fall 2019

1. **Basic Concepts**

   (a) **[10 points]** What is an atomic instruction? What would happen if multiple CPUs/cores execute their atomic instructions? **Make sure you will explain atomic instructions fully. Otherwise, you may receive a lower or very low score.**

   Answer: An atomic instruction is a machine instruction that executes as one *uninterruptible* unit without interleaving and cannot be split by other instructions. When an atomic instruction is recognized by the CPU, we have the following:

   - All other instructions being executed in various stages by the CPUs are suspended (and perhaps re-issued later) until this atomic instruction finishes.
   - No interrupts can occur.

   If two such instructions are issued at the same time on different CPUs/cores, they will be executed sequentially in an arbitrary order determined by the hardware.

   See pp. 12–13 of `02-Hardware-OS.pdf`. ∎

   (b) **[10 points]** Explain *interrupts* and *traps*, and provide a detailed account of the procedure that an operating system handles an interrupt.

   Answer: An *interrupt* is an event that requires the attention of the operating system. These events include the completion of an I/O, a key press, the alarm clock going off, division by zero, accessing a memory area that does not belong to the running program, and so on. Interrupts may be generated by hardware or software. A *trap* is an interrupt generated by software (*e.g.*, division by 0 and system call).

   When an interrupt occurs, the following steps will take place to handle the interrupt:

   - The executing program is suspended and control is transferred to the operating system. Mode switch may be needed.
   - A general routine in the operating system examines the received interrupt and calls the interrupt-specific handler.
   - After the interrupt is processed, a context switch transfers control back to a suspended process. Of course, mode switch may be needed.

   See pp. 6–7 of `02-Hardware-OS.pdf`. ∎

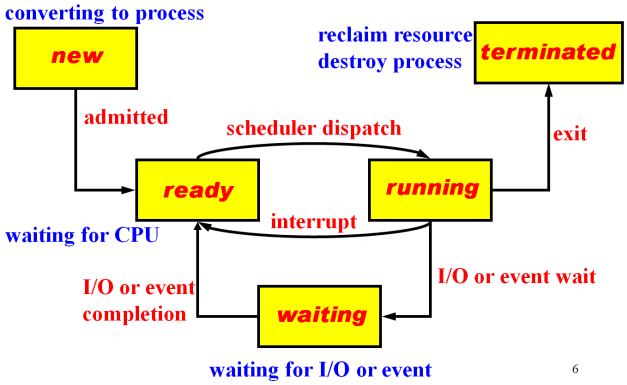
2. **Processes**

   (a) **[10 points]** Draw the state diagram of a process from its creation to termination, including all transitions. Make sure you will elaborate **every state** and **every transition** in the diagram.

   Answer: The following state diagram is taken from my class note.

   There are five states: **new**, **ready**, **running**, **waiting**, and **terminated**.

   - **New**: The process is being created.
   - **Ready**: The process has everything but the CPU, and is waiting to be assigned to a processor.
   - **Running**: The process is executing on a CPU.
   - **Waiting**: The process is waiting for some event to occur (*e.g.*, I/O completion or some resource).
   - **Terminated**: The process has finished execution.



The transitions between states are as follows:

- **New→Ready**: The process has been created and is ready to run.
- **Ready→Running**: The process is selected by the CPU scheduler and runs on a CPU/core.
- **Running→Ready**: An interrupt has occurred forcing the process to wait for the CPU.
- **Running→Waiting**: The process must wait for an event (*e.g.*, I/O completion or a resource).
- **Waiting→Ready**: The event the process is waiting has occurred, and the process is now ready for execution.
- **Running→Terminated**: The process exits.

See pp. 5–6 of `03-Process.pdf`. ∎

(b) **[10 points]** What is a *context*? Provide a detail description of *all* activities of a *context switch*.

Answer: A process needs some system resources (*e.g.*, memory and files) to run properly. These system resources and other information of a process include process ID, process state, registers, memory areas (for instructions, local and global variables, stack and so on), various tables (*e.g.*,, PCB), a program counter to indicate the next instruction to be executed, etc. They form the *environment* or *context* of a process. The steps of switching process *A* to process *B* are as follows:

- The operating system suspends *A*'s execution. A CPU mode switch may be needed.
- Transfer the control to the CPU scheduler.
- Save *A*'s context to its PCB and other tables.
- Load *B*'s context to register, etc. from *B*'s PCB.
- Resume *B*'s execution of the instruction at *B*'s program counter. A CPU mode switch may be needed.

See pp. 10–11 of `03-Process.pdf`. ∎

3. **Threads**

(a) **[10 points]** Enumerate the major differences between kernel-supported threads and user-level threads.

Answer: Kernel-supported threads are threads directly handled by the kernel. The kernel does thread creation, termination, joining, memory allocation, and scheduling in kernel space. User threads are supported at the user level and are not recognized by the kernel, and thread creation, termination, joining, memory allocation, and scheduling are done in the user space. Usually, a library running in user space provides all support. Due to the kernel involvement, the overhead of managing kernel-supported threads is higher than that of user threads.

Since there is no kernel intervention, user threads are more efficient than kernel threads. On the other hand, in a multiprocessor environment, the kernel may schedule kernel-supported threads to run on multiple processors, which is impossible for user threads because the kernel does not schedule user threads. Additionally, since the kernel does not recognize and schedule user threads, if the containing process or its associated kernel-supported thread is blocked, all user threads of that process (or kernel thread) are also blocked. However, blocking a kernel-supported thread will not cause all threads of the containing process to be blocked.

See pp. 5–6 and pp. 9–12 `04-Thread.pdf`. ∎

4. **Synchronization**

(a) **[10 points]** Define the meaning of a *race condition*? Answer the question first and use execution sequences with a clear and convincing argument to illustrate your answer. **You must explain step-by-step why your example causes a race condition.**

Answer: A *race condition* is a situation in which *more than one* processes or threads manipulate a shared resource *concurrently*, and the result depends on *the order of execution*.

The following is a simple counter updating example discussed in class. The value of `count` may be 9, 10 or 11, depending on the order of execution of the **machine instructions** of `count++` and `count--`.

```
int         count = 10;  // shared variable

Process 1                Process 2

count++;                 count--;
```

The following execution sequence shows a race condition. Two processes run concurrently (condition 1). Both processes access the shared variable `count` concurrently (condition 2) because `count` is accessed in an interleaved way. Finally, the computation result depends on the order of execution of the `SAVE` instructions (condition 3). The execution sequence below shows the result being 9; however, switching the two `SAVE` instructions yields 11. Since all conditions are met, we have a race condition. **Note that you have to provide TWO execution sequences, one for each possible result, to justify the existence of a race condition.**

`Thread_1`	`Thread_2`	*Comment*
do somthing	do somthing	`count = 10` initially
`LOAD count`		`Thread_1` executes `count++`
`ADD #1`		
	`LOAD count`	`Thread_2` executes `count--`
	`SUB #1`	
`SAVE count`		`count` is 11 in memory
	`SAVE count`	Now, `count` is 9 in memory

Stating that "`count++` followed by `count--`" or "`count--` followed by `count++`", even using machine instructions, produces different results and hence a race condition is **incomplete**, because the two processes do not access the shared variable `count` concurrently. Note that the use of higher-level language statement interleaved execution may not reveal the key concept of "sharing" as discussed in class. Therefore, use instruction level interleaved instead.

See pp. 5–12 of `05-Sync-Basics.pdf`. ∎

5. **Problem Solving:**

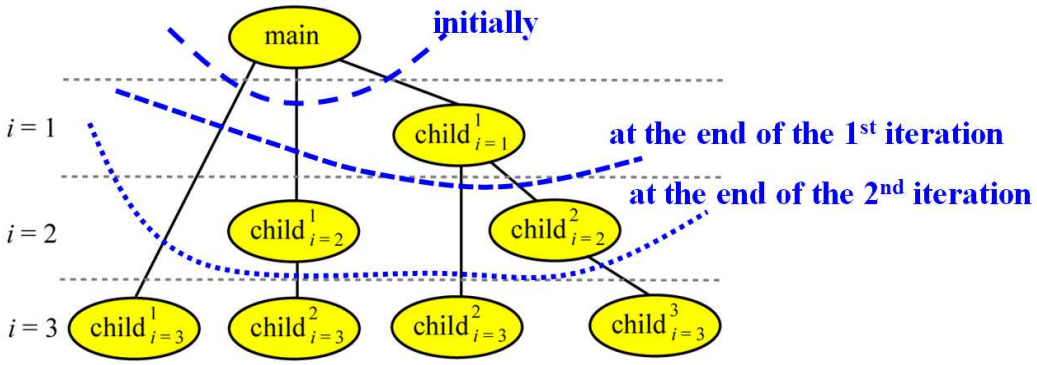
(a) **[10 points]** Consider the following program segment. Suppose all `fork()` calls are successful. Answer the following questions: **(1)** Draw a diagram showing the parent-child relationship of *all* involved processes, the main program included, and provide an explanation how this relationship is obtained. **Vague and not convincing arguments receive zero point. (2)** This program segment uses `n = 3`. How many processes will be created if `n` is set to a positive integer k? You don't need to draw a diagram; but, you still have to provide a justification for your answer. **Vague and not convincing argument receive zero point.**

```
int i, n = 3;

for (i = 1; i <= n; i++)
     fork();
```

Answer: The following has the answer to each question:

Question (1): Keep in mind that each `fork()` creates a child process and the child process receives a separate address space that is identical to of that of the parent. As a result, after a child process is created, the value of `i` of this child is the same as that of the parent. After the `fork()` call, the parent and its child run concurrently. Initially, only the parent process runs. See the diagram below.



The creation steps are as follows:

- **i = 1**: the `main` creates a child, $\text{child}^1_{i=1}$, where the superscript 1 indicates the "generation" and the subscript has the value of i. Now we have two processes running, `main` and $\text{child}^1_{i=1}$. Both of them go back to the beginning of the `for` loop and advance to $i = 2$.
- **i = 2**: Both `main` and $\text{child}^1_{i=1}$ reach the `fork()` call and both create a child. The `main` creates a child $\text{child}^1_{i=2}$, which is still the first generation with $i = 2$ as shown in the diagram. Process $\text{child}^1_{i=1}$ also executes `fork()` to creates a child, $\text{child}^2_{i=2}$, which is the second generation in the hierarchy with $i = 2$. Now, we have four processes running, `main`, $\text{child}^1_{i=1}$ (`main`'s first child created when $i = 1$), $\text{child}^1_{i=2}$ (`main`'s second child created when $i = 2$), and $\text{child}^2_{i=2}$ ($\text{child}^1_{i=1}$'s child created when $i = 2$).
- **i = 3**: Each of the above mentioned four processes executes the `fork()` call, and creates one child process. As a result, there are eight processes in total after the `for` loop completes as shown in the diagram.

Question (2): Since each process creates one and only one child process in each iteration, the number of processes is doubled. We have the `main` initially. At the end of $i = 1$, we have $2 = 2 \times 1 = 2^1$ processes; at the end of $i = 2$, we have $4 = 2^1 \times 2 = 2^2$ processes; and at the end of $i = 3$, we have $8 = 2^2 \times 2 = 2^3$ processes. Continuing this observation, at the end of $i = k$, we should have 2^k processes.

This observation can easily be proved with mathematical induction.

- **Base Case:** If $k = 0$, we have `main` only and hence $2^0 = 1$ process.
- **Induction Step:** Assume that the proposition holds for $k-1$. That is, we assume that at the end of $i = k-1$ there are 2^{k-1} processes. Since each process only creates one child process in each iteration, when $i = k$ each of the 2^{k-1} processes creates one child process, making the total number of processes $2 \times 2^{k-1} = 2^k$. Therefore, at the end of $i = k$ we have 2^k processes.

From the above induction proof, the number of processes at the end of iteration k is 2^k. ∎

(b) **[15 points]** Consider the following two processes, *A* and *B*, to be run concurrently using a shared memory for the `int` variable `x`.

```
Process A                      Process B
---------                      ---------
for (i = 1; i <= 2; i++)       x = 2*x;
   x++;
```

Assume that `x` is initialized to 0, and `x` must be loaded into a register before further computations can take place. What are **all possible** values of `x` after both processes have terminated. **You must use clear step-by-step execution sequences of the above processes with a convincing argument. Any vague and unconvincing argument receives no points.**

Answer: Obviously, the answer must be in the range of 0 and 4. It is non-negative, because the initial value is 0 and no subtraction is used. It cannot be larger than 4, because the two `x++` statements and `x = 2*x` together can at most double the value of `x` twice.

The easiest answers are 2, 3 and 4 if `x = 2*x` executes before, between and after the two `x++` statements, respectively. The following shows the possible execution sequences.

`x = 2*x` **is before both** `x++`		
Process 1	Process 2	`x` in memory
	`x = 2*x`	0
`x++`		1
`x++`		2

`x = 2*x` **is between the two** `x++`		
Process 1	Process 2	`x` in memory
`x++`		1
	`x = 2*x`	2
`x++`		3

`x = 2*x` **is after both** `x++`		
Process 1	Process 2	`x` in memory
`x++`		1
`x++`		2
	`x = 2*x`	4

The situation is a bit more complex with instruction interleaving. Process B's `x = 2*x` may be translated to the following machine instructions:

```
LOAD x
MUL  #2
SAVE x
```

Because the `LOAD` retrieves the value of `x`, and the `SAVE` may change the current value of `x`, the results depend on the positions of `LOAD` and `SAVE`. The following shows the result being 0. In this case, `LOAD` loads 0 *before* both `x++` statements, and the result 0 is saved *after* both `x++` statements.

Process 1	Process 2	`x` in memory	Comments
	`LOAD x`	0	Load $x = 2$ into register
	`MUL #2`	0	Process 2's register is 0
`x++`		1	Process 1 adds 1 to x
`x++`		2	Process 1 adds 1 to x
	`SAVE x`	0	Process 2 saves 0 to x

If the `SAVE` executes between the two `x++` statements, the result is 1.

Process 1	Process 2	`x` in memory	Comments
	`LOAD x`	0	Load $x = 2$ into register
	`MUL #2`	0	Process 2's register is 0
`x++`		1	Process 1 adds 1 to x
	`SAVE x`	0	Process 2 saves 0 to x
`x++`		1	Process 1 adds 1 to x

You may try other instruction interleaving possibilities (*e.g.*, replacing the `x++` with machine instructions) and the answers should still be in the range of 0 and 4. ■

(c) **[15 points]** The following solution to the critical section problem has a global array `idea[2]` and an `int` variable `switch`.

```
 1. bool  idea[2];   // global flags
 2. int   switch;    // global variable, initially 0 or 1

    Process 0                Process 1
    ---------                ---------
 3. idea[0] = TRUE;          idea[1] = TRUE;         // I am interested
 4. while (idea[1]) {        while (idea[0]) {       // wait if you are interested
 5.   if (switch == 1) {      if (switch == 0) {    //   my switch?
 6.     idea[0] = FALSE;        idea[1] = FALSE;    //     I am no more interested
 7.     while (switch != 0)     while (switch != 1) //     wait for my switch
 8.       ;                       ;
 9.     idea[0] = TRUE;         idea[1] = TRUE;     //   let me try again
10.   }                       }
11. }                       }
                // in critical section
12. switch = 1;             switch = 0;             // it is your turn now
13. idea[0] = FALSE;        idea[1] = FALSE;        // I am not interested
```

Prove rigorously that this solution satisfies the mutual exclusion condition. *You will receive* **zero** *point if* **(1)** *you prove by example, or* **(2)** *your proof is vague and/or unconvincing.*

Answer: This is a very simple problem and an exercise in a weekly reading list. In fact, it is an extension of Attempt II of `06-Sync-Soft-Hardware.pdf` (pp. 6–10) and its proof follows exactly the same logic.

For P_0 to enter its critical section, it sets `idea[0]` to `TRUE`. Once P_0 reaches its `while` loop, we have the following cases to consider:

- If P_0 is lucky and sees `idea[1]` being `FALSE` the first time, then P_0 enters its critical section. In this case, `idea[0]` and `idea[1]` are `TRUE` and `FALSE`, respectively.
- If P_0 enters the `while` loop because `idea[1]` is `TRUE`, then we have two cases:
  - If P_0 finds out `switch` being not 1, then P_0 goes back for the next iteration without changing the value of `idea[0]`.
  - If P_0 finds out `switch` being 1, then P_0 executes the "then" part of the `if` statement. Before it leaves the "then" part, P_0 sets `idea[0]` to `TRUE` again.

  Hence, no matter what has happened to the `if` statement, at the end of the `while` loop, `idea[0]` is always `TRUE`. If P_0 eventually enters its critical section, `idea[0]` is always `TRUE` and `idea[1]` is always `FALSE`.

In summary, if P_0 is in its critical section, we have `idea[0] = TRUE` and `idea[1] = FALSE`. Similarly, if P_1 is in its critical section, we have `idea[1] = TRUE` and `idea[0] = FALSE`.

Now, if P_0 and P_1 are *both* in their critical sections, `idea[0]` is set to `TRUE` by P_0 and sees `idea[1]` being `FALSE` from its `while` loop. By the same reason, P_1 sets `idea[1]` to `TRUE` and sees `idea[1]` being `FALSE` from its `while` loop. As a result, `idea[0]` (and `idea[1]`) must be both `TRUE` and `FALSE` at the same time. Because a variable can only hold one value, we have a contradiction. Consequently, the mutual exclusion condition is satisfied.

Note that no matter what the value of `switch` is, a process can enter its critical section only if it can pass its `while` loop. Therefore, the control variables are `idea[0]` and `idea[1]` rather than `switch`. You should read the code carefully. ∎