Slide 1

Administrivia

- Remember that Homework 1 is due by 5pm today.

Slide 2

Minute Essay From Last Lecture

- Question: In a system with 8 CPUs and 100 processes, what's the maximum number of processes that can be running? ready? blocked?
- Answer:
  - 8 processes can be running (assuming there are 8 that are runnable).
  - 92 processes can be ready (if there are more, some will be running).
  - 100 processes can be blocked.

CSCI 4320 September 16, 2003

Slide 3

Recap — Processes

- Process abstraction — "program running on virtual CPU" (virtual program counter, virtual registers, etc.).
- Apparent concurrency (in almost all respects identical to real concurrency) provided by interleaving / context switches.
- Context switch — switch between virtual CPUs, triggered by interrupts (I/O, error, system call, timer).
- Process can also be a way of grouping together other resources needed by a running program, e.g., "address space", list of open files.

  These resources may form part of the "context" that must be saved / restored on a context switch.

Slide 4

Recap — Process States

- Three basic states for processes – running, ready, blocked.
- Some transitions are obvious, others require decision-making (ready to running); for now, assume existence of "scheduler" to make decisions.

Slide 5

Recap — Threads

- Processes versus threads:
  - Process implements "program on virtual CPU" abstraction, has its own group of resources.
  - Thread implements "program on virtual CPU" abstraction, shares group of resources with (some) other threads.
- Threads are in a way "processes within processes".
- Compare context switching between processes with context switching between threads within process.
- Two basic approaches to implementing threads — "in user space" and "in kernel space".

Slide 6

Interprocess Communication

- Processes almost always need to interact with other processes:
  - "Ordering constraints" – e.g., process B uses as input some data produced by process A.
  - Use of shared resources — files, shared memory locations, etc.
- Use of shared resources can lead to "race conditions" — output depends on details of interleaving.
- Processes must communicate to avoid race conditions and otherwise synchronize.

CSCI 4320 September 16, 2003

Slide 7

Mutual Exclusion Problem

- In many situations, we want only one process at a time to have access to some shared resource.
- Generic/abstract version — multiple processes, each with a "critical region" ("critical section"):

```
while (true) {
    do_cr();        // must be "finite"
    do_non_cr();    // need not be "finite"
}
```

- Goal is to add something to this code such that:
  1. No more than one process at a time can be "in its critical region".
  2. No process not in its critical region can block another process.
  3. No process waits forever to enter its critical region.
  4. No assumptions are made about how many CPUs, their speeds.

Slide 8

Mutual Exclusion Problem, Continued

- We'll look at various solutions (some correct, some not):
  - Using only hardware features always present (some notion of shared variable).
  - Using optional hardware features.
  - Using "synchronization primitives" (abstractions that help solve this and other problems).
- Recall that a correct solution
  - Must work for more than 1 CPU.
  - Must work even in the face of unpredictable context switches — whatever we're doing, another process can pull the rug out from under us between "atomic operations" (machine instructions).

Slide 9

Atomic Operations

- Which of the following are "atomic"?
  - `x = 1;`
  - `x = x + 1;`
  - `++x;`
  - `if (x = 0) x = 1;`

Slide 10

Proposed Solution — Disable Interrupts

- Pseudocode for each process:

```
while (true) {
    disable_interrupts();
    do_cr();
    enable_interrupts();
    do_non_cr();
}
```

- Does it work? reviewing the criteria . . .

CSCI 4320 September 16, 2003

Slide 11

Proposed Solution — Simple Lock Variable

- Shared variables:

```
int lock = 0;
```

Pseudocode for each process:

```
while (true) {
    while (lock != 0);
    lock = 1;
    do_cr();
    lock = 0;
    do_non_cr();
}
```

- Does it work? reviewing the criteria . . . No.

Slide 12

Proposed Solution — Strict Alternation

- Shared variables:

```
int turn = 0;
```

Pseudocode for process p0:

```
while (true) {
    while (turn != 0);
    do_cr();
    turn = 1;
    do_non_cr();
}
```

Pseudocode for process p1:

```
while (true) {
    while (turn != 1);
    do_cr();
    turn = 0;
    do_non_cr();
}
```

- Does it work? reviewing the criteria . . . No.

Slide 13

Proposed Solution — Peterson's Algorithm

- Shared variables:

```
int turn = 0;   // "who tried most recently"
bool interested0 = false, interested1 = false;
```

Pseudocode for process p0:

```
while (true) {
    interested0 = true;
    turn = 0;
    while ((turn == 0)
        && interested1);
    do_cr();
    interested0 = false;
    do_non_cr();
}
```

Pseudocode for process p1:

```
while (true) {
    interested1 = true;
    turn = 1;
    while ((turn == 1)
        && interested0);
    do_cr();
    interested1 = false;
    do_non_cr();
}
```

- Does it work? reviewing the criteria . . . Yes.

Slide 14

Proposed Solution — TSL Instruction

- A key problem in concurrent algorithms is the idea of "atomicity" (operations guaranteed to execute without interference from another CPU/process). Hardware can provide some help with this.
- E.g., "test and set lock" (TSL) instruction:

  `TSL registerX, lockVar`

  (1) copies `lockVar` to `registerX` and (2) sets `lockVar` to non-zero, all as one atomic operation.

  How to make this work is the hardware designers' problem!

CSCI 4320 September 16, 2003

Slide 15

Proposed Solution — TSL Instruction, Continued

- Shared variables:

```
int lock = 0;
```

Pseudocode for each process:

```
while (true) {
    enter_cr();
    do_cr();
    leave_cr();
    do_non_cr();
}
```

Assembly-language routines:

```
enter_cr:
    TSL regX, lock
    compare regX with 0
    if not equal
        jump to enter_cr
    return
leave_cr:
    store 0 in lock
    return
```

- Does it work? reviewing the criteria . . . Yes.

Slide 16

Mutual Exclusion Solutions So Far

- Solutions so far have some problems: inefficient, dependent on whether scheduler/etc. guarantees fairness.
- Also, they're very low-level, so might be hard to use for more complicated problems.
- So, people have proposed various "synchronization mechanisms" . . .

Minute Essay

- Do you see why the various solutions to the mutual exclusion problem so far work / don't work?
- Give an example (other than those discussed) of a situation in which you think a solution to this problem would be needed.