Parallel Programming: Background Information and Tips

Three Reasons to Study Parallel Programming

1. Increase performance: do more work in the same amount of time
2. Increase performance: take less time to do the same amount of work
3. Make some programming tasks more convenient to implement

Example:
- Decrease the time to compute a simulation
- Create a web browser where the tasks of monitoring the user interface, downloading text, and downloading multiple images are happening simultaneously
- Increase the resolution, and thus the accuracy, of a simulation

Three Types of Parallelism:

1. Instruction Level Parallelism (ILP)
   A program might consist of a continuous stream of assembly instructions, but it is not necessarily executed continuously. Sometimes it has "pauses," waiting for something to be ready so that it can proceed.

   \[
   A = B + 1; \quad C = 3;
   \]

   Load \( B, r0 \)
   \[
   \text{Add } 1, r0 \]
   \[
   \text{Store } r0, A
   \]

   Out-of-order execution capability will slide instructions up if they can be executed while waiting for the block to end

   If \( B \) is not already in cache, this instruction will block while \( B \) is fetched from memory

   Prefetch \( B \)

   Load \( r3, r1 \)

   Store \( r1, C \)

   If a compiler does this, it's called Static ILP
   If the CPU chip does this, it's called Dynamic ILP

   This is all good to know, but it's nothing we can control much of

2. Data Level Parallelism (DLP)
   Executing the same instructions on different parts of the data

   \[
   \text{for} (i = 0; i < \text{NUM}; i++)
   \]
   \[
   B[i] = \sqrt{A[i]};
   \]

3. Thread Level Parallelism (TLP)
   Executing different instructions

   Example: processing a variety of incoming transaction requests

   In general, TLP implies that you have more threads than cores

   Thread execution switches when a thread blocks or uses up its time slice

Flynn’s Taxonomy
### Von Neumann Architecture

**Basically the fundamental pieces of a CPU have not changed since the 1960s**

- Memory
- Arithmetic Logic Unit
- Control Unit
- Accumulator

Other elements:
- Clock
- Registers
- Program Counter
- Stack Pointer

These together are the 'state' of the processor

### What Exactly is a Process?

Processes execute a program in memory. The process keeps a state (program counter, registers, and stack).

- Program and Data in Memory (the heap is here too)
- Registers
- Program Counter
- Stack Pointer

### What Exactly is a Thread?

Threads are separate independent processes, all executing a common program and sharing memory. Each thread has its own state (program counter, registers, and stack pointer).

- Program and Data in Shared Memory (the heap is shared too)
- Registers
- Program Counter
- Stack Pointer

### Memory Allocation in a Multithreaded Program

- One-thread
  - Stack
  - Program Executable
  - Global
  - Heap

- Multiple-threads
  - Stack
  - Common Program Executable
  - Common Global
  - Common Heap

Don't take this completely literally. The exact arrangement depends on the operating system and the compiler. For example, sometimes the stack and heap are arranged so that they grow towards each other.

### What Exactly is a Thread?

A 'thread' is an independent path through the program code. Each thread has its own Program Counter, Registers, and Stack Pointer. But, since each thread is executing some part of the same program, each thread has access to the same global data in memory. Each thread is scheduled and swapped just like any other process.

Threads can share time on a single processor. You don’t have to have multiple processors (although you can – the multicore topic is coming soon!).

This is useful, for example, in a web browser when you want several things to happen autonomously:

- User interface
- Communication with an external web server
- Web page display
- Image loading
- Animation
When is it Good to use Multithreading?

- Where specific operations can become blocked, waiting for something else to happen
- Where specific operations can be CPU-intensive
- Where specific operations must respond to asynchronous I/O, including the user interface (UI)
- Where specific operations have higher or lower priority than other operations
- To manage independent behaviors in interactive simulations
- When you want to accelerate a single program on multicore CPU chips

Threads can make it easier to have many things going on in your program at one time, and can absorb the dead-time of other threads.

Two Ways to Decompose your Problem into Parallelizable Pieces

**Functional (or Task) Decomposition**

Breaking a task into sub-tasks that represent separate functions. A web browser is a good example. So is a climate modeling program:

```
Land
Ocean
Air
```

**Domain (or Data) Decomposition**

Breaking a task into sub-tasks that represent separate sections of the data. An example is a large diagonally-dominant matrix solution:

```
[ 0 1 0 ]
[ 0 2 0 ]
[ 0 0 3 ]
```

Data Decomposition Reduces the Problem Size per Thread

Example: A diagonally-dominant matrix solution

- Break the problem into blocks
- Solve within the block
- Handle borders separately after a Barrier

```
[ 0 1 ] * [ 2 ] = [ 0 ]
[ 0 2 ] * [ 1 ] = [ 0 ]
[ 0 0 ] * [ 2 ] = [ 0 ]
```

Share results across boundaries

Some Definitions

- **Atomic**: An operation that takes place to completion with no chance of being interrupted by another thread
- **Barrier**: A point in the program where all threads must reach before any of them are allowed to proceed
- **Coarse-grained parallelism**: Breaking a task up into a small number of large tasks
- **Deterministic**: The same set of inputs always gives the same outputs
- **Dynamic scheduling**: Dividing the total number of tasks \( T \) up so that each of \( N \) available threads has less than \( T/N \) sub-tasks to do, and then doling out the remaining tasks to threads as they become available
- **Fine-grained parallelism**: Breaking a task up into lots of small tasks
- **Fork-join**: An operation where multiple threads are created from a main thread. All of those forked threads are expected to eventually finish and thus "join back up" with the main thread.

Parallel Programming Tips

- **Private variable**: After a fork operation, a variable which has a private copy within each thread
- **Reduction**: Combining the results from multiple threads into a single sum or product, continuing to use multithreading. Typically this is performed so that it takes \( O(\log N) \) time instead of \( O(N) \) time:
- **Shared variable**: After a fork operation, a variable which is shared among threads, i.e., has a single value
- **Speed-up** \( (T_1 / T_N) \)
- **Speed-up Efficiency** \( Speed-up(N) / N \)
- **Static Scheduling**: Dividing the total number of tasks \( T \) up so that each of \( N \) available threads has \( T/N \) sub-tasks to do.
If you do keep internal state between calls, there is a chance that a second thread will hop in and change it, then the first thread will use that state thinking it has not been changed.

Ironically, some of the standard C functions that we use all the time (e.g., `strtok`) keep internal state:

```c
char * strtok ( char * str, const char * delims );
```

**Tip #1 -- Don't Keep Internal State**

```c
int GetLastPositiveNumber( int x )
{
    static int savedX;
    if( x >= 0 )
        savedX = x;
    return savedX;
}
```

1. Thread #1 sets the internal character array pointer to somewhere in Line1[ ].
2. Thread #2 resets the same internal character array pointer to somewhere in Line2[ ].
3. Thread #1 uses that internal character array pointer, but it is not pointing into Line1[ ] where Thread #1 thinks it left it.

**Execution order no longer matters!**

**Tip #2 – Avoid Deadlock**

Deadlock is when two threads are each waiting for the other to do something.

```
if( x >= 0 )
    savedX = x;
return savedX;
```

```
if( x >= 0 )
    savedX = x;
return savedX;
```

```
if( x >= 0 )
    savedX = x;
return savedX;
```

```
if( x >= 0 )
    savedX = x;
return savedX;
```

**Worst of all, the way these problems occur is not always deterministic!**
A Race Condition is where it matters which thread gets to a particular piece of code first. This often comes about when one thread is modifying a variable while the other thread is in the midst of using it. A good example is maintaining and using the pointer in a stack data structure:

Thread #1: Pushing:
```
p++;
*p = incoming;
p++;
```

Thread #2: Popping:
```
outgoing = *p;
p--;
```

Execution order:

Worst of all, the way these problems occur is not always deterministic!

**Tip #3 – Avoid Race Conditions**

BTW, Race Conditions can often be fixed through the use of Mutual Exclusion Locks (Mutexes):

```
Thread #1: Pushing:
MutexLock A
```

```
Thread #2: Popping:
MutexLock A
```

Execution order:

While solving a race condition, we can accidentally create a deadlock condition if the thread that owns the lock is waiting for the other thread to do something. Mutex Locks are usually named somehow so that you can have multiple ones with no ambiguity.

**Tip #4 -- Sending a Message to the Optimizer:**

The **volatile** Keyword

The **volatile** keyword is used to let the compiler know that another thread might be changing a variable “in the background”, so don’t make any assumptions about what can be optimized away.

```
int val = 0;
...
while( val != 0 ) ;
```

A good compiler optimizer will eliminate this code because it “knows” that, for all time, val == 0

```
volatile int val = 0;
...
while( val != 0 ) ;
```

The **volatile** keyword tells the compiler optimizer that it cannot count on val being == 0 here.

**Tip #5 -- Sending a Message to the Optimizer:**

The **restrict** Keyword

Remember our Instruction Level Parallelism example?

```
A = B + 1;
C = 3;
```

```
Load B, r0
Add $1, r0
Store r0, A
```

```
Load $3, r1
Store r1, C
```

Prefetch B

```
Load $3, r1
Store r1, C
```

Load B, r0
Add $1, r0
Store r0, A

What’s really happening

Uh-oh! B is being loaded at the same time it is being stored into. Who gets there first? Which value is correct?

```
int * restrict p;
int * restrict q;
...
p = &B;
q = &C;
A = *p + 1;
*t = 3;
```

Here the example has been changed slightly. This is what worries the out-of-order mechanisms and keeps them from optimizing as much as they could.

```
Assembly language
```

```
Using the pointers, and using out-of-order processing
```

```
Sending a Message to the Optimizer:
The restrict Keyword
```

```
What’s really happening
```

```
int * restrict p;
int * restrict q;
...
p = &B;
q = &C;
A = "p" + 1;
t = 3;
```

This is us promising that p and q will never point to the same memory location.

Now there is no conflict
Tip #6 – Beware of False Sharing Caching Issues

We will get to this in the Caching notes!