OpenMP Multithreaded Programming

- OpenMP stands for “Open Multi-Processor”
- OpenMP is a multi-vendor (see next page) standard to perform shared-memory multithreading
- OpenMP is both compiler-directive- and library-based
- OpenMP threads share a single executable, a single global memory, and a single heap (malloc, new)
- Each OpenMP thread has its own stack (function arguments, function return address, local variables)
- Using OpenMP usually requires no dramatic code changes
- OpenMP probably gives you the biggest multithread benefit per amount of work you have to put in to using it

Much of your use of OpenMP will be accomplished by issuing C/C++ “pragmas” to tell the compiler how to build the threads into your executable, like this:

```
#pragma omp directive [clause]
```
Who is in the OpenMP Consortium?

- AMD
- ARM
- CRAY
- FUJITSU
- Hewlett Packard Enterprise
- IBM
- Intel
- Micron
- NEC
- NVIDIA
- ORACLE
- Red Hat
- National Institute of Standards and Technology (NIST)
- Argonne National Laboratory
- EPCC
- INRIA
- Los Alamos National Laboratory
- NASA
- University of Houston
- University of Illinois at Urbana-Champaign (UIUC)
- Texas Advanced Computing Center (TACC)
- National Renewable Energy Laboratory (NREL)
- European Centre for Medium-Range Weather Forecasts (ECMWF)
- United Kingdom's Science and Technology Facilities Council (STFC)

What OpenMP Isn't:

- OpenMP doesn't check for data dependencies, data conflicts, deadlocks, or race conditions. You are responsible for avoiding those yourself.
- OpenMP doesn't check for non-conforming code sequences (we'll talk about what this means).
- OpenMP doesn't guarantee identical behavior across vendors or hardware, or even between multiple runs on the same vendor's hardware.
- OpenMP doesn't guarantee the order in which threads execute, just that they do execute.
- OpenMP is not overhead-free.
- OpenMP does not prevent you from writing code that triggers cache performance problems (such as in false-sharing), in fact, it makes it really easy.

We will get to “false sharing” in the cache notes.
Memory Allocation in a Multithreaded Program

One-thread

Multiple-threads

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.

Using OpenMP on Linux

```bash
g++ -o proj proj.cpp -lm -fopenmp
```

Using OpenMP in Microsoft Visual Studio

1. Go to the Project menu → Project Properties

2. Change the setting Configuration Properties → C/C++ → Language → OpenMP Support to "Yes (/openmp)"

If you are using Visual Studio and get a compile message that looks like this:

```
1>c1xx: error C2338: two-phase name lookup is not supported for C++/CLI, C++/CX, or OpenMP; use /Zc:twoPhase-
```

then do this:

1. Go to "Project Properties" → "C/C++" → "Command Line"
2. Add `/Zc:twoPhase-` in "Additional Options" in the bottom section
3. Press OK
### Seeing if OpenMP is Supported on Your System

```c
#ifdef _OPENMP
    fprintf(stderr, "OpenMP release %d is supported here\n", _OPENMP);
#else
    fprintf(stderr, "OpenMP is not supported here – sorry!\n");
    return 1;
#endif
```

Printing `_OPENMP` gives you a year and month of the OpenMP release that you are using.

To get the OpenMP version number from the year and month, check here:

<table>
<thead>
<tr>
<th>OpenMP Version</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>November 2018</td>
</tr>
<tr>
<td>4.5</td>
<td>November 2015</td>
</tr>
<tr>
<td>4.0</td>
<td>July 2013</td>
</tr>
<tr>
<td>3.1</td>
<td>July 2011</td>
</tr>
<tr>
<td>2.0</td>
<td>March 2002</td>
</tr>
<tr>
<td>1.0</td>
<td>October 1998</td>
</tr>
</tbody>
</table>

- By default, flip uses g++ 11.4, which uses OpenMP version 4.5
- Visual Studio 2022 uses OpenMP 2.0

### Numbers of OpenMP threads

#### How to specify how many OpenMP threads you want to have available:

```c
omp_set_num_threads( num );
```

#### Asking how many cores this program has access to:

```c
num = omp_get_num_procs( );
```

Actually returns the number of hyperthreads, not the number of physical cores.

#### Setting the number of available threads to the exact number of cores available:

```c
omp_set_num_threads( omp_get_num_procs( ) );
```

#### Asking how many OpenMP threads this program is using right now:

```c
num = omp_get_num_threads( );
```

#### Asking which thread number this one is:

```c
me = omp_get_thread_num( );
```
Creating an OpenMP Team of Threads

```c
#pragma omp parallel default(none)
{
    . . .
}
```

This creates a team of threads. Each thread then executes all lines of code in this block.

Think of it this way:

```c
#pragma omp parallel default(none)
```

The OpenMP Thread Team Prints a Friendly Message

```c
#include <stdio.h>
#include <omp.h>

int main()
{
    omp_set_num_threads(8);
    #pragma omp parallel default(none)
    {
        printf("Hello, World, from thread %d !\n", omp_get_thread_num( ) );
    }
    return 0;
}
```

Hint: run it several times in a row. What do you see? Why?
Uh-oh...

<table>
<thead>
<tr>
<th>First Run</th>
<th>Second Run</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hello, World, from thread #6 !</td>
<td>Hello, World, from thread #0 !</td>
</tr>
<tr>
<td>Hello, World, from thread #1 !</td>
<td>Hello, World, from thread #7 !</td>
</tr>
<tr>
<td>Hello, World, from thread #7 !</td>
<td>Hello, World, from thread #4 !</td>
</tr>
<tr>
<td>Hello, World, from thread #5 !</td>
<td>Hello, World, from thread #6 !</td>
</tr>
<tr>
<td>Hello, World, from thread #4 !</td>
<td>Hello, World, from thread #1 !</td>
</tr>
<tr>
<td>Hello, World, from thread #3 !</td>
<td>Hello, World, from thread #3 !</td>
</tr>
<tr>
<td>Hello, World, from thread #2 !</td>
<td>Hello, World, from thread #5 !</td>
</tr>
<tr>
<td>Hello, World, from thread #0 !</td>
<td>Hello, World, from thread #2 !</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Third Run</th>
<th>Fourth Run</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hello, World, from thread #2 !</td>
<td>Hello, World, from thread #1 !</td>
</tr>
<tr>
<td>Hello, World, from thread #5 !</td>
<td>Hello, World, from thread #3 !</td>
</tr>
<tr>
<td>Hello, World, from thread #0 !</td>
<td>Hello, World, from thread #7 !</td>
</tr>
<tr>
<td>Hello, World, from thread #4 !</td>
<td>Hello, World, from thread #5 !</td>
</tr>
<tr>
<td>Hello, World, from thread #7 !</td>
<td>Hello, World, from thread #2 !</td>
</tr>
<tr>
<td>Hello, World, from thread #1 !</td>
<td>Hello, World, from thread #4 !</td>
</tr>
<tr>
<td>Hello, World, from thread #3 !</td>
<td>Hello, World, from thread #7 !</td>
</tr>
<tr>
<td>Hello, World, from thread #4 !</td>
<td>Hello, World, from thread #6 !</td>
</tr>
<tr>
<td>Hello, World, from thread #6 !</td>
<td>Hello, World, from thread #0 !</td>
</tr>
</tbody>
</table>

There is no guarantee of thread execution order!

---

Creating OpenMP Threads to Process Loop Passes

```c
#include <omp.h>

omp_set_num_threads( NUMT );

#pragma omp parallel for default(none) for( int i = 0; i < arraySize; i++ )
{
    // Code here
}
```

The code starts out executing in a single thread.

This sets how many threads will be created. It doesn't create them yet; it just says how many will be used the next time you ask for them.

Here we ask for them. This creates a team of threads and divides the for-loop passes up among those threads.

There is an “implied barrier” at the end where each thread waits until all threads are done, then the code continues in a single thread.

This tells the compiler to parallelize the for-loop into multiple threads. Each thread automatically gets its own personal copy of the variable `i` because it is defined within the for-loop body.

The `default(none)` directive forces you to explicitly declare all variables declared outside the parallel region to be either private or shared while they are in the parallel region. Variables declared within the for-loop are automatically private.
OpenMP for-Loop Rules

```c
#pragma omp parallel for default(none), shared(...), private(…)
for( int index = start ; index terminate condition; index changed )
```

- The index must be an int or a pointer
- The start and terminate conditions must have compatible types
- Neither the start nor the terminate conditions can be changed during the execution of the loop
- The index can only be modified by the changed expression (i.e., not modified inside the loop itself)
- You cannot use a break or a goto to get out of the loop
- There can be no inter-loop data dependencies such as:
  ```c
  a[ i ] = a[ i-1 ] + 1.;
  ```

```
// what if this is the last line of thread #0's work?

```c
a[101] = a[100] + 1.;
```

```
// what if this is the first line of thread #1's work?

```

OpenMP For-Loop Rules

```c
for( index = start ;
  index < end
  index <= end
  index > end
  index >= end
  ;
  
  index++
  ++index
  index--
  --index
  index += incr
  index = index + incr
  index = incr + index
  index -= decr
  index = index - decr
)```
What to do about Variables Declared Before the for-loop Starts?

- **private(x)**
  - Means that each thread will get its own version of the variable
  - Example:
    ```
    float x = 0.;
    #pragma omp parallel for default(none), private(x)
    for( int i = 0; i < N; i++ )
    {
      x = (float) i;
      float y = x*x;
      << more code... >
    }
    ```
  - i and y are automatically private because they are defined within the loop.
  - Good practice demands that x be explicitly declared to be shared or private!

- **shared(x)**
  - Means that all threads will share a common version of the variable

- **default(none)**
  - I recommend that you include this in your OpenMP for-loop directive. This will force you to explicitly flag all of your externally-declared variables as shared or private. Don't make a mistake by leaving it up to the default!

---

For-loop "Fission"

Because of the loop dependency, this whole thing is not parallelizable:

```
float x[ 0 ] = 0.;
y[ 0 ] *= 2.;
for( int i = 1; i < N; i++ )
{
  x[ i ] = x[ i-1 ] + 1.;
  y[ i ] *= 2.;
}
```

But it can be broken into one loop that is not parallelizable, plus one that is:

```
x[ 0 ] = 0.;
for( int i = 1; i < N; i++ )
{
  x[ i ] = x[ i-1 ] + 1.;
}
```

```
#pragma omp parallel for shared(y)
for( int i = 0; i < N; i++ )
{
  y[ i ] *= 2.;
}
```
For-loop “Collapsing”

Uh-oh, which for-loop do you put the #pragma on?

```c
for( int i = 1; i < N; i++ )
{
    for( int j = 0; j < M; j++ )
    {
        ...
    }
}
```

Ah-ha – trick question. You put it on both!

```c
#pragma omp parallel for collapse(2)
for( int i = 1; i < N; i++ )
{
    for( int j = 0; j < M; j++ )
    {
        ...
    }
}
```

How many for-loops to collapse into one loop

Single Program Multiple Data (SPMD) in OpenMP

```c
#define NUM 1000000
float A[NUM], B[NUM], C[NUM];
...
int total = omp_get_num_threads( );
#pragma omp parallel default(none),shared(total)
{
    int me = omp_get_thread_num( );
    DoWork( me, total );
}
```

```c
void DoWork( int m, int t )
{
    int first = NUM * m / t;
    int last = NUM * (m+1)/t - 1;
    for( int i = first; i <= last; i++ )
    {
        C[ i ] = A[ i ] * B[ i ];
    }
}
**OpenMP Allocation of Work to Threads**

**Static Threads**
- All work is allocated and assigned at runtime

**Dynamic Threads**
- The pool is statically assigned some of the work at runtime, but not all of it
- When a thread from the pool becomes idle, it gets a new assignment
- “Round-robin assignments”

**OpenMP Scheduling**

```c
schedule(static [,chunksize])
schedule(dynamic [,chunksize])
```

Defaults to static
chunksize defaults to 1

---

**OpenMP Allocation of Work to Threads**

```
#pragma omp parallel for default(none),schedule(static,chunksize)
for ( int index = 0 ; index < 12 ; index++ )
```

<table>
<thead>
<tr>
<th>Chunksize</th>
<th>Thread 0</th>
<th>Thread 1</th>
<th>Thread 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0,3,6,9</td>
<td>1,4,7,10</td>
<td>2,5,8,11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Each thread is assigned one iteration, then</td>
<td>the assignments start over</td>
</tr>
<tr>
<td>2</td>
<td>0,1,6,7</td>
<td>2,3,8,9</td>
<td>4,5,10,11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Each thread is assigned two iterations, then</td>
<td>the assignments start over</td>
</tr>
<tr>
<td>4</td>
<td>0,1,2,3</td>
<td>4,5,6,7</td>
<td>8,9,10,11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Each thread is assigned four iterations, then</td>
<td>the assignments start over</td>
</tr>
</tbody>
</table>

Think of dealing for-loop passes to threads the same way as dogs deal cards.
Arithmetic Operations Among Threads – A Problem

```c
float sum = 0.;
#pragma omp parallel for default(none), shared(sum)
for( int i = 0; i < N; i++ )
{
    float myPartialSum = ...
    sum = sum + myPartialSum;
}
```

• There is no guarantee when each thread will execute this line

• There is not even a guarantee that each thread will finish this line before some other thread interrupts it. (Remember that each line of code usually generates multiple lines of assembly.)

• This is non-deterministic!

Assembly code:

<table>
<thead>
<tr>
<th>Load sum</th>
<th>Add myPartialSum</th>
<th>Store sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>sum</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What if the scheduler decides to switch threads right here?

Conclusion: Don’t do it this way!

Here’s a trapezoid integration example.
The partial sums are added up, as shown on the previous page.
The integration was done 30 times.
The answer is supposed to be exactly 2.
None of the 30 answers is even close.
And not only are the answers bad, but they are not even consistently bad!

Don’t do it this way! We’ll talk about how to do it correctly in the Trapezoid Integration notestet.
Mutual Exclusion Locks (Mutexes)

- `omp_init_lock(omp_lock_t *);`
- `omp_set_lock(omp_lock_t *);`
- `omp_unset_lock(omp_lock_t *);`
- `omp_test_lock(omp_lock_t *);`

(omp_lock_t is really an array of 4 unsigned chars)

Critical sections

- `#pragma omp critical`
  - Restricts execution to one thread at a time

- `#pragma omp single`
  - Restricts execution to a single thread ever

Barriers

- `#pragma omp barrier`
  - Forces each thread to wait here until all threads arrive

(Note: there is an implied barrier after parallel for loops and OpenMP sections, unless the `nowait` clause is used)

Synchronization Example

```c
omp_lock_t Sync;

omp_init_lock(&Sync);

omp_set_lock(&Sync);
<< code that needs the mutual exclusion >>
omp_unset_lock(&Sync);

Thread #0:
omp_set_lock(&Sync);
<< code that needs the mutual exclusion >>
omp_unset_lock(&Sync);

Thread #1:
omp_set_lock(&Sync);
<< code that needs the mutual exclusion >>
omp_unset_lock(&Sync);
```
Synchronization Example

```c
omp_lock_t Sync;
...
omp_init_lock( &Sync );
...
Thread #0:
while( omp_test_lock( &Sync ) == 0 )
{
  DoSomeUsefulWork_0();
}
Thread #1:
while( omp_test_lock( &Sync ) == 0 )
{
  DoSomeUsefulWork_1();
}
```

Single-thread-execution Synchronization

```c
#pragma omp single
```
Restricts execution to a single thread ever. This is used when an operation only makes sense for one thread to do. Reading data from a file is a good example.
Creating Sections of OpenMP Code

Sections are independent blocks of code, able to be assigned to separate threads if they are available.

```c
#pragma omp parallel sections
{
    #pragma omp section
    { Task 1 }
    #pragma omp section
    { Task 2 }
}
```

(Note: there is an implied barrier after parallel for loops and OpenMP sections, unless the nowait clause is used)

A Functional Decomposition Sections Example

```c
omp_set_num_threads(3);
#pragma omp parallel sections
{
    #pragma omp section
    { Watcher( ); }

    #pragma omp section
    { Animals( ); }

    #pragma omp section
    { Plants( ); }

} // implied barrier -- all functions must return to get past here
```
A Potential OpenMP/Visual Studio Compiler Problem

If you print to standard error (stderr) from inside a for-loop, like I do, then you think that you need to include stderr in the shared list because, well, the loops share it:

#pragma omp parallel for default(none) shared(a,b,stderr)

This turns out to be true for g++/gcc only.

If you are using Visual Studio, then do not include stderr in the list. If you do, you will get this error:

1>Y:ICS575S22\robertw5-01\Project1\Project1.cpp(113,98): error C2059: syntax error: '('

This is because:
• In g++/gcc, stderr is a variable
• In Visual Studio, stderr is a defined macro