High Performance Computing Modernization Program (HPCMP) Summer 2011 Puerto Rico Workshop on
Intermediate Parallel Programming & Cluster Computing
in conjunction with
the National Computational Science Institute (NCSI)/ SC11 Conference

Jointly hosted at Polytechnic U of Puerto Rico and U Oklahoma
and available live via videoconferencing
(streaming video recordings coming soon)

Sponsored by
DOD HPCMP,
SC11/ACM,
NCSI and
OK EPSCoR

Intermediate Parallel Programming & Cluster Computing

Shared Memory Multithreading

Josh Alexander, University of Oklahoma
Ivan Babic, Earlham College
Ken Gamradt, South Dakota State University
Andrew Fitz Gibbon, Amazon.com
Mobeen Ludin, Earlham College
Tom Murphy, Contra Costa College
Henry Neeman, University of Oklahoma
Charlie Peck, Earlham College
Stephen Providence, Hampton University
Jeff Rufinus, Widener University
Luis Vicente, Polytechnic University of Puerto Rico
Aaron Weeden, Earlham College
Sunday July 31 – Saturday August 6 2011

SHODOR

EARLHAM COLLEGE

POLYTECHNIC UNIVERSITY PUERTO RICO

CCC

HAMPTON UNIVERSITY

Widener University
1821

SDSU

OU SUPERCOMPUTING CENTER for EDUCATION & RESEARCH UNIVERSITY OF OKLAHOMA OSCER

OU IT INFORMATION TECHNOLOGY THE UNIVERSITY OF OKLAHOMA

This is an experiment!

It's the nature of these kinds of videoconferences that **FAILURES ARE GUARANTEED TO HAPPEN! NO PROMISES!**

So, please bear with us. Hopefully everything will work out well enough.

If you lose your connection, you can retry the same kind of connection, or try connecting another way.

Remember, if all else fails, you always have the toll free phone bridge to fall back on.

H.323 (Polycom etc)

If you want to use H.323 videoconferencing – for example, Polycom – then:

- If you ARE already registered with the OneNet gatekeeper, dial 2500409.

- If you AREN'T registered with the OneNet gatekeeper (which is probably the case), then:
  - Dial `164.58.250.47`
  - When asked for the conference ID, enter: `#0409#`

Many thanks to Roger Holder and OneNet for providing this.

H.323 from Internet Explorer

From a Windows PC running Internet Explorer:

1. You **MUST** have the ability to install software on the PC (or have someone install it for you).
2. Download and install the latest Java Runtime Environment (JRE) from here (click on the Java Download icon, because that install package includes both the JRE and other components).
3. Download and install this video decoder.
4. Start Internet Explorer.
5. Copy-and-paste this URL into your IE window:
   http://164.58.250.47/
6. When that webpage loads, in the upper left, click on "Streaming".
7. In the textbox labeled Sign-in Name, type your name.
8. In the textbox labeled Conference ID, type this:
   0409
9. Click on "Stream this conference".
10. When that webpage loads, you may see, at the very top, a bar offering you options. If so, click on it and choose "Install this add-on."

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

EVO

There's a quick description of how to use EVO on the workshop logistics webpage.

Phone Bridge

If all else fails, you can call into our toll free phone bridge:

1-800-832-0736

* 623 2874 #

Please mute yourself and use the phone to listen.

Don't worry, we'll call out slide numbers as we go.

Please use the phone bridge **<u>ONLY</u>** if you cannot connect any other way: the phone bridge is charged per connection per minute, so our preference is to minimize the number of connections.

Many thanks to OU Information Technology for providing the toll free phone bridge.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Please Mute Yourself

No matter how you connect, please mute yourself, so that we cannot hear you.

At ISU and UW, we will turn off the sound on all conferencing technologies.

That way, we won't have problems with echo cancellation.

Of course, that means we cannot hear questions.

So for questions, you'll need to send some kind of text.

Questions via Text: Piazzza

Ask questions via:

http://www.piazza.com/

All questions will be read out loud and then answered out loud.

NOTE: Because of image-and-likeness rules, people attending remotely offsite via videoconferencing **CANNOT** ask questions via voice.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Thanks for helping and sponsoring!

- OSCER operations staff (Brandon George, Dave Akin, Brett Zimmerman, Josh Alexander, Patrick Calhoun)
- Debi Gentis, OU
- Kevin Blake, OU IT (videographer)
- OU School of Electrical & Computer Engineering (LittleFe buildout)
- James Deaton and Roger Holder, OneNet
- Luis Vicente and Alfredo Cruz, Polytechnic U of Puerto Rico
- Omar Padron, Kean U
- Scott Lathrop, SC11 General Chair
- Donna Cappo, ACM
- Bob Panoff, Jack Parkin, Joyce South, Shodor Education Foundation Inc
- Jerry Malayer and Jim Wicksted, Oklahoma EPSCoR
- Dept of Defense High Performance Computing Modernization Program

This is an experiment!

It's the nature of these kinds of videoconferences that **FAILURES ARE GUARANTEED TO HAPPEN! NO PROMISES!**

So, please bear with us. Hopefully everything will work out well enough.

If you lose your connection, you can retry the same kind of connection, or try connecting another way.

Remember, if all else fails, you always have the toll free phone bridge to fall back on.

Outline

- Parallelism

- Shared Memory Multithreading
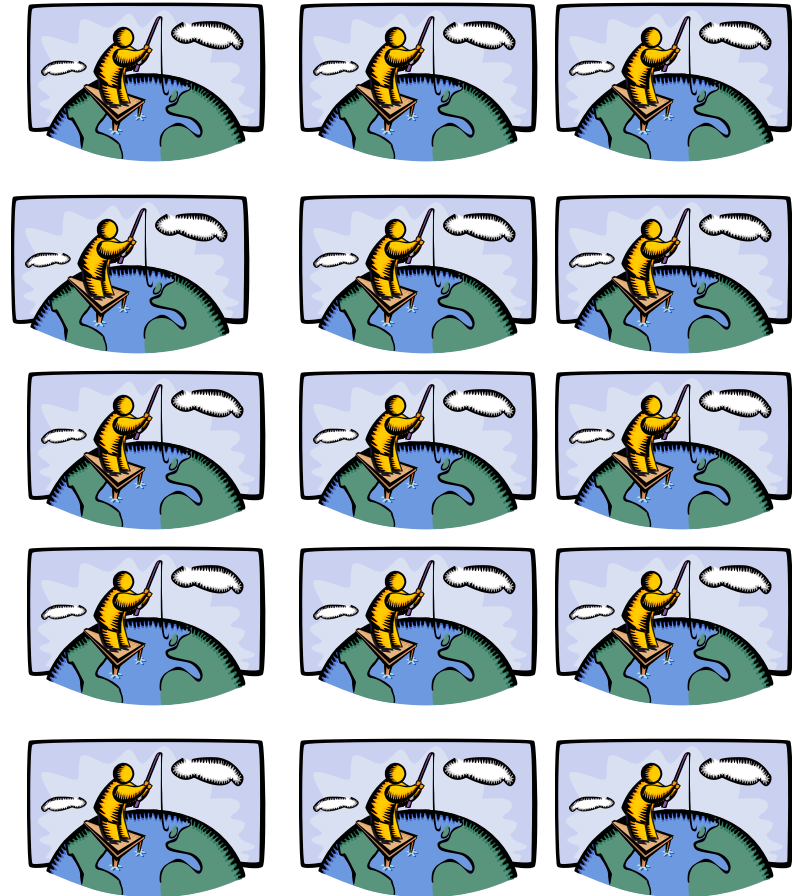
- OpenMP

Parallelism

Parallelism

__Parallelism__ means doing multiple things at the same time: you can get more work done in the same time.

Less fish …

More fish!

What Is Parallelism?

Parallelism is the use of multiple processing units – either processors or parts of an individual processor – to solve a problem, and in particular the use of multiple processing units operating concurrently on different parts of a problem.

The different parts could be different tasks, or the same task on different pieces of the problem's data.

Common Kinds of Parallelism

- Instruction Level Parallelism

- Shared Memory Multithreading (for example, OpenMP)

- Distributed Multiprocessing (for example, MPI)

- GPU Parallelism (for example, CUDA)

- Hybrid Parallelism

  - Distributed + Shared (for example, MPI + OpenMP)

  - Shared + GPU (for example, OpenMP + CUDA)

  - Distributed + GPU (for example, MPI + CUDA)

Why Parallelism Is Good

- **<u>The Trees</u>**: We like parallelism because, as the number of processing units working on a problem grows, we can solve **<u>the same problem in less time</u>**.

- **<u>The Forest</u>**: We like parallelism because, as the number of processing units working on a problem grows, we can solve **<u>bigger problems</u>**.

Parallelism Jargon

- ***Threads*** are execution sequences that share a single memory area ("***address space***")

- ***Processes*** are execution sequences with their own independent, private memory areas

… and thus:

- ***Multithreading***: parallelism via multiple **threads**

- ***Multiprocessing***: parallelism via multiple **processes**

Generally:

- Shared Memory Parallelism is concerned with **threads**, and

- Distributed Parallelism is concerned with **processes**.

Jargon Alert!

In principle:

- "shared memory parallelism" ➜ "multithreading"
- "distributed parallelism"   ➜ "multiprocessing"

In practice, sadly, the following terms are often used interchangeably:

- Parallelism
- ***Concurrency*** (not as popular these days)
- Multithreading
- Multiprocessing

Typically, you have to figure out what is meant based on the context.

Amdahl's Law

In 1967, Gene Amdahl came up with an idea so crucial to our understanding of parallelism that they named a **Law** for him:

$$S = \frac{1}{(1 - F_p) + \dfrac{F_p}{S_p}}$$

where S is the overall speedup achieved by parallelizing a code, F_p is the fraction of the code that's parallelizable, and S_p is the speedup achieved in the parallel part.[1]

Amdahl's Law: Huh?

What does Amdahl's Law tell us?

Imagine that you run your code on a zillion processors. The parallel part of the code could speed up by as much as a factor of a zillion.

For sufficiently large values of a zillion, the **parallel part would take zero time**!

But, the **serial** (non-parallel) part would take the **same amount of time as on a single processor**.

So running your code on infinitely many processors would still take at least as much time as it takes to run just the serial part.

Max Speedup by Serial %

Sponsored by
DOD HPCMP,
SC11/ACM,
NCSI and
OK EPSCoR
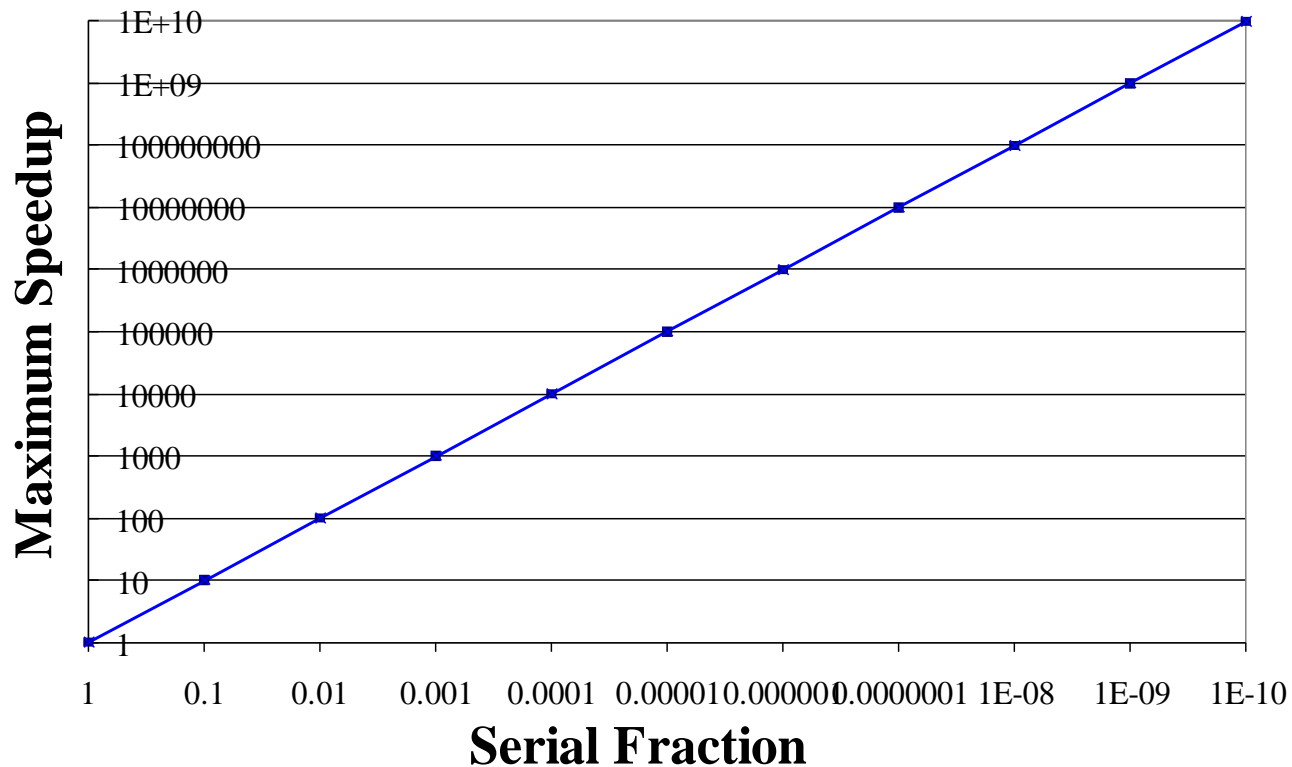
Amdahl's Law Example (F90)

```
PROGRAM amdahl_test
  IMPLICIT NONE
  REAL,DIMENSION(a_lot) :: array
  REAL     :: scalar
  INTEGER :: index

  READ *, scalar        !! Serial part
  DO index = 1, a_lot !! Parallel part
    array(index) = scalar * index
  END DO
END PROGRAM amdahl_test
```

If we run this program on infinitely many CPUs, then the total run time will still be at least as much as the time it takes to perform the **READ**.

Amdahl's Law Example (C)

```c
int main ()
{
  float array[a_lot];
  float scalar;
  int   index;

  scanf("%f", scalar); /* Serial part */
  /* Parallel part */
  for (index = 0; index < a_lot; index++) {
    array(index) = scalar * index
  }
}
```

If we run this program on infinitely many CPUs, then the total run time will still be at least as much as the time it takes to perform the **scanf**.

The Point of Amdahl's Law

<u>Rule of Thumb</u>: When you write a parallel code, try to make as much of the code parallel as possible, because the <strong style="color:red"><u>serial part will be the limiting factor</u> on parallel speedup.

Note that this rule will not hold when the <strong style="color:red"><u>overhead</u> cost of parallelizing exceeds the parallel speedup. More on this presently.

Speedup

The goal in parallelism is *__linear speedup__*: getting the speed of the job to increase by a factor equal to the number of processors.

Very few programs actually exhibit linear speedup, but some come close.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

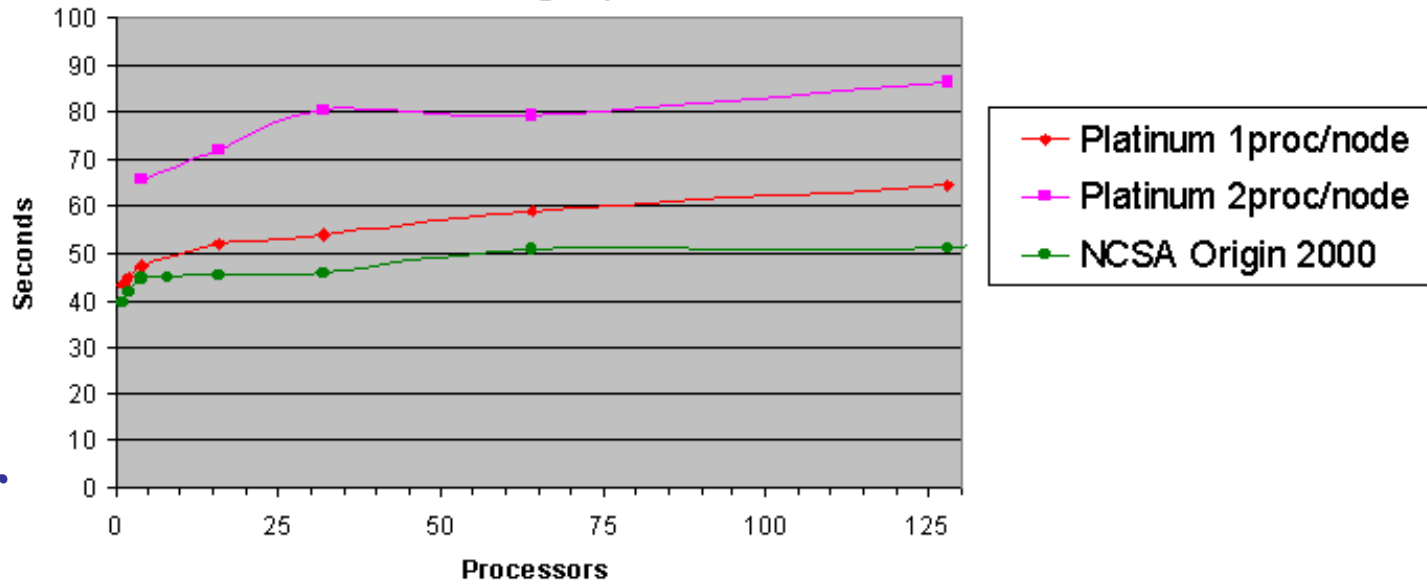# Scalability

Scalable means "performs just as well regardless of how big the problem is." A scalable code has near linear speedup.

Better

ARPS Benchmark Timings
19x19x43 3km grid/processor

Platinum = NCSA 1024 processor PIII/1GHZ Linux Cluster
Note: NCSA Origin timings are scaled from 19x19x53 domains.

Strong vs Weak Scalability

- ***Strong Scalability***: If you double the number of processors, but you **keep the problem size constant**, then the problem takes **half as long** to complete.

- ***Weak Scalability***: If you double the number of processors, and **double the problem size**, then the problem takes the **same amount of time** to complete.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Scalability

This benchmark shows **<u>weak</u>** scalability.

ARPS Benchmark Timings
19x19x43 3km grid/processor

Platinum 1proc/node
Platinum 2proc/node
NCSA Origin 2000

Better

Platinum = NCSA 1024 processor PIII/1GHZ Linux Cluster
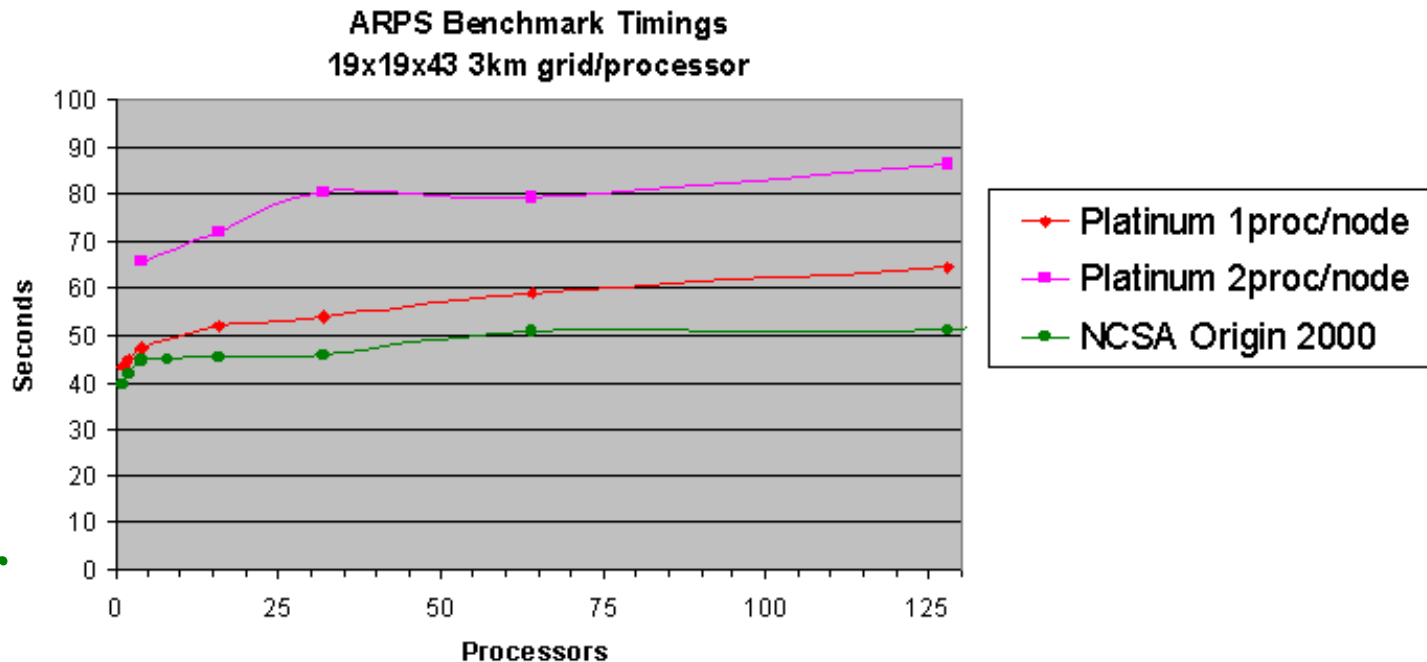Note: NCSA Origin timings are scaled from 19x19x53 domains.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Granularity

Granularity is the size of the subproblem that each thread or process works on, and in particular the size that it works on between communicating or synchronizing with the others.

Some codes are ***coarse grain*** (a few very large parallel parts) and some are ***fine grain*** (many small parallel parts).

Usually, **coarse grain codes are more scalable** than fine grain codes, because less of the runtime is spent managing the parallelism, so a higher proportion of the runtime is spent getting the work done.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Parallel Overhead

Parallelism isn't free. Behind the scenes, the compiler and the hardware have to do a lot of ***overhead*** work to make parallelism happen.

The overhead typically includes:

- **Managing** the multiple threads/processes
- **Communication** among threads/processes
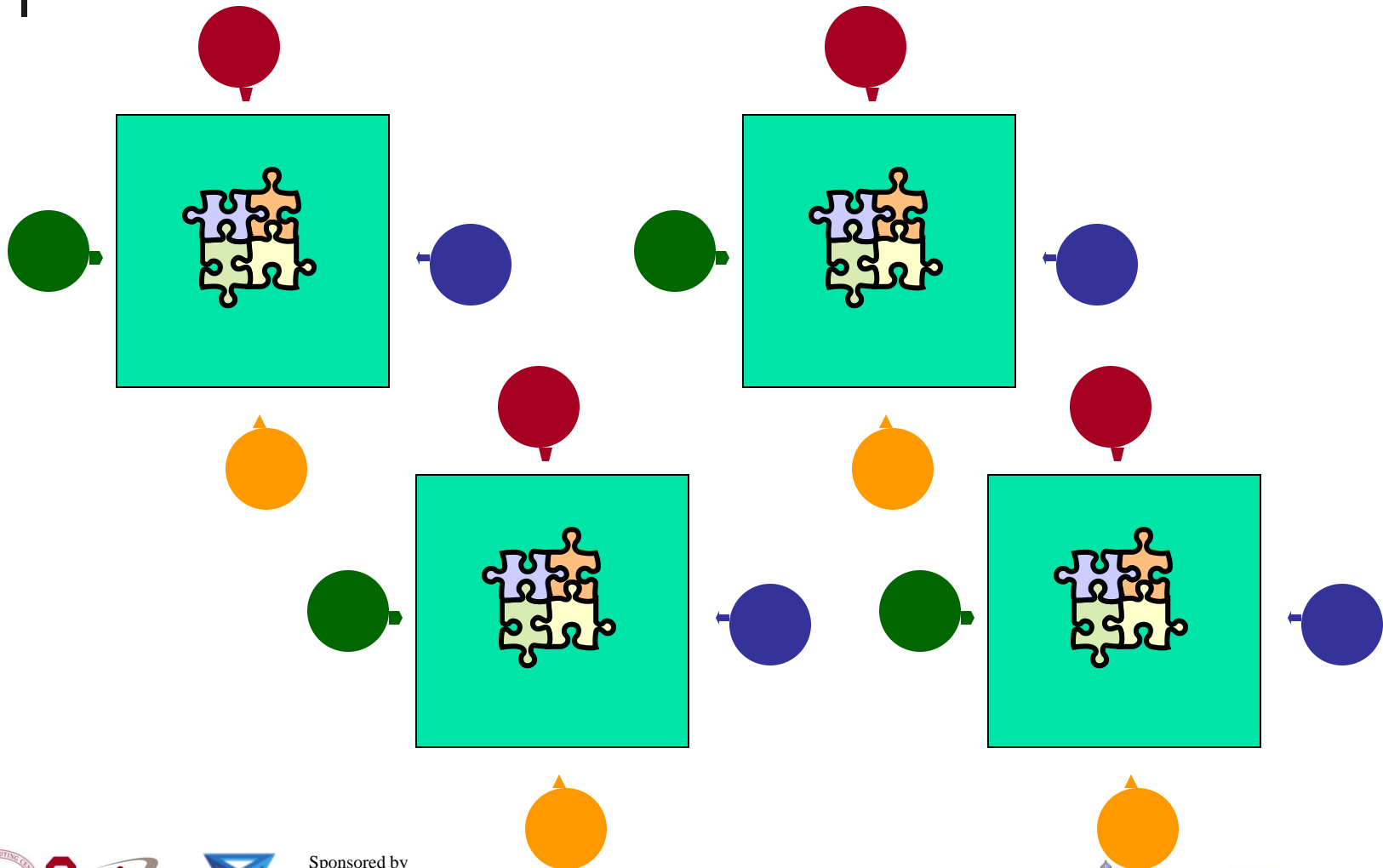- ***Synchronization*** (described later)

Shared Memory Multithreading

The Jigsaw Puzzle Analogy

Sponsored by
DOD HPCMP,
SC11/ACM,
NCSI and
OK EPSCoR
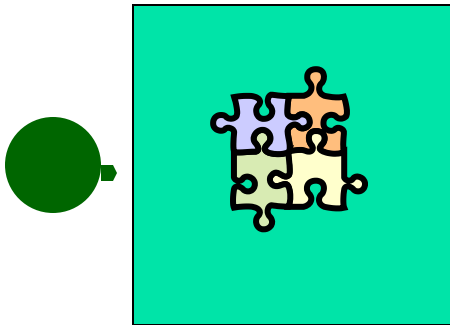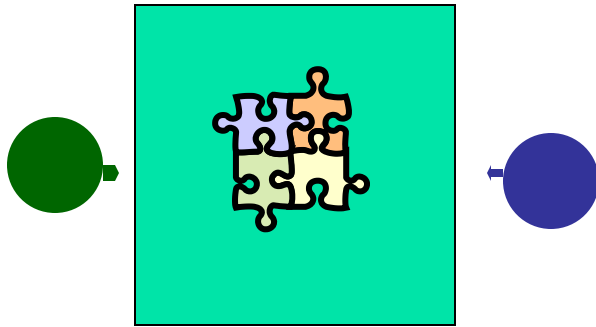
Serial Computing

Suppose you want to do a jigsaw puzzle that has, say, a thousand pieces.

We can imagine that it'll take you a certain amount of time. Let's say that you can put the puzzle together in an hour.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Shared Memory Parallelism

If Scott sits across the table from you, then he can work on his half of the puzzle and you can work on yours. Once in a while, you'll both reach into the pile of pieces at the same time (you'll **_contend_** for the same resource), which will cause a little bit of slowdown. And from time to time you'll have to work together (**_communicate_**) at the interface between his half and yours. The speedup will be nearly 2-to-1: y'all might take 35 minutes instead of 30.

The More the Merrier?
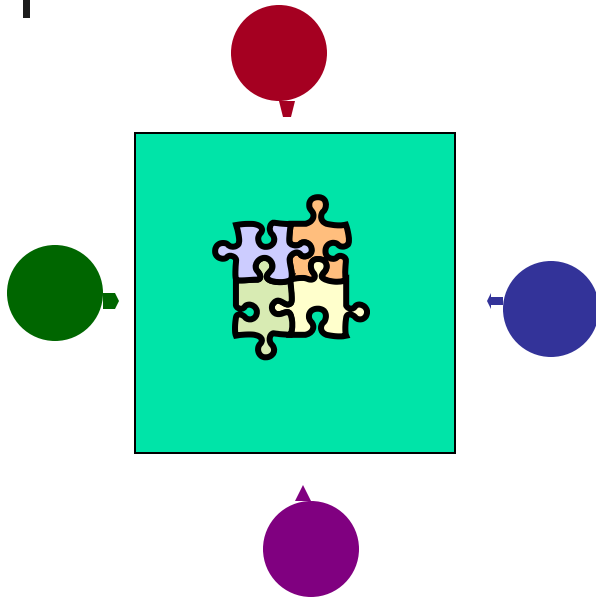
Now let's put Bob and Charlie on the other two sides of the table. Each of you can work on a part of the puzzle, but there'll be a lot more contention for the shared resource (the pile of puzzle pieces) and a lot more communication at the interfaces. So y'all will get noticeably less than a 4-to-1 speedup, but you'll still have an improvement, maybe something like 3-to-1: the four of you can get it done in 20 minutes instead of an hour.

Diminishing Returns
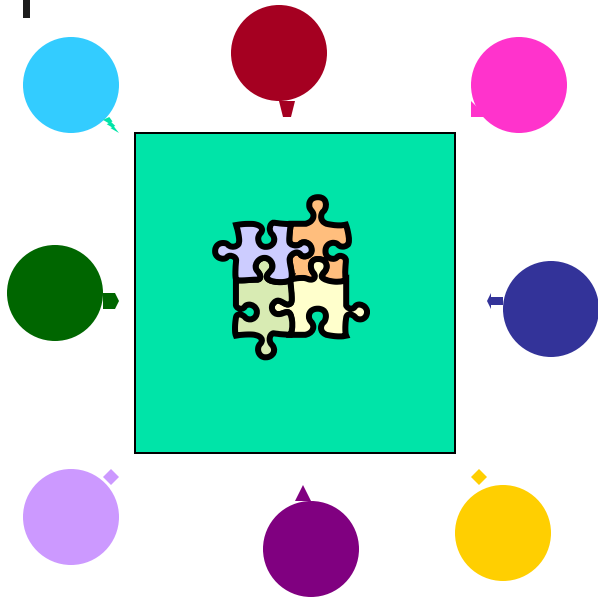
If we now put Dave and Tom and Dan and Paul on the corners of the table, there's going to be a whole lot of contention for the shared resource, and a lot of communication at the many interfaces. So the speedup y'all get will be much less than we'd like; you'll be lucky to get 5-to-1.

So we can see that adding more and more workers onto a shared resource is eventually going to have a diminishing return.

Distributed Parallelism

Now let's try something a little different. Let's set up two tables, and let's put you at one of them and Scott at the other. Let's put half of the puzzle pieces on your table and the other half of the pieces on Scott's. Now y'all can work completely independently, without any contention for a shared resource. **BUT**, the cost per communication is **MUCH** higher (you have to scootch your tables together), and you need the ability to split up (***decompose***) the puzzle pieces reasonably evenly, which may be tricky to do for some puzzles.

More Distributed Processors
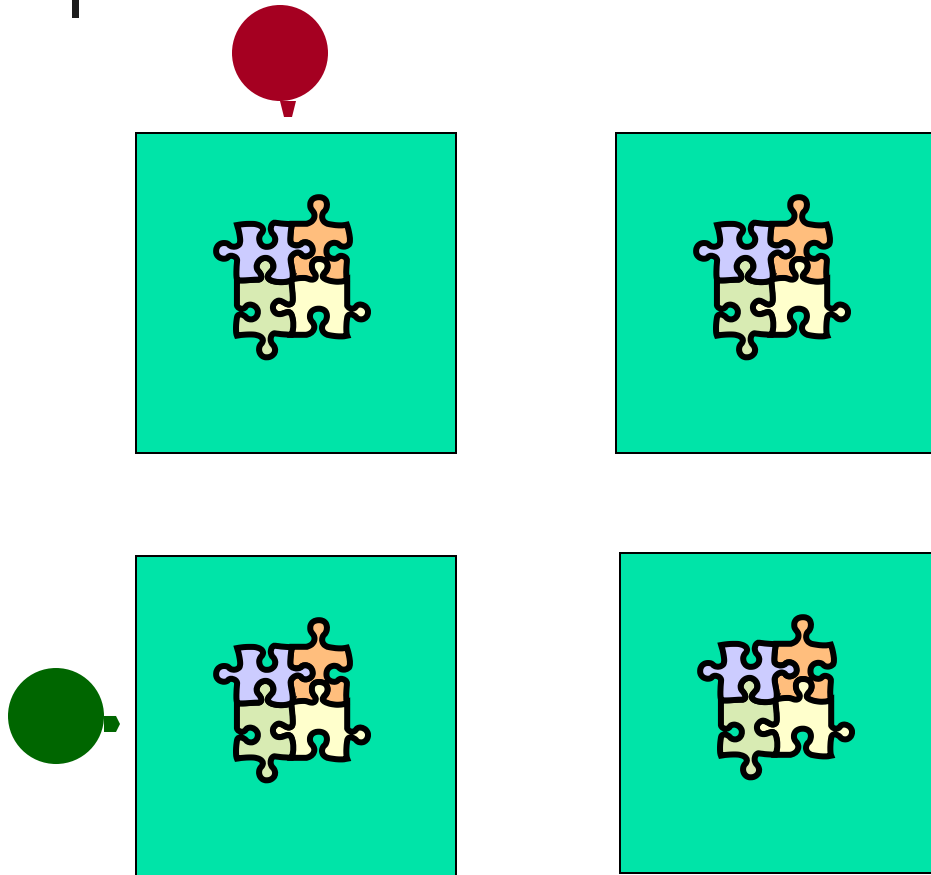
It's a lot easier to add more processors in distributed parallelism. But, you always have to be aware of the need to decompose the problem and to communicate among the processors. Also, as you add more processors, it may be harder to **_load balance_** the amount of work that each processor gets.

Load Balancing
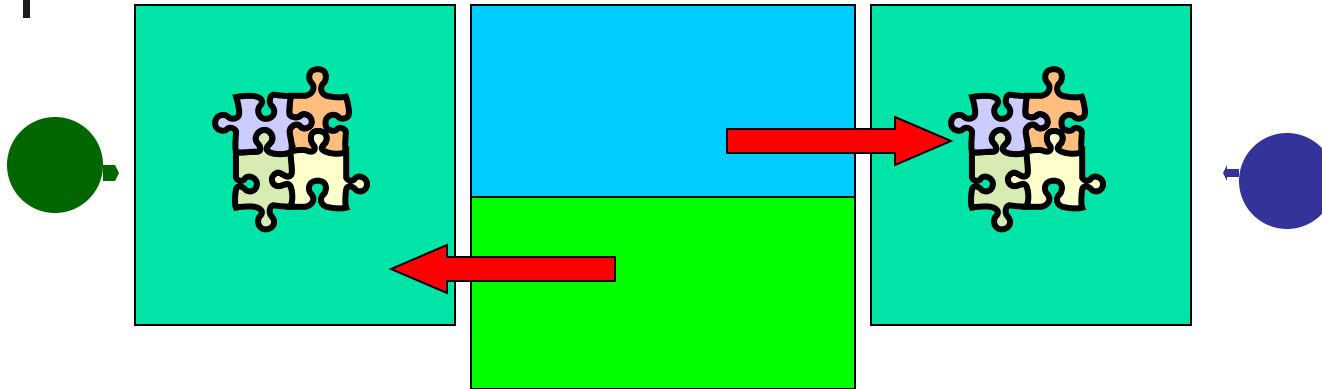
Load balancing means ensuring that everyone completes their workload at roughly the same time.

For example, if the jigsaw puzzle is half grass and half sky, then you can do the grass and Scott can do the sky, and then y'all only have to communicate at the horizon – and the amount of work that each of you does on your own is roughly equal. So you'll get pretty good speedup.

Load Balancing
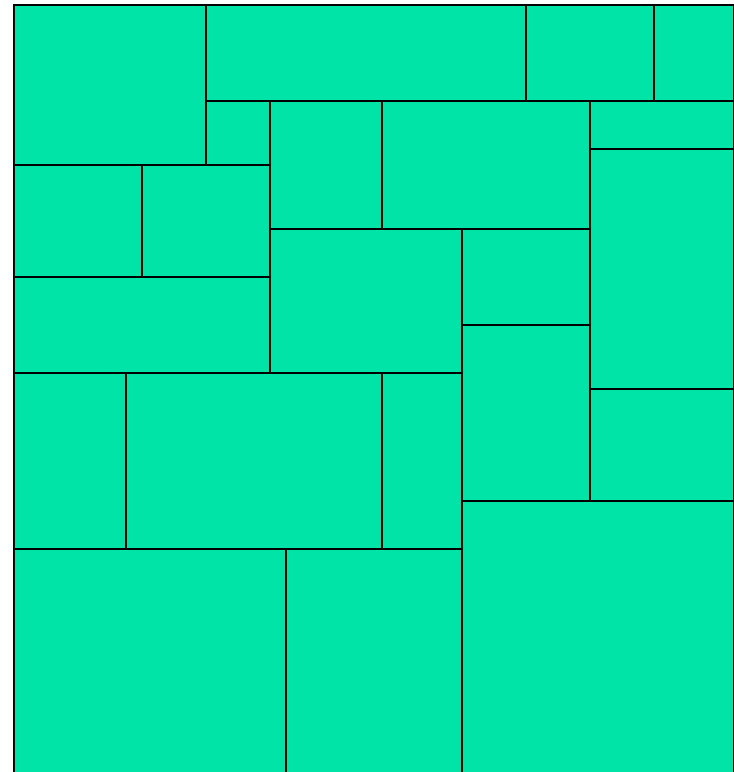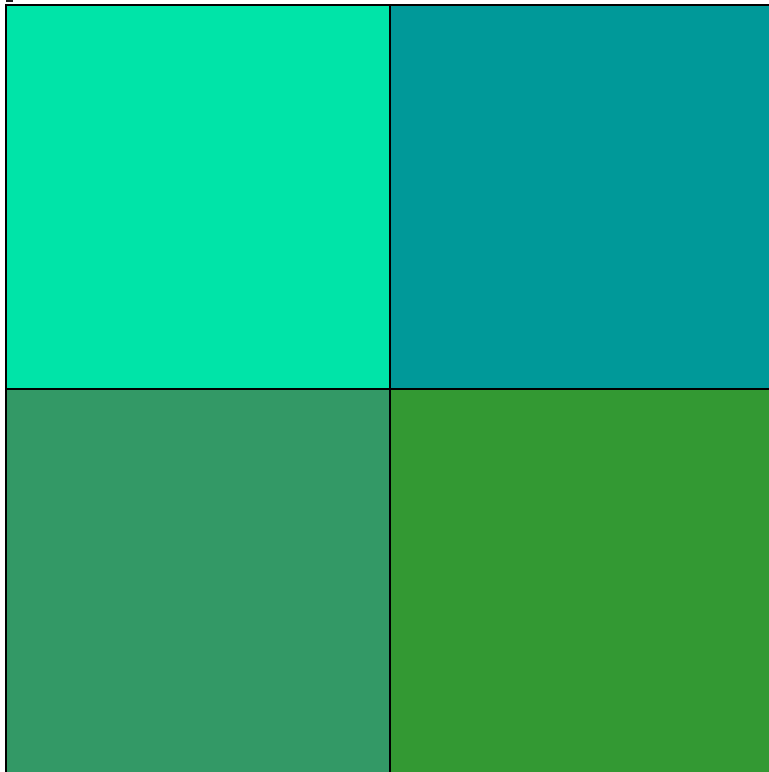
Load balancing can be easy, if the problem splits up into chunks of roughly equal size, with one chunk per processor. Or load balancing can be very hard.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Load Balancing

Load balancing can be easy, if the problem splits up into chunks of roughly equal size, with one chunk per processor. Or load balancing can be very hard.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Load Balancing

Load balancing can be easy, if the problem splits up into chunks of roughly equal size, with one chunk per processor. Or load balancing can be very hard.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011
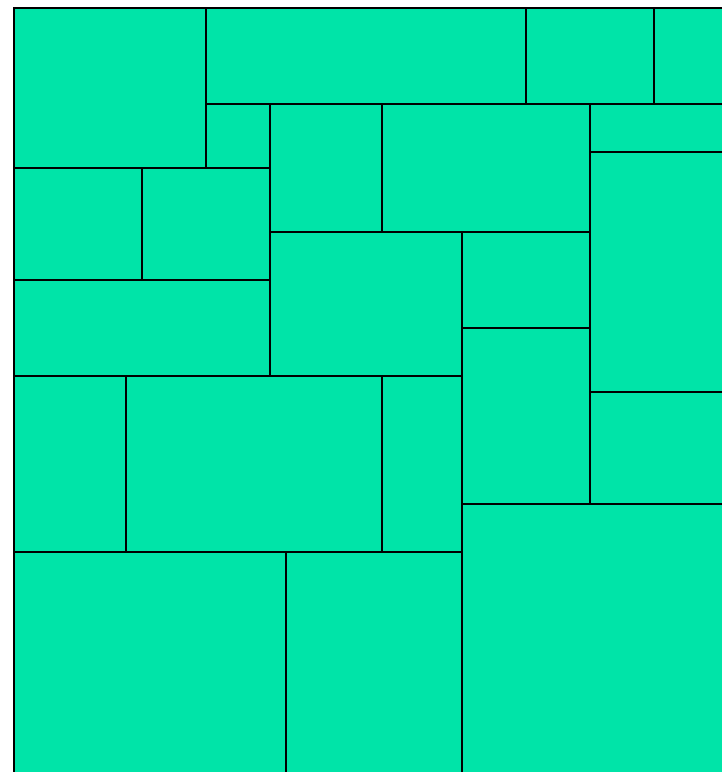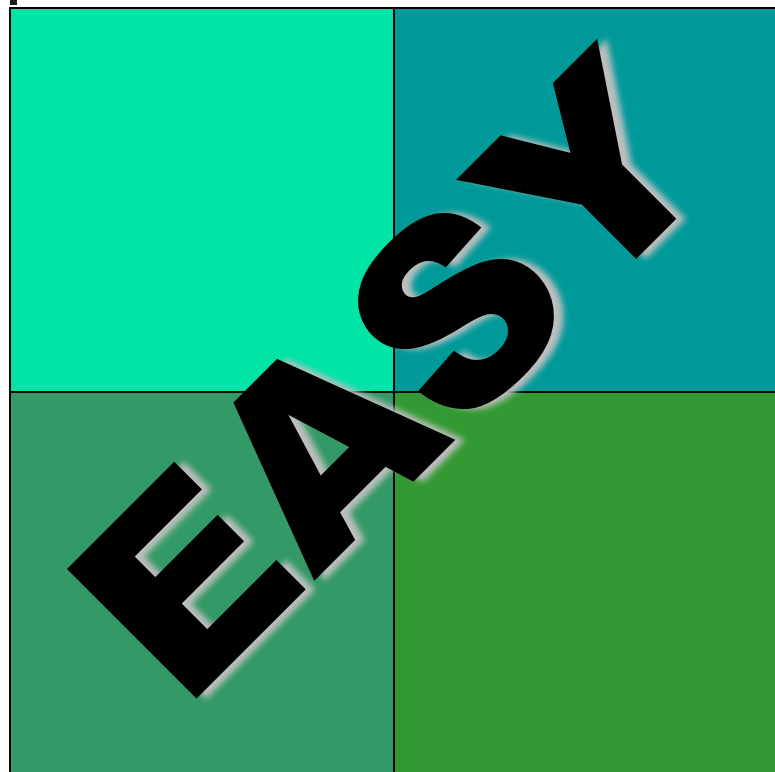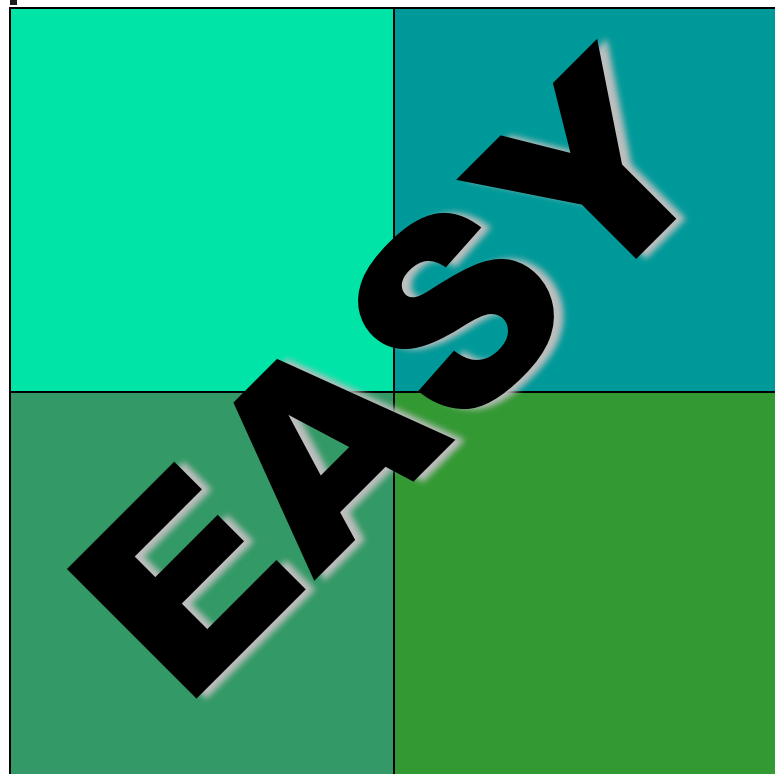
How Shared Memory Parallelism Behaves

The Fork/Join Model

Many shared memory parallel systems use a programming model called *Fork*/*Join*. Each program begins executing on just a single thread, called the *parent*.

Fork: When a parallel region is reached, the **parent** thread *spawns* additional *child* threads as needed.

Join: When the parallel region ends, the **child threads shut down**, leaving only the parent still running.

The Fork/Join Model (cont'd)

Parent Thread

Start

Fork Overhead Child Threads

Compute time

Join Overhead

End
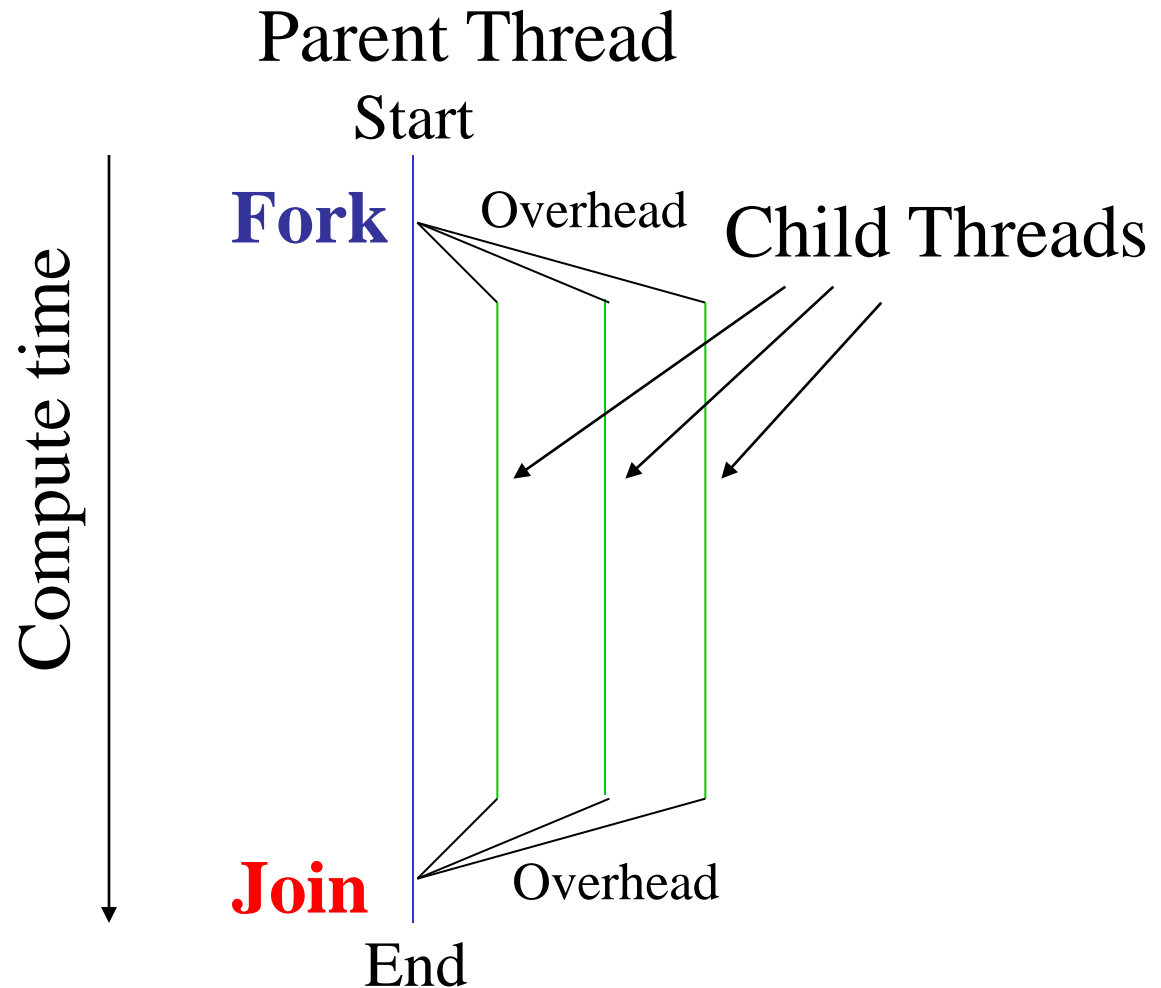
The Fork/Join Model (cont'd)

In principle, as a parallel section completes, the child threads shut down (join the parent), forking off again when the parent reaches another parallel section.

In practice, the child threads often continue to exist but are ___idle___.

Why?

Sponsored by
DOD HPCMP,
SC11/ACM,
NCSI and
OK EPSCoR

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

47

Principle vs. Practice

Start

Fork

Join

End

Start

Fork

Idle

Join

End
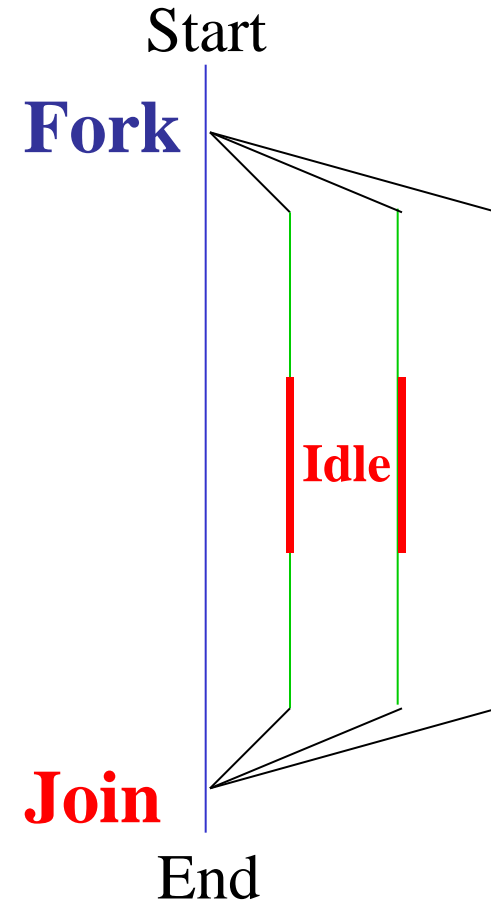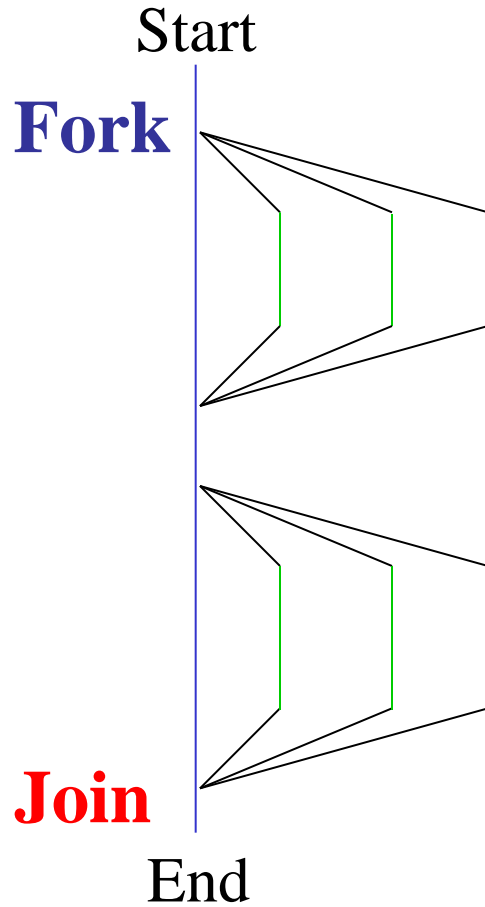
Why Idle?

- On some shared memory multithreading computers, the **overhead cost** of forking and joining is **high** compared to the cost of computing, so rather than waste time on overhead, the **children sit idle** until the next parallel section.

- On some computers, joining threads releases a program's control over the child processors, so they may not be available for more parallel work later in the run. ***Gang scheduling*** is preferable, because then all of the processors are guaranteed to be available for the whole run.

Standards and Nonstandards

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Standards and Nonstandards

In computing, there are standards and nonstandards.

Standards are established by independent organizations and made public, so that anyone can produce a standard-compliant implementation.

Example standards organizations include:

- International Organization for Standardization (ISO)
  - "'ISO' [is] derived from the Greek *isos*, meaning 'equal'." [2]
- American National Standards Institute (ANSI)
- Ecma International

Nonstandards are produced by a single organization or consortium, with no requirement for external input and no recognized standard.

Standards and Nonstandards

In practice, there are:

- **standard standards**, which both are common and have been accepted as official standards – for example: C, TCP/IP, HTML;

- **nonstandard standards**, which aren't common but have been accepted as official standards – for example: Myrinet;

- **standard nonstandards**, which are common but haven't been accepted as official standard – for example: PDF, Windows;

- **nonstandard nonstandards**, which aren't common and haven't been accepted as official standards – for example: WordStar.

OpenMP

Most of this discussion is from [3], with a little bit from [4].

What Is OpenMP?

OpenMP is a ***standard*** way of expressing shared memory parallelism.

OpenMP consists of ***compiler directives***, **functions** and **environment variables**.

When you compile a program that has OpenMP in it, then:

- if your compiler knows OpenMP, then you get an executable that can run in parallel;

- otherwise, the compiler ignores the OpenMP stuff and you get a purely serial executable.

OpenMP can be used in Fortran, C and C++, but only if your preferred compiler explicitly supports it.

Sponsored by
DOD HPCMP,
SC11/ACM,
NCSI and
OK EPSCoR

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

54

Compiler Directives

A **_compiler directive_** is a line of source code that gives the compiler special information about the statement or block of code that immediately follows.

C++ and C programmers already know about compiler directives:

```
#include "MyClass.h"
```

Many Fortran programmers already have seen at least one compiler directive:

```
INCLUDE 'mycommon.inc'
```

OR

```
INCLUDE "mycommon.inc"
```

Sponsored by
DOD HPCMP,
SC11/ACM,
NCSI and
OK EPSCoR

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

55

OpenMP Compiler Directives

OpenMP compiler directives in Fortran look like this:

`!$OMP` ...*stuff*...

In C++ and C, OpenMP directives look like:

`#pragma omp` ...*stuff*...

Both directive forms mean "the rest of this line contains OpenMP information."

Aside: "***pragma***" is the Greek word for "thing." Go figure.

Example OpenMP Directives

Fortran	C++/C
`!$OMP PARALLEL DO`	`#pragma omp parallel for`
`!$OMP CRITICAL`	`#pragma omp critical`
`!$OMP MASTER`	`#pragma omp master`
`!$OMP BARRIER`	`#pragma omp barrier`
`!$OMP SINGLE`	`#pragma omp single`
`!$OMP ATOMIC`	`#pragma omp atomic`
`!$OMP SECTION`	`#pragma omp section`
`!$OMP FLUSH`	`#pragma omp flush`
`!$OMP ORDERED`	`#pragma omp ordered`

Note that we won't cover all of these.

A First OpenMP Program (F90)

```fortran
PROGRAM hello_world
  IMPLICIT NONE
  INTEGER :: number_of_threads, this_thread, iteration
  INTEGER,EXTERNAL :: omp_get_max_threads,
                      omp_get_thread_num

  number_of_threads = omp_get_max_threads()
  WRITE (0,"(I2,A)") number_of_threads, " threads"
!$OMP PARALLEL DO DEFAULT(PRIVATE) &
!$OMP                 SHARED(number_of_threads)
  DO iteration = 0, number_of_threads - 1
    this_thread = omp_get_thread_num()
    WRITE (0,"(A,I2,A,I2,A) ")"Iteration ", &
 &    iteration, ", thread ", this_thread, &
 &    ": Hello, world!"
  END DO
END PROGRAM hello_world
```

A First OpenMP Program (C)

```c
int main ()
{
  int number_of_threads, this_thread, iteration;
  int omp_get_max_threads(), omp_get_thread_num();

  number_of_threads = omp_get_max_threads();
  fprintf(stderr, "%2d threads\n", number_of_threads);
# pragma omp parallel for default(private) \
                          shared(number_of_threads)
  for (iteration = 0;
       iteration < number_of_threads; iteration++) {
    this_thread = omp_get_thread_num();
    fprintf(stderr, "Iteration %2d, thread %2d: Hello, world!\n",
        iteration, this_thread);
  }
}
```

Sponsored by
DOD HPCMP,
SC11/ACM,
NCSI and
OK EPSCoR

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

59

Running `hello_world`

```
% setenv  OMP_NUM_THREADS  4
% hello_world
 4 threads
Iteration  0, thread  0: Hello, world!
Iteration  1, thread  1: Hello, world!
Iteration  3, thread  3: Hello, world!
Iteration  2, thread  2: Hello, world!
% hello_world
 4 threads
Iteration  2, thread  2: Hello, world!
Iteration  1, thread  1: Hello, world!
Iteration  0, thread  0: Hello, world!
Iteration  3, thread  3: Hello, world!
% hello_world
 4 threads
Iteration  1, thread  1: Hello, world!
Iteration  2, thread  2: Hello, world!
Iteration  0, thread  0: Hello, world!
Iteration  3, thread  3: Hello, world!
```

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

OpenMP Issues Observed

From the `hello_world` program, we learn that:

- At some point before running an OpenMP program, you must set an environment variable

  OMP_NUM_THREADS

  that represents the number of threads to use.

- The order in which the threads execute is **nondeterministic**.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

The **PARALLEL** DO Directive (F90)

The **PARALLEL DO** directive tells the compiler that the **DO** loop immediately after the directive should be executed in parallel; for example:

```
!$OMP PARALLEL DO
  DO index = 1, length
    array(index) = index * index
  END DO
```

The iterations of the loop will be computed in parallel (note that they are independent of one another).

The `parallel for` Directive (C)

The **`parallel for`** directive tells the compiler that the **`for`** loop immediately after the directive should be executed in parallel; for example:

```
# pragma omp parallel for
  for (index = 0; index < length; index++) {
    array[index] = index * index;
  }
```

The iterations of the loop will be computed in parallel (note that they are independent of one another).

A Change to `hello_world`

Suppose we do 3 loop iterations per thread:

`DO iteration = 0, number_of_threads * 3 - 1`

```
% hello_world
 4 threads
Iteration  9, thread  3: Hello, world!
Iteration  0, thread  0: Hello, world!
Iteration 10, thread  3: Hello, world!
Iteration 11, thread  3: Hello, world!
Iteration  1, thread  0: Hello, world!
Iteration  2, thread  0: Hello, world!
Iteration  3, thread  1: Hello, world!
Iteration  6, thread  2: Hello, world!
Iteration  7, thread  2: Hello, world!
Iteration  8, thread  2: Hello, world!
Iteration  4, thread  1: Hello, world!
Iteration  5, thread  1: Hello, world!
```

Notice that the iterations are split into contiguous *__chunks__*, and each thread gets one chunk of iterations.

Chunks

By default, OpenMP splits the iterations of a loop into chunks of equal (or roughly equal) size, assigns each chunk to a thread, and lets each thread loop through its subset of the iterations.

So, for example, if you have 4 threads and 12 iterations, then each thread gets three iterations:

- Thread 0: iterations 0, 1, 2
- Thread 1: iterations 3, 4, 5
- Thread 2: iterations 6, 7, 8
- Thread 3: iterations 9, 10, 11

Notice that each thread performs its own chunk in deterministic order, but that the overall order is nondeterministic.

Private and Shared Data

Private data are data that are owned by, and only visible to, a single individual thread.

Shared data are data that are owned by and visible to all threads.

(Note: In distributed parallelism, all data are private, as we'll see next time.)

Should All Data Be Shared?

In our example program, we saw this:

`!$OMP PARALLEL DO DEFAULT(PRIVATE) SHARED(number_of_threads)`

What do `DEFAULT(PRIVATE)` and `SHARED` mean?

We said that OpenMP uses shared memory parallelism. So `PRIVATE` and `SHARED` refer to memory.

Would it make sense for all data within a parallel loop to be shared?

A Private Variable (F90)

Consider this loop:

```
!$OMP PARALLEL DO …
  DO iteration = 0, number_of_threads - 1
    this_thread = omp_get_thread_num()
    WRITE (0,"(A,I2,A,I2,A) ") "Iteration ", iteration, &
 &     ", thread ", this_thread, ": Hello, world!"
  END DO
```

Notice that, if the iterations of the loop are executed concurrently, then the loop index variable named **iteration** will be wrong for all but one of the threads.

Each thread should get its own copy of the variable named **iteration**.

Sponsored by
DOD HPCMP,
SC11/ACM,
NCSI and
OK EPSCoR

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

68

A Private Variable (C)

Consider this loop:

```
#pragma omp parallel for …
  for (iteration = 0;
       iteration < number_of_threads; iteration++) {
    this_thread = omp_get_thread_num();
    printf("Iteration %d, thread %d: Hello, world!\n",
       iteration, this_thread);
}
```

Notice that, if the iterations of the loop are executed concurrently, then the loop index variable named **iteration** will be wrong for all but one of the threads.

Each thread should get its own copy of the variable named **iteration**.

Another Private Variable (F90)

```fortran
!$OMP PARALLEL DO …
  DO iteration = 0, number_of_threads - 1
    this_thread = omp_get_thread_num()
    WRITE (0,"(A,I2,A,I2,A)") "Iteration ", iteration, &
 &     ", thread ", this_thread, ": Hello, world!"
  END DO
```

Notice that, if the iterations of the loop are executed concurrently, then **this_thread** will be wrong for all but one of the threads.

Each thread should get its own copy of the variable named **this_thread**.

Another Private Variable (C)

```
#pragma omp parallel for …
  for (iteration = 0;
       iteration < number_of_threads; iteration++) {
    this_thread = omp_get_thread_num();
    printf("Iteration %d, thread %d: Hello, world!\n",
       iteration, this_thread);
}
```

Notice that, if the iterations of the loop are executed concurrently, then **this_thread** will be wrong for all but one of the threads.

Each thread should get its own copy of the variable named **this_thread**.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

A Shared Variable (F90)

```
!$OMP PARALLEL DO …
  DO iteration = 0, number_of_threads - 1
    this_thread = omp_get_thread_num()
    WRITE (0,"(A,I2,A,I2,A)""") "Iteration ", iteration, &
 &     ", thread ", this_thread, ": Hello, world!"
  END DO
```

Notice that, regardless of whether the iterations of the loop are executed serially or in parallel, **number_of_threads** will be correct for all of the threads.

All threads should share a single instance of **number_of_threads**.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

A Shared Variable (C)

```
#pragma omp parallel for …
  for (iteration = 0;
       iteration < number_of_threads; iteration++) {
    this_thread = omp_get_thread_num();
    printf("Iteration %d, thread %d: Hello, world!\n",
       iteration, thread);
}
```

Notice that, regardless of whether the iterations of the loop are executed serially or in parallel, **number_of_threads** will be correct for all of the threads.

All threads should share a single instance of **number_of_threads**.

SHARED & PRIVATE Clauses

The **PARALLEL DO** directive allows extra ***clauses*** to be appended that tell the compiler which variables are shared and which are private:

```
!$OMP PARALLEL DO PRIVATE(iteration,this_thread) &
!$OMP                    SHARED (number_of_threads)
```

This tells that compiler that **iteration** and **this_thread** are private but that **number_of_threads** is shared.

(Note the syntax for continuing a directive in Fortran90.)

DEFAULT Clause

If your loop has lots of variables, it may be cumbersome to put all of them into **SHARED** and **PRIVATE** clauses.

So, OpenMP allows you to declare one kind of data to be the default, and then you only need to explicitly declare variables of the other kind:

```
!$OMP PARALLEL DO DEFAULT(PRIVATE) &
!$OMP                 SHARED(number_of_threads)
```

The default **DEFAULT** (so to speak) is **SHARED**, except for the loop index variable, which by default is **PRIVATE**.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Different Workloads

What happens if the threads have different amounts of work to do?

```fortran
!$OMP PARALLEL DO
  DO index = 1, length
    x(index) = index / 3.0
    IF (x(index) < 0) THEN
      y(index) = LOG(x(index))
    ELSE
      y(index) = 1.0 - x(index)
    END IF
  END DO
```

The threads that finish early have to wait.

Chunks

By default, OpenMP splits the iterations of a loop into chunks of equal (or roughly equal) size, assigns each chunk to a thread, and lets each thread loop through its subset of the iterations.

So, for example, if you have 4 threads and 12 iterations, then each thread gets three iterations:

- Thread 0: iterations 0, 1, 2
- Thread 1: iterations 3, 4, 5
- Thread 2: iterations 6, 7, 8
- Thread 3: iterations 9, 10, 11

Notice that each thread performs its own chunk in deterministic order, but that the overall order is nondeterministic.

Scheduling Strategies

OpenMP supports three scheduling strategies:

- ***Static***: The default, as described in the previous slides – good for iterations that are inherently load balanced.

- ***Dynamic***: Each thread gets a chunk of a few iterations, and when it finishes that chunk it goes back for more, and so on until all of the iterations are done – good when iterations aren't load balanced at all.

- ***Guided***: Each thread gets smaller and smaller chunks over time – a compromise.

Static Scheduling

For N_i iterations and N_t threads, each thread gets one chunk of N_i/N_t loop iterations:
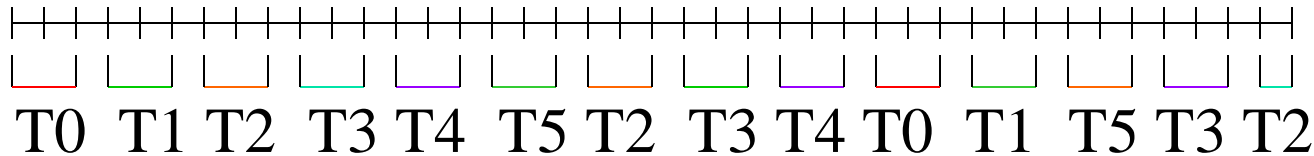

T0   T1   T2   T3   T4   T5
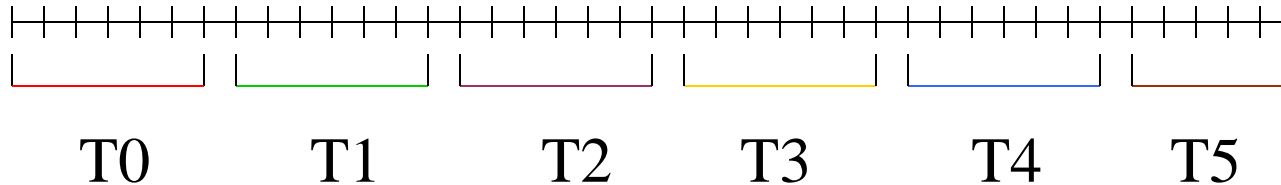
- Thread #0: iterations   0 through N_i/N_t-1
- Thread #1: iterations N_i/N_t through $2N_i/N_t$-1
- Thread #2: iterations $2N_i/N_t$ through $3N_i/N_t$-1

…

- Thread #N_t-1: iterations $(N_t$-1$)N_i/N_t$ through N_i-1

Dynamic Scheduling

For N_i iterations and N_t threads, each thread gets a fixed-size chunk of k loop iterations:

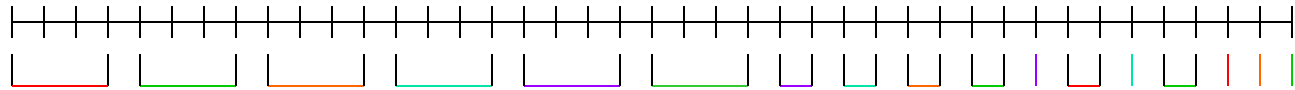T0 T1 T2 T3 T4 T5 T2 T3 T4 T0 T1 T5 T3 T2

When a particular thread finishes its chunk of iterations, it gets assigned a new chunk. So, the relationship between iterations and threads is nondeterministic.

- Advantage: very flexible
- Disadvantage: high overhead – lots of decision making about which thread gets each chunk

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Guided Scheduling

For N_i iterations and N_t threads, initially each thread gets a fixed-size chunk of $k < N_i/N_t$ loop iterations:

T0   T1   T2   T3   T4   T5 2 3 4 1 0 2 5 4 2 3 1

After each thread finishes its chunk of k iterations, it gets a chunk of $k/2$ iterations, then $k/4$, etc. Chunks are assigned dynamically, as threads finish their previous chunks.

- Advantage over static: can handle imbalanced load
- Advantage over dynamic: fewer decisions, so less overhead

How to Know Which Schedule?

Test all three using a typical case as a ***benchmark***.

Whichever wins is probably the one you want to use most of the time on that particular platform.

This may vary depending on problem size, new versions of the compiler, who's on the machine, what day of the week it is, etc, so you may want to benchmark the three schedules from time to time.

SCHEDULE Clause

The **PARALLEL DO** directive allows a **SCHEDULE** clause to be appended that tell the compiler which variables are shared and which are private:

!\$OMP PARALLEL DO ... **SCHEDULE(STATIC)**

This tells that compiler that the schedule will be static.

Likewise, the schedule could be **GUIDED** or **DYNAMIC**.

However, the very best schedule to put in the **SCHEDULE** clause is **RUNTIME**.

You can then set the environment variable **OMP_SCHEDULE** to **STATIC** or **GUIDED** or **DYNAMIC** at runtime – great for benchmarking!

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Synchronization

Jargon: Waiting for other threads to finish a parallel loop (or other parallel section) before going on to the work after the parallel section is called ***synchronization***.

Synchronization is **BAD**, because when a thread is waiting for the others to finish, it isn't getting any work done, so it isn't contributing to speedup.

So why would anyone ever synchronize?

Why Synchronize? (F90)

Synchronizing is necessary when the code that follows a parallel section needs all threads to have their final answers.

```fortran
!$OMP PARALLEL DO
  DO index = 1, length
    x(index) = index / 1024.0
    IF ((index / 1000) < 1) THEN
      y(index) = LOG(x(index))
    ELSE
      y(index) = x(index) + 2
    END IF
  END DO
! Need to synchronize here!
  DO index = 1, length
    z(index) = y(index) + y(length - index + 1)
  END DO
```

Why Synchronize?

Synchronizing is necessary when the code that follows a parallel
section needs all threads to have their final answers.

```
#pragma omp parallel for
  for (index = 0; index < length; index++) {
    x[index] = index / 1024.0;
    if ((index / 1000) < 1) {
      y[index] = log(x[index]);
    }
    else {
      y[index] = x[index] + 2;
    }
  }
/* Need to synchronize here! */
  for (index = 0; index < length; index++) {
    z[index] = y[index] + y[length - index + 1];
  }
```

Sponsored by
DOD HPCMP,
SC11/ACM,
NCSI and
OK EPSCoR

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

86

Barriers

A ***barrier*** is a place where synchronization is forced to occur; that is, where faster threads have to wait for slower ones.

The **PARALLEL DO** directive automatically puts an invisible, implied barrier at the end of its **DO** loop:

```
!$OMP PARALLEL DO
  DO index = 1, length
    ... parallel stuff ...
  END DO
! Implied barrier
  ... serial stuff ...
```

OpenMP also has an explicit **BARRIER** directive, but most people don't need it.

Critical Sections

A ***critical section*** is a piece of code that any thread can execute, but that only one thread can execute at a time.

```
!$OMP PARALLEL DO
  DO index = 1, length
    … parallel stuff …
!$OMP CRITICAL(summing)
    sum = sum + x(index) * y(index)
!$OMP END CRITICAL(summing)
    … more parallel stuff …
  END DO
```

What's the point?

Why Have Critical Sections?

If only one thread at a time can execute a critical section, that slows the code down, because the other threads may be waiting to enter the critical section.

But, for certain statements, if you don't ensure ***mutual exclusion***, then you can get nondeterministic results.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

If No Critical Section

```
!$OMP CRITICAL(summing)
    sum = sum + x(index) * y(index)
!$OMP END CRITICAL(summing)
```

Suppose for thread #0, **index** is 27, and for thread #1, **index** is 92.

If the two threads execute the above statement at the same time, **sum** could be

- the value after adding **x(27) * y(27)**, or
- the value after adding **x(92) * y(92)**, or
- garbage!

This is called a *__race condition__*: the result depends on who wins the race.

Pen Game #1: Take the Pen

We need two volunteers for this game.

1. I'll hold a pen in my hand.

2. You win by taking the pen from my hand.

3. One, two, three, go!

Can we predict the outcome? Therefore, can we guarantee that we get the correct outcome?

Pen Game #2: Look at the Pen

We need two volunteers for this game.

1. I'll hold a pen in my hand.

2. You win by looking at the pen.

3. One, two, three, go!

Can we predict the outcome? Therefore, can we guarantee that we get the correct outcome?

Race Conditions

A ***race condition*** is a situation in which multiple processes can change the value of a variable at the same time.

As in Pen Game #1 (Take the Pen), a race condition can lead to unpredictable results.

So, race conditions are **BAD**.

Reductions

A *reduction* converts an array to a scalar: sum, product, minimum value, maximum value, location of minimum value, location of maximum value, Boolean AND, Boolean OR, number of occurrences, etc.

Reductions are so common, and so important, that OpenMP has a specific construct to handle them: the `REDUCTION` clause in a `PARALLEL DO` directive.

Reduction Clause

```
total_mass = 0
!$OMP PARALLEL DO REDUCTION(+:total_mass)
  DO index = 1, length
    total_mass = total_mass + mass(index)
  END DO !! index
```

This is equivalent to:

```
  DO thread = 0, number_of_threads - 1
    thread_mass(thread) = 0
  END DO !! thread
$OMP PARALLEL DO
  DO index = 1, length
    thread = omp_get_thread_num()
    thread_mass(thread) = thread_mass(thread) + mass(index)
  END DO !! index
  total_mass = 0
  DO thread = 0, number_of_threads - 1
    total_mass = total_mass + thread_mass(thread)
  END DO !! thread
```
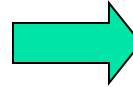
Parallelizing a Serial Code #1

```
PROGRAM big_science
  ... declarations ...

  DO ...
  ... parallelizable work ...
  END DO
... serial work ...

  DO ...
  ... more parallelizable work ...
  END DO
... serial work ...
... etc ...
END PROGRAM big_science
```

```
PROGRAM big_science
  ... declarations ...
  !$OMP PARALLEL DO …
  DO ...
  ... parallelizable work ...
  END DO
... serial work ...
  !$OMP PARALLEL DO …
  DO ...
  ... more parallelizable work ...
  END DO
... serial work ...
... etc ...
END PROGRAM big_science
```

This way may have lots of **synchronization** overhead.

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Parallelizing a Serial Code #2

```fortran
PROGRAM big_science
  ... declarations ...

  DO task = 1, numtasks
    CALL science_task(…)
  END DO
END PROGRAM big_science


SUBROUTINE science_task (…)
 ... parallelizable work ...

 ... serial work ...

 ... more parallelizable work ...

 ... serial work ...

 ... etc ...
END PROGRAM big_science
```
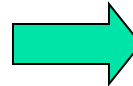
```fortran
PROGRAM big_science
  ... declarations ...
!$OMP PARALLEL DO …
  DO task = 1, numtasks
    CALL science_task(…)
  END DO
END PROGRAM big_science


SUBROUTINE science_task (…)
 ... parallelizable work ...
!$OMP MASTER
 ... serial work ...
!$OMP END MASTER
 ... more parallelizable work ...
!$OMP MASTER
 ... serial work ...
!$OMP END MASTER
 ... etc ...
END PROGRAM big_science
```

Shared Memory Multithreading
Intermediate Parallel, July 31 – Aug 6 2011

Thanks for your attention!

Questions?

References

[1] Amdahl, G.M. "Validity of the single-processor approach to achieving large scale computing capabilities." In *AFIPS Conference Proceedings* vol. 30 (Atlantic City, N.J., Apr. 18-20). AFIPS Press, Reston VA, 1967, pp. 483-485. Cited in http://www.scl.ameslab.gov/Publications/AmdahlsLaw/Amdahls.html

[2] http://www.iso.org/iso/about/discover-iso_isos-name.htm

[3] R. Chandra, L. Dagum, D. Kohr, D. Maydan, J. McDonald and R. Menon, *Parallel Programming in OpenMP*. Morgan Kaufmann, 2001.

[4] Kevin Dowd and Charles Severance, *High Performance Computing,* 2nd ed. O'Reilly, 1998.