



# An Introduction to Parallel Programming with OpenMP

**Tim Mattson**  
Intel Corp.

Download tutorial materials onto your laptop:  
git clone <https://github.com/tgmattso/ParProgForPhys.git>

# Introduction

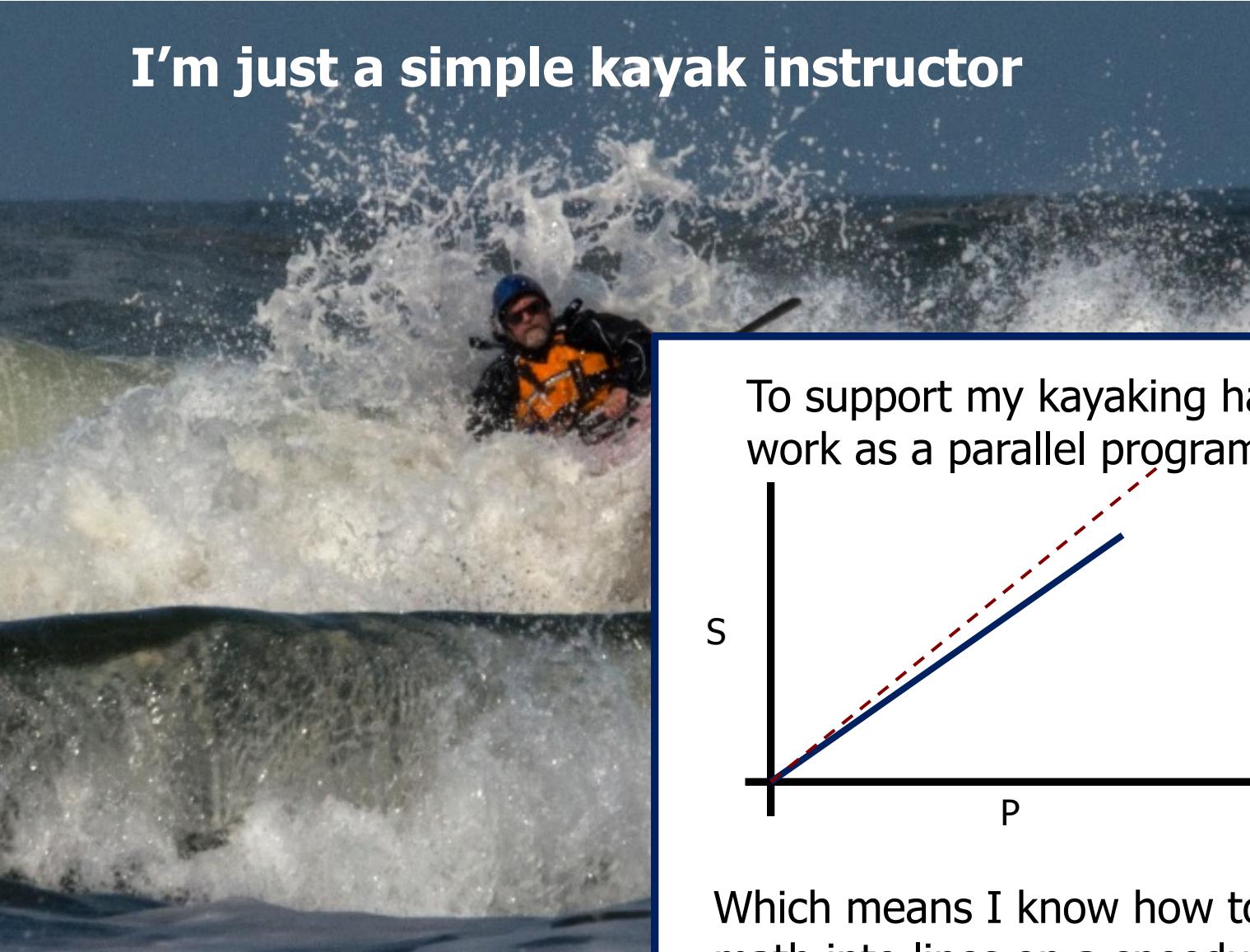
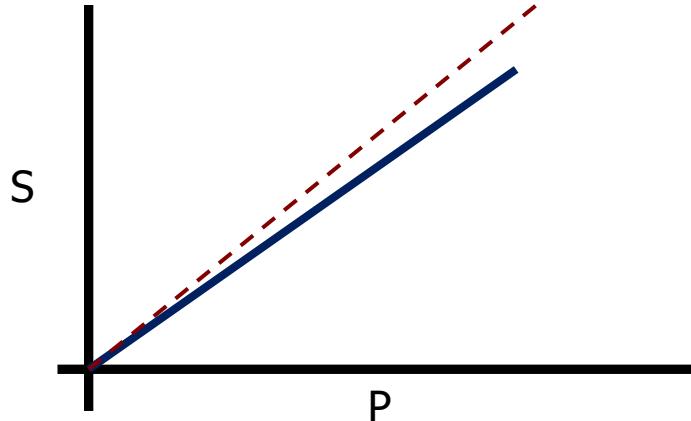


Photo © by Greg Clopton, 2014

To support my kayaking habit, I work as a parallel programmer



Which means I know how to turn math into lines on a speedup plot

# Preliminaries: Part 1

- Disclosures
  - The views expressed in this tutorial are those of the people delivering the tutorial.
    - We are not speaking for our employers.
    - We are not speaking for the OpenMP ARB
- We take these tutorials VERY seriously:
  - Help us improve ... tell us how you would make this tutorial better.

Download tutorial materials onto your laptop:  
git clone <https://github.com/tgmattso/ParProgForPhys.git>

# Preliminaries: Part 2

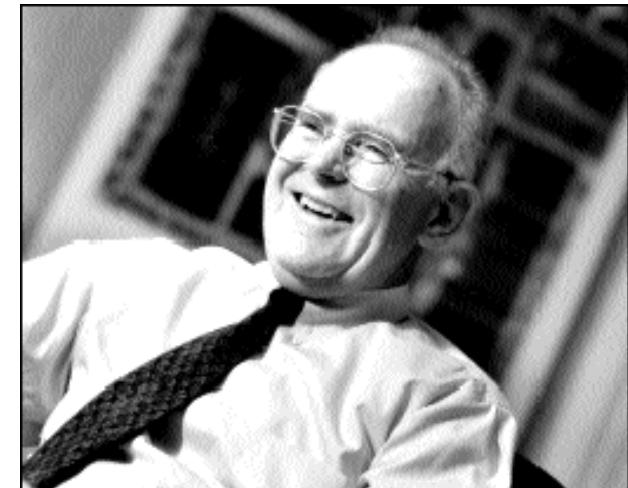
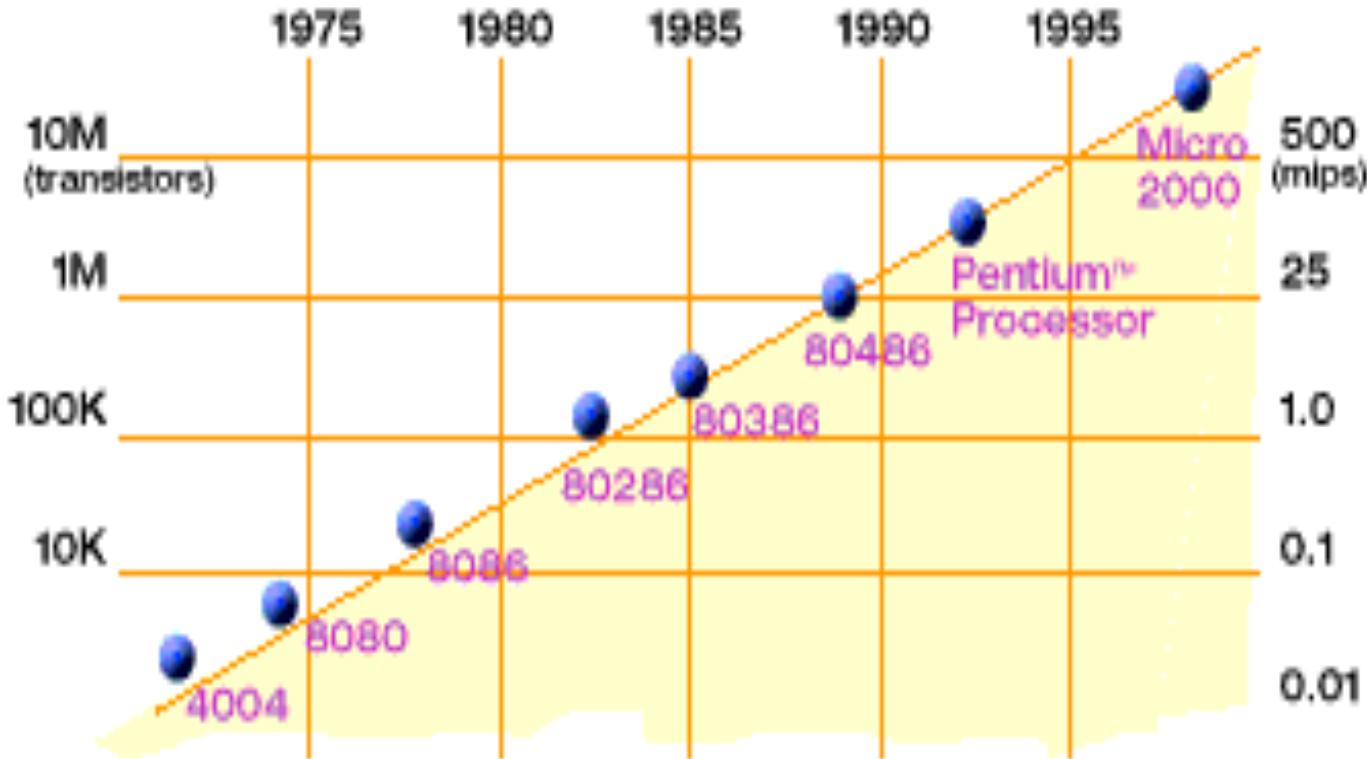
- Our plan ... Active learning!
  - We will mix short lectures with short exercises.
  - You will use your laptop to connect to a multiprocessor server.
- Please follow these simple rules
  - Do the exercises that we assign and then change things around and experiment.
  - Embrace active learning!
  - Don't cheat: Do **Not** look at the solutions before you complete an exercise ... even if you get really frustrated.

Download tutorial materials onto your laptop:  
git clone <https://github.com/tgmattso/ParProgForPhys.git>

# Outline

- ➡ • Introduction to parallel computing and OpenMP
  - Creating Threads
  - Synchronization
  - Parallel Loops
  - Data Environment
  - Memory Model
  - Irregular Parallelism and Tasks
  - Recap
  - Beyond the Common Core:
    - Worksharing Revisited
    - Synchronization Revisited: Options for Mutual exclusion
    - Memory models and point-to-point Synchronization
    - Programming your GPU with OpenMP
    - Thread Affinity and Data Locality
    - Thread Private Data

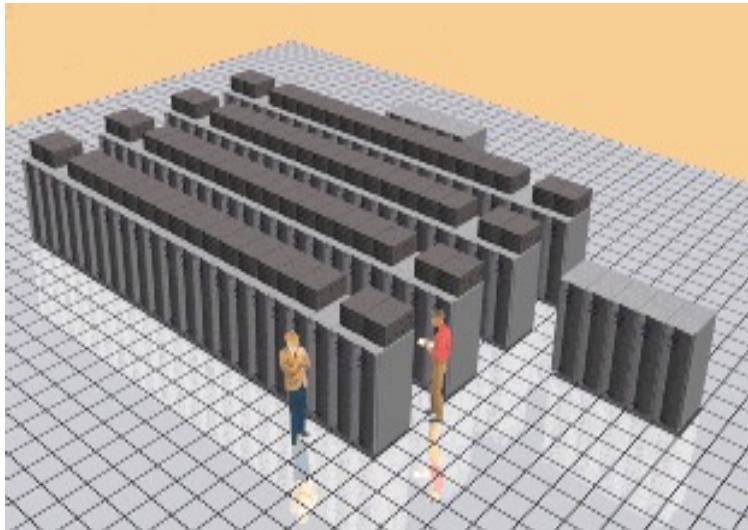
# Moore's Law



- In 1965, Intel co-founder Gordon Moore predicted (from just 3 data points!) that semiconductor density would double every 18 months.
  - ***He was right!*** Over the last 50 years, transistor densities have increased as he predicted.

# Moore's Law: A personal perspective

First TeraScale\* computer: 1997



Intel's ASCI Option Red

## Intel's ASCI Red Supercomputer

9000 CPUs

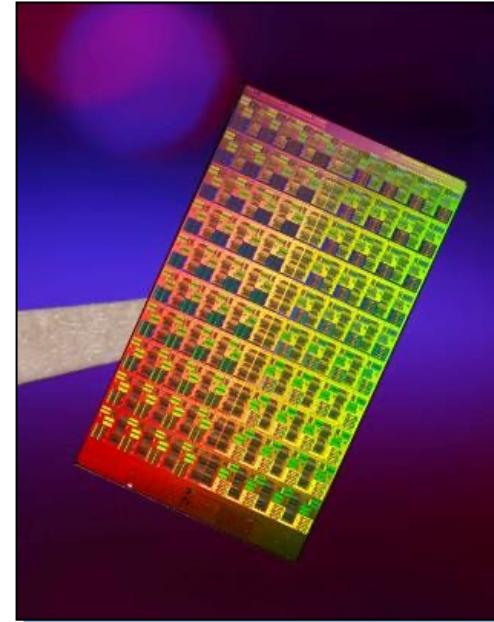
one megawatt of electricity.

1600 square feet of floor space.

\*Double Precision TFLOPS running MP-Linpack

A TeraFLOP in 1996: The ASCI TeraFLOP Supercomputer,  
Proceedings of the International Parallel Processing  
Symposium (1996), T.G. Mattson, D. Scott and S. Wheat.

First TeraScale% chip: 2007



## Intel's 80 core teraScale Chip

1 CPU

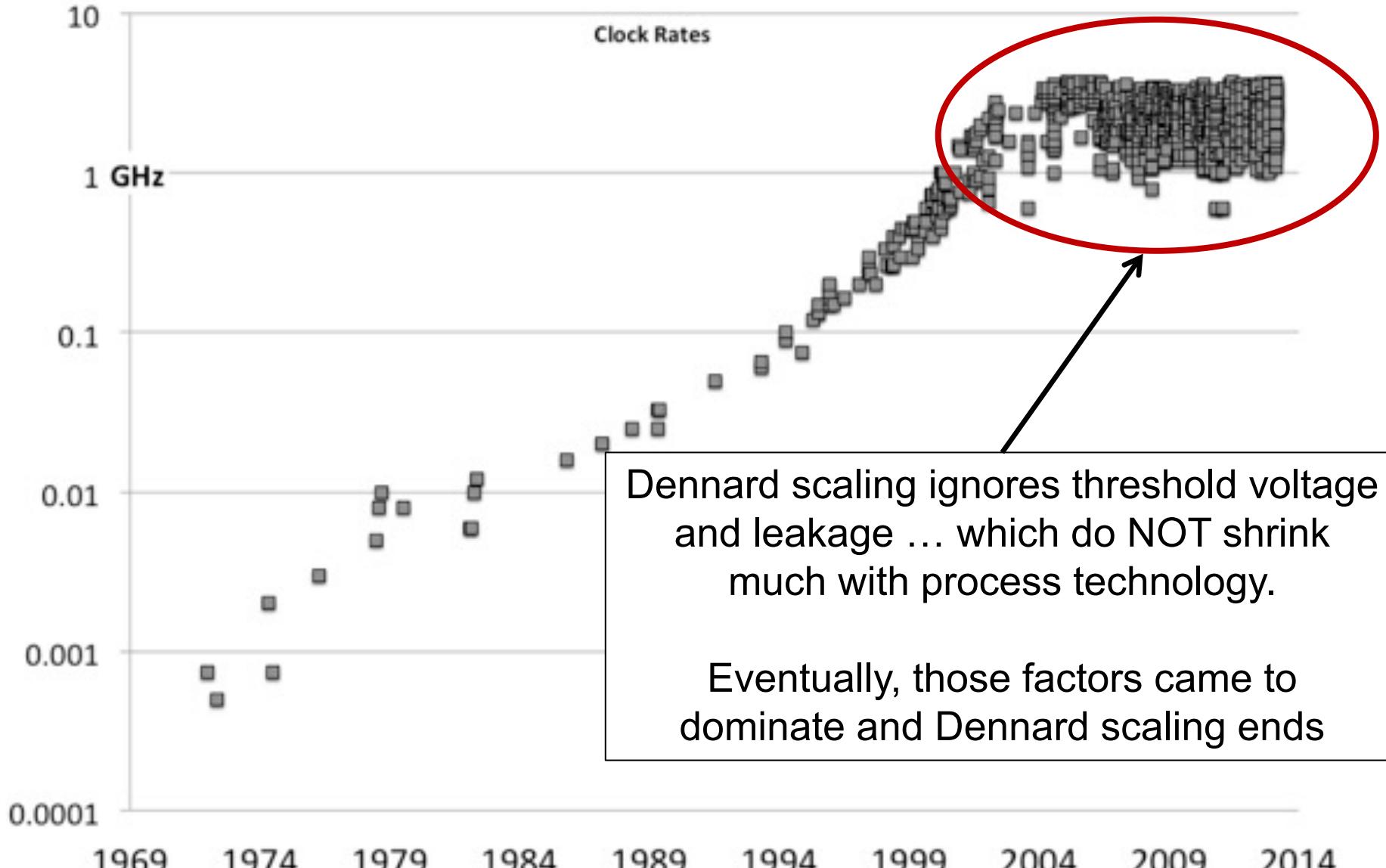
97 watt

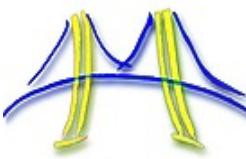
275 mm<sup>2</sup>

%Single Precision TFLOPS running stencil

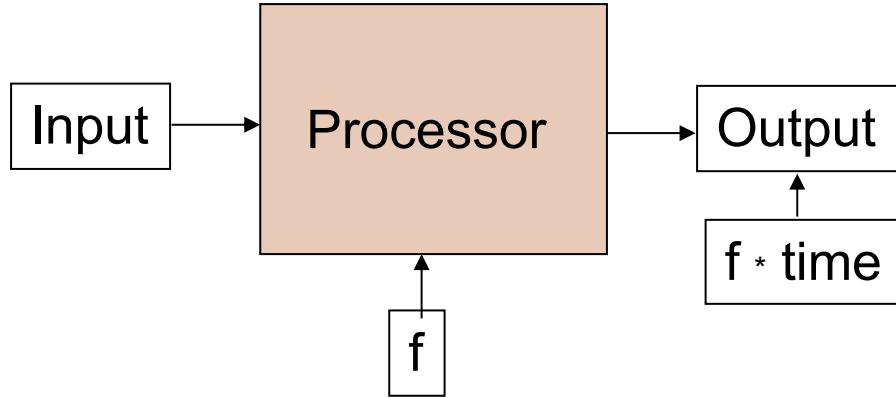
Programming Intel's 80 core terascale processor  
SC08, Austin Texas, Nov. 2008, Tim Mattson,  
Rob van der Wijngaart, Michael Frumkin

# CPU Frequency (GHz) over time (years)





# Consider power in a chip ...



Capacitance = C  
Voltage = V  
Frequency = f  
Power =  $CV^2f$

C = capacitance ... it measures the ability of a circuit to store energy:

$$C = q/V \rightarrow q = CV$$

Work is pushing something (charge or q) across a “distance” ... in electrostatic terms pushing q from 0 to V:

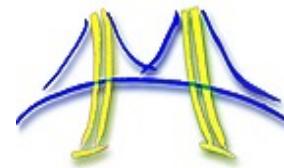
$$V * q = W.$$

But for a circuit  $q = CV$  so

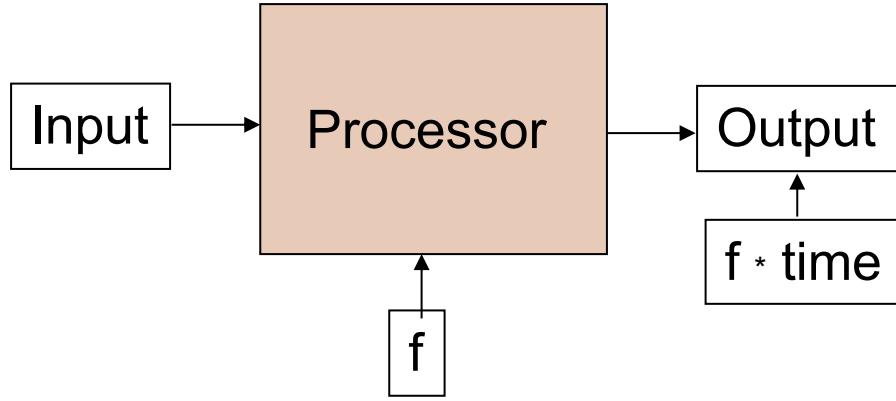
$$W = CV^2$$

power is work over time ... or how many times per second we oscillate the circuit

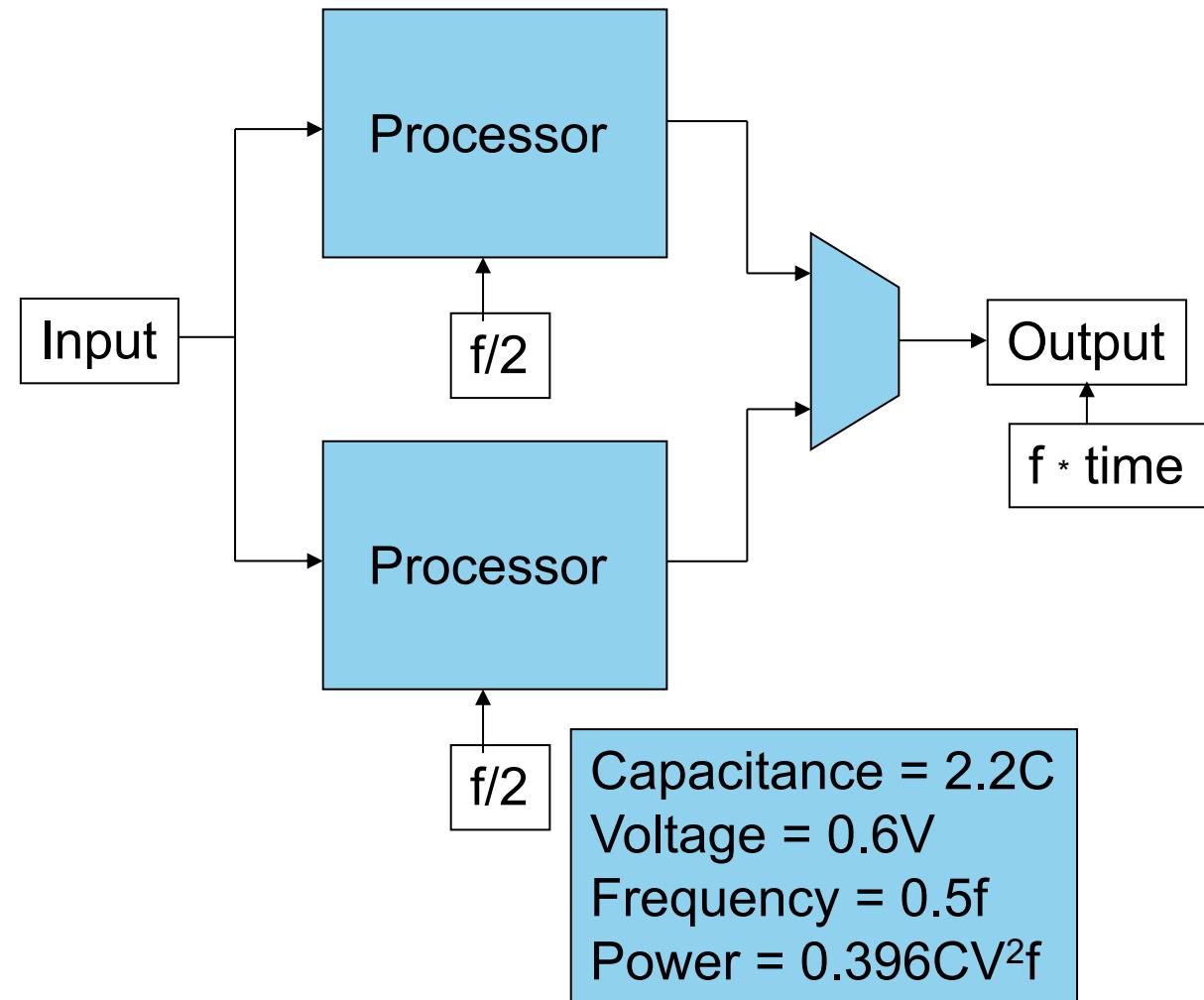
$$\text{Power} = W * F \rightarrow \text{Power} = CV^2f$$



# ... Reduce power by adding cores



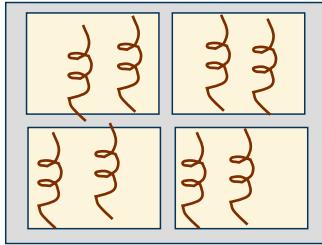
Capacitance = C  
Voltage = V  
Frequency = f  
Power =  $CV^2f$



Capacitance =  $2.2C$   
Voltage =  $0.6V$   
Frequency =  $0.5f$   
Power =  $0.396CV^2f$

# For hardware ... parallelism is the path to performance

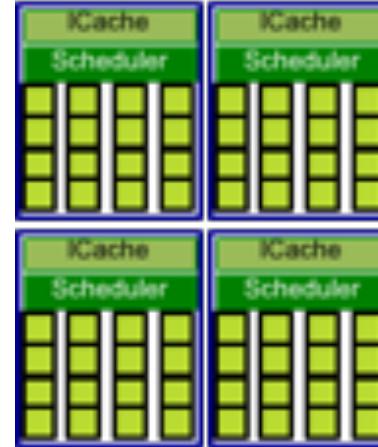
All hardware vendors are in the game ... parallelism is ubiquitous so if you care about getting the most from your hardware, you will need to create parallel software.



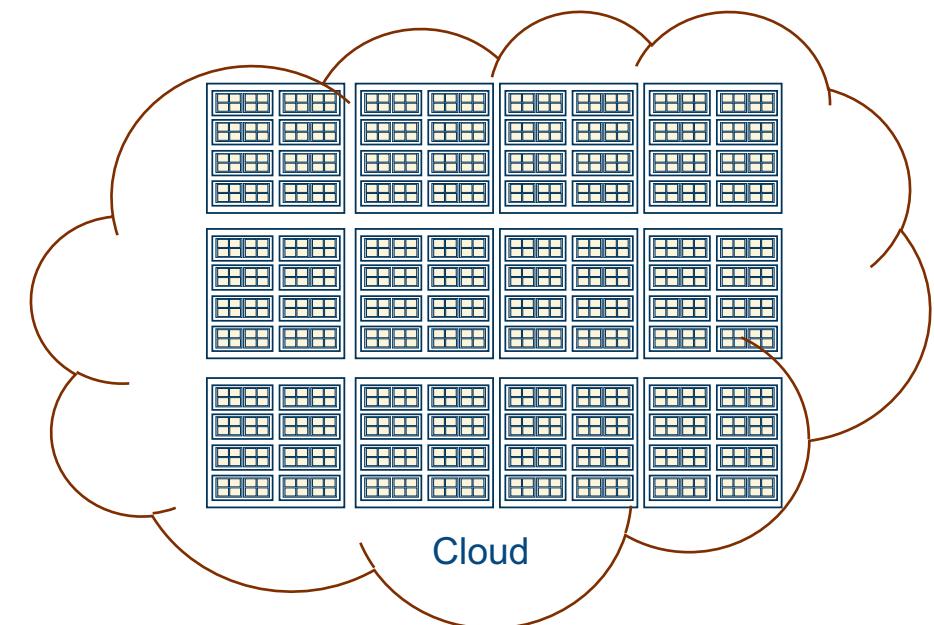
CPU



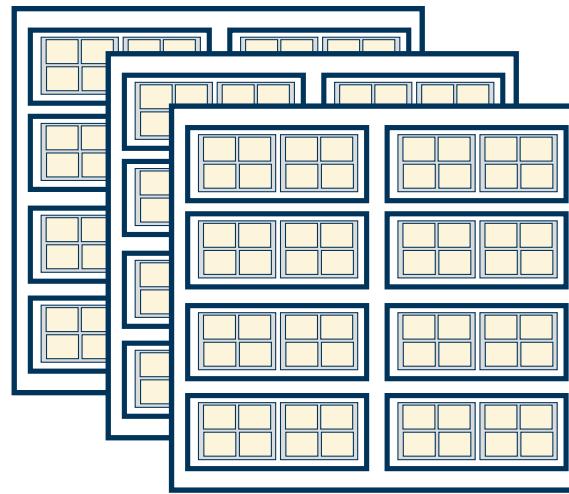
SIMD/Vector



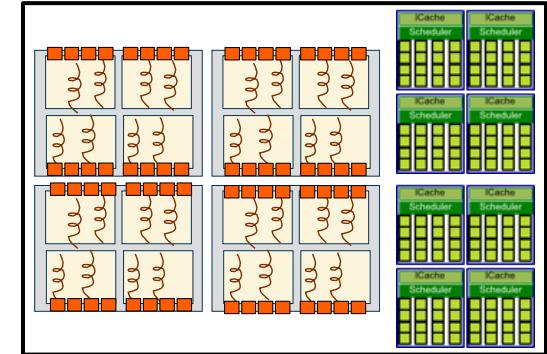
GPU



Cloud



Cluster



Heterogeneous node

# We'll cover programming for the major categories of parallel systems (**shared memory**, **cluster** and **GPU**)

| duration | US EDT | August 22 Monday   | August 23 Tuesday                                 | August 24 Wednesday                                  | August 25 Thursday             | August 26 Friday              | August 27 Saturday |
|----------|--------|--|---|--|--------------------------------|-------------------------------|--------------------|
| 90'      | 9      | Welcome<br>(Cristina/Paul/Dennis)                                | Parallel Programming OpenMP<br>T. Mattson (Intel) | Parallel Programming MPI<br>T. Mattson (Intel)       | Surrogate modeling<br>and AI   | Introduction to GPU computing |                    |
|          | 9:30   | Lightning Talks  |   |  |                                |                               |                    |
| 30'      | 10:30  | Coffee Break   | Coffee Break                                      | Coffee Break   |                                |                               |                    |
| 90'      | 11     | Practical HPC, architectures and workflows<br>A. Reuther (MITII) | Parallel Programming OpenMP<br>T. Mattson (Intel) | Open source and best practices<br>N. Murphy (CfA)    | Hardware                       |                               |                    |
|          | 12:30  | Lunch  | Lunch   | Lunch  |                                |                               |                    |
| 120'     | 13:30  | Intro to PSFC/MIT computing<br>J. Wright (PSFC)                  |   |  | Surrogate modeling and AI      |                               |                    |
|          | 14:00  | Intro to Parallel Programming OpenMP/MPI<br>T. Mattson (Intel)   | Parallel Programming MPI<br>T. Mattson (Intel)    | GPU computing and accelerators<br>T. Mattson (Intel) |                                |                               |                    |
| 30'      | 15:30  | Coffee Break   | Coffee Break                                      | Coffee Break   |                                |                               |                    |
| 90'      | 16     | John & Pablo (PSFC)<br>RF+transport/AI/num modeling              | W7X/AI/modeling to be confirmed                   | Anshu Dubay (ANL)                                    | GPU computing and accelerators |                               |                    |
| 60'      | 17:30  | SPARC - CFS Q&A  |   |  | Session 1                      |                               |                    |

This schedule is my best guess for how we will proceed.

This course is “student driven” not “presenter driven”. The material I cover and how much time I spend on any topic depends on you and your progress with the material.

So don’t be surprised if I make big changes to the schedule. For example, I may squeeze MPI a bit to give us time to fully cover OpenMP and to make room (if I think we’ll need it) for more time with the GPU.

**The best way to master parallel computing ...**

**start with a simple approach to parallelism and build  
an intellectual foundation by writing parallel code.**

**... and the simplest API for parallelism is?**

# OpenMP\* Overview

C\$OMP FLUSH

#pragma omp critical

#pragma omp single

C\$OMP THREADPRIVATE (/ABC/)

C\$OMP ATOMIC

CALL OMP\_SET\_NUM\_THREADS(10)

## *OpenMP: An API for Writing Parallel Applications*

cal

- A set of compiler directives and library routines for parallel application programmers
- Greatly simplifies writing multi-threaded (MT) programs in Fortran, C and C++
- Also supports non-uniform memories, vectorization and GPU programming

RED

#pragma omp parallel for private(A, B)

C\$OMP PARALLEL REDUCTION (+: A, B)

C\$OMP PARALLEL COPYIN(/blk/)

C\$OMP DO lastprivate(XX)

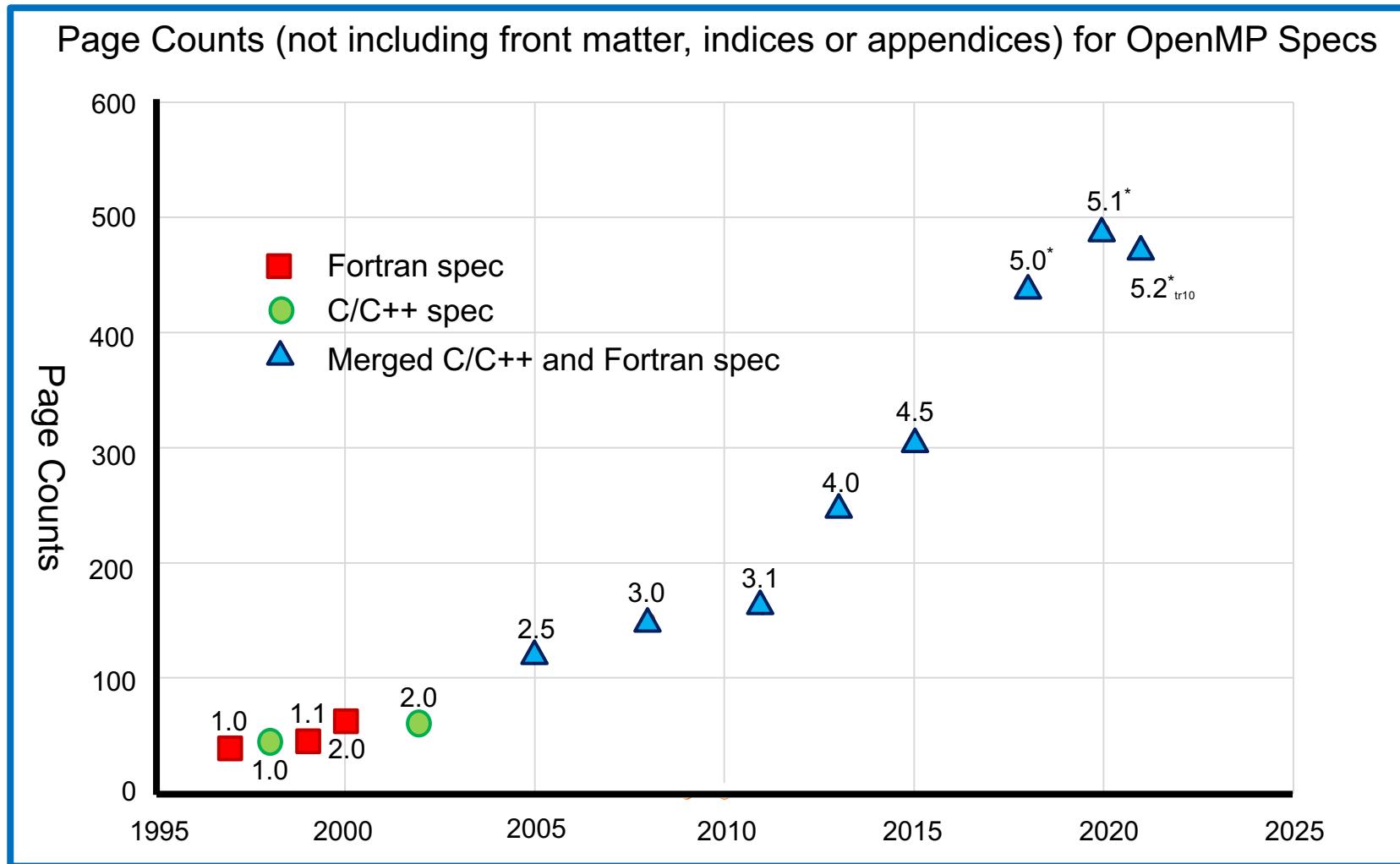
#pragma omp atomic seq\_cst

Nthrds = OMP\_GET\_NUM\_PROCS()

omp\_set\_lock(lck)

# The Growth of Complexity in OpenMP

Our goal in 1997 ... A simple interface for application programmers

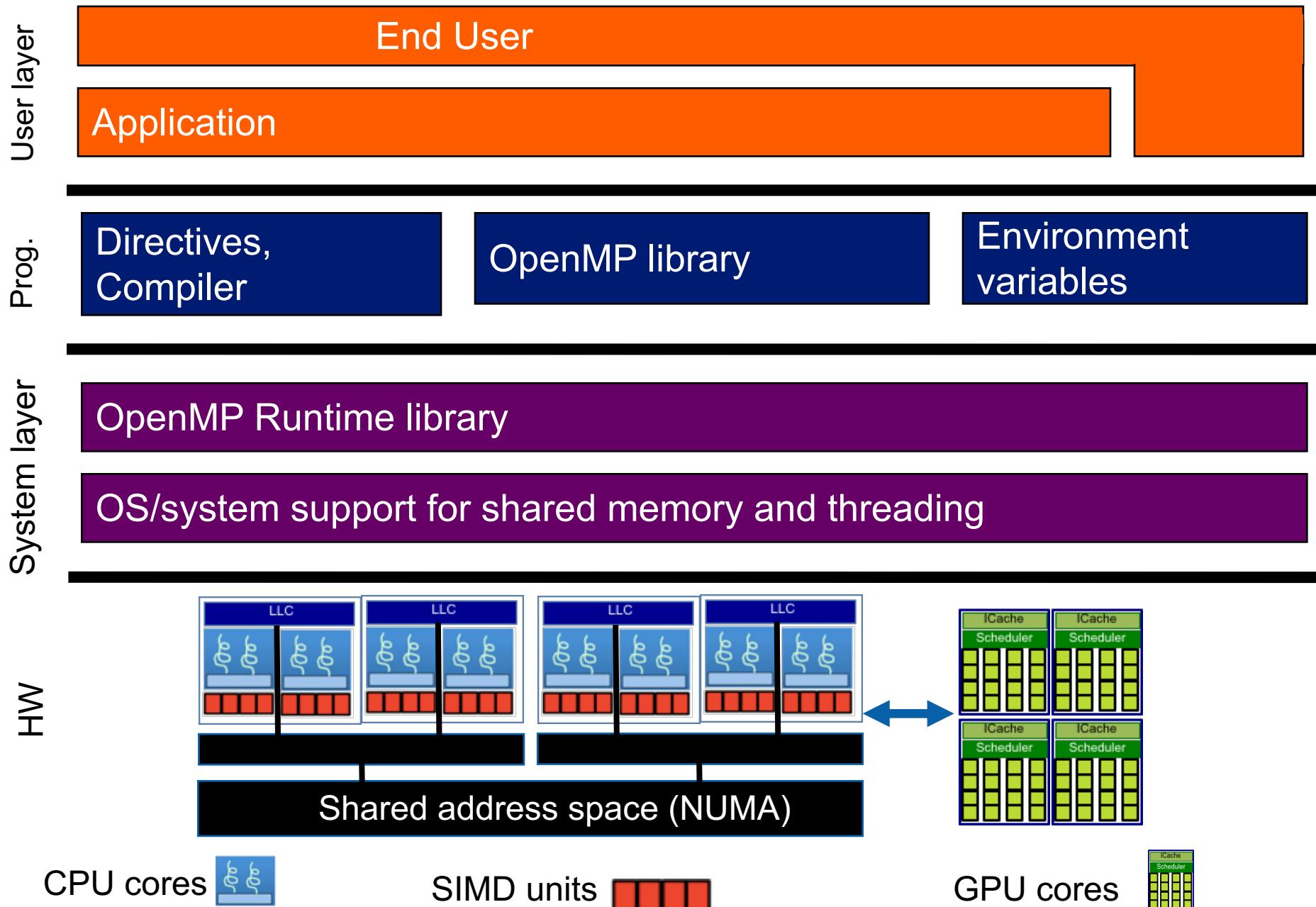


The full spec is overwhelming. We focus on the Common Core: the 21 items most people restrict themselves to

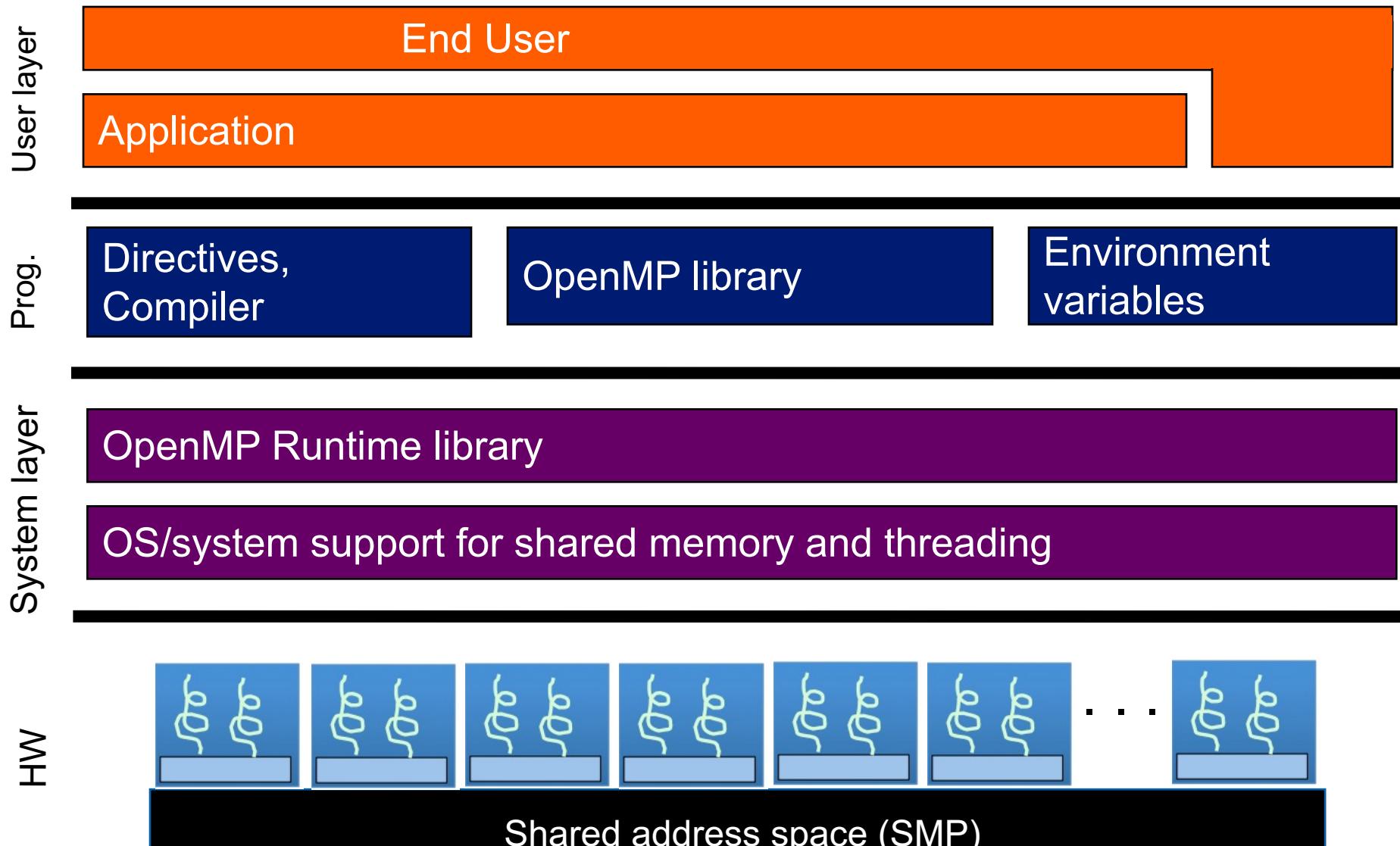
# The OpenMP Common Core: Most OpenMP programs only use these 21 items

| OpenMP pragma, function, or clause   | Concepts  |
|--|---|
| #pragma omp parallel   | Parallel region, teams of threads, structured block, interleaved execution across threads.  |
| void omp_set_thread_num()<br>int omp_get_thread_num()<br>int omp_get_num_threads() | Default number of threads and internal control variables.<br>SPMD pattern: Create threads with a parallel region and split up the work using the number of threads and the thread ID. |
| double omp_get_wtime()   | Speedup and Amdahl's law. False sharing and other performance issues.   |
| setenv OMP_NUM_THREADS N   | Setting the internal control variable for the default number of threads with an environment variable  |
| #pragma omp barrier<br>#pragma omp critical  | Synchronization and race conditions.<br>Revisit interleaved execution.  |
| #pragma omp for<br>#pragma omp parallel for  | Worksharing, parallel loops, loop carried dependencies.   |
| reduction(op:list)   | Reductions of values across a team of threads.  |
| schedule (static [,chunk])<br>schedule(dynamic [,chunk])                           | Loop schedules, loop overheads, and load balance.   |
| shared(list), private(list), firstprivate(list)                                    | Data environment.   |
| default(None)  | Force explicit definition of each variable's storage attribute  |
| nowait   | Disabling implied barriers on workshare constructs, the high cost of barriers, and the flush concept (but not the flush directive).   |
| #pragma omp single   | Workshare with a single thread.   |
| #pragma omp task<br>#pragma omp taskwait   | Tasks including the data environment for tasks.   |

# OpenMP Basic Definitions: Basic Solution Stack



# OpenMP Basic Definitions: Basic Solution Stack



For the OpenMP Common Core, we focus on Symmetric Multiprocessor Case ....  
i.e., lots of threads with “equal cost access” to memory

# OpenMP Basic Syntax

- Most of the constructs in OpenMP are compiler directives.

| C and C++  | Fortran   |
|--|---|
| Compiler directives                                  |   |
| <b>#pragma omp construct [clause [clause]...]</b>    | <b>!\$OMP construct [clause [clause] ...]</b>                       |
| Example  |   |
| <b>#pragma omp parallel private(x)</b><br>{<br><br>} | <b>!\$OMP PARALLEL PRIVATE(X)</b><br><br><b>!\$OMP END PARALLEL</b> |
| Function prototypes and types:                       |   |
| <b>#include &lt;omp.h&gt;</b>                        | <b>use OMP_LIB</b>  |

- Most OpenMP constructs apply to a “structured block”.
  - Structured block: a block of one or more statements with one point of entry at the top and one point of exit at the bottom.
  - It’s OK to have an exit() within the structured block.

# Exercise, Part A: Hello World

## Verify that your environment works

- Write a program that prints “hello world”.

git clone <https://github.com/tgmattso/OpenMPCommonCore.git>

```
#include<stdio.h>
int main()
{
    printf(" hello ");
    printf(" world \n");
}
```

Download tutorial materials onto your laptop:

git clone <https://github.com/tgmattso/ParProgForPhys.git>

# Exercise, Part B: Hello World

## Verify that your OpenMP environment works

- Write a multithreaded program that prints “hello world”.

```
git clone https://github.com/tgmattso/OpenMP_Common_Core.git
```

```
#include <omp.h>
#include <stdio.h>
int main()
{
    #pragma omp parallel
    {
        printf(" hello ");
        printf(" world \n");
    }
}
```

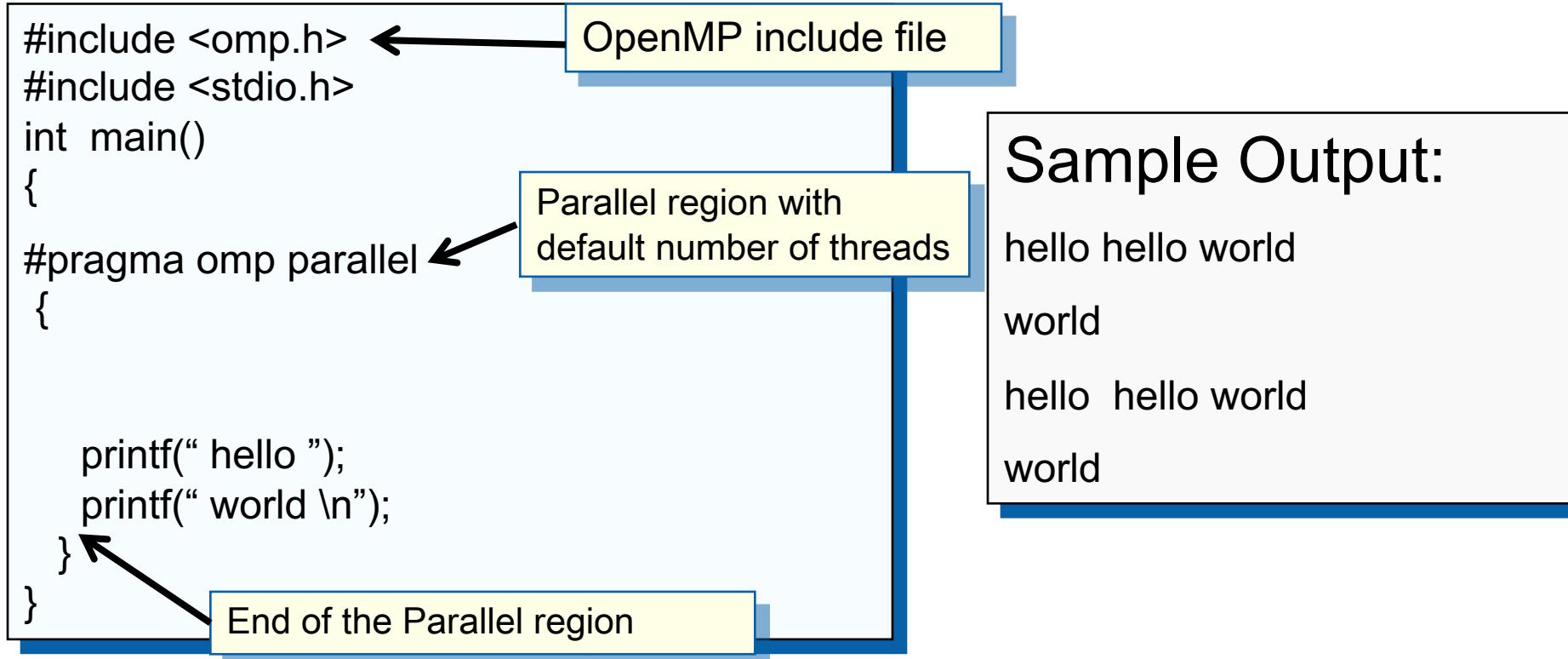
### Switches for compiling and linking

|              |                     |
|--------------|---------------------|
| gcc -fopenmp | Gnu (Linux, OSX)    |
| cc -qopenmp  | Intel (Linux@NERSC) |
| icl /Qopenmp | Intel (windows)     |
| icc -fopenmp | Intel (Linux, OSX)  |

# Solution

## A Multi-Threaded “Hello World” Program

- Write a multithreaded program where each thread prints “hello world”.



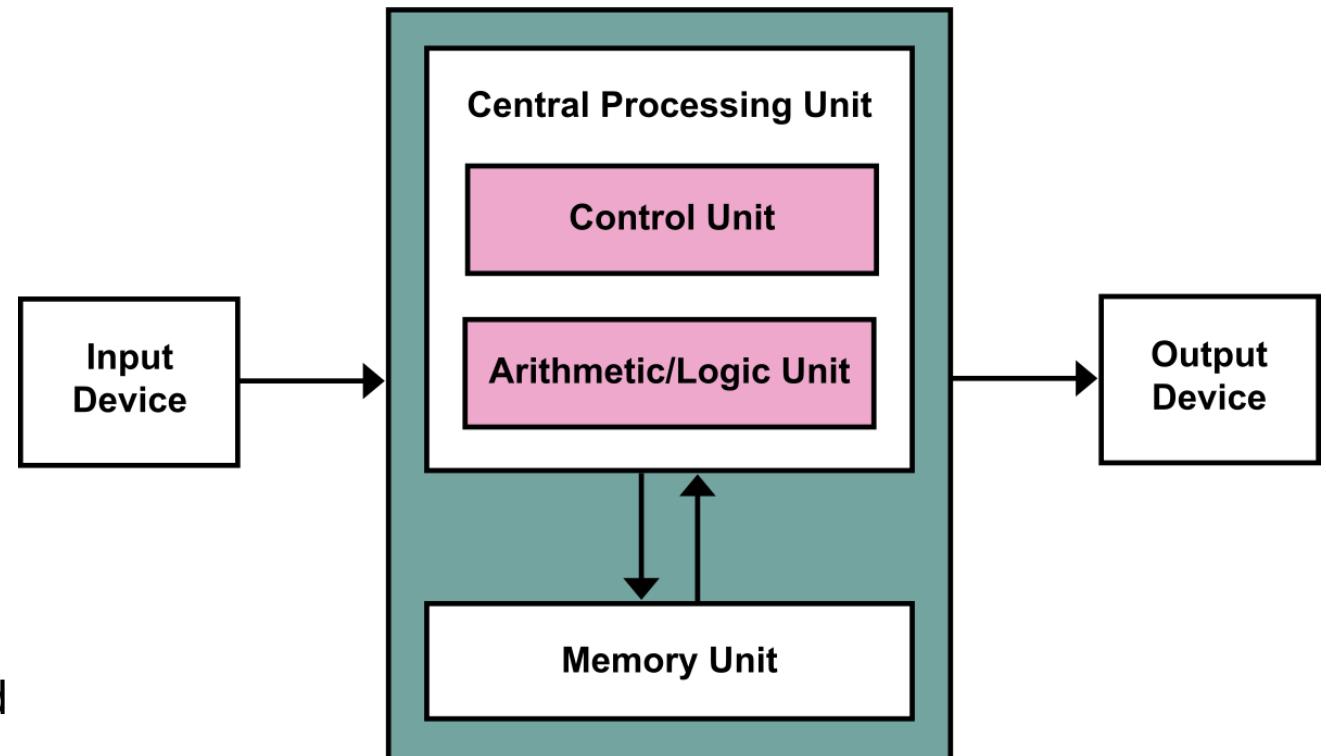
The statements are interleaved based on how the operating system schedules the threads

# **A brief digression on the terminology of parallel computing**

# Let's agree on a few definitions:

- **Computer:**

- A machine that transforms *input data* into *output data*.
- Typically, a computer consists of Control, Arithmetic/Logic, and Memory units.
- The transformation is defined by a stored **program** (von Neumann architecture).



- **Task:**

- A specific sequence of instructions plus a data environment. A program is composed of one or more tasks.

- **Active task:**

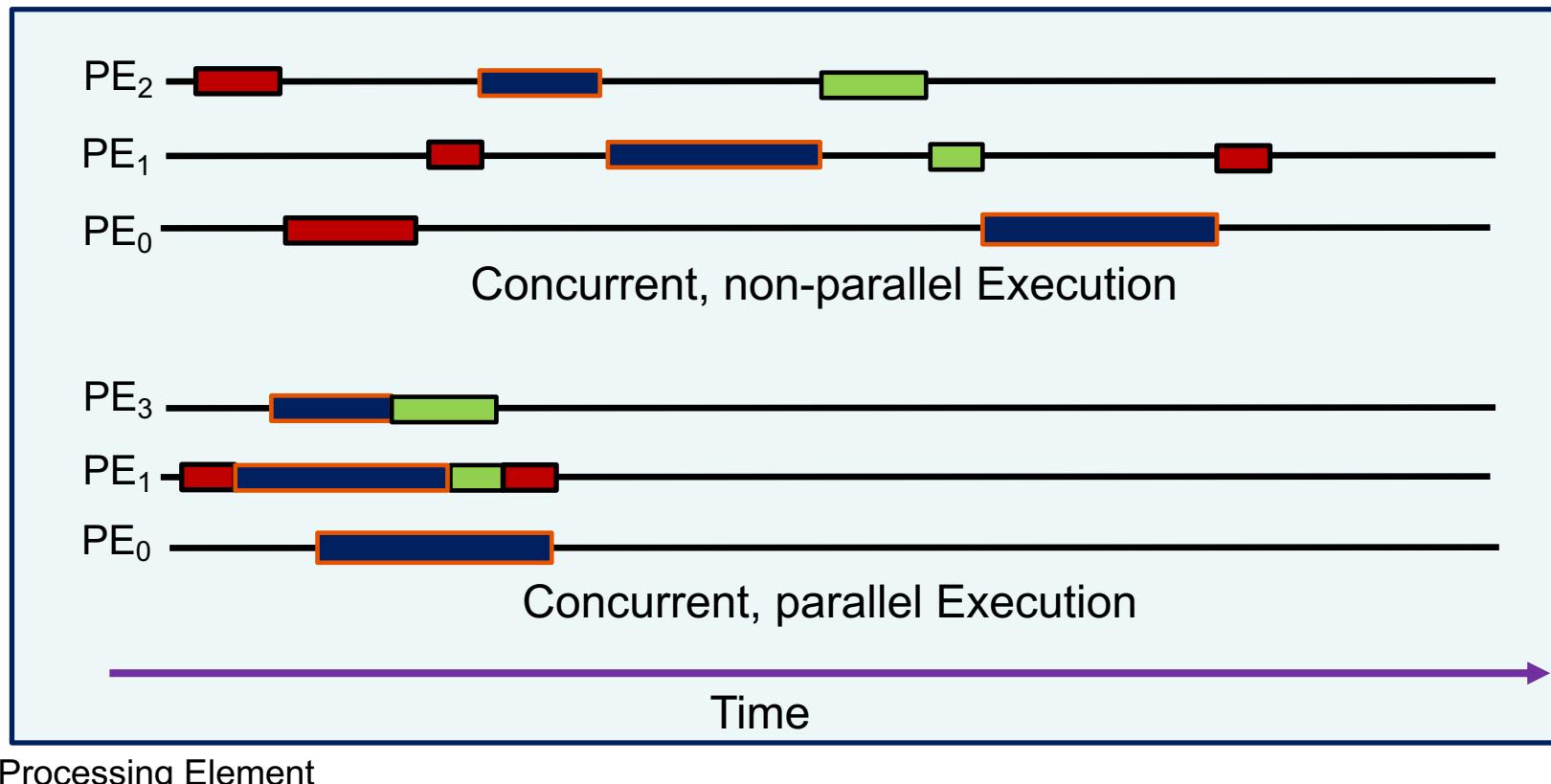
- A task that is available to be scheduled for execution. When the task is moving through its sequence of instructions, we say it is making **forward progress**

- **Fair scheduling:**

- When a scheduler gives each active task an equal *opportunity* for execution.

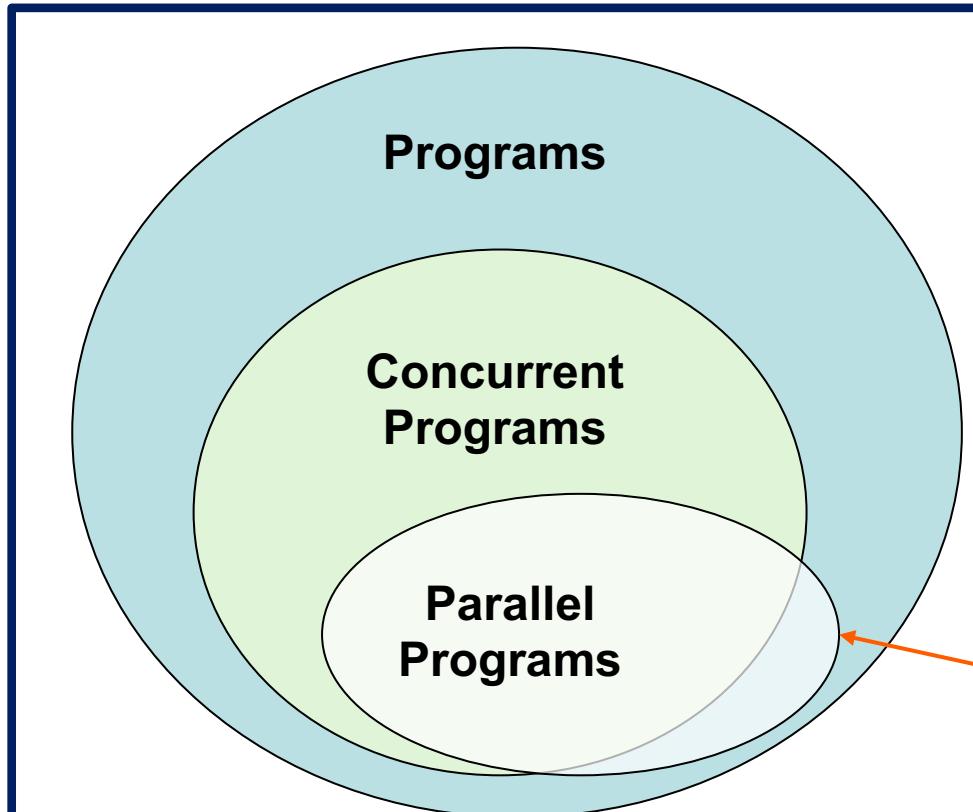
# Concurrency vs. Parallelism

- Two important definitions:
  - Concurrency: A condition of a system in which multiple tasks are active and unordered. If **scheduled fairly**, they can be described as logically making **forward progress** at the same time.
  - Parallelism: A condition of a system in which multiple tasks are actually making **forward progress** at the same time.



# Concurrency vs. Parallelism

- Two important definitions:
  - Concurrency: A condition of a system in which multiple tasks are active and unordered. If **scheduled fairly**, they can be described as logically making **forward progress** at the same time.
  - Parallelism: A condition of a system in which multiple tasks are actually making **forward progress** at the same time.



In most cases, parallel programs exploit concurrency in a problem to run tasks on multiple processing elements

We use Parallelism to:

- Do more work in less time
- Work with larger problems

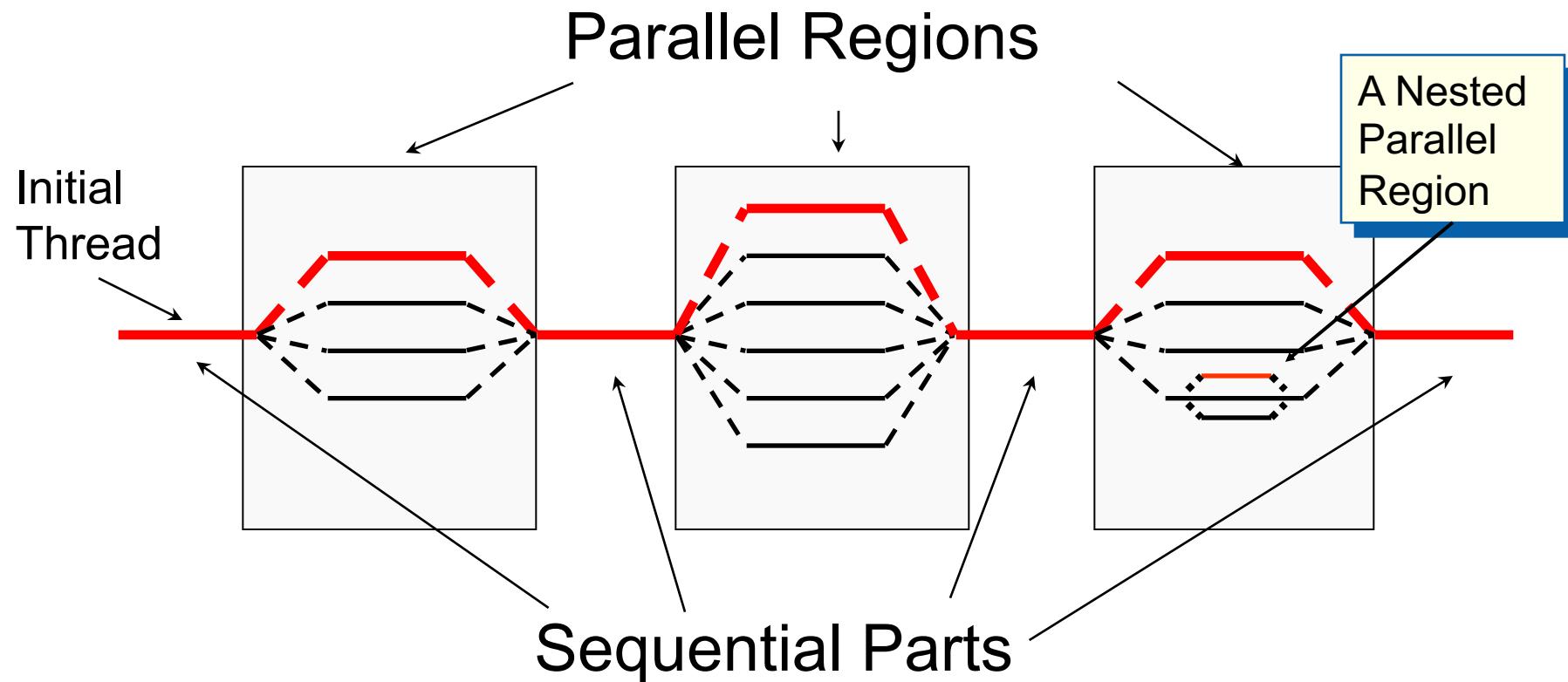
If tasks execute in “lock step” they are not concurrent, but they are still parallel.  
Example ... a SIMD unit.

- Introduction to OpenMP
- • Creating Threads
- Synchronization
- Parallel Loops
- Data Environment
- Memory Model
- Irregular Parallelism and Tasks
- Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data

# OpenMP Execution model:

## Fork-Join Parallelism:

- ◆ Initial thread spawns a team of threads as needed.
- ◆ Parallelism added incrementally until performance goals are met, i.e., the sequential program evolves into a parallel program.



# Thread Creation: Parallel Regions

- You create threads in OpenMP\* with the parallel construct.
- For example, to create a 4 thread Parallel region:

Each thread executes a copy of the code within the structured block

```
double A[1000];
omp_set_num_threads(4); ←
#pragma omp parallel
{
    int ID = omp_get_thread_num();
    pooh(ID,A);
}
```

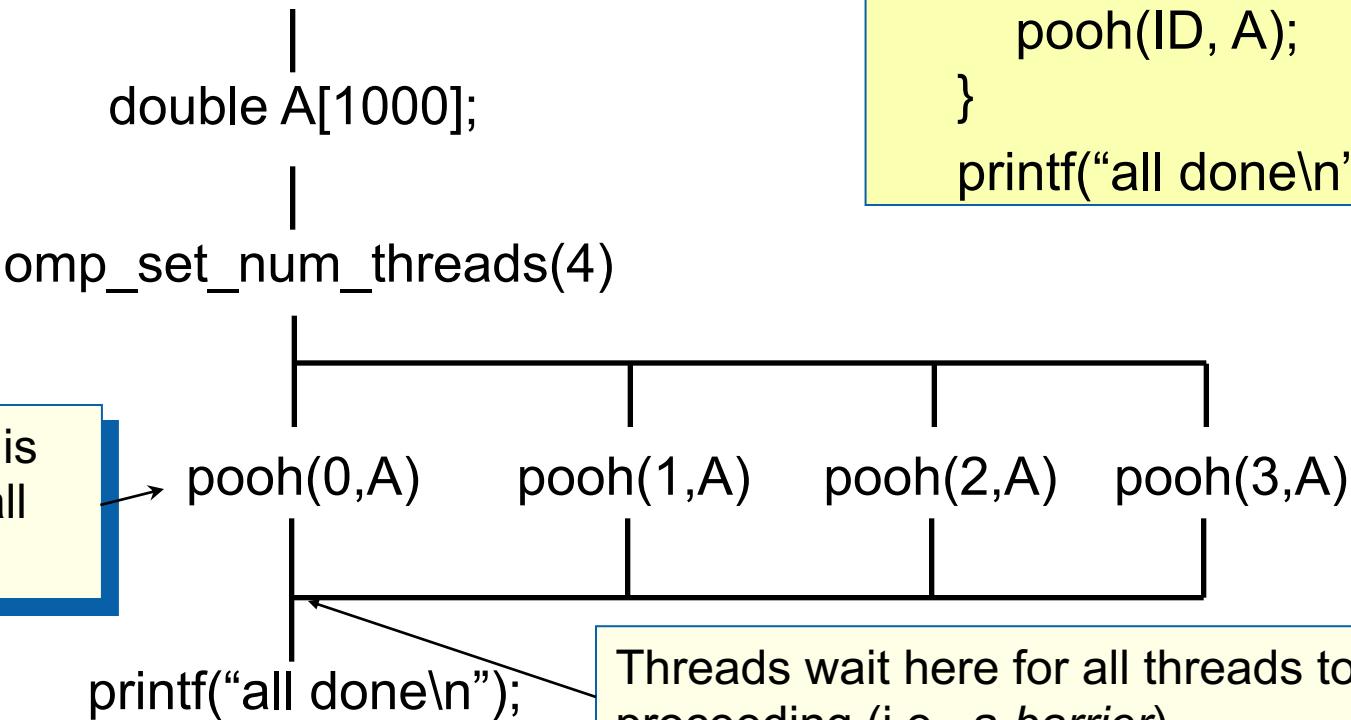
Runtime function to request a certain number of threads

Runtime function returning a thread ID

- Each thread calls pooh(ID,A) for ID = 0 to 3

# Thread Creation: Parallel Regions Example

- Each thread executes the same code redundantly.



# Thread creation: How many threads did you actually get?

- Request a number of threads with `omp_set_num_threads()`
- The number requested may not be the number you actually get.
  - An implementation may silently give you fewer threads than you requested.
  - Once a team of threads has launched, it will not be reduced.

Each thread executes a copy of the code within the structured block

```
double A[1000];
omp_set_num_threads(4); ←
#pragma omp parallel
{
    int ID      = omp_get_thread_num();
    int nthrds = omp_get_num_threads();
    pooh(ID,A);
}
```

Runtime function to request a certain number of threads

Runtime function to return actual number of threads in the team

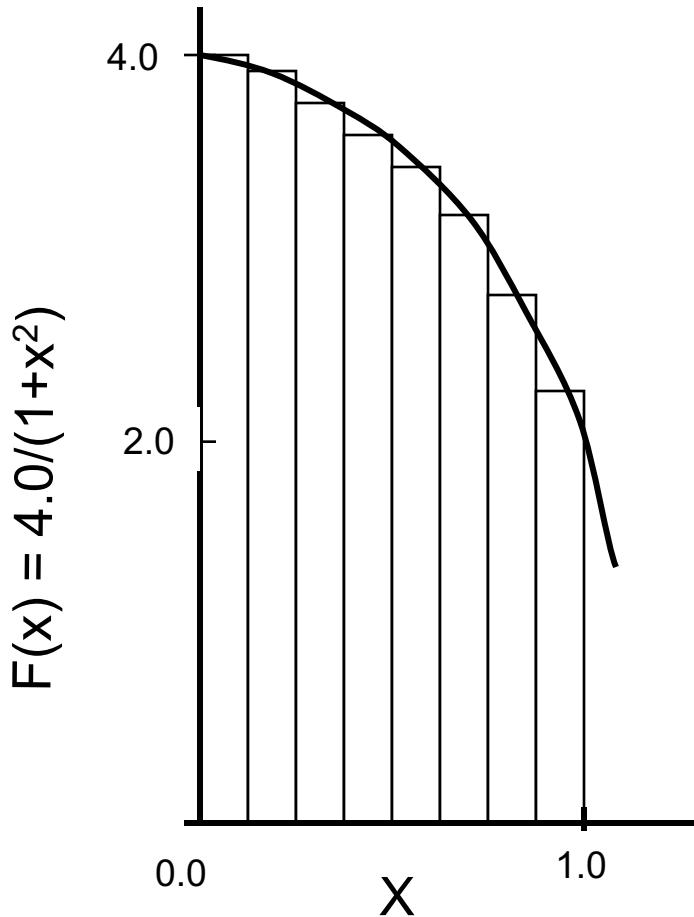
- Each thread calls `pooh(ID,A)` for  $ID = 0$  to  $nthrds-1$

# An Interesting Problem to Play With

## Numerical Integration

Mathematically, we know that:

$$\int_0^1 \frac{4.0}{(1+x^2)} dx = \pi$$



We can approximate the integral as a sum of rectangles:

$$\sum_{i=0}^N F(x_i) \Delta x = \Delta x \sum_{i=0}^N F(x_i) \approx \pi$$

Where each rectangle has width  $\Delta x$  and height  $F(x_i)$  at the middle of interval  $i$ .

# Serial PI Program

```
static long num_steps = 100000;
double step;
int main ()
{
    int i;    double x, pi, sum = 0.0;

    step = 1.0/(double) num_steps;

    for (i=0;i< num_steps; i++){
        x = (i+0.5)*step;
        sum = sum + 4.0/(1.0+x*x);
    }
    pi = step * sum;
}
```

See ParProgForPhys/OMP\_Exercises/pi.c

# Serial PI Program

```
#include <omp.h>
static long num_steps = 100000;
double step;
int main ()
{
    int i;    double x, pi, sum = 0.0;

    step = 1.0/(double) num_steps;
    double tdata = omp_get_wtime();
    for (i=0;i< num_steps; i++){
        x = (i+0.5)*step;
        sum = sum + 4.0/(1.0+x*x);
    }
    pi = step * sum;
    tdata = omp_get_wtime() - tdata;
    printf(" pi = %f in %f secs\n",pi, tdata);
}
```

The library routine `get_omp_wtime()` is used to find the elapsed “wall time” for blocks of code

See ParProgForPhys/OMP\_Exercises /pi.c

# Exercise: the Parallel Pi Program

- Create a parallel version of the pi program using a parallel construct:  
`#pragma omp parallel`
- Pay close attention to shared versus private variables.
- In addition to a parallel construct, you will need the runtime library routines

- `int omp_get_num_threads();` ← Number of threads in the team
- `int omp_get_thread_num();` → Thread ID or rank
- `double omp_get_wtime();` ← Time in seconds since a fixed point in the past
- `omp_set_num_threads();`

Request a number of threads in the team

# Hints: the Parallel Pi Program

- Use a parallel construct:

```
#pragma omp parallel
```

- The challenge is to:
  - divide loop iterations between threads (use the thread ID and the number of threads).
  - Create an accumulator for each thread to hold partial sums that you can later combine to generate the global sum.
- In addition to a parallel construct, you will need the runtime library routines
  - int omp\_set\_num\_threads();
  - int omp\_get\_num\_threads();
  - int omp\_get\_thread\_num();
  - double omp\_get\_wtime();

# Example: A simple SPMD pi program

```
#include <omp.h>
static long num_steps = 100000;      double step;
#define NUM_THREADS 2
void main ()
{
    int i, nthreads; double pi, sum[NUM_THREADS];
    step = 1.0/(double) num_steps;
    omp_set_num_threads(NUM_THREADS);
    #pragma omp parallel
    {
        int i, id, numthrds;
        double x;
        id = omp_get_thread_num();
        numthrds = omp_get_num_threads();
        if (id == 0) nthreads = numthrds;
        for (i=id, sum[id]=0.0; i< num_steps; i=i+numthrds) {
            x = (i+0.5)*step;
            sum[id] += 4.0/(1.0+x*x);
        }
    }
    for(i=0, pi=0.0; i<nthreads; i++) pi += sum[i] * step;
}
```

Promote scalar to an array dimensioned by number of threads to avoid race condition.

Only one thread should copy the number of threads to the global value to make sure multiple threads writing to the same address don't conflict.

This is a common trick in SPMD programs to create a cyclic distribution of loop iterations

# Results\*

- Original Serial pi program with 100000000 steps ran in 1.83 seconds.

```
#include <omp.h>
static long num_steps = 100000;      double step;
#define NUM_THREADS 2
void main ()
{
    int i, nthreads; double pi, sum[NUM_THREADS];
    step = 1.0/(double) num_steps;
    omp_set_num_threads(NUM_THREADS);
    #pragma omp parallel
    {
        int i, id,nthrds;
        double x;
        id = omp_get_thread_num();
        nthrds = omp_get_num_threads();
        if (id == 0)  nthreads = nthrds;
        for (i=id, sum[id]=0.0;i< num_steps; i=i+nthrds) {
            x = (i+0.5)*step;
            sum[id] += 4.0/(1.0+x*x);
        }
    }
    for(i=0, pi=0.0;i<nthreads;i++)pi += sum[i] * step;
}
```

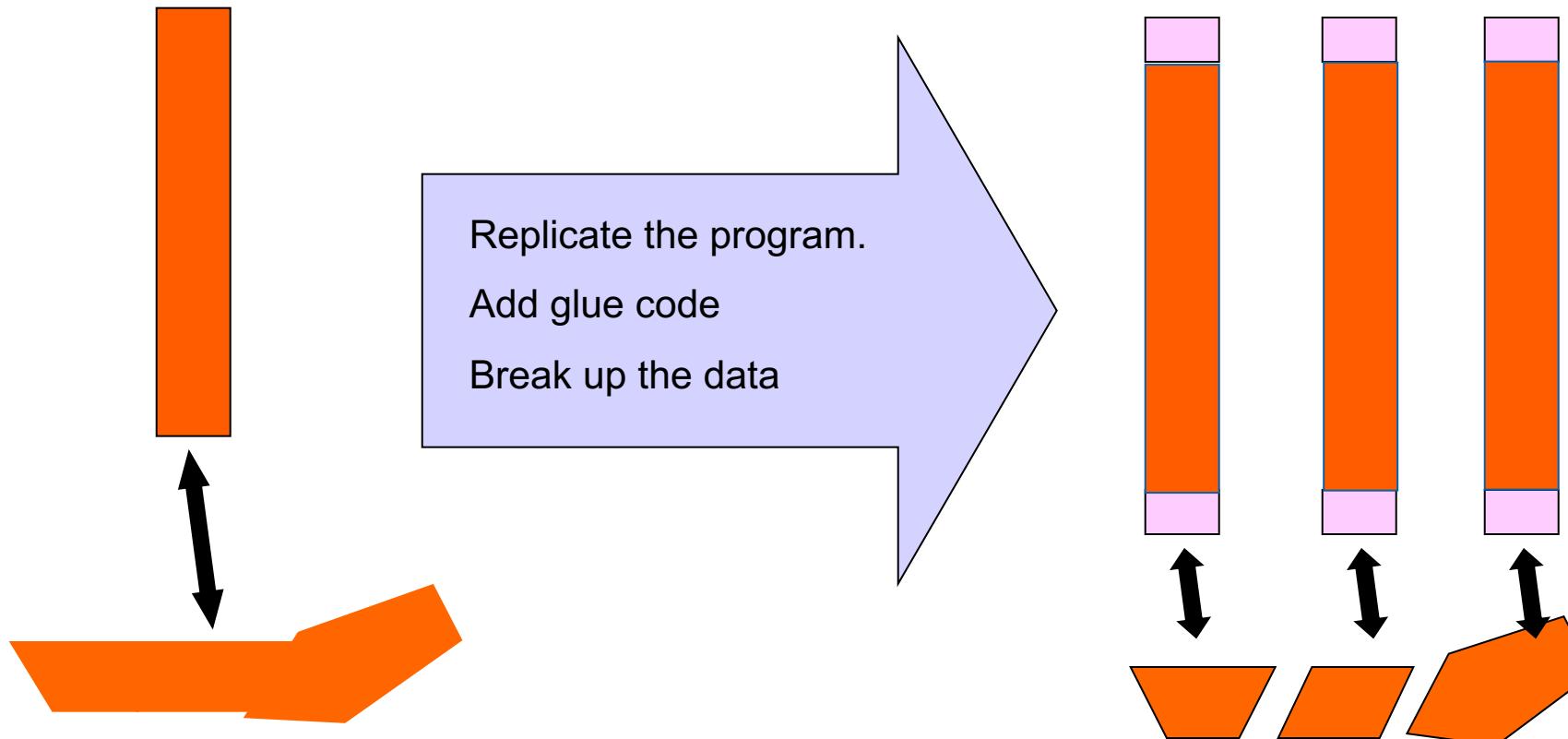
| threads | 1 <sup>st</sup><br>SPMD* |
|---------|--------------------------|
| 1       | 1.86                     |
| 2       | 1.03                     |
| 3       | 1.08                     |
| 4       | 0.97                     |

Intel compiler (icpc) with default optimization level (O2) on Apple OS X 10.7.3 with a dual core (four HW thread)  
Intel® Core™ i5 processor at 1.7 Ghz and 4 Gbyte DDR3 memory at 1.333 Ghz.

\*SPMD: Single Program Multiple Data

# SPMD: Single Program Multiple Data

- Run the same program on  $P$  processing elements where  $P$  can be arbitrarily large.



- Use the rank ... an ID ranging from 0 to  $(P-1)$  ... to select between a set of tasks and to manage any shared data structures.

MPI programs almost always use this pattern ... it is probably the most commonly used pattern in the history of parallel programming.

**A brief digression to talk about  
performance issues in parallel  
programs**

# Consider performance of parallel programs

Compute N independent tasks on one processor

Load Data

Compute  $T_1$

...

Compute  $T_N$

Consume Results

$$Time_{seq}(1) = T_{load} + N*T_{task} + T_{consume}$$

Compute N independent tasks with P processors

Load Data

Compute  $T_1$

...

Consume Results

Compute  $T_N$

Ideally Cut  
runtime by  $\sim 1/P$

(Note: Parallelism  
only speeds-up the  
concurrent part)

$$Time_{par}(P) = T_{load} + (N/P)*T_{task} + T_{consume}$$

# Talking about performance

- Speedup: the increased performance from running on  $P$  processors.
- Perfect Linear Speedup: happens when no parallel overhead and algorithm is 100% parallel.
- Super-linear Speedup: typically due to cache effects ... i.e. as  $P$  grows, aggregate cache size grows so more of the problem fits in cache

$$S(P) = \frac{Time_{seq}(1)}{Time_{par}(P)}$$

$$S(P) = P$$

$$S(P) > P$$

# Amdahl's Law

- What is the maximum speedup you can expect from a parallel program?
- Approximate the runtime as a part that can be sped up with additional processors and a part that is fundamentally serial.

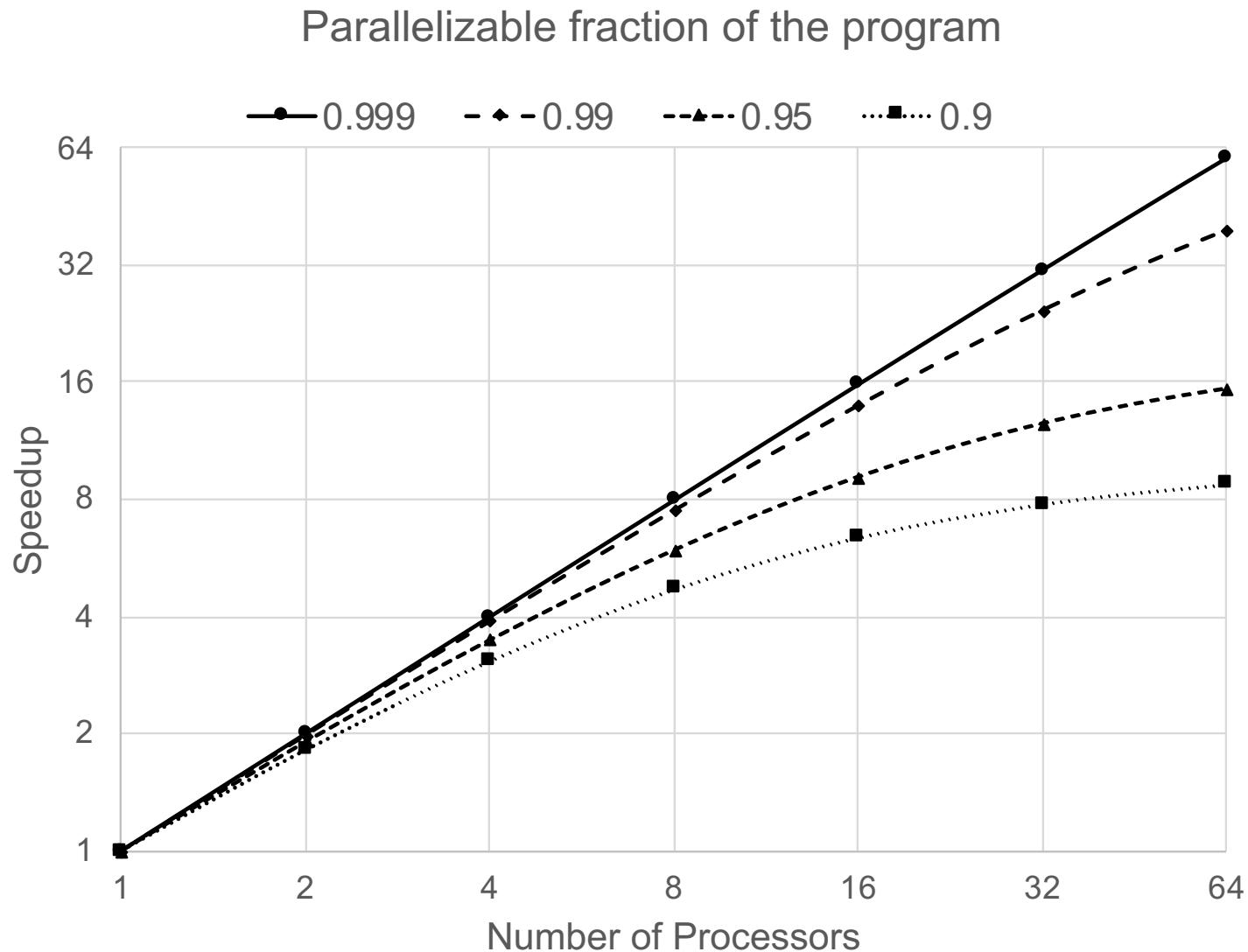
$$Time_{par}(P) = (serial\_fraction + \frac{parallel\_fraction}{P}) * Time_{seq}$$

- If the serial fraction is  $\alpha$  and the parallel fraction is  $(1 - \alpha)$  then the speedup is:

$$S(P) = \frac{Time_{seq}}{Time_{par}(P)} = \frac{Time_{seq}}{(\alpha + \frac{1-\alpha}{P}) * Time_{seq}} = \frac{1}{\alpha + \frac{1-\alpha}{P}}$$

- If you had an unlimited number of processors:  $P \rightarrow \infty$
- The maximum possible speedup is:  $S = \frac{1}{\alpha} \leftarrow$  Amdahl's Law

# Amdahl's Law



# So now you should understand my silly introduction slide.

## Introduction

I'm just a simple kayak instructor

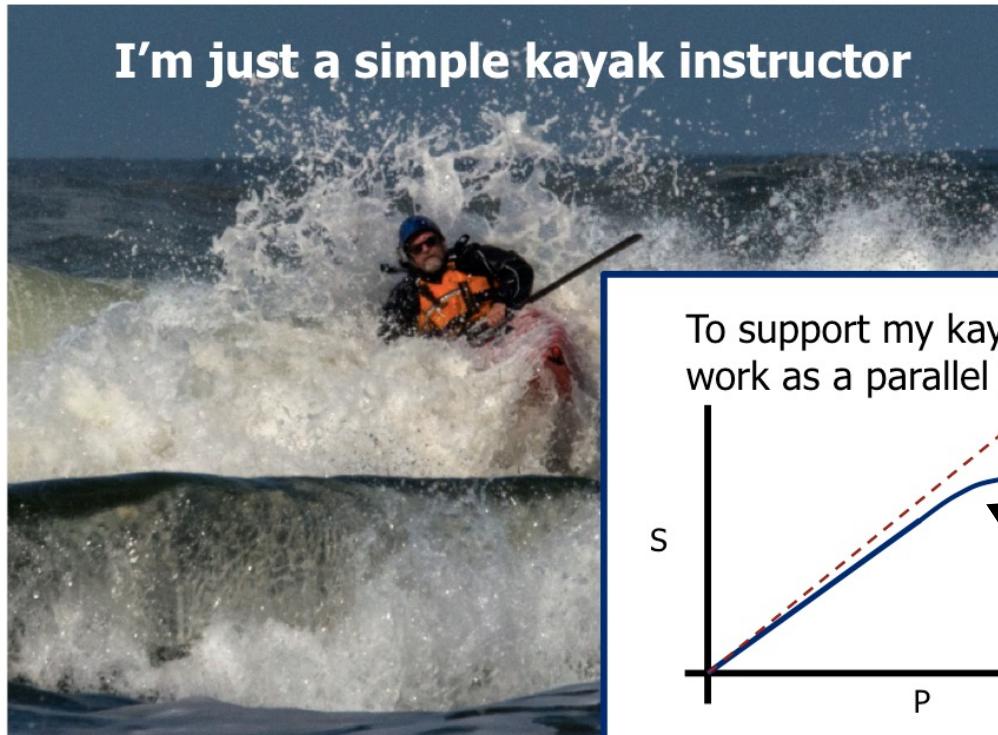
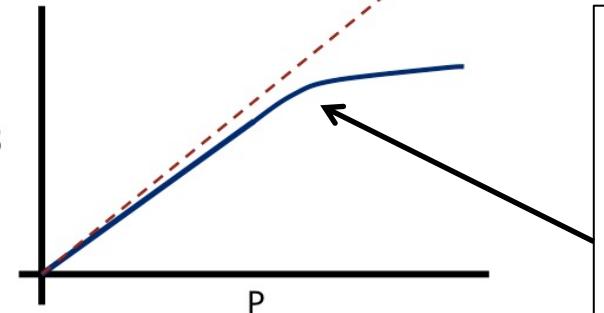


Photo © by Greg Clopton, 2014

We measure our success as parallel programmers by how close we come to ideal linear speedup.

To support my kayaking habit I work as a parallel programmer



Which means I know how to turn math into lines on a speedup plot

A good parallel programmer always figures out when you fall off the linear speedup curve and why that has occurred.

**Now that you understand how to  
think about parallel performance,  
lets get back to OpenMP**

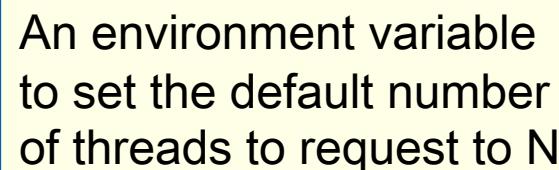
# Internal control variables and how to control the number of threads in a team

- We've used the following construct to control the number of threads. (e.g. to request 12 threads):
  - `omp_set_num_threads(12)`
- What does `omp_set_num_threads()` actually do?
  - It resets an “internal control variable” the system queries to select the default number of threads to request on subsequent parallel constructs.
- Is there an easier way to change this internal control variable ... perhaps one that doesn't require re-compilation? Yes.
  - When an OpenMP program starts up, it queries an environment variable `OMP_NUM_THREADS` and sets the appropriate internal control variable to the value of **OMP\_NUM\_THREADS**
  - For example, to set the initial, default number of threads to request in OpenMP from my apple laptop
    - > **export OMP\_NUM\_THREADS=12**

# Exercise

- Go back to your parallel pi program and explore how well it scales with the number of threads.
- Can you explain your performance with Amdahl's law? If not what else might be going on?

- `int omp_get_num_threads();`
- `int omp_get_thread_num();`
- `double omp_get_wtime();`
- `omp_set_num_threads();`
- `export OMP_NUM_THREADS = N`



An environment variable  
to set the default number  
of threads to request to N

# Results\*

- Original Serial pi program with 100000000 steps ran in 1.83 seconds.

```
#include <omp.h>
static long num_steps = 100000;      double step;
#define NUM_THREADS 2
void main ()
{
    int i, nthreads; double pi, sum[NUM_THREADS];
    step = 1.0/(double) num_steps;
    omp_set_num_threads(NUM_THREADS);
    #pragma omp parallel
    {
        int i, id,nthrds;
        double x;
        id = omp_get_thread_num();
        nthrds = omp_get_num_threads();
        if (id == 0)  nthreads = nthrds;
        for (i=id, sum[id]=0.0;i< num_steps; i=i+nthrds) {
            x = (i+0.5)*step;
            sum[id] += 4.0/(1.0+x*x);
        }
    }
    for(i=0, pi=0.0;i<nthreads;i++)pi += sum[i] * step;
}
```

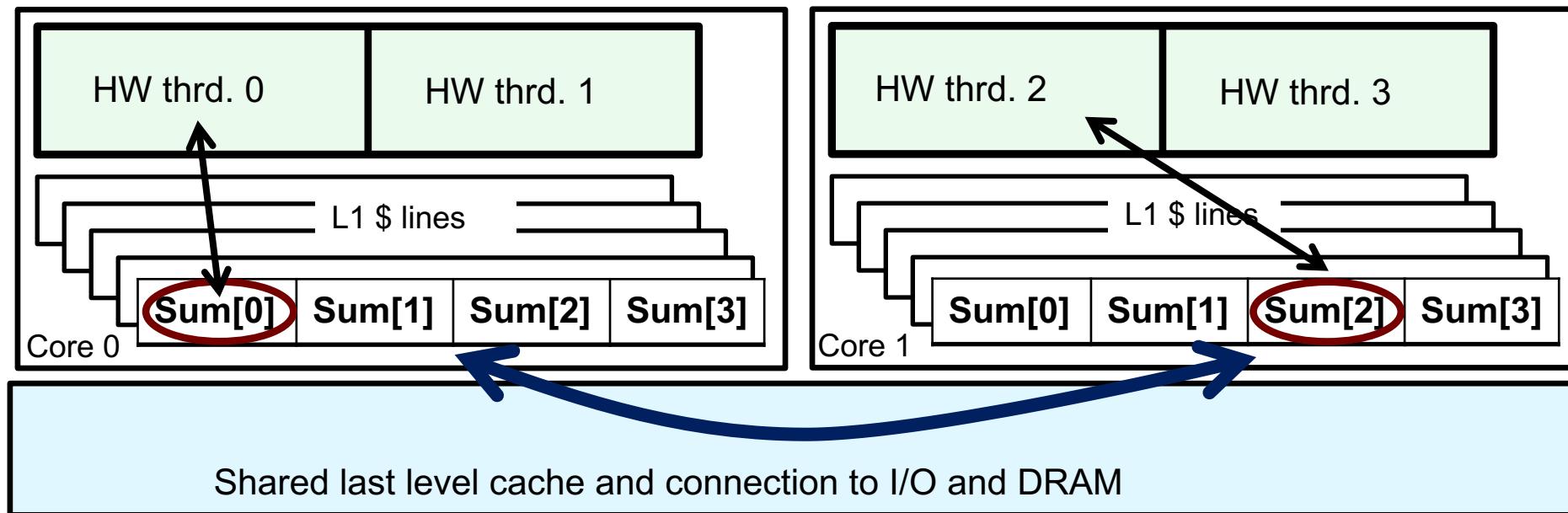
| threads | 1 <sup>st</sup><br>SPMD* |
|---------|--------------------------|
| 1       | 1.86                     |
| 2       | 1.03                     |
| 3       | 1.08                     |
| 4       | 0.97                     |

Intel compiler (icpc) with default optimization level (O2) on Apple OS X 10.7.3 with a dual core (four HW thread)  
Intel® Core™ i5 processor at 1.7 Ghz and 4 Gbyte DDR3 memory at 1.333 Ghz.

\*SPMD: Single Program Multiple Data

# Why Such Poor Scaling? False Sharing

- If independent data elements happen to sit on the same cache line, each update will cause the cache lines to “slosh back and forth” between threads ... This is called “**false sharing**”.



- If you promote scalars to an array to support creation of an SPMD program, the array elements are contiguous in memory and hence share cache lines ... Results in poor scalability.
- Solution: Pad arrays so elements you use are on distinct cache lines.

## Example: Eliminate false sharing by padding the sum array

```
#include <omp.h>
static long num_steps = 100000;      double step;
#define NUM_THREADS 2
#define PAD 8      // assume 64 byte L1 cache line size
void main ()
{   int i, nthreads;  double pi, sum[NUM_THREADS][PAD];
    step = 1.0/(double) num_steps;
    omp_set_num_threads(NUM_THREADS);
    #pragma omp parallel
    {
        int i, id,nthrds;
        double x;
        id = omp_get_thread_num();
        nthrds = omp_get_num_threads();
        if (id == 0)  nthreads = nthrds;
        for (i=id, sum[id]=0.0;i< num_steps; i=i+nthrds) {
            x = (i+0.5)*step;
            sum[id][0] += 4.0/(1.0+x*x);
        }
    }
    for(i=0, pi=0.0;i<nthreads;i++)pi += sum[i][0] * step;
}
```

Pad the array so each  
sum value is in a  
different cache line

# Results\*: PI Program, Padded Accumulator

- Original Serial pi program with 100000000 steps ran in 1.83 seconds.

```
#include <omp.h>
static long num_steps = 100000;      double step;
#define NUM_THREADS 2
#define PAD 8    // assume 64 byte L1 cache line size
void main ()
{
    int i, nthreads; double pi, sum[NUM_THREADS][PAD];
    step = 1.0/(double) num_steps;
    omp_set_num_threads(NUM_THREADS);
    #pragma omp parallel
    {
        int i, id,nthrds;
        double x;
        id = omp_get_thread_num();
        nthrds = omp_get_num_threads();
        if (id == 0)  nthreads = nthrds;
        for (i=id, sum[id]=0.0;i< num_steps; i=i+nthrds) {
            x = (i+0.5)*step;
            sum[id][0] += 4.0/(1.0+x*x);
        }
    }
    for(i=0, pi=0.0;i<nthreads;i++)pi += sum[i][0] * step;
}
```

| threads | 1 <sup>st</sup><br>SPMD | 1 <sup>st</sup><br>SPMD<br>padded |
|---------|-------------------------|-----------------------------------|
| 1       | 1.86                    | 1.86                              |
| 2       | 1.03                    | 1.01                              |
| 3       | 1.08                    | 0.69                              |
| 4       | 0.97                    | 0.53                              |

\*Intel compiler (icpc) with default optimization level (O2) on Apple OS X 10.7.3 with a dual core (four HW thread) Intel® Core™ i5 processor at 1.7 Ghz and 4 Gbyte DDR3 memory at 1.333 Ghz.

# Outline

OpenMP®

- Introduction to OpenMP
- Creating Threads
- • Synchronization
- Parallel Loops
- Data Environment
- Memory Model
- Irregular Parallelism and Tasks
- Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data

# Synchronization

Synchronization is used to impose order constraints and to protect access to shared data

- High level synchronization included in the common core:
  - critical
  - barrier
- Other, more advanced, synchronization operations:
  - atomic
  - ordered
  - flush
  - locks (both simple and nested)

# Synchronization: critical

- Mutual exclusion: Only one thread at a time can enter a **critical** region.

Threads wait their turn  
– only one thread at a  
time calls consume()

```
float res;  
#pragma omp parallel  
{    float B;    int i, id, nthrds;  
    id = omp_get_thread_num();  
    nthrds = omp_get_num_threads();  
    B = big_SPMD_job(id, nthrds);  
#pragma omp critical  
    res += consume (B);  
}
```

# Synchronization: barrier

- Barrier: a point in a program all threads must reach before any threads are allowed to proceed.
- It is a “stand alone” pragma meaning it is not associated with user code ... it is an executable statement.

```
double Arr[8], Brr[8];          int numthrds;  
omp_set_num_threads(8)  
#pragma omp parallel  
{  int id, nthrds;  
    id = omp_get_thread_num();  
    nthrds = omp_get_num_threads();  
    if (id==0) numthrds = nthrds;  
    Arr[id] = big_ugly_calc(id, nthrds);  
#pragma omp barrier  
    Brr[id] = really_big_and_ugly(id, nthrds, Arr);  
}
```

Threads wait until all  
threads hit the barrier.  
Then they can go on.



# Exercise

- In your first Pi program, you probably used an array to create space for each thread to store its partial sum.
- If array elements happen to share a cache line, this leads to false sharing.
  - Non-shared data in the same cache line so each update invalidates the cache line ... in essence “sloshing independent data” back and forth between threads.
- Modify your “pi program” to avoid false sharing due to the partial sum array.

```
int omp_get_num_threads();
int omp_get_thread_num();
double omp_get_wtime();
omp_set_num_threads();
#pragma parallel
#pragma critical
```

# PI Program with False Sharing

- Original Serial pi program with 100000000 steps ran in 1.83 seconds.

```
#include <omp.h>
static long num_steps = 100000;      double step;
#define NUM_THREADS 2
void main ()
{
    int i, nthreads; double pi, sum[NUM_THREADS];
    step = 1.0/(double) num_steps;
    omp_set_num_threads(NUM_THREADS);
#pragma omp parallel
{
    int i, id,nthrds;
    double x;
    id = omp_get_thread_num();
    nthrds = omp_get_num_threads();
    if (id == 0)  nthreads = nthrds;
    for (i=id, sum[id]=0.0;i< num_steps; i=i+nthrds) {
        x = (i+0.5)*step;
        sum[id] += 4.0/(1.0+x*x);
    }
}
for(i=0, pi=0.0;i<nthreads;i++)pi += sum[i] * step;
}
```

Recall that promoting sum to an array made the coding easy, but led to false sharing and poor performance.

| threads | 1 <sup>st</sup><br>SPMD |
|---------|-------------------------|
| 1       | 1.86                    |
| 2       | 1.03                    |
| 3       | 1.08                    |
| 4       | 0.97                    |

\*Intel compiler (icpc) with no optimization on Apple OS X 10.7.3 with a dual core (four HW thread)  
Intel® Core™ i5 processor at 1.7 Ghz and 4 Gbyte DDR3 memory at 1.333 Ghz.

## Example: Using a critical section to remove impact of false sharing

```
#include <omp.h>
static long num_steps = 100000;      double step;
#define NUM_THREADS 2
void main ()
{ int nthreads; double pi=0.0;      step = 1.0/(double) num_steps;
  omp_set_num_threads(NUM_THREADS);
  #pragma omp parallel
  {
    int i, id, nthrds;  double x, sum; ← Create a scalar local to each
    id = omp_get_thread_num();                                thread to accumulate partial sums.
    nthrds = omp_get_num_threads();
    if (id == 0)  nthreads = nthrds;
    for (i=id, sum=0.0;i< num_steps; i=i+nthrds) {
      x = (i+0.5)*step;
      sum += 4.0/(1.0+x*x); ← No array, so no false sharing.
    }
    #pragma omp critical
    pi += sum * step; ← Sum goes “out of scope” beyond the parallel region ...
  }                                         so you must sum it in here. Must protect summation
}                                         into pi in a critical region so updates don’t conflict
```

# Results\*: pi program critical section

- Original Serial pi program with 100000000 steps ran in 1.83 seconds.

```
#include <omp.h>
static long num_steps = 100000;      double step;
#define NUM_THREADS 2
void main ()
{ int nthreads; double pi=0.0;      step = 1.0/(double) num_steps;
  omp_set_num_threads(NUM_THREADS);
  #pragma omp parallel
  {
    int i, id, nthrds;  double x, sum;
    id = omp_get_thread_num();
    nthrds = omp_get_num_threads();
    if (id == 0)  nthreads = nthrds;
    for (i=id, sum=0.0;i< num_steps; i=i+nthrds) {
      x = (i+0.5)*step;
      sum += 4.0/(1.0+x*x);
    }
    #pragma omp critical
      pi += sum * step;
  }
}
```

| threads | 1st<br>SPMD | 1st<br>SPMD<br>padded | SPMD<br>critical |
|---------|-------------|-----------------------|------------------|
| 1       | 1.86        | 1.86                  | 1.87             |
| 2       | 1.03        | 1.01                  | 1.00             |
| 3       | 1.08        | 0.69                  | 0.68             |
| 4       | 0.97        | 0.53                  | 0.53             |

\*Intel compiler (icpc) with no optimization on Apple OS X 10.7.3 with a dual core (four HW thread) Intel® Core™ i5 processor at 1.7 Ghz and 4 Gbyte DDR3 memory at 1.333 Ghz.

## Example: Using a critical section to remove impact of false sharing

```
#include <omp.h>
static long num_steps = 100000;      double step;
#define NUM_THREADS 2
void main ()
{ int nthreads; double pi=0.0;      step = 1.0/(double) num_steps;
  omp_set_num_threads(NUM_THREADS);
  #pragma omp parallel
  {
    int i, id, nthrds;  double x, sum;
    id = omp_get_thread_num();
    nthrds = omp_get_num_threads();
    if (id == 0)  nthreads = nthrds;
    for (i=id, sum=0.0;i< num_steps; i=i+nthrds) {
      x = (i+0.5)*step;
      #pragma omp critical
      sum += 4.0/(1.0+x*x);
    }
  }
}
```

What would happen if you put the critical section inside the loop?

- Introduction to OpenMP
- Creating Threads
- Synchronization
- • Parallel Loops
- Data Environment
- Memory Model
- Irregular Parallelism and Tasks
- Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data

# The Loop Worksharing Construct

- The loop worksharing construct splits up loop iterations among the threads in a team

```
#pragma omp parallel
```

```
{
```

```
#pragma omp for
```

```
for (I=0;I<N;I++){
```

```
    NEAT_STUFF(I);
```

```
}
```

The loop control index I is made  
“private” to each thread by default.

Threads wait here until all  
threads are finished with the  
parallel loop before any proceed  
past the end of the loop

Loop construct name:

- C/C++: for
- Fortran: do

# Loop Worksharing Construct

## A motivating example

Sequential code

```
for(i=0;i<N;i++) { a[i] = a[i] + b[i];}
```

OpenMP parallel region  
(SPMD Pattern)

```
#pragma omp parallel
{
    int id, i, Nthrds, istart, iend;
    id = omp_get_thread_num();
    Nthrds = omp_get_num_threads();
    istart = id * N / Nthrds;
    iend = (id+1) * (N / Nthrds)-1;
    if (id == Nthrds-1)iend = N;
    for(i=istart;i<iend;i++) { a[i] = a[i] + b[i];}
}
```

OpenMP parallel region and  
a worksharing for construct

```
#pragma omp parallel
#pragma omp for
for(i=0;i<N;i++) { a[i] = a[i] + b[i];}
```

# Loop Worksharing Constructs: The schedule clause

- The schedule clause affects how loop iterations are mapped onto threads
  - **schedule(static [,chunk])**
    - Deal-out blocks of iterations of size “chunk” to each thread.
  - **schedule(dynamic[,chunk])**
    - Each thread grabs “chunk” iterations off a queue until all iterations have been handled.
- Example:
  - `#pragma omp for schedule(dynamic, 10)`

| Schedule Clause | When To Use  |  |
|-----------------|--|--|
| <b>STATIC</b>   | <b>Pre-determined and predictable by the programmer</b>  | Least work at runtime : scheduling done at compile-time          |
| <b>DYNAMIC</b>  | <b>Unpredictable, highly variable work per iteration</b> | Most work at runtime : complex scheduling logic used at run-time |

# Combined Parallel/Worksharing Construct

- OpenMP shortcut: Put the “parallel” and the worksharing directive on the same line

```
double res[MAX]; int i;  
#pragma omp parallel  
{  
    #pragma omp for  
    for (i=0;i< MAX; i++) {  
        res[i] = huge();  
    }  
}
```

```
double res[MAX]; int i;  
#pragma omp parallel for  
for (i=0;i< MAX; i++) {  
    res[i] = huge();  
}
```

These are equivalent

# Working with loops

- Basic approach
  - Find compute intensive loops
  - Make the loop iterations independent ... So they can safely execute in any order without loop-carried dependencies
  - Place the appropriate OpenMP directive and test

```
int i, j, A[MAX];
j = 5;
for (i=0;i< MAX; i++) {
    j +=2;
    A[i] = big(j);
}
```

Note: loop index  
“i” is private by  
default

Remove loop  
carried  
dependence

```
int i, A[MAX];
#pragma omp parallel for
for (i=0;i< MAX; i++) {
    int j = 5 + 2*(i+1);
    A[i] = big(j);
}
```

# Reduction

- How do we handle this case?

```
double ave=0.0, A[MAX];
int i;
for (i=0;i< MAX; i++) {
    ave += A[i];
}
ave = ave/MAX;
```

- We are combining values into a single accumulation variable (ave) ... there is a true dependence between loop iterations that can't be trivially removed.
- This is a very common situation ... it is called a “reduction”.
- Support for reduction operations is included in most parallel programming environments.

# Reduction

- OpenMP reduction clause:

reduction (op : list)

- Inside a parallel or a work-sharing construct:

- A local copy of each list variable is made and initialized depending on the “op” (e.g. 0 for “+”).
  - Updates occur on the local copy.
  - Local copies are reduced into a single value and combined with the original global value.

- The variables in “list” must be shared in the enclosing parallel region.

```
double ave=0.0, A[MAX];  int i;  
#pragma omp parallel for reduction (+:ave)  
for (i=0;i< MAX; i++) {  
    ave += A[i];  
}  
ave = ave/MAX;
```

# OpenMP: Reduction operands/initial-values

- Many different associative operands can be used with reduction:
- Initial values are the ones that make sense mathematically.

| Operator | Initial value       |
|----------|---------------------|
| +        | 0                   |
| *        | 1                   |
| -        | 0                   |
| min      | Largest pos. number |
| max      | Most neg. number    |

| C/C++ only |               |
|------------|---------------|
| Operator   | Initial value |
| &          | $\sim 0$      |
|            | 0             |
| ^          | 0             |
| &&         | 1             |
|            | 0             |

| Fortran Only |               |
|--------------|---------------|
| Operator     | Initial value |
| .AND.        | .true.        |
| .OR.         | .false.       |
| .NEQV.       | .false.       |
| .IEOR.       | 0             |
| .IOR.        | 0             |
| .IAND.       | All bits on   |
| .EQV.        | .true.        |

OpenMP includes user defined reductions and array-sections as reduction variables (we just don't cover those topics here)

# Exercise: PI with loops

- Go back to the serial pi program and parallelize it with a loop construct
- Your goal is to minimize the number of changes made to the serial program.

```
#pragma omp parallel
#pragma omp for
#pragma omp parallel for
#pragma omp for reduction(op:list)
#pragma omp critical
int omp_get_num_threads();
int omp_get_thread_num();
double omp_get_wtime();
```

# Example: PI with a loop and a reduction

```
#include <omp.h>
static long num_steps = 100000;      double step;
void main ()
{   int i;           double x, pi, sum = 0.0;
    step = 1.0/(double) num_steps;
    #pragma omp parallel
    {
        double x;           ← Create a scalar local to each thread to hold
        #pragma omp for reduction(+:sum)
        for (i=0;i< num_steps; i++){
            x = (i+0.5)*step;
            sum = sum + 4.0/(1.0+x*x), ← Break up loop iterations
        }                                and assign them to
    }                                    threads ... setting up a
    pi = step * sum;                   reduction into sum.
}                                     Note ... the loop index is
                                   local to a thread by default.
```

# Example: PI with a loop and a reduction

```
#include <omp.h>
static long num_steps = 100000;      double step;
void main ()
{
    double pi, sum = 0.0;
    step = 1.0/(double) num_steps;

#pragma omp parallel for reduction(+:sum)
for (int i=0;i< num_steps; i++){
    double x = (i+0.5)*step;
    sum = sum + 4.0/(1.0+x*x);
}
pi = step * sum;
}
```

Using modern C style, we put declarations close to where they are used ... which lets me use the parallel for construct.

# Results\*: PI with a loop and a reduction

- Original Serial pi program with 100000000 steps ran in 1.83 seconds.

## Example: Pi with a

```
#include <omp.h>
static long num_steps = 100000000;
void main ()
{
    int i;      double x, pi, sum;
    step = 1.0/(double) num_steps;
    #pragma omp parallel
    {
        double x;
        #pragma omp for reduction(+:sum)
        for (i=0;i< num_steps; i++){
            x = (i+0.5)*step;
            sum = sum + 4.0/(1.0+x*x);
        }
    }
    pi = step * sum;
}
```

| threads | 1 <sup>st</sup><br>SPMD | 1 <sup>st</sup><br>SPMD<br>padded | SPMD<br>critical | PI Loop |
|---------|-------------------------|-----------------------------------|------------------|---------|
| 1       | 1.86                    | 1.86                              | 1.87             | 1.91    |
| 2       | 1.03                    | 1.01                              | 1.00             | 1.02    |
| 3       | 1.08                    | 0.69                              | 0.68             | 0.80    |
| 4       | 0.97                    | 0.53                              | 0.53             | 0.68    |

\*Intel compiler (icpc) with no optimization on Apple OS X 10.7.3 with a dual core (four HW thread) Intel® Core™ i5 processor at 1.7 Ghz and 4 Gbyte DDR3 memory at 1.333 Ghz.

# The nowait clause

- Barriers are really expensive. You need to understand when they are implied and how to skip them when it's safe to do so.

```
double A[big], B[big], C[big];  
  
#pragma omp parallel  
{  
    int id=omp_get_thread_num();  
    A[id] = big_calc1(id);  
#pragma omp barrier  
#pragma omp for  
    for(i=0;i<N;i++){C[i]=big_calc3(i,A);}  
#pragma omp for nowait  
    for(i=0;i<N;i++){ B[i]=big_calc2(C, i); }  
    A[id] = big_calc4(id);  
}
```

implicit barrier at the end of a for worksharing construct

implicit barrier at the end of a parallel region

no implicit barrier due to nowait

- Introduction to OpenMP
- Creating Threads
- Synchronization
- Parallel Loops
- ➡ • Data Environment
  - Memory Model
  - Irregular Parallelism and Tasks
  - Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data

# Data Environment: Default storage attributes

- Shared memory programming model:
  - Most variables are shared by default
- Global variables are SHARED among threads
  - Fortran: COMMON blocks, SAVE variables, MODULE variables
  - C: File scope variables, static
  - Both: dynamically allocated memory (ALLOCATE, malloc, new)
- But not everything is shared...
  - Stack variables in subprograms(Fortran) or functions(C) called from parallel regions are PRIVATE
  - Automatic variables within a statement block are PRIVATE.

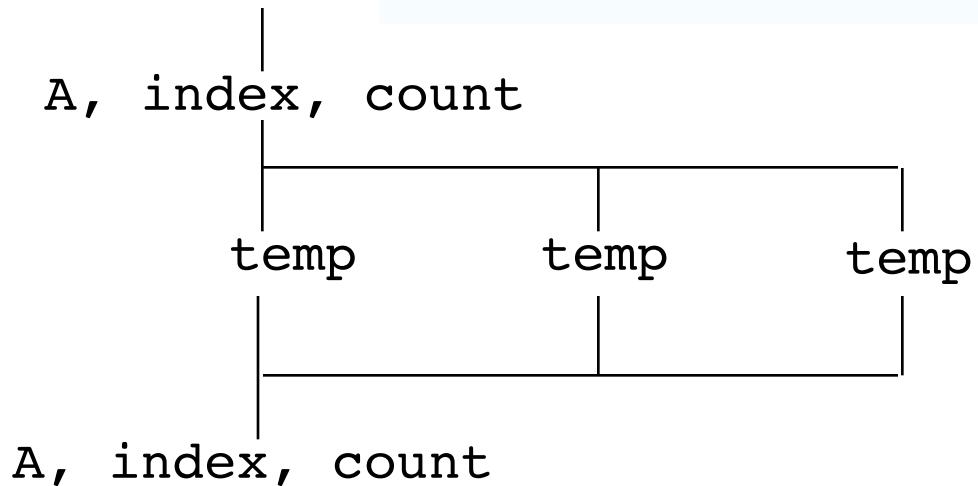
# Data Sharing: Examples

```
double A[10];
int main() {
    int index[10];
    #pragma omp parallel
        work(index);
    printf("%d\n", index[0]);
}
```

A, index and count are shared by all threads.

temp is local to each thread

```
extern double A[10];
void work(int *index) {
    double temp[10];
    static int count;
    ...
}
```



# Data Sharing: Changing storage attributes

- One can selectively change storage attributes for constructs using the following clauses\* (note: *list* is a comma-separated list of variables)
  - shared(*list*)
  - private(*list*)
  - firstprivate(*list*)
- These can be used on parallel and for constructs ... other than shared which can only be used on a parallel construct
- Force the programmer to explicitly define storage attributes
  - default (none)

default() can only be used  
on parallel constructs

# Data Sharing: Private clause

- `private(var)` creates a new local copy of var for each thread.

```
int N = 1000;  
extern void init_arrays(int N, double *A, double *B, double *C);
```

```
void example () {  
    int i, j;  
    double A[N][N], B[N][N], C[N][N];  
    init_arrays(N, *A, *B, *C);  
  
    #pragma omp parallel for private(j)  
    for (i = 0; i < 1000; i++)  
        for( j = 0; j<1000; j++)  
            C[i][j] = A[i][j] + B[i][j];  
}
```

OpenMP makes the loop control index on the parallel loop (i) private by default ... but not for the second loop (j)

# Data Sharing: Private clause

- `private(var)` creates a new local copy of var for each thread.
  - The value of the private copies is uninitialized
  - The value of the original variable is unchanged after the region

```
void wrong() {  
    int tmp = 0;  
#pragma omp parallel for private(tmp)  
    for (int j = 0; j < 1000; ++j)  
        tmp += j;  
    printf("%d\n", tmp);  
}
```

When you need to refer to the variable `tmp` that exists prior to the construct, we call it the **original variable**.

`tmp` was not initialized

`tmp` is 0 here

# Data Sharing: Private and the original variable

- The original variable's value is unspecified if it is referenced outside of the construct
  - Implementations may reference the original variable or a copy ..... a dangerous programming practice!
  - For example, consider what would happen if the compiler inlined work()?

```
int tmp;  
void danger() {  
    tmp = 0;  
#pragma omp parallel private(tmp)  
    work();  
    printf("%d\n", tmp);  
}
```

tmp has unspecified value

```
extern int tmp;  
void work() {  
    tmp = 5;  
}
```

unspecified which  
copy of tmp

# Firstprivate clause

- Variables initialized from a shared variable
- C++ objects are copy-constructed

```
incr = 0;  
#pragma omp parallel for firstprivate(incr)  
for (i = 0; i <= MAX; i++) {  
    if ((i%2)==0) incr++;  
    A[i] = incr;  
}
```

Each thread gets its own copy of  
incr with an initial value of 0

# Data sharing: A data environment test

- Consider this example of PRIVATE and FIRSTPRIVATE

```
variables: A = 1, B = 1, C = 1  
#pragma omp parallel private(B) firstprivate(C)
```

- Are A,B,C private to each thread or shared inside the parallel region?
- What are their initial values inside and values after the parallel region?

Inside this parallel region ...

- “A” is shared by all threads; equals 1
- “B” and “C” are private to each thread.
  - B’s initial value is undefined
  - C’s initial value equals 1

Following the parallel region ...

- B and C revert to their original values of 1
- A is either 1 or the value it was set to inside the parallel region

# Data Sharing: Default clause

- **default(none)**: Forces you to define the storage attributes for variables that appear inside the static extent of the construct ... if you fail the compiler will complain. Good programming practice!
- You can put the default clause on parallel and parallel + workshare constructs.

The static extent is the code in the compilation unit that contains the construct.

```
#include <omp.h>
int main()
{
    int i, j=5;    double x=1.0, y=42.0;
    #pragma omp parallel for default(none) reduction(*:x)
    for (i=0;i<N;i++){
        for(j=0; j<3; j++)
            x+= foobar(i, j, y);
    }
    printf(" x is %f\n", (float)x);
}
```

The compiler would complain about j and y, which is important since you don't want j to be shared

The full OpenMP specification has other versions of the default clause, but they are not used very often so we skip them in the common core

# Exercise: Mandelbrot set area

- The supplied program (mandel.c) computes the area of a Mandelbrot set.
- The program has been parallelized with OpenMP, but we were lazy and didn't do it right.
- Find and fix the errors (hint ... the problem is with the data environment).
- Once you have a working version, try to optimize the program.
  - Try different schedules on the parallel loop.
  - Try different mechanisms to support mutual exclusion ... do the efficiencies change?

# The Mandelbrot Set Area Program

```
#include <omp.h>
#define NPOINTS 1000
#define MXITR 1000
struct d_complex{
    double r;    double i;
};
void testpoint(struct d_complex);
struct d_complex c;
int numoutside = 0;

int main(){
    int i, j;
    double area, error, eps = 1.0e-5;
#pragma omp parallel for private(c, j) firstprivate(eps)
    for (i=0; i<NPOINTS; i++) {
        for (j=0; j<NPOINTS; j++) {
            c.r = -2.0+2.5*(double)(i)/(double)(NPOINTS)+eps;
            c.i = 1.125*(double)(j)/(double)(NPOINTS)+eps;
            testpoint(c);
        }
    }
    area=2.0*2.5*1.125*(double)(NPOINTS*NPOINTS-
    numoutside)/(double)(NPOINTS*NPOINTS);
    error=area/(double)NPOINTS;
}
```

```
void testpoint(struct d_complex c){
    struct d_complex z;
    int iter;
    double temp;

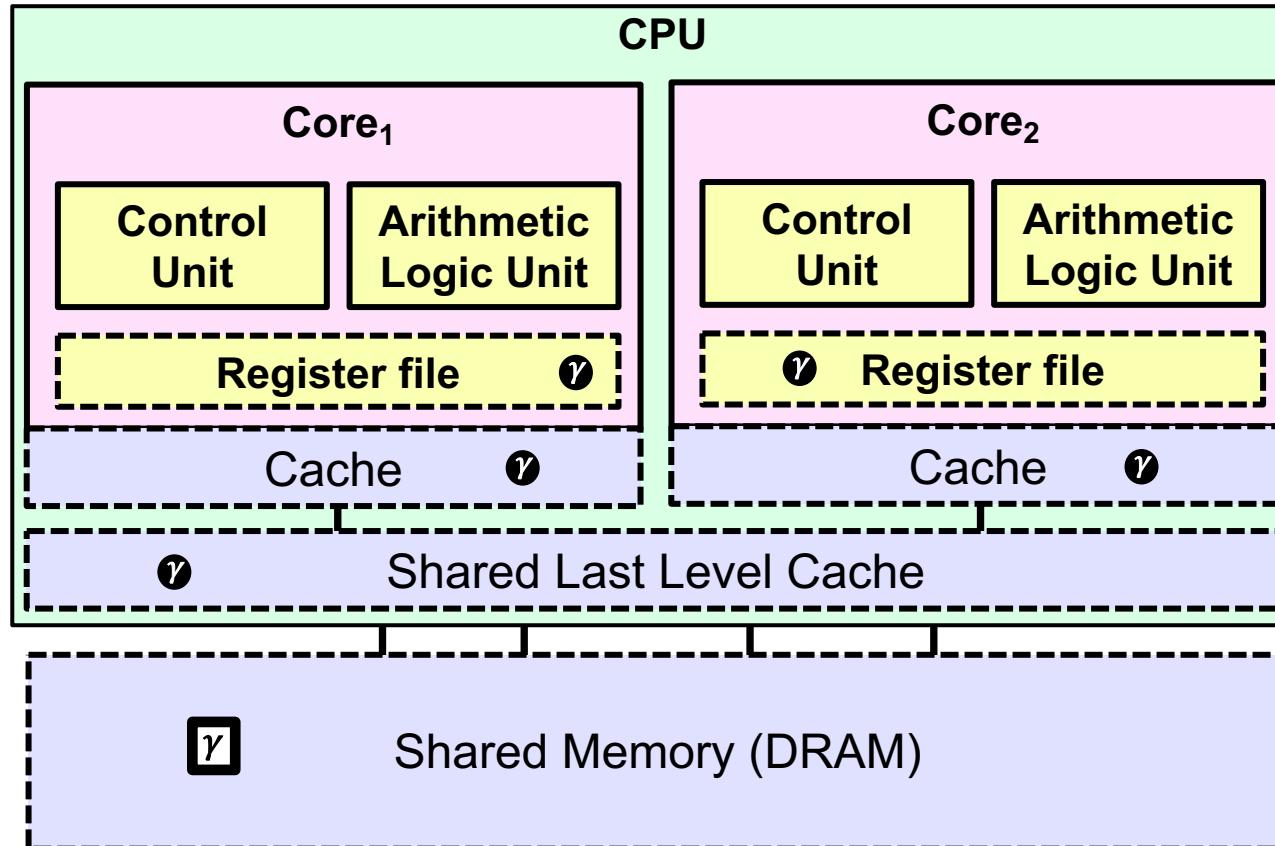
    z=c;
    for (iter=0; iter<MXITR; iter++){
        temp = (z.r*z.r)-(z.i*z.i)+c.r;
        z.i = z.r*z.i*2+c.i;
        z.r = temp;
        if ((z.r*z.r+z.i*z.i)>4.0) {
            #pragma omp critical
                numoutside++;
            break;
        }
    }
}
```

- eps was not initialized
- Protect updates of numoutside
- Which value of c does testpoint() see? Global or private?

- Introduction to OpenMP
- Creating Threads
- Synchronization
- Parallel Loops
- Data Environment
- Memory Model
- Irregular Parallelism and Tasks
- Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data

# Memory Models ...

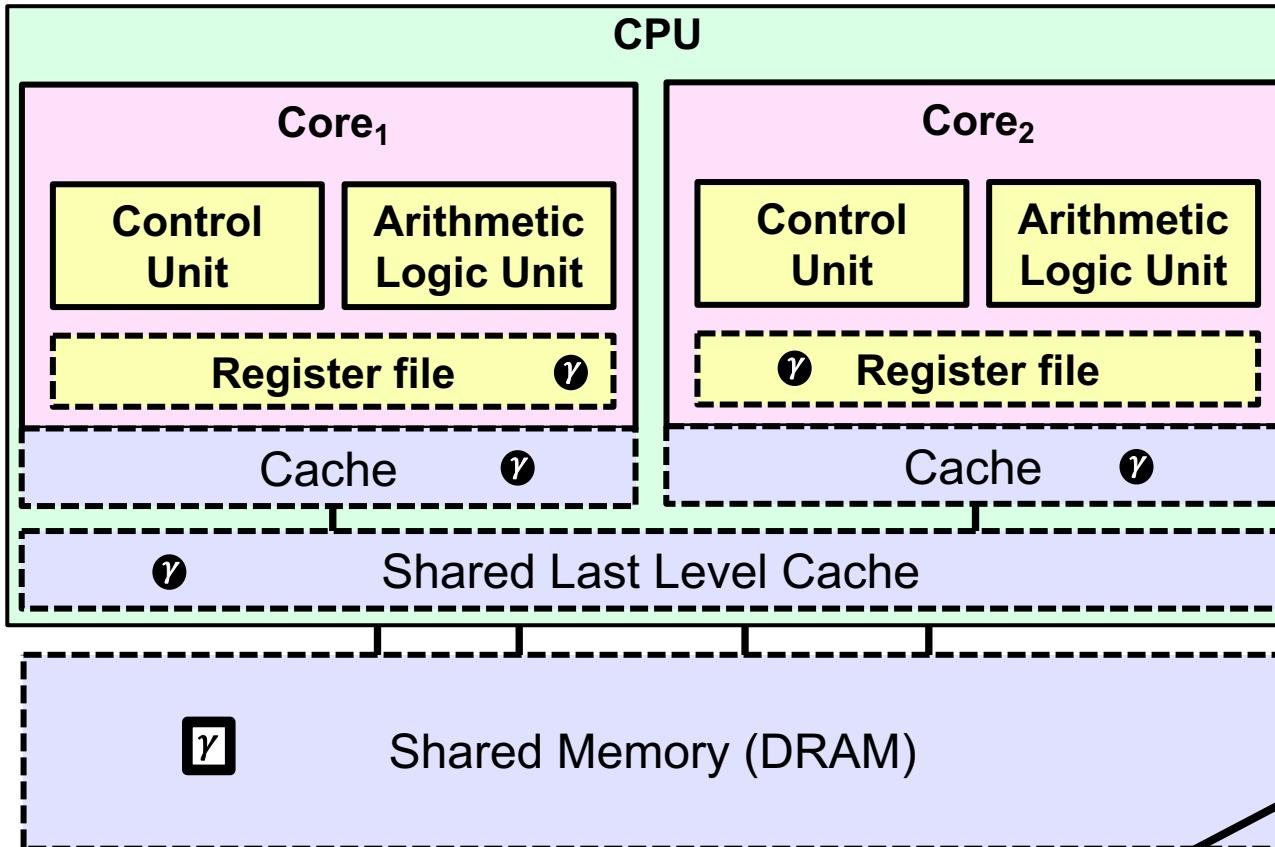
- Programming models for Multithreading support shared memory.
- All threads share an address space ... but consider the variable  $\gamma$



- Multiple copies of a variable (such as  $\gamma$ ) may be present at various levels of cache, or in registers and they may ALL have different values.
- So which value of  $\gamma$  is the one a thread should see at any point in a computation?

# Memory Models ...

- Programming models for Multithreading support shared memory.
- All threads share an address space ... but consider the variable  $\gamma$



A memory consistency model (or “memory model” for short) provides the rules needed to answer this question.

- Multiple copies of a variable (such as  $\gamma$ ) may be present at various levels of cache, or in registers and they may ALL have different values.
- So which value of  $\gamma$  is the one a thread should see at any point in a computation?

# OpenMP and Relaxed Consistency

- Most (if not all) multithreading programming models (including OpenMP) supports a **relaxed-consistency** memory model
  - Threads can maintain a **temporary view** of shared memory that is not consistent with that of other threads
  - These temporary views are made consistent only at certain points in the program
  - The operation that enforces consistency is called the **flush operation\***

\*Note: in OpenMP 5.0 the name for the flush described here was changed to a "strong flush". This was done so we could distinguish the traditional OpenMP flush (the strong flush) from the new synchronizing flushes (acquire flush and release flush).

# Flush Operation

- Defines a sequence point at which a thread is guaranteed to see a consistent view of memory\*
  - Previous read/writes by this thread have completed and are visible to other threads
  - No subsequent read/writes by this thread have occurred
- A flush operation is analogous to a **fence** in other shared memory APIs

\* This applies to the set of shared variables visible to a thread at the point the flush is encountered. We call this “**the flush set**”

# Flush Example

- Flush forces data to be updated in memory so other threads see the most recent value\*

```
double A;  
A = compute();  
#pragma omp flush(A)  
    // flush to memory to make sure other  
    // threads can pick up the right value
```

Note: OpenMP's flush is analogous to a fence in other shared memory APIs

\* If you pass a list of variables to the flush directive, then that list is “**the flush set**”

# What is the BIG DEAL with Flush?

- Compilers routinely reorder instructions implementing a program
  - Can better exploit the functional units, keep the machine busy, hide memory latencies, etc.
- Compilers generally cannot move instructions:
  - Past a barrier
  - Past a flush on all variables
- But it can move them past a flush with a list of variables so long as those variables are not accessed
- Keeping track of consistency when flushes are used can be confusing ... especially if “flush(list)” is used.

Warning: the flush operation (a strong flush) does not actually synchronize different threads. It just ensures that a thread's variables are made consistent with main memory

# Flush and Synchronization

- A flush operation is implied by OpenMP synchronizations, e.g.,
    - at entry/exit of parallel regions
    - at implicit and explicit barriers
    - at entry/exit of critical regions
    - ....
- (but not on entry to worksharing regions)

**WARNING:**

If you find yourself wanting to write code with explicit flushes, stop and get help. It is very difficult to manage flushes on your own. Even experts often get them wrong.

This is why we defined OpenMP constructs to automatically apply flushes most places where you really need them.

- Introduction to OpenMP
- Creating Threads
- Synchronization
- Parallel Loops
- Data Environment
- Memory Model
- • Irregular Parallelism and Tasks
- Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data

# Irregular Parallelism

- Let's call a problem "irregular" when one or both of the following hold:
  - Data Structures are sparse
  - Control structures are not basic for-loops
- Example: Traversing Linked lists:

```
p = listhead ;
while (p) {
    process(p) ;
    p=p->next;
}
```

- Using what we've learned so far, traversing a linked list in parallel using OpenMP is difficult.

# Exercise: Traversing linked lists

- Consider the program linked.c
  - Traverses a linked list computing a sequence of Fibonacci numbers at each node.
- Parallelize this program selecting from the following list of constructs:

```
#pragma omp parallel
#pragma omp for
#pragma omp parallel for
#pragma omp for reduction(op:list)
#pragma omp critical
int omp_get_num_threads();
int omp_get_thread_num();
double omp_get_wtime();
schedule(static[,chunk]) or schedule(dynamic[,chunk])
private(), firstprivate(), default(none)
```

- Hint: Just worry about the while loop that is timed inside main(). You don't need to make any changes to the "list functions"

# Linked Lists with OpenMP (without tasks)

- See the file solutions/linked\_notasks.c

```
while (p != NULL) {
    p = p->next;
    count++;
}

struct node *parr = (struct node*) malloc(count*sizeof(struct node));
p = head;
for(i=0; i<count; i++) {
    parr[i] = p;
    p = p->next;
}
#pragma omp parallel
{
    #pragma omp for schedule(static,1)
    for(i=0; i<count; i++)
        processwork(parr[i]);
}
```

Count number of items in the linked list

Copy pointer to each node into an array

Process nodes in parallel with a for loop

| Number of threads | Schedule   |            |
|-------------------|------------|------------|
|                   | Default    | Static, 1  |
| 1                 | 48 seconds | 45 seconds |
| 2                 | 39 seconds | 28 seconds |

# Linked Lists with OpenMP (without tasks)

- See the file solutions/linked\_notasks.c

```
while (p != NULL) {
    p = p->next;
    count++;
}
struct node *parr = (struct node*) malloc(count*sizeof(struct node));
p = head;
for(i=0; i<count; i++) {
    parr[i] = p;
    p = p->next;
}
#pragma omp parallel
{
    #pragma omp for schedule(static,1)
    for(i=0; i<count; i++)
        processwork(parr[i]);
}
```

Count number of items in the linked list

Copy pointer to each node into an array

Process nodes in parallel with a for loop

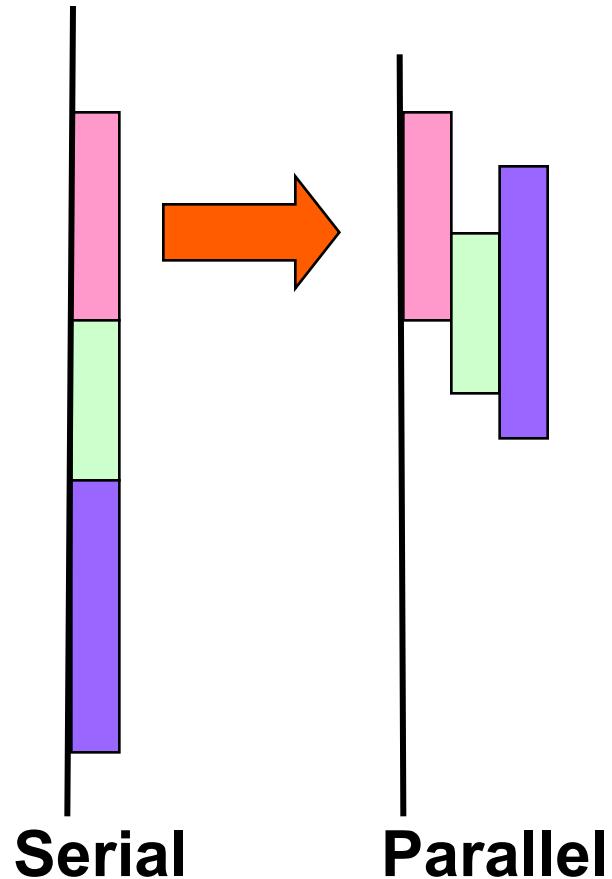
With so much code to add and three passes through the data, this is really ugly.

There has got to be a better way to do this

| Number of threads | Schedule   |            |
|-------------------|------------|------------|
|                   | Default    | Static, 1  |
| 1                 | 48 seconds | 45 seconds |
| 2                 | 39 seconds | 28 seconds |

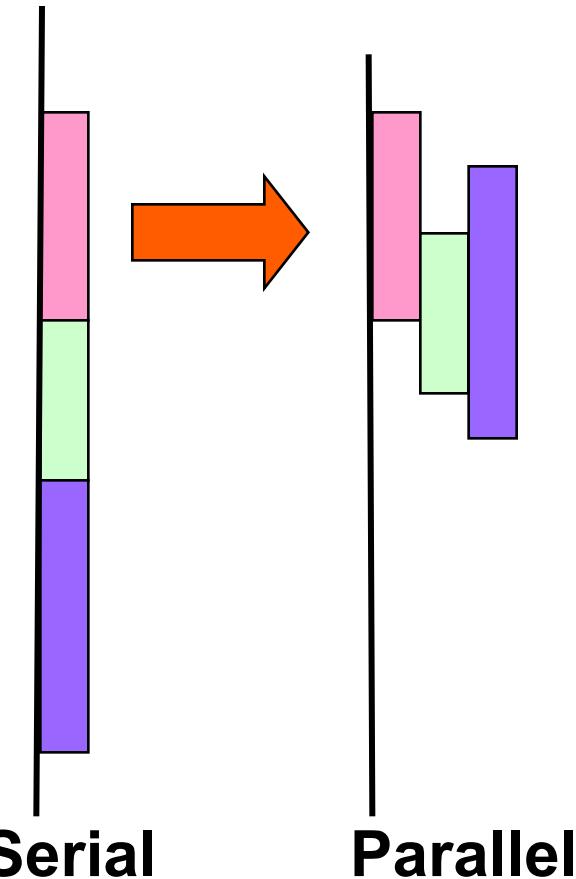
# What are Tasks?

- Tasks are independent units of work
- Tasks are composed of:
  - code to execute
  - data to compute with
- Threads are assigned to perform the work of each task.
  - The thread that encounters the task construct may execute the task immediately.
  - The threads may defer execution until later



# What are Tasks?

- The task construct includes a structured block of code
- Inside a parallel region, a thread encountering a task construct will package up the code block and its data for execution
- Tasks can be nested: i.e. a task may itself generate tasks.



A common Pattern is to have one thread create the tasks while the other threads wait at a barrier and execute the tasks

# Single Worksharing Construct

- The **single** construct denotes a block of code that is executed by only one thread (not necessarily the primary\* thread).
- A barrier is implied at the end of the single block (can remove the barrier with a *nowait* clause).

```
#pragma omp parallel
{
    do_many_things();
#pragma omp single
    {   exchange_boundaries(); }
    do_many_other_things();
}
```

\*This used to be called the “master thread”. The term “master” has been deprecated in OpenMP 5.1 and replaced with the term “primary”.

# Task Directive

```
#pragma omp task [clauses]
```

*structured-block*

---

```
#pragma omp parallel ← Create some threads
{
```

```
    #pragma omp single ← One Thread
    {                                packages tasks
```

```
        #pragma omp task
            fred();
```

```
        #pragma omp task
            daisy();
```

```
        #pragma omp task
            billy();
```

```
}
```

All tasks complete before this barrier is released

# Exercise: Simple tasks

- Write a program using tasks that will “randomly” generate one of two strings:
  - “I think “ “race” “car” “s are fun”
  - “I think “ “car” “race” “s are fun”
- Hint: use tasks to print the indeterminate part of the output (i.e. the “race” or “car” parts).
- This is called a “Race Condition”. It occurs when the result of a program depends on how the OS schedules the threads.
- NOTE: A “data race” is when threads “race to update a shared variable”. They produce race conditions. Programs containing data races are undefined (in OpenMP but also ANSI standards C++'11 and beyond).

```
#pragma omp parallel  
#pragma omp task  
#pragma omp single
```

# Racey Cars: Solution

```
#include <stdio.h>
#include <omp.h>
int main()
{ printf("I think");
  #pragma omp parallel
  {
    #pragma omp single
    {
      #pragma omp task
      printf(" car");
      #pragma omp task
      printf(" race");
    }
  }
  printf("s");
  printf(" are fun!\n");
}
```

# Data Scoping with Tasks

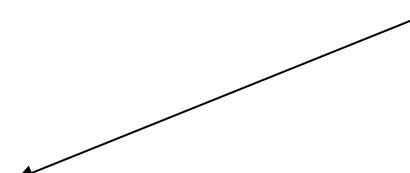
- Variables can be shared, private or firstprivate with respect to task
- These concepts are a little bit different compared with threads:
  - If a variable is **shared** on a task construct, the references to it inside the construct are to the storage with that name at the point where the task was encountered
  - If a variable is **private** on a task construct, the references to it inside the construct are to new uninitialized storage that is created when the task is executed
  - If a variable is **firstprivate** on a construct, the references to it inside the construct are to new storage that is created and initialized with the value of the existing storage of that name when the task is encountered

# Data Scoping Defaults

- The behavior you want for tasks is usually firstprivate, because the task may not be executed until later (and variables may have gone out of scope)
  - Variables that are private when the task construct is encountered are firstprivate by default
- Variables that are shared in all constructs starting from the innermost enclosing parallel construct are shared by default

```
#pragma omp parallel shared(A) private(B)
{
    ...
#pragma omp task
    {
        int C;
        compute(A, B, C);
    }
}
```

A is shared  
B is firstprivate  
C is private



# Exercise: Traversing linked lists

- Consider the program linked.c
  - Traverses a linked list computing a sequence of Fibonacci numbers at each node.
- Parallelize this program selecting from the following list of constructs:

```
#pragma omp parallel
#pragma omp single
#pragma omp task
int omp_get_num_threads();
int omp_get_thread_num();
double omp_get_wtime();
private(), firstprivate()
```

- Hint: Just worry about the contents of main(). You don't need to make any changes to the "list functions"

# Parallel Linked List Traversal

```
#pragma omp parallel
{
    #pragma omp single
    {
        p = listhead ;
        while (p) {
            #pragma omp task firstprivate(p)
            {
                process (p) ;
            }
            p=next (p) ;
        }
    }
}
```

Only one thread packages tasks

makes a copy of p  
when the task is  
packaged

# When/Where are Tasks Complete?

- At thread barriers (explicit or implicit)
  - all tasks generated inside a region must complete at the next barrier encountered by the threads in that region. Common examples:
    - **Tasks generated inside a single construct:** all tasks complete before exiting the barrier on the single.
    - **Tasks generated inside a parallel region:** all tasks complete before exiting the barrier at the end of the parallel region.
- At taskwait directive
  - i.e. Wait until all tasks defined in the current task have completed.  
`#pragma omp taskwait`
  - Note: applies only to tasks generated in the current task, not to “descendants” .

# Example

```
#pragma omp parallel
{
    #pragma omp single
    {
        #pragma omp task
        fred();
        #pragma omp task
        daisy();
#pragma omp taskwait
        #pragma omp task
        billy();
    }
}
```

**fred()** and **daisy()** must complete before **billy()** starts, but this does not include tasks created inside **fred()** and **daisy()**

All tasks including those created inside **fred()** and **daisy()** must complete before exiting this barrier

# Example

```
#pragma omp parallel
{
    #pragma omp single nowait
    {
        #pragma omp task
        fred();
        #pragma omp task
        daisy();
        #pragma omp taskwait
        #pragma omp task
        billy();
    }
}
```

The barrier at the end of the single is expensive and not needed since you get the barrier at the end of the parallel region. So use nowait to turn it off.

All tasks including those created inside **fred()** and **daisy()** must complete before exiting this barrier

# Example: Fibonacci numbers

```
int fib (int n)
{
    int x,y;
    if (n < 2) return n;

    x = fib(n-1);
    y = fib (n-2);
    return (x+y);
}

Int main()
{
    int NW = 5000;
    fib(NW);
}
```

- $F_n = F_{n-1} + F_{n-2}$
- Inefficient  $O(n^2)$  recursive implementation!

# Parallel Fibonacci

```
int fib (int n)
{
    int x,y;
    if (n < 2) return n;

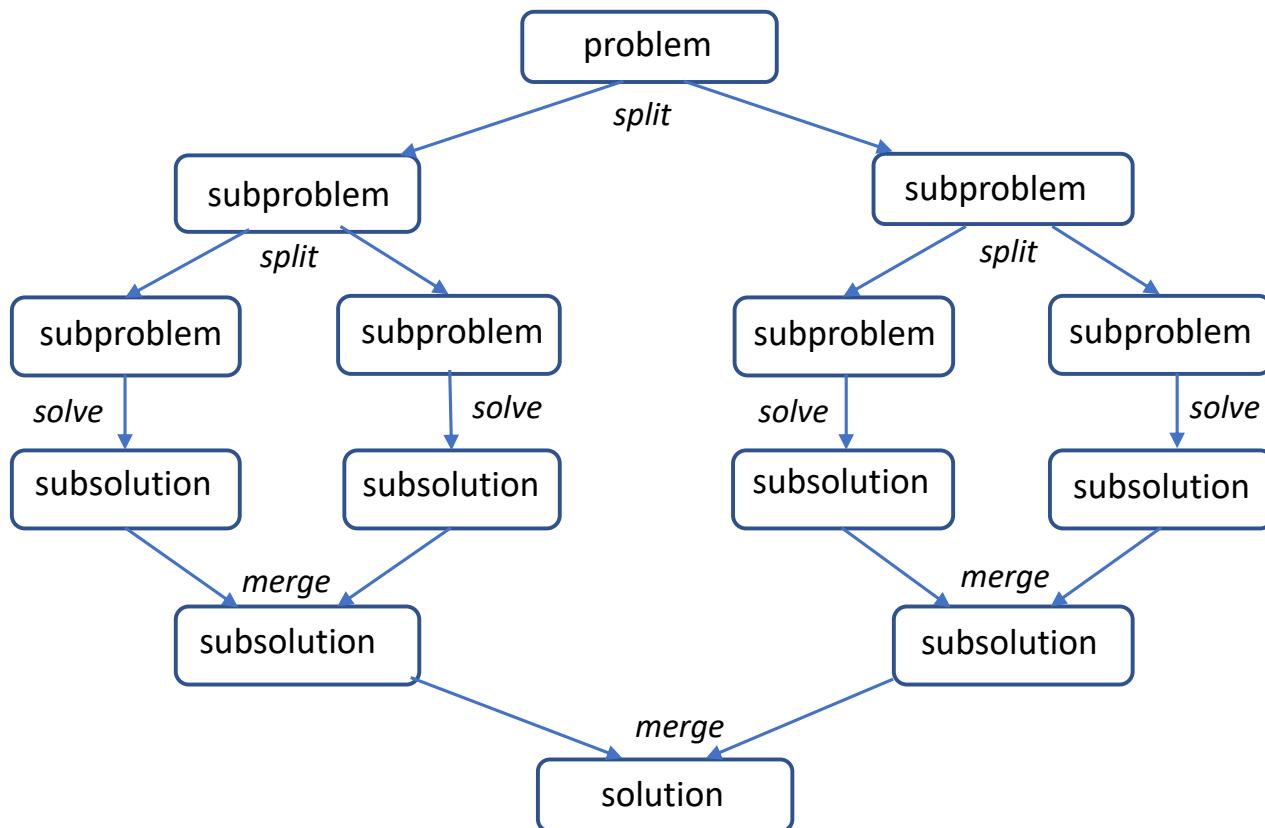
#pragma omp task shared(x)
    x = fib(n-1);
#pragma omp task shared(y)
    y = fib (n-2);
#pragma omp taskwait
    return (x+y);
}

Int main()
{
    int NW = 5000;
#pragma omp parallel
{
    #pragma omp single
        fib(NW);
}
}
```

- Binary tree of tasks
- Traversed using a recursive function
- A task cannot complete until all tasks below it in the tree are complete (enforced with taskwait)
- **x, y** are local, and so by default they are private to current task
  - must be shared on child tasks so they don't create their own firstprivate copies at this level!

# Divide and Conquer

- Split the problem into smaller sub-problems; continue until the sub-problems can be solved directly



- 3 Options for parallelism:
  - Do work as you split into sub-problems
  - Do work only at the leaves
  - Do work as you recombine

# Exercise: PI with tasks

- Go back to the original pi.c program
  - Parallelize this program using OpenMP tasks

```
#pragma omp parallel
#pragma omp task
#pragma omp taskwait
#pragma omp single
double omp_get_wtime()
int omp_get_thread_num();
int omp_get_num_threads();
```

- Hint: first create a recursive pi program and verify that it works. **Think about the computation you want to do at the leaves. If you go all the way down to one iteration per leaf-node, won't you just swamp the system with tasks?**

# Program: OpenMP tasks

```
include <omp.h>
static long num_steps = 100000000;
#define MIN_BLK 10000000
double pi_comp(int Nstart,int Nfinish,double step)
{ int i,iblk;
  double x, sum = 0.0,sum1, sum2;
  if (Nfinish-Nstart < MIN_BLK){
    for (i=Nstart;i< Nfinish; i++){
      x = (i+0.5)*step;
      sum = sum + 4.0/(1.0+x*x);
    }
  }
  else{
    iblk = Nfinish-Nstart;
    #pragma omp task shared(sum1)
      sum1 = pi_comp(Nstart,      Nfinish-iblk/2,step);
    #pragma omp task shared(sum2)
      sum2 = pi_comp(Nfinish-iblk/2, Nfinish,      step);
    #pragma omp taskwait
      sum = sum1 + sum2;
  }
  return sum;
}
```

```
int main ()
{
  int i;
  double step, pi, sum;
  step = 1.0/(double) num_steps;
  #pragma omp parallel
  {
    #pragma omp single
      sum =
        pi_comp(0,num_steps,step);
    }
    pi = step * sum;
}
```

# Results\*: Pi with tasks

| threads | 1 <sup>st</sup> SPMD | SPMD critical | PI Loop | Pi tasks |
|---------|----------------------|---------------|---------|----------|
| 1       | 1.86                 | 1.87          | 1.91    | 1.87     |
| 2       | 1.03                 | 1.00          | 1.02    | 1.00     |
| 3       | 1.08                 | 0.68          | 0.80    | 0.76     |
| 4       | 0.97                 | 0.53          | 0.68    | 0.52     |

\*Intel compiler (icpc) with no optimization on Apple OS X 10.7.3 with a dual core (four HW thread) Intel® Core™ i5 processor at 1.7 Ghz and 4 Gbyte DDR3 memory at 1.333 Ghz.

# Using Tasks

- Don't use tasks for things already well supported by OpenMP
  - e.g. standard do/for loops
  - the overhead of using tasks is greater
- Don't expect miracles from the runtime
  - best results usually obtained where the user controls the number and granularity of tasks

- Introduction to OpenMP
- Creating Threads
- Synchronization
- Parallel Loops
- Data Environment
- Memory Model
- Irregular Parallelism and Tasks
- • Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data

# The OpenMP Common Core: Most OpenMP programs only use these 21 items

| OpenMP pragma, function, or clause   | Concepts  |
|--|---|
| #pragma omp parallel   | Parallel region, teams of threads, structured block, interleaved execution across threads.  |
| void omp_set_thread_num()<br>int omp_get_thread_num()<br>int omp_get_num_threads() | Default number of threads and internal control variables.<br>SPMD pattern: Create threads with a parallel region and split up the work using the number of threads and the thread ID. |
| double omp_get_wtime()   | Speedup and Amdahl's law. False sharing and other performance issues.   |
| setenv OMP_NUM_THREADS N   | Setting the internal control variable for the default number of threads with an environment variable  |
| #pragma omp barrier<br>#pragma omp critical  | Synchronization and race conditions.<br>Revisit interleaved execution.  |
| #pragma omp for<br>#pragma omp parallel for  | Worksharing, parallel loops, loop carried dependencies.   |
| reduction(op:list)   | Reductions of values across a team of threads.  |
| schedule (static [,chunk])<br>schedule(dynamic [,chunk])                           | Loop schedules, loop overheads, and load balance.   |
| shared(list), private(list), firstprivate(list)                                    | Data environment.   |
| default(None)  | Force explicit definition of each variable's storage attribute  |
| nowait   | Disabling implied barriers on workshare constructs, the high cost of barriers, and the flush concept (but not the flush directive).   |
| #pragma omp single   | Workshare with a single thread.   |
| #pragma omp task<br>#pragma omp taskwait   | Tasks including the data environment for tasks.   |

# There is Much More to OpenMP than the Common Core

- Synchronization mechanisms
  - locks, synchronizing flushes and several forms of atomic
- Data environment
  - lastprivate, threadprivate, default(private|shared)
- Fine grained task control
  - dependencies, tied vs. untied tasks, task groups, task loops ...
- Vectorization constructs
  - simd, uniform, simdlen, inbranch vs. nobranch, ....
- Map work onto an attached device (such as a GPU)
  - target, teams distribute parallel for, target data ...
- ... and much more. The OpenMP 5.0 specification is over 618 pages!!!

Don't become overwhelmed. Master the common core and move on to other constructs when you encounter problems that require them.

# OpenMP Organizations

- OpenMP Architecture Review Board (ARB) URL, the “owner” of the OpenMP specification:

[www.openmp.org](http://www.openmp.org)

- OpenMP User’s Group (cOMPunity) URL:

[www.community.org](http://www.community.org)

Get involved, join the ARB and cOMPunity.

Help define the future of OpenMP

# Resources

- [www.openmp.org](http://www.openmp.org) has a wealth of helpful resources

The screenshot shows the OpenMP website's "Specifications" page. The header features the "OpenMP®" logo with the tagline "Enabling HPC since 1997" and the subtitle "The OpenMP API specification for parallel programming". The navigation bar includes links for Home, Specifications (which is highlighted in orange), Blog, Community, Resources, News & Events, About, and a search icon. Below the navigation is a breadcrumb trail: Home > Specifications. The main content area is titled "Specifications" and contains two sections: "OpenMP 5.0 Specifications" and "OpenMP 4.5 Specifications". Each section includes a circular icon with a document symbol, a title, and a bulleted list of links. A blue arrow points from the "Supplementary Source Code" link under OpenMP 5.0 to the "OpenMP 4.5 Examples Discussion Forum" link under OpenMP 4.5. A callout box at the bottom left states: "Including a comprehensive collection of examples of code using the OpenMP constructs".

**OpenMP®**  
Enabling HPC since 1997

*The OpenMP API specification for parallel programming*

Home Specifications Blog Community ▾ Resources ▾ News & Events ▾ About ▾

Home > Specifications

## Specifications

### OpenMP 5.0 Specifications

- [OpenMP 5.0 Specification \(PDF\)](#) – Nov 2018 – [HTML Version](#)
- Softcover Version – Purchase from [Amazon](#)
- [OpenMP 5.0 Discussion Forum](#)
- [OpenMP 5.0 Reference Guides](#)
- [OpenMP 5.0 Context Definitions Public Comment Draft](#)
- [Supplementary Source Code](#) – ([GitHub Repository](#))

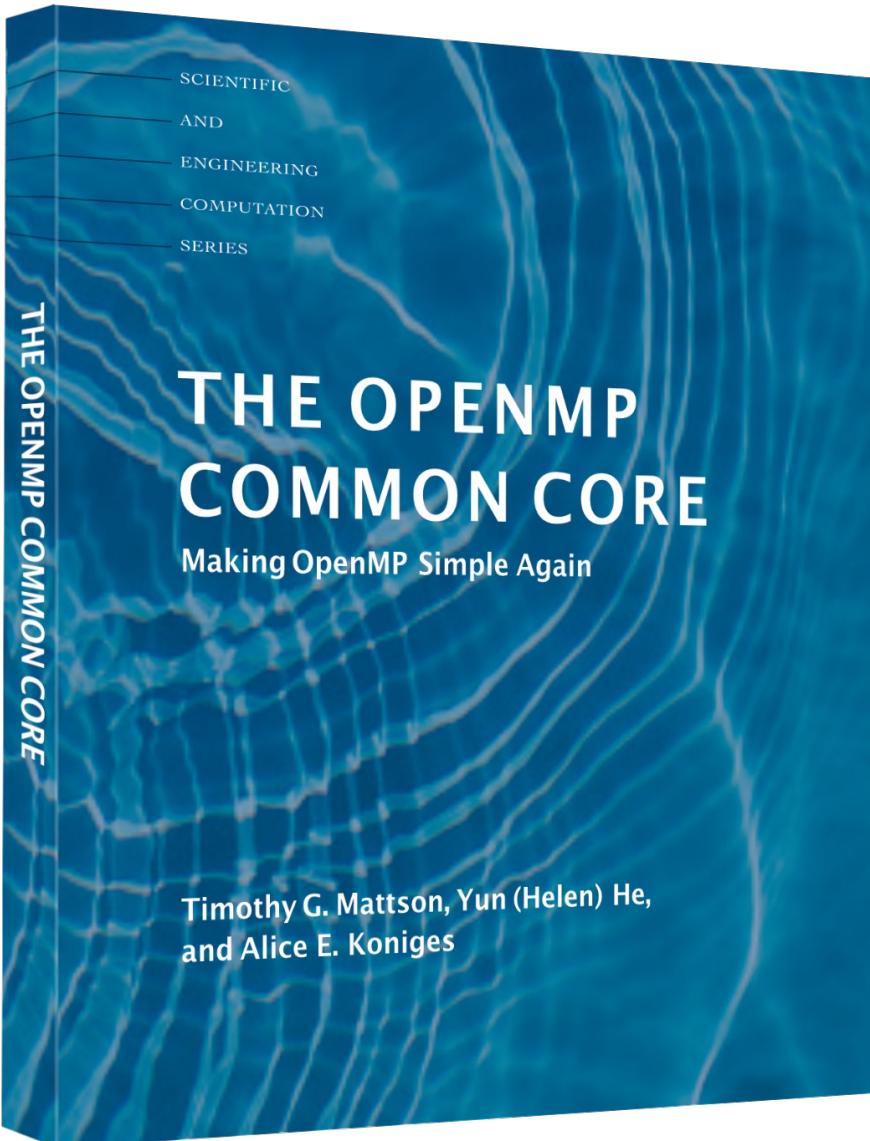
### OpenMP 4.5 Specifications

- [OpenMP 4.5 Complete Specifications \(Nov 2015\) pdf](#)
- [OpenMP 4.5 Discussion Forum](#)
- [OpenMP 4.5 Reference Guide – C/C++ \(Nov 2015\) pdf](#)
- [OpenMP 4.5 Reference Guide – Fortran \(Nov 2015\) pdf](#)
- [OpenMP 4.5 Examples \(Nov 2016\) pdf](#)
- [OpenMP 4.5 Examples Discussion Forum](#)

Including a comprehensive collection of examples of code using the OpenMP constructs

# To learn OpenMP:

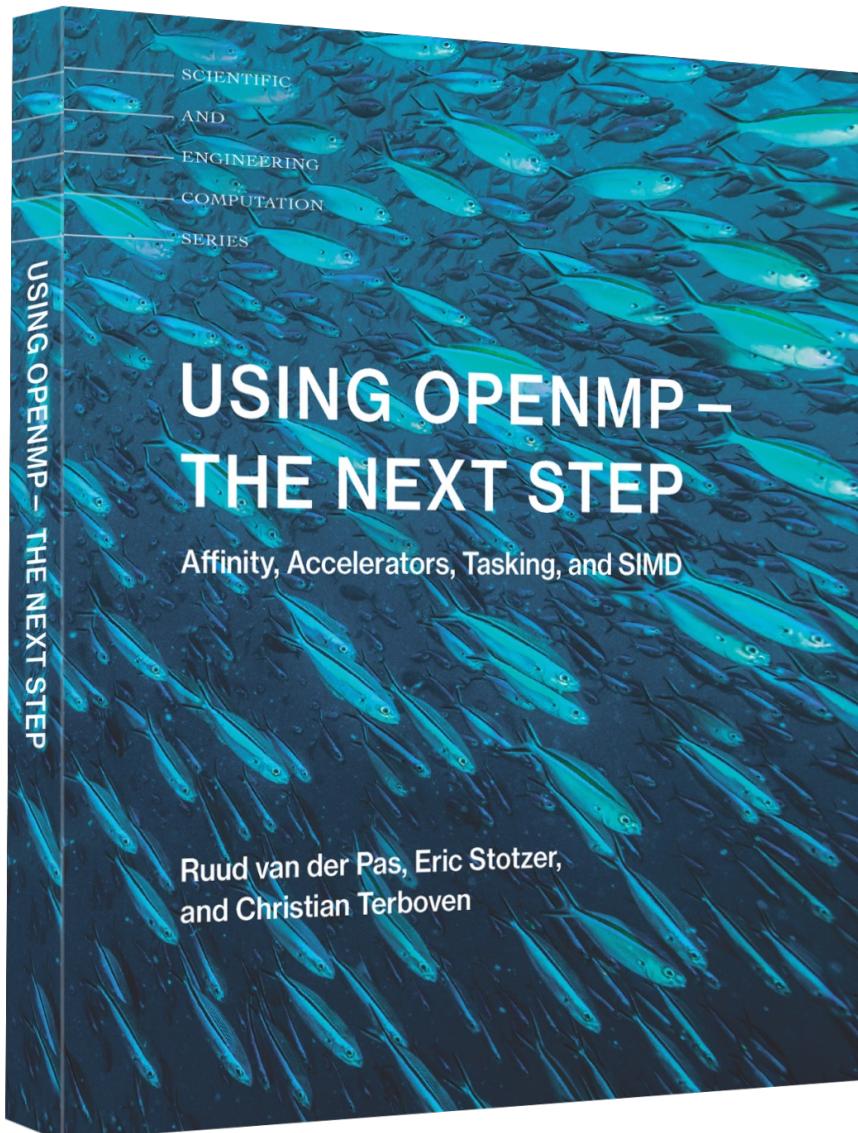
- An exciting new book that Covers the Common Core of OpenMP plus a few key features beyond the common core that people frequently use
- It's geared towards people learning OpenMP, but as one commentator put it ... **everyone at any skill level should read the memory model chapters.**
- Available from MIT Press



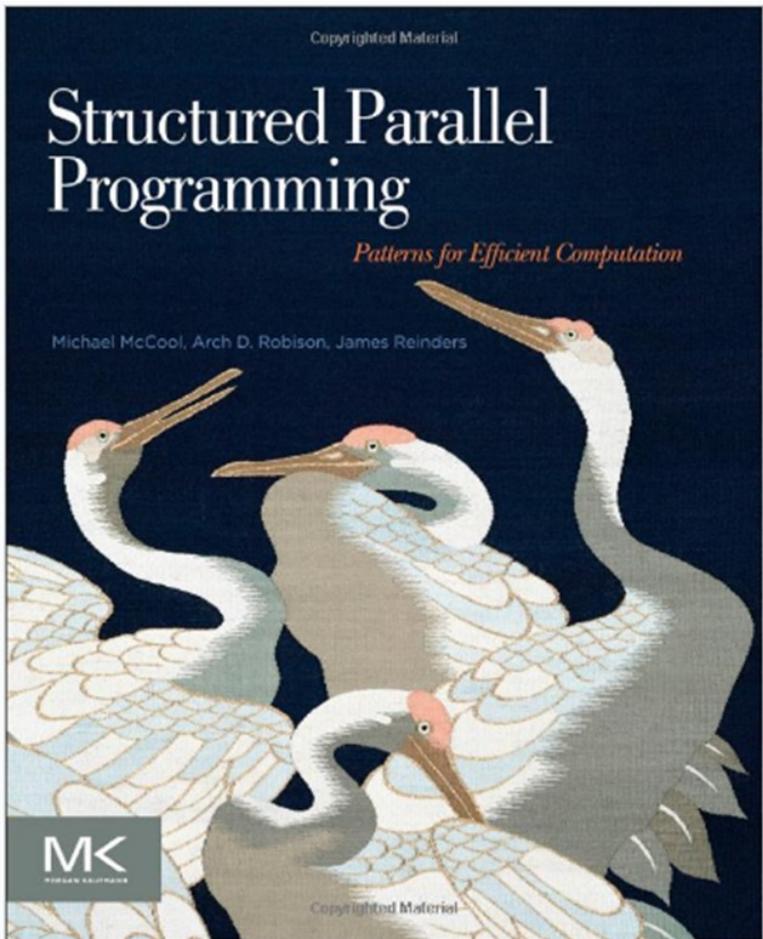
[www.ompcore.com](http://www.ompcore.com) for code samples and the Fortran supplement

# Books about OpenMP

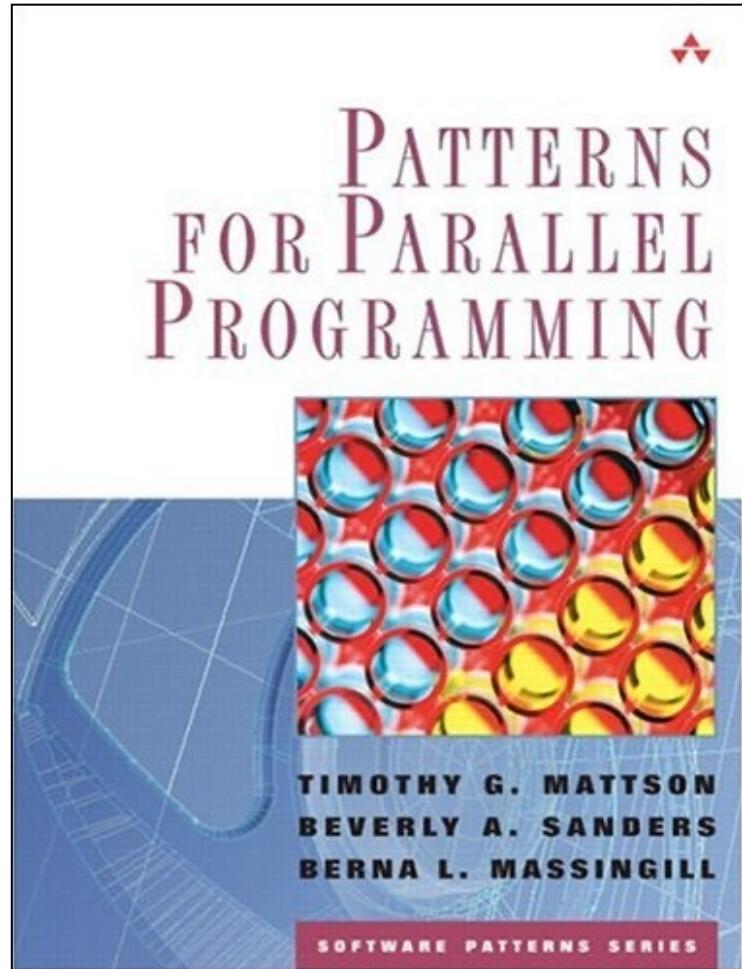
A great book that covers  
OpenMP features beyond  
OpenMP 2.5



# Background references



A great book that explores key patterns with Cilk, TBB, OpenCL, and OpenMP (by McCool, Robison, and Reinders)



- A book about how to “think parallel” with examples in OpenMP, MPI and java

- Introduction to OpenMP
- Creating Threads
- Synchronization
- Parallel Loops
- Data Environment
- Memory Model
- Irregular Parallelism and Tasks
- Recap
- Beyond the Common Core:
  - – Worksharing Revisited
    - Synchronization Revisited: Options for Mutual exclusion
    - Memory models and point-to-point Synchronization
    - Programming your GPU with OpenMP
    - Thread Affinity and Data Locality
    - Thread Private Data

# The Loop Worksharing Constructs

- The loop worksharing construct splits up loop iterations among the threads in a team

```
#pragma omp parallel
{
    #pragma omp for
    for (I=0;I<N;I++){
        NEAT_STUFF(I);
    }
}
```

The variable I is made “private” to each thread by default. You could do this explicitly with a “private(I)” clause

Loop construct name:

- C/C++: for
- Fortran: do

# Loop Worksharing Constructs: The *schedule* clause

- The schedule clause affects how loop iterations are mapped onto threads
  - **schedule(static [,chunk])**
    - Deal-out blocks of iterations of size “chunk” to each thread.
  - **schedule(dynamic[,chunk])**
    - Each thread grabs “chunk” iterations off a queue until all iterations have been handled.
  - **schedule(guided[,chunk])**
    - Threads dynamically grab blocks of iterations. The size of the block starts large and shrinks down to size “chunk” as the calculation proceeds.
  - **schedule(runtime)**
    - Schedule and chunk size taken from the OMP\_SCHEDULE environment variable (or the runtime library) ... vary schedule without a recompile!
  - **Schedule(auto)**
    - Schedule is left up to the runtime to choose (does not have to be any of the above).

OpenMP 4.5 added modifiers monotonic, nonmonotonic and simd.

# Loop Worksharing Constructs: The schedule clause

| Schedule Clause | When To Use   |  |
|-----------------|---|--|
| <b>STATIC</b>   | <b>Pre-determined and predictable by the programmer</b>                       | Least work at runtime : scheduling done at compile-time          |
| <b>DYNAMIC</b>  | <b>Unpredictable, highly variable work per iteration</b>                      | Most work at runtime : complex scheduling logic used at run-time |
| <b>GUIDED</b>   | <b>Special case of dynamic to reduce scheduling overhead</b>                  |  |
| <b>AUTO</b>     | <b>When the runtime can “learn” from previous executions of the same loop</b> |  |

# Nested Loops

- For perfectly nested rectangular loops we can parallelize multiple loops in the nest with the collapse clause:

```
#pragma omp parallel for collapse(2)
```

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

Number of loops  
to be  
parallelized,  
counting from  
the outside

- Will form a single loop of length NxM and then parallelize that.
- Useful if N is O(no. of threads) so parallelizing the outer loop makes balancing the load difficult.

# Sections Worksharing Construct

- The *Sections* worksharing construct gives a different structured block to each thread.

```
#pragma omp parallel
{
    #pragma omp sections
    {
        #pragma omp section
            x_calculation();
        #pragma omp section
            y_calculation();
        #pragma omp section
            z_calculation();
    }
}
```

By default, there is a barrier at the end of the “omp sections”. Use the “nowait” clause to turn off the barrier.

# Array Sections with Reduce

```
#include <stdio.h>
#define N 100
void init(int n, float (*b)[N]);
int main(){
    int i,j; float a[N], b[N][N]; init(N,b);
    for(i=0; i<N; i++) a[i]=0.0e0;
```

Works the same as any other reduce ... a private array is formed for each thread, element wise combination across threads and then with original array at the end

```
#pragma omp parallel for reduction(+:a[0:N]) private(j)
for(i=0; i<N; i++){
    for(j=0; j<N; j++){
        a[j] += b[i][j];
    }
}
printf(" a[0] a[N-1]: %f %f\n", a[0], a[N-1]);
return 0;
```

# Exercise

- Go back to your parallel mandel.c program.
- Using what we've learned in this block of slides can you improve the runtime?

# Optimizing mandel.c

```
wtime = omp_get_wtime();
#pragma omp parallel for collapse(2) schedule(runtime) firstprivate(eps) private(j,c)
for (i=0; i<NPOINTS; i++) {
    for (j=0; j<NPOINTS; j++) {
        c.r = -2.0+2.5*(double)(i)/(double)(NPOINTS)+eps;
        c.i = 1.125*(double)(j)/(double)(NPOINTS)+eps;
        testpoint(c);
    }
}
wtime = omp_get_wtime() - wtime;
```

```
$ export OMP_SCHEDULE="dynamic,100"
$ ./mandel_par
```

|                                   |           |
|-----------------------------------|-----------|
| default schedule                  | 0.48 secs |
| schedule(dynamic,100)             | 0.39 secs |
| collapse(2) schedule(dynamic,100) | 0.34 secs |

Four threads on a dual core Apple laptop (Macbook air ... 2.2 Ghz Intel Core i7 with 8 GB memory)  
and the gcc version 9.1. Times are the minimum time from three runs

- Introduction to OpenMP
- Creating Threads
- Synchronization
- Parallel Loops
- Data Environment
- Memory Model
- Irregular Parallelism and Tasks
- Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data

# Synchronization

Synchronization is used to impose order constraints between threads and to protect access to shared data

- High level synchronization included in the common core:

- critical
- barrier

Covered earlier

- Other, more advanced, synchronization operations:

- atomic
- ordered
- flush
- locks (both simple and nested)

Covered in this section

# Synchronization: Atomic

- Atomic provides mutual exclusion but only applies to the update of a memory location (the update of X in the following example)

```
#pragma omp parallel
```

```
{
```

```
    double B;
```

```
    B = DOIT();
```

```
#pragma omp atomic
```

```
    X += big_ugly(B);
```

```
}
```

# Synchronization: Atomic

- Atomic provides mutual exclusion but only applies to the update of a memory location (the update of X in the following example)

```
#pragma omp parallel  
{  
    double B, tmp;  
    B = DOIT();  
    tmp = big_ugly(B);  
#pragma omp atomic  
    X += tmp;  
}
```

Atomic only protects the  
read/update of X

# The OpenMP 3.1 Atomics (1 of 2)

- Atomic was expanded to cover the full range of common scenarios where you need to protect a memory operation so it occurs atomically:

```
# pragma omp atomic [read | write | update | capture]
```

- Atomic can protect loads

```
# pragma omp atomic read
```

```
v = x;
```

- Atomic can protect stores

```
# pragma omp atomic write
```

```
x = expr;
```

- Atomic can protect updates to a storage location (this is the default behavior ... i.e. when you don't provide a clause)

```
# pragma omp atomic update
```

```
x++; or ++x; or x--; or -x; or
```

```
x binop= expr; or x = x binop expr;
```

This is the  
original OpenMP  
atomic

# The OpenMP 3.1 Atomics (2 of 2)

- Atomic can protect the assignment of a value (its capture) AND an associated update operation:

```
# pragma omp atomic capture  
statement or structured block
```

- Where the statement is one of the following forms:

**v = x++;**    **v = ++x;**    **v = x--;**    **v = -x;**    **v = x binop expr;**

- Where the structured block is one of the following forms:

{v = x; x binop = expr;}

{v=x; x=x binop expr;}

{v = x; x++;}

{++x; v=x:}

{v = x; x--;}

{--x; v = x;}

{x binop = expr; v = x;}

{X = x binop expr; v = x;}

{v=x; ++x:}

{x++; v = x;}

{v = x; --x;}

{x--; v = x;}

The capture semantics in atomic were added to map onto common hardware supported atomic operations and to support modern lock free algorithms

# Synchronization: Lock Routines

- Simple Lock routines:
  - A simple lock is available if it is unset.
    - `omp_init_lock()`, `omp_set_lock()`,  
`omp_unset_lock()`, `omp_test_lock()`, `omp_destroy_lock()`
- Nested Locks
  - A nested lock is available if it is unset or if it is set but owned by the thread executing the nested lock function
    - `omp_init_nest_lock()`, `omp_set_nest_lock()`, `omp_unset_nest_lock()`,  
`omp_test_nest_lock()`, `omp_destroy_nest_lock()`

A lock implies a memory fence (a “flush”) of all thread visible variables

Note: a thread always accesses the most recent copy of the lock, so you don't need to use a flush on the lock variable.

Locks with hints were added in OpenMP 4.5 to suggest a lock strategy based on intended use (e.g. contended, uncontended, speculative, unspeculative)

# Synchronization: Simple Locks Example

- Count odds and evens in an input array(x) of N random values.

```
int i, ix, even_count = 0, odd_count = 0;  
omp_lock_t odd_lck, even_lck;  
omp_init_lock(&odd_lck);  
omp_init_lock(&even_lck);
```

One lock per case ... even and odd

```
#pragma omp parallel for private(ix) shared(even_count, odd_count)  
for(i=0; i<N; i++){  
    ix = (int) x[i]; //truncate to int
```

```
    if((int) x[i])%2 == 0 {  
        omp_set_lock(&even_lck);  
        even_count++;  
        omp_unset_lock(&even_lck);  
    }  
    else{  
        omp_set_lock(&odd_lck);  
        odd_count++;  
        omp_unset_lock(&odd_lck);  
    }  
}  
omp_destroy_lock(&odd_lck);  
omp_destroy_lock(&even_lck);  
}
```

Enforce mutual exclusion updates,  
but in parallel for each case.

Free-up storage when done.

# Exercise

- In the file hist.c, we provide a program that generates a large array of random numbers and then generates a histogram of values.
- This is a "quick and informal" way to test a random number generator ... if all goes well the bins of the histogram should be the same size.
- Parallelize the filling of the histogram. You must assure that your program is race free and gets the same result as the sequential program.
- Using everything we've covered today, **manage updates to shared data in two different ways.** Try to minimize the time to generate the histogram.
- Time ONLY the assignment to the histogram. Can you beat the sequential time?

# Histogram Program: Critical section

- A critical section means that only one thread at a time can update a histogram bin ... but this effectively serializes the loops and adds huge overhead as the runtime manages all the threads waiting for their turn for the update.

```
#pragma omp parallel for
for(i=0;i<NVALS;i++){
    ival = (int) x[i];
    #pragma omp critical
        hist[ival]++;
}
```

Easy to write and  
correct, but terrible  
performance

# Histogram program: one lock per histogram bin

- Example: conflicts are rare, but to play it safe, we must assure mutual exclusion for updates to histogram elements.

```
#pragma omp parallel for
for(i=0;i<NBUCKETS; i++){
    omp_init_lock(&hist_locks[i]);
    hist[i] = 0;
}

#pragma omp parallel for
for(i=0;i<NVALS;i++){
    ival = (int) x[i];
    omp_set_lock(&hist_locks[ival]);
    hist[ival]++;
    omp_unset_lock(&hist_locks[ival]);
}

#pragma omp parallel for
for(i=0;i<NBUCKETS; i++)
    omp_destroy_lock(&hist_locks[i]);
```

One lock per element of hist

Enforce mutual exclusion on update to hist array

Free-up storage when done.

# Histogram program: reduction with an array

- We can give each thread a copy of the histogram, they can fill them in parallel, and then combine them when done

```
#pragma omp parallel for reduction(+:hist[0:Nbins])
for(i=0;i<NVALS;i++){
    ival = (int) x[i];
    hist[ival]++;
}
```

Easy to write and correct, Uses a lot of memory on the stack, but its fast ... sometimes faster than the serial method.

|                                       |              |
|---------------------------------------|--------------|
| sequential                            | 0.0019 secs  |
| critical                              | 0.079 secs   |
| Locks per bin                         | 0.029 secs   |
| Reduction, replicated histogram array | 0.00097 secs |

1000000 random values in X sorted into 50 bins. Four threads on a dual core Apple laptop (Macbook air ... 2.2 Ghz Intel Core i7 with 8 GB memory) and the gcc version 9.1. Times are for the above loop only (we do not time set-up for locks, destruction of locks or anything else)

# Sometimes when working with multiple interacting locks, you have to pay attention to the locking orders

Lock Example from Gafort (SpecOMP'2001)

- Genetic algorithm in Fortran
- Most “interesting” loop: shuffle the population.
  - Original loop is not parallel; performs pair-wise swap of an array element with another, randomly selected element. There are 40,000 elements.
  - Parallelization idea:
    - Perform the swaps in parallel
    - Need to prevent simultaneous access to same array element: use one lock per array element → 40,000 locks.

# Parallel Loop In shuffle.f of Gafort

Exclusive access  
to array  
elements.

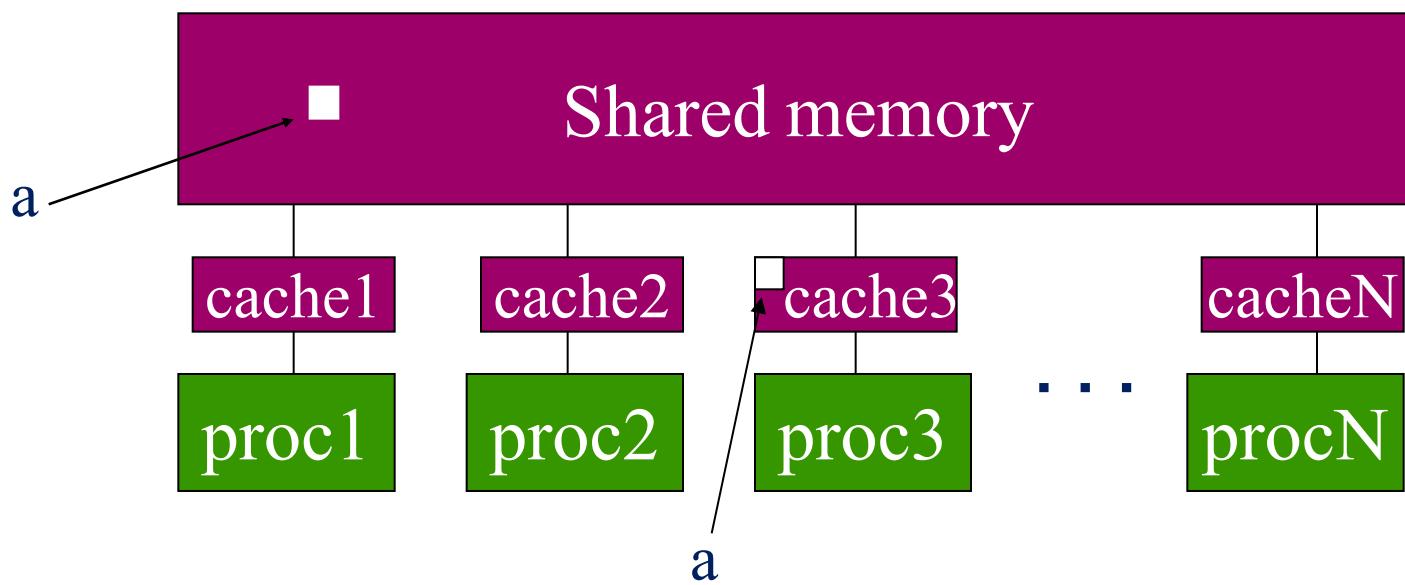
Ordered locking  
prevents  
deadlock.

```
!$OMP PARALLEL PRIVATE(rand, iother, itemp, temp, my_cpu_id)
    my_cpu_id = 1
!$ my_cpu_id = omp_get_thread_num() + 1
!$OMP DO
    DO j=1,npopsiz-1
        CALL ran3(1,rand,my_cpu_id,0)
        iother=j+1+DINT(DBLE(npopsiz-j)*rand)
        !$ IF (j < iother) THEN
        !$     CALL omp_set_lock(lck(j))
        !$     CALL omp_set_lock(lck(iother))
        !$ ELSE
        !$     CALL omp_set_lock(lck(iother))
        !$     CALL omp_set_lock(lck(j))
        !$ END IF
        itemp(1:nchrome)=iparent(1:nchrome,iother)
        iparent(1:nchrome,iother)=iparent(1:nchrome,j)
        iparent(1:nchrome,j)=itemp(1:nchrome)
        temp=fitness(iother)
        fitness(iother)=fitness(j)
        fitness(j)=temp
        !$ IF (j < iother) THEN
        !$     CALL omp_unset_lock(lck(iother))
        !$     CALL omp_unset_lock(lck(j))
        !$ ELSE
        !$     CALL omp_unset_lock(lck(j))
        !$     CALL omp_unset_lock(lck(iother))
        !$ END IF
    END DO
!$OMP END DO
!$OMP END PARALLEL
```

- Introduction to OpenMP
- Creating Threads
- Synchronization
- Parallel Loops
- Data Environment
- Memory Model
- Irregular Parallelism and Tasks
- Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data

# OpenMP Memory Model

- OpenMP supports a shared memory model
- All threads share an address space, where variable can be stored or retrieved:



- Threads maintain their own temporary view of memory as well ... the details of which are not defined in OpenMP but this temporary view typically resides in caches, registers, write-buffers, etc.

# Flush Operation

- Defines a sequence point at which a thread enforces a consistent view of memory.
- For variables visible to other threads and associated with the flush operation (**the flush-set**)
  - The compiler can't move loads/stores of the flush-set around a flush:
    - All previous read/writes of the flush-set by this thread have completed
    - No subsequent read/writes of the flush-set by this thread have occurred
  - Variables in the flush set are moved from temporary storage to shared memory.
  - Reads of variables in the flush set following the flush are loaded from shared memory.

**IMPORTANT POINT:** The flush makes the calling threads temporary view match the view in shared memory. Flush by itself does not force synchronization.

# Memory Consistency: Flush Example

- Flush forces data to be updated in memory so other threads see the most recent value

```
double A;  
A = compute();  
#pragma omp flush(A)  
// flush to memory to make sure other  
// threads can pick up the right value
```

Flush without a list: flush set is all thread visible variables

Flush with a list: flush set is the list of variables

Note: OpenMP's flush is analogous to a fence in other shared memory APIs

# Flush and Synchronization

- A flush operation is implied by OpenMP synchronizations, e.g.,
    - at entry/exit of parallel regions
    - at implicit and explicit barriers
    - at entry/exit of critical regions
    - whenever a lock is set or unset
    - ....
- (but not at entry to worksharing regions or entry/exit of primary\* regions)

\*the term “master” has been deprecated in OpenMP 5.1 and replaced with the term “primary”.

# Example: prod\_cons.c

- Parallelize a producer/consumer program
    - One thread produces values that another thread consumes.
    - Often used with a stream of produced values to implement “pipeline parallelism”
    - The key is to implement pairwise synchronization between threads
- ```
int main()
{
    double *A, sum, runtime;    int flag = 0;

    A = (double *) malloc(N*sizeof(double));

    runtime = omp_get_wtime();

    fill_rand(N, A);    // Producer: fill an array of data

    sum = Sum_array(N, A); // Consumer: sum the array

    runtime = omp_get_wtime() - runtime;

    printf(" In %lf secs, The sum is %lf \n",runtime,sum);
}
```

# Pairwise Synchronization in OpenMP

- OpenMP lacks synchronization constructs that work between pairs of threads.
- When needed, you have to build it yourself.
- Pairwise synchronization
  - Use a shared flag variable
  - Reader spins waiting for the new flag value
  - Use flushes to force updates to and from memory

# Exercise: Producer/Consumer

```
int main()
{
    double *A, sum, runtime;    int numthreads, flag = 0;
    A = (double *)malloc(N*sizeof(double));
#pragma omp parallel sections
{
    #pragma omp section
    {
        fill_rand(N, A);

        flag = 1;
    }

    #pragma omp section
    {
        while (flag == 0){

        }

        sum = Sum_array(N, A);
    }
}
```

Put the flushes in the right places to make this program race-free.

Do you need any other synchronization constructs to make this work?

# Solution (try 1): Producer/Consumer

```
int main()
{
    double *A, sum, runtime;    int numthreads, flag = 0;
    A = (double *)malloc(N*sizeof(double));
    #pragma omp parallel sections
    {
        #pragma omp section
        {
            fill_rand(N, A);
            #pragma omp flush
            flag = 1;
            #pragma omp flush (flag)
        }
        #pragma omp section
        {
            #pragma omp flush (flag)
            while (flag == 0){
                #pragma omp flush (flag)
            }
            #pragma omp flush
            sum = Sum_array(N, A);
        }
    }
}
```

Use flag to Signal when the “produced” value is ready

Flush forces refresh to memory; guarantees that the other thread sees the new value of A

Flush needed on both “reader” and “writer” sides of the communication

Notice you must put the flush inside the while loop to make sure the updated flag variable is seen

This program works with the x86 memory model (loads and stores use relaxed atomics), but it technically has a race ... on the store and later load of flag

# The OpenMP 3.1 Atomics (1 of 2)

- Atomic was expanded to cover the full range of common scenarios where you need to protect a memory operation so it occurs atomically:

**# pragma omp atomic [read | write | update | capture]**

- Atomic can protect loads

**# pragma omp atomic read**

**v = x;**

- Atomic can protect stores

**# pragma omp atomic write**

**x = expr;**

- Atomic can protect updates to a storage location (this is the default behavior ... i.e. when you don't provide a clause)

**# pragma omp atomic update**

**x++; or ++x; or x--; or -x; or**

**x binop= expr; or x = x binop expr;**

This is the  
original OpenMP  
atomic

# The OpenMP 3.1 Atomics (2 of 2)

- Atomic can protect the assignment of a value (its capture) AND an associated update operation:

```
# pragma omp atomic capture  
    statement or structured block
```

- Where the statement is one of the following forms:

**v = x++;**    **v = ++x;**    **v = x--;**    **v = -x;**    **v = x binop expr;**

- Where the structured block is one of the following forms:

{v = x; x binop = expr;}

{v=x; x=x binop expr;}

{v = x; x++;}

{++x; v=x:}

{v = x; x--;}

{--x; v = x;}

{x binop = expr; v = x;}

{X = x binop expr; v = x;}

{v=x; ++x:}

{x++; v = x;}

{v= x; --x;}

{x--; v = x;}

The capture semantics in atomic were added to map onto common hardware supported atomic operations and to support modern lock free algorithms

# Atomics and Synchronization Flags

```
int main()
{
    double *A, sum, runtime;
    int numthreads, flag = 0, flg_tmp;
    A = (double *)malloc(N*sizeof(double));
    #pragma omp parallel sections
    {
        #pragma omp section
        {
            fill_rand(N, A);
            #pragma omp flush
            #pragma omp atomic write
            flag = 1;
            #pragma omp flush (flag)
        }
        #pragma omp section
        {
            while (1){
                #pragma omp flush(flag)
                #pragma omp atomic read
                flg_tmp= flag;
                if (flg_tmp==1) break;
            }
            #pragma omp flush
            sum = Sum_array(N, A);
        }
    }
}
```

This program is truly race free ... the reads and writes of flag are protected so the two threads cannot conflict

Still painful and error prone due to all of the flushes that are required

# OpenMP 4.0 Atomic: Sequential consistency



- Sequential consistency:
  - The order of loads and stores in a race-free program appear in some interleaved order and all threads in the team see this same order.
- OpenMP 4.0 added an optional clause to atomics
  - #pragma omp atomic [read | write | update | capture] [**seq\_cst**]
- In more pragmatic terms:
  - If the seq\_cst clause is included, OpenMP adds a flush without an argument list to the atomic operation so you don't need to.
- In terms of the C++'11 memory model:
  - Use of the seq\_cst clause makes atomics follow the sequentially consistent memory order.
  - Leaving off the seq\_cst clause makes the atomics relaxed.

Advice to programmers: save yourself a world of hurt ... let OpenMP take care of your flushes for you whenever possible ... use seq\_cst

# Atomics and Synchronization Flags (4.0)

```
int main()
{
    double *A, sum, runtime;
    int numthreads, flag = 0, flg_tmp;
    A = (double *)malloc(N*sizeof(double));
#pragma omp parallel sections
{
    #pragma omp section
    {
        fill_rand(N, A);

        #pragma omp atomic write seq_cst
        flag = 1;

    }
    #pragma omp section
    {
        while (1){

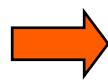
            #pragma omp atomic read seq_cst
            flg_tmp= flag;
            if (flg_tmp==1) break;
        }

        sum = Sum_array(N, A);
    }
}
```

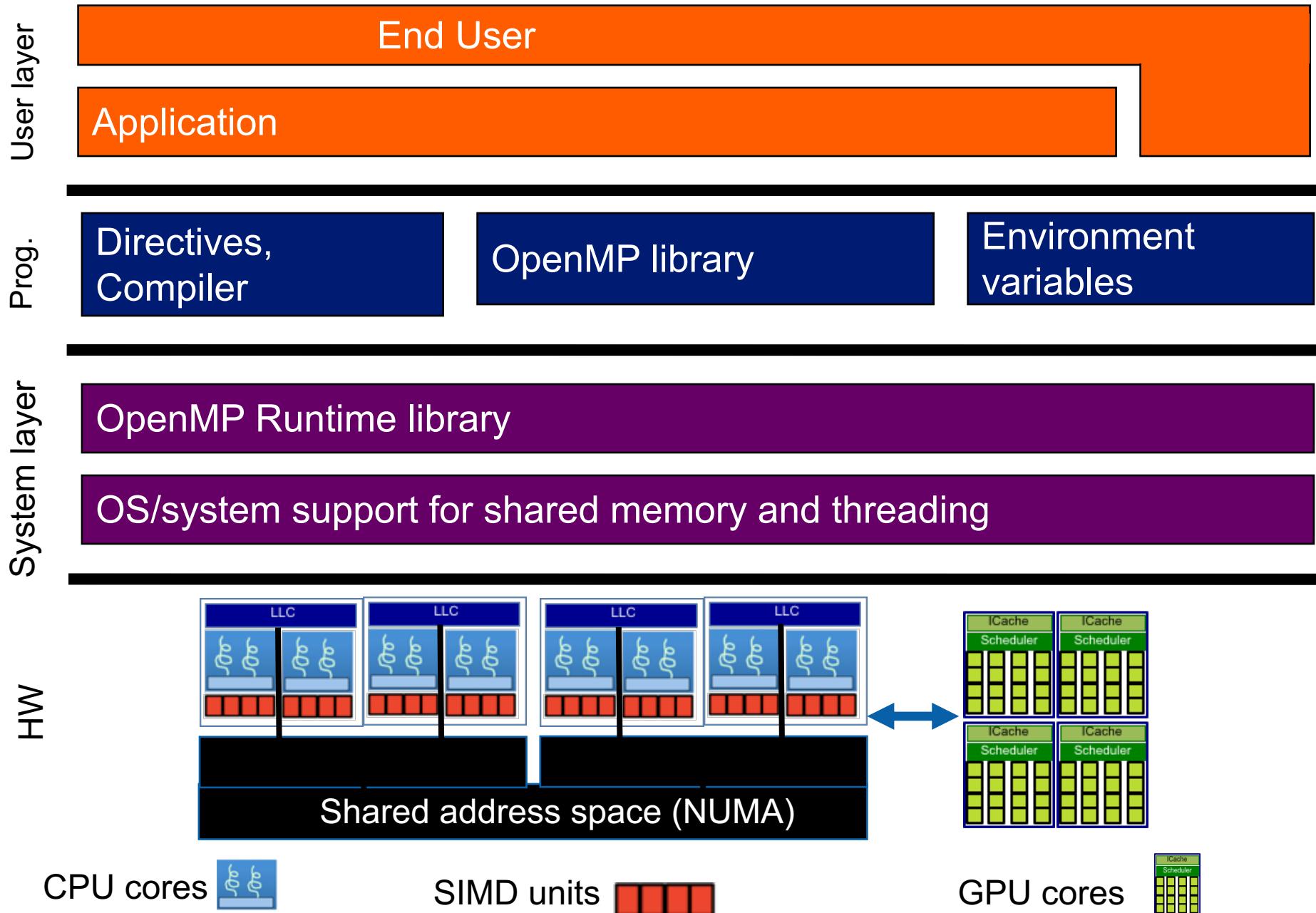
This program is truly race free ... the reads and writes of flag are protected so the two threads cannot conflict – and you do not use any explicit flush constructs (OpenMP does them for you)

# Outline

- Introduction to OpenMP
- Creating Threads
- Synchronization
- Parallel Loops
- Data Environment
- Memory Model
- Irregular Parallelism and Tasks
- Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data



# OpenMP Basic Definitions: Solution stack



# The “BIG idea” of GPU programming

## Traditional loops

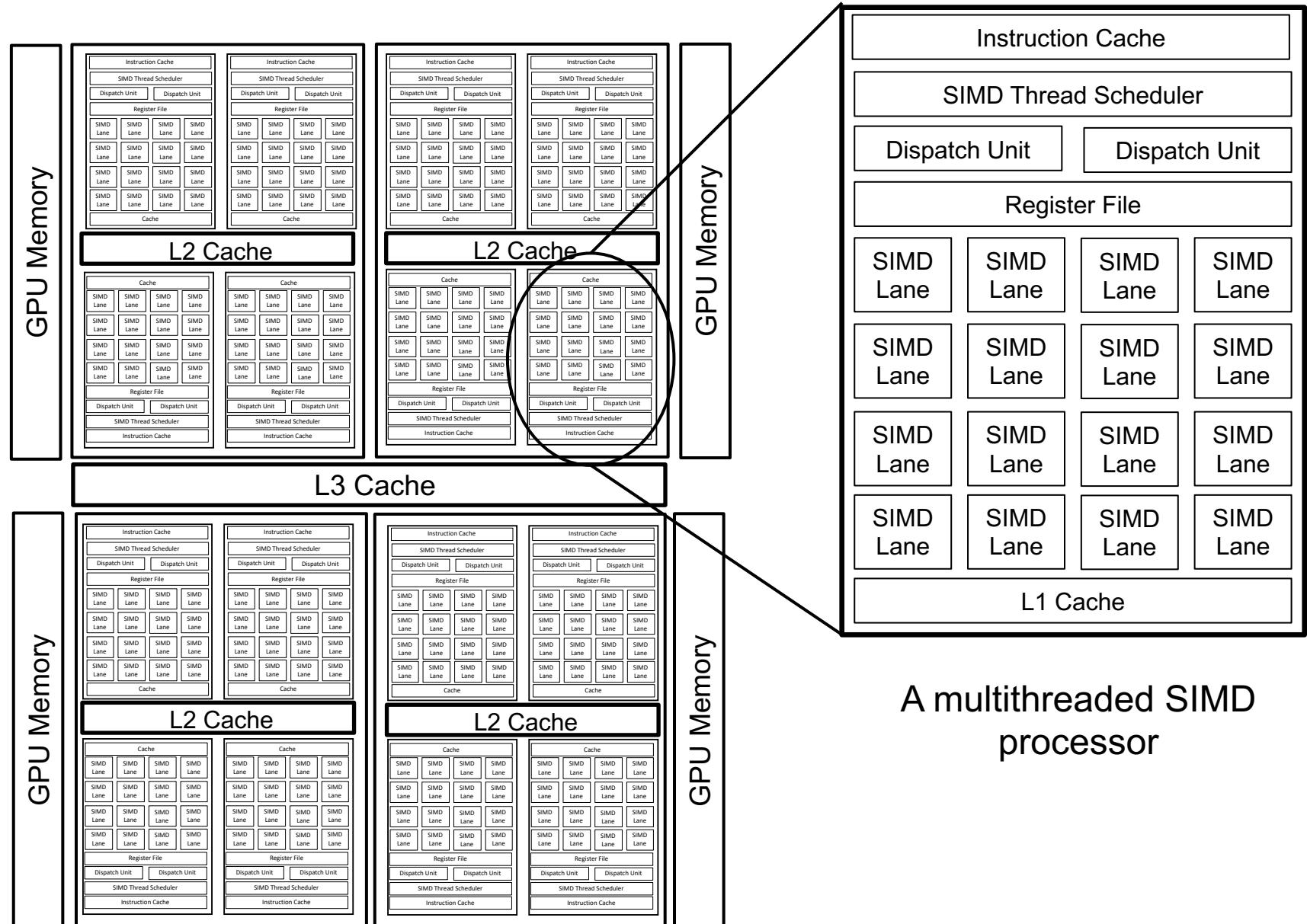
```
void  
trad_mul(int n,  
          const float *a,  
          const float *b,  
          float *c)  
{  
    int i;  
    for (i=0; i<n; i++)  
        c[i] = a[i] * b[i];  
}
```

## Data Parallel OpenCL

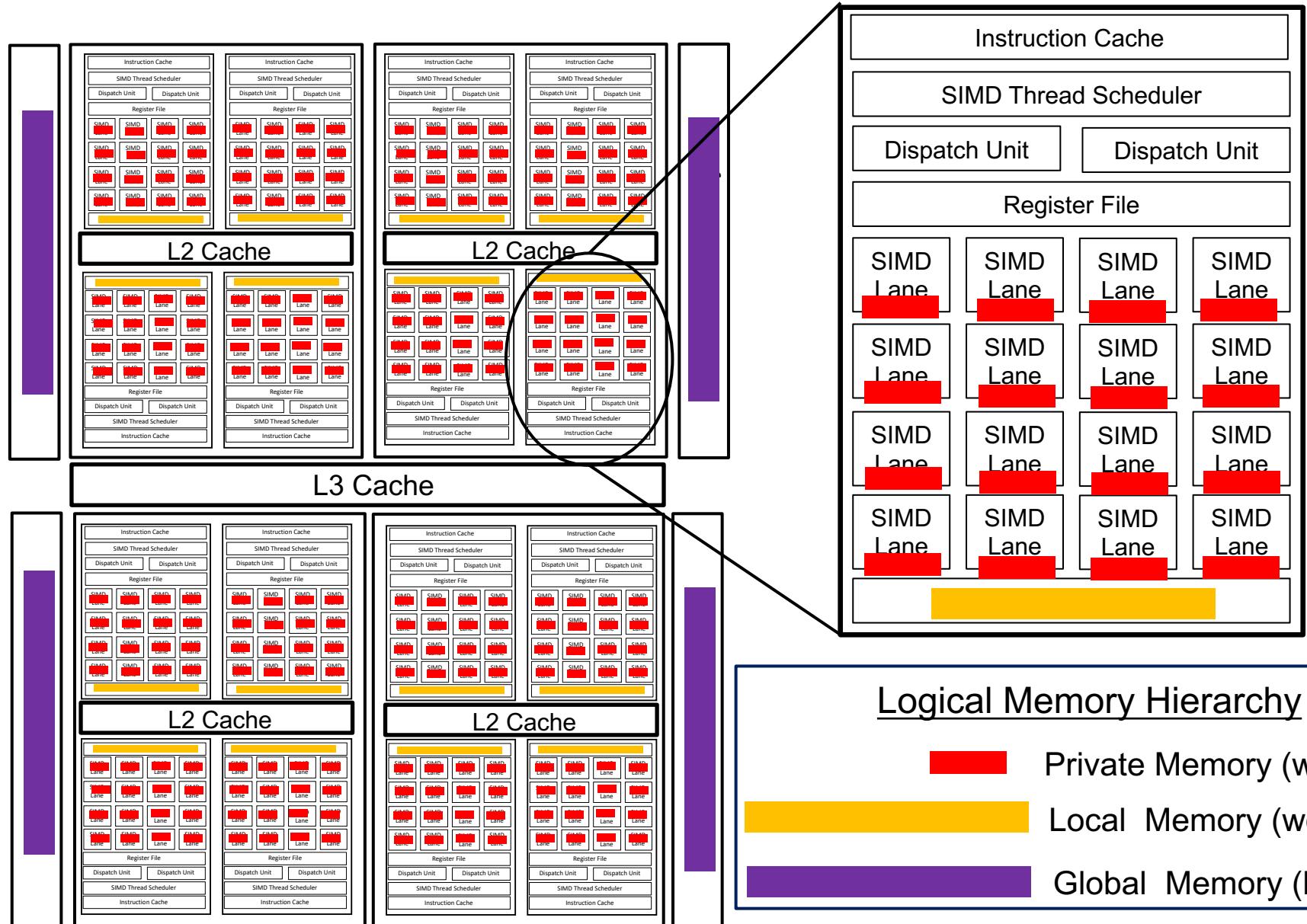
```
kernel void  
dp_mul(global const float *a,  
       global const float *b,  
       global float *c)  
{  
    int id = get_global_id(0);  
  
    c[id] = a[id] * b[id];  
  
} // execute over "n" work-items
```



# A Generic GPU (following Hennessy and Patterson)



# A Generic GPU (following Hennessy and Patterson)



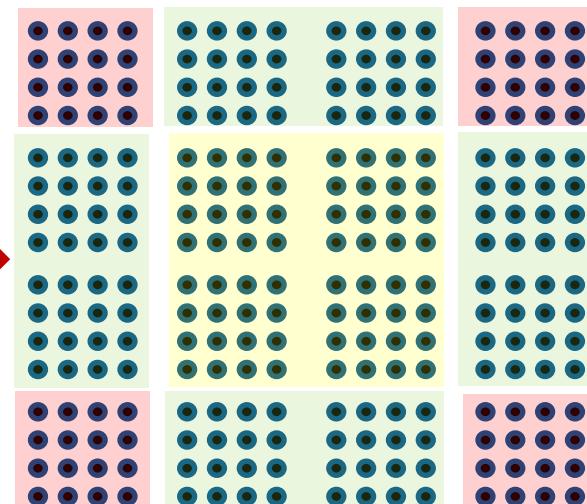
# How do we execute code on a GPU: The SIMT model (Single Instruction Multiple Thread)

1. Turn source code into a scalar work-item

```
extern void reduce( __local float*, __global float*);  
  
__kernel void pi( const int niters, float step_size,  
                 __local float* l_sums, __global float* p_sums)  
{  
    int n_wrk_items = get_local_size(0);  
    int loc_id    = get_local_id(0);  
    int grp_id   = get_group_id(0);  
    float x, accum = 0.0f;  int i,start,iend;  
  
    start = (grp_id * n_wrk_items + loc_id) * niters;  
    iend  = start+niters;  
  
    for(i= start; i<iend; i++){  
        x = (i+0.5f)*step_size;  accum += 4.0f/(1.0f+x*x);}  
  
    l_sums[loc_id] = accum;  
    barrier(CLK_LOCAL_MEM_FENCE);  
    reduce(l_sums, p_sums);  
}
```

This is OpenCL kernel code ... the sort of code the OpenMP compiler generates on your behalf

2. Map work-items onto an N dim index space.



4. Run on hardware designed around the same SIMT execution model



3. Map data structures onto the same index space

# GPU terminology is Broken (sorry about that)

| <b>Hennessy and Patterson</b>    | <b>CUDA</b>              | <b>OpenCL</b>        |
|----------------------------------|--------------------------|----------------------|
| Multithreaded SIMD Processor     | Streaming multiprocessor | Compute Unit         |
| SIMD Thread Scheduler            | Warp Scheduler           | Work-group scheduler |
| SIMD Lane                        | CUDA Core                | Processing Element   |
| GPU Memory                       | Global Memory            | Global Memory        |
| Private Memory                   | Local Memory             | Private Memory       |
| Local Memory                     | Shared Memory            | Local Memory         |
| Vectorizable Loop                | Grid                     | NDRange              |
| Sequence of SIMD Lane operations | CUDA Thread              | work-item            |
| A thread of SIMD instructions    | Warp                     | sub-group            |

# How do we execute code on a GPU: OpenCL and CUDA nomenclature

Turn source code into a scalar  
**work-item** (a CUDA **thread**)

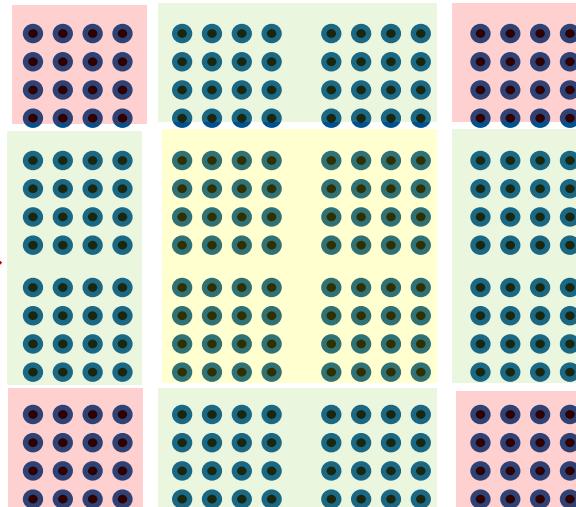
```
extern void reduce( __local float*, __global float*);  
  
__kernel void pi( const int niters, float step_size,  
    __local float* l_sums, __global float* p_sums)  
{  
    int n_wrk_items = get_local_size(0);  
    int loc_id = get_local_id(0);  
    int grp_id = get_group_id(0);  
    float x, accum = 0.0f; int i,start,iend;  
  
    start = (grp_id * n_wrk_items + loc_id) * niters;  
    iend = start+niters;  
  
    for(i= start; i<iend; i++){  
        x = (i+0.5f)*step_size; accum += 4.0f/(1.0f+x*x); }  
  
    l_sums[local_id] = accum;  
    barrier(CLK_LOCAL_MEM_FENCE);  
    reduce(l_sums, p_sums);  
}
```

This code defines a **kernel**

It's called SIMD, but GPUs are really vector-architectures with a block of work-items executing together (a **subgroup** in OpenCL or a **warp** with CUDA)

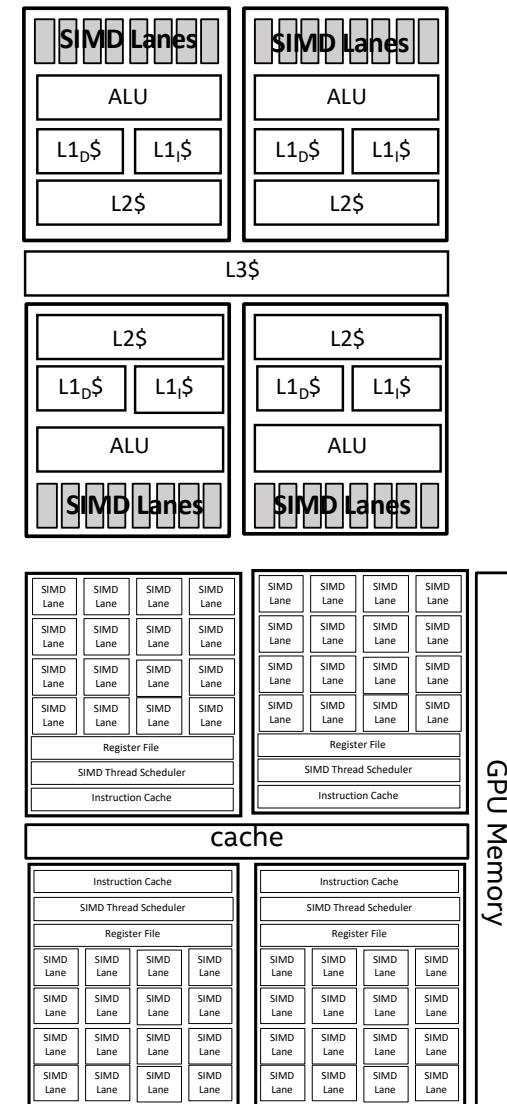
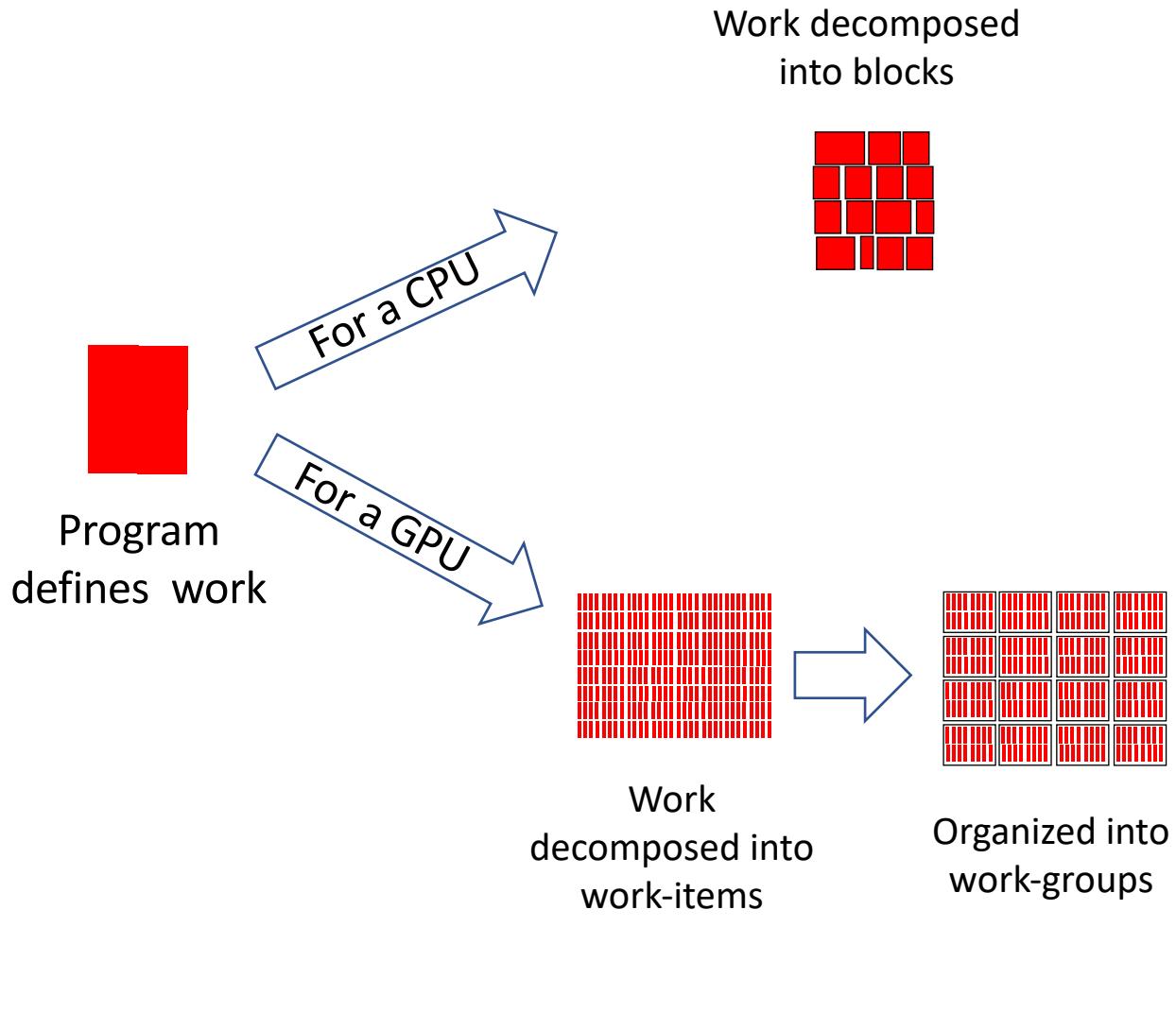
Submit a kernel to an OpenCL **command queue** or a CUDA **stream**

Organize work-items into **work-groups** and map onto an an N dim index space. CUDA calls a work-group a **thread-block**



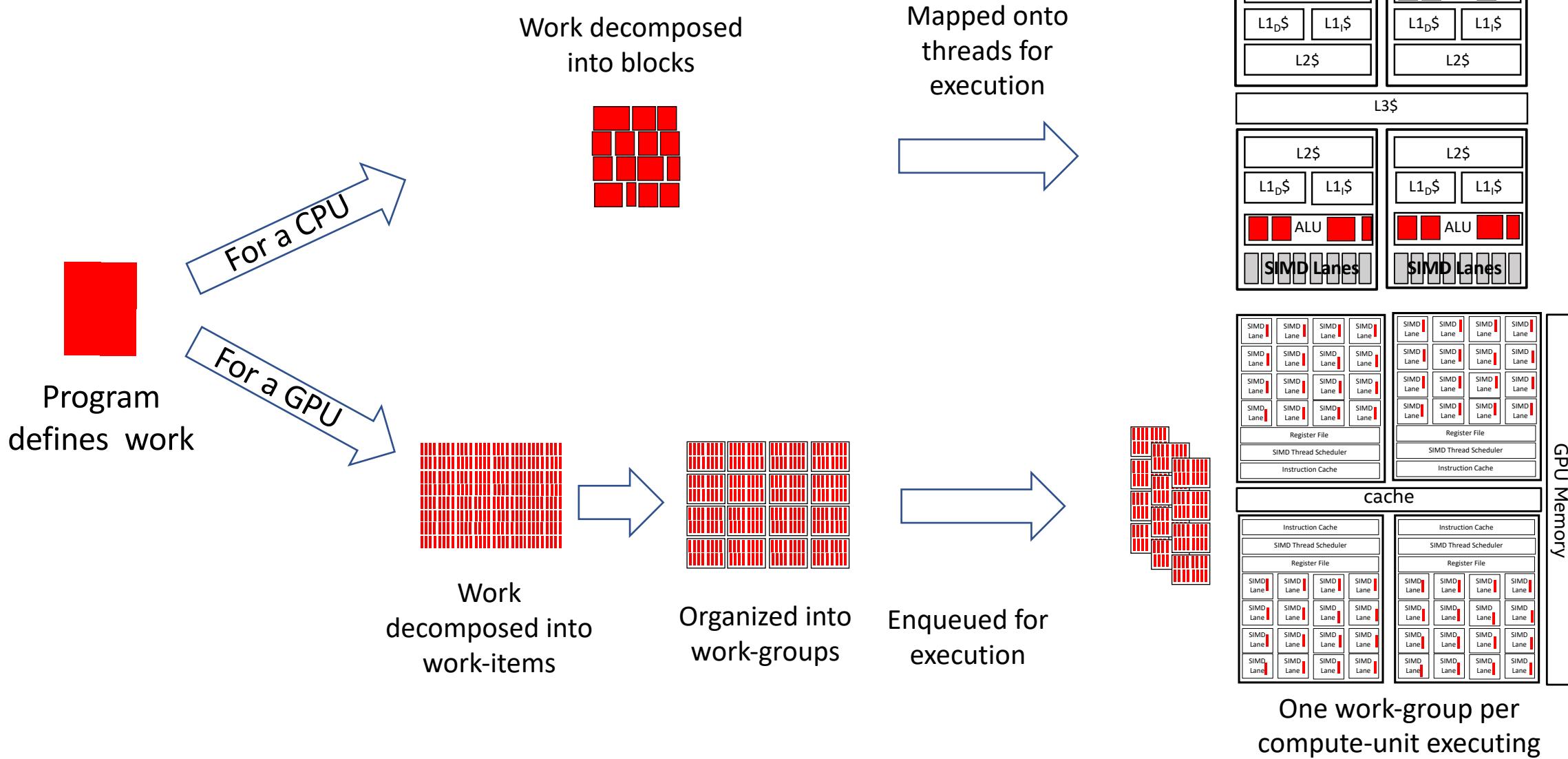
OpenCL index space is called an **NDRange**.  
CUDA calls this a **Grid**

# Executing a program on CPUs and GPUs



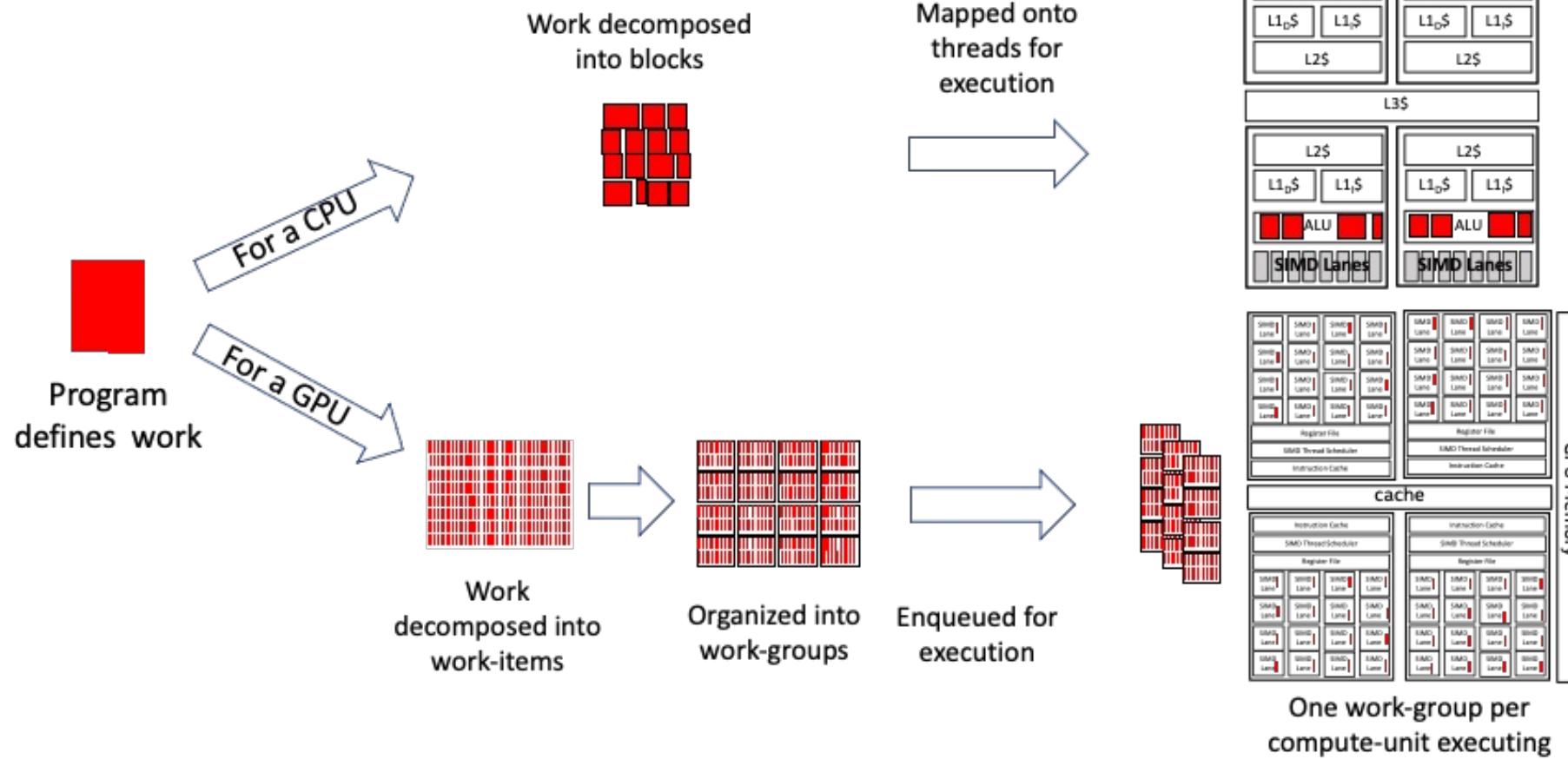
One work-group per  
compute-unit executing

# Executing a program on CPUs and GPUs



# CPU/GPU execution models

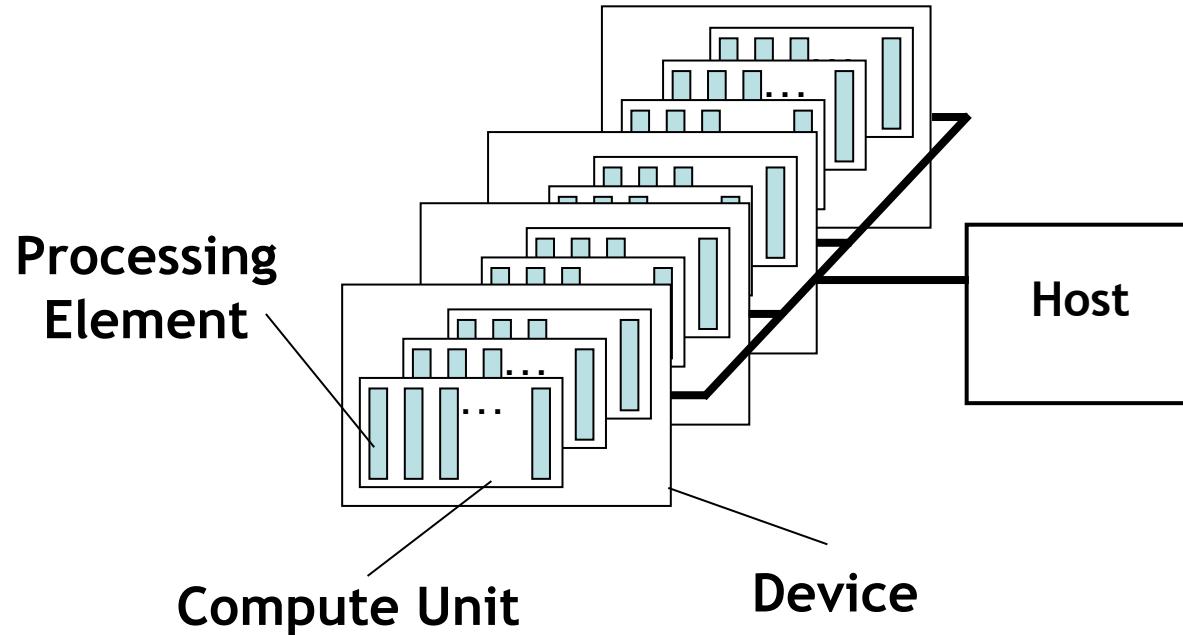
## Executing a program on CPUs and GPUs



For a CPU, the threads are all active and able to make forward progress.

For a GPU, any given work-group might be in the queue waiting to execute.

# A Generic Host/Device Platform Model



- One **Host** and one or more **Devices**
  - Each Device is composed of one or more Compute Units
  - Each Compute Unit is divided into one or more **Processing Elements**
- Memory divided into **host memory** and **device memory**

# The target data environment

- Remember: distinct memory spaces on host and device.
- OpenMP uses a combination of *implicit* and *explicit* memory movement.
- Data may move between the host and the device in well defined places:
  - Firstly, at the beginning and end of a **target** region:

```
#pragma omp target
{ // Data may move here
  ...
} // and here
```

- We'll discuss the other places later...

# Default Data Mapping: implicit movement with a target region

- Scalar variables:
  - Examples:
    - int N; double x;
  - OpenMP implicitly maps scalar variables as **firstprivate**
    - A new value per work-item initialized with the original value (in OpenCL nomenclature, the firstprivate goes in private memory).
  - The variable *is not* copied back to the host at the end of the target region.
  - OpenMP target regions for GPUs execute with CUDA/OpenCL, and a firstprivate scalar can be launched as a parameter to a kernel function without the overhead of setting up a variable in device memory.

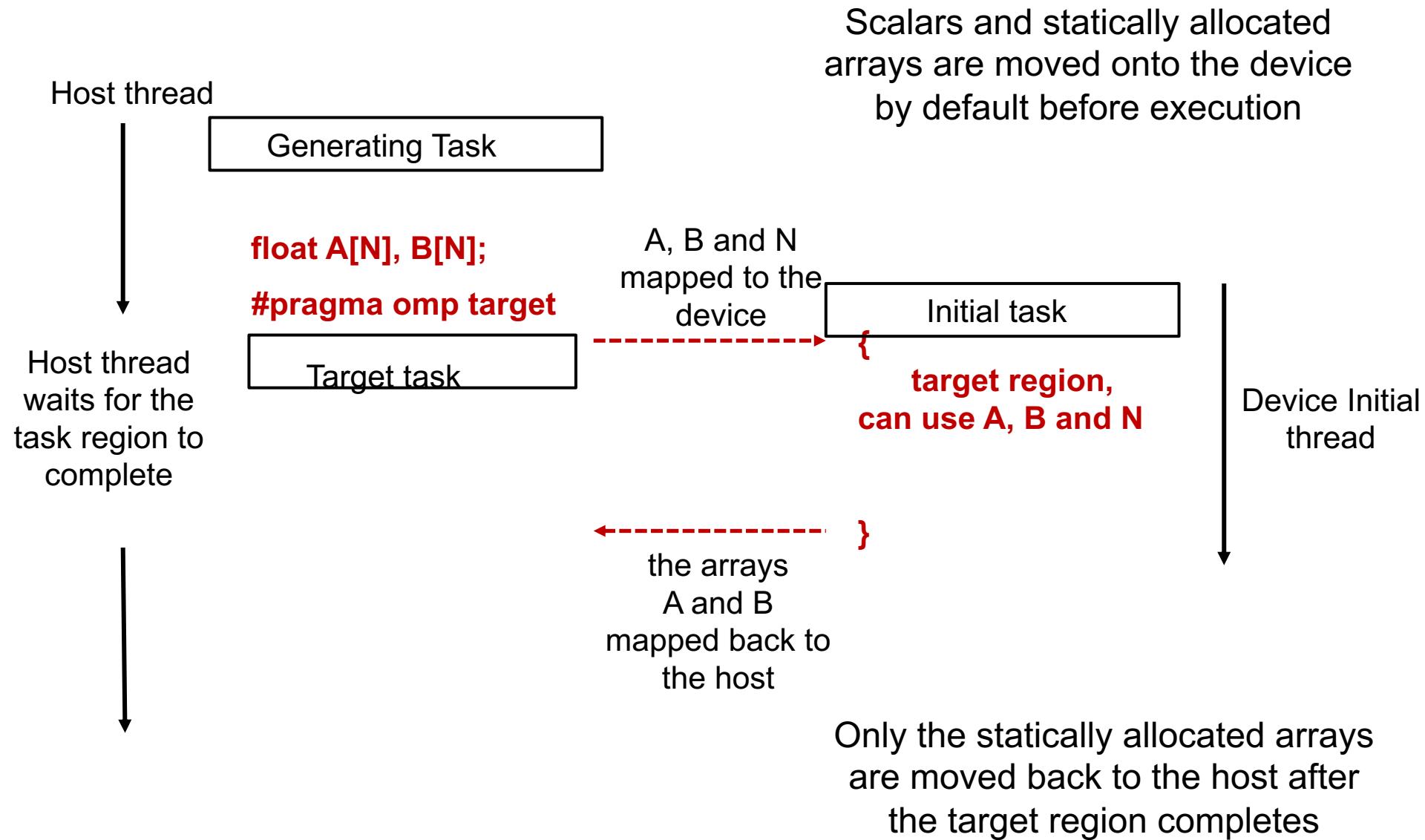
# Default Data Mapping: implicit movement with a target region

- Non-scalar variables:
  - Must have a *complete type*.
  - Example: fixed sized (stack) array:
    - `double A[1000];`
  - Copied to the device at the start of the **target** region, *and* copied back at the end. In OpenCL nomenclature, these are placed in device global memory.
  - A new value is created in the target region and initialized with the original data, but it is shared between threads on the device. Data is copied back to the host at the end of the target region.
  - OpenMP calls this mapping **tofrom**

# Default Data Mapping: implicit movement with a target region

- Pointers and their data:
  - *Example: arrays allocated on the heap*
  - `double *A = malloc(sizeof(double)*1000);`
  - The pointer value will be mapped.
  - But the data it points to ***will not*** be mapped by default.

# The target data environment



# Default Data Sharing: example

```
int main(void) {  
    int N = 1024;  
    double A[N], B[N];  
  
    #pragma omp target  
    {  
  
        for (int ii = 0; ii < N; ++ii) {  
  
            A[ii] = A[ii] + B[ii];  
  
        }  
    } // end of target region  
}
```

1. Variables created in host memory.

2. Scalar **N** and stack arrays **A** and **B** are copied *to* device memory. Execution transferred to device.

3. **ii** is **private** on the device as it's declared within the target region

4. Execution on the device.

5. stack arrays **A** and **B** are copied *from* device memory back to the host. Host resumes execution.

# So lets run code in parallel on the device

```
int main(void) {  
    int N = 1024;  
    double A[N], B[N];  
  
    #pragma omp target  
    {  
        #pragma omp loop  
        for (int ii = 0; ii < N; ++ii) {  
            A[ii] = A[ii] + B[ii];  
  
        }  
    } // end of target region  
}
```

The loop construct tells the compiler:  
*"this loop will execute correctly if the loop iterations run in any order. You can safely run them concurrently. And the loop-body doesn't contain any OpenMP constructs. So do whatever you can to make the code run fast"*

The loop construct is a declarative construct. You tell the compiler what you want done but you DO NOT tell it how to "do it". This is new for OpenMP

# Explicit Data Sharing

- Previously, we described the rules for *implicit* data movement.
- We *explicitly* control the movement of data using the **map** clause.
- Data allocated on the heap needs to explicitly copied to/from the device:

```
int main(void) {
    int ii=0, N = 1024;
    int* A = malloc(sizeof(int)*N);

#pragma omp target
{
    // N, ii and A all exist here
    // The data that A points to (*A , A[ii]) DOES NOT exist here!
}
```

# Controlling data movement

```
int i, a[N], b[N], c[N];  
#pragma omp target map(to:a,b) map(tofrom:c)
```

Data movement defined from the *host* perspective.

- The various forms of the map clause
  - **map(to:list)**: On entering the region, variables in the list are initialized on the device using the original values from the host (host to device copy).
  - **map(from:list)**: At the end of the target region, the values from variables in the list are copied into the original variables (device to host copy). On entering the region, initial value of the variable is not initialized.
  - **map(tofrom:list)**: the effect of both a map-to and a map-from (host to device copy at start of region, device to host copy at end)
  - **map(alloc:list)**: On entering the region, data is allocated and uninitialized on the device.
  - **map(list)**: equivalent to **map(tofrom:list)**.
- For pointers you must use array section notation ..
  - **map(to:a[0:N])**. Notation is **A[lower-bound : length]**

# Moving arrays with the map clause

```
int main(void) {  
    int N = 1024;  
    int* A = malloc(sizeof(int)*N);  
  
    #pragma omp target map(A[0:N])  
    {  
        // N, ii and A all exist here  
        // The data that A points to DOES exist here!  
    }  
}
```

Default mapping  
**map(tofrom: A[0:N])**

Copy at start and end of  
**target** region.

# Our running example: Jacobi solver

- An iterative method to solve a system of linear equations
  - Given a matrix A and a vector b find the vector x such that  $Ax=b$
- The basic algorithm:
  - Write A as a lower triangular (L), upper triangular (U) and diagonal matrix
$$Ax = (L+D+U)x = b$$
  - Carry out multiplications and rearrange
$$Dx = b - (L+U)x \rightarrow x = (b - (L+U)x)/D$$
  - Iteratively compute a new x using the x from the previous iteration
$$X_{\text{new}} = (b - (L+U)x_{\text{old}})/D$$
- Advantage: we can easily test if the answer is correct by multiplying our final x by A and comparing to b
- Disadvantage: It takes many iterations and only works for diagonally dominant matrices

# Jacobi Solver

Iteratively update xnew until the value stabilizes (i.e. change less than a preset TOL)

```
<<< allocate and initialize the matrix A >>>
<<< and vectors x1, x2 and b      >>>

while((conv > TOL) && (iters<MAX_ITERS))
{
    iters++;
    for (i=0; i<Ndim; i++){
        xnew[i] = (TYPE) 0.0;
        for (j=0; j<Ndim;j++){
            if(i!=j)
                xnew[i]+= A[i*Ndim + j]*xold[j];
        }
        xnew[i] = (b[i]-xnew[i])/A[i*Ndim+i];
    }

    // test convergence
    conv = 0.0;
    for (i=0; i<Ndim; i++){
        tmp  = xnew[i]-xold[i];
        conv += tmp*tmp;
    }
    conv = sqrt((double)conv);

    // swap pointers for next
    // iteration
    TYPE* tmp = xold;
    xold = xnew;
    xnew = tmp;

} // end while loop
```

# Jacobi Solver (Par Targ, 1/2)

```
while((conv > TOL) && (iters<MAX_ITERS))
{
    iters++;
#pragma omp target map(tofrom:xnew[0:Ndim],xold[0:Ndim]) \
map(to:A[0:Ndim*Ndim], b[0:Ndim])
#pragma omp loop
for (i=0; i<Ndim; i++){
    xnew[i] = (TYPE) 0.0;
    for (j=0; j<Ndim;j++){
        if(i!=j)
            xnew[i]+= A[i*Ndim + j]*xold[j];
    }
    xnew[i] = (b[i]-xnew[i])/A[i*Ndim+i];
}
```

# Jacobi Solver (Par Targ, 2/2)

```
//  
// test convergence  
//  
conv = 0.0;  
#pragma omp target map(to:xnew[0:Ndim],xold[0:Ndim]) \  
          map(tofrom:conv)  
#pragma omp loop private(i,tmp) reduction(+:conv)  
for (i=0; i<Ndim; i++){  
    tmp = xnew[i]-xold[i];  
    conv += tmp*tmp;  
}  
conv = sqrt((double)conv);  
TYPE* tmp = xold;  
xold = xnew;  
xnew = tmp;  
} // end while loop
```

This worked but the performance was awful. Why?

| System            | Implementation      | Ndim = 4096 |
|-------------------|---------------------|-------------|
| NVIDIA® K20X™ GPU | Target dir per loop | 131.94 secs |

Cray® XC40™ Supercomputer running Cray® Compiling Environment 8.5.3.  
Intel® Xeon ® CPU E5-2697 v2 @ 2.70GHz with 32 GB DDR3. NVIDIA® Tesla® K20X, 6GB.

# Data movement dominates!!!

```
while((conv > TOLERANCE) && (iters<MAX_ITERS))  
{ iters++;  
xnew = iters % s ? x2 : x1;  
xold = iters % s ? x1 : x2;
```

Typically over 4000 iterations!

```
#pragma omp target map(tofrom:xnew[0:Ndim],xold[0:Ndim]) \  
map(to:A[0:Ndim*Ndim], b[0:Ndim] )
```

```
#pragma omp loop private(i,j)  
for (i=0; i<Ndim; i++){  
    xnew[i] = (TYPE) 0.0;  
    for (j=0; j<Ndim;j++){  
        if(i!=j)  
            xnew[i]+= A[i*Ndim + j]*xold[j];  
    }  
    xnew[i] = (b[i]-xnew[i])/A[i*Ndim+i];  
}
```

```
// test convergence  
conv = 0.0;
```

```
#pragma omp target map(to:xnew[0:Ndim],xold[0:Ndim]) \  
map(tofrom:conv)
```

```
#pragma loop reduction(+:conv)  
for (i=0; i<Ndim; i++){  
    tmp = xnew[i]-xold[i];  
    conv += tmp*tmp;  
}
```

```
conv = sqrt((double)conv);  
}
```

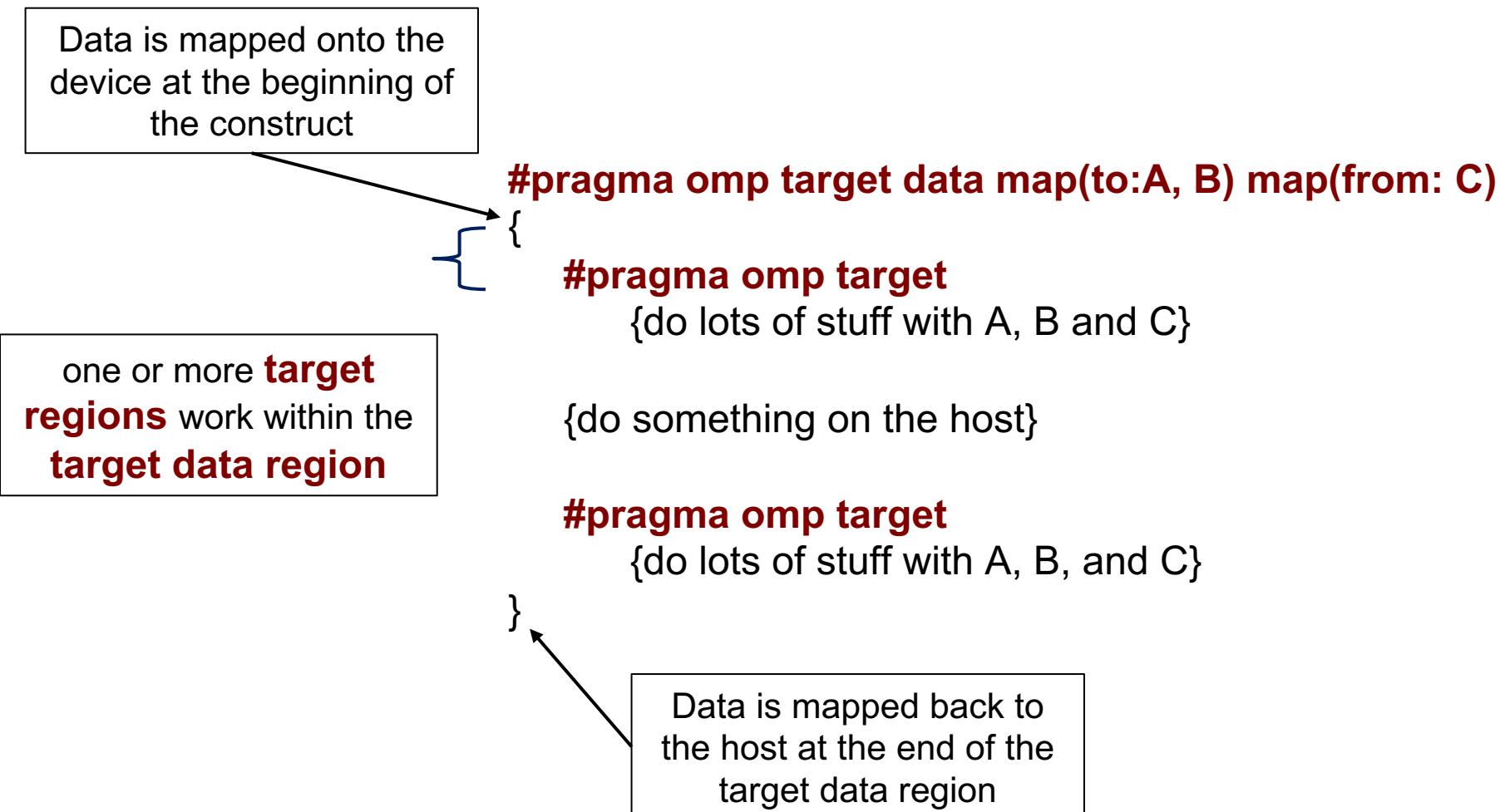
For each iteration, **copy to device**  
 $(3*Ndim+Ndim^2)*\text{sizeof}(\text{TYPE})$  bytes

For each iteration, **copy from device**  
 $2*Ndim*\text{sizeof}(\text{TYPE})$  bytes

For each iteration, **copy to device**  
 $2*Ndim*\text{sizeof}(\text{TYPE})$  bytes

# Target data directive

- The **target data** construct creates a target data region
  - ... use **map** clauses for explicit data management



# Jacobi Solver (Par Target Data, 1/2)

```
#pragma omp target data map(tofrom:x1[0:Ndim],x2[0:Ndim]) \
    map(to:A[0:Ndim*Ndim], b[0:Ndim] ,Ndim)
while((conv > TOL) && (iters<MAX_ITERS))
{ iters++;
  #pragma omp target
  #pragma omp loop private(j) firstprivate(xnew,xold)
  for (i=0; i<Ndim; i++){
    xnew[i] = (TYPE) 0.0;
    for (j=0; j<Ndim;j++){
      if(i!=j)
        xnew[i]+= A[i*Ndim + j]*xold[j];
    }
    xnew[i] = (b[i]-xnew[i])/A[i*Ndim+i];
  }
}
```

# Jacobi Solver (Par Target Data, 2/2)

```
// test convergence
conv = 0.0;
#pragma omp target map(tofrom: conv)
#pragma omp loop private(tmp) firstprivate(xnew,xold) reduction(+:conv)

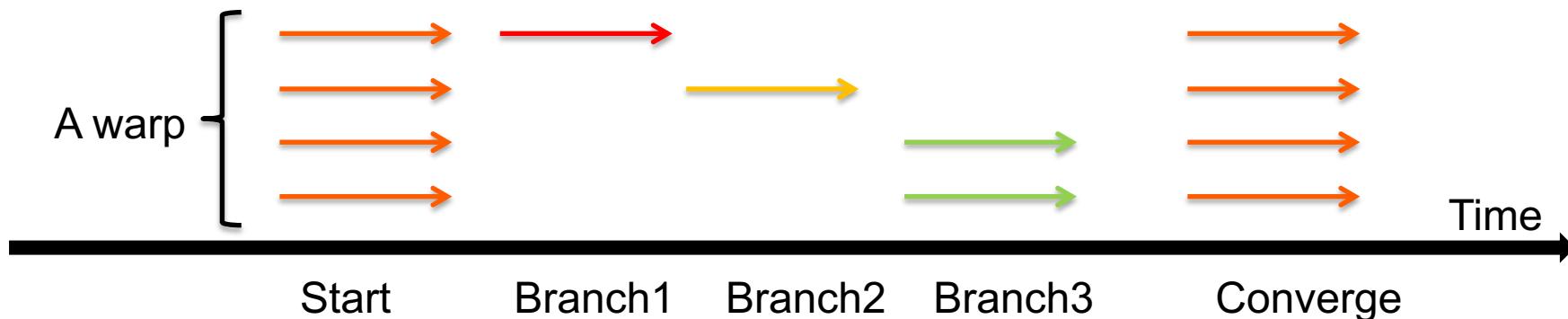
for (i=0; i<Ndim; i++){
    tmp = xnew[i]-xold[i];
    conv += tmp*tmp;
}
// end target region
conv = sqrt((double)conv);

TYPE* tmp = xold;
xold = xnew;
xnew = tmp;
} // end while loop
```

| System            | Implementation                | Ndim = 4096 |
|-------------------|-------------------------------|-------------|
| NVIDIA® K20X™ GPU | Target dir per loop           | 131.94 secs |
|                   | Above plus target data region | 18.37 secs  |

# Single Instruction Multiple Data

- Individual work-items of a warp start together at the same program address
- Each work-item has its own instruction address counter and register state
  - Each work-item is free to branch and execute independently
  - Supports the SPMD pattern.
- Branch behavior
  - Each branch will be executed serially
  - Work-items not following the current branch will be disabled



# Branching

## Conditional execution

```
// Only evaluate expression  
// if condition is met  
if (a > b)  
{  
    acc += (a - b*c);  
}
```

## Selection and masking

```
// Always evaluate expression  
// and mask result  
temp = (a - b*c);  
mask = (a > b ? 1.f : 0.f);  
acc += (mask * temp);
```

# Coalescence

- Coalesce - to combine into one
- Coalesced memory accesses are key for high bandwidth
- Simply, it means, if thread  $i$  accesses memory location  $n$  then thread  $i+1$  accesses memory location  $n+1$
- In practice, it's not quite as strict...

```
for (int id = 0; id < size; id++)
{
    // ideal
    float val1 = memA[id];

    // still pretty good
    const int c = 3;
    float val2 = memA[id + c];

    // stride size is not so good
    float val3 = memA[c*id];

    // terrible
    const int loc =
        some_strange_func(id);

    float val4 = memA[loc];
}
```

# Jacobi Solver (Target Data/branchless/coalesced mem, 1/2)

```
#pragma omp target data map(tofrom:x1[0:Ndim],x2[0:Ndim]) \
    map(to:A[0:Ndim*Ndim], b[0:Ndim] ,Ndim)
while((conv > TOL) && (iters<MAX_ITERS))
{ iters++;
#pragma omp target
    #pragma omp loop private(j)
    for (i=0; i<Ndim; i++){
        xnew[i] = (TYPE) 0.0;
        for (j=0; j<Ndim;j++){
            xnew[i]+= (A[j*Ndim + i]*xold[j])*((TYPE)(i != j));
        }
        xnew[i] = (b[i]-xnew[i])/A[i*Ndim+i];
    }
}
```

We replaced the original code with a poor memory access pattern  
 $xnew[i]+= (A[i*Ndim + j]*xold[j])$   
With the more efficient  
 $xnew[i]+= (A[j*Ndim + i]*xold[j])$

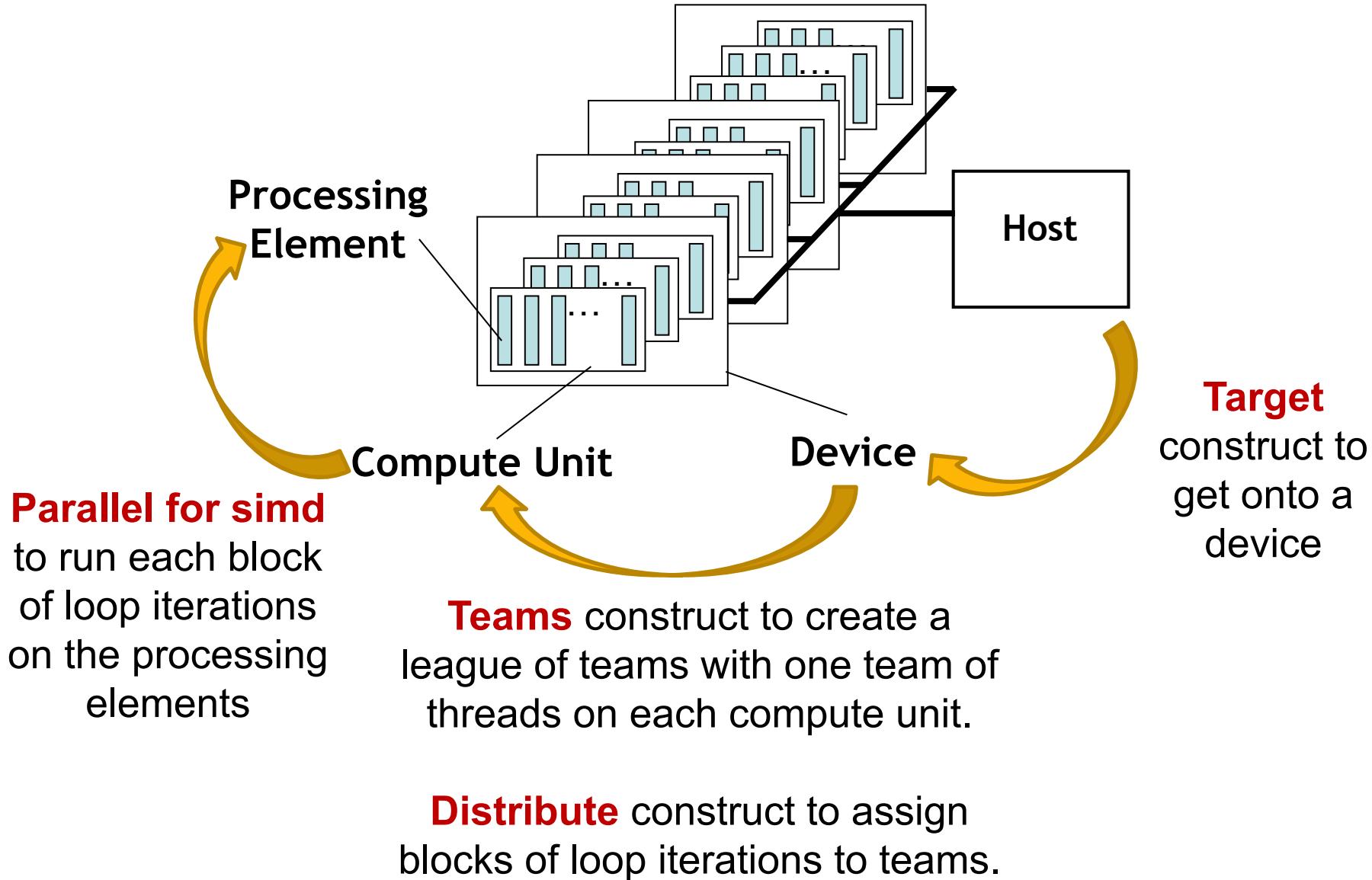
# Jacobi Solver (Target Data/branchless/coalesced mem, 2/2)

```
//  
// test convergence  
conv = 0.0;  
#pragma omp target map(tofrom: conv)  
#pragma omp loop private(tmp) reduction(+:conv)  
for (i=0; i<Ndim; i++){  
    tmp = xnew[i]-xold[i];  
    conv += tmp*tmp;  
}  
conv = sqrt((double)conv);  
TYPE* tmp = xold;  
xold = xnew;  
xnew = tmp;  
} // end while loop
```

| System            | Implementation                 | Ndim = 4096 |
|-------------------|--------------------------------|-------------|
| NVIDIA® K20X™ GPU | Target dir per loop            | 131.94 secs |
|                   | Above plus target data region  | 18.37 secs  |
|                   | Above plus reduced branching   | 13.74 secs  |
|                   | Above plus improved mem access | 7.64 secs   |

**The loop construct is great, but sometimes you want more control.**

# Our host/device Platform Model and OpenMP



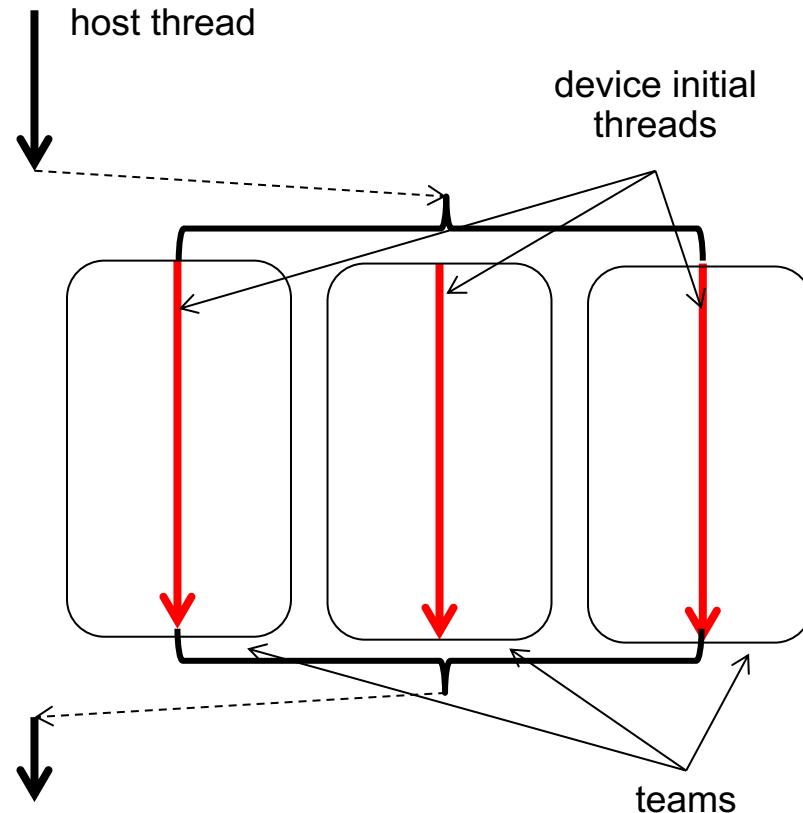
# teams and distribute constructs

- The **teams** construct
  - Similar to the **parallel** construct
  - It starts a league of thread teams
  - Each team in the league starts as one initial thread – a team of one
  - Threads in different teams cannot synchronize with each other
  - The construct must be “perfectly” nested in a **target** construct
- The **distribute** construct
  - Similar to the **for** construct
  - Loop iterations are workshared across the initial threads in a league
  - No implicit barrier at the end of the construct
  - **dist\_schedule(*kind*[, *chunk\_size*])**
    - If specified, scheduling kind must be static
    - Chunks are distributed in round-robin fashion in chunks of size ***chunk\_size***
    - If no chunk size specified, chunks are of (almost) equal size; each team receives at least one chunk

# Accelerated workshare v2.0

- teams construct
- distribute construct

```
#pragma omp target
#pragma omp teams
#pragma omp distribute
for (i=0;i<N;i++)
...
...
```

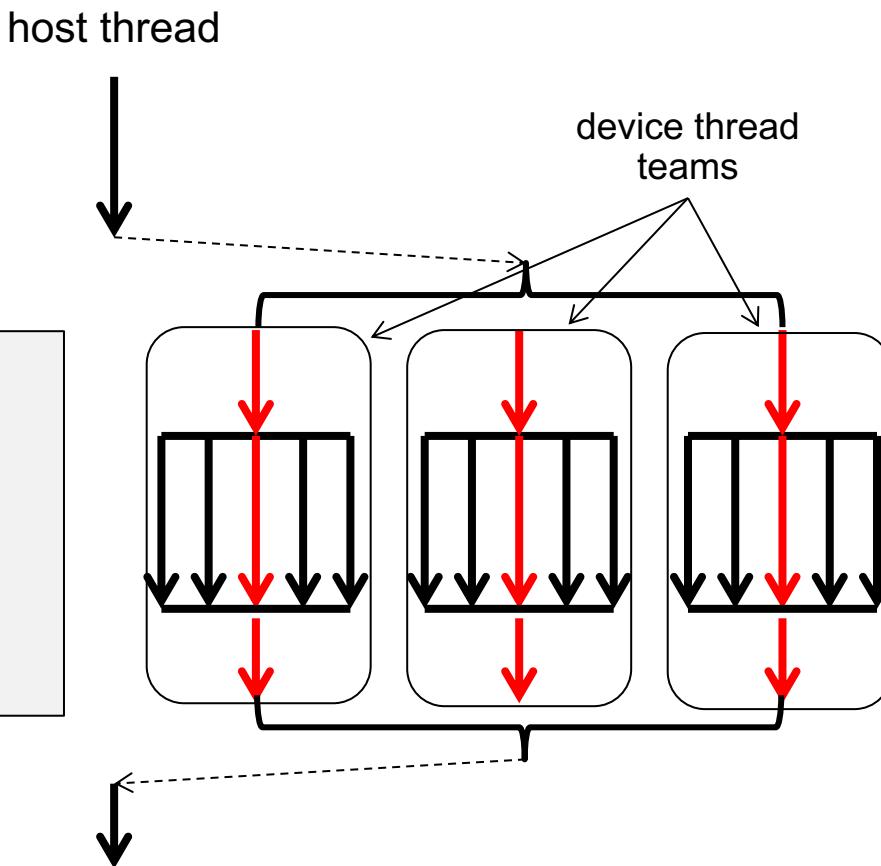


- Transfer execution control to **MULTIPLE** device initial threads
- Workshare loop iterations across the initial threads.

# Accelerate workshare v3.0

- teams distribute
- parallel for simd

```
#pragma omp target
#pragma omp teams
#pragma omp distribute
#pragma omp parallel for simd
for (i=0;i<N;i++)
...
```



- Transfer execution control to **MULTIPLE** device initial threads
  - Workshare loop iterations across the initial threads (teams distribute)
- Each initial thread becomes the primary\* thread in a thread team
  - Workshare loop iterations across the threads in a team (parallel for)

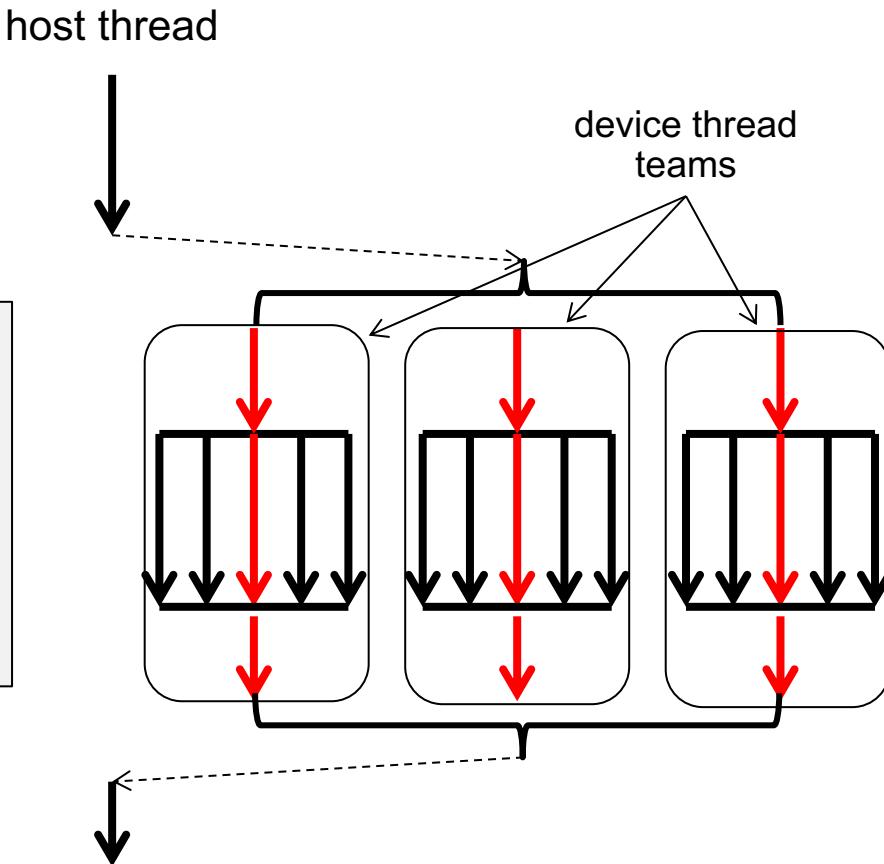
\*the term “master” has been deprecated in OpenMP 5.1 and replaced with the term “primary”.

# Accelerate workshare v3.0

- teams distribute
- parallel for simd

Works with  
nested loops  
as well

```
#pragma omp target
#pragma omp teams distribute
for (i=0;i<N;i++)
#pragma omp parallel for simd
for (j=0;j<M;j++)
...
...
```



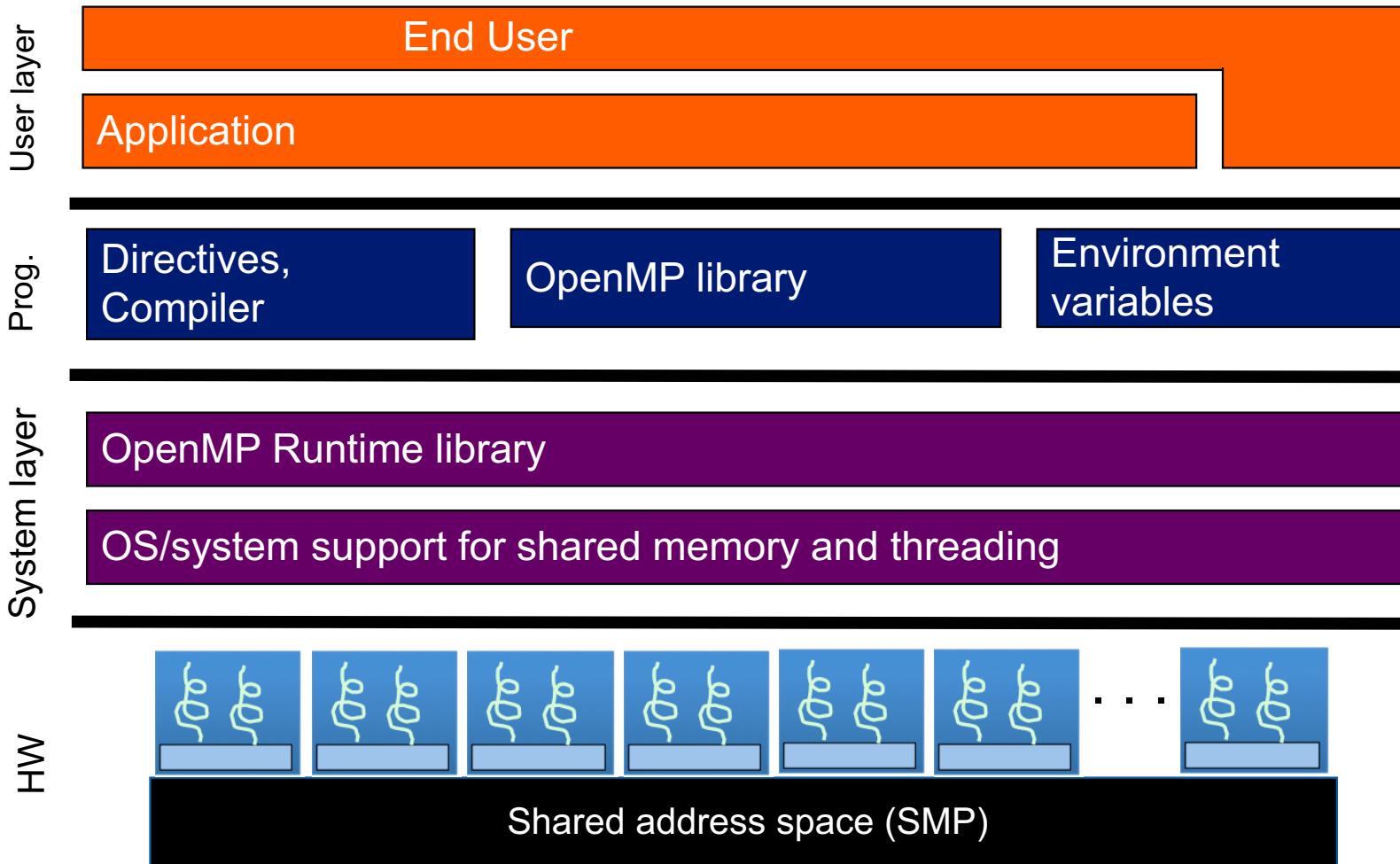
- Transfer execution control to **MULTIPLE** device initial threads
  - Workshare loop iterations across the initial threads (teams distribute)
- Each initial thread becomes the primary\* thread in a thread team
  - Workshare loop iterations across the threads in a team (parallel for)

\*the term “master” has been deprecated in OpenMP 5.1 and replaced with the term “primary”.

- Introduction to OpenMP
- Creating Threads
- Synchronization
- Parallel Loops
- Data Environment
- Memory Model
- Irregular Parallelism and Tasks
- Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data



# OpenMP basic definitions: Basic Solution stack



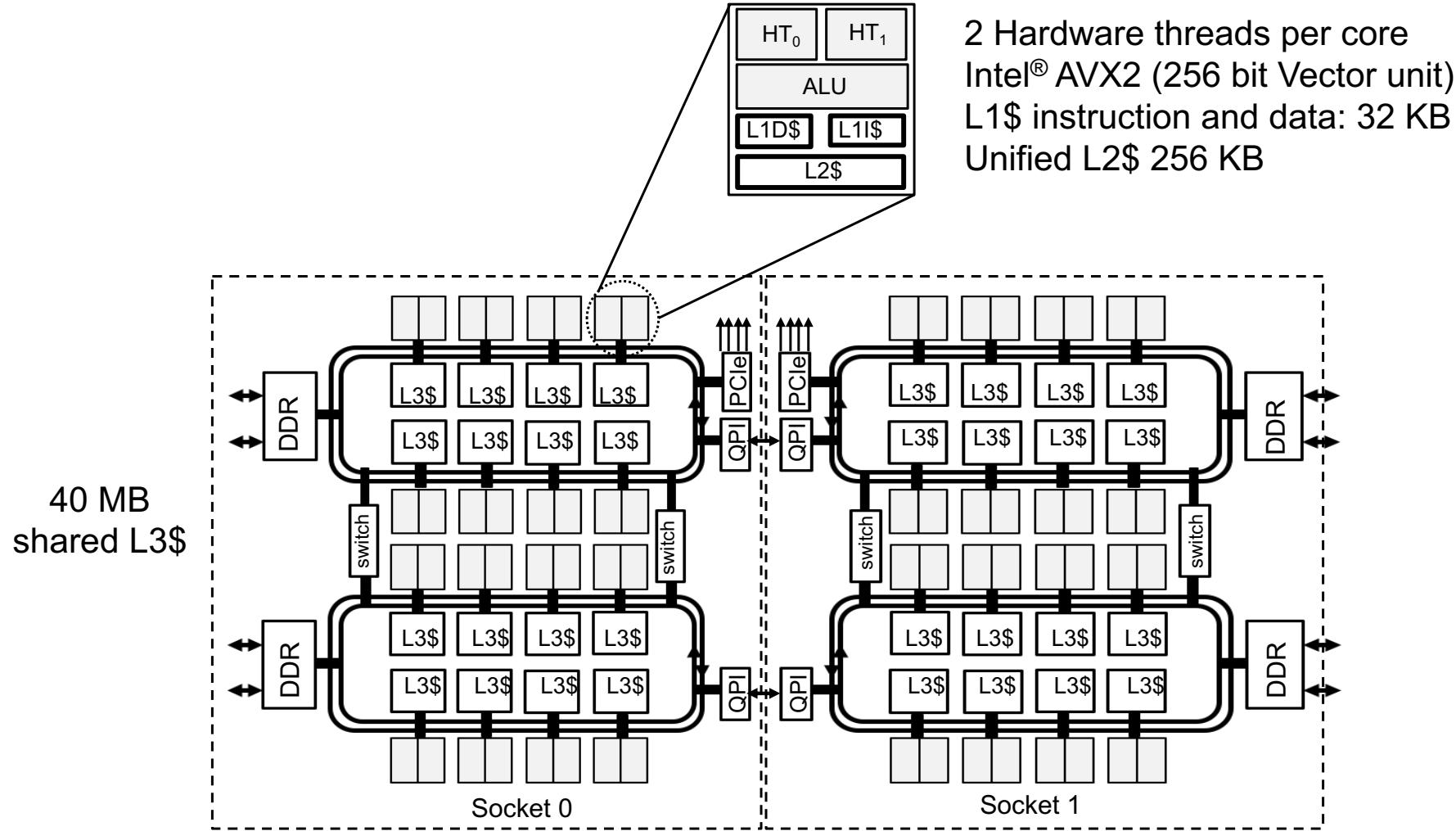
In learning OpenMP, you consider a Symmetric Multiprocessor (SMP) ....  
i.e. lots of threads with "**equal cost access**" to memory

# CPU Architecture Trend

- Multi-socket nodes with rapidly increasing core counts
  - Memory per core decreases
  - Memory bandwidth per core decreases
  - Network bandwidth per core decreases
- Applications often use a hybrid programming model with three levels of parallelism
  - MPI between nodes or sockets
  - Shared memory (such as OpenMP) on the nodes/sockets
  - Increase vectorization for lower level loop structures

# A Typical CPU Node in an HPC System

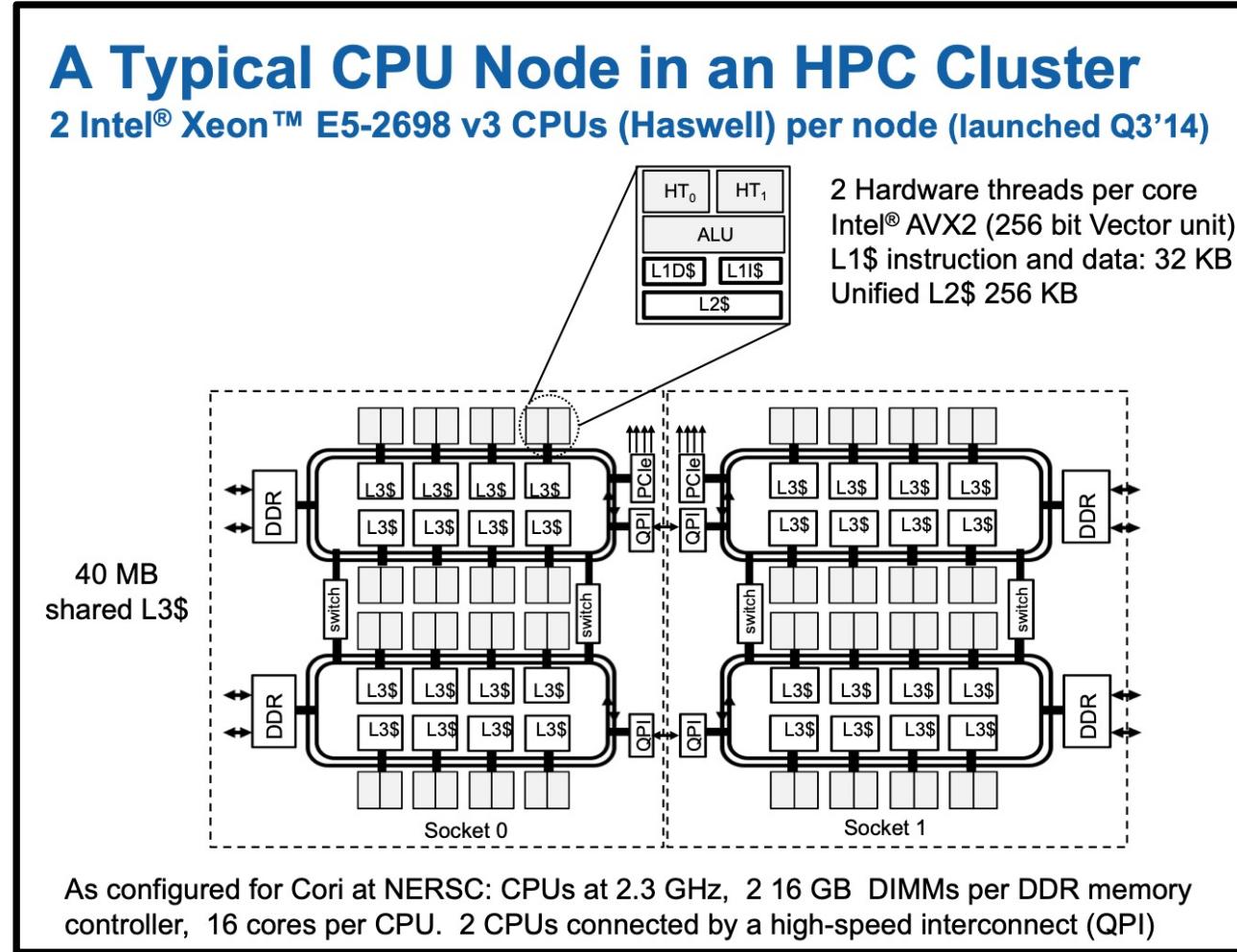
2 Intel® Xeon™ E5-2698 v3 CPUs (Haswell) per node (launched Q3'14)



As configured for Cori at NERSC: CPUs at 2.3 GHz, 2 16 GB DIMMs per DDR memory controller, 16 cores per CPU. 2 CPUs connected by a high-speed interconnect (QPI)

# Does this look like an SMP node to you?

There may be a single address space, but there are multiple levels of non-uniformity to the memory. This is a **Non-Uniform Memory Architecture** (NUMA)



Even a single CPU is properly considered a NUMA architecture

# NUMA Systems

- Most systems today are Non-Uniform Memory Access (NUMA)
- Accessing memory in remote NUMA is slower than accessing memory in local NUMA
- Accessing High Bandwidth Memory is faster than DDR

*A Generic Contemporary NUMA System*

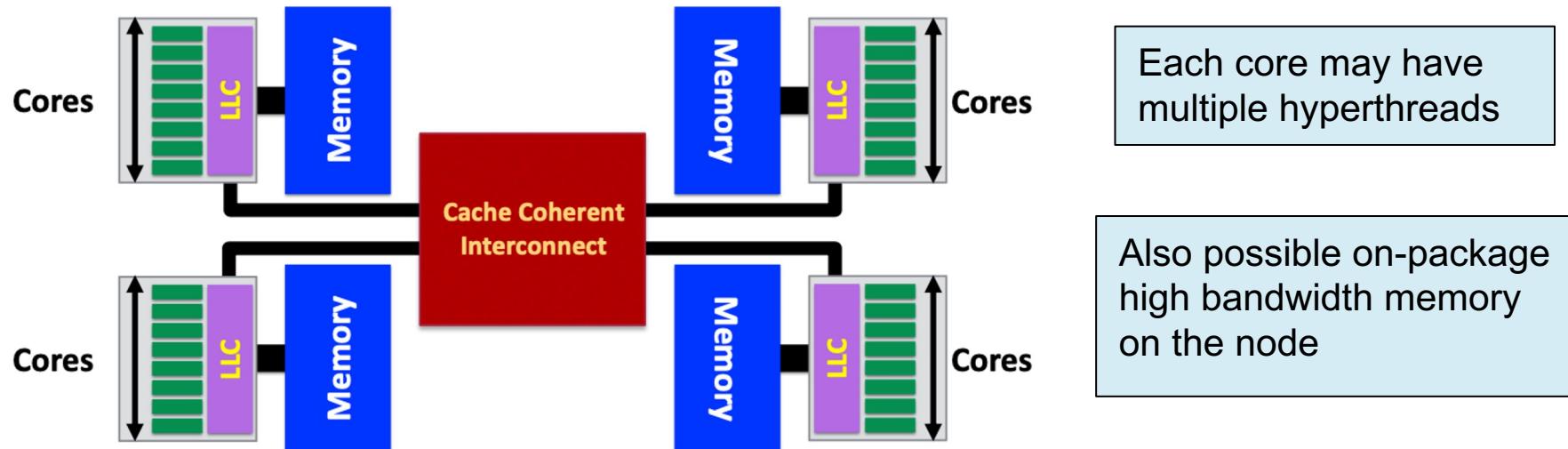


Diagram courtesy Ruud van der Pas

# Memory Locality

- Most systems today are Non-Uniform Memory Access (NUMA)
- Example, the Intel® Xeon Phi™ processor

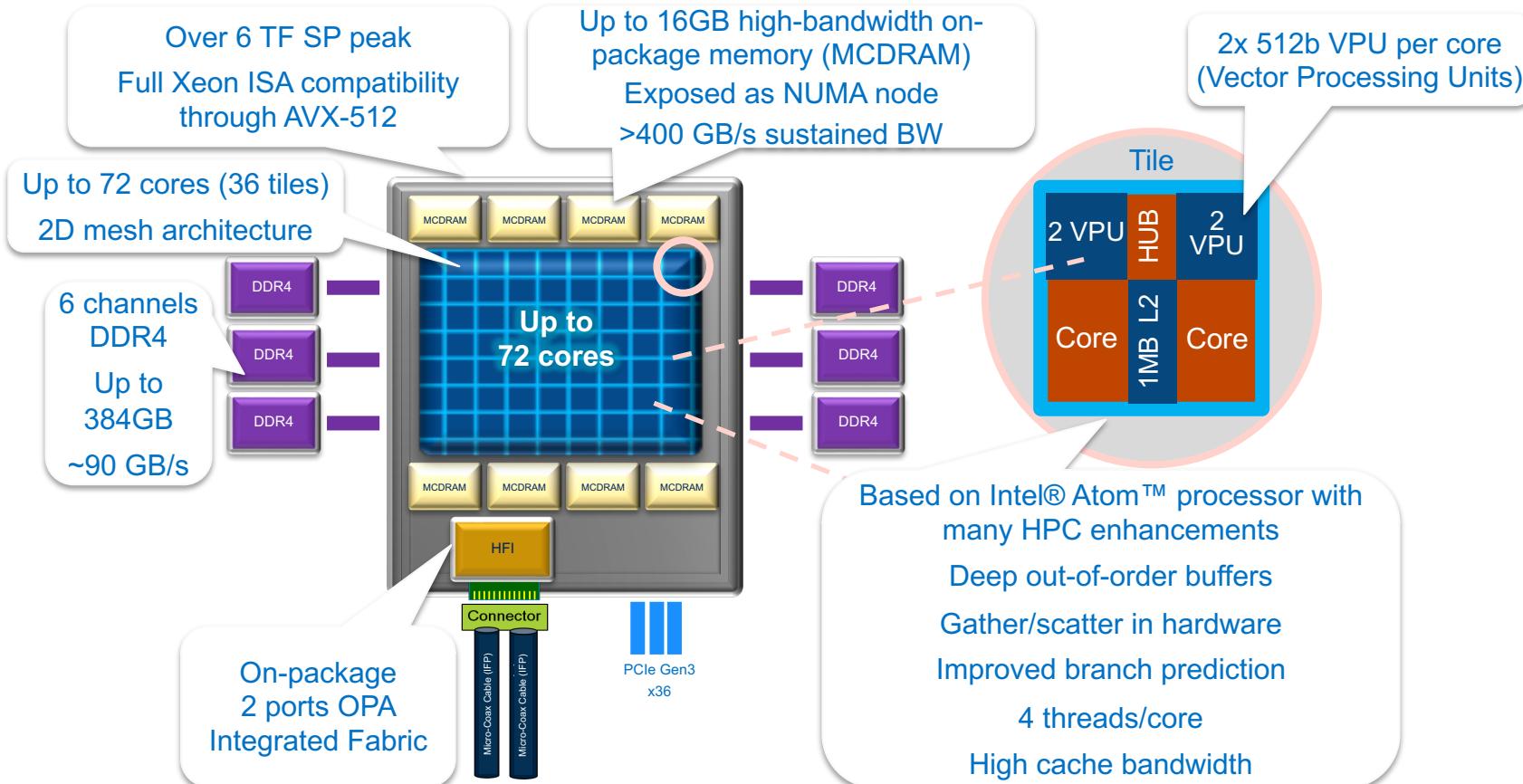


Diagram is for conceptual purposes only and only illustrates a CPU and memory – it is not to scale and does not include all functional areas of the CPU, nor does it represent actual component layout.

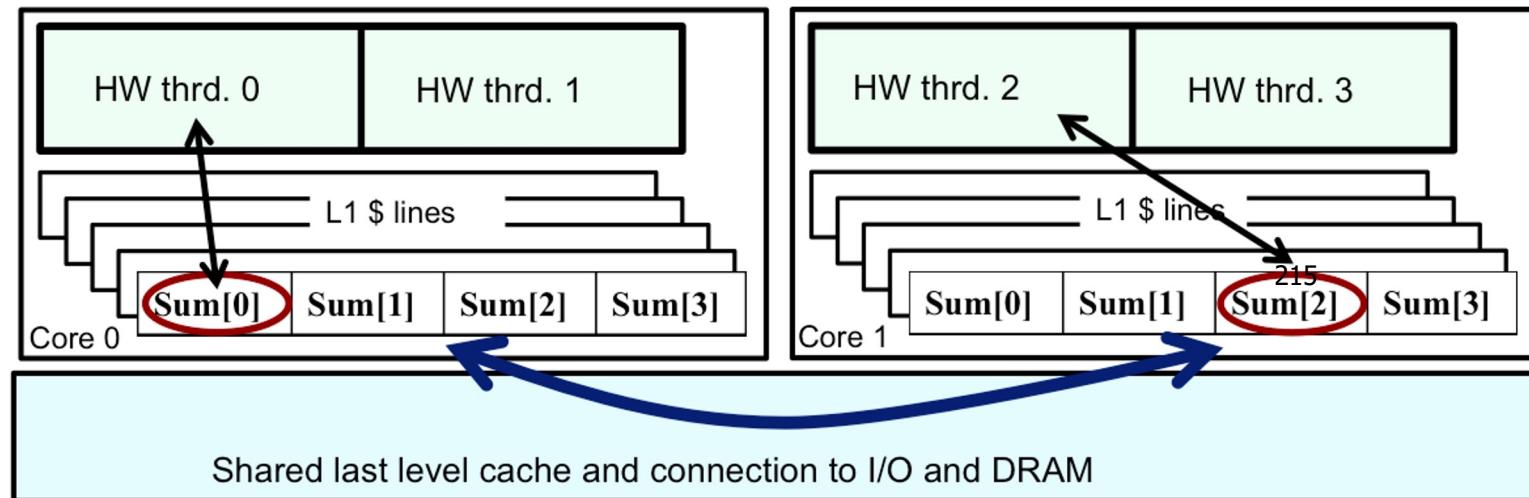
# Memory Locality

- Memory access in different NUMA domains are different
  - Accessing memory in remote NUMA is slower than accessing memory in local NUMA
  - Accessing High Bandwidth Memory on KNL\* is faster than DDR
- OpenMP does not explicitly map data across shared memories
- Memory locality is important since it impacts both memory and intra-node performance

\*KNL: Intel® Xeon Phi™ processor 7250 with 68 cores @ 1.4 Ghz ...  
the “bootable” version that sits in a socket, not a co-processor

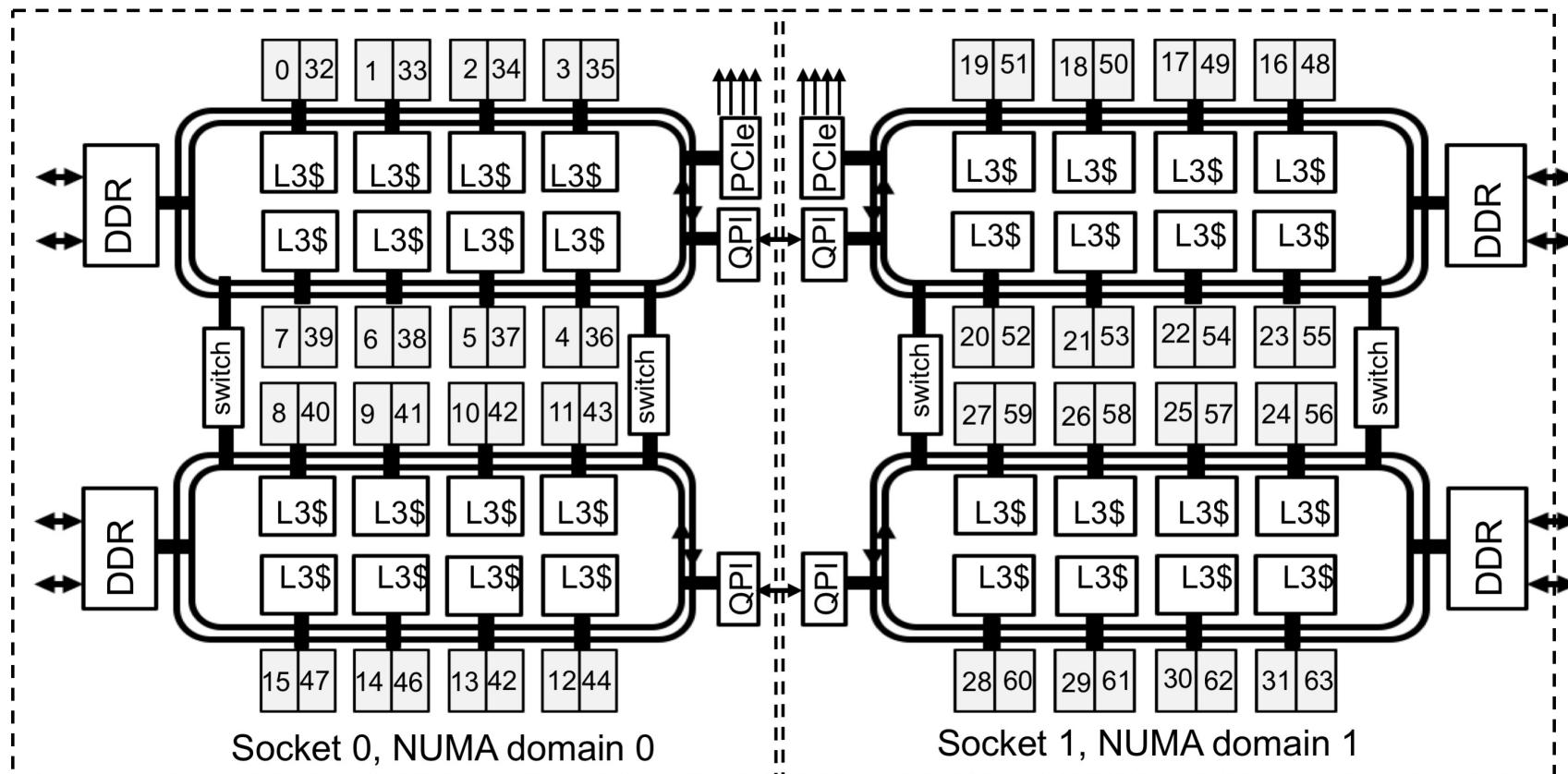
# Cache Coherence and False Sharing

- ccNUMA node: cache-coherence NUMA node.
- Data from memory are accessed via cache lines.
- Multiple threads hold local copies of the same (global) data in their caches. Cache coherence ensures the local copy to be consistent with the global data.
- Main copy needs to be updated when a thread writes to local copy.
- Writes to same cache line from different threads is called false sharing or cache thrashing, since it needs to be done in serial. Use atomic or critical or private variables to avoid race condition.



# Exploring your NUMA world: numactl

- numactl shows you how the OS processor-numbers map onto the physical cores of the chip:



2 Intel® Xeon™ E5-2698 v3 CPUs (Haswell) per node (launched Q3'14)

# Tool to Check NUMA Node Information: numactl

- **numactl**: controls NUMA policy for processes or shared memory
  - **numactl -H**: provides NUMA info of the CPUs

**Haswell node example  
32 cores, 2 sockets**

```
% numactl -H
available: 2 nodes (0-1)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
node 0 size: 64430 MB
node 0 free: 63002 MB
node 1 cpus: 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 48 49 50 51 52 53 54 55 56 57 58 59 60 61
62 63
node 1 size: 64635 MB
node 1 free: 63395 MB
node distances:
node  0    1
 0: 10 21 } ←
 1: 21 10 }
```

Shows relative costs .... In this case, there's a factor of two in the cost of the local (on CPU) DRAM vs going to the other socket

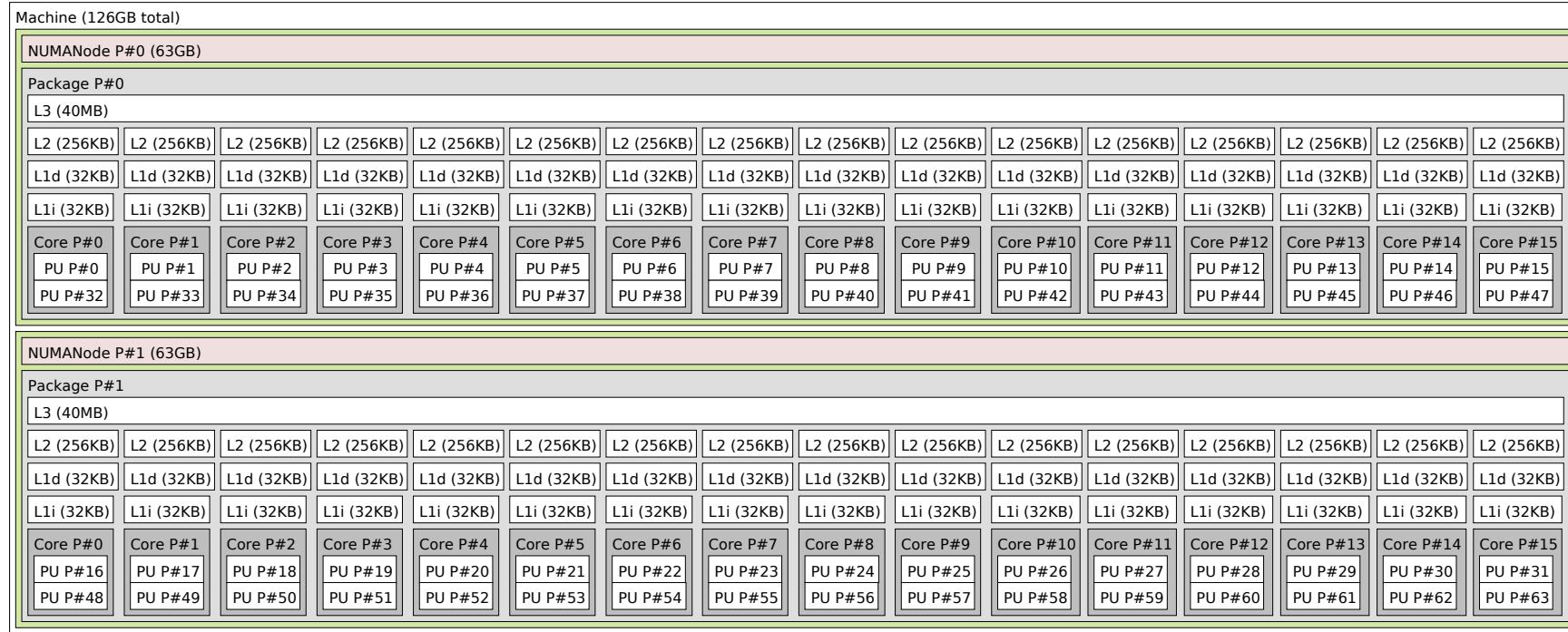
\*Haswell: 16-core Intel® Xeon™ Processor E5-2698 v3 at 2.3 GHz

# Use numactl Command Line Tool

- **numactl** is a Linux tool to investigate and handle NUMA
- Can be used to request CPU or memory binding
  - Use “**numactl <options> ./myapp**” as the executable (instead of “**./myapp**”)
- CPU binding example:
  - **% numactl --cpunodebind 0,1 ./code.exe**  
only use cores of NUMA nodes 0 and 1
- Memory binding example:
  - **% numactl --membind 1 ./code.exe**  
only use memory in NUMA nodes 1, such as the MCDRAM (High Bandwidth Memory) in KNL quad, flat mode

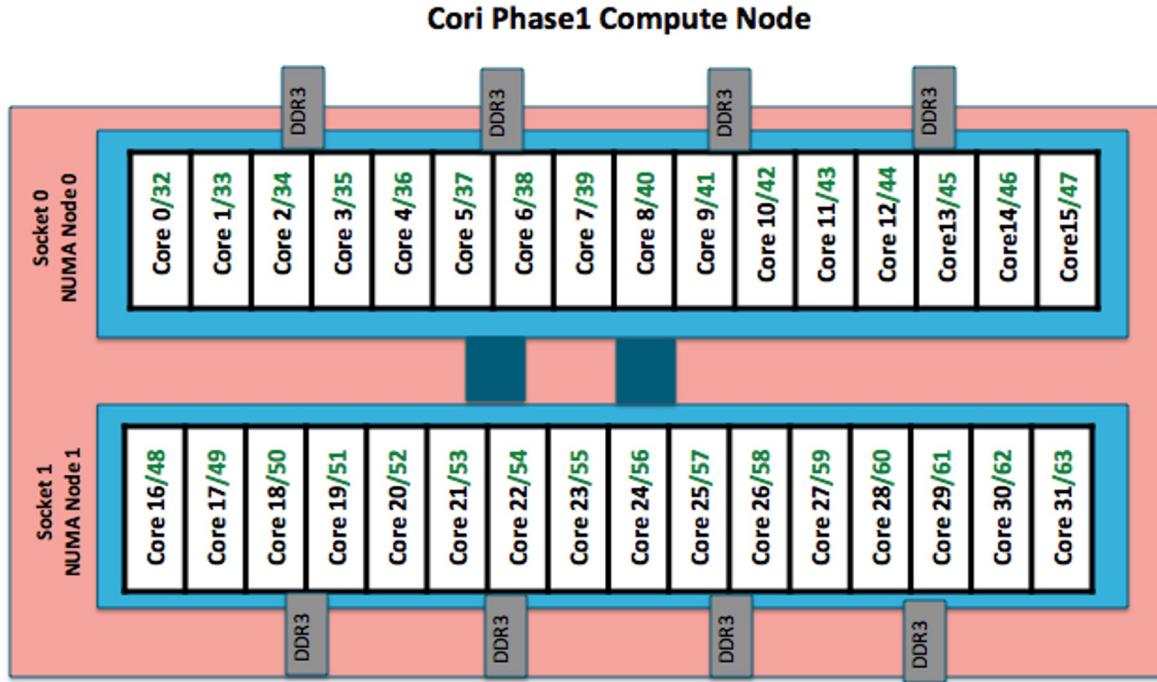
# Tools to Check Node Information: hwloc

- Portable Hardware Locality (hwloc)
  - `hwloc-ls` and `lstopo`: provides a **text** and **graphical** representation of the system topology, NUMA nodes, cache info, and the mapping of procs.



**Haswell node example  
32 cores, 2 sockets**

# Haswell Compute Nodes Example



To obtain processor info:

Get on a compute node:

```
% salloc -N 1 -C ...
```

Then:

```
% numactl -H
```

```
or % cat /proc/cpuinfo
```

```
or % hwloc-ls
```

- Each Haswell node has 2 Intel Xeon 16-core Haswell processors
  - 2 NUMA domains (sockets) per node, 16 cores per NUMA domain. 2 hardware threads per physical core.
  - NUMA Domain 0: physical cores 0-15 (and logical cores 32-47)  
NUMA Domain 1: physical cores 16-31 (and logical cores 48-63)
- Memory bandwidth is non-homogeneous among NUMA domains

# Find Processor Info on a Mac Laptop

```
$ sysctl -n machdep.cpu.brand_string  
Intel(R) Core(TM) i7-8569U CPU @ 2.80GHz
```

```
$ system_profiler |grep Processor
```

...

Processor Name: Quad-Core Intel Core i7

Processor Speed: 2.8 GHz

Number of Processors: 1

...

## Exercise: Node Information

- Characterize the processor/memory layout of your system
- Try on a Cori login node, a Cori Haswell and a Cori KNL node, and find out the differences

# Process / Thread / Memory Affinity (1)

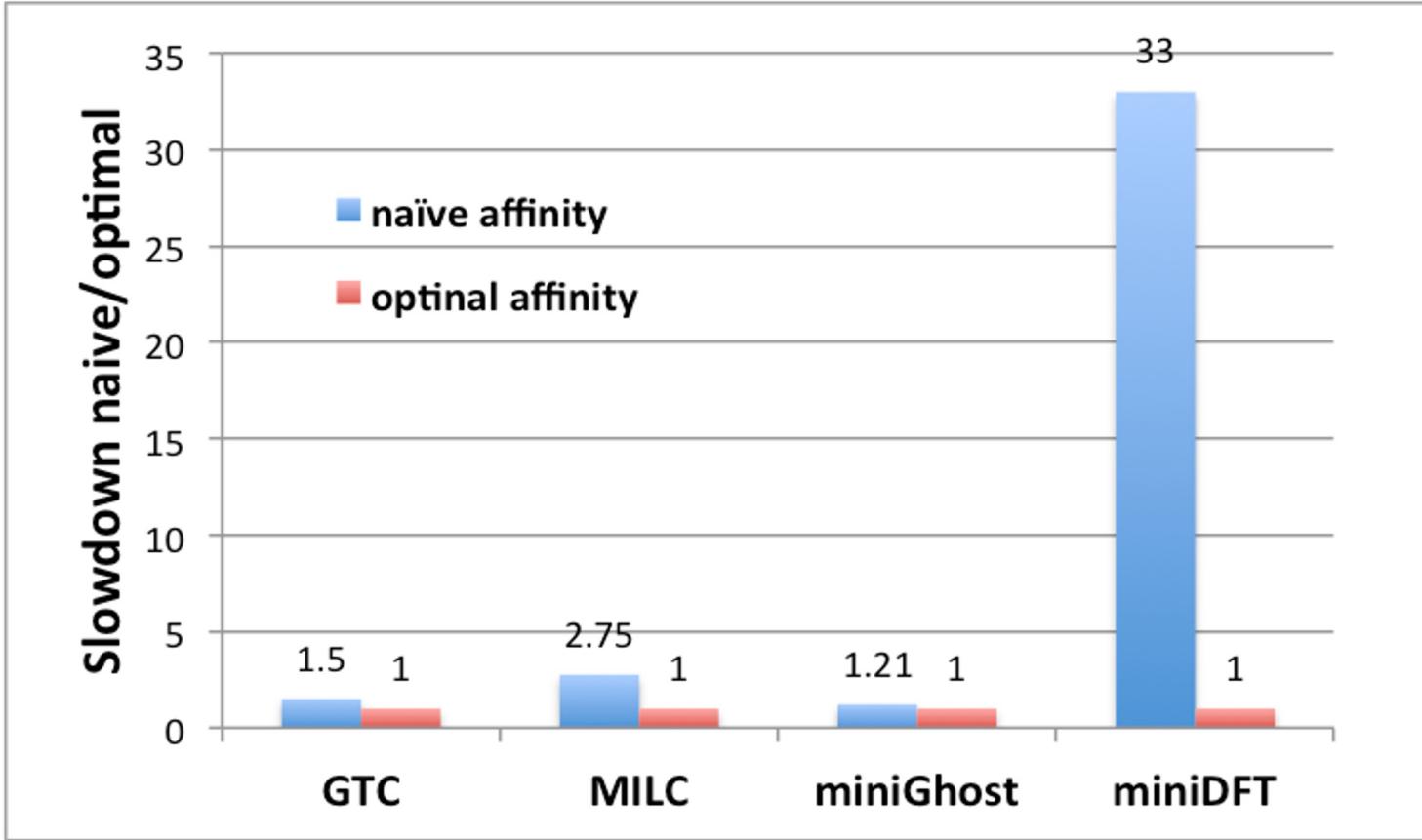
- **Process Affinity**: also called "CPU pinning", binds processes (MPI tasks, etc.) to a CPU or a range of CPUs on a node
  - It is important to spread MPI ranks evenly onto cores in different NUMA domains
- **Thread Affinity**: further binding threads to CPUs that are allocated to their parent process
  - **Thread affinity should be based on achieving process affinity first**
  - Threads forked by a certain MPI task have thread affinity binding close to the process affinity binding of their parent MPI task
  - **Do not over schedule CPUs for threads**

# Process / Thread / Memory Affinity (2)

- **Memory Locality**: allocate memory as close as possible to the core on which the task that requested the memory is running
  - Applications should **use memory from local NUMA domain as much as possible**
- **Cache Locality**: reuse data in cache as much as possible
- Our goal is to promote **OpenMP standard settings for portability**
  - OMP\_PLACES and OMP\_PROC\_BIND are preferred to vendor specific settings
- Correct process, thread and memory affinity is the basis for getting optimal performance. It is also essential for guiding further performance optimizations.

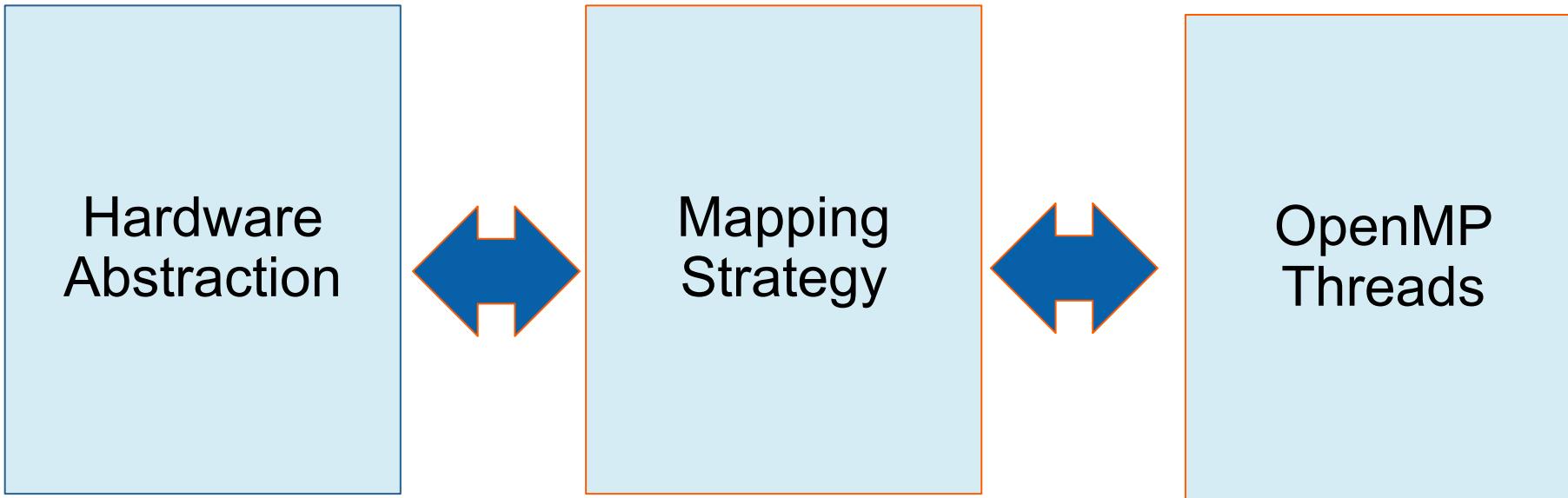
# Naïve vs. Optimal Affinity

Application Benchmark Performance on Cori



# OpenMP Thread Affinity

- Three main concepts:



**OMP\_PLACES**  
Environment Variable  
(e.g. threads, cores,  
sockets)

**OMP\_PROC\_BIND**  
Environment Variable  
or  
**proc\_bind()** clause  
of parallel region

**OMP\_NUM\_THREADS**  
Environment Variable  
or  
**num\_threads()** clause  
of parallel region

*Courtesy of Oscar Hernandez, ORNL*

# Writing NUMA-aware OpenMP Code

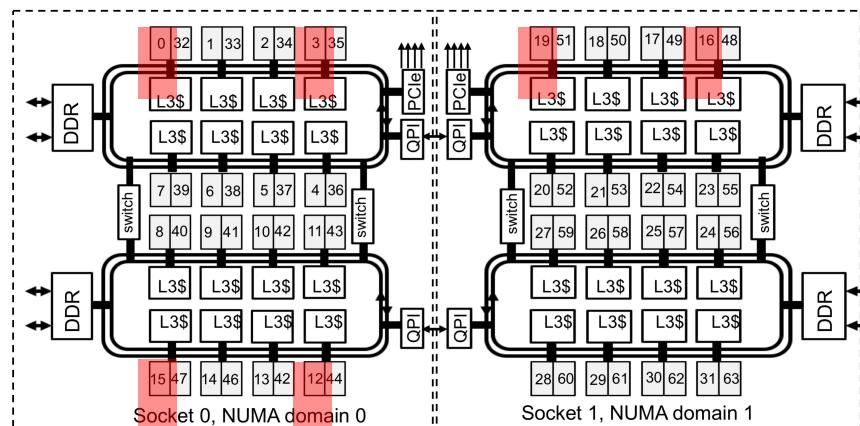
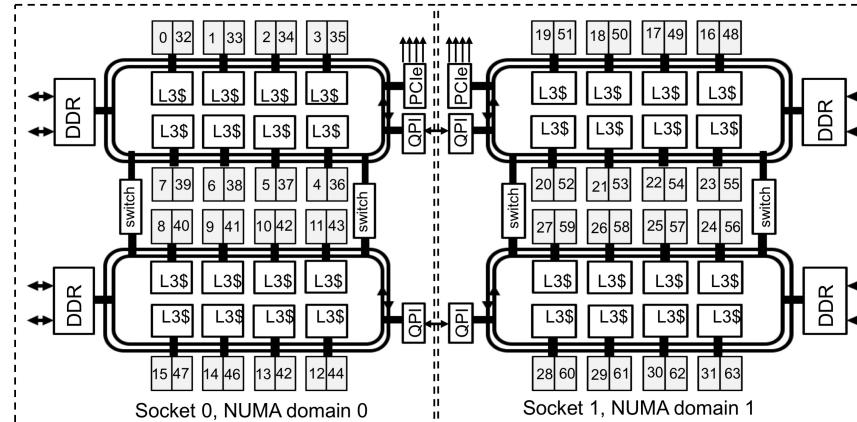
- 
- Control the places where threads are mapped
    - Place threads onto cores to optimize performance
    - Keep threads working on similar data close to each other
    - Maximize utilization of memory controllers by spreading threads out
  - Processor binding ... Disable thread migration
    - By Default, an OS migrates threads to maximize utilization of resources on the chip.
    - To Optimize for NUMA, we need to turn off thread migration ... bind threads to a processor/core
  - Memory Affinity
    - Maximize reuse of data in the cache hierarchy
    - Maximize reuse of data in memory pages

# The Concept of Places

- The Operating System assigns logical CPU IDs to hardware threads.
- Recall ... the linux command `numactl -H` returns those numbers.
- A place: numbers between {}:  
`export OMP_PLACES="{}0,1,2,3{}"`
- A place defines where threads can run

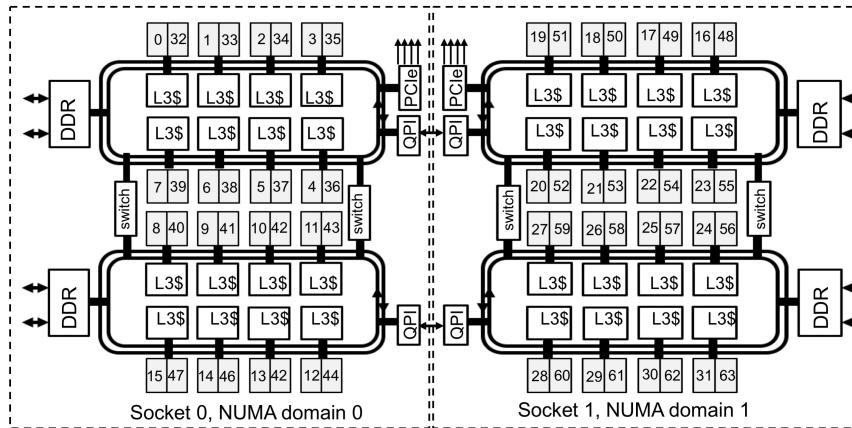
```
> export OMP_PLACES "{0, 3, 15, 12, 19, 16, 28, 31}"
> export NUM_THREADS= 6
```

```
#pragma omp parallel
{
    // do a bunch of cool stuff
}
```



# The Concept of Places

- The Operating System assigns logical CPU IDs to hardware threads.
- Recall ... the linux command `numactl -H` returns those numbers.
- Set with an environment variable:  
`export OMP_PLACES="0,1,2,3"`
- Can also specify with `{lower-bound:length:stride}`
- Can define multiple places:



Default  
Stride is 1

`OMP_PLACES="0,1,2,3" → OMP_PLACES="0:4:1" → OMP_PLACES="0:4"`

`OMP_PLACES="0,1,2,3},{4,6,8},{9,10,11,12}"`

`OMP_PLACES="0,4},{4,3:2},{9:4}"`

These are  
equivalent

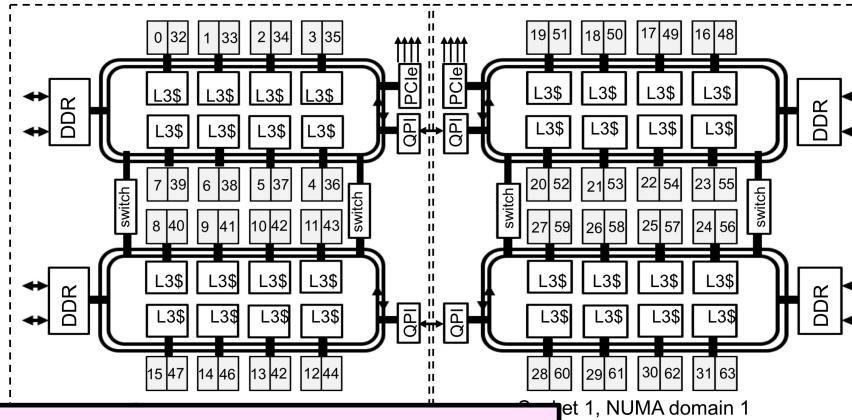
# The Concept of Places

- The Operating System assigns logical CPU IDs to hardware threads.
- Recall ... the linux command `numactl -H` returns those numbers.

- Set with ~~an environment variable~~

Programmers can use OMP\_PLACES for detailed control over the execution-units threads utilize. BUT ...

- Can a  
OMP
  - The rules for mapping onto physical execution units are complicated.
  - PLACES expressed as numbers is non-portable
- There has to be an easier and more portable way to describe places
- Can



`OMP_PLACES="0,1,2,3},{4,6,8},{9,10,11,12}"`

These are equivalent

`OMP_PLACES="0,4},{4,3:2},{9:4}"`

# Hardware Abstraction: OMP\_PLACES

- OMP\_PLACES environment variable
  - controls thread allocation
  - defines a series of places to which the threads are assigned
- It can be an abstract name or a specific list
  - **threads**: each place corresponds to a single hardware thread
  - **cores**: each place corresponds to a single core (having one or more hardware threads)
  - **sockets**: each place corresponds to a single socket (consisting of one or more cores)
  - a list with explicit place values of CPU ids, such as:
    - `export OMP_PLACES=" {0:4:2},{1:4:2}"` (equivalent to "`{0,2,4,6},{1,3,5,7}`")

- Examples:
  - `export OMP_PLACES=threads`
  - `export OMP_PLACES=cores`

# Writing NUMA-aware OpenMP Code

- Control the places where threads are mapped
  - Place threads onto cores to optimize performance
  - Keep threads working on similar data close to each other
  - Maximize utilization of memory controllers by spreading threads out
- Processor binding ... Disable thread migration
  - By Default, an OS migrates threads to maximize utilization of resources on the chip.
  - To Optimize for NUMA, we need to turn off thread migration ... bind threads to a processor/core
- Memory Affinity
  - Maximize reuse of data in the cache hierarchy
  - Maximize reuse of data in memory pages

# Mapping Strategy: OMP\_PROC\_BIND (1)

- Controls thread affinity within and between OpenMP places
- Allowed values:
  - **true**: the runtime will not move threads around between processors
  - **false**: the runtime may move threads around between processors
  - **close**: bind threads close to the master thread
  - **spread**: bind threads as evenly distributed (spreaded) as possible
  - **primary\***: bind threads to the same place as the master thread
- The values **primary\***, **close**, and **spread** imply the value **true**

Examples:

```
export OMP_PROC_BIND=spread  
export OMP_PROC_BIND=spread,close (for nested levels)
```

\*the term “master” has been deprecated in OpenMP 5.1 and replaced with the term “primary”.

## Mapping Strategy: OMP\_PROC\_BIND (2)

- Put threads far apart (spread) may improve aggregated memory bandwidth and available cache size for your application, but may also increase synchronization overhead
- Put threads “close” have the reverse impact as “spread”

# Mapping Strategy: OMP\_PROC\_BIND (2)

Prototype example: 4 cores total, 2 hyperthreads per core, 4 OpenMP threads

- **none**: no affinity setting
- **close**: Bind threads as close to each other as possible

| Node   | Core 0 |     | Core 1 |     | Core 2 |     | Core 3 |     |
|--------|--------|-----|--------|-----|--------|-----|--------|-----|
|        | HT1    | HT2 | HT1    | HT2 | HT1    | HT2 | HT1    | HT2 |
| Thread | 0      | 1   | 2      | 3   |        |     |        |     |

- **spread**: Bind threads as far apart as possible

| Node   | Core 0 |     | Core 1 |     | Core 2 |     | Core 3 |     |
|--------|--------|-----|--------|-----|--------|-----|--------|-----|
|        | HT1    | HT2 | HT1    | HT2 | HT1    | HT2 | HT1    | HT2 |
| Thread | 0      |     | 1      |     | 2      |     | 3      |     |

- **master**: bind threads to the same place as the master thread

# Various Methods to Set Number of Threads

## 1) Use num\_threads clause

```
#pragma omp parallel num_threads(4)
{
    int ID = omp_get_thread_num();
    pooh(ID,A);
}
```

## 2) Call omp\_set\_num\_threads API

```
omp_set_num_threads(4);
#pragma omp parallel
{
    int ID = omp_get_thread_num();
    pooh(ID,A);
}
```

## 3) Set runtime environment

```
export OMP_NUM_THREADS=4
#pragma omp parallel
{
    int ID = omp_get_thread_num();
    pooh(ID,A);
}
```

## 4) Do none of the three above.

Code will use an implementation dependent default number of threads defined by the compiler.

- Precedence: 1) > 2) > 3) > 4)
- You may get fewer threads than you requested, check with `omp_get_num_threads()`

# Affinity Clauses for OpenMP Parallel Construct

- The `num_threads` and `proc_bind` clauses can be used
  - The values set with these clauses take precedence over values set by runtime environment variables
- Helps code portability

- Examples:

- C/C++:

```
#pragma omp parallel num_threads(2) proc_bind(spread)
```

- Fortran:

```
!$omp parallel num_threads (2) proc_bind (spread)
```

```
...
```

```
!$omp end parallel
```

# Affinity Verification Methods

- NERSC provides pre-built binaries from a Cray code (xthi.c) to display process thread affinity

```
% srun -n 32 -c 8 --cpu-bind=cores check-mpi.intel.cori | sort -nk 4
```

```
Hello from rank 0, on nid02305. (core affinity = 0,1,68,69,136,137,204,205)
```

```
Hello from rank 1, on nid02305. (core affinity = 2,3,70,71,138,139,206,207)
```

- Use portable OpenMP environment variables [OMP\\_DISPLAY\\_AFFINITY](#) and [OMP\\_AFFINITY\\_FORMAT](#) (in OpenMP 5.0)

- Automatically displays affinity info when [OMP\\_DISPLAY\\_AFFINITY=true](#)
- Can set custom [OMP\\_DISPLAY\\_AFFINITY\\_FORMAT](#)
- Also has runtime APIs such as [omp\\_display\\_affinity](#) and [omp\\_capture\\_affinity](#)

# OMP\_AFFINITY\_FORMAT Fields

| Short Name | Long name       | Meaning                                                                                                                                       |
|------------|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| L          | thread_level    | from omp_get_level()                                                                                                                          |
| n          | thread_num      | from omp_get_thread_num()                                                                                                                     |
| a          | thread_affinity | the numerical identifiers of the processors the current thread is binding to, in the format of a comma separated list of OpenMP thread places |
| h          | host            | host or node name                                                                                                                             |
| p          | process_id      | process id used by the implementation (such as the process id for the MPI process)                                                            |
| N          | num_threads     | from omp_get_num_threads()                                                                                                                    |
| A          | ancestor_tnum   | from omp_get_ancestor_thread_num(). One level up only.                                                                                        |

```
% export OMP_DISPLAY_AFFINITY=true
% export OMP_AFFINITY_FORMAT="host=%h, pid=%p, thread_num=%n, thread affinity=%a"
host=nid02496, pid=150147, thread_num=0, thread affinity=0
host=nid02496, pid=150147, thread_num=1, thread affinity=4
% export OMP_AFFINITY_FORMAT="Thread Affinity: %0.3L %.10n %.20{thread_affinity} %.15h"
Thread Affinity: 001      0      0-1,16-17      nid003
Thread Affinity: 001      1      2-3,18-19      nid003
```

# Sample Nested OpenMP Program

```
#include <omp.h>
#include <stdio.h>
void report_num_threads(int level)
{
    #pragma omp single {
        printf("Level %d: number of threads in the
team: %d\n", level, omp_get_num_threads());
    }
}
int main()
{
    omp_set_dynamic(0);
    #pragma omp parallel num_threads(2) {
        report_num_threads(1);
        #pragma omp parallel num_threads(2) {
            report_num_threads(2);
            #pragma omp parallel num_threads(2) {
                report_num_threads(3);
            }
        }
    }
    return(0);
}
```

% ./a.out

Level 1: number of threads in the team: 2

Level 2: number of threads in the team: 1

Level 3: number of threads in the team: 1

Level 2: number of threads in the team: 1

Level 3: number of threads in the team: 1

% export OMP\_NESTED=true

% export OMP\_MAX\_ACTIVE\_LEVELS=3

% ./a.out

Level 1: number of threads in the team: 2

Level 2: number of threads in the team: 2

Level 2: number of threads in the team: 2

Level 3: number of threads in the team: 2

Level 3: number of threads in the team: 2

Level 3: number of threads in the team: 2

Level 3: number of threads in the team: 2

Level 0: P0

Level 1: P0 P1

Level 2: P0 P2; P1 P3

Level 3: P0 P4; P2 P5; P1 P6; P3 P7

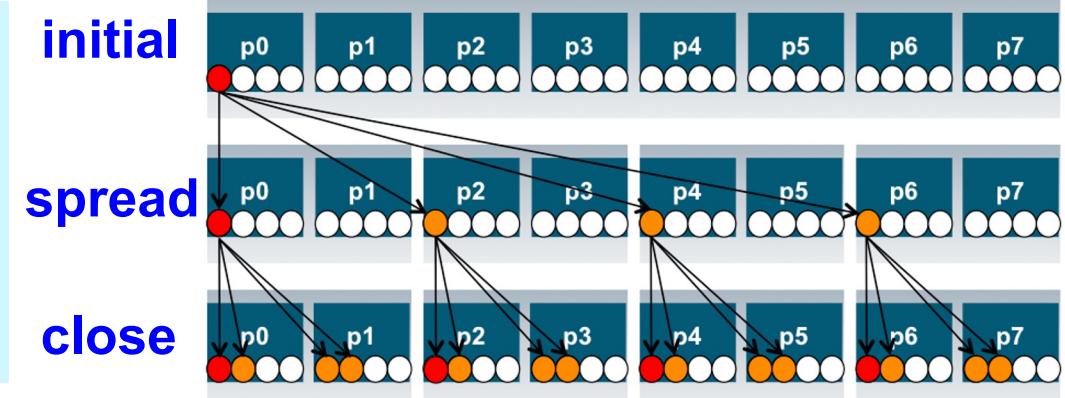
# Process and Thread Affinity in Nested OpenMP

- A combination of OpenMP environment variables and runtime flags are needed for different compilers and different batch schedulers on different systems

```
#pragma omp parallel proc_bind(spread)  
#pragma omp parallel proc_bind(close)
```

**Example: Use Intel compiler with SLURM on Cori Haswell:**  
export OMP\_NESTED=true  
export OMP\_MAX\_ACTIVE\_LEVELS=2  
**export OMP\_NUM\_THREADS=4,4**  
**export OMP\_PROC\_BIND=spread,close**  
**export OMP\_PLACES=threads**  
srun -n 4 -c 16 --cpu\_bind=cores ./code.exe

Illustration of a system with:  
2 sockets, 4 cores per socket,  
4 hyper-threads per core



- Use num\_threads clause in source codes to set threads for nested regions
- For most other non-nested regions, use OMP\_NUM\_THREADS environment variable for simplicity and flexibility

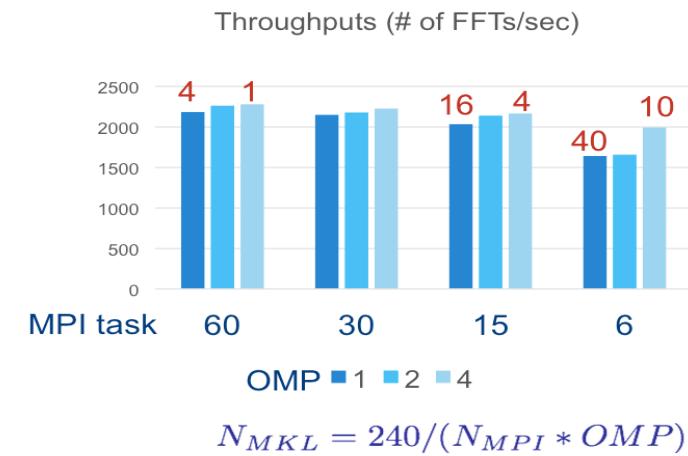
# When to Use Nested OpenMP

- Beneficial to use nested OpenMP to allow more fine-grained thread parallelism
- Some application teams are exploring with nested OpenMP to allow more fine-grained thread parallelism
  - Hybrid MPI/OpenMP not using node fully packed
  - Top level OpenMP loop does not use all available threads
  - Multiple levels of OpenMP loops are not easily collapsed
  - Certain computational intensive kernels could use more threads
  - MKL can use extra cores with nested OpenMP
- Nested level can be arbitrarily deep

# Use Multiple Threads in MKL

- By Default, in OpenMP parallel regions, only 1 thread will be used for MKL calls.
  - MKL\_DYNAMICS is true by default
- Nested OpenMP can be used to enable multiple threads for MKL calls. **Treat MKL as a nested inner OpenMP region.**
- Sample settings

```
export OMP_NESTED=true
export OMP_PLACES=cores
export OMP_PROC_BIND=sprad,close
export OMP_NUM_THREADS=6,4
export MKL_DYNAMICS=false
export OMP_MAX_ACTIVE_LEVELS=2
```



FFT3D on KNC, Ng=64<sup>3</sup> example  
Courtesy of Jeongnim Kim, Intel

\*KNC: Intel® Xeon Phi™ processor (Knights Corner) ... the first generation co-processor version of the chip.

## Exercise: Affinity Verification

- Run the “Hello World” code, use OMP\_DISPLAY\_AFFINITY to observe affinity status
- Change thread binding and number of threads and see how affinity status changes

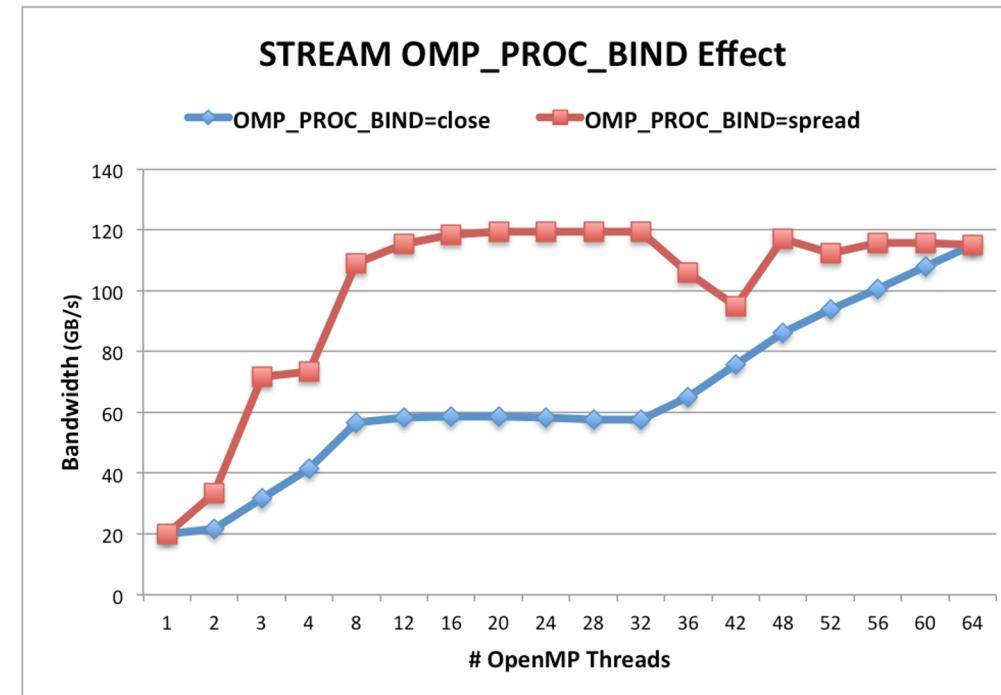
# OMP\_PROC\_BIND Choices for STREAM Benchmark

**OMP\_NUM\_THREADS=32**  
**OMP\_PLACES=threads**

**OMP\_PROC\_BIND=close**  
Threads 0 to 31 bind to CPUs 0,32,1,33,2,34,...15,47. All threads are in the first socket. The second socket is idle. Not optimal.

**OMP\_PROC\_BIND=spread**  
Threads 0 to 31 bind to CPUs 0,1,2,... to 31. Both sockets and memory are used to maximize memory bandwidth.

Blue: OMP\_PROC\_BIND=close  
Red: OMP\_PROC\_BIND=spread  
Both with First Touch



# Exercise: STREAM Benchmark

- Use the STREAM benchmark code: C/affinity/stream.c
  - Sample batch script: “run\_stream\_sample.sh”  
% sbatch <job\_script>
  - STREAM memory bandwidth results: check “Best Rate” for “Triad” in the output
  - Experiment with different OMP\_NUM\_THREADS, OMP\_PROC\_BIND, and OMP\_PLACES, and OMP\_DISPLAY\_AFFINITY settings to check thread affinity output and performance result
  - Run with 8, 16, 32, 48, 64 threads, and OMP\_PROC\_BIND=spread or close
- Compare your results with the previous STREAM plot

# Writing NUMA-aware OpenMP Code

- Control the places where threads are mapped
  - Place threads onto cores to optimize performance
  - Keep threads working on similar data close to each other
  - Maximize utilization of memory controllers by spreading threads out
- Processor binding ... Disable thread migration
  - By Default, an OS migrates threads to maximize utilization of resources on the chip.
  - To Optimize for NUMA, we need to turn off thread migration ... bind threads to a processor/core
- Memory Affinity
  - Maximize reuse of data in the cache hierarchy
  - Maximize reuse of data in memory pages



# Memory Affinity: “First Touch” memory

## *Step 1.1 Initialization by master thread only*

```
for (j=0; j<VectorSize; j++) {  
    a[j] = 1.0; b[j] = 2.0; c[j] = 0.0;}
```

## *Step 1.2 Initialization by all threads*

```
#pragma omp parallel for  
for (j=0; j<VectorSize; j++) {  
    a[j] = 1.0; b[j] = 2.0; c[j] = 0.0;}
```

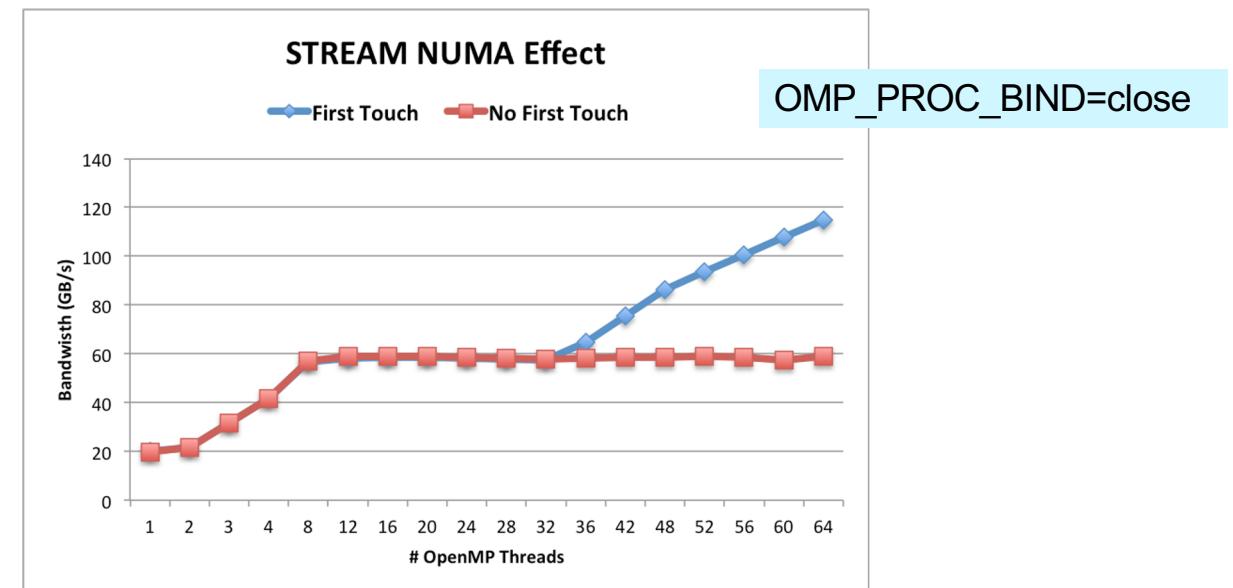
## *Step 2 Compute*

```
#pragma omp parallel for  
for (j=0; j<VectorSize; j++) {  
    a[j]=b[j]+d*c[j];}
```

- Memory affinity is not defined when memory was allocated, instead it will be defined at initialization.
- Memory will be local to the thread which initializes it. This is called **first touch** policy.
- Hard to do “perfect touch” for real applications. General recommendation is to [use number of threads fewer than number of CPUs \(one or more MPI tasks\) per NUMA domain](#).

Red: step 1.1 + step 2. No First Touch

Blue: step 1.2 + step 2. First Touch

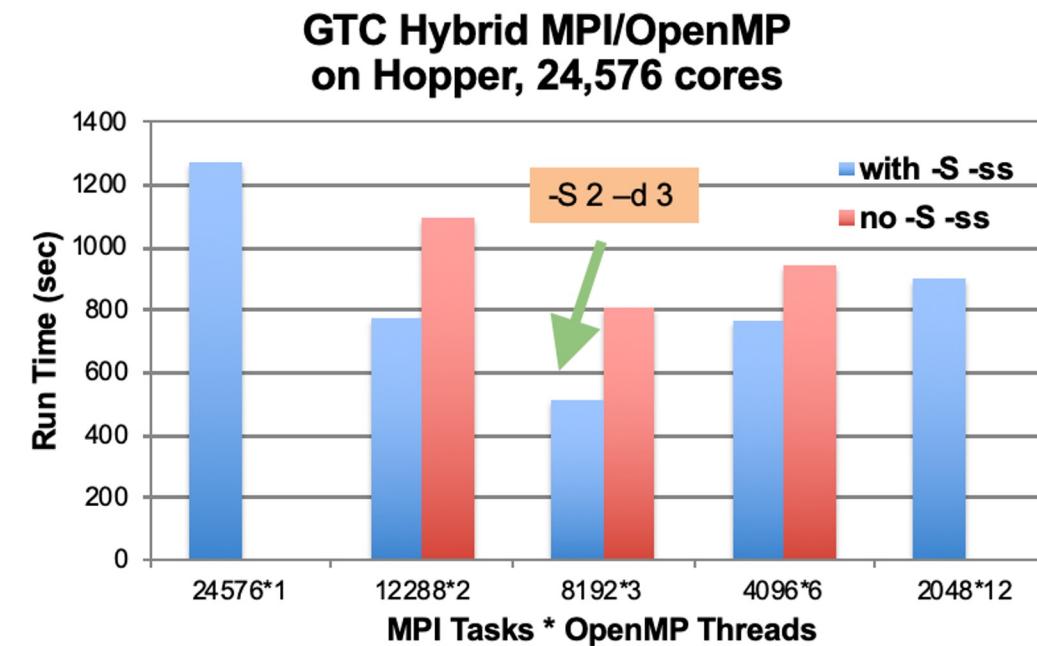
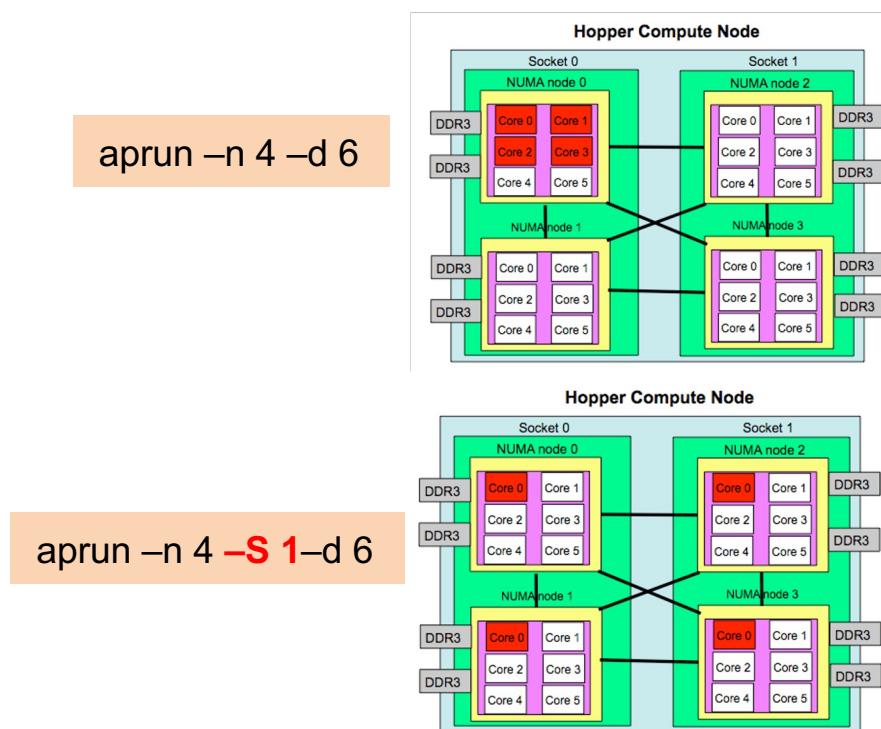


# “Perfect Touch” is Hard

- Hard to do “perfect touch” for real applications
- General recommendation: use number of threads fewer than number of CPUs per NUMA domain
- In the previous example, there are 16 cores (32 CPUs) per NUMA domain. Sample run options:
  - 2 MPI tasks, 1 MPI task per NUMA domain, with 32 OpenMP threads (if using hyperthreads) or 16 OpenMP threads (if not using hyperthreads) per MPI task
  - 4 MPI tasks, 2 MPI tasks per NUMA domain, with 16 OpenMP threads (if using hyperthreads) or 8 OpenMP threads (if not using hyperthreads) per MPI task
  - ...

# MPI Process Affinity Example: aprun “-S” Option

- Important to spread MPI ranks evenly onto different NUMA nodes
- Use the “-S” option: [specify #MPI\\_tasks per NUMA domain](#)
- The example below was from an XE6 system (NERSC Hopper)



# Exercise: Importance of First Touch

- Do the same STREAM experiments with the no first touch code: “stream\_nft.c” to understand the impact of first touch
  - Experiment with different OMP\_NUM\_THREADS, OMP\_PROC\_BIND, and OMP\_PLACES, and OMP\_DISPLAY\_AFFINITY settings to check thread affinity output and performance result
  - Run with 8, 16, 32, 48, 64 threads, and OMP\_PROC\_BIND=spread or close
- Compare your results with the previous STREAM plot

# OpenMP task-to-data Affinity (in OpenMP 5.0)

- Affinity hints can be provided for OpenMP tasks, resulting data to be closer to tasks
- Useful for multi-socket systems

```
void task_affinity() {  
    double* B;  
#pragma omp task shared(B) affinity(A[0:N])  
    B = init_B_and_important_computation(A);  
  
#pragma omp task firstprivate(B) affinity(B[0:N])  
    important_computation_too(B);  
  
#pragma omp taskwait  
}
```

# Memory Allocators (in OpenMP 5.0)

| Allocator name             | Storage selection intent                                                                                     |
|----------------------------|--------------------------------------------------------------------------------------------------------------|
| omp_default_mem_alloc      | use default storage                                                                                          |
| omp_large_cap_mem_alloc    | use storage with large capacity                                                                              |
| omp_const_mem_alloc        | use storage optimized for read-only variables                                                                |
| omp_high_bw_mem_alloc      | use storage with high bandwidth                                                                              |
| omp_low_lat_mem_alloc      | use storage with low latency                                                                                 |
| omp_cgroup_mem_alloc       | use storage close to all threads in the contention group of the thread requesting the allocation             |
| omp_pteam_mem_alloc        | use storage that is close to all threads in the same parallel region of the thread requesting the allocation |
| omp_thread_local_mem_alloc | use storage that is close to the thread requesting the allocation                                            |

- Support versatile types of memory available on current and future systems: DDR, High-Bandwidth Memory (HBM), non-volatile memory, constant memory
- Memory allocators define types of memory that variables can be allocated to, such as large capacity, low latency, cgroup, thread local, etc.

# Using Memory Allocators

```
void allocator_example(omp_allocator_t *my_allocator) {  
    int a[M], b[N];  
    #pragma omp allocate(a) allocator(omp_high_bw_mem_alloc)  
    #pragma omp allocate(b) // use default OMP_ALLOCATOR  
  
    double *p = (double *) omp_alloc(N*M*sizeof(*p), my_allocator);  
  
    #pragma omp parallel private(a) allocate(omp_low_lat_mem_alloc:a)  
    {  
        some_parallel_code();  
    }  
    omp_free(p);  
}
```

## A NUMA Case study

# Benchmarking ... I Must Control Everything!

- Goal: To compare different programming systems applied to the same problem:
  - We must control everything we can to make sure any observed differences are due to the different programming systems.
- We need to know exactly which cores we are using and how thread IDs map onto cores ... so we can understand data detailed memory movement and make sure it's the same between the different test cases.

# Step 1: Know Your System

- My system did not have numactl or Hwloc. So I went with my third option .... lscpu (note: I'm only showing a subset of the actual output):

```
$ lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
Address sizes:         46 bits physical, 48 bits virtual
CPU(s):                72
On-line CPU(s) list:  0-71
Thread(s) per core:   2
Core(s) per socket:   18
Socket(s):             2
NUMA node(s):          2
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 63
Model name:            Intel(R) Xeon(R) CPU E5-2699 v3 @ 2.30GHz
Stepping:               2
CPU MHz:               1197.539
CPU max MHz:           3600.0000
CPU min MHz:           1200.0000
L1d cache:              1.1 MiB
L1i cache:              1.1 MiB
L2 cache:               9 MiB
L3 cache:               90 MiB
NUMA node0 CPU(s):     0-17,36-53
NUMA node1 CPU(s):     18-35,54-71
```

SMT enabled ... two HW threads per core

2 CPUs (sockets) with 18 physical cores per CPU

Note: a HW thread is a CPU (or core) as far as the OS is concerned. These two lines show you the numbering of these “cores”.

# Step 1: Know Your System

- My system did not have numactl or Hwloc. So I went with my third option .... lscpu (note: I'm only showing a subset of the actual output):

```
$ lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
Address sizes:         46 bits physical, 48 bits virtual
CPU(s):                72
On-line CPU(s) list:   0-71
Thread(s) per core:    2
Core(s) per socket:    18
Socket(s):             2
NUMA node(s):          2
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 63
Model name:            Intel(R) Xeon(R) CPU E5-2699 v3 @ 2.30GHz
Stepping:               2
CPU MHz:               1197.539
CPU max MHz:           3600.0000
CPU min MHz:           1200.0000
L1d cache:              1.1 MiB
L1i cache:              1.1 MiB
L2 cache:               9 MiB
L3 cache:               90 MiB
NUMA node0 CPU(s):      0-17,36-53
NUMA node1 CPU(s):      18-35,54-71
```

SMT enabled ... two HW threads per core

2 CPUs (sockets) with 18 physical cores per CPU

The numbering of these “cores” (in 2 sockets).

|       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0/36  | 1/37  | 2/38  | 3/39  | 4/40  | 5/41  | 6/42  | 7/43  | 8/44  |
| 9/45  | 10/46 | 11/47 | 12/48 | 13/49 | 14/50 | 15/51 | 16/52 | 17/53 |
| 18/54 | 19/55 | 20/56 | 21/57 | 22/58 | 23/59 | 24/60 | 25/61 | 26/62 |
| 27/63 | 28/64 | 29/65 | 30/66 | 31/67 | 32/68 | 33/69 | 34/70 | 35/71 |

# Setup a Runscript (so you can reproduce the computations later)

```
#!/usr/bin/env bash
# Run script for DGEMM with C and OpenMP

# Define shared parameters for the calculations we will run
BLOCK=0
ORDER=1000
ITERS=5

# setup environment for the intel compilers
source /opt/intel/compilers_and_libraries_2020.4.304/linux/bin/compilervars.sh -arch intel64

# Setup display of mapping from OpenMP threads to "hardware" threads.
export OMP_DISPLAY_AFFINITY=true
export OMP_AFFINITY_FORMAT="Thrd Lev=%3L, thrd_num=%5n, thrd_aff=%15A"

# Enable explicit affinity control.
export OMP_PLACES="{0},{1},{2},{3},{4},{5},{6},{7},{8},{9},{10},{11},{12},{13},{14},{15},{16}"
export OMP_PROC_BIND=close

./dgemm 8 $ITERS $ORDER $BLOCK
./dgemm 16 $ITERS $ORDER $BLOCK
```

# A NUMA Case Study: Results

Parallel Research Kernels version 2.17

OpenMP Dense matrix-matrix multiplication

Thrd Lev=1 , thrd\_num=0 , thrd\_aff=0

Thrd Lev=1 , thrd\_num=4 , thrd\_aff=4

Thrd Lev=1 , thrd\_num=3 , thrd\_aff=3

Thrd Lev=1 , thrd\_num=5 , thrd\_aff=5

Thrd Lev=1 , thrd\_num=1 , thrd\_aff=1

Thrd Lev=1 , thrd\_num=2 , thrd\_aff=2

Thrd Lev=1 , thrd\_num=6 , thrd\_aff=6

Thrd Lev=1 , thrd\_num=7 , thrd\_aff=7

Matrix order = 1000

Number of threads = 8

Rate : 21650.601956 +/- 1589.413250 MFlops/s

Notice the one-to-one mapping of thread ID onto hardware thread.

Normally, this is going too far, but for benchmarking, this is a handy trick.

Parallel Research Kernels version 2.17

OpenMP Dense matrix-matrix multiplication

Thrd Lev=1 , thrd\_num=0 , thrd\_aff=0

Thrd Lev=1 , thrd\_num=13 , thrd\_aff=13

Thrd Lev=1 , thrd\_num=4 , thrd\_aff=4

Thrd Lev=1 , thrd\_num=11 , thrd\_aff=11

Thrd Lev=1 , thrd\_num=10 , thrd\_aff=10

Thrd Lev=1 , thrd\_num=8 , thrd\_aff=8

Thrd Lev=1 , thrd\_num=9 , thrd\_aff=9

Thrd Lev=1 , thrd\_num=1 , thrd\_aff=1

Thrd Lev=1 , thrd\_num=3 , thrd\_aff=3

Thrd Lev=1 , thrd\_num=2 , thrd\_aff=2

Thrd Lev=1 , thrd\_num=12 , thrd\_aff=12

Thrd Lev=1 , thrd\_num=7 , thrd\_aff=7

Thrd Lev=1 , thrd\_num=6 , thrd\_aff=6

Thrd Lev=1 , thrd\_num=5 , thrd\_aff=5

Thrd Lev=1 , thrd\_num=14 , thrd\_aff=14

Thrd Lev=1 , thrd\_num=15 , thrd\_aff=15

Matrix order = 1000

Number of threads = 16

Rate : 38765.867067 +/- 3303.460980 MFlops/s

## Obtain Optimal Affinity on Cori KNL Example

# KNL Compute Nodes

A Cori KNL node has 68 cores/272 CPUs, 96 GB DDR memory, 16 GB high bandwidth on package memory (MCDRAM)

| Arrangement of Hardware Threads for 68 Core KNL |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-------------------------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Core # →                                        | 0   | 1   | 2   | 3   | ... | 16  | 17  | 18  | ... | 33  | 34  | 35  | ... | 50  | 51  | 52  | ... | 65  | 66  | 67  |
| HW Thread #                                     | 0   | 1   | 2   | 3   | ... | 16  | 17  | 18  | ... | 33  | 34  | 35  | ... | 50  | 51  | 52  | ... | 65  | 66  | 67  |
|                                                 | 68  | 69  | 70  | 71  | ... | 84  | 85  | 86  | ... | 101 | 102 | 103 | ... | 118 | 119 | 120 | ... | 133 | 134 | 135 |
|                                                 | 136 | 137 | 138 | 139 | ... | 152 | 153 | 154 | ... | 169 | 170 | 171 | ... | 186 | 187 | 188 | ... | 201 | 202 | 203 |
|                                                 | 204 | 205 | 206 | 207 | ... | 220 | 221 | 222 | ... | 237 | 238 | 239 | ... | 254 | 255 | 256 | ... | 269 | 270 | 271 |

A [quad,cache](#) node (default setting) has only [1 NUMA node](#) with all CPUs on the NUMA node 0 (DDR memory). MCDRAM is hidden from the “numactl -H” result since it is a cache.

# Can We Just Do a Naive `srun`?

Example: 16 MPI tasks x 8 OpenMP threads per task on a single 68-core KNL quad,cache node:

```
% export OMP_NUM_THREADS=8  
% export OMP_PROC_BIND=spread      (other choice are "close","master","true","false")  
% export OMP_PLACES=threads       (other choices are: cores, sockets, and various ways to specify explicit lists, etc.)
```

```
% srun -n 16 ./xthi |sort -k4n,6n      or % mpirun -n 16 ./xthi  
Hello from rank 0, thread 0, on nid02304. (core affinity = 0)  
Hello from rank 0, thread 1, on nid02304. (core affinity = 144)      (on physical core 8)  
Hello from rank 0, thread 2, on nid02304. (core affinity = 17)  
Hello from rank 0, thread 3, on nid02304. (core affinity = 161)      (on physical core 25)  
...  
Hello from rank 1, thread 0, on nid02304. (core affinity = 0)  
Hello from rank 1, thread 1, on nid02304. (core affinity = 144)
```

**It is a mess!** e.g., thread 0 for rank 0, and thread 1 for rank 1 are on same physical core 0

# MPI Process Affinity: Selected Slurm `srun` Options

- `--cpu-bind=threads`  
Automatically generate masks binding tasks to threads
- `--cpu-bind=cores`  
Automatically generate masks binding tasks to cores
- `--cpu-bind=sockets`  
Automatically generate masks binding tasks to sockets
- `--cpu-bind=map_cpu:<cpulist>`  
Bind by setting CPU masks on tasks (or ranks)
- `--cpu-bind=map_ldom:<NUMA_domain_list>`  
Bind by mapping NUMA locality domain IDs to tasks  
(ldom means logical domain)

# Example mpirun or srun Commands: Fix the Problem

- The reason is #MPI tasks is not divisible by 68!
    - Each MPI task is getting  $68 \times 4 / \# \text{MPI tasks}$  of logical cores as the domain size
    - MPI tasks are crossing tile boundaries
  - Let's set number of logical cores per MPI task manually by wasting extra 4 cores on purpose, which is  $256 / \# \text{MPI tasks}$
- Cray MPICH with Aries network using native SLURM
    - `% srun -n 16 -c 16 --cpu_bind=cores ./code.exe`  
Notes: Here the value for `-c` is also set to number of logical cores per MPI task, i.e.,  $256 / \# \text{MPI tasks}$ .
  - Intel MPI with Omni Path using mpirun:
    - `% export I_MPI_PIN_DOMAIN=16`
    - `% mpirun -n 16 ./code.exe`

# Now It Looks Good!

**Process/thread affinity are good! (Marked first 6 and last MPI tasks only)**

# Intel KNL Quad,Flat Node Example

Cori KNL quad,flat node example  
68 cores (272 CPUs)

% numactl –H

```
available: 2 nodes (0-1)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82
83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117
118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147
148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177
178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207
208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237
238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267
268 269 270 271
node 0 size: 96723 MB
node 0 free: 93924 MB
node 1 cpus:
node 1 size: 16157 MB
node 1 free: 16088 MB
node distances:
node 0 1
0: 10 31
1: 31 10
```

- The quad,flat mode has only 2 NUMA nodes with all CPUs on the NUMA node 0 (DDR memory).
- And NUMA node 1 has MCDRAM (high bandwidth memory).

# Essential Runtime Settings for KNL MCDRAM Memory Affinity

- In quad, cache mode, no special setting is needed to use MCDRAM
- In quad,flat mode, using quad,flat as an example
  - NUMA node 1 is MCDRAM
- Enforced memory mapping to MCDRAM
  - If using >16 GB, malloc will fail
  - Use “**numactl -m 1 ./myapp**” as the executable  
(instead of “./myapp”)
- Preferred memory mapping to MCDRAM
  - If using >16 GB, malloc will spill to DDR
  - Use “**numactl -p 1 ./myapp**” as the executable  
(instead of “./myapp”)

# Process and Thread Affinity Best Practices

- Achieving best data locality, and optimal process and thread affinity is crucial in getting good performance with MPI/OpenMP, yet **not straightforward**
  - Understand the [node architecture](#) with tools such as “numactl -H” first
  - Set [correct cpu-bind and OMP\\_PLACES options](#)
  - Always use simple examples with the same settings for your real application to [verify affinity](#) first or check with `OMP_DISPLAY_AFFINITY`
  - For nested OpenMP, set `OMP_PROC_BIND=spread,close` is recommended
- Optimize code for memory affinity
  - Pay special attention to [avoid false sharing](#)
  - Exploit first touch data policy, or use [at least 1 MPI task per NUMA domain](#)
  - [Optimize code for cache locality](#)
  - Compare performance with [put threads close or far apart \(spread\)](#)
  - Use [omp\\_allocator](#)
  - Use [numactl -m](#) option to explicitly request memory allocation in specific NUMA domain (such as high bandwidth memory in KNL)

- Introduction to OpenMP
- Creating Threads
- Synchronization
- Parallel Loops
- Data Environment
- Memory Model
- Irregular Parallelism and Tasks
- Recap
- Beyond the Common Core:
  - Worksharing Revisited
  - Synchronization Revisited: Options for Mutual exclusion
  - Memory models and point-to-point Synchronization
  - Programming your GPU with OpenMP
  - Thread Affinity and Data Locality
  - Thread Private Data



# Data Sharing: Threadprivate

- Makes global data private to a thread
  - Fortran: **COMMON** blocks
  - C: File scope and static variables, static class members
- Different from making them **PRIVATE**
  - with **PRIVATE** global variables are masked.
  - **THREADPRIVATE** preserves global scope within each thread
- Threadprivate variables can be initialized using **COPYIN** or at time of definition (using language-defined initialization capabilities)

# A Threadprivate Example (C)

Use `threadprivate` to create a counter for each thread.

```
int counter = 0;  
#pragma omp threadprivate(counter)  
  
int increment_counter()  
{  
    counter++;  
    return (counter);  
}
```

# Data Copying: Copyin

You initialize threadprivate data using a copyin clause.

```
parameter (N=1000)
common/buf/A(N)
!$OMP THREADPRIVATE(/buf/)

!$ Initialize the A array
call init_data(N,A)

!$OMP PARALLEL COPYIN(A)
... Now each thread sees threadprivate array A initialized
... to the global value set in the subroutine init_data()

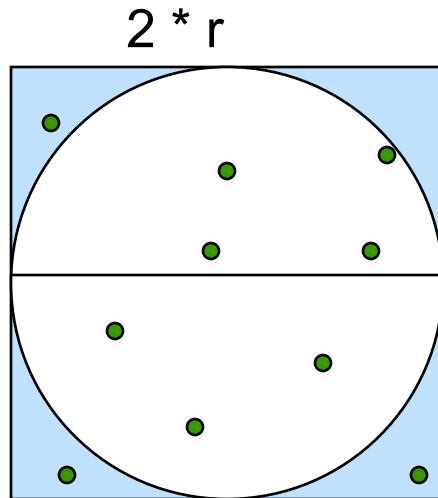
!$OMP END PARALLEL

end
```

# Exercise: Monte Carlo Calculations

Using random numbers to solve tough problems

- Sample a problem domain to estimate areas, compute probabilities, find optimal values, etc.
- Example: Computing  $\pi$  with a digital dart board:



|           |               |
|-----------|---------------|
| $N = 10$  | $\pi = 2.8$   |
| $N=100$   | $\pi = 3.16$  |
| $N= 1000$ | $\pi = 3.148$ |

- Throw darts at the circle/square.
- Chance of falling in circle is proportional to ratio of areas:
$$A_c = r^2 * \pi$$
$$A_s = (2 * r) * (2 * r) = 4 * r^2$$
$$P = A_c / A_s = \pi / 4$$
- Compute  $\pi$  by randomly choosing points;  $\pi$  is four times the fraction that falls in the circle

# Exercise: Monte Carlo pi (cont)

- We provide three files for this exercise
  - pi\_mc.c: the Monte Carlo method pi program
  - random.c: a simple random number generator
  - random.h: include file for random number generator
- Create a parallel version of this program.
- Run it multiple times with varying numbers of threads.
- Is the program working correctly? Is there anything wrong?

# Parallel Programmers love Monte Carlo algorithms

```
#include "omp.h"
static long num_trials = 10000;
int main ()
{
    long i;    long Ncirc = 0;    double pi, x, y;
    double r = 1.0; // radius of circle. Side of square is 2*r
    seed(0,-r, r); // The circle and square are centered at the origin
    #pragma omp parallel for private (x, y) reduction (+:Ncirc)
    for(i=0;i<num_trials; i++)
    {
        x = random();      y = random();
        if ( x*x + y*y ) <= r*r) Ncirc++;
    }

    pi = 4.0 * ((double)Ncirc/(double)num_trials);
    printf("\n %d trials, pi is %f \n",num_trials, pi);
}
```

Embarrassingly parallel: the parallelism is so easy its embarrassing.

Add two lines and you have a parallel program.

# Random Numbers: Linear Congruential Generator (LCG)

- LCG: Easy to write, cheap to compute, portable, OK quality

```
random_next = (MULTIPLIER * random_last + ADDEND)% PMOD;  
random_last = random_next;
```

- If you pick the multiplier and addend correctly, LCG has a period of PMOD.
- Picking good LCG parameters is complicated, so look it up (Numerical Recipes is a good source). I used the following:
  - ◆ MULTIPLIER = 1366
  - ◆ ADDEND = 150889
  - ◆ PMOD = 714025

# LCG code

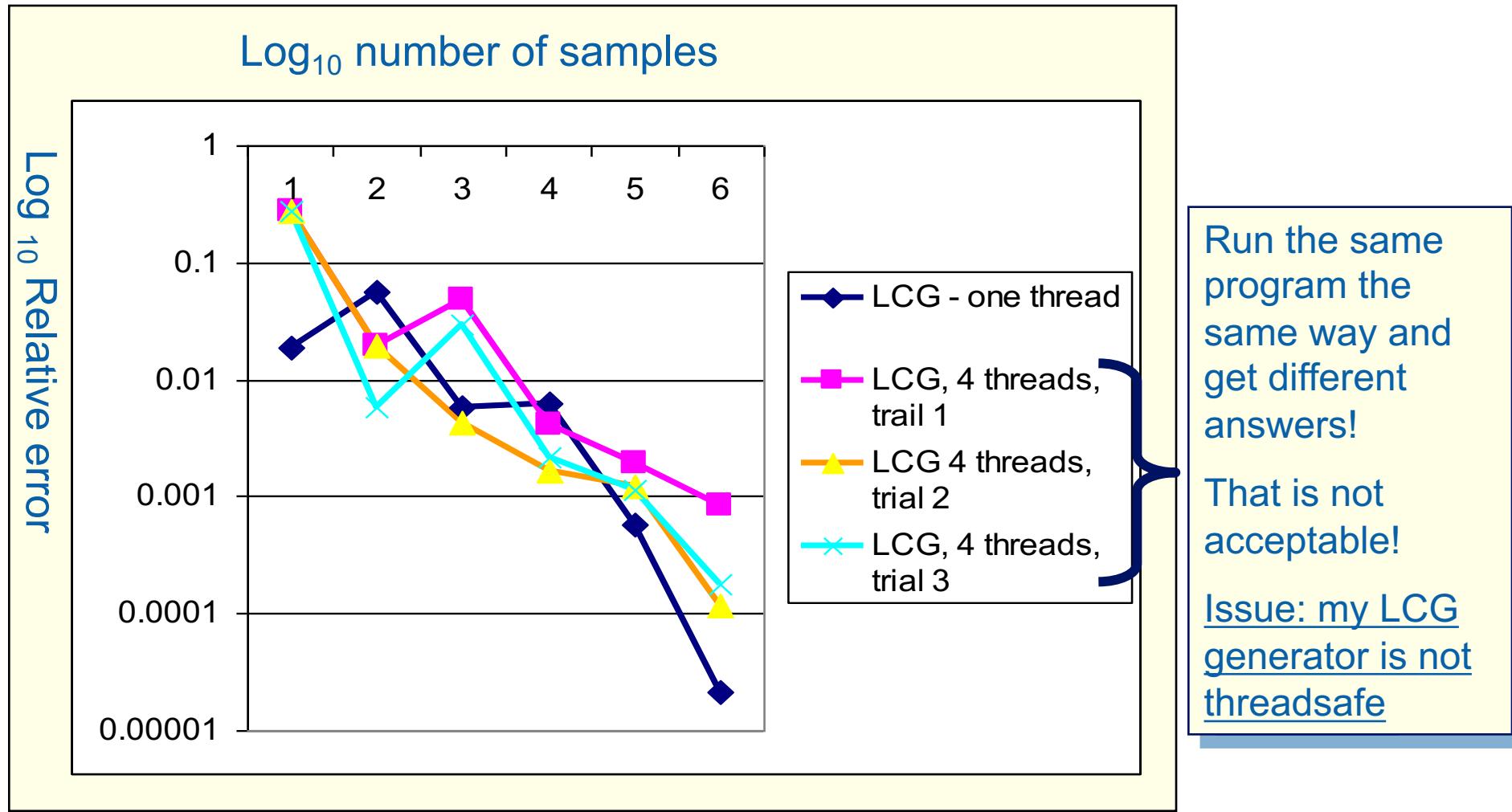
```
static long MULTIPLIER = 1366;
static long ADDEND    = 150889;
static long PMOD      = 714025;
long random_last = 0;
double random ()
{
    long random_next;

    random_next = (MULTIPLIER * random_last + ADDEND)% PMOD;
    random_last = random_next;

    return ((double)random_next/(double)PMOD);
}
```

Seed the pseudo random sequence by setting random\_last

# Running the PI\_MC program with LCG generator



Program written using the Intel C/C++ compiler (10.0.659.2005) in Microsoft Visual studio 2005 (8.0.50727.42) and running on a dual-core laptop (Intel T2400 @ 1.83 Ghz with 2 GB RAM) running Microsoft Windows XP.

# Exercise: Monte Carlo pi (cont)

- Create a threadsafe version of the monte carlo pi program
- Do not change the interfaces to functions in random.c
  - This is an exercise in modular software ... why should a user of your parallel random number generator have to know any details of the generator or make any changes to how the generator is called?
  - The random number generator must be thread-safe
- Verify that the program is thread safe by running multiple times for a fixed number of threads.
- Any concerns with the program behavior?

# LCG code: threadsafe version

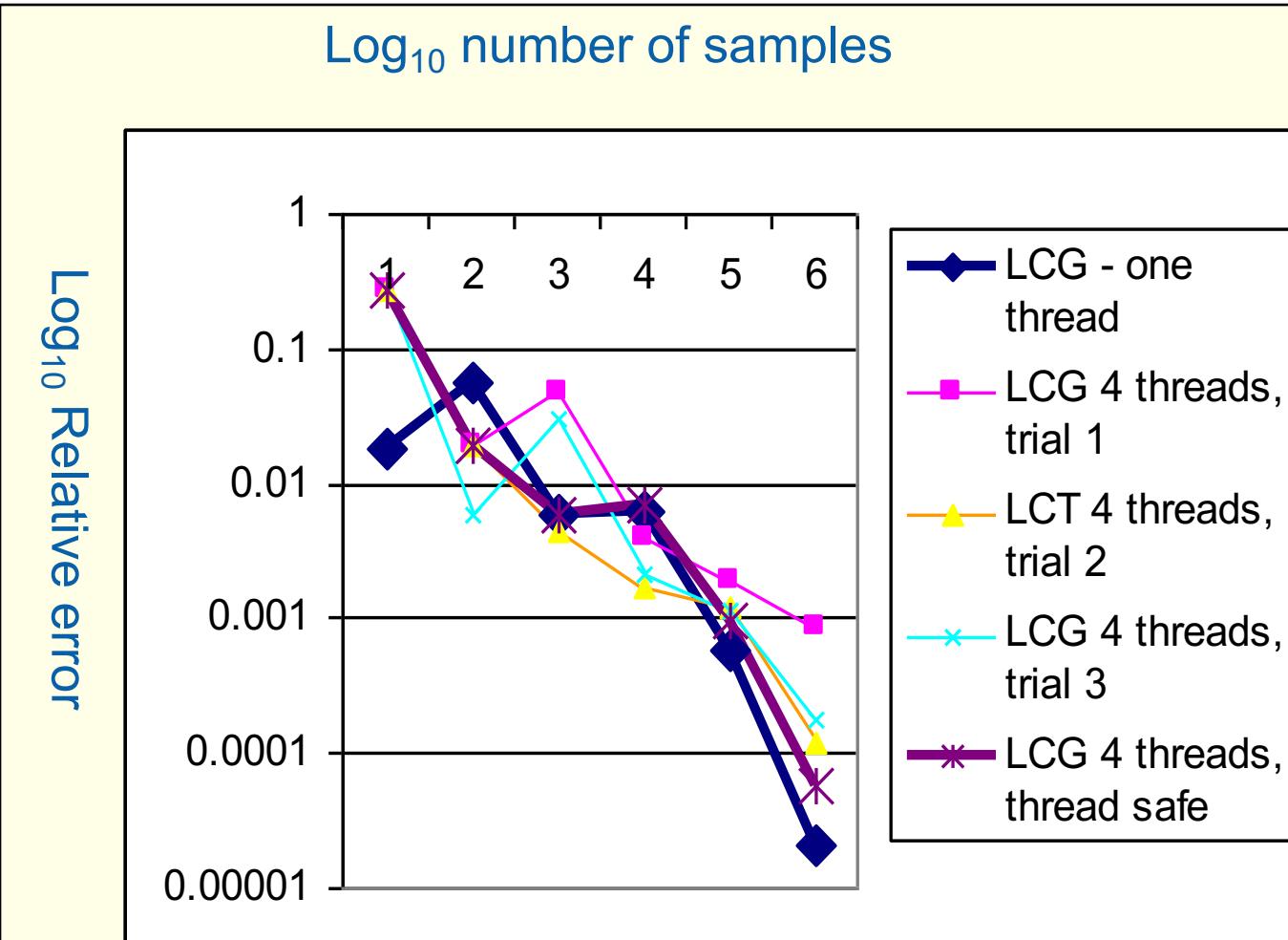
```
static long MULTIPLIER = 1366;
static long ADDEND    = 150889;
static long PMOD      = 714025;
long random_last = 0;
#pragma omp threadprivate(random_last)
double random ()
{
    long random_next;

    random_next = (MULTIPLIER * random_last + ADDEND)% PMOD;
    random_last = random_next;

    return ((double)random_next/(double)PMOD);
}
```

random\_last carries state between random number computations, To make the generator threadsafe, make random\_last threadprivate so each thread has its own copy.

# Thread Safe Random Number Generators



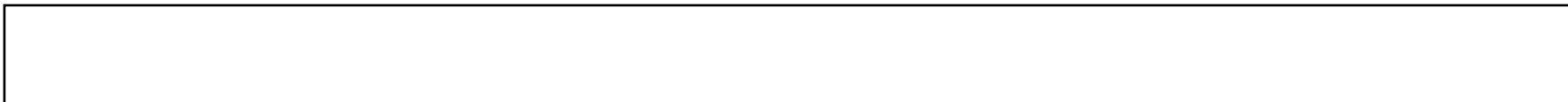
Thread safe version gives the same answer each time you run the program.

But for large number of samples, its quality is lower than the one thread result!

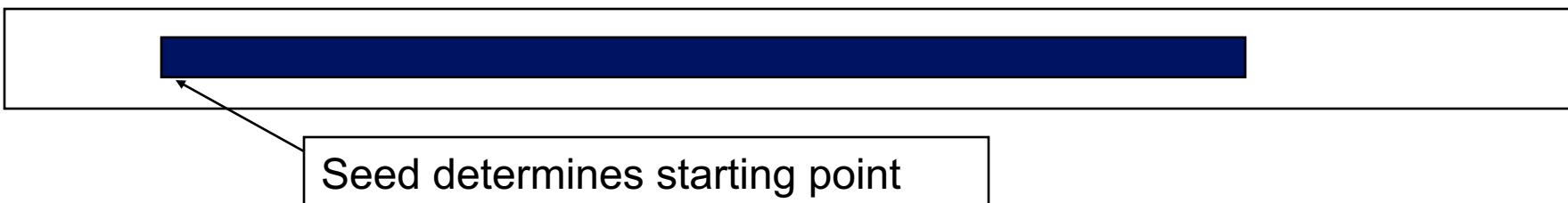
Why?

# Pseudo Random Sequences

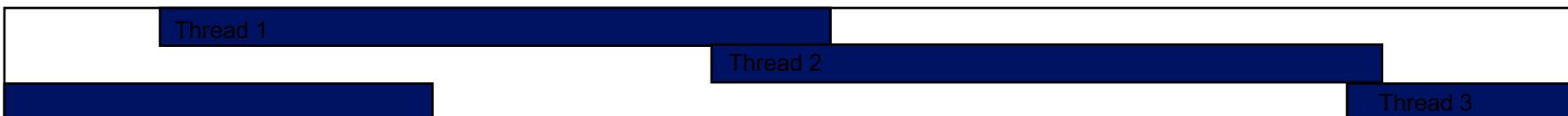
- Random number Generators (RNGs) define a sequence of pseudo-random numbers of length equal to the period of the RNG



- In a typical problem, you grab a subsequence of the RNG range



- Grab arbitrary seeds and you may generate overlapping sequences
  - ◆ E.g. three sequences ... last one wraps at the end of the RNG period.



- Overlapping sequences = over-sampling and bad statistics ... lower quality or even wrong answers!

# Parallel random number generators

- Multiple threads cooperate to generate and use random numbers.
- Solutions:
  - Replicate and Pray
  - Give each thread a separate, independent generator
  - Have one thread generate all the numbers.
  - Leapfrog ... deal out sequence values “round robin” as if dealing a deck of cards.
  - Block method ... pick your seed so each threads gets a distinct contiguous block.
- Other than “replicate and pray”, these are difficult to implement. Be smart ... get a math library that does it right.

If done right, can generate the same sequence regardless of the number of threads ...

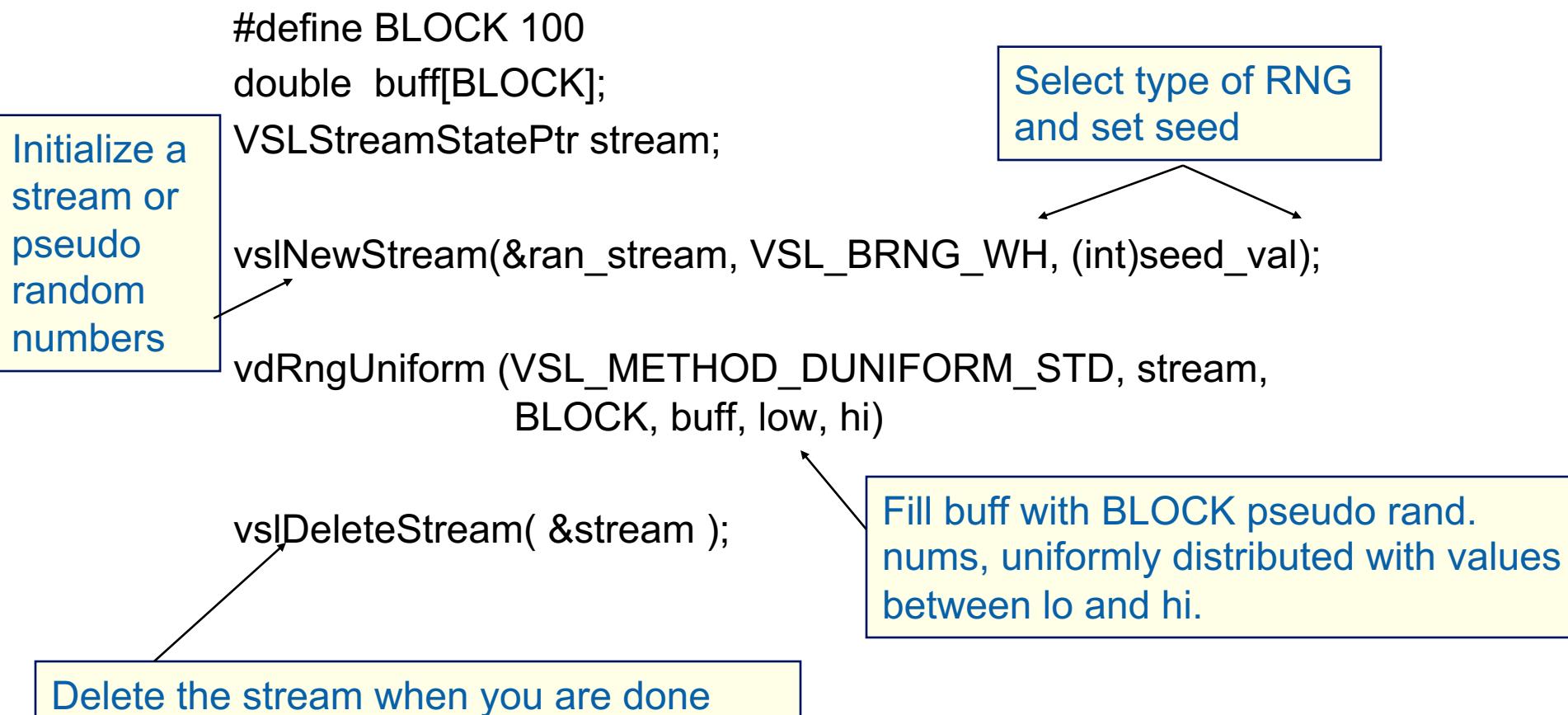
Nice for debugging, but not really needed scientifically.

Intel's Math kernel Library supports a wide range of parallel random number generators.

For an open alternative, the state of the art is the Scalable Parallel Random Number Generators Library (SPRNG): <http://www.sprng.org/> from Michael Mascagni's group at Florida State University.

# MKL Random Number Generators (RNG)

- MKL includes several families of RNGs in its vector statistics library.
- Specialized to efficiently generate vectors of random numbers

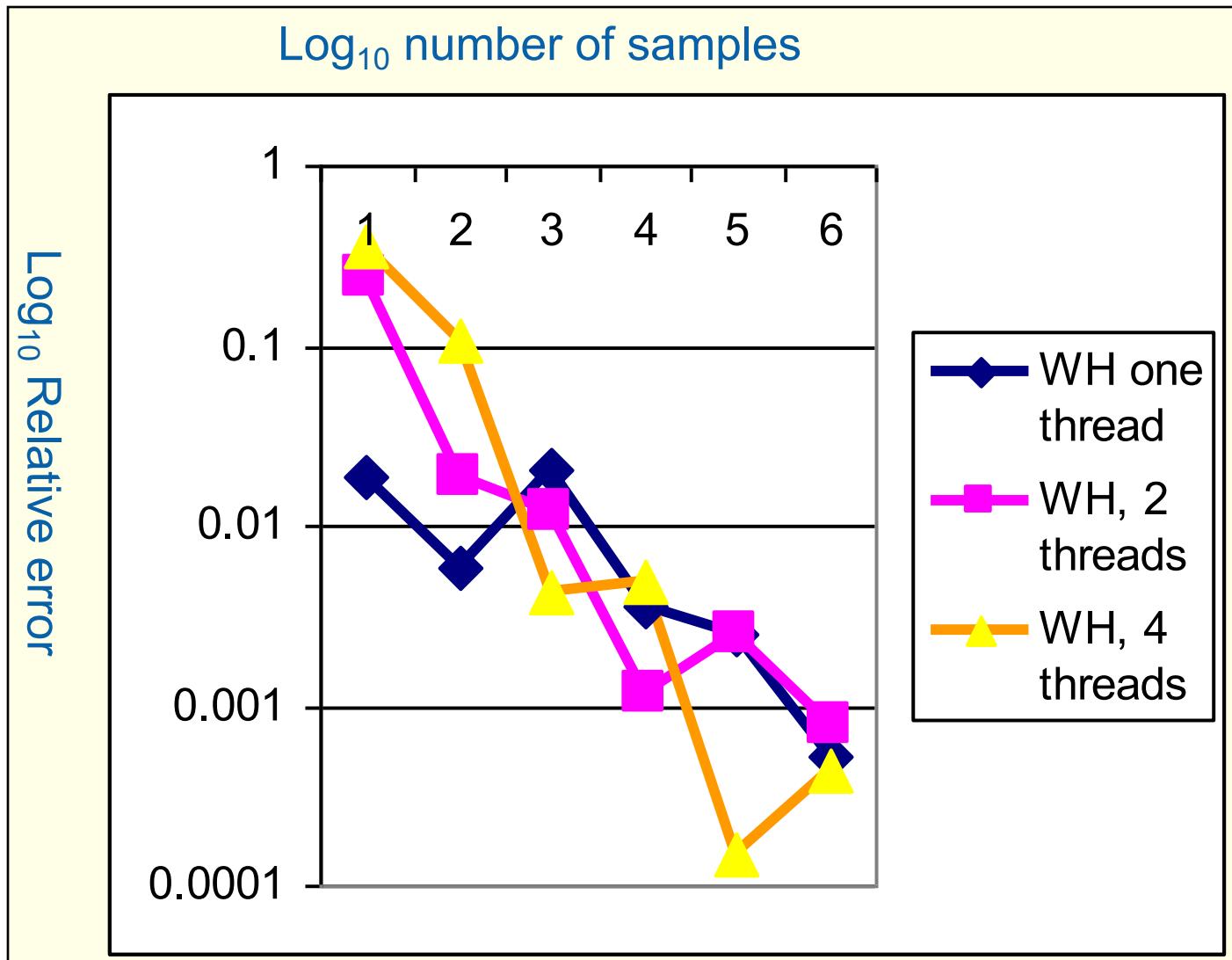


# Wichmann-Hill Generators (WH)

- WH is a family of 273 parameter sets each defining a non-overlapping and independent RNG.
- Easy to use, just make each stream threadprivate and initiate RNG stream so each thread gets a unique WG RNG.

```
VSLStreamStatePtr stream;  
#pragma omp threadprivate(stream)  
  
....  
  
vslNewStream(&ran_stream, VSL_BRNG_WH+Thrd_ID, (int)seed);
```

# Independent Generator for each thread



Notice that once you get beyond the high error, small sample count range, adding threads doesn't decrease quality of random sampling.

# Leap Frog Method

- Interleave samples in the sequence of pseudo random numbers:
  - Thread  $i$  starts at the  $i^{\text{th}}$  number in the sequence
  - Stride through sequence, stride length = number of threads.
- Result ... the same sequence of values regardless of the number of threads.

```
#pragma omp single
{  nthreads = omp_get_num_threads();
   iseed = PMOD/MULTIPLIER;    // just pick a seed
   pseed[0] = iseed;
   mult_n = MULTIPLIER;
   for (i = 1; i < nthreads; ++i)
   {
      iseed = (unsigned long long)((MULTIPLIER * iseed) % PMOD);
      pseed[i] = iseed;
      mult_n = (mult_n * MULTIPLIER) % PMOD;
   }
}
random_last = (unsigned long long) pseed[id];
```

One thread computes offsets and strided multiplier

LCG with Addend = 0 just to keep things simple

Each thread stores offset starting point into its threadprivate “last random” value

# Same sequence with many threads.

- We can use the leapfrog method to generate the same answer for any number of threads

| <b>Steps</b> | <b>One thread</b> | <b>2 threads</b> | <b>4 threads</b> |
|--------------|-------------------|------------------|------------------|
| 1000         | 3.156             | 3.156            | 3.156            |
| 10000        | 3.1168            | 3.1168           | 3.1168           |
| 100000       | 3.13964           | 3.13964          | 3.13964          |
| 1000000      | 3.140348          | 3.140348         | 3.140348         |
| 10000000     | 3.141658          | 3.141658         | 3.141658         |

Used the MKL library with two generator streams per computation: one for the x values (WH) and one for the y values (WH+1). Also used the leapfrog method to deal out iterations among threads.



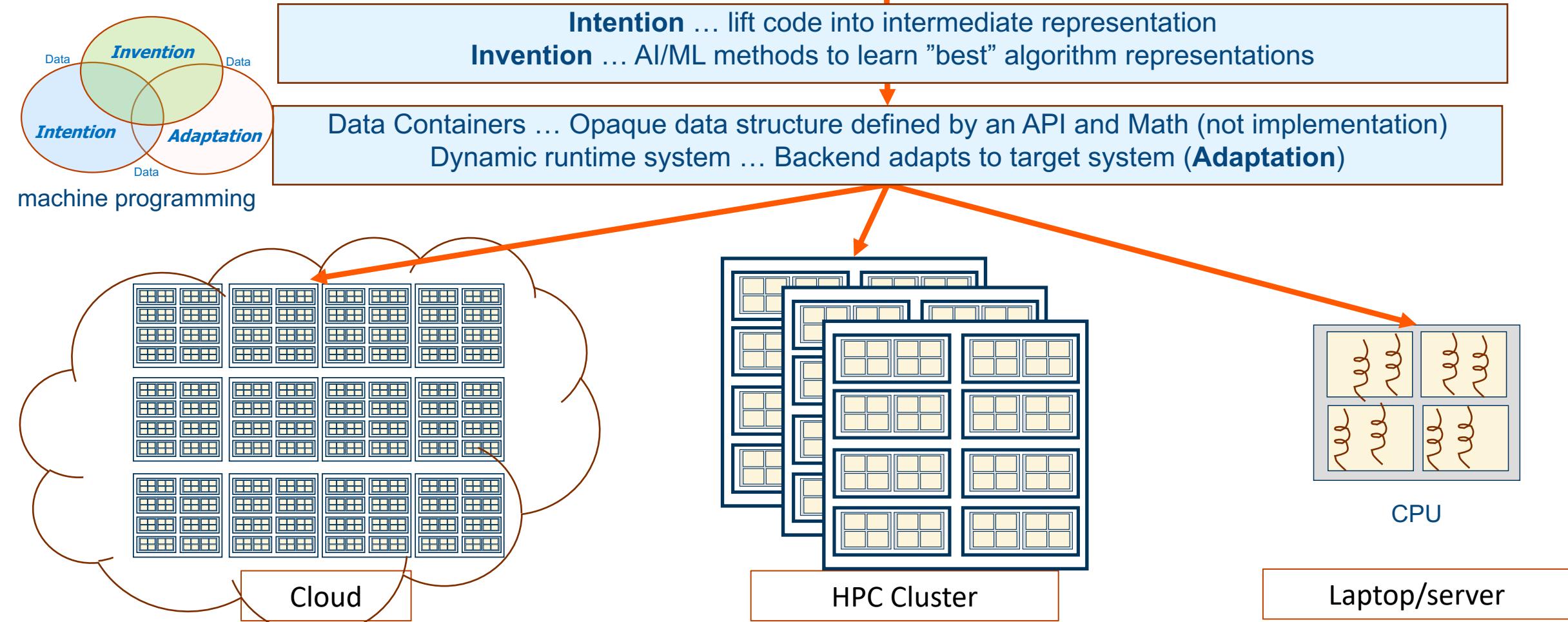


# Camelot: the software perspective

Compelling endpoint

One source code mapped and optimized onto  
any hardware backend ... including cloud native

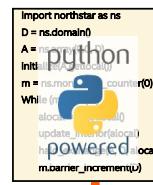
Scalability limited by the  
algorithm and the hardware



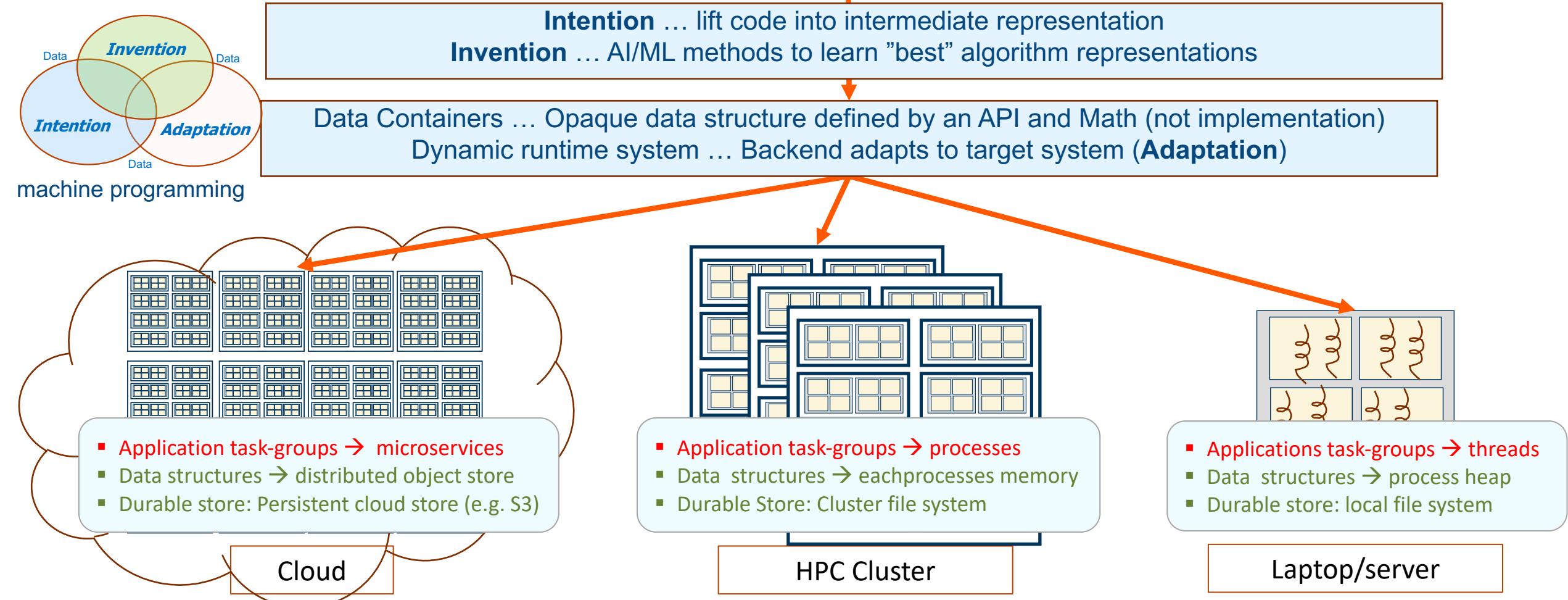
# Camelot: the software perspective

Compelling endpoint

One source code mapped and optimized onto  
any hardware backend ... including cloud native



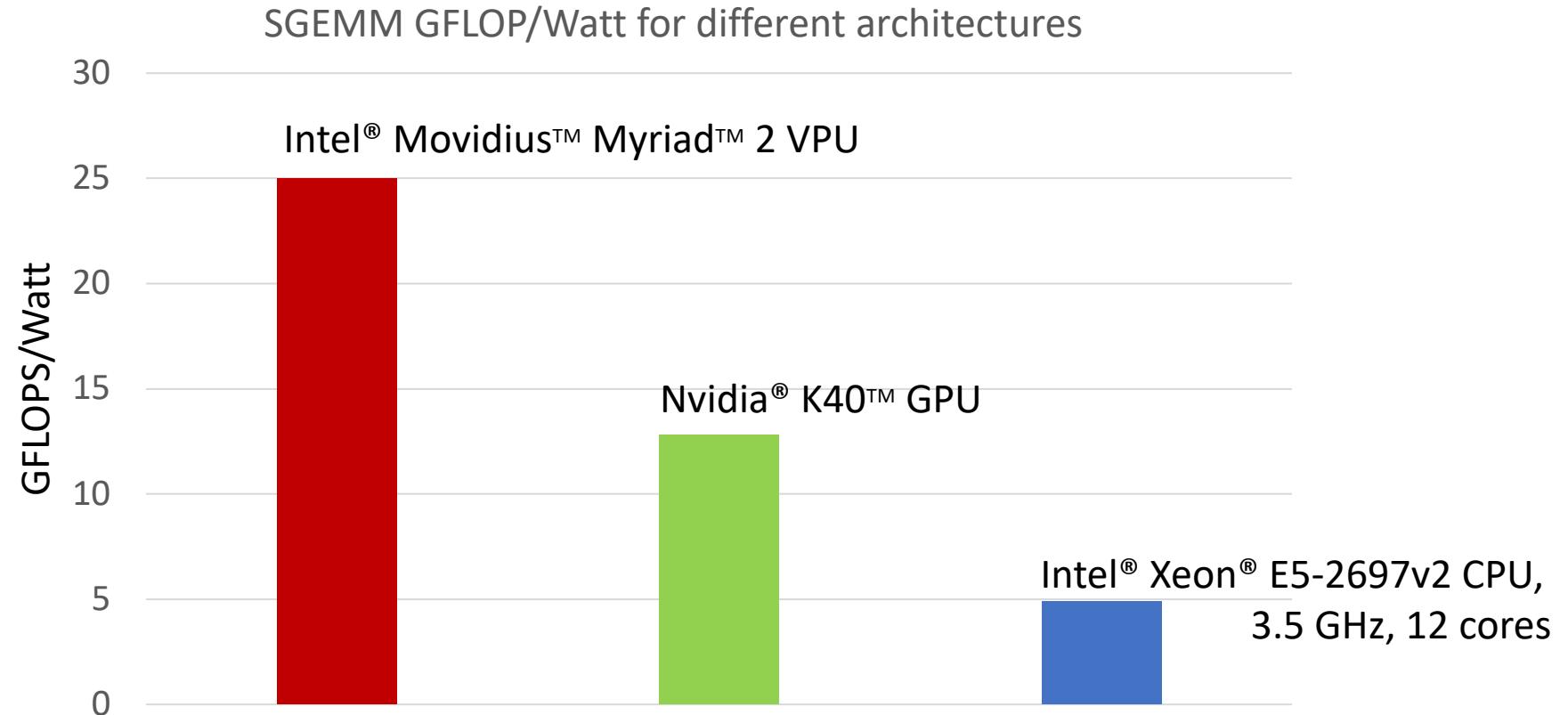
Scalability limited by the  
algorithm and the hardware





# If you care about power, the world is heterogeneous?

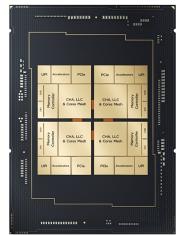
Specialized  
processors doing  
operations suited to  
their architecture  
are more efficient  
than general  
purpose processors.



Hence, future systems will be increasingly heterogeneous ... GPUs, CPUs,  
FPGAs, and a wide range of accelerators

# A New Golden Age for Computer Architecture

## Products



Novel On-Die Accelerators



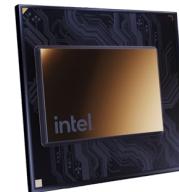
GPU/Data Parallel



Spatial/  
Dataflow



Deep Learning  
Optimized

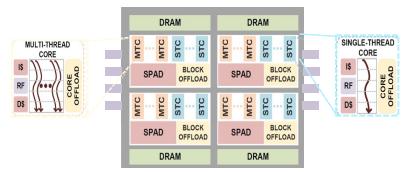


Blockchain

## Research



Neuromorphic



Graph Analytics

“Strong growth in demand for data center services continues to be mostly offset by ongoing efficiency improvements...”

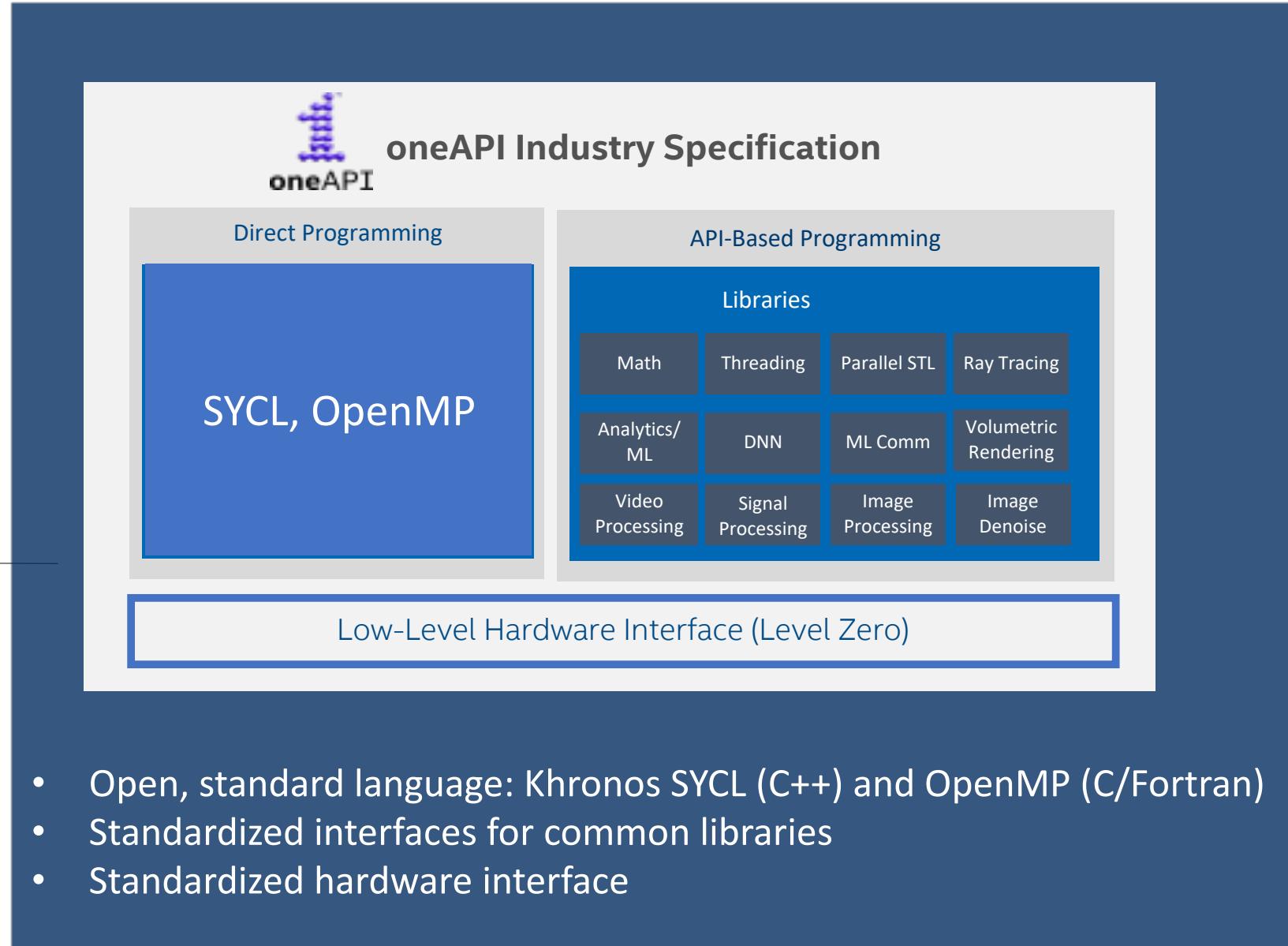
# oneAPI: One Name, Two Distinct Objectives



- Open industry specification
- Open-source repo and development
- Community driven
- Supports multi-vendor implementation



- Intel's implementation
- Toolkits optimized for Intel HW
- Available for free download



# Why SYCL (if you are a C++ programmer)?

---

```
sycl::queue q(cpu_selector{});
```

```
auto A = sycl::malloc_shared<float>(n, q);
auto B = sycl::malloc_shared<float>(n, q);
```

```
q.parallel_for( sycl::range<1>{n},
    [=] (sycl::id<1> i) {
        B[i] += A[i] * A[i];
    }
).wait();
```

```
sycl::queue q(gpu_selector{});
```

```
auto A = sycl::malloc_shared<float>(n, q);
auto B = sycl::malloc_shared<float>(n, q);
```

```
q.parallel_for( sycl::range<1>{n},
    [=] (sycl::id<1> i) {
        B[i] += A[i] * A[i];
    }
).wait();
```

- 
- Standard C++17 aids time to developer productivity
  - Syntax for accelerators (device selection, offload, memory transfer)
  - Unified shared memory
  - Single source (host and device code)
  - Multi-Architecture (CPU, GPU, FPGA, and other targets)
  - Stack based on standards and open specifications (CLANG, LLVM, SPIR-V, Level Zero)

# Why OpenMP (if you are a C or Fortran programmer)?

---

```
// CPU  
float A[n];  
float B[n];  
  
#pragma omp parallel for  
for(int i=0; i<n; i++)  
    B[i] += A[i] * A[i];
```

```
//GPU  
float A[n];  
float B[n];  
  
#pragma omp target teams loop  
for(int i=0; i<n; i++)  
    B[i] += A[i] * A[i];
```

- 
- Standard C and Fortran aids time to developer productivity
  - Syntax for accelerators (device selection, offload, memory transfer)
  - Unified shared memory
  - Single source (host and device code)
  - Multi-Architecture (CPU, GPU)
  - Stack based on standards and open specifications (CLANG, LLVM, SPIR-V, Level Zero)

# Why PyOMP (if you are a Python programmer)?

---

```
From numba import njit
From numba.openmp import openmp_context as openmp

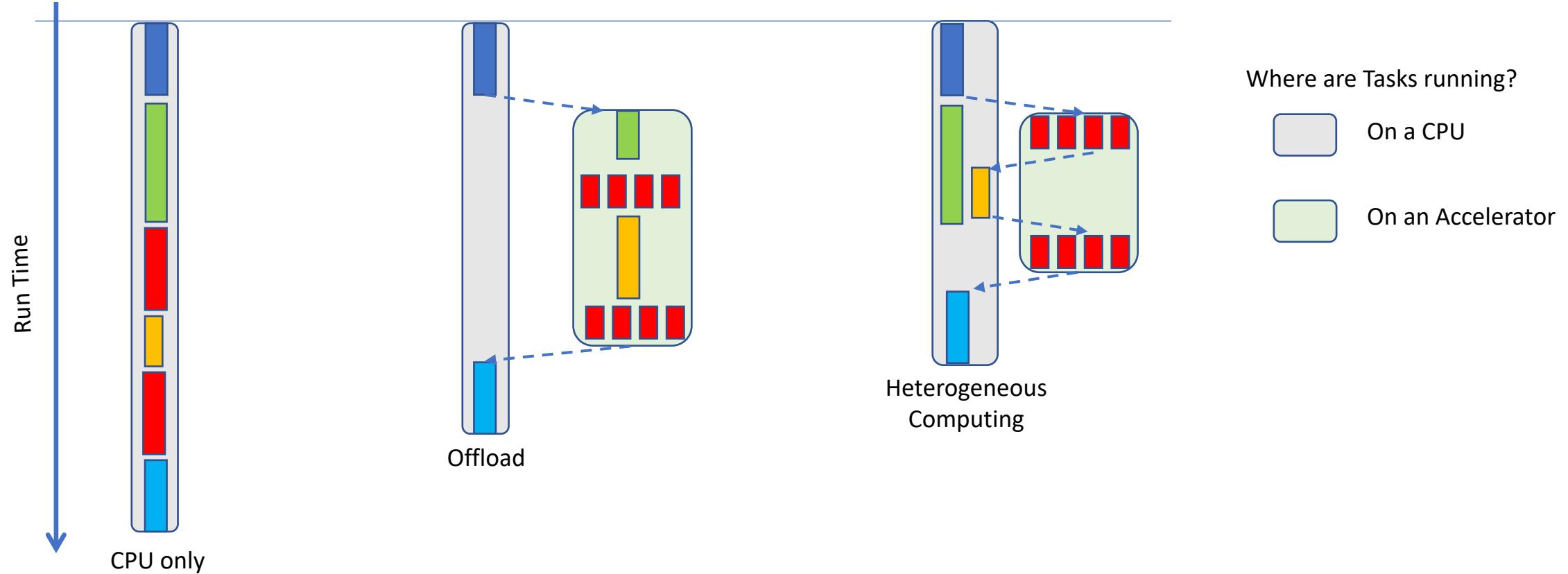
@njit
def vsqr(n, A, B):
    with openmp("parallel for"):
        for i in range(n):
            B[i] = A[i] * A[i]

    return
```

- 
- Uses numba JIT compiler to get around the global interpreter lock (enables parallel multithreading)
  - Other than one-time JIT cost, performance approaches that from C and OpenMP
  - CPU only today ... GPU support expected later this year
  - Stack based on standards and open specifications (Python ecosystems including numba and LLVM)

# Offload vs. Heterogeneous computing

- **Offload:** The CPU moves work to an accelerator and waits for the answer.
- **Heterogeneous Computing:** Run sub-problems in parallel on the hardware best suited to them.



# Example: Single-cell RNA-Seq benchmark (SCANPY)

- SCANPY ... a widely used tool for studying gene expression. All data are elapsed time in seconds
- We started with results from an Nvidia blog (Example 2 from [link](#)), optimized code for one socket of Intel® Xeon® 8380 CPU and then “simulated” heterogeneous computing result by taking the faster of CPU and GPU execution times.

| Pipeline stages              | 64 vCPUs<br>n1-highmem-64<br>(off-the-shelf Python) | A100 40Gb<br>(Clara Parabricks) | ICX-1s, 40 cores<br><b>(optimized by Intel)</b> | "Simulated"<br>heterogeneous<br>A100 & ICS-1s 40 cores |
|------------------------------|-----------------------------------------------------|---------------------------------|-------------------------------------------------|--------------------------------------------------------|
| Data Loading & Preprocessing | 1120                                                | 475                             | 15.7                                            | 15.7                                                   |
| PCA                          | 44                                                  | 17.8                            | 5.0                                             | 5.0                                                    |
| T-SNE                        | 6509                                                | 37                              | 205.6                                           | 37                                                     |
| K-means (single iteration)   | 148                                                 | 2                               | 7.1                                             | 7.1                                                    |
| KNN                          | 154                                                 | 62                              | 59.8                                            | 59.8                                                   |
| UMAP                         | 2571                                                | 21                              | 84.5                                            | 84.5                                                   |
| Louvain clustering           | 1153                                                | 2.4                             | 6.0                                             | 6.0                                                    |
| Leiden clustering            | 6345                                                | 1.7                             | 28.4                                            | 28.4                                                   |
| Reanalysis of subgroup       | 255                                                 | 17.9                            | 22.5                                            | 22.5                                                   |
| Rest                         | 39                                                  | 49.2                            | 49.0                                            | 49.0                                                   |
| <b>End-to-End runtime</b>    | <b>18338</b>                                        | <b>686</b>                      | <b>483.6</b>                                    | <b>211.5</b>                                           |

Clara Parabricks: Nvidia solution stack built on RAPIDS for healthcare applications

<https://github.com/clara-parabricks/rapids-single-cell-examples>

github repository as of Dec 16, 2020

## Lessons learned:

- Be careful comparing unoptimized python to hand-tuned CUDA code
- GPUs are great. So are CPUs if you fully utilize all the cores and vector units.
- What you really want is the best of both worlds. **You want heterogeneous computing!**

See Backup for workloads and configurations. Results may vary.

This column shows the potential of heterogeneous computing. We ignored extra communication and synchronization overhead, so actual runtimes would be slightly greater.

Source: Github repository as of Dec 16, 2020 - Example 2: Single-cell RNA-seq of 1.3 Million Mouse Brain Cells comparing CPU (n1-highmem-64 64 vCPUs) vs GPU (n1-highmem-16). <https://github.com/clara-parabricks/rapids-single-cell-examples>. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

1S Ice Lake: See Backup for workloads and configurations. Results may vary.

Third party names are the property of their owners



# Five Epochs of Distributed Computing\*

| Epoch starting date | Defining limitations                      | Application                              | Interaction time and Network performance | Capability           |
|---------------------|-------------------------------------------|------------------------------------------|------------------------------------------|----------------------|
| First<br>1970       | Rare connections to expensive computers   | FTP, telnet, email                       | 100 ms<br>Low bandwidth high latency     | People to computers  |
| Second<br>1984      | I/O wall, disks can't keep up             | RPC,<br>Client Server                    | 10 ms<br>10 mbps                         | Computer to computer |
| Third<br>1990       | Networking wall                           | MPP HPC, three-tier datacenter networks  | 1 ms<br>100 mbs → 1 Gbs                  | Services to services |
| Fourth<br>2000      | Dennard scaling wall ... per core plateau | Web search, planet-scale services        | 100 $\mu$ s<br>10 Gbps flash             | People to people     |
| Fifth<br>2015       | Per socket wall ... accelerators take off | Machine Learning, data centric computing | 10 $\mu$ s<br>200 Gbps → 1 Tbps          | People to insights   |

\*The five Epochs of distributed computing, Amin Vahdat of Google: SIGCOMM Lifetime achievement award keynote, 2020.

# The Eight Fallacies of Distributed Computing

(Peter Deutsch of Sun Microsystems, 1994 ... item 8 added in 1997 by James Gosling)

Essentially everyone, when they first build a distributed application, makes the following eight assumptions. All prove to be false in the long run and all cause *big* trouble and *painful* learning experiences.

1. The network is reliable
2. Latency is zero
3. Bandwidth is infinite
4. The network is secure
5. Topology doesn't change
6. There is one administrator
7. Transport cost is zero
8. The network is homogeneous

# The Eight Fallacies of Distributed Computing

(Peter Deutsch of Sun Microsystems, 1994 ... item 8 added in 1997 by James Gosling)

Essentially everyone, when they first build a distributed application, makes the following eight assumptions. All prove to be false in the long run and all cause *big* trouble and *painful* learning experiences.

1. The network is reliable
2. Latency is **low and fixed**
3. Bandwidth is **high and fixed**
4. The network is secure
5. Topology doesn't change
6. There is one administrator
7. Transport cost is **negligible**
8. The network is homogeneous

# The Eight Fallacies of Distributed Computing

(Peter Deutsch of Sun Microsystems, 1994 ... item 8 added in 1997 by James Gosling)

Essentially everyone, when they first build a distributed application, makes the following eight assumptions. All prove to be false in the long run and all cause *big* trouble and *painful* learning experiences.

## Cloud

- ~~X. The network is reliable~~
- ~~X. Latency is **low and fixed**~~
- ~~X. Bandwidth is **high and fixed**~~
- ~~X. The network is secure~~
- ~~X. Topology doesn't change~~
- ~~X. There is one administrator~~
- ~~X. Transport cost is **negligible**~~
- ~~X. The network is homogeneous~~

## HPC Cluster

- ✓. The network is reliable**
- ✓. Latency is **low and fixed****
- ✓. Bandwidth is **high and fixed****
- ✓. The network is secure**
- ✓. Topology doesn't change**
- ✓. There is one administrator**
- ~~X. Transport cost is **negligible**~~
- ✓. The network is homogeneous**

# The three domains of parallel programming

|                           |                      |                                                                |                                                                                      |                                                                            |
|---------------------------|----------------------|----------------------------------------------------------------|--------------------------------------------------------------------------------------|----------------------------------------------------------------------------|
| Platform*                 | Laptop or server     | HPC Cluster                                                    |  | Cloud                                                                      |
| Execution Agent           | Threads              | Processes                                                      |   | Microservices                                                              |
| Memory                    | Single Address Space | Distributed memory, local memory owned by individual processes |   | Distributed object store (in memory) backed by a persistent storage system |
| Typical Execution Pattern | Fork-join            | SPMD                                                           |  | Event driven tasks, FaaS, and Actors                                       |

Laptop/server and cluster models work well together.

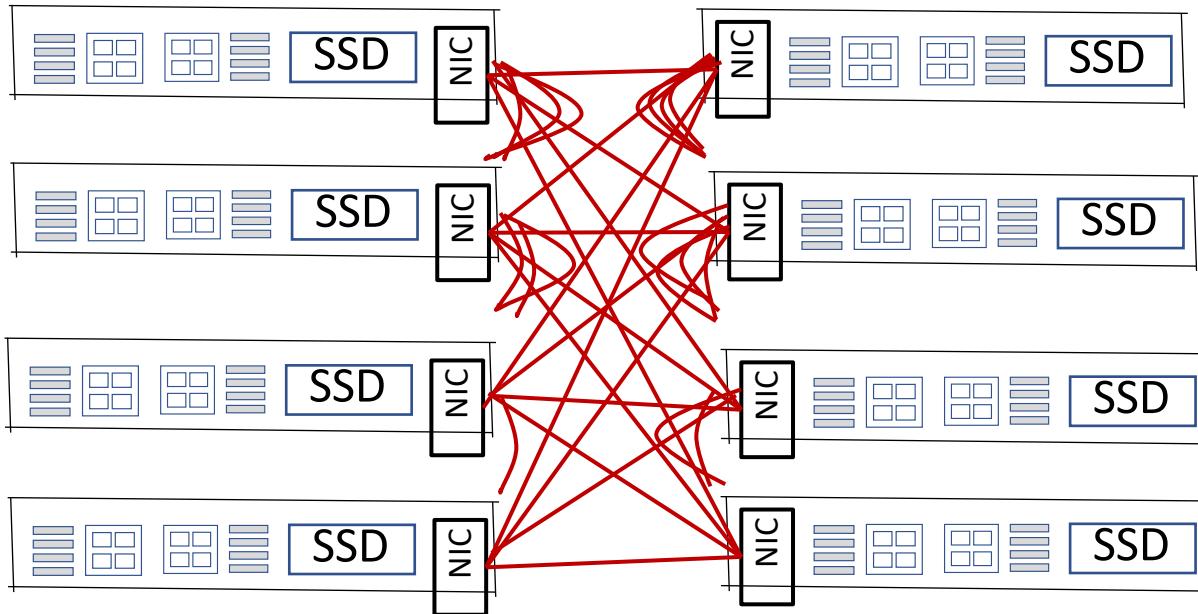
An impenetrable wall separates them from the cloud-native world

# The sixth Epoch of Distributed Computing

| Epoch starting date | Defining limitations                      | Application                                                              | Interaction time and Network performance | Capability            |
|---------------------|-------------------------------------------|--------------------------------------------------------------------------|------------------------------------------|-----------------------|
| First<br>1970       | Rare connections to expensive computers   | FTP, telnet, email                                                       | 100 ms<br>Low bandwidth high latency     | People to computers   |
| Second<br>1984      | I/O wall, disks can't keep up             | RPC,<br>Client Server                                                    | 10 ms<br>10 mbps                         | Computer to computer  |
| Third<br>1990       | Networking wall                           | MPP HPC, three-tier datacenter networks                                  | 1 ms<br>100 mbs → 1 Gbs                  | Services to services  |
| Fourth<br>2000      | Dennard Scaling Wall ... per core plateau | Web search, planet-scale services                                        | 100 $\mu$ s<br>10 Gbps<br>flash          | People to people      |
| Fifth<br>2015       | Per socket wall ... accelerators take off | Machine Learning, data centric computing                                 | 10 $\mu$ s<br>200 Gbps → 1 Tbps          | People to insights    |
| Sixth<br>2025       | Speed of light                            | Dynamic, real-time AI, integrated from data-center to the edge with SDE* | 100 ns<br>10 Tbs                         | People to experiences |

\* SDE: Software defined Everything, i.e. software defined networking, software defined infrastructure, software defined servers ... All at the same time ... to dynamically construct systems to meet the needs of workloads.

# Networking technology... replace generic data center network with a cluster of cliques



A clique: A graph where every vertex is connected to every other vertex

A Clique: a network of diameter one with

$O(\frac{1}{4}N^2)$  bisection bandwidth

Combine with next generation optical networks to hit latencies of 100 ns

# Latencies every engineer should know ...

L1 cache reference 1.5 ns

L2 cache reference 5 ns

Branch misprediction 6 ns

Uncontented mutex lock/unlock 20 ns

L3 cache reference 25 ns

Main memory reference 100 ns

“Far memory”/Fast NVM reference 1,000 ns (1us)

Read 1 MB sequentially from memory 12,000 ns (12 us)

SSD Random Read 100,000 ns (100 us)

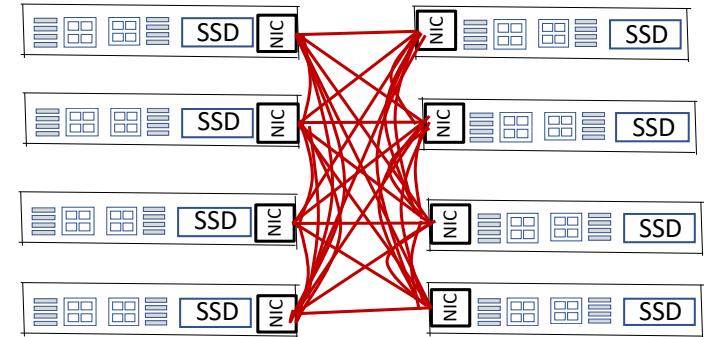
Read 1 MB bytes sequentially from SSD 500,000 ns (500 us)

Read 1 MB sequentially from 10Gbps network 1,000,000 ns (1 ms)

Read 1 MB sequentially from disk 10,000,000 ns (10 ms)

Disk seek 10,000,000 ns (10 ms)

Send packet California → Netherlands → California (150 ms)



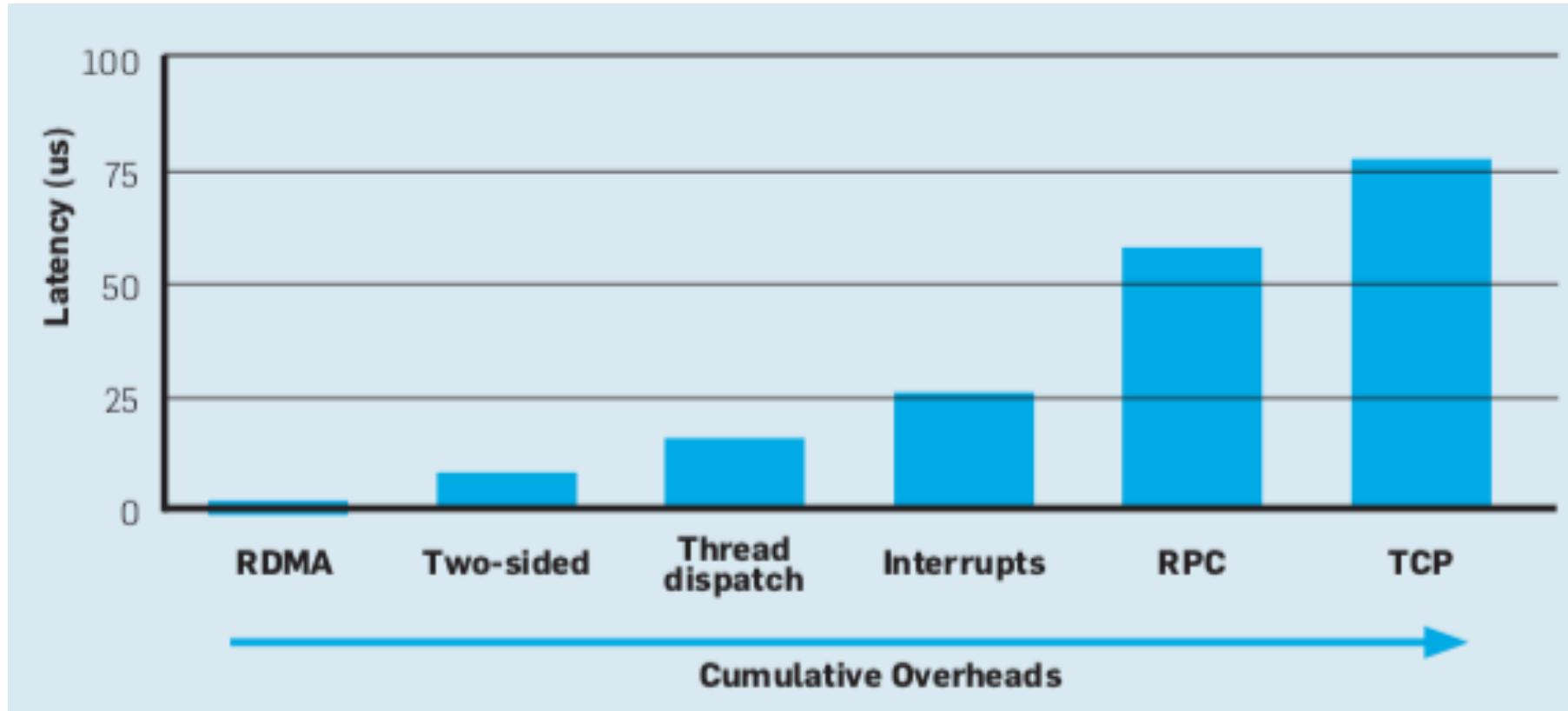
A cluster of nodes with a Clique network topology and low latency optical network...

Yields one hop network latencies on par with DRAM access latencies.

Source: **The Datacenter as a Computer: Designing Warehouse-Scale Machines**, Luiz Andre Barroso, Urs Holzle, Parthasarathy Ranganathan, 3<sup>rd</sup> edition, Morgan & Claypool, 2019.

# Take out the big stuff & you're left with lots of $\mu$ s overheads

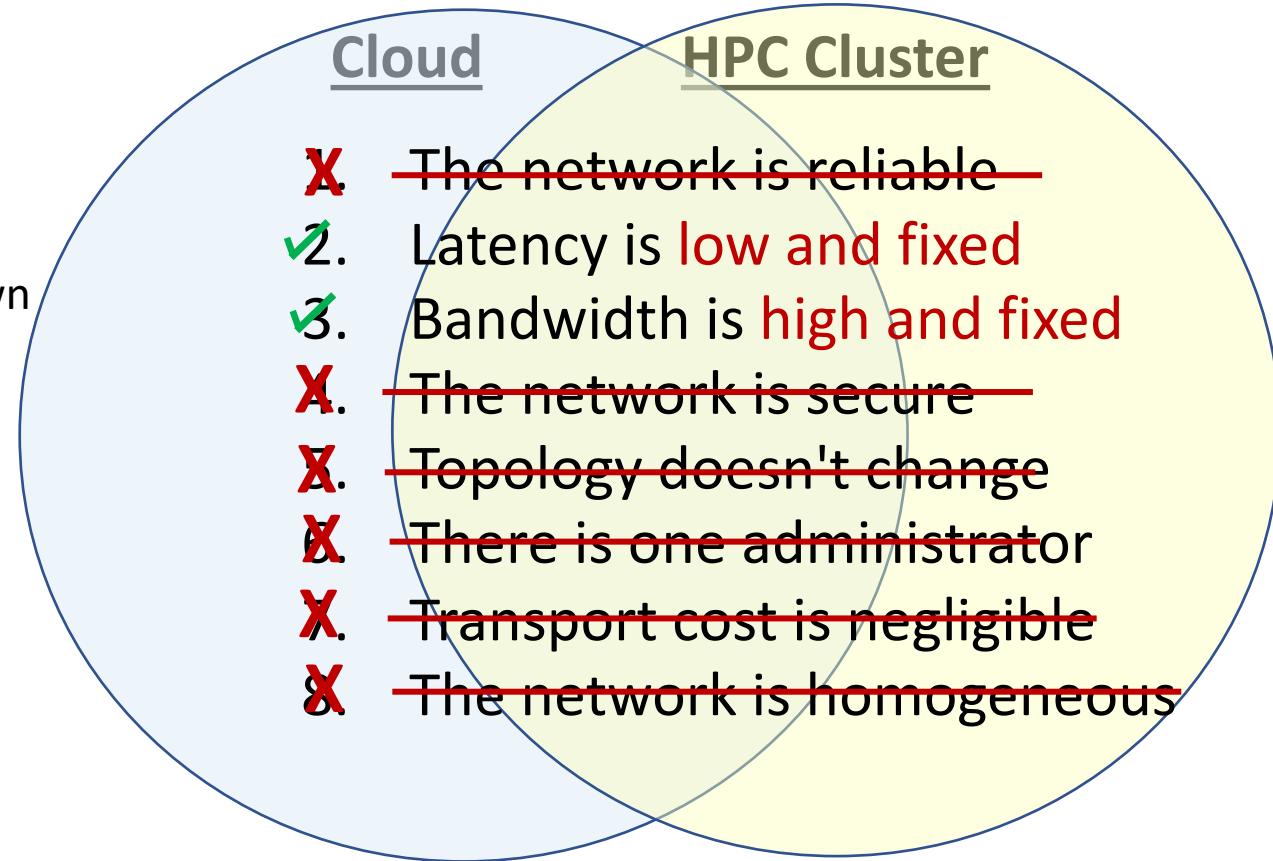
All those SW overheads add up ... like bricks that combine to build a networking-wall ...  
turning a 2  $\mu$ s network into a 100  $\mu$ s network...



Computer Scientists need to rethink system SW stacks to minimize latencies ... fast RDMA, reduce sync contention, low latency interrupt handlers, and more .... All to hit  $O(\mu\text{s})$  latencies.

# In the sixth Epoch of Distributed Computing, cloud and cluster overlap ... or even merge!

Chip-to-chip optical networks push latency down and bandwidth up



Data Streaming Accelerator reduces tail latency.

P4/P5/P6 + Infrastructure Processing Units drive down latency and reduces jitter

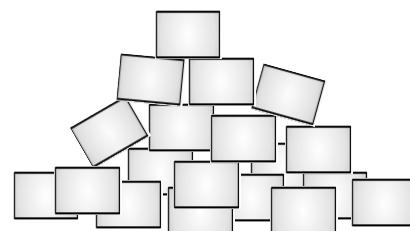
With Low Latencies, high bandwidths and stable performance, we can do loosely synchronous and synchronous applications in the cloud. The economics of the cloud vs dedicated HPC clusters means the cloud will dominate HPC

HPC applications will need to change to deal with reliability and network inhomogeneities.

# The three domains of parallel programming

| Platform*                 | Laptop or server     | HPC Cluster                                                    | Cloud                                                                      |
|---------------------------|----------------------|----------------------------------------------------------------|----------------------------------------------------------------------------|
| Execution Agent           | Threads              | Processes                                                      | Microservices                                                              |
| Memory                    | Single Address Space | Distributed memory, local memory owned by individual processes | Distributed object store (in memory) backed by a persistent storage system |
| Typical Execution Pattern | Fork-join            | SPMD                                                           | Event driven tasks, FaaS, and Actors                                       |

Advances in networking technology plus low-overhead software stacks optimized to reduce tail-latency will shatter this wall



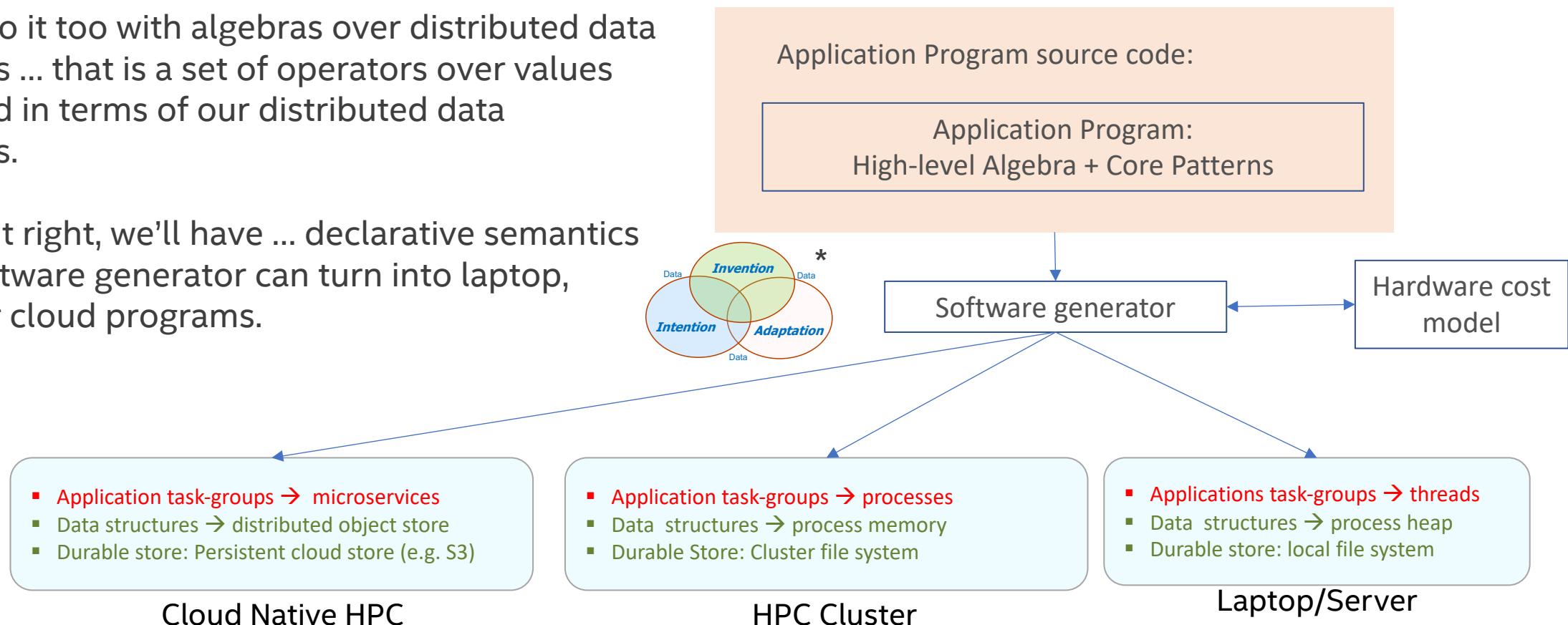
# The three domains of parallel programming

|                           |                      |                                                                |                                                                            |
|---------------------------|----------------------|----------------------------------------------------------------|----------------------------------------------------------------------------|
| Platform*                 | Laptop or server     | HPC Cluster                                                    | Cloud                                                                      |
| Execution Agent           | Threads              | Processes                                                      | Microservices                                                              |
| Memory                    | Single Address Space | Distributed memory, local memory owned by individual processes | Distributed object store (in memory) backed by a persistent storage system |
| Typical Execution Pattern | Fork-join            | SPMD                                                           | Event driven tasks, FaaS, and Actors                                       |

There will always be a need for top-end scalable systems in supercomputer centers, but economics will push the bulk of scientific computing into the cloud.

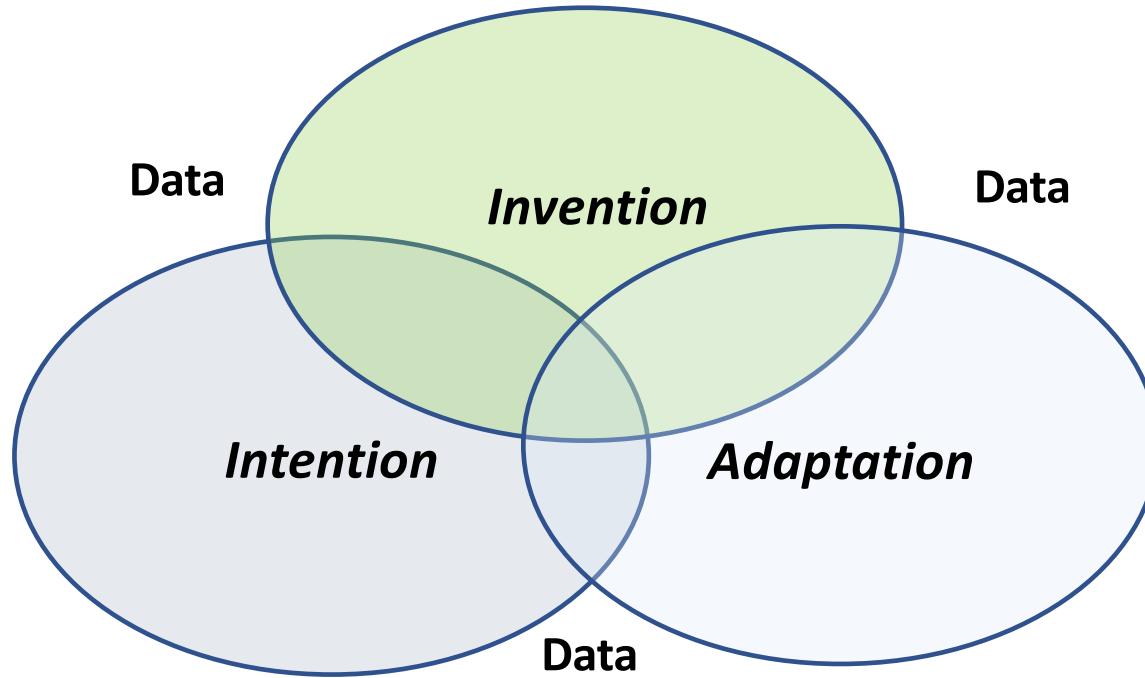
# One codebase → many systems

- Performance, Productivity AND Portability ... the database people “did it” with relational algebras and SQL.
- We can do it too with algebras over distributed data structures ... that is a set of operators over values expressed in terms of our distributed data structures.
- If we get it right, we’ll have ... declarative semantics that a software generator can turn into laptop, cluster or cloud programs.



\*This is the logo of the machine programming research program I help lead inside Intel Labs

# The Three Pillars of Machine Programming (MP)



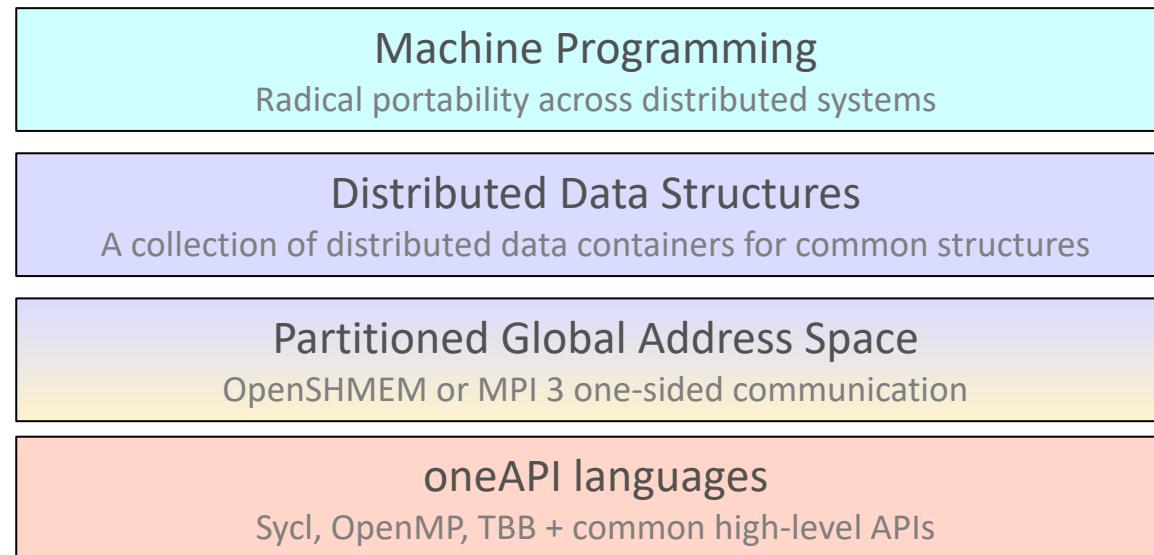
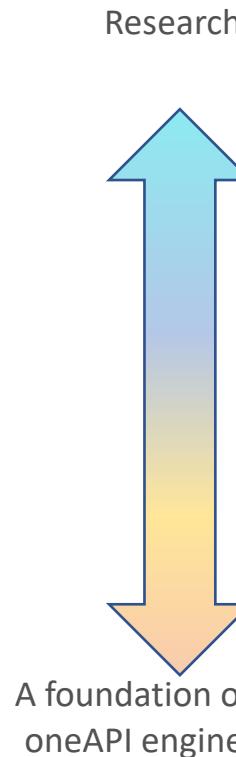
- **MP is the automation of software development**
  - **Intention:** Discover the intent of a programmer
  - **Invention:** Create new algorithms and data structures
  - **Adaptation:** Evolve in a changing hardware/software world

Justin Gottschlich, Intel Labs  
Armando Solar-Lezama, MIT  
Nesime Tatbul, Intel Labs  
Michael Carbin, MIT  
Martin Rinard, MIT  
Regina Barzilay, MIT  
Saman Amarasinghe, MIT  
Joshua B Tenenbaum, MIT  
Tim Mattson, Intel Labs

Summarized ~90 works.  
Key efforts by Berkeley, Google, Microsoft, MIT, Stanford, UW and others.

# oneAPI: A bridge to our heterogeneous/Distributed Future

My vision for how we bring oneAPI into a future dominated by power-optimized heterogenous chips organized into distributed systems.



The key to making this work ... the programmer is in control and chooses the level of abstraction based on the programming task.