

Ohio Supercomputer Center

An **OH**·**TECH** Consortium Member

Performance Tuning Workshop

Dr. Judy Gardiner July 27, 2017

www.osc.edu



Workshop Philosophy

- Aim for "reasonably good" performance
- Discuss performance tuning techniques common to most HPC architectures
 - Compiler options
 - Code modification
- Focus on serial performance
 - ***Most important: Unit stride memory access
- Touch on parallel processing
 - Multithreading
 - MPI





More Important than Performance!

- Correctness of results
- Code readability / maintainability
- Portability future systems
- Time to solution vs. execution time





Factors Affecting Performance

- Effective use of processor features
 - High degree of internal concurrency in a single core
- Data locality
 - Memory access is slow compared to computation
- File I/O
 - Use an appropriate file system
- Scalable algorithms
- Compiler optimizations
 - Modern compilers are amazing!
- Explicit parallelism



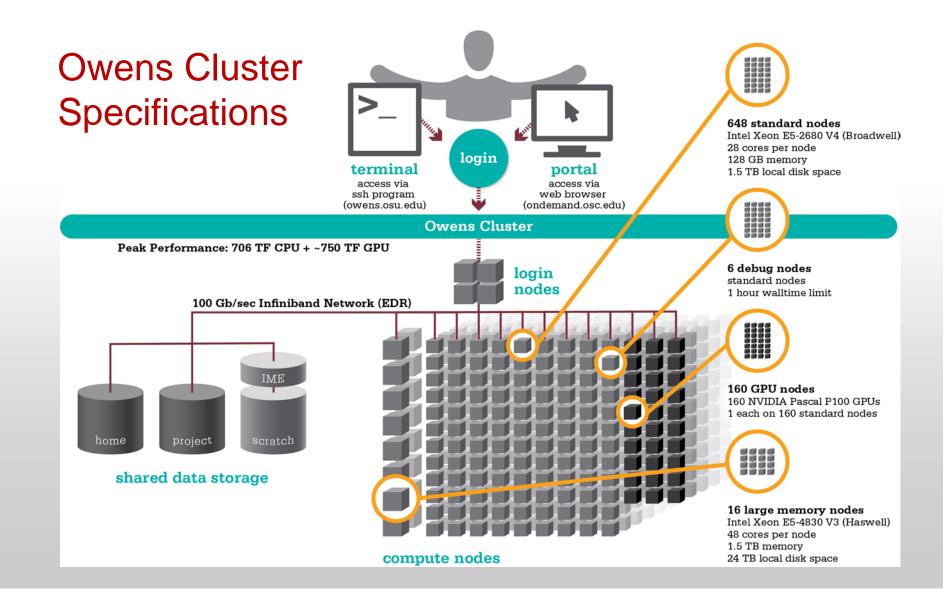


Outline

- Hardware overview
- Performance measurement and analysis
- Help from the compiler
- Code tuning / optimization
- Parallel computing
- Optional hands-on exercises







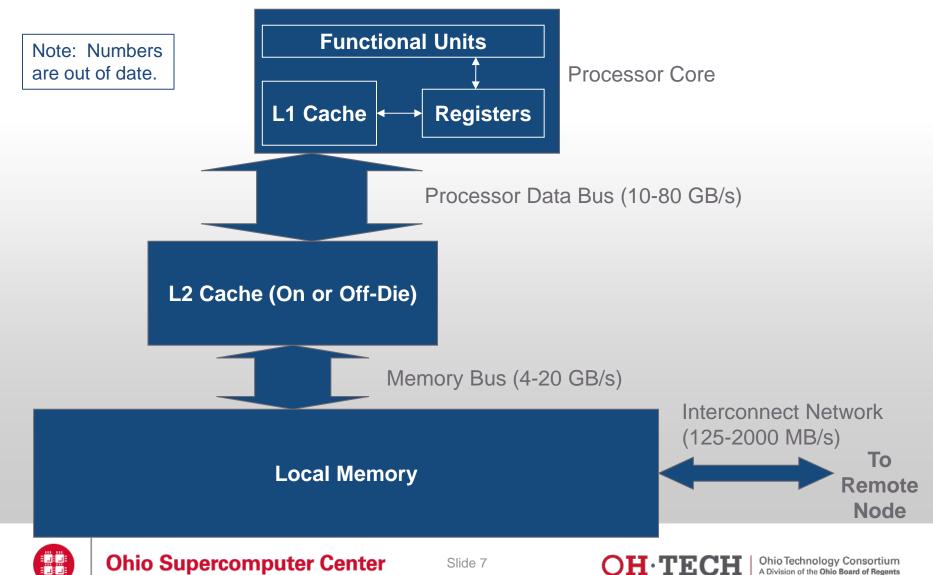


Ohio Supercomputer Center

Slide 6



Hierarchical Memory



Some Processor Features

- 28 cores per node
 - 14 cores per socket * 2 sockets per node
- Vector unit
 - Supports AVX 2
 - Vector length 4 double or 8 single precision values
 - Fused multiply-add
- Hyperthreading
 - Hardware support for 2 threads per core
 - Not currently enabled on OSC systems





Keep data close to the processor – file systems

• NEVER DO HEAVY I/O IN YOUR HOME DIRECTORY!

- Home directories are for long-term storage, not scratch files
- One user's heavy I/O load can affect all users
- For I/O-intensive jobs
 - Local disk on compute node (not shared)
 - Stage files to and from home directory into \$TMPDIR
 - Execute program in \$TMPDIR
 - Scratch file system
 - /fs/scratch/username or \$PFSDIR
 - Faster than other file systems
 - Good for parallel jobs
 - May be faster than local disk





What is good performance?

- FLOPS
 - Floating Point OPerations per Second
- Peak performance
 - Theoretical maximum (all cores fully utilized)
 - Owens 859 trillion FLOPS (859 teraflops)
- Sustained performance
 - LINPACK benchmark
 - Owens 706 teraflops
- Application performance is typically much less





Performance Measurement and Analysis

- Wallclock time
 - How long the program takes to run
- Performance reports
 - Easy, brief summary
- Profiling
 - Detailed information, more involved





Timing – command line

- Time a program
 - /usr/bin/time command

```
/usr/bin/time j3
5415.03user 13.75system 1:30:29elapsed 99%CPU \
(0avgtext+0avgdata 0maxresident)k \
0inputs+0outputs (255major+509333minor)pagefaults 0swaps
```

- Note: Hardcode the path less information otherwise
- /usr/bin/time gives results for •
 - user time (CPU time spent running your program)
 - system time (CPU time spent by your program in system calls)
 - elapsed time (wallclock)
 - % CPU -- (user+system)/elapsed
 - memory, pagefault, and swap statistics
 - I/O statistics





Timing routines embedded in code

- Time portions of your code
 - C/C++
 - Wallclock: time(2), difftime(3), getrusage(2)
 - CPU: times(2)
 - Fortran 77/90
 - Wallclock: SYSTEM_CLOCK(3)
 - CPU: DTIME(3), ETIME(3)
 - MPI (C/C++/Fortran)
 - Wallclock: MPI_Wtime(3)





Profiling Tools Available at OSC

- Profiling tools
 - Allinea Performance Reports
 - Allinea MAP
 - Intel VTune
 - Intel Trace Analyzer and Collector (ITAC)
 - Intel Advisor





What can a profiler show you?

- Whether code is
 - compute-bound
 - memory-bound
 - communication-bound
- How well the code uses available resources
 - Multiple cores
 - Vectorization
- How much time is spent in different parts of the code
- Profilers use hardware performance counters





Compilation flags for profiling

- For profiling
 - Use -g flag
 - Explicitly specify optimization level -On
 - Example: icc -g -O3 -o mycode mycode.c
- Use the same level of optimization you normally do
 - Bad example: icc -g -o mycode mycode.c
 - Equivalent to -O0



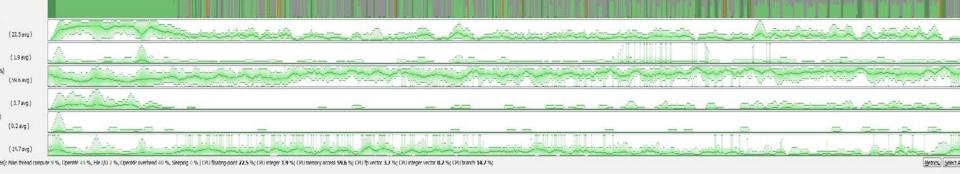


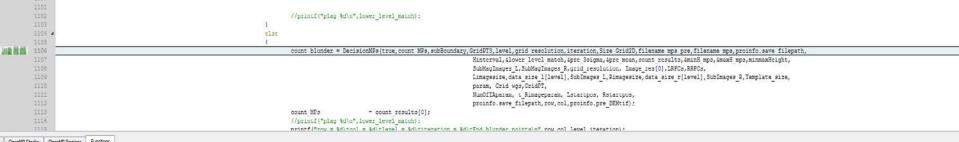
Allinea MAP

- Interpretation of profile requires some expertise
- Gives details about your code's performance
- For a non-MPI program:
 - module load allinea
 - map --profile --no-mpi ./mycode [args]
- For an MPI program:
 - map --profile -np num_procs ./mycode [args]
- View and explore resulting profile using Allinea client









OpenMP Stades	OpenMP Regions	Functions
---------------	----------------	-----------

Total	Child (Overhead	Fuclin
and 15.3%	V	16.39%	Imp Jourds worker(wold)
16.3%		15.3%	Imp Jourch Treed
16.3%	6 3	16.3%	mak (OpenPP Overlead)
(16.3%	Υ	16.3%	long first barrier(nr. nr)
16.3%	1	16.3%	Imp hyper barrier release/barrier type, imp into", int, int, int, int, int, int, int, int
	0.1%		Verticalinations Surder (DeerMP region 1)
14.3%		14.3%	Imp wait yield 1
		9.2%	Into wail. (Enclose (Over MP Overhead)
7.7%	1	7,7%	superd [hined] (OpenPO Vieted)
		1.7%	timp suspend 64 (Dirent/PD Direntend)
7.7%		7.7%	Imp support temptate [inited] (DeerVP Overhead)
		1.7%	pthread_cond_wait (DentRP Overhead)
17.1%		7.1%	tonp_wait_template [nined] (DpenVP Overhead)
6.2%	6	6,2%	_imp dispath pert
6.2%		h.2%	is the new set of the set
5.4%			
10.0%			Orientation [minos]
7.9%			Section: The section of the section
3.6%			Heritalinetuse (Jopen Prepin 1)
2.9%		2,9%	
3.3%			
4,5%	2.3%		10 yficarf internal
2,0%	222		
1.7%		1.7%	Louis and the second
	<0.1%		SetHeyhRange, Blunder (DoerNP region 0)
1.6%			cite prove
1.5%			_sched_yeld
1.5%			
	3.6%		and a provide [mined]
1.4%		£4%	Jerogramme (minor) mort_set_passe (minor) (Deenty Overhead)
1.7%		1.7%	
1.0%		2763-65	intree internet in the second s
1.0%			::Greatelingsoffynamid(void) (OperMP region 1)
11 1.5%			
0.8%			
0.8%		0.8%	
	16.7%		Installations burde (DeeMP)
0.7%		102207	20Dexode
0.7%			dan 2
0.6%			
0.6%			in make
0.0%			_IR_mov (between the second se
	1.11		

Profiling – What do I look for?

- Hot spots where most of the time is spent
 - This is where we'll focus our optimization effort
- Excessive number of calls to short functions
 - Use inlining! (compiler flags)
- Memory usage
 - Swapping, thrashing not allowed at OSC (job gets killed)
- CPU time vs. wall time (% CPU)
 - Low CPU utilization may mean excessive I/O delays



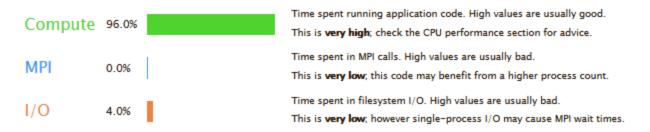


Allinea Performance Reports

- Easy to use
 - "-g" flag not needed works on precompiled binaries
- Gives a summary of your code's performance
 - view report with browser
- For a non-MPI program:
 - module load allinea
 - perf-report --no-mpi ./mycode [args]
- For an MPI program:
 - perf-report -np num_procs ./mycode [args]

nessures.	i node (co physical, co logical cores per node)
Memory:	63 GB per node
Tasks:	1 process, OMP_NUM_THREADS was 0
Machine:	r0019.ten.osc.edu
Start time:	Fri Aug 28 11:03:22 2015
Total time:	667 seconds (11 minutes)
Full path:	/nfs/14/judithg/howat/SETSM_c_0224
Input file:	
Notes:	

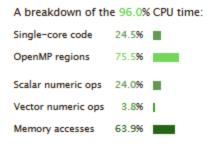
Summary: setsm is Compute-bound in this configuration



This application run was Compute-bound. A breakdown of this time and advice for investigating further is in the CPU section below.

As very little time is spent in MPI calls, this code may also benefit from running at larger scales.

CPU



MPI

A breakdown of the 0.0% MPI time:			
Time in collective calls	0.0%		
Time in point-to-point calls	0.0%		
Effective process collective rate	0.00 bytes/s		
Effective process point-to-point rate	0.00 bytes/s		

No time is spent in MPI operations. There's nothing to optimize here!

The per-core performance is memory-bound. Use a profiler to identify time-consuming loops and check their cache performance.

.

More Information about Allinea Tools

- www.osc.edu/documentation/software_list/allinea
- <u>www.allinea.com</u>





Compiler and Language Choice

- HPC software traditionally written in Fortran or C
- OSC supports several compiler families
 - Intel (icc, icpc, ifort)
 - Usually gives fastest code on Intel architecture
 - Portland Group (PGI pgcc, pgc++, pgf90)
 - Good for GPU programming, OpenACC
 - GNU (gcc, g++, gfortran)
 - Open source, universally available





Compiler Options for Performance Tuning

- Why use compiler options?
 - Processors have a high degree of internal concurrency
 - Compilers do an amazing job at optimization
 - Easy to use Let the compiler do the work!
 - Reasonably portable performance
- Optimization options
 - Let you control aspects of the optimization
- Warning:
 - Different compilers have different default values for options





Compiler Optimizations

- Function inlining
 - Eliminate function calls
- Interprocedural optimization/analysis (ipo/ipa)
 - Same file or multiple files
- Loop transformations
 - Unrolling, interchange, splitting, tiling
- Vectorization
 - Operate on arrays of operands
- Automatic parallelization of loops
 - Very conservative multithreading





What compiler flags do I try first?

- General optimization flags (-O2, -O3, -fast)
- Fast math
- Interprocedural optimization / analysis
- Profile again, look for changes
- Look for new problems / opportunities





Floating Point Speed vs. Accuracy

- Faster operations are sometimes less accurate
- Some algorithms are okay, some quite sensitive
- Intel compilers
 - Fast math by default with -O2 and -O3
 - Use -fp-model precise if you have a problem (slower)
- GNU compilers
 - Precise math by default with -O2 and -O3 (slower)
 - Use -ffast-math for faster performance





Interprocedural Optimization / Inlining

- Inlining
 - Replace a subroutine or function call with the actual body of the subprogram
- Advantages
 - Overhead of calling the subprogram is eliminated
 - More loop optimizations are possible if calls are eliminated
- One source file
 - Typically automatic with -O2 and -O3
- Multiple source files compiled separately
 - Use compiler option for compile and link phases





Optimization Compiler Options – Intel compilers

-fast	Common optimizations
-0 <i>n</i>	Set optimization level (0, 1, 2, 3)
-ipo	Interprocedural optimization, multiple files
-03	Loop transforms
-xHost	Use highest instruction set available
-parallel	Loop auto- parallelizaton

- Don't use -fast for MPI programs with Intel compilers
- Use same compiler command to link for -ipo with separate compilation
- Many other optimization options are available
- See man pages for details
- Recommended options:
 -03 -xHost
- Example: ifort -03 program.f90



Optimization Compiler Options – PGI compilers

-fast	Common optimizations
-0 <i>n</i>	Set optimization level (0, 1, 2, 3, 4)
-Mipa	Interprocedural analysis
-Mconcur	Loop auto- parallelizaton

- Many other optimization options are available
- Use same compiler command to link for -Mipa with separate compilation
- See man pages for details
- Recommended options:
 -fast
- Example:

pgf90 -fast program.f90



Optimization Compiler Options – GNU compilers

-0 <i>n</i>	Set optimization level (0, 1, 2, 3)
N/A for separate compilation	Interprocedural optimization
-03	Loop transforms
-ffast-math	Potentially unsafe float pt optimizations
-march=native	Use highest instruction set available

- Many other optimization options are available
- See man pages for details
- Recommended options:
 -03 -ffast-math
 -march=native
- Example: gfortran -03 program.f90



Compiler Optimization Reports

- Let you understand
 - how well the compiler is doing at optimizing your code
 - what parts of code need work
- Generated at compile time
 - Describe what optimizations were applied at various points in the source code
 - May tell you why optimizations could not be performed





Options to Generate Optimization Reports

- Intel compilers
 - -opt-report
 - Output to a file
- Portland Group compilers
 - -Minfo
 - Output to stderr
- GNU compilers
 - -fopt-info
 - Output to stderr by default





Sample from an Optimization Report

```
LOOP BEGIN at laplace-good.f(10,7)
   remark #15542: loop was not vectorized: inner loop was already vectorized
  LOOP BEGIN at laplace-good.f(11,10)
   <Peeled loop for vectorization>
   LOOP END
  LOOP BEGIN at laplace-good.f(11,10)
      remark #15300: LOOP WAS VECTORIZED
   LOOP END
  LOOP BEGIN at laplace-good.f(11,10)
   <Remainder loop for vectorization>
      remark #15301: REMAINDER LOOP WAS VECTORIZED
   LOOP END
  LOOP BEGIN at laplace-good.f(11,10)
   <Remainder loop for vectorization>
  LOOP END
LOOP END
```





A word about algorithms

- Problem-dependent can't generalize
- Scalability is important
 - How computational time increases with problem size
- Replace with an equivalent algorithm of lower complexity
 - computational geometry: change from vertex representation to half-plane representation
 - $O(2^n) \rightarrow O(n)$
 - replace 2D convolutions with 2D FFTs
 - $O(n^4) \rightarrow O(n^2 \log(n))$
- Replace home-grown algorithm with call to optimized library



Code Modifications for Optimization

- Memory optimizations
 - Unit stride memory access
 - Efficient cache usage
- Vectorization
 - Vectorizable loops
 - Vectorization inhibitors





Unit Stride Memory Access

- The most important factor in your code's performance!!!
- Loops that work with arrays should use a stride of one whenever possible
- C, C++ are *row-major*; in a 2D array, they store elements consecutively by row:
 - First array index should be outermost loop
 - Last array index should be innermost loop
- Fortran is *column-major*, so the reverse is true:
 - Last array index should be outermost loop
 - First array index should be innermost loop
- Avoid arrays of derived data types, structs, or classes





Data Layout: Object-Oriented Languages

- Arrays of objects may give poor performance on HPC systems if used naively
 - C structs
 - C++ classes
 - Fortran 90 user-defined types
- Inefficient use of cache not unit stride
 - Can often get factor of 3 or 4 speedup just by fixing it
- You can use them efficiently! Be aware of data layout.
- Data layout may be the only thing modern compilers can't optimize





Efficient Cache Usage

- Cache lines
 - 8 words (64 bytes) of consecutive memory
 - Entire cache line is loaded when a piece of data is fetched
- Good example Entire cache line used
 - 2 cache lines used for every 8 loop iterations
 - Unit stride

real*8 a(N), b(N)
do i=1,N
 a(i)=a(i)+b(i)
end do

2 cache lines: a(1),a(2),a(3), ... a(8) b(1),b(2),b(3), ... b(8)





Efficient Cache Usage – Cache Lines (cont.)

- Bad example Unneeded data loaded
 - 1 cache line loaded for each loop iteration
 - 8 words loaded, only 2 words used
 - Not unit stride

```
TYPE :: node
  real*8 a, b, c, d, w, x, y, z
END TYPE node
TYPE(node) :: s(N)
do i=1,N
  s(i)%a = s(i)%a + s(i)%b
end do
```

cache line: a(1),b(1),c(1),d(1),w(1),x(1),y(1),z(1)





Vectorization / Streaming

- Code is structured to operate on arrays of operands
 - Single Instruction, Multiple Data (SIMD)
- Vector instructions built into processor (AVX, SSE, etc.)
 - Vector length 8 single or 4 double precision on Owens
- Requires unit stride
- Fortran 90, MATLAB have this idea built in
- A vectorizable loop:

```
do i=1,N
a(i)=b(i)+x(i)*c(i)
end do
```





Vectorization Inhibitors

- Not unit stride
 - Loops in wrong order (column-major vs. row major)
 - Usually fixed by the compiler
 - Loops over derived types
- Function calls
 - Sometimes fixed by inlining
 - Can split loop into two loops
- Too many conditionals
 - "if" statements





Optimized Mathematical Libraries

- MKL (Intel Math Kernel Library)
 - BLAS
 - LAPACK
 - FFT
 - Vectorized transcendental functions (sin, cos, exp)
- FFTW
- ScaLAPACK
- SuperLU
- ... and many others





Parallel Computing

- Multithreading
 - Shared-memory model (single node)
 - OpenMP support in compilers
- Message Passing Interface (MPI)
 - Distributed-memory model (single or multiple nodes)
 - Several available libraries
- Accelerators / Coprocessors
 - GPUs
 - Intel Xeon Phi (not currently available at OSC)





What is OpenMP?

- Shared-memory, threaded parallel programming model
- Portable standard
- A set of compiler directives
- A library of support functions
- Supported by vendors' compilers
 - Intel
 - Portland Group
 - GNU
 - Cray





Parallel loop execution – Fortran

- Inner loop vectorizes
- Outer loop executes on multiple threads

```
PROGRAM omploop
INTEGER, PARAMETER :: N = 1000
INTEGER i, j
REAL, DIMENSION(N,N) :: a, b, c, x
... ! Initialize arrays
!$OMP PARALLEL DO
do j=1,N
    do i=1,N
        a(i,j)=b(i,j)+x(i,j)*c(i,j)
        end do
end do
!$OMP END PARALLEL DO
END PROGRAM omploop
```





Parallel loop execution – C

- Inner loop vectorizes
- Outer loop executes on multiple threads





Compiling a program with OpenMP

- Intel compilers
 - Add the -gopenmp option

[owens-login01]\$ ifort -qopenmp ompex.f90 -o ompex

- gnu compilers
 - Add the -fopenmp option

[owens-login01]\$ gcc -fopenmp ompex.c -o ompex

- Portland Group compilers
 - Add the -mp option

[owens-login01]\$ pgf90 -mp ompex.f90 -o ompex





Running an OpenMP program

- Request multiple processors through PBS
 - Example: nodes=1:ppn=28
- Set the OMP_NUM_THREADS environment variable
 - Default: Use all available cores
- For best performance run at most one thread per core
 - Otherwise too much overhead
 - Applies to typical HPC workload, exceptions exist





Running an OpenMP program – Example

```
[owens-login01]$ cat omploop.pbs
#PBS -N omploop
#PBS −j oe
```

```
#PBS -1 nodes=1:ppn=28
```

```
#PBS -1 walltime=1:00
```

```
cd $PBS_O_WORKDIR
export OMP NUM THREADS=28
/usr/bin/time ./omploop
```





More Information about OpenMP

- <u>www.openmp.org</u>
- OpenMP Application Program Interface
 - Version 3.1, July 2011
 - <u>http://www.openmp.org/wp-content/uploads/OpenMP3.1.pdf</u>





What is MPI?

- Message Passing Interface
 - Multiple processes run on one or more nodes
 - Distributed-memory model
- A message passing library
- A run-time environment
 - mpiexec
- Compiler wrappers
- Supported by all major parallel machine manufacturers

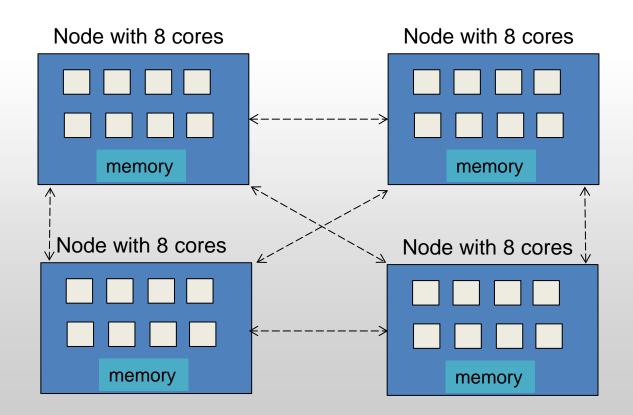




OpenMP vs. MPI

Infiniband network

←----→





Slide 53



A simple MPI program

```
[owens-login01]$ cat hello.c
#include <mpi.h>
#include <stdio.h>
int main(int argc, char *argv[])
ł
    int rank,size;
    MPI_Init(&argc,&argv);
    MPI Comm rank(MPI COMM WORLD,&rank);
    MPI Comm size(MPI COMM WORLD,&size);
    printf("Hello from node %d of %d\n",rank,size);
    MPI Finalize();
    return(0);
```





MPI Implementations Available at OSC

- mvapich2
 - default
- IntelMPI
 - available only with Intel compilers
- OpenMPI





Compiling MPI programs

- Compile with the MPI compiler wrappers
 - mpicc, mpicxx and mpif90
 - Accept the same arguments as the compilers they wrap

[owens-login01]\$ mpicc -o hello hello.c

Compiler and MPI implementation depend on modules
 loaded





Running MPI programs

- MPI programs must run in batch only
 - Debugging runs may be done with interactive batch jobs
- mpiexec
 - Automatically determines execution nodes from PBS
 - Starts the program running, 2x28=56 copies

```
[owens-login01]$ cat hello.pbs
#PBS -N mpi hello
#PBS −j oe
#PBS -1 nodes=2:ppn=28
#PBS -1 walltime=1:00
cd $PBS O WORKDIR
mpiexec ./hello
```





More Information about MPI

- <u>www.mpi-forum.org</u>
- MPI: A Message-Passing Interface Standard
 - Version 3.1, June 4, 2015
 - <u>http://mpi-forum.org/docs/mpi-3.1/mpi31-report.pdf</u>





GPU-Accelerated Computing

- GPU = Graphics Processing Unit
 - Can be used to accelerate computation
- OSC clusters have some nodes with NVIDIA GPUs
- Manycore processors
 - more cores than multicore
- Can be programmed with CUDA
 - low level
- PGI and gnu compilers support OpenACC
 - easier than CUDA
 - similar to OpenMP





Summary: What would I do with your code?

- Profile it
- Experiment with compiler optimization flags
- Analyze data layout, memory access patterns
- Examine algorithms
 - Complexity
 - Availability of optimized version
- Look for potential parallelism, inhibitors to parallelism
 - Including vectorization





Other Sources of Information

- Online manuals
 - man ifort
 - man pgc++
 - man gcc
- Related workshop courses
 - www.osc.edu/supercomputing/training
- Online tutorials from Cornell
 - <u>https://cvw.cac.cornell.edu/</u>
- <u>oschelp@osc.edu</u>

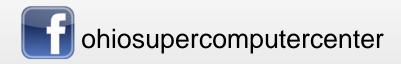




Questions

Judy Gardiner Scientific Applications Engineer Ohio Supercomputer Center judithg@osc.edu

1224 Kinnear Road Columbus, OH 43212 Phone: (614) 292-9623





ohiosupercomputerctr





