Hybrid MPI & OpenMP Parallel Programming

MPI + OpenMP and other models on clusters of SMP nodes

Rolf Rabenseifner¹) Georg Hager²) Gabriele Jost³)
Rabenseifner@hlrs.de Georg.Hager@rrze.uni-erlangen.de Gabriele.Jost@amd.com

¹) High Performance Computing Center (HLRS), University of Stuttgart, Germany
²) Regional Computing Center (RRZE), University of Erlangen, Germany
³) AMD, USA, formerly Texas Advanced Computing Center, USA

PRACE Winter School: “Hybrid Programming on Massively Parallel Architectures”
Feb. 6–10, 2011, CINECA, Bologna, Italy
Outline

- Introduction / Motivation
- Programming models on clusters of SMP nodes
- Case Studies / pure MPI vs hybrid MPI+OpenMP
- Practical “How-To” on hybrid programming
- Mismatch Problems
- Opportunities: Application categories that can benefit from hybrid parallelization
- Thread-safety quality of MPI libraries
- Tools for debugging and profiling MPI+OpenMP
- Other options on clusters of SMP nodes
- Summary
- Appendix
- Content (detailed)
Motivation

- Efficient programming of clusters of SMP nodes
  SMP nodes:
  - Dual/multi core CPUs
  - Multi CPU shared memory
  - Multi CPU ccNUMA
  - Any mixture with shared memory programming model

- Hardware range
  - mini-cluster with dual-core CPUs
  - ...
  - large constellations with large SMP nodes
    ... with several sockets (CPUs) per SMP node
    ... with several cores per socket
    → Hierarchical system layout

- Hybrid MPI/OpenMP programming seems natural
  - MPI between the nodes
  - OpenMP inside of each SMP node
Motivation

- Which programming model is fastest?
- MPI everywhere?
- Fully hybrid MPI & OpenMP?
- Something between? (Mixed model)
- Often hybrid programming slower than pure MPI – Examples, Reasons, ...

Hybrid Parallel Programming
Rabenseifner, Hager, Jost
Goals of this tutorial

- Sensitize to problems on clusters of SMP nodes
  see sections → Case studies → Mismatch problems

- Technical aspects of hybrid programming
  see sections → Programming models on clusters → Examples on hybrid programming

- Opportunities with hybrid programming
  see section → Opportunities: Application categories that can benefit from hybrid paralleliz.

- Issues and their Solutions
  with sections → Thread-safety quality of MPI libraries → Tools for debugging and profiling for MPI+OpenMP

• Less frustration & • More success
with your parallel program on clusters of SMP nodes
Outline

• Introduction / Motivation

• Programming models on clusters of SMP nodes
  • Case Studies / pure MPI vs hybrid MPI+OpenMP
  • Practical “How-To” on hybrid programming
  • Mismatch Problems
  • Opportunities:
    Application categories that can benefit from hybrid parallelization
  • Thread-safety quality of MPI libraries
  • Tools for debugging and profiling MPI+OpenMP
  • Other options on clusters of SMP nodes
  • Summary
Major Programming models on hybrid systems

- Pure MPI (one MPI process on each core)
- Hybrid MPI+OpenMP
  - shared memory OpenMP
  - distributed memory MPI
- Other: Virtual shared memory systems, PGAS, HPF, …
- Often **hybrid programming (MPI+OpenMP)** slower than pure MPI
  - why?

**MPI**
- Sequential program on each core
- Explicit **Message Passing** by calling MPI Send & MPI Recv
- local data in each process

**OpenMP** (shared data)
- some_serial_code
- #pragma omp parallel for
- for (j=...;...; j++)
- block_to_be_parallelized
- again_some_serial_code
- Master thread, other threads
- sleeping

OpenMP inside of the SMP nodes
MPI between the nodes via node interconnect
Node Interconnect
Parallel Programming Models on Hybrid Platforms

- **pure MPI**
  - one MPI process on each core

- **hybrid MPI+OpenMP**
  - MPI: inter-node communication
  - OpenMP: inside of each SMP node

- **OpenMP only**
  - distributed virtual shared memory

- **No overlap of Comm. + Comp.**
  - MPI only outside of parallel regions of the numerical application code

- **Overlapping Comm. + Comp.**
  - MPI communication by one or a few threads while other threads are computing

- **Masteronly**
  - MPI only outside of parallel regions
Pure MPI

Advantages
- No modifications on existing MPI codes
- MPI library need not to support multiple threads

Major problems
- Does MPI library uses internally different protocols?
  - Shared memory inside of the SMP nodes
  - Network communication between the nodes
- Does application topology fit on hardware topology?
- Unnecessary MPI-communication inside of SMP nodes!

Discussed in detail later on in the section Mismatch Problems
Hybrid Masteronly

**Advantages**
- No message passing inside of the SMP nodes
- No topology problem

```c
for (iteration ....) {
    #pragma omp parallel
    numerical code
    /*end omp parallel */
    /* on master thread only */
    MPI_Send (original data to halo areas in other SMP nodes)
    MPI_Recv (halo data from the neighbors)
} /*end for loop*/
```

**Major Problems**
- All other threads are sleeping while master thread communicates!
- Which inter-node bandwidth?
- MPI-lib must support at least MPI_THREAD_FUNNELED

→ Section Thread-safety quality of MPI libraries
Overlapping Communication and Computation

MPI communication by one or a few threads while other threads are computing

```c
if (my_thread_rank < ...) {
    MPI_Send/Recv....
    i.e., communicate all halo data
} else {
    Execute those parts of the application
    that do not need halo data
    (on non-communicating threads)
}

Execute those parts of the application
that need halo data
(on all threads)
```
Pure OpenMP (on the cluster)

- Distributed shared virtual memory system needed
- Must support clusters of SMP nodes
- e.g., Intel® Cluster OpenMP
  - Shared memory parallel inside of SMP nodes
  - Communication of modified parts of pages at OpenMP flush (part of each OpenMP barrier)

i.e., the OpenMP memory and parallelization model is prepared for clusters!

Experience: → Mismatch section
Outline

- Introduction / Motivation
- Programming models on clusters of SMP nodes

**Case Studies / pure MPI vs hybrid MPI+OpenMP**
- The Multi-Zone NAS Parallel Benchmarks
- For each application we discuss:
  - Benchmark implementations based on different strategies and programming paradigms
  - Performance results and analysis on different hardware architectures
- Compilation and Execution Summary

Gabriele Jost (University of Texas, TACC/Naval Postgraduate School, Monterey CA)

- Practical “How-To” on hybrid programming
- Mismatch Problems
- Opportunities: Application categories that can benefit from hybrid parallelism
- Thread-safety quality of MPI libraries
- Tools for debugging and profiling MPI+OpenMP
- Other options on clusters of SMP nodes
- Summary
The Multi-Zone NAS Parallel Benchmarks

- Multi-zone versions of the NAS Parallel Benchmarks LU, SP, and BT
- Two hybrid sample implementations
- Load balance heuristics part of sample codes
- [www.nas.nasa.gov/Resources/Software/software.html](http://www.nas.nasa.gov/Resources/Software/software.html)
call omp_set_numthreads (weight)
do step = 1, itmax
    call exch_qbc(u, qbc, nx,...)
end do
... call mpi_send/recv

do zone = 1, num_zones
    if (iam .eq. pzone_id(zone)) then
        call zsolve(u, rsd,...)
    end if
end do

end do

...
Pipelined Thread Execution in SSOR

subroutine ssor
  !$OMP PARALLEL DEFAULT(SHARED)
  !$OMP& PRIVATE(m,i,j,k...)
    call sync1()
    do k = 2, nz-1
      !$OMP DO
        do j = 2, ny-1
          do i = 2, nx-1
            do m = 1, 5
              rsd(m,i,j,k) =
              dt*rsd(m,i,j,k-1) + ...
              end do
            end do
          end do
        end do
      !$OMP END DO nowait
    end do
    call sync2()
  !$OMP END PARALLEL
  ...

subroutine sync1
  ...neigh = iam -1
  do while (isync(neigh) .eq. 0)
    !$OMP FLUSH(isync)
  end do
  isync(neigh) = 0
  !$OMP FLUSH(isync)
  ...

subroutine sync2
  ...
  neigh = iam -1
  do while (isync(neigh) .eq. 1)
    !$OMP FLUSH(isync)
  end do
  isync(neigh) = 1
  !$OMP FLUSH(isync)
Benchmark Characteristics

- Aggregate sizes:
  - Class D: 1632 x 1216 x 34 grid points
  - Class E: 4224 x 3456 x 92 grid points

- **BT-MZ:** *(Block tridiagonal simulated CFD application)*
  - Alternative Directions Implicit (ADI) method
  - #Zones: 1024 (D), 4096 (E)
  - Size of the zones varies widely:
    - large/small about 20
    - requires multi-level parallelism to achieve a good load-balance

- **LU-MZ:** *(LU decomposition simulated CFD application)*
  - SSOR method (2D pipelined method)
  - #Zones: 16 (all Classes)
  - Size of the zones identical:
    - no load-balancing required
    - limited parallelism on outer level

- **SP-MZ:** *(Scalar Pentadiagonal simulated CFD application)*
  - #Zones: 1024 (D), 4096 (E)
  - Size of zones identical
    - no load-balancing required

Expectations:

- Pure MPI: Load-balancing problems!
- Good candidate for MPI+OpenMP
- Limited MPI Parallelism: \( \rightarrow \) MPI+OpenMP increases Parallelism
- Load-balanced on MPI level: Pure MPI should perform best
Benchmark Architectures

- Sun Constellation (Ranger)
- Cray XT5 (skipped)
- IBM Power 6
Sun Constellation Cluster Ranger (1)

- Located at the Texas Advanced Computing Center (TACC), University of Texas at Austin (http://www.tacc.utexas.edu)
- 3936 Sun Blades, 4 AMD Quad-core 64bit 2.3GHz processors per node (blade), 62976 cores total
- 123TB aggregate memory
- Peak Performance 579 Tflops
- InfiniBand Switch interconnect
- Sun Blade x6420 Compute Node:
  - 4 Sockets per node
  - 4 cores per socket
  - HyperTransport System Bus
  - 32GB memory
Sun Constellation Cluster Ranger (2)

- **Compilation:**
  - PGI pgf90 7.1
  - mpif90 -tp barcelona-64 -r8 -mp

- **Cache optimized benchmarks Execution:**
  - MPI MVAPICH
  - setenv OMP_NUM_THREADS `nthreads`
  - lbrun numactl bt-mz.exe

- **numactl** controls
  - Socket affinity: select sockets to run
  - Core affinity: select cores within socket
  - Memory policy: where to allocate memory
  - [http://www.halobates.de/numaapi3.pdf](http://www.halobates.de/numaapi3.pdf) i.e., with OpenMP
SUN: Running hybrid on Sun Constellation Cluster Ranger

- Highly hierarchical
- Shared Memory:
  - Cache-coherent, Non-uniform memory access (ccNUMA) 16-way Node (Blade)
- Distributed memory:
  - Network of ccNUMA blades
    - Core-to-Core
    - Socket-to-Socket
    - Blade-to-Blade
    - Chassis-to-Chassis

network
SUN: NPB-MZ Class E Scalability on Ranger

- Scalability in MFlops
- MPI/OpenMP outperforms pure MPI
- Use of numactl essential to achieve scalability

**BT**
- Significant improvement (235%):
  - Load-balancing issues solved with MPI+OpenMP

**SP**
- Pure MPI is already load-balanced.
  - But hybrid 9.6% faster, due to smaller message rate at NIC
  - Cannot be build for 8192 processes!

**Hybrid**
- **SP:** still scales
- **BT:** does not scale
NUMA Control: Process Placement

- Affinity and Policy can be changed externally through `numactl` at the socket and core level.

**Command:** `numactl <options> ./a.out`

**Example:**
- `numactl -N 1 ./a.out`
- `numactl -c 0,1 ./a.out`
NUMA Operations: Memory Placement

Memory allocation:
- MPI
  - local allocation is best
- OpenMP
  - Interleave best for large, completely shared arrays that are randomly accessed by different threads
  - local best for private arrays
- Once allocated, a memory-structure is fixed

Example: `numactl -N 1 -1 ./a.out`
### NUMA Operations (cont. 3)

<table>
<thead>
<tr>
<th>cmd</th>
<th>option</th>
<th>arguments</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socket Affinity</td>
<td>numactl</td>
<td>-N</td>
<td>{0,1,2,3} Only execute process on cores of this (these) socket(s).</td>
</tr>
<tr>
<td>Memory Policy</td>
<td>numactl</td>
<td>-l</td>
<td>{no argument} Allocate on current socket.</td>
</tr>
<tr>
<td>Memory Policy</td>
<td>numactl</td>
<td>-i</td>
<td>{0,1,2,3} Allocate round robin (interleave) on these sockets.</td>
</tr>
<tr>
<td>Memory Policy</td>
<td>numactl</td>
<td>--preferred=</td>
<td>{0,1,2,3} select only one Allocate on this socket; fallback to any other if full.</td>
</tr>
<tr>
<td>Memory Policy</td>
<td>numactl</td>
<td>-m</td>
<td>{0,1,2,3} Only allocate on this (these) socket(s).</td>
</tr>
<tr>
<td>Core Affinity</td>
<td>numactl</td>
<td>-C</td>
<td>{0,1,2,3, 4,5,6,7, 8,9,10,11, 12,13,14,15} Only execute process on this (these) Core(s).</td>
</tr>
</tbody>
</table>
### Hybrid Batch Script: 4 tasks, 4 threads/task

<table>
<thead>
<tr>
<th><strong>job script</strong> (Bourne shell)</th>
<th><strong>job script</strong> (C shell)</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>#! -pe 4way 32</td>
<td>#! -pe 4way 32</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>export OMP_NUM_THREADS=4</td>
<td>setenv OMP_NUM_THREADS 4</td>
</tr>
<tr>
<td>ibrun numa.sh</td>
<td>ibrun numa.csh</td>
</tr>
</tbody>
</table>

#### numa.sh

```bash
#!/bin/bash
export MV2_USE_AFFINITY=0
export MV2_ENABLE_AFFINITY=0
export VIADEV_USE_AFFINITY=0

#TasksPerNode
TPN=`echo $PE | sed 's/way//'`
[ ! $TPN ] && echo TPN NOT defined!
[ ! $TPN ] && exit 1

socket=$(( $PMI_RANK % $TPN ))
numactl -N $socket -m $socket ./a.out
```

#### numa.csh

```bash
#!/bin/tcsh
setenv MV2_USE_AFFINITY 0
setenv MV2_ENABLE_AFFINITY 0
setenv VIADEV_USE_AFFINITY 0

#TasksPerNode
set TPN = `echo $PE | sed 's/way//'`
if( !( ${%TPN}) ) echo TPN NOT defined!
if( !( ${%TPN}) ) exit 0

@ socket = $PMI_RANK % $TPN
numactl -N $socket -m $socket ./a.out
```
Numactl – Pitfalls: Using Threads across Sockets

bt-mz.1024x8 yields best load-balance

```bash
-pe 2way 8192
export OMP_NUM_THREADS=8
my_rank=$PMI_RANK
local_rank=$(( $my_rank % $myway ))
numnode=$(( $local_rank + 1 ))
```

**Original:**
numactl -N $numnode -m $numnode *

**Bad performance!**
- Each process runs 8 threads on 4 cores
- Memory allocated on one socket

**Network Diagram:**

**Rank 0:**
- Rank 0
- Rank 1

**Network:**

- 0
- 1
- 2
- 3

---

Hybrid Parallel Programming
Rabenseifner, Hager, Jost

Slide 27 / 154
**Numactl – Pitfalls:**

**Using Threads across Sockets**

```bash
bt-mz.1024x8
export OMP_NUM_THREADS=8

my_rank=$PMI_RANK
local_rank=$(( $my_rank % $myway ))
numnode=$(( $local_rank + 1 ))

**Original:**
numactl -N $numnode -m $numnode *

**Modified:**
if [ $local_rank -eq 0 ]; then
    numactl -N 0,3 -m 0,3 *
else
    numactl -N 1,2 -m 1,2 *
fi

**Achieves Scalability!**
- Process uses cores and memory across 2 sockets
- Suitable for 8 threads
Cray XT5

- Results obtained by the courtesy of the HPCMO Program and the Engineer Research and Development Center Major Shared Resource Center, Vicksburg, MS (http://www.erdc.hpc.mil/index)
- Cray XT5 is located at the Arctic Region Supercomputing Center (ARSC) (http://www.arsc.edu/resources/pingo)
  - 432- Cray XT5 compute nodes with
    - 32 GB of shared memory per node (4 GB per core)
    - 2 - quad core 2.3 GHz AMD Opteron processors per node.
    - 1 - Seastar2+ Interconnect Module per node.
  - Cray Seastar2+ Interconnect between all compute and login nodes
Cray XT5: NPB-MZ Class D Scalability

Results reported for Class D on 256-2048 cores

- SP-MZ pure MPI scales up to 1024 cores
- SP-MZ MPI/OpenMP scales to 2048 cores
- SP-MZ MPI/OpenMP outperforms pure MPI for 1024 cores
- BT-MZ MPI does not scale
- BT-MZ MPI/OpenMP scales to 2048 cores, outperforms pure MPI

Expected: #MPI processes limited

Unexpected!
Cray XT5: CrayPat Performance Analysis

- module load xt-craypat
- Compilation:
  - ftn –fastsse –tp barcelona–64 –r8 –mp=nonuma,[trace ]
- Instrument:
  - pat_build –w –T TraceOmp, –g mpi,omp bt.exe bt.exe.pat
- Execution:
  - (export PAT_RT_HWPC {0,1,2,..})
  - export OMP_NUM_THREADS 4
  - aprun –n NPROCS –S 1 –d 4 ./bt.exe.pat
- Generate report:
  - pat_report –O
    load_balance,thread_times,program_time,mpi_callers –O
    profile_pe.th $1
Cray XT5: BT-MZ 32x4 Function Profile

```c
!$OMP PARALLEL DEFAULT(SHARED) PRIVATE(n,m,k,i,j,ksize)
!$OMP&  SHARED(dz5,dz4,dz3,dz2,dz1,tz2,tz1,dt,c1345,c4,c3,con43,c3c4,c1,
    c2,nx,ny,nz)
  ksize = nz-1

  c Compute the indices for storing the block-diagonal matrix;
  c determine c (labeled f) and s jacobians

!$OMP DO
  do j = 1, ny-2
    do i = 1, nx-2
      do k = 0, ksize
        tmp1 = 1.d0 / u(1,i,j,k)
        tmp2 = tmp1 * tmp1
        tmp3 = tmp1 * tmp2
        fjac(1,1,k) = 0.d0
        fjac(1,2,k) = 0.d0
        fjac(1,3,k) = 0.d0
        fjac(1,4,k) = 1.d0
        fjac(1,5,k) = 0.d0
```
Cray XT5: BT-MZ Load-Balance 32x4 vs 128x1

- maximum, median, minimum PE are shown
- bt-mz.C.128x1 shows large imbalance in User and MPI time
- bt-mz.C.32x4 shows well balanced times

<table>
<thead>
<tr>
<th>Time %</th>
<th>Time</th>
<th>Calls</th>
<th>Group</th>
<th>PE[mmm]</th>
<th>Thread</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.0%</td>
<td>1.782603</td>
<td>18662</td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>86.1%</td>
<td>1.535163</td>
<td>7783</td>
<td>USER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.7%</td>
<td>1.535987</td>
<td>6813</td>
<td>lpe.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7%</td>
<td>1.535987</td>
<td>6188</td>
<td>lthread.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7%</td>
<td>1.535987</td>
<td>6188</td>
<td>lthread.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7%</td>
<td>1.535987</td>
<td>6188</td>
<td>lthread.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7%</td>
<td>1.466954</td>
<td>6813</td>
<td>lthread.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.7%</td>
<td>1.535147</td>
<td>7783</td>
<td>lpe.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7%</td>
<td>1.535147</td>
<td>7072</td>
<td>lthread.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7%</td>
<td>1.534995</td>
<td>7072</td>
<td>lthread.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7%</td>
<td>1.534968</td>
<td>7072</td>
<td>lthread.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6%</td>
<td>1.290502</td>
<td>7783</td>
<td>lthread.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.7%</td>
<td>1.534239</td>
<td>7783</td>
<td>lpe.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7%</td>
<td>1.534239</td>
<td>7072</td>
<td>lthread.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7%</td>
<td>1.534101</td>
<td>7072</td>
<td>lthread.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7%</td>
<td>1.534076</td>
<td>7072</td>
<td>lthread.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6%</td>
<td>1.268085</td>
<td>7783</td>
<td>lthread.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
IBM Power 6

- Results obtained by the courtesy of the HPCMO Program and the Engineer Research and Development Center Major Shared Resource Center, Vicksburg, MS (http://www.erdc.hpc.mil/index)
- The IBM Power 6 System is located at (http://www.navo.hpc.mil/davinci_about.html)
- 150 Compute Nodes
- 32 4.7GHz Power6 Cores per Node (4800 cores total)
- 64 GBytes of dedicated memory per node
- QLOGOC Infiniband DDR interconnect
- IBM MPI: MPI 1.2 + MPI-IO
  - mpxlf_r -O4 -qarch=pwr6 -qtune=pwr6 -qsmp=omp
- Execution:
  - poe launch $PBS_O_WORKDIR./sp.C.16x4.exe

Flag was essential to achieve full compiler optimization in presence of OMP directives!
NPB-MZ Class D on IBM Power 6: Exploiting SMT for 2048 Core Results

- Results for 128-2048 cores
- Only 1024 cores were available for the experiments
- BT-MZ and SP-MZ show benefit from Simultaneous Multithreading (SMT): 2048 threads on 1024 cores

Doubling the number of threads through hyperthreading (SMT):

```bash
#!/bin/csh
PBS -l select=32:ncpus=64:mpiprocs=NP:ompthreads=NT
```

Hybrid Parallel Programming

Slide 35 / 154  Rabenseifner, Hager, Jost
Conventional Multi-Threading

- Threads alternate
  - Nothing shared

Charles Grassl, IBM
Simultaneous Multi-Threading

- Simultaneous execution
  - Shared registers
  - Shared functional units

Charles Grassl, IBM
AMD OPTERON™ 6200 SERIES PROCESSOR ("INTERLAGOS")

- Multi-Chip Module (MCM) Package
  - 8, 12, & 16 core models
  - 4 DDR3 memory channels supporting LRDIMM, ULV-DIMM, UDIMM, & RDIMM
  - 16M L3 cache (Up to 32M L2+L3 cache)

FPUs are shared between two cores

Same platform as AMD Opteron™ 6100 Series processor.

Note: Graphic may not be fully representative of actual layout

From: AMD “Bulldozer” Technology, ©2011 AMD
Performance Analysis on IBM Power 6

- Compilation:
  - `mpxlf_r -O4 -qarch=pwr6 -qtune=pwr6 -qsmp=omp -pg`
- Execution:
  - `export OMP_NUM_THREADS 4`
  - `poe launch $PBS_O_WORKDIR./sp.C.16x4.exe`
  - Generates a file `gmount.MPI_RANK.out` for each MPI Process
- Generate report:
  - `gprof sp.C.16x4.exe gmon*`

<table>
<thead>
<tr>
<th></th>
<th>cumulative</th>
<th>self</th>
<th>self</th>
<th>total</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>seconds</td>
<td>seconds</td>
<td>calls</td>
<td>ms/call</td>
<td>ms/call</td>
</tr>
<tr>
<td>16.7</td>
<td>117.94</td>
<td>117.94</td>
<td>205245</td>
<td>0.57</td>
<td>0.57 .@10@x_solve@OL@1 [2]</td>
</tr>
<tr>
<td>14.6</td>
<td>221.14</td>
<td>103.20</td>
<td>205064</td>
<td>0.50</td>
<td>0.50 .@15@z_solve@OL@1 [3]</td>
</tr>
<tr>
<td>12.1</td>
<td>307.14</td>
<td>86.00</td>
<td>205200</td>
<td>0.42</td>
<td>0.42 .@12@y_solve@OL@1 [4]</td>
</tr>
<tr>
<td>6.2</td>
<td>350.83</td>
<td>43.69</td>
<td>205300</td>
<td>0.21</td>
<td>0.21 .@8@compute_rhs@OL@1@OL@6 [5]</td>
</tr>
</tbody>
</table>
Conclusions:

• **BT-MZ:**
  - Inherent workload imbalance on MPI level
  - \#nprocs = \#nzones yields poor performance
  - \#nprocs < \#zones => better workload balance, but decreases parallelism
  - Hybrid MPI/OpenMP yields better load-balance, maintains amount of parallelism

• **SP-MZ:**
  - No workload imbalance on MPI level, pure MPI should perform best
  - MPI/OpenMP outperforms MPI on some platforms due contention to network access within a node

• **LU-MZ:**
  - Hybrid MPI/OpenMP increases level of parallelism

• **“Best of category” depends on many factors**
  - Depends on many factors
  - Hard to predict
  - Good thread affinity is essential
Outline

• Introduction / Motivation
• Programming models on clusters of SMP nodes
• Case Studies / pure MPI vs hybrid MPI+OpenMP

• Practical “How-To” on hybrid programming
  Georg Hager, Regionales Rechenzentrum Erlangen (RRZE)

• Mismatch Problems
• Application categories that can benefit from hybrid parallelization
• Thread-safety quality of MPI libraries
• Tools for debugging and profiling MPI+OpenMP
• Other options on clusters of SMP nodes
• Summary
Hybrid Programming How-To: Overview

• A practical introduction to hybrid programming
  – How to compile and link
  – Getting a hybrid program to run on a cluster

• Running hybrid programs efficiently on multi-core clusters
  – Affinity issues
    • ccNUMA
    • Bandwidth bottlenecks
  – Intra-node MPI/OpenMP anisotropy
    • MPI communication characteristics
    • OpenMP loop startup overhead
  – Thread/process binding
How to compile, link and run

- Use appropriate OpenMP compiler switch (-openmp, -xopenmp, -mp, -qsmp=openmp, …) and MPI compiler script (if available)
- Link with MPI library
  - Usually wrapped in MPI compiler script
  - If required, specify to link against thread-safe MPI library
    - Often automatic when OpenMP or auto-parallelization is switched on
- Running the code
  - Highly non-portable! Consult system docs! (if available…)
  - If you are on your own, consider the following points
  - Make sure OMP_NUM_THREADS etc. is available on all MPI processes
    - Start “env VAR=VALUE … <YOUR BINARY>” instead of your binary alone
    - Use Pete Wyckoff’s mpiexec MPI launcher (see below): http://www.osc.edu/~pw/mpiexec
  - Figure out how to start less MPI processes than cores on your nodes
Some examples for compilation and execution (1)

- **NEC SX9**
  - NEC SX9 compiler
  - `mpif90 -C hopt -P openmp ... # -ftrace for profiling info`
  - Execution:

    ```
    $ export OMP_NUM_THREADS=<num_threads>
    $ MPIEXPORT="OMP_NUM_THREADS"
    $ mpirun -nn <# MPI procs per node> -nnp <# of nodes> a.out
    ```

- **Standard Intel Xeon cluster (e.g. @HLRS):**
  - Intel Compiler
  - `mpif90 -openmp ...`
  - Execution (handling of OMP_NUM_THREADS, see next slide):

    ```
    $ mpirun_ssh -np <num MPI procs> -hostfile machines a.out
    ```
Handling of OMP_NUM_THREADS

• **without** any support by mpirun:
  – E.g. with mpich-1
  – Problem: mpirun has no features to export environment variables to the via ssh automatically started MPI processes
  – Solution: Set
    ```bash
    export OMP_NUM_THREADS=<# threads per MPI process>
    ```
in ~/.bashrc (if a bash is used as login shell)
  – If you want to set OMP_NUM_THREADS individually when starting the MPI processes:
    • Add
      ```bash
      test -s ~/myexports && . ~/myexports
      ```
in your ~/.bashrc
    • Add
      ```bash
      echo '$OMP_NUM_THREADS=<# threads per MPI process>' > ~/myexports
      ```
      before invoking mpirun
    • Caution: Several invocations of mpirun cannot be executed at the same time with this trick!
Some examples for compilation and execution (3)

Handling of OMP_NUM_THREADS (continued)

- **with** support by OpenMPI – x option:

  ```bash
  export OMP_NUM_THREADS= <# threads per MPI process>
  mpiexec -x OMP_NUM_THREADS -n <# MPI processes> ./executable
  ```
Some examples for compilation and execution (4)

- Sun Constellation Cluster:
  - `mpif90 -fastsse -tp barcelona-64 -mp ...
  - SGE Batch System
  - `setenv OMP_NUM_THREADS`
  - `ibrun numactl.sh a.out`
  - Details see TACC Ranger User Guide (www.tacc.utexas.edu/services/userguides/ranger/#numactl)

- Cray XT5:
  - `ftn -fastsse -tp barcelona-64 -mp=nonuma ...
  - `aprun -n nprocs -N nprocs_per_node a.out`
Interlude: Advantages of mpiexec or similar mechanisms

- Uses PBS/Torque Task Manager (“TM”) interface to spawn MPI processes on nodes
  - As opposed to starting remote processes with ssh/rsh:
    - Correct CPU time accounting in batch system
    - Faster startup
    - Safe process termination
    - Understands PBS per-job nodefile
    - Allowing password-less user login not required between nodes
  - Support for many different types of MPI
    - All MPICHs, MVAPICHs, Intel MPI, …
  - Interfaces directly with batch system to determine number of procs
  - Downside: If you don’t use PBS or Torque, you’re out of luck…
- Provisions for starting less processes per node than available cores
  - Required for hybrid programming
  - “-pernode” and “-npernode #” options – does not require messing around with nodefiles
Running the code

Examples with mpiexec

• Example for using mpiexec on a dual-socket quad-core cluster:

  $ export OMP_NUM_THREADS=8
  $ mpiexec -pernode ./a.out

• Same but 2 MPI processes per node:

  $ export OMP_NUM_THREADS=4
  $ mpiexec -npernode 2 ./a.out

• Pure MPI:

  $ export OMP_NUM_THREADS=1 # or nothing if serial code
  $ mpiexec ./a.out
Running the code **efficiently**?

- Symmetric, UMA-type compute nodes have become rare animals
  - NEC SX
  - Intel 1-socket ("Port Townsend/Melstone/Lynnfield") – see case studies
  - Hitachi SR8000, IBM SP2, single-core multi-socket Intel Xeon… (all dead)
- Instead, systems have become “non-isotropic” on the node level
  - ccNUMA (AMD Opteron, SGI Altix, IBM Power6 (p575), Intel Nehalem)
- Multi-core, multi-socket
  - Shared vs. separate caches
  - Multi-chip vs. single-chip
  - Separate/shared buses
Issues for running code efficiently on “non-isotropic” nodes

• **ccNUMA** locality effects
  – Penalties for inter-LD access
  – Impact of contention
  – Consequences of file I/O for page placement
  – Placement of MPI buffers

• Multi-core / multi-socket *anisotropy* effects
  – Bandwidth bottlenecks, shared caches
  – Intra-node MPI performance
    • Core ↔ core vs. socket ↔ socket
  – OpenMP loop overhead depends on mutual position of threads in team
A short introduction to ccNUMA

- ccNUMA:
  - whole memory is *transparency accessible* by all processors
  - but *physically distributed*
  - with varying bandwidth and latency
  - and *potential contention* (shared memory paths)
Example: HP DL585 G5
4-socket ccNUMA Opteron 8220 Server

- **CPU**
  - 64 kB L1 per core
  - 1 MB L2 per core
  - No shared caches
  - On-chip memory controller (MI)
  - 10.6 GB/s local memory bandwidth

- **HyperTransport** 1000 network
  - 4 GB/s per link per direction

- **3 distance categories** for core-to-memory connections:
  - same LD
  - 1 hop
  - 2 hops

- **Q1**: What are the real penalties for non-local accesses?
- **Q2**: What is the impact of contention?
Effect of non-local access on HP DL585 G5:
Serial vector triad $A(:) = B(:) + C(:) \times D(:)$

Local

1 hop

2 hops

MFlops/s
Contestion vs. parallel access on HP DL585 G5:

OpenMP vector triad $A(:)=B(:)+C(:)*D(:)$

- In-cache performance unharmed by ccNUMA
- Single LD saturated by 2 cores!
- Perfect scaling across LDs

Affinity matters!
ccNUMA Memory Locality Problems

- **Locality of reference** is key to scalable performance on ccNUMA
  - Less of a problem with pure MPI, but see below
- What factors can destroy locality?
- MPI programming:
  - processes lose their association with the CPU the mapping took place on originally
  - OS kernel tries to maintain strong affinity, but sometimes fails
- **Shared Memory Programming** (OpenMP, hybrid):
  - threads losing association with the CPU the mapping took place on originally
  - improper initialization of distributed data
  - Lots of extra threads are running on a node, especially for hybrid
- All cases:
  - Other agents (e.g., OS kernel) may fill memory with data that prevents optimal placement of user data
Avoiding locality problems

• How can we make sure that memory ends up where it is close to the CPU that uses it?
  – See the following slides

• How can we make sure that it stays that way throughout program execution?
  – See end of section
Solving Memory Locality Problems: First Touch

- "Golden Rule" of ccNUMA:
  A memory page gets mapped into the local memory of the processor that first touches it!
  - Except if there is not enough local memory available
  - this might be a problem, see later
  - Some OSs allow to influence placement in more direct ways
    - cf. libnuma (Linux), MPO (Solaris), ...

- Caveat: "touch" means "write", not "allocate"
- Example:

  ```c
  double *huge = (double*)malloc(N*sizeof(double));
  // memory not mapped yet
  for(i=0; i<N; i++) // or i+=PAGE_SIZE
    huge[i] = 0.0; // mapping takes place here!
  ```

- It is sufficient to touch a single item to map the entire page
ccNUMA problems beyond first touch

- OS uses part of main memory for disk buffer (FS) cache
  - If FS cache fills part of memory, apps will probably allocate from foreign domains
  - → non-local access!
  - Locality problem even on hybrid and pure MPI with “asymmetric” file I/O, i.e. if not all MPI processes perform I/O

- Remedies
  - Drop FS cache pages after user job has run (admin’s job)
    - Only prevents cross-job buffer cache “heritage”
  - “Sweeper” code (run by user)
  - Flush buffer cache after I/O if necessary (“sync” is not sufficient!)
ccNUMA problems beyond first touch

- Real-world example: ccNUMA vs. UMA and the Linux buffer cache
- Compare two 4-way systems: AMD Opteron ccNUMA vs. Intel UMA, 4 GB main memory
- Run 4 concurrent triads (512 MB each) after writing a large file
- Report performance vs. file size
- Drop FS cache after each data point

![Graph showing performance vs. disk cache size]
Intra-node MPI characteristics: IMB Ping-Pong benchmark

- Code (to be run on 2 processors):

```fortran
wc = MPI_WTIME()

do i=1,NREPEAT
    if(rank.eq.0) then
        MPI_SEND(buffer,N,MPI_BYTE,1,0,MPI_COMM_WORLD,ierr)
        MPI_RECV(buffer,N,MPI_BYTE,1,0,MPI_COMM_WORLD,
                  status,ierr)
    else
        MPI_RECV(...)
        MPI_SEND(...)
    endif
endo
dc = MPI_WTIME() - wc
```

- Intranode (1S): `mpirun -np 2 -pin "1 3" ./a.out`
- Intranode (2S): `mpirun -np 2 -pin "2 3" ./a.out`
- Internode: `mpirun -np 2 -pernode ./a.out`
IMB Ping-Pong: Latency
_Intra-node vs. Inter-node on Woodcrest DDR-IB cluster_ (Intel MPI 3.1)

Affinity matters!

<table>
<thead>
<tr>
<th>Latency [µs]</th>
<th>IB internode</th>
<th>IB intranode 2S</th>
<th>IB intranode 1S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.24</td>
<td>0.55</td>
<td>0.31</td>
</tr>
</tbody>
</table>
IMB Ping-Pong: Bandwidth Characteristics

*Intra-node vs. Inter-node on Woodcrest DDR-IB cluster* (Intel MPI 3.1)

- inter-node
- inter-socket
- revolving buffers
- intra-socket

**Shared cache advantage**

- Between two cores of one socket
- Between two sockets of one node
- Between two nodes via InfiniBand

**Affinity matters!**
OpenMP Overhead

- As with intra-node MPI, OpenMP loop start overhead varies with the mutual position of threads in a team
- Possible variations
  - Intra-socket vs. inter-socket
  - Different overhead for “parallel for” vs. plain “for”
  - If one multi-threaded MPI process spans multiple sockets,
    - ... are neighboring threads on neighboring cores?
    - ... or are threads distributed “round-robin” across cores?

- Test benchmark: Vector triad

```c
#pragma omp parallel
for(int j=0; j < NITER; j++){
  #pragma omp (parallel) for
  for(i=0; i < N; ++i)
    a[i]=b[i]+c[i]*d[i];
  if(OBSCURE)
    dummy(a,b,c,d);
}
```

Look at performance for small array sizes!
OpenMP Overhead

Nomenclature:

- 1S/2S: 1-/2-socket
- RR: round-robin
- SS: socket-socket
- inner: parallel on inner loop

OMP overhead can be comparable to MPI latency!

Affinity matters!
Most of the OpenMP overhead is barrier sync!

*But how much is it exactly, and does it depend on the topology?*

**Overhead in cycles:**

<table>
<thead>
<tr>
<th></th>
<th>Q9550</th>
<th>i7 920 (shared L3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Threads (pthreads_barrier_wait)</td>
<td>42533</td>
<td>9820</td>
</tr>
<tr>
<td>omp barrier (icc 11.0)</td>
<td>977</td>
<td>814</td>
</tr>
<tr>
<td>gcc 4.4.3</td>
<td>41154</td>
<td>8075</td>
</tr>
</tbody>
</table>

pthreads/gcc → OS kernel call 😞

OpenMP & Intel compiler 😊

<table>
<thead>
<tr>
<th>Nehalem 2 Threads</th>
<th>Shared SMT threads</th>
<th>shared L3</th>
<th>different socket</th>
</tr>
</thead>
<tbody>
<tr>
<td>(pthreads_barrier_wait)</td>
<td>23352</td>
<td>4796</td>
<td>49237</td>
</tr>
<tr>
<td>omp barrier (icc 11.0)</td>
<td>2761</td>
<td>479</td>
<td>1206</td>
</tr>
</tbody>
</table>

- SMT can be a performance problem for synchronizing threads
- Topology has an influence on overhead!
Thread/Process Affinity ("Pinning")

- Highly OS-dependent system calls
  - But available on all systems
    - Linux: `sched_setaffinity()`, PLPA (see below) \(\rightarrow\) `hwloc`
    - Solaris: `processor_bind()`
    - Windows: `SetThreadAffinityMask()`
    - ...
- Support for "semi-automatic" pinning in some compilers/environments
  - Intel compilers > V9.1 (`KMP_AFFINITY` environment variable)
  - Pathscale
  - SGI Altix `dplace` (works with logical CPU numbers!)
  - Generic Linux: `taskset`, `numactl`, `likwid-pin` (see below)
- Affinity awareness in MPI libraries
  - SGI MPT
  - OpenMPI
  - Intel MPI
  - ...

Seen on SUN Ranger slides

Widely usable example: Using PLPA under Linux!
Explicit Process/Thread Binding With PLPA on Linux:
http://www.open-mpi.org/software/plpa/

- **Portable Linux Processor Affinity**
- Wrapper library for `sched_*affinity()` functions
  - Robust against changes in kernel API
- Example for pure OpenMP: Pinning of threads

```c
#include <plpa.h>
...
#pragma omp parallel
{
    #pragma omp critical
    {
        if(PLPA_NAME(api_probe)() != PLPA_PROBE_OK) {
            cerr << "PLPA failed!" << endl; exit(1);
        }
        plpa_cpu_set_t msk;
        PLPA_CPU_ZERO(&msk);
        int cpu = omp_get_thread_num();
        PLPA_CPU_SET(cpu, &msk);
        PLPA_NAME(sched_setaffinity)((pid_t)0, sizeof(cpu_set_t), &msk);
    }
}
```

**Care about correct core numbering!**
0…N-1 is not always contiguous! If required, reorder by a map:
```c
cpu = map[cpu];
```

**Pinning available?**

**Which CPU to run on?**

**Pin “me”**
Process/Thread Binding With PLPA

- Example for pure MPI: Process pinning
  - Bind MPI processes to cores in a cluster of 2x2-core machines

```c
MPI_Comm_rank(MPI_COMM_WORLD, &rank);
int mask = (rank % 4);
PLPA_CPU_SET(mask, &msk);
PLPA_NAME(sched_setaffinity)((pid_t)0,
    sizeof(cpu_set_t), &msk);
```

- Hybrid case:

```c
MPI_Comm_rank(MPI_COMM_WORLD, &rank);
#pragma omp parallel
{
    plpa_cpu_set_t msk;
    PLPA_CPU_ZERO(&msk);
    int cpu = (rank % MPI_PROCESSES_PER_NODE)*omp_num_threads
        +omp_get_thread_num();
    PLPA_CPU_SET(cpu, &msk);
    PLPA_NAME(sched_setaffinity)((pid_t)0, sizeof(cpu_set_t), &msk);
}
```
How do we figure out the topology?

• … and how do we enforce the mapping without changing the code?
• Compilers and MPI libs may still give you ways to do that

• But LIKWID supports all sorts of combinations:

  Like
  I
  Knew
  What
  I’m
  Doing

• Open source tool collection (developed at RRZE):

  http://code.google.com/p/likwid
Likwid Tool Suite

- Command line tools for Linux:
  - works with standard Linux 2.6 kernel
  - supports Intel and AMD CPUs
  - Supports all compilers whose OpenMP implementation is based on pthreads

- Current tools:
  - `likwid-topology`: Print thread and cache topology (similar to lstopo from the hwloc package)
  - `likwid-pin`: Pin threaded application without touching code
  - `likwid-perfctr`: Measure performance counters
  - `likwid-perfscope`: Performance oscilloscope w/ real-time display
  - `likwid-powermeter`: Current power consumption of chip (alpha stage)
  - `likwid-features`: View and enable/disable hardware prefetchers
  - `likwid-bench`: Low-level bandwidth benchmark generator tool
  - `likwid-mpirun`: mpirun wrapper script for easy LIKWID integration
likwid-topology – Topology information

• Based on cpuid information

• Functionality:
  – Measured clock frequency
  – Thread topology
  – Cache topology
  – Cache parameters (-c command line switch)
  – ASCII art output (-g command line switch)

• Currently supported:
  – Intel Core 2 (45nm + 65 nm)
  – Intel Nehalem
  – AMD K10 (Quadcore and Hexacore)
  – AMD K8
Output of likwid-topology

**CPU name:** Intel Core i7 processor  
**CPU clock:** 2666683826 Hz

****************************
**Hardware Thread Topology**
****************************

**Sockets:** 2  
**Cores per socket:** 4  
**Threads per core:** 2

<table>
<thead>
<tr>
<th>HWThread</th>
<th>Thread</th>
<th>Core</th>
<th>Socket</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
likwid-topology continued

Socket 0: ( 0 1 2 3 4 5 6 7 )
Socket 1: ( 8 9 10 11 12 13 14 15 )

*************************************************************
Cache Topology
*************************************************************

Level: 1
Size: 32 kB
Cache groups: ( 0 1 ) ( 2 3 ) ( 4 5 ) ( 6 7 ) ( 8 9 ) ( 10 11 ) ( 12 13 ) ( 14 15 )

Level: 2
Size: 256 kB
Cache groups: ( 0 1 ) ( 2 3 ) ( 4 5 ) ( 6 7 ) ( 8 9 ) ( 10 11 ) ( 12 13 ) ( 14 15 )

Level: 3
Size: 8 MB
Cache groups: ( 0 1 2 3 4 5 6 7 ) ( 8 9 10 11 12 13 14 15 )

• ... and also try the ultra-cool \(-g\) option!
likwid-pin

- Inspired and based on `ptoverride` (Michael Meier, RRZE) and `taskset`
- Pins process and threads to specific cores without touching code
- Directly supports pthreads, gcc OpenMP, Intel OpenMP
- Allows user to specify skip mask (i.e., supports many different compiler/MPI combinations)
- Can also be used as replacement for `taskset`
- Uses logical (contiguous) core numbering when running inside a restricted set of cores
- Supports logical core numbering inside node, socket, core

Usage examples:

- `env OMP_NUM_THREADS=6 likwid-pin -t intel -c 0,2,4-6 ./myApp parameters`
- `env OMP_NUM_THREADS=6 likwid-pin -c S0:0-2@S1:0-2 ./myApp`
- `env OMP_NUM_THREADS=2 mpirun -npernode 2 \likwid-pin -s 0x3 -c 0,1 ./myApp parameters`
Example: STREAM benchmark on 12-core Intel Westmere. Anarchy vs. thread pinning.
Topology ("mapping") choices with MPI+OpenMP: More examples using Intel MPI+compiler & home-grown mpirun

One MPI process per node

env OMP_NUM_THREADS=8 mpirun -pernode \likwid-pin -t intel -c 0-7 ./a.out

One MPI process per socket

env OMP_NUM_THREADS=4 mpirun -nperrnode 2 \-pin "0,1,2,3_4,5,6,7" ./a.out

OpenMP threads pinned "round robin" across cores in node

env OMP_NUM_THREADS=4 mpirun -nperrnode 2 \-pin "0,1,4,5_2,3,6,7" \likwid-pin -t intel -c 0,2,1,3 ./a.out

Two MPI processes per socket

env OMP_NUM_THREADS=2 mpirun -nperrnode 4 \-pin "0,1_2,3_4,5_6,7" \likwid-pin -t intel -c 0,1 ./a.out
MPI/OpenMP hybrid “how-to”: Take-home messages

• Do not use hybrid if the pure MPI code scales ok

• Be aware of intranode MPI behavior
• Always observe the topology dependence of
  – Intranode MPI
  – OpenMP overheads
• Enforce proper thread/process to core binding, using appropriate tools (whatever you use, but use SOMETHING)
• Multi-LD OpenMP processes on ccNUMA nodes require correct page placement

• Finally: Always compare the best pure MPI code with the best OpenMP code!
Outline

• Introduction / Motivation
• Programming models on clusters of SMP nodes
• Case Studies / pure MPI vs hybrid MPI+OpenMP
• Practical “How-To” on hybrid programming

• Mismatch Problems

• Opportunities:
  Application categories that can benefit from hybrid parallelization
• Thread-safety quality of MPI libraries
• Tools for debugging and profiling MPI+OpenMP
• Other options on clusters of SMP nodes
• Summary
Mismatch Problems

- None of the programming models fits to the hierarchical hardware (cluster of SMP nodes)
- Several mismatch problems → following slides
- Benefit through hybrid programming → Opportunities, see next section
- Quantitative implications → depends on your application

<table>
<thead>
<tr>
<th>Examples</th>
<th>No.1</th>
<th>No.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benefit through hybrid (see next section)</td>
<td>30%</td>
<td>10%</td>
</tr>
<tr>
<td>Loss by mismatch problems</td>
<td>−10%</td>
<td>−25%</td>
</tr>
<tr>
<td>Total</td>
<td>+20%</td>
<td>−15%</td>
</tr>
</tbody>
</table>

In most cases: Both categories!
The Topology Problem with pure MPI

Application example on 80 cores:
- Cartesian application with $5 \times 16 = 80$ sub-domains
- On system with $10 \times$ dual socket $\times$ quad-core

17 x inter-node connections per node
1 x inter-socket connection per node

Sequential ranking of MPI_COMM_WORLD

Does it matter?
The Topology Problem with

Application example on 80 cores:
- Cartesian application with $5 \times 16 = 80$ sub-domains
- On system with $10 \times$ dual socket $\times$ quad-core

pure MPI
one MPI process on each core

Round robin ranking of MPI_COMM_WORLD

\begin{itemize}
  \item 32 x inter-node connections per node
  \item 0 x inter-socket connection per node
\end{itemize}
The Topology Problem with pure MPI
one MPI process on each core

Application example on 80 cores:
• Cartesian application with $5 \times 16 = 80$ sub-domains
• On system with $10 \times$ dual socket $\times$ quad-core

Two levels of domain decomposition

12 x inter-node connections per node
4 x inter-socket connection per node

Bad affinity of cores to thread ranks
The Topology Problem with

Application example on 80 cores:
- Cartesian application with $5 \times 16 = 80$ sub-domains
- On system with $10 \times$ dual socket $\times$ quad-core

One MPI process on each core

12 x inter-node connections per node
2 x inter-socket connection per node

Good affinity of cores to thread ranks
The Topology Problem with hybrid MPI+OpenMP

MPI: inter-node communication
OpenMP: inside of each SMP node

Exa.: 2 SMP nodes, 8 cores/node

Problem
- Does application topology inside of SMP parallelization fit on inner hardware topology of each SMP node?

Solutions:
- Domain decomposition inside of each thread-parallel MPI process, and
- first touch strategy with OpenMP

Successful examples:
- Multi-Zone NAS Parallel Benchmarks (MZ-NPB)
The Topology Problem with hybrid MPI+OpenMP

Application example:
• Same Cartesian application aspect ratio: 5 x 16
• On system with 10 x dual socket x quad-core
• 2 x 5 domain decomposition

MPI: inter-node communication
OpenMP: inside of each SMP node

3 x inter-node connections per node, but ~ 4 x more traffic
2 x inter-socket connection per node

Affinity of cores to thread ranks !!!
Numerical Optimization inside of an SMP node

2nd level of domain decomposition: OpenMP

3rd level: 2nd level cache

4th level: 1st level cache

Optimizing the numerical performance
The Mapping Problem with mixed model

Several multi-threaded MPI process per SMP node:

Problem
- Where are your processes and threads really located?

Solutions:
- Depends on your platform,
- e.g., with `numactl`

→ Case study on Sun Constellation Cluster Ranger with BT-MZ and SP-MZ

Further questions:
- Where is the NIC\(^1\) located?
- Which cores share caches?

\(^1\) NIC = Network Interface Card
Unnecessary intra-node communication

Problem:
- If several MPI processes on each SMP node
  → unnecessary intra-node communication

Solution:
- Only one MPI process per SMP node

Remarks:
- MPI library must use appropriate fabrics / protocol for intra-node communication
- Intra-node bandwidth higher than inter-node bandwidth
  → problem may be small
- MPI implementation may cause unnecessary data copying
  → waste of memory bandwidth

Quality aspects of the MPI library

Mixed model
(several multi-threaded MPI processes per SMP node)
Problem 1:
- Can the master thread saturate the network?
Solution:
- If not, use mixed model
  - i.e., several MPI processes per SMP node

Problem 2:
- Sleeping threads are wasting CPU time
Solution:
- Overlapping of computation and communication

Problem 1&2 together:
- Producing more idle time through lousy bandwidth of master thread

for (iteration ....) {
    #pragma omp parallel
    numerical code
    /*end omp parallel */

    /* on master thread only */
    MPI_Send (original data to halo areas in other SMP nodes)
    MPI_Recv (halo data from the neighbors)
} /*end for loop
OpenMP: Additional Overhead & Pitfalls

- Using OpenMP
  - may prohibit compiler optimization
  - may cause significant loss of computational performance

- Thread fork / join overhead

- On ccNUMA SMP nodes:
  - Loss of performance due to missing memory page locality or missing first touch strategy
  - E.g. with the masteronly scheme:
    - One thread produces data
    - Master thread sends the data with MPI
    - data may be internally communicated from one memory to the other one

- Amdahl’s law for each level of parallelism

- Using MPI-parallel application libraries? → Are they prepared for hybrid?

See, e.g., the necessary –O4 flag with mpxlf_r on IBM Power6 systems
Three problems:

- **the application problem:**
  - one must separate application into:
    - code that can run before the halo data is received
    - code that needs halo data
  ➔ very hard to do !!!

- **the thread-rank problem:**
  - comm. / comp. via thread-rank
  - cannot use work-sharing directives
  ➔ loss of major OpenMP support

- **the load balancing problem**

```c
if (my_thread_rank < 1) {
    MPI_Send/Recv….
} else {
    my_range = (high-low-1) / (num_threads-1) + 1;
    my_low = low + (my_thread_rank+1)*my_range;
    my_high=high+ (my_thread_rank+1+1)*my_range;
    my_high = max(high, my_high)
    for (i=my_low; i<my_high; i++) {
        …. 
    }
}
```
Overlapping Communication and Computation

MPI communication by one or a few threads while other threads are computing

Subteams

• Important proposal for OpenMP 3.x or OpenMP 4.x

Barbara Chapman et al.: Toward Enhancing OpenMP’s Work-Sharing Directives.

```c
#pragma omp parallel
{
#pragma omp single onthreads( 0 )
    {
        MPI_Send/Recv… .
    }
#pragma omp for onthreads( 1 : omp_get_numthreads()-1 )
    for (………)
        {
            /* work without halo information */
            } /* barrier at the end is only inside of the subteam */
    …
#pragma omp barrier
#pragma omp for
    for (………)
        {
            /* work based on halo information */
            }
} /* end omp parallel */
```
Parallel Programming Models on Hybrid Platforms

- **pure MPI**
  - one MPI process on each core

- **hybrid MPI+OpenMP**
  - MPI: inter-node communication
  - OpenMP: inside of each SMP node

- **OpenMP only**
  - distributed virtual shared memory

- **No overlap of Comm. + Comp.**
  - MPI only outside of parallel regions of the numerical application code

- **Overlapping Comm. + Comp.**
  - MPI communication by one or a few threads while other threads are computing

- **Masteronly**
  - MPI only outside of parallel regions

- **Multiple/only**
  - appl. threads
  - inside of MPI

- **Funneled**
  - MPI only on master-thread

- **Multiple**
  - more than one thread may communicate

- **Funneled & Reserved**
  - reserved thread for communication

- **Funneled with Full Load Balancing**

- **Multiple & Reserved**
  - reserved threads for communication

- **Multiple with Full Load Balancing**

- **Different strategies to simplify the load balancing**

Hybrid Parallel Programming
Slide 94 / 154
Rabenseifner, Hager, Jost
Experiment: Matrix-vector-multiply (MVM)

- Jacobi-Davidson-Solver on IBM SP Power3 nodes with 16 CPUs per node
- funneled&reserved is always faster in this experiments
- Reason: Memory bandwidth is already saturated by 15 CPUs, see inset
- Inset: Speedup on 1 SMP node using different number of threads

NEW OpenMP Tasking Model gives a new way to achieve more parallelism form hybrid computation.

Case study: Communication and Computation in Gyrokinetic Tokamak Simulation (GTS) shift routine

Work on particle array (packing for sending, reordering, adding after sending) can be overlapped with data independent MPI communication using OpenMP tasks.

Slides, courtesy of Alice Koniges, NERSC, LBNL
Overlapping can be achieved with OpenMP tasks (1st part)

Overlapping MPI_Allreduce with particle work

• **Overlap**: Master thread encounters (!$omp master) tasking statements and creates work for the thread team for deferred execution. MPI Allreduce call is immediately executed.
• MPI implementation has to support at least MPI_THREAD_FUNNELED
• Subdividing tasks into smaller chunks to allow better *load balancing* and *scalability* among threads.
Overlapping can be achieved with OpenMP tasks (2nd part)

Particle reordering of remaining particles (above) and adding sent particles into array (right) & sending or receiving of shifted particles can be independently executed.

Overlapping particle reordering

Overlapping remaining MPI_Sendrecv

Slides, courtesy of Alice Koniges, NERSC, LBNL
OpenMP tasking version outperforms original shifter, especially in larger poloidal domains

- Performance breakdown of GTS shifter routine using 4 OpenMP threads per MPI process with varying domain decomposition and particles per cell on Franklin Cray XT4.
- MPI communication in the shift phase uses a **toroidal MPI communicator** (constantly 128).
- Large performance differences in the 256 MPI run compared to 2048 MPI run!
- Speed-Up is expected to be higher on larger GTS runs with hundreds of thousands CPUs since MPI communication is more expensive.
OpenMP/DSM

- Distributed shared memory (DSM)
- Distributed virtual shared memory (DVSM)
- Shared virtual memory (SVM)

Principles
- emulates a shared memory
- on distributed memory hardware

Implementations
- e.g., Intel® Cluster OpenMP
Basic idea:

- Between OpenMP barriers, data exchange is not necessary, i.e., visibility of data modifications to other threads only after synchronization.
- When a page of sharable memory is not up-to-date, it becomes *protected*.
- Any access then faults (SIGSEGV) into Cluster OpenMP runtime library, which requests info from remote nodes and updates the page.
- Protection is removed from page.
- Instruction causing the fault is re-started, this time successfully accessing the data.
Comparison:

MPI based parallelization ↔ DSM

- MPI based:
  - Potential of boundary exchange between two domains in one large message
    - Dominated by **bandwidth** of the network
- DSM based (e.g. Intel® Cluster OpenMP):
  - Additional latency based overhead in each barrier
    - May be marginal
  - Communication of **updated data of pages**
    - Not all of this data may be needed
    - i.e., too much data is transferred
    - Packages may be too small
    - Significant latency
  - Communication not oriented on boundaries of a domain decomposition
    - Probably more data must be transferred than necessary

*by rule of thumb:*

Communication may be 10 times slower than with MPI
Comparing results with heat example

- Normal OpenMP on shared memory (ccNUMA) NEC TX-7

heat_x.c / heatc2_x.c with OpenMP on NEC TX-7

- Speedup
- Threads
- Ideal speedup
Heat example: Cluster OpenMP Efficiency

- Cluster OpenMP on a Dual-Xeon cluster

Efficiency only with small communication foot-print

Up to 3 CPUs with 3000x3000

Terrible with non-default schedule

No speedup with 1000x1000

Second CPU only usable in small cases
Back to the mixed model – an Example

Topology-problem solved:
Only horizontal inter-node comm.

Still intra-node communication

Several threads per SMP node are communicating in parallel:
→ network saturation is possible

Additional OpenMP overhead

With Masteronly style:
75% of the threads sleep while master thread communicates

With Overlapping Comm.& Comp.:
Master thread should be reserved for communication only partially – otherwise too expensive

MPI library must support
– Multiple threads
– Two fabrics (shmem + internode)
No silver bullet

- The analyzed programming models do not fit on hybrid architectures
  - whether drawbacks are minor or major
    - depends on applications’ needs
  - But there are major opportunities → next section

- In the NPB-MZ case-studies
  - We tried to use optimal parallel environment
    - for pure MPI
    - for hybrid MPI+OpenMP
  - i.e., the developers of the MZ codes and we tried to minimize the mismatch problems
  → the opportunities in next section dominated the comparisons
Outline

- Introduction / Motivation
- Programming models on clusters of SMP nodes
- Case Studies / pure MPI vs hybrid MPI+OpenMP
- Practical “How-To” on hybrid programming
- Mismatch Problems

**Opportunities:**
Application categories that can benefit from hybrid parallelization

- Thread-safety quality of MPI libraries
- Tools for debugging and profiling MPI+OpenMP
- Other options on clusters of SMP nodes
- Summary
Nested Parallelism

- Example NPB: BT-MZ *(Block tridiagonal simulated CFD application)*
  - Outer loop:
    - limited number of zones \(\rightarrow\) limited parallelism
    - zones with different workload \(\rightarrow\) speedup \(\frac{\text{Sum of workload of all zones}}{\text{Max workload of a zone}}\)
  - Inner loop:
    - OpenMP parallelized (static schedule)
    - Not suitable for distributed memory parallelization

- Principles:
  - Limited parallelism on outer level
  - Additional inner level of parallelism
  - Inner level not suitable for MPI
  - Inner level may be suitable for static OpenMP worksharing
Load-Balancing
(on same or different level of parallelism)

- **OpenMP enables**
  - Cheap *dynamic* and *guided* load-balancing
  - Just a parallelization option (clause on omp for / do directive)
  - Without additional software effort
  - Without explicit data movement

- **On MPI level**
  - *Dynamic load balancing* requires moving of parts of the data structure through the network
  - Significant runtime overhead
  - Complicated software / therefore not implemented

- **MPI & OpenMP**
  - Simple static load-balancing on MPI level,
    dynamic or guided on OpenMP level  
    \[ \text{medium quality} \]
    \[ \text{cheap implementation} \]
Memory consumption

- Shared nothing
  - Heroic theory
  - In practice: Some data is duplicated

- **MPI & OpenMP**
  With n threads per MPI process:
  - Duplicated data may be reduced by factor n
Using more OpenMP threads could reduce the memory usage substantially, up to five times on Hopper Cray XT5 (eight-core nodes).

Memory consumption (continued)

• Future:
  With 100+ cores per chip the memory per core is limited.
  – Data reduction through usage of shared memory may be a key issue
  – Domain decomposition on each hardware level
    • Maximizes
      – Data locality
      – Cache reuse
    • Minimizes
      – ccNUMA accesses
      – Message passing
  – No halos between domains inside of SMP node
    • Minimizes
      – Memory consumption
How many threads per MPI process?

- SMP node = with **m sockets** and **n cores/socket**
- How many threads (i.e., cores) per MPI process?
  - Too many threads per MPI process
    → overlapping of MPI and computation may be necessary,
    → some NICs unused?
  - Too few threads
    → too much memory consumption (see previous slides)
- Optimum
  - somewhere between 1 and m x n threads per MPI process,
  - Typically:
    - **Optimum** = n, i.e., 1 MPI process per socket
    - Sometimes = n/2 i.e., 2 MPI processes per socket
    - Seldom = 2n, i.e., each MPI process on 2 sockets
Opportunities, if MPI speedup is limited due to algorithmic problems

- Algorithmic opportunities due to larger physical domains inside of each MPI process
  → If multigrid algorithm only inside of MPI processes
  → If separate preconditioning inside of MPI nodes and between MPI nodes
  → If MPI domain decomposition is based on physical zones
To overcome MPI scaling problems

- Reduced number of MPI messages, reduced aggregated message size compared to pure MPI
- MPI has a few scaling problems
  - Handling of more than 10,000 MPI processes
  - Irregular Collectives: MPI_...v(), e.g. MPI_Gatherv()
    - **Scaling applications should not use MPI_...v() routines**
  - MPI-2.1 Graph topology (MPI_Graph_create)
    - MPI-2.2 MPI_Dist_graph_create_adjacent
  - Creation of sub-communicators with MPI_Comm_create
    - **MPI-2.2 introduces a new scaling meaning of MPI_Comm_create**
  - ... see P. Balaji, et al.: **MPI on a Million Processors.** Proceedings EuroPVM/MPI 2009.
- Hybrid programming reduces all these problems (due to a smaller number of processes)
Summary: Opportunities of hybrid parallelization (MPI & OpenMP)

- Nested Parallelism
  → Outer loop with MPI / inner loop with OpenMP
- Load-Balancing
  → Using OpenMP *dynamic* and *guided* worksharing
- Memory consumption
  → Significantly reduction of replicated data on MPI level
- Opportunities, if MPI speedup is limited due to algorithmic problem
  → Significantly reduced number of MPI processes
- Reduced MPI scaling problems
  → Significantly reduced number of MPI processes
Outline

- Introduction / Motivation
- Programming models on clusters of SMP nodes
- Case Studies / pure MPI vs hybrid MPI+OpenMP
- Practical “How-To” on hybrid programming
- Mismatch Problems
- Opportunities:
  Application categories that can benefit from hybrid parallelization

- **Thread-safety quality of MPI libraries**
  - Tools for debugging and profiling MPI+OpenMP
  - Other options on clusters of SMP nodes
  - Summary
MPI rules with OpenMP / Automatic SMP-parallelization

• Special MPI-2 Init for multi-threaded MPI processes:

```c
int MPI_Init_thread( int * argc, char ** argv[],
            int thread_level_required,
            int * thread_level_provided);
int MPI_Query_thread( int * thread_level_provided);
int MPI_Is_main_thread(int * flag);
```

• REQUIRED values (increasing order):
  - `MPI_THREAD_SINGLE`: Only one thread will execute
  - `THREAD_MASTERONLY`:
    - (virtual value, not part of the standard)
    - MPI processes may be multi-threaded, but only master thread will make MPI-calls AND only while other threads are sleeping
  - `MPI_THREAD_FUNNELED`: Only master thread will make MPI-calls
  - `MPI_THREAD_SERIALIZED`: Multiple threads may make MPI-calls, but only one at a time
  - `MPI_THREAD_MULTIPLE`: Multiple threads may call MPI, with no restrictions

• returned `provided` may be less than REQUIRED by the application
Calling MPI inside of OMP MASTER

- Inside of a parallel region, with “OMP MASTER”
- Requires MPI_THREAD_FUNNELED, i.e., only master thread will make MPI-calls
- **Caution:** There isn’t any synchronization with “OMP MASTER”! Therefore, “OMP BARRIER” normally necessary to guarantee, that data or buffer space from/for other threads is available before/after the MPI call!

```
!$OMP BARRIER  #pragma omp barrier
!$OMP MASTER   #pragma omp master
    call MPI_Xxx(...)  MPI_Xxx(...);
!$OMP END MASTER
!$OMP BARRIER  #pragma omp barrier
```

- But this implies that all other threads are sleeping!
- The additional barrier implies also the necessary cache flush!
... the barrier is necessary –
example with MPI_Recv

```c
!$OMP PARALLEL
!$OMP DO
   do i=1,1000
      a(i) = buf(i)
   end do
!$OMP END DO NOWAIT
!$OMP BARRIER
!$OMP MASTER
   call MPI_RECV(buf,...)
!$OMP END MASTER
!$OMP BARRIER
!$OMP END PARALLEL

#pragma omp parallel
{  
#pragma omp for nowait
    for (i=0; i<1000; i++)
      a[i] = buf[i];

#pragma omp barrier
#pragma omp master
    MPI_Recv(buf,...);
#pragma omp barrier
#pragma omp for nowait
    for (i=0; i<1000; i++)
      c[i] = buf[i];
}
/* omp end parallel */
```
Thread support in MPI libraries

The following MPI libraries offer thread support:

<table>
<thead>
<tr>
<th>Implementation</th>
<th>Thread support level</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPIch-1.2.7p1</td>
<td>Always announces MPI_THREAD_FUNNELED.</td>
</tr>
<tr>
<td>MPIch2-1.0.8</td>
<td>ch3:sock supports MPI_THREAD_MULTIPLE</td>
</tr>
<tr>
<td>ch3:nemesis</td>
<td>ch:nemesis has “Initial Thread-support”</td>
</tr>
<tr>
<td>MPIch2-1.1.0a2</td>
<td>ch3:nemesis (default) has MPI_THREAD_MULTIPLE</td>
</tr>
<tr>
<td>Intel MPI 3.1</td>
<td>Full MPI_THREAD_MULTIPLE</td>
</tr>
<tr>
<td>SciCortex MPI</td>
<td>MPI_THREAD_FUNNELED</td>
</tr>
<tr>
<td>HP MPI-2.2.7</td>
<td>Full MPI_THREAD_MULTIPLE (with libmtmpi)</td>
</tr>
<tr>
<td>SGI MPT-1.14</td>
<td>Not thread-safe?</td>
</tr>
<tr>
<td>IBM MPI</td>
<td>Full MPI_THREAD_MULTIPLE</td>
</tr>
<tr>
<td>Nec MPI/SX</td>
<td>MPI_THREAD_SERIALIZED</td>
</tr>
</tbody>
</table>

Testsuites for thread-safety may still discover bugs in the MPI libraries.
Thread support within Open MPI

• In order to enable thread support in Open MPI, configure with:

```
configure --enable-mpi-threads
```

• This turns on:
  – Support for full MPI_THREAD_MULTIPLE
  – Internal checks when run with threads (``--enable-debug``)

```
configure --enable-mpi-threads --enable-progress-threads
```

• This (additionally) turns on:
  – Progress threads to asynchronously transfer/receive data per network BTL.

• Additional Feature:
  – Compiling **with** debugging support, but **without** threads will check for recursive locking
Outline

- Introduction / Motivation
- Programming models on clusters of SMP nodes
- Case Studies / pure MPI vs hybrid MPI+OpenMP
- Practical “How-To” on hybrid programming
- Mismatch Problems
- Opportunities: Application categories that can benefit from hybrid parallelization
- Thread-safety quality of MPI libraries

- Tools for debugging and profiling MPI+OpenMP

- Other options on clusters of SMP nodes
- Summary

This section is skipped, see talks on tools on Thursday
Thread Correctness – Intel ThreadChecker  1/3

- Intel ThreadChecker operates in a similar fashion to helgrind,
- Compile with `-tcheck`, then run program using `tcheck_cl`:

```
Application finished
```

<table>
<thead>
<tr>
<th>ID</th>
<th>Short Description</th>
<th>Severity</th>
<th>Context</th>
<th>Description</th>
<th>1st Acc</th>
<th>2nd Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Write -&gt; Error</td>
<td>1</td>
<td>pthread</td>
<td>Memory write of global_variable at pthread</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ad_race</td>
<td>pthread_race.c:31 conflicts with d_race</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

With new Intel Inspector XE 2011:
Command line interface for `mpirun` is undocumented.
Thread Correctness – Intel ThreadChecker  2/3

- One may output to HTML:

```
tcheck_cl --format HTML --report pthread_race.html pthread_race
```
Thread Correctness – Intel ThreadChecker 3/3

- If one wants to compile with threaded Open MPI (option for IB):

```
classification --enable-mpi-threads
--enable-debug
--enable-mca-no-build=memory-ptmalloc2
CC=icc F77=ifort FC=ifort
CFLAGS='--debug all -inline-debug-info tcheck'
CXXFLAGS='--debug all -inline-debug-info tcheck'
FFLAGS='--debug all -tcheck' LDFLAGS='tcheck'
```

- Then run with:

```
mpirun --mca tcp,sm,self -np 2 tcheck_cl
  --reinstrument -u full --format html
  --cache_dir '/tmp/my_username_${__tc_cl_cache}'
  --report 'tc_mpi_test_suite_${}'
  --options 'file=tc_my_executable_%H_%I,'
    pad=128, delay=2, stall=2'
  ./my_executable my_arg1 my_arg2 ...
```

 Courtesy of Rainer Keller, HLRS and ORNL — skipped —
Performance Tools Support for Hybrid Code

- Paraver examples have already been shown, tracing is done with linking against (closed-source) `omptrace` or `ompitrace`.

- For Vampir/Vampirtrace performance analysis:
  ```
  ./configure --enable-omp
  --enable-hyb
  --with-mpi-dir=/opt/OpenMPI/1.3-icc
  CC=icc F77=ifort FC=ifort
  (Attention: does not wrap MPI_Init_thread!)
  ```

Courtesy of Rainer Keller, HLRS and ORNL — skipped —
Scalasca – Example “Wait at Barrier”

Indication of non-optimal load balance
Scalasca – Example “Wait at Barrier”, Solution

Better load balancing with dynamic loop schedule
Outline

- Introduction / Motivation
- Programming models on clusters of SMP nodes
- Case Studies / pure MPI vs hybrid MPI+OpenMP
- Practical “How-To” on hybrid programming
- Mismatch Problems
- Opportunities:
  Application categories that can benefit from hybrid parallelization
- Thread-safety quality of MPI libraries
- Tools for debugging and profiling MPI+OpenMP

Other options on clusters of SMP nodes

- Summary
Pure MPI – multi-core aware

• Hierarchical domain decomposition (or distribution of Cartesian arrays)

Domain decomposition: 1 sub-domain / SMP node

Further partitioning: 1 sub-domain / socket

1 / core

Cache optimization: Blocking inside of each core, block size relates to cache size. 1-3 cache levels.

Example on 10 nodes, each with 4 sockets, each with 6 cores.
How to achieve a hierarchical domain decomposition (DD)?

- **Cartesian grids:**
  - Several levels of subdivide
  - Ranking of MPI_COMM_WORLD — three choices:
    a) Sequential ranks through original data structure
       + locating these ranks correctly on the hardware
       ➢ can be achieved with one-level DD on finest grid
       + special startup (mpiexec) with optimized rank-mapping
    b) Sequential ranks in comm_cart (from MPI_CART_CREATE)
       ➢ requires optimized MPI_CART_CREATE, or special startup (mpiexec) with optimized rank-mapping
    c) Sequential ranks in MPI_COMM_WORLD
       + additional communicator with sequential ranks in the data structure
       + self-written and optimized rank mapping.

- **Unstructured grids:**
  ➔ next slide
How to achieve a hierarchical domain decomposition (DD)?

• Unstructured grids:
  – Multi-level DD:
    • Top-down: Several levels of (Par)Metis \(\rightarrow\) not recommended
    • Bottom-up: Low level DD + higher level recombination
  – Single-level DD (finest level)
    • Analysis of the communication pattern in a first run (with only a few iterations)
    • Optimized rank mapping to the hardware before production run
    • E.g., with CrayPAT + CrayApprentice
Top-down – several levels of (Par)Metis
(not recommended)

Steps:

- Load-balancing (e.g., with ParMetis) on outer level, i.e., between all SMP nodes
- Independent (Par)Metis inside of each node
- Metis inside of each socket

➤ Subdivide does not care on balancing of the outer boundary

➤ processes can get a lot of neighbors with inter-node communication

➤ unbalanced communication
Bottom-up – Multi-level DD through recombination

1. Core-level DD: partitioning of application’s data grid
2. Socket-level DD: recombining of core-domains
3. SMP node level DD: recombining of socket-domains

- **Problem:** Recombination must **not** calculate patches that are smaller or larger than the average
- In this example the load-balancer **must** combine always
  - 6 cores, and
  - 4 sockets
- **Advantage:** Communication is balanced!
Profiling solution

- First run with profiling
  - Analysis of the communication pattern
- Optimization step
  - Calculation of an optimal mapping of ranks in MPI_COMM_WORLD to the hardware grid (physical cores / sockets / SMP nodes)
- Restart of the application with this optimized locating of the ranks on the hardware grid

- Example: CrayPat and CrayApprentice
Scalability of MPI to hundreds of thousands …

Weak scalability of pure MPI

- As long as the application does not use
  - MPI_ALLTOALL
  - MPI_<collectives>V  (i.e., with length arrays)

and application
- distributes all data arrays

one can expect:
- Significant, but still scalable memory overhead for halo cells.
- MPI library is internally scalable:
  - E.g., mapping ranks → hardware grid
    - Centralized storing in shared memory (OS level)
    - In each MPI process, only used neighbor ranks are stored (cached) in process-local memory.
  - Tree based algorithm with O(log N)
    - From 1000 to 1000,000 process O(Log N) only doubles!

The vendors will (or must) deliver scalable MPI libraries for their largest systems!
Remarks on Cache Optimization

- **After** all parallelization domain decompositions (DD, up to 3 levels) are done:
- Additional DD into data blocks
  - that fit to 2\textsuperscript{nd} or 3\textsuperscript{rd} level cache.
  - It is done inside of each MPI process (on each core).
  - Outer loops over these blocks
  - Inner loops inside of a block
- Cartesian example: 3-dim loop is split into

  \[
  \begin{align*}
  \text{do } i_{\text{block}} &= 1, n_i, \text{stride}_i \\
  \text{do } j_{\text{block}} &= 1, n_j, \text{stride}_j \\
  \text{do } k_{\text{block}} &= 1, n_k, \text{stride}_k \\
  \text{do } i &= i_{\text{block}}, \min(i_{\text{block}} + \text{stride}_i - 1, n_i) \\
  \text{do } j &= j_{\text{block}}, \min(j_{\text{block}} + \text{stride}_j - 1, n_j) \\
  \text{do } k &= k_{\text{block}}, \min(k_{\text{block}} + \text{stride}_k - 1, n_k) \\
  a(i, j, k) &= \mathcal{f}( b(i\pm0,1,2, j\pm0,1,2, k\pm0,1,2) )
  \end{align*}
  \]

  Access to 13-point stencil
Remarks on Cost-Benefit Calculation

Costs
• for optimization effort
  – e.g., additional OpenMP parallelization
  – e.g., 3 person month x 5,000 € = 15,000 € (full costs)

Benefit
• from reduced CPU utilization
  – e.g., Example 1:
    100,000 € hardware costs of the cluster
    x 20% used by this application over whole lifetime of the cluster
    x 7% performance win through the optimization
    = 1,400 € \(\Rightarrow\) total loss = 13,600 €
  – e.g., Example 2:
    10 Mio € system x 5% used x 8% performance win
    = 40,000 € \(\Rightarrow\) total win = 25,000 €
Remarks on MPI and PGAS (UPC & CAF)

- Parallelization always means
  - expressing locality.

- If the application has no locality,
  - Then the parallelization needs not to model locality
  → UPC with its round robin data distribution may fit

- If the application has locality,
  - then it must be expressed in the parallelization

- Coarray Fortran (CAF) expresses data locality explicitly through “co-dimension”:
  - $A(17,15)[3]$
  - = element $A(17,13)$ in the distributed array $A$ in process with rank 3
Remarks on MPI and PGAS (UPC & CAF)

• Future shrinking of memory per core implies
  – Communication time becomes a bottleneck
  → Computation and communication must be overlapped,
    i.e., latency hiding is needed

• With PGAS, halos are not needed.
  – But it is hard for the compiler to access data such early that the
    transfer can be overlapped with enough computation.

• With MPI, typically too large message chunks are transferred.
  – This problem also complicates overlapping.

• Strided transfer is expected to be slower than contiguous transfers
  – Typical packing strategies do not work for PGAS on compiler level
  – Only with MPI, or with explicit application programming with PGAS
Remarks on MPI and PGAS (UPC & CAF)

- Point-to-point neighbor communication
  - PGAS or MPI nonblocking may fit if message size makes sense for overlapping.

- Collective communication
  - Library routines are best optimized
  - Non-blocking collectives (comes with MPI-3.0) versus calling MPI from additional communication thread
  - Only blocking collectives in PGAS library?
Remarks on MPI and PGAS (UPC & CAF)

• For extreme HPC (many nodes × many cores)
  – Most parallelization may still use MPI
  – Parts are optimized with PGAS, e.g., for better latency hiding
  – PGAS efficiency is less portable than MPI
  – #ifdef … PGAS
  – Requires mixed programming PGAS & MPI
    → will be addressed by MPI-3.0
Outline

• Introduction / Motivation
• Programming models on clusters of SMP nodes
• Case Studies / pure MPI vs hybrid MPI+OpenMP
• Practical “How-To” on hybrid programming
• Mismatch Problems
• Opportunities:
  Application categories that can benefit from hybrid parallelization
• Thread-safety quality of MPI libraries
• Tools for debugging and profiling MPI+OpenMP
• Other options on clusters of SMP nodes

• Summary
Acknowledgements

- We want to thank
  - Co-authors who could not be here for presenting their slides:
    - Gabriele Jost
    - Georg Hager
  - Other contributors:
    - Gerhard Wellein, RRZE
    - Alice Koniges, NERSC, LBNL
    - Rainer Keller, HLRS and ORNL
    - Jim Cownie, Intel
    - KOJAK project at JSC, Research Center Jülich
    - HPCMO Program and the Engineer Research and Development Center Major Shared Resource Center, Vicksburg, MS (http://www.erdc.hpc.mil/index)
Summary – the good news

MPI + OpenMP

• Significant opportunity → higher performance on smaller number of threads
• Seen with NPB-MZ examples
  – BT-MZ → strong improvement (as expected)
  – SP-MZ → small improvement (none was expected)
• Usable on higher number of cores
• Advantages
  – Load balancing
  – Memory consumption
  – Two levels of parallelism
    • Outer → distributed memory → halo data transfer → MPI
    • Inner → shared memory → ease of SMP parallelization → OpenMP
• You can do it → “How To”
Summary – the bad news

MPI+OpenMP: There is a huge amount of pitfalls

• Pitfalls of MPI
• Pitfalls of OpenMP
  – On ccNUMA → e.g., first touch
  – Pinning of threads on cores
• Pitfalls through combination of MPI & OpenMP
  – E.g., topology and mapping problems
  – Many mismatch problems
• Tools are available 😊
  – It is not easier than analyzing pure MPI programs 😞
• Most hybrid programs → Masteronly style
• Overlapping communication and computation with several threads
  – Requires thread-safety quality of MPI library
  – Loss of OpenMP worksharing support → using OpenMP tasks as workaround
Summary – good and bad

- Optimization
  - 1 MPI process per core ................................. 1 MPI process per SMP node
  \(^\sim \text{somewhere between may be the optimum}\)

- 😏 Efficiency of MPI+OpenMP is not for free:
  - The efficiency strongly depends on
  - 😊 the amount of work in the source code development
Summary – Alternatives

Pure MPI

+ Ease of use
  – Topology and mapping problems may need to be solved (depends on loss of efficiency with these problems)
  – Number of cores may be more limited than with MPI+OpenMP
+ Good candidate for perfectly load-balanced applications

Pure OpenMP

+ Ease of use
  – Limited to problems with tiny communication footprint
  – Source code modifications are necessary (Variables that are used with “shared” data scope must be allocated as “sharable”)
± (Only) for the appropriate application a suitable tool
Summary

• This tutorial tried to
  – help to negotiate obstacles with hybrid parallelization,
  – give hints for the design of a hybrid parallelization,
  – and technical hints for the implementation → “How To”,
  – show tools if the application does not work as designed.

• This tutorial was not an introduction into other parallelization models:
  – Partitioned Global Address Space (PGAS) languages
    (Unified Parallel C (UPC), Co-array Fortran (CAF), Chapel, Fortress, Titanium,
    and X10).
  – High Performance Fortran (HPF)
  → Many rocks in the cluster-of-SMP-sea do not vanish
    into thin air by using new parallelization models
  → Area of interesting research in next years
Conclusions

- Future hardware will be more complicated
  - Heterogeneous → GPU, FPGA, ...
  - ccNUMA quality may be lost on cluster nodes
  - ...
- High-end programming → more complex
- Medium number of cores → more simple
  (if \#cores / SMP-node will not shrink)
- MPI+OpenMP → work horse on large systems
- Pure MPI → still on smaller cluster
- OpenMP → on large ccNUMA nodes
  (not ClusterOpenMP)

Thank you for your interest

Q & A

Please fill in the feedback sheet – Thank you
Appendix

- Abstract
- Authors
- References (with direct relation to the content of this tutorial)
- Further references
Abstract

Half-Day Tutorial  (Level: 20% Introductory, 50% Intermediate, 30% Advanced)

Authors.  Rolf Rabenseifner, HLRS, University of Stuttgart, Germany
Georg Hager, University of Erlangen-Nuremberg, Germany
Gabriele Jost, Texas Advanced Computing Center, The University of Texas at Austin, USA

Abstract.  Most HPC systems are clusters of shared memory nodes. Such systems can be PC clusters with single/multi-socket and multi-core SMP nodes, but also "constellation" type systems with large SMP nodes. Parallel programming may combine the distributed memory parallelization on the node inter-connect with the shared memory parallelization inside of each node.

This tutorial analyzes the strength and weakness of several parallel programming models on clusters of SMP nodes. Various hybrid MPI+OpenMP programming models are compared with pure MPI. Benchmark results of several platforms are presented. The thread-safety quality of several existing MPI libraries is also discussed. Case studies will be provided to demonstrate various aspects of hybrid MPI/OpenMP programming. Another option is the use of distributed virtual shared-memory technologies. Application categories that can take advantage of hybrid programming are identified. Multi-socket-multi-core systems in highly parallel environments are given special consideration.

Rolf Rabenseifner

Dr. Rolf Rabenseifner studied mathematics and physics at the University of Stuttgart. Since 1984, he has worked at the High-Performance Computing-Center Stuttgart (HLRS). He led the projects DFN-RPC, a remote procedure call tool, and MPI-GLUE, the first metacomputing MPI combining different vendor's MPIs without losing the full MPI interface. In his dissertation, he developed a controlled logical clock as global time for trace-based profiling of parallel and distributed applications. Since 1996, he has been a member of the MPI-2 Forum and since Dec. 2007, he is in the steering committee of the MPI-3 Forum. From January to April 1999, he was an invited researcher at the Center for High-Performance Computing at Dresden University of Technology. Currently, he is head of Parallel Computing - Training and Application Services at HLRS. He is involved in MPI profiling and benchmarking, e.g., in the HPC Challenge Benchmark Suite. In recent projects, he studied parallel I/O, parallel programming models for clusters of SMP nodes, and optimization of MPI collective routines. In workshops and summer schools, he teaches parallel programming models in many universities and labs in Germany.
Georg Hager

Georg Hager holds a PhD in computational physics from the University of Greifswald. He has been working with high performance systems since 1995, and is now a senior research scientist in the HPC group at Erlangen Regional Computing Center (RRZE). His daily work encompasses all aspects of HPC user support and training, assessment of novel system and processor architectures, and supervision of student projects and theses. Recent research includes architecture-specific optimization for current microprocessors, performance modeling on processor and system levels, and the efficient use of hybrid parallel systems. A full list of publications, talks, and other HPC-related stuff he is interested in can be found in his blog: http://blogs.fau.de/hager.
Gabriele Jost obtained her doctorate in Applied Mathematics from the University of Göttingen, Germany. For more than a decade she worked for various vendors (Suprenum GmbH, Thinking Machines Corporation, and NEC) of high performance parallel computers in the areas of vectorization, parallelization, performance analysis and optimization of scientific and engineering applications.

In 2005 she moved from California to the Pacific Northwest and joined Sun Microsystems as a staff engineer in the Compiler Performance Engineering team, analyzing compiler generated code and providing feedback and suggestions for improvement to the compiler group. She then decided to explore the world beyond scientific computing and joined Oracle as a Principal Engineer working on performance analysis for application server software. That was fun, but she realized that her real passions remains in area of performance analysis and evaluation of programming paradigms for high performance computing and that she really liked California. She is now a Research Scientist at the Texas Advanced Computing Center (TACC), working remotely from Monterey, CA on all sorts of exciting projects related to large scale parallel processing for scientific computing.
Book (with direct relation to the content of this tutorial)


References (with direct relation to the content of this tutorial)

- **NAS Parallel Benchmarks:**
  http://www.nas.nasa.gov/Resources/Software/npb.html

- R.v.d. Wijngaart and H. Jin,
  *NAS Parallel Benchmarks, Multi-Zone Versions,*
  NAS Technical Report NAS-03-010, 2003

- H. Jin and R. v.d.Wijngaart,
  *Performance Characteristics of the multi-zone NAS Parallel Benchmarks,*
  Proceedings IPDPS 2004

- G. Jost, H. Jin, D. an Mey and F. Hatay,
  *Comparing OpenMP, MPI, and Hybrid Programming,*
  Proc. Of the 5th European Workshop on OpenMP, 2003

- E. Ayguade, M. Gonzalez, X. Martorell, and G. Jost,
  *Employing Nested OpenMP for the Parallelization of Multi-Zone CFD Applications,*
  Proc. Of IPDPS 2004
References


- Rolf Rabenseifner, *Comparison of Parallel Programming Models on Clusters of SMP Nodes.* In proceedings of the 45nd Cray User Group Conference, CUG SUMMIT 2003, May 12-16, Columbus, Ohio, USA.


References


References

References

• J. Treibig, G. Hager and G. Wellein:
  **LIKWID: A lightweight performance-oriented tool suite for x86 multicore environments.**

• H. Stengel:
  **Parallel programming on hybrid hardware: Models and applications.**
  http://www.hpc.rrze.uni-erlangen.de/Projekte/hybrid.shtml
Further references

- Sergio Briguglio, Beniamino Di Martino, Giuliana Fogaccia and Gregorio Vlad, *Hierarchical MPI+OpenMP implementation of parallel PIC applications on clusters of Symmetric MultiProcessors*, 10th European PVM/MPI Users' Group Conference (EuroPVM/MPI'03), Venice, Italy, 29 Sep - 2 Oct, 2003

- Barbara Chapman, *Parallel Application Development with the Hybrid MPI+OpenMP Programming Model*, Tutorial, 9th EuroPVM/MPI & 4th DAPSYS Conference, Johannes Kepler University Linz, Austria September 29-October 02, 2002


Further references

• Holger Brunst and Bernd Mohr,  
  **Performance Analysis of Large-scale OpenMP and Hybrid MPI/OpenMP Applications with VampirNG**  
  http://www.fz-juelich.de/zam/kojak/documentation/publications/

• Felix Wolf and Bernd Mohr,  
  **Automatic performance analysis of hybrid MPI/OpenMP applications**  
  http://www.fz-juelich.de/zam/kojak/documentation/publications/

• Felix Wolf and Bernd Mohr,  
  **Automatic Performance Analysis of Hybrid MPI/OpenMP Applications**  
  http://www.fz-juelich.de/zam/kojak/documentation/publications/
Further references


Further references


Further references


Further references


## Content

<table>
<thead>
<tr>
<th>Section</th>
<th>slide</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction / Motivation</strong></td>
<td>1</td>
</tr>
<tr>
<td><strong>Programming models on clusters of SMP nodes</strong></td>
<td>6</td>
</tr>
<tr>
<td>– Major programming models</td>
<td>7</td>
</tr>
<tr>
<td>– Pure MPI</td>
<td>9</td>
</tr>
<tr>
<td>– Hybrid Masteronly Style</td>
<td>10</td>
</tr>
<tr>
<td>– Overlapping Communication and Computation</td>
<td>11</td>
</tr>
<tr>
<td>– Pure OpenMP</td>
<td>12</td>
</tr>
<tr>
<td><strong>Case Studies / pure MPI vs. hybrid MPI+OpenMP</strong></td>
<td>13</td>
</tr>
<tr>
<td>– The Multi-Zone NAS Parallel Benchmarks</td>
<td>14</td>
</tr>
<tr>
<td>– Benchmark Architectures</td>
<td>17</td>
</tr>
<tr>
<td>– On the Sun Constellation Cluster Ranger</td>
<td>18</td>
</tr>
<tr>
<td>– NUMA Control (numactl)</td>
<td>23</td>
</tr>
<tr>
<td>– On a Cray XT5 cluster</td>
<td>29</td>
</tr>
<tr>
<td>– On a IBM Power6 system</td>
<td>34</td>
</tr>
<tr>
<td>– Conclusions</td>
<td>40</td>
</tr>
<tr>
<td><strong>Practical “How-To” on hybrid programming</strong></td>
<td>41</td>
</tr>
<tr>
<td>– How to compile, link and run</td>
<td>43</td>
</tr>
<tr>
<td>– Running the code <strong>efficiently</strong>?</td>
<td>50</td>
</tr>
<tr>
<td>– A short introduction to ccNUMA</td>
<td>52</td>
</tr>
<tr>
<td>– ccNUMA Memory Locality Problems / First Touch</td>
<td>56</td>
</tr>
<tr>
<td>– ccNUMA problems beyond first touch</td>
<td>59</td>
</tr>
<tr>
<td>– Bandwidth and latency</td>
<td>61</td>
</tr>
<tr>
<td>– OpenMP and Threading overhead</td>
<td>64</td>
</tr>
<tr>
<td>– Thread/Process Affinity (“Pinning”)</td>
<td>67</td>
</tr>
<tr>
<td>– Hybrid MPI/OpenMP: “how-to”</td>
<td>78</td>
</tr>
<tr>
<td><strong>Mismatch Problems</strong></td>
<td>79</td>
</tr>
<tr>
<td>– Topology problem</td>
<td>81</td>
</tr>
<tr>
<td>– Mapping problem with mixed model</td>
<td>88</td>
</tr>
<tr>
<td>– Unnecessary intra-node communication</td>
<td>89</td>
</tr>
<tr>
<td>– Sleeping threads and network saturation problem</td>
<td>90</td>
</tr>
<tr>
<td>– Additional OpenMP overhead</td>
<td>91</td>
</tr>
<tr>
<td>– Overlapping communication and computation</td>
<td>92</td>
</tr>
<tr>
<td>– Communication overhead with DSM</td>
<td>101</td>
</tr>
<tr>
<td>– Back to the mixed model</td>
<td>106</td>
</tr>
<tr>
<td>– No silver bullet</td>
<td>107</td>
</tr>
<tr>
<td><strong>Opportunities: Application categories that can benefit from hybrid parallelization</strong></td>
<td>108</td>
</tr>
<tr>
<td>– Nested Parallelism</td>
<td>109</td>
</tr>
<tr>
<td>– Load-Balancing</td>
<td>110</td>
</tr>
<tr>
<td>– Memory consumption</td>
<td>111</td>
</tr>
<tr>
<td>– Opportunities, if MPI speedup is limited due to algorithmic problem</td>
<td>115</td>
</tr>
<tr>
<td>– To overcome MPI scaling problems</td>
<td>116</td>
</tr>
<tr>
<td>– Summary</td>
<td>117</td>
</tr>
</tbody>
</table>
## Content

- **Thread-safety quality of MPI libraries** 118
  - MPI rules with OpenMP 119
  - Thread support of MPI libraries 122
  - Thread Support within OpenMPI 123
- **Tools for debugging and profiling MPI+OpenMP** 124
  - Intel ThreadChecker 125
  - Performance Tools Support for Hybrid Code 127
- **Other options on clusters of SMP nodes** 131
  - Pure MPI – multi-core aware 132
  - Hierarchical domain decomposition 133
  - Scalability of MPI to hundreds of thousands 138
  - Remarks on Cache Optimization 139
  - Remarks on Cost-Benefit Calculation 140
  - Remarks on MPI and PGAS (UPC & CAF) 141
- **Summary** 145
  - Acknowledgements 146
  - Summaries 147
  - Conclusions 152
- **Appendix** 153
  - Abstract 154
  - Authors 155
  - References (with direct relation to the content of this tutorial) 158
  - Further references 163
- **Content** 170