Virtual Institute – High Productivity Supercomputing

Virtual Institute –
High Productivity Supercomputing

Code Tuning Tutorial

18 May 2012

Brian Wylie
Jülich Supercomputing Centre
b.wylie@fz-juelich.de
Friday 18 May

- 09:00 Start
- Welcome & introduction to VI-HPS
- Overview of parallel performance analysis tools
- Profile analysis with Scalasca
- 11:00-11:30 (break)
- Trace analysis with Vampir (Holger Brunst, TU Dresden)
- 13:00-14:30 (lunch)
- Assisted analysis of participants' codes
- 17:00 Closing of the Spring School
We'd like to know a little about you, your application(s), and your expectations and desires from this tutorial

- **What programming paradigms do you use in your app(s)?**
  - only MPI, only OpenMP, mixed-mode/hybrid OpenMP/MPI, ...
  - Fortran, C, C++, multi-language, ...

- **What platforms/systems must your app(s) run well on?**
  - Cray XT/XE/XK, IBM BlueGene, SGI Altix, Linux cluster™, ...

- **Who's already familiar with **serial** performance analysis?**
  - Which tools have you used?
    - time, print/printf, prof/gprof, ...

- **Who's already familiar with **parallel** performance analysis?**
  - Which tools have you used?
    - time, print/printf, prof/gprof, Periscope, Scalasca, TAU, Vampir, ...
Prepare to analyse your own application codes

- Ensure your application codes build and run to completion with appropriate datasets
  - initial configuration should ideally run in less than 15 minutes with 1-4 compute nodes (up to 128 processes/threads)
    - to facilitate rapid turnaround and quick experimentation
  - larger/longer scalability configurations are also interesting
    - turnaround may be limited due to busyness of batch queues

- Compare your application performance on other systems
  - VI-HPS tools already installed on a number of HPC systems
    - if not, ask your system administrator to install them
      (or install a personal copy yourself)
Goal: Improve the quality and accelerate the development process of complex simulation codes running on highly-parallel computer systems

- Start-up funding (2006-2011) by Helmholtz Association of German Research Centres

- Activities
  - Development and integration of HPC programming tools
    - Correctness checking & performance analysis
  - Training workshops
  - Service
    - Support email lists
    - Application engagement
  - Academic workshops

www.vi-hps.org
VI-HPS partners (founders)

- Forschungszentrum Jülich
  - Jülich Supercomputing Centre
- RWTH Aachen University
  - Centre for Computing & Communication
- Technical University of Dresden
  - Centre for Information Services & HPC
- University of Tennessee (Knoxville)
  - Innovative Computing Laboratory
German Research School
  ■ Laboratory of Parallel Programming
Technical University of Munich
  ■ Chair for Computer Architecture
University of Oregon
  ■ Performance Research Laboratory
University of Stuttgart
  ■ HPC Centre
University of Versailles St-Quentin
  ■ LRC ITACA

Barcelona Supercomputing Center
  ■ Centro Nacional de Supercomputación
Lawrence Livermore National Laboratory
  ■ Centre for Applied Scientific Computing
VI-HPS productivity tools

- **Marmot/MUST**
  - MPI correctness checking

- **PAPI**
  - Interfacing to hardware performance counters

- **Periscope**
  - Automatic analysis driven by on-line distributed search

- **Scalasca**
  - Large-scale parallel performance analysis

- **TAU**
  - Integrated parallel performance system

- **Vampir/VampirTrace**
  - Event tracing and graphical trace visualization & analysis
VI-HPS productivity tools (cont.)

- **KCachegrind**
  - Callgraph-based cache analysis [x86 only]

- **MAQAO**
  - Assembly instrumentation & optimization [x86 only]

- **ompP**
  - OpenMP profiling tool

- **OpenMPI**
  - Memory checking

- **Paraver/Extrae**
  - Event tracing and graphical trace visualization & analysis

- **Score-P**
  - Common instrumentation & measurement infrastructure

- **SIONlib**
  - Optimized native parallel file I/O
Technologies and their integration

- **Technologies**
  - PAPI
  - KCACHEGRIND
  - MARMOT / MUST
  - SIONLIB / OPENMPI
  - VAMPIR / PARAVER
  - SCALASCA
  - TAU
  - MAQAO

- **Integration Processes**
  - Error detection
  - Hardware monitoring
  - Automatic profile & trace analysis
  - Visual trace analysis
  - Execution
  - Optimization
Tools will *not* automatically make you, your applications or computer systems more *productive*. However, they can help you understand *how* your parallel code executes and *when* / *where* it's necessary to work on *correctness* and *performance* issues.
• Goals
  ■ Give an overview of the programming tools suite
  ■ Explain the functionality of individual tools
  ■ Teach how to use the tools effectively
  ■ Offer hands-on experience and expert assistance using tools
  ■ Receive feedback from users to guide future development

• For best results, bring & analyse/tune your own code(s)!

• VI-HPS Tutorial series
  ■ SC'08, ICCS'09, SC'09, Cluster'10, SC'10, SC'11

• VI-HPS Tuning Workshop series
  ■ 2008 (Aachen & Dresden), 2009 (Jülich & Bremen), 2010 (Garching & Amsterdam), 2011 (Stuttgart & Aachen)
  ■ 2012/04/23-27 (St-Quentin), 2012/10/15-19 (Garching)
Upcoming VI-HPS training events

- EU/US HPC Summer School (25-28 June 2012)
  - Dublin, Ireland
  - Half-day hands-on tutorial with PAPI, Scalasca, TAU, Vampir
- CSCS Summer School (6-10 August 2012)
  - Lugano, Switzerland
  - Two day hands-on tutorial on Cray XK with Scalasca + Vampir
- EuroMPI conference hands-on tutorial (23 Sep 2012)
  - Vienna, Austria
  - Periscope, Scalasca, TAU, Vampir (using Score-P)

Check [www.vi-hps.org/training](http://www.vi-hps.org/training) for announced events

- Contact us if you might be interested in hosting an event
Upcoming VI-HPS training events

- 10th VI-HPS Tuning Workshop (16-19 Oct 2012)
  - hosted by LRZ, Garching-bei-Muenchen, Germany
  - using PRACE Tier-0 SuperMUC iDataPlex system
  - Scalasca, Vampir, TAU, Periscope, KCachegrind, MAQAO, ...

- Further events to be determined
  - (one-day) tutorials
    - with guided exercises usually using Live DVD
  - (multi-day) training workshops
    - with your own applications on real HPC systems

Check [www.vi-hps.org/training](http://www.vi-hps.org/training) for announced events

- Contact us if you might be interested in hosting an event
• Bootable Linux installation ISO (on DVD or USB drive)
• Includes everything needed to try out our parallel tools on an x86-architecture notebook computer
  ■ VI-HPS tools: KCachegrind, Marmot, PAPI, Periscope, Scalasca, TAU, VT/Vampir*
  ■ Also: Eclipse/PTP, TotalView*, etc.
    ► * time/capability-limited evaluation licences provided for commercial products
  ■ GCC (w/ OpenMP), OpenMPI
  ■ Manuals/User Guides
  ■ Tutorial exercises & examples
• Produced by U. Oregon PRL
  ■ Sameer Shende
• ISO image approximately 4GB
  ■ distributed on DVD or USB drive
  ■ or download from website

• Boot directly from disk
  ■ enables hardware counter access and offers best performance

• Boot within virtual machine
  ■ faster boot time and can save/resume state, but no hardware counter access

• Boots into Linux environment
  ■ supports building and running provided MPI and/or OpenMP parallel application codes
  ■ and experimentation with VI-HPS (and other) tools
Cachegrind: cache analysis by simple cache simulation
- Captures dynamic callgraph
- Based on valgrind dynamic binary instrumentation
- Runs on x86/PowerPC/ARM unmodified binaries
  - No root access required
- ASCII reports produced

[KQ] Cachegrind GUI
- Visualization of cachegrind output

Developed by TU Munich
- Released as GPL open-source
- [http://kcachegrind.sf.net/](http://kcachegrind.sf.net/)
KCacheGrind GUI

- Event cost tree map
- Source code view
- Call graph view
- Machine code annotation
Tool to check for correct MPI usage at runtime

- Checks conformance to MPI standard
  - Supports Fortran & C bindings of MPI-1.2
- Checks parameters passed to MPI
- Monitors MPI resource usage

Implementation

- C++ library gets linked to the application
- Does not require source code modifications
- Additional process used as DebugServer
- Results written in a log file (ASCII/HTML/CUBE)

Developed by HLRS & TU Dresden

- Released as open-source
- http://www.hlrs.de/organization/av/amt/projects/marmot
Next generation MPI runtime error detection tool

- Successor of the Marmot and Umpire tools
- Initial merge of Marmot's many local checks with Umpire's non-local checks
- Improved scalability expected in future
- Exploits CMake, GTI & PnMPI infrastructure

Developed by TU Dresden, LLNL & LANL

- BSD license open-source initial release in November 2011
- http://tu-dresden.de/zih/must/
Portable performance counter library & utilities

- Configures and accesses hardware/system counters
- Predefined events derived from available native counters
- Core component for CPU/processor counters
  - instructions, floating point operations, branches predicted/taken, cache accesses/misses, TLB misses, cycles, stall cycles, ...
  - performs transparent multiplexing when required
- Extensible components for off-processor counters
  - InfiniBand network, Lustre filesystem, system hardware health, ...
- Used by multi-platform performance measurement tools
  - Periscope, Scalasca, TAU, VampirTrace, ...

Developed by UTK-ICL

- Available as open-source for most modern processors

http://icl.cs.utk.edu/papi/
Available events and hardware information.

<table>
<thead>
<tr>
<th>Name</th>
<th>Code</th>
<th>Avail</th>
<th>Deriv</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAPI_L1_DCM</td>
<td>0x80000000</td>
<td>Yes</td>
<td>No</td>
<td>Level 1 data cache misses</td>
</tr>
<tr>
<td>PAPI_L1_ICM</td>
<td>0x80000001</td>
<td>Yes</td>
<td>No</td>
<td>Level 1 instruction cache misses</td>
</tr>
</tbody>
</table>

Of 107 possible events, 35 are available, of which 9 are derived.
Available native events and hardware information.

<table>
<thead>
<tr>
<th>Event Code</th>
<th>Symbol</th>
<th>Long Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x40000000</td>
<td>UNHALTED_CORE_CYCLES</td>
<td>count core clock cycles whenever the clock signal on the specific core is running (not halted). Alias to event CPU_CLK_UNHALTED:THREAD</td>
</tr>
<tr>
<td>0x40000001</td>
<td>INSTRUCTION_RETIRED</td>
<td>count the number of instructions at retirement. Alias to event INST_RETIRED:ANY_P</td>
</tr>
<tr>
<td>0x40000086</td>
<td>UNC_SNP_RESP_TO_REMOTE_HOME</td>
<td>Remote home snoop response - LLC does not have cache line</td>
</tr>
<tr>
<td>4000486</td>
<td>:I_STATE</td>
<td>Remote home snoop response - LLC does not have cache line</td>
</tr>
<tr>
<td>40000886</td>
<td>:S_STATE</td>
<td>Remote home snoop response - LLC has cache line in S state</td>
</tr>
<tr>
<td>40001086</td>
<td>:FWD_S_STATE</td>
<td>Remote home snoop response - LLC forwarding cache line</td>
</tr>
<tr>
<td>40002086</td>
<td>:FWD_I_STATE</td>
<td>Remote home snoop response - LLC has forwarded a modified cache line</td>
</tr>
<tr>
<td>40004086</td>
<td>:CONFLICT</td>
<td>Remote home conflict snoop response</td>
</tr>
<tr>
<td>40008086</td>
<td>:WB</td>
<td>Remote home snoop response - LLC has cache line in the M state</td>
</tr>
<tr>
<td>40010086</td>
<td>:HITM</td>
<td>Remote home snoop response - LLC HITM</td>
</tr>
</tbody>
</table>

Total events reported: 135
Automated profile-based performance analysis

- Iterative on-line performance analysis
  - Multiple distributed hierarchical agents
- Automatic search for bottlenecks based on properties formalizing expert knowledge
  - MPI wait states, OpenMP overheads and imbalances
  - Processor utilization hardware counters
- Clustering of processes/threads with similar properties
- Eclipse-based integrated environment

Supports

- SGI Altix Itanium2, IBM Power and x86-based architectures

Developed by TU Munich

- Released as open-source
- http://www.lrr.in.tum.de/periscope
Periscope properties & strategies (examples)

**MPI**
- Excessive MPI communication time
- Excessive MPI time due to many small messages
- Excessive MPI time in receive due to late sender
- ...

**OpenMP**
- Load imbalance in parallel region/section
- Sequential computation in master/single/ordered region
- ...

**Hardware performance counters (platform-specific)**
- Cycles lost due to cache misses
  - High L1/L2/L3 demand load miss rate
- Cycles lost due to no instruction to dispatch
- ...

Periscope plug-in to Eclipse environment

Source code view

SIR outline view

Project view

Properties view
Automatic performance analysis toolset

- Scalable performance analysis of large-scale applications
  - particularly focused on MPI & OpenMP paradigms
  - analysis of communication & synchronization overheads
- Automatic and manual instrumentation capabilities
- Runtime summarization and/or event trace analyses
- Automatic search of event traces for patterns of inefficiency
  - Scalable trace analysis based on parallel replay
- Interactive exploration GUI and algebra utilities for XML callpath profile analysis reports

Developed by JSC & GRS

- Released as open-source
- http://www.scalasca.org/
Scalasca automatic trace analysis report
Scalasca automatic trace analysis report

**Late Sender Time**

Description:
Refers to the time lost waiting caused by a blocking receive operation (e.g., MPI_Recv()) or MPI_Wait()) that is posted earlier than the corresponding send operation.

**Online description**

- Coor: (0, 0, 1)
- Node: c9-1c2s4n0
- Name: Thread 0
- MPI rank: 1
- Thread id: 0
- Value: 0.24 (0.24%)
- Absolutes: 0.00 (0.24%)
Integrated performance toolkit

- Instrumentation, measurement, analysis & visualization
  - Highly customizable installation, API, envvars & GUI
  - Supports multiple profiling & tracing capabilities
- Performance data management & data mining
- Targets all parallel programming/execution paradigms
  - Ported to a wide range of computer systems
- Performance problem solving framework for HPC
- Extensive bridges to/from other performance tools
  - PerfSuite, Scalasca, Vampir, ...

Developed by U. Oregon/PRL

- Broadly deployed open-source software
- http://tau.uoregon.edu/
TAU ParaProf GUI displays (selected)
TAU PerfExplorer data mining
Interactive event trace analysis

- Alternative & supplement to automatic trace analysis
- Visual presentation of dynamic runtime behaviour
  - event timeline chart for states & interactions of processes/threads
  - communication statistics, summaries & more
- Interactive browsing, zooming, selecting
  - linked displays & statistics adapt to selected time interval (zoom)
  - scalable server runs in parallel to handle larger traces

Developed by TU Dresden ZIH

- Open-source VampirTrace library bundled with OpenMPI 1.3
- [http://www.tu-dresden.de/zh/vampirtrace/](http://www.tu-dresden.de/zh/vampirtrace/)
- Vampir Server & GUI have a commercial license
- [http://www.vampir.eu/](http://www.vampir.eu/)
Vampir interactive trace analysis GUI
Vampir interactive trace analysis GUI
Interactive event trace analysis
  - Visual presentation of dynamic runtime behaviour
    - event timeline chart for states & interactions of processes
    - Interactive browsing, zooming, selecting
  - Large variety of highly configurable analyses & displays

Developed by Barcelona Supercomputing Center
  - Paraver trace analyser and Extrae measurement library
  - Open source available from http://www.bsc.es/paraver/
• Modular Assembler Quality Analyzer & Optimizer
  ▪ Framework for binary manipulation
    ▶ using plugins and scripting language
  ▪ Tool exploiting framework to produce reports
    ▶ fast prototyping and batch interface
  ▪ STAN static performance model
  ▪ MIL instrumentation language for dynamic analysis
    ▶ building custom performance evaluation tools using HWCs
    ▶ instrumentation of functions, loops, blocks & instructions
• Developed by UVSQ Exascale Computing Research lab
  ▪ Supports Intel x86_64 microarchitecture
  ▪ Available from www.maqao.org
Key tool components also provided as open-source

- Program development environment
  - Eclipse PTP ETFw, UNITE

- Program/library instrumentation
  - COBI, OPARI, PDToolkit

- Runtime measurement systems
  - Score-P, UniMCI

- Scalable I/O
  - SIONlib

- Libraries & tools for handling (and converting) traces
  - EPILOG, OTF, PEARL

- Analysis algebra & hierarchical/topological presentation
  - CUBE
Scalable performance measurement infrastructure

- Supports instrumentation, profiling & trace collection, as well as online analysis of HPC parallel applications
- Works with Periscope, Scalasca, TAU & Vampir prototypes
- Based on updated tool components
  - CUBE4 profile data utilities & GUI
  - OA online access interface to performance measurements
  - OPARI2 OpenMP & pragma instrumenter
  - OTF2 open trace format

Created by German BMBF SILC & US DOE PRIMA projects

- JSC, RWTH, TUD, TUM, GNS, GRS, GWT & UO PRL
- Available as BSD open-source from http://www.score-p.org/
Portable native parallel I/O library & utilities

- Scalable massively-parallel I/O to task-local files
- Manages single or multiple physical files on disk
  - optimizes bandwidth available from I/O servers by matching blocksizes/alignment, reduces metadata-server contention
- POSIX-I/O-compatible sequential & parallel API
  - adoption requires minimal source-code changes
- Tuned for common parallel filesystems
  - GPFS (BlueGene), Lustre (Cray), ...
- Convenient for application I/O, checkpointing,
  - Used by Scalasca tracing (when configured)

Developed by JSC

- Available as open-source from http://www.fz-juelich.de/jsc/sionlib/
Uniform integrated tool environment

- Manages installation & access to program development tools
  - based on software environment management “modules”
  - commonly used on most cluster and HPC systems
  - configurable for multiple MPI libraries & compiler suites
- Specifies how & where tools packages get installed
  - including integrating tools where possible
- Defines standard module names and different versions
- Supplies pre-defined module files
- Configurable to co-exist with local installations & policies

Developed by JSC, RWTH & TUD

- Available as open-source from http://www.vi-hps.org/projects/unite/