PERFORMANCE

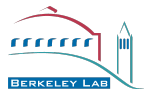
Facilitators

Adolfy Hoisie

Committee members: A. Hoisie, C. Iancu, J. Vetter

Technical Committee Report
to the Hybrid Multicore Consortium

First HMC Roadmap Workshop
January 19-22, San Francisco



Breakout Participants

- AdolfyHoisie LANL hoisie@lanl.gov
- Osni Marques DOE oamarques@ascr.doe.gov
- Sudhakar Yalamanchilei GIT sudha@ece.gatech.edu
- Jeffrey Vetter ORNL/GT vetter@computer.org
- Nick Wright LBNL nwright@lbl.gov
- Roger Golliver Intel roger.a.golliver@intel.com
- Karen Karavanic PSU karavan@cs.pdx.edu
- Costin Iancu LBL Cciancu@lbl.gov
- Rich Lethin Reservoir lethin@reservoir.com
- Doug Doerfler Sandia dwdoerf@sandia.gov
- Subhash Saini NASA subhash.saini@nasa.gov
- AGK agk@multicorewareinc.com

CHARGE TO BREAKOUT SESSIONS

- Goal of Roadmap:
  - Identify technologies that need to be developed to make next generation, large-scale, accelerator-based systems "production ready"
  - Provide community input needed to prioritize and support activities
- Focus is near term, while keeping an eye toward to long term (avoid box canyons)
- Work with the other TCs to support the overall co-design of applications, architectures, programming, and performance and to build ties with and provide feedback to vendors.
- Develop strategies for early and broader access to these accelerator-based or future hybrid multicore systems.

SUMMARY OF TC

- Areas of interest to this TC
  - Tools
  - Modeling
  - Code optimization
- Relation to other TCs
  - Architecture
  - Programming models
  - Apps

Performance is at the boundaries of all these areas, and spans the lifecycle/spectrum from R&D to design to implementation to optimization

Performance and Analysis Topics

- Monitoring, observation and Analysis Tools for systems and applications
  - Memory, node, interconnect, apps
- Code optimization
  - Autotuning, compilation
- Predictive modeling
  - Optimal application-architecture mapping for hybrid
  - Application/architecture co-design
  - Methodology development (modeling of many flavors, simulation)
  - Dynamic (runtime) model-driven system/application optimization

GRADING CRITERIA

Urgency How soon is it needed?	**Duration** How long will it be useful?	**Responsive** How much will money help?	**Applicability** How broadly can it be used?	**Timeline** How soon can we expect it?
Critical Needed now	**Long** Useful for the foreseeable future	**High** Funding enables significant progress	**Broad** Applicable beyond HPC	**Immediate** Results within 1-2 years
Important Needed within 3 years	**Medium** Useful for Exascale	**Moderate** Funding enables progress	**HPC** Applicable to all of HPC	**Soon** Results within 2-5 years
Useful Needed after 3 years	**Near** Only useful for immediate systems	**Low** Funding has little affect on progress	**Narrow** Only applicable to immediate systems	**Eventually** Results after 5 years

Predictive Modeling

- Optimal app-arch mapping, app/arch co-design , methodology development, dynamic model-driven sys/app optimization.
- Modeling power, reliability, performance in concert rather than independently
- Power modeling for software
- Methodology development (modeling, simulation)
- "Should I port my code to hybrid? Is it worth it?"
- Representation for hybrid codes – programming. model ;
- Modeling hybrid applications - multiphysics
- Statistical techniques?
- Predict very large scale performance based on small scale measurements
- What is the measure of success for a model? (eg how precise to be useful? don't always need more than coarse grained answer –"yes, porting is worthwhile")
- Simulation: interoperability of simulators
- Fault modeling, prediction and detection; reliability modeling; error propagation. Focus on tools for this – what do we need, specific to accelerator based systems? How do accelerators influence reliability?
- Validation methodologies

- Relations to other TCs
  - Architecture, runtime SW, programming environment, apps
- Related Projects
  - LANL/PAL
  - ORNL
  - PMAC/SDSC
  - LBL/Roofline
  - P-Bound (ANL)
  - Rice
  - Sandia/SST

Urgency	Duration	Responsive	Applicability	Timeline
Critical	Long	High	HPC	Immediate

INSTRUMENTATION

- Description: performance instrumentation for accelerators
- Binary / dynamic instrumentation for mixed codes
- Measuring buses
- HW (counters) & SW (system, application)
- Common interface for counters
- Memory subsystem analysis/ diagnosis
- MPI profile-like feedback at different levels (whole system, node level) about data movement
- Event tracing (clock; buffer)

- Relations to other TCs
  - Hooks into architecture and runtime system
- Related Projects
  - NVIDIA
  - PGI/TAU
  - UIUC
  - MIT
  - UC Berkeley

Urgency	Duration	Responsive	Applicability	Timeline
Important	Medium	Moderate	Broad	Soon

Integrated Measurements

- Infrastructure for migrating applications (performance portability)
- Tool perturbation
- Power consumption – sensors
- Diagnosis and attribution of root cause
- Resource contention and allocation / partitioning
- Mapping measurements to instructions or source code
- Performance variation / Noise for heterogeneous systems
- Comparison-based performance analysis
- Data management & representation & volume
- Tool interoperability/composition/frameworks: hierarchy (intra- vs inter-node performance) and heterogeneity
- Scalability

- Relations to other TCs
  - Hooks into architecture and runtime system
- Related Projects
  - TAU
  - PGI
  - Dimemas

Urgency	Duration	Responsive	Applicability	Timeline
Important	Medium	High	HPC	Immediate

Tools for Code Optimization

- Auto-tuning
- Dynamic Compilation
- "rules of thumb" "lessons learned" "Design Patterns" for hybrid devt, porting decisions ("should I port my code to GPU cluster?")
- Mixed precision: interactions with dynamic compilation; specifications for precision?
- Implications for correctness debugging – performance debugging interface
- Relations to other TCs
  - Fill this in
- Related Projects
  - Atlas/Magma
  - R-stream (Reservoir)
  - Ocelot
  - Parlab (Berkeley)
  - SCOUT
  - Rich Vuduc/GTech

Urgency	Duration	Responsive	Applicability	Timeline
Important	Long	Moderate	Broad	Soon

Breakout Summary

Topic	Urgency	Duration	Responsive	Applicability	Timeline
Performance Instrumentation	Important	Medium	Moderate	Broad	Soon
Integrated measurements	Important	Medium	High	HPC	Immediate
Tools for code optimization	Important	Long	Moderate	Broad	Soon
Predictive modeling	Critical	Long	High	HPC	Immediate

Notes and Recommendations

- Contact Adolfy Hoisie hoisie@lanl.gov with any comments/ recommendations