HMC Architectures and Metrics Technical Committee

Facilitators

Steve Poole, Ken Koch, Jeff Broughton

Technical Committee Report
to the Hybrid Multicore Consortium

First HMC Roadmap Workshop
January 19-22, San Francisco

Breakout Participants

- Steve Poole, ORNL
- Ken Koch, LANL
- Jeff Broughton, LBNL
- Jeff Kuehn, ORNL
- James Laros, Sandia
- John Daly, DoD/CEC
- Bill Dally, Nvidia
- Allan Cantle, Nallatech
- Benoit Meister, Reservoir Labs
- Jim Ang, Sandia
- Fred Johnson, SAIC
- Prasanna Sundararajan, Xilinx
- Bill Brantley, AMD
- Bob Ciotti, NASA
- Ada Garvrilovska, Georgia Tech
- Galen Shipman, ORNL
- Jakub Kurzak, UTenn
- Victor Lee, Intel
- Fabrizio Petrini, IBM
- John Leidel, Convey
- Glenn Lupton, HP

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.

Charge to Architecture & Metrics

- Identify and report on approaches for building large-scale accelerator-based hybrid computer systems in the near term and in the future.
- Identify the types and degrees of parallelism provided by hybrid cores and to define key architectural metrics of this class of hybrid machine.

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	**Science** Applicable to all of scientific computing	**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

Summary of Architecture & Metrics

- Areas of interest to this TC
  - Accelerator/System Interfaces
  - Accelerator Design
  - System Software
  - System Design
  - Simulation & Modeling
  - Metrics
- Relation to other TCs
  - Programming Models: Ease of programming & debugging
  - Performance: Enhance throughput & provide measurement tools
  - Applications & Libraries: same

Accelerator/System Interface

- Description
  - [1] Enhance bandwidth, latency between CPU & Accel. e.g. On-die or in-socket or in-stack (peer or parent)]. Improve power eff.
  - [2] Shared address space between accelerator and host CPU
  - [3] Enhance BW, latency and power eff. between nodes required for balanced performance in hybrid systems
  - [4] Efficient synchronization
  - [5] Well-defined end-to-end error detection/correction
  - [6] Global RMA or shared address space across nodes

- Notes from Discussion
  - Interfaces to CPU, Memory, Interconnect
  - Work to focus on eng. studies / risk reduction
- Relations to other TCs
  - Improve performance
  - Improve programmability
- Related Projects
  - PCI-e Gen3
  - Larabee
  - Grand Fusion
  - QPI / HT
  - Networking research projects (photonics, etc)

Urgency	Duration	Responsive	Applicability	Timeline
Critical	Long	High	Broad	Soon

Accelerator Design

- Description
  - [1] Memory architecture (coherent plus hooks, increased addressable memory)
  - [2] Scalar performance
  - [3] Flexible synchronization
  - Improved thread scheduling
  - [4] Improve fault detection and recovery
  - [5] System-level counters & status (performance & beyond)
  - User-level exception handling
  - Latency hiding techniques
  - Time correlation
  - Power/frequency scaling
- Notes from Discussion
  - Homogeneity v. Heterogeneity
- Relations to other TCs
  - Enhance performance
  - Simplify programming
- Related Projects
  - Nvidia
  - ATI
  - Intel
  - IBM
  - FPGAs
  - Tilera

Urgency	Duration	Responsive	Applicability	Timeline
Critical	Long	High	Broad	Soon

System Software

- Description
  - [1] Direct data transfer (network, I/O, etc.) to/from accelerator
  - [2] Optimization of data locality
  - [3] Scalable, Heterogeneous-aware OS
  - [4] Flexible management, monitoring, and scheduling of heterogeneous resources
  - [5] RAS system
  - Scalable I/O in hybrid system
- Notes from Discussion

- Relations to other TCs
  - All
- Related Projects
  - Lightweight OS projects
  - Heterogeneous-aware (e.g HyVM, Helios, Barrelfish)
  - OS/IO Function Shipping projects
  - Service Isolation projects

Urgency	Duration	Responsive	Applicability	Timeline
Critical	Long	High	HPC	Immediate

System Design

- Description
  - [1] Investigate proper balance of accelerators to cores, host interconnect, etc.
  - [2] Repeatable performance: eliminating sources of system variability
  - [3] Packaging, density, scaling, cooling methods, serviceability
- Notes from Discussion

- Relations to other TCs
  - Performance
  - Programming models
- Related Projects
  - Can learn from existing hybrid systems: Roadrunner, Lincoln, etc.

Urgency	Duration	Responsive	Applicability	Timeline
Important	Medium	High	Broad	Immediate

SIMULATION & MODELING

- Description
  - [1] Support system level simulation using component level black box emulators and simulators (both cycle-accurate and not)
  - [2] Validation of accuracy of simulators against real systems/ applications
  - [3] Predictive, science based models and UQ methods for modeling and optimization of power, performance and reliability
- Notes from Discussion

- Relations to other TCs
  - Predictive performance models
- Related Projects
  - Ocelot (PTX emulation), SST, QEMU, MacSim, McPAT DRAMsim, Sim-Panalyzer, etc., Mambo (IBM - System Simulator), IAA (SNL, ORNL – Simulator project), BigSim, DiskSim
  - Etc.

Urgency	Duration	Responsive	Applicability	Timeline
Important	Long	High	Broad	Immediate

METRICS

- Description
  - [1] Establish common terminology
  - [2] Quantify power efficiency, performance, and reliability
  - [3] Collect data to inform architecture evaluation and system/accelerator design
  - [4] Enhance usable instrumental in accelerators and hybrid systems
  - Examples:
    - Connectivity, BW, latency, I/O BW
    - Power efficiency, W/flop, MB/flop, Science/W (or throughput/W), Simulation Results/Joule
    - Reliability & availability of individual hybrid components and groups of components

- Notes from Discussion
- Relations to other TCs
  - Enable predictive performance models
  - Establish common terminology across TCs
- Related Projects
  - UHPC
  - HP Exascale group

Urgency	Duration	Responsive	Applicability	Timeline
Important	Long	High	Broad	Soon

Breakout Summary

Topic	Urgency	Duration	Responsive	Applicability	Timeline
Accelerator/ System Interface	Critical	Long	High	Broad	Soon
Accelerator Design	Critical	Long	High	Broad	Soon
System software	Critical	Long	High	HPC	Immediate
Simulation & Modeling	Important	Long	High	Broad	Immediate
Metrics	Important	Long	High	Broad	Soon
System Design	Important	Medium	High	Broad	Immediate

TESTBEDS

- Open access to entire community
- Multiple sites
- Application and system software development

- Production systems of significant scale (100-1000TF)
- Hardware evaluation systems (~50TF) x 1-3
- Four-node systems x N for specific, small-scale development use

Thank You

- Steve Poole
  - spoole@ornl.gov
- Ken Koch
  - krk@lanl.gov
- Jeff Broughton
  - jbroughton@lbnl.gov