Vancouver: A Software Stack for Productive Heterogeneous Exascale Computing

Presented by

Jeffrey Vetter (PI)
Oak Ridge National Laboratory

Wen-mei Hwu
University of Illinois at Urbana-Champaign

Allen D. Malony
University of Oregon

Rich Vuduc
Georgia Institute of Technology
Vancouver overview

• Large-scale heterogeneous system deployments are becoming more common

• Many challenges remain in using these systems
  – Programmer productivity
  – Lack of tools, libraries
  – Sensitive performance stability
  – Lack of constructs to span parallelism levels

• The Vancouver project is addressing these deficiencies with a three-tiered approach
  – Low-level libraries and runtime systems
  – Programming, development, and performance tools
  – High-level systems and abstractions
**Language tools**

- **Goal:** use static analysis and transformations to generate efficient heterogeneous executables

- **Approach:** *Pyon* to combine productivity of Python and efficiency of CUDA/OpenCL, and *Gluon* to automate optimizations
Runtime data orchestration

- **Goal:** create a runtime system to orchestrate data movement with little or no input from the application

- **Approach:** a new runtime system, Maestro, will combine task queue management, data movement, load balancing, and robustness for programmers
Autotuned libraries

• **Goal:** support more irregularly structured computation (sparse matrix, tree based) than has been considered for heterogeneous architectures

• **Approach:** investigate model-driven autotuning frameworks for libraries with respect to algorithmic, data, and architectural parameters
Partitioned global address space

- **Goal:** create models to facilitate many-node programming

- **Approach:** two paths to develop a prototype global view programming model for heterogeneous memory systems
  - Utilize on Pyon/Gluon in the context of multiple global array data-parallel operations
  - Library-based solution with new interfaces for spawning asynchronous GPU computation
Performance measurement

**Goal:** provide an integrated view of all information in a heterogeneous system

**Approach:** leverage and expand the TAU performance system to support CUDA, OpenCL, and GPU accelerator code instrumentation to capture performance data
Performance prediction

- **Goal:** generate predictions that account for different instruction set architectures and data orchestration costs

- **Approach:** the ADAPT tool will combine static interpretation, machine specification, and runtime instrumentation to generate accurate performance predictions

![CPU Traces](image)

**TAUcuda trace of GPU-accelerated Linpack**
Benchmarks

- **Goal:** provide quantitative guidance to users, tools, and developers about costs of computation and data movement on heterogeneous systems

- **Approach:** enhance Scalable Heterogeneous Computing (SHOC) benchmark suite
  - New and expanded tests
  - Result sharing website
References


Contacts

Jeffrey Vetter
Future Technologies Group
Computer Science and Mathematics Division
(865) 356-1649
vetter@ornl.gov

Jeremy Meredith
jsmeredith@ornl.gov

Kyle Spafford
spaffordkl@ornl.gov

Vinod Tipparaju
tipparajuv@ornl.gov