

Workshop on Performance & Productivity of Extreme-Scale Parallel Systems

Tuesday, October 17th, 2006

Organized by: Adolfo Hoisie, Los Alamos National Laboratory
Darren J. Kerbyson, Los Alamos National Laboratory
Dan Reed, University of North Carolina / Renaissance Computing Institute

Invited Keynote Presentation

9:00 [*Architecture, uses, and future of Cell Broadband Engine\(TM\)* technology*](#)

H. Peter Hofstee, Cell BE Chief Scientist, IBM Austin

Session 1: Software Issues in the Extreme

10:00 [*The Impact of Multicore on Math Software and Exploiting Single Precision Computing to Obtain Double Precision Results*](#), Jack Dongarra, University of Tennessee

10:30 Coffee Break

11:00 [*Performance Analysis, Modeling and Tuning at Scale*](#), Katherine Yelick, U.C. Berkeley and Lawrence Berkeley National Laboratory

11:30 [*Reducing HPC network requirements – exploiting overlap between Computation and Communication*](#), Jose Sancho, Los Alamos National Laboratory

12:00 [*It's not too hot for penguins: a status check of Linux on the BlueGenes*](#), George Almasi, IBM TJ Watson

12:30 Lunch (on your own)

Session 2: Performance Analysis

2:00 [*Experiences with Modeling of the performance of the Krak Hydrodynamics Application*](#), Kevin J. Barker, Los Alamos National Laboratory

2:30 [*Towards the Incorporation of Dynamic Adaptation into Operating Systems: Adaptive Disk I/O*](#), Patricia J. Teller, University of Texas-El Paso

3:00 [*From Timing Programs to Programmers: Measuring HPC Productivity*](#), Jeff Hollingsworth, University of Maryland

3:30 Coffee Break

Session 3: Performance Tools and Experiences

4:00 [*A case study in performance tuning*](#), John Feo, Cray

4:30 [*Scalable Performance Analysis of Large-Scale Parallel Applications*](#), Brian Wylie, John von Neumann Institute for Computing, Julich, Germany.

- 5:00 [Scalable System Measurement and Performance Analysis: Recent Progress](#), Robert J. Fowler, Renaissance Computing Institute, University of North Carolina
- 5:30 [The Red Storm System: Architecture, System Update and Performance Analysis](#), Doug Doerfler, Sandia National Laboratory
- 6:00 [Closing Remarks](#)

7th Symposium of the Los Alamos Computer Science
Institute: LACSI 2006, 17-19 October
Eldorado Hotel, Santa Fe, New Mexico

Architecture, uses, and future of Cell Broadband Engine(TM)* technology

H. Peter Hofstee

Cell BE Chief Scientist

IBM Systems and Technology Group, 11500 Burnet Rd, Austin, TX 78758

This talk will very briefly summarize the major decisions that have led to the Cell BE architecture. We will show both the chip and IBM systems roadmap for Cell, and discuss how Cell BE based systems can maintain their competitive edge over time. Next we will discuss some application examples and their performance. The talk will end with some observations on hybrid systems and approaches to programming such systems.

* Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc.

[Slides](#)

The Impact of Multicore on Math Software and Exploiting Single Precision Computing to Obtain Double Precision Results

Jack Dongarra

Innovative Computing Laboratory; Computer Science Dept, 1122 Volunteer Blvd; University of Tennessee; Knoxville TN, 37996-3450

Recent versions of microprocessors exhibit performance characteristics for 32 bit floating point arithmetic (single precision) that is substantially higher than 64 bit floating point arithmetic (double precision). Examples include the Intel's Pentium IV and M processors, AMD's Opteron architectures and the IBM's Cell Broad Engine processor. When working in single precision, floating point operations can be performed up to two times faster on the Pentium and up to ten times faster on the Cell over double precision. The performance enhancements in these architectures are derived by accessing extensions to the basic architecture, such as SSE2 in the case of the Pentium and the vector functions on the IBM Cell. The motivation for this talk is to exploit single precision operations

whenever possible and resort to double precision at critical stages while attempting to provide the full double precision results. The results described here are fairly general and can be applied to various problems in linear algebra such as solving large sparse systems, using direct or iterative methods and some eigenvalue problems.

[Slides](#)

Performance Analysis, Modeling and Tuning at Scale

Katherine Yelick

U.C. Berkeley and Lawrence Berkeley National Laboratory

The petascale computing systems that are promised over the next few years offer enormous potential to computational science and engineering communities, along with new challenges in the degree and kinds of parallelism that need to be exploited for high performance. To fully utilize the system capabilities, programmers will have to take advantage of multicore processors, heterogeneous processor accelerators, and latency tolerance in the memory and network system, in addition to the ever-increasing degree of parallelism between computational nodes. The Berkeley Institute for Performance Studies, a collaboration between U.C. Berkeley and Lawrence Berkeley National Laboratory, has research efforts in the architectural design, analysis, modeling and tuning of scientific computations. In this talk I will describe some of the major efforts in understanding system bottlenecks through microbenchmarking and analysis, as well as large-scale application studies to see how computational methods map onto various architectures. I will also describe some of our efforts in automatic performance optimization and the use of novel architectural features to address some of the fundamental limitations of high end computing systems, such as memory and network latencies, as well as effects of system scale on the performance and predictability of the interconnect networks.

On the other hand, it is not too difficult to tune programs I will describe work by members of the Berkeley Institute for Performance Studies, including Leonid Oliker, Erich Strohmaier, Hongzhang Shan, John Shalf, James Demmel, Parry Husbands, Rich Vuduc, Sam Williams, Shoaib Kamil, Kaushik Datta, and Rajesh Nishtala.

[Slides](#)

Reducing HPC network requirements – exploiting overlap between Computation and Communication

Jose Sancho

Performance and Architecture Lab, Los Alamos National Laboratory, NM 87545.

The design and implementation of a high performance communication network are critical factors in

determining the performance and cost-effectiveness of a large-scale computing system. One promising technique for extracting maximum application performance given limited network resources is based on overlapping computation with communication, which partially or entirely hides communication delays. In this talk, first we describe the method developed to analyze the potential communication and computation overlap in terms of 'dependent' computation on communication data, and also that which it is computation 'independent' of. And second, we explore the potential overlap for some large-scale production scientific codes for various network profiles. The experimental results obtained for these codes running on a large cluster containing 1,024 processors indicate that reduced network bandwidths and/or increased latencies can be tolerated without negatively impacting application performance. This allows for a potentially significant relaxation of network requirements, and thus suggesting that lower cost networks could be used to build more cost-effective HPC systems in the future if overlapping is exploited in the applications.

[Slides](#)

It's not too hot for penguins: a status check of Linux on the BlueGenes

George Almasi

IBM TJ Watson Research Center, Yorktown Heights, NY 10598

In this talk I will describe recent efforts at IBM TJ Watson to make the Linux O/S suitable for running high performance jobs on BlueGene. Building on the achievements of the Colony project (LLNL) and ZeptoOS (ANL), we focused on porting Linux to the BlueGene compute nodes, implementing function shipping and other performance tweaks to achieve performance comparable to that of the lightweight Compute Node Kernel.

[Slides](#)

Experiences with Modeling the performance of the Krak Hydrodynamics Application

Kevin J. Barker

Performance and Architecture Lab, Los Alamos National Laboratory, NM 87545.

In this talk, we present an analytic performance model of a large-scale hydrodynamics code developed at Los Alamos National Laboratory and describe the modeling team's experience with its development. This modeling work is part of an ongoing effort to develop models and modeling techniques for large-scale codes and systems of interest to Los Alamos and the national laboratory community. The Krak application comprises over 270,000 lines of source code and is capable of executing on a large number of parallel processors. The development of an accurate performance model is complicated by the irregular partitioning of input spatial grid cells and the various material properties assigned to each cell. We relate the method used to simplify model development through

the application of several careful approximations concerning subgrid size, subgrid shape, and material composition. Although such approximations allow for the development of a simple and powerful performance model, we are able to demonstrate that they do not adversely affect prediction accuracy. We validate our model on several spatial grid sizes and processor configurations and demonstrate an accuracy at the largest scale on 512 processor to within a 3% error.

[Slides](#)

Towards the Incorporation of Dynamic Adaptation into Operating Systems: Adaptive Disk I/O

Patricia J. Teller

University of Texas-El Paso, Department of Computer Science, El-Paso, TX

In the context of the DAiSES (Dynamic Adaptability in Support of Extreme Scale) research project, we are investigating ways to incorporate adaptation into operating systems, either by varying parameter values or policies at runtime. In response to changing workload characteristics or requirements, operating system (OS) adaptation is meant to dynamically “customize” the OS in an attempt to provide “best” service, based on predefined criteria, for the active workload. Current DAiSES research activities focus on three adaptation targets: disk scheduling, virtual memory management, and file I/O. Thus far, disk scheduling has received most of our attention, and will be the focus of this talk. The questions that we are trying to address in this research include the following:

- What is the best metric to use to fairly allocate disk resources?
- How should multiple different data requirements be satisfied?
- Can I/O performance be predicted and can performance isolation be ensured?
- Is the community measuring I/O performance correctly?
- Is programming for I/O performance necessary?

As you will see, although we partially answer these questions, the answers give rise to yet more questions!

[Slides](#)

From Timing Programs to Programmers: Measuring HPC Productivity

Jeff Hollingsworth

University of Maryland

[Slides](#)

A case study in performance tuning

John Feo,

Cray Incorporated, San Diego Supercomputer Center, University of California, San Diego, La Jolla, California 92093-0505

Performance tuning is a complex requirement of large scale computing. It involves many interrelated issues such as algorithm design, programming language, compilers, runtime systems, and architecture. Tuning for parallel computer systems with thousands of processors and deep memory hierarchies is an almost impossible task. It's difficult to see how peta-scale systems of the same design will make the problem easier.

On the other hand, it is not too difficult to tune programs written for shared-memory parallel systems that use parallelism rather than deep memory hierarchies to hide latencies. In this talk I present a case study of optimizing programs for such a machine. While it is unlikely that future peta-scale systems of this class will retain the simplicity of current systems, at least we begin with a problem we can solve

[Slides](#)

Scalable Performance Analysis of Large-Scale Parallel Applications

Brian Wylie

John von Neumann Institute for Computing (NIC), Forschungszentrum Jülich GmbH, D-52425 Jülich, Germany

The SCALASCA project is developing a new generation of tools for scalable performance analysis of large-scale parallel applications, building on extensive experience gained with the KOJAK toolset developed by Forschungszentrum Jülich and the University of Tennessee. First steps have extended KOJAK to significantly improve the scalability of trace capture and analysis of thousands of MPI processes, by avoiding trace re-writing and exploiting a replay-based parallel analysis. As demonstrated with Top-10 IBM Blue Gene/L and Cray XT3 systems, previously impractical performance analysis has now been enabled.

[Slides](#)

Scalable System Measurement and Performance Analysis: Recent Progress

Robert J. Fowler

Director of HPC Research, Renaissance Computing Institute, The University of North Carolina at Chapel Hill, 100 Europa Dr, Suite 540, Chapel Hill, NC 27517

We will describe recent work at RENCi and Rice U. towards taming the issue of scalability in the

capture, and management of performance data on very large parallel systems, and on analyzing that data for the purpose of problem diagnosis and remediation. The methods used include signal compression and adaptive statistical sampling and clustering strategies. The requirements on future systems to support scalable methods will also be discussed.

[Slides](#)

The Red Storm System: Architecture, System Update and Performance Analysis

Doug Doerfler

Sandia National Laboratory, Albuquerque, NM

The Red Storm Architecture is the basis of the first instantiation of the the Red Storm System and in general the Cray XT3 product line. The Red Storm System has been in use for over year, so it is now possible to analyze the performance of a real application workload and contrast it to other HPC architectures. The Red Storm System has also recently undergone a significant upgrade of its processors and interconnect. In this talk we will give a brief overview of the Red Storm Architecture, the Red Storm System in it's original and upgraded configurations, and an analysis of its performance on actual NNSA ASC applications.

[Slides](#)
