Computer Science Research Institute

Sandia Petaflops Workshop

June 18, 2002

Sandia National Laboratories

A Department of Energy National Laboratory

Agenda

Monday, June 17, 2002

6:00-8:00 PM Hospitality (optional)

Tuesday, June 18, 2002

8:00 Arrival at Sandia Badge Office

8:20 Neil Pundit – Welcome

8:40 Erik DeBenedictis – Workshop Introduction

9:00 Self Introductions & Comments on the Intended Purpose

TECHNOLOGY ISSUES

9:20 Carl Diegert – Hitting the Kilowatt per Square Foot Wall

9:40 Erik DeBenedictis – Sandia Petaflops Planner

10:00 Break

COMPUTER ARCHITECTURE ISSUES

10:20 Thomas Sterling – MIND: A PIM Strategy for Petaflops Computing

10:40 Doug Burger – Polymorphic Scientific Computing on the UT-Austin TRIPS Processor

11:00 Jim Tomkins – Future Directions for ASCI Clusters/MPPs

APPLICATIONS PERFORMANANCE

11:20 Thomas Christopher – How should existing applications influence future supercomputer designs?

11:40 Darren J. Kerbyson – Predicting Achievable Application Performance on Future Systems

12:00 Lunch

RUN-TIME SYSTEMS ISSUES

1:20 Vitus Leung – Node Allocation on Network-Bound Parallel Computers

1:40 Barney Maccabe/Ron Brightwell – FAST-OS: Scalable Technology for Runtime and Operating Systems

APPLICATIONS

2:00 Michael A. Bender – Cache-Oblivious Data Structures

2:20 Larry Rudolph – The Importance of Metrics While Keeping the End-Goal in Sight

2:40 Break

WORKING SESSION

3:00 The workshop will choose one of the three organizations below. The participants will split into groups according to the selected organization and convene into separate rooms. Each group will then separately discuss and prepare an analysis of the stated question.

Organizational Choice 1: Architecture

For each of the candidate supercomputer architectures below, what problems does the architecture have that need further research or development? Which other architectures address those problems?

Group 1: PIM architecture.
Group 2: Discrete MPP architecture.
Group 3: Cluster architecture.

Organizational Choice 2: Business versus Technology

Group 1: How can the cost and performance of future supercomputers be evaluated objectively across all architectures?	Group 2: What criteria should the Government use in procuring supercomputers? Consider leveraging procurement funds in addition to technical issues.

Organizational Choice 3: Research Issues

Group 1: What hardware and technology research issues must be addressed on the way to Petaflops-level supercomputers?	Group 2: What software and applications issues must be addressed on the way to Petaflops-level supercomputers?

4:40 Presentation of Group Results to Entire Workshop

5:20 Erik DeBenedictis - Wrap Up

5:30 End

Workshop Introduction

Erik DeBenedictis
Sandia National Laboratories

Technology changes: The speed of light does not obey Moore's Law and increase exponentially. This has broad implications for supercomputers several generation out: individual chips must pipeline data movement on their centimeter scales and MPP/clusters suffer higher costs for less effective interprocessor interconnects. If one extrapolates supercomputer trends, these factors will cause progressively larger and unacceptable loss of machine efficiency. To reverse this efficiency loss will require a systems-level approach that considers

- technology (CMOS & Moore's Law)
- computer architecture and implementation styles (packaging)
- applications
- operating systems
- costs

This is a Petaflops Workshop organized by the Government. We seek to build on previous work, but focused the Government's computing needs. The Government needs new and faster computers to meet a focused but fairly broad range of computing needs rather than unfocused basic research in computer science.

We hope that the group of experts assembled at this workshop can make a roadmap for supercomputers that can run applications of interest to the Government at the 1-10 petaflops level.

Hitting the Kilowatt per Square Foot Wall

Carl Diegert
Sandia National Laboratories

The total cost of high performance computing platforms includes the cost of the building to house them, and cost of cabinets and interconnects to provide physical infrastructure. Machine rooms to house ASCI computers are constructed very much like typical collocation facilities, using access flooring and fancoil coolers. While commercial collocation facilities are straining to accommodate equipment like dense blade servers that demand power and cooling at 200 watts per square foot, ASCI platforms have reached higher density with custom packaging of their electronics. The ASCI RED machine at Sandia National Laboratories occupies about 1600 square feet (85 cabinets) and consumes about 1.6 megawatts of electrical power, or about a kilowatt per square foot. RED has been operating reliably since it cleared a benchmark at 1.34 teraflops in the spring of 1997. The IBM Blue Gene now in design will also occupy 1600 square feet (40 by 40 feet), and consume fewer than two megawatts (IBM Systems Journal, v.40, n.2, p.322, 2001). While performance has progressed from RED's teraflop to Blue Gene's petaflop level, the density remains at about a kilowatt per square foot. We will examine the packaging designs that reached the kilowatt per square foot density and provide some analysis that shows why density has not progressed beyond this level.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

Sandia Petaflops Planner

Erik DeBenedictis
Sandia National Laboratories

The Sandia Petaflops Planner is a tool for projecting the design and performance of parallel supercomputers into the future. The mathematical basis of these projections is the International Technology Roadmap for Semiconductors (ITRS, or an detailed version of Moore's Law) and DOE balance factors for supercomputer procurements. The planner is capable of various forms of scenario analysis, cost estimation, and technology analysis. The tool is described along with technology conclusions regarding PFLOPS-level supercomputers in the upcoming decade.

MIND: A PIM Strategy for Petaflops Computing

Dr. Thomas Sterling
California Institute of Technology

In spite of the advances in systems capable of Teraflops scale peak performance, many factors related to their architecture limit their efficiency, programmability, reliability, and scalability. It is possible that the conventional strategies embodied by MPPs and commodity clusters may have to be significantly modified to provide a viable approach for effective Petaflops-scale computing by the end of the decade. Processor-in-Memory (PIM) technology and architecture provides the opportunity for realizing new structures that may enable an innovative class of high end computing architecture delivering dramatic improvements in performance, power, size, cost, and reliability. MIND (Memory, Intelligence, and Networking Devices) is a new PIM architecture under development to support general purpose high end computing both in homogeneous standalone arrays and as smart memory for heterogeneous distributed shared memory systems. MIND/PIM addresses critical aspects of parallel computing including latency, memory bandwidth, overhead, contention, and parallelism. It also provides a framework to address the challenge of active fault tolerance. Surprisingly, it may simplify rather than complicate the problem of parallel programming by facilitating efficient dynamic adaptive resource management. This brief talk will discuss the potential opportunity of exploiting PIM to greatly enhance the capability of high end computing systems and describe the specific attributes of the MIND architecture devised to achieve this objective.

Polymorphic Scientific Computing on the UT-Austin TRIPS Processor

Doug Burger
University of Texas at Austin

Commodity microprocessors are not ideal for scientific computing. In the UT-Austin TRIPS processor, we are designing hardware support for a number of execution modes, called morphs. Each morph serves a major class of applications, from single-thread, control-bound codes, to multithreaded server codes, to highly parallel scientific codes. In this talk, I will describe the features in the TRIPS S-morph, which explicitly supports scientific computation.

Future Directions for ASCI Clusters/MPPs

Jim Tomkins
Sandia National Laboratories

To Be Determined

How should existing applications influence future supercomputer designs?

Thomas Christopher
Consultant to Sandia National Laboratories

There is a growing collection of applications written for current massively parallel processors, so in considering future supercomputer designs, it would behoove us to consider how well these applications would run on them. Starting with particle transport algorithms, we are using mathematical and simulation models to help predict the performance of the algorithms on possible future hardware. Many questions arise in how to evaluate radically different designs, especially involving whether and to what extent programmers might adapt the codes to the hardware.

Predicting Achievable Application Performance on Future Systems

Darren J. Kerbyson
Los Alamos National Laboratory

Performance Modeling is an important tool that can give information on application performance prior to system availability. Incorporating key characteristics of both code and systems in a model, a range of performance scenarios can be examined, and what-if questions answered. Models of several ASCI applications have already been developed at Los Alamos. These have found use in: explaining currently achieved performance on existing machines, exploring alternatives in code implementation strategies, predicting performance for procurement, and predicting achievable performance on future hypothesized systems.

Node Allocation on Network-Bound Parallel Computers

Vitus Leung
Sandia National Laboratories

The lightening-fast custom network for the ASCI Red supercomputer made node allocation trivial: any placement was about as good as any other. However, commodity-based supercomputers such as Cplant are network limited, and petaflop-scale machines are likely to have bandwidth and network-speed issues also. To obtain maximum throughput in network-limited systems, jobs should be allocated to localized clusters of processors. This minimizes communication costs and avoids bandwidth contention caused by overlapping jobs.

We consider three processor topologies for future supercomputers: 2D meshes, 3D meshes, and everything else. We will present a strategy for node allocation on general processor topologies. In particular, we order the processors so that processors that have similar ranks in the order are physically close in the network. We then solve a one-dimensional allocation problem.

Ultimately the quality of any node-allocation strategy is determined by its performance on a real system. However, it is extremely expensive to rigorously test strategies on production (and research) machines and future systems don't exist. We will discuss issues of how to effectively use simulation to evaluate node allocation strategies. In particular, how can we use simulation when we cannot determine the exact effect of allocation environment on running time?

Joint work with:
Esther Arkin (SUNY Stony Brook)
Michael Bender, SUNY Stony Brook
David Bunde (University of Illinois)
Jeanette Johnston (Sandia National Laboratories)
Alok Lal (Tufts University)
Joseph S.B. Mitchell (SUNY Stony Brook)
Cynthia Phillips, Sandia National Laboratories
Steven Seiden (Lousiana State University).

FAST-OS: Scalable Technology for Runtime and Operating Systems

Barney Maccabe
University of New Mexico

Ron Brightwell
Sandia National Laboratories

To Be Determined

Cache-Oblivious Data Structures

Michael A. Bender
SUNY Stony Brook

A new promising line of research is to develop data structures and algorithms that run efficiently on a hierarchical memory, even though they avoid any memory-specific parameterization (e.g., block sizes or access times). Such platform-independent algorithms are said to be cache-oblivious. If a cache-oblivious algorithm works optimally on a two-level hierarchy, then it works optimally on all levels of a multilevel memory hierarchy; cache-oblivious algorithms automatically tune to arbitrary memory architectures.

This talk summarizes the recent results in cache-oblivious data structures. First we present cache-oblivious B-trees, which match the performance of standard B-trees. Then we summarize other cache-oblivious structures, including cache-oblivious priority queues, tries, and dynamic structures supporting efficient scans.

Joint work with L. Arge, R. Cole, E. Demaine, Z. Duan, J. Iacono, M. Farach-Colton, B. Holland-Minkley, I. Munro, and J. Wu.

==========

Michael A. Bender is an Assistant Professor of Computer Science at SUNY Stony Brook. He received his BA in Applied Mathematics from Harvard University in 1992 and obtained a DEA in Computer Science from the Ecole Normale Superieure de Lyon, France in 1993. He completed a PhD on Scheduling Algorithms from Harvard University in 1998.

The Importance of Metrics While Keeping the End-Goal in Sight

Larry Rudolph
MIT

One of the few general theorems in scheduling tells us that the load can be increased up to a certain point with little adverse affects. But when increased beyond that point, response time dramatically increases. This is true for batch or interactive jobs and for people as well. The design of a large-scale petaflop supercomputer must consider the job completion time of the real human job part of which consists of multiple computer jobs. This talk will present some metrics for a job scheduling, some suggestions for latency reduction, and some observations about supercomputer design.

Attendees

David Bader	University of New Mexico
Brian Barrett	Sandia National Laboratories
Robert Balance	University of New Mexico
Michael Bender	State University of New York at Stony Brook
Bill McLendon	Sandia National Laboratories
Bill Camp	Sandia National Laboratories
Ron Brightwell	Sandia National Laboratories
Maciej Brodowicz	Caltech
David Bunde	University of Illinois, Urbana Champaign
Doug Burger	University of Texas, Austin
John Busch	Sun Microsystems
Thomas Christopher	Sandia National Laboratories
George Davidson	Sandia National Laboratories
Erik DeBenedictis	Sandia National Laboratories
Carl Diegert	Sandia National Laboratories
Doug Doerfler	Sandia National Laboratories
Eitan Frachtenberg	Los Alamos National Laboratory
Adolfy Hoisie	Los Alamos National Laboratory
David Jackson	Ames Laboratory
Jeanette Johnston	Sandia National Laboratories
Laxmikant Kale	University of Illinois, Urbana Champaign
Roman Kaluzniacki	Department of Defense
Richard Kaufmann	Hewlett Packard
Vitus Leung	Sandia National Laboratories
Barney Mccabe	University of New Mexico
Scott Pakin	Los Alamos National Laboratory
DK Panda	Ohio State University
Fabrizio Petrini	Los Alamos National Laboratory
Cindy Phillips	Sandia National Laboratories
Steve Plimpton	Sandia National Laboratories
Neil Pundit	Sandia National Laboratories
Arnold Rosenberg	University of Massachusetts
Larry Rudolph	Massachusetts Institute of Technology
Steve Seiden	Louisiana State University
Thomas Sterling	Caltech
Jim Tomkins	Sandia National Laboratories
Shukri Wakid	Hewlett Packard

This workshop is sponsored by the
Computer Science Research Institute at
Sandia National Laboratories.

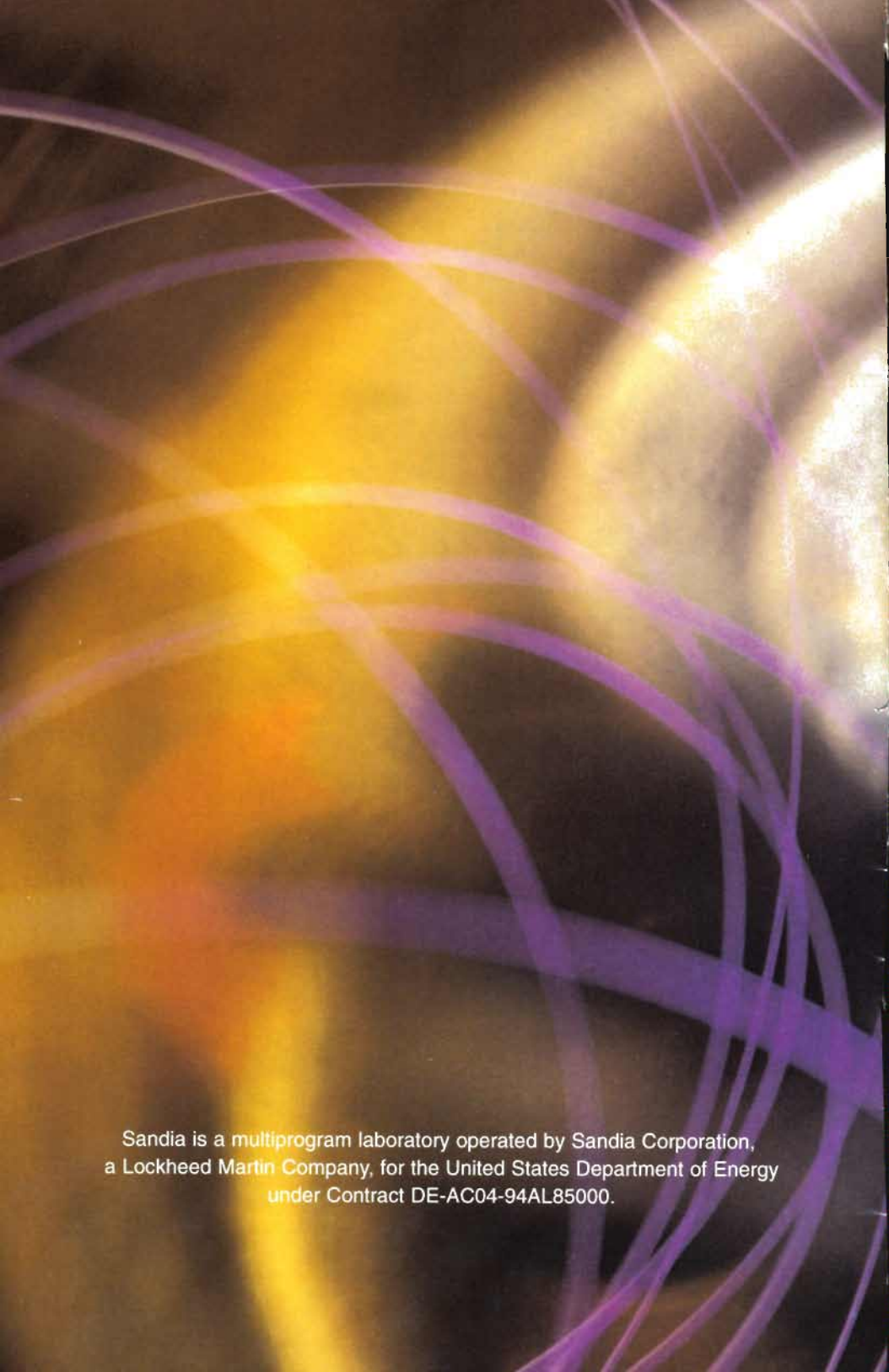
Financial support also provided by
Hewlett-Packard Corporation.

Sandia Petaflops Panel

Ron Brightwell
George Davidson
Erik DeBenedictis
Carl Diegert
Doug Doerfler
Barney Maccabe
Steve Plimpton
Jim Tomkins

Administrative Support

Deanna Ceballos
Barbara DeLap



Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy
under Contract DE-AC04-94AL85000.