Sandia’s Programs in Supercomputing

and Nanotechnology

October 23, 2007

Sudip Dosanjh
Computer and Software Systems
Sandia National Laboratories
sudip@sandia.gov
Science and Engineering Apps

- **Continuum**
  - Computational fluid dynamics
  - Shock physics (CTH)
  - Arbitrary Lagrangian Eulerian (Alegra)
  - Structural mechanics
  - Combustion
  - Device simulations
  - E&M

- **Radiation**
  - Enclosure radiation

- **DAE**
  - Circuit Modeling

- **Particles**
  - Molecular dynamics (LAMMPS)
  - Particle-in-cell
Informatics is an Emerging App

Pattern
Person
- reside
- rent
- observe

Factory
House
Truck
Fertilizer

21 West St
Honda
34 East St

Tom
Richard
Harry

Acme, Inc.

Bill
Ted

123 Main St

Gasoline
Ben
Jennifer

Bentley

Phone call

Image Source:
T. Coffman, S. Greenblatt, S. Marcus,
Graph-based technologies for intelligence analysis,
CACM, 47 (3, March 2004): pp 45-47
# Red Storm

## Before Upgrade

- 10,880 2.0 GHz single-core AMD Opteron CPUs
  - 43.52 TF/s peak
- SeaStar 1.2
- 2-4 GB per socket
- #9 on June 2006 Top 500 list
- Catamount LWK

## After Upgrade

- 13,600 2.4 GHz dual-core AMD Opteron CPUs
  - 130.56 TF/s peak
- SeaStar 2.1 network
  - Doubled NIC bandwidth
- 2-4 GB per socket
- #3 on current Top 500 list
- Catamount LWK with virtual node mode support

Link bandwidth/flop is still reasonable (approx. 1)

Some concerns about memory bandwidth/flop
Catamount Virtual Node LWK Performs Well on 7X Applications
Need Better Modeling

• Better prediction of application performance on new architectures
• Trade-off studies to determine sensitivities to key parameters
  – Improved investment of NRE
• Design of future supercomputers
Structural Simulation Toolkit (SST)

**Motivation**
- Currently developing a simulation environment to ...
  - Provide validated baseline for future exploration
  - Answer “What If” questions to guide future design efforts
  - Understand complex system-level interactions

**Goals**
- Focus on parallel systems: HW & SW
- Quick turnaround
- Flexibility
  - Multiple front-ends
    - Execution driven
    - Trace driven
  - Multiple back-ends
    - Explore novel architectures (e.g. Multi-core, NIC, Memory)
    - Support conventional architectures (e.g. Single core, DDR)
- Reusable, Extensible, & Parallelizable

**Customers**
- Micro-architects
- System Architects
- Application Performance Analysis
• Front-Ends & Back-Ends Joined by Processor/Thread Interface
• Enkidu “glues” back-end components
SST: Capabilities and Components

Processor-in-Memory
Multithreaded Processor
EDRAM
DRAM
FBDIMM Channels
PIM Network Interface
Memory Controller

Conventional Processor
SMP/CMP Processors
Heterogeneous Proc

Programmable NIC
Simple Network
2D/3D Mesh Router
PIM NIC Processor
DMA Engine
NIC BUS
Applying SST: Red Storm SeaStar NIC

- **Architectural Features**
  - Embedded 500 Mhz PPC440, local SRAM, DMA Engines, NIC Bus
  - High speed network interface to 3D mesh router
  - 800Mhz HyperTransport interface to CPU
  - Host/NIC communicate through memory

- **HyperTransport Modeling**
  - HyperTransport connection modeled at two components
  - HTLink models latency
  - HTLink_bw models link bandwidth
    - Models contention
    - Tracks backlog of requests w/ simple BW counting scheme
    - Implements flow-control with finite request queue
    - Queue depth set to cover round-trip times and allow full bandwidth

- **NIC Modeling**
  - PPC 440: Used SimpleScalar
  - Local SRAM: Existing SST component
  - Tx/Rx DMA engines
    - Existing component
    - Respond to same commands at RS DMA
    - Flow controlled
  - HT Interface
    - Connects CPU/NIC
  - NIC Bus
    - Connects internal NIC components (PPC, SRAM, etc...)
Validating SST: Latency & Bandwidth

- Used MPI “ping-pong” and OSU streaming BW
- Compared with real Seastar 1.2 and 2.1 chips
- Latency, message rate, and bandwidth
  - within 5% for range of sizes
## Validating SST: Primitives

<table>
<thead>
<tr>
<th>Routine</th>
<th>Simulated</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUT Command</td>
<td>0.486</td>
<td>0.592</td>
</tr>
<tr>
<td>tx_complete USER</td>
<td>0.196</td>
<td>0.154</td>
</tr>
<tr>
<td>rx_message ACK</td>
<td>0.959</td>
<td>1.002</td>
</tr>
<tr>
<td>rx_complete ACK</td>
<td>0.127</td>
<td>0.242</td>
</tr>
<tr>
<td>POST Command</td>
<td>0.477</td>
<td>0.442</td>
</tr>
<tr>
<td>rx_message USER</td>
<td>1.936</td>
<td>1.686</td>
</tr>
<tr>
<td>tx_complete ACK</td>
<td>0.114</td>
<td>0.118</td>
</tr>
<tr>
<td>rx_complete USER</td>
<td>0.230</td>
<td>0.378</td>
</tr>
</tbody>
</table>

### Sources of Error
- Small message optimization in Red Storm (<16 bytes)
- Lack of cache-line invalidation instruction
- Processor model?
Design Space Exploration

- Varying NIC Clk Freq:
  - 2X NIC Clock -> 30% improvement in latency
  - 4X NIC Clock -> 50% improvement

- Varying HT BW:
  - No effect to small messages
  - Does effect peak bandwidth

- Varying NIC Bus Latency:
  - 1/2 Bus Latency -> 8% latency reduction

- Varying HT Latency:
  - MPI latency increases linearly with HT latency
    - 4 HT transactions per MPI message
Finding the Bottleneck: Computation, Branches, or Memory

- In the node, Memory performance is key bottleneck.
- Even perfect branch prediction and infinite FUs would be less valuable than improving memory latency.
- Prefetching, caches don’t help emerging applications.
Latency/Bandwidth Sensitivity

Scientific Applications

Emerging applications more sensitive to Latency and Bandwidth

Informatics Applications

Latency & Bandwidth are both constraining performance
Memory Operations Dominate

- FP ops (“Real work”) < 10% of Sandia codes
- Several Integer calculations, loads for each FP load
- Memory and Integer Ops dominate
  - ...and most integer ops are computing memory addresses
- Theme: processing is now cheap, data movement is expensive
Application Characteristics

Benchmark Suite Mean Temporal vs. Spatial Locality

- Sandia FP
- LINPACK
- SPEC FP
- SPEC Int
- Sandia Int
- RandomAccess
We Need a Change of Mindset

- FLOPS are “free”. In most cases we can now compute on the data as fast as we can move it.
- CPUs (cores) must be optimized for efficient coordinated data movement.
- Compilers/tools must enable applications to benefit from multi-core CPUs
- Applications should be designed to minimize data movement.
Issues

• Opportunity cost associated with building such a machine
• Industry interest in investigating different packaging technologies at Sandia
Example Prototype Machine

- 3D Stacked Homogeneous Processing-In-Memory (PIM) Array
  - Hardware support for multithreading/thread migration
  - Enhanced Synchronization
  - Low latency/high bandwidth 3D stacked memory system
  - Highly scalable
    - Tight integration with network
    - Short vector processing
- Small Array (10’s-100’s of chips, 100’s of GBs of memory), boards, software
- Industry collaboration for the memory system
Technical Challenges

• Architecture
  – New Multithreaded Architecture
  – New Synchronization Mechanisms
  – New ISA

• System Software
  – Thread and Global Address Space Management

• VLSI Implementation
  – New (but simple!) architecture, power, validation

• Fabrication and Packaging
  – 3D integration, network implementation (SERDES or optics)

• Algorithms and Applications
  – Mapping to new architecture/programming model
  – New Application Classes (e.g., informatics)

• Compilers and Programming Models
  – Expressing multilevel parallelism and synchronization
  – Lack of easy infrastructure for targeting new architectures

• System Integration
  – Actually bringing a machine up in the lab
MESA Complex

System Engineering

Weapons Integration Facility
374 people
162,000 GSF
Construction: $77M
Equipment: $16M

Science

MicroLab
274 people
131,000 GSF
Construction: $55M
Equipment: $13M

MicroFab
0 people
98,000 GSF

Components

Integrated, Co-located Capability for Design, Fabrication, Packaging

TOTALS: 391,000 GSF
648 People

Construction: $246M
Contingency: $48M
Equipment: $168M
TEC: $462M
Relevant Sandia Capabilities

Micro Ion Trap Fabrication
• Design of micro ion traps
• Microfabrication of MEMS-based micro ion traps
• Simulation of ion trap potentials and ion trajectories
• Robust packaging of micro ion trap arrays

Integrated Micro-optic Elements
• Design, modeling, and fabrication of MEMS-based micro mirrors for micro-optic applications
• Integration of micro mirrors and solid-state waveguides
• Control algorithms for micro-mirror operation
Center for Integrated Nanotechnologies
Sandia National Laboratories • Los Alamos National Laboratory

“One scientific community focused on nanoscience integration”

- World-class scientific staff
- Vibrant user community
- State-of-the-art facilities
- A focused attack on nanoscience integration challenges
- Leveraging Laboratories’ capabilities
- Developing & deploying innovative approaches to nanoscale integration
- Discovery through application with a diverse portfolio of customers
CINT Thrust Areas provide broad base of expertise

Nanoelectronics & Nanophotonics: Precise control of electronic and photonic wavefunctions

Nano-Bio-Micro Interfaces: Biological principles & functions imported into artificial bio-mimetic systems

Complex Functional Nanomaterials: Relationships between synthesis, structure and complex and emergent properties

Nanomechanics: Understanding the mechanical behavior of nanostructured materials

Theory & Simulation: Theoretical, modeling and simulation techniques for multiple length and time scales and functionality
Future challenges

• Data locality on chip
• Impact of programming models
• Accelerators