Red Storm Overview

Ron Brightwell
(for James L. Tomkins)
Sandia National Laboratories
Albuquerque, NM
System Architecture Goals

- Balanced System Performance: CPU, Memory, Interconnect and I/O
- Usability: Functionality of hardware and software meets needs of users for Massively Parallel Computing
- Scalability: System Hardware and Software scale, single cabinet system to 32K processor system
- Reliability: Machine stays up long enough between interrupts to make real progress on completing application runs (at least 50 hours MTBI), requires full system RAS capability
- Upgradeability: System can be upgraded with a processor swap and additional cabinets to 100T or greater
- Red/Black switching: Capability to switch major portions of the machine between classified and unclassified computing environments
- Space, Power, Cooling: High density, low power system
- Price/Performance: Excellent performance per dollar, use high volume commodity parts where feasible
Red Storm Architecture

- True MPP, designed to be a single system
- Distributed memory MIMD parallel supercomputer
- Fully connected 3D mesh interconnect. Each compute node processor has a bi-directional connection to the primary communication network
- 108 compute node cabinets and 10,368 compute node processors (AMD Sledgehammer @ 2.0 GHz)
- ~30 TB of DDR memory
- Red/Black switching: ~1/4, ~1/2, ~1/4
- 8 Service and I/O cabinets on each end (256 processors for each color)
- > 240 TB of disk storage (> 120 TB per color)
Red Storm Architecture

- Functional hardware partitioning: service and I/O nodes, compute nodes, and RAS nodes
- Partitioned Operating System (OS): LINUX on service and I/O nodes, LWK (Catamount) on compute nodes, stripped down LINUX on RAS nodes
- Separate RAS and system management network (Ethernet)
- Router table-based routing in the interconnect
- Less than 2 MW total power and cooling
- Less than 3,000 ft² of floor space for machine
Red Storm Topology

- Compute node topology:
  - 27 x 16 x 24 (x, y, z) – Red/Black split: 2,688 – 4,992 – 2,688

- Service and I/O node topology
  - 2 x 8 x 16 (x, y, z) on each end (network is 2 x 16 x 16)
  - 256 full bandwidth links to Compute Node Mesh (384 available)
Red Storm Layout
(27 × 16 × 24 mesh)

Normally Classified
Switchable Nodes
Normally Unclassified

I/O and Service Nodes
Disconnect Cabinets
I/O and Service Nodes

Disk storage system not shown
Red Storm Cabinet Layout

- **Compute Node Cabinet**
  - 3 Card Cages per Cabinet
  - 8 Boards per Card Cage
  - 4 Processors per Board
  - 4 NIC/Router Chips per Board
  - N + 1 Power Supplies
  - Passive Backplane

- **Service and I/O Node Cabinet**
  - 2 Card Cages per Cabinet
  - 8 Boards per Card Cage
  - 2 Processors per Board
  - 4 NIC/Router Chips per Board
  - PCI-X for each processor
  - N + 1 Power Supplies
  - Passive Backplane
Red Storm Architecture

- RAS Workstations
  - Separate and redundant RAS workstations for Red and Black ends of machine
  - System administration and monitoring interface
  - Error logging and monitoring for major system components including processors, memory, NIC/Router, power supplies, fans, disk controllers, and disks

- RAS Network: Dedicated Ethernet network for connecting RAS nodes to RAS workstations

- RAS Nodes
  - One for each compute board
  - One for each cabinet
Red Storm System Software

- Operating Systems
  - LINUX on service and I/O nodes
  - LWK (Catamount) on compute nodes
  - LINUX on RAS nodes

- Run-Time System
  - Logarithmic loader
  - Node allocator
  - Batch system – PBS
  - Libraries – MPI, I/O, Math

- File Systems - Lustre for both UFS and Parallel
Red Storm System Software

- Tools
  - ANSI Standard Compilers – Fortran, C, C++
  - Debugger – TotalView
  - Performance Monitor - PAPI

- System Management and Administration
  - Accounting
  - RAS GUI Interface
Red Storm Performance

- Peak of ~40 TF based on 2 floating point instruction issues per clock. Expected performance is ~10 times faster than ASCI Red.
- MP-Linpack performance: >14 TF (Expect to get ~30TF)
- Aggregate system memory bandwidth: ~55 TB/s
- Aggregate sustained interconnect bandwidth: >100 TB/s
Red Storm Performance
Processors and Memory

- Processors
  - AMD Sledgehammer (Opteron)
  - 2.0 GHz
  - 64 Bit extension to IA32 instruction set
  - 64 KB L1 instruction and data caches on chip
  - 1 MB L2 shared (Data and Instruction) cache on chip
  - Integrated dual DDR memory controllers @ 333 MHz
  - Integrated 3 Hyper Transport Interfaces @ 3.2 GB/s each direction

- Node memory system
  - Page miss latency to local processor memory is ~80 ns
  - Peak memory bandwidth of ~5.3 GB/s for each processor
Red Storm Performance
Interconnect and I/O

- Sandia/UNM Portals 3.3 programming interface
- Interconnect performance
  - MPI Latency requirements <2 µs (neighbor), <5 µs (full machine)
  - Peak link bandwidth 3.84 GB/s each direction
  - Bi-section bandwidth ~2.95 TB/s Y-Z, ~4.98 TB/s X-Z, ~6.64 TB/s X-Y
- I/O system performance
  - Sustained file system bandwidth of 50 GB/s for each color
  - Sustained external network bandwidth of 25 GB/s for each color
Red Storm Network Status

The Multiple Networks to Support Red Storm Are Installed and Operational
Red Storm Data Network

Network - A
Service Partition

Compute Partition

Network - B
Service Partition

Compute Partition

Network - A
SWITCH/Router

100/1000/10000 Ethernet Connections

Multi-Mode, Copper

Network - B
SWITCH/Router

100/1000/10000 Ethernet Connections

Multi-Mode, Copper

Bldg 880

SNLA
Network - A
Data Network

SNLA
Network - B
Data Network

10 Gig Ethernet Connections

Single-Mode

10 Gig Ethernet Connections

Single-Mode

Red Storm Data Network

10 Gig Ethernet Connections

81x216

Red Storm Data Network

10 Gig Ethernet Connections

Red Storm Data Network

10 Gig Ethernet Connections

Red Storm Data Network

10 Gig Ethernet Connections
Red Storm Connections To Production Networks

Sandia Production Networks

10 GigE

JCEL

2/10 GigE

HPSS/VIZ

8/10 GigE

HPC CORE

Tera Bit Switch/Router

Red Storm

17/10 GigE

DISCOM

10 GigE

Institutional Computing

10 GigE

(TBD) High End Drivers & Sinks

10 GigE
DisCom Access To Red Storm

ESNET

Secure Net

OC-3

ATM Switch

OC-48

LLNL & SNLL

OC-48

LANL

OC-48

Ultra Fastlane

OC-48

Fastlane

OC-3

SNLA SCN

10 GigE

SNLA Red Storm

10 GigE

DisCom

SNLA SCN

OC-3

SNLA SCN

OC-3

SNLA SCN

OC-3

SNLA SCN

OC-3

SNLA SCN

OC-3
Red Storm Project Status

- **Hardware**
  - Full system installed and integrated
- **System software is a joint project between Cray and Sandia**
  - Sandia Catamount software (Run-time and LWK) is functional and has been tested at scale
  - Currently (3/17) able to boot 2x20
  - Working toward 3x20 and 3x27
  - Limited I/O capability – Lustre not fully operational
- **Network**
  - Portals firmware is under active development
    - Currently takes interrupts on every new message
    - Latency is $\sim8.5$us
    - Bandwidth is 1.1 – 1.6 GB/s
MPI Ping-Pong

![Graph showing latency versus message size for MPICH2 and MPICH 1.2.6]
Red Storm Application Status

Nodes

Applications

<table>
<thead>
<tr>
<th>Applications</th>
<th>3/18/05 Sandia</th>
<th>3/18/05 Cray</th>
<th>3/24/05 Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOOT</td>
<td>2562</td>
<td>3827</td>
<td>3456</td>
</tr>
<tr>
<td>HPL</td>
<td>1900</td>
<td>625</td>
<td>3456</td>
</tr>
<tr>
<td>CTH</td>
<td>1900</td>
<td>2048</td>
<td>3456</td>
</tr>
<tr>
<td>ITS</td>
<td>1872</td>
<td>3827</td>
<td>3456</td>
</tr>
<tr>
<td>Sage</td>
<td>1900</td>
<td>1850</td>
<td>1344</td>
</tr>
<tr>
<td>Partsin</td>
<td>1900</td>
<td>512</td>
<td>1344</td>
</tr>
<tr>
<td>UMT2K</td>
<td>1900</td>
<td>3000</td>
<td>1344</td>
</tr>
<tr>
<td>sPPM</td>
<td>1872</td>
<td>3600</td>
<td>1344</td>
</tr>
<tr>
<td>Salinas</td>
<td>1000</td>
<td>125</td>
<td>1344</td>
</tr>
<tr>
<td>Alegra</td>
<td>512</td>
<td>256</td>
<td>265</td>
</tr>
<tr>
<td>Presto</td>
<td>1536</td>
<td>512</td>
<td>4</td>
</tr>
<tr>
<td>Calore</td>
<td>1536</td>
<td>512</td>
<td>4</td>
</tr>
</tbody>
</table>

(Chart showing the number of applications for different dates and systems)