Red Storm

Jim Tomkins

SOS7 Workshop

Durango, CO

March 3-6, 2003

Outline

Sandia Experience in Paralell Computing

Sandia Application Code Characteristics

Red Storm Overview

- Design Goals
- Hardware and System Software
- Performance

Unique Aspects of Red Storm

Sandia Experience in Parallel Computing

Computer Systems - 1024 processor nCUBE 10 (1987), 16K processor CM 2 (1989),1024 processor nCUBE 2 (1990), ~3600 processor Intel Paragon (1993), ~9500 processor Intel ASCI Red (1997), 64 processor SGI O2K (1997), Cplant.

Programming Model - Explicit Message Passing

Sandia Application Code Characteristics

Most Codes are 3-D Meshes

Structured Grids
Unstructured Grids - Indirect Addressing
Adaptive Mesh Refinement - Move lots of data around machine

Sparse Matrices - Low computation to memory access ratio

Complex Equations of State - Lots of wasted cache lines

Solvers

Explicit
Implicit
Monte Carlo
Mostly Transient

Sandia Application Code Characteristics

Memory Access

Codes go through most of the node memory each time step
A lot of indirect addressing
Poor cache reuse for data
Bandwidth and Latency are extremely important to performance

Node to Node Communication

Most Codes are tightly synchronized
Lots of communication
Latency and Bandwidth are extremely important to scalability

Red Storm Design 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 ~30,000 processor system.

Reliability - Machine stays up long enough between interrupts to make real progress on completing application run (at least 50 hours MTBI), requires full system RAS capability.

Upgradability - 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 Design Parameters

True MPP, designed to be a single system.

Fully connected high performance 3-D mesh interconnect.

Topology - 27 X 16 X 24 compute nodes and 2 X 8 X 16 service and I/O nodes

108 compute node cabinets and 10,368 compute node processors. (AMD Sledgehammer @ 2.0 GHz)

~10 TB of DDR memory @ 333 MHz (1.0 GB per processor)

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 Design Parameters

Functional hardware partitioning - service and I/O nodes, compute nodes, and RAS nodes.

Functional system software partitioning - 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).

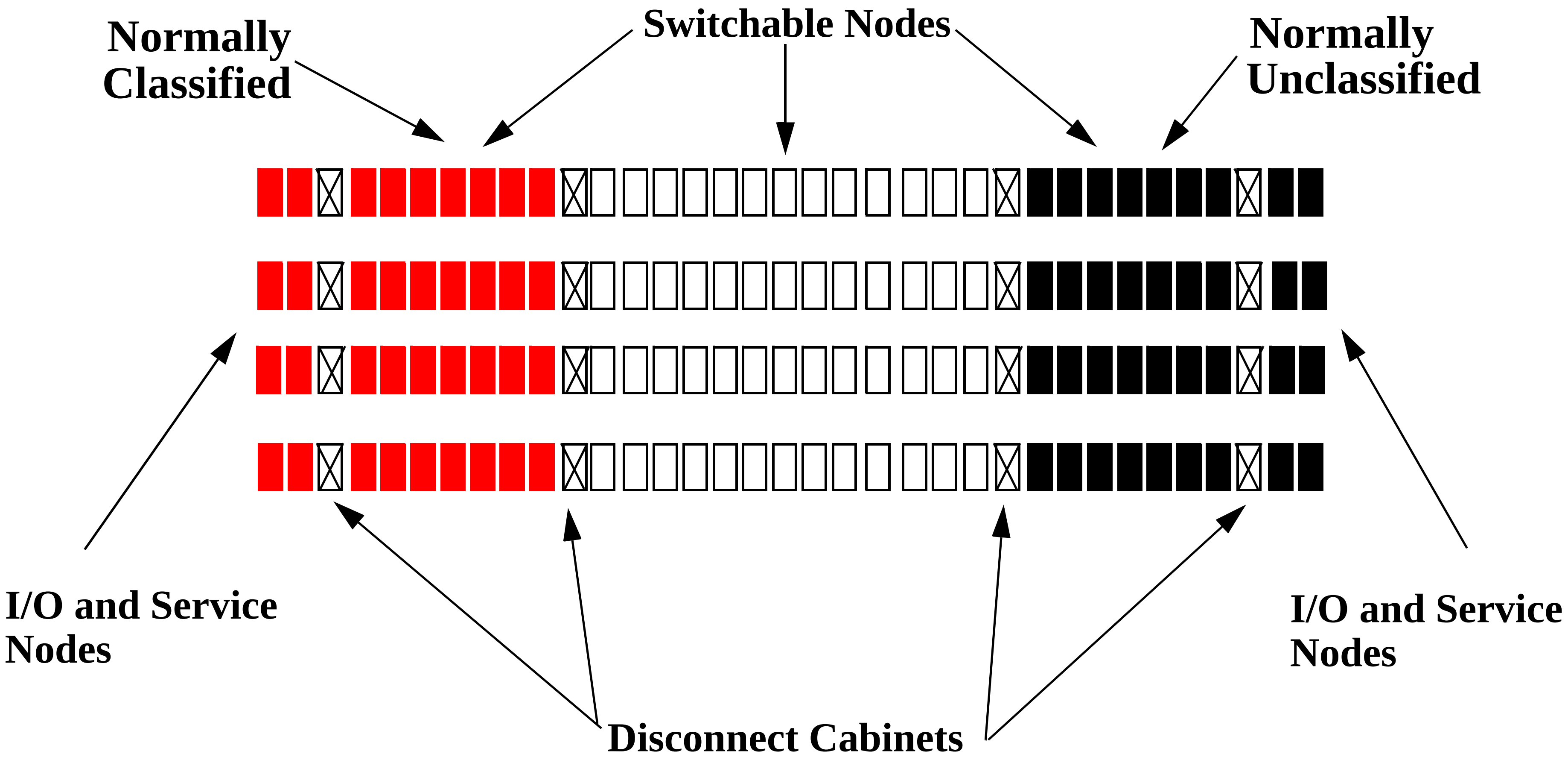
Router table based routing in the interconnect.

Less than 2 MW total power and cooling.

Less than 3,000 square feet of floor space.

Red Storm Layout

(27 X 16 X 24 mesh)



Normally Classified

Switchable Nodes

Normally Unclassified

I/O and Service Nodes

I/O and Service Nodes

Disconnect Cabinets

Disk storage system not shown

Red Storm System Software

Operating Systems

- Compute nodes - LWK (Catamount)
- Service and I/O nodes - LINUX
- RAS nodes - LINUX

Compilers - Fortran, C, C++

Debugger - TotalView

Performance Monitor

Libraries - MPI-2, Math, I/O

Red Storm RAS System

RAS Workstations

- Separate and redundant RAS workstations for Red and Black ends of machine.
- System administration and monitoring interface.
- Communicates with operating system.
- 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 Performance

Peak of ~ 40 TF

Expected MP-Linpack performance >20 TF

Aggregate system memory bandwidth - ~55 TB/s

Interconnect

- Aggregate sustained interconnect bandwidth > 100 TB/s
- MPI Latency - 2 μs neighbor, 5 μs across machine
- Bi-Section bandwidth ~2.3 TB/s
- Link bandwidth ~3.0 GB/s in each direction

I/O System

- Sustained 50 GB/s disk I/O bandwidth for each color.
- Sustained 25 GB/s external network bandwidth for each color.

Unique Aspects of Red Storm

1. **Rebirth of the tightly integrated, micro-processor based MPP.**

2. System interconnect performance.

3. Linear scalability of system from a single cabinet to 30,000+ processors.

34 The level of functional partitioning of hardware and system software.

5. Full system RAS.

6. Red/Black switching.