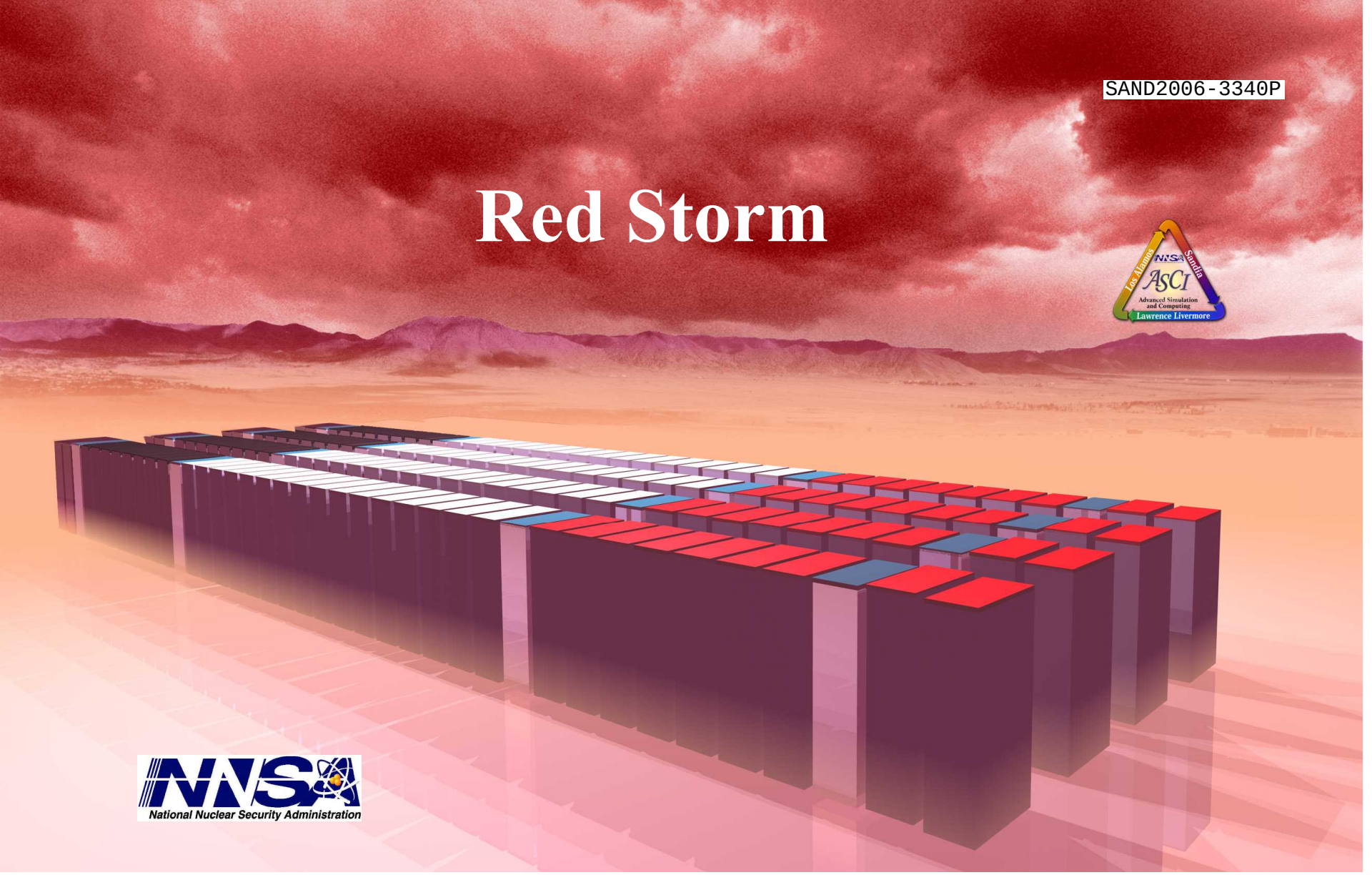
SAND2006-3340P

Red Storm



National Nuclear Security Administration

Sandia National Laboratories

Red Storm Design Goals

- Balanced System Performance: CPU, Memory, Interconnect and I/O
- Scalability: System Hardware and System Software scale, a single cabinet system to 32K processor system
- Functional Partitioning: Hardware and System Software
- Reliability: Full system Reliability, Availability, Serviceability (RAS) Designed into Architecture
- Upgradeability: Designed in path for system upgrade
- Red/Black Switching: Flexible support for both classified and unclassified Computing in a single system
- Custom Packaging: High density, relatively low power system
- Price/Performance: Excellent performance per dollar, use high volume commodity parts where feasible

Red Storm System

- True MPP, designed to be a single system
- Distributed memory MIMD parallel supercomputer
- Fully connected 3-D mesh interconnect. Each node has its own high bandwidth, bi-directional connection to the primary communication network.
- 108 compute node cabinets and 10,368 compute node processors (AMD *Opteron* @ 2.0 GHz)
- ~30 TB of compute node memory
- Red/Black Switching - ~1/4, ~1/2, ~1/4
- 8 Service and I/O cabinets on each end (256 nodes for each color)
- 400+ TB of disk storage (200+ TB per color)

Red Storm System

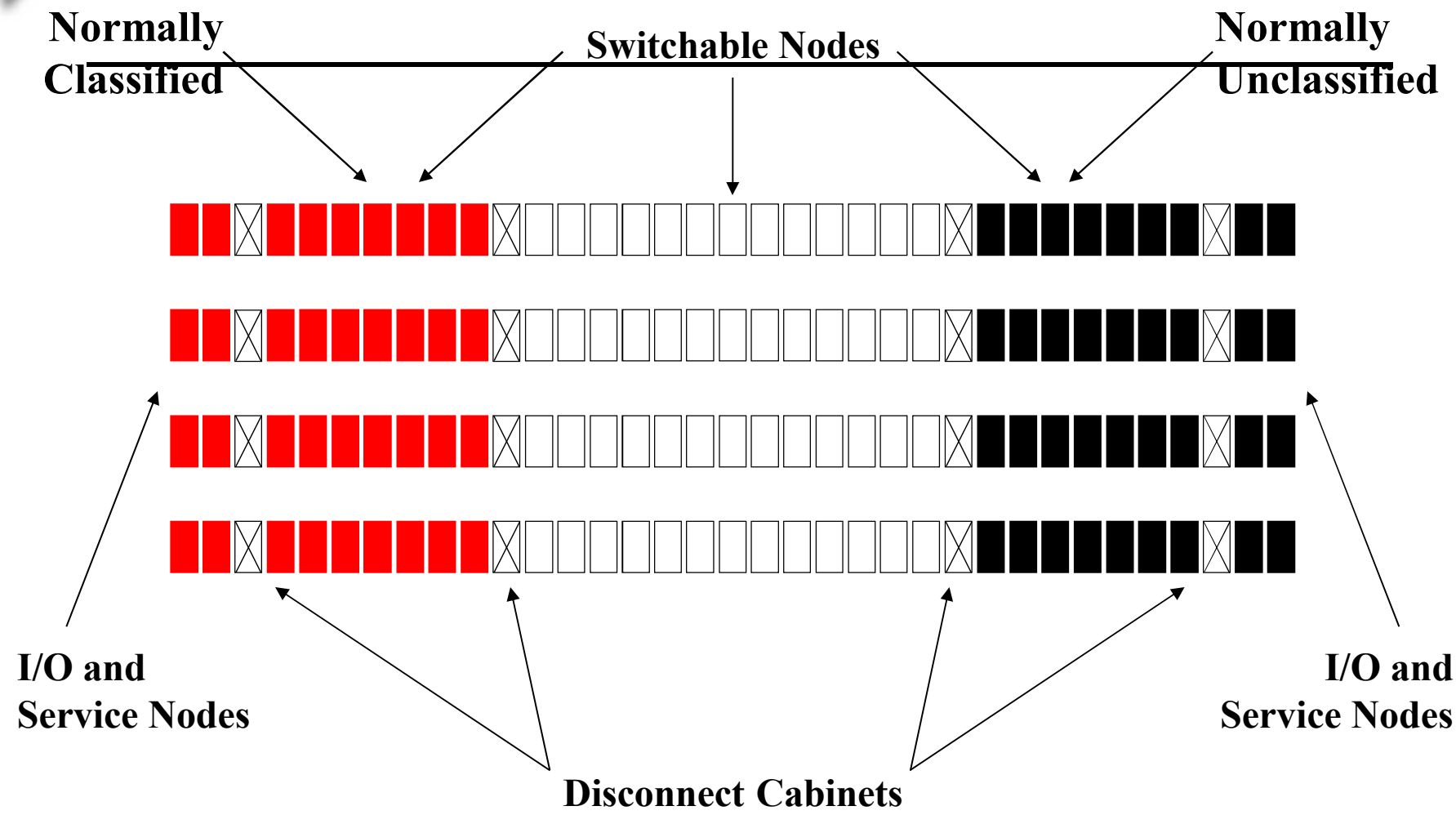
- Functional Hardware Partitioning - service and I/O nodes, compute nodes, RAS nodes
- Partitioned Operating System (OS) - LINUX on service and I/O nodes, LWK (Catamount) on compute nodes, stripped down LINUX on RAS nodes.
- Separate RAS and system management network (Ethernet).
- Less than 2 MW total power and cooling.
- Less than 3000 sq ft.

Red Storm Topology

- Compute Nodes
  - 27 x 16 x 24 (X, Y, Z) - Red/Black split 2688 - 4992 - 2688
- Service and I/O Nodes
  - 2 x 8 x 16 (X, Y, Z) nodes on each end
  - Mesh is 2 x 16 x 16 (X, Y, Z) on each end
  - Center card cage is empty

Red Storm Layout

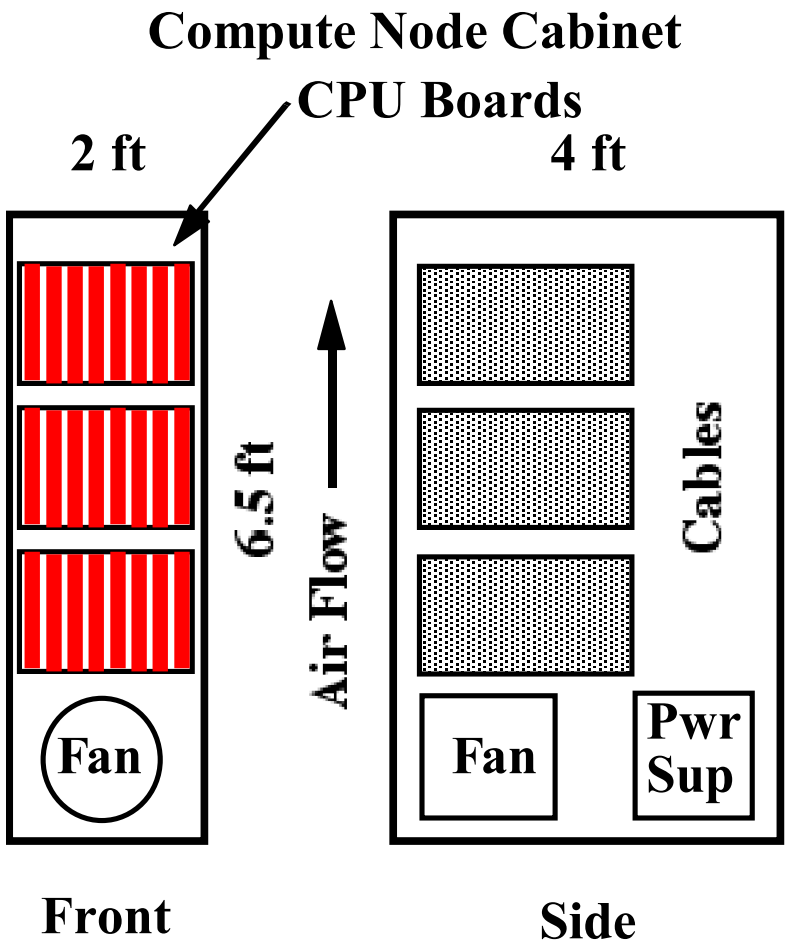
(27 × 16 × 24 mesh)



Disk storage system not shown

Sandia National Laboratories

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+2 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
- 2 PCI-X slots per Node
- N+2 Power Supplies
- Passive Backplane

Sandia National Laboratories

Red Storm System Management and RAS

- 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
- File Systems
  - Parallel File System - *Lustre*
  - Unix File System- *Lustre*
  - NFS
- Run-Time System
  - Logarithmic loader
  - Node allocator
  - Batch system – PBS
  - Libraries – MPI, I/O, Math
- Single System View

Red Storm System Software

- Programming Model
  - Message Passing
  - Support for Heterogeneous Applications
- Tools
  - ANSI Standard Compilers - Fortran, C, C++
  - Debugger - *TotalView*
  - Performance Monitor - Cray Apprentice and PAPI
- System Management and Administration
  - Accounting
  - RAS GUI Interface for monitoring system
  - Single System View

Red Storm Performance - Interconnect and I/O

- Interconnect performance
  - Latency <5 (2, 3.6) μs (neighbor) <8 (5) μs (full machine)
  - Peak Link bandwidth - 3.84 GB/s each direction
  - Peak HT bandwidth - 3.2 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 System Performance

- ~41 TF Peak - Upgradeable to well over 100TF
- Delivered Performance - Based on application code testing Red Storm is delivering over 10× performance improvement over ASCI Red on Sandia's suite of application codes.
- HPL performance - 36.19 TF (6)
- Aggregate system memory bandwidth: ~55 TB/s
- Aggregate sustained interconnect bandwidth: >100 TB/s

Sandia National Laboratories