HPC Top 10 InfiniBand Machine …
A 3D Torus IB Interconnect on Red Sky

Marcus Epperson

John Naegle **Jim Schutt**

Matthew Bohnsack **Steve Monk**

Mahesh Rajan **Doug Doerfler**

March 2010

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

Agenda

- Red Sky Background
- 3D Torus Interconnect Concepts
- Difficulties of Torus in IB
- New Routing Code for IB a 3D Torus
- Red Sky 3D Torus Implementation
- Managing a Large IB Machine

Sandia National Laboratories

Computing at Sandia

- Capability Computing
  - Designed for scaling of single large runs
  - Usually proprietary for maximum performance
  - Red Storm is Sandia's current capability machine
- Capacity Computing
  - Computing for the masses
  - 100s of jobs and 100s of users
  - Extreme reliability required
  - Flexibility for changing workload
  - Thunderbird will be decommissioned this quarter
  - Red Sky is our future capacity computing platform
  - Red Mesa machine for National Renewable Energy Lab

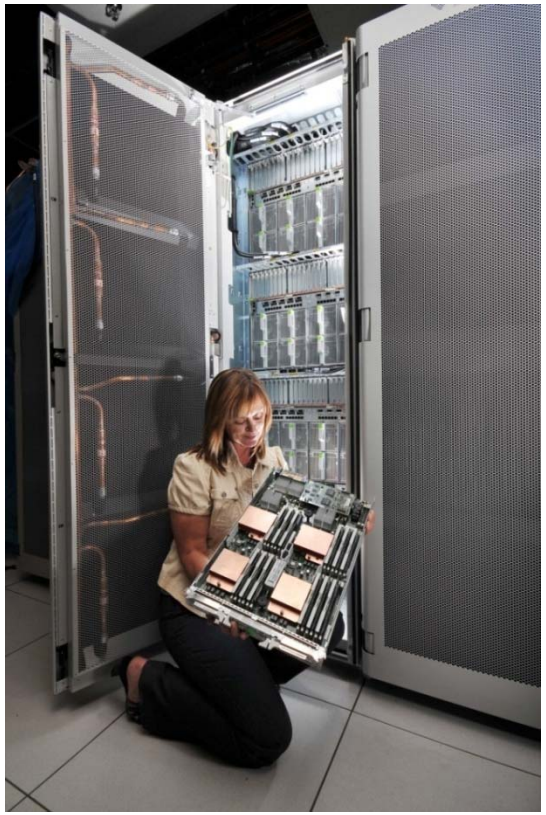
Red Sky Main themes

- **Cheaper**
  - 5X capacity of Tbird at 2/3 the cost
  - Substantially cheaper per flop than our last large capacity machine purchase
- **Leaner**
  - Lower operational costs
  - Three security environments via modular fabric
  - Expandable, upgradeable, extensible
  - Designed for 6yr. life cycle
- **Greener**
  - 15% less power … 1/6th power per flop
  - 40% less water … 5M gallons saved annually
  - 10X better cooling efficiency
  - 4x denser footprint

Sandia National Laboratories

Major innovations

- **Bridging from capacity to capability**
  - Many “Red Storm” characteristics (scaling) at commodity price
  - 2-3X *faster* than Red Storm in mid range
  - 1/3 operational costs
- **Top Red Sky innovations**
  - Petascale midrange architecture
  - Intel Nehalem processor
  - QDR Infiniband
  - 3D mesh/torus
  - Optical cabling
  - Optical Red/Black switching
  - Refrigerant cooling/glacier doors
  - Power distribution
  - Routing & Interconnect resiliency
  - Minimal Ethernet (RAS & mgmt. only)
  - Boot over IB



Sandia National Laboratories

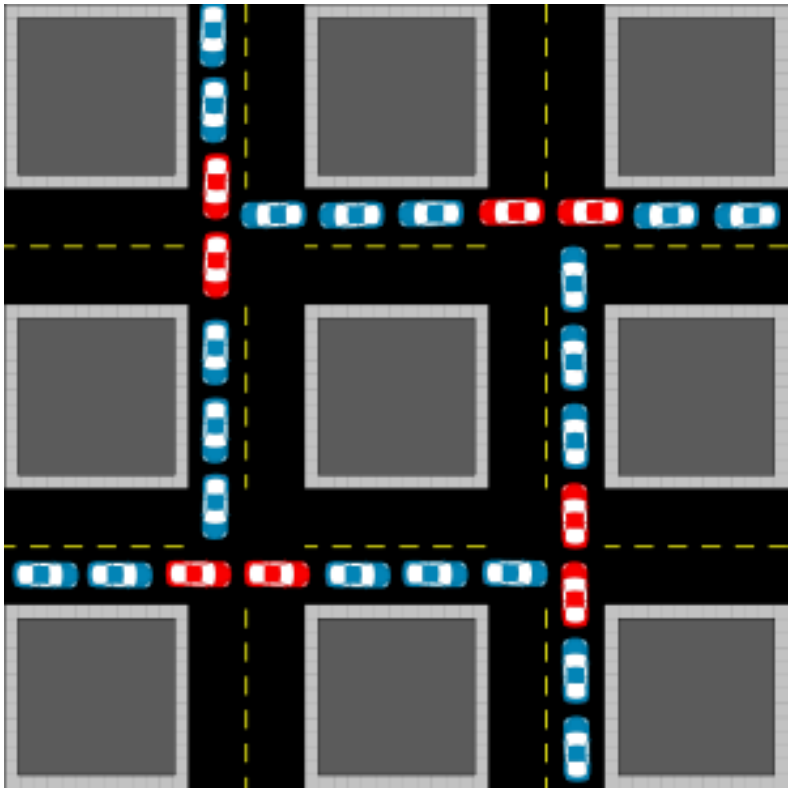
Benefits of the Torus Architecture

- 12x QDR paths in each dimension maintains reasonable bisection bandwidth/FLOP ratio
- Regular wiring enables Red/Black switching
- Physically scales linearly
- Works well for localized communication, particularly in capacity environment
- Save cost, power, cooling, and cabling of external fat-tree IB switches
- Save cost, power, cooling, and cabling of high-speed Ethernet infrastructure

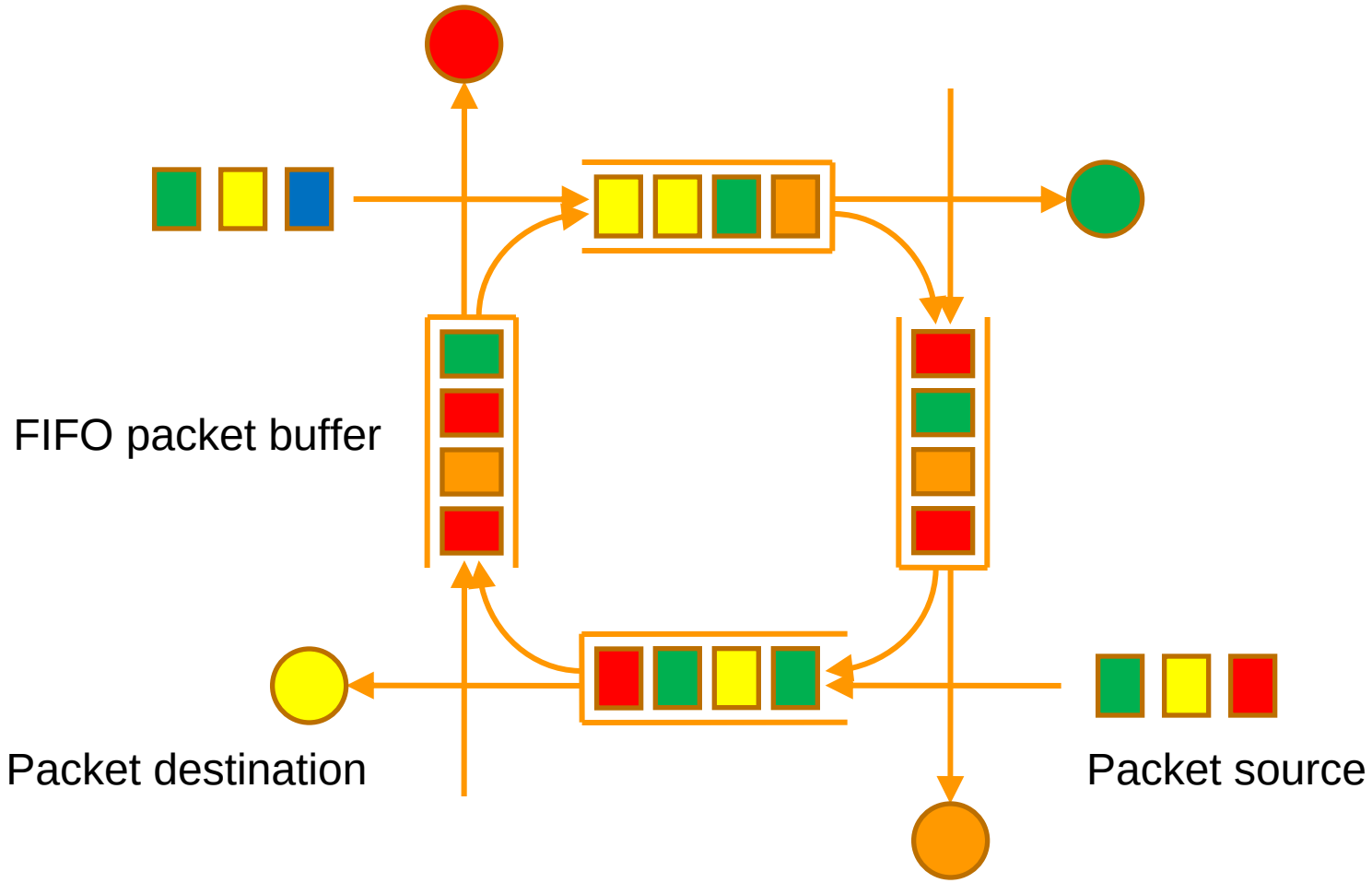
Difficulties of IB Torus

- Torus susceptible to message deadlock (credit loops)
- The IBA makes deadlock free routing difficult
  - Traditional dateline method of deadlock avoidance not possible because output Virtual Lane (VL) is not a function of input VL
  - Limited by Service Level (SL) to VL mapping and fixed sizes
  - Must use constant SL determined at source
  - Must share SL function with QoS implementation
  - Must use Path Record Queries for connection setup
  - Resiliency to switch or link failures very difficult

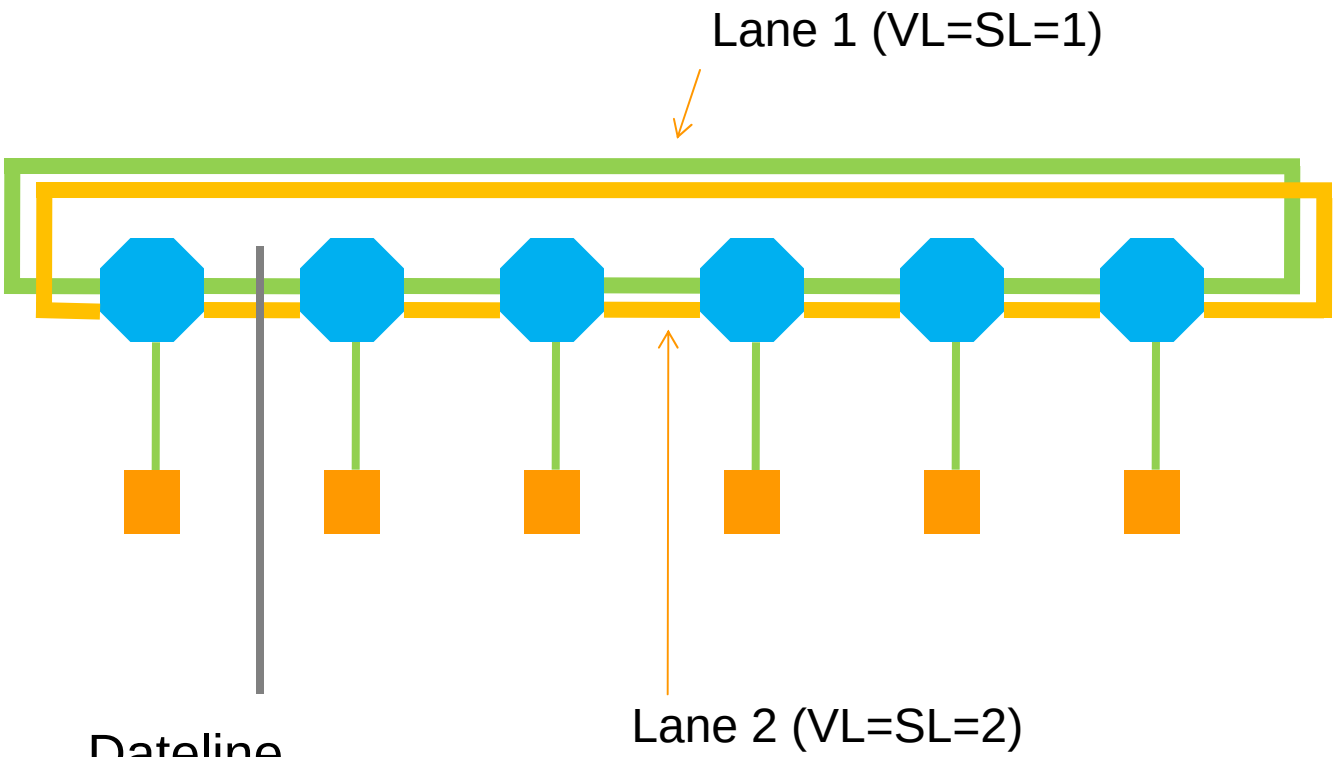
Deadlock is Gridlock for Bits



Simple Deadlock Credit Loop



Deadlock Avoidance



Dateline
Path Crosses -> SL=2
Path Does Not Cross -> SL=1

Sandia National Laboratories

Difficulties of IB Torus

- Torus susceptible to message deadlock (credit loops)
- The IBA makes deadlock free routing difficult
  - Traditional dateline method of deadlock avoidance not possible because output Virtual Lane (VL) is not a function of input VL
  - Limited by Service Level (SL) to VL mapping and fixed sizes
  - Must use constant SL determined at source
  - Must share SL function with QoS implementation
  - Must use Path Record Queries for connection setup
  - Resiliency to switch or link failures very difficult

Potential Deadlock Free Routing Algorithms

- Dimension Order Routing (DOR) with Dateline checking in Torus
  - Provably deadlock free with 2 VLs and 8 SLs
  - Not resilient to any switch or link failures
  - Requires Path Record Query to get the path SL for each connection
- LASH
  - Algorithm to map each route and add SLs when needed to avoid loops
  - Modified existing algorithm to utilize basic DOR concept to minimize number of required SLs/VLs
  - Also requires Path Record Query to get the path SL for each connection
  - A single failure may result in many path SL changes
  - Requires non-existent Path Record Update implementation for resilience to failures after initial Path Record Query
  - Route computation time scales poorly with increasing system size

Resilient Deadlock Free Algorithm (torus-2QoS)

- Novel technique to use input port to determine illegal turns and utilize secondary routing
- Heuristically demonstrated, no mathematical proof
- Passes OFED tool for deadlock checking (ibdmchk)
- Much more resilient to failures
  - Can reroute any single switch failure
  - Can reroute multiple switch failures *iff* they are adjacent in the last DOR dimension
  - Can reroute multiple link failures as long as no disjoint 1D rings are created
  - Rerouting does not change path SL
- Does NOT require Path Record Updates
- Multicast required modifications to maintain deadlock free routing in conjunction with unicast routes
- Implementation delivered to OFED mailing list
  - Implemented in OFED OpenSM
  - Standard tools for fabric simulation, route checking, and credit loops
  - Working at scale on Red Sky platform

Sandia National Laboratories

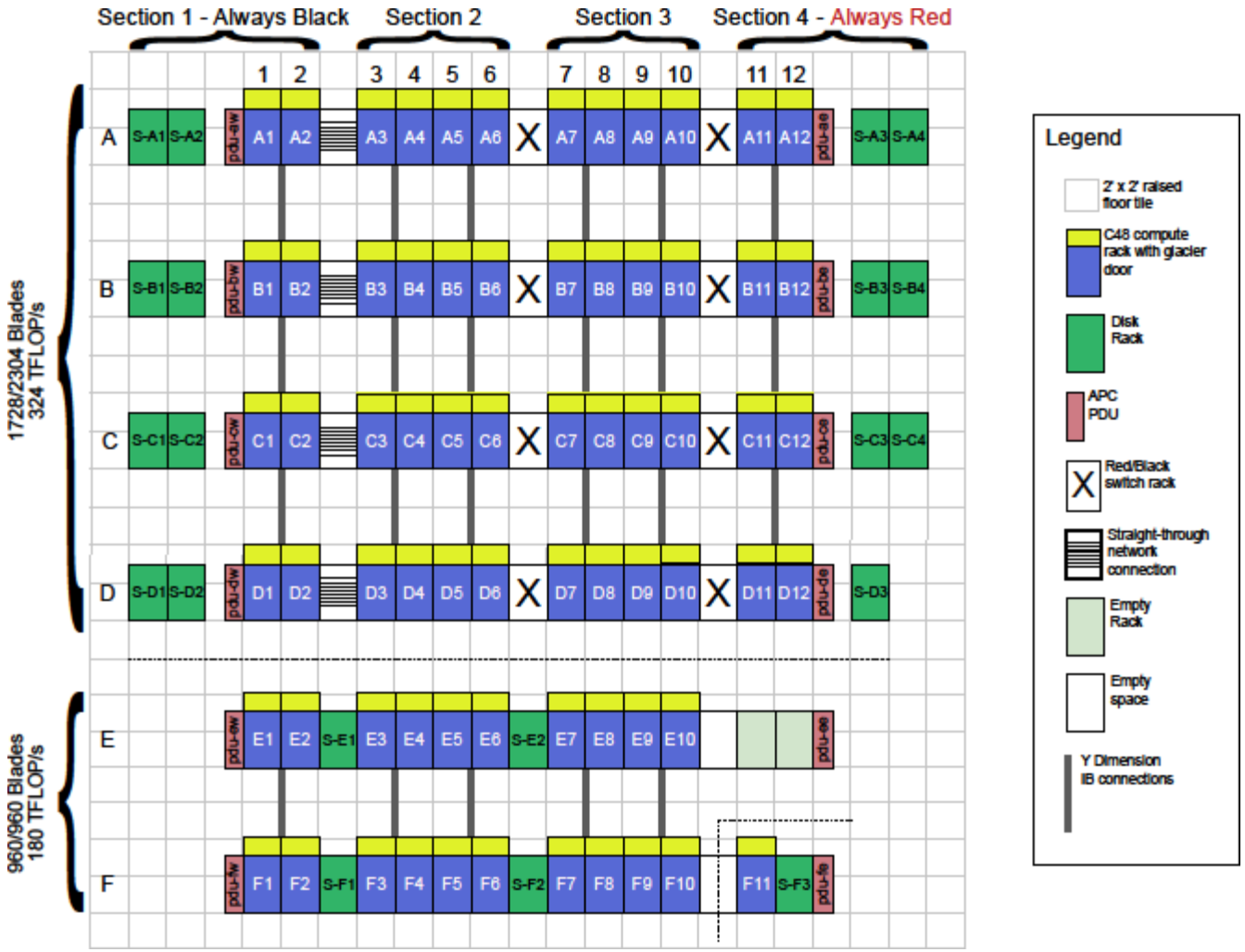
Other Issues

- Requires applications to use Path Record Queries (PRQs) to determine launching SL
  - Inherently difficult to scale, investigating options such as static tables, caching, or distributed SM
  - OpenMPI RDMA-CM implementation fixed
    - Scaling is still an issue
  - OpenMPI OOB connection setup (using IP) modified to use PRQs
    - Worked for LINPACK top 10 scale runs
  - MVAPICH2 not working with PRQs for us yet
- QoS also uses SLs, combined solution limits number of QoS levels to two
- Many of the basic IB management tools did not work with SL != 0 and needed to be fixed

Red Sky Hardware architecture

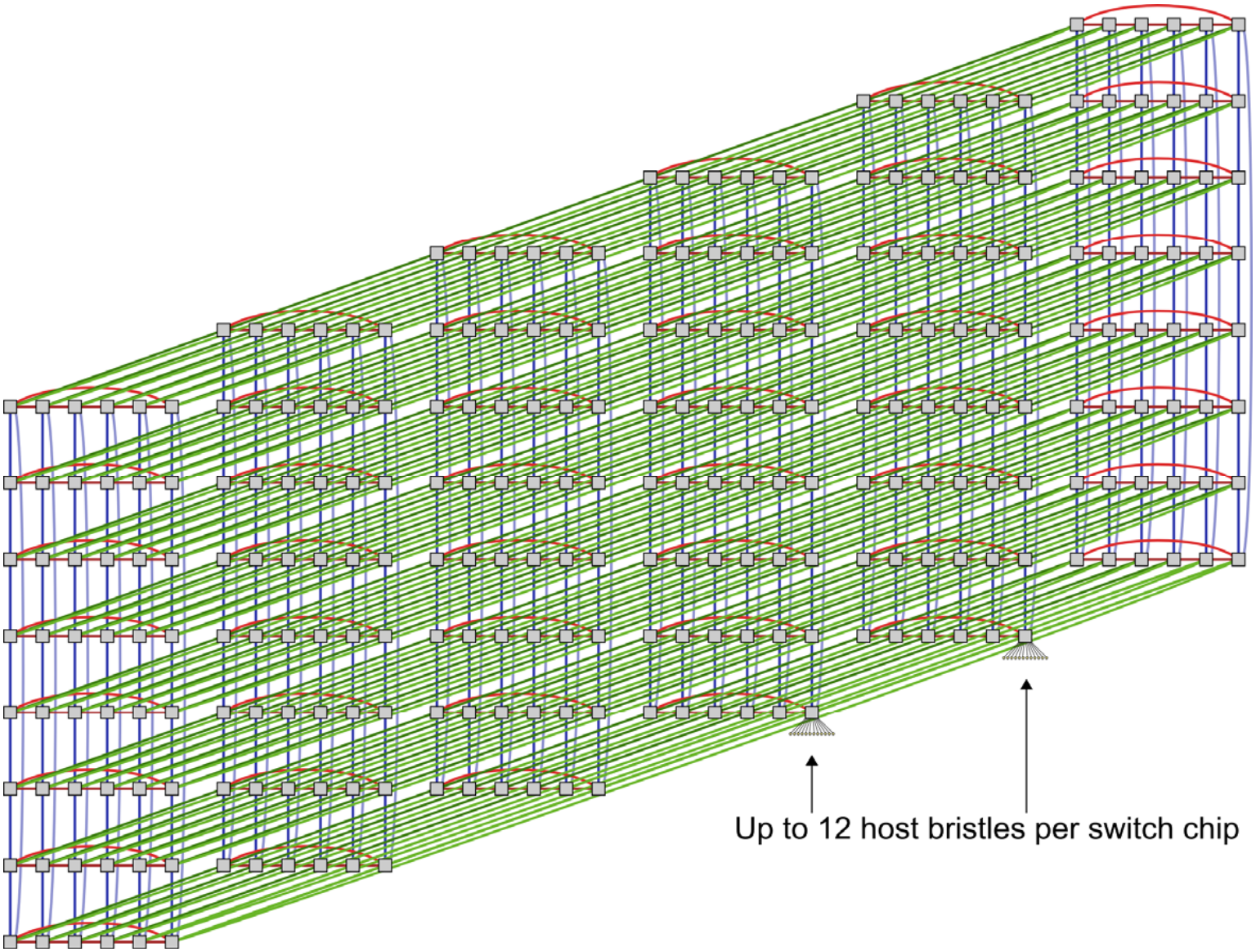
- **Topology**
  - 3D Torus with Red/Black switching
  - Mellanox 36-port InfiniBand/QDR switches
  - 3 by 4x QDR ports per dimension
  - 12 “bristled” node connections per switch
- **Racks**
  - Sun Constellation (C48) -- 48 blade slots
  - 4 Chassis per rack, 12 blades per chassis
  - Switches integrated with backplane (NEM)
  - Integrated liquid cooling doors
- **Node**
  - 2-way SMP
  - Sun Vayu Blades for Compute & Service partitions
  - 2 nodes per compute blade
  - 2 Sockets per node, Intel Nehalem CPUs w/QPI
  - 8 Cores per node
  - Mellanox ConnectX HBA

Red Sky Layout For Operational flexibility & agility



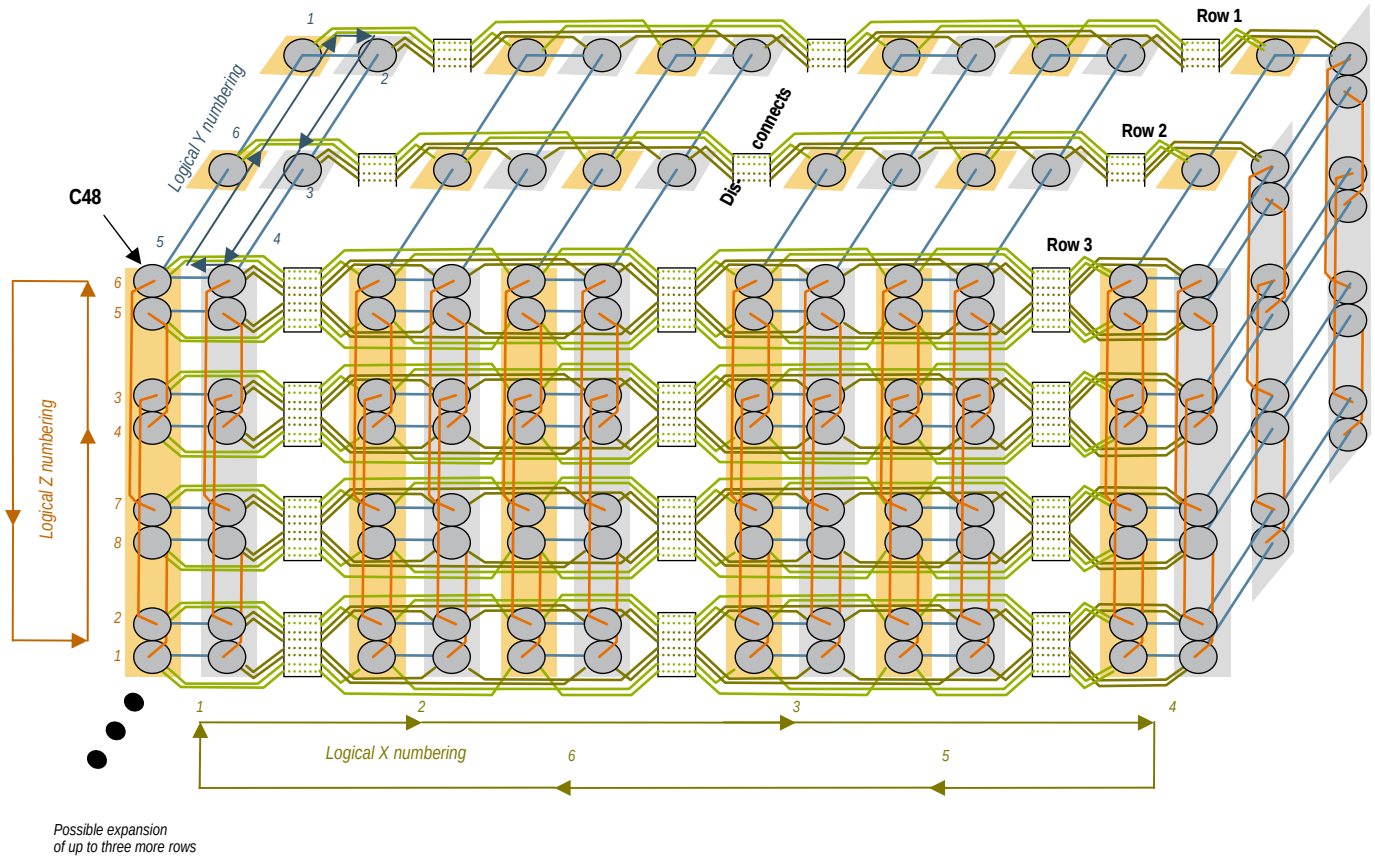
Sandia National Laboratories

Logical Layout of RedSky



Sandia National Laboratories

Logical and physical layout related



Logical: 6 x 6 x 8 (x, y, z) Physical: 12 x 3 x 8 (x, y, z)

Logical x dimension skips every other rack

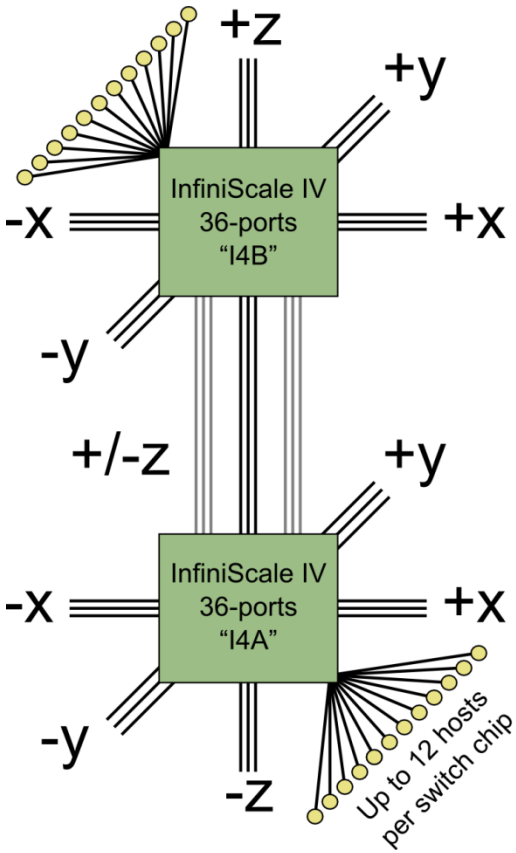
Logical y dimension “folds over” at the last physical row

Logical z dimension is self contained within a rack

Sandia National Laboratories

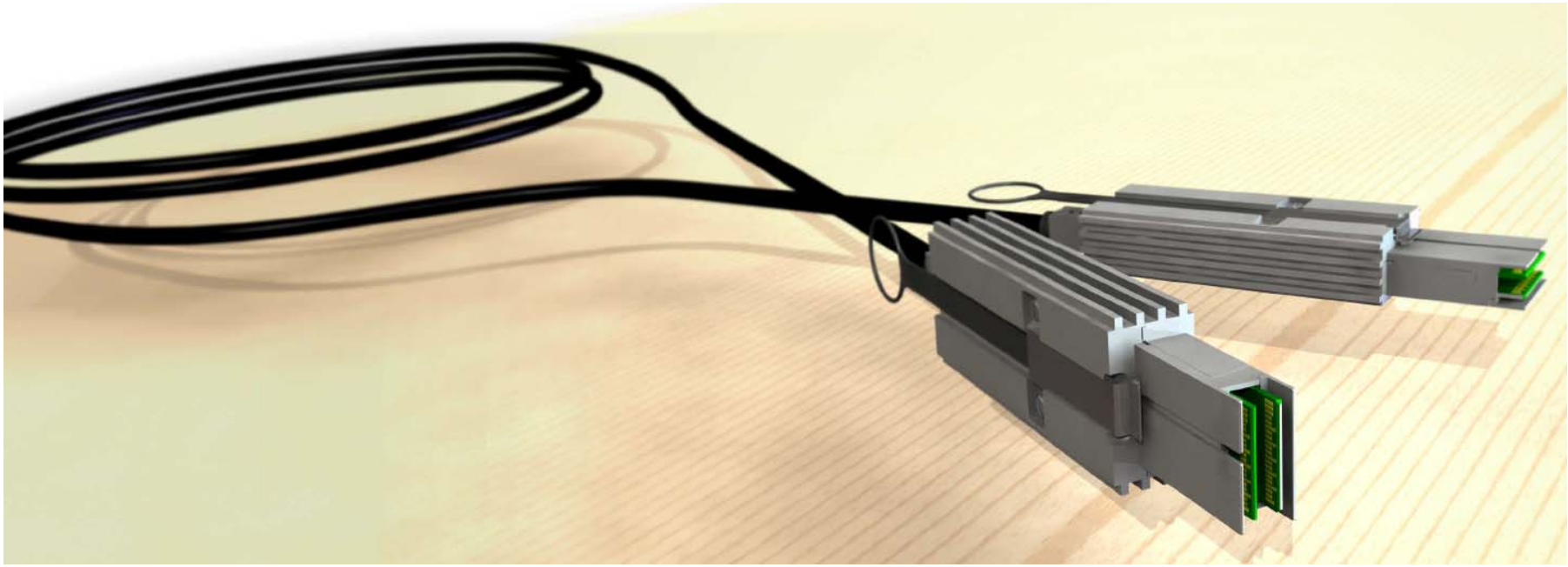
QNEM: 3D Torus Building Block

- QDR Network Express Module (QNEM)
- Two vertices per shelf, intra-shelf Z connectivity “on PCB”
- Four in each blade rack (one per shelf)



Sandia National Laboratories

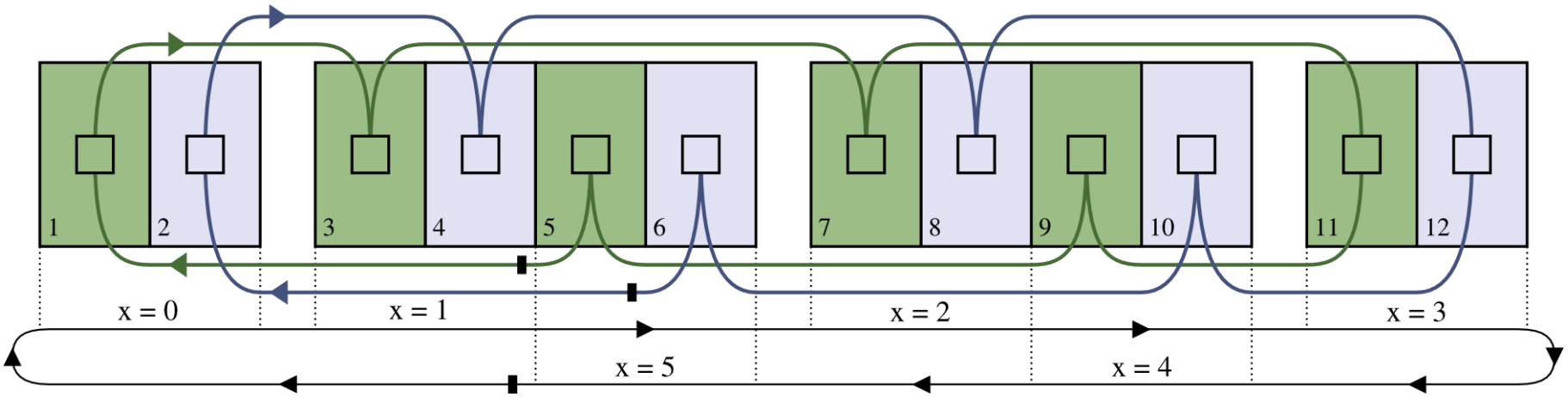
Optical cables – the long awaited jump

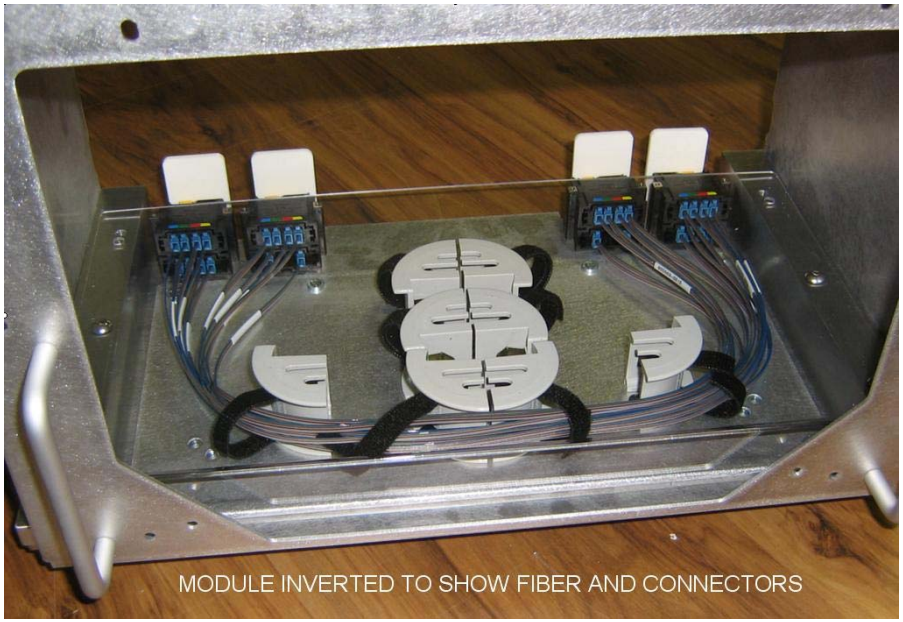


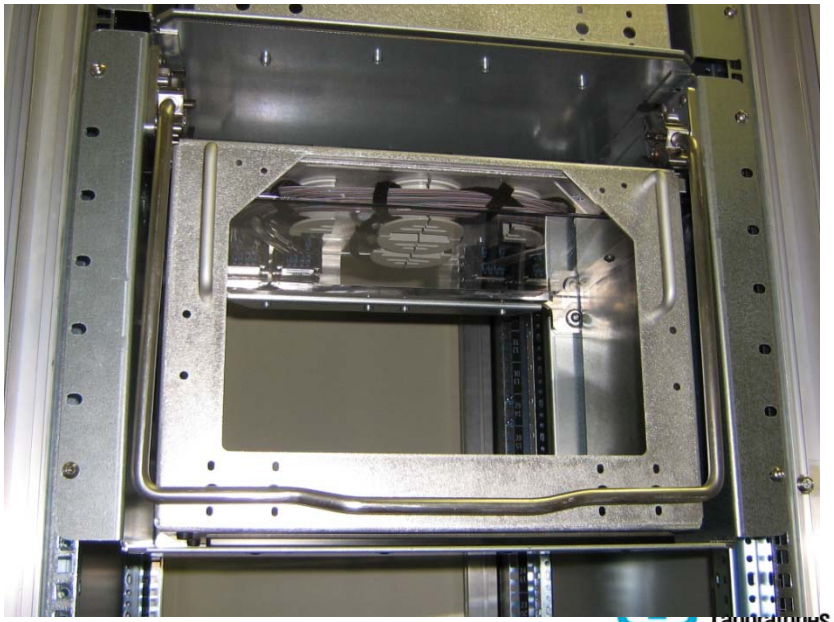
- **First deployment**
  - 12X QDR
- **High agility**
  - 5 mm x 5 mm
  - 22g per meter
  - Min. long term bending radius 5cm

Sandia National Laboratories

Optical red/black switch innovation




MODULE INVERTED TO SHOW FIBER AND CONNECTORS



Software overview

- **CentOS Linux Distribution**
- **OFED 1.4.2+, OpenMPI 1.4.1+**
- **gPXE & Boot over IB**
  - Ethernet for RAS & management only
- **OpenSM and torus-2QoS: deadlock-free routing for a torus**
- **Filesystems**
  - Lustre /home and /projects
  - Root fs via NFS over IB
- **SLURM + Moab**
  - Topology-aware job placement
- **RAS and Management**
  - SNL-developed system management toolset and RAS system
  - oneSIS + git – diskless/stateless booting with revision-controlled shared image
  - rpower - power management
  - ConMan - console access
  - FreeIPMI, IPMItool - IPMI protocol over the RAS network

Sandia National Laboratories

Management tools for IB fabric

- **Error collection and analysis**
  - **Using ibqueryerrors to gather/clear errors at regular interval**
  - **Created tool to parse raw output, simplify it, and inject topological information:**

```
** START -- Thu Feb 18 13:42:08 MST 2010 ***
d12-nem1-a,-x (14) <--> d8-nem1-a,+x (11)  --  RcvErrors == 8 , SymbolErrors == 13
d12-nem1-a,-y (8)  <--> d11-nem1-a,+y (5)  --  RcvErrors == 35 , SymbolErrors == 79
d12-nem1-b,in (21) <--> rs3284             --  SymbolErrors == 5
a12-nem4-a,-y (8)  <--> b12-nem4-a,+y (5)  --  RcvErrors == 1 , SymbolErrors == 5
c11-nem4-b,in (19) <--> rs3004             --  RcvErrors == 7 , SymbolErrors == 1117
*** END ***
```

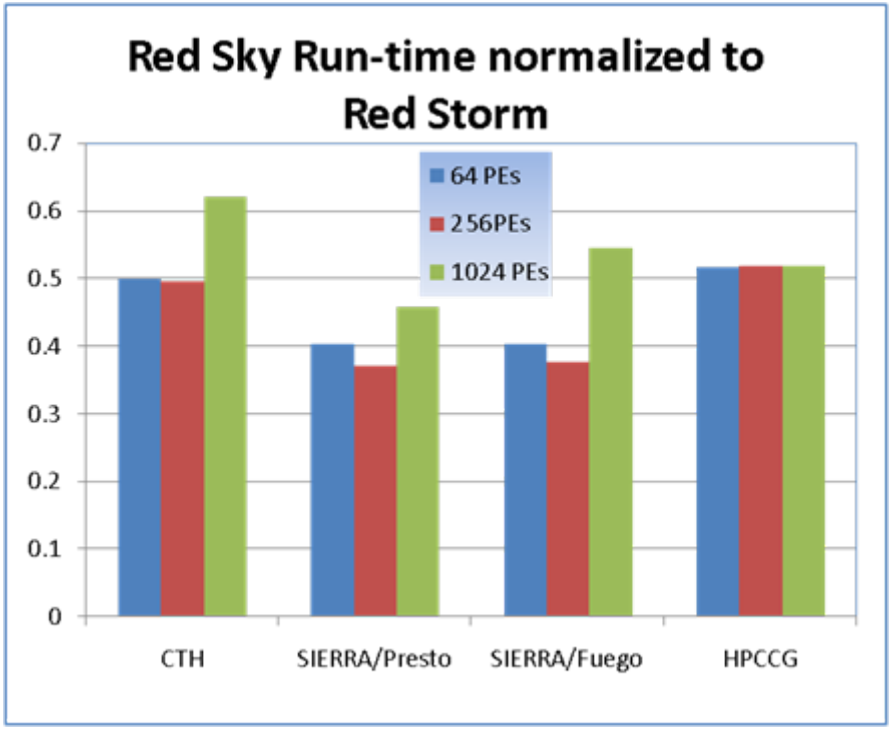
- **Additional tools utilize this enhanced output for further diagnostics**

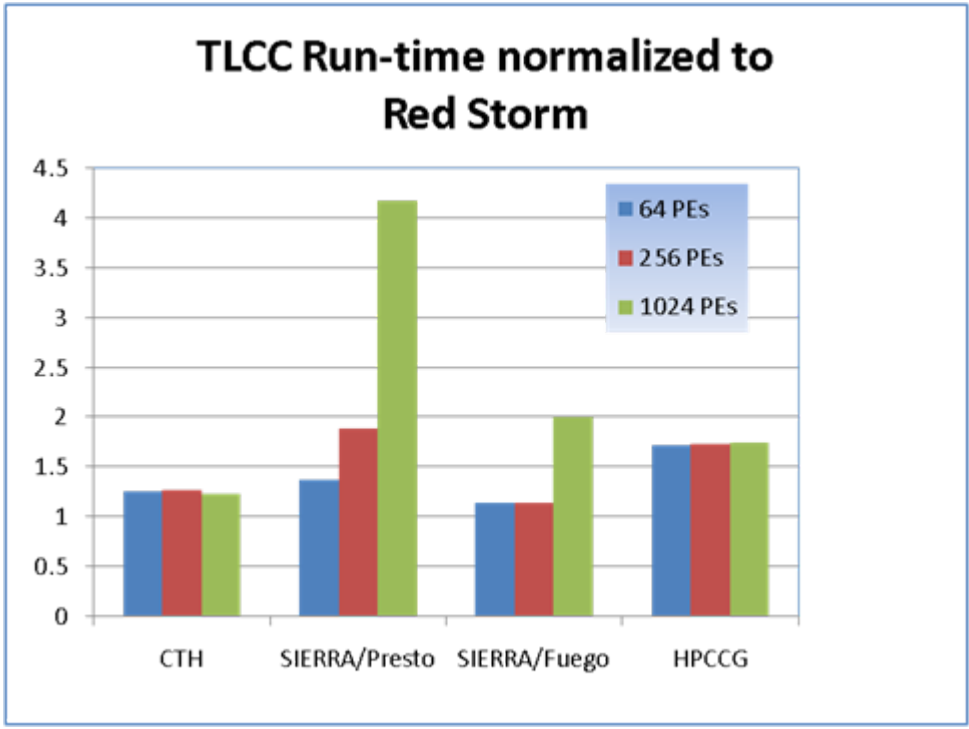
Sandia National Laboratories

Management tools for IB fabric (continued...)

- OpenSM log monitoring
  - torus-2QoS provides additional error and link state change reporting (fewer links in a torus, so each link is more important)
  - Wrote tool to watch log and inject useful high-level information in-line
    - Translate GUIDs and Directed Routes (DRs) to human-readable device names
    - Color output for certain message classes
- Error handling
  - When a redundant link shows many errors, better to disable it and reduce bandwidth than to have it cause retries
  - Tools for link state management via DRs

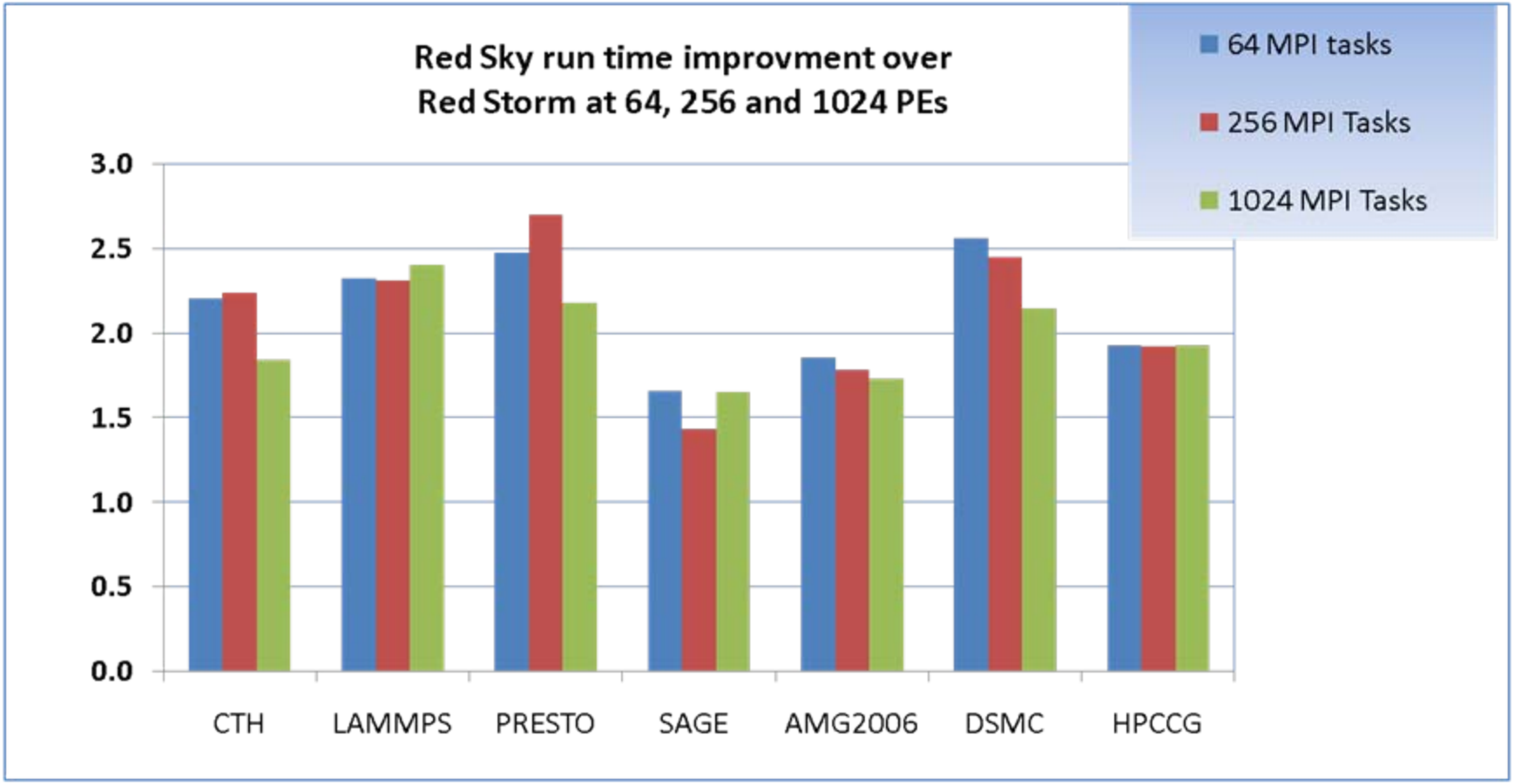
Red Sky performance and scaling





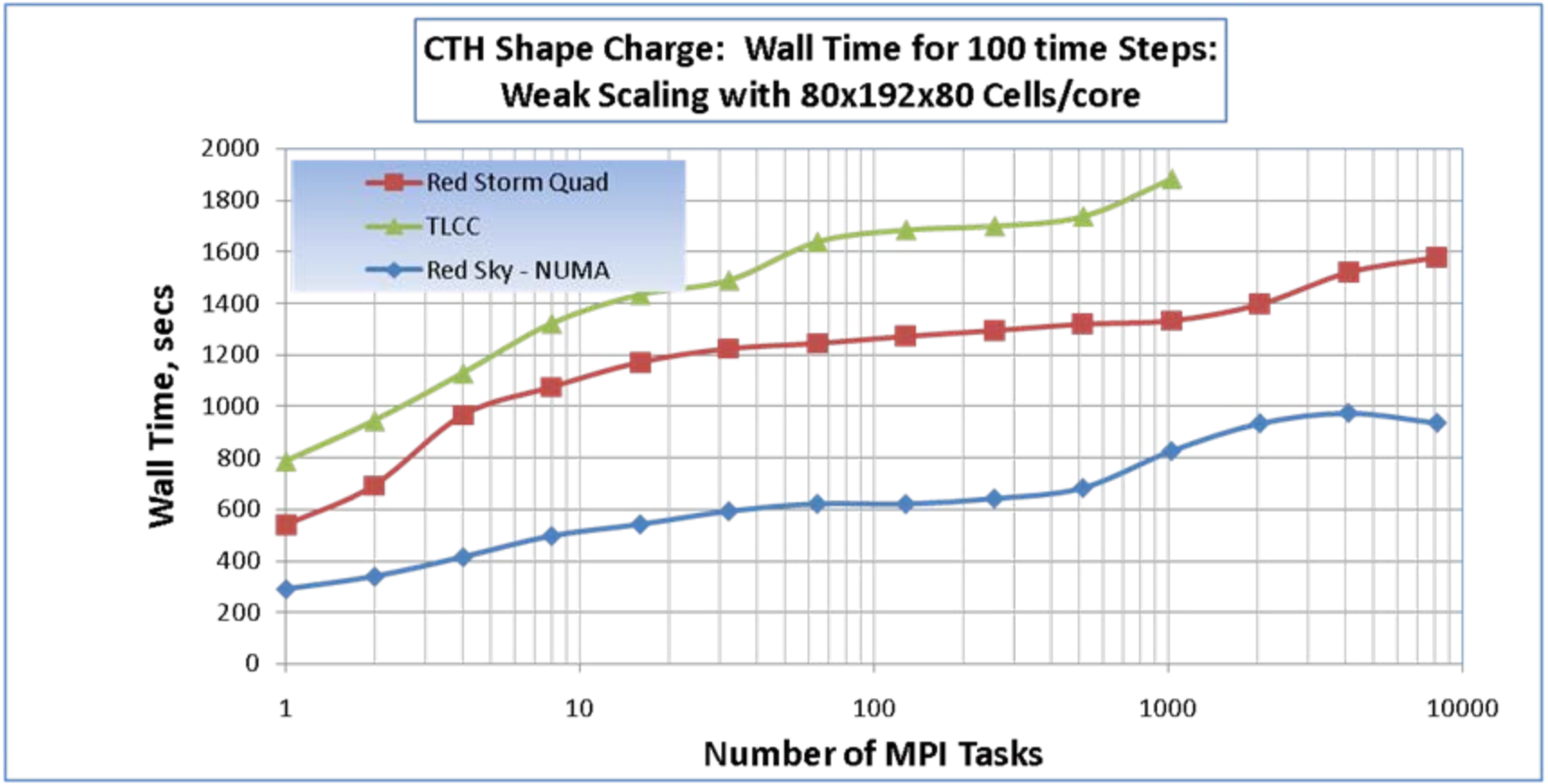
Sandia National Laboratories

2x-3x faster than Red Storm over the midrange



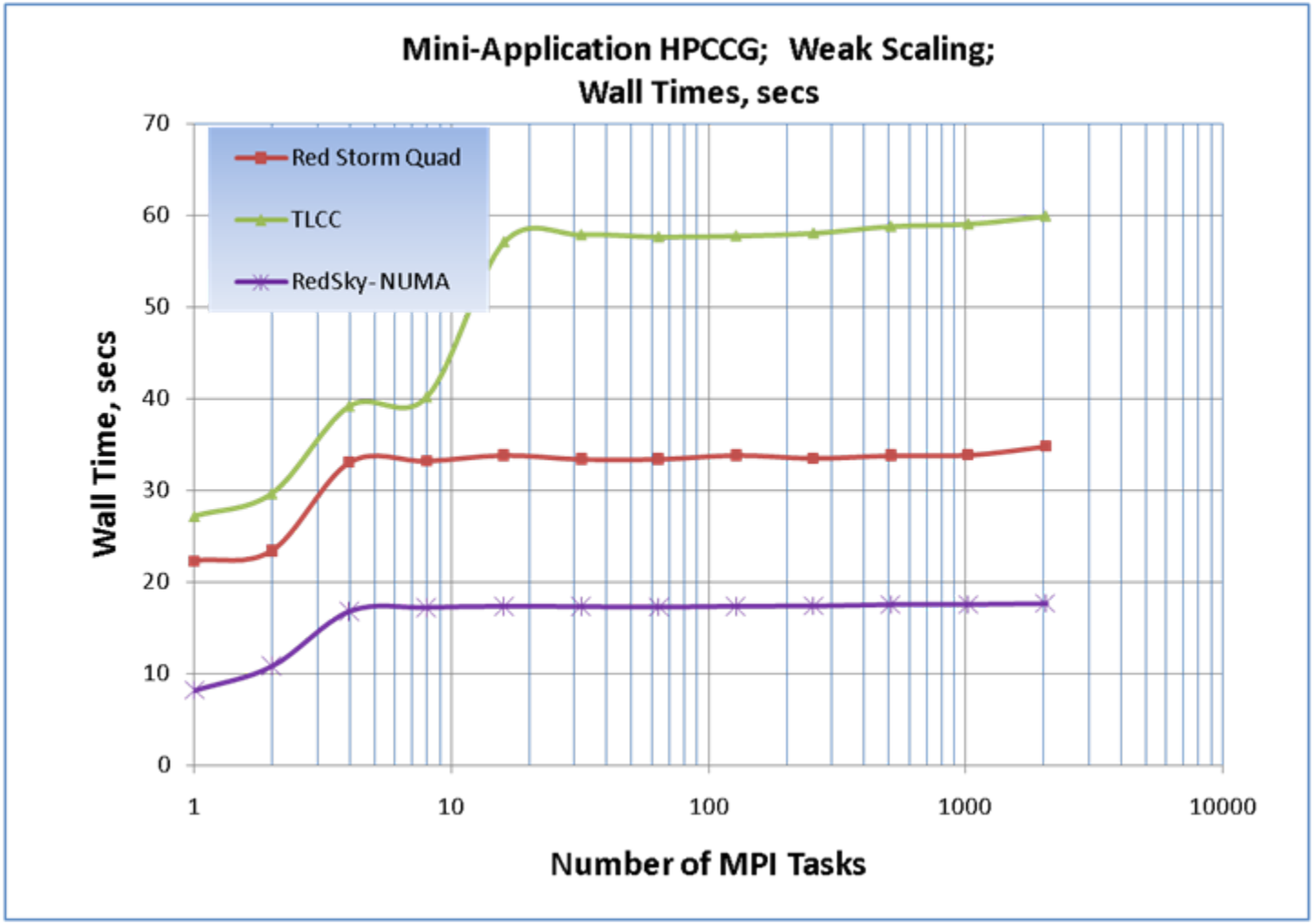
Sandia National Laboratories

CTH



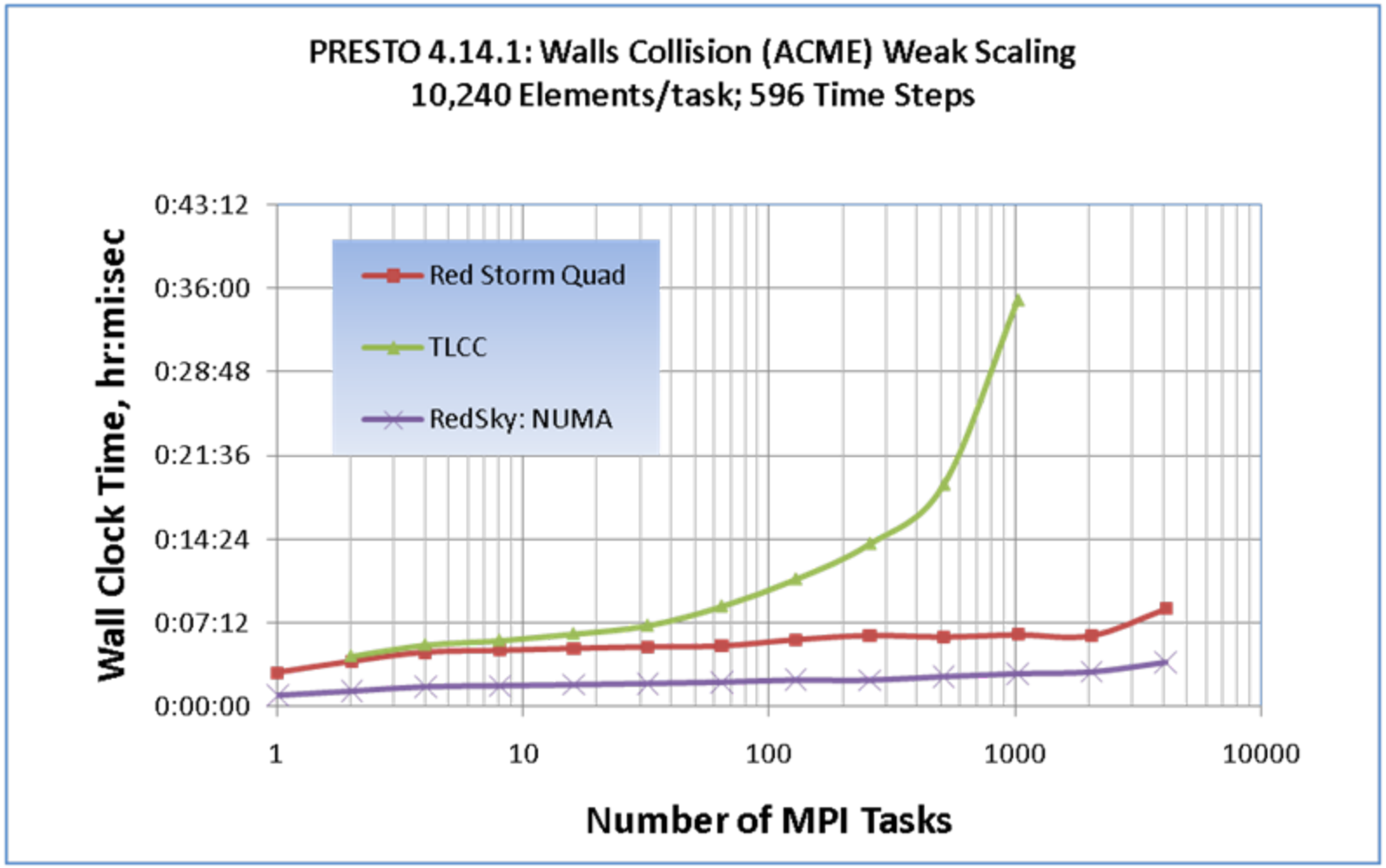
Sandia National Laboratories

HPCCG



Sandia National Laboratories

Presto



Sandia National Laboratories

Summary

- RedSky is pushing IB in many dimensions
- Torus routing in IB was a very difficult effort
- We are making great progress
- We have demonstrated robust operation at scale
- We are looking for others to join us

Sandia National Laboratories

User comments about Red Sky

- “Red Sky is an extremely fast machine that is robust for jobs in the range of four thousand cores and smaller. It performs especially well for job sizes in the range of 1024 cores and smaller and provides great turnaround time. Although issues are still being worked out, it will be a wonderful capacity computational platform for SNL code developers and analysts. Red Sky is also backed by a superb and very helpful support team.”
- “Red Sky is performing very well. I had no trouble getting started. I got my first job running in less than 10 minutes after porting input/restart files. The machine has been extremely stable, jobs have been running for a number of days with no system interruptions. Overall experience has been great.”
- “Based on my runs thus far, Red Sky will be a very attractive platform for Sandia.”
- “Red Sky is great, very fast machine. The transition to Red Sky was easy.”
- “My Red Sky jobs are running twice as fast as Tbird. The transition was painless.”
- “Machine is excellent so far. The jobs run to completion with no problems. Jobs are running more than twice as fast as Tbird on the same number of processors. I ported my application to Red Sky without any problems.”
- “My experience on Red Sky set the record for me as far as minimum time required from originally getting on a machine to being able to run.”