iWARP and Adaptive Routing in a 10G Ethernet Cluster

John Naegle

Jim Brandt, Helen Chen, Cathy Houf, Don Rudish, Jim Schutt

Open Fabrics Sonoma Workshop

May 5 - May 8, 2008

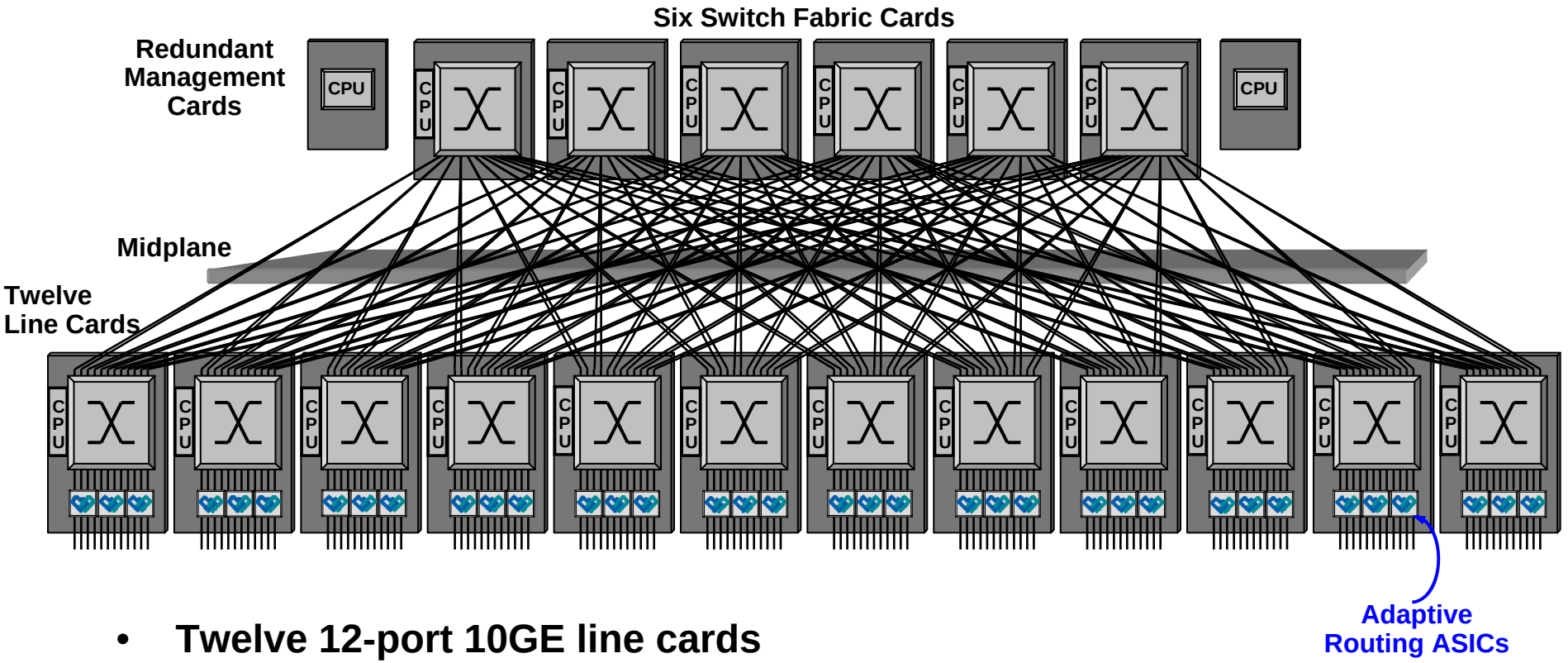
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

The Problems

- Static routing is limiting the performance of interconnected switches
  - Most high port density systems use multiple connections between small radix switches
  - IB uses strict static routing
  - Ethernet uses static Hashes for LAG
  - Link oversubscription reduces performance
- Processing the TCP stack still consumes significant system resources
  - CPU, memory bandwidth, etc.
  - Particular problem for “gateway” nodes

Adaptive Routing Ethernet Switch

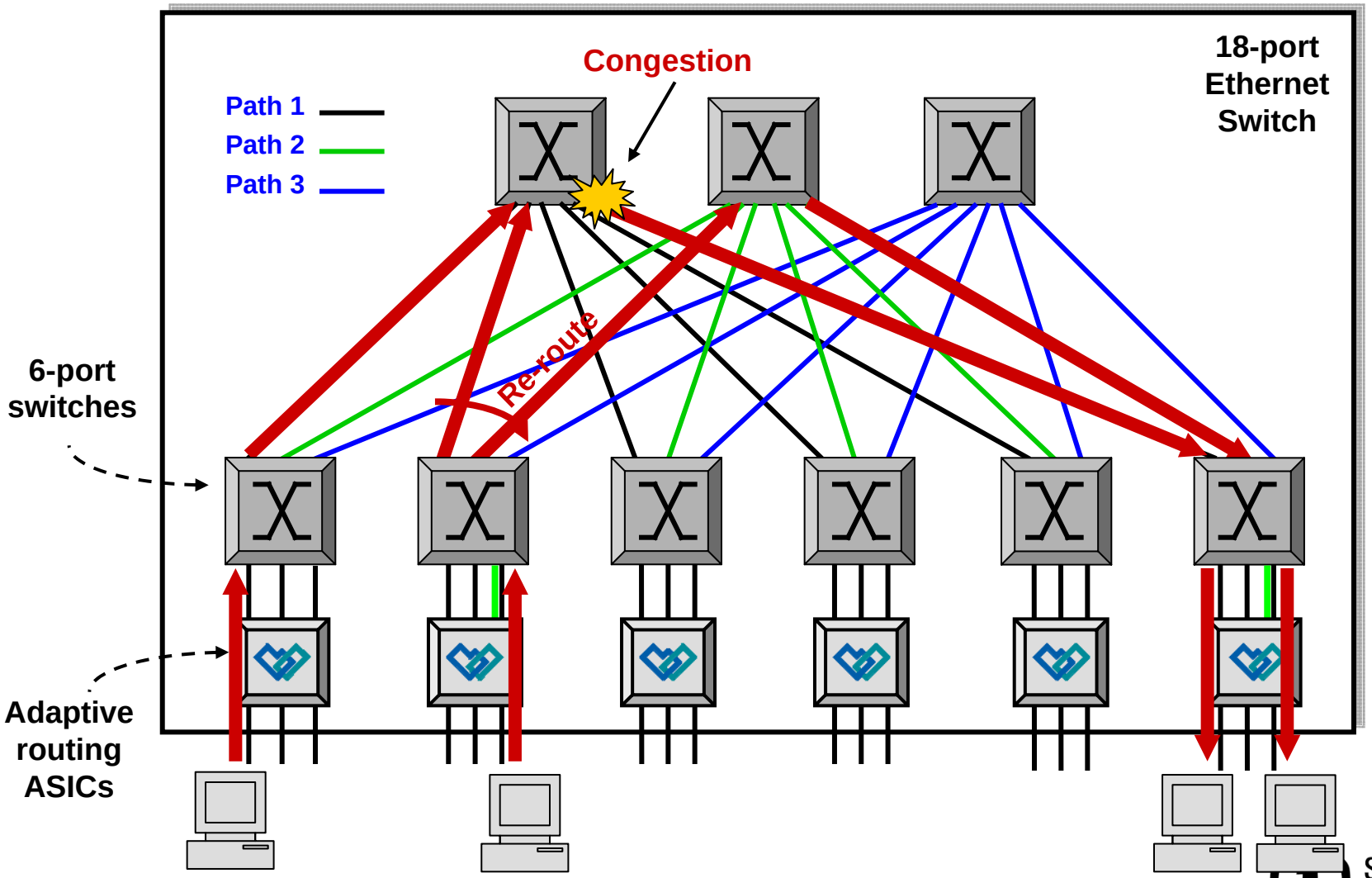


- **Twelve 12-port 10GE line cards**
- **Fat-tree architecture**
- **ASICs at edge of switch perform adaptive routing: detects congestion and reroutes traffic around hot spots**

NNSA National Nuclear Security Administration

Sandia National Laboratories

Ethernet Fabric using Multiple Paths and Adaptive Routing to Avoid Hotspots



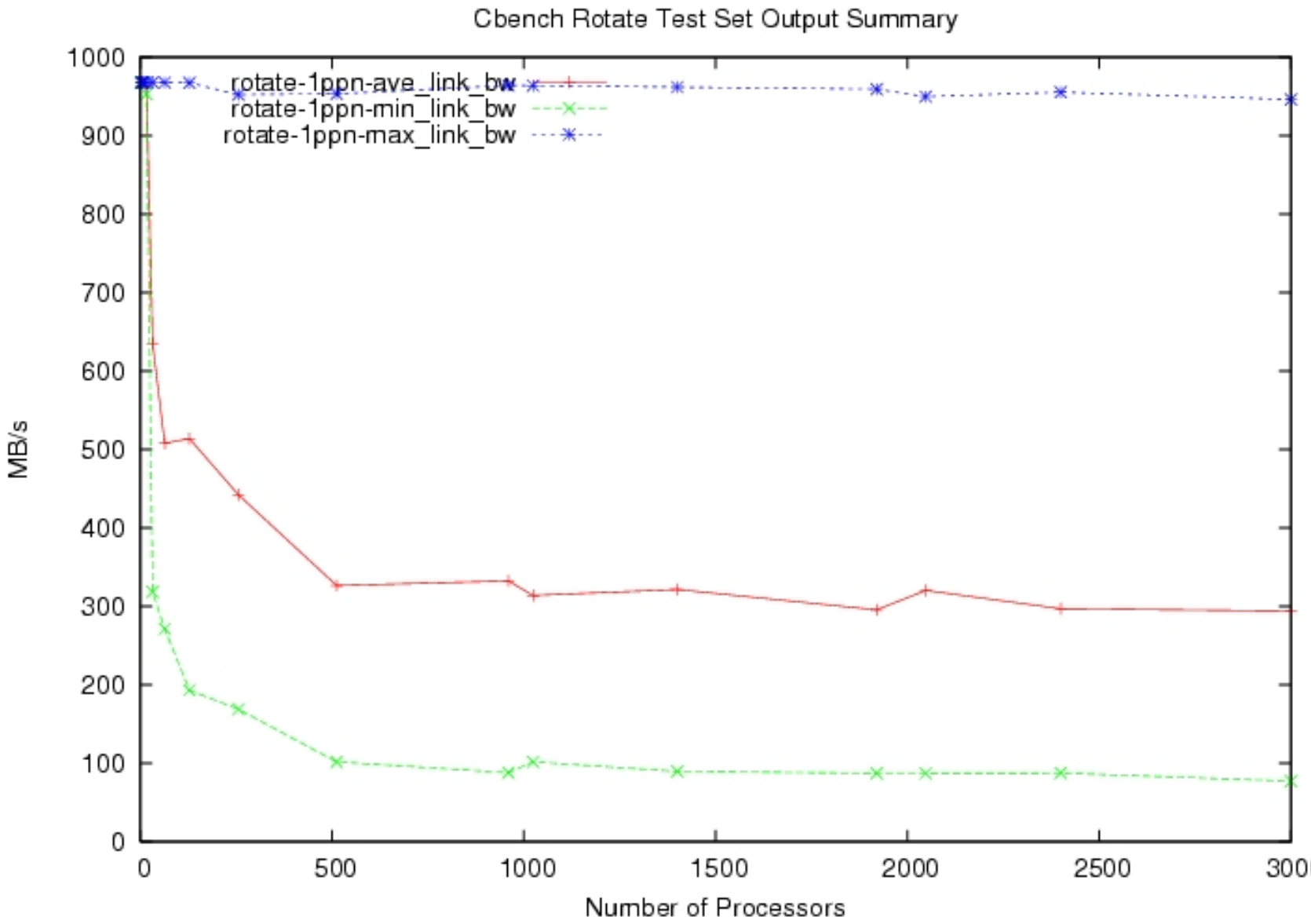
NNSA National Nuclear Security Administration

WovenSystems

Sandia National Laboratories

Per Pair Bisectional Bandwidth

~4000 Node IB Cluster

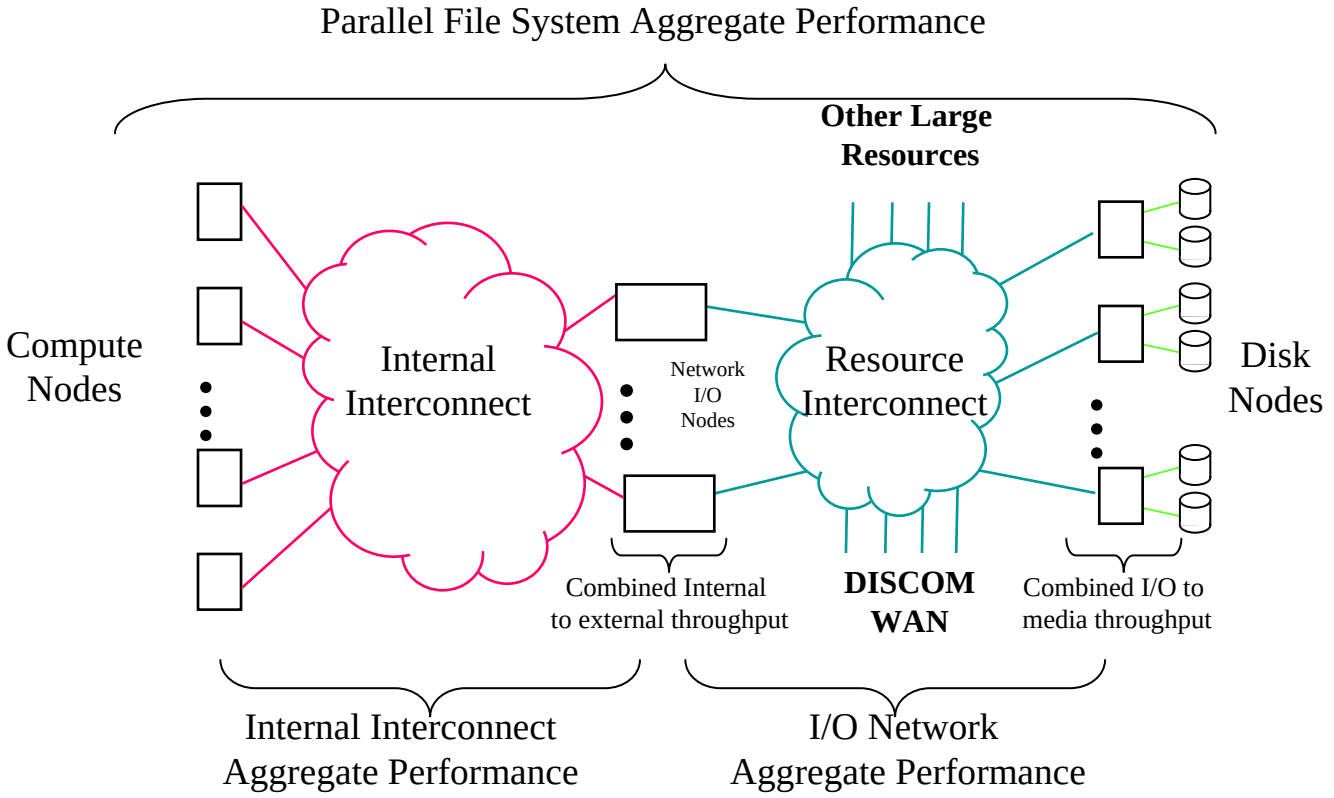


Sandia National Laboratories

Where Do We See The Problems?

- Computational Clusters
  - Synchronous data flows limited by slowest link
- Supercomputer to parallel File Systems
  - Sustained data flows to/from disk also limited by slowest link
  - PetaScale File Systems pushing 2000 ports
- Large Server Farms

Generic Global Parallel File System Architecture



What Are We Doing?

- Investigating and supporting dynamic routing implementations
- Investigating and supporting OpenFabrics iWarp RDMA scaled implementations
- Formed a collaboration to demonstrate one particularly promising environment
  - Woven active congestion management
  - Chelsio 10GE RNICs using OpenFabrics stack
  - Sandia 128 node cluster (Talon)

Goals of the Collaboration

- Demonstrate scalability of a high-density 10GbE switching infrastructure
- Demonstrate effectiveness of dynamic routing over static routing for low radix switch interconnects
- Evaluate Low Latency 10 GbE with RDMA as an alternative for deploying:
  - Common I/O infrastructure between PetaScale resources (compute, vis, disk, tape, etc)
  - Cluster interconnect
- Utilize simulation and analysis to project results to larger scales

Sandia CBench Suite

- Sandia's CBench Suite includes industry standards:
  - HPCC, Intel MPI Benchmarks, OSU, NAS, etc.
- Also includes benchmarks developed to stress bi-sectional bandwidth and latency
  - "Rotate Bandwidth" pair-wise transmits 80MB of data from half of the nodes to the other half. Repeat that test for many different bi-sections and report Min, Average, and Max individual throughput
  - "Rotate Latency" performs similar strategy as Rotate Bandwidth but tests simultaneous small packet latency instead of throughput

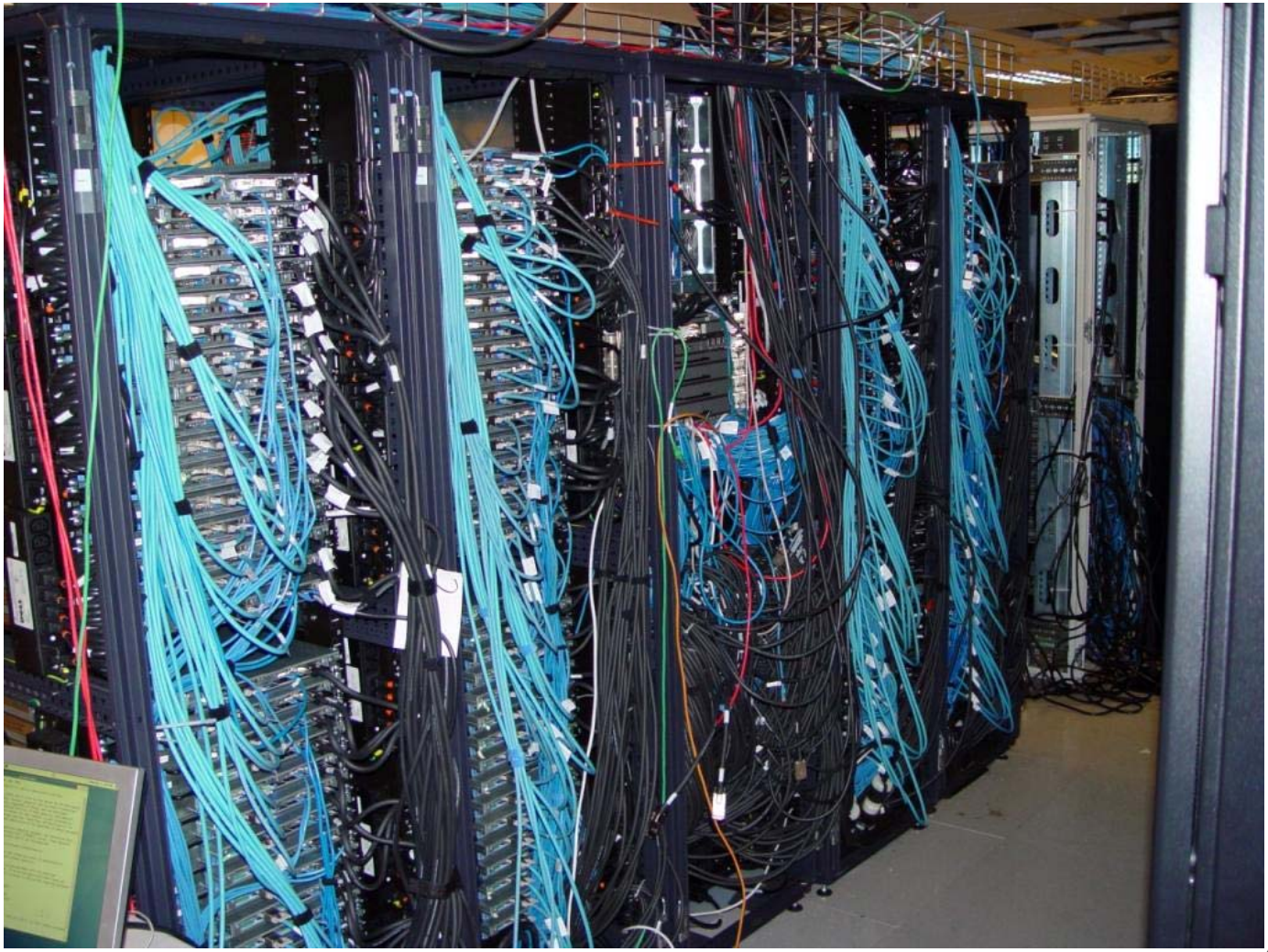
Testbed Configuration and Startup

- Testbed Details
  - 128 Dell 1850 Nodes running 2.6.9-55.0.9 Linux kernel
  - Chelsio T3 RNICs using the 1.0.109 driver + patches
  - OFED 1.2.5 with included MVAPICH2
  - 1 Topspin SDR IB switch with SDR HCA per host
  - 1 Woven 144 port EFX-1000 switch
- Many issues to work through
  - Significant manpower to maintain cluster
  - Bugs in new implementations of switch/NICs
  - Bugs in scaling OpenFabrics RDMA implementation
  - Many knobs to tweak in switch/NIC tuning
  - HP Linpack and benchmark tuning always time consuming

NNSA National Nuclear Security Administration

Sandia National Laboratories

Talon Cluster



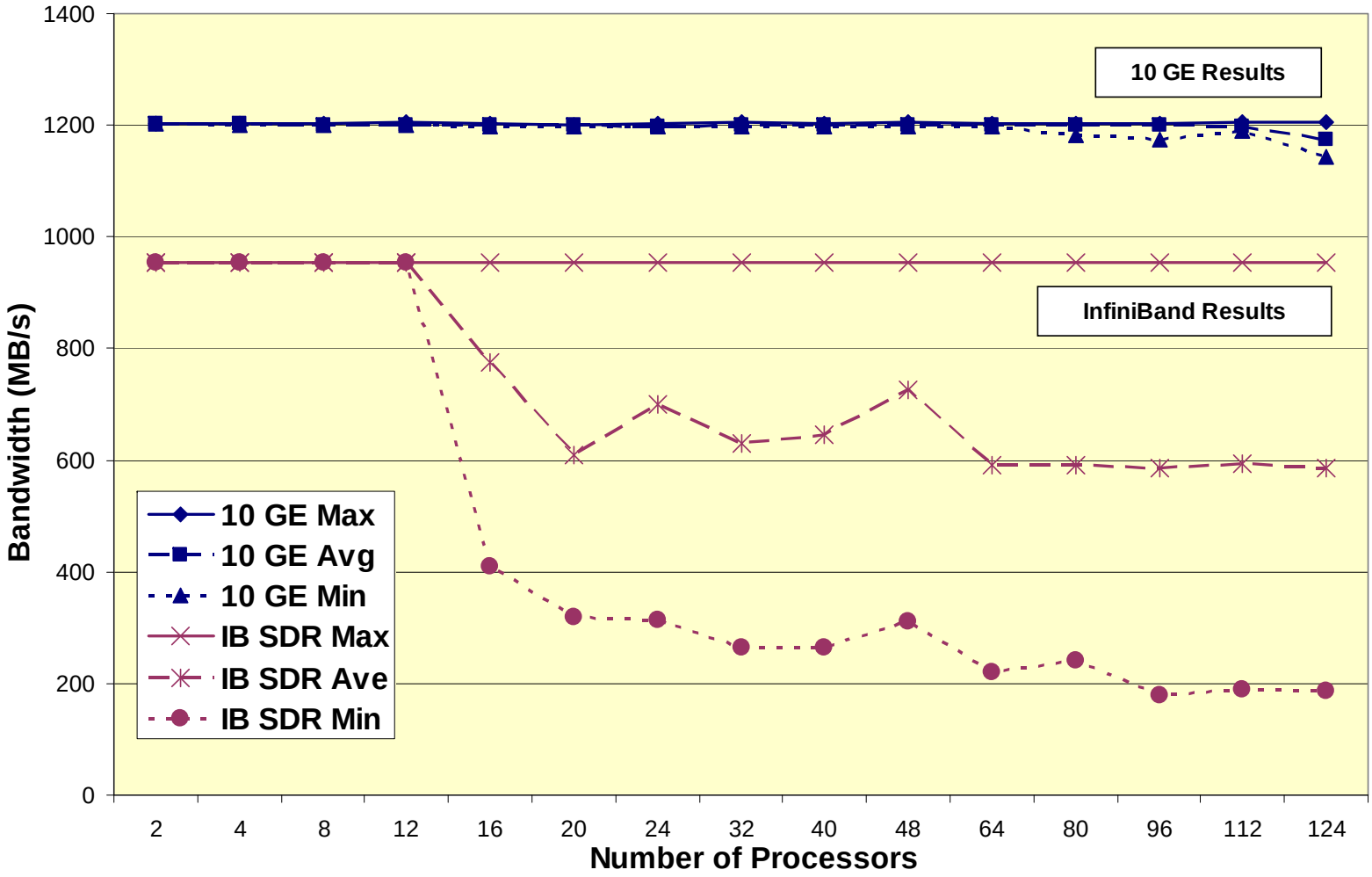
Tuning and Configuration

- Utilized 9000 Byte MTU
- Enabled/disabled adaptive routing functionality
  - Not normally exposed to users
- Compared strict versus relaxed packet ordering
- Enabled/disabled RX/TX Pause frames on edge ports
- Chelsio enhanced driver to tune the hardware stack
  - Traffic scheduling and management; fast error recovery
- Tuned several switch and NIC internal parameters to optimize for interconnect performance
  - Many were incorporated in to vendors' default configurations, contact vendors for help with performance issues

Significant Issues

- Port Sharing between Kernel and offload stacks
  - TCP ports used by offload stack are not known by the kernel stack and vice versa, could cause application failure
  - Proper fix requires overcoming resistance from kernel stack maintainers
  - Used a patch by Steve Wise to address this issue
- ARP scaling problem was discovered and fixed
  - Connections weren't being accepted for stale ARP entries
- Completion Queue
  - Mvapich2 would hang due to completion queue overrun
  - Used MV2_DEFAULT_MAX_CQ_SIZE=6000 for benchmarks
  - Must overcome contiguous memory limitation to scale to larger clusters

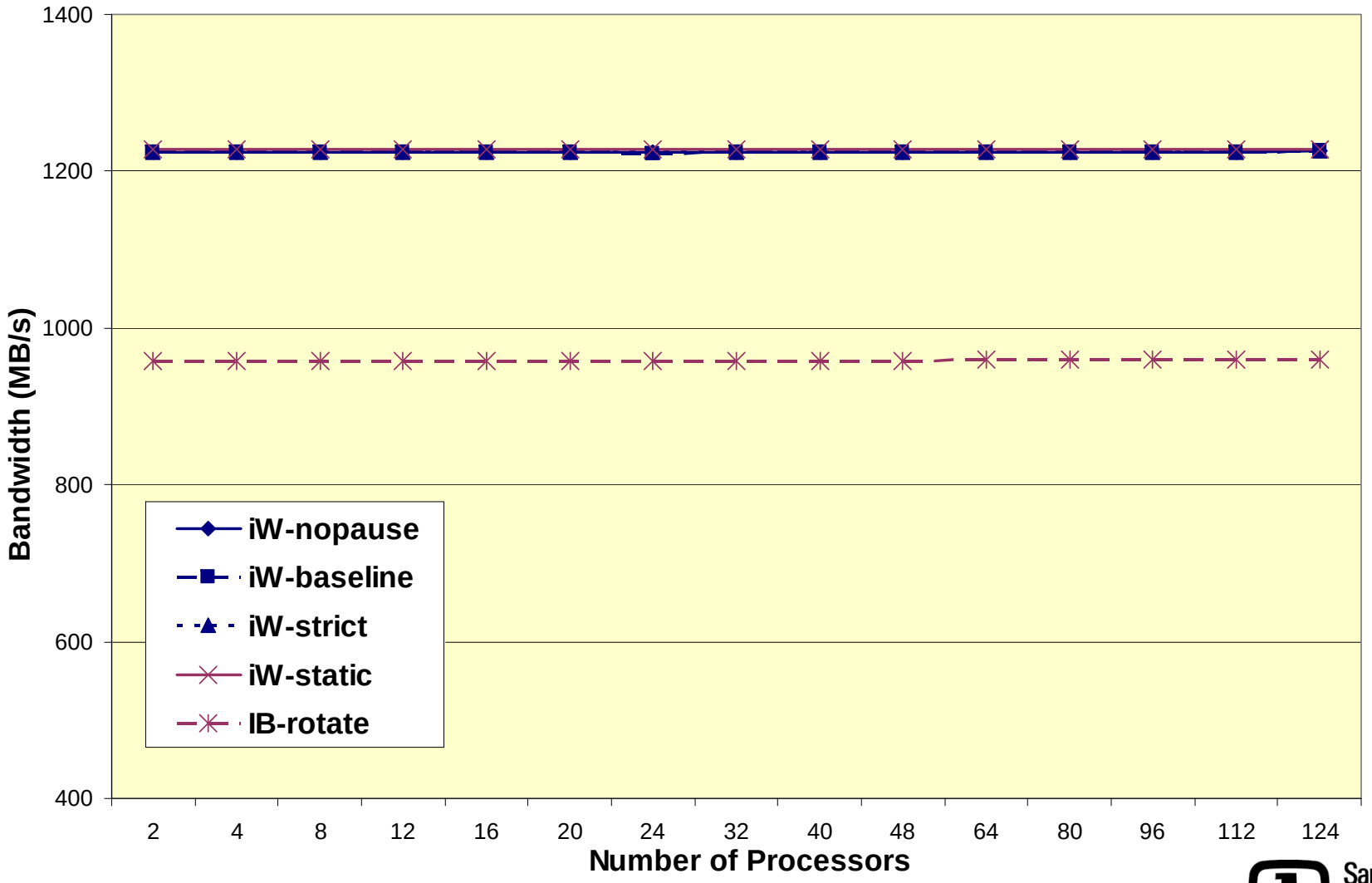
Cbench Rotate Benchmark Test (Relaxed Ordering)



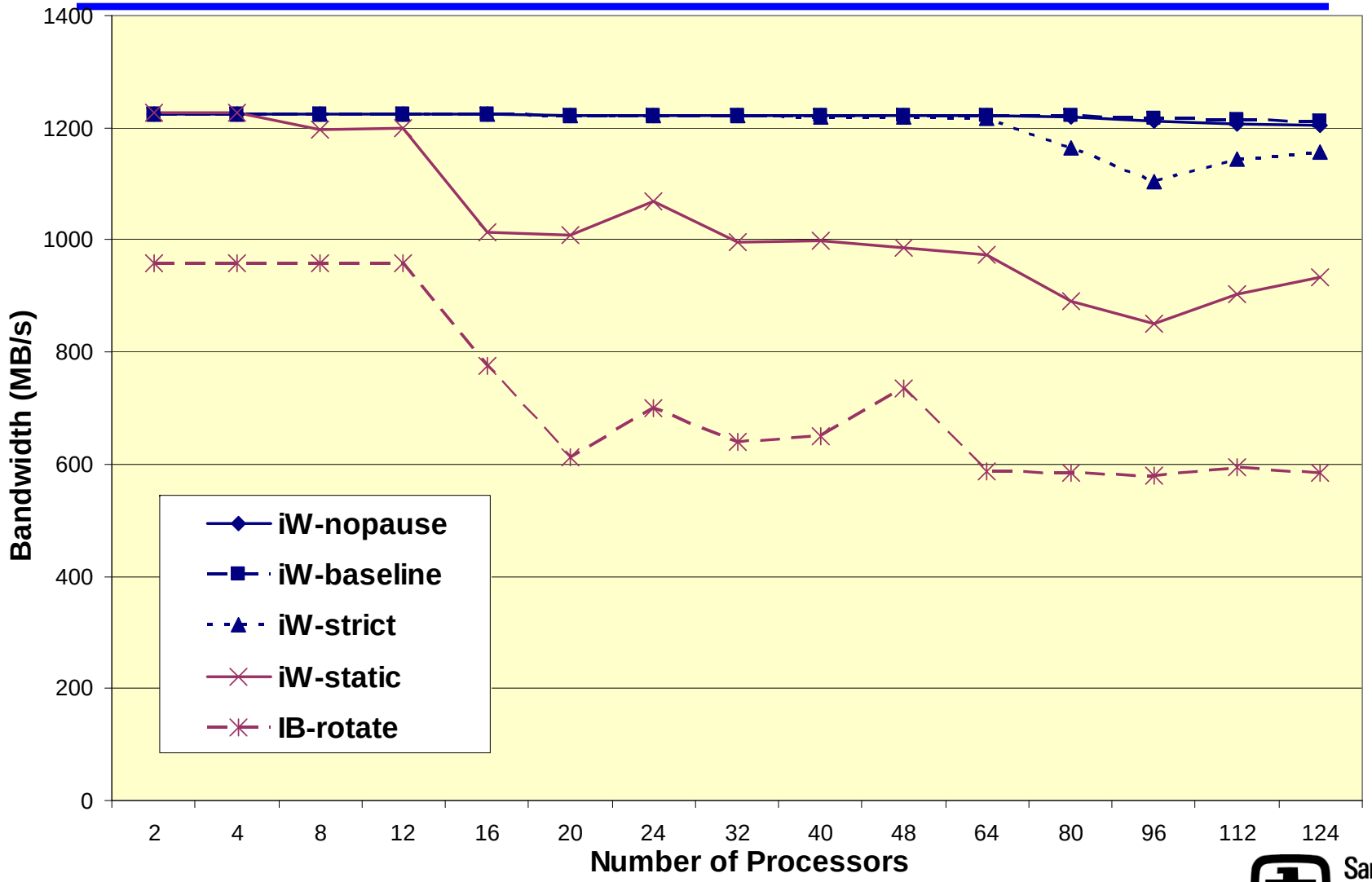
NNSA National Nuclear Security Administration

Sandia National Laboratories

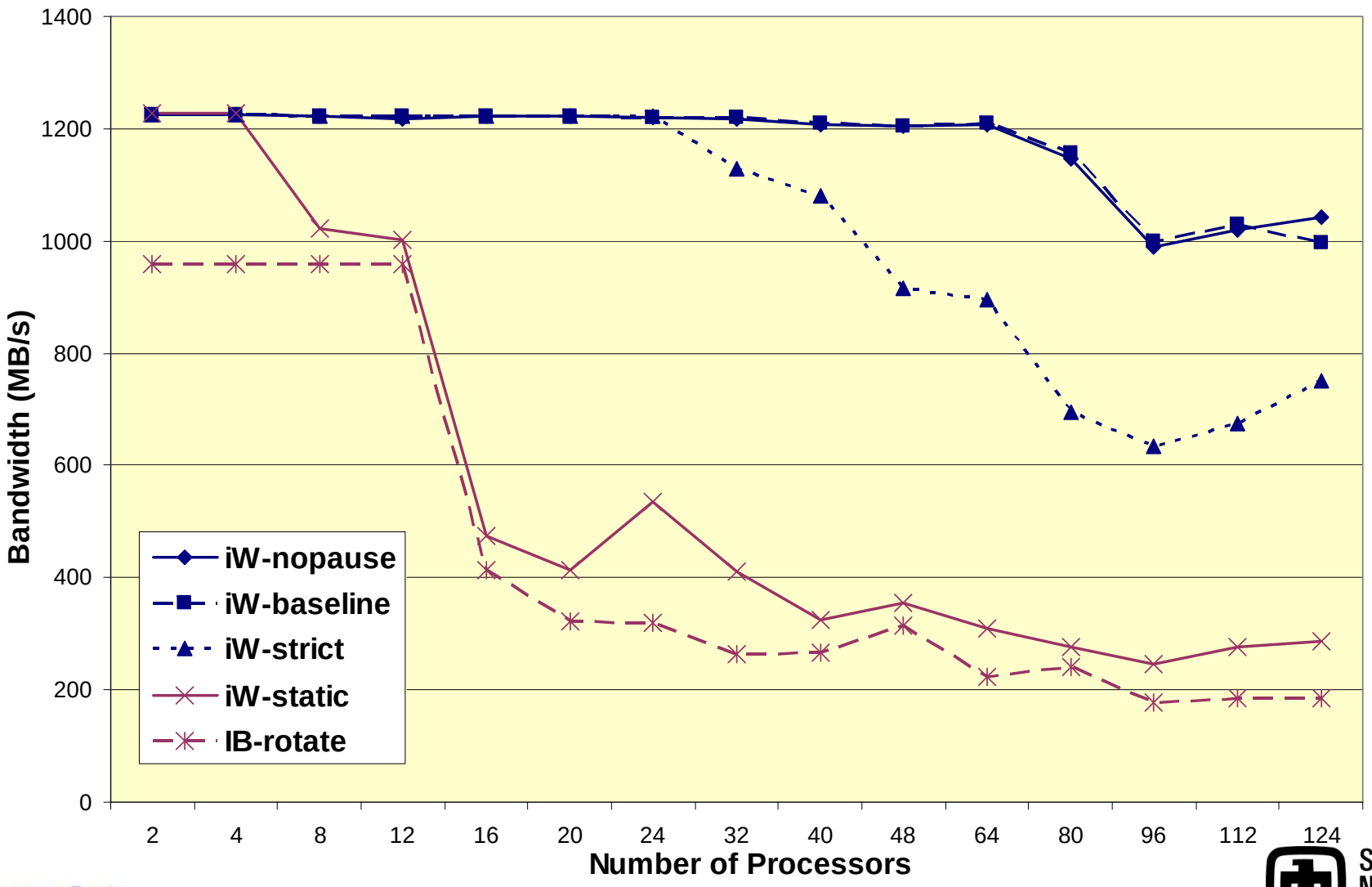
Maximum Bandwidth Rotate Test



Average Bandwidth Rotate Test



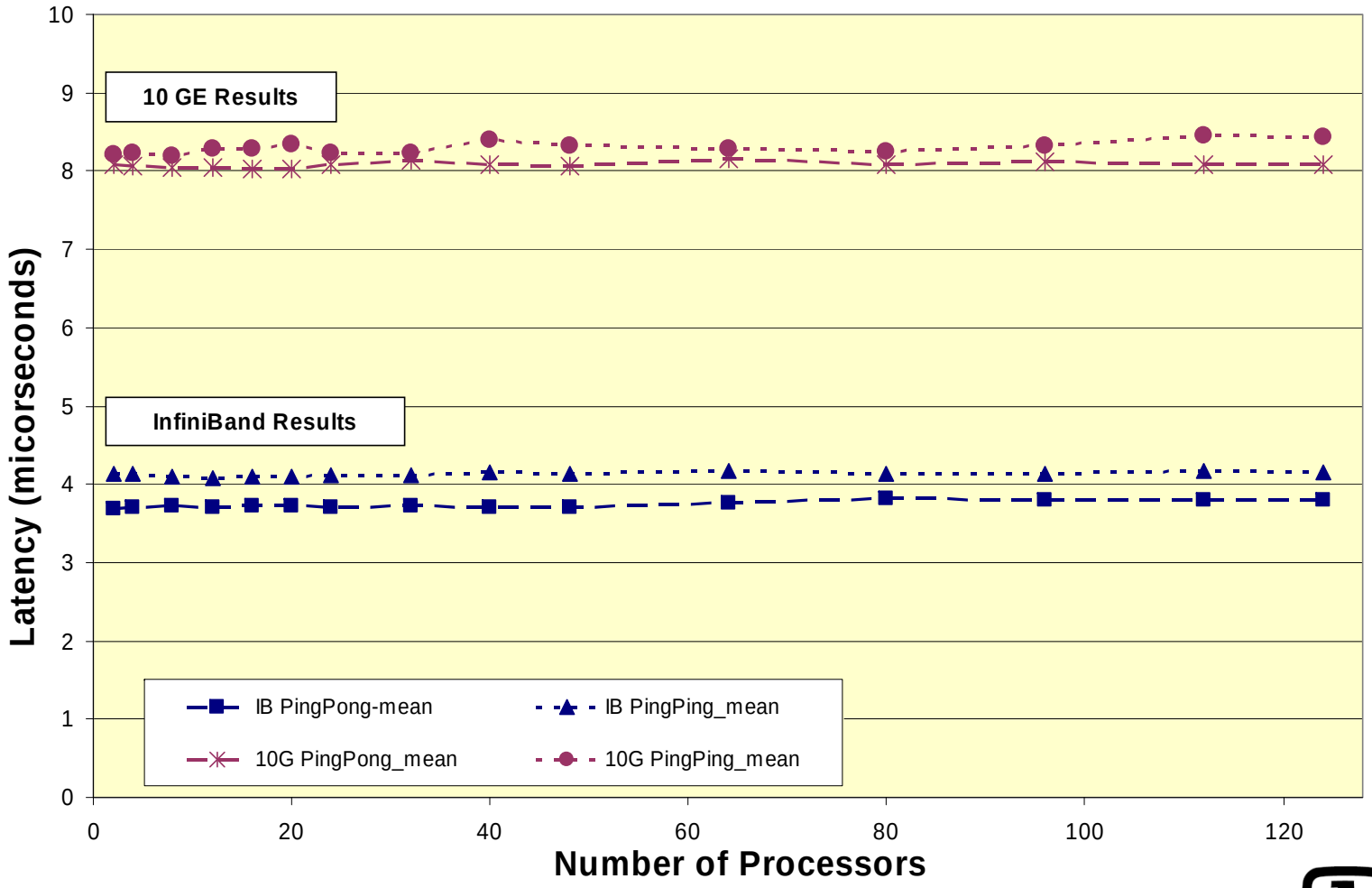
Minimum Bandwidth Rotate Test



NNSA National Nuclear Security Administration

Sandia National Laboratories

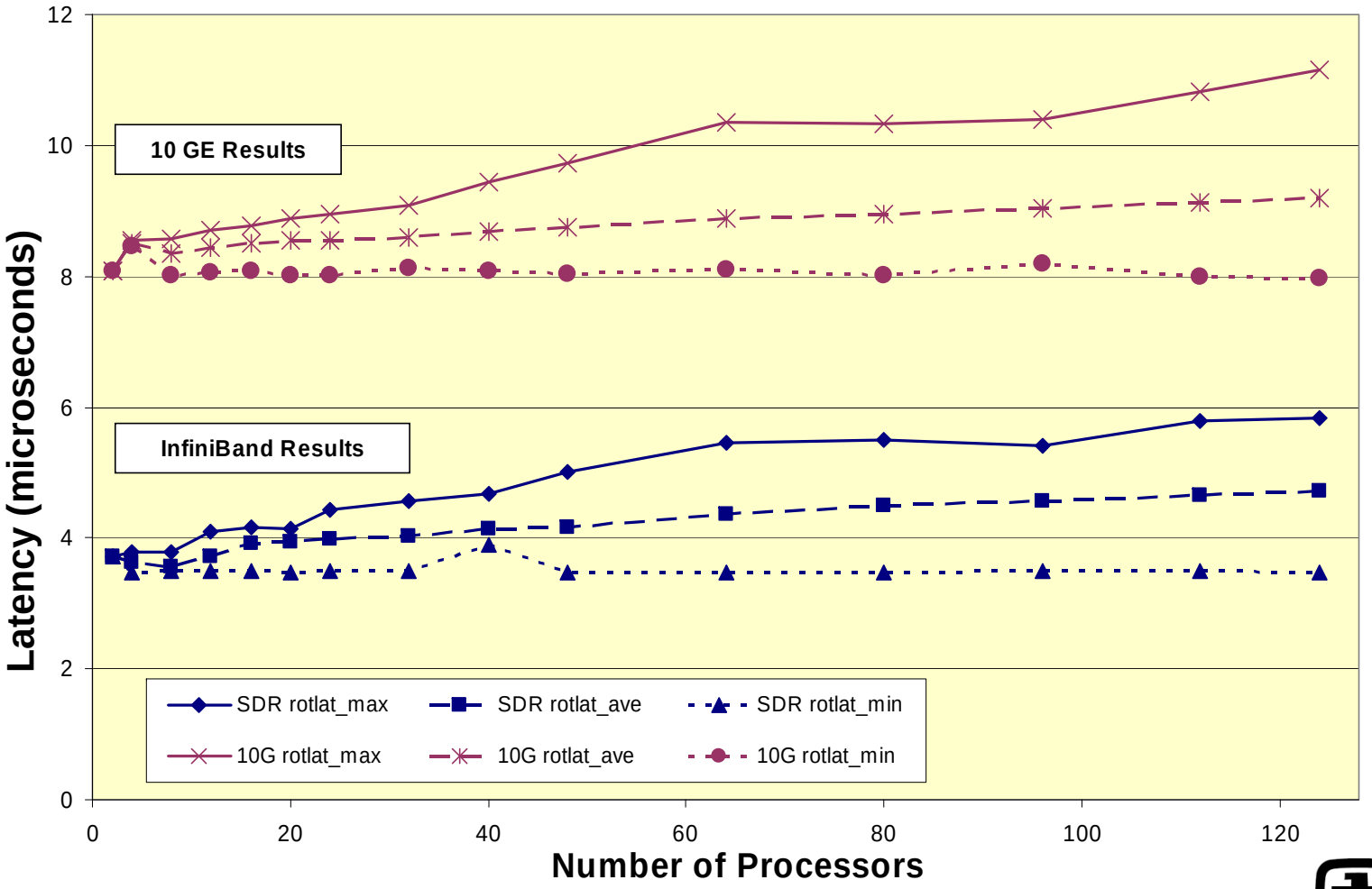
Intel MPI Latency Benchmark Test



NNSA National Nuclear Security Administration

Sandia National Laboratories

Cbench Rotate Latency Benchmark Test



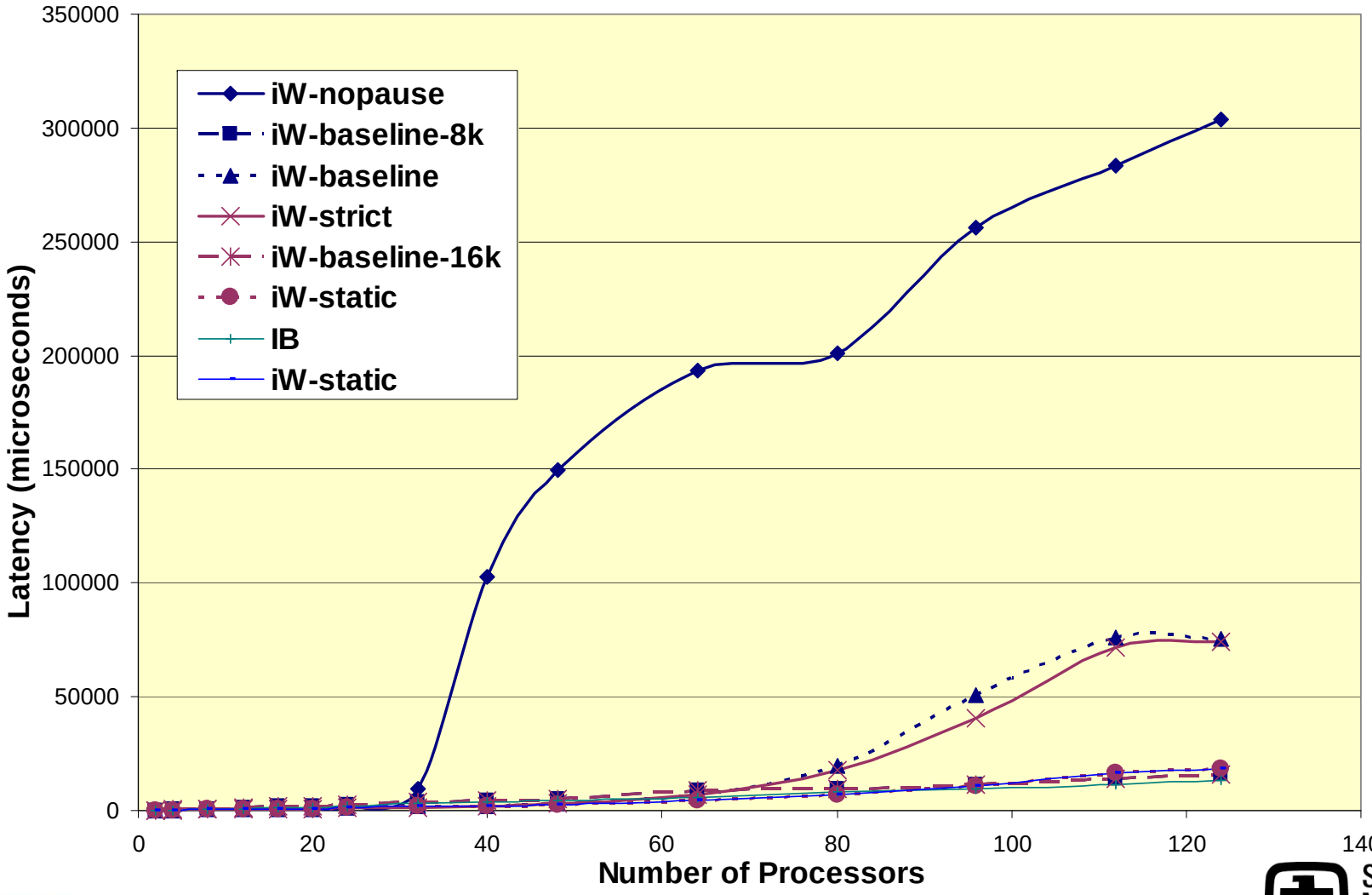
NNSA National Nuclear Security Administration

Sandia National Laboratories

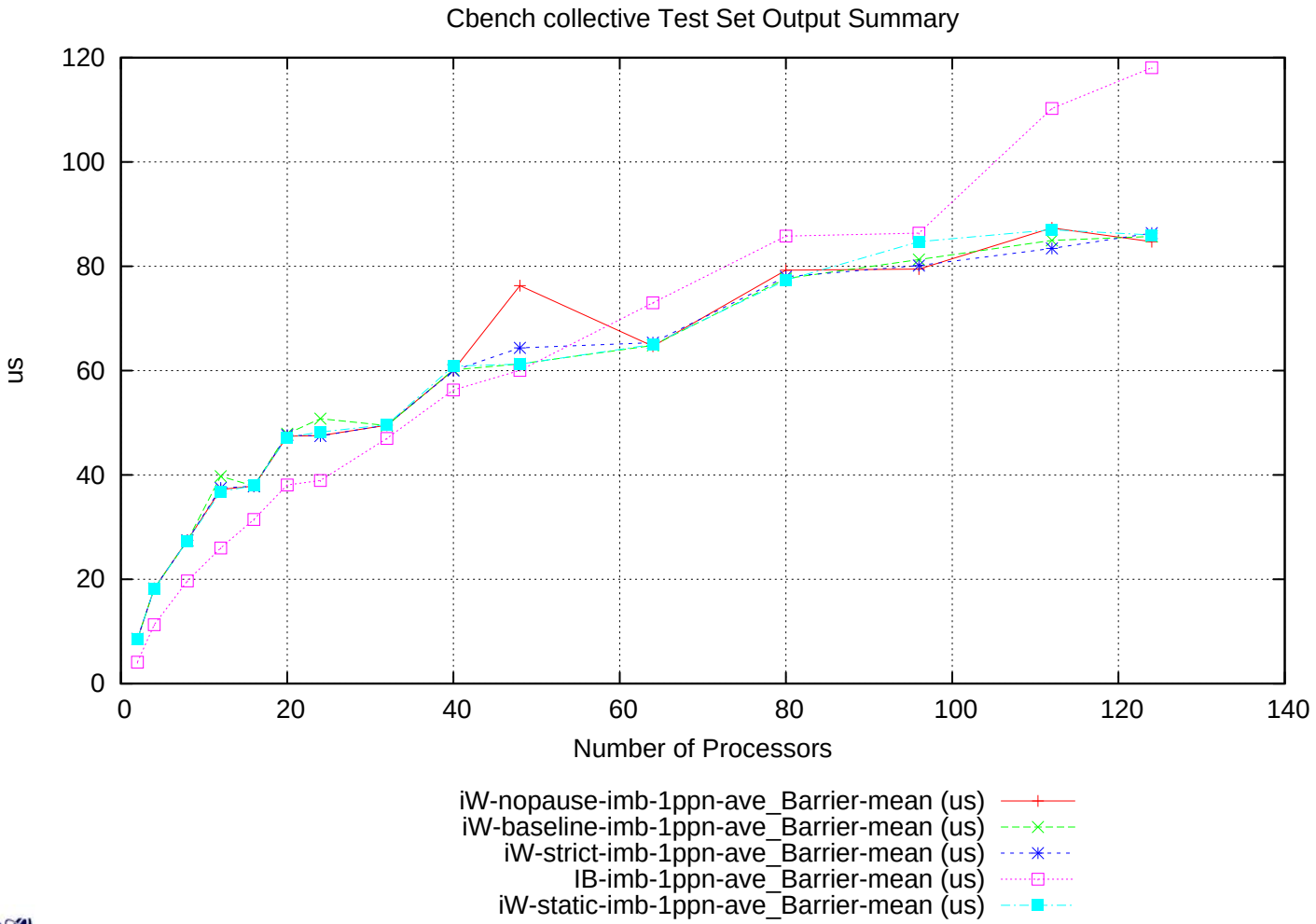
Switch, NIC, and Application Tuning

- Switch required significant tuning
  - Buffer management
  - Rerouting timing
  - Pause functionality
- NIC tuning
  - Tuned to provide good performance across all of the applications we investigated
- Most tuning is part of standard deployment now
- Refer to vendors for more details
- MPI tuning avoided some pathological cases

All-to-All 32KB Message Latency Results



Barrier Collective Latency Result



NNSA National Nuclear Security Administration

Sandia National Laboratories

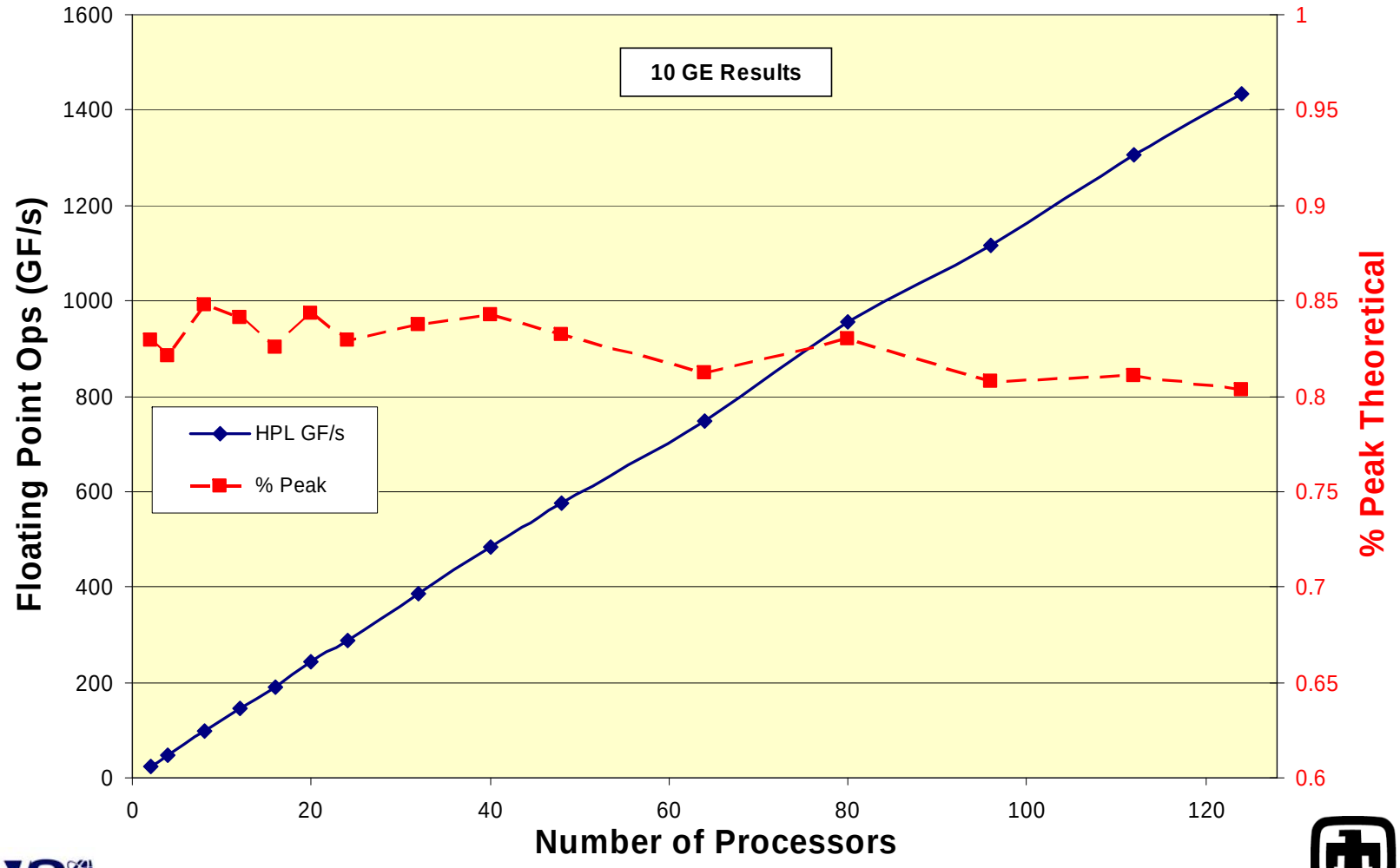
Congestion Management Issues

- Long-term contention: multiple flows contending for single port
  - Time scale much greater than switch latency
  - Example: flows for parallel file system, w/more clients than servers
  - Using pause results in head-of-line blocking
  - Must reduce per-flow offered load – need congestion indication
    - Dropped packets can be recovered via TCP fast retransmit
    - Explicit congestion notification (ECN) might help

- Short-term congestion: synchronized, bursty traffic
  - Time scale similar to switch latency
  - Example: flows for all-to-all algorithms where all send to all
  - Short messages mean dropped packets recovered via TCP retransmit timeout
  - Use pause to prevent packet loss

- Need both pause and per-flow congestion events to handle all traffic profiles
- Current pause implementations problematic
- IEEE 802.1au congestion notification could help as well

HP Linpack Benchmark Test

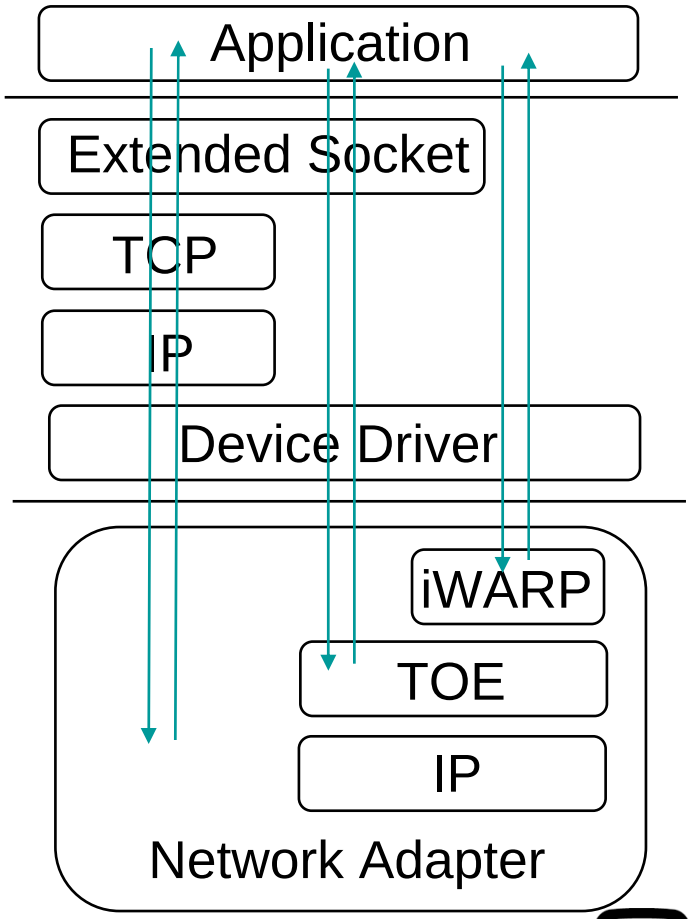


NNSA National Nuclear Security Administration

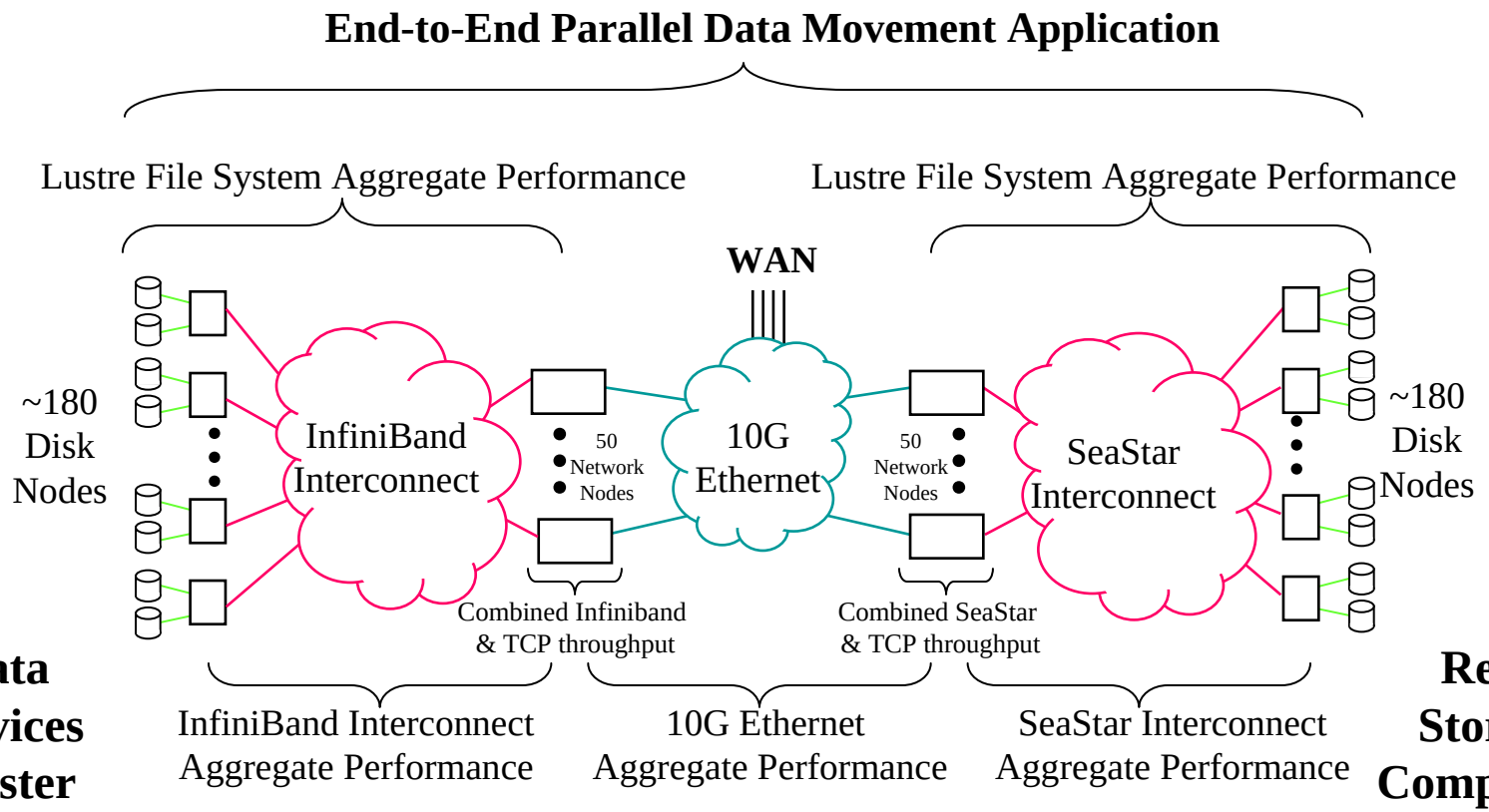
Sandia National Laboratories

iWARP - RDMA protocol for TCP/IP

- iWARP is the suite of RDMA protocols for TCP/IP
- RNIC is a RDMA capable NIC with offloaded iWARP as well as TCP/IP (TOE)
- RNIC typically exposes NIC, TOE and iWARP interfaces to upper layer applications



Example Red Storm Architecture



NNSA National Nuclear Security Administration

Sandia National Laboratories

Example of Inefficiency Impact in Gateway Nodes

- 32 10GE paths from RedStorm to viz. Cluster
  - Planned for 50% efficiency: 16GBytes/s
  - Network sustained full 16GB/s using real HPSS application from memory to memory
    - ~95% CPU utilization, ~500MB/s per path
  - When using real disk, performance dropped to ~270MB/s with 100% CPU utilization
  - With sendfile (no user memory copy): ~340MB/s
  - **With TOE: >600MB/s with ~15% CPU!**

RDMA Results

- Interaction with Linux Kernel contentious
  - TCP stack in hardware similar to TOE
  - Some interaction with Kernel (ARP, IP port sharing, etc.) still required
- OpenFabrics concept working well
  - Same CBench executables run on IB or 10GE
  - Some IB capabilities not implemented for 10GE
- Yet to run detailed CPU comparison tests
  - Preliminary indications looking good

Major Results

- Bi-sectional bandwidth scaling looks excellent
- Latency is getting close to SDR IB RDMA
- Linpack is in same efficiency range as IB
- Switch strict-order delivery impacts bandwidth
  - Relaxed ordering working well
- Both Pause and per flow congestion management important for optimum performance
- Tuning of switch, NIC and application required to achieve maximum performance
- Issues that appear only at scale are difficult to debug

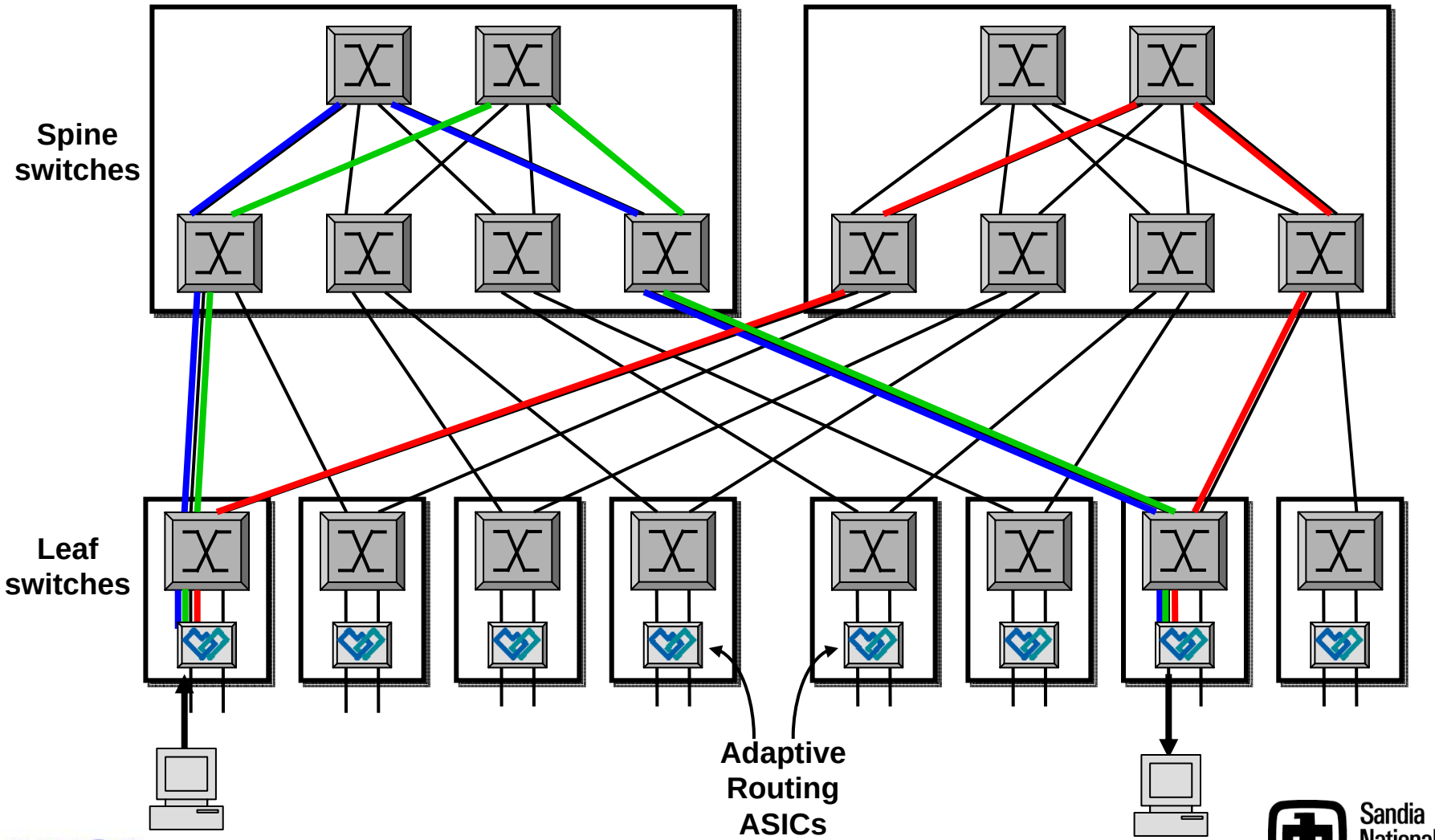
Summary

- Dynamic routing significantly improves the measurable bi-sectional bandwidth
- RDMA over 10G Ethernet seems to be very efficient and effective for a cluster interconnect
- We need to pay close attention to system scaling issues
  - Build and debug is inefficient and expensive
- The iWarp RDMA over 10G Ethernet was very stable after debugging and tuning completed
  - Ran thousands of batch jobs over several days with no failures

Future Work

- Production interoperability testing of RNICs
  - NetEffect NICs coming on line now
- Multi-tier switch operation
- Other flows (small packet, etc)
- Comparison with DDR
- Comparison with TOE
- Lustre/pNFS etc. testing with RDMA
- Open MPI testing

Adaptive Routing over a Multi-tier Fat Tree Topology



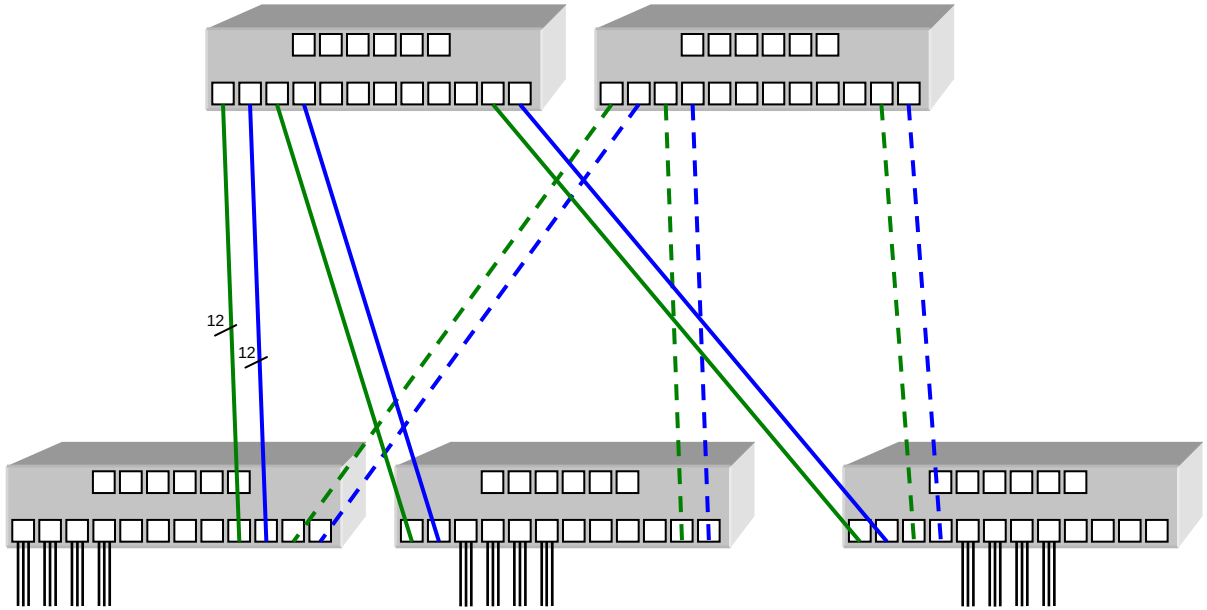
NNSA National Nuclear Security Administration

WovenSystems

Sandia National Laboratories

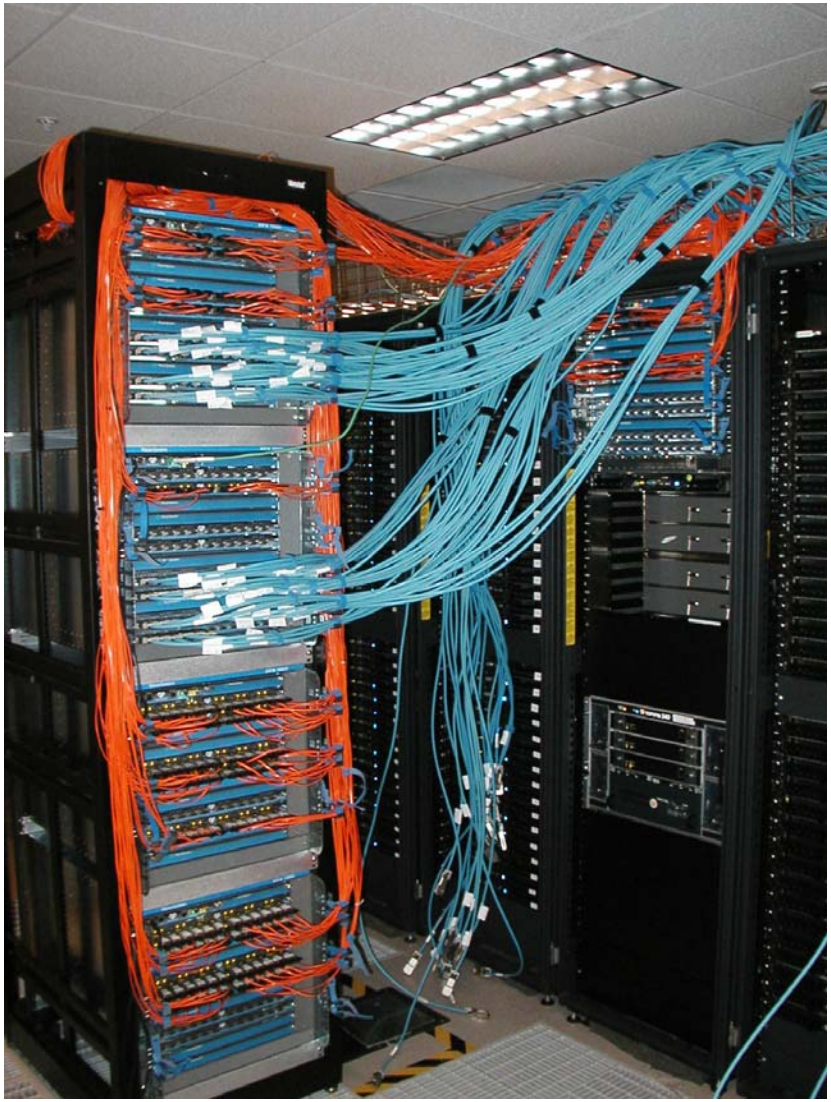
Multi-Chassis Configuration

2 Core Switches



3 Edge Switches

Multi-Chassis Hardware



NNSA National Nuclear Security Administration

Sandia National Laboratories