SAND2007-7270C

An Adaptive Routing Implementation for Efficient Cluster Interconnects

John Naegle, Jim Brandt, Helen Chen,
Josh England, Jim Schutt

SC07 Conference

November 13, 2007

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.

National Nuclear Security Administration

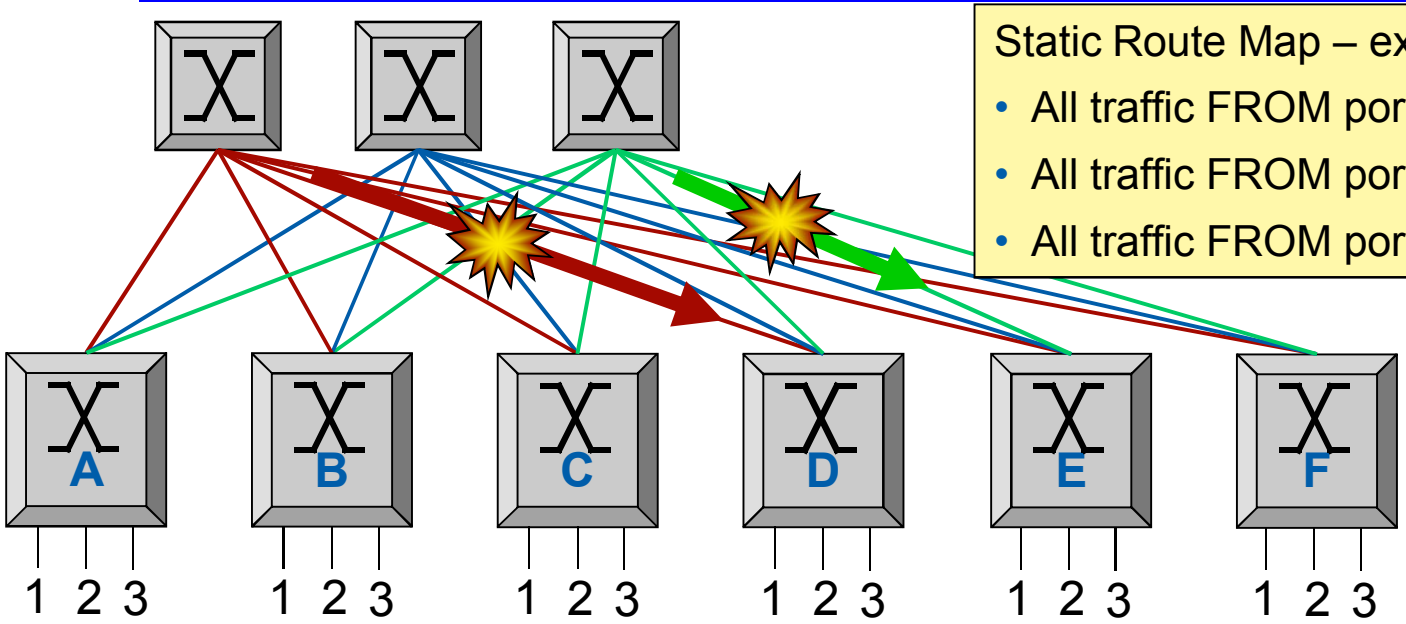
Sandia National Laboratories

The Problem

- 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

Performance Bottlenecks with Statically Routed Protocols

IP ECMP, InfiniBand, Ethernet LAG, etc.



Static Route Map – example:

- All traffic FROM port 1 mapped on **red**
- All traffic FROM port 2 mapped on **blue**
- All traffic FROM port 3 mapped on **green**

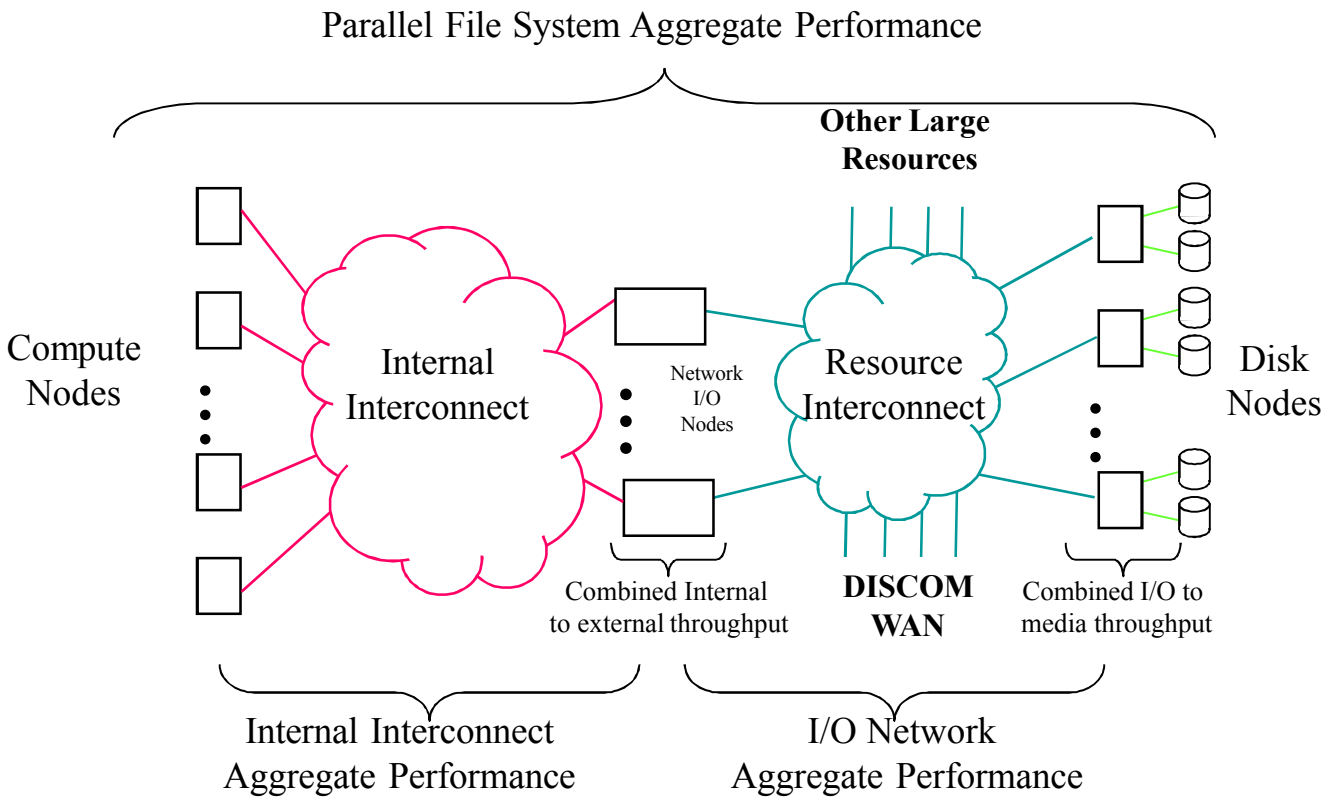
Input	Output	Path	BW State
A1	D1	Red	Full
A2	D2	Blue	Full
A3	D3	Green	Full
B1	E1	Red	Full
B2	E2	Blue	Full
B3	E3	Green	Full

Input	Output	Path	BW State
A1	D1	Red	1/3
B1	D2	Red	1/3
C1	D3	Red	1/3
B2	E1	Blue	Full
B3	E2	Green	1/2
C3	E3	Green	1/2

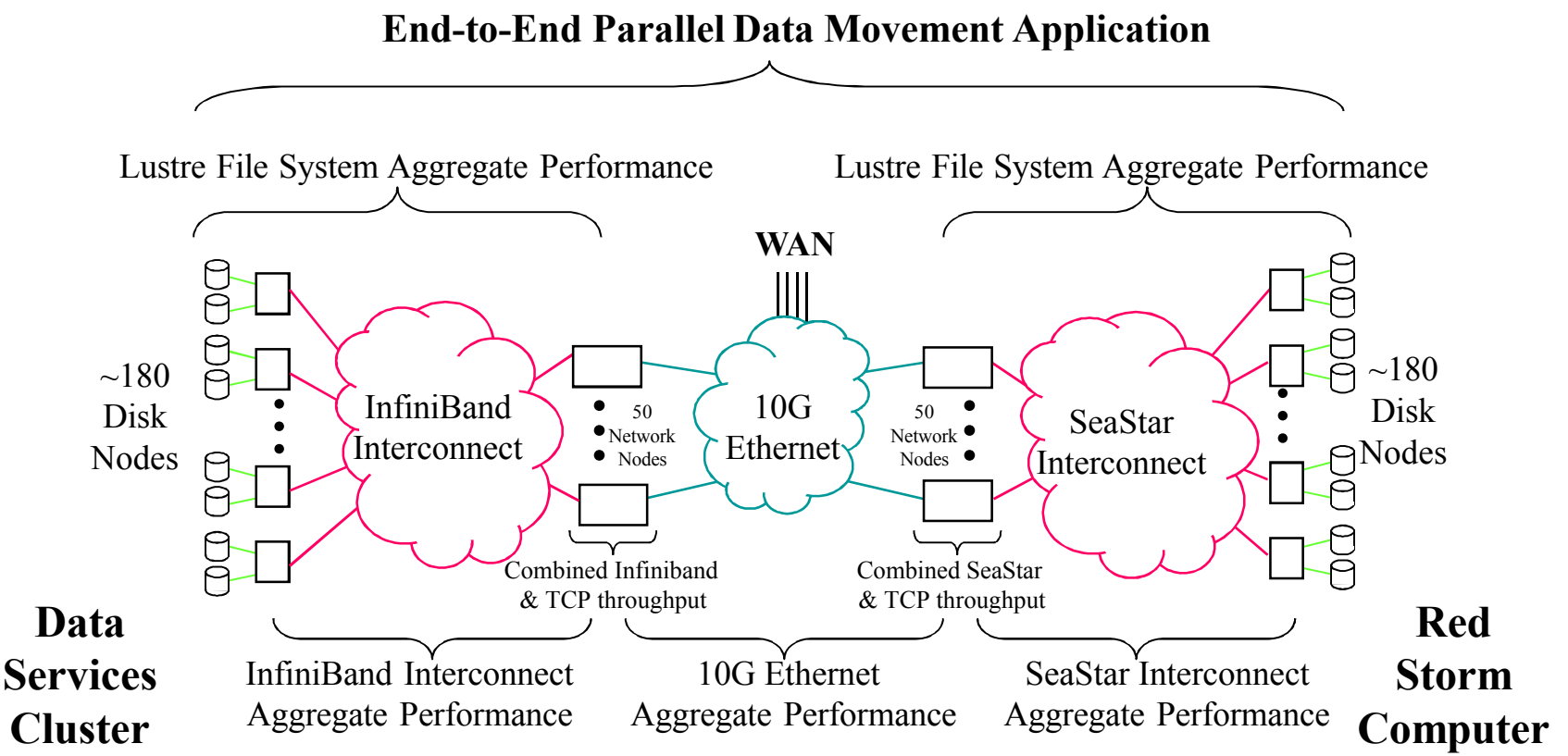
Where Is This a Problem?

- 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



Example Red Storm Architecture



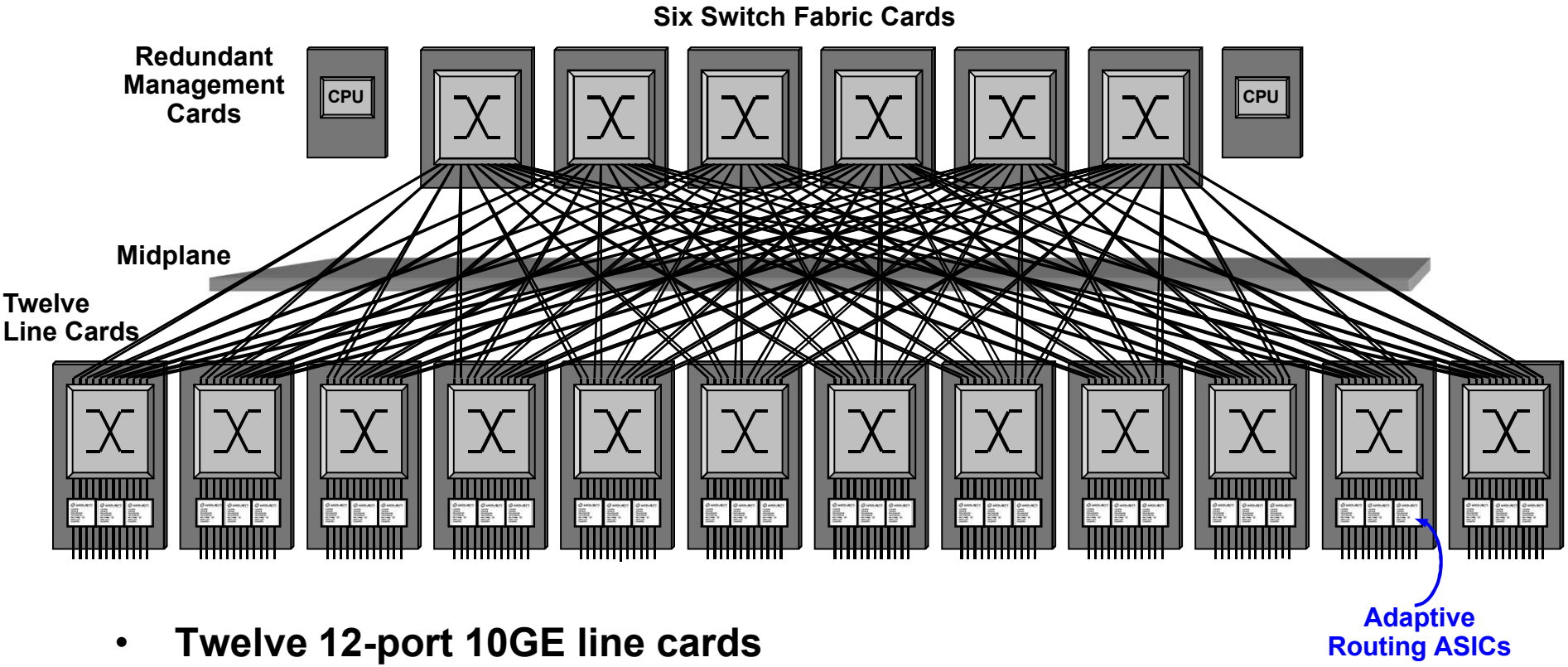
What Are We Doing?

- Investigating and supporting dynamic routing implementations
- Formed a collaboration to demonstrate one particularly promising implementation
  - Woven active congestion management
  - Chelsio and NetEffect 10GE RNICs
  - Sandia 128 node cluster (Talon)

Goals of the Collaboration

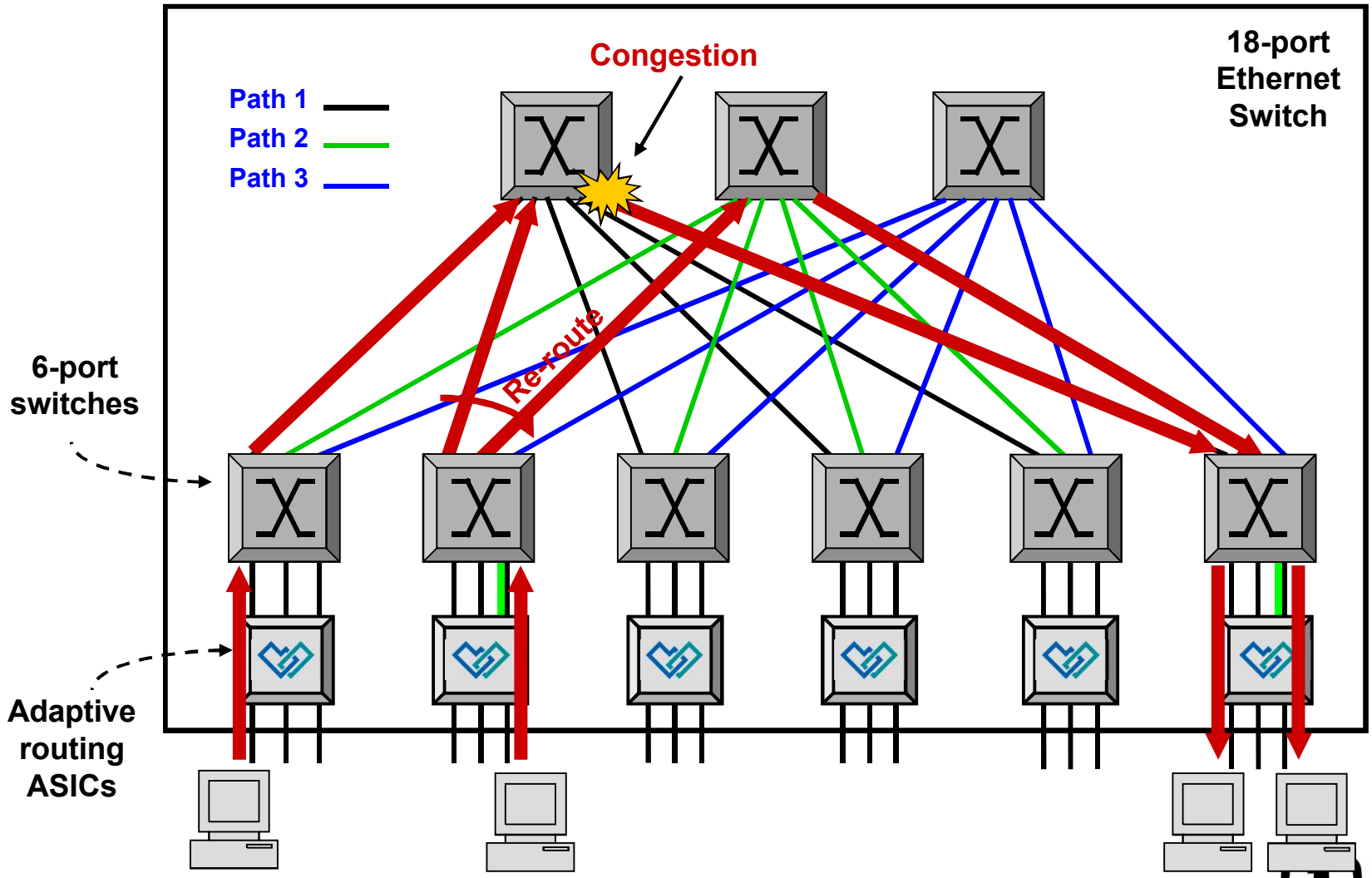
- Demonstrate scalability of a high-density 10GbE switching infrastructure
- Demonstrate effectiveness of dynamic routing over static routing for low radix switch interconnects
- Evaluate Low Latency 10 GbE with RDMA as an alternative for deploying:
  - Common I/O infrastructure between PetaScale resources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

Adaptive Routing Ethernet Switch

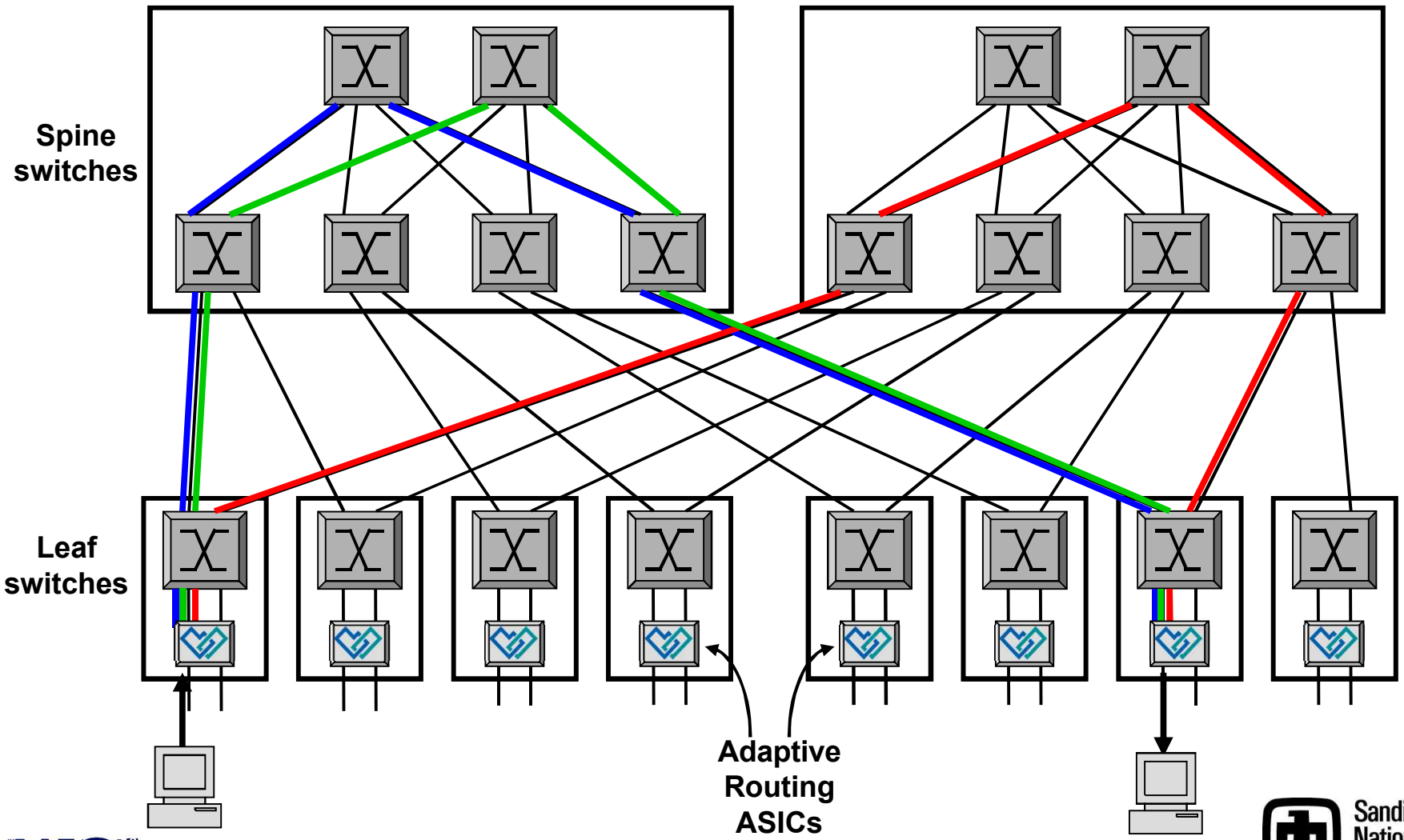


- Twelve 12-port 10GE line cards
- Fat-tree architecture
- ASICs at edge of switch perform adaptive routing: detects congestion and reroutes traffic around hot spots

Ethernet Fabric using Multiple Paths and Adaptive Routing to Avoid Hotspots



Adaptive Routing over a Multi-tier Fat Tree Topology



Progress

- 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 tuning
  - HP Linpack tuning always time consuming
- Fully functioning 128 Dell 1750 nodes with MVAPICH2 and Chelsio RNICs
- OFED 1.2.5 working well
- Bringing up a 12 node cluster with NetEffect RNICs
- Still working issues and producing results

Sandia CBench Suite

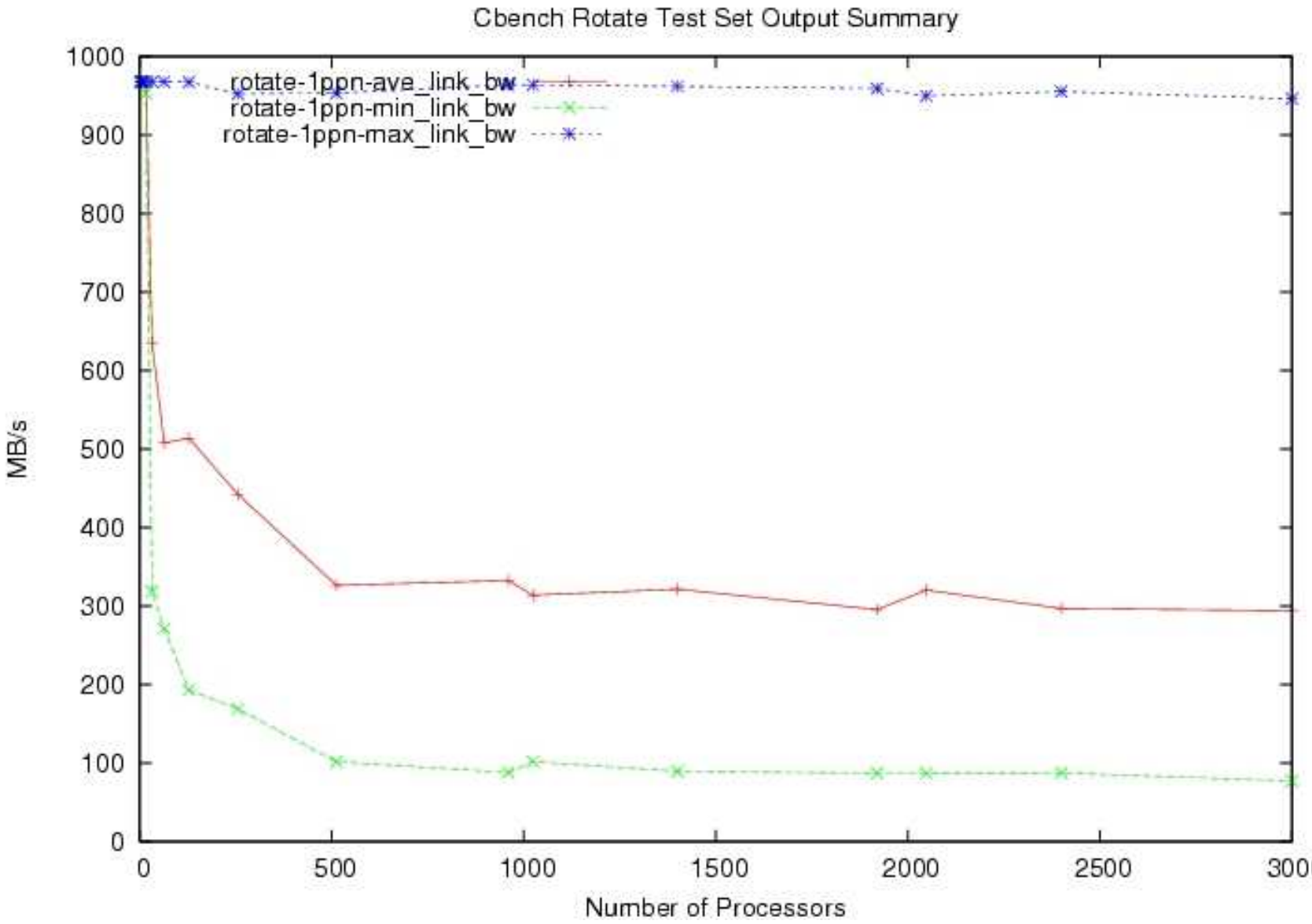
- Sandia's CBench Suite includes industry standards:
  - HPCC, Intel MPI Benchmarks, OSU, NAS, etc.
- Also includes benchmarks developed to stress bi-sectional bandwidth and latency
  - "Rotate Bandwidth" pair-wise transmits 80MB of data from half of the nodes to the other half. Repeat that test for 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

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 may impact bandwidth
  - Relaxed ordering working well
- Issues that appear only at scale are difficult to debug

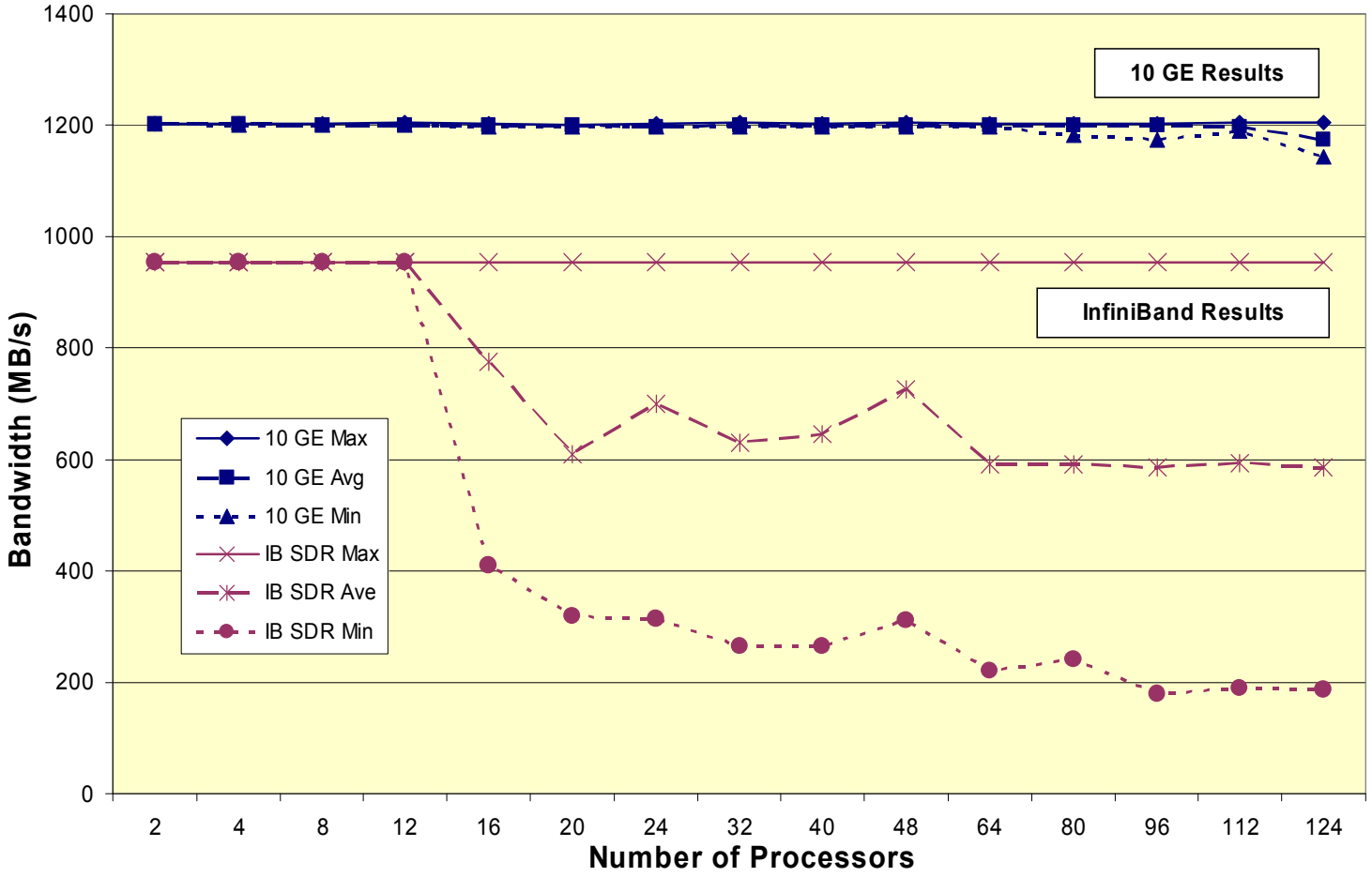
Per Pair Bisectional Bandwidth

~4000 Node IB Cluster

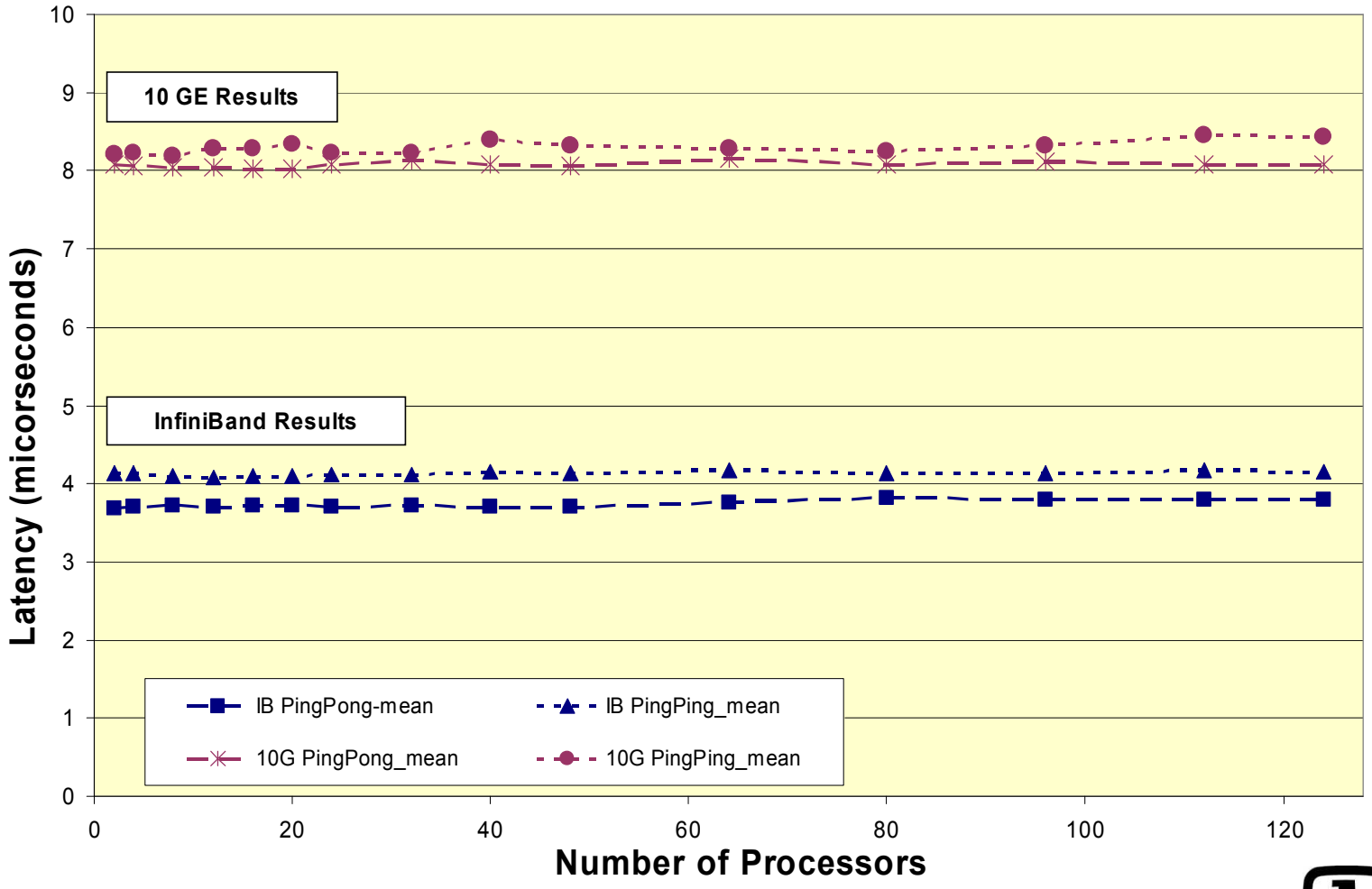


Sandia National Laboratories

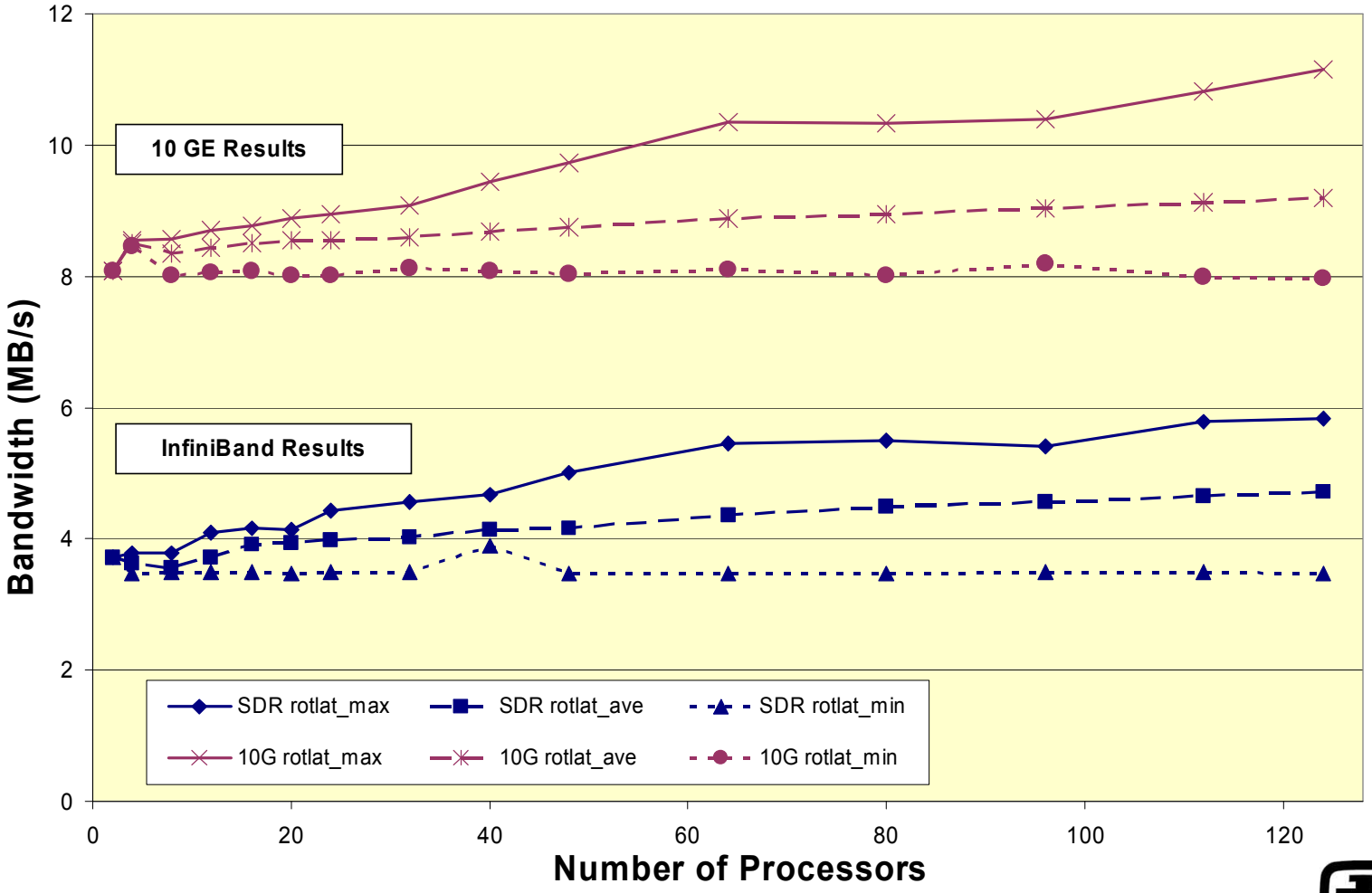
Cbench Rotate Benchmark Test (Relaxed Ordering)



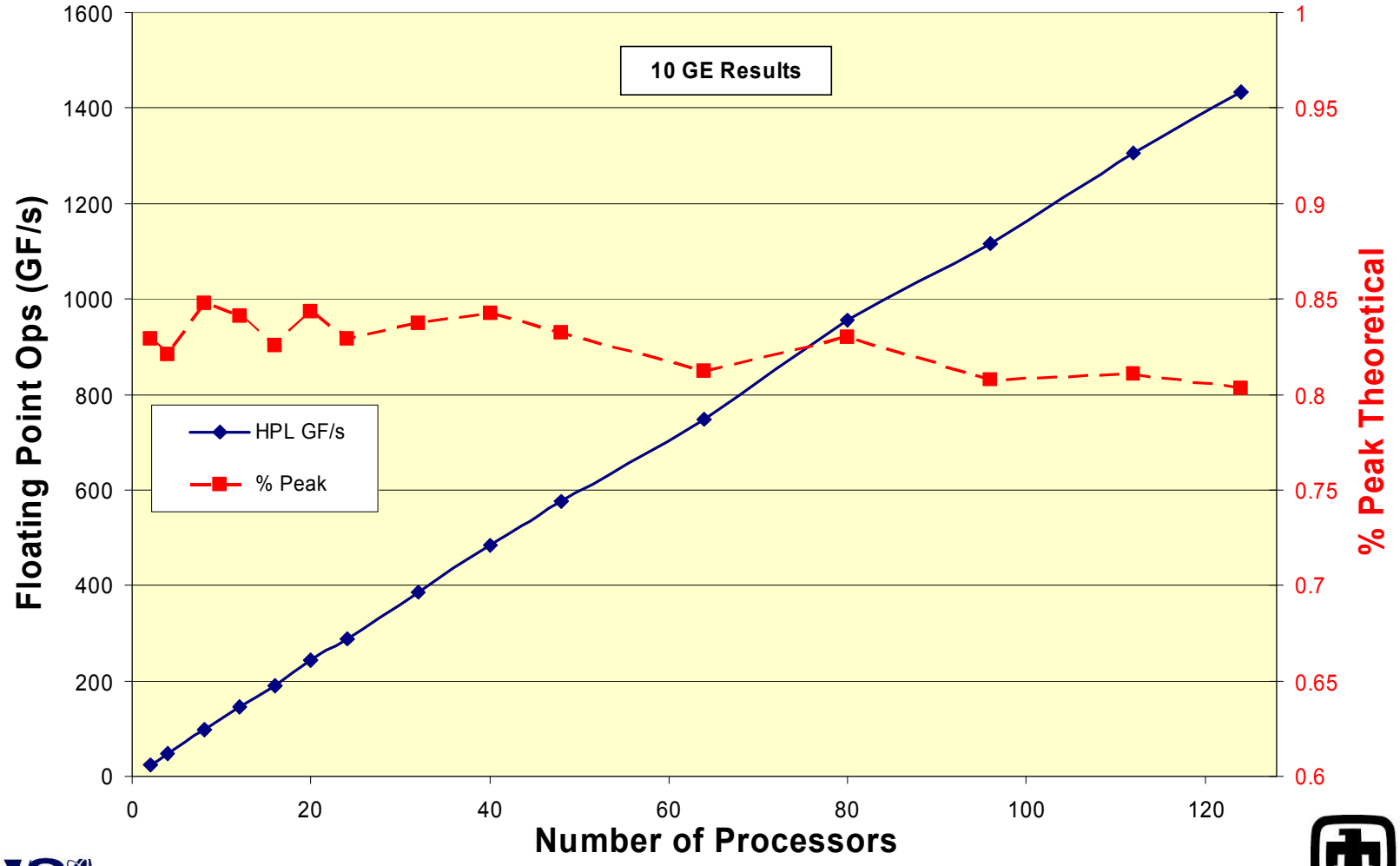
Intel MPI Latency Benchmark Test



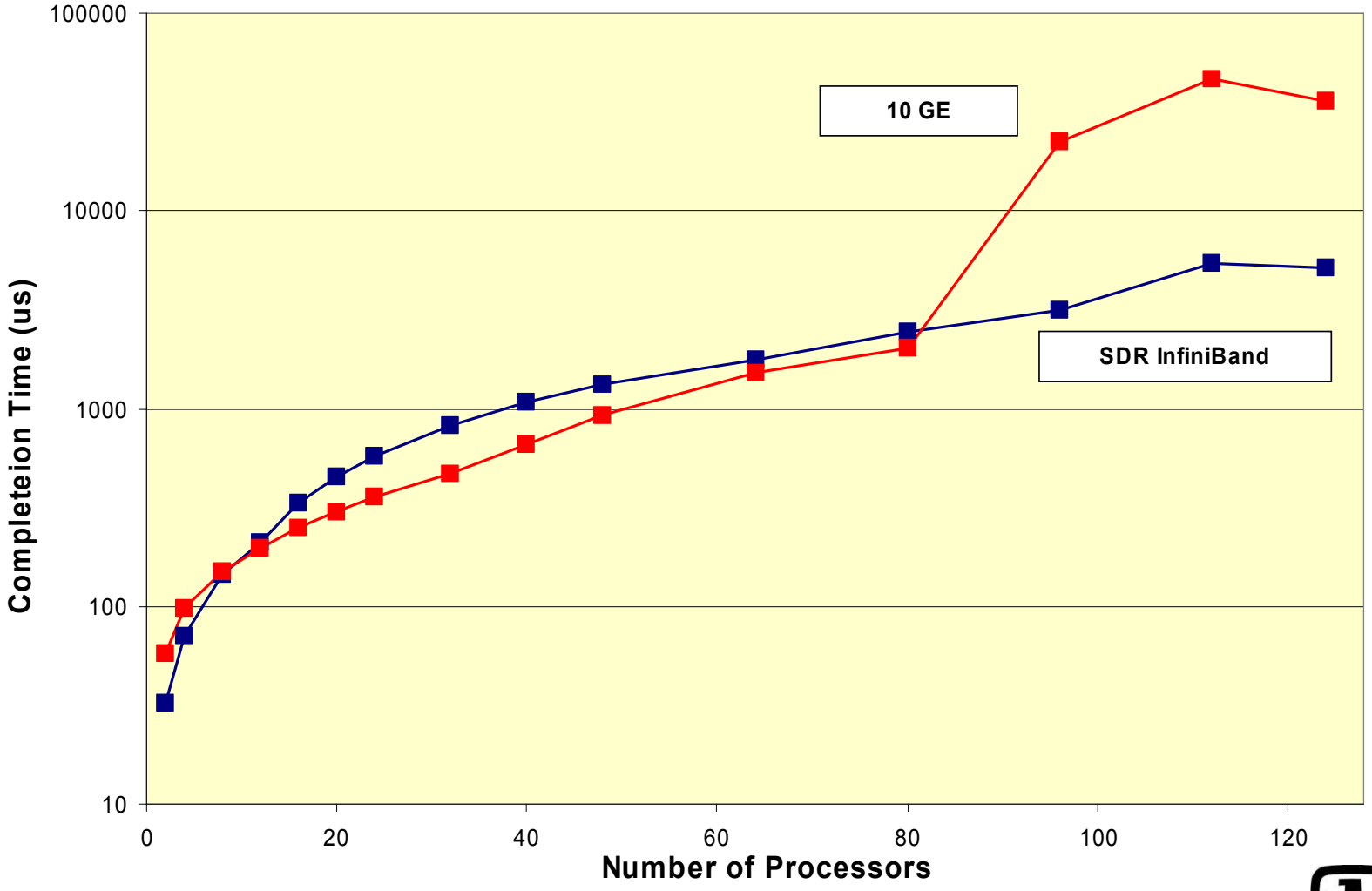
Cbench Rotate Latency Benchmark Test



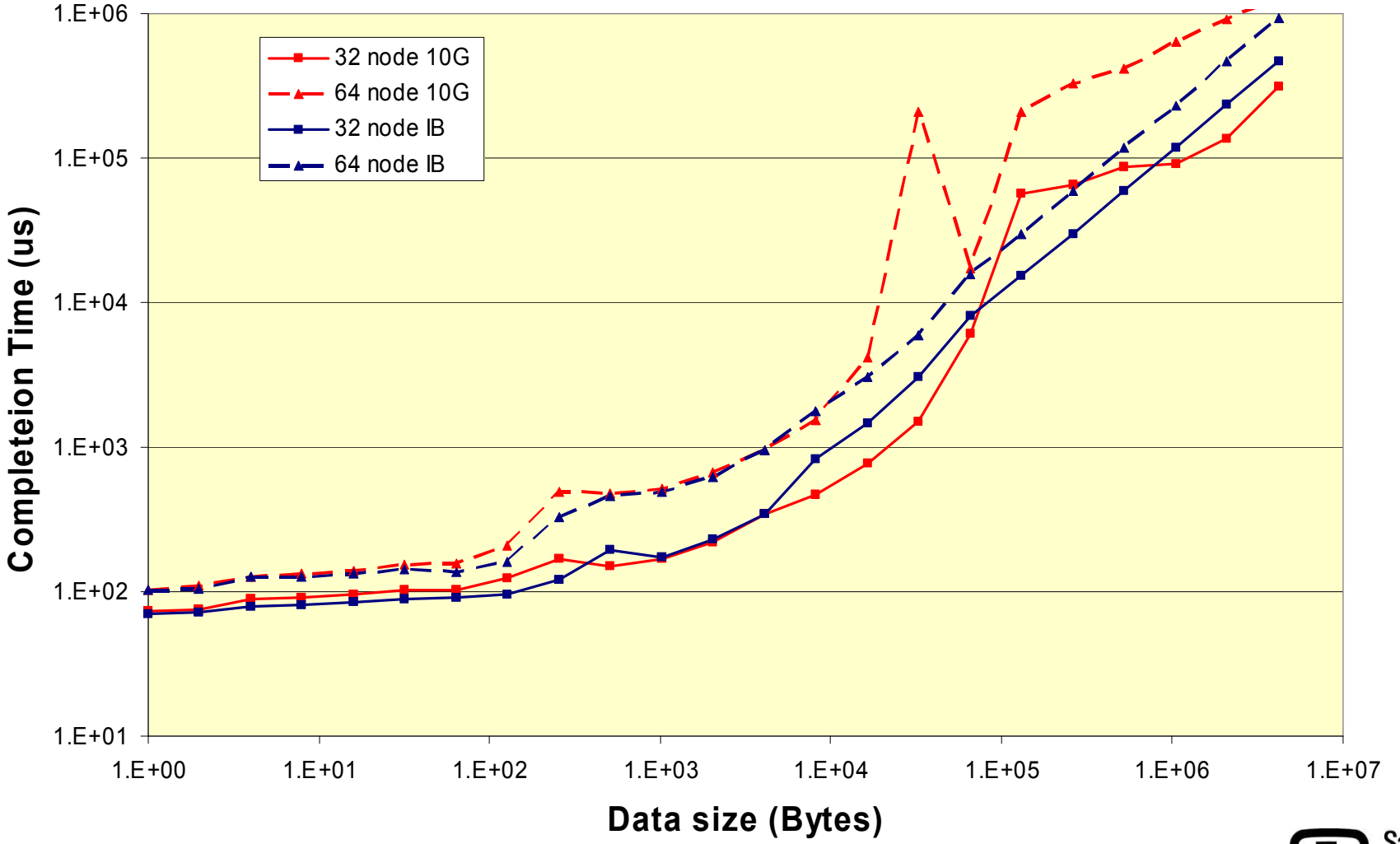
HP Linpack Benchmark Test



Intel MPI All to All Benchmark 8KB Data Size



Intel MPI All-to-All Benchmark



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

Future Work

- Complete debugging
- Production interoperability testing of RNICs
- Multi-tier switch operation
- Comparison with DDR
- Comparison with TOE
- Lustre/pNFS etc. testing with RDMA
- Open MPI testing