

Future Routing Schemes in Petascale clusters

Gilad Shainer, Mellanox, USA

Ola Torudbakken, Sun Microsystems, Norway

Richard Graham, Oak Ridge National Laboratory, USA

Birds of a Feather Presentation



INTERNATIONAL
SUPERCOMPUTING CONFERENCE

'08



Future HPC systems will span tens-of-thousands of nodes, all connected together via high-speed connectivity solutions to form multi-Petaflop clusters. With the growing size of clusters and CPU cores per cluster node, not only the traditional demands from the cluster interconnect increase dramatically, but new demands are introduced. The interconnect needs to provide balanced throughput and latency, to address IO requirements of each CPU core, while maintaining high network utilization. Moreover, the overall number of communication links grows with the size of the cluster and link data errors have become a growing concern for large-scale platforms, as they tend to have an adverse affect on the performance. The session will drive a discussion on the needed communication capabilities, static and dynamic routing, congestion control and handling networks errors. The session will also present models and simulations results on the new adaptive routing implementation for InfiniBand networks.

InfiniBand Clusters - Present (2007)



1280 server nodes



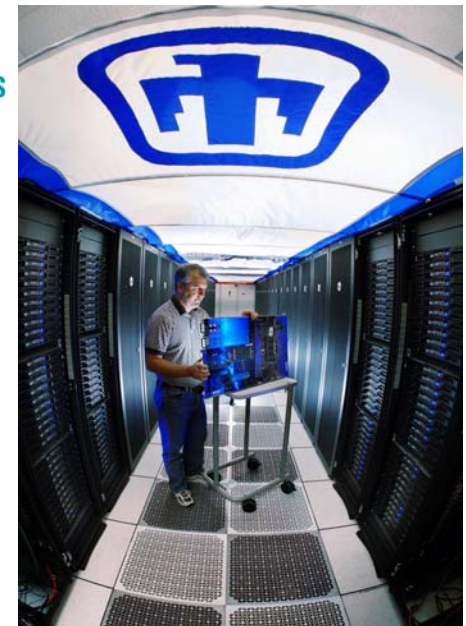
1300 server nodes



4500 server nodes

Teraflop Era

ISC'07



2300 server nodes



1400 server nodes



InfiniBand Clusters - Present (2008)



3456 nodes



3936 nodes



Petaflop Era
ISC'08

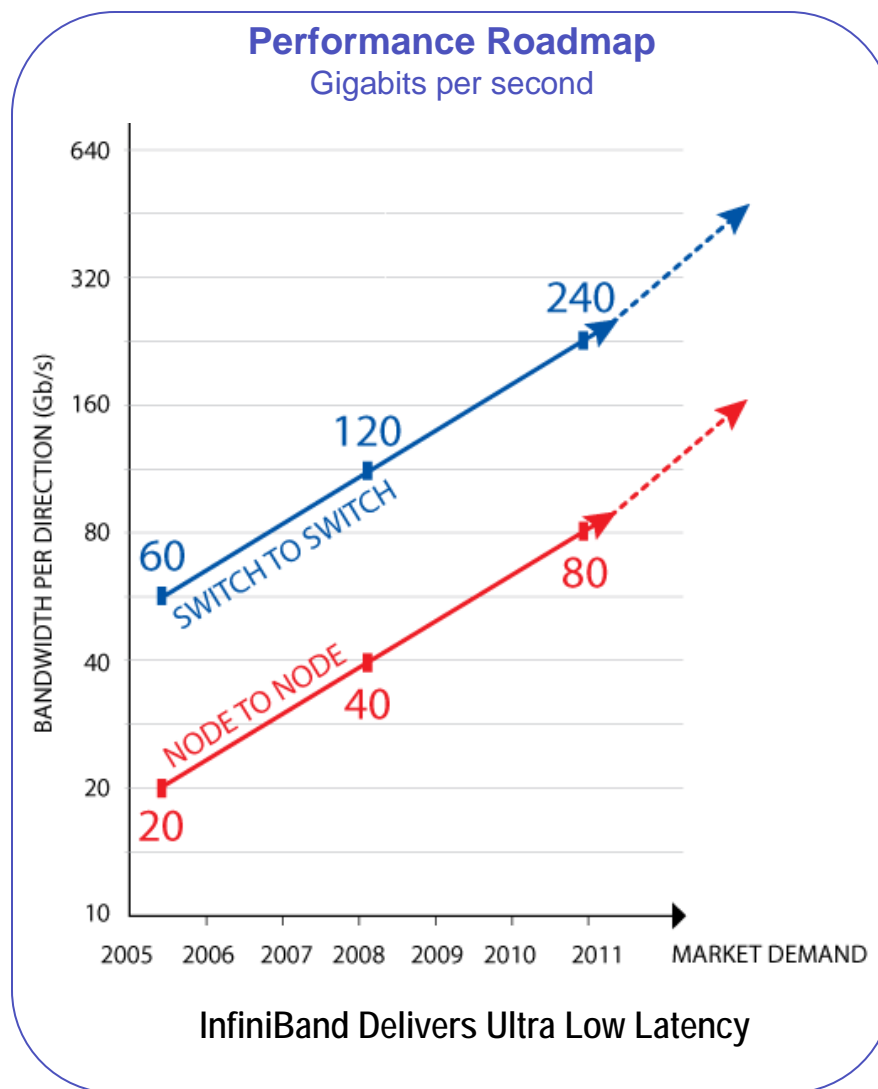
- More cores, more servers
- Higher speed, scalable and reliable interconnect



InfiniBand Technology Leadership



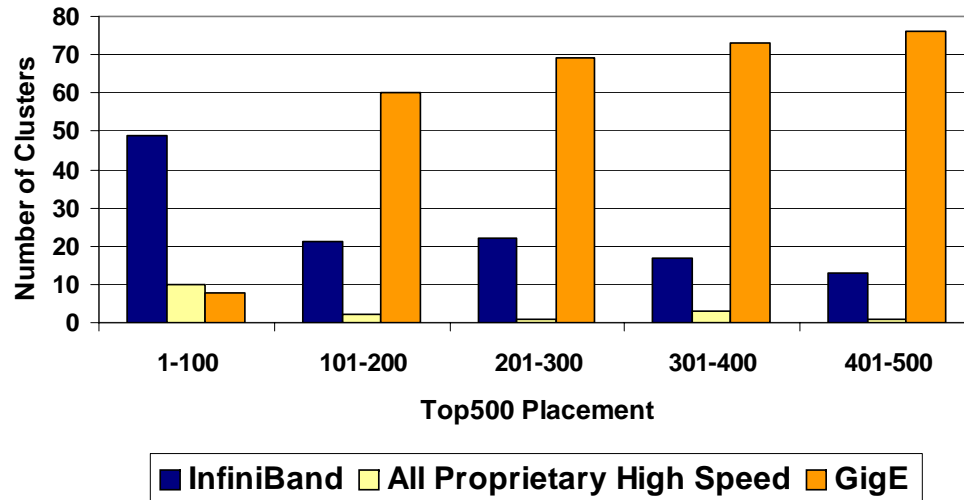
- **Industry Standard**
 - Hardware, software, cabling, management
 - Design for clustering and storage interconnect
- **Price and Performance**
 - 40Gb/s node-to-node
 - 120Gb/s switch-to-switch
 - 1us application latency
 - Most aggressive roadmap in the industry
- **Reliable with congestion management**
- **Efficient**
 - RDMA and Transport Offload
 - Kernel bypass
 - CPU focuses on application processing
- **Scalable for Petascale computing & beyond**
- **End-to-end quality of service**
- **Virtualization acceleration**
- **I/O consolidation Including storage**



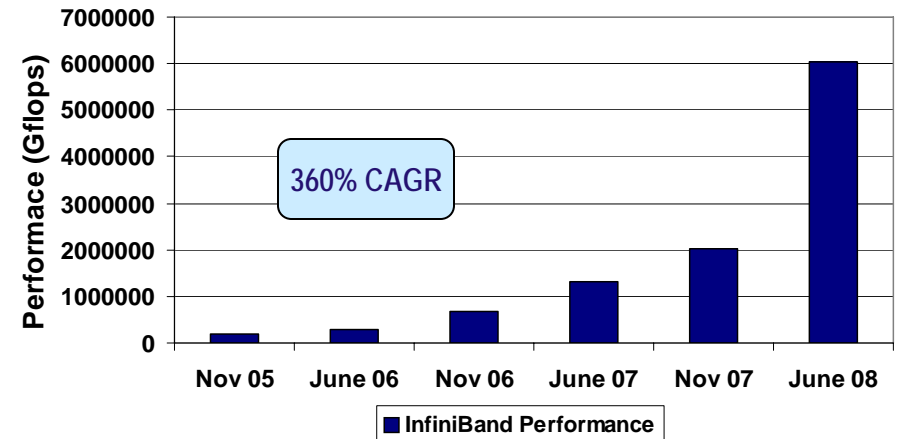
InfiniBand in the TOP500



Top500 Interconnect Placement



InfiniBand Clusters - Performance



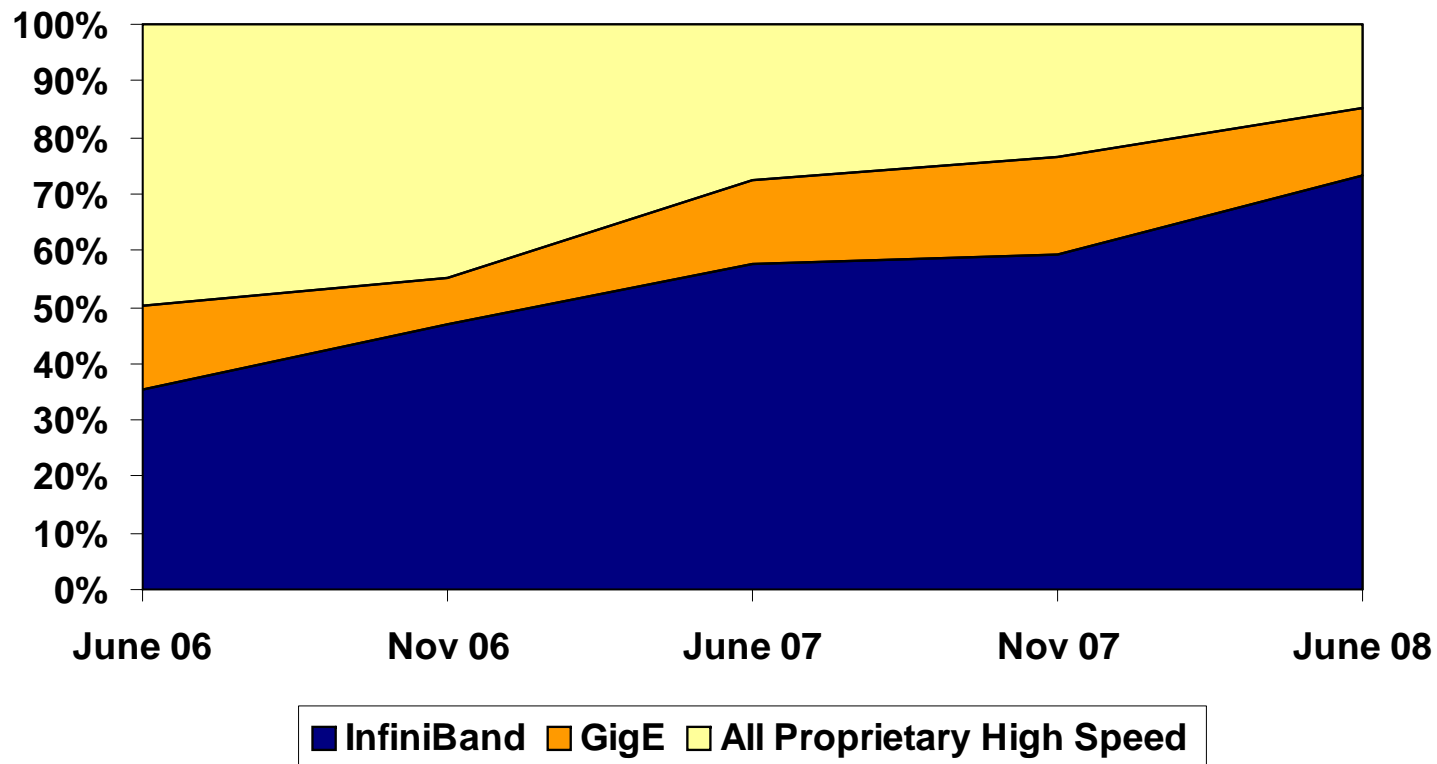
- **InfiniBand makes the most powerful clusters**
 - 5 of the top 10 (#1, #4, #7, #8, #10), 49 of the Top100
 - The leading interconnect for the Top200
 - Responsible for ~40% of the total Top500 performance
- **InfiniBand enables the most power efficient clusters**
- **InfiniBand QDR expected Nov 2008**
- **No 10GigE clusters exist on the list**



Top100 Trends Over Time



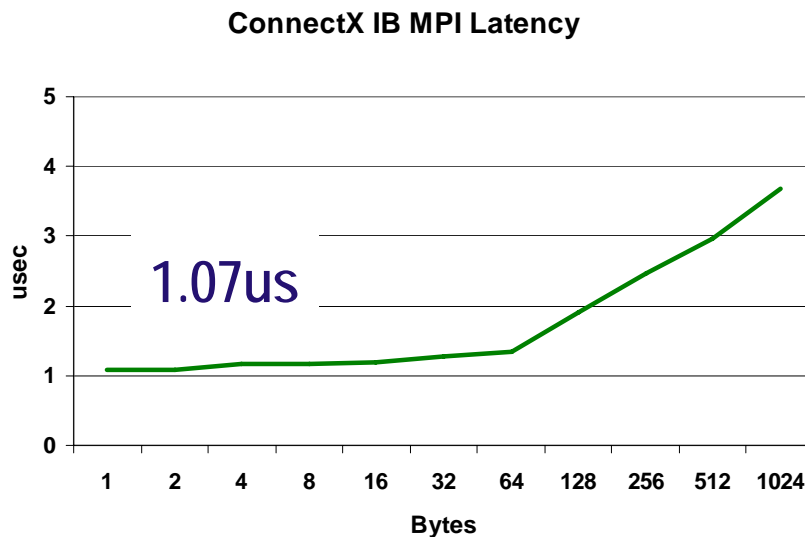
TOP100 Clustering Interconnect Share Over Time



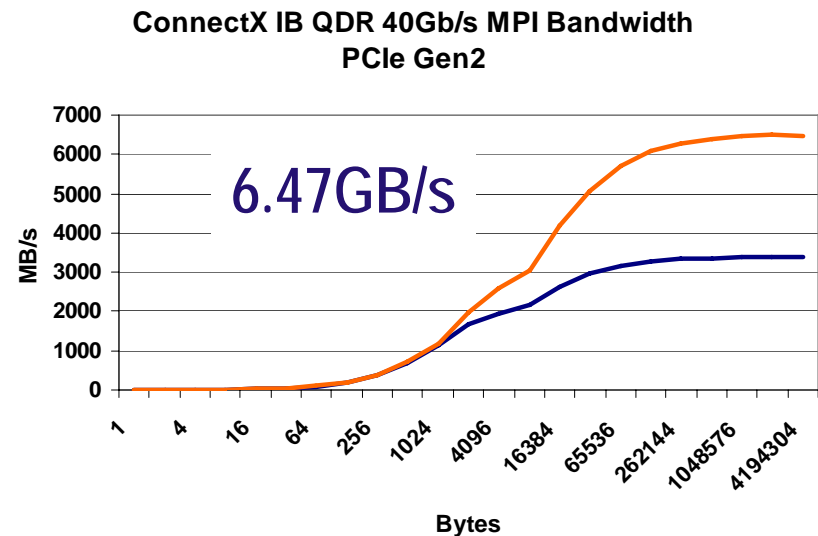
ConnectX - Fastest InfiniBand Technology



- **Performance driven architecture**
 - MPI latency 1us, ~6.5GB/s with 40Gb/s InfiniBand (bi-directional)
 - MPI message rate of >40 Million/sec
- **Superior real application performance**
 - Engineering Automotive, oil & gas, financial analysis, etc.



— PCIe Gen2 IB QDR Latency

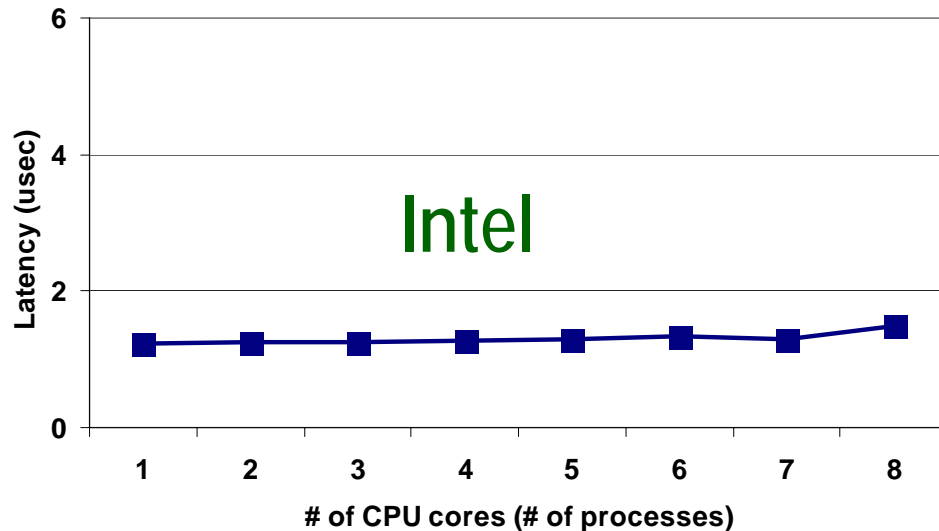


— IB QDR Uni-dir — IB QDR Bi-dir

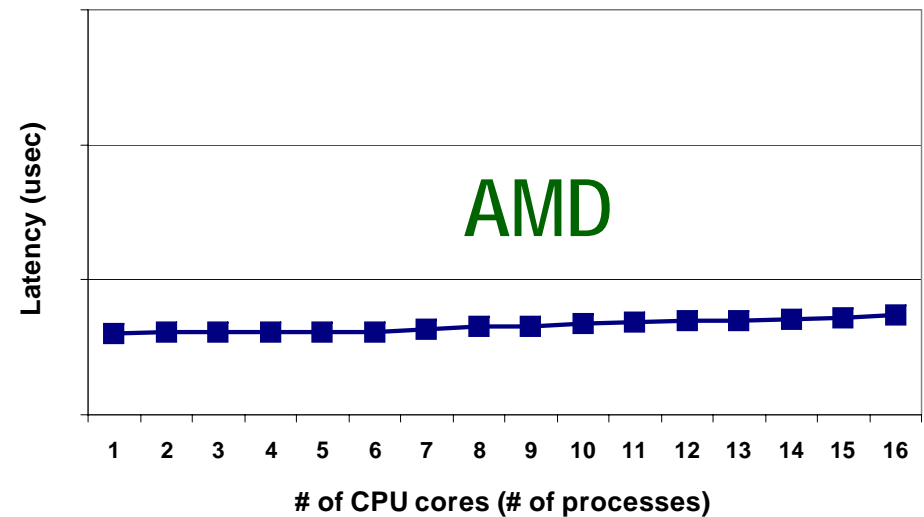
ConnectX Multi-core MPI Scalability



Mellanox ConnectX
MPI Latency - Multi-core Scaling



Mellanox ConnectX
MPI Latency - Multi-core Scaling



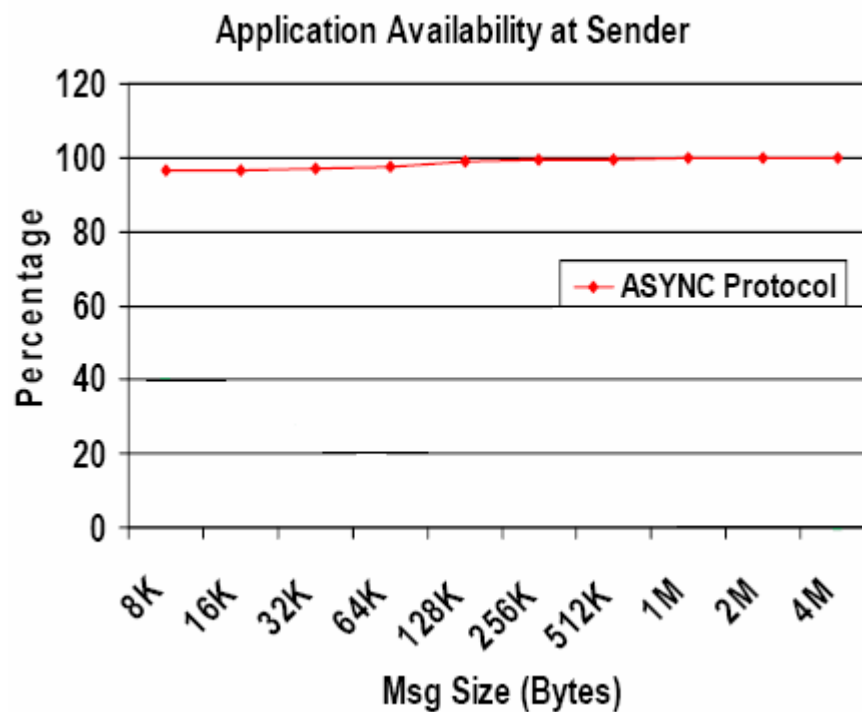
- Scalability to 64+ cores per node, to 20K+ nodes per subnet
- Guarantees same low latency regardless of the number of cores
- Guarantees linear scalability for real applications



Overlapping Communication/Computation



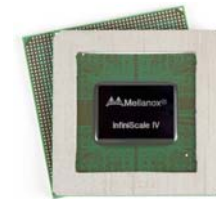
- Transport offload
Interconnect maximize applications efficiency
- Maximum CPU cycles dedicated to application
- Asynchronous Progress will be available with MVAPICH 1.1



InfiniScale IV Switch: Unprecedented Scalability



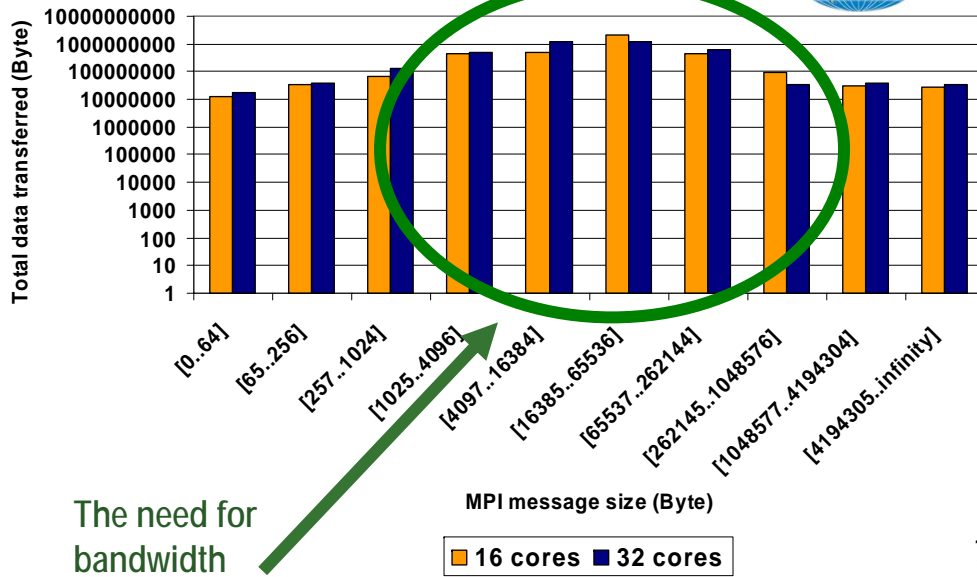
- **36 40Gb/s or 12 120Gb/s InfiniBand Ports**
 - Adaptive routing and congestion control
 - Virtual Subnet Partitioning
- **6X switching and data capacity**
 - Vs. using 24-port 10GigE Ethernet switch devices
- **4X storage I/O throughput**
 - Critical for backup, snapshot and quickly loading large datasets
 - Vs. deploying 8Gb/s Fibre Channel SANs
- **10X lower end-to-end latency performance**
 - Vs. using 10GigE/DCE switches and iWARP-based adapters
- **3X the server and storage node cluster scalability when building a 3-tier CLOS fabric**
 - Vs. using 24-port 10GigE Ethernet switch devices



HPC Applications Demand Highest Throughput

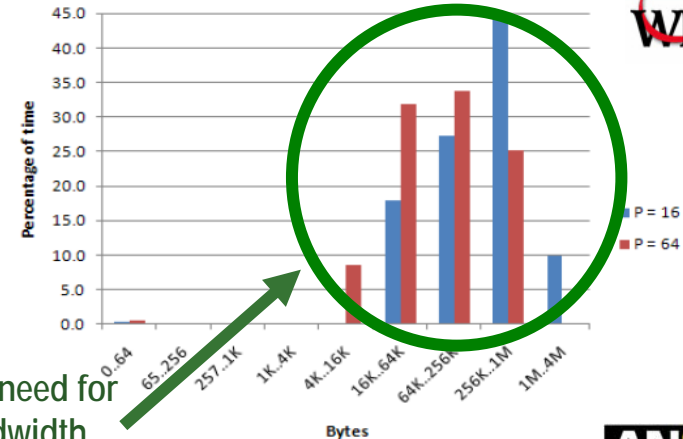


LS-DYNA Profiling



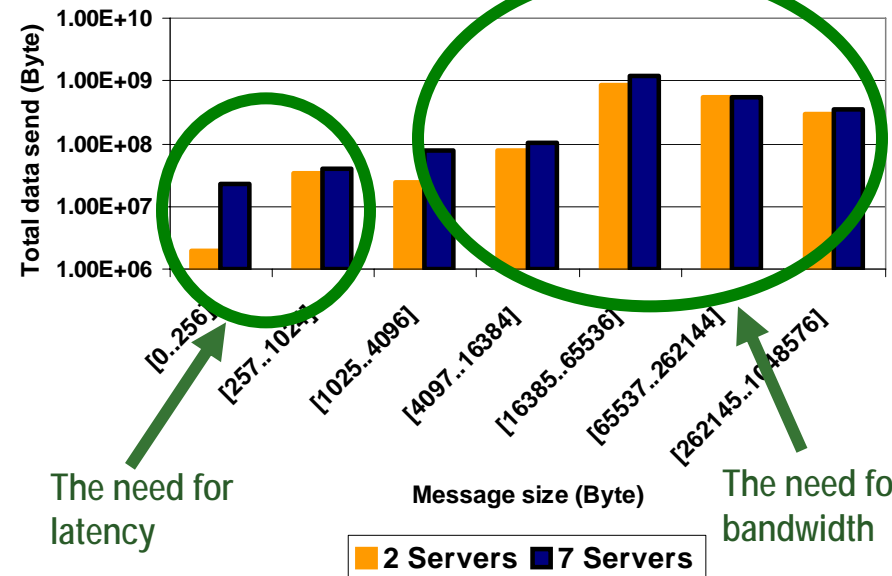
The need for bandwidth

Message size distribution %



The need for bandwidth

Fluent Message Size Profiling



The need for latency

The need for bandwidth

Scalability Mandates
Highest Bandwidth
Lowest Latency

Addressing the Needs for Petascale Computing



- **Faster network streaming propagation**
 - Network speed capabilities
 - Solution: InfiniBand QDR
- **Large clusters**
 - Scaling to many nodes, many cores per node
 - Solution: High density InfiniBand switch
- **Balanced random network streaming**
 - "One to One" random streaming
 - Solution: Adaptive routing
- **Balanced known network streaming**
 - "One to One" known streaming
 - Solution: Static routing
- **Un-balanced network streaming**
 - "Many to one" streaming
 - Solution: Congestion control

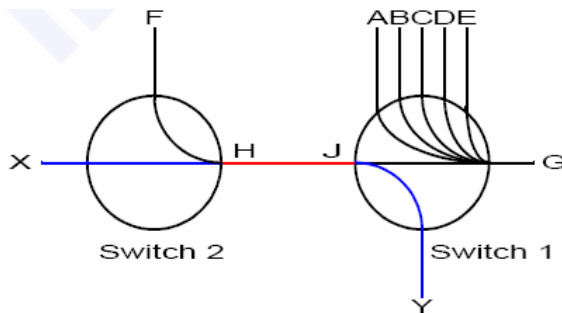


Designed to handle all communications in HW

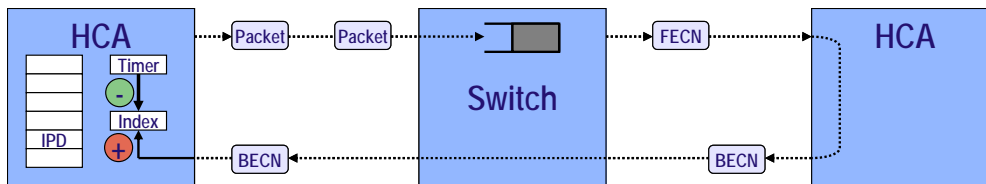
Hardware Congestion Control



- Congestion spots → catastrophic loss of throughput
 - Old techniques are not adequate today



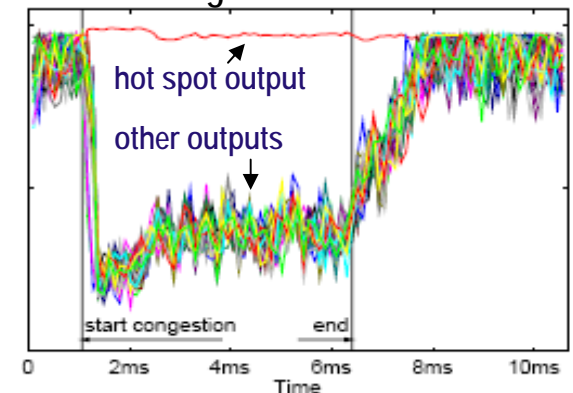
- InfiniBand HW congestion control
 - No a priori network assumptions needed
 - Automatic hot spots discovery
 - Data traffics adjustments
 - No bandwidth oscillation or other stability side effects
 - SM receives notices of congestion
- Ensures maximum effective bandwidth



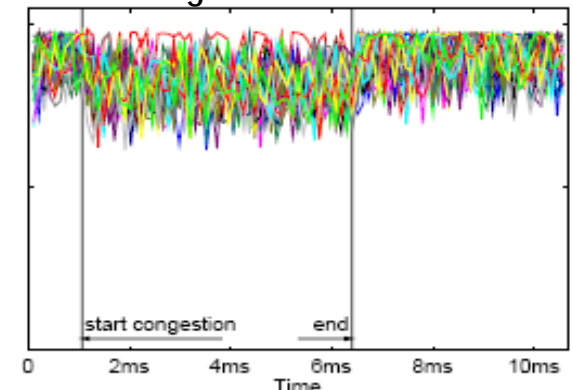
Simulation results

32-port 3 stage fat-tree network
High input load, large hotspot degree

Before congestion control



After congestion control



"Solving Hot Spot Contention Using InfiniBand Architecture Congestion Control
IBM Research; IBM Systems and Technology Group; Technical University of Valencia, Spain

- **A capability to overcome data corruptions is required**
 - Hardware or Software
- **High cluster efficiency/scalability require hardware mechanism**
 - Re-transmission mechanism
 - Forward error correction (FEC) mechanism
- **FEC requires the sender to add redundant data to the data packet**
 - Allow the receiver (switch or end-node) to correct BER related errors
 - FEC can avoid Re-transmission of data
 - But require high overhead bandwidth (up to 10s% of the transmitted data)
 - Significantly reduces the effective bandwidth (the "true" bandwidth)
 - Reduces the overall cluster performance
 - Can cause additional congestion problems
- **FEC overhead bandwidth increases with the data traffic**
 - Does not provide a good scaling solution.

InfiniBand Reliability Mechanism



- **InfiniBand provides a scalable and reliable high-speed interconnect**
 - for servers and storage
- **InfiniBand uses an end-to-end hardware reliability mechanism**
 - For data integrity and to guarantee reliable data transfer between end-nodes
- **InfiniBand packets contain two Cyclic Redundancy Checks (CRCs)**
 - The Invariant CRC (ICRC)
 - covers all fields which do not change as the packet traverses the fabric
 - The Variant CRC (VCRC)
 - covers the entire packet
- **The two CRCs allows switches (and routers) to modify appropriate fields and still maintain end-to-end data integrity**

The Petaflop Era Model – Simulations Results



- The model was presented at ISC07
- Model parameters – worse case
 - BER - 10^{-13}
 - Node bandwidth - 300Gb/s
 - The maximum number of hops between two nodes - 20
 - Number of nodes >50,000
 - CTP period - 5msec
- Latency overhead
 - 1.2% for BER of $10e-13$
- Bandwidth overhead
 - 3×10^{-5} bits per bit of nominal bandwidth
- InfiniBand re-transmission - scalable cost-effective mechanism
 - Does not affect the overall application performance
 - Show no scaling limitations.

HPC Council Advisory



- Distinguished HPC alliance (OEMs, IHVs, ISVs, end-users)
- Members activities
 - Qualify and optimize HPC solutions
 - Early access to new technology, and mutual development of future solutions
 - Explore new opportunities within the HPC market
 - HPC targeted joint marketing programs
- A community effort support center for HPC end-users
 - Mellanox Cluster Center
 - Latest InfiniBand and the HPC Advisory Council member technology
 - Development, testing, benchmarking and optimization environment
 - End- user support center - HPCHelp@mellanox.com
- For details – HPC@mellanox.com