

# Ranger As The Intro For Future Extreme Scale Computing

*Birds of a Feather Presentation*



Gilad Shainer (Mellanox Technologies)

Tommy Minyard (Texas Advanced Computing Center)

John Fragalla (Sun Microsystems)



# Abstract



"Ranger" is the largest computing system in the world for open science research. Located at the Texas Advanced Supercomputing Center, Ranger serves NSF TeraGrid researchers and academic institutions. It is the most powerful commodity-based system that does not utilize specialized accelerators; only off-the-shelf CPUs and InfiniBand interconnect technology (Sun Data Center Switch 3456) to provide the 579 Teraflops of compute power, and therefore does not require new application development. The system consists of more than 15K sockets and centralized networking infrastructure that connects all the sockets in a full fat-tree configuration. As we enter the exascale computing era, the numbers of expected sockets will grow in a magnitude of order and the networking infrastructure could evolve into a mesh or hybrid one. Lessons learned from Ranger will provide a solid foundation for building future extreme scale computing infrastructures and will be the main focus of this session.

# The Teraflop Era – Ancient History ....



1280 server nodes

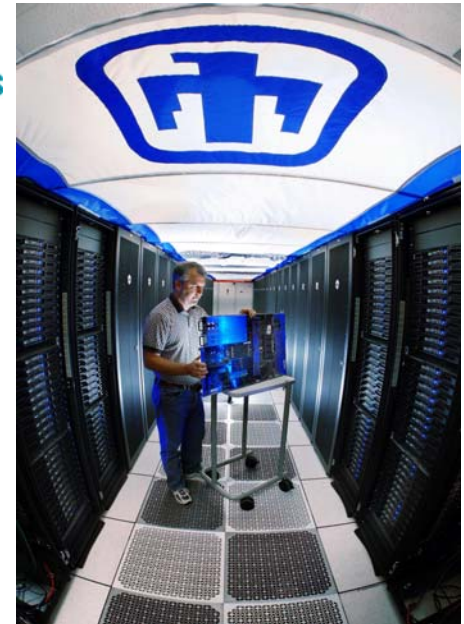


1300 server nodes



4500 server nodes

Teraflop Era



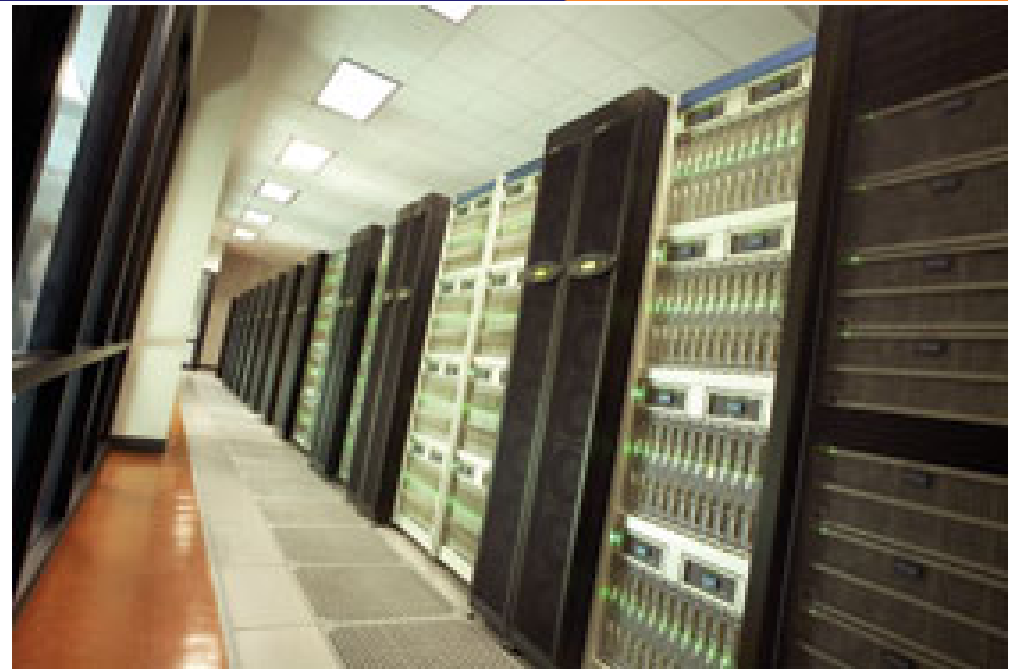
2300 server nodes



1400 server nodes



# The Petaflop Era Has Arrived!



3936 nodes

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

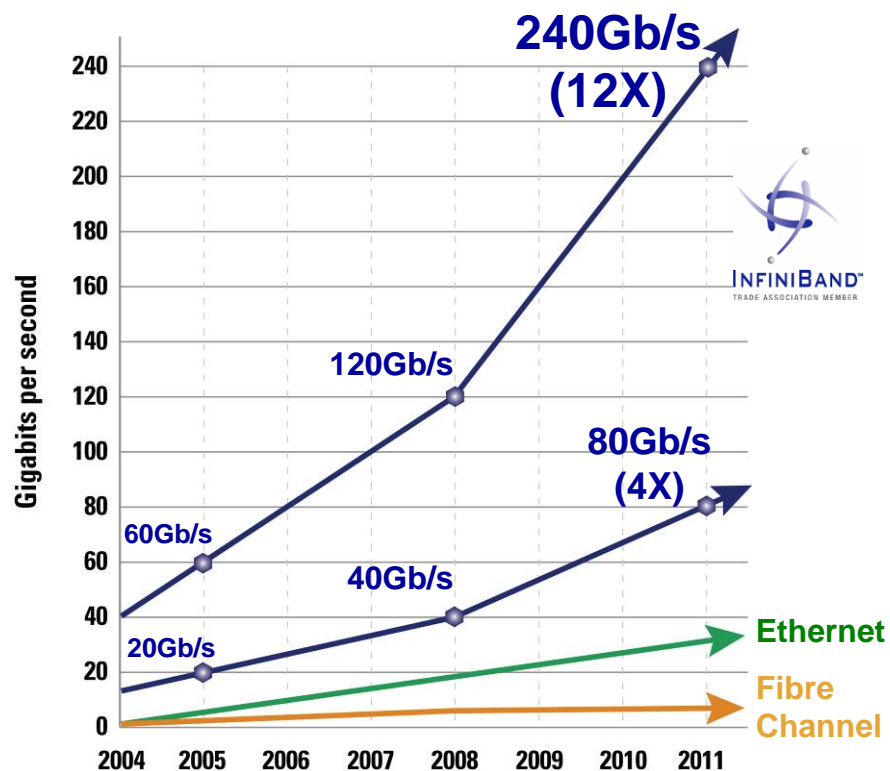


# InfiniBand Technology Leadership



- **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**

## The InfiniBand Performance Gap is Increasing

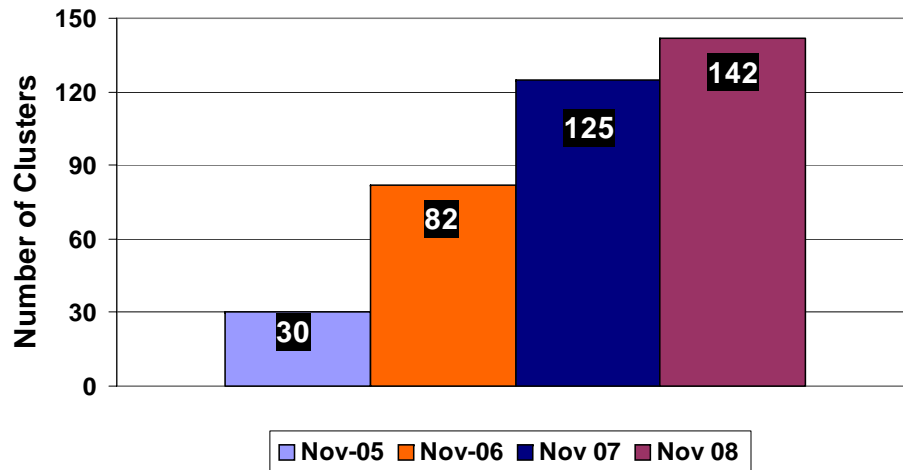


InfiniBand Delivers the Lowest Latency

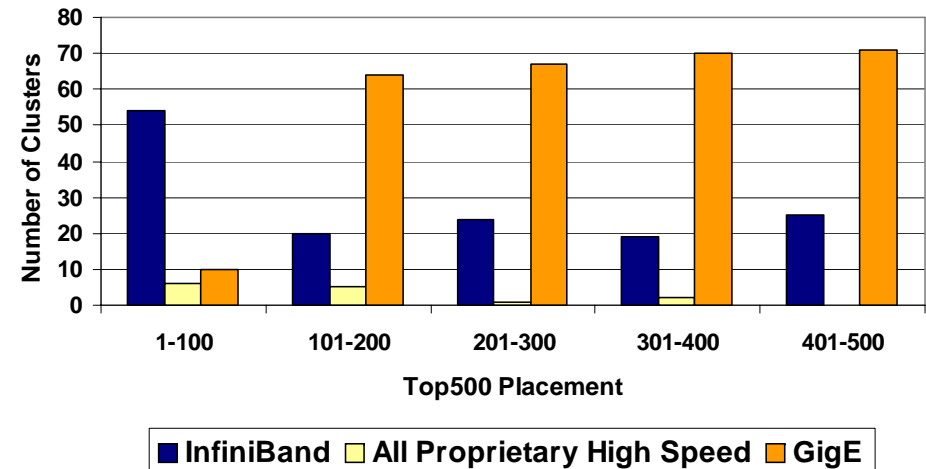
# InfiniBand in the TOP500



### Top500 InfiniBand Trends



### Top500 Interconnect Placement



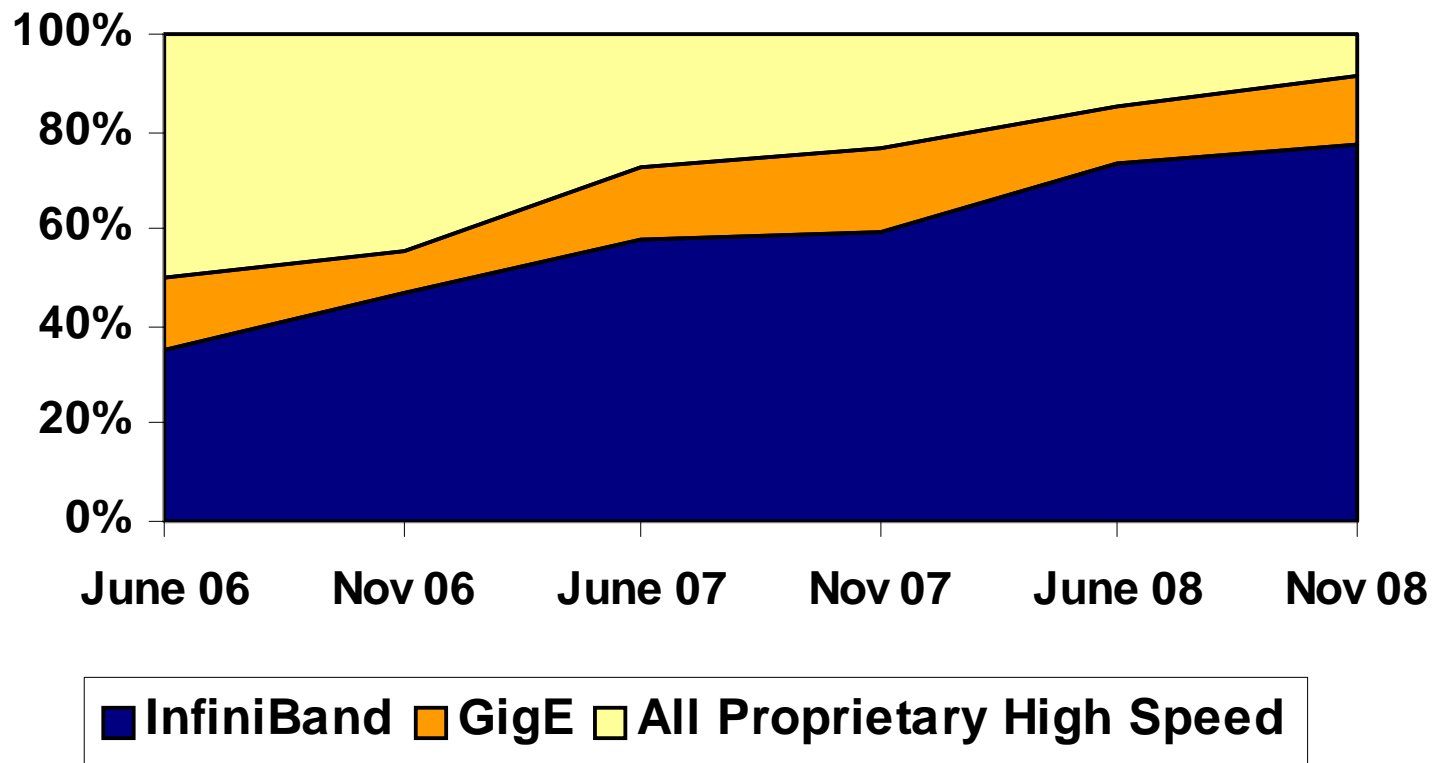
- **Mellanox InfiniBand makes the most powerful clusters**
  - 4 of the top 10 (#1, #3, #6, #10) and 54 of the Top100
- **All InfiniBand clusters use Mellanox switch silicon**
  - 139 out of the 142 clusters use Mellanox InfiniBand HCA adapters
- **InfiniBand enables the most power efficient clusters**
  - The only growing high speed interconnect solutions
- **Mellanox 40Gb/s InfiniBand end-to-end the only proven technology**



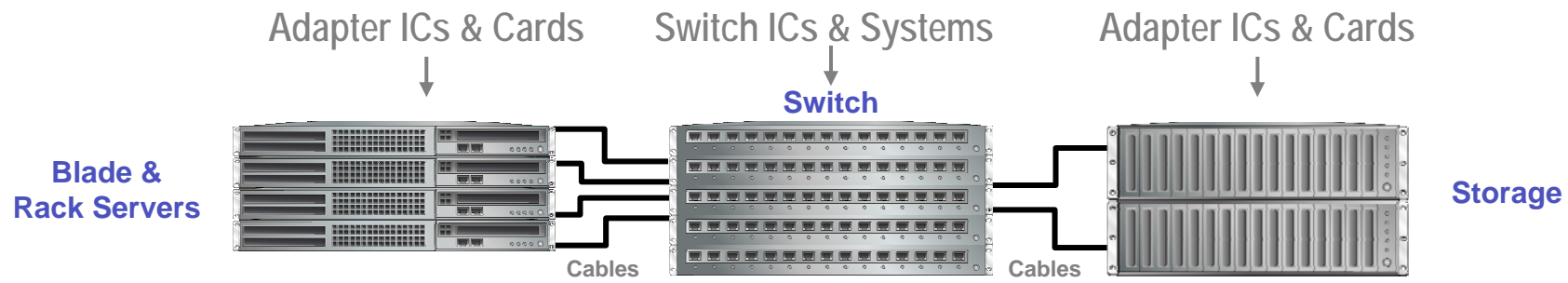
# Top100 Trends Over Time



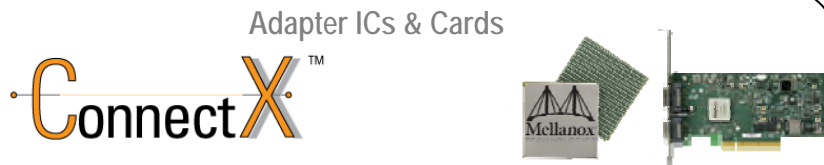
## Top100 Clustering Interconnect Share Over Time



# InfiniBand End-to-End Products



Standard & Custom Form Factors, Enabling Software, Reference Designs



- Single chip/slot optimizes cost, power, footprint, end-to-end reliability
- Dual-Port 10/20/40Gb/s InfiniBand, 10GigE & Data Center Ethernet
- I/O offload engines
- Virtualization acceleration

Switch ICs

## InfiniScale® IV



- Industry's fastest single-chip switch with the highest scalability
- 36 40Gb/s or 12 120Gb/s IB Ports
- Advanced traffic management
- Ultra-low latency

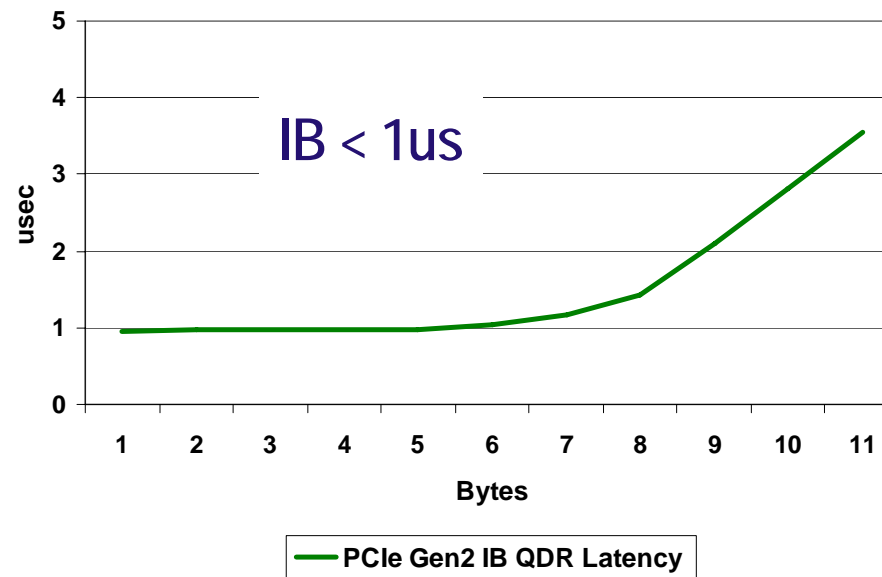


# ConnectX InfiniBand Performance



- **Performance driven architecture**
  - MPI latency  $< 1\mu\text{s}$ ,  $> 6.6\text{GB/s}$  with 40Gb/s InfiniBand (bi-directional)
  - MPI message rate of  $> 40$  Million/sec
- **Superior real application performance**
  - Scalability, efficiency, productivity

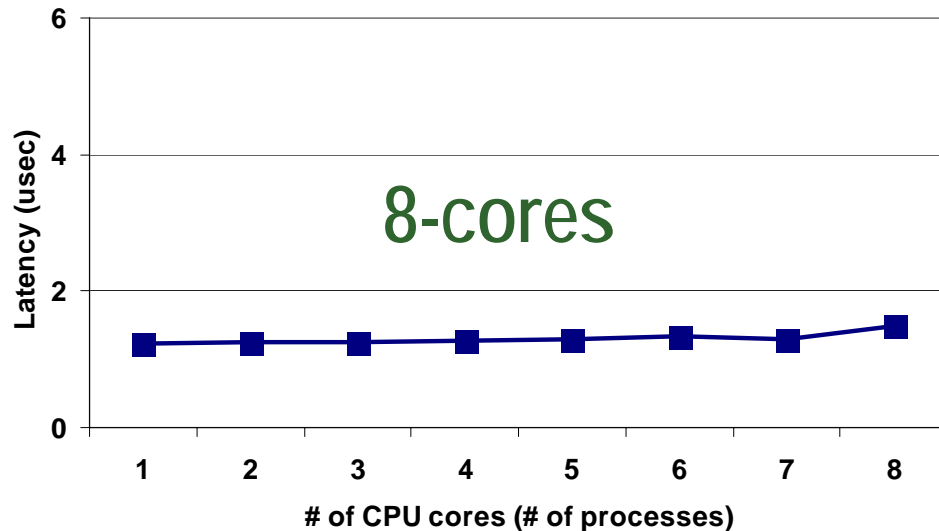
ConnectX IB MPI Latency



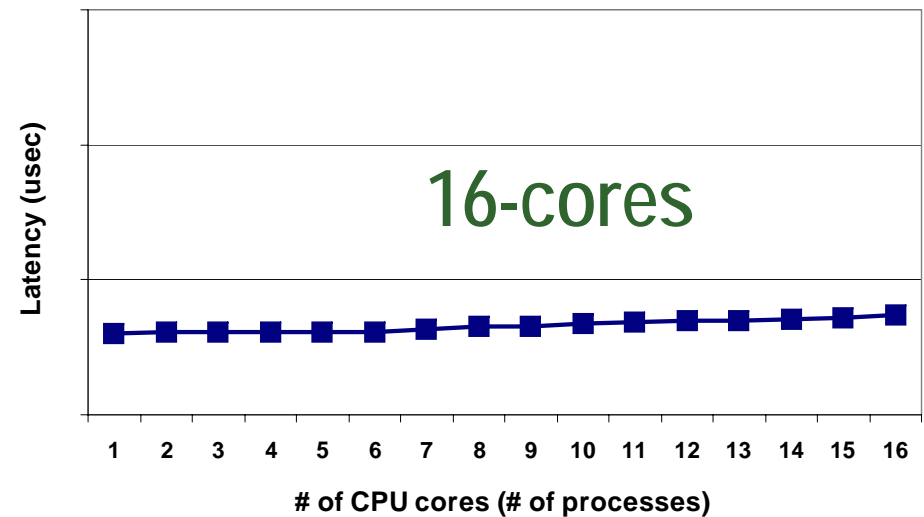
# ConnectX Multi-core MPI Scalability



Mellanox ConnectX  
MPI Latency - Multi-core Scaling



Mellanox ConnectX  
MPI Latency - Multi-core Scaling



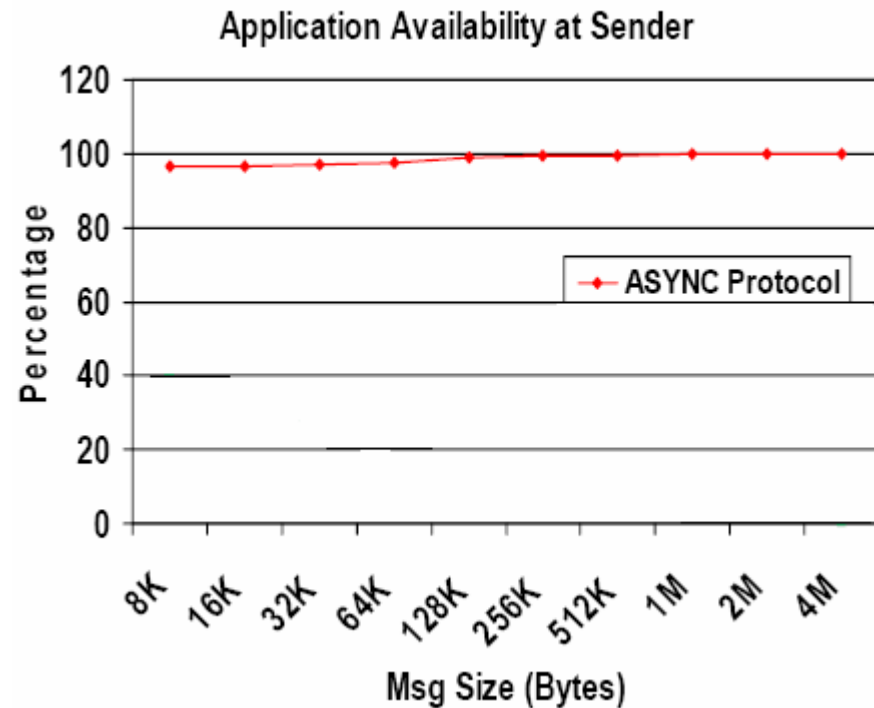
- Scalability to 64+ cores per node, to 20K+ nodes per subnet
- Guarantees same low latency regardless of the number of cores
- Guarantees linear scalability for real applications



# Overlapping Communication/Computation



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



# Minimizing Systems Noise Effect



- For Extreme Scale Scientific Simulation Through Function Delegation
- Utilizing Mellanox full transport-offload architecture
  - Essential for minimizing systems jitter
- Multi-faceted approach to address the effects of system noise
  - Application performance
- Isolating collective communication operations from the effects of system noise
  - Retarget collective operations to use enhanced InfiniBand HCAs
  - Application ability to execute asynchronously
  - Collective communications to applications via non-blocking communication functions



# Addressing Petascale Interconnect reliability

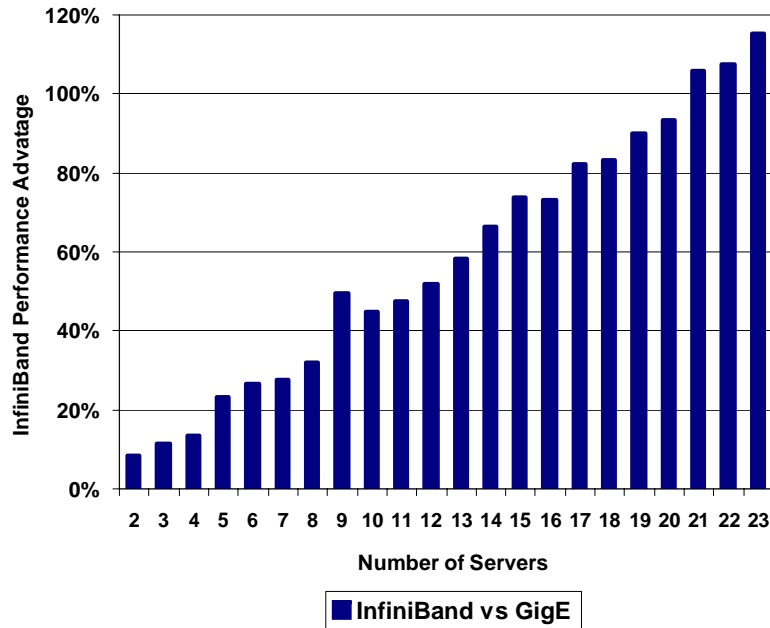


- **InfiniBand Architecture for highest reliability**
  - Reliable transport, HW data integrity, Link level flow control (no packet drops)
    - Two Cyclic Redundancy Checks (CRCs)
  - Congestion control, Automatic Path Migration (APM)
- **Mellanox products qualified for a minimum BER of  $10^{-15}$**
- **Product samples are tested for higher BER**
  - Require large clusters and longer testing times (in order of months)
- **Testing shows products meet BER of lower than  $10^{-17}$** 
  - Mellanox expects to confirm lower BER in the future
- **Full ECC protection throughout the HCAs and switches**

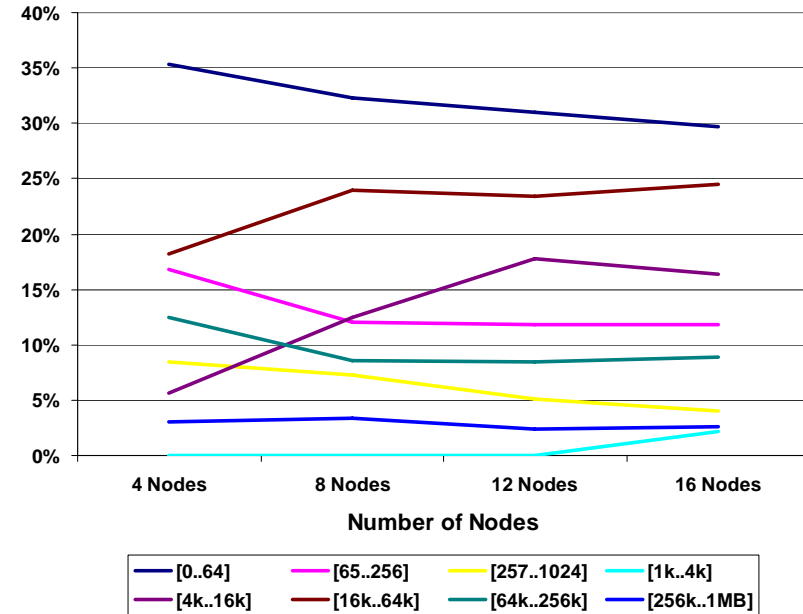
# HPC Applications Profiling - Weather



WRF Benchmark Results - InfiniBand vs GigE

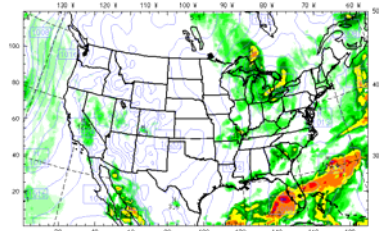


WRF MPI Profiling Message Distribution



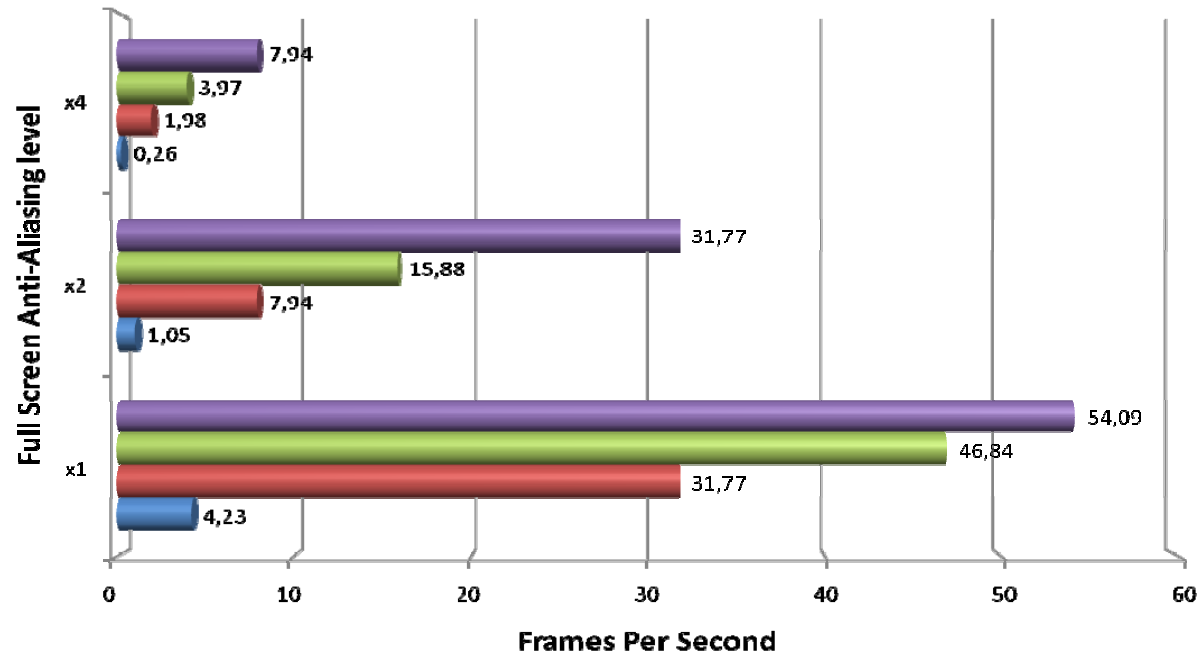
- Bottlenecks investigations
- Optimization in all layers
  - SW, middleware, HW
- Roadmap features definition

ARW WRF - 30KM-NEST - NCAR/MPPI  
 Exec = 19 h  
 Total precip. since h: 0  
 Total precip. since h: 0  
 Sep 08 11:20:00  
 Valid: 18 UTC Tue 30 Sep 08 11:20:00  
 Tue 30 Sep 08



# Direct Transport Compositor

Impact of the interconnect on the rendering speed (sort-last, 8 nodes, 1920x1200, local = 64 fps)



- QDR IB
- DDR IB
- SDR IB
- GigE



- 40Gb/s InfiniBand deliver up to 100% faster rendering
  - Ability to render 3D models at the required frame rates for high image quality

# 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

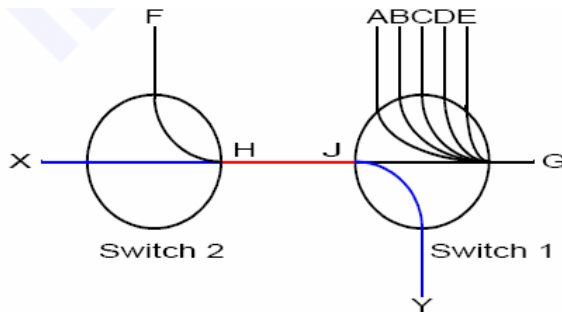


**Designed to handle all communications in HW**

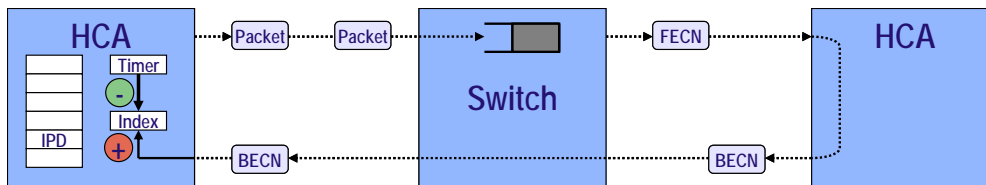
# Hardware Congestion Control



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



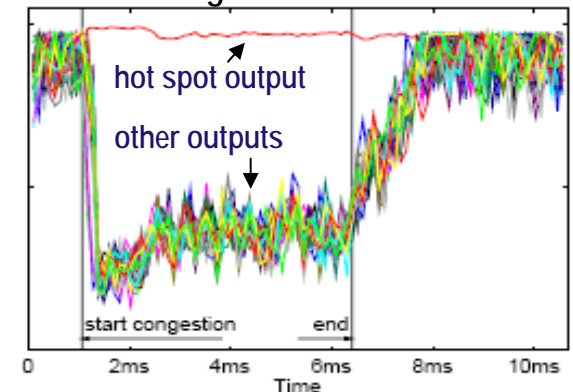
- InfiniBand HW congestion control
  - No a priori network assumptions needed
  - Automatic hot spots discovery
  - Data traffics adjustments
  - No bandwidth oscillation or other stability side effects
  - SM receives notices of congestion
- Ensures maximum effective bandwidth



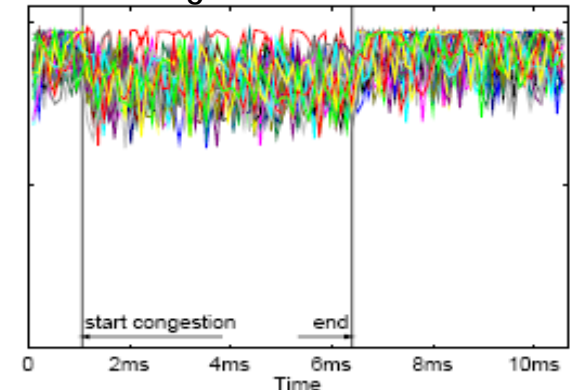
## Simulation results

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
  - $3 \times 10^{-5}$  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 Advisory Council



- 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
- More than 60 members
- A community effort support center for HPC end-users
  - Cluster Centers
  - End- user support center
- For details – [HPC@mellanox.com](mailto:HPC@mellanox.com), [HPCAdvisoryCouncil.mellanox.com](http://HPCAdvisoryCouncil.mellanox.com)
  - Participate in the council BoF session – Thu, 12:15pm, room 18A/18B/18C/18D

# Thank You

Participate in the council BoF session – Thu, 12:15pm, room 18A/18B/18C/18D

Join Mellanox Wednesday night event – visit Mellanox booth #343 for details



Join the HPC Advisory Council  
[HPC@mellanox.com](mailto:HPC@mellanox.com)