

# High Speed Networking Present and Futures

HUO Zhigang, The National Research Center for Intelligent Computing Systems (NCIC)  
Gilad Shainer, HPC Advisory Council and Mellanox Technologies



HPC Advisory Council Workshop

October 28th, 2009

Changsha, Hunan, China

# Exascale Networking Requirements

**High Throughput  
Low Latency**

**Enhanced  
Scalability**

**Network  
Reliability**

**Transport  
Offload**



**Congestion  
Avoidance**

**Network  
Adaptation**

**Advanced  
Quality of Service**

**Multiple Network  
Topologies**

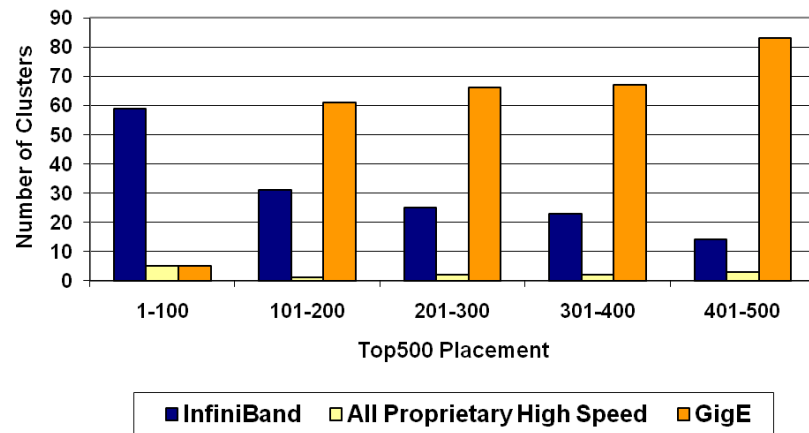
# HPC Interconnect Technologies

	Standard	Software Development	Technology Pace	Main Focus	Market Adoption	Performance
Quadrics	No	Close	Medium	HPC	Mid-range High-end	High
Myricom	No	Close	Medium	HPC	Mid-range	High
Cray	No	Close	Medium	HPC	High-end	High
Ethernet	Yes	Open	Slow	EDC	Low-end Mid-range	Low-Mid
InfiniBand	Yes	Open	Fast	HPC/EDC	Low-end Mid-range High-end	High

- InfiniBand is clearly becoming the de-facto interconnect technology for HPC
- 25% increase in InfiniBand systems on the Top500 list between June 2008 and June 2009



Top500 Interconnect Placement



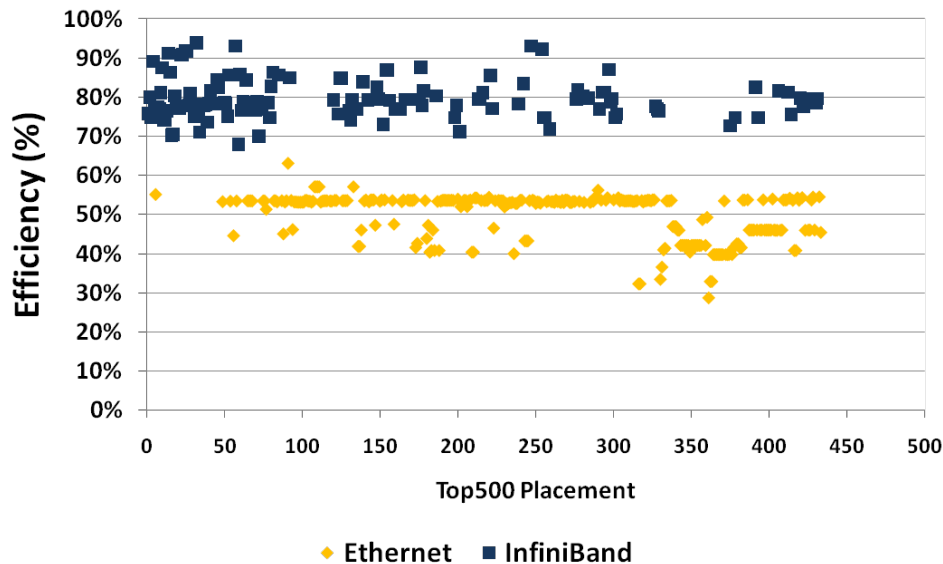
# InfiniBand Technology

- **Industry Standard**
  - Hardware, software, cabling, management
  - Design for clustering and storage interconnect
- **Performance**
  - 40Gb/s node-to-node
  - 1us application latency
  - Most aggressive roadmap in the industry
- **Reliability and efficiency**
  - RDMA, Transport Offload and Kernel bypass
  - CPU focuses on application processing
- **Scalability for Petascale computing**
  - Congestion management
  - End-to-end quality of service
- **Virtualization acceleration**
  - SRIOV, vSwitch offloads
- **I/O consolidation**
  - IPC, storage and management
  - FCoIB, iSCSI, IPoIB, Ethernet over IB
  - High availability and self healing resilience

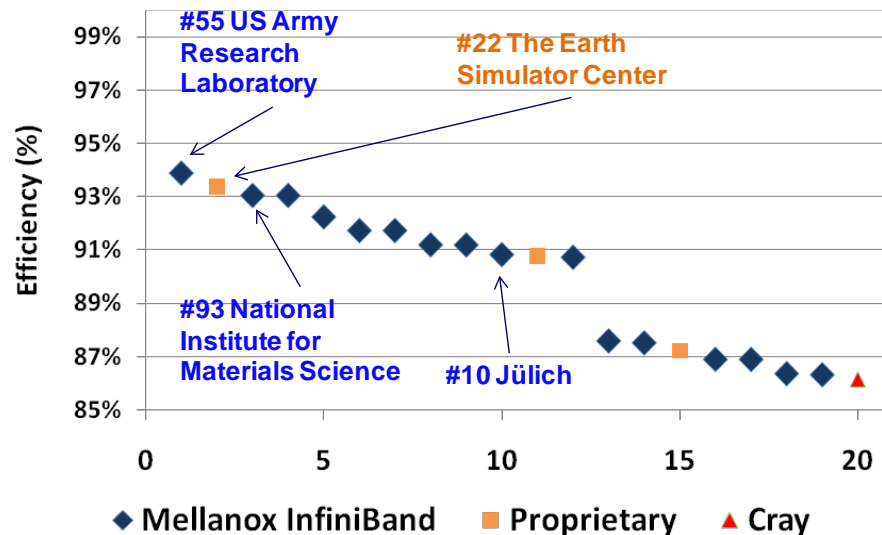


# InfiniBand Unsurpassed System Efficiency

## Top500 Efficiency Comparison



## The 20 Most Efficient Top500 Systems



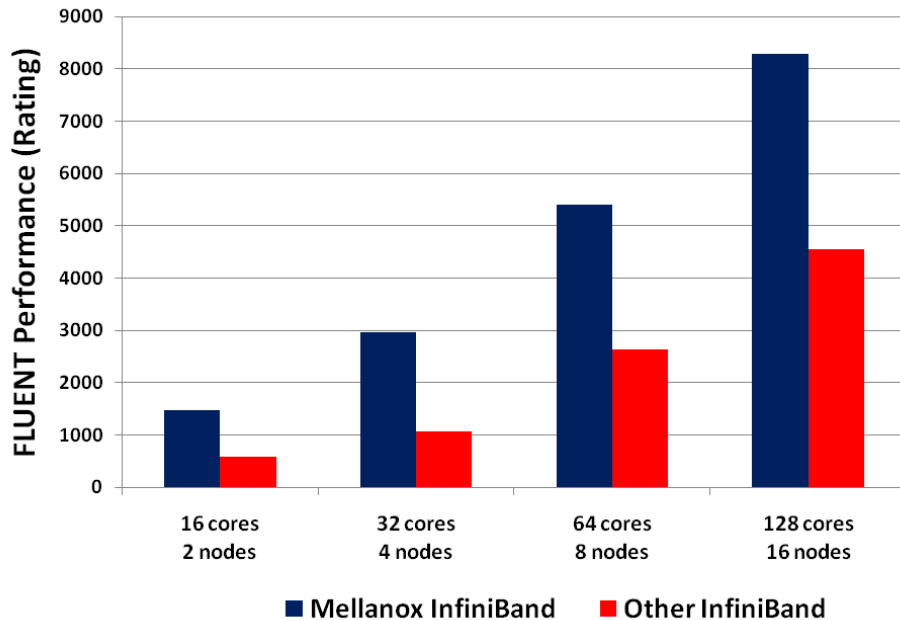
- Top500 systems listed according to their efficiency
- InfiniBand is the key element responsible for the highest systems efficiency
- The only standard interconnect solution in the top 100 highest utilization systems
- All InfiniBand system use Mellanox InfiniBand solutions

# Mellanox Technologies InfiniBand Leadership

- **Highest performance and efficiency**
  - 40Gb/s bandwidth, 1usec latency end-to-end latency
  - 100nsec switch latency (40% faster than any other switch)
  - 50M MPI messages per second
  - Highest dense switches – up to 51.8TB in a single non-blocking enclosure
- **The ONLY full hardware offload solution**
  - Only solution to provide transport offload, RDMA and zero-copy
  - Ensures highest CPU efficiency and applications availability
- **Only proven scalability for 10s and 100s of thousands of nodes**
  - LANL Roadrunner 1PFlops system, Jülich Germany, Shanghai Supercomputing center
- **Highest reliability for large scale systems**
  - Only solution with hardware based reliability and failover
- **Ease-of-use and automatic installation and troubleshooting**
  - Full fabric management, performance optimizations, systems design support
- **One stop shop for end-to-end InfiniBand infrastructure**
  - From silicon to Adapters, switches, cables, software
  - Certified by all tier-1 OEMs, ISV, OSV and system integrators

# Mellanox InfiniBand Advantage

ANSYS FLUENT Aircraft\_2M Performance

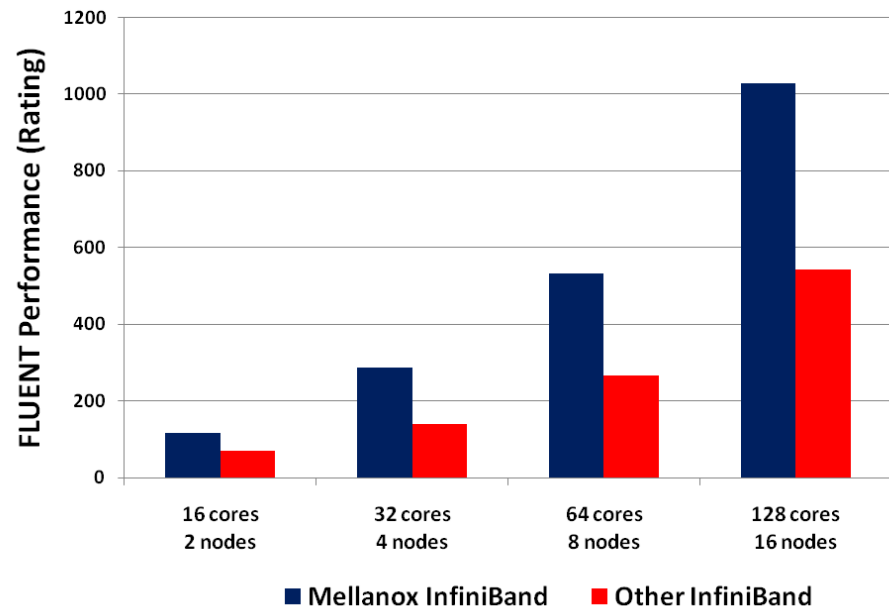


- Mellanox InfiniBand transport offload enables the highest CPU efficiency and availability for applications
- Results in highest application performance and cluster reliability

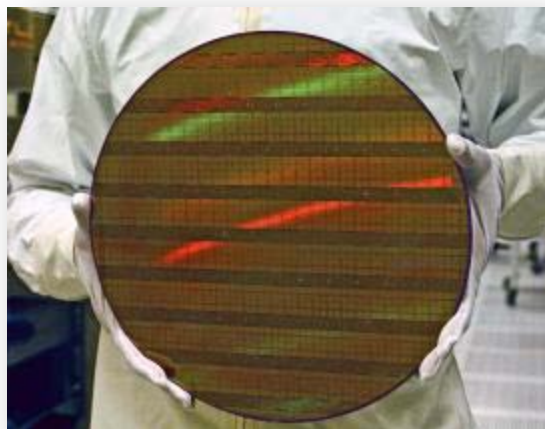
- Application performance example – ANSYS FLUNET CFD application
- Same behavior expected/seen in other application cases



ANSYS FLUENT Truck\_14M Performance



# The New HPC Roadblocks



## Multi-core CPU/GPU – Interconnect Ratio

- Multi-core: Intend to reduce power consumption
- Require a changed the used ratio of server to Interconnect bandwidth
- 40Gb/s networking not enough for HPC applications any more!
- Interconnect needs to provide 4x improvements between different platforms generations



## CPU/GPU/System Efficiencies

- CPU/system efficiency **DEPENDS** on the networking offloads
- Offloads options: transport, applications
- Applications complexity, increase in simulation capability and HPC systems size demands advanced applications offloading
- New networking topologies will effect systems efficiency as well

# The HPC Evolution

- Supercomputers are growing at an unrelenting rate
  - To meet the needs of scientific research
- As supercomputers increase in size from mere thousands to hundreds-of-thousands of processor cores, new performance and scalability challenges have emerged
- Past performance tuning of a parallel application by using algorithms optimizations has reach its limit
- New technologies/approaches are needed to address efficiency and scalability for future large scale machines

# Collective Communications - Overview

- **Collective communications are frequently used by scientific simulation**
  - **Examples**
    - Broadcast to send around initial input data
    - Reduction in operations to consolidate data from multiple sources
    - Barriers for global synchronization
- **Collective operations couple all processes in a system**
  - Tend to be the part of the simulations that most negatively impact application scalability
- **The communication coupling used in HPC implementations of collective algorithms tends to magnify the effects of system-noise on application performance**
  - Estimations shows 20-30% scalability loss

# Mellanox/ORNL Applications Offloads Technology

- **US Department of Energy (DOE) funded project – ORNL and Mellanox**
  - Mellanox - InfiniBand hardware solutions, ORNL – MPI software
- **Mellanox new ConnectX-2 HPC adapters addresses collective communication scalability by offloading it**
  - Will be announced at SC09 (November)
- **Offloading collectives operation**
  - Eliminate system noise and jitter issue
  - Increase the CPU cycles available for applications
  - Allow the processes communications to progress asynchronously
- **ConnectX-2 also includes a floating point operation unit**
  - Enables offloading collectives operations data manipulation as well
  - Provide full offload capability
- **ConnectX-2 provides build-in support for future non-blocking collective operations as defined by the MPI-3 specifications**
  - Allowing overlap collective operations with application computation

# Collectives Offloads – Initial Performance

## ■ First testing results

- Present the overlapping benefit of collective offloads

## ■ Non-blocking collective implementation

## ■ Benchmark: Non-blocking barrier

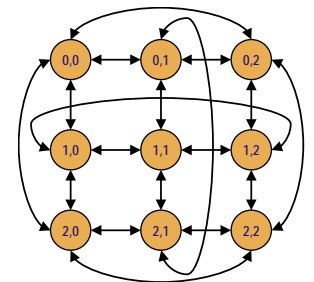
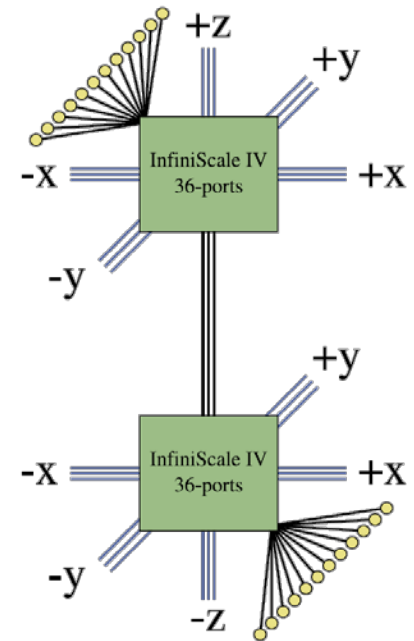
- Initiate non-blocking MPI barrier
- CPU to performance applications calculations
- Wait for non blocking barrier to complete

## ■ Results

- Using collective offload – latency 9.75us
- Using pt2pt collective – latency 12.89us
- **25% reduction in applications run time!**
- 2 node system, gap can be bigger in larger systems!

# Fastest InfiniBand Solutions - Mellanox 120G IB

- **InfiniBand 120G switch family – 120Gb/s**
  - 3x bandwidth increase versus current solutions
  - 12-ports, 36-ports, 72-ports
  - Hybrid 120G and 40G – 18 ports 40G and 6-ports 120G
- **Delivering highest throughput for switch connectivity**
  - 50% higher bandwidth per port versus other solutions
- **Reducing network congestion and enhancing efficiency**
- **Reducing the number of cables by a factor of 3**
- **Providing optimized switch solutions for 3D-Torus systems**
  - Supports multiple server nodes per cube junction
  - Enable building systems at scale with smallest 3D cube size
  - Lowest latency between remote nodes
  - Built in support for adaptive routing, congestion control, QoS



# Mellanox End-to-end For PetaScale and Beyond

- Highest performance - highest throughput, lowest latency
- HPC applications offload - ensures highest efficiency
- Self recovery - ensures highest reliability
- Scalability - the solution for Peta/Exa flops systems
- On-demand resources - allocation per demand
- Support for Fat-tree, mesh and 3D-Torus system architectures
- The only solution to provide hardware based adaptive routing and congestion control
- Green HPC – lowest system power consumption
- Future proof and the most aggressive roadmap

