

Mellanox InfiniBand QDR 40Gb/s

## The Fabric of Choice for High Performance Computing

Gilad Shainer, [shainer@mellanox.com](mailto:shainer@mellanox.com)

June 2008

### *Birds of a Feather Presentation*



INTERNATIONAL  
SUPERCOMPUTING CONFERENCE

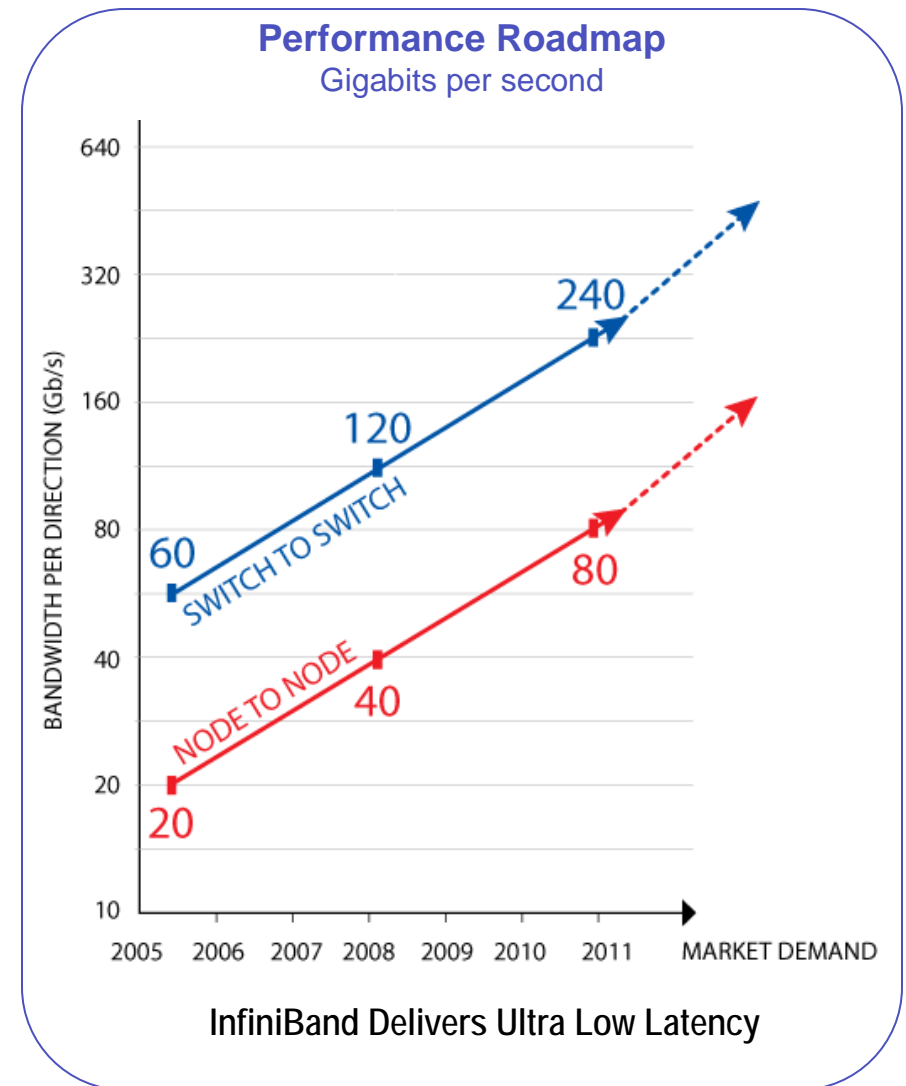
'08



# InfiniBand Technology Leadership



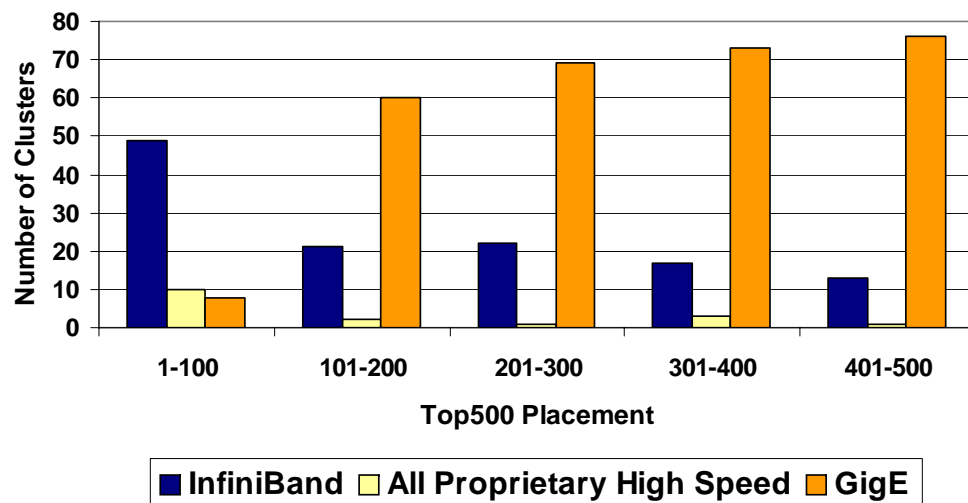
- **Industry Standard**
  - Hardware, software, cabling, management
  - Design for clustering and storage interconnect
- **Price and Performance**
  - 40Gb/s node-to-node
  - 120Gb/s switch-to-switch
  - 1us application latency
  - Most aggressive roadmap in the industry
- **Reliable with congestion 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**



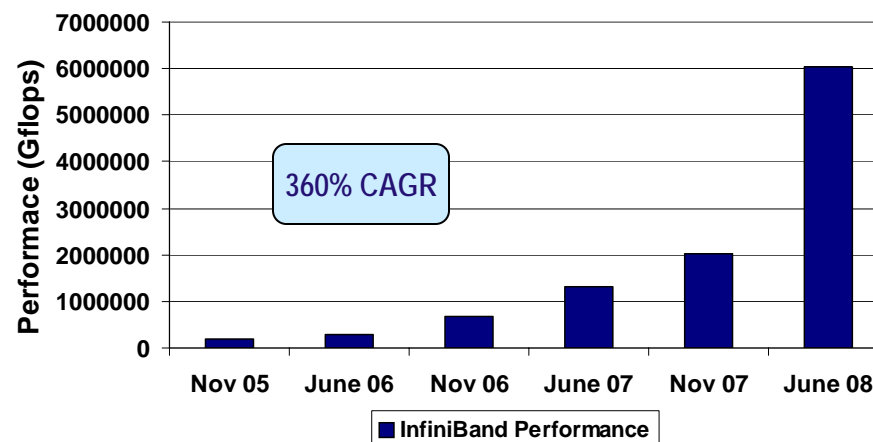
# InfiniBand in the TOP500



### Top500 Interconnect Placement

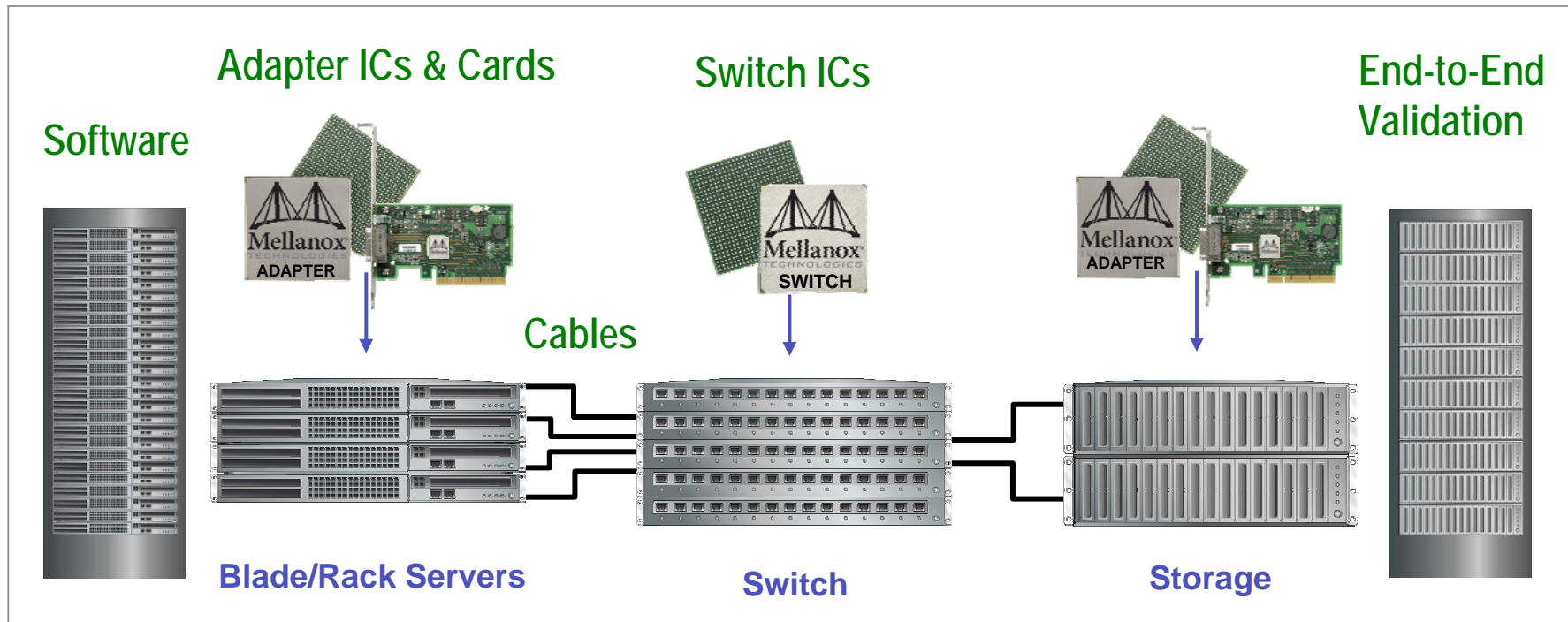


### InfiniBand Clusters - Performance



- **InfiniBand makes the most powerful clusters**
  - 5 of the top 10 (#1, #4, #7, #8, #10) and 49 of the Top100
  - The leading interconnect for the Top200
  - InfiniBand clusters 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**

# Mellanox InfiniBand End-to-End Products



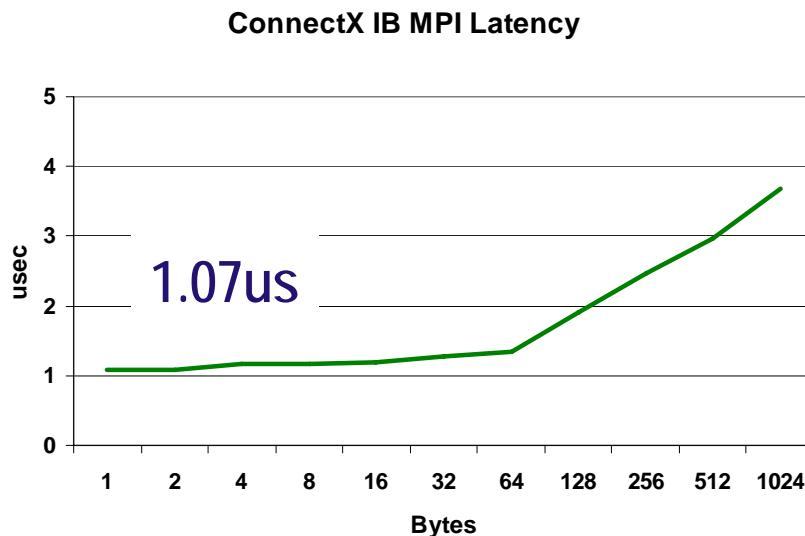
- High Throughput - 40Gb/s
- Low latency - 1us
- Low CPU overhead
- Kernel bypass
- Remote DMA (RDMA)
- Reliability

**Maximum Productivity**

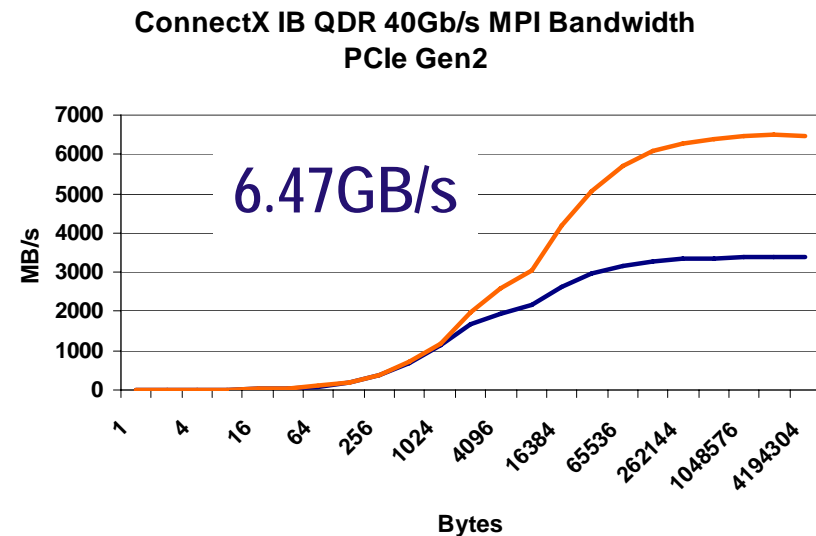
# ConnectX - Fastest InfiniBand Technology



- 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

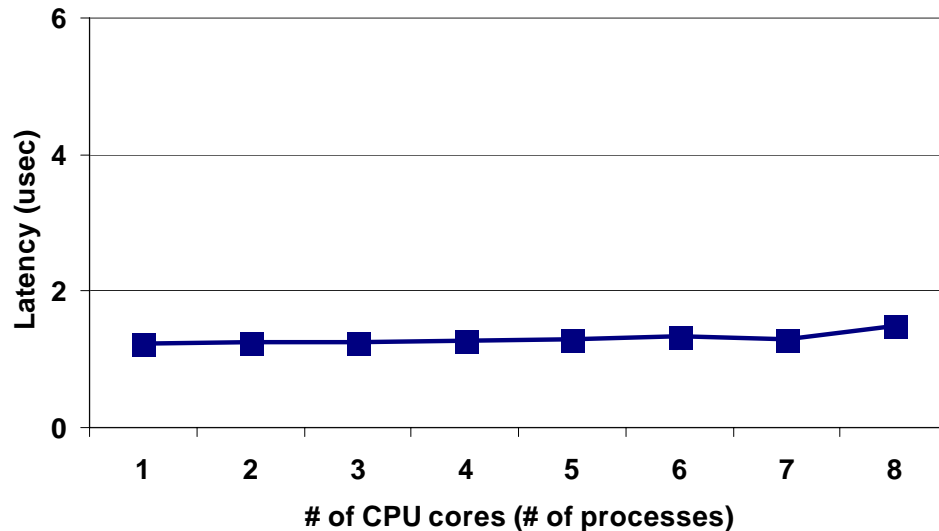


— IB QDR Uni-dir — IB QDR Bi-dir

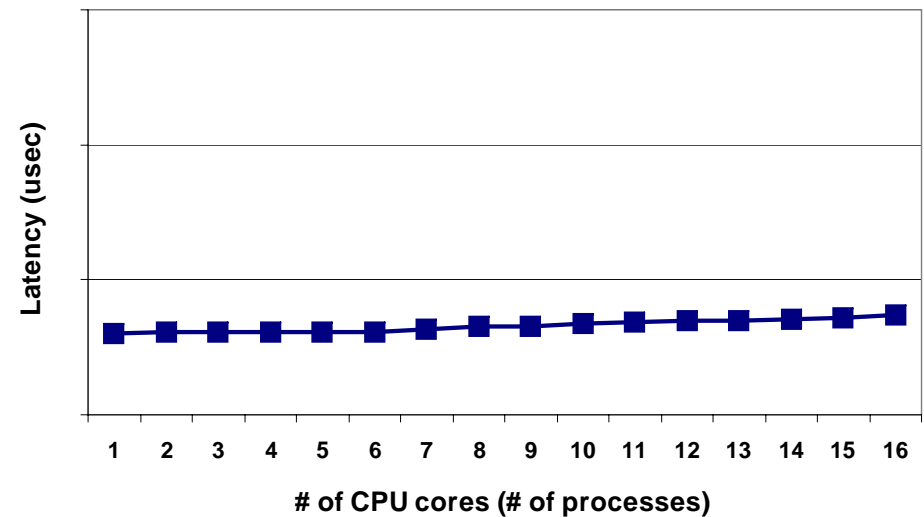
# ConnectX Multi-core MPI Scalability



Mellanox ConnectX  
MPI Latency - Multi-core Scaling



Mellanox ConnectX  
MPI Latency - Multi-core Scaling



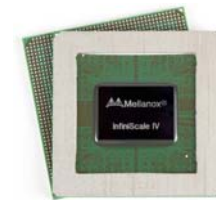
- Scalability to 64+ cores per node, to 20K+ nodes per subnet
- Guarantees same low latency regardless of the number of cores
- Guarantees linear scalability for real applications



# InfiniScale IV Switch: Unprecedented Scalability



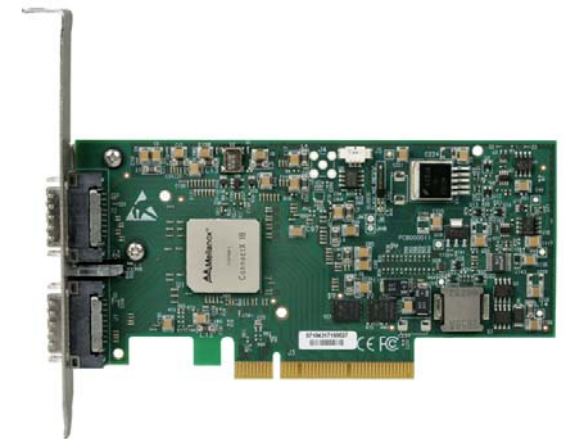
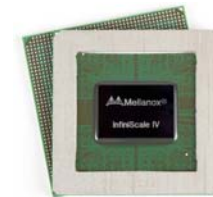
- **36 40Gb/s or 12 120Gb/s InfiniBand Ports**
  - Adaptive routing and congestion control
  - Virtual Subnet Partitioning
- **6X switching and data capacity**
  - Vs. using 24-port 10GigE Ethernet switch devices
- **4X storage I/O throughput**
  - Critical for backup, snapshot and quickly loading large 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



# 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

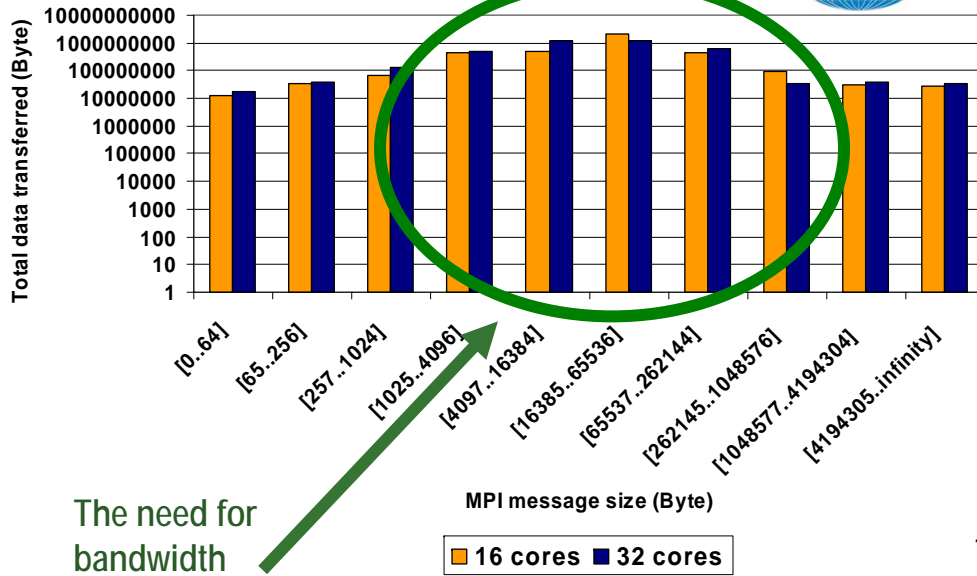


**Designed to handle all communications in HW**

# HPC Applications Demand Highest Throughput

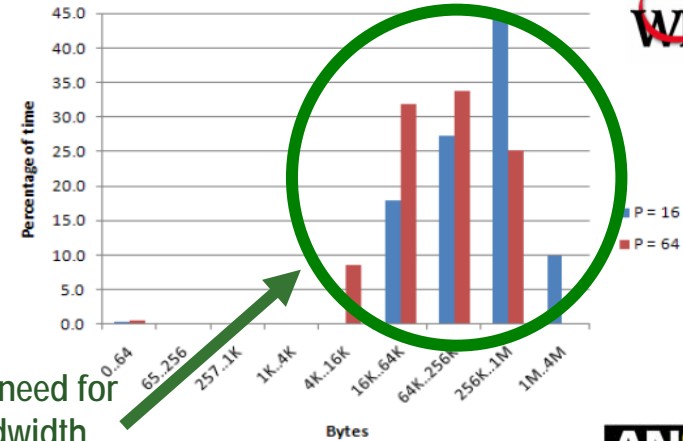


LS-DYNA Profiling



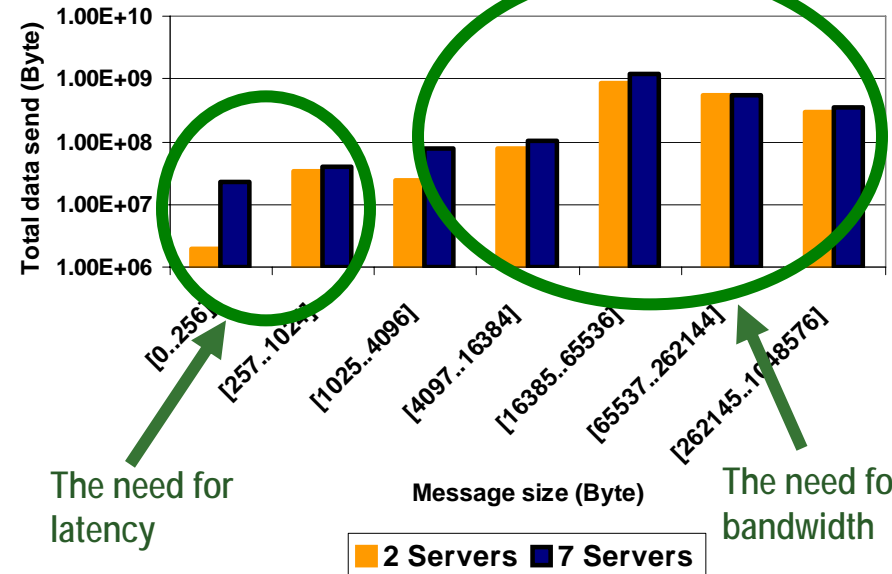
The need for bandwidth

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

# 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](mailto:HPCHelp@mellanox.com)
- For details – [HPC@mellanox.com](mailto:HPC@mellanox.com)

*Providing advanced, powerful, and stable  
high performance computing solutions*

