

Driving InfiniBand to Petascale Computing and Beyond

Dror Goldenberg, VP Architecture
Gilad Shainer, Senior Director, HPC and Technical Computing

HPC Cetraro, June 21st, 2010



■ Leading connectivity solutions provider for data center servers and storage systems

- Foundation for the world's most powerful and energy-efficient systems
- >5.8M ports shipped as of March '10

■ Company headquarters:

- Yokneam, Israel; Sunnyvale, California
- 375+ employees; worldwide sales & support

■ Solid financial position

- Record Revenue in Q1'10; \$36.2M
- Record Revenue in FY'09; \$116.0M
- \$220.6M cash / no debt

Recent Awards



Connectivity Solutions for Efficient Computing



Enterprise HPC



High-end HPC



HPC Clouds

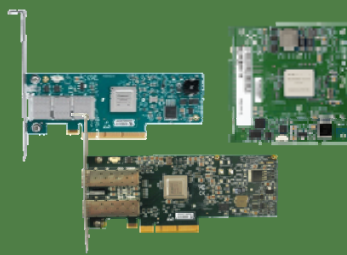


Mellanox Interconnect Networking Solutions

ICs



Adapter Cards



Host/Fabric Software



Switches/Gateways



Cables



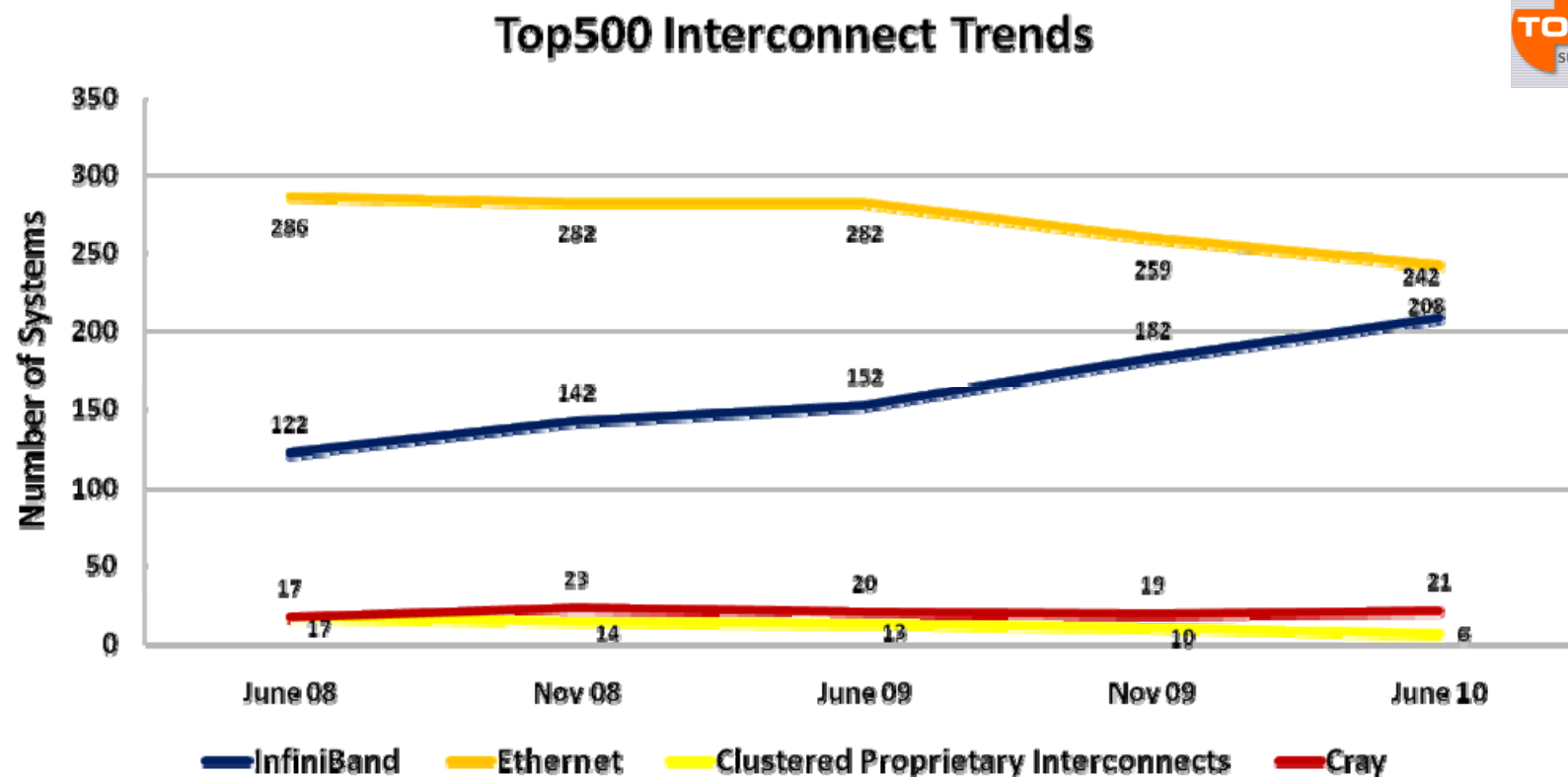
Connecting World Leading Large-Scale Systems



- **Mellanox InfiniBand solutions are proven for Petascale computing**
 - Connecting the first Petaflop clusters in the world
- **Connecting the leading supercomputing in the world**
 - 5 systems from the world Top10 systems, 63 out of the Top100
 - Fat-tree, 4K nodes, 130K cores - LANL "Roadrunner"
 - Fat-tree, 9.2K nodes, 82K cores - NASA
 - Fat-tree, 3K nodes, 72K cores – NUDT "TianHe"
 - Fat-tree, 3K nodes, 30K cores - Jülich
 - Fat-tree 4K nodes, 63K cores – TACC
 - 3D-Torus, 5.4K nodes, 43K cores – Sandia "Red Sky"
- **Connecting the world's Top10 systems since 2003**



Interconnect Trends – Top500

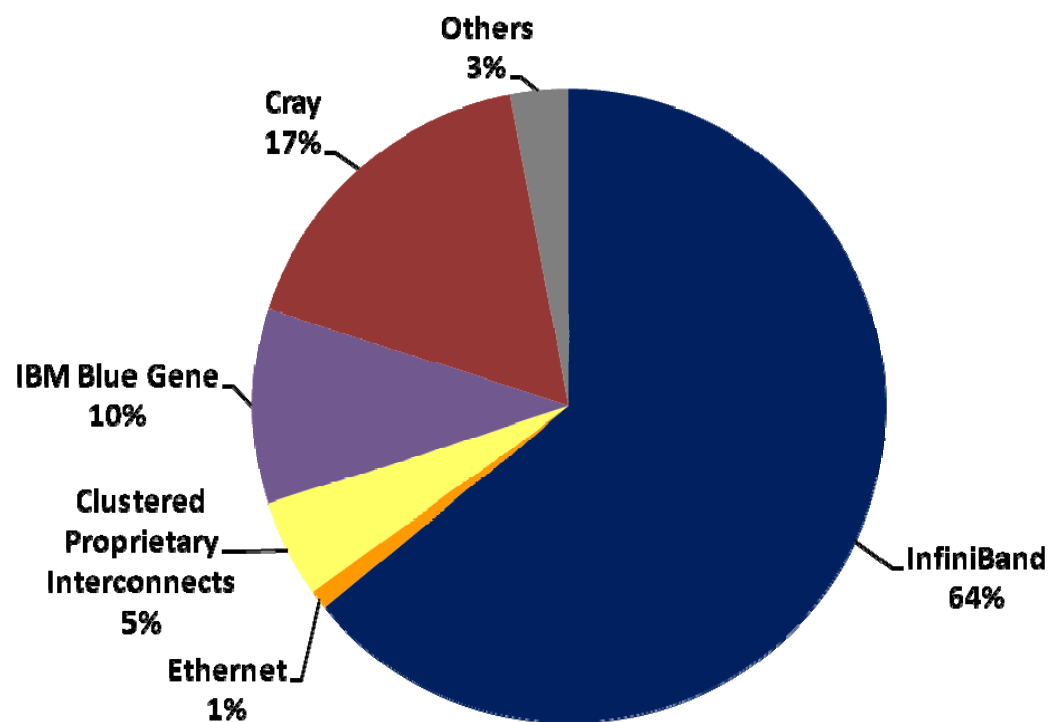


- InfiniBand is the only growing high speed clustering interconnect
 - 208 systems on the June 10 list, 37% increase since June 2009
- InfiniBand is the HPC interconnect of choice
 - Connecting 41.6% of the Top500 systems

Interconnect Trends – Top100



Top100 Systems

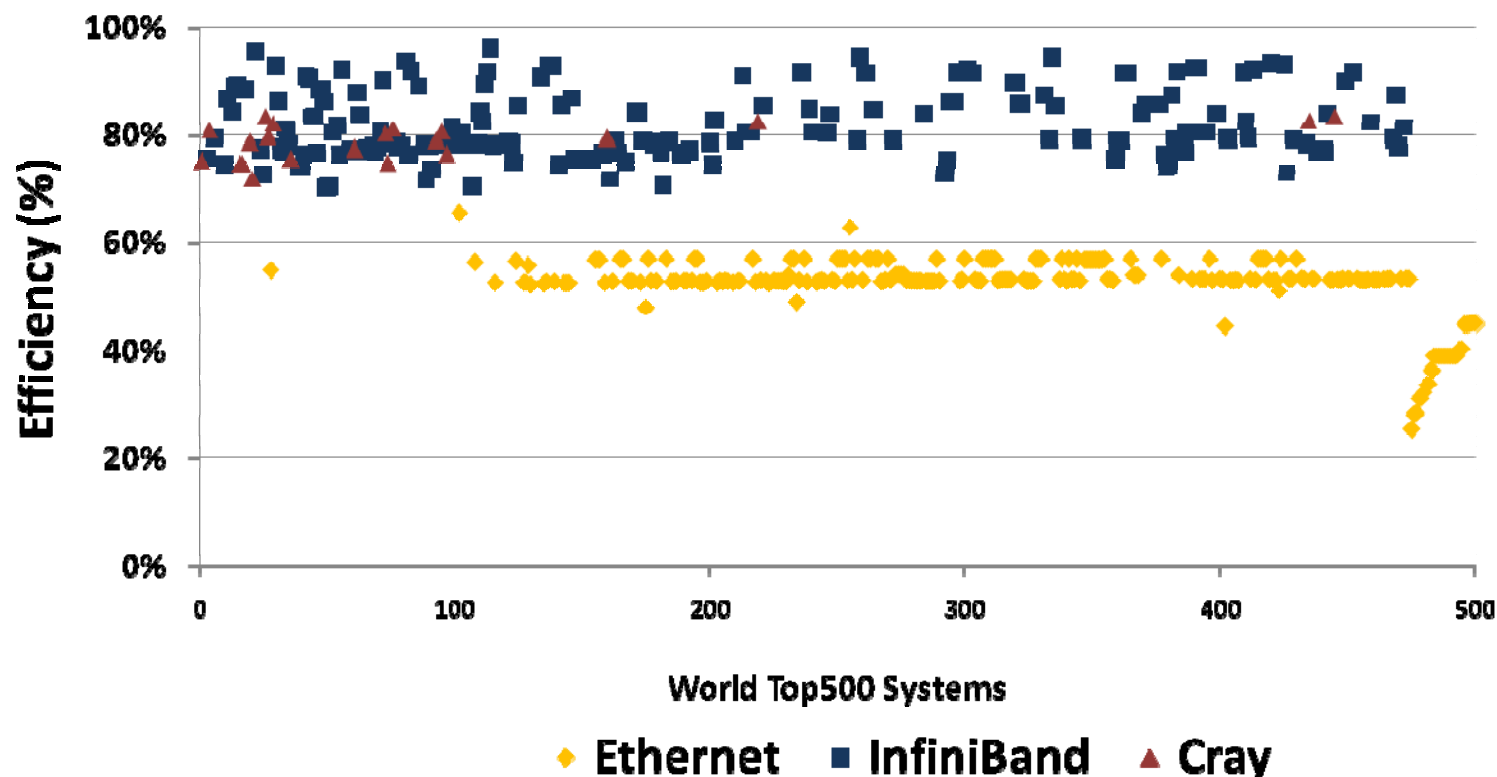


- Mellanox InfiniBand builds the most powerful supercomputers
 - 5 of the Top10 (#2, #3, #6, #7, #10) and 64 of the Top100
- The natural choice for world leading supercomputers
 - Performance, Efficiency, Scalability

InfiniBand Unsurpassed System Efficiency



World Leading Compute Systems Efficiency Comparison



- Top500 systems listed according to their efficiency
- InfiniBand is the key element responsible for the highest systems efficiency

High Performance Computing Challenges



Performance

- Latency
- Throughput
- Message Rate



Scalability

- Many Cores (CPU/GPU)
- Many Nodes
- Distributed Computing (cloud)



Reliability

- Failover
- Redundancy
- Up time



Efficiency

- Core Availability
- Effective Flops
- Power/Flop



Mellanox HPC Technology Solutions



Performance

- <1usec Latency
- 40Gb/s bandwidth
- 50M MPI msg/sec
- RDMA



Scalability

- CORE-Direct
- GPUDirect
- Transport Offload
- Topologies (Fat-Tree, 3D Torus etc.)



Reliability

- Auto Negotiation
- Automatic Failover
- High Availability
- Lowest Bit Error Rate (BER)



Efficiency

- Congestion Control
- Adaptive Routing
- Low Power
- Virtual Protocol Interconnect (VPI)



- Congestion control
 - Eliminates network congestions (hot-spots)
 - Related to many senders and a single receiver
- Adaptive routing
 - Eliminates network congestions
 - Related to point-to-point communications sharing the same network path
- GPU-Direct
- CORE-Direct (Collectives Offload Resource Engine)

CORE-Direct

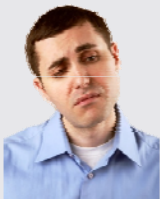
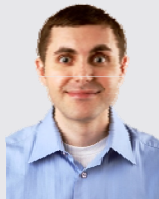


- **Collectives communications have a crucial impact on the application's scalability and performance**
 - Used for sending around initial input data
 - Reductions for consolidating data from multiple sources
 - Barriers for global synchronization
- **Every collective communication executes global communication operation by coupling all processes in a given group**
- **Collectives operations**
 - Must be executed as fast as possible
 - Each local node delay will impact the entire cluster performance
 - Consume high percentage of CPU cycles

Efficient Execution of Collectives Operation



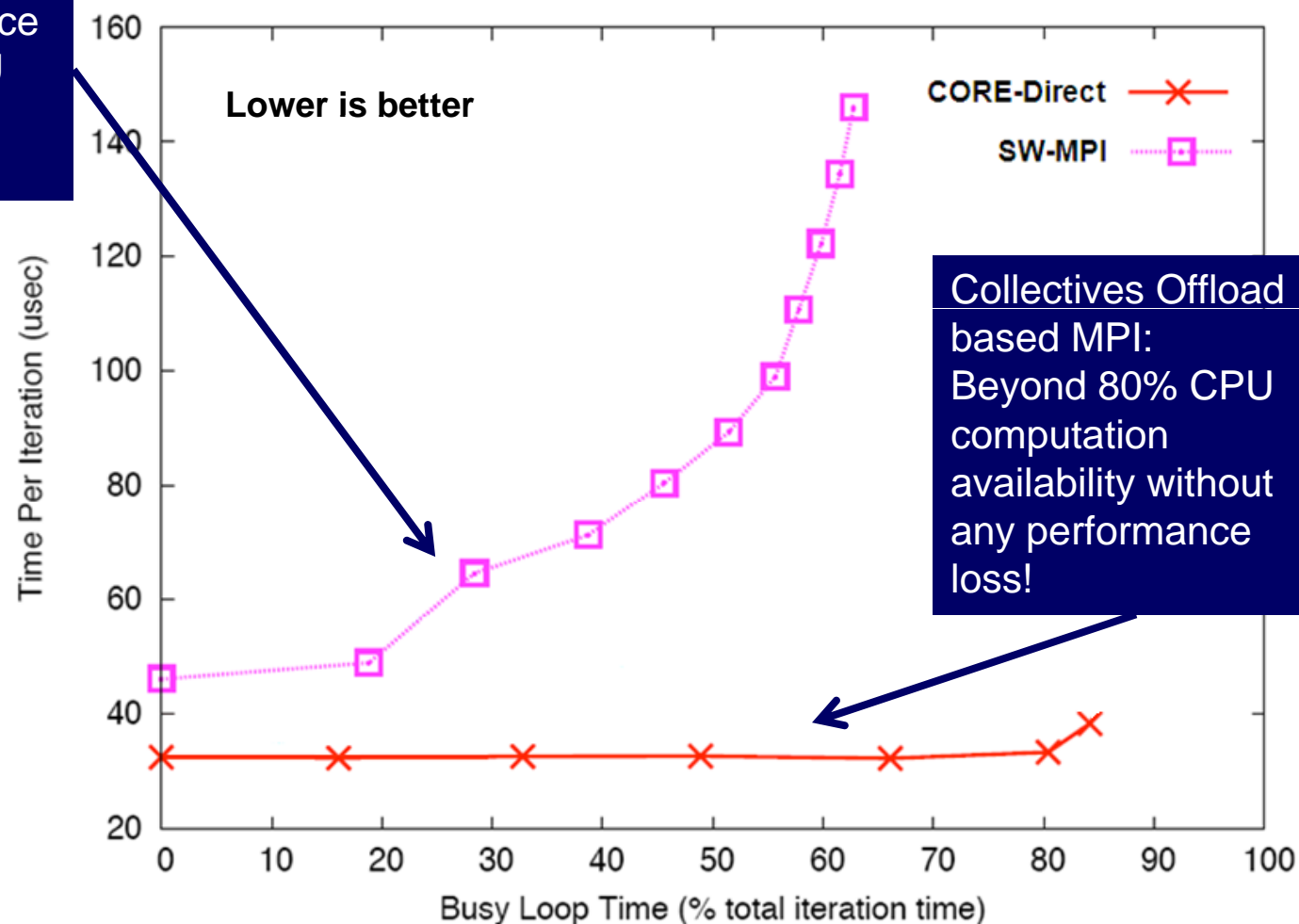
- Server components capable of managing operations – CPU, interconnect

	CPU Executes Collectives	Interconnect Executes Collectives
Fast propagation throughout the system	Fast	Fastest
Negative effect of a single node on the entire system (system noise/jitter)	Maximizing the effect	Minimizing the effect
Reducing CPU overhead and maximizing CPU availability for the application	Maximum CPU overhead	Minimum CPU overhead, allowing overlap between computations and communications
Best place for executing collectives operations		

CORE-Direct Performance – CPU Availability



Software MPI:
Losing performance
beyond 20% CPU
computation
availability



Collectives Offload
based MPI:
Beyond 80% CPU
availability without
any performance
loss!

GPUDirect

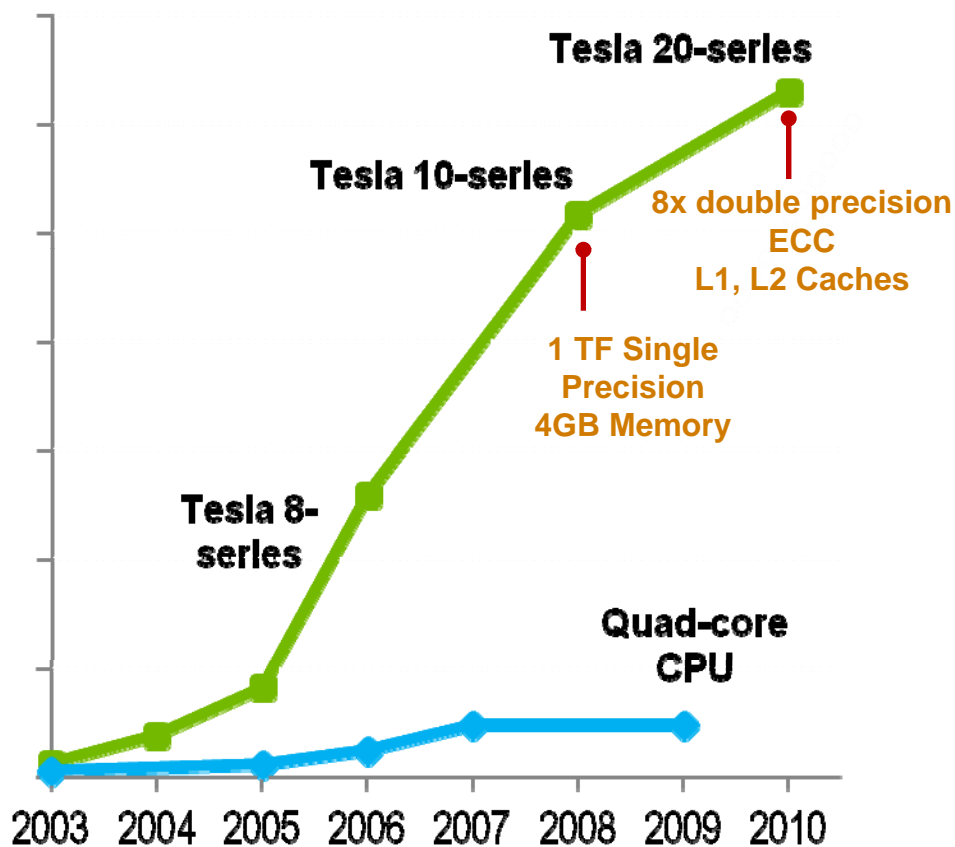


GPU Performance (NVIDIA)



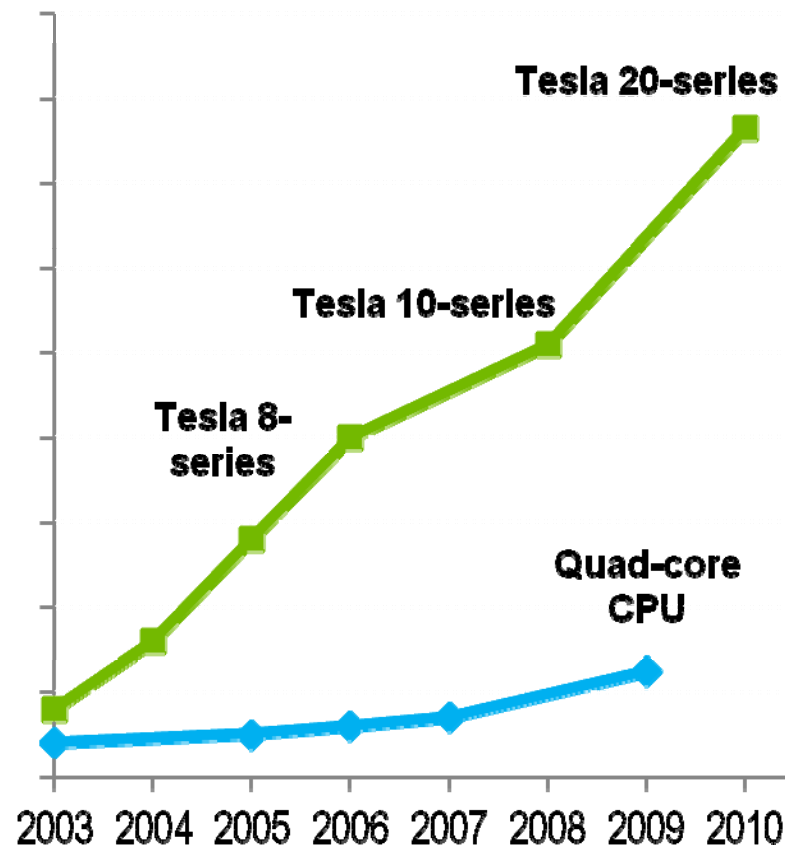
Peak Single Precision Performance

GFlops/sec



Peak Memory Bandwidth

GB/sec



■ NVIDIA GPU
◆ X86 CPU

GPU – InfiniBand Based Supercomputers



- **Cost/effective supercomputers architecture**
 - Lower system cost, less space, lower power/cooling costs

Mellanox IB – GPU Supercomputer

**National Supercomputing
Centre in Shenzhen (NSCS)**



5K end-points (nodes)

Proprietary Supercomputer

**Jaguar
Oak Ridge National Lab**

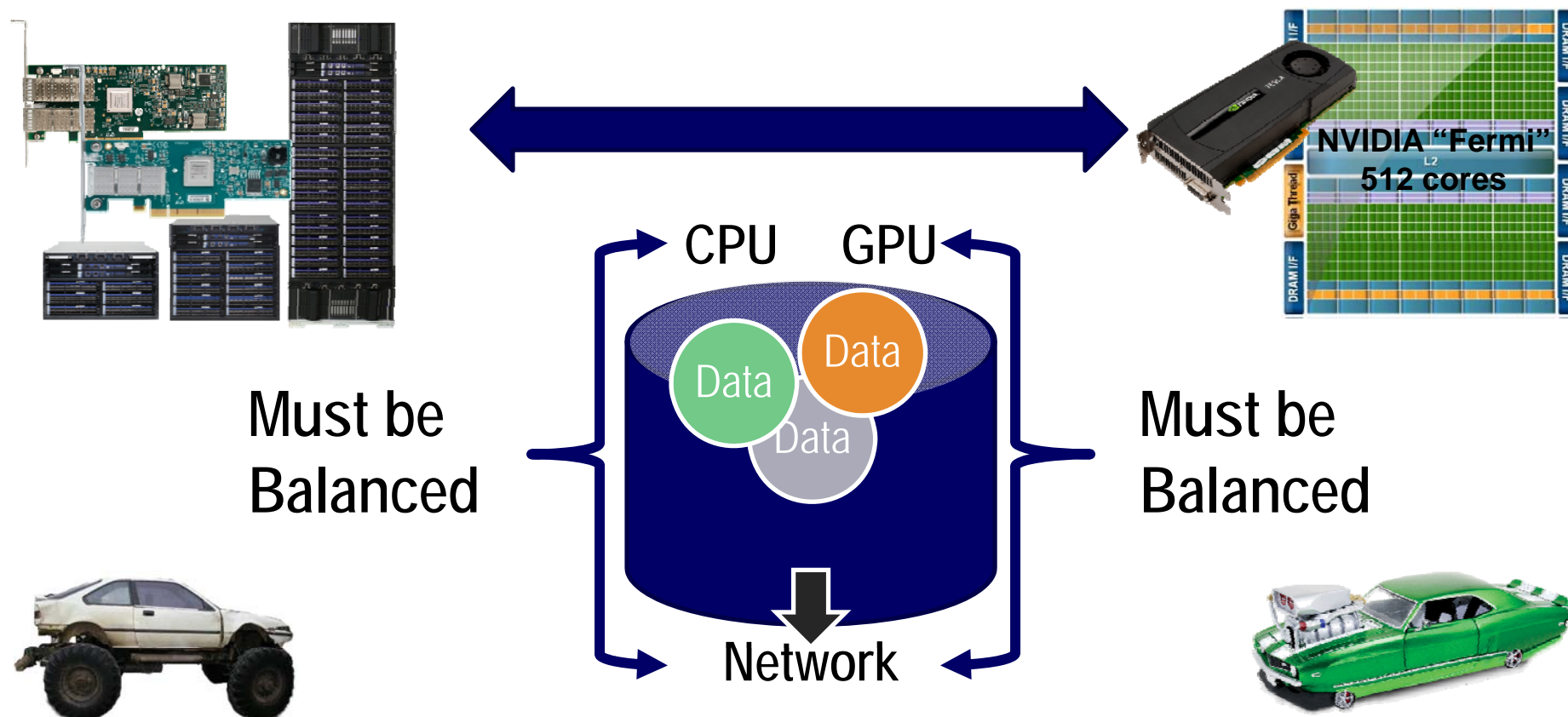


20K end-points (nodes)

GPUs Based Clustering



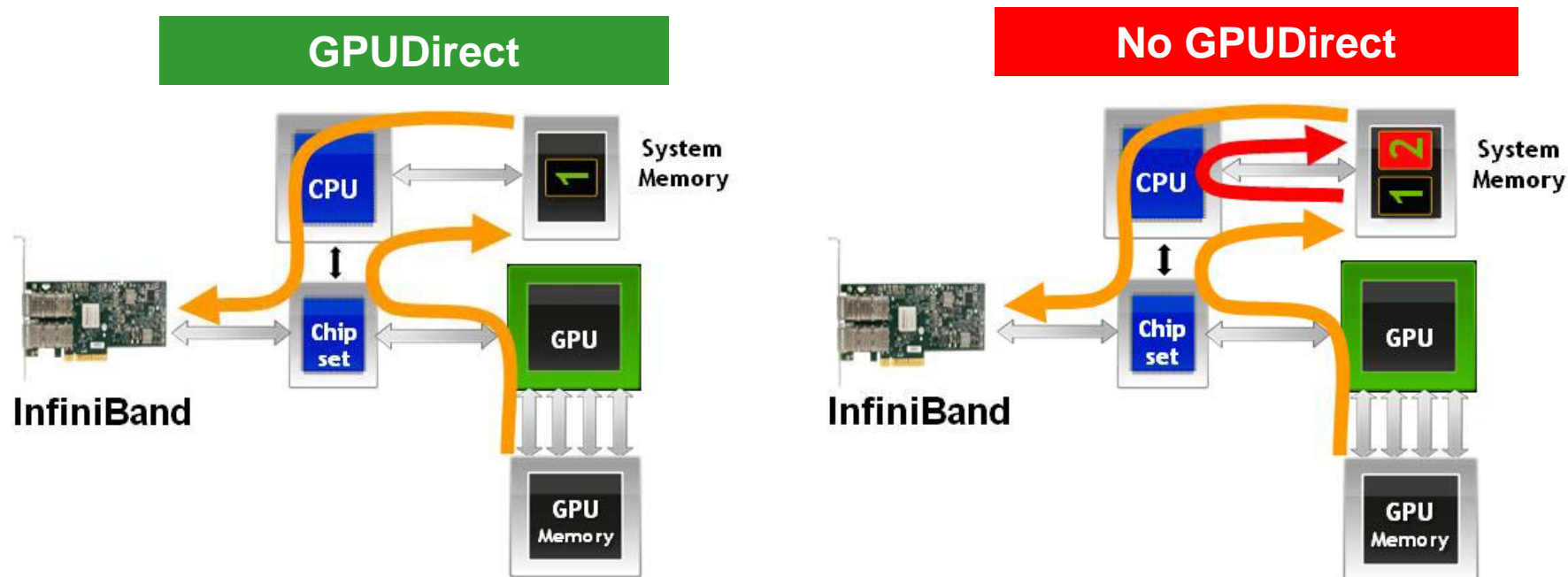
- GPUs introduce high load on the cluster communication
 - Highest speed interconnect is a must – Mellanox InfiniBand



Mellanox – NVIDIA GPUDirect Technology



- **Allows Mellanox InfiniBand and NVIDIA GPU to communicate faster**
 - Eliminates memory copies between InfiniBand and GPU
 - Reduces latency by 30% for GPUs communication
 - HPC applications with up to 42% performance increase

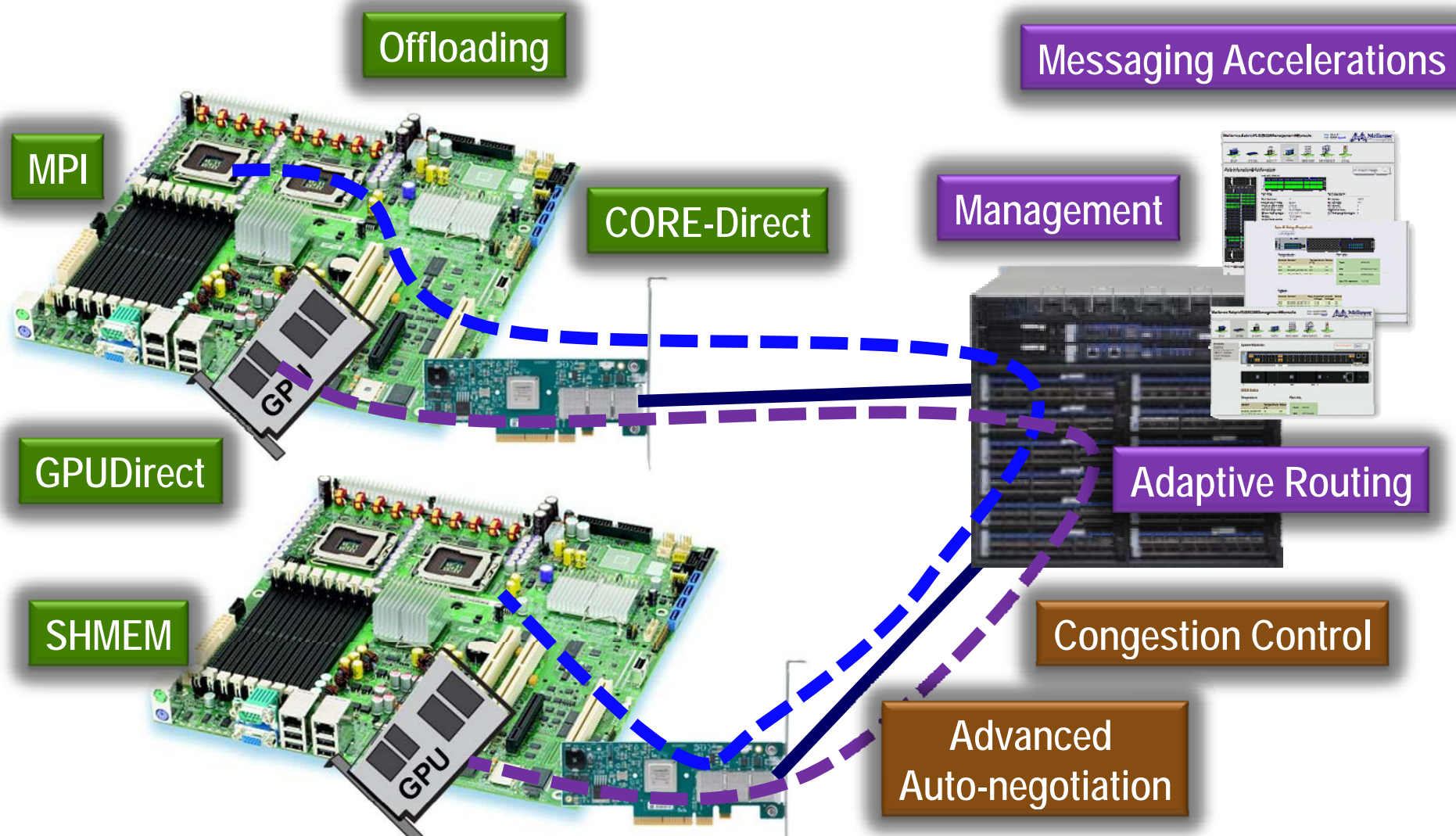


Mellanox-NVIDIA GPUDirect Enables Fastest GPU-to-GPU Communications

Paving The Road to Exascale Computing

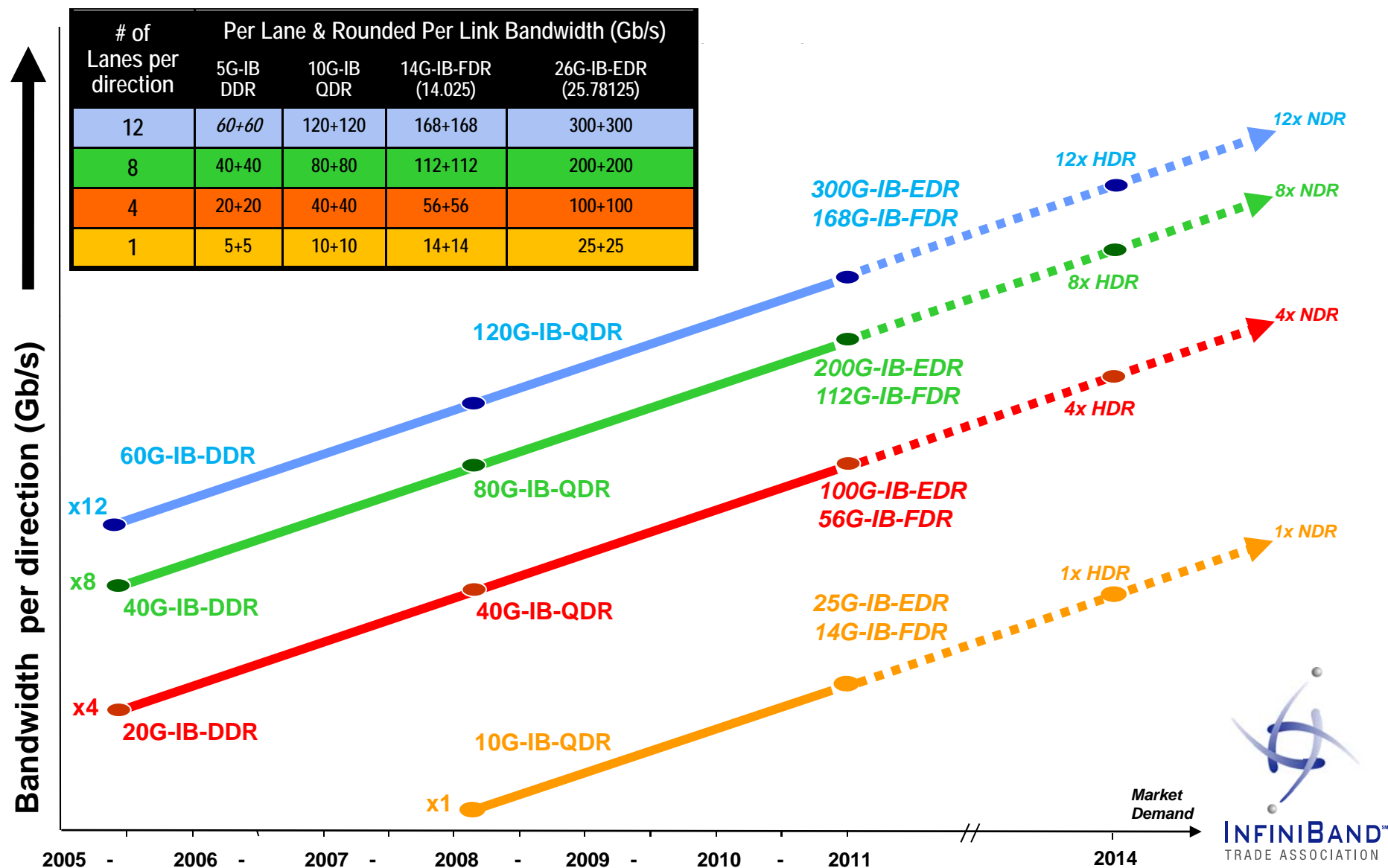


Bringing HPC to the Exascale



The Only Scalable Networking for Exascale HPC

InfiniBand Link Speed Roadmap



Mellanox Connectivity: Taking HPC to New Heights



World Highest Efficiency

- The world's only full transport-offload
- CORE-Direct - MPI and SHMEM offloads
- GPU-Direct - direct connectivity GPU-IB

World Fastest InfiniBand

- Lowest applications latency - 1us
- Highest dense switch solutions - 51.8TB in a single switch
- World's lowest switch latency at 100% load -100ns

HPC Topologies for Scale

- Fat-tree, mesh, 3D-Torus, Hybrid
- Advanced adaptive routing capabilities
- Highest reliability, lowest bit error rate, real-time adjustments



Paving The Road to Exascale

HPC@mellanox.com

