The Road to ExaScale

Advances in High-Performance Interconnect Infrastructure

September 2011

diego@mellanox.com

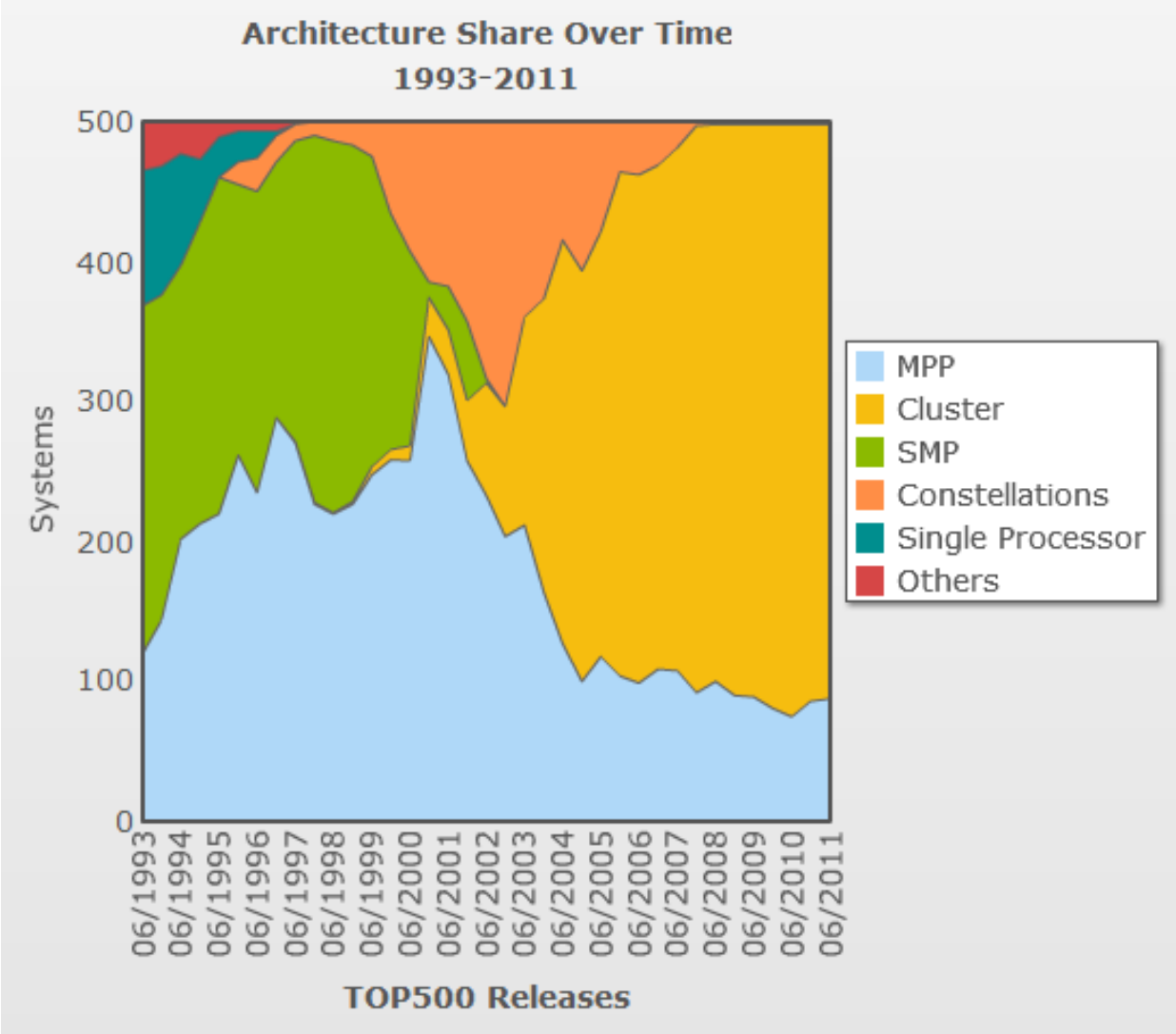
ExaScale Computing

- Ambitious Challenges Foster Progress

- Demand
  - Research Institutes, Universities and Government Labs
  - Commercial Space
    - Automotive, aerospace, oil and gas explorations, digital media, financial simulation
    - Mechanical simulation, package designs, silicon manufacturing etc.
  - Clouds

- Affordable Commoditization
  - Positive Feedback (massive market helps lower prices even more)

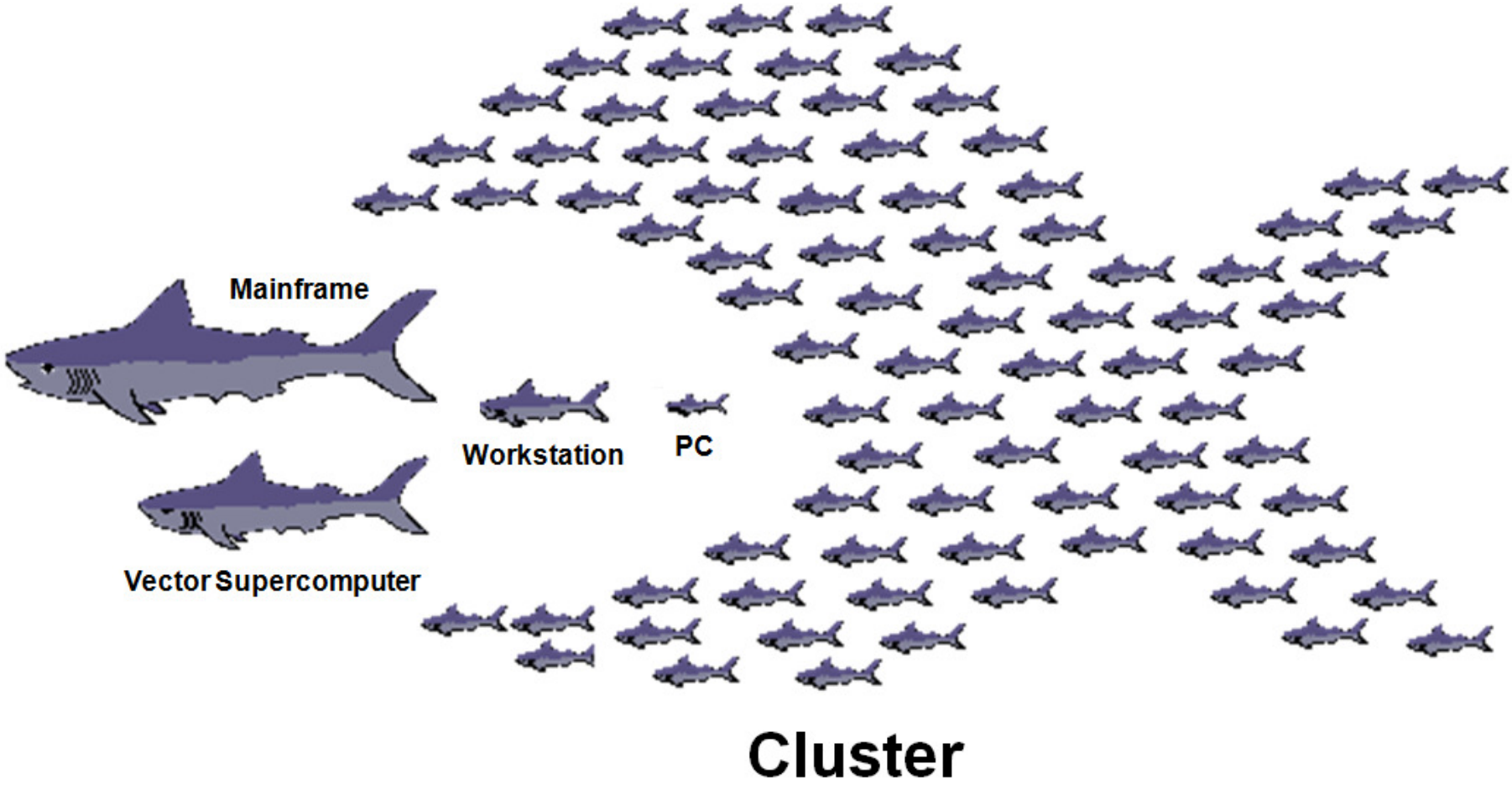
Top500 Supercomputers List – System Architecture



Clusters have become the most used HPC system architecture

More than 80% of Top500 systems are clusters

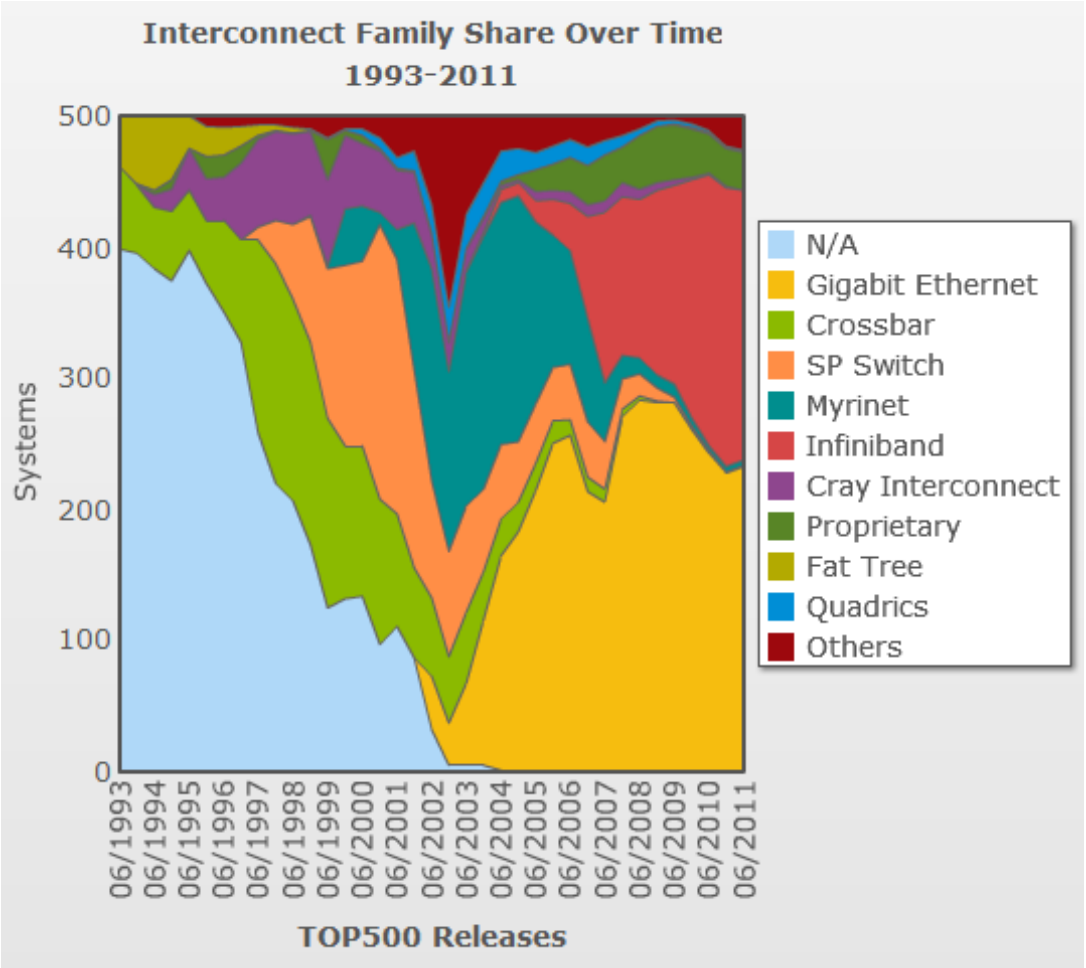
Affordable and Efficient Scalability – The Demise of Specialized Systems



Parallel Computing on a Large Number of Servers is More Efficient than using Specialized Systems

Not a Cluster unless it Scales…

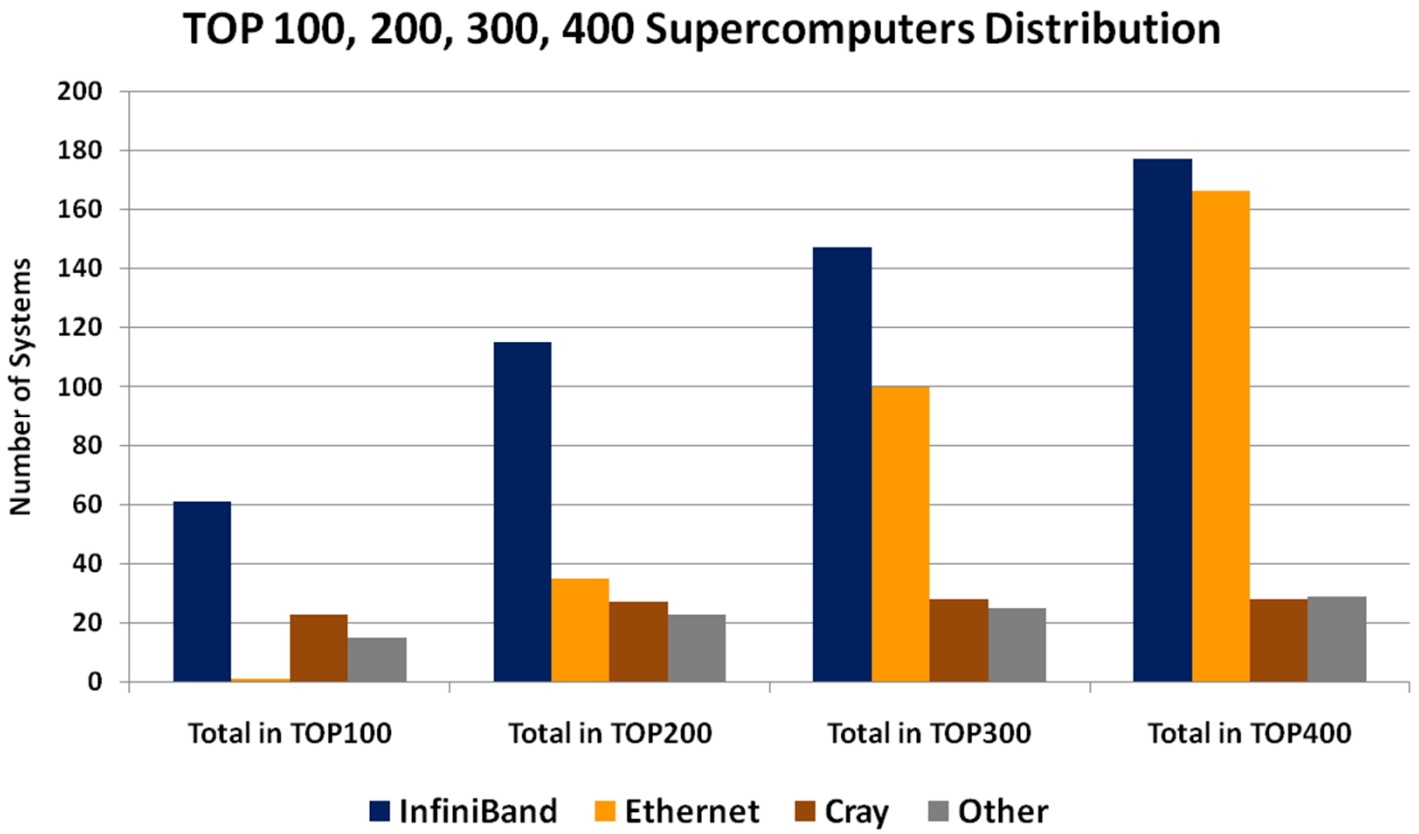
- Cluster vs JBCN (Just a Bunch of Compute Nodes)
  - Performance
  - Efficiency



InfiniBand and Ethernet

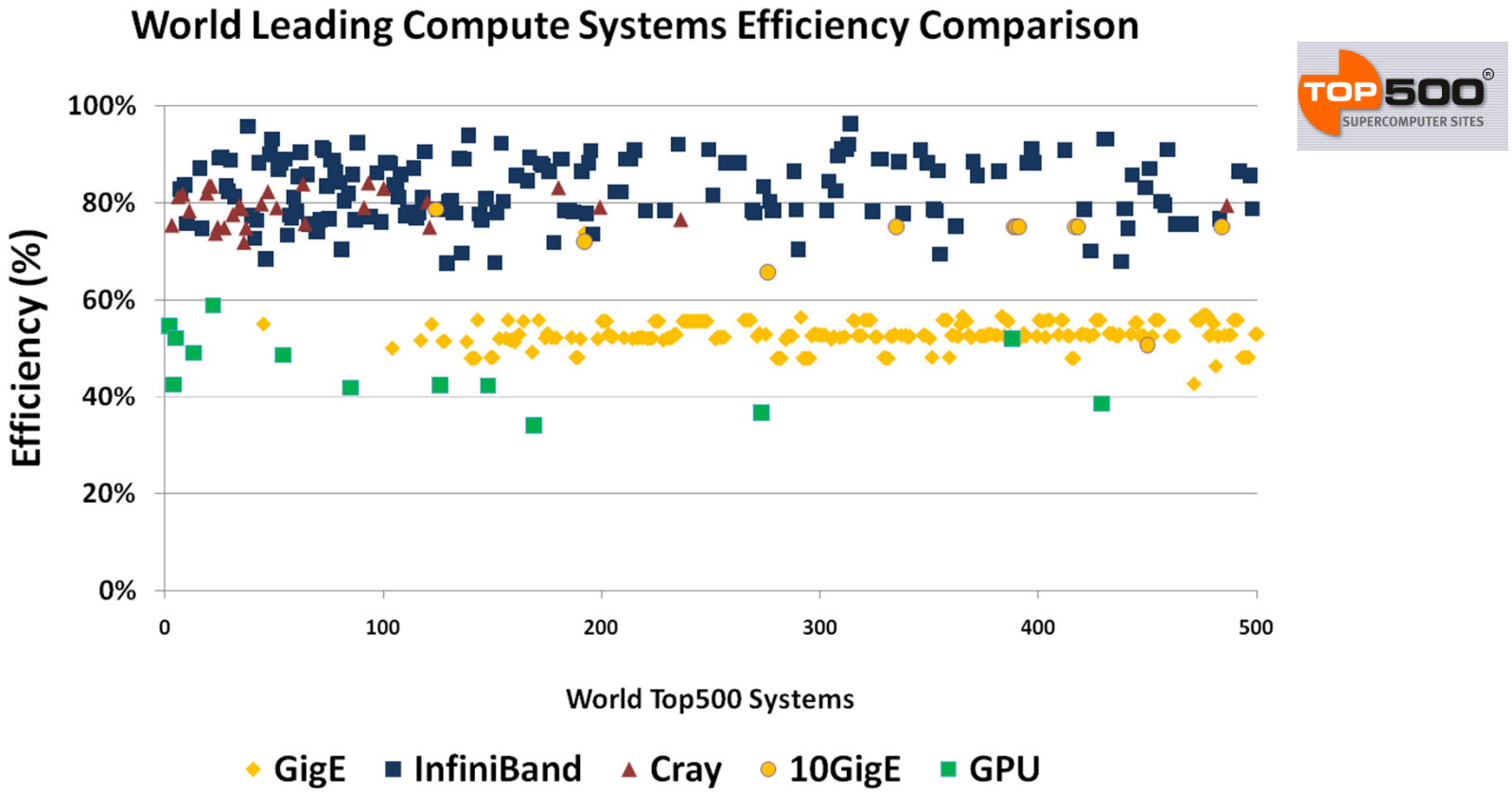
Feature	Ethernet	Ethernet w/RoCE	InfiniBand
Data rate	10Gb/s and 40Gb/s		56Gb/s InfiniBand FDR
Latency	5 to 10 usec	~2usec	Less than 1 usec
Lossless	Incipient / Pause Based		Credit Based Flow Control
CLOS/Fat Tree Scalability	No. Spanning Tree. Some proprietary schemes available. Incipient Stds.		Yes. In deployment for years. Proven 10000+ nodes deployments.
Congestion Management	Software (TCP) based. Incipient Std for L2 Congestion Management.		HW based congestion management
Stateful Offloads	TOE. Very Limited to date. Power, scaling and Linux community adoption challenges.	InfiniBand Transport offload has been mainstream for almost a decade.	
RDMA	Limited availability. Not mature. TOE issues.	Yes	
Management	Ethernet Network Management		Centralized IB Management

Interconnect Trends – Top100, Top200, Top300, Top400

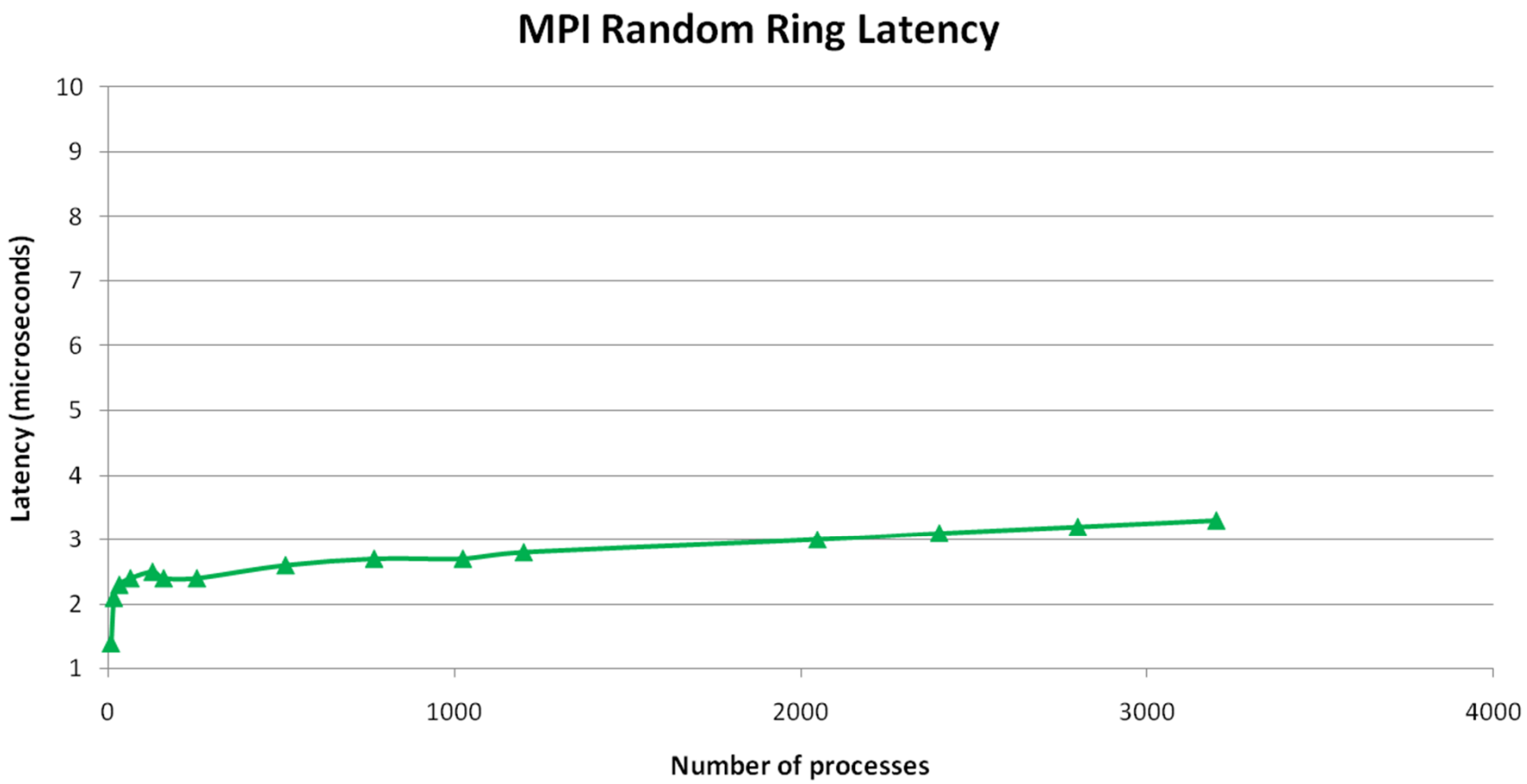


- InfiniBand connects the majority of the TOP100, 200, 300, 400 supercomputers

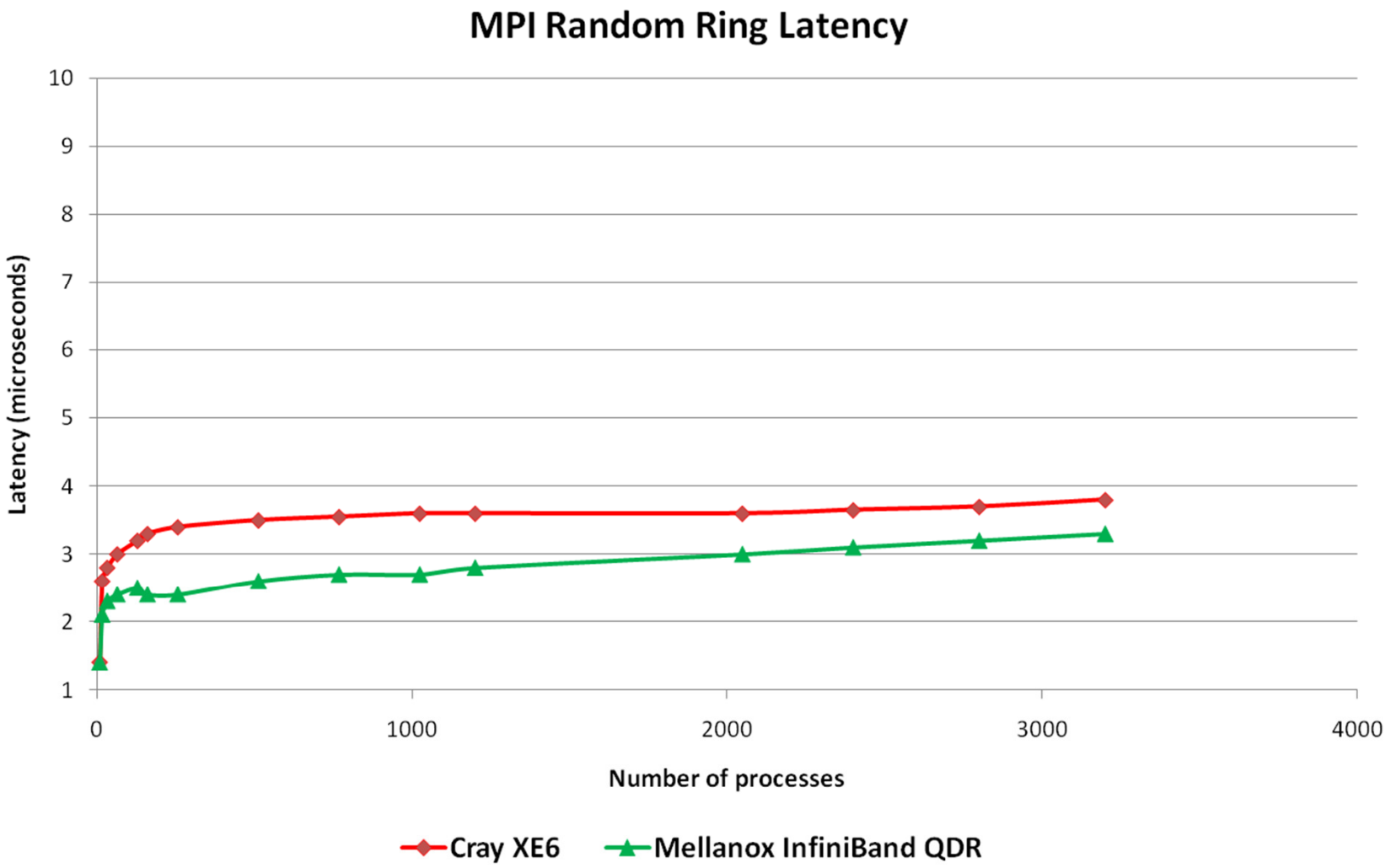
System Efficiency



Mellanox MPI Optimization – MPI Random Ring

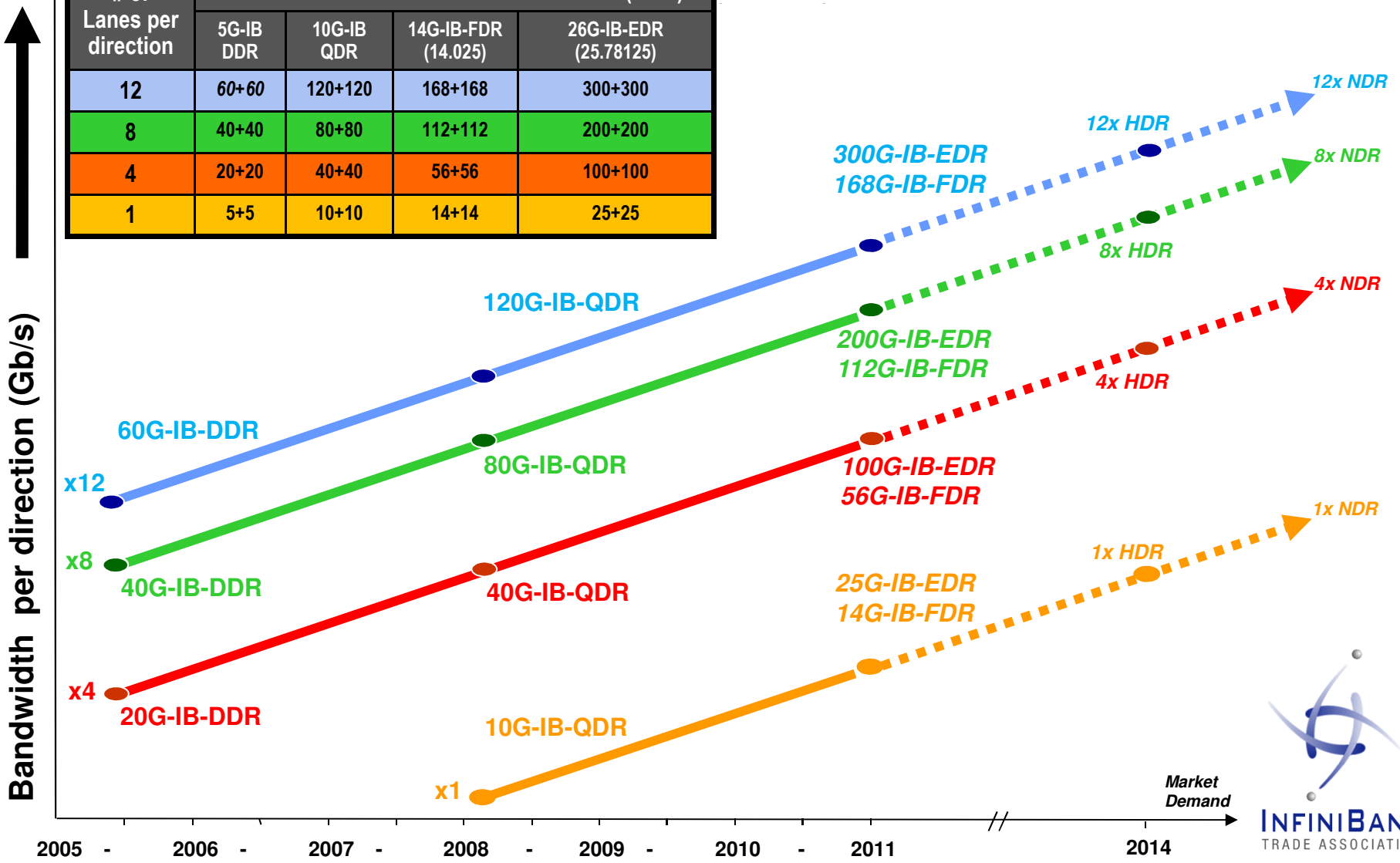


Mellanox InfiniBand versus Cray

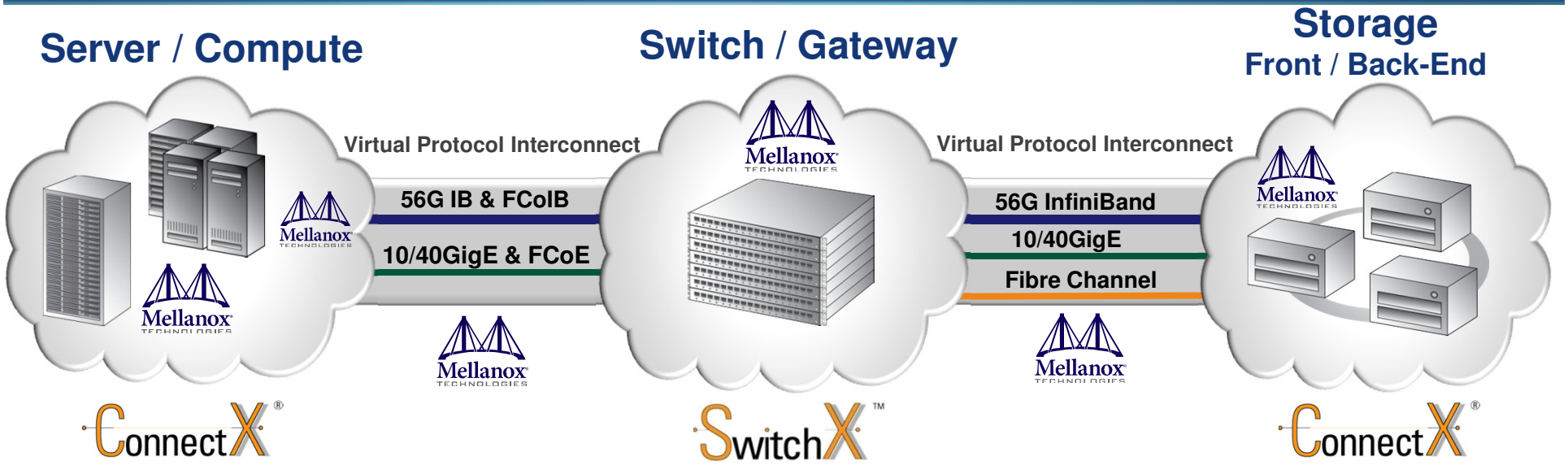


InfiniBand Link Speed Roadmap

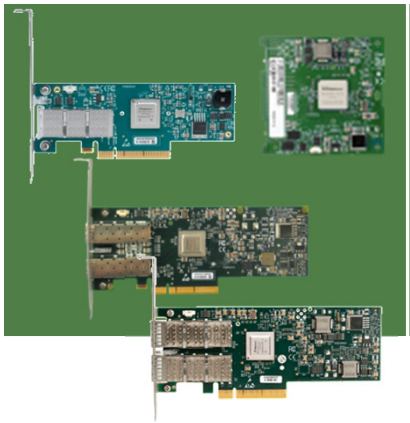
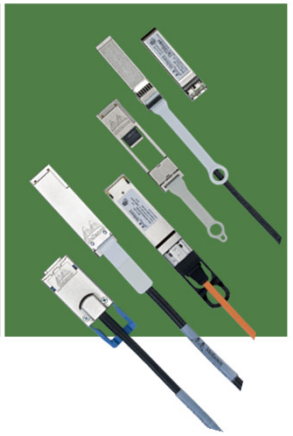
# of Lanes per direction	Per Lane & Rounded Per Link Bandwidth (Gb/s)			
	5G-IB DDR	10G-IB QDR	14G-IB-FDR (14.025)	26G-IB-EDR (25.78125)
12	60+60	120+120	168+168	300+300
8	40+40	80+80	112+112	200+200
4	20+20	40+40	56+56	100+100
1	5+5	10+10	14+14	25+25



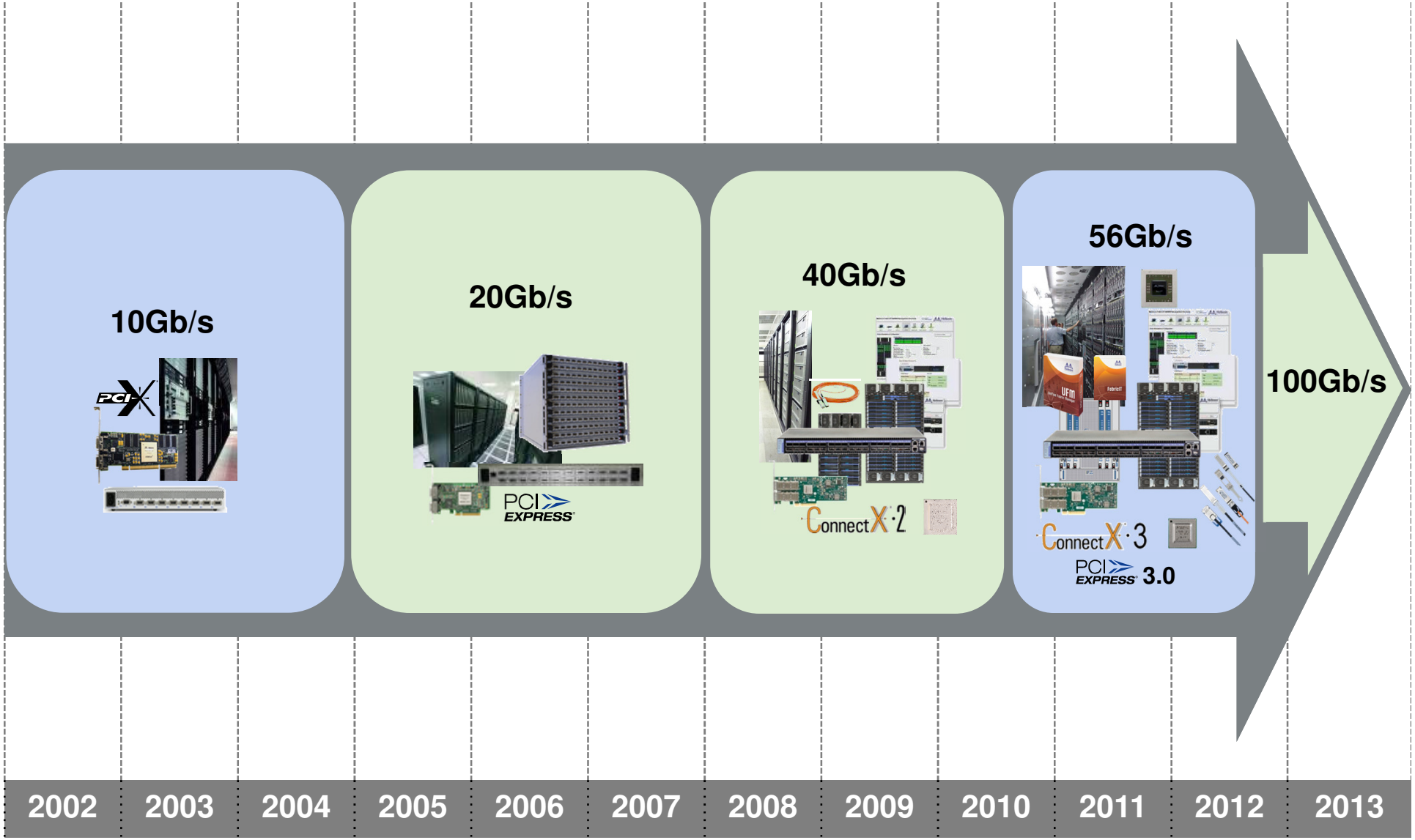
Mellanox End-to-End Connectivity Solutions for Servers and Storage



Industry's Only End-to-End InfiniBand and Ethernet Portfolio

ICs	Adapter Cards	Host/Fabric Software	Switches/Gateways	Cables
				

Mellanox Technology/Solutions Roadmap



Paving The Road to Exascale Computing

Dawning (China)

TSUBAME (Japan)

NASA (USA)
>11K nodes

LANL (USA)

Los Alamos NATIONAL LABORATORY

PetaScale

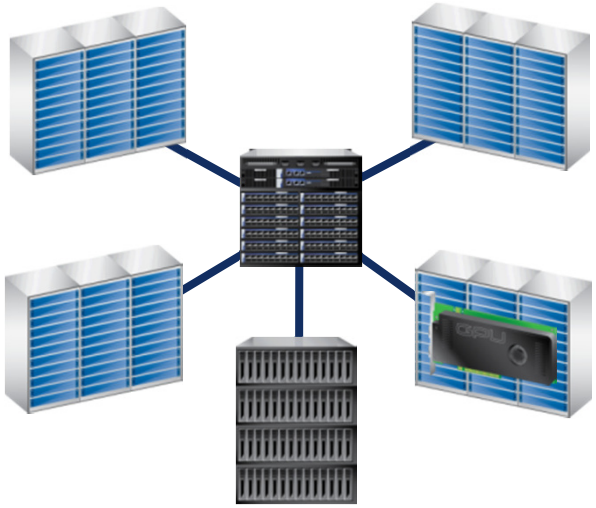
Mellanox Connected

CEA (France)

- Mellanox InfiniBand is the interconnect of choice for PetaScale computing
  - Accelerating 50% of the sustained PetaScale systems (5 systems out of 10)

HPC Advisory Council

- Multiple compute clusters, open environment
- Enable remote access for development, testing, benchmarking
- Join at http://www.hpcadvisorycouncil.com/

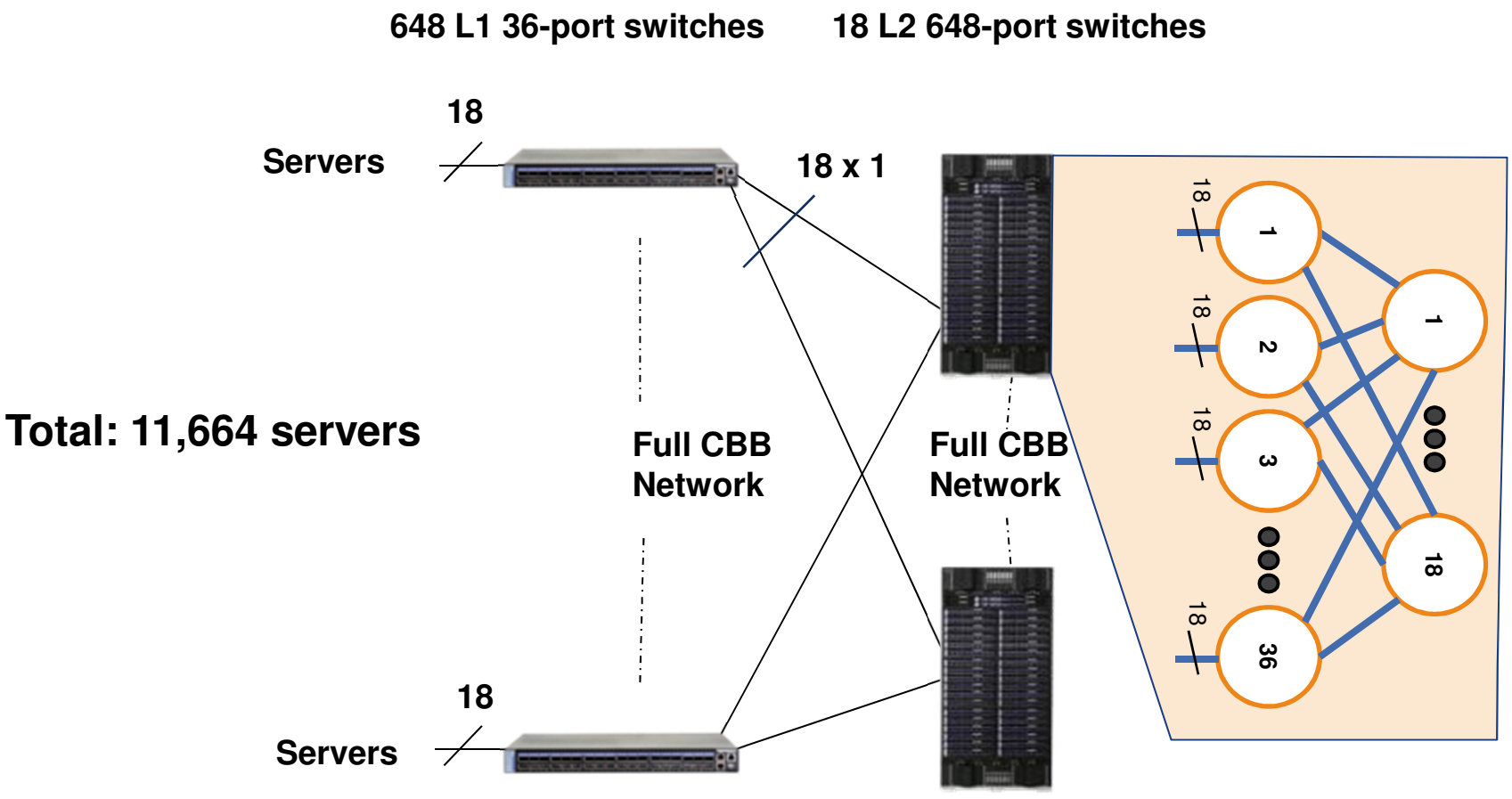


HPC ADVISORY COUNCIL

The Road to ExaScale

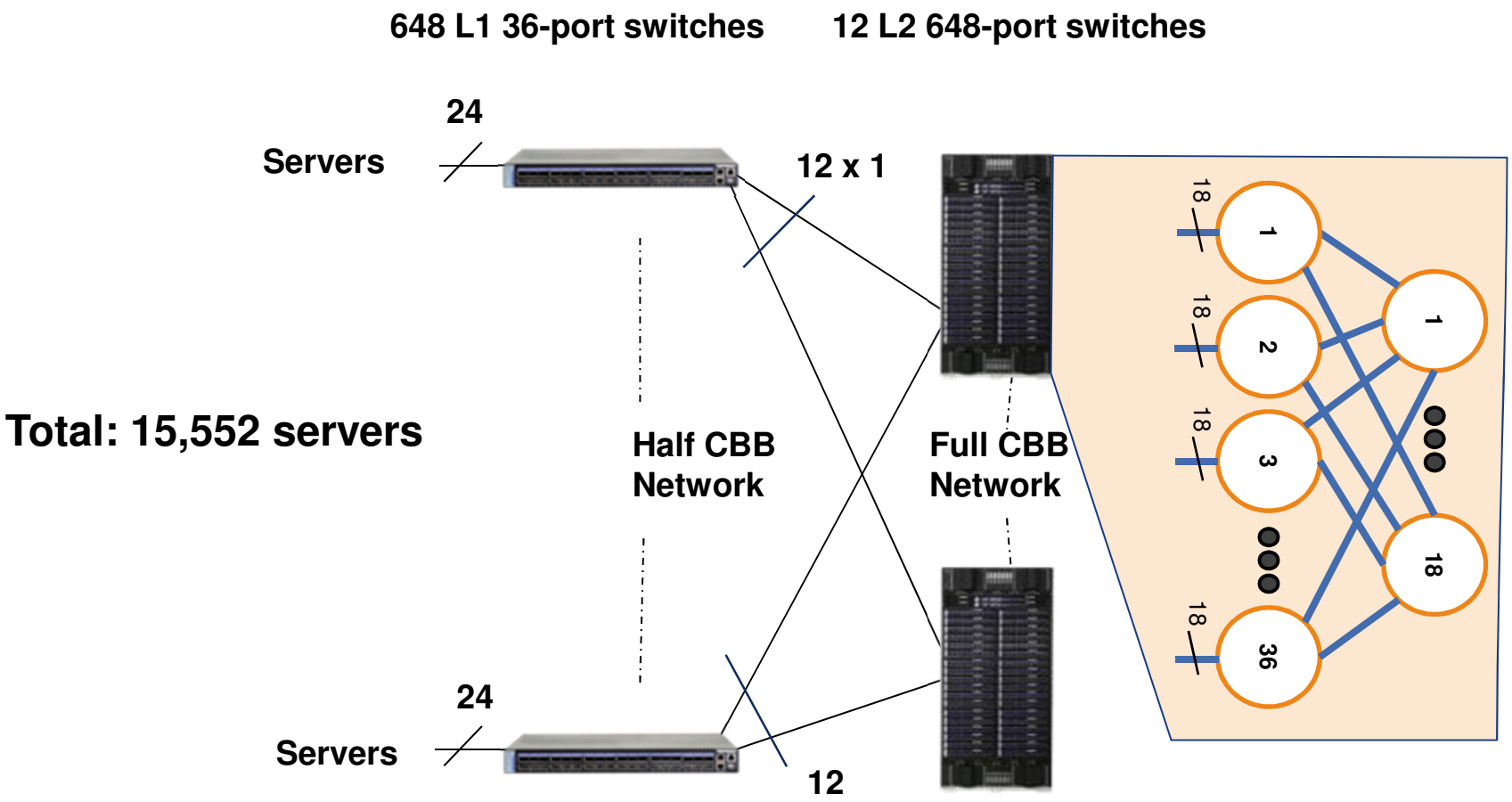
Cluster Topologies

3 Level Full CBB (1:1) Fat-Tree

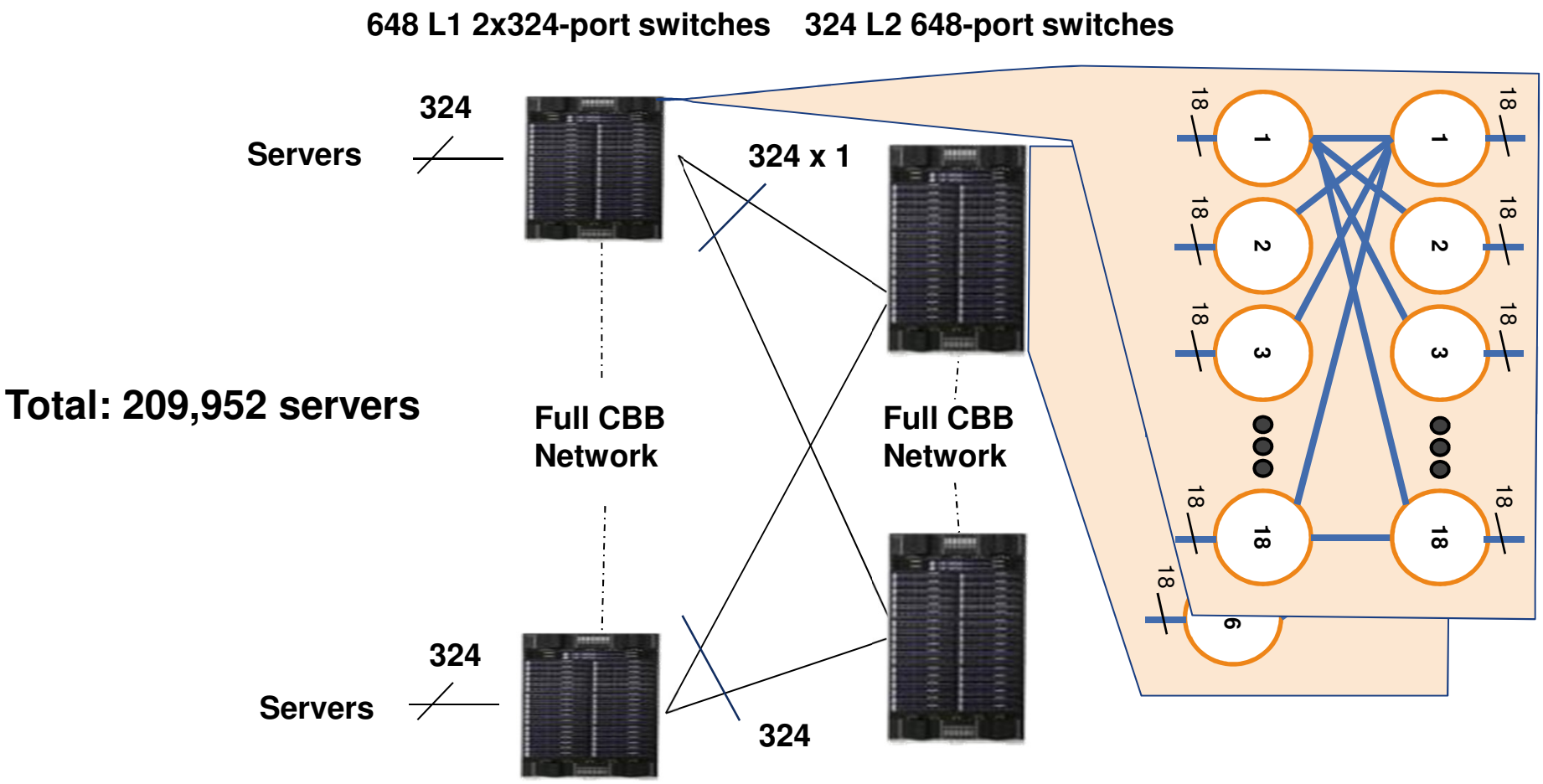


Total: 11,664 servers

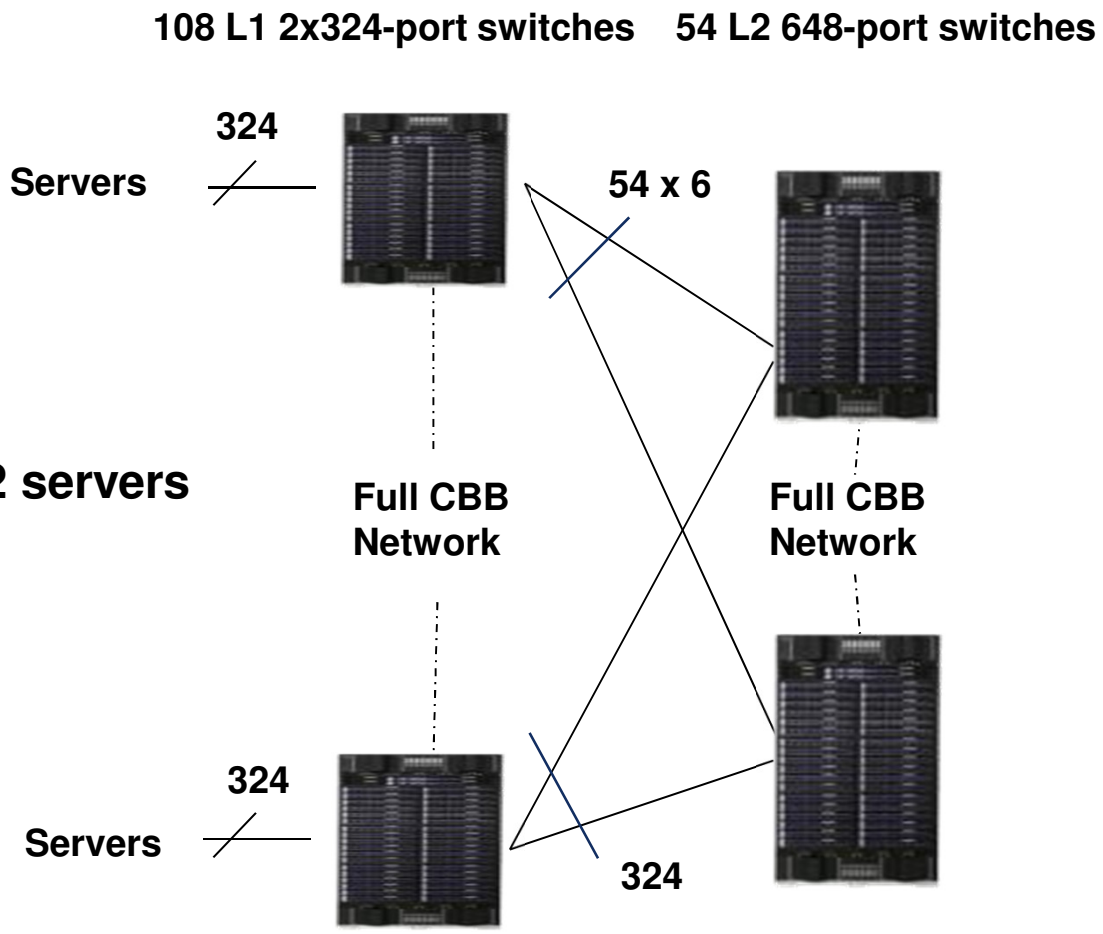
3 Level Half CBB Fat-Tree



4 Level Full CBB Fat-Tree



A 35K nodes 4 Level Full CBB Fat-Tree



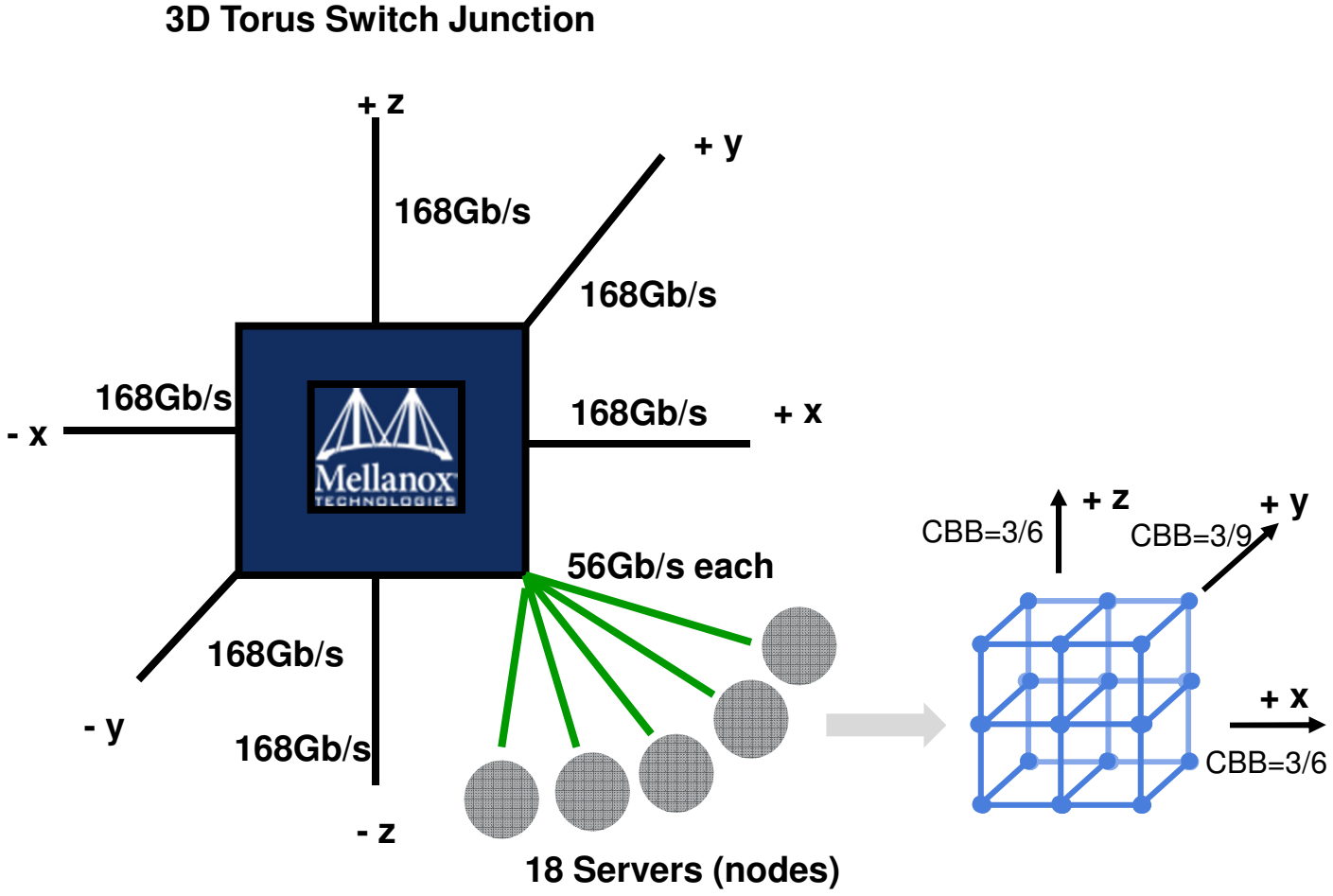
Total: 34,992 servers

Cables: Short = Hosts; Long = Hosts

Summary – Fat Trees Scaling

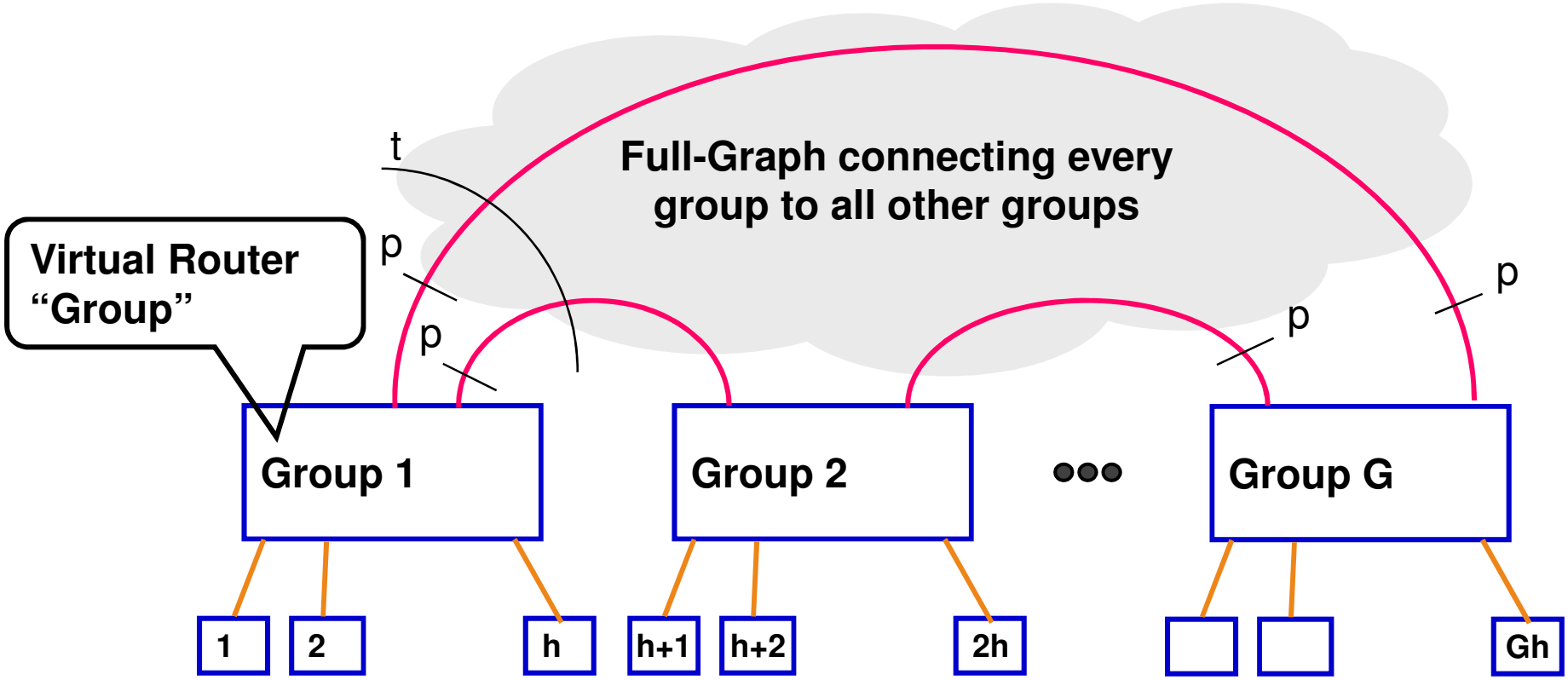
- Fat Trees provide a simple Cost/CBB tradeoff
  - Applied at each level of the tree
- At 4 levels may scale to 210K nodes for CBB=1
- Number of “long cables” is Num-Host / CBB
  - For 3 and 4 levels topologies

3(or more)D Torus

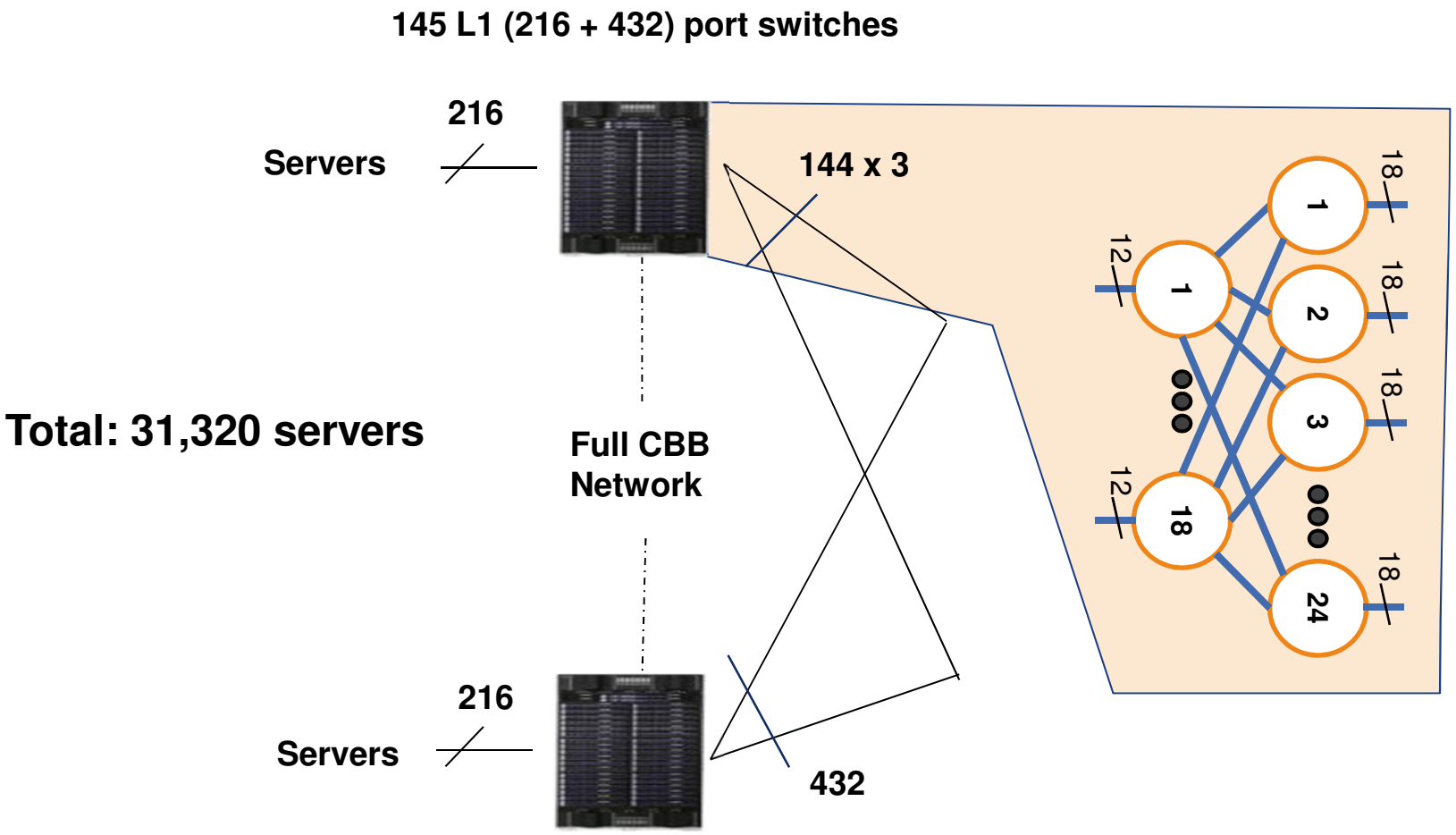


3D Torus size: 8x8x8 (512 36-port switches)
Total number of servers: 9216

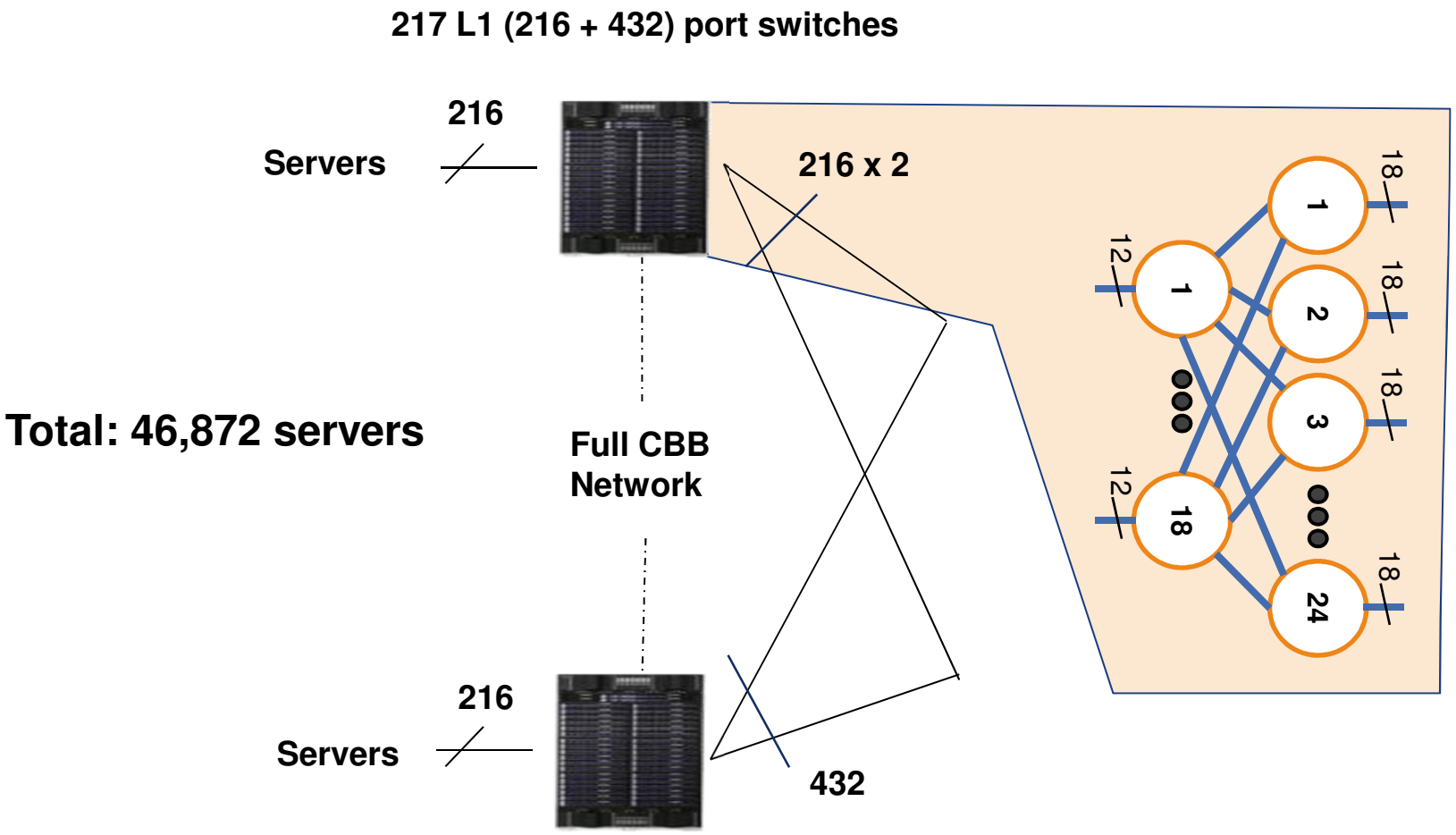
DragonFly Topology – The "Global" View



Example: A ~31K nodes Full CBB DragonFly



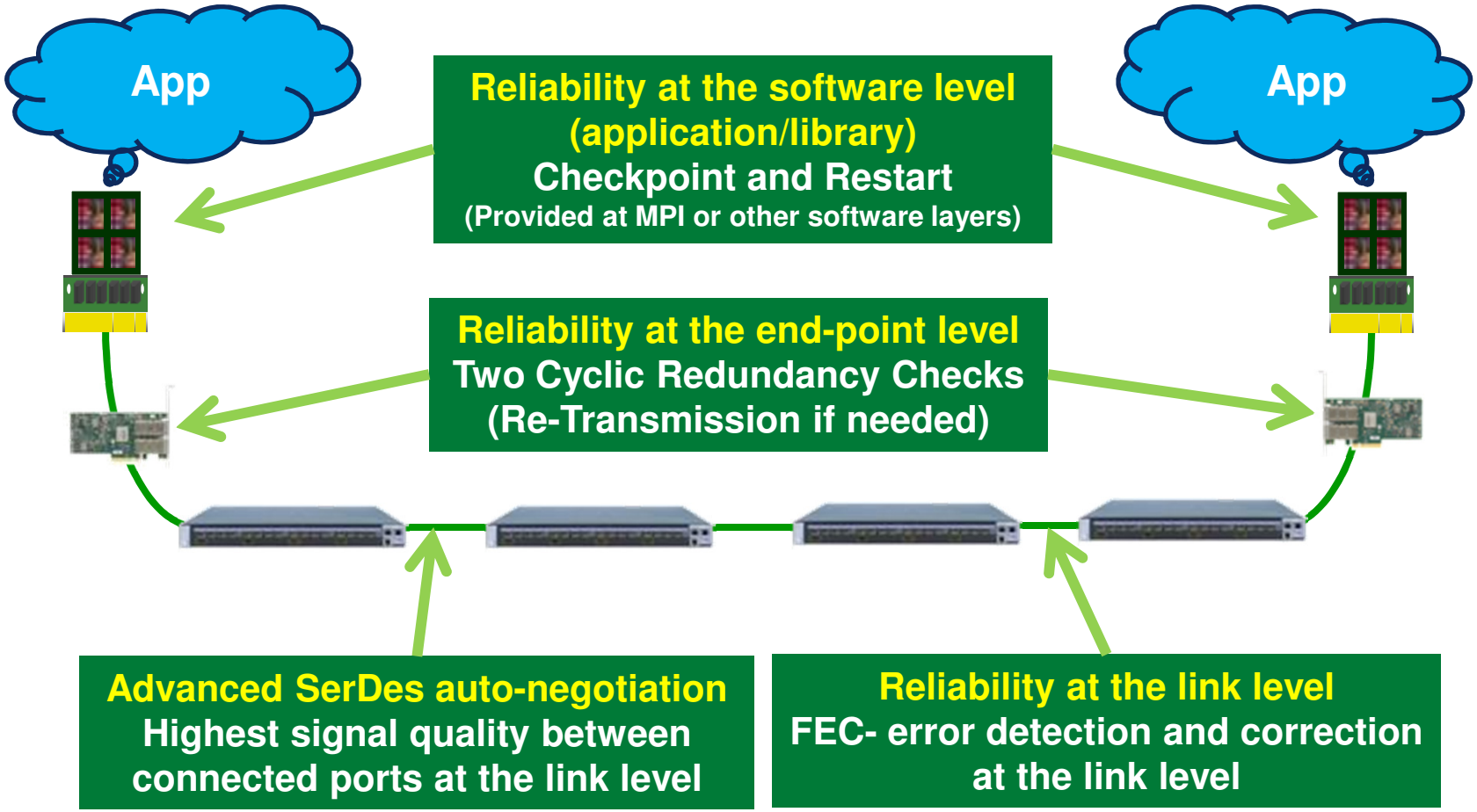
Example: A ~47K nodes Full CBB DragonFly



The Road to ExaScale

High Availability and Fault Tolerance

Complete End-to-End Reliability

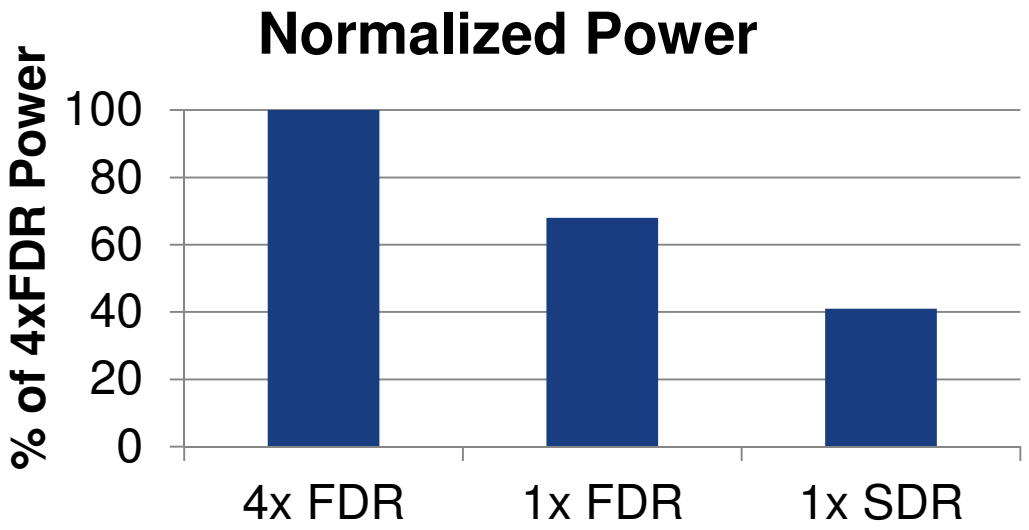


The Road to ExaScale

Power Efficiency

Link Level - Energy Proportional IO

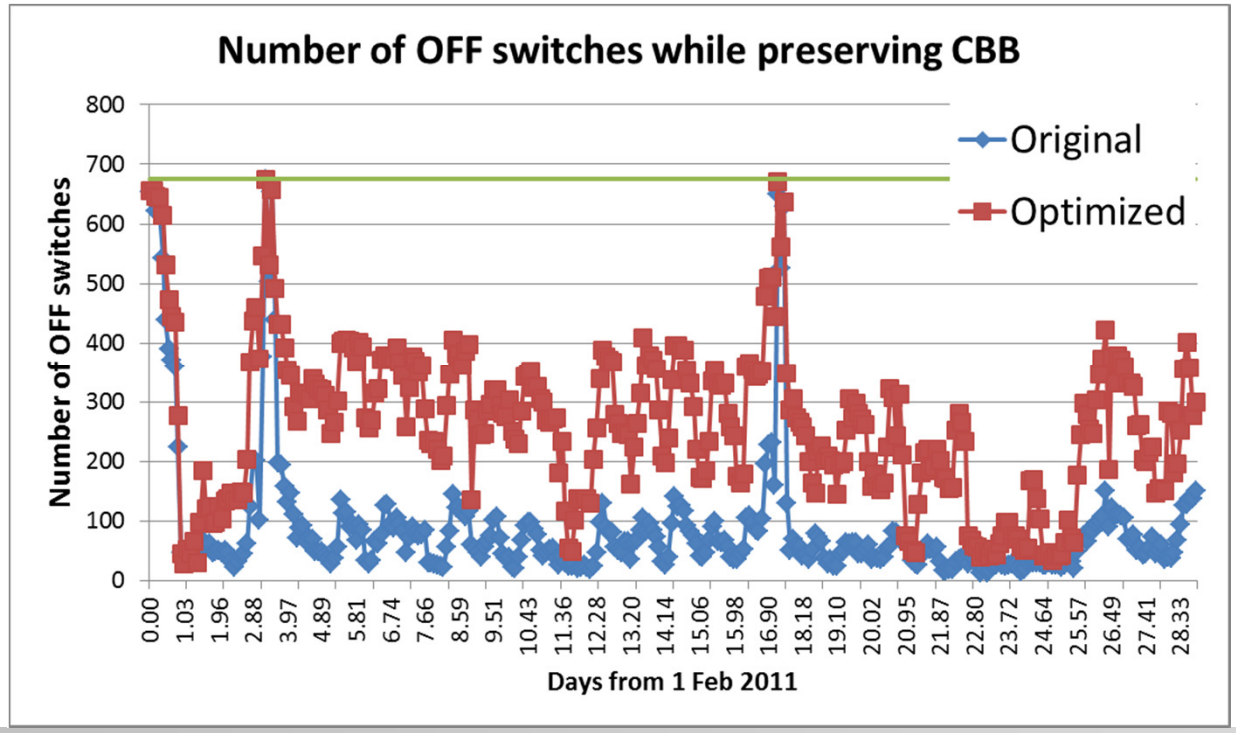
- Much of the cluster power can be saved if the link's power is made proportional to the BW [3]
  - Mellanox InfiniBand™ provides this feature
  - Scale link speed, chip frequency and voltage
- Self detection and transparent power reduction by link adaptation



[3] D. Abts, M. R. Marty, P. M. Wells, P. Klausler, and H. Liu, "Energy proportional datacenter networks," ACM SIGARCH Computer Architecture News, vol. 38, p. 338–347, Jun. 2010.

Job and Task Scheduling for Power

- Traffic planning enables planned switch turn OFF
- Job placement can maximize saving while CBB is preserved
- A new algorithm was simulated on EUROPA Moab logs
- On average 38% of the switches could be turned OFF
  - While maintaining cluster utilization of ~74%



The Road to ExaScale

Scalable Collectives Acceleration

MPI Collective Operations

- MPI provides messaging interface for parallel computing
  - Communications options include send/receive and collectives
  - Used by applications processes for communications
    - One-to-one (one process to another), many-to-one, one-to-many
- Collectives communications
  - Have a crucial impact on the application's scalability and performance
  - Communications used for one-to-many or many-to-one
    - Used for sending around initial input data
    - Reductions for consolidating data from multiple sources
    - Barriers for global synchronization
- 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
- Offloading MPI collectives to the network
  - Increases performance, increase CPU efficiency, provides overlapping
  - Critical for ensuring high scalability

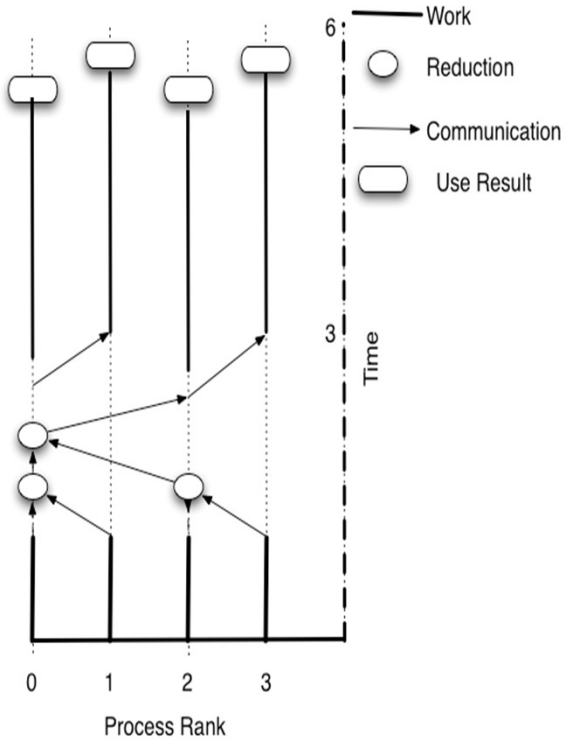
OS Noise and Jitter

- As time between communication drops so rises the Collective Communication frequency
- Collectives suffer from OS Noise
  - As they synchronize at least once
  - Some collectives accumulate OS Noise through their critical path
- Offloading MPI collectives to the network
  - Increases performance, increase CPU efficiency, provides overlapping
  - Critical for ensuring high scalability
- Mellanox FCA product reduce OS Jitter and improves communication computation overlap [4]

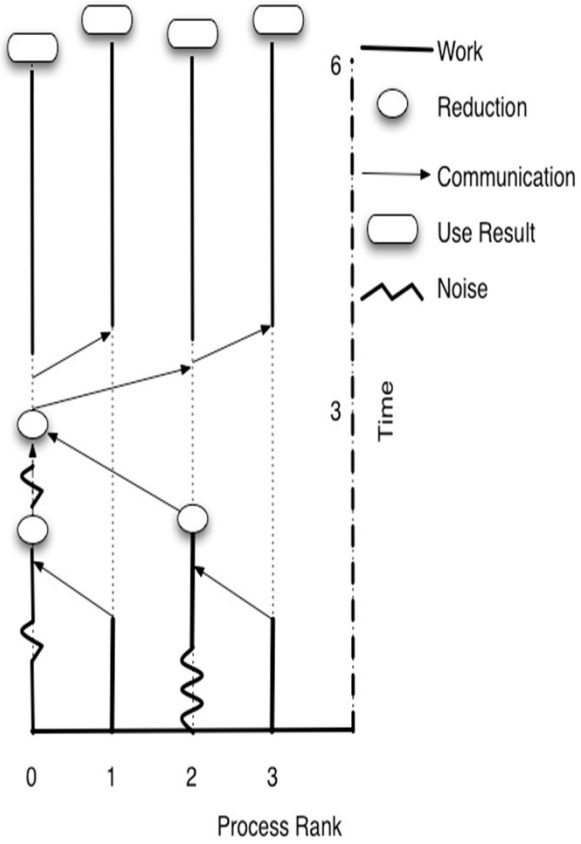
[4] M. G. Venkata, R. Graham, J. Ladd, P. Shamis, I. Rabinovitz, V. Filipov, G. Shainer “ConnectX-2 CORE-Direct Enabled Asynchronous Broadcast Collective Communications”, Communication Architecture for Scalable Systems Workshop IPDPS 2011

The Effects of System Noise on Applications Performance

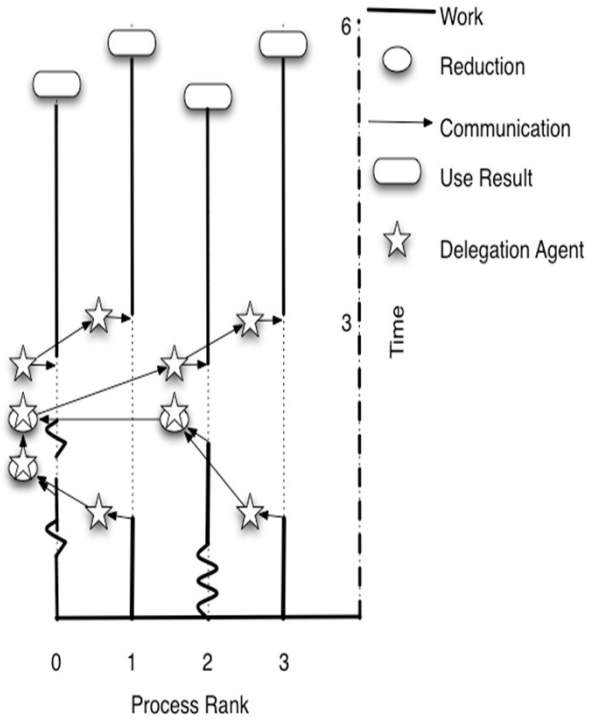
- Minimizing the impact of system noise on applications – critical for scalability



Ideal



System noise



CORE-Direct (Offload)

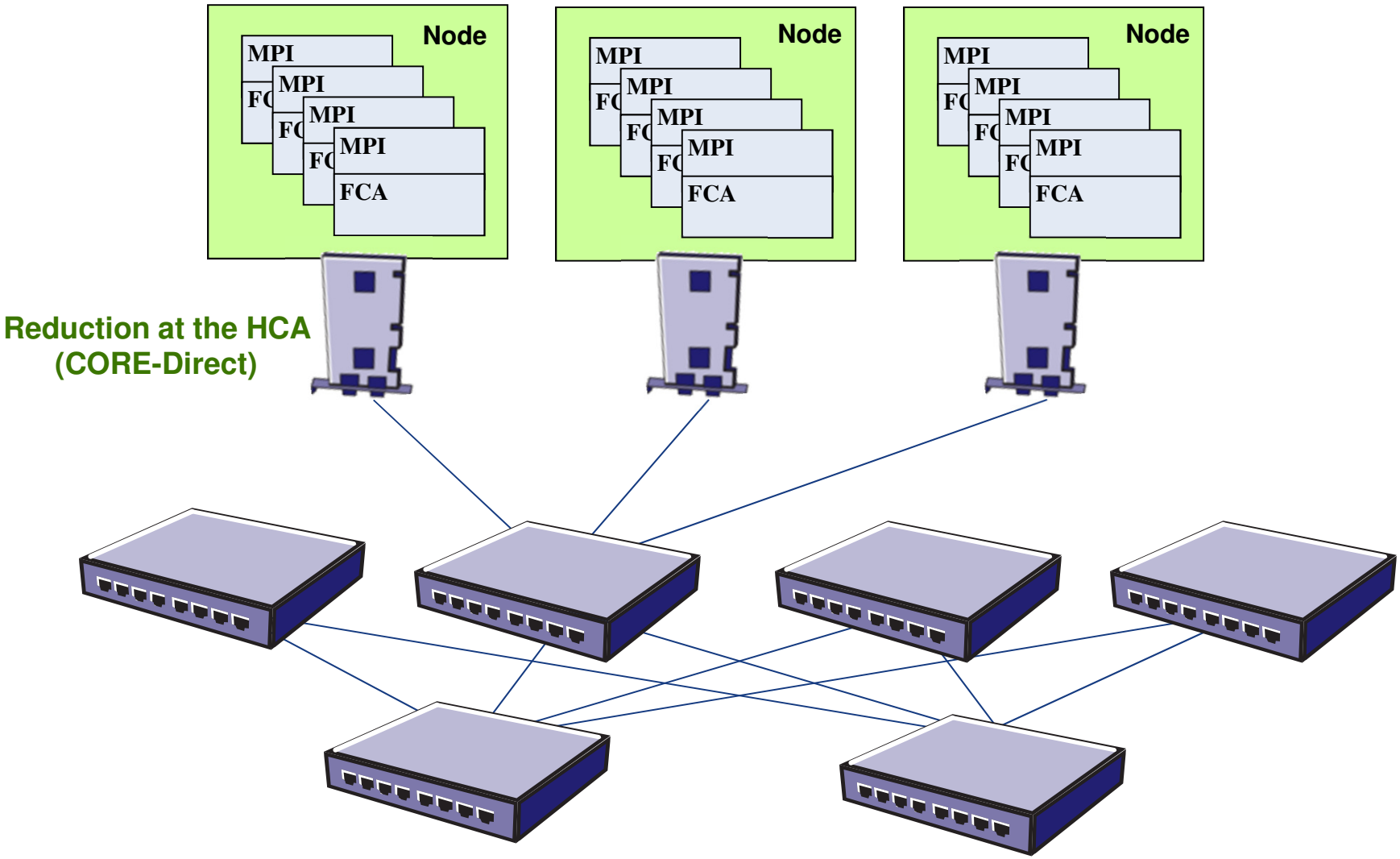
Mellanox Collectives Acceleration Solution

- Hardware-based Acceleration technologies

- CORE-Direct
  - Adapter-based hardware offloading for collectives operations
  - Includes floating-point capability on the adapter for data reductions
  - CORE-Direct API is exposed through the Mellanox drivers
    - Available for 3rd party libraries/software protocols

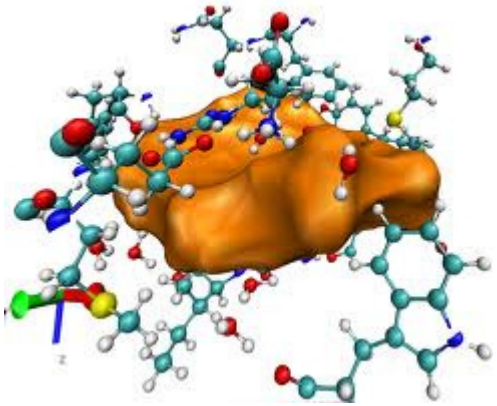
- FCA
  - A software plug-in package that integrates into available MPIs
  - FCA replaces the MPI software library code for collective communications
  - FCA implements MPI collectives operations using the hardware accelerations
  - FCA includes support for sophisticated collectives algorithms
  - FCA is available through licensing

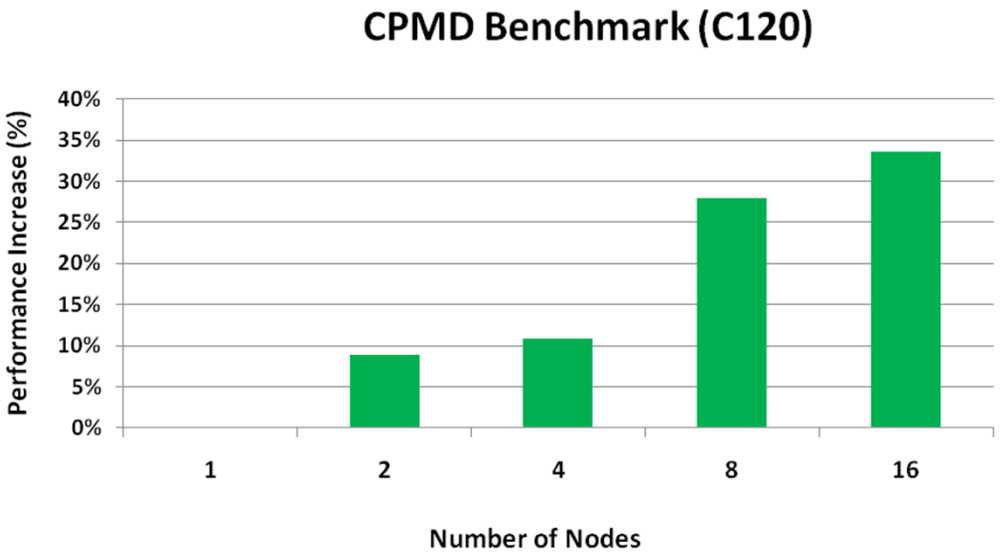
Scalable Collectives Acceleration



Application Example: CPMD (Molecular Dynamics)

- CPMD is a leading molecular dynamics applications
- Result: FCA accelerates CPMD by nearly 35%
  - At 16 nodes, 192 cores
  - Performance benefit increases with cluster size – higher benefit expected at larger scale



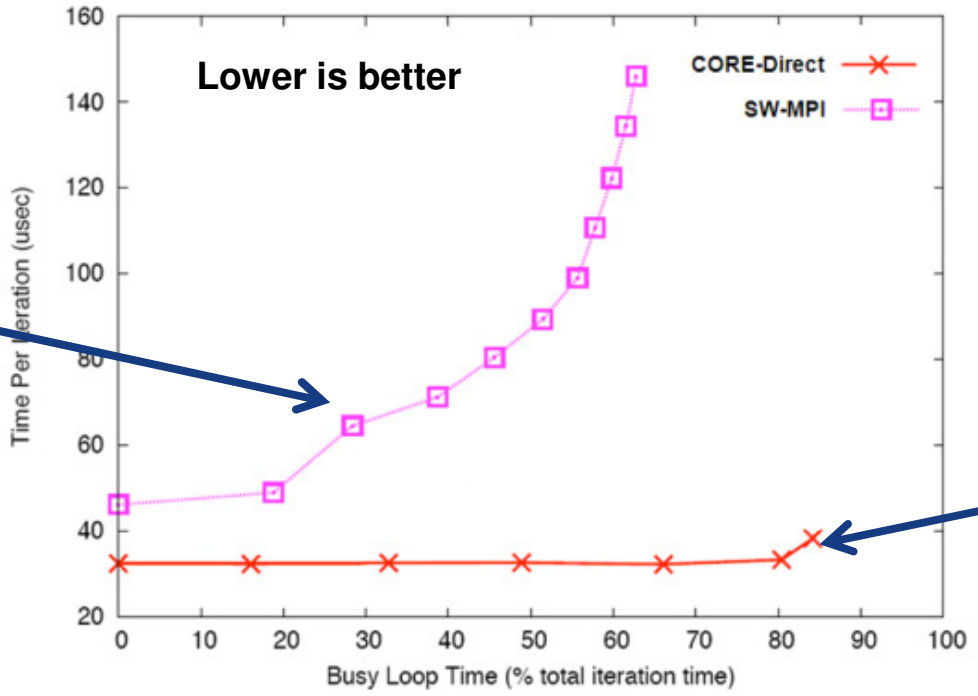


Higher is better

***Acknowledgment: HPC Advisory Council for providing the performance results**

Collectives Offloads - Non Blocking Collectives

- Presents the overlapping benefit of collective offloads
  - Non-blocking collective implementation - non-blocking barrier
    - Initiate non-blocking MPI barrier
    - CPU to performance application calculations
    - Wait for non-blocking barrier to complete



Software MPI: Losing performance beyond 20% CPU computation availability

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

* Data provided by Oakridge National Lab

The Road to ExaScale

GPU Accelerations with GPUDirect

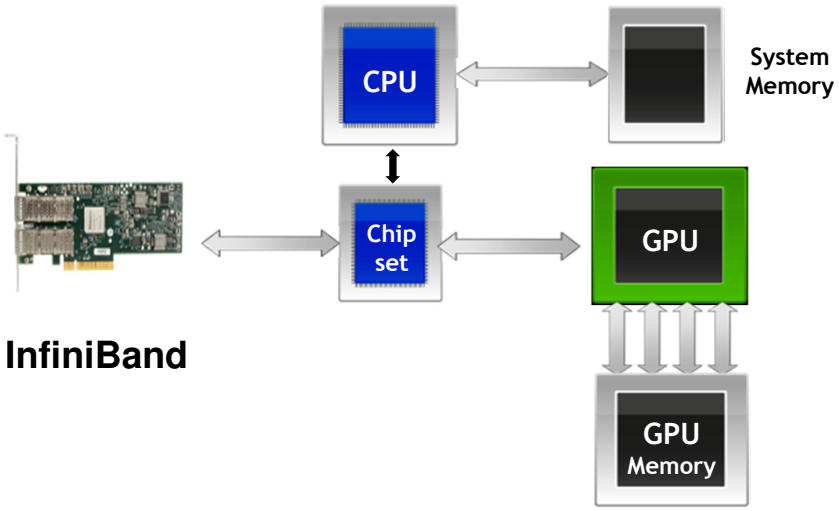
GPUDirect Overview

- The GPUDirect project
  - “NVIDIA Tesla GPUs To Communicate Faster Over Mellanox InfiniBand Networks”, http://www.nvidia.com/object/io_1258539409179.html
- GPUDirect was developed together by Mellanox and NVIDIA
  - New interface (API) within the Tesla GPU driver
  - New interface within the Mellanox InfiniBand drivers
  - Linux kernel modification to allow direct communication between drivers

GPU-InfiniBand Bottleneck (pre-GPUDirect)

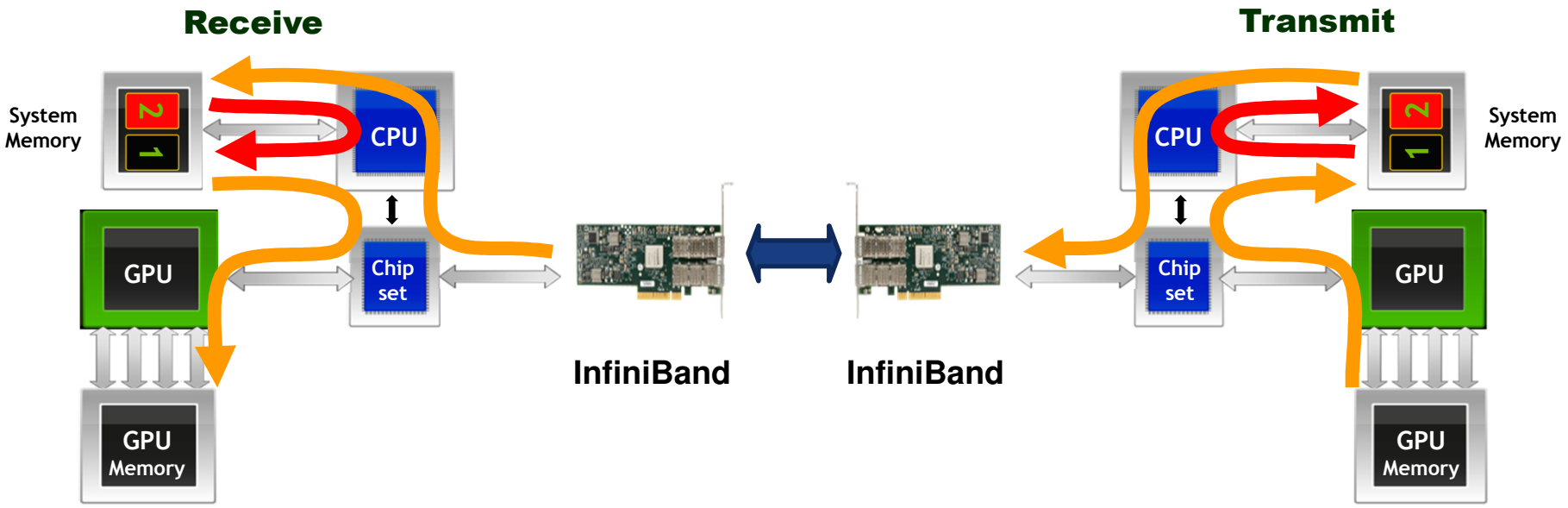
- GPU communications uses “pinned” buffers for data movement
  - A section in the host memory that is dedicated for the GPU
  - Allows optimizations such as write-combining and overlapping GPU computation and data transfer for best performance

- InfiniBand uses “pinned” buffers for efficient RDMA transactions
  - Zero-copy data transfers, Kernel bypass
  - Reduces CPU overhead



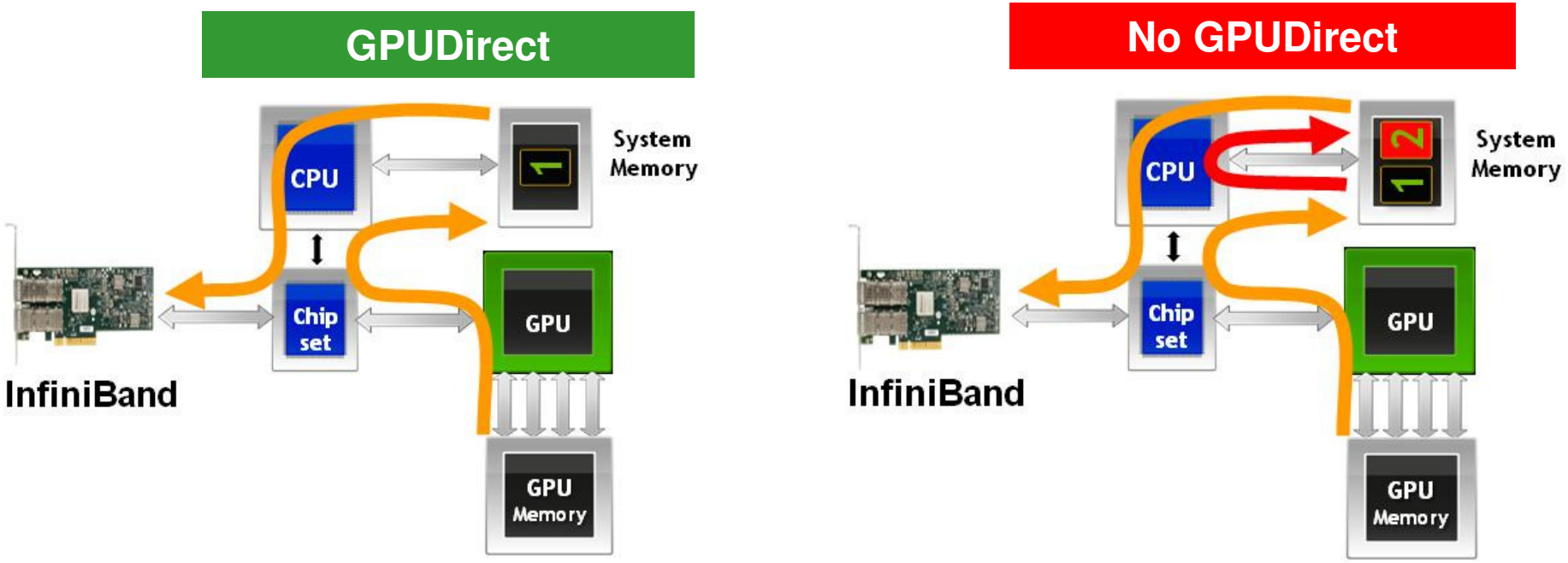
GPU-InfiniBand Bottleneck (pre-GPUDirect)

- Pre-GPUDirect, GPU communications required CPU involvement in the data path
  - Memory copies between the different “pinned buffers”
  - Slow down the GPU communications and creates communication bottleneck



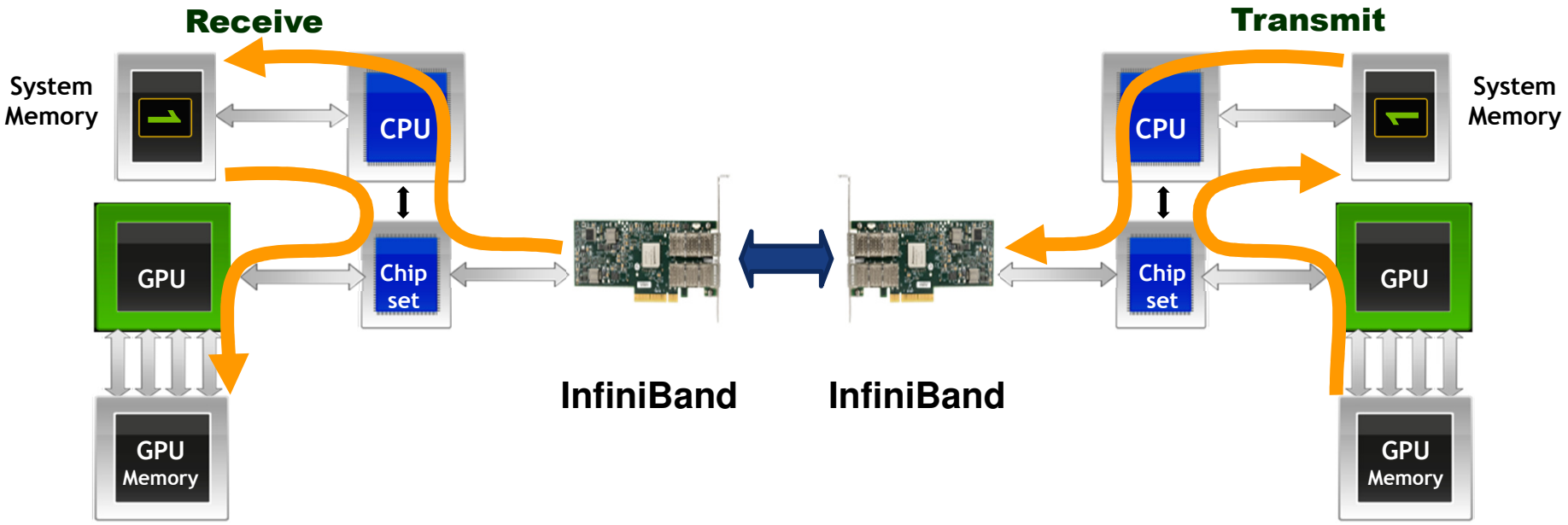
GPUDirect Technology

- Allows Mellanox InfiniBand and NVIDIA GPU to communicate faster
  - Eliminates memory copies between InfiniBand and GPU
  - Eliminate CPU involvement in the GPU data path
  - Note: Only offloading InfiniBand can deliver GPUDirect

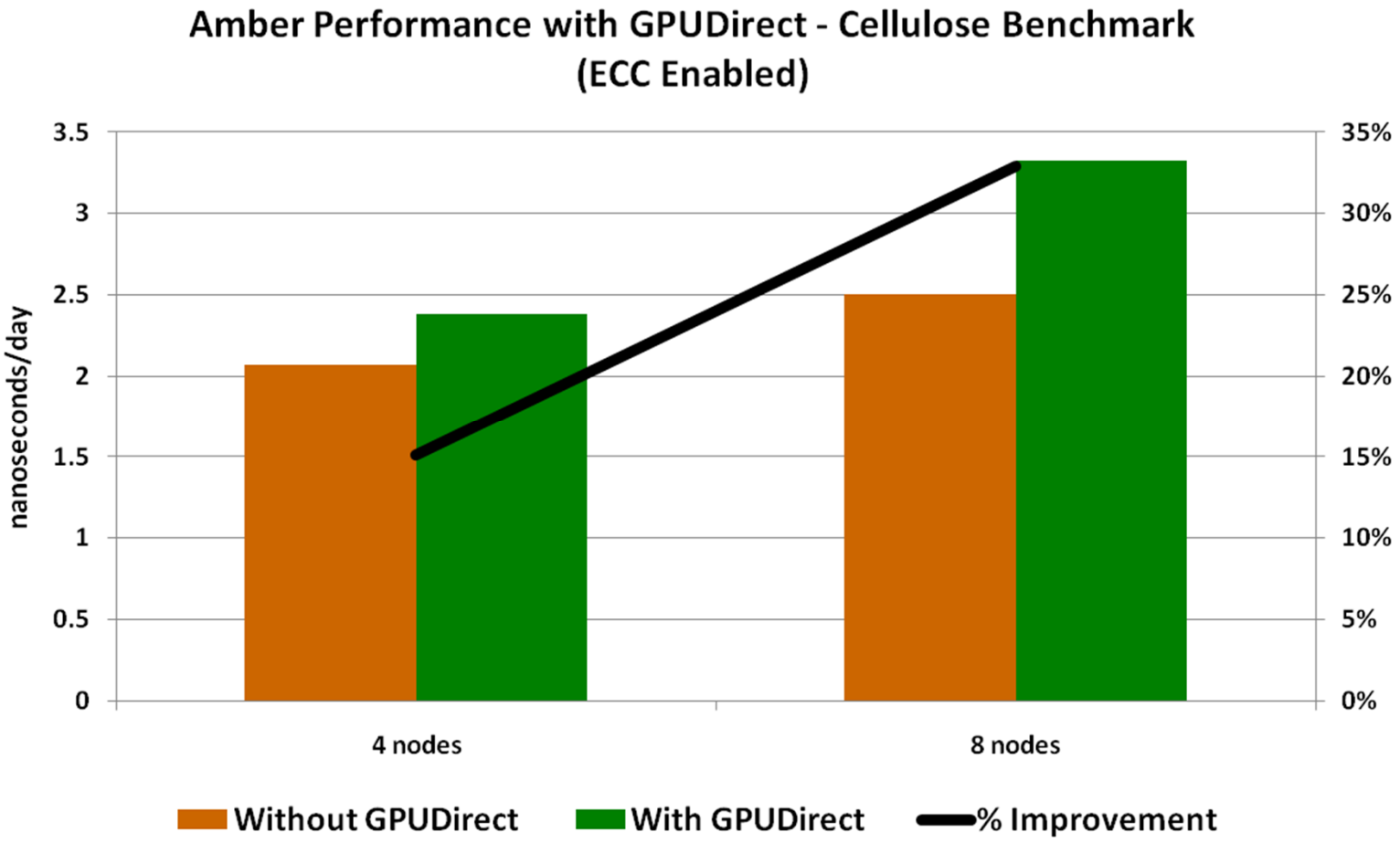


Accelerating GPU Based Supercomputing

- Fast GPU to GPU communications
- Native RDMA for efficient data transfer
- Reduces latency by 30% for GPUs communication

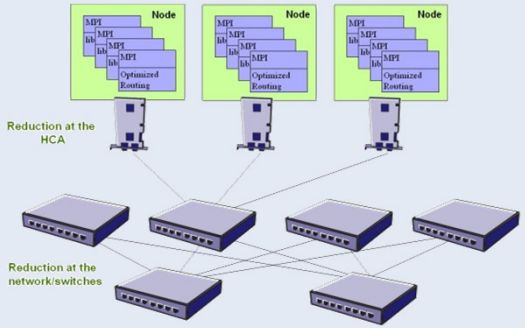


Amber Performance with GPUDirect
Cellulose Benchmark



- 33% performance increase with GPUDirect
- Performance benefit increases with cluster size

Paving The Road to Exascale Computing



OAK RIDGE National Laboratory

FCA Fabric Collective Accelerator

10s-100s% Boost

Scalable Offloading for MPI/SHMEM

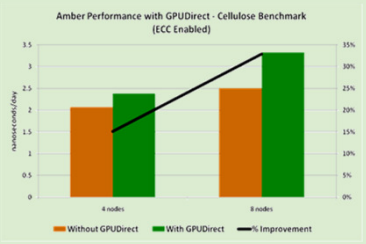
CSCS

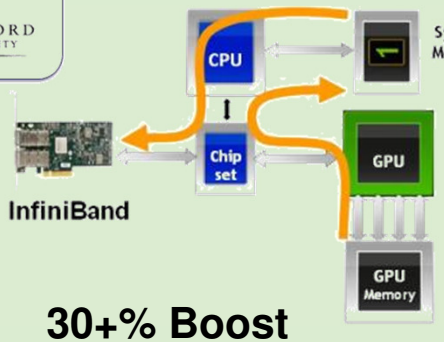
NERSC

Sandia National Laboratories

STANFORD UNIVERSITY

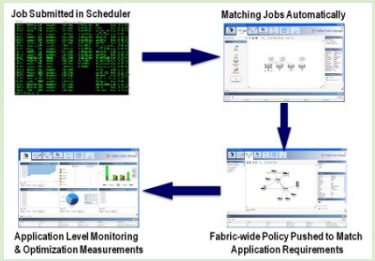
GPUDirect

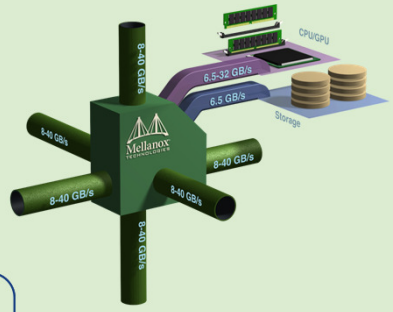




30+% Boost

Accelerating GPU Communications





simula research laboratory

Platform Computing

Adaptive COMPUTING

Altair Engineering

80+% Boost

Maximizing Network Utilization Through Routing & Management (3D-Torus, Fat-Tree)

ANNOUNCING

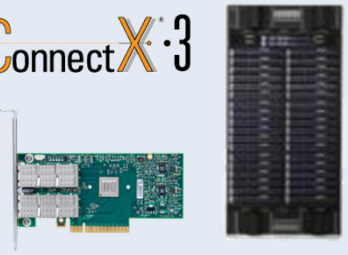
FDR INFINIBAND TECHNOLOGY

THE NEXT GENERATION OF HIGH-PERFORMANCE SCALABLE CONNECTIVITY

- 15%-45% lower latency
- 56Gb/s throughput
- Higher scalability
- Maximum Reliability

SwitchX™

ConnectX®-3



Highest Throughput and Scalability With InfiniBand FDR

Thank You

PAVING THE ROAD TO EXASCALE

ADVANCING NETWORK PERFORMANCE, EFFICIENCY, AND SCALABILITY.