

Optics in Future Clusters

Michael Kagan
CTO, Mellanox

Hot Interconnects, August 2009



InfiniBand Roadmap Development



TOP 500[®] - #1
SUPERCOMPUTER SITES



TOP 500[®] - #2
SUPERCOMPUTER SITES

ConnectX[®] PCI EXPRESS[®] 2.0



TOP 500[®] - #3
SUPERCOMPUTER SITES



PCI EXPRESS[®]

PCI EXPRESS[®]



Enabling next generation HPC systems
Scalable and high efficient networking

2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011

InfiniBand - Connecting The Most Powerful Systems



4K nodes
130K cores



4K nodes
63K cores



National Aeronautics and Space Administration



2K nodes
30K cores



6.5K nodes
51K cores

4.5K nodes



3K nodes
26K cores



Networking Speeds and Optics



Fiber
(On chip)



3m copper
>3m Fiber



8m copper
>10m Fiber



20m copper
>20m Fiber

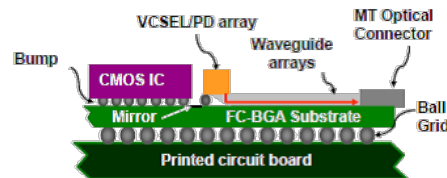
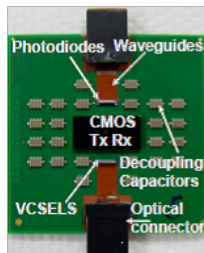


Copper

Enabling next generation HPC systems
Requires cost-effective optics solutions

2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013

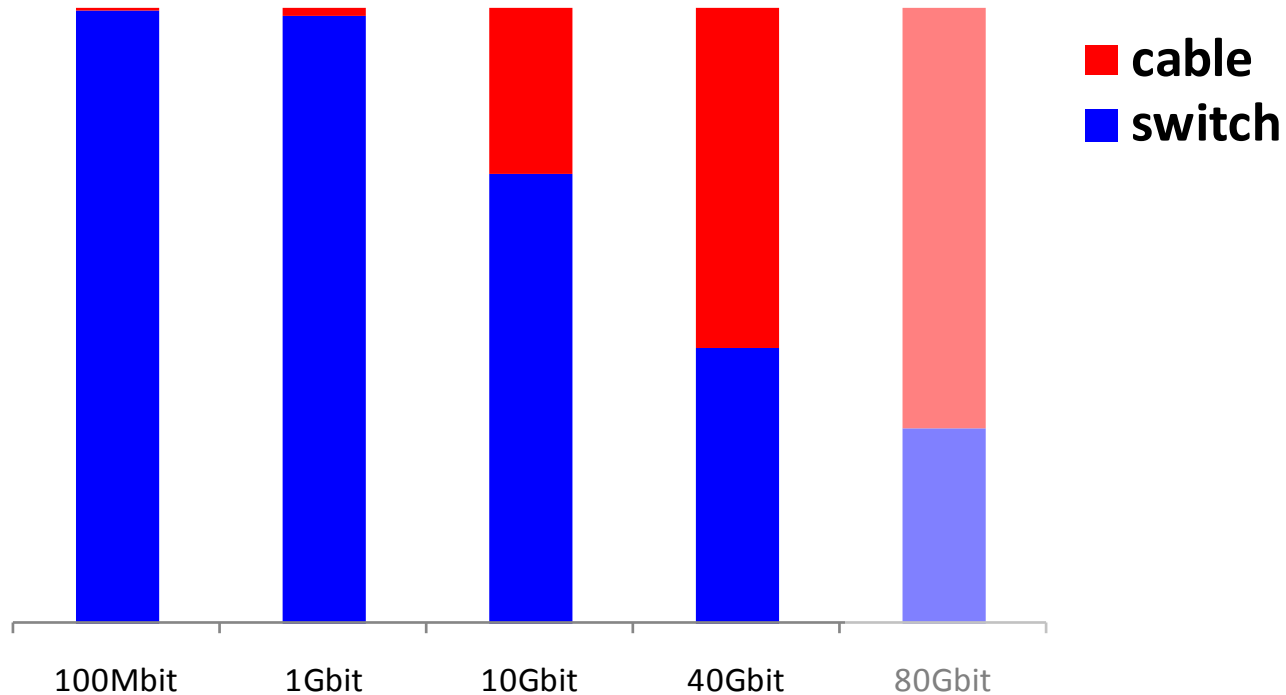
- The interconnect is a critical factor for high system utilization
 - At any scale
- Future networking speeds will require optics-dominant infrastructure
 - Cables cost becomes significant part of infrastructure
 - Optical PCBs, On-chip optics



- Old dogs must learn new tricks
 - Brain to compensate for increased cost
 - Adaptive routing
 - Congestion control



Cabling – the impact



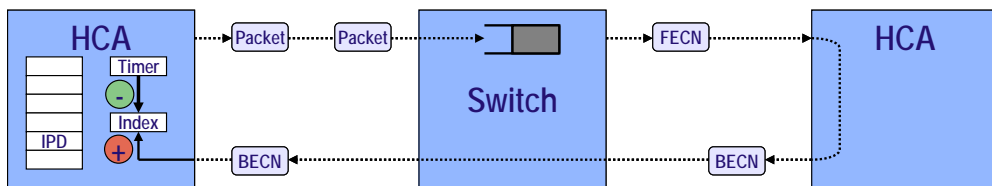
Optics
Copper

Optics
Copper

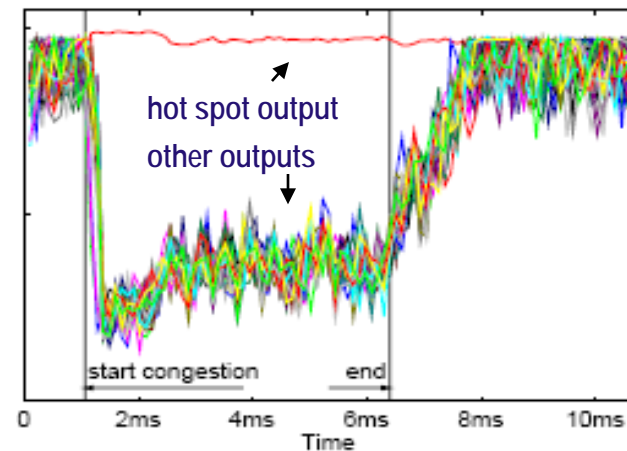
Optics
Copper

InfiniBand Hardware Congestion Control

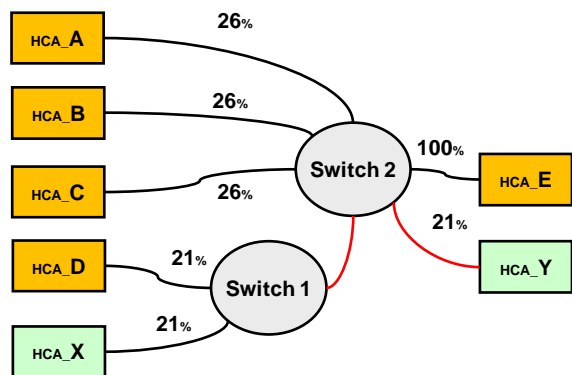
- HW congestion control allows real-time response
- Ensuring maximum network efficiency



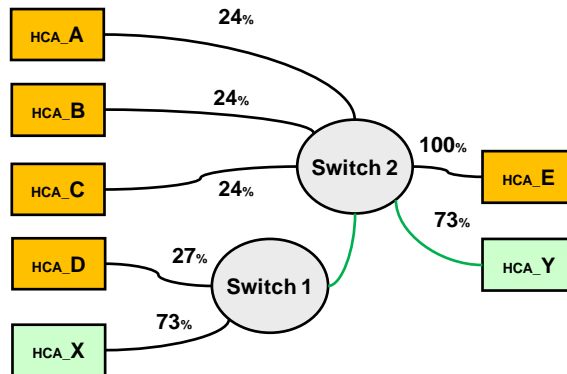
Before congestion control



Before Congestion control

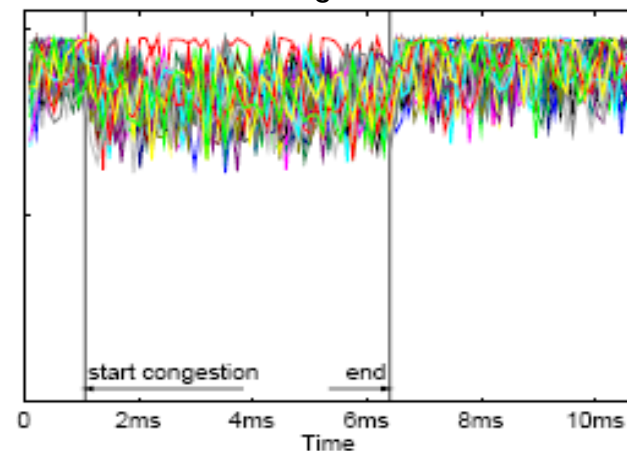


After Congestion control



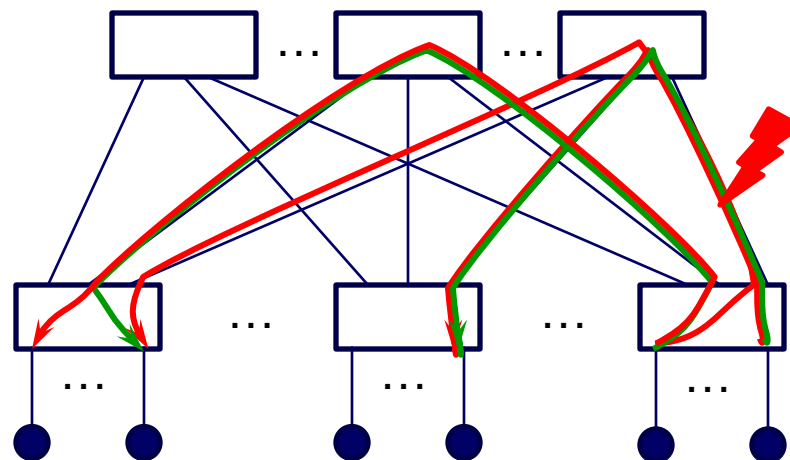
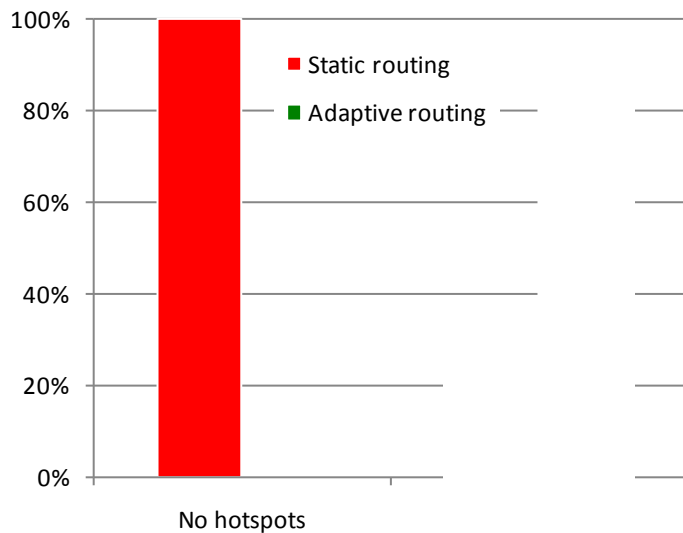
Actual Results

After congestion control



"Solving Hot Spot Contention Using InfiniBand Architecture Congestion Control
IBM Research; IBM Systems and Technology Group; Technical University of Valencia, Spain

- 220 nodes, 2-stage CLOS network
- Mellanox InfiniBand HCAs and switches



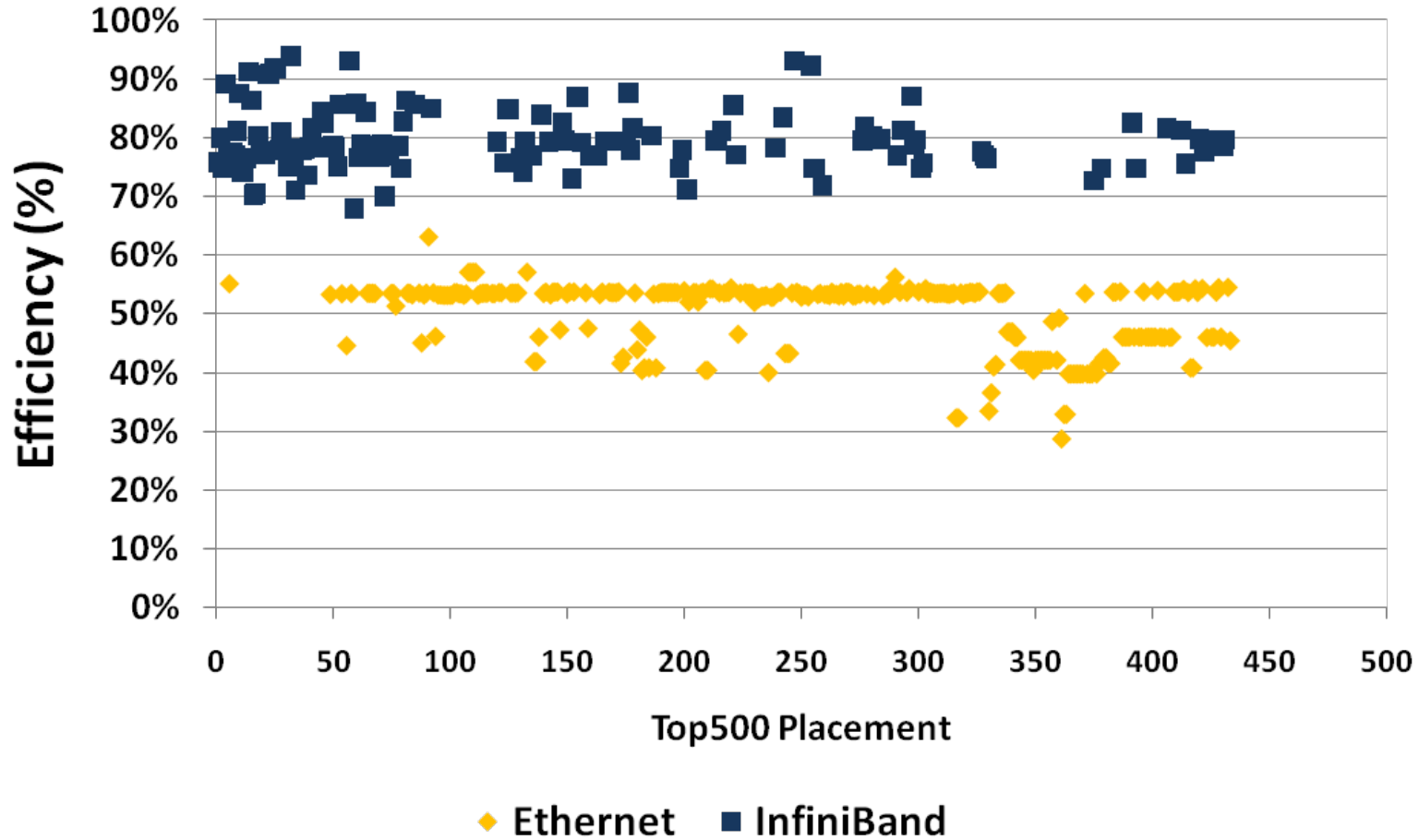
- Maximizes network efficiency
 - For random traffic or multiple application scenarios
 - Switches dynamically re-routes traffic to alleviate congested ports
- Fast path modifications, no overhead throughput

220 server node system, Mellanox InfiniBand HCAs and switches

Hot Spots Configuration	No Adaptive Routing		With Adaptive Routing	
	Average Bandwidth	Minimum Bandwidth	Average Bandwidth	Minimum Bandwidth
None	100.0%		NA	NA
2:41,3:4,4:1	79.4%	37.3%	99.7%	99.6%
2:35,3:5,4:1	80.3%	37.1%	99.7%	99.0%
2:30,3:10,4:1	75.9%	36.5%	99.7%	99.5%
2:41,3:5,4:1	82.4%	37.7%	99.7%	99.4%

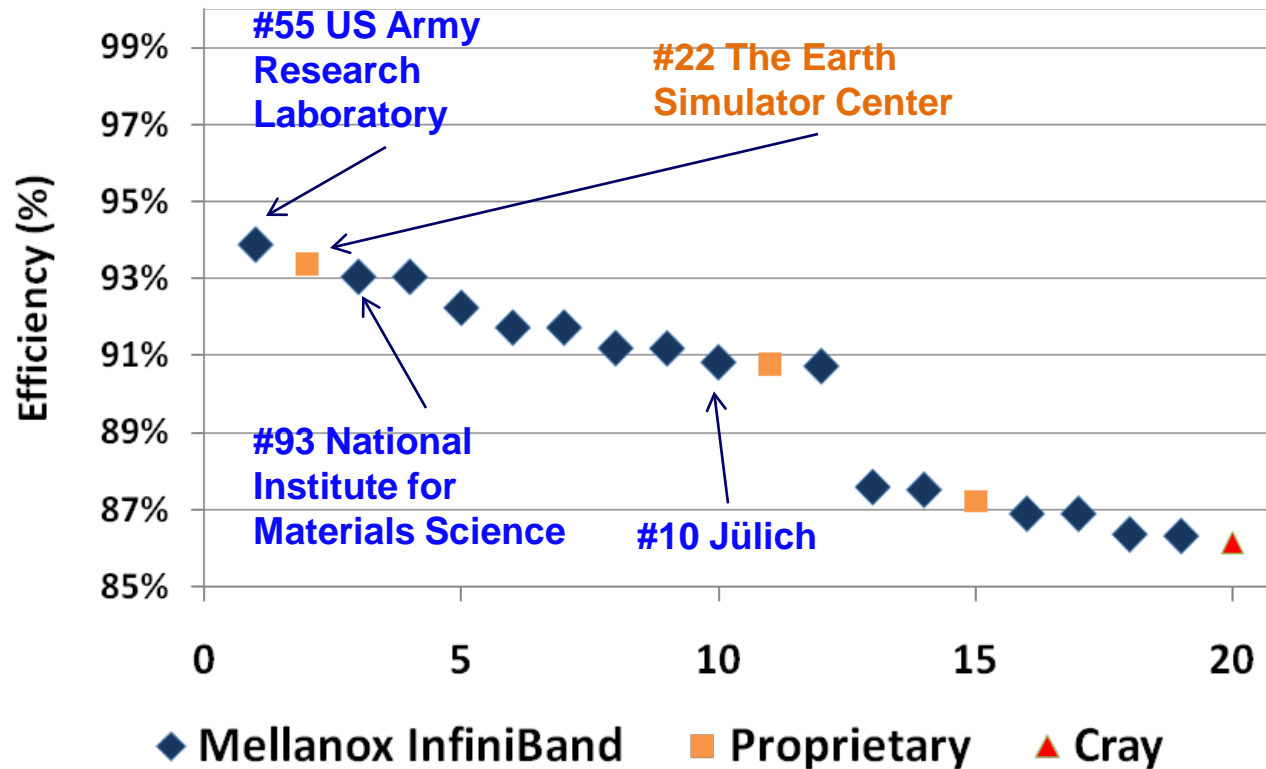
- Each case represent a different congestion scenario
 - For example 2:10 means 10 links where each shares 2 connections)
- Adaptive routing increases the network efficiency by an average of 25.5% for the average bandwidth and the minimum bandwidth by an average of 167.5%

Top500 Efficiency Comparison



The 20 Most Efficient Top500 Systems

The 20 Most Efficient Top500 Systems



InfiniBand – The Most Efficient Systems at Any Scale



Automotive



**Computational Chemistry
Cancer research**



CH-53 HEAVY LIFT

**Sikorsky CH-53K program
Reducing simulations
duration from 4 days to
several hours**



Education



**Personal
supercomputing**

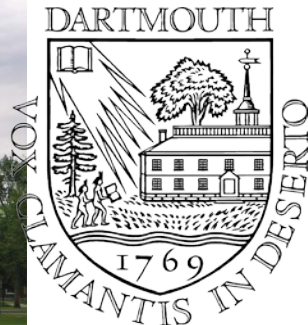
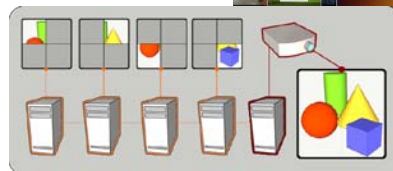


HP

Visualization, rendering



Cray



**Energy dissipation in silicon nano-
resonators by using molecular
dynamics**

High-end HPC



Enterprise HPC



Entry-level HPC



Performance

100%
Increase

Complete Scalable I/O Consolidation Solutions

TCO

50%
Reduction

Energy Costs

67%
Reduction

Infrastructure

62%
Saving

Thank You

