Maximize Clusters' Performance

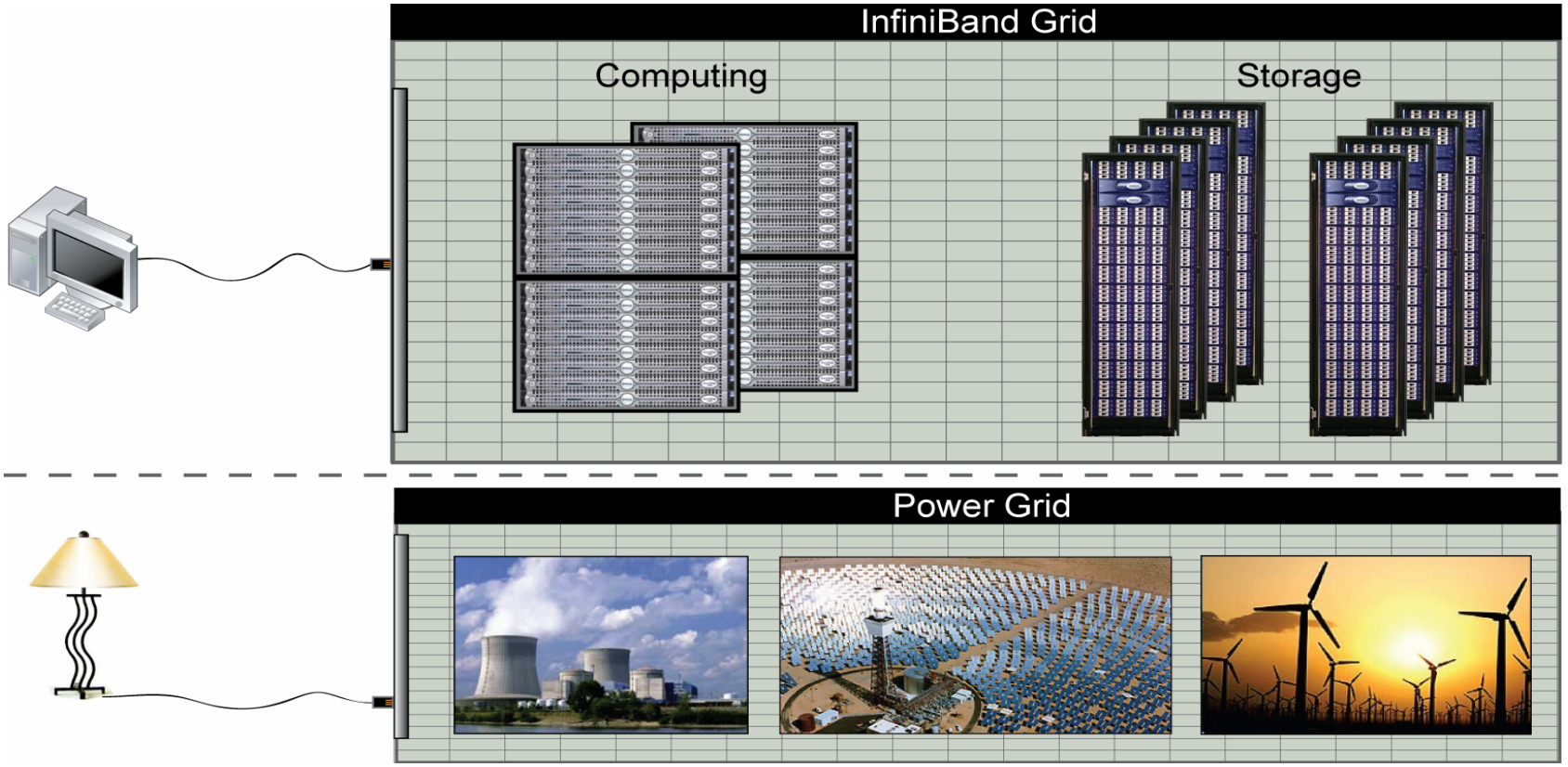
The Next Level of Clustering

Michael Kagan
VP Architecture
Mellanox Technologies

Agenda

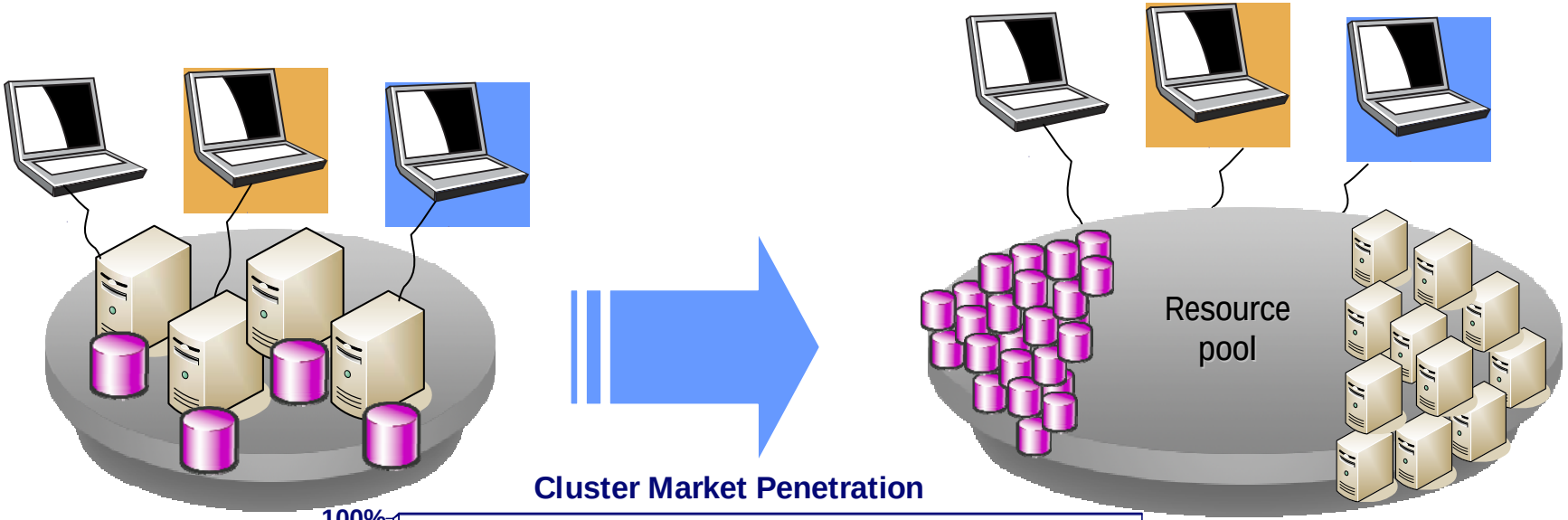
- The vision
- The interconnect challenge
- Industry solution
- Products and roadmap

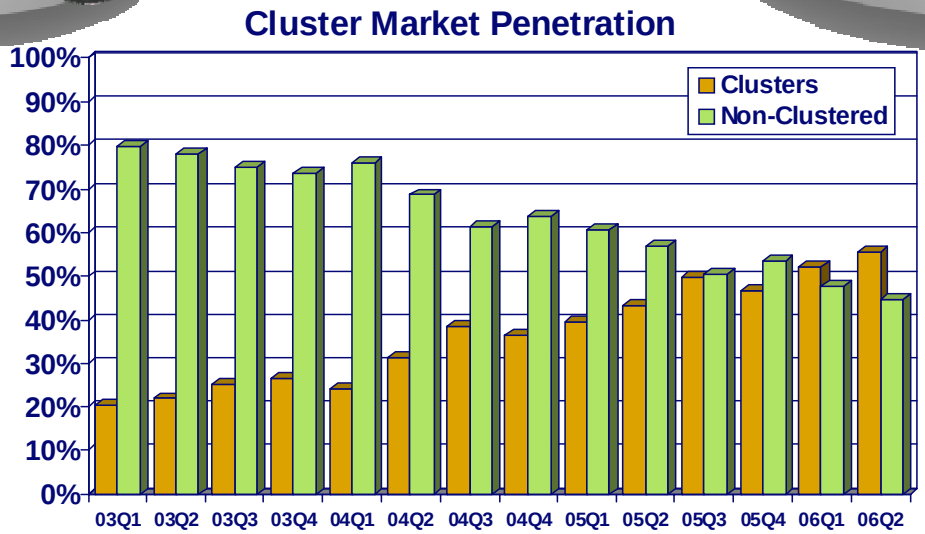
The Vision – Clustered Grid



Deliver Services to Clients

From equipment warehouse to service provider





Clustering – the grid backbone technology

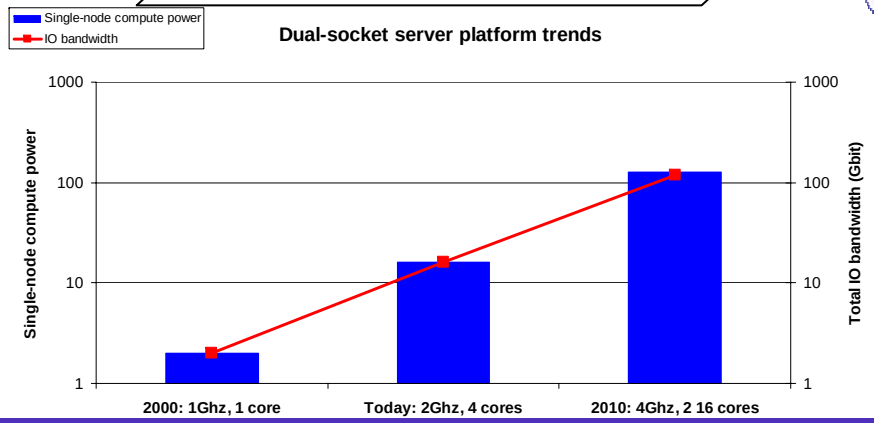
IO Performance Growth

Multi Core CPUs → More applications per server and I/O

SAN Adoption → More traffic per I/O with server I/O consolidation

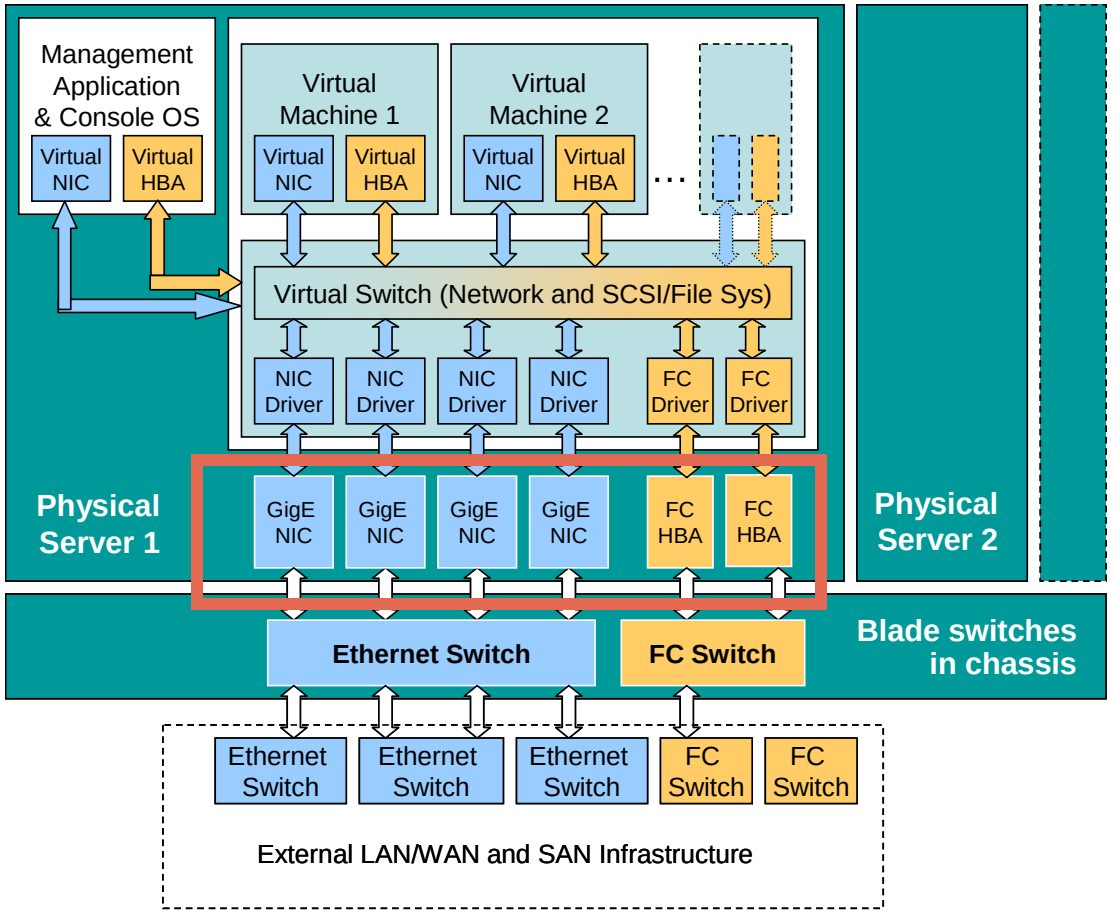
Shared Resources → I/O capacity per server dictated by the most demanding apps

10Gb/s+ connectivity for each server



Multi-core CPUs mandating 10Gb/s+ connectivity

Typical Deployments With GigE and FC



- 6-8 I/O adapters per server
  - Bandwidth
  - Functions
- Not feasible Blade Servers
  - Limited PCI slots, backplane traces
- High CAPEX, OPEX
  - More ports/server
- Management challenge
  - Multiple networks
  - Multiple management domains

Traditional approach does not scale

I/O Overloaded?

IO Delivery challenges

- High bandwidth, low latency
- Scalability
- Low power consumption
- Virtualization, dynamic load balancing
- Agility, quicker business results

Server and storage I/O takes on a new role

IO Infrastructure – Requirements

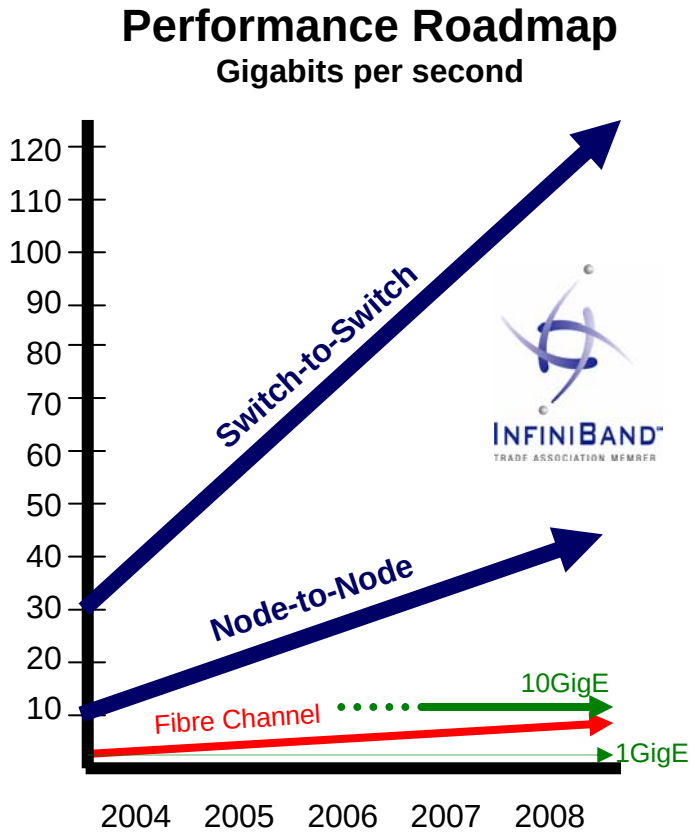
Network Layer	Data Center requirement	Ethernet	TCP	InfiniBand
Transport (layer 4)	**Scalability to 100Gbit** **Virtual interface support** **Point to multipoint communication** **QoS-aware interface to client** **High Availability features**		**Hard** **No** **No** **No** **Poor**	**Easy** **Yes** **Yes** **Yes** **Good**
Data (layer 2)	**Network convergence (virtual networks)** **Quality of Service** **Lossless network** **Congestion management** **Optimal routing**	**No** **No** **No** **No** **No**		**Yes** **Yes** **Yes** **Yes** **Yes**

Source: Intel/CISCO, August 2005

Legacy interconnect does not fit Grid requirements

InfiniBand – The Grid Interconnect

- Top Performance at lowest price
  - Defined for low-cost implementation
  - Up to 120Gbit port speed
- Scalable
  - Tens-of-thousands of nodes
  - Multi-core servers
- Low CPU overhead
  - RDMA and Transport Offload
- Service Oriented I/O
  - Quality-of-Service
  - Virtualization
  - I/O consolidation
- Established software ecosystem



Industry standard for grid interconnect

Delivering Service Oriented I/O

End-to-End Quality Of Service	▪ Congestion control @ source ▪ Resource allocation
I/O Consolidation	▪ Multiple traffic types ▪ Up to 40% power savings
Optimal Path Management	▪ Packet drop prevention ▪ Scale @ wire speed
Dedicated Virtual Machine Services	▪ Virtual machine partitioning ▪ Near native performance

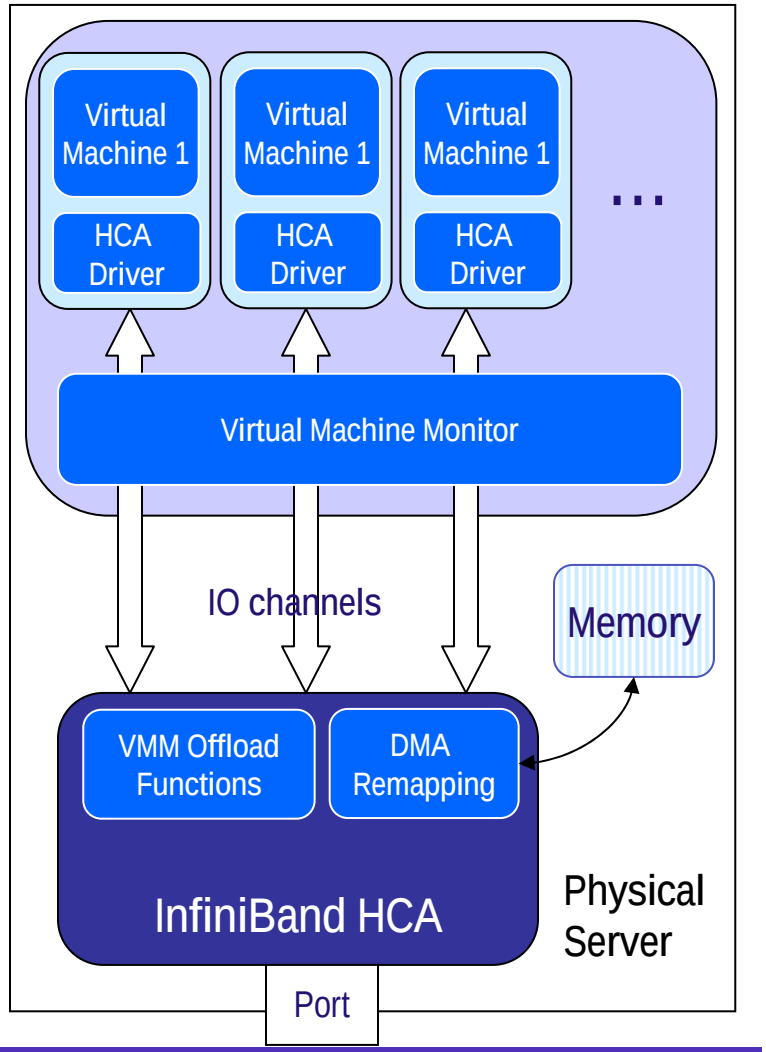
- Clustering
- Communications
- Storage
- Management

Provisioned IO services on a single wire

Channel I/O Virtualization on Server

- Hardware-based I/O virtualization
  - Isolation & protection per VM
  - DMA remapping / virtual address translation
  - Resource provisioning
  - Hypervisor offload (switching, traffic steering)
- Supports current and future servers
  - Programmable
  - IOTLB* for future I/O MMU chipsets
  - Intel VT-d IOV, PCI-SIG IOV
- Better resource utilization
  - Frees up CPU through hypervisor offload
  - Enables significantly more VMs per CPU
- Native OS performance for VMs
  - Eliminates VMM overheads

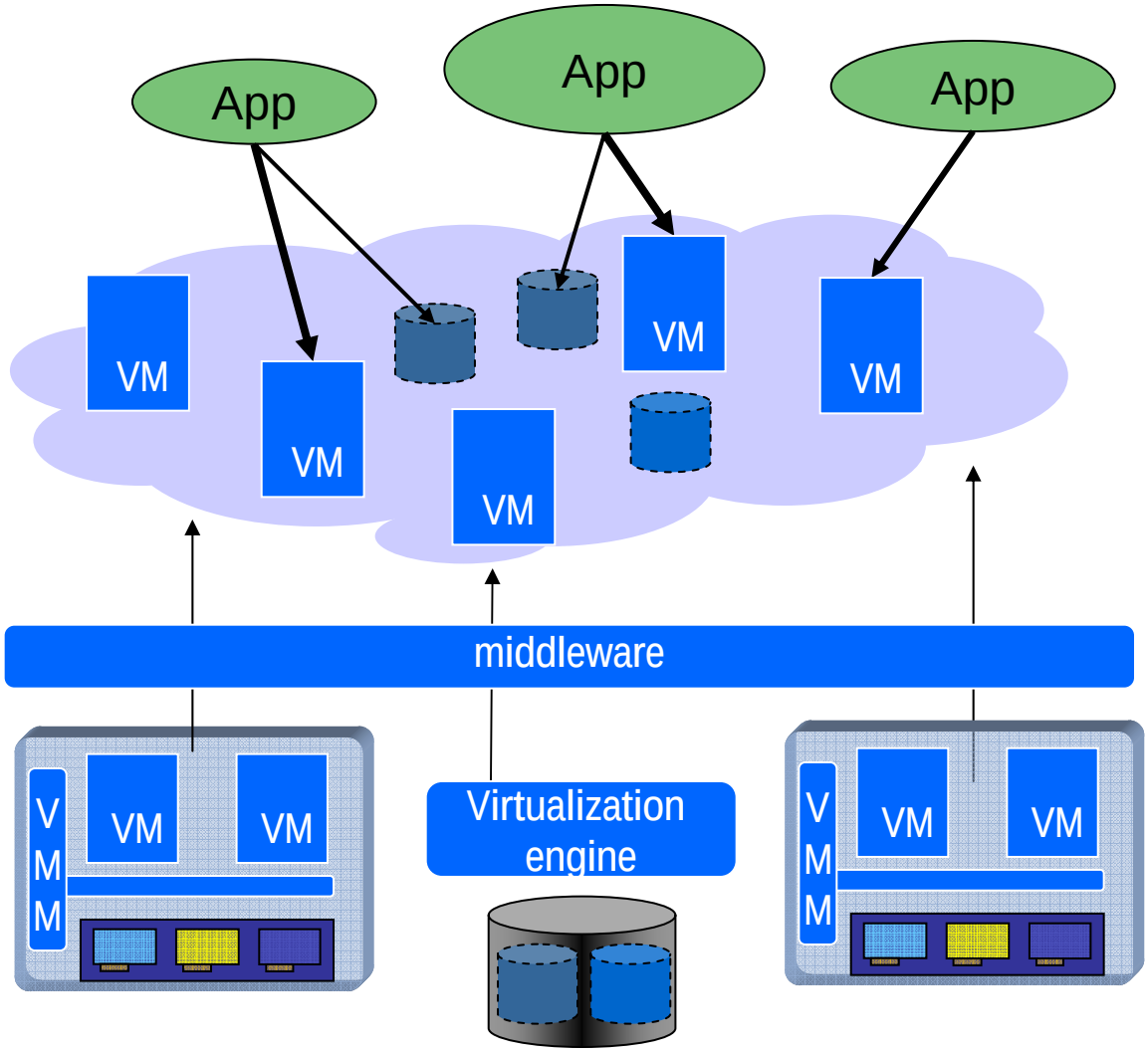
* I/O Translation Look-aside Buffer



Channel IO– deliver IO services to consumer

Cluster as a Pool of Resources

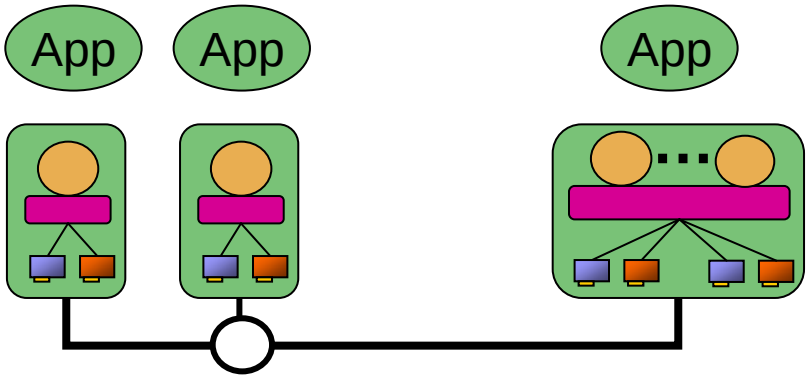
- Each VM (“container”) represents available compute resource
- Applications are run in any available container
- Storage is disassociated from the platform
- Applications become readily transportable



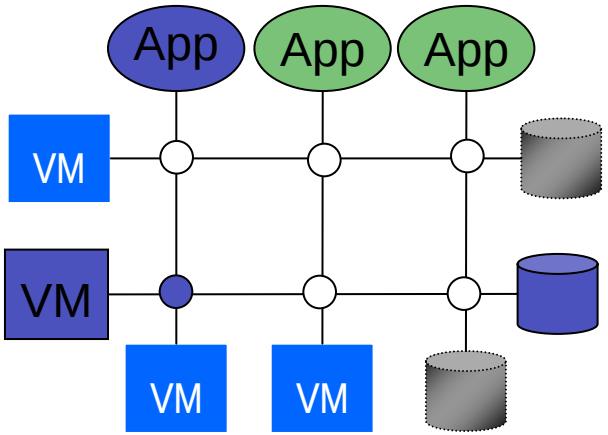
Unconstrained service delivery

Clusters – Deliver Service to Consumer

Server-centric view: servers interconnected by networks

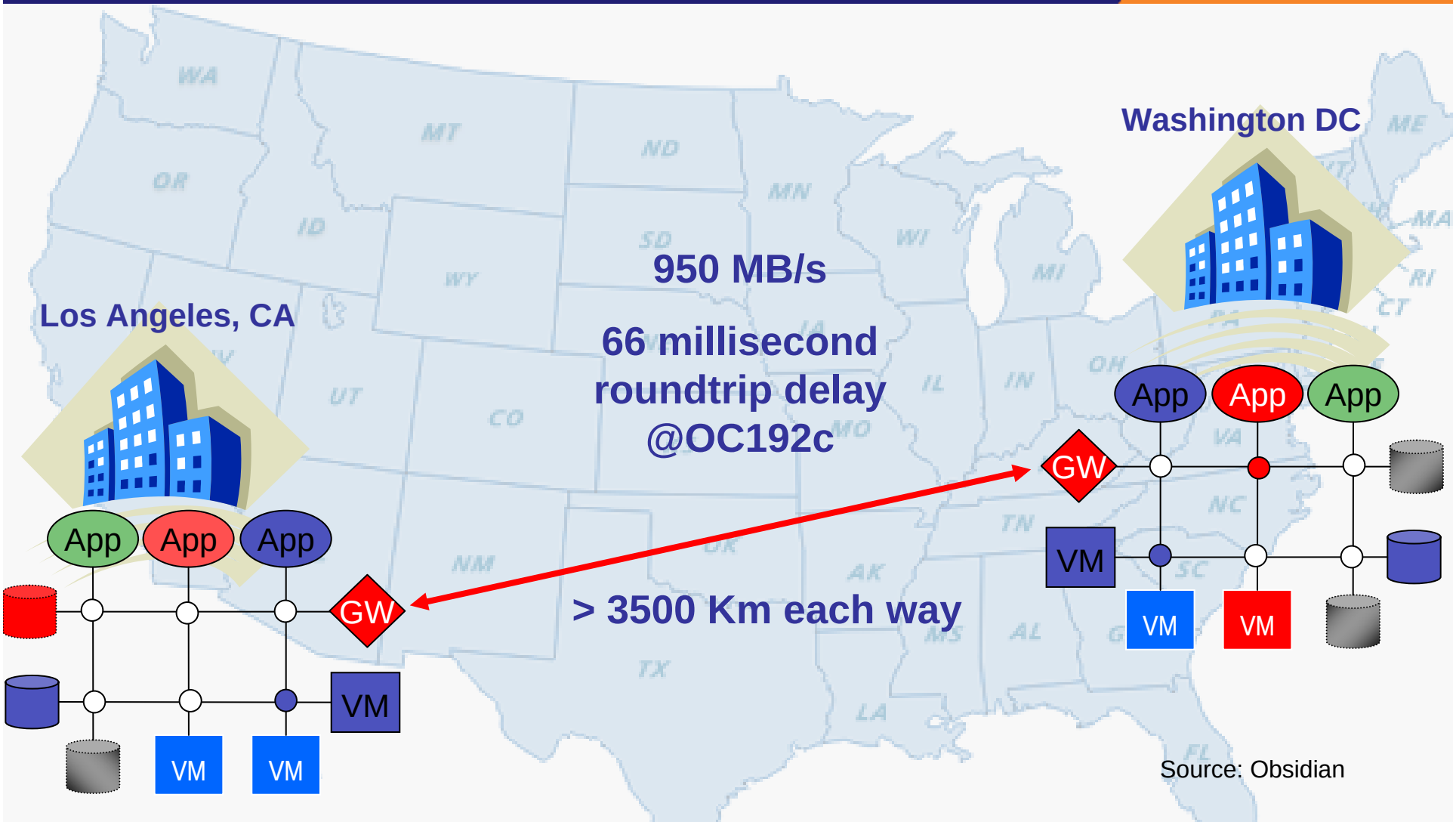


Service-centric view: applications are provisioned from a pool of resources



Interconnect – service delivery enabler

“Global architecture”



Location-agnostic solution – from microns to miles

Mellanox Technologies

- A global leader in semiconductor solutions for server, storage and embedded connectivity
- Leading provider of low-latency and high-bandwidth InfiniBand solutions
  - Up to 20Gb/s NIC, up to 60Gb/s Switch
  - 1.7M ports shipped (Dec 2006)
  - 2.25us latency production, ~1us latency in 2007
  - 3W power consumption per HCA port
- Efficient and scalable I/O
  - 4500-node cluster in production
  - 10K+ nodes clusters being installed
- Price-performance-power leader
- Converges clustering, communications, storage solutions

Grid interconnect building blocks available today

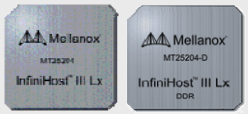
Product Roadmap

4th Generation

ConnectX™ PCI EXPRESS Gen2 Adapter

Two 10/20/40 Gb/s IB or Two 1/10Gb/s Ethernet

3rd Generation



20, 40Gb/s Total PCI EXPRESS Adapter

40, 80Gb/s Total PCI EXPRESS Adapter

480, 960Gb/s Total

Switch

2nd Generation

40Gb/s Total PCI-X **Adapter**

160Gb/s Total

Switch

1st Gen.

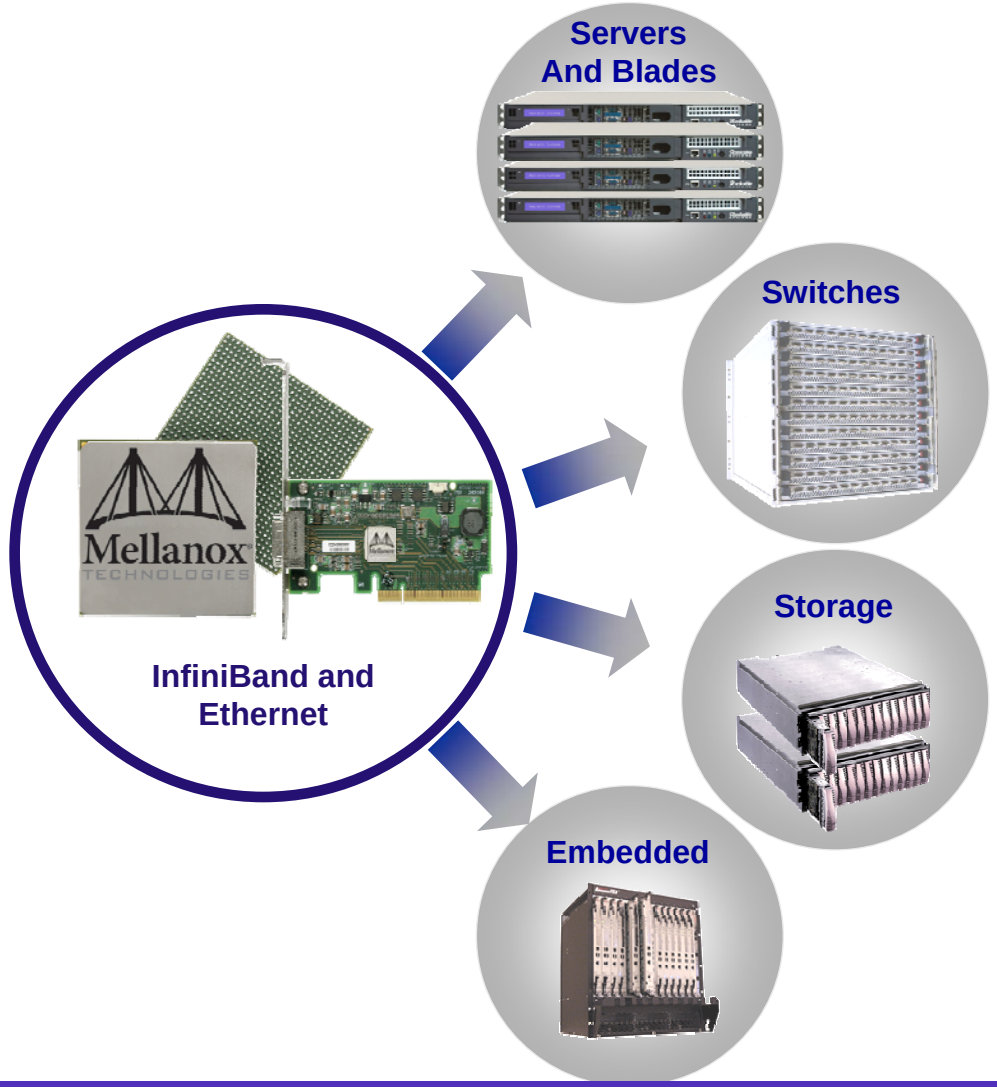
40Gb/s Total PCI CONVENTIONAL

Adapter + Switch

2000 2001 2002 2003 2004 2005 2006 2007

Consistent execution on the roadmap

Interconnect: A Competitive Advantage



End-Users

Enterprise Data Centers

- Clustered Database
- Customer Relationship Management
- eCommerce and Retail
- Financial
- Web Services

High-Performance Computing

- Biosciences and Geosciences
- Computer Automated Engineering
- Digital Content Creation
- Electronic Design Automation
- Government and Defense

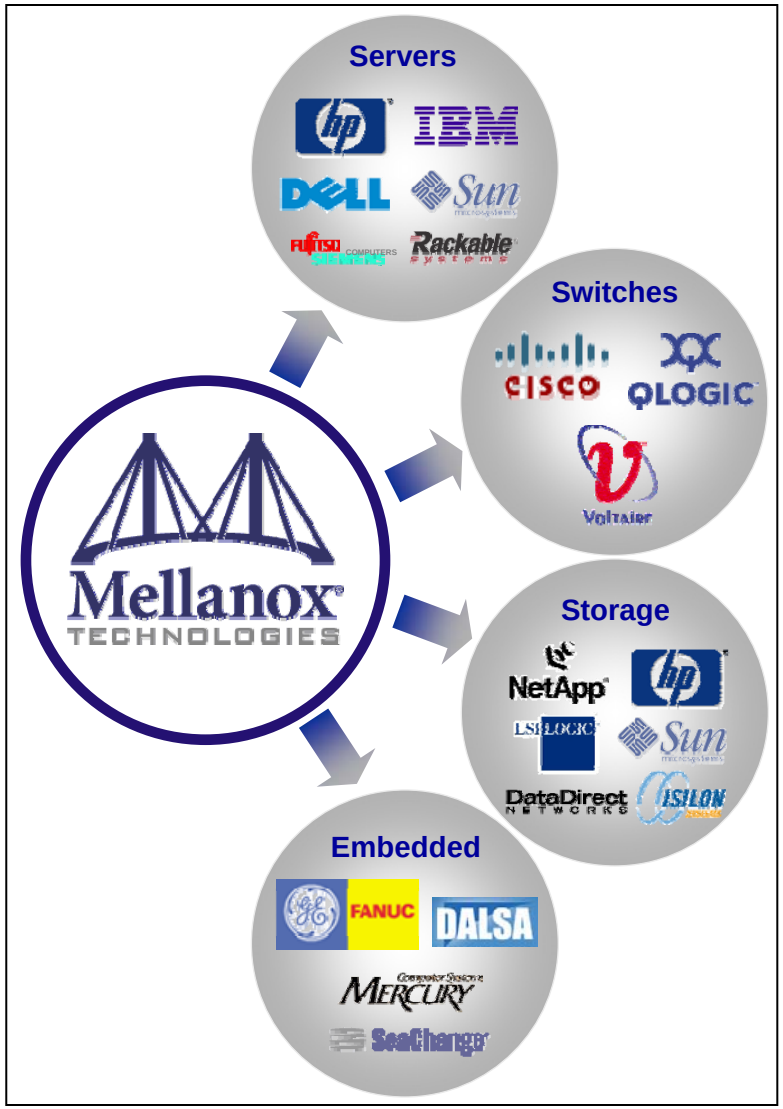
Embedded

- Communications
- Computing and Storage Aggregation
- Industrial
- Medical
- Military

Complete system solutions for all market segments

Leading Customers / Growing Markets

Hardware OEMs



Software Partners

ORACLE
IBM DB2
Microsoft
vmware
Novell
redhat
SOLARIS
HP-UX 11i
AIX L
Autodesk
ANSYS
LSTC Livermore Software Technology Corp.

End-Users

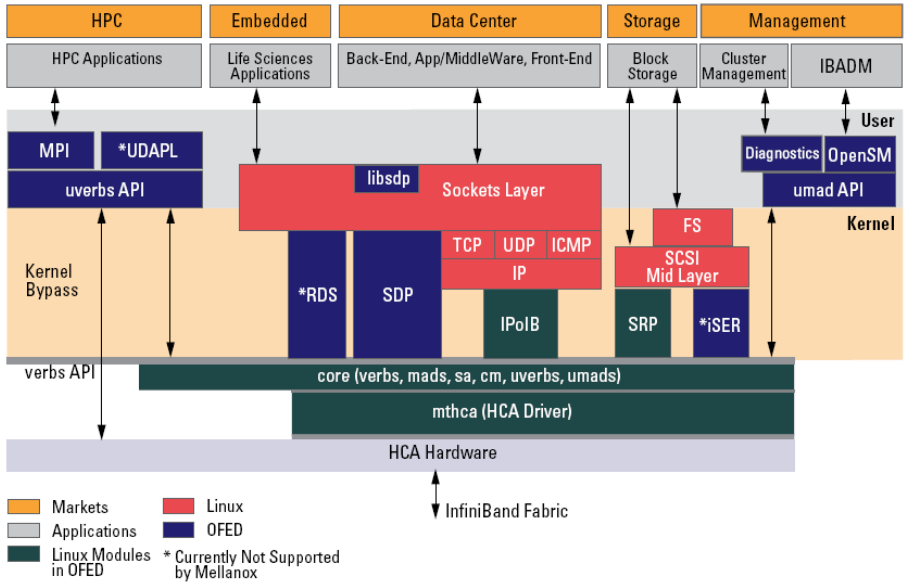
Enterprise Data Centers

High-Performance Computing

Embedded

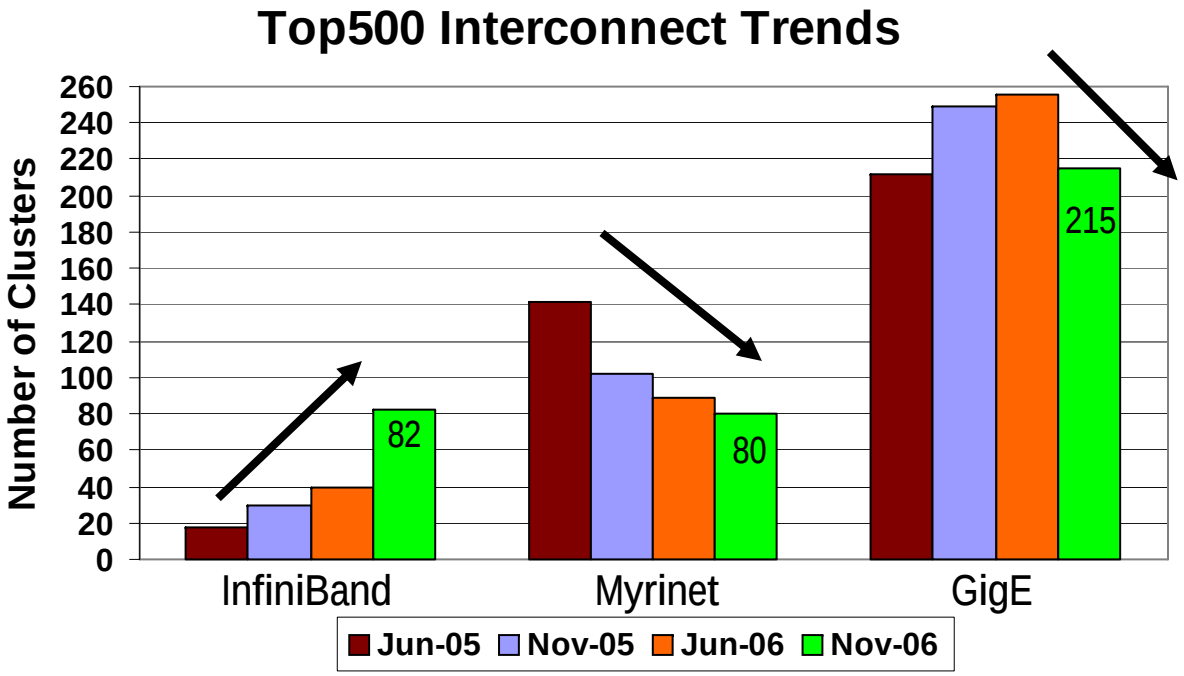
InfiniBand Software Support

- Industry-wide development of standard SW stack
- Supported by all Linux distributions
- Granted WQHL by Microsoft



Full IO solution on all major OS distributions

HPC – the Early InfiniBand Adopters



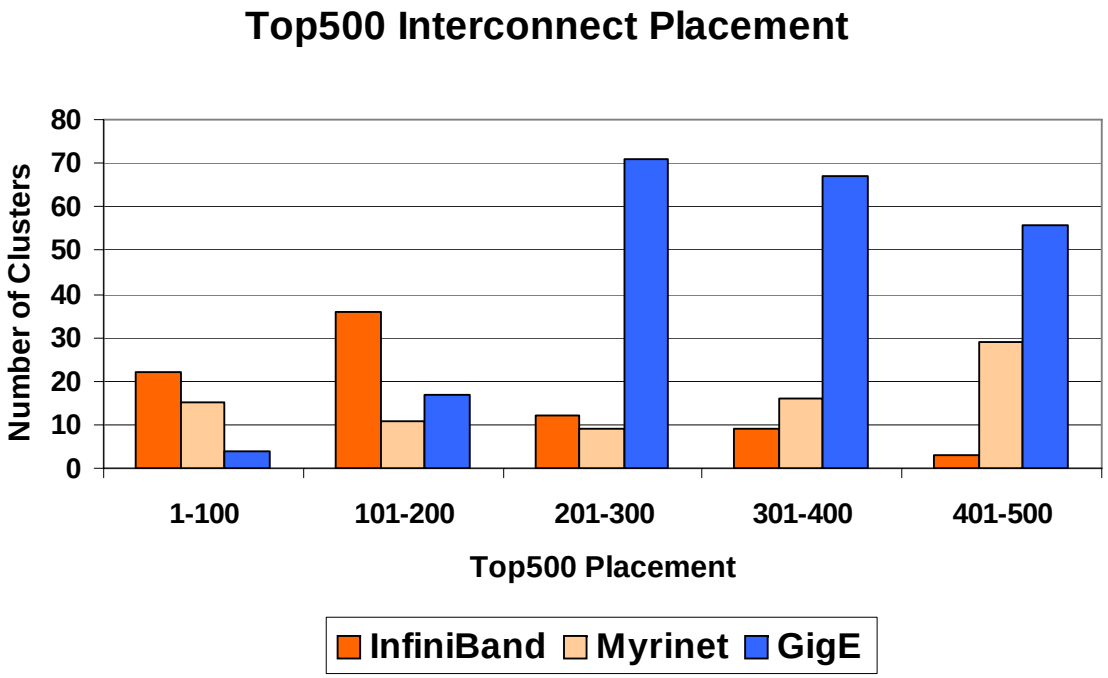
Growth rate from June 06 to Nov 06

- InfiniBand: +105%
- Myrinet: -10%
- GigE: -16%

- 105% growth from June 2006
- 173% growth from Nov 2005

InfiniBand – the only growing interconnect

Top500 Interconnect Placement Nov 06

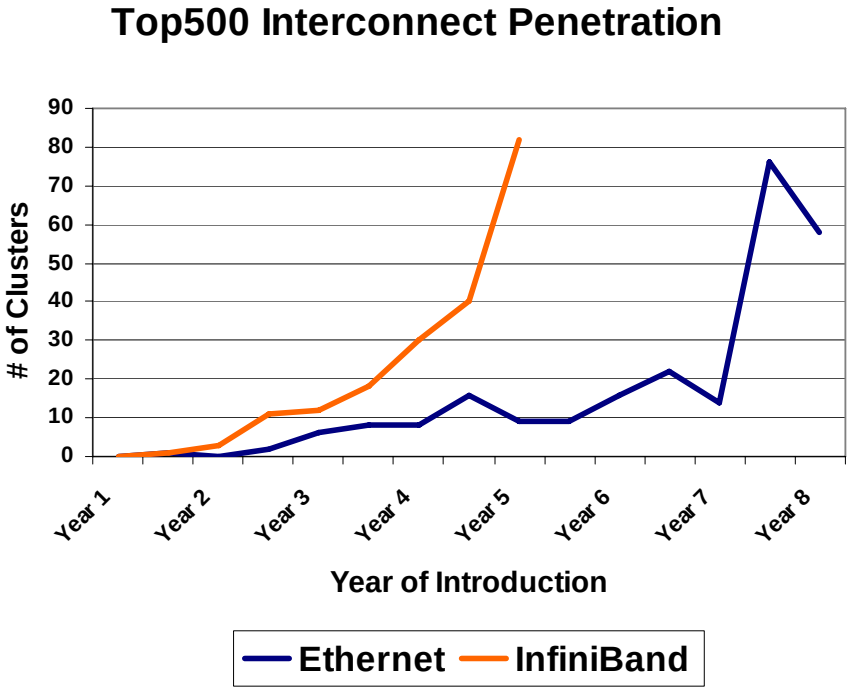


Source: TOP500 SUPERCOMPUTER SITES

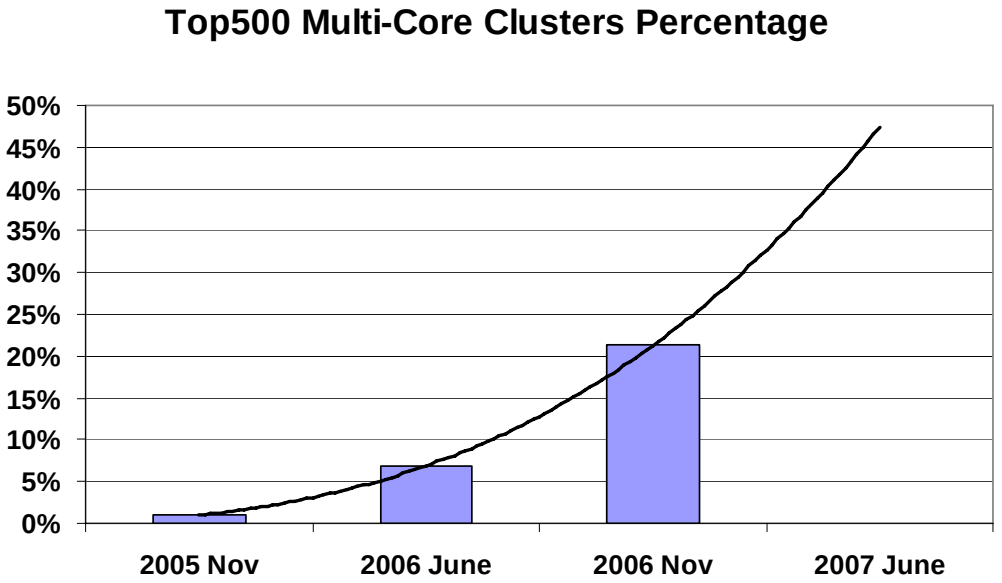
- InfiniBand is the preferred high performance interconnect
  - Connecting the most powerful clusters
  - InfiniBand is the best price/performance connectivity for Petascale clusters

InfiniBand – the dominant HPC interconnect

Top500 Statistics



- InfiniBand adoption is faster than Ethernet
  - Ethernet year 1 – June 1996
  - InfiniBand year 1 – June 2003



- Multi-core will dominate the list in 2007
  - Native multi-core

The fastest growth in top500 interconnect history

Mellanox Superior Application Performance

Automotive	LSTC LS-DYNA	26% better than Qlogic 85% better than Myrinet 115% better than GigE
	ESI PAM-CRASH	300% better than GigE
Oil and Gas	Schlumberger Eclipse	55% better than Myrinet
Fluid Dynamics	Fluent	145%-1400% better than GigE 15% better than Qlogic
	Exa PowerFLOW	24% better than Myrinet
	CD-Adapco STAR CD	29% better than Myrinet
Mathematical Modeling	Wolfram gridMathematica	57% better than GigE
Digital Media	Autodesk	470% better than GigE
Financial	Wombat	120% better than GigE
Data Base	Oracle	300% better than GigE

Mellanox InfiniBand wins on real application

Personal Supercomputers (PSC)

From top 500 to technical computing

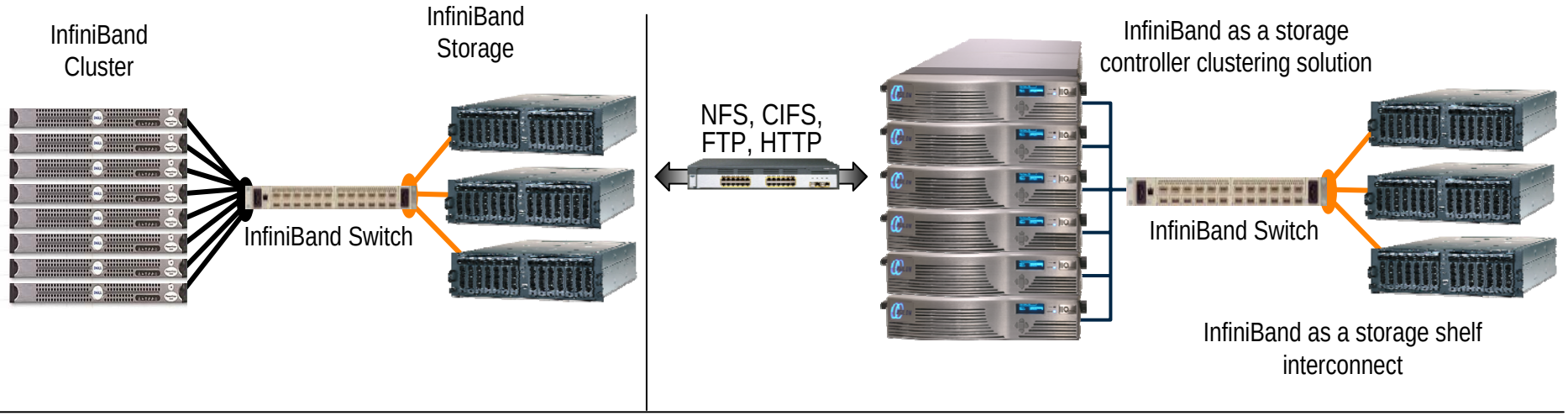
- Driving supercomputing to the masses
  - Maximum performance, Minimum cost
  - Easy to use, Turnkey cluster
  - Fits into “cubicle” environment
  - Standard power, quiet operation
  - Ability to scale efficiently

Maximum Performance
At Your Fingertips

Personal Supercomputing

Cluster waterfall from compute room to desktop

InfiniBand Storage Interconnect



- Native InfiniBand storage servers
  - Optimal Performance, Power Savings and TCO
  - No gateway bottlenecks
  - Ultimate scalability
  - Service Oriented I/O for consolidation
- Storage controller clustering and failover solution
- Used to connect to the storage disk arrays

Storage servers and backend clustering

InfiniBand Storage Solutions

InfiniBand Backend Clustering and Failover

Native InfiniBand Visualization Storage System

Native InfiniBand Solid State Storage

Native InfiniBand Block Storage Systems

HP StorageWorks
Scalable File Share (HP SFS)

Native InfiniBand Clustered File System

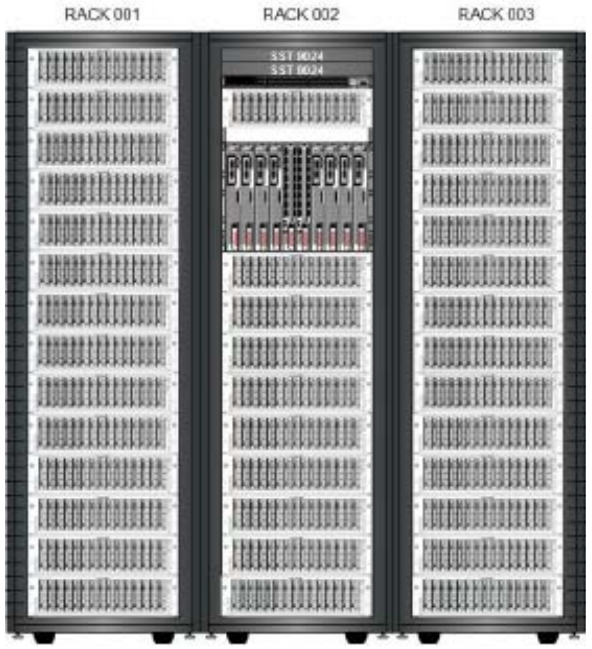
Native InfiniBand Clustered File Storage Software

Native InfiniBand Block Storage Software

Multiple storage vendors deploy InfiniBand today

IB Storage = Best Price/Performance

- TPC-H –Industry Standard Benchmark
  - Actual database transaction profiling
  - Results must be qualified and approved
- InfiniBand Storage delivers best price/performance at 1TB class
- 8 4-way dual-core AMD Opteron Servers
- 37 x PANTA Systems Storage Array Modules
  - 518 x 250GB 7200rpm SATA HDD
  - Total Storage : 129500 GB
- Turnkey server/storage solution for data warehousing applications



ORACLE DATABASE 10g

1,000 GB Results

Rank	Company	System	QphH	Price/QphH	System Availability	Database	Operating System	Date Submitted
1	PANTA	PANTA Systems PANTAmatrix	59,353	24.94 US $	04/15/07	Oracle Database 10g release2 Enterpise Editi	Red Hat Enterprise Linux 4 AS	10/23/06

InfiniBand solutions win TPC-H 3rd year in a row

Cluster Evaluation Center

- Essential platform for customer evaluation
  - Latest and greatest InfiniBand hardware and software
  - InfiniBand based storage
    - NFS over RDMA
  - AMD and Intel based platforms
- Available clusters
  - Mellanox cluster center: Intel quad core, AMD dual core rev E
  - Colfax International: AMD dual core rev F
- Available for customers' evaluation
  - Developments, testing and evaluation
  - Free of charge

InfiniBand clusters available for evaluation

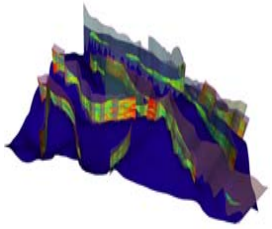
Summary

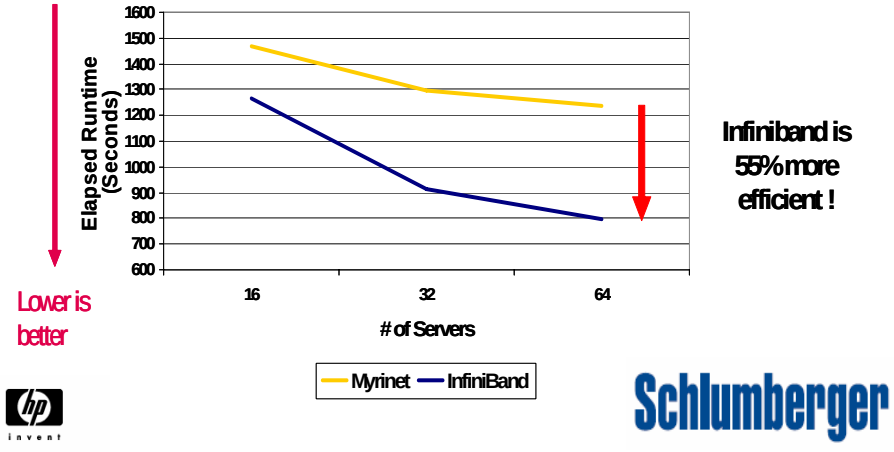
- InfiniBand is a clustering interconnect of choice
  - The only growing interconnect on top500
  - Dominates the top part of the top500
- InfiniBand supports major industry trend
  - Industry standard, supported by major OEMs and ISVs
  - Leading cost/performance solutions
  - Multi-core and multi-node scalability
  - Service-oriented IO delivery
  - Compute and storage services on a single wire
  - Transition from equipment warehouse to service provider
  - Clusters – the utility computing backbone
- InfiniBand clusters available for customers' evaluation
  - Free of charge
  - Multiple platforms

InfiniBand – the standard cluster interconnect

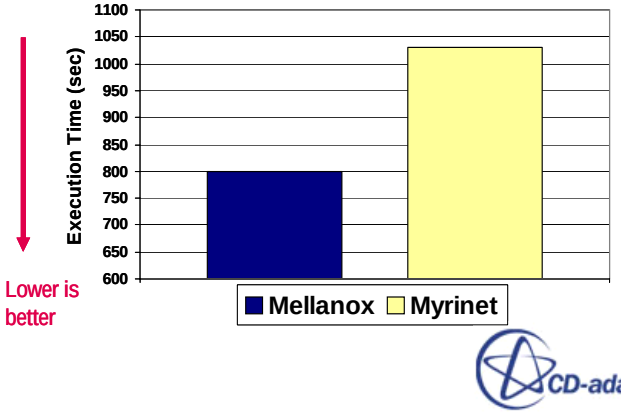
Real-life Applications

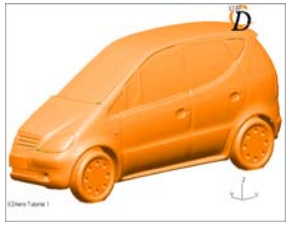
- ECLIPSE million cell model
  - HP DL145 2.6Ghz servers, single CPU
  - OS: SUSE 9





- CFD – CD-Adapco Star CD

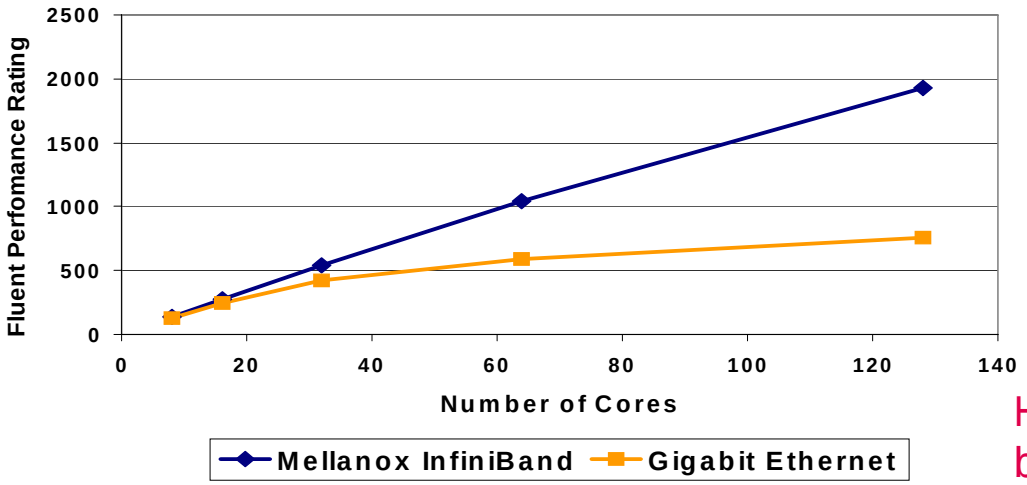




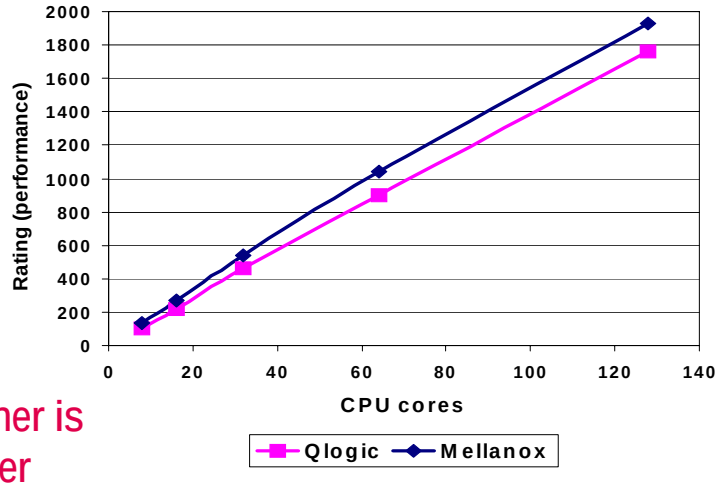
Benchmarks STAR-CD V3240/V3260

InfiniBand wins on real application

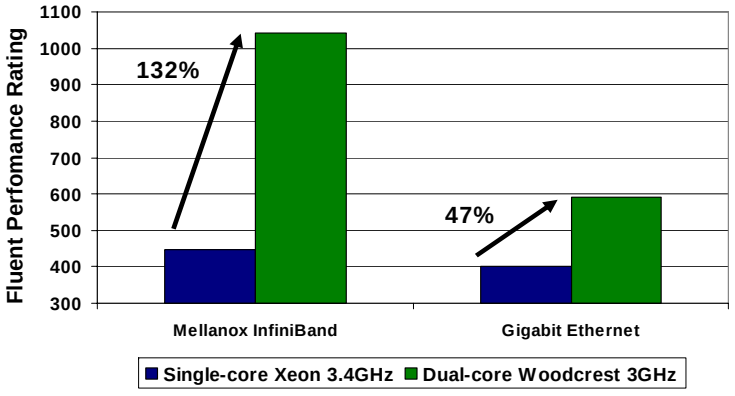
CFD – Fluent



Higher is better



- Great scaling from small to large cluster
- Multi-core environment demands InfiniBand

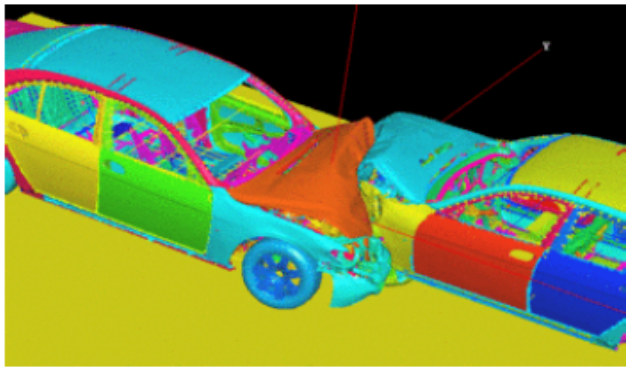


Highest performance, best scalability

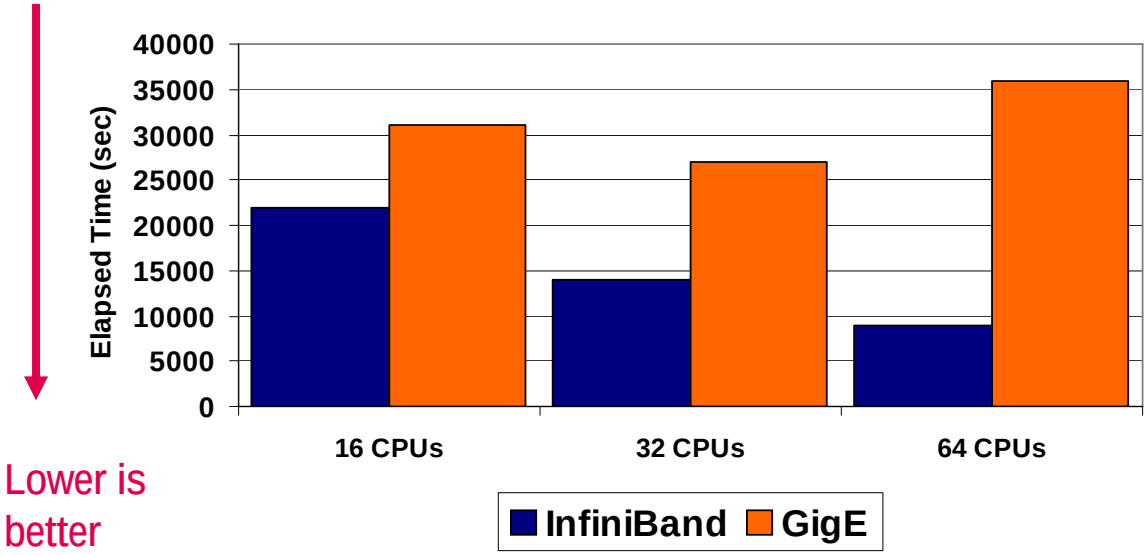
Crash – ESI PAM-CRASH

Bavarian Car-To-Car Model1.1 M elements, 145000 cycles

ESI GROUP



PAM-CRASH



Lower is better

"Gigabit Ethernet becomes ineffective with cluster size growth while Mellanox InfiniBand allows continued scalable speed up"

Highest performance, best scaling