MPI over InfiniBand

Gil Bloch, Architect

gil@mellanox.co.il

InfiniBand Trade Association

INFINIBAND℠
Trade Association

Sponsor Roster

Steering Roster

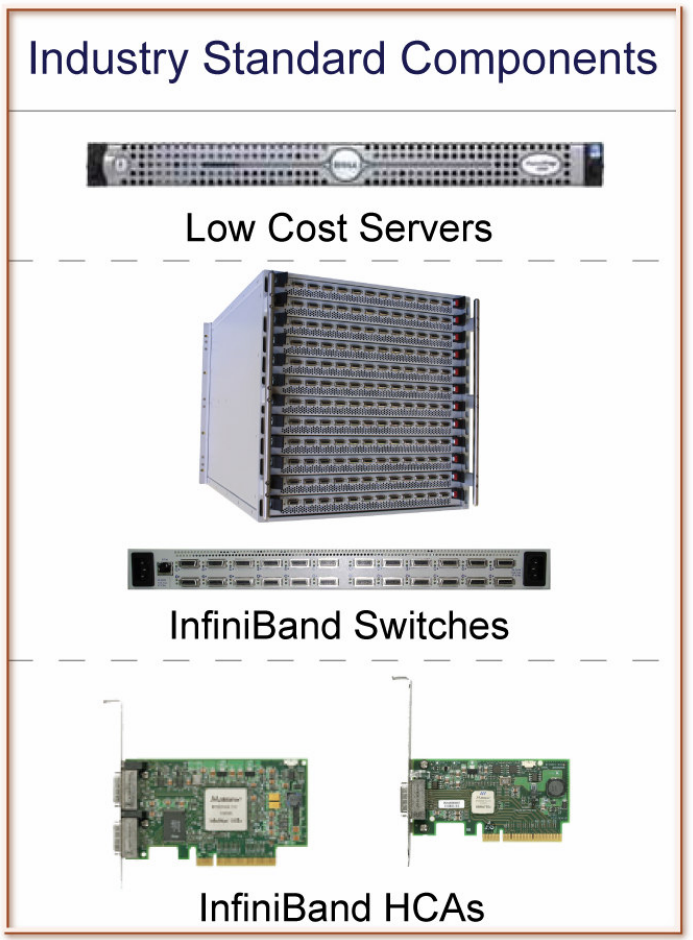
Member Roster

- Amphenol Interconnect Products
- Brocade Communications Systems, Inc.
- Fabric Networks, Inc.
- Foxconn
- Fujikura America, Inc.
- Fujitsu Components America, Inc.
- Fujitsu Limited
- Golden Bridge Electech, Inc.
- Hewlett-Packard
- JAE Electronics, Inc
- KeyEye Communications
- Lamprey Networks, Inc.
- LSI Logic
- Molex Inc.
- Network Appliance Inc.
- PathScale, Inc.
- Quake Technologies Inc.
- Sandia Corporation
- SBS Technologies, Inc.
- SilverStorm
- Tektronix
- Tyco Electronics Corporation
- W.L. Gore & Associates, Inc.

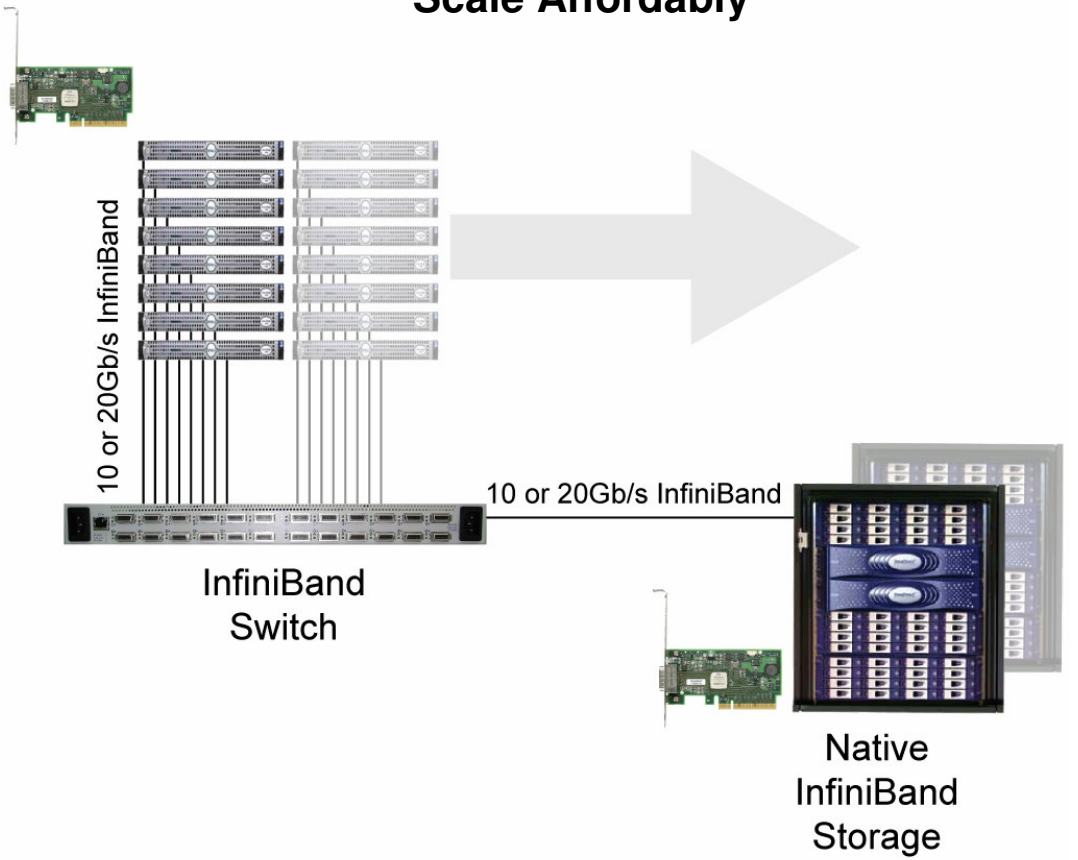
InfiniBand Technology

- High-performance fabric for server and storage clustering
- Industry-Standard
- Low-latency and high bandwidth
  - From 2.5Gb/s to 120Gb/s
  - Low latency
  - Reliability
  - Scalability
- Converges communications, computing, management and storage onto a single link with Quality of Service

InfiniBand Creates Affordable Supercomputing



InfiniBand Enables Industry-Standard Servers to Scale Affordably



Mass Market InfiniBand Solutions

Mellanox TECHNOLOGIES

InfiniBand
Landed on Motherboard

InfiniBand Blade Servers

Servers

Switches and Infrastructure

Clustering and Failover

Native InfiniBand Storage

Storage

Embedded, Communications, Military, and Industrial

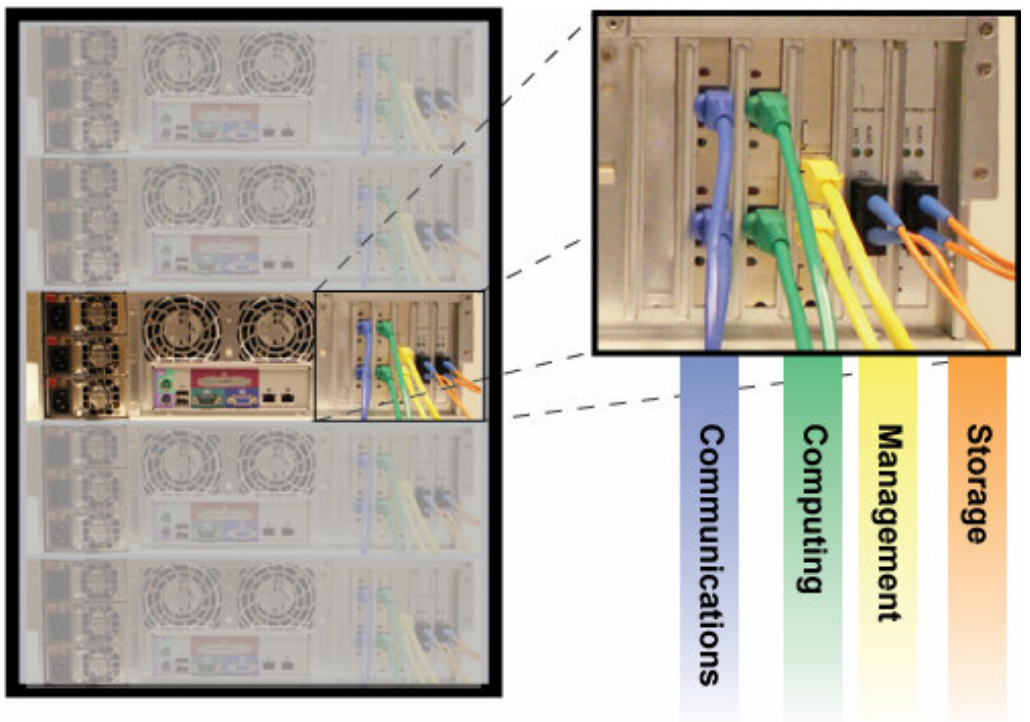
*Partial Lists

Data Centers Converge with InfiniBand

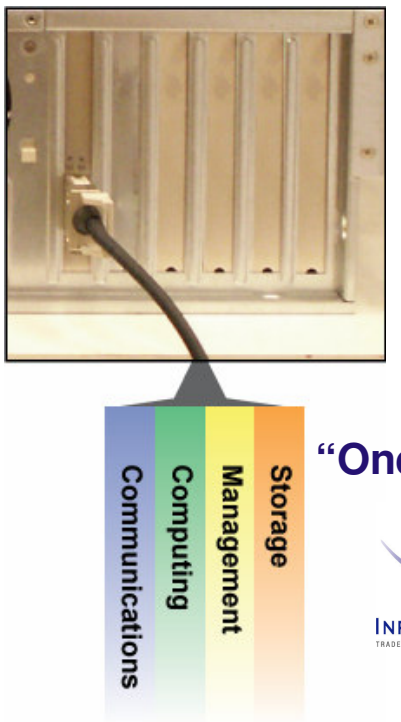
Mellanox TECHNOLOGIES

Cluster of Servers

Multiple Fabrics
High CapEx and High TCO



Single InfiniBand Fabric
Low CapEx and Optimal TCO



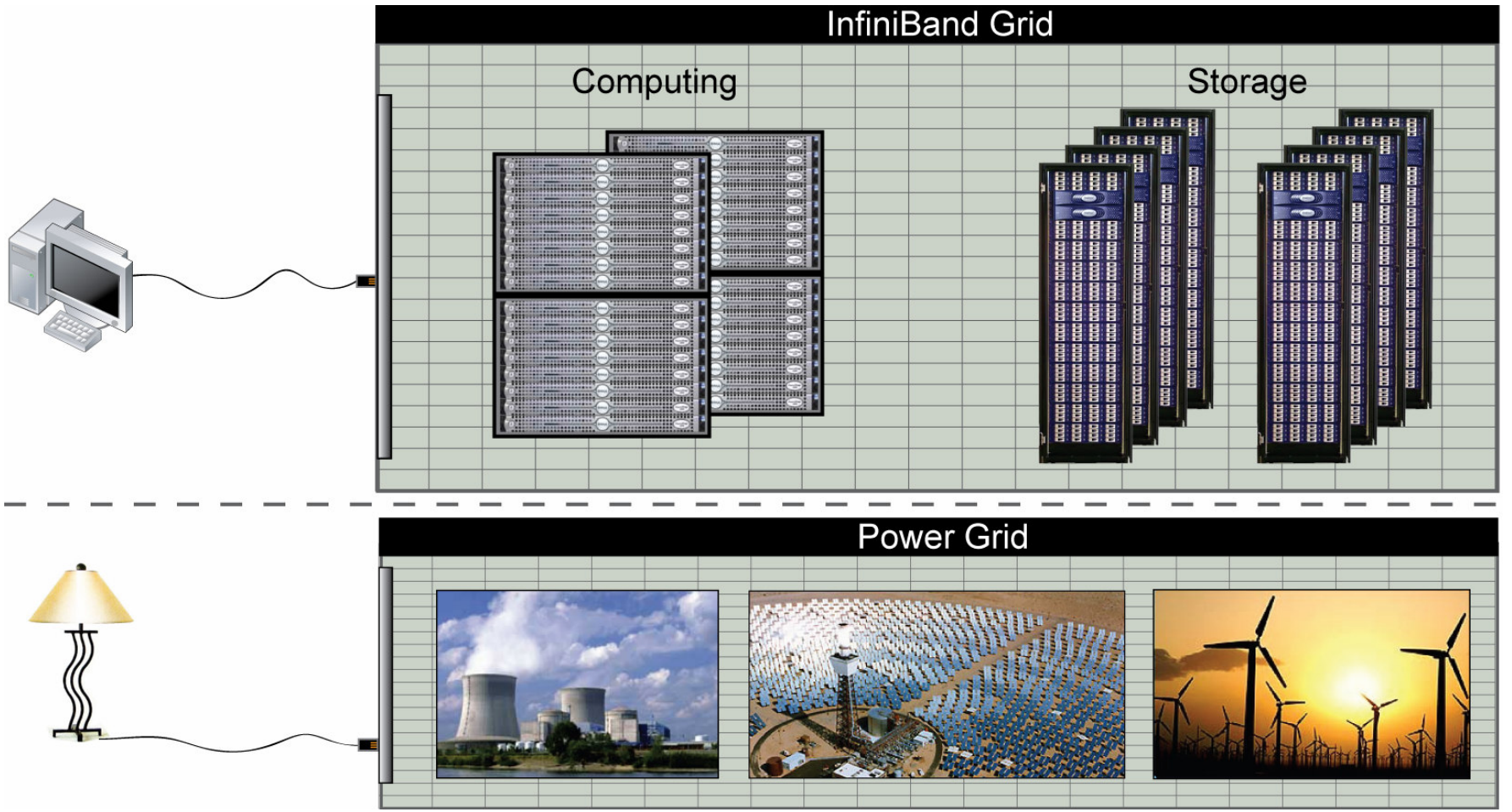
"One Wire"

InfiniBand
TRADE ASSOCIATION MEMBER

InfiniBand is the only fabric technology that can efficiently and cost-effectively converge the data center

The Vision

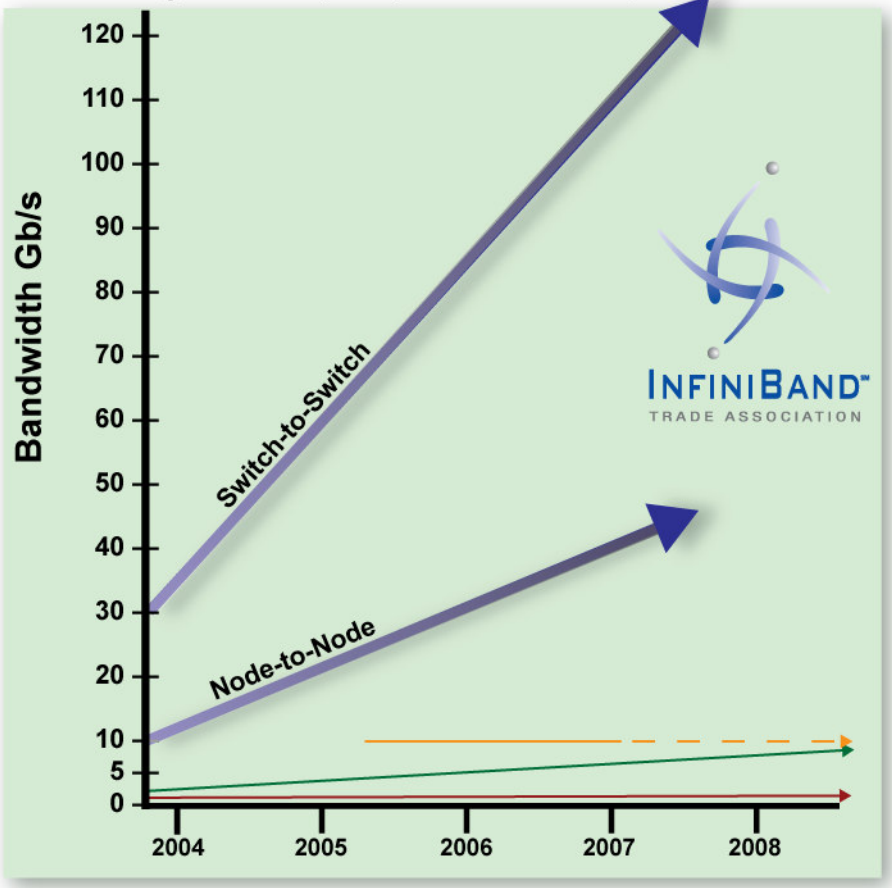
- Computing and storage as a utility
  - Exactly the same as electricity



InfiniBand Roadmap

InfiniBand Roadmap

InfiniBand's roadmap outpaces all proprietary and standard based I/O technologies in both pure performance and price/performance



Number of IB Lanes	Per Lane Bandwidth		
	SDR 2.5Gb/s	DDR 5Gb/s	QDR 10Gb/s
4X	10Gb/s	20Gb/s	40Gb/s
8X	20Gb/s	40Gb/s	80Gb/s
12X	30Gb/s	60Gb/s	120Gb/s

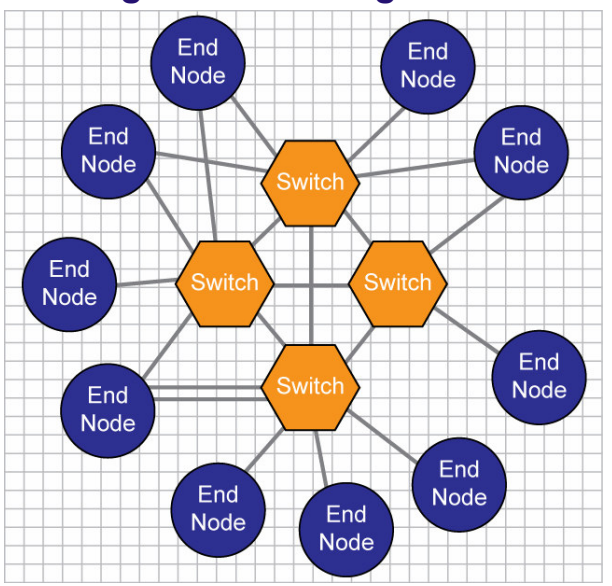
- InfiniBand
- 10GigE, 10G iSCSI, Proprietary
- Fibre Channel*
- GigE, iSCSI

**FCIA estimates 2007/8 for 8Gb/s FC*

†Roadmap based on silicon availability

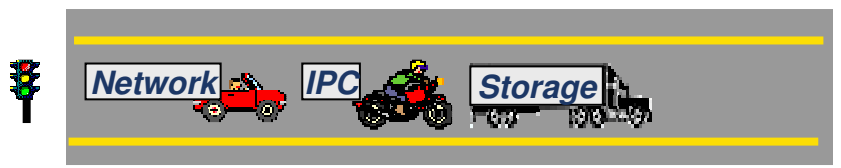
The InfiniBand Fabric

- Switch Fabric Links
  - 2.5 to 120Gb/s
- Transport Protocol
  - Reliable and Unreliable
  - Connected and Datagram
  - Send / Receive
  - RDMA read / write
  - Remote atomic operations
  - Automatic path migration
- Kernel bypass
  - Memory translation and protection tables
- HW Assisted Protection and inter-Process isolation
  - OS Virtualization
- Quality Of Service
  - Process at host CPU level
  - QP at the adapter level
  - Virtual Lane at the link level
- Scalability/flexibility
  - Up to 48K LIDs in subnet, up to 2^{128} in network
- Network partitioning
  - Multiple networks on a single wire
- Reliable, lossless, self-managing fabric
  - Link Level Flow Control
  - Multicast Support
  - Congestion management



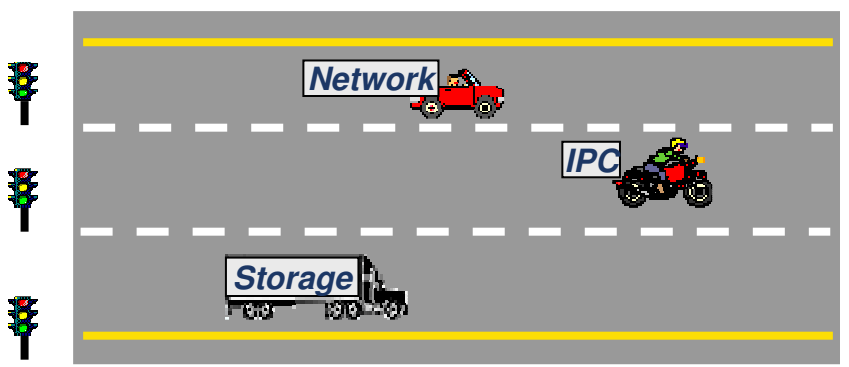
The Fabric – Virtual Lanes

Other networks



- Single lane shared by all traffic
  - Performance impact
- Cross-traffic dependency
- Force to use multiple fabrics

InfiniBand



- Independent virtual fabrics
  - Up to 16 Virtual lanes per link
- Separate flow control per lane
- No cross-traffic dependency

InfiniBand Network

- High-performance fabric
  - Link speed up to 120Gbit
- Multiple independent virtual fabrics
  - Flexible management
- Host off-load
  - Transport, RDMA, QOS
- Industry standard
  - Defined for low-cost implementation

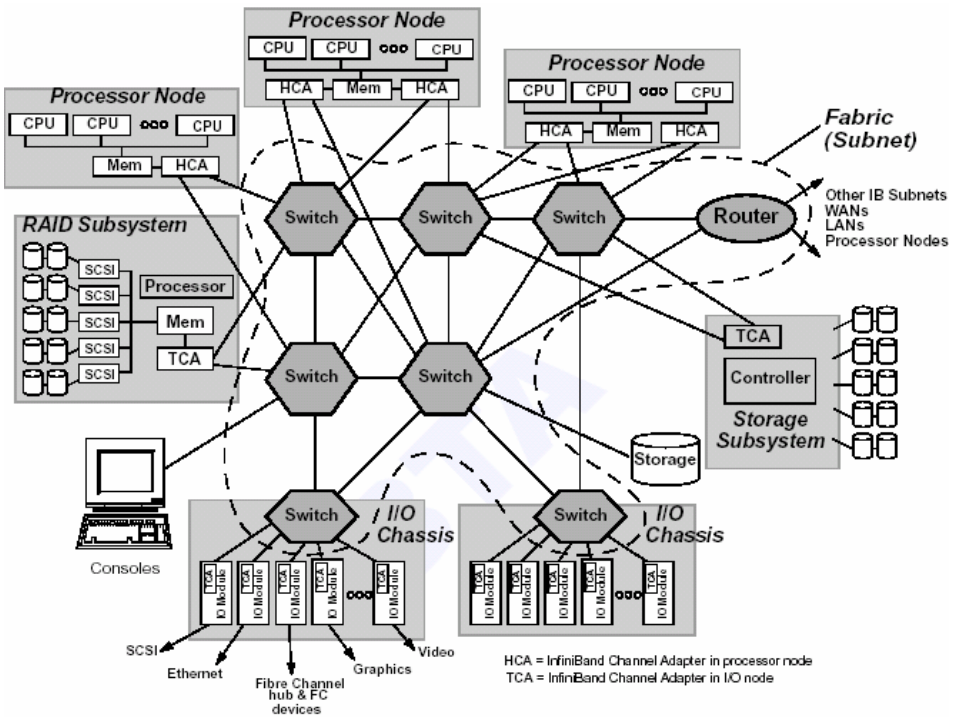


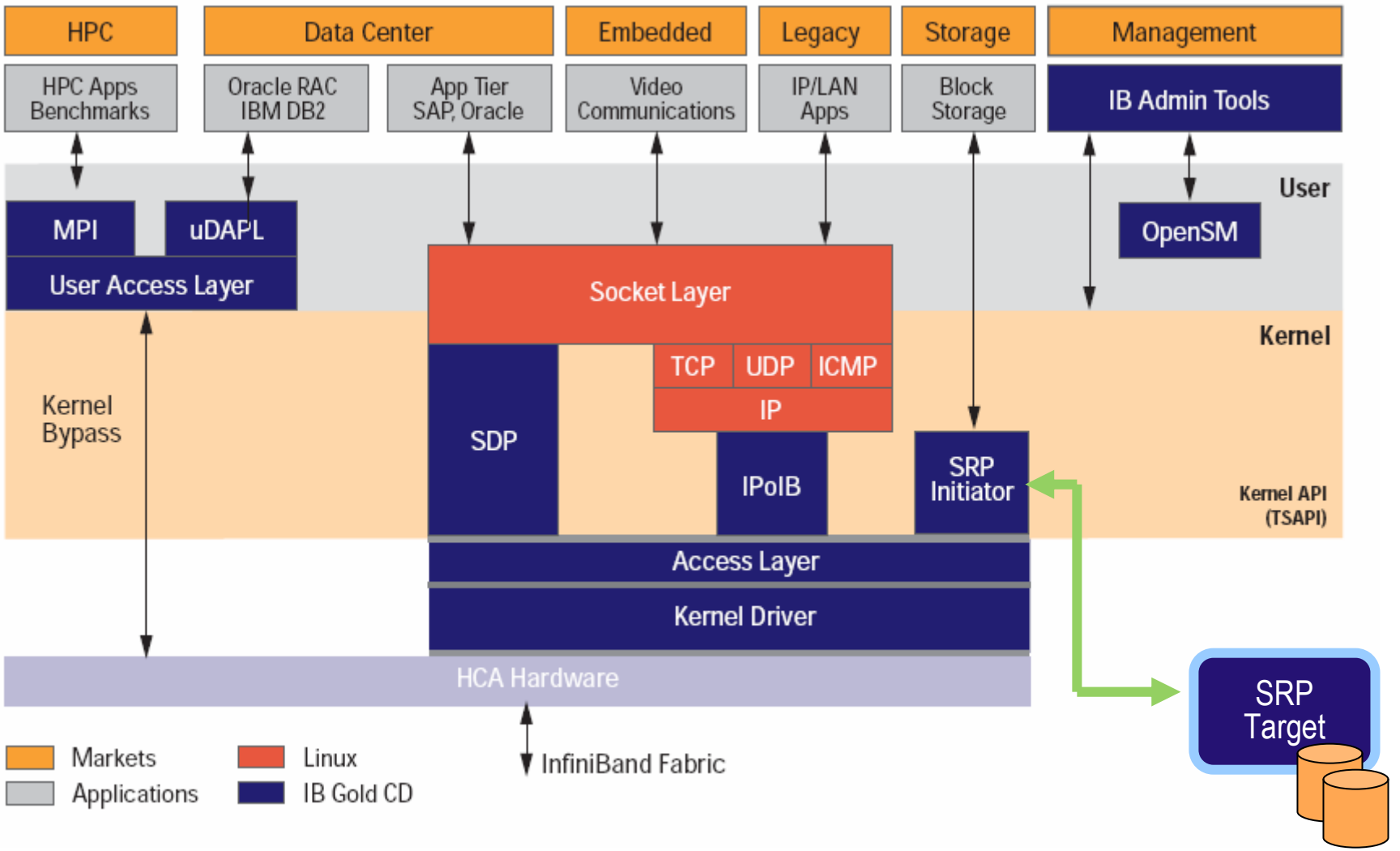
Figure 6 IBA System Area Network

InfiniBand Software Stack

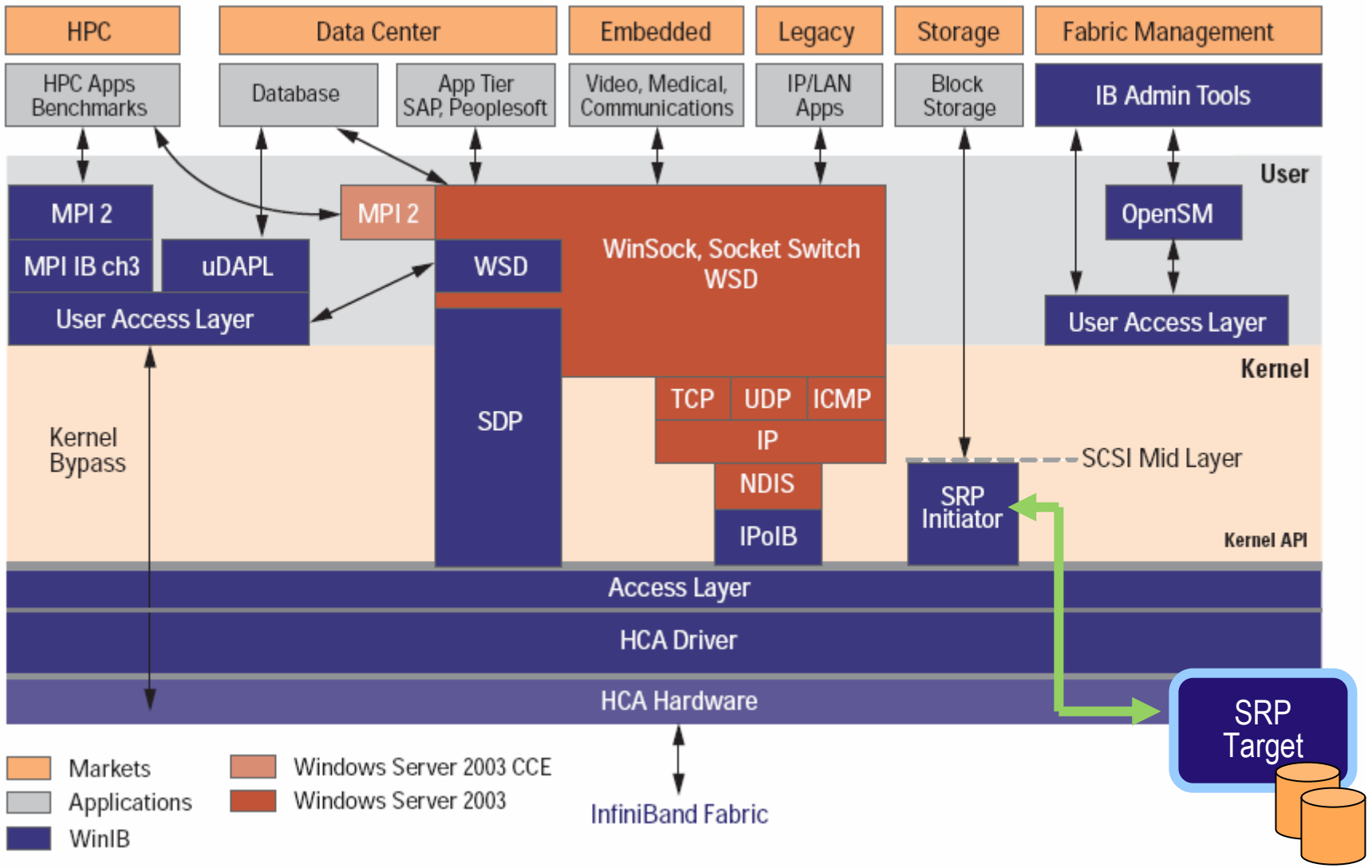
InfiniBand Driver Development

- **InfiniBand Linux and Windows drivers development is being done in an open source development community – OpenIB.org**
  - **OpenIB.org members**
    - Appro, Cisco, Data Direct Networks, Dell, LSI, Intel, Linux Networx, Mellanox Technologies, Network Appliance,Oracle, PathScale, Rackable Systems, Silicon Graphics, SilverStorm Technologies, Sun Microsystems,Tyan, Veritas, Voltaire, Sandia National Laboratories, Los Alamos National Laboratory, Lawrence Livermore Laboratory

InfiniBand Linux Stack



InfiniBand Windows Stack



InfiniBand Complete Platform Support

- **Complete Platform Support**
  - Itanium, Xeon, Opteron, PowerPC, SPARC
- **OS Support**
  - Linux, Microsoft, AIX, HPUX, OS-X, Solaris, VxWorks
- **Linux OpenIB.org drivers**
  - Linux Kernel
  - Redhat
  - SuSE
- **Windows OpenIB.org drivers**
  - Microsoft certification

Linux

The 2.6.11 kernel is out

Thursday March 03, 2005 (05:30 PM GMT)

"Linus has, at long last, released the 2.6.11 kernel. Only a small set of fixes has gone in since the last release candidate. Changes since 2.6.10 include **InfiniBand support**…

Software Protocols Details

- MPI
  - De facto standard in High Performance Computing applications
  - Supports advanced communication schemes like Asynchronous IO and "collect" semantics
- SDP
  - Transparent to the application
  - Standardized wire protocol
  - Maintains socket semantics
  - Leverage InfiniBand capabilities
    - Transport offload – Reliable connection
    - Zero Copy – using RDMA
    - Standardized wire protocol
- IPoIB
  - Simplest IB NIC driver
  - Transparent to the application
    - IB is selected by the system networking through IP mask
  - Standardized wire protocol

	MPI	SDP	IPoIB
API	MPI	Sockets (TCP only)	Sockets
Latency	Lowest	Lower	Low
BW	Highest	Higher	High

Storage

InfiniBand Storage Solutions

InfiniBand Backend Clustering and Failover

Native InfiniBand Visualization Storage System

Native InfiniBand Solid State Storage

Native InfiniBand Block Storage Systems

Native InfiniBand Clustered File Storage System

Native InfiniBand Clustered File Storage Software

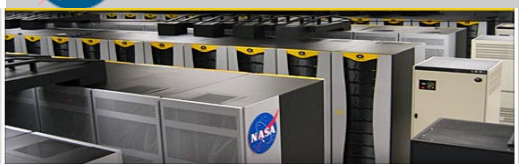
Native InfiniBand Block Storage Management Software

High Performance Computing

High Performance Computing

- InfiniBand growth in Top500 represents continued strong ramp in HPC
  - InfiniBand shows ~200% growth in 1 year
- The Highest two ranked industry-standard clusters use InfiniBand
- #4 NASA
  - 52 TFlops, 10K Itanium-2 CPUs
  - #2 in 2004, #3 in June 2005
- #5 Sandia National Laboratories
  - 38 Tflops, 4500 nodes, 9000 CPUs
  - New entry in Nov 2005 List
- #130 Glactic Computing
  - 3.413 TFlops , 562 Intel Xeon EM64T CPUs
  - InfiniBand MemFree
  - 820MB/s storage throughput
  - 84.35% Linpack efficiency - WORLD'S HIGHEST REPORTED RATING on x86!

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION



蚬壳星盈 GALACTIC COMPUTING

MPI over InfiniBand

MPI needs

Point to point (Blocking / Non-blocking) Collectives

Process / Fabric failover Eager

Rendezvous Tag Matching Ordering

Scalability Bla Bla Bla

InfiniBand Features

QoS HW Multicast Transport Offload

Send / Receive RDMA Write / Read Atomic (CS / FA)

Multi-path Transport Service (RC/UC/RD/UD)

Write w. Immediate Read fence CM etc.

MPI – Mellanox experience

- MVAPICH and MPICH-VMI for Linux / Unix
- MPICH-2 for Windows
  - MPI library scalability
  - Congestion avoidance and control
  - Soft failures
  - Robustness
  - MPI library tuning
  - Communication and computation overlap
  - MPI cleanup
  - Ease of installation, ease of use

MPI Library Scalability

- **Memory allocation**
  - **Registration failures fallback**
    - MPI should not break if registration fails
  - **Dynamic resource allocation**
    - On-demand connection management
    - Adaptive receive queue size
    - Adaptive SQ size
  - **Resource sharing**
    - Shared Receive Queue
  - **Fine grained resource allocation (buffers going to unexpected queue)**
    - Eager buffers smaller then eager threshold (can be pipelined)
    - Use receive buffers of different sizes

MPI Library Scalability – Cont.

- Connection management
  - On demand connection
  - Unreliable Datagram
- mpirun scalability
  - Using ssh to thousands of nodes is not a good idea
- MPI_Init
  - Moving a huge amount of data on the OOB channel – use the fast interconnect (InfiniBand)

Congestion Avoidance

- **Congestion avoidance**
  - **Hardware congestion management**
    - Simulation results published by IBM
    - Should be implemented close to the fabric
  - **Network topology awareness**
    - Job scheduler and resource allocation
    - MPI rank mapping to processors
    - Topology aware collectives

Soft Failures

- **Soft failures**
  - **Server failures**
    - Host memory data protection
    - Process fault tolerance (e.g. check points and restart)
  - **NICs failures**
    - Failover NIC
    - NIC restart
  - **Fabric failures**
    - APM (Automatic Path Migration)

Robustness

- Industry support for Open MPI is good
  - Commercial release will help in testing and QA
- Testing resources are expansive – use it wisely
  - Good test-plan, testing collaboration

MPI library tuning

- Many parameters (and keep growing)
- Default are not always the best
  - Platform oriented (CPU type, memory architecture…)
  - Cluster size dependent
  - Interconnect dependent (SDR / DDR)
- Most users are not capable of tuning all the parameters
- Provide a test that can measure and provide best parameters for specific environment
  - MVAPICH have some mechanism to measure SRQ parameters

Asynchronous Progress

- HCA transport offload allows communication and computation overlap

- MPI should allow asynchronous progress using progress thread
  - Should only be used for rendezvous progress
  - Eager messages can be consumed by the user thread
  - Use InfiniBand notification mechanism, no need to use OS timers

Interesting Ideas

- RDMA is NOT the only way to do zero copy
  - Send / Receive might do better in *some* cases
- Transport Service
  - Reliable connection is *not the only* high performance transport service in InfiniBand
- QoS – InfiniBand supports QoS
  - *Not* all messages were born equal.
- FIN messages optimizations
  - Use read fence to post read completion
  - Write w. Immediate can save FIN message for PUT operations

What I would like to see

- We know very little about applications demand
  - Buffer reuse
  - Queue usage

- Internal profiling framework
  - Defined interface for statistics collection
  - Defined interface for the data (database?)
  - Could later be accessed for run-time parameter tuning based on history data
    - MPICH-VMI (NCSA) is doing some sort of history-based decision taking.

InfiniBand Contacts

- If you are not sure
- If things behaves not as you expected
- If some InfiniBand feature is not implemented (and you need it)

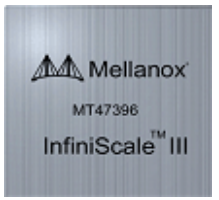
ASK

– Open MPI mailing lists	www.open-mpi.org
– OpenIB mailing list	www.openib.org
– Me	gil@mellanox.co.il

The Building Blocks

HCA and Switch Silicon Devices

Market Name	InfiniHost™	InfiniHost™ III Lx	InfiniHost™ III Lx	InfiniHost™ III Ex
# IB Ports	Dual 4X SDR	Single 4X SDR	Single 4X **DDR**	Dual 4X SDR
Host I/F	PCI-X	PCI Express x8	PCI Express x8	PCI Express x8



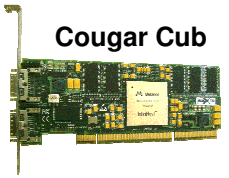
Market Name	InfiniScale™	InfiniScale™ III	InfiniScale™ III
# IB Ports	8-port 4X SDR	24-port 4X or 8-port 12X SDR	24-port 4X or 8-port 12X **DDR**

HCA Cards

InfiniHost

Dual 4X IB

Cougar Cub



MHXL-CFXXXT
128/256MB Memory Down

InfiniHost III Lx SDR/DDR

Single 4X IB

Tiger — PCI EXPRESS x4

MHES14-XT
MemFree, Media adapter

Cheetah — PCI EXPRESS x8

MHES18-XT
MemFree, Media Adapter

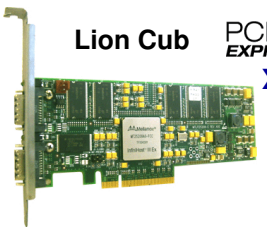
Cheetah DDR — PCI EXPRESS x8

MHGS18-XT
MemFree, Media Adapter

InfiniHost III Ex SDR/DDR

Dual 4X IB

Lion Cub — PCI EXPRESS x8



MHEL-CFXXXT
128/256MB Memory Down

Lion Mini — PCI EXPRESS x8

MHEA28-XT
MemFree, Media Adapter

Lion Cub DDR — PCI EXPRESS x8

MHGA28-1T
128MB, Media Adapter

Lion Mini DDR — PCI EXPRESS x8

MHGA28-XT
MemFree, Media Adapter

SDR **DDR**

InfiniBand Value Proposition

Mellanox InfiniBand – Driving Supercomputing to Mainstream Usage

- Mature, Industry-Standard Technology
  - Over 1 Million Ports Installed Base
  - Mass market availability of affordable solutions
- Convergence
  - Communications, computing, management and storage onto one fabric for lowest CapEx and optimal total-cost-of-ownership
- HPC and Data Centers
  - Mass scalability for Enterprise, Departmental and Personal environments
- Mass availability
  - Standard servers, Switches, Storage
- InfiniBand World class Price/Performance
  - Top500 growth, Low-cost Switches, Single-port HCA, InfiniBand DDR, Personal Supercomputing
- Software
  - OpenIB, Redhat, SuSE, Windows
- InfiniBand Ease-of-Use
  - Industry standard Of-the-shelf components, standard software interfaces, Cat6, Fiber cable and WAN solutions

Mellanox awarded "Most Innovative HPC Networking Solution" by HPC Wire at SC|05