InfiniBand℠

TRADE ASSOCIATION

Key Measures of InfiniBand™ Performance in the Data Center

Driving Metrics for
End User Benefits

INFINIBAND℠
TRADE ASSOCIATION

Benchmark Subgroup

Benchmark Subgroup Charter

The InfiniBand™ Benchmarking Subgroup has been chartered by the InfiniBand Trade Association Marketing Work Group to propose and support InfiniBand Trade Association programs targeted at benchmarking and reporting InfiniBand Architecture related data. Subgroup deliverables will include: facilitation, collection, publication and promotion of InfiniBand Trade Association Member company InfiniBand benchmarks and related data.

Member Companies

Agilent Technologies, Banderacom, Crossroads Systems, InfiniCon Systems, InfiniSwitch, Lane15 Software, Mellanox Technologies, Vieo

Abstract

This presentation highlights the performance benefits of InfiniBand Architecture deployment in the data center. Key application performance drivers are introduced and InfiniBand Architecture's role in optimizing these drivers discussed. Actual InfiniBand technology performance measurements are presented by individual InfiniBand Trade Association member companies. The session concludes with a general Q&A with the presenters.

Agenda

Introduction to InfiniBand Performance

Chris Petty (Banderacom)

Member Company Performance Presentations

Chris Petty (Banderacom)

Chris Eddington (Mellanox)

Brad Benton (Lane15)

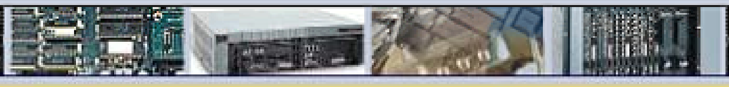
Thomas Dippon (Agilent)

Lawrence Didsbury (Auspex)

General Q&A

InfiniBand℠

TRADE ASSOCIATION



InfiniBand Performance Measurement

Driving Metrics for End User Benefits

InfiniBand Silicon Solutions

Agenda

Application and system level performance issues

- Where IB will be deployed in the data center
- Performance attributes

Performance impact of IB in the data center

- Theoretical impact
- Performance impact of real applications

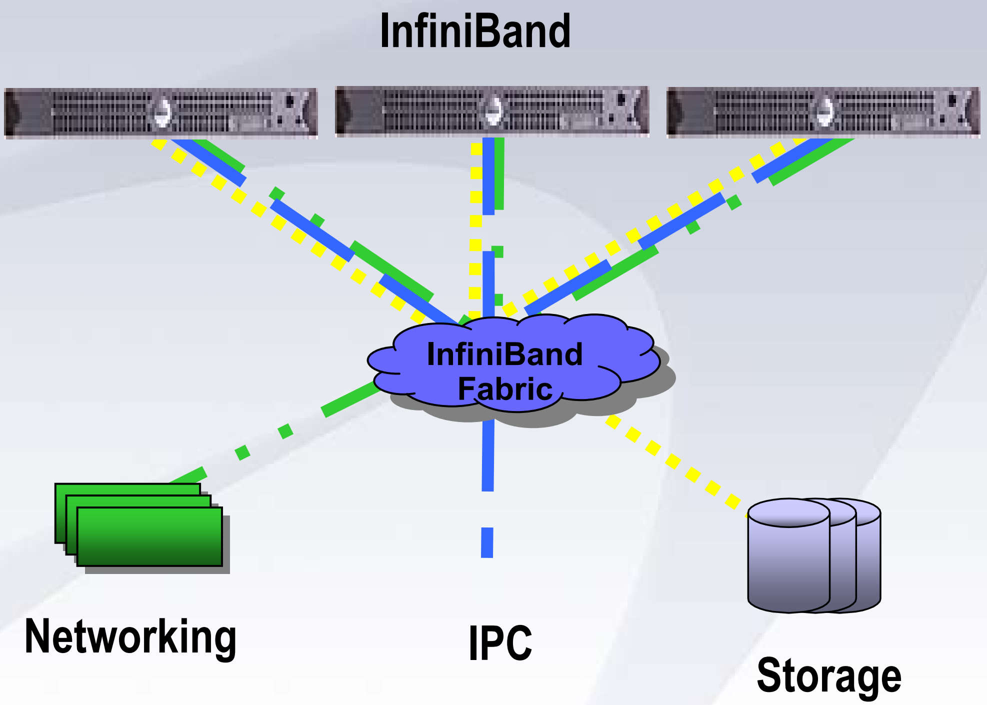
Why a New I/O Interconnect

Current Implementations: A Separate Fabric for *Each* Communication type - *Costly, Harder to Manage/Administer & With Considerable Host Processor Overhead*

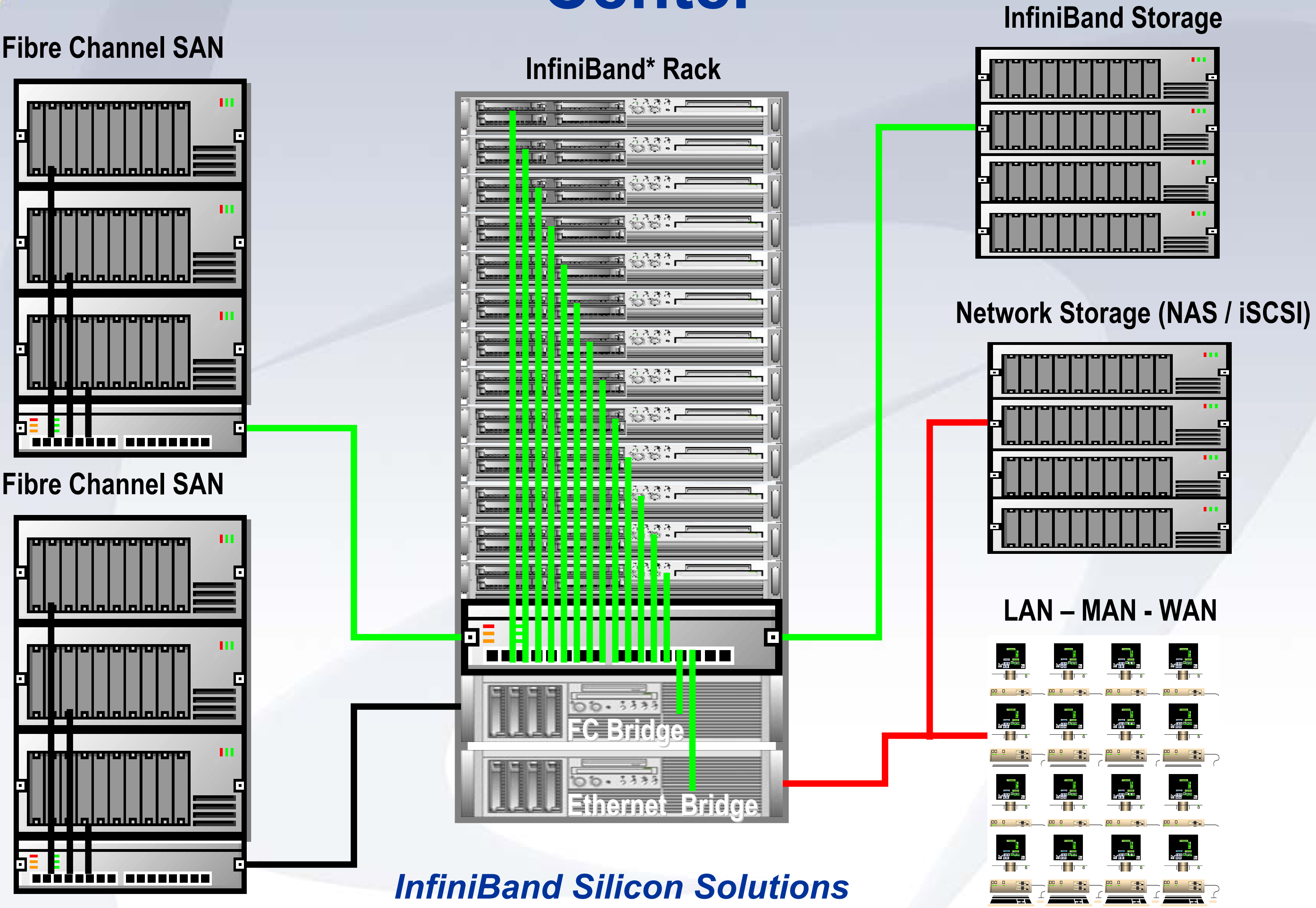
InfiniBand offers *One Integrated* Fabric for all types of Communication - *Converges:* IPC, Storage I/O & Network I/O

Data Center Performance Impact

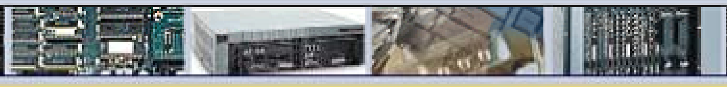
- All three types of I/O exist in a server and data center
- I/O fabric must provide performance for each type
- InfiniBand delivers on performance metrics for all I/O types
  - CPU utilization
  - Scalable throughput
  - Extremely low latency



InfiniBand Architecture in the Data Center



InfiniBand Silicon Solutions



With InfiniBand …

- Data center scales for all I/O
  - Bandwidth ready for large installs
  - Multiple 10GigE and 10GigFC easily handled
- Applications receive more CPU MIPS
  - CPU usage returned to application from I/O
- High performance clustering
  - Low latency enables large cluster size
  - Service partitions divide I/O and IPC on the same fabric

Define Drivers

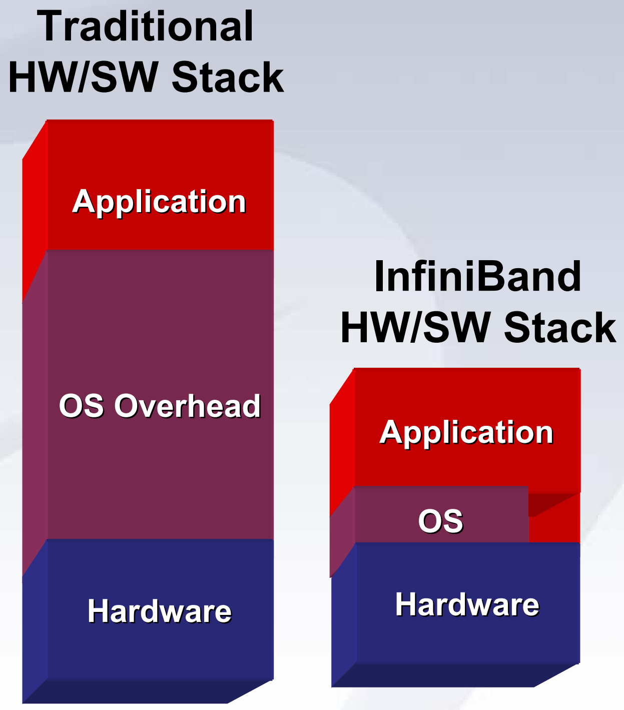
- CPU Load
  - Time spent by application processors on I/O protocol
  - Current GigE NIC and FC HBA evaluation
- IOPS
  - Number of discrete I/O's the system can drive through the I/O device
  - Current FC HBA evaluation
- Bandwidth
  - The total throughput of data an application can drive through the I/O
  - Current GigE NIC and Switch and FC HCA and Switch evaluation
- Latency
  - The time taken from an input of a packet or I/O to the output of that packet or I/O
  - Current GigE Switch and FC Switch evaluation

Application Performance Drivers

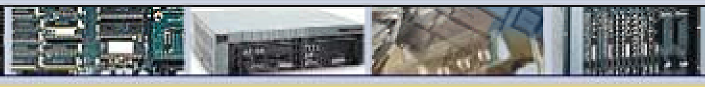
	Storage	Networking	IPC
CPU Load *processing metric*	Moderate	Important	Important
IOPS *processing metric*	Important	Important	Moderate
Bandwidth *data metric*	Important	Moderate	Low
Latency *data metric*	Low	Low	Important

Impact of InfiniBand on Processing

- Full Offload of reliable communication
  - No SW required to ensure data delivery
  - Linear growth of CPU MIPS to I/O capability
- User mode and Virtual Addressing
  - No OS protection required to access fabric
  - Zero copy for data structures
- Multiple Channels and Priorities
  - Independent resources eliminate sharing
  - “Fire and Forget” for applications



(Not to Scale)

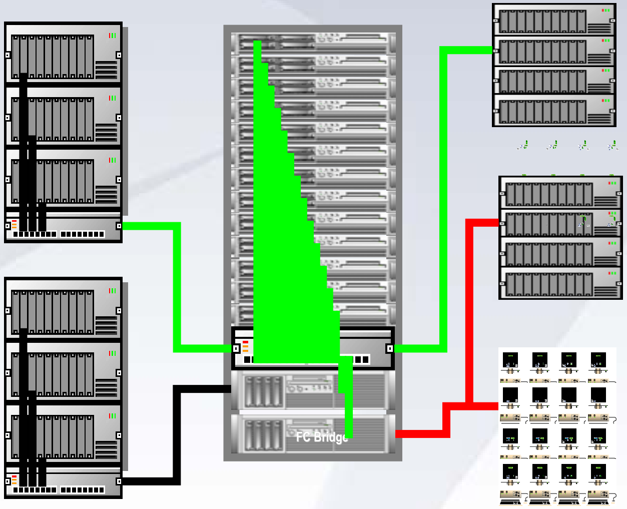


Impact of InfiniBand on Data Movement

- User Mode Access to Fabric
  - Applications move data without Kernel transition
  - Multiple Applications access in parallel
- Independent resources for Priority Levels
  - Different traffic streams have fabric paths
  - Cluster traffic not blocked by I/O traffic
- Low Hop Delay through Fabric Components
  - Stacked switches don’t introduce large delay
  - Fabric level delays constant as fabric grows
- Scalable Link Speed
  - Addition of link width transparent to fabric
  - Seamless addition of bandwidth anywhere

Application Benefits of InfiniBand Performance

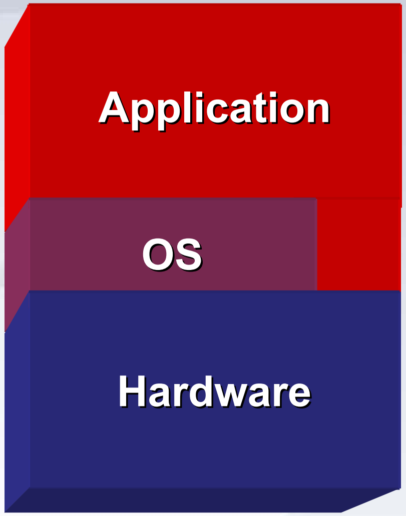
- Storage
  - Increased bandwidth yields faster access to data
  - Increased IOP allows more simultaneous accesses
- Networking
  - Lower processing overhead increases throughput to LAN clients
  - Increased IOP allows more efficient small packet web traffic
- Clustering
  - Lower processing overhead gives more CPU to the application
  - Lower latency allows clusters to scale efficiently



Data Center

Metrics for InfiniBand Performance

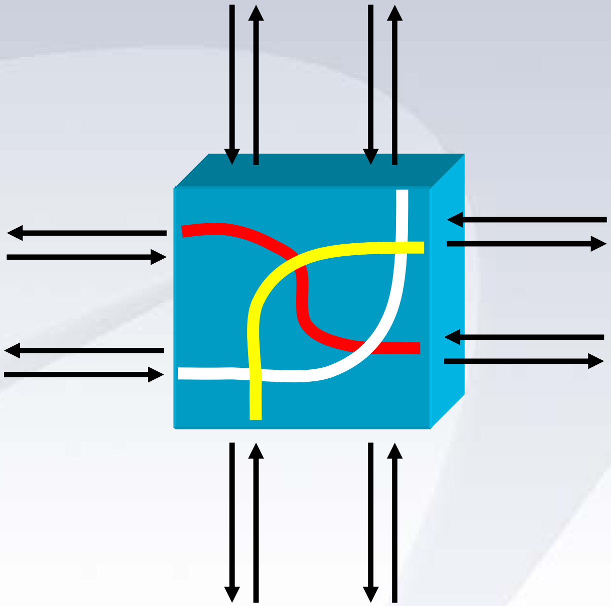
- CPU Utilization
  - Channel Services interface measurement
  - Price paid for accessing the fabric
- Messages Per Second
  - Channel Adapter Send/Receive work measurement
  - Transport message processing capability



InfiniBand Software

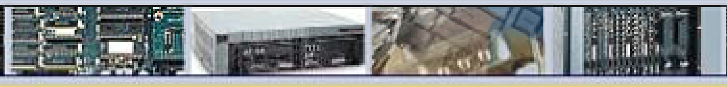
Metrics for InfiniBand Performance

- Bandwidth
  - Bus and wire level Gbit rate measurement
  - Efficiency of interfaces and internal structure
- Latency
  - Bus and wire level measurement in nano seconds
  - Speed of internal structures in "cut-through" mode



8p4x InfiniBand Switch

Company Examples



Banderacom

InfiniBand Silicon Solutions

InfiniBand Solutions Conference

April 2002

Chris Pettey, CTO

Metrics for InfiniBand

	Storage	Networking	IPC
CPU Load *processing metric*	Moderate	Important	Important
IOPS *processing metric*	Important	Important	Moderate
Bandwidth *data metric*	Important	Moderate	Low
Latency *data metric*	Low	Low	Important

Example Requirements for IB Performance - IOPs

IB-FC Bridge

- SRP protocol
  - 3^+ IB IOPs per FC IOP
  - 2 SEND + 1^+ RDMA
- 50K to 100K FC IOPs
- 150K to 300K IB IOPs

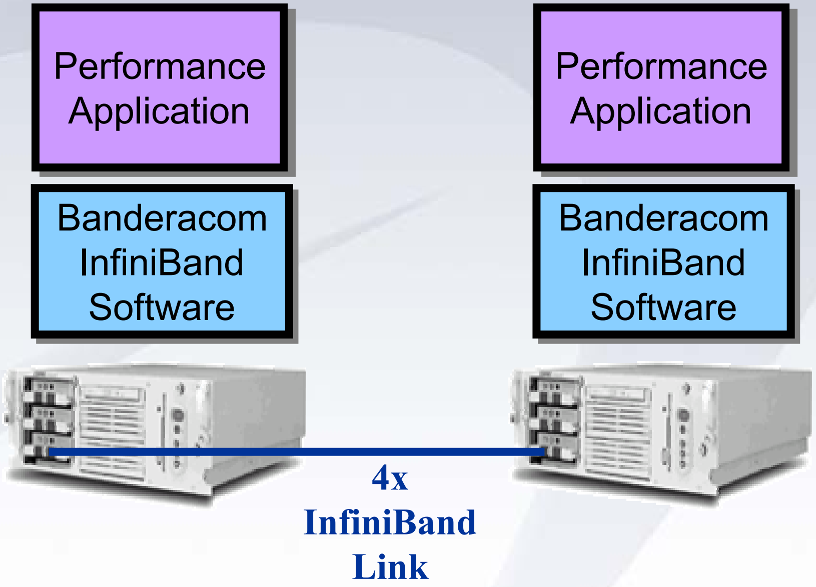
IB-Ethernet Bridge

- vNIC protocol
  - 1 IB IOPs per packet
  - Either SEND or RDMA
- 100K to 500K GigE packets per second
- 100K to 500K IB IOPs

Performance Application

Hardware

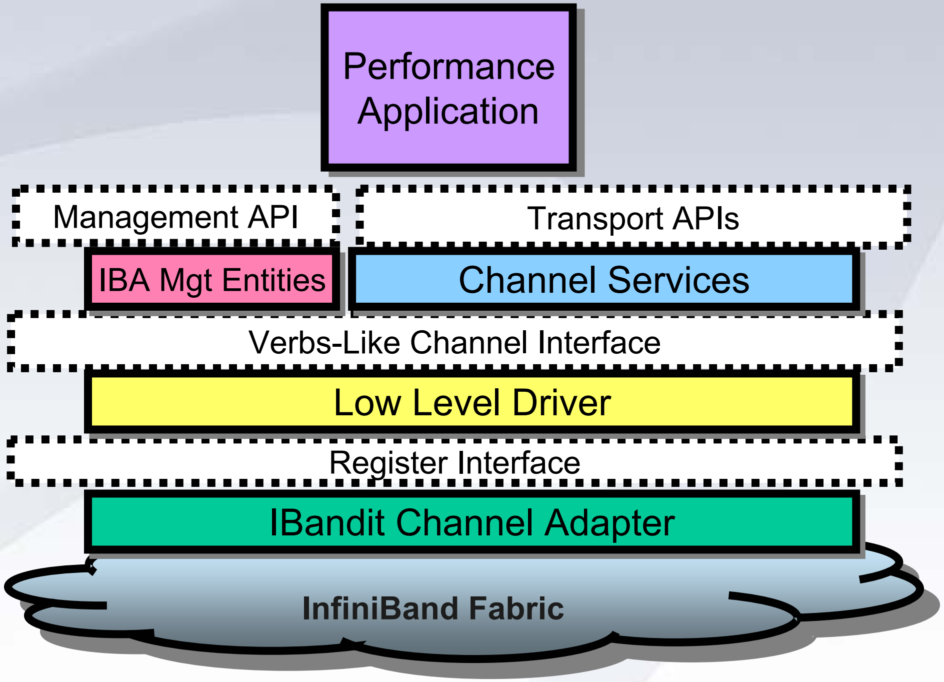
- ***Prototype PCI-X system at 133MHz***
- ***Single Pentium III at 1 GHz***
- ***IBandit-TCA operating in 4x mode and point-to-point interconnect***



Performance Application

Software

- ***Performance application stressing at the channel services interface using RC***
- ***Banderacom software running on Windows 2000***



IBandit-TCA Performance

- IBandit-TCA saturates 133MHz PCI-X bus
  - > 6 Gbps bi-directional transfer rate
  - 70-80% PCI-X bus utilization using prototype PCI-X chipset
- IBandit-TCA is capable of driving >350 KIOPS
  - Currently limited by bus and CPU capabilities, IBandit-TCA can scale even higher
- IBandit-TCA performance scales linearly across QPs

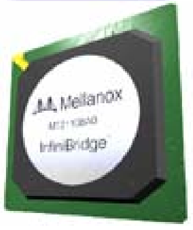
InfiniBand℠

TRADE ASSOCIATION

Mellanox Technologies, Inc
InfiniBand Solution Conference
April, 2002

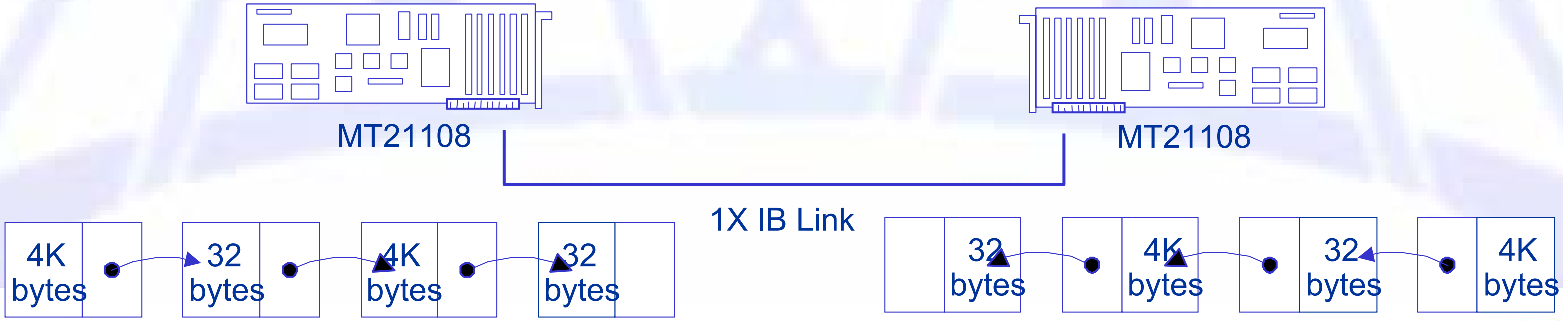
Chris Eddington

Director of Technical Marketing

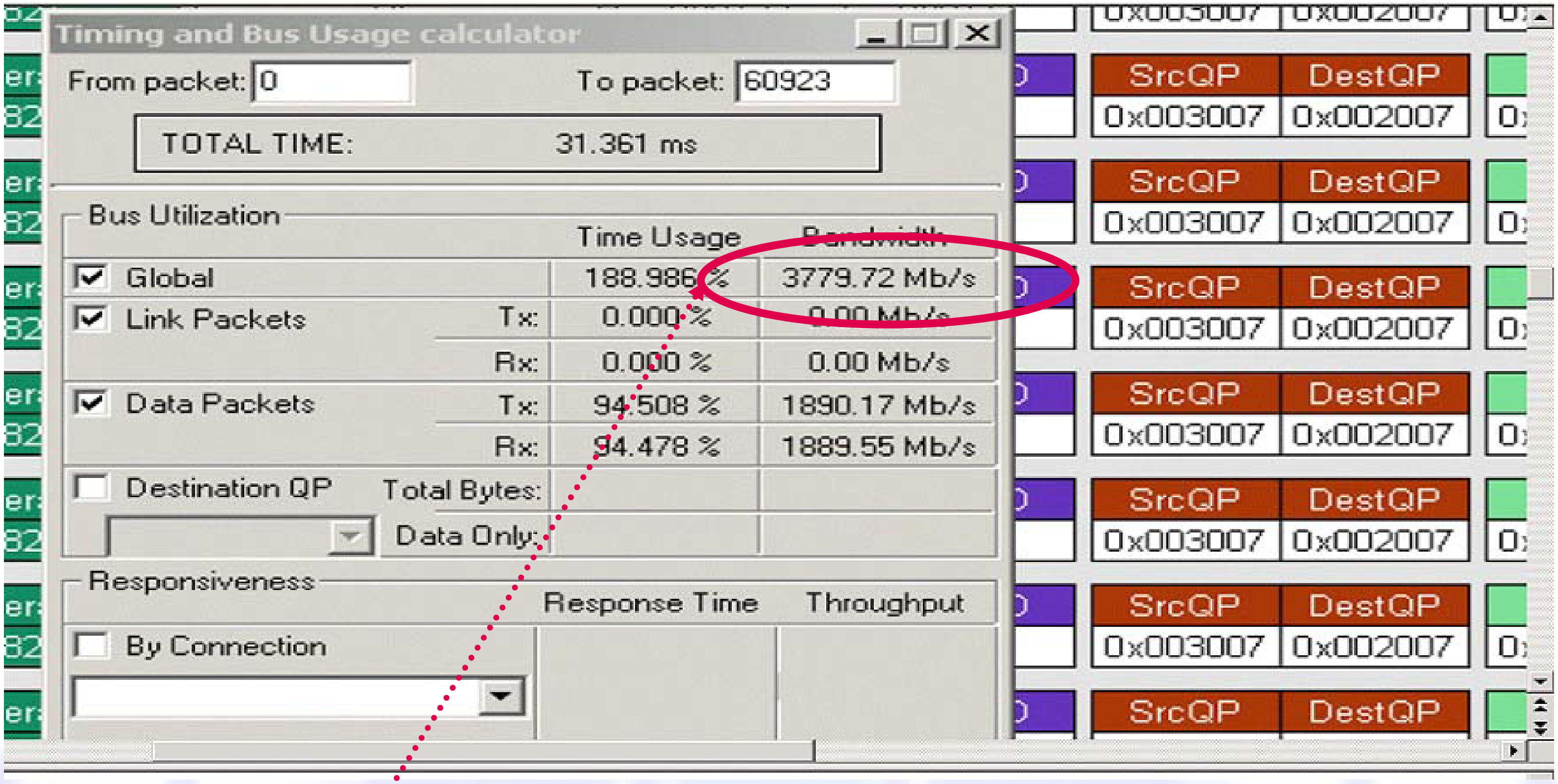


Test Configuration

- InfiniBridge™ MT21108 in a storage application that pushes for the highest InfiniBand 1X bandwidth
- Throughput measurements from storage application under development
- Systems: 800 MHz Pentium III ® CPU 64 bit 66 MHz PCI bus
- Even mix of large (4KB) and small (32B) packets
- Fully pipelined descriptor posting and completion
- Takes advantage of completion aggregation



InfiniBridge™ Link Performance Update



Timing and Bus Usage calculator

From packet: 0 To packet: 60923

TOTAL TIME: 31.361 ms

Bus Utilization

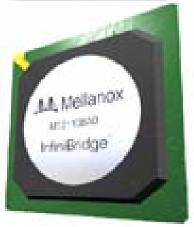
		Time Usage	Bandwidth
☑ Global		188.986 %	3779.72 Mb/s
☑ Link Packets	Tx:	0.000 %	0.00 Mb/s
	Rx:	0.000 %	0.00 Mb/s
☑ Data Packets	Tx:	94.508 %	1890.17 Mb/s
	Rx:	94.478 %	1889.55 Mb/s
☐ Destination QP	Total Bytes:		
	Data Only:		

Responsiveness

	Response Time	Throughput
☐ By Connection		

SrcQP	DestQP
0x003007	0x002007
SrcQP	DestQP
0x003007	0x002007
SrcQP	DestQP
0x003007	0x002007
SrcQP	DestQP
0x003007	0x002007
SrcQP	DestQP
0x003007	0x002007
SrcQP	DestQP
0x003007	0x002007
SrcQP	DestQP

- Result: ~3.8 Gb/s or 94% of maximum bandwidth achieved with only 7% processor utilization



InfiniBand℠

TRADE ASSOCIATION

LANE 15

SOFTWARE

The Leader in InfiniBand™ Management Software

IBSC Performance Presentation

Brad Benton

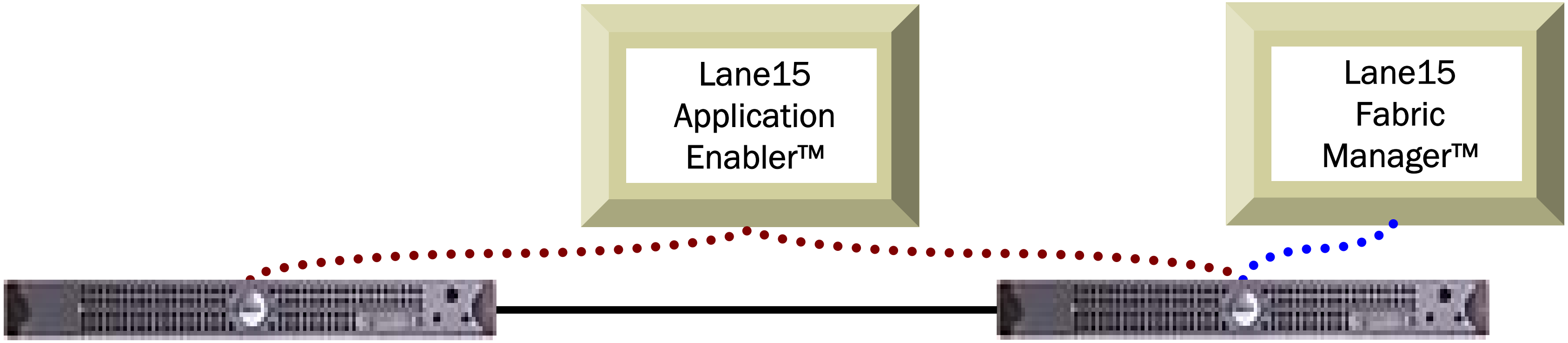
Technical Relationship Manager

Lane15 Application Performance

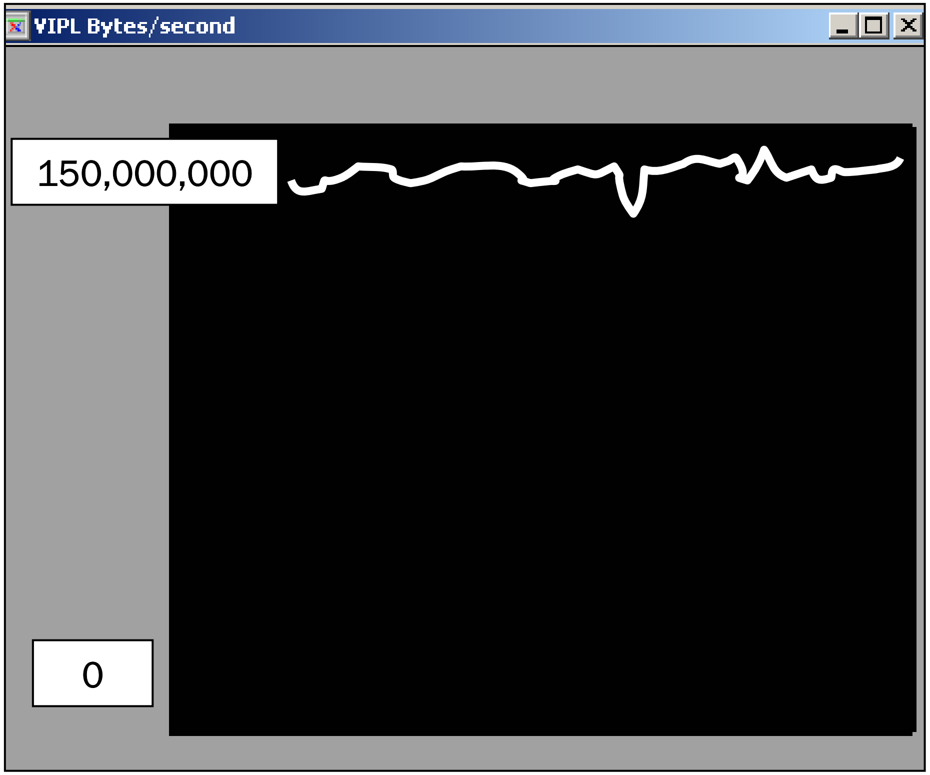
Demonstration

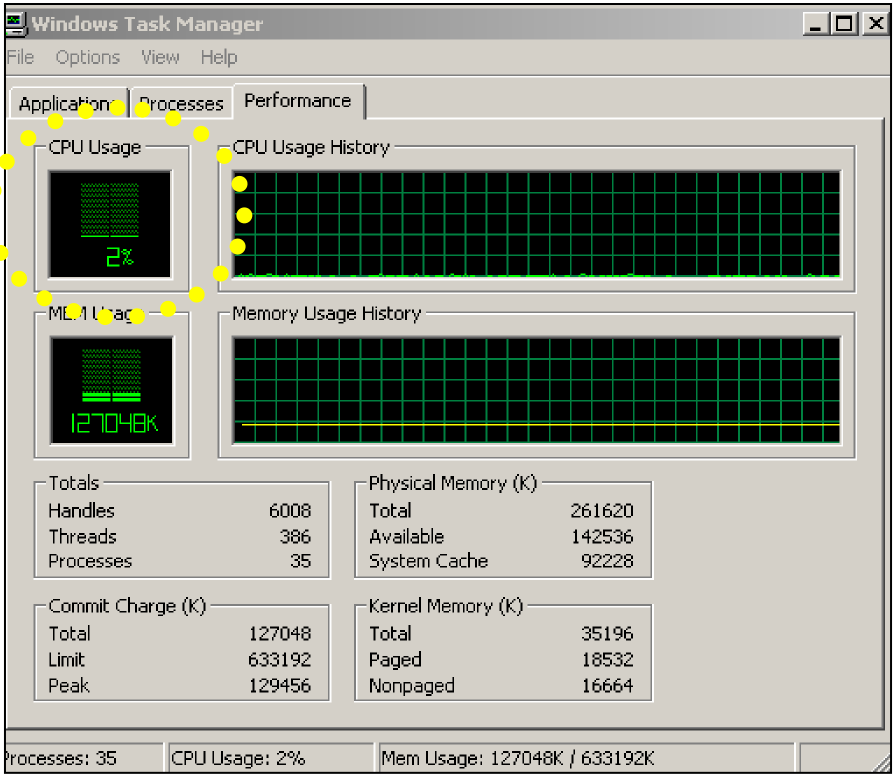
- VIPL-based data pump
- High-speed data transfers
- Low protocol overhead
- OS bypass technology
- Yield high data rates
- Minimal CPU overhead

150,000,000 bytes per second with < 2% CPU Utilization!



High data rate with low CPU overhead





Totals	
Handles	6008
Threads	386
Processes	35

Physical Memory (K)	
Total	261620
Available	142536
System Cache	92228

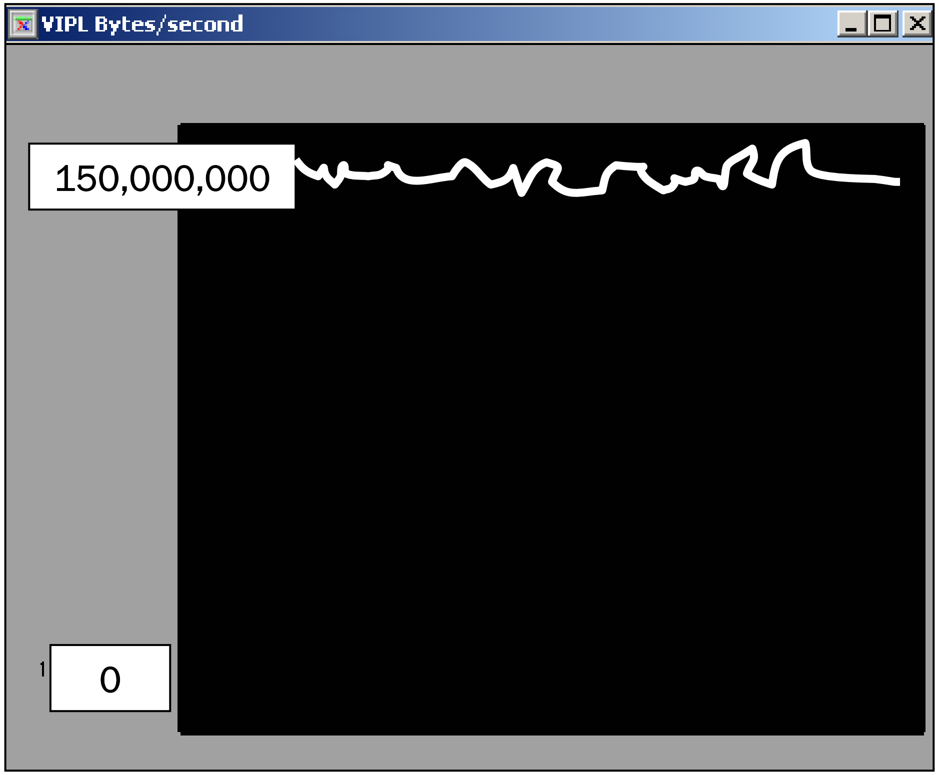
Commit Charge (K)	
Total	127048
Limit	633192
Peak	129456

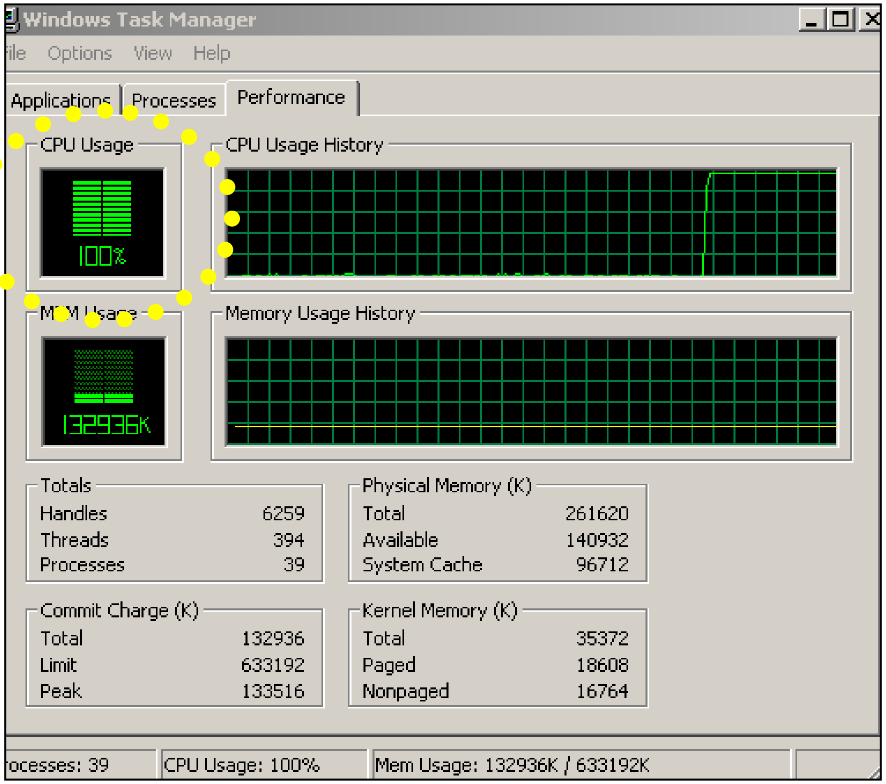
Kernel Memory (K)	
Total	35196
Paged	18532
Nonpaged	16664

Processes: 35 | CPU Usage: 2% | Mem Usage: 127048K / 633192K

CPU Utilization < 2%

Data rate with high application load





Totals

Handles	6259
Threads	394
Processes	39

Physical Memory (K)

Total	261620
Available	140932
System Cache	96712

Commit Charge (K)

Total	132936
Limit	633192
Peak	133516

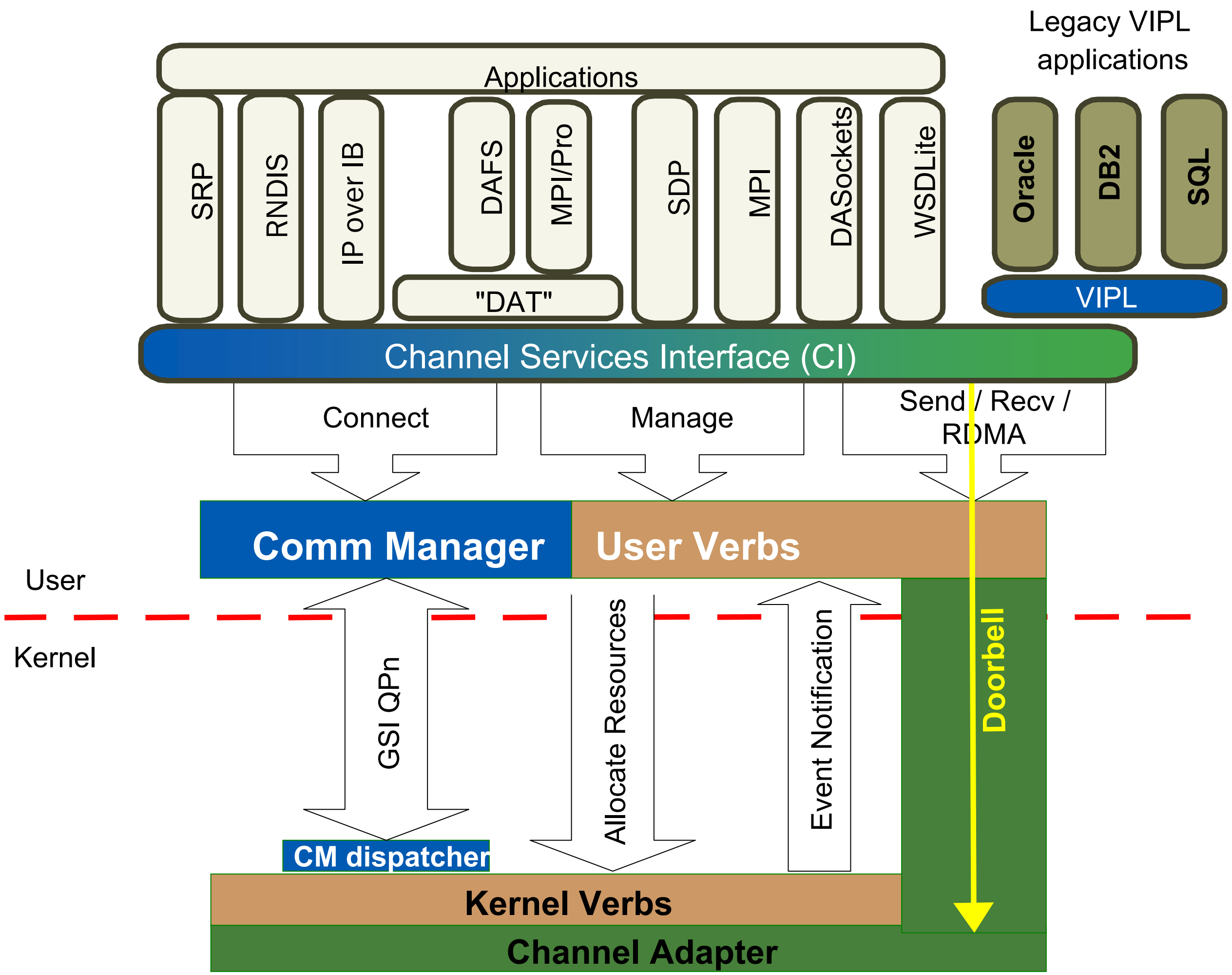
Kernel Memory (K)

Total	35372
Paged	18608
Nonpaged	16764

ocesses: 39 CPU Usage: 100% Mem Usage: 132936K / 633192K

Data rates are the same!

High Performance Architecture



Summary

- InfiniBand™ Performance is REAL
- VI based applications run unmodified on InfiniBand Fabrics today
- Native InfiniBand Applications are coming!

InfiniBand Architecture gives you your CPU back!

INFINIBAND[SM]

TRADE ASSOCIATION

Roland Scherzinger / Thomas Dippon

Performance Analysis of InfiniBand™ Switches

- A Case Study -

Agilent Technologies

Contents

- Overview
- Definition of Terms
- Test method and setup
- Measurement results
- Summary

Overview

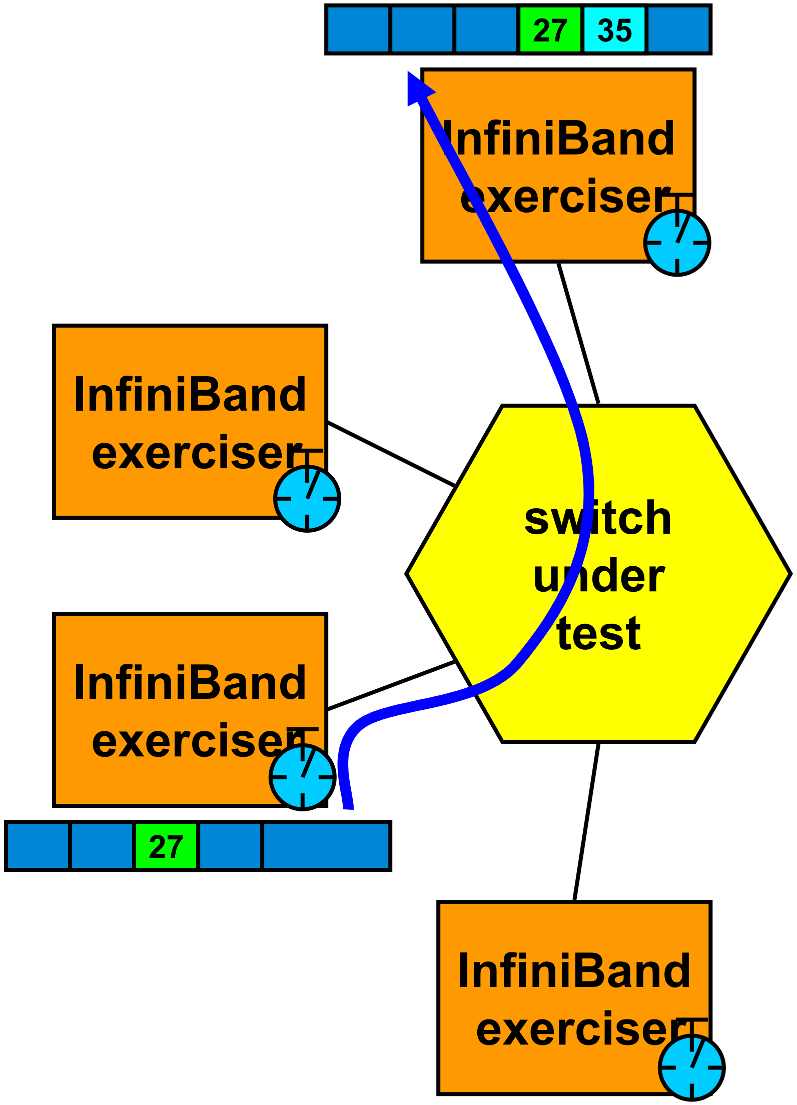
- The performance parameters of switches are key to overall performance of fabric topologies such as InfiniBand™ Architecture
- Due to the number of input factors and the complexity of switch cores it is hard to predict actual performance under real-life load conditions
- Therefore measurement of switch performance parameters is critical for manufacturers to verify simulated results under real-life and worst-case load conditions
- End-users can benefit from measurements to compare various switch implementations and verify claims of manufacturers
- The goals of this presentation are
  - To show a method for measuring InfiniBand switch performance
  - To show how well InfiniBand switches work - even in this early stage of the technology
  - NOT a competitive analysis

Definition of Terms

- Stream
  - A continuous sequence of packets with certain characteristics (load, range of packet sizes, VL, SL, etc.) transmitted from one port of a switch to another
- Throughput
  - The number of bytes per second that are actually transferred through each port of a switch depending on various operating conditions (load on each port, packet sizes, size of forwarding tables, etc.).
- Latency
  - The difference between the time when the first byte of a given packet enters a switch and the time when the first byte of the same packet leaves the switch

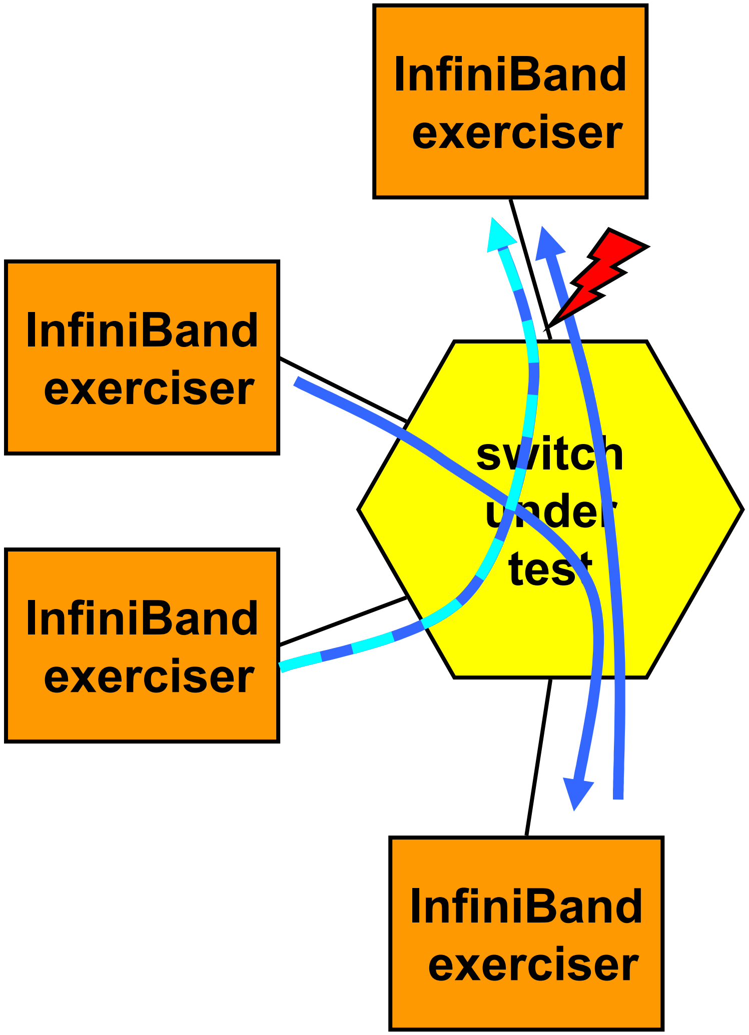
Test Method (1)

- InfiniBand(TM) exercisers are connected to the ports of a switch
- All exercisers have precisely synchronized timestamp clocks (+/- few nanoseconds)
- Exercisers generate "instrumented packets" by inserting a timestamp in the payload of each packet just before transmission
- Upon reception, the difference between the receivers clock and the timestamp value in the packet indicates the latency
- In addition there are counters keep track of the number of transmitted/received bytes/packets

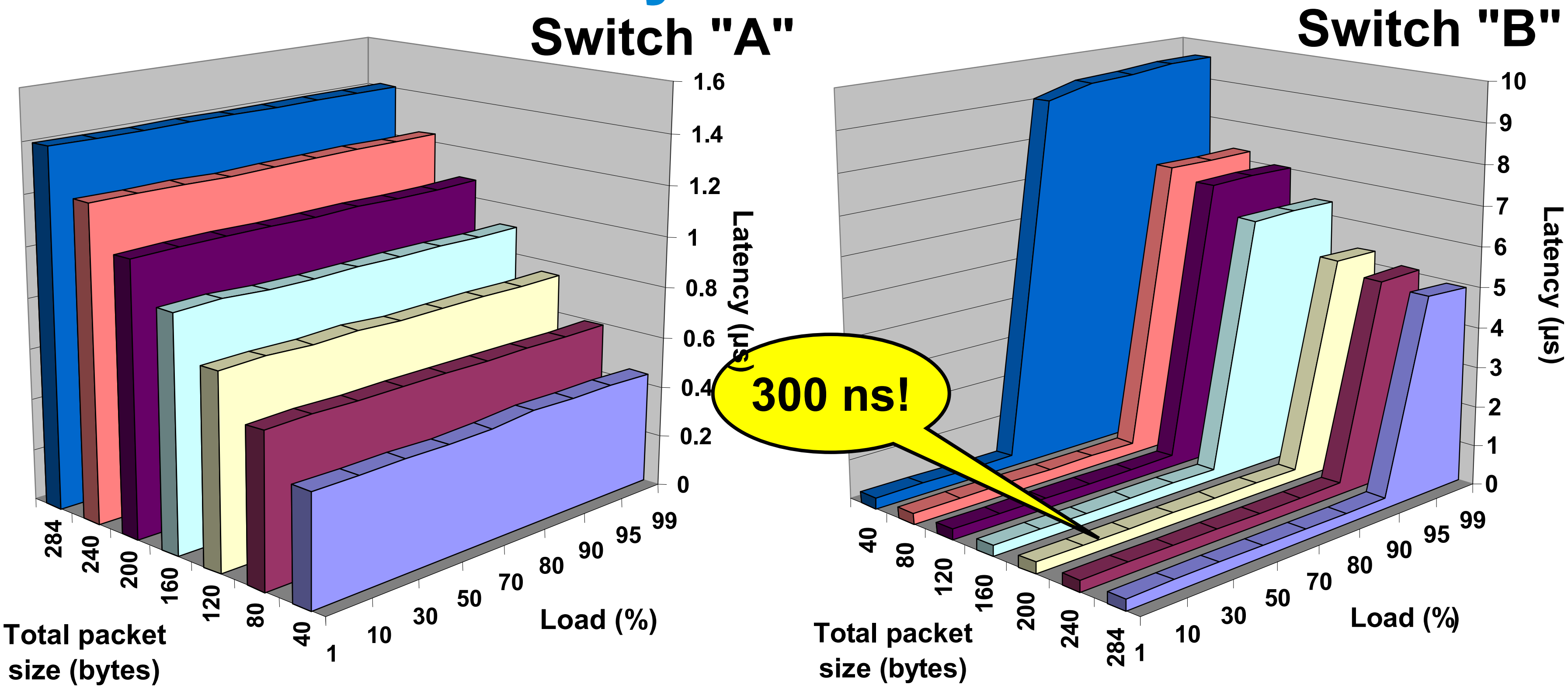


Test Method (2)

- Performance parameters are measured with one or more simultaneous streams
  - Single stream
  - With other background traffic
  - With traffic congestion (two or more streams destined for the same port)
  - Etc…
- Stream parameters can be varied
  - Packet size (or range of packet sizes)
  - Inter-packet delay (used bandwidth)
  - Burst of packets or continuous stream



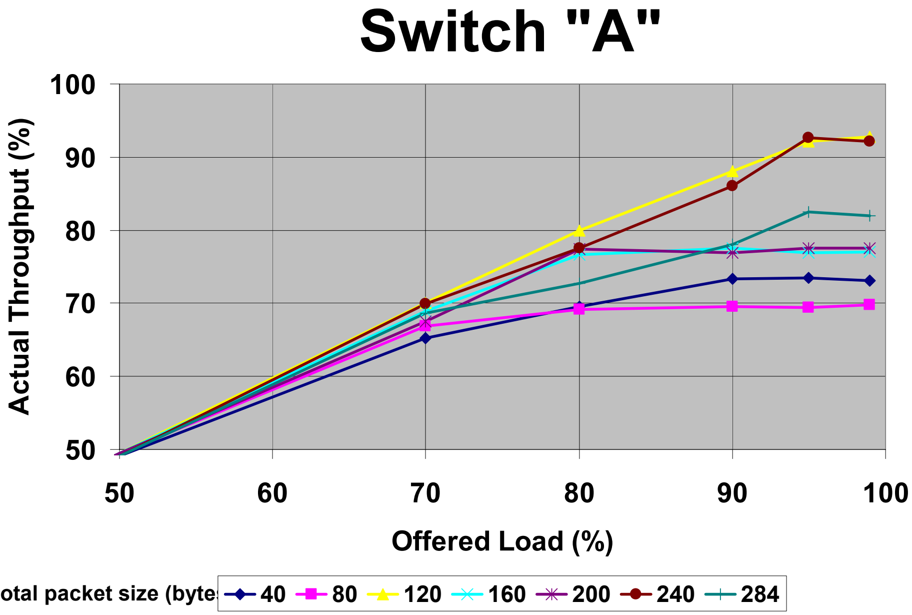
Results of Latency Measurements

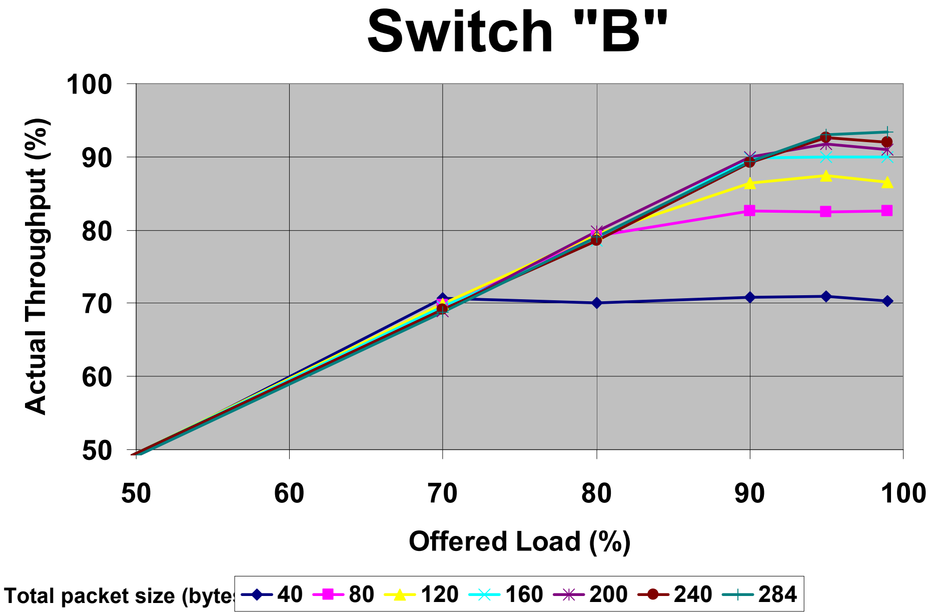


- Latency increases linearly with packet size: store & forward
    - ~1µs per 256 byte packet size
- No increased latency even at extreme load

- Latency in wide range independent of packet size: cut through switching
- Latency step function at a certain load level indicates throughput limitation on the output

Results of Throughput Measurements





- Early testing shows excellent results up to ~70…80% load
- Throughput limitations are typically seen with small packets

Summary

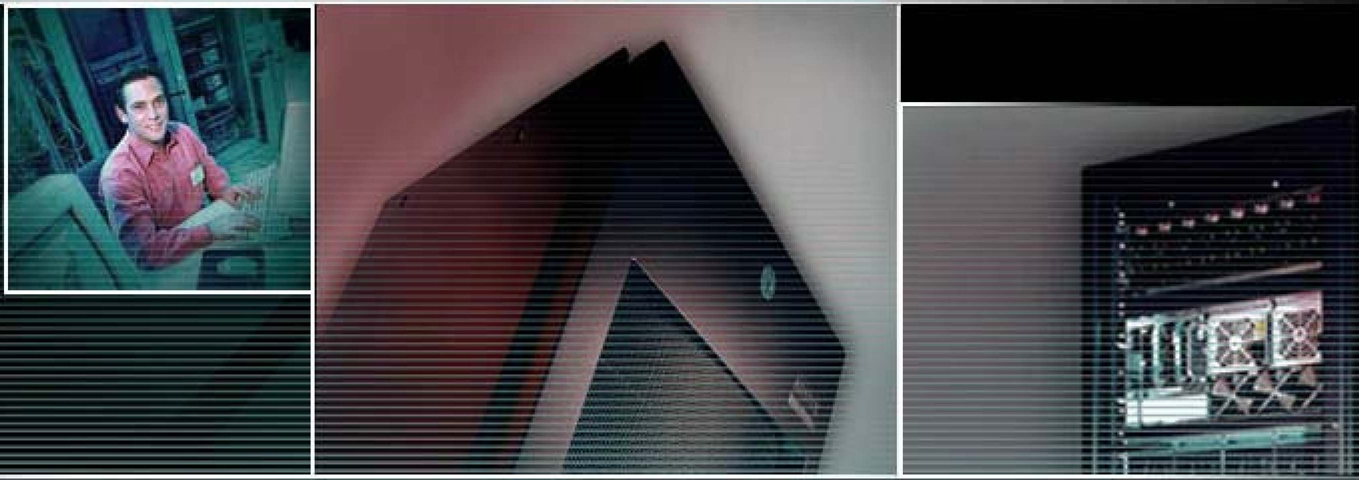
- Even 1st generation InfiniBand switches show very impressive performance numbers
  - 300 ns cut-through latency!
- It is important to understand the performance characteristics and limitations of switches
  - Even early switches perform excellent up to ~70…80% load
- Many other scenarios (multiple streams, multiple partitions, P_Key checking, SL to VL mapping, weighted VL priorities etc.) have to be taken into account to get a realistic picture
- InfiniBand 4x silicon and switches are now available - stay tuned for great performance numbers

InfiniBand℠

TRADE ASSOCIATION

AUSPEX

How Data Gets Delivered



Auspex InfiniBand™ Architecture Project

Using an InfiniBand Architecture point solution to deliver even more enterprise data storage performance!

InfiniBand™ Architecture Meets NAS

AUSPEX

- Improved NFS mounts and CIFS authentication by moving to an InfiniBand interconnect (higher number of mounts more rapidly)
  - Zero load on the CPU doing IO transfers
- Increase the scalability of the overall multi-node system immediately
  - from 3 to 4 I/O Nodes
  - eventually to N+1 nodes with an InfiniBand switch
- Increased interconnect bandwidth with InfiniBand Architecture
  - high-speed, point-to-point, bi-directional communication

InfiniBand™ Architecture Meets Enterprise Storage

- Removal of existing software delays on both the network and host processors
- Significant load reduction on both network and host processors (processor load observed using PCI bus analyzer with various throughput quantities)

InfiniBand™ Architecture Meets NAS-SAN Convergence

- Virtually unlimited I/O sizes on each DMA element
- NSc3000 provides NAS file services for SANs, extending the Infiniband benefit to the SAN data delivery
- Future interconnect to SAN? Within SAN? Other point solutions?

Some Specific Numbers

AUSPEX

Before
InfiniBandTM Architecture

- 60 MB/sec PCI bus throughput causes 89% CPU utilization
- 90 MB/sec PCI bus throughput causes 100% CPU utilization

After
InfiniBandTM Architecture

- 60 MB/sec PCI bus throughput causes 12% CPU utilization
- 90 MB/sec PCI bus throughput causes 18% CPU utilization

InfiniBand℠

TRADE ASSOCIATION