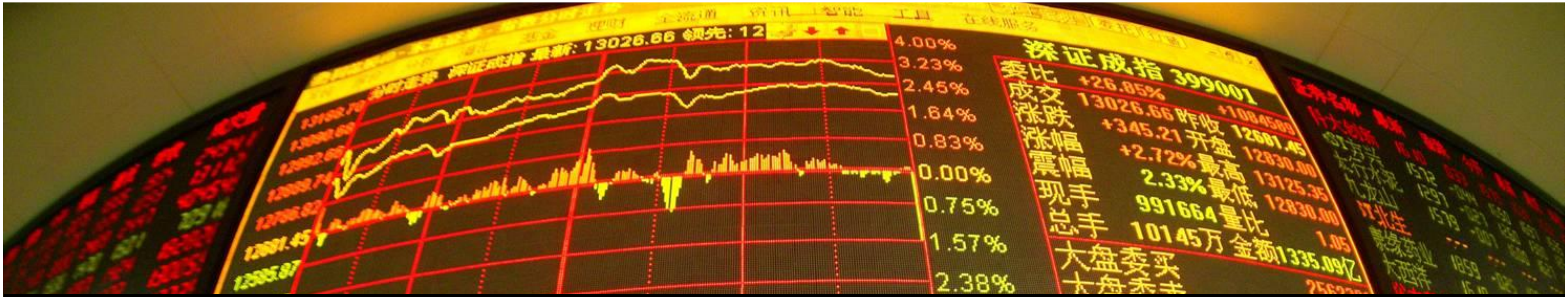
citihub™



▸ eTrading & Market Data ▸ Agile infrastructure ▸ Telecoms ▸ Data Center ▸ Grid

Creating an agile infrastructure with Virtualized I/O

Richard Croucher

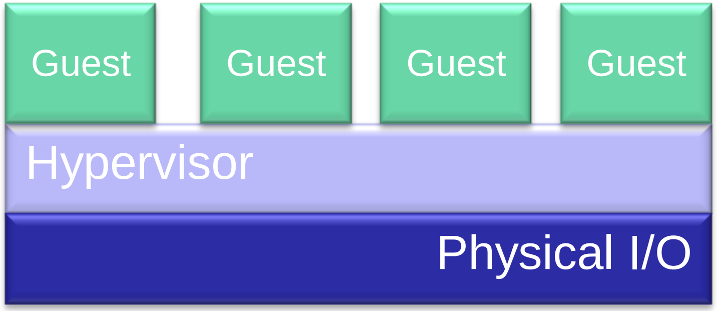
May 2009

Smart Infrastructure Solutions
London ☎ New York ☎ Singapore
www.citihub.com

Problem description

- Majority of current servers virtualized have been low utilization windows servers

- To push further and make Virtualised the default we need to be able to economically accommodate more I/O intensive workloads



- Issues which virtualized I/O help to address:
  - Lack of intrinsic I/O capacity management within existing Virtualized OS environments
  - Plethora of I/O cards and cables which need to purchased, fitted and maintained
  - Multiplicity of different I/O configurations on physical servers which limit mobility options

Virtualized I/O options

- Block level solutions
  - PVSCSI and VM Direct Path I/O
  - iSCSI
  - FibreChannel over Ethernet (FCoE)
  - InfiniBand

There are also NAS solutions, ie. Logical level solutions which we won’t go into in this session. These are commonly used today but do not provide the same level of throughput as block level solutions and make diskless support non-standard (Windows) or complex (UNIX)

PVSCSI and VMDirect Path

citihub™

- VMware continue to add capabilities to their flagship ESX platform
- Paravirtualized SCSI
  - Separate virtualised adapter used by Guest for high performance block level access over existing physical I/O paths
  - Intended for Guests demanding high performance I/O – still dependent on underlying physical connectivity
  - Requires Guest OS support – currently limited to WinSvr2003/2008 and RH Ent5
- VMDirect path I/O for storage
  - Maps a physical HBA to a single guest
  - One for one, no sharing
  - Limited to specific physical adapters

iSCSI

- Enables block level access to storage across TCP/IP networks
- Initiators supported by most OS environments out of the box, and can be used for guests via IP. Supported by VMware ESX.
- Free to use on servers with existing Ethernet ports and uses existing Ethernet infrastructure.   Fully routable protocol. Relies on TCP for data integrity and in-order delivery
- Complexity of setup, particularly setting up and maintenance of the CHAP authentication
- No intrinsic boot support in Windows. Boot disk support dependent on INT13 provided with HW iSCSI cards and not accessible to Guests. Guests can attach to storage via iSCSI after booting or boot from an iSCSI disk presented to it via it's host.
- High TCP/IP overhead.   Can be handled using TOE cards but drives cost up.   Bigger problem on storage arrays (targets) where it limits maximum I/O throughput compared with other protocols.
- Storage vendors typically only target SMB and avoid Enterprise due to the overhead issues on arrays and increased CPU load on servers
- A server running iSCSI target and with FibreChannel can act as a gateway, however adding in resiliency increases complexity significantly.

FC over Ethernet

- Initiative led by Cisco and others under ANSI T11.3 (see www.fcoe.com) to deliver FibreChannel protocol directly over Ethernet.
- Creates no TCP load on server and storage arrays
- Encapsulates FC frame with Header and SCSI command/data inside Ethernet data field. Simplifies de-encapsulation onto physical FC networks
- Requires guaranteed, in-order delivery demanding priority based flow control be added to the Ethernet standard
- All Ethernet components in path need to be FCoE compliant, i.e. Switches need to support Converged Enhanced Ethernet (CEE) to prevent problems.
- Relevant proposed standards are:
  - IEEE 802.1Qbb Priority Flow control
  - IEEE 802.1Qaz Enhance Transmission Selection
  - IEEE 802.1Qau Congestion Notification
- Issues:
  - Standards are still draft, expect to be ratified late 2009 or early 2010
  - Non-routable, server and storage need to be in same subnet, fabric extenders create a STP free L2 fabric

FCoE Product Examples

Converged Network Adapters

- Qlogic QL8100 2x 10Ge *note
- Emulex LP2100 – 2x 10Ge *note
- Brocade 1020 – 2x 10Ge
- ServerEngine – 2x 10Ge *note

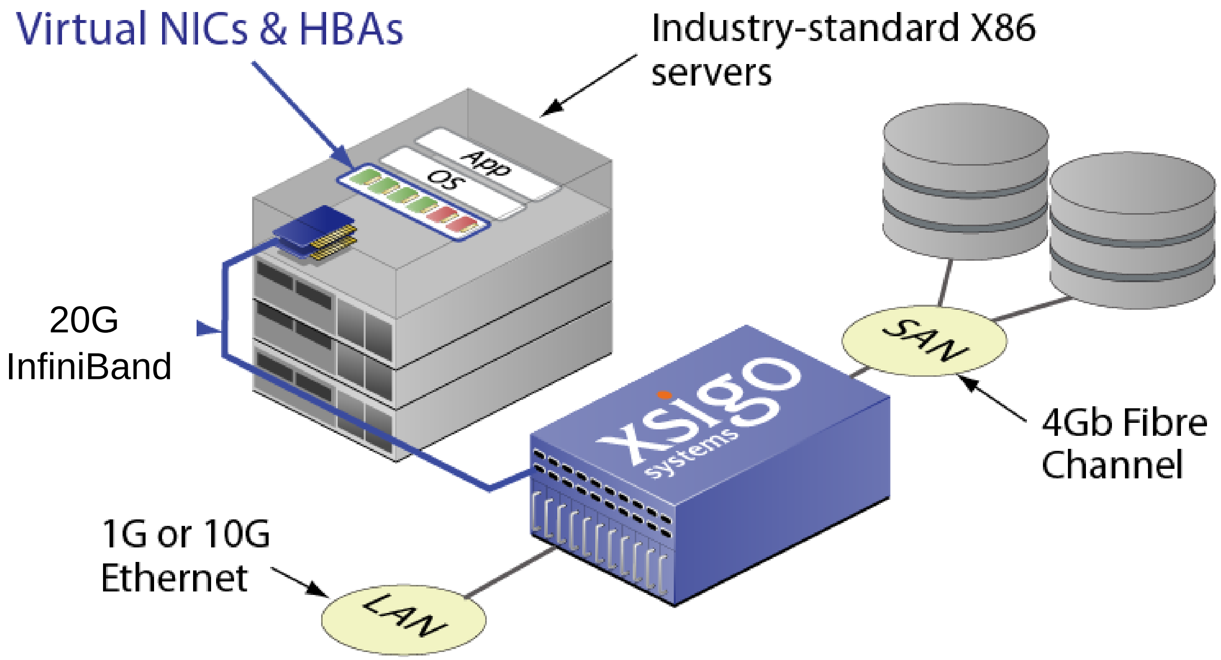
*Note: iSCSI TOE included for backward compatibility

Converged Enhanced Ethernet Switches (pre standard products)

- Cisco Nexus 5010 with 20x 10Ge + 1 expansion slots
- Cisco Nexus 5020 with 40x 10Ge + 2 expansion slots
  - Expansion cards
    - 8x 4G FC, 6x 10Ge, 4x10Ge+ 4x 4G FC
- Brocade 8000 24x 10Ge+ 8x 8G FC
- Blade Network Technologies 24x 10Ge SFP+

Xsigo Virtualized I/O gateway

citihub™



- Virtualized I/O over 20Gb/s InfiniBand
- Guests run vHBA and vNIC, with CIR/PIR based QoS, no need for InfiniBand drivers
- VMware ESX runs InfiniBand stack and allocates vHBA's/vNICs to guests
- I/O director provides physical connectivity into LAN/SAN environments
- 1 physical connector/cable (2 for resiliency) per physical server providing 40Gb/s
- Xsigo Management Server integrated into VMware Virtual Center
- Resold by Dell

Xsigo VP780 I/O Director

Hardware-based architecture

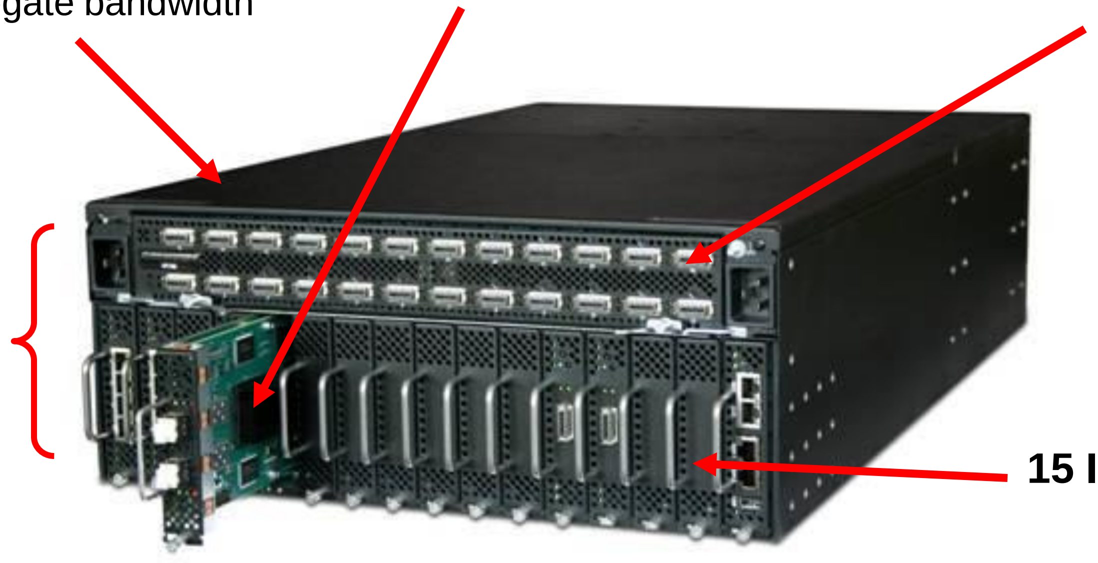
- Fully non-blocking fabric
- 780 Gb/s aggregate bandwidth

Custom silicon

Line rate throughput

24 Server ports

- Expansion switch available for connection to hundreds of servers



4U height

15 I/O module

4x 1Ge

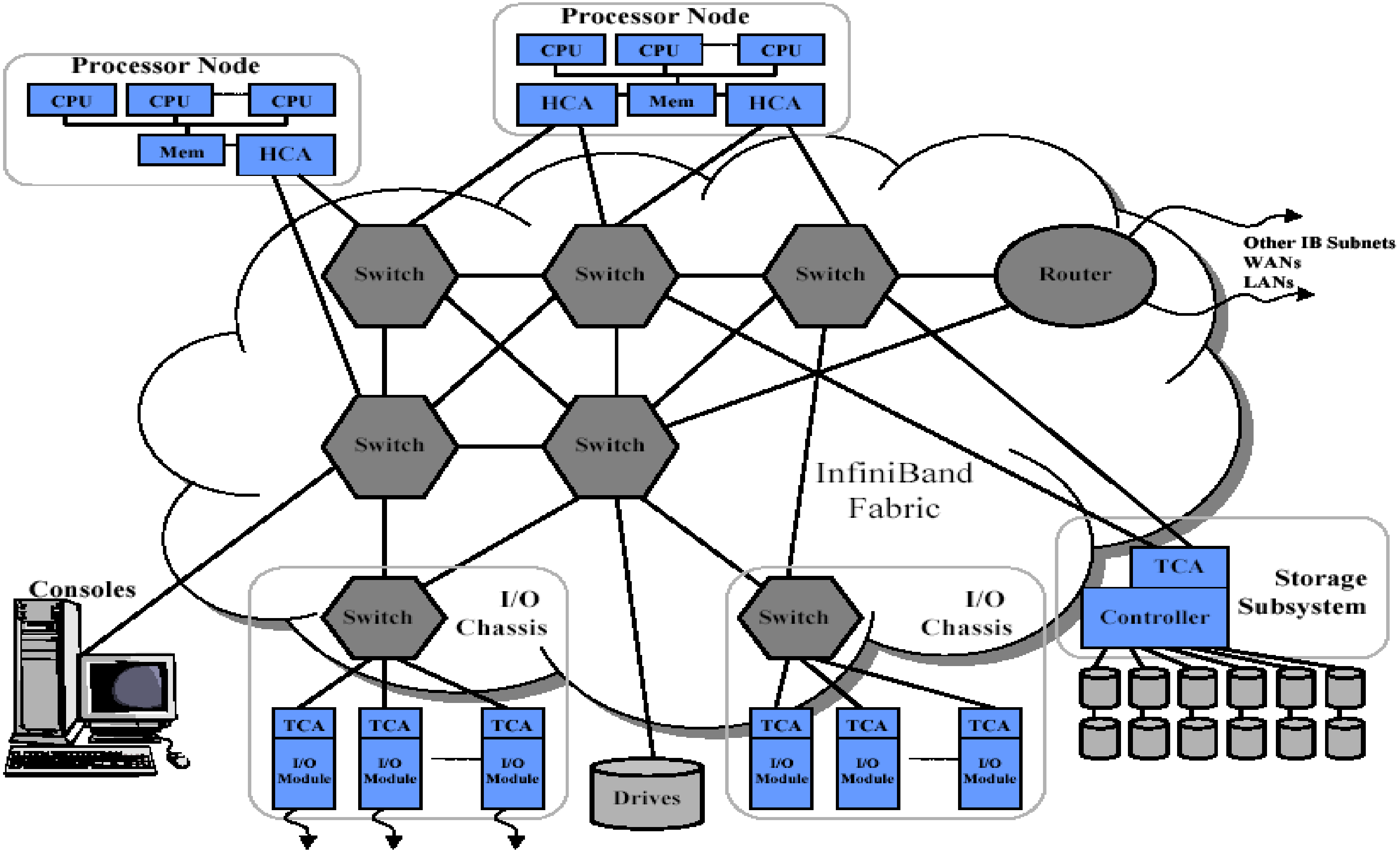
10x 1Ge

1x 10Ge

2x 4G FC

InfiniBand fabric

citihub™



What is InfiniBand

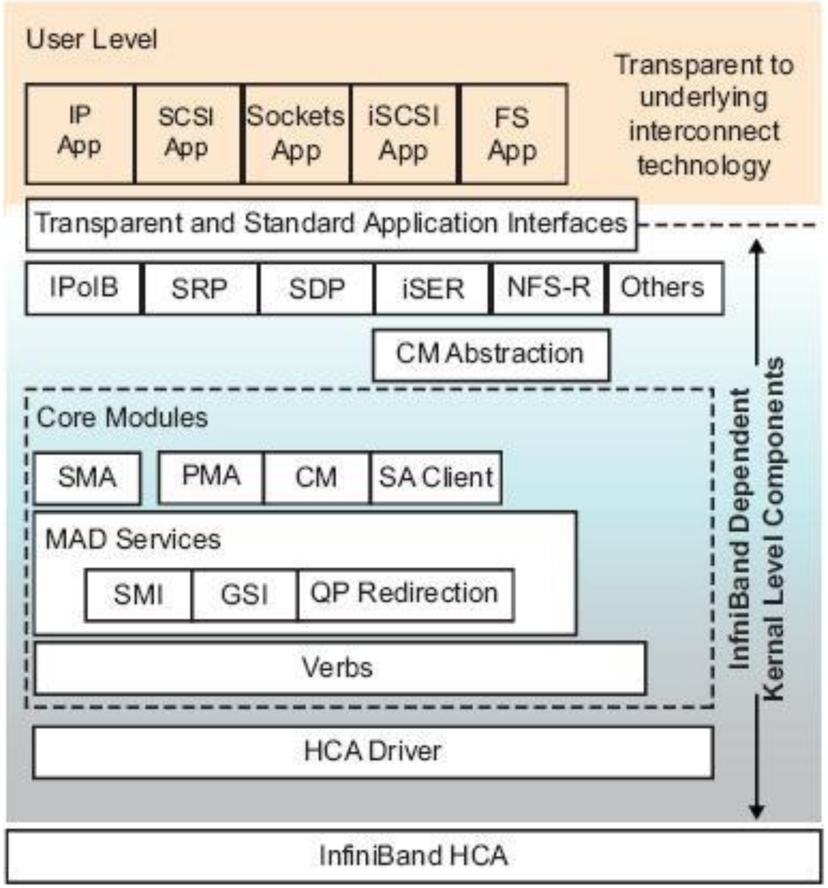
InfiniBand is a I/O protocol designed to provide high bandwidth, low-latency interconnect for clustering. It has been designed to offload CPU overhead by incorporating powerful RDMA (Remote Direct Memory Access) engines.

Mature standard with proven vendor interoperability

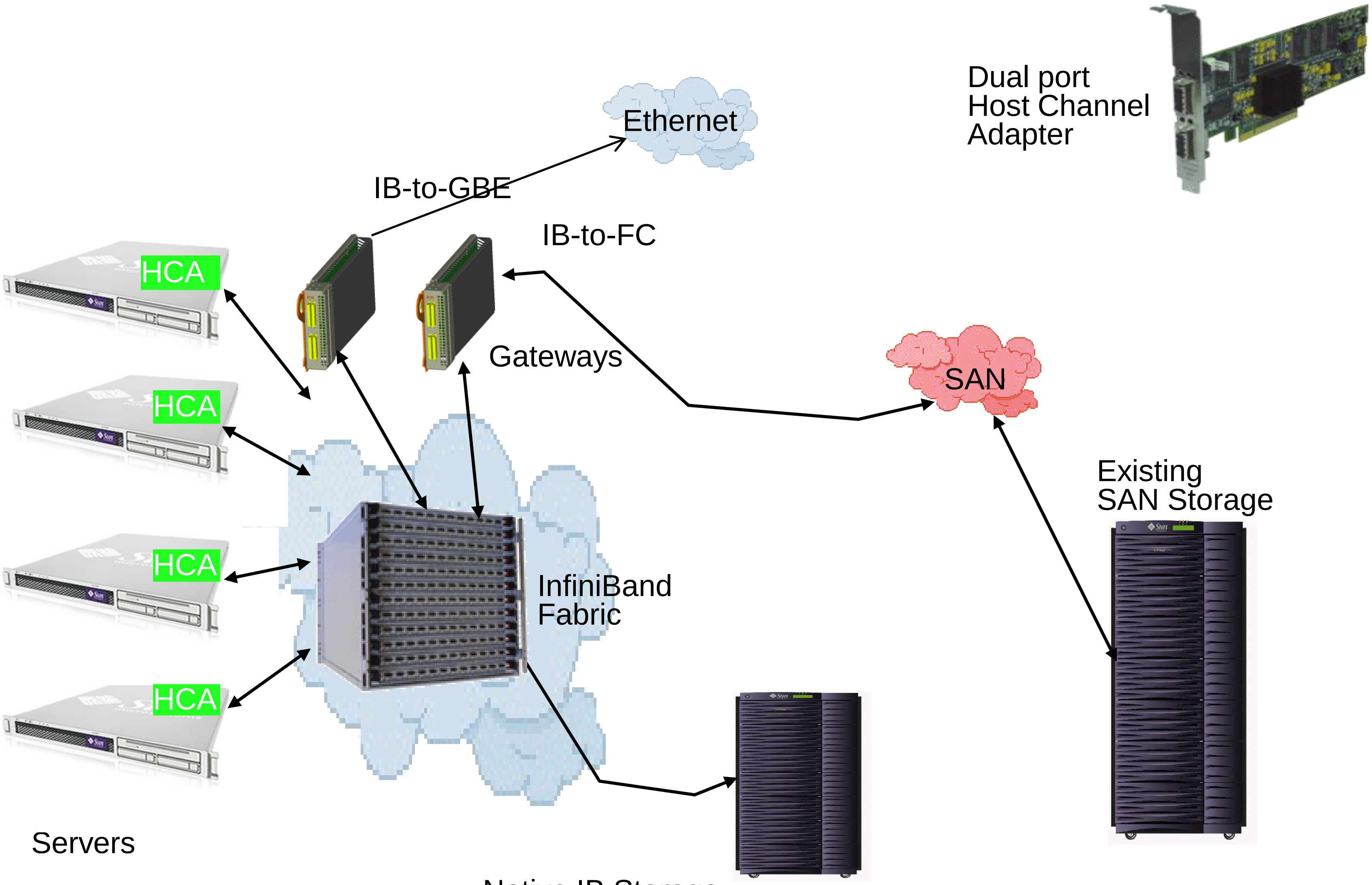
Historically deployed for HPC Grids but now entering Enterprise to support Virtualization rollouts

IB Switch functionality allows cut-through packets at wire speed with link and end-to-end data integrity

Implicit trunking across multiple serial lanes provides protocol independent speed improvements
1x = 2.5 Gbps 10B/8B = 2.0 Gbps data
4x = 10 Gbps == 1G Byte/second data transfer
12x = 30 Gbps == 3G Byte/second data transfer
Double Data Rate = 5 Gbps
4xDDR = 20 Gbps == 2G Byte/second data transfer
Quad Data Rate = 10 Gbps
4xQDR == 4G Byte/second transfer



InfiniBand as a virtual interconnect



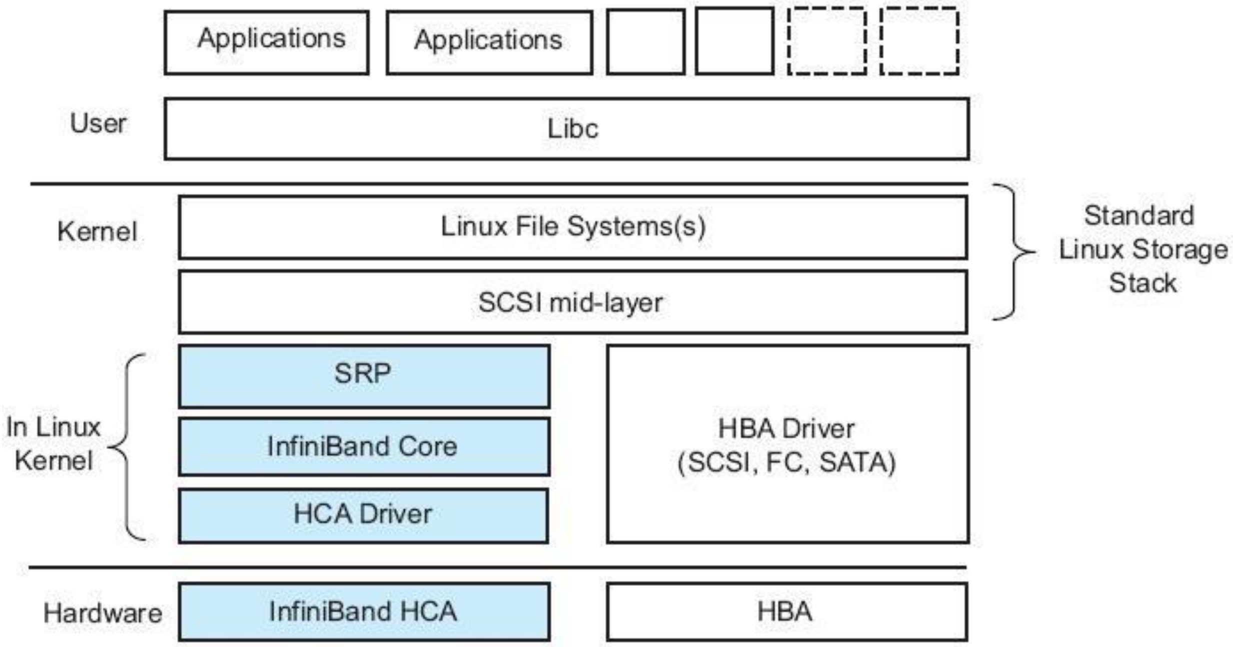
InfiniBand Ethernet and TCP/IP Integration

- IP over IB
  - Included with all InfiniBand implementations
  - Runs full TCP/IP stack including TCP/UDP and multicast
  - Connects InfiniBand attached nodes or through gateway to Ethernet
  - Does not leverage RDMA or support raw Ethernet
- vNIC
  - Layer 2 interface support raw Ethernet
  - Connects via compatible gateway to Ethernet
  - Does not leverage RDMA
- Sockets Direct Protocol
  - Bypasses TCP stack/overhead
  - Supports existing TCP socket applications
  - Preload libraries avoid recompilation
  - Connects InfiniBand attached nodes

InfiniBand Storage standards

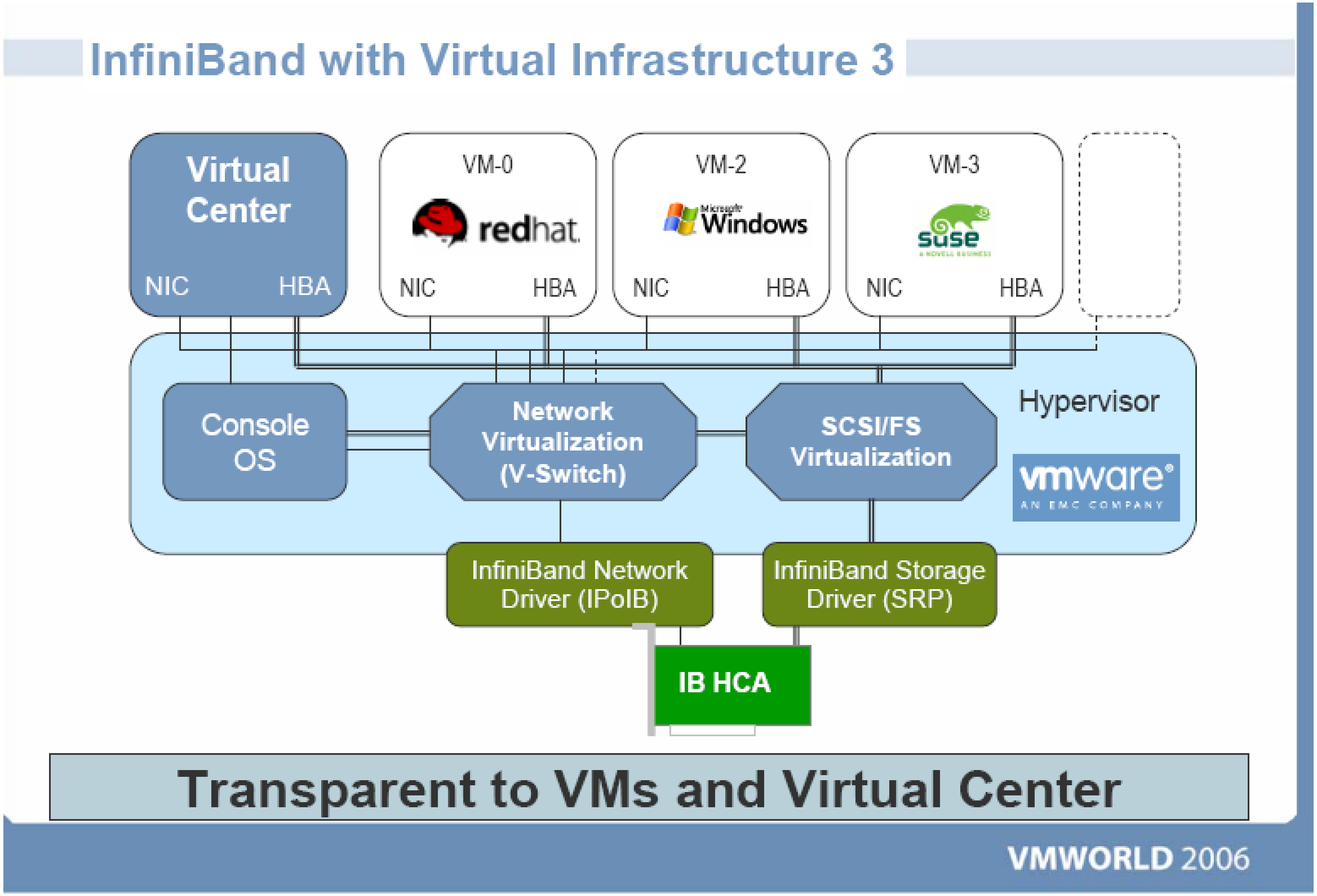
- SCSI RDMA Protocol (SRP)
  - Defined by ANSI T.10
  - Bypasses TCP/IP stack
  - All InfiniBand switch vendors offer gateways to FibreChannel
  - Driver support for Linux, Solaris, Windows, VMware ESX
- iSCSI extension for RDMA (iSER)
  - Initially defined by Voltaire, now IETF draft
  - Now open sourced and accepted by OpenIB into Linux implementation
  - Bypasses TCP/IP stack
- vHBA
  - Non standardized, Gateway specific, available for Qlogic, Mellanox and Xsigo

SCSI RDMA Protocol (SRP)



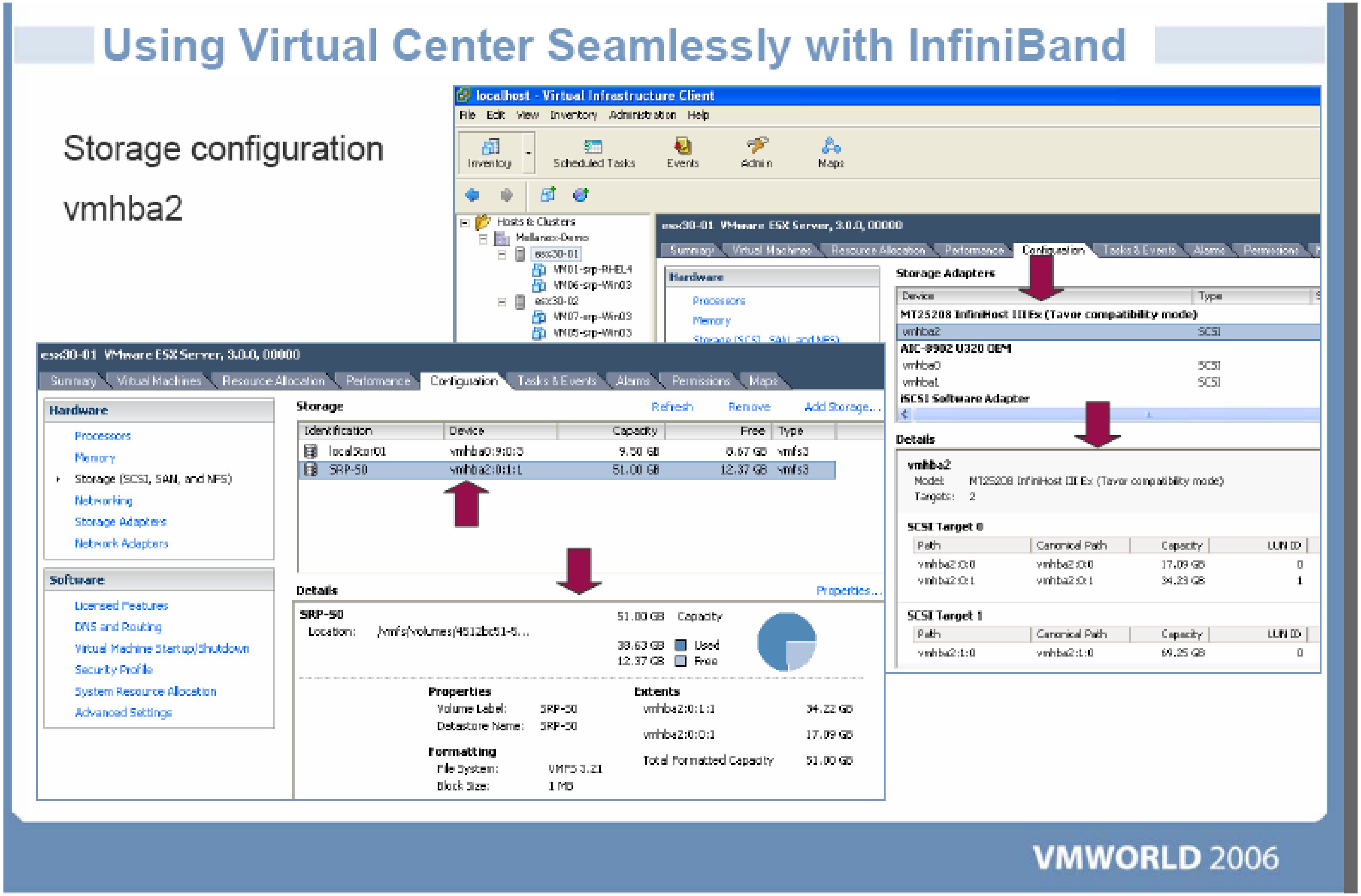
- Provides block level support
- Interposes below existing SCSI driver, just appears as a block level device
- Uses RDMA, requiring less CPU than FibreChannel
- Supported by VMware ESX, Windows, Linux, Solaris
- Host can connect through SRP and present block level device to Guest

VMware InfiniBand support - 1


InfiniBand with Virtual Infrastructure 3
Virtual Center
NIC
HBA
VM-0
redhat
NIC
HBA
VM-2
Microsoft Windows
NIC
HBA
VM-3
suse
NIC
HBA
Console OS
Network Virtualization (V-Switch)
SCSI/FS Virtualization
Hypervisor
vmware AN EMC COMPANY
InfiniBand Network Driver (IPoIB)
InfiniBand Storage Driver (SRP)
IB HCA
Transparent to VMs and Virtual Center
VMWORLD 2006

InfiniBand: Enhancing Virtualization ROI with new Data Center Efficiencies, Sujal Das, Mellanox

VMware InfiniBand support - 2



Using Virtual Center Seamlessly with InfiniBand

Storage configuration

vmhba2

localhost - Virtual Infrastructure Client

Storage

Identification	Device	Capacity	Free	Type
localStor01	vmhba0:9:0:3	9.50 GB	8.67 GB	vmfs3
SRP-50	vmhba2:0:1:1	51.00 GB	12.37 GB	vmfs3

Details

SRP-50
Location: /vmfs/volumes/4512bc51-5...

51.00 GB Capacity
38.63 GB ■ Used
12.37 GB □ Free

Properties
Volume Label: SRP-50
Datastore Name: SRP-50

Formatting
File System: VMFS 3.21
Block Size: 1 MB

Extents

vmhba2:0:1:1	34.22 GB
vmhba2:0:0:1	17.09 GB
Total Formatted Capacity	51.00 GB

Storage Adapters

Device	Type
MT25208 InfiniHost IIIEx (Tavor compatibility mode)	
vmhba2	SCSI
AIC-8902 U320 OEM	
vmhba0	SCSI
vmhba1	SCSI
iSCSI Software Adapter	

Details

vmhba2
Model: MT25208 InfiniHost III Ex (Tavor compatibility mode)
Targets: 2

SCSI Target 0

Path	Canonical Path	Capacity	LUN ID
vmhba2:0:0	vmhba2:0:0	17.09 GB	0
vmhba2:0:1	vmhba2:0:1	34.23 GB	1

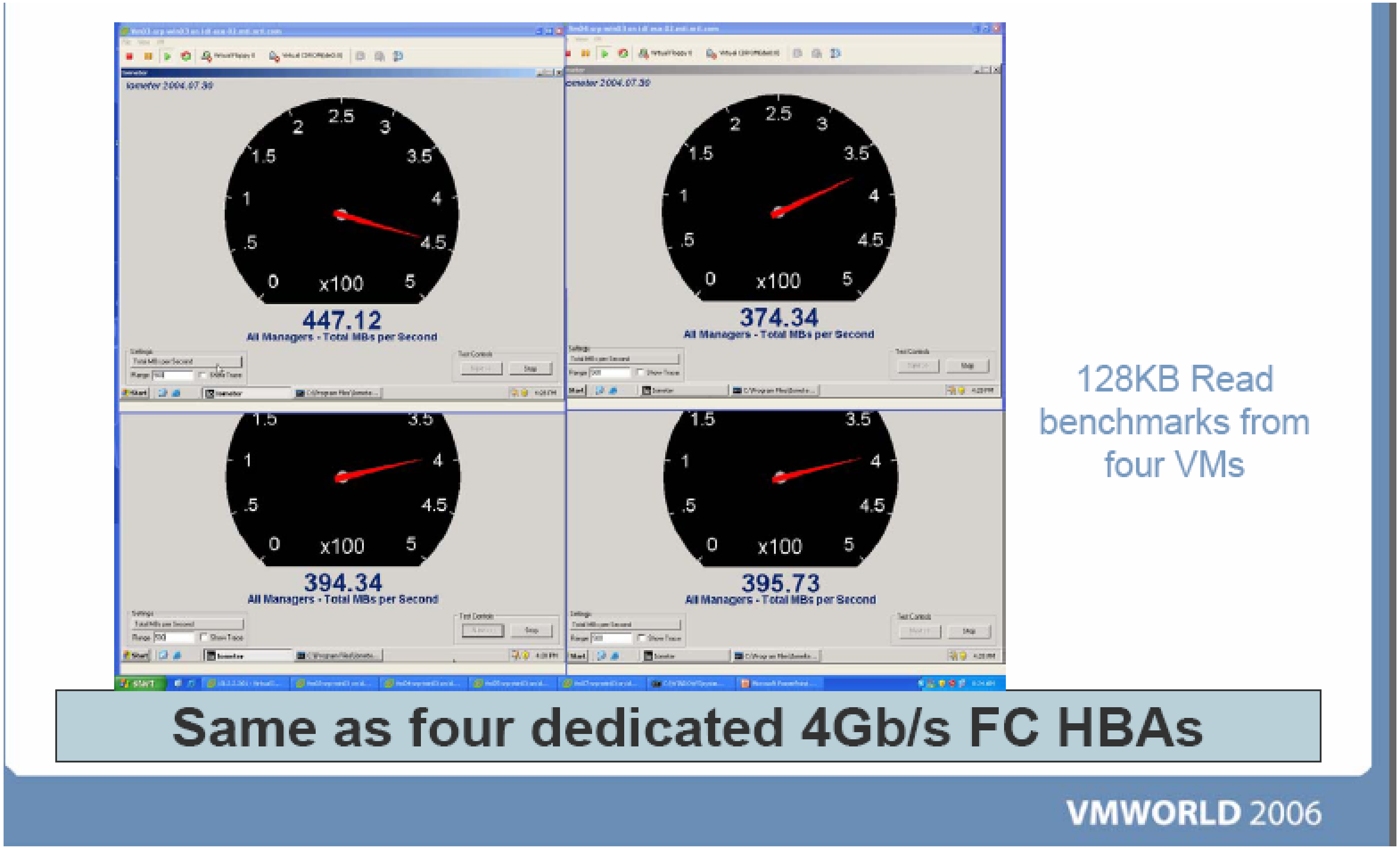
SCSI Target 1

Path	Canonical Path	Capacity	LUN ID
vmhba2:1:0	vmhba2:1:0	69.25 GB	0

VMWORLD 2006

InfiniBand: Enhancing Virtualization ROI with new Data Center Efficiencies, Sujal Das, Mellanox

VMware InfiniBand support - 3



Iometer 2004.07.30

447.12
All Managers - Total MBs per Second

374.34
All Managers - Total MBs per Second

394.34
All Managers - Total MBs per Second

395.73
All Managers - Total MBs per Second

128KB Read benchmarks from four VMs

Same as four dedicated 4Gb/s FC HBAs

VMWORLD 2006

InfiniBand: Enhancing Virtualization ROI with new Data Center Efficiencies, Sujal Das, Mellanox

Qlogic Ethernet and FibreChannel gateways

Fibre Channel Leaf Module

Ethernet Leaf Module

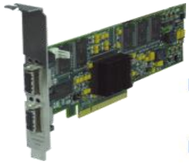
Fibre Channel VIC (vHBA)

- 10-port 20G InfiniBand
- 8-port 1/2/4 Gb FC

10 Gigabit Ethernet VIC (vNIC)

- 10 port 20G InfiniBand
- 2-port 10 GbE

- Modules fit inside Qlogics director class InfiniBand switches
- Directors for 4,8,12, 24 modules
- Also 12 port 20G InfiniBand modules



QLOGIC®

- Latest range is 36 – 864x 40G InfiniBand switches
- 40G gateway planned for 2010

Mellanox InfiniBand switches and gateways

BridgeX BX4000
4x 40G InfiniBand
12x 10Ge (vNIC and IPoIB)
Or
16x 8G FC (vHBA)

ConnectX
2x 40G IB or 2x 10Ge

MTS3600
32x 40G InfiniBand

10Ge ports FCoE (draft) compatible
Full driver stack for VMware, Linux and Windows
Boot over InfiniBand for diskless host servers

Voltaire InfiniBand

SR4G

FibreChannel gateway
4x 4G FC + 2x 20G InfiniBand
1U server profile
iSER + iSCSI support

sRB-20210G

Ethernet gateway
2x 10Ge + 22x 20G InfiniBand
Module option for Director Class switches
Directors for 4, 6, 12 modules
Also
24 port 20G InfiniBand modules
32 port 40G standalone switch

Voltaire GridVision Virtualization Management

Automatically discover physical objects

- **Switches, routers, storage, physical servers, …**

Define lo

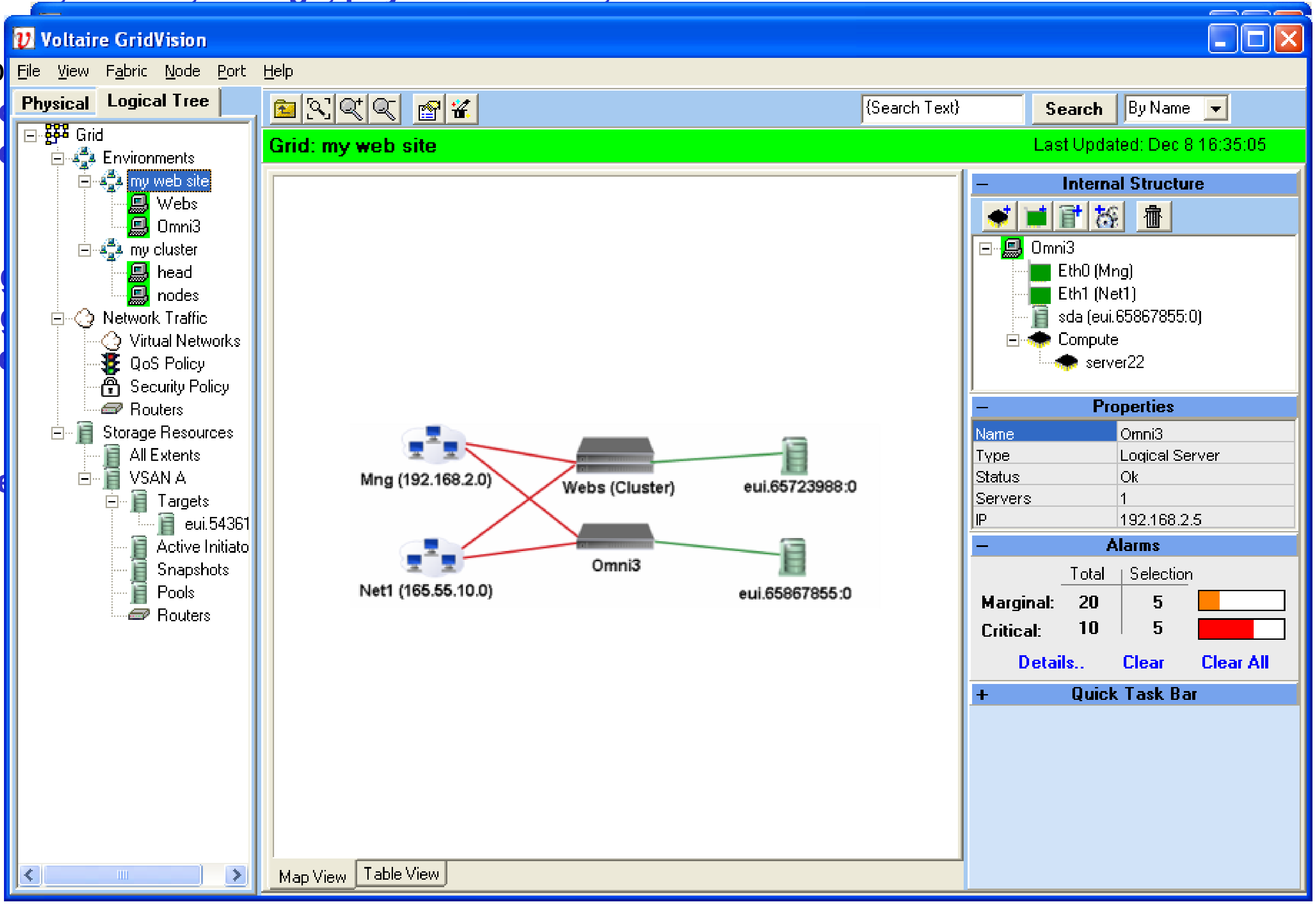
- **Envir**
- **Define**

Bind the

- **Confi**
- **Confi**
- **Provi**

Abstract

- **Single**



Grid: my web site — Last Updated: Dec 8 16:35:05

Properties

Name	Omni3
Type	Logical Server
Status	Ok
Servers	1
IP	192.168.2.5

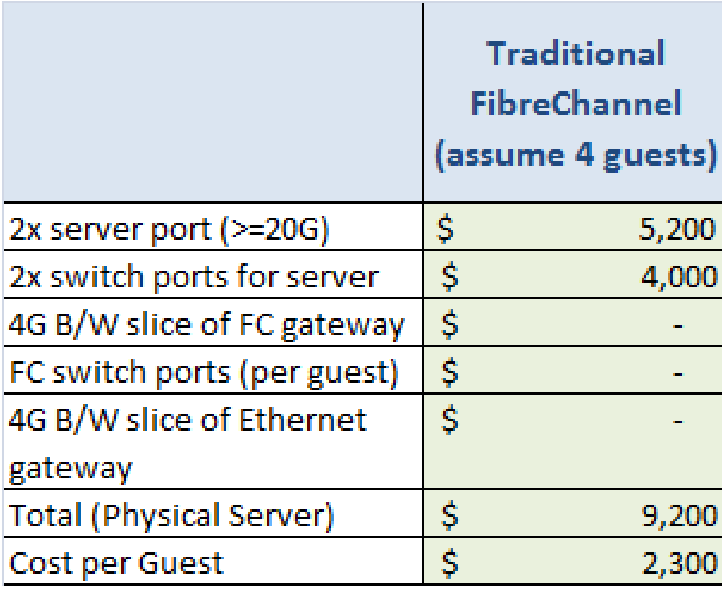
Alarms

	Total	Selection
Marginal:	20	5
Critical:	10	5

InfiniBand based virtual I/O deployment example

- Can use low profile servers since only one dual port PCIe card provides up to (80Gb/s)
- InfiniBand drivers to Manager only
- Guests use vNICs, vHBA's
- Connect multiple racks into a single virtual I/O cluster
  - Leverages high bandwidth
  - Shares cost of gateways across multiple servers particularly for resilient configurations

Cost comparison to existing FC storage

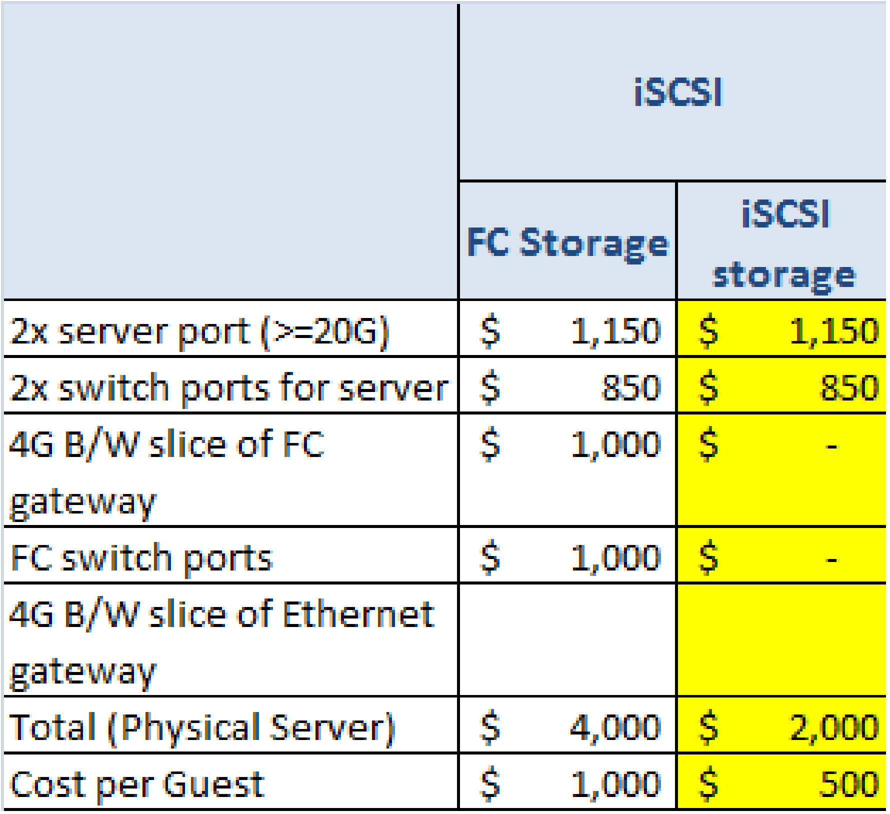


	Traditional FibreChannel (assume 4 guests)
2x server port (>=20G)	$ 5,200
2x switch ports for server	$ 4,000
4G B/W slice of FC gateway	$ -
FC switch ports (per guest)	$ -
4G B/W slice of Ethernet gateway	$ -
Total (Physical Server)	$ 9,200
Cost per Guest	$ 2,300

Excludes cost of server and storage

Your prices may vary

Cost comparison to iSCSI or FCoE storage



	iSCSI	
	FC Storage	iSCSI storage
2x server port (>=20G)	$ 1,150	$ 1,150
2x switch ports for server	$ 850	$ 850
4G B/W slice of FC gateway	$ 1,000	$ -
FC switch ports	$ 1,000	$ -
4G B/W slice of Ethernet gateway		
Total (Physical Server)	$ 4,000	$ 2,000
Cost per Guest	$ 1,000	$ 500

Your prices may vary

Excludes cost of server and storage
note increased storage cost of iSCSI due to TCP overhead

Conclusions

- I/O virtualization simplifies virtual server deployment and enables a wider range of applications to be supported
- iSCSI is low end solution due to server CPU and storage array TCP/IP overheads
- FCoE is the strategic solution but:
  - ALL products are still pre-standard – demand free upgrade from suppliers
  - Recommend separate storage Ethernet infrastructure
  - Still missing routing and long distance solutions
  - Real savings only after storage migrated from FibreChannel to direct Ethernet
- InfiniBand is available today, well supported, lower cost, higher bandwidth server attach than both FibreChannel or 10Ge
  - Tactical solution until Ethernet standards settle. May need to wait for next generation 40/100G Ethernet
  - Routing and long distance solutions available today for InfiniBand
  - Can be deployed without adding InfiniBand drivers to Guests

Richard.Croucher@Citihub.com

Independent advice for Data Centre Infrastructure
London – New York - Singapore