



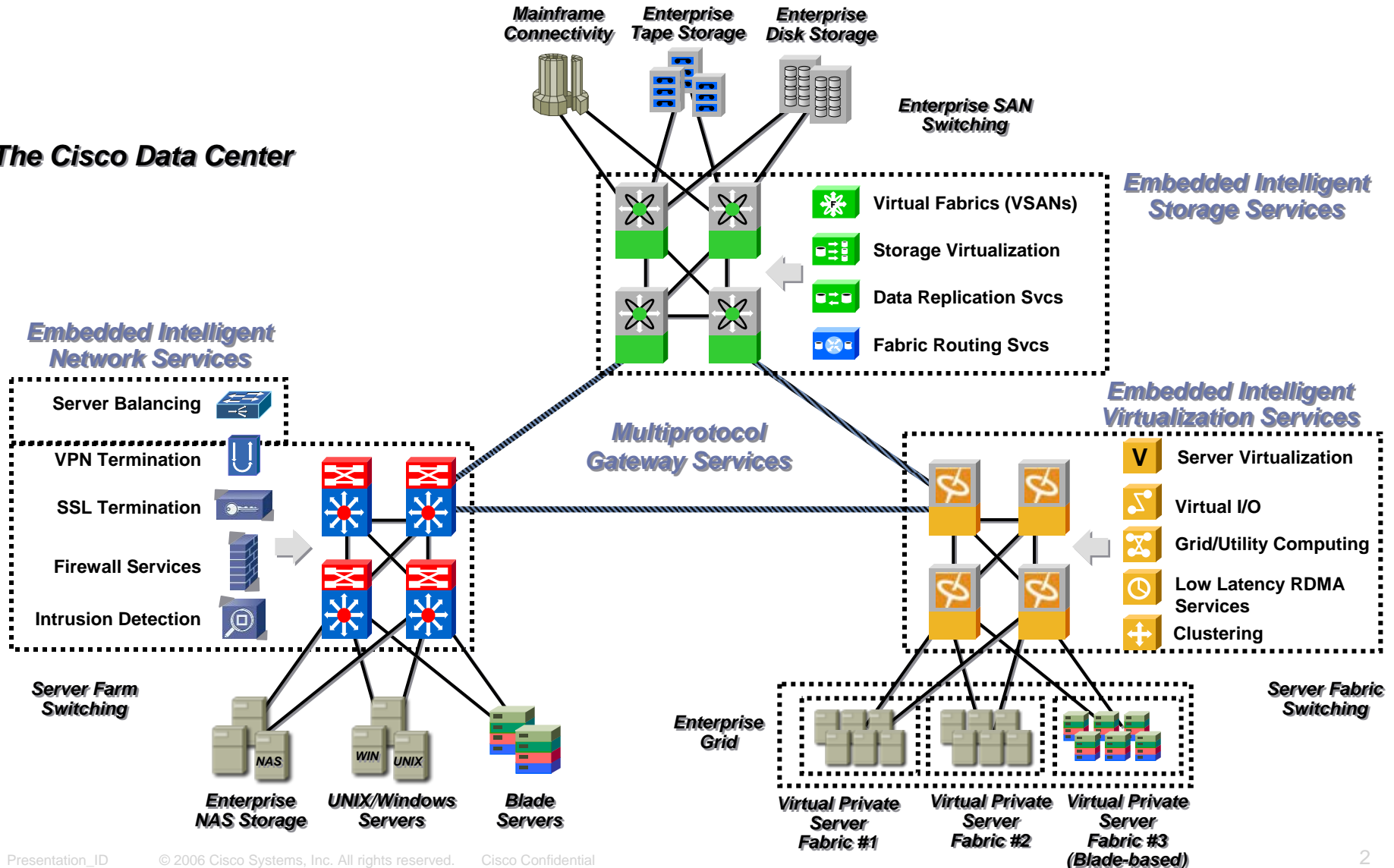
Infiniband Solutions in Enterprise Datacenters



Walter Dey
Distinguished System Engineer PhD
Cisco Systems Emerging Markets
wdey@cisco.com

The Big Picture - The Cisco Data Center

The Cisco Data Center

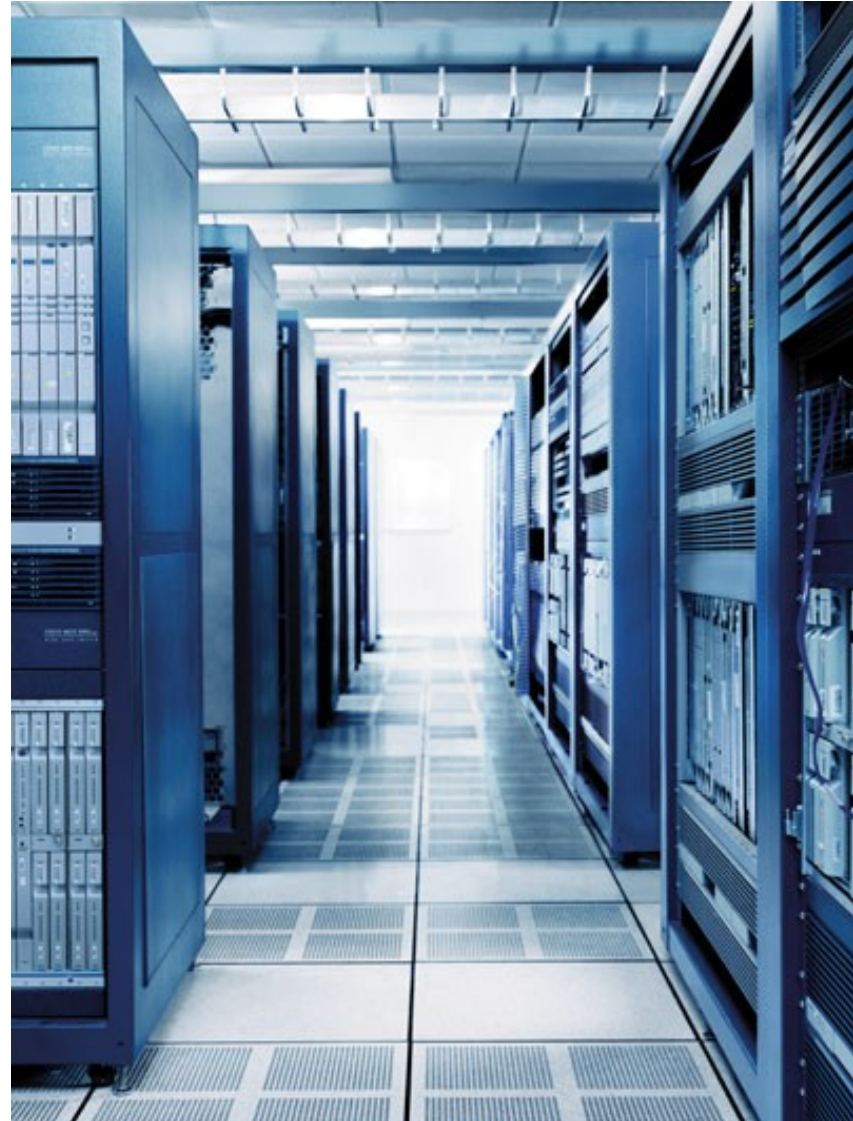


InfiniBand Technology

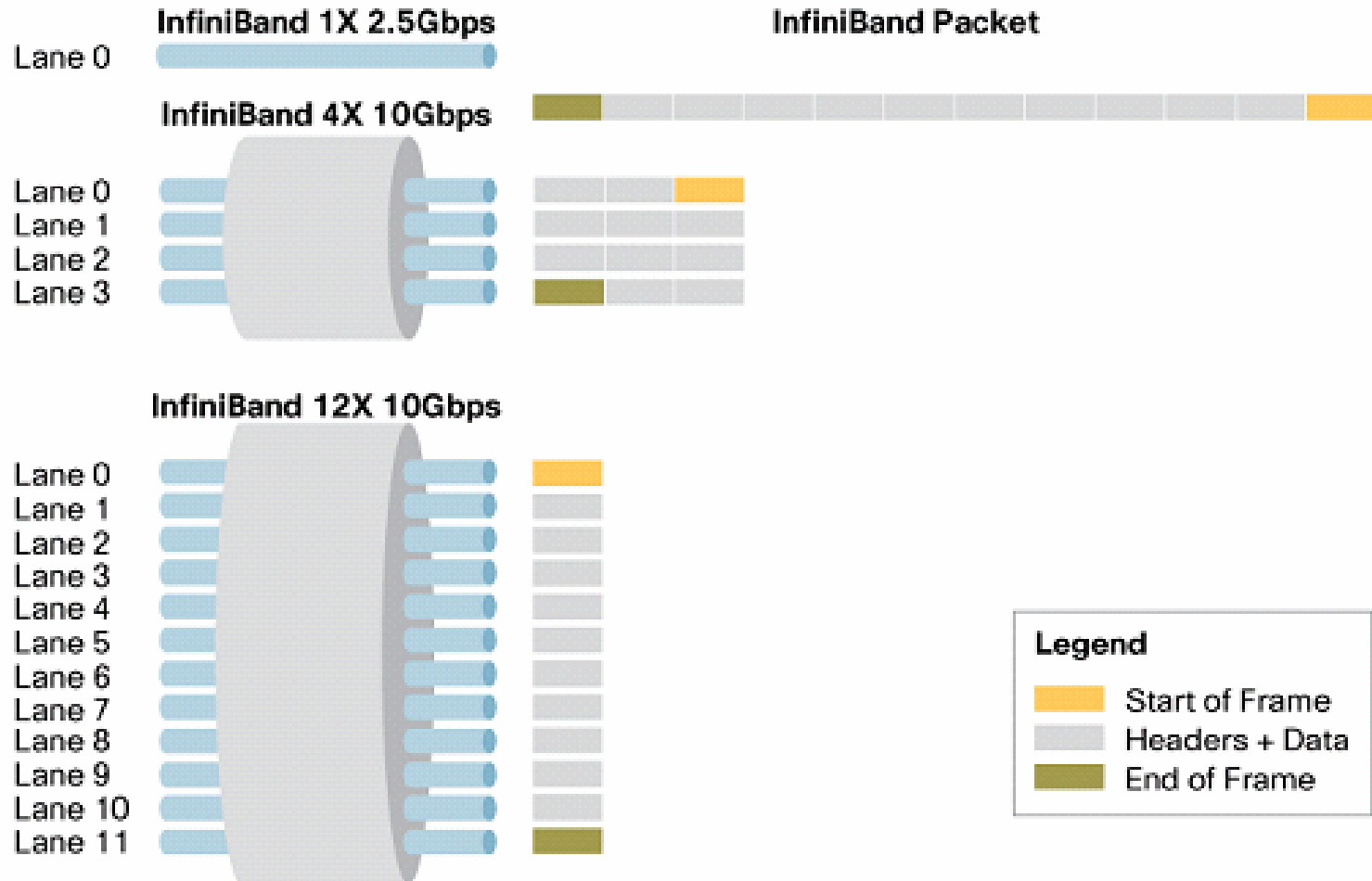


Infiniband Overview

- Standards-based interconnect
<http://www.infinibandta.org>
- Channelized, connection-based interconnect optimized for high performance computing
- Supports server and storage attachments
- Bandwidth Capabilities (SDR/DDR)
 - 1x—2.5/5 Gbps: 2/4 Gbps actual data rate (base rate for InfiniBand)
 - 4x—10/20 Gbps: 8/16 Gbps actual data rate
 - 12x—30/60 Gbps: 24/28 Gbps actual data rate
- Built-in RDMA as core capability for inter-CPU communication



Parallel Transmission and “Virtual Lanes”



Signaling / Data Rates

InfiniBand Link	Signal Pairs	Signaling Rate	Data Rate (Full Duplex)
1X-SDR	2	2.5 Gbps	2.0 Gbps
4X-SDR	8	10 Gbps (4 x 2.5 Gbps)	8 Gbps
12X-SDR	24	30 Gbps (12 x 2.5 Gbps)	24 Gbps
1X-DDR	2	5 Gbps	4.0 Gbps
4X-DDR	8	20 Gbps (4 x 5 Gbps)	16 Gbps
12X-DDR	24	60 Gbps (12 x 5 Gbps)	48 Gbps
1X-QDR	2	10 Gbps	8.0 Gbps
4X-QDR	8	40 Gbps (4 x 10 Gbps)	32 Gbps
12XQDDR	24	120 Gbps (12 x 10 Gbps)	96 Gbps

Note: Although the signaling rate is 2.5 Gbps, the effective data rate is limited to 2 Gbps because of the 8B/10B encoding scheme: $(2.5 \times 8) \div 10 = 2$ Gbps

InfiniBand Connections

- **Copper and Fibre interfaces are specified in the InfiniBand Standards**

- **Copper**

 - Up to 8m for 4x DDR connections

 - Up to 10m for 12x SDR connections

- **Optical**

 - Requires an additional transceiver

 - Up to 150m for all connections

 - Long Haul possible, leverages DWDM infrastructure (Obsidian)

1X Connector



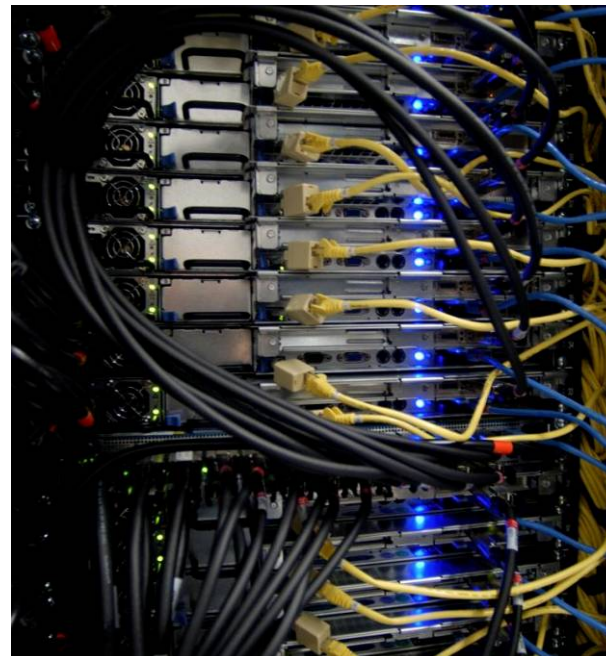
4X Connector



12X Connector



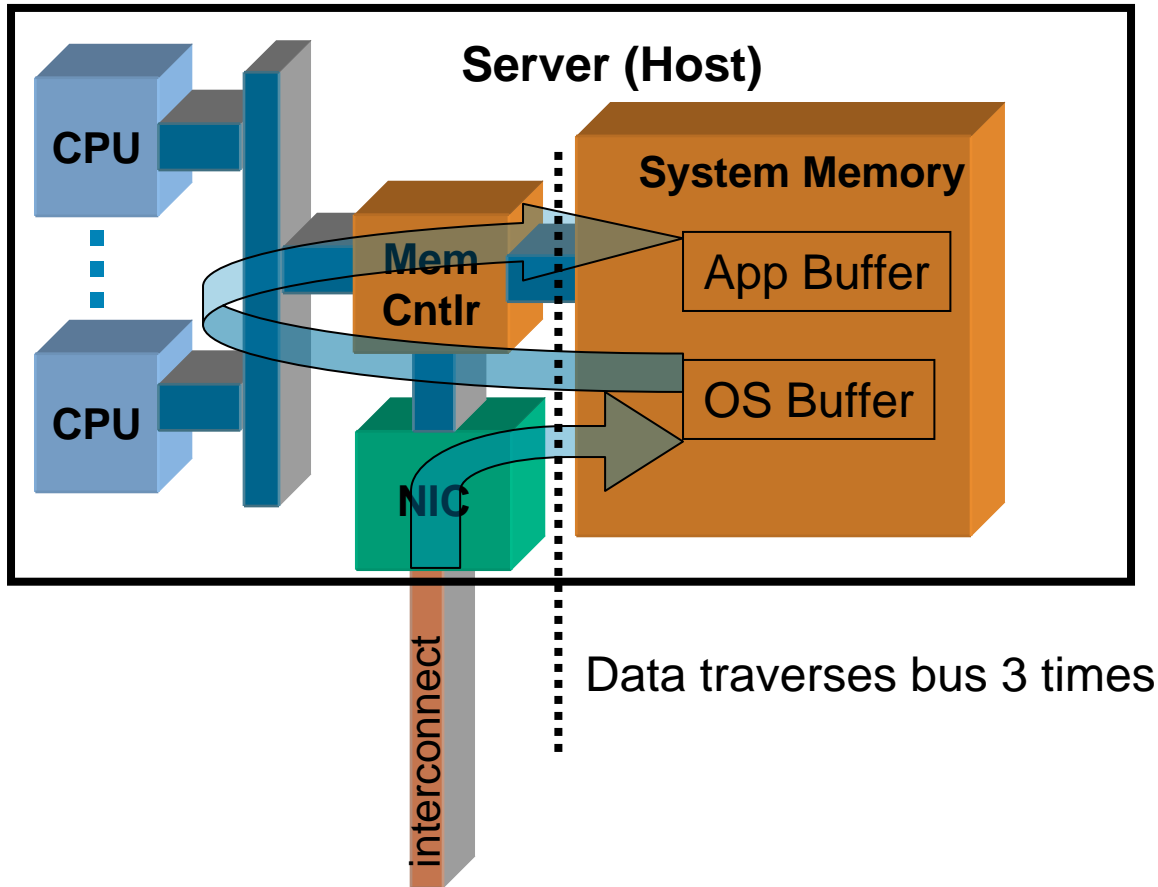
4X Infiniband Cluster Copper Cabling



InfiniBand Protocol Summary

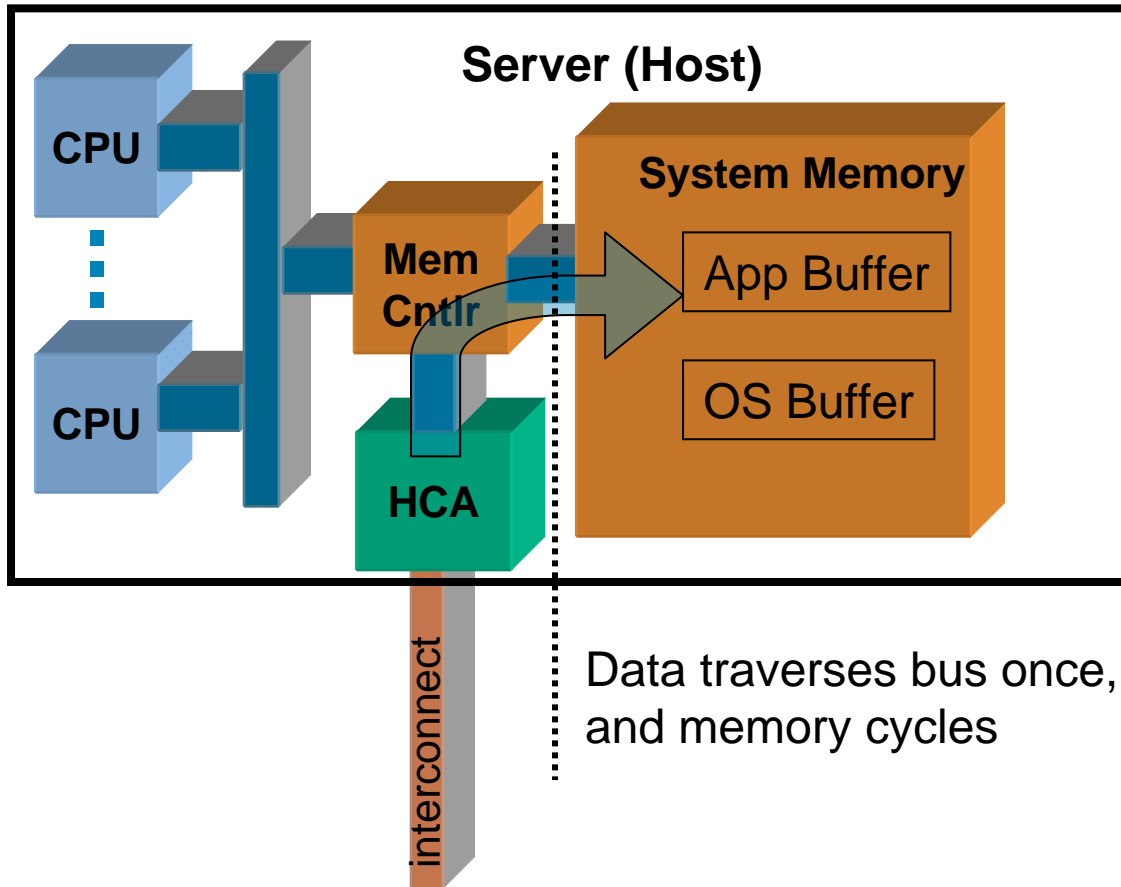
Protocol / Application	Summary	Application Example
IPoIB (IP over InfiniBand)	Enables IP-based applications to run over InfiniBand transport	Standard IP-based applications - when used in conjunction with Ethernet Gateway, allows connectivity between IB network and LAN
SDP (Sockets Direct Protocol)	Accelerates sockets-based applications using RC and/or RDMA	Communication between database nodes and application nodes, as well as between database instances
SRP (SCSI RDMA Protocol)	Allows InfiniBand-attached servers to utilize block storage devices	When used in conjunction with the Fibre Channel gateway, allows connectivity between IB network and SAN
uDAPL (Direct Access Programming Library)	Enables maximum advantage of RDMA flexible programming API	Used for IPC communication between cluster nodes for Oracle 10G RAC
MPI (Message Passing Interface)	Low latency protocol used widely in HPC environments	HPC applications

Standard NIC Architecture



- Multiple context switches robs CPU cycles from actual work
- Memory bandwidth and per packet interrupts limit max throughput
- OS manages end-to-end communications path

With RDMA and OS Bypass



- Secure Memory – Memory transfers with no CPU overhead
- PCI-X/PCI-e becomes the bottleneck for network data transmission
- **HCA manages remote data transmission**

IP over InfiniBand (IPoIB)

- Transmission of IP over InfiniBand
 - Use IB as a link layer for IP
 - Use InfiniBand UD transport mode
 - Used for other protocol address resolution
 - Encapsulation for ARP and IPv4
 - Transport IP multicast over IB
- Provides highest level of application compatibility
- Applications do not need to be rewritten or recompiled
- Standard IP utilities and applications work as usual

Sockets Direct Protocol (SDP)

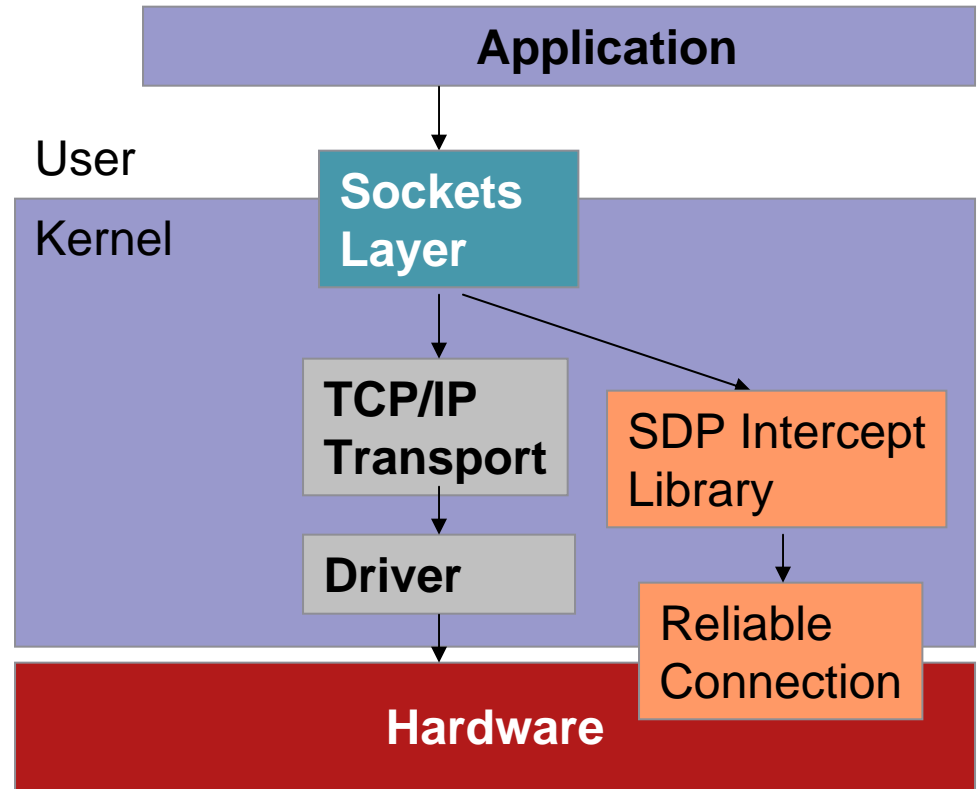
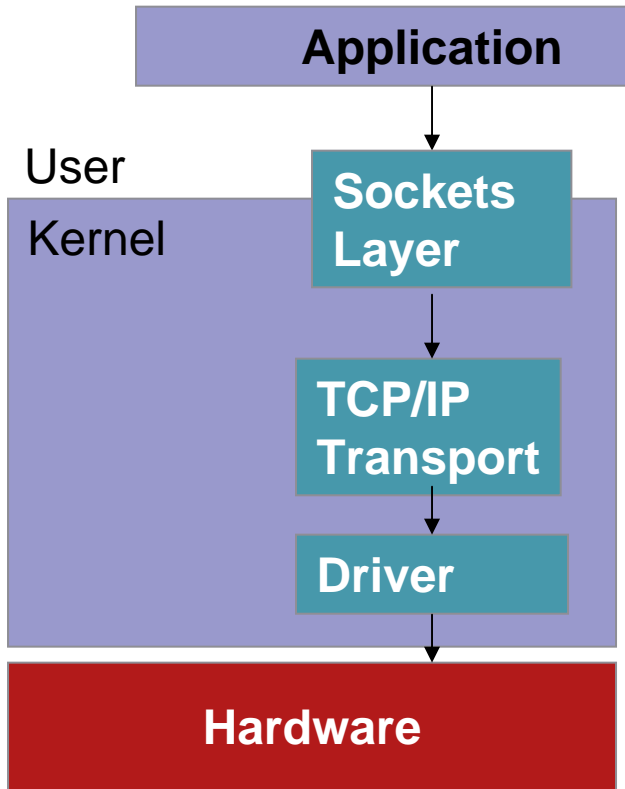
- Sockets Direct Protocol
- Runs socket based TCP/IP traffic with TCP and copy offload
- Highly configurable:
 - By process
 - By port
 - By destination
 - By environment variable
- No application recompile or rework necessary
- Zero copy capability using Asynchronous I/O (AIO)
 - AIO requires application rework

Sockets Direct Protocol (SDP)

Traditional Model



SDP Library Model



SCSI RDMA Protocol (SRP)

- SCSI Semantics over RDMA fabric
- Not IB specific
- Host drivers tie into standard SCSI/Disk interfaces in kernel/OS
- Can be used for end-to-end IB storage (implemented today!)

InfiniBand Solutions



InfiniBand Vertical Solutions



Vertical Industry Solutions for HPC



Financial Services

Low-latency, High-message rate market data environments

Real-time analytics

JPMC – 2000+ Servers in Global Deployment

Citi – Fixed Income Trading



Oil & Gas

Increase accuracy of Reservoir Modeling and Seismic Analysis

Deliver large datasets optimally

Statoil – Multiple Clusters

ONGC

ENI

Occidental



Manufacturing

Reduce time to market for new products

Better Safety & Product Design through Simulation

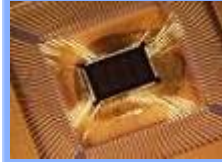
Honda

Ferrari – F1

RedBull Racing

Airbus

Boeing



EDA

Shorten Time for Tape-Out

Improve Yield

Intel

Motorola

TSMC

Altis Semiconductor



Federal Govt.

DoE – Faster Time to Science

DoD – Data Mining, Time to Scientific Research

Sandia National Labs

Maui HPC

ARL

Fleet Numeric



Academic Research

Expand Research Capabilities

Complex Research Problems

Greater Industry Outreach

NCSA @ UIUC

Stanford Univ

MIT

Harvard Univ

UNC Chapel Hill



Biotech

Accelerate time to market

Molecular Modeling and Protein folding experiments for drug discovery

DE Shaw R&D

Cedar Sinai

Stanford BioX

Scripps Institute

High Performance SFS InfiniBand Networking Solutions

InfiniBand Horizontal Solutions



Horizontal Solutions for SFS InfiniBand



HPC

High Performance server to server communication

Latency Sensitive

High Message Rate and throughput desirable

Financial Services

Oil & Gas

Manufacturing

Academic and Research Labs

Biotech



Oracle RAC Clusters

Accelerating 10g RAC Environments



High Performance Storage

- High Performance SAN Access

- High Performance NAS Access

- Support for Parallel and Clustered File Systems

- ISVs – IBRIX, CFS

- NAS – NetApp, EMC, Isilon, Panasas

- SAN – IBM, EMC, HDS, NetApp, DDN, Engenio



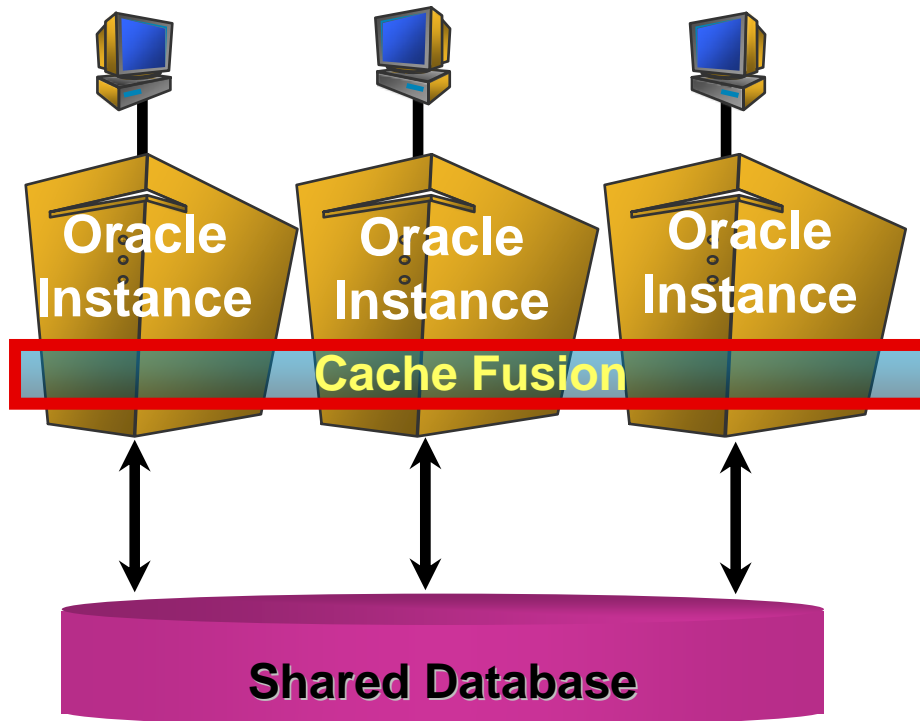
Hypervisor Virtualization

Adding Support for VMWare and Xen

Oracle RAC



Oracle Database 10g RAC

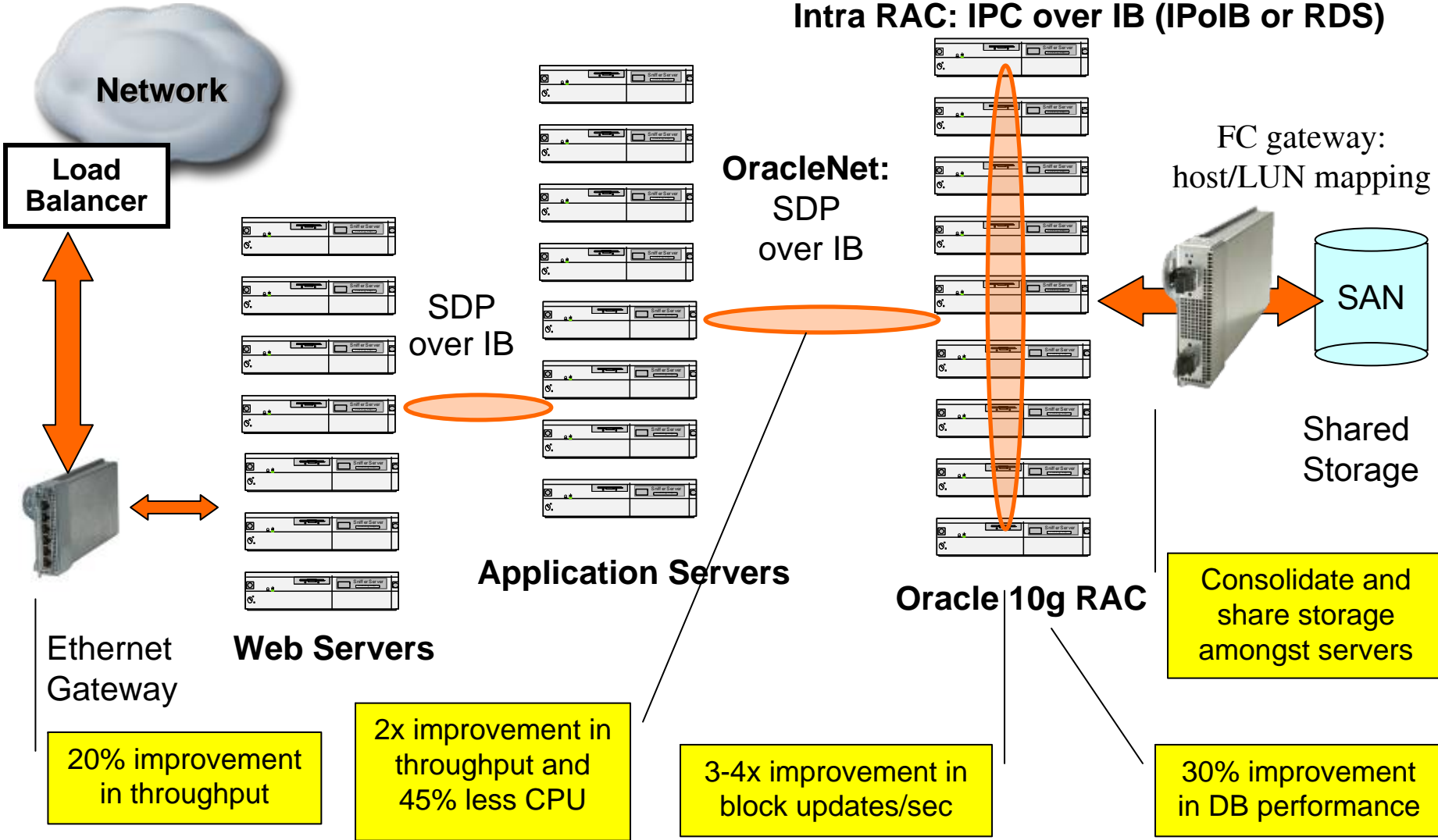


Operating Systems:

AIX, HP-UX, RHEL/SLES Linux, Solaris

- Oracle Database 10g Real Application Clusters (RAC) provides the ability to build an database platform from multiple clustered servers
- Highly Scalable
 - Cache-to-cache data shipping
 - Scales off-the-shelf applications with no changes
 - Easily add and delete nodes
- Highly Available
 - Eliminates a node as single point of failure
 - Node failure is transparent to applications
 - Recovers from node failure in 17 seconds - workload independent
 - Pre-warmed cache speeds restart

Oracle 10g RAC and Infiniband

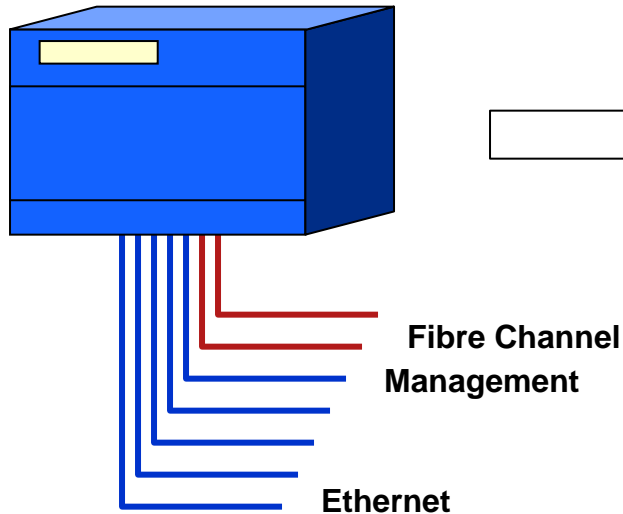


I/O Consolidation (Multi Fabric I/O)



Multi Fabric I/O (I/O Virtualization)

Traditional Datacenter I/O



Ethernet (2-5 connections)

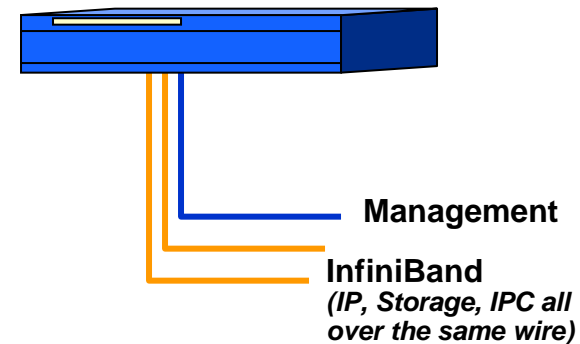
- Web/Client
- Backup/Restore
- Management
- DMZ
- Database

Fibre Channel (0-4connections)

- up to four for I/O intensive apps like Oracle

Proprietary cluster interconnects

Unified Fabric I/O



Two InfiniBand connections

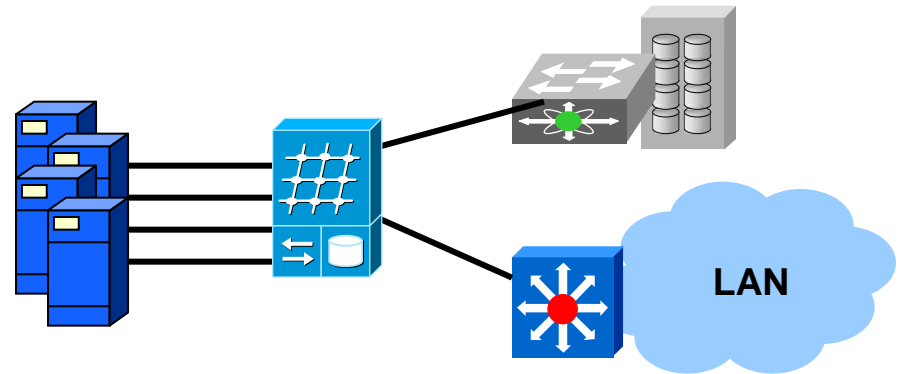
- IP, FC and IPC traffic all run over one or two 10Gbps IB pipes
- Data Center shares IP and FC access through central I/O gateways

One Ethernet for management

Physical vs. Logical View

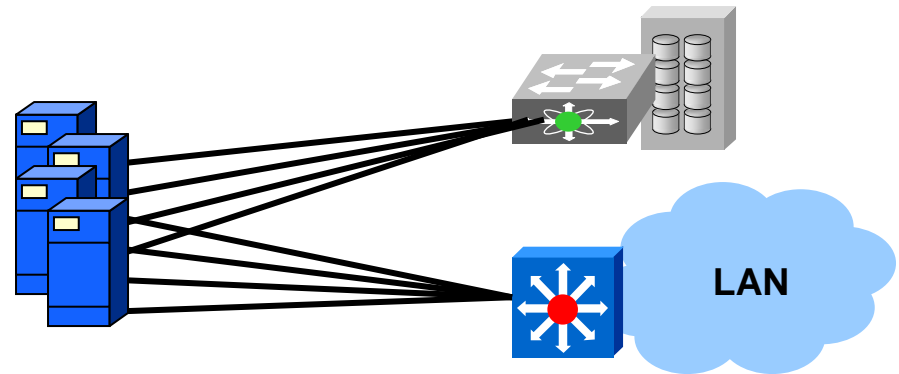
Physical View

- Servers connected via IB
- SAN attached via public AL
- Ethernet attached via Gig Etherchannel

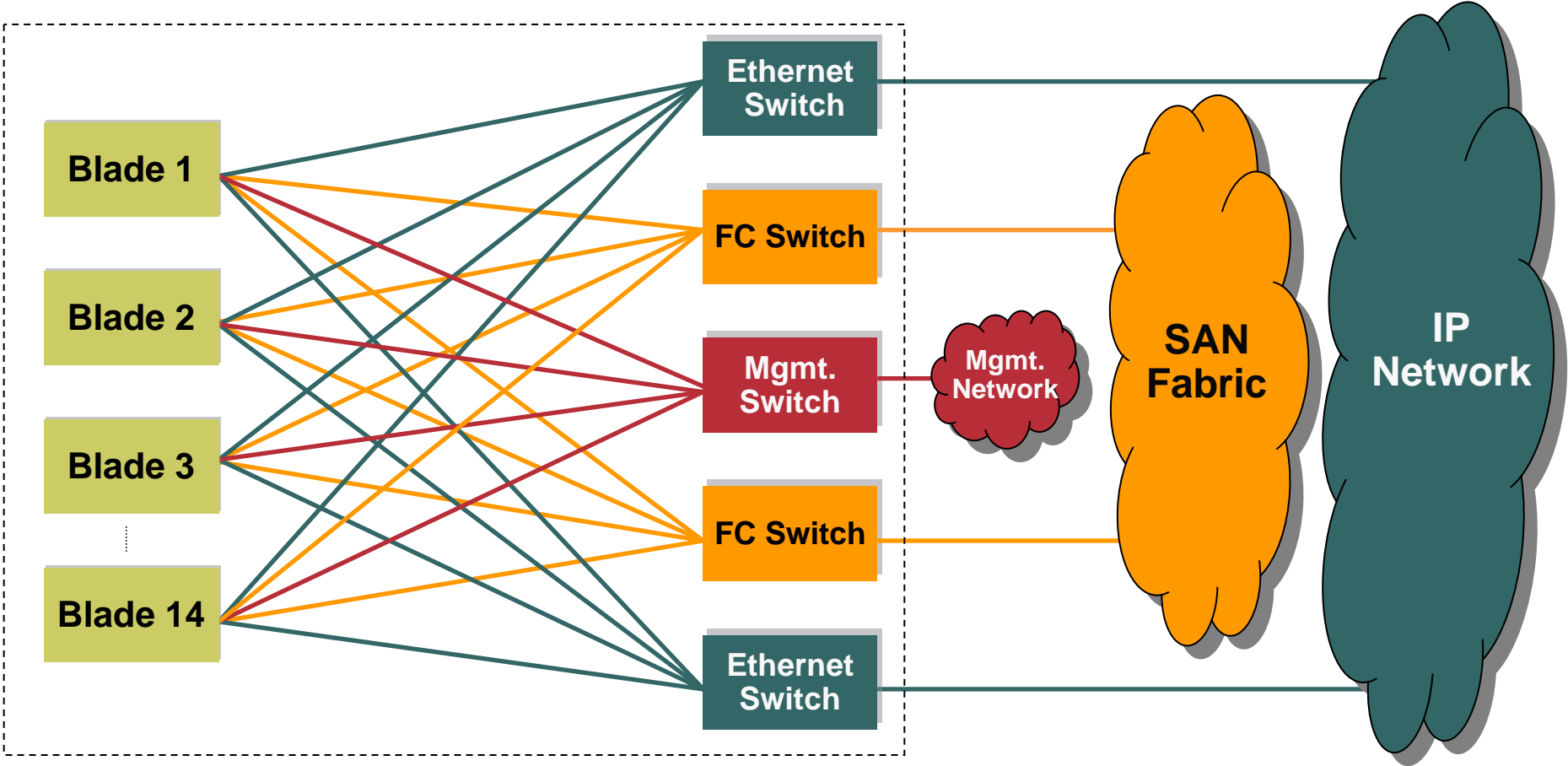


Logical View

- Hosts present WWNN on SAN
- Hosts present IP address on VLAN

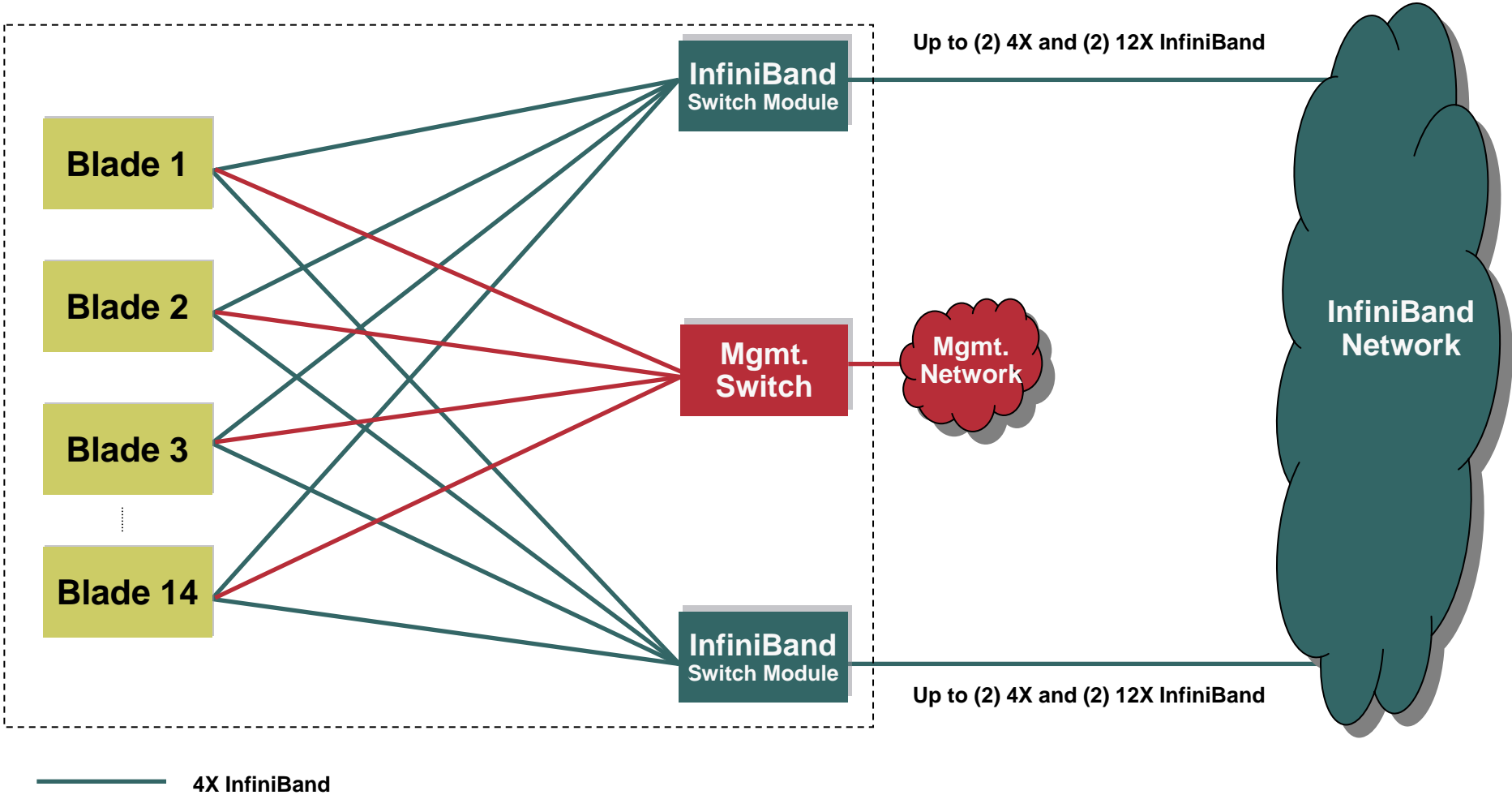


Typical Blade Switch Topology

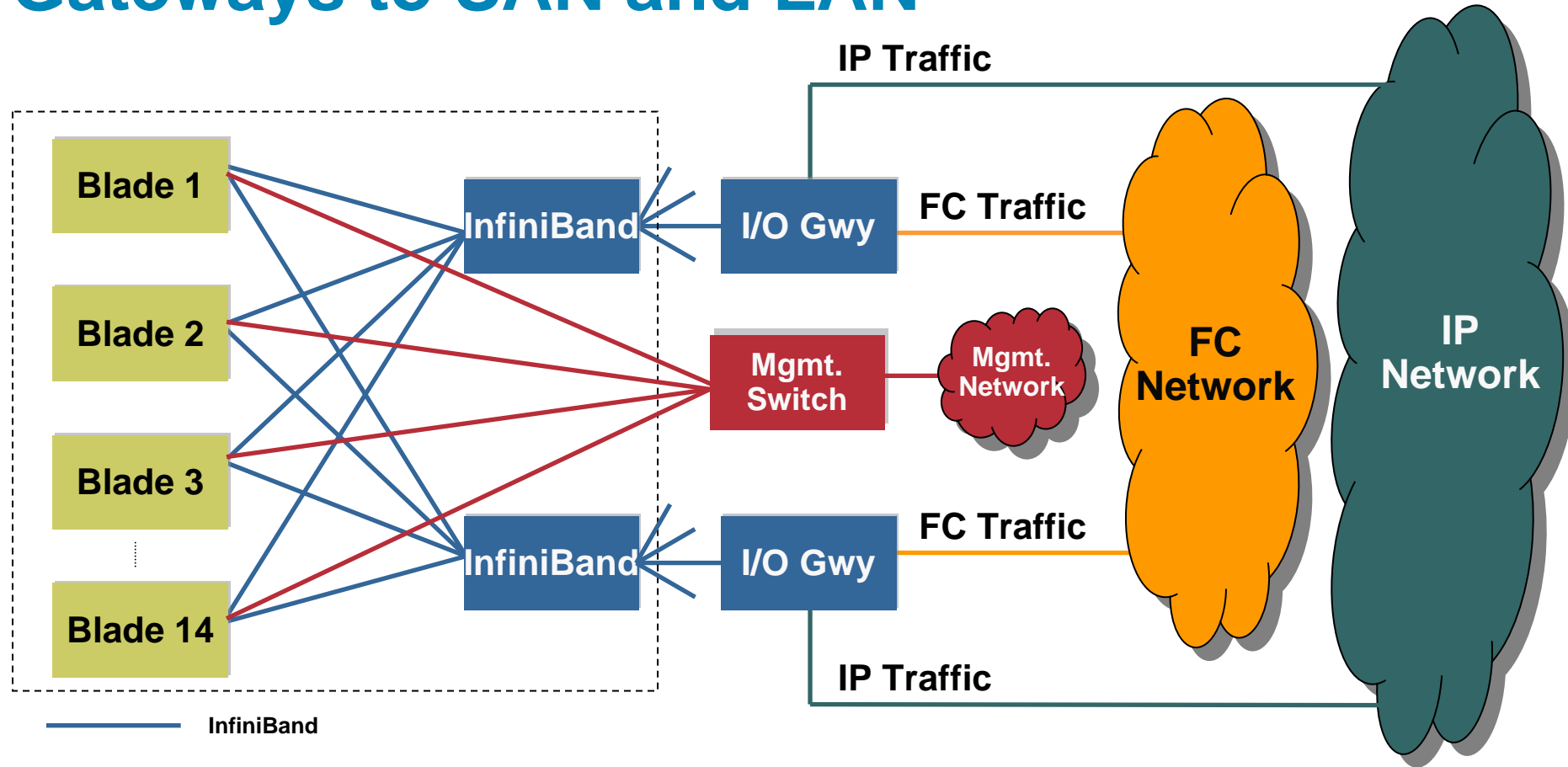


- GigE / 100Mb Ethernet
- 4 Gb Fiber Channel

I/O Consolidation with InfiniBand

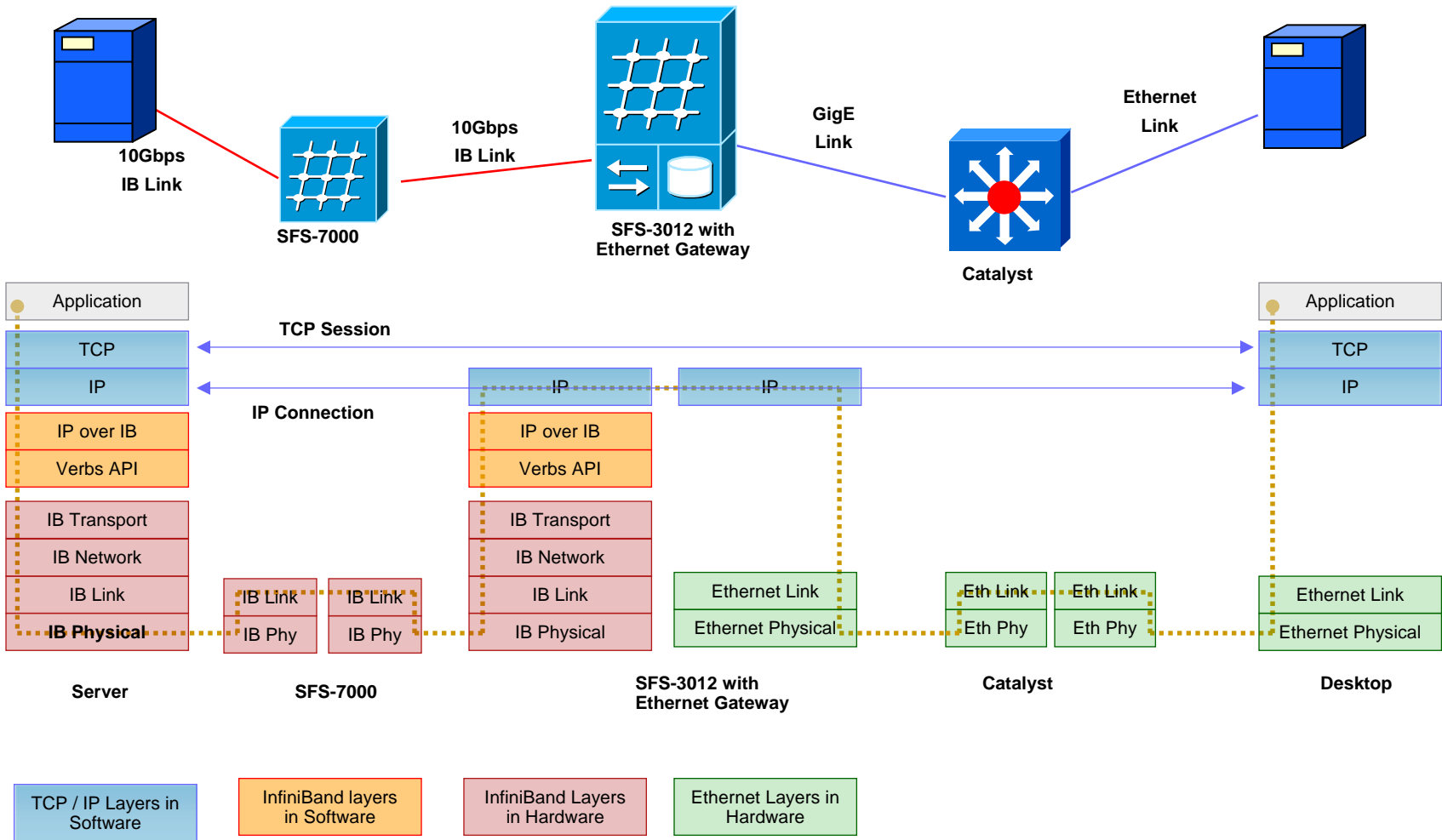


I/O Consolidation with Infiniband and Gateways to SAN and LAN

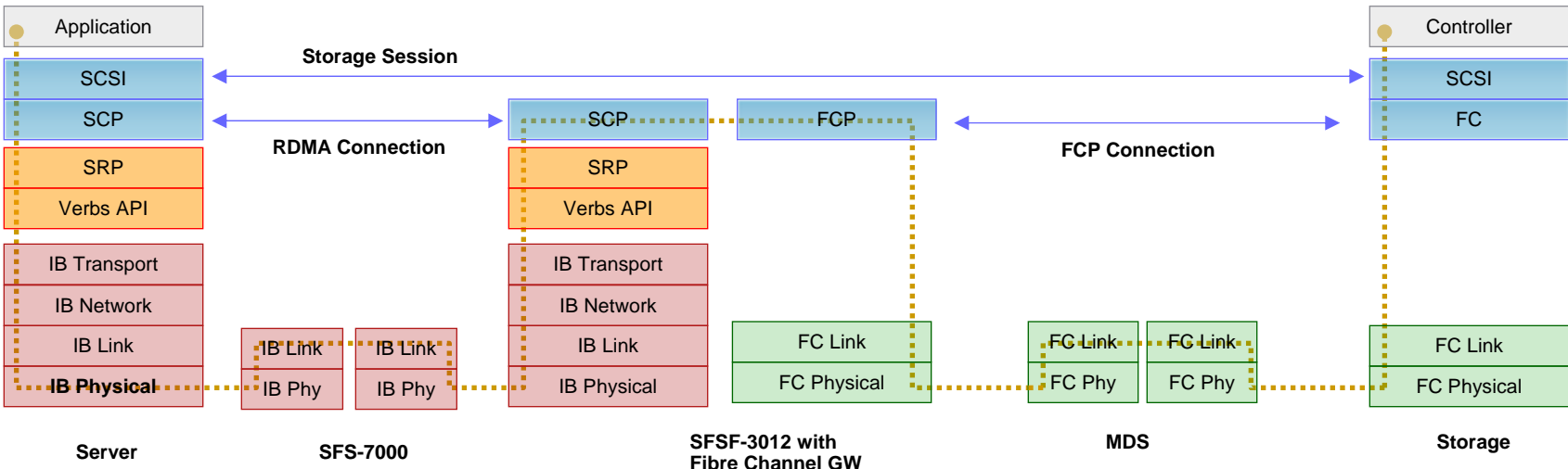
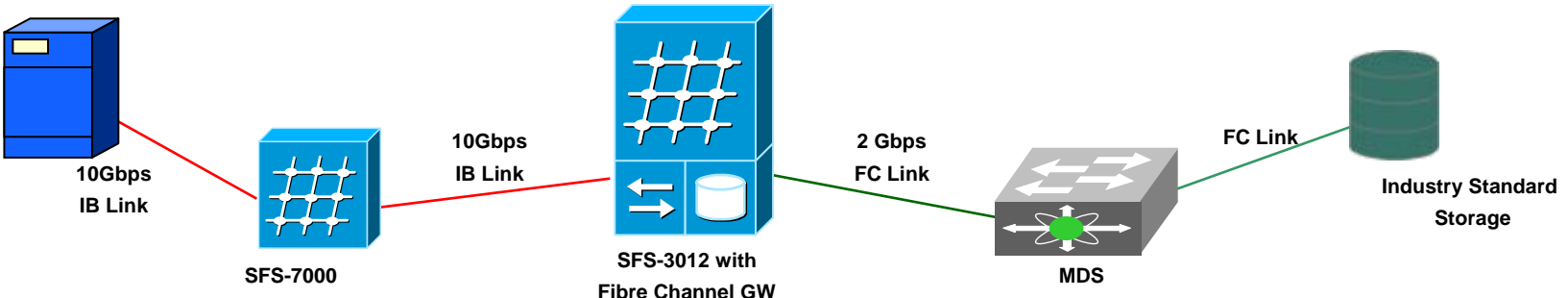


- RDMA Enabled
- Higher interconnect speeds
- I/O Consolidation
- Less port complexity in Blade chassis

IPoIB Gateway (Ethernet)



SRP Gateway (Fibrechannel)



Native InfiniBand Storage

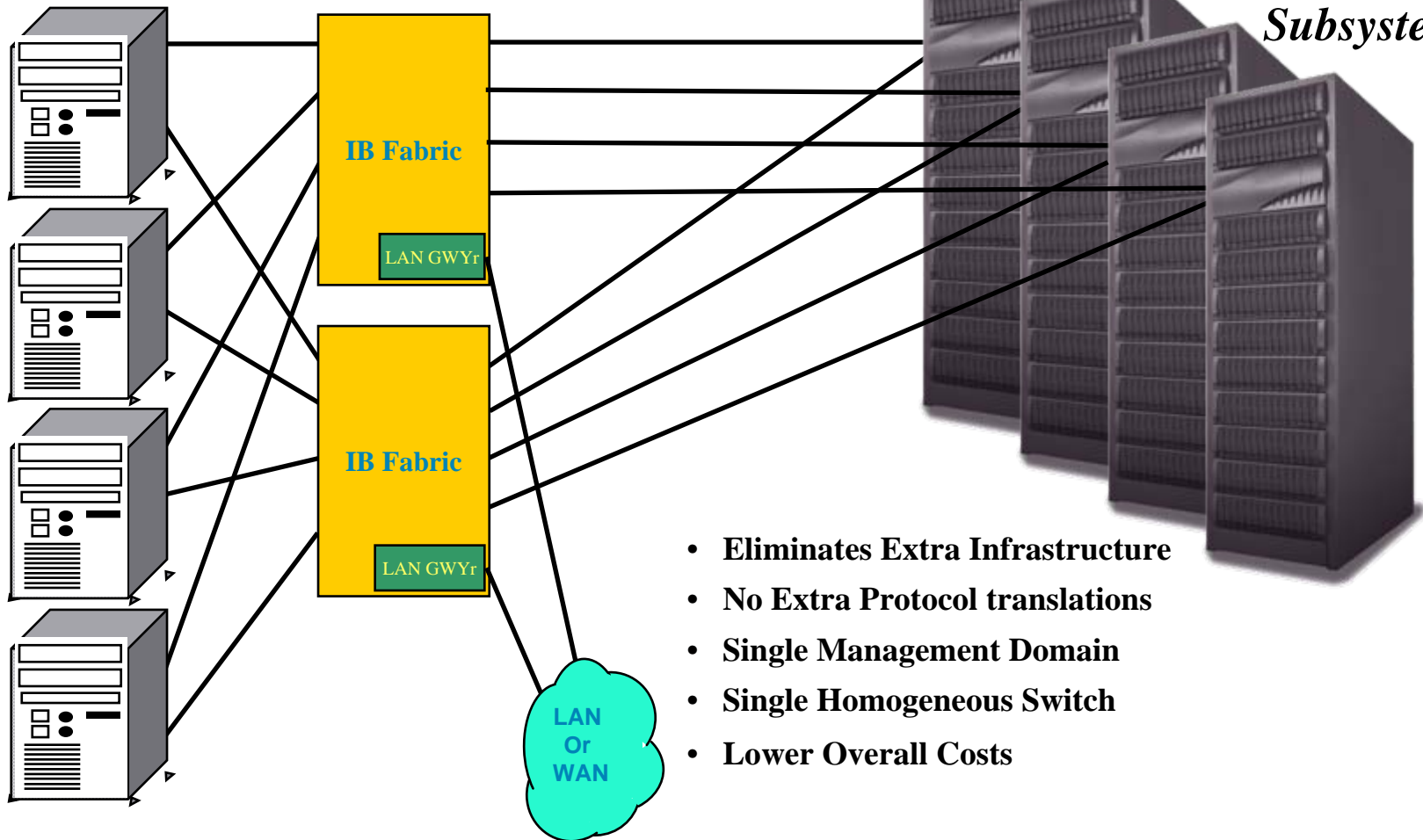


Native SCSI over InfiniBand

*InfiniBand
Native
Storage
Subsystems*

Servers

IB Fabric

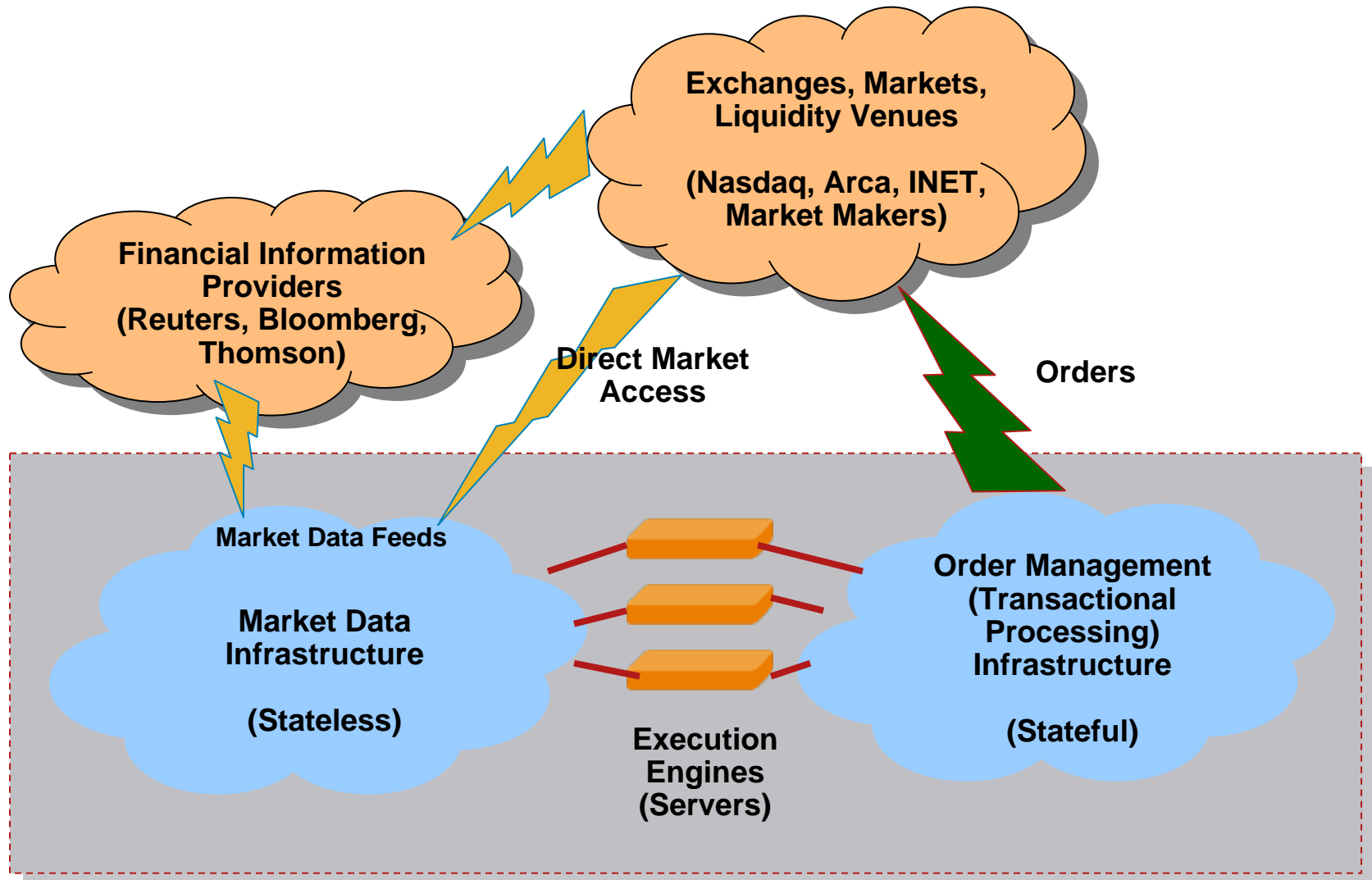


- **Eliminates Extra Infrastructure**
- **No Extra Protocol translations**
- **Single Management Domain**
- **Single Homogeneous Switch**
- **Lower Overall Costs**

High End Low Latency Server Farms

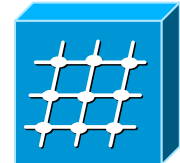
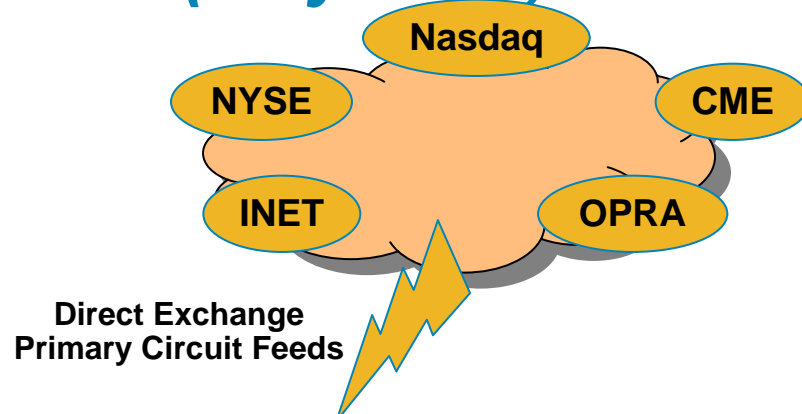


Automated Trading - High Level Architecture (Buy Side)



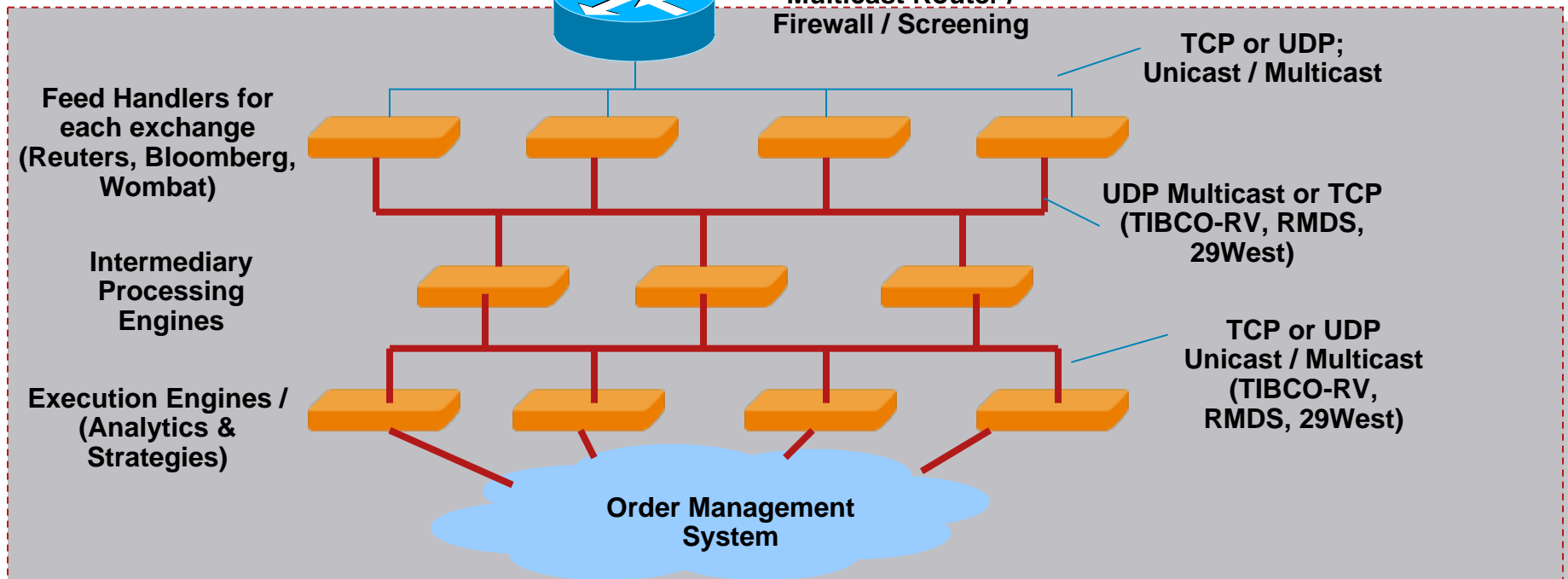
Boundary of the Firm

Market Data Infrastructure Architecture (Buy Side)

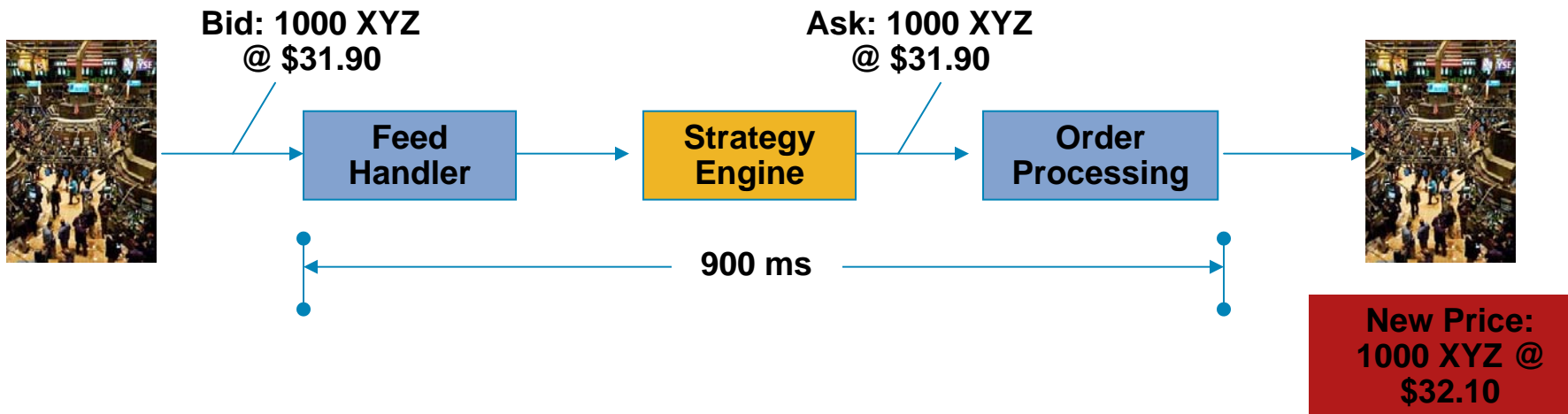
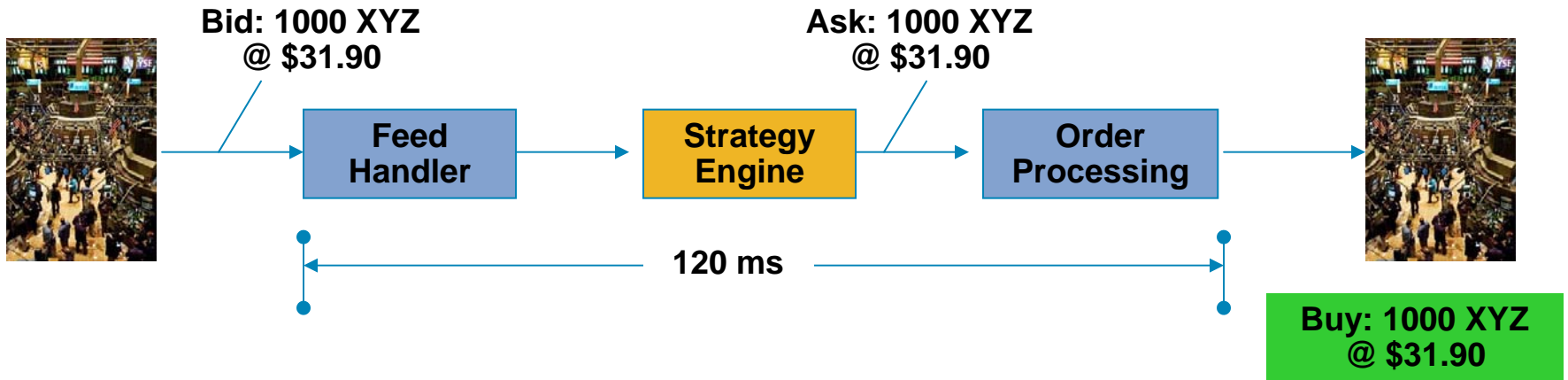


IB Fabric supports multicast in HW; SDP & DAL provide RDMA performance for applications using TCP and UDP

Servers

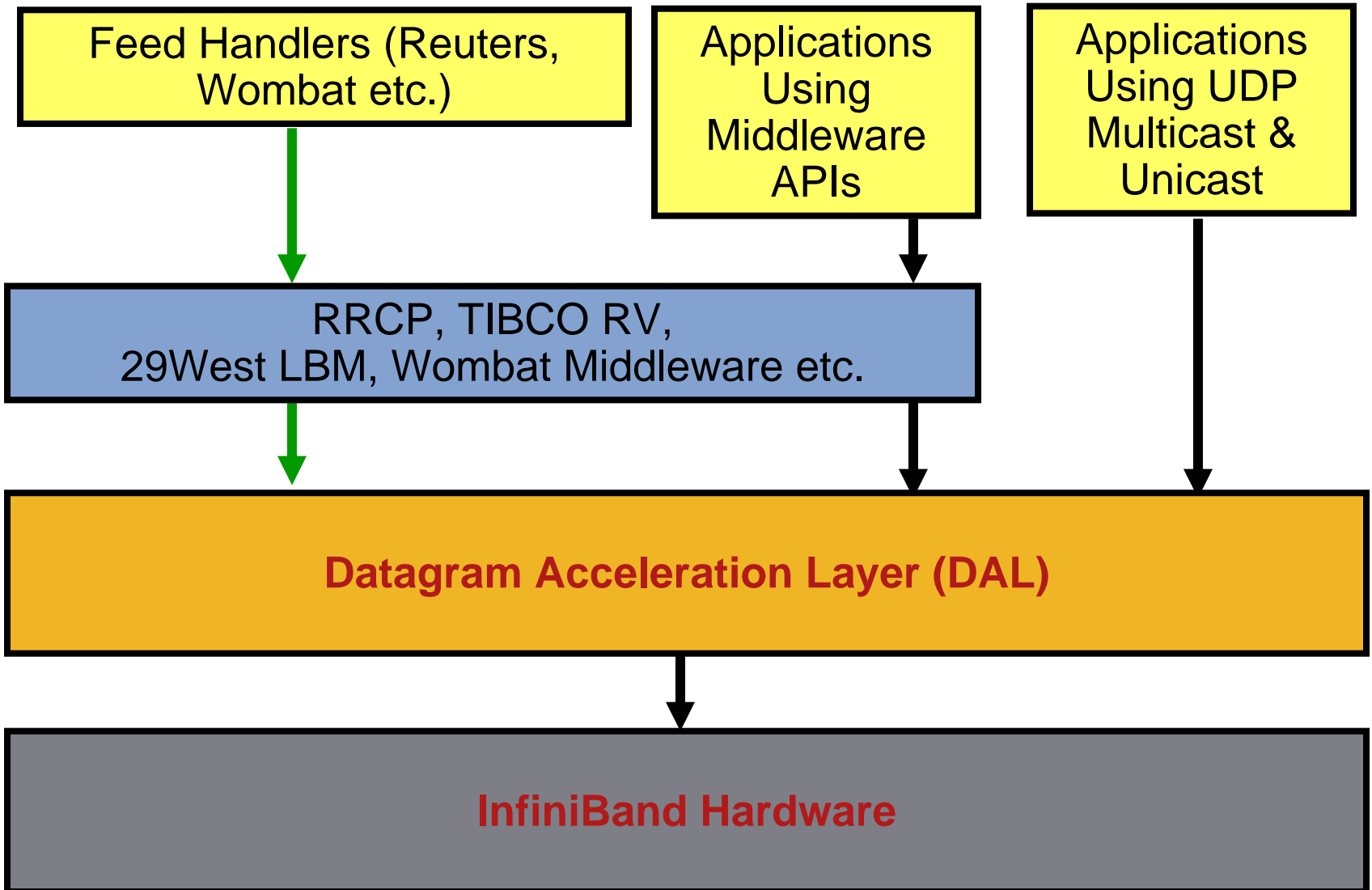


Impact of Latency on Proprietary Trades Desk



Source: TradingMetrics, Inc. <http://www.tradingmetrics.com/>

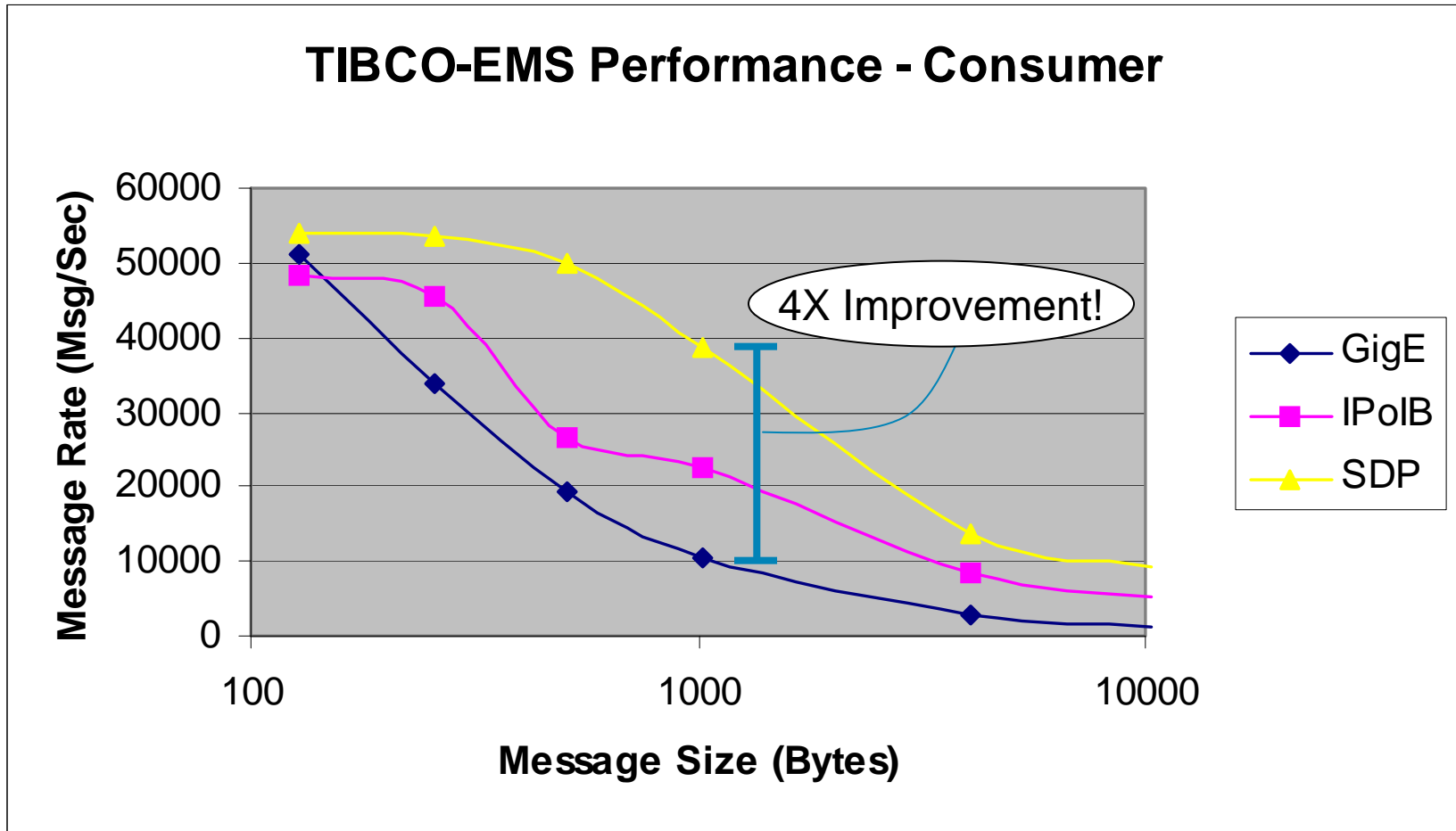
DAL Positioning



Example: TIBCO-EMS Installations

- Cisco SFS offers the 'SDP' protocol to accelerate applications using TCP
 - Offer's reliable in-order delivery for application using TCP sockets
- By utilizing the benefits of RDMA in InfiniBand technology, SDP can offer the following benefits
 - High Throughput – From IB rates of 10Gbps and 20Gbps
 - Low Latency – From ZCopy, fragmentation in hardware, eliminating context switches, async-IO
 - Low CPU Utilization – By eliminating TCP stack traversal
 - High Message Rate – By implementing layers 2-4 in hardware

TIBCO-EMS Benchmarks (Consumer)



Cisco InfiniBand Products



Cisco InfiniBand Hardware



M Bladecenter H
InfiniBand 4X
Switch & HCA



Cisco High
Performance
Subnet Manager

SFS-3504



CiscoWorks - LAN
Management System



SFS-3012
Gateways

SFS-3012P



Device Fault
Manager



SFS-3012

SFS-7024P



SFS-7012P

Resource Manager
Essentials



4X DDR
PCI-Ex HCA

SFS-7000D
Family

Cisco PCI-X & PCI-Ex
2*4X HCA



SFS-7008P



IBM Bladecenter
InfiniBand 1X
Switch & HCA

SFS-7000P



Cisco PCI-X & PCI-EX
1*4X HCA



SFS 3504: IB to Fibrechannel and Ethernet Gateway

Feature	Benefit
4U/4S chassis with 60Gbps/Slot	Future ready: Upto 60Gbps per gateway slot.
Integrated Switch and DDR IB Port Card	Cost optimized architecture. Single chip architecture.
Cisco VSAN Support on FC Gateway	Connectivity to multiple FC Sub Fabrics.
IPV6 ready Ethernet gateway	Compliant with Federal Requirements
10GbE Ethernet Gateway	Within two quarters of FCS
Availability	Q4CY07



SFS-3504

- ✓ E-Port Interoperable with leading FC Switches
- ✓ 10GbE Link Speed Capable
- ✓ DDR speeds for external and internal IB Ports

SFS 3504 I/O Gateways



- **Six 100/1000 Ethernet ports per module, up to 4 gateways per chassis.**
- **Fully compliant Cisco Ethernet Feature set**
- **High-availability within chassis / across chassis**

11M Packets Per Second



- **4 x 4-Gbps Fibre Channel ports; upto 4 gateways per chassis.**
- **E-Port support**
- **Cisco VSAN Support**
- **Multi-pathing support**
- **Port load-balancing capabilities**

400K I/O Per Second

InfiniBand – Current Blade Switch Offerings



- **InfiniBand High Speed Switch Module – 10Gig SDR**
- **InfiniBand 10Gbps SDR Daughter Card**
- **Switch and Daughter Card are manufactured by Cisco available exclusively through IBM**



- **4x InfiniBand Passthru Module (10Gbps) - SDR**
- **InfiniBand 10Gbps SDR Daughter Card**
- **Manufactured by Cisco & available through Dell**



- **Unmanaged DDR Switch – manufactured by Mellanox**
- **DDR (20G) HCA Daughter Card manufactured by Mellanox**
- **Cisco validates and offers software package for HCA card**
- **Available as a Cisco HCA package from HP**

InfiniBand – Roadmap for Blade Chassis



- Cisco / IBM to support DDR Passthru from Voltaire & DDR Connect X HCAs from Mellanox
- Passthru is branded IBM
- Cisco switches validated with pass-thru module
- Cisco HCA driver package will be available from IBM branded as Cisco DDR HCA
- Drivers are only OFED – Linux
- Available in Oct 2007
- DDR IB Blade Switch from Cisco for Dell Noble Chassis
- Unmanaged switch (requires external subnet manager)
- 50% Blocking switch (16 blades / 8 uplinks)
- ConnectX HCA Daughter card from Mellanox
- Driver support by Cisco based on OFED
- Available in Jan 2008

