Advanced Understanding of InfiniBand Technology

BRKDCT-3009

Walter Dey

Cisco Networkers 2007

HOUSEKEEPING

- We value your feedback, don't forget to complete your online session evaluations after each session and complete the Overall Conference Evaluation which will be available online from Friday.
- Visit the World of Solutions on Level -01!
- Please remember this is a 'No Smoking' venue!
- Please switch off your mobile phones!
- Please remember to wear your badge at all times including the Party!
- Do you have a question? Feel free to ask them during the Q&A section or write your question on the Question form given to you and hand it to the Room Monitor when you see them holding up the Q&A sign.

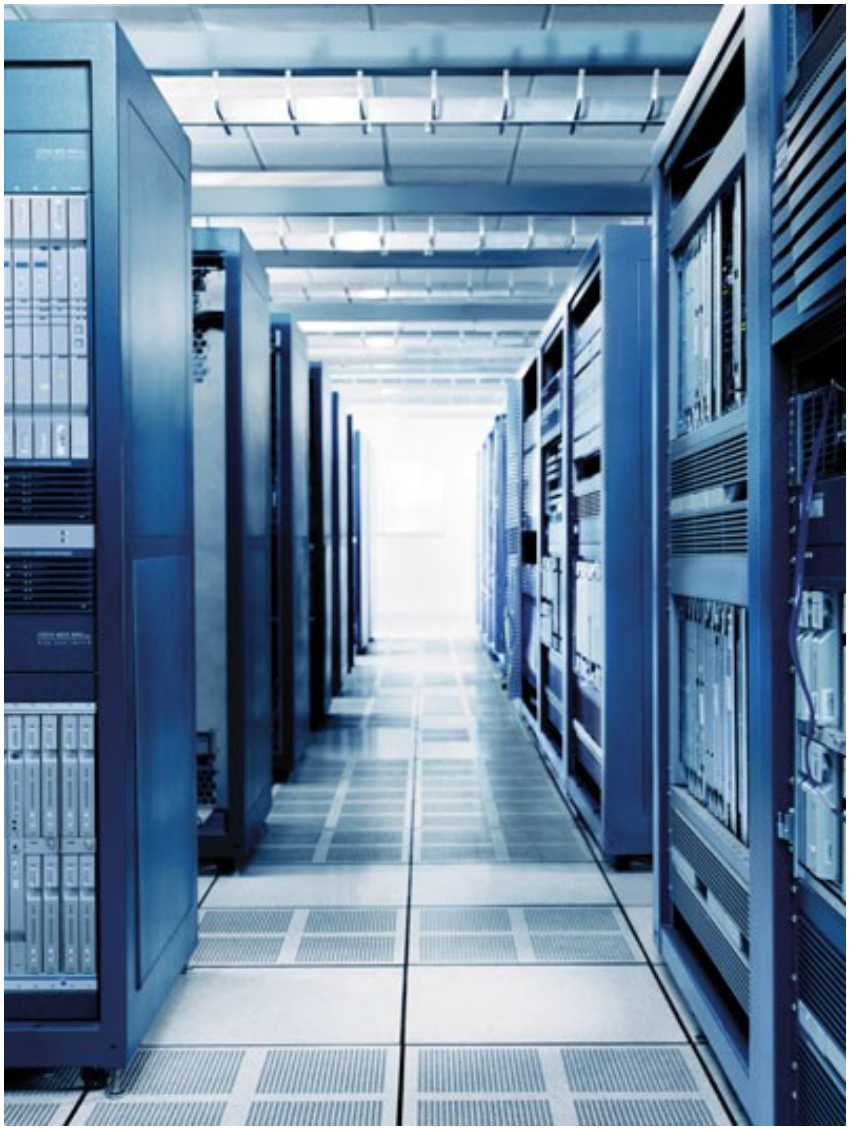
Agenda

- InfiniBand System Overview
- RDMA and Upper Layer Protocols
- InfiniBand Hardware Overview
- InfiniBand Applications
  - High Performance Computing (HPC)
  - I/O Virtualization
  - Case Studies

InfiniBand System Overview

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**



InfiniBand Architecture (IBA)

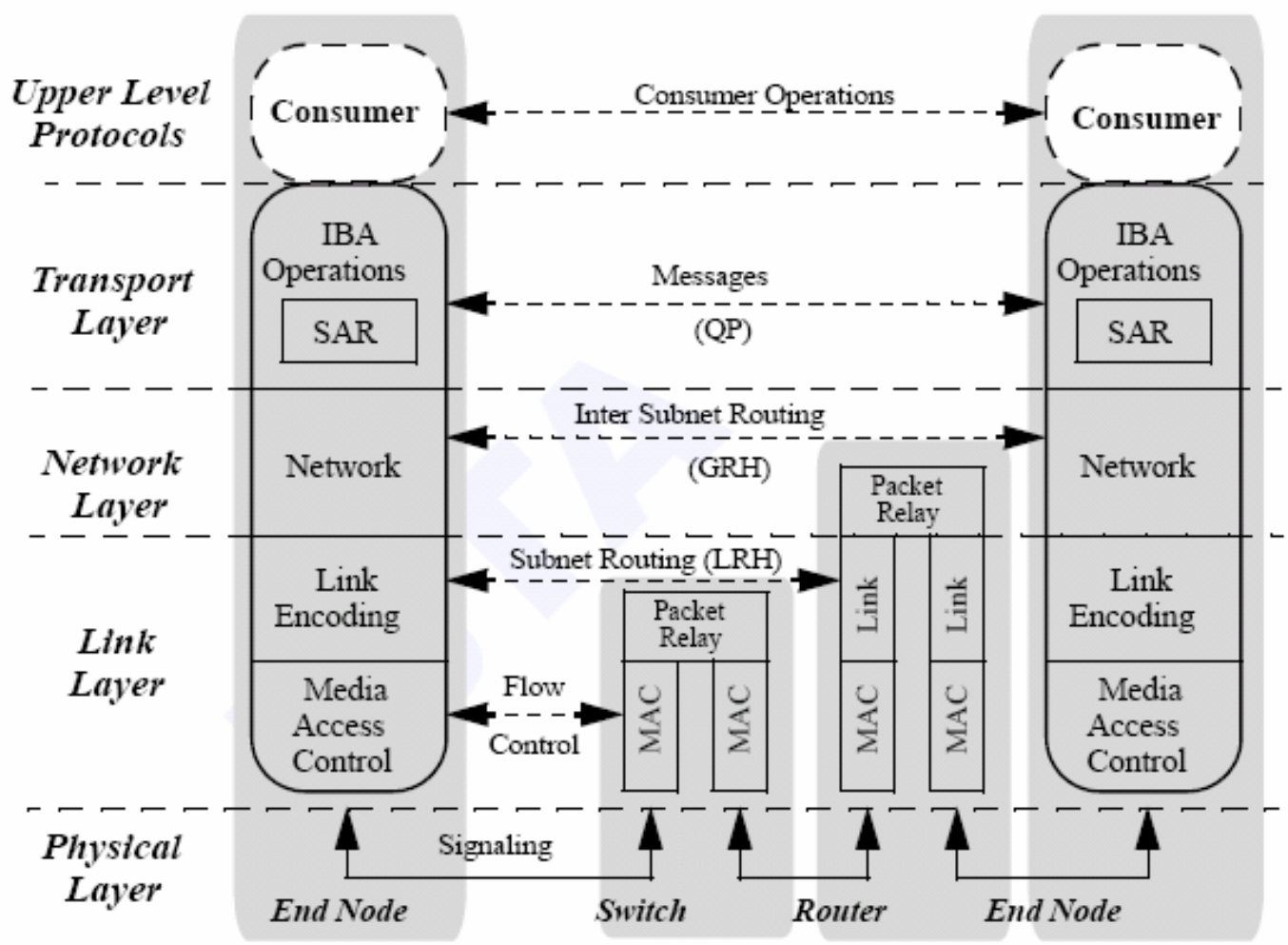
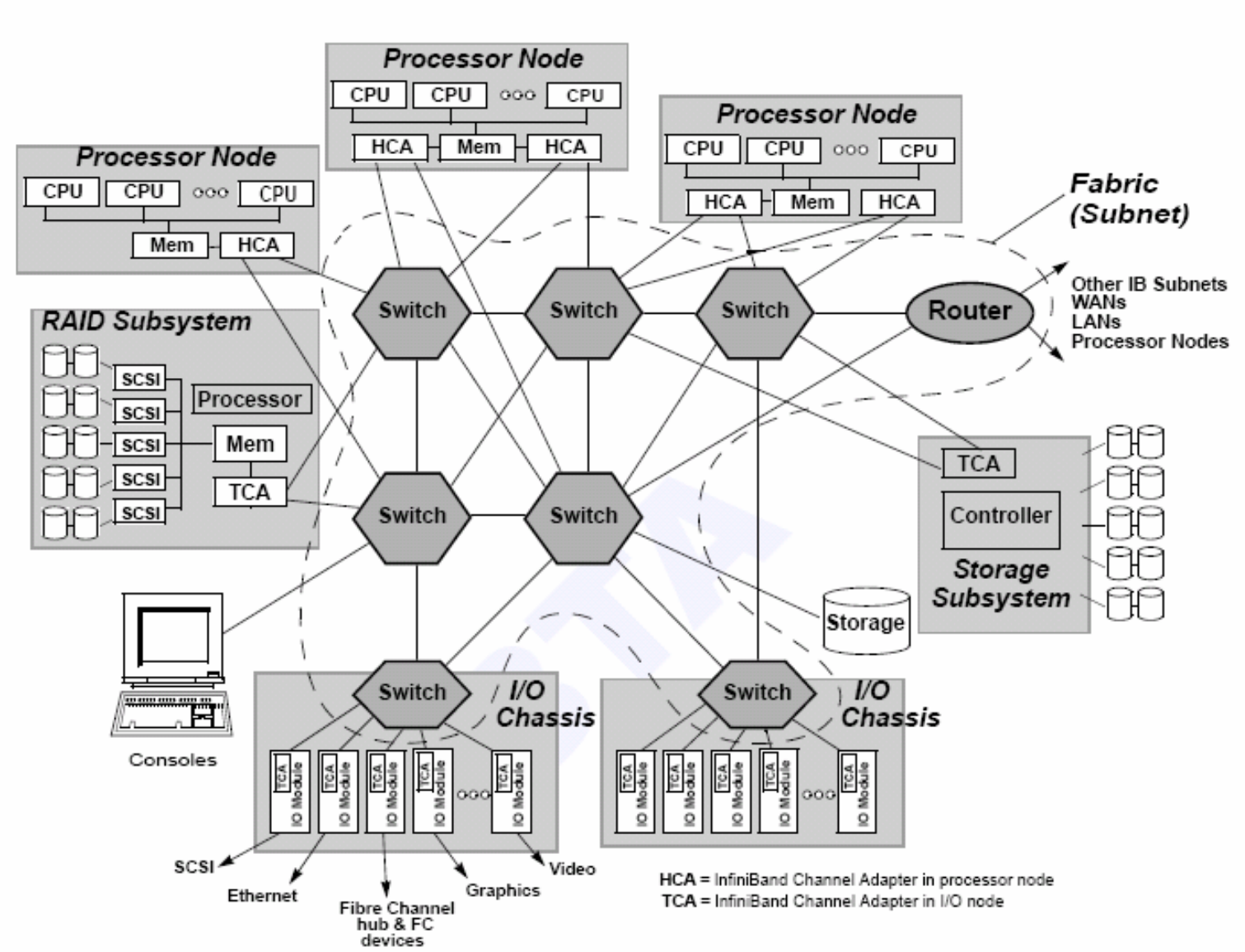


Figure 25 IBA Architecture Layers

InfiniBand System Area Network



Terms and Components

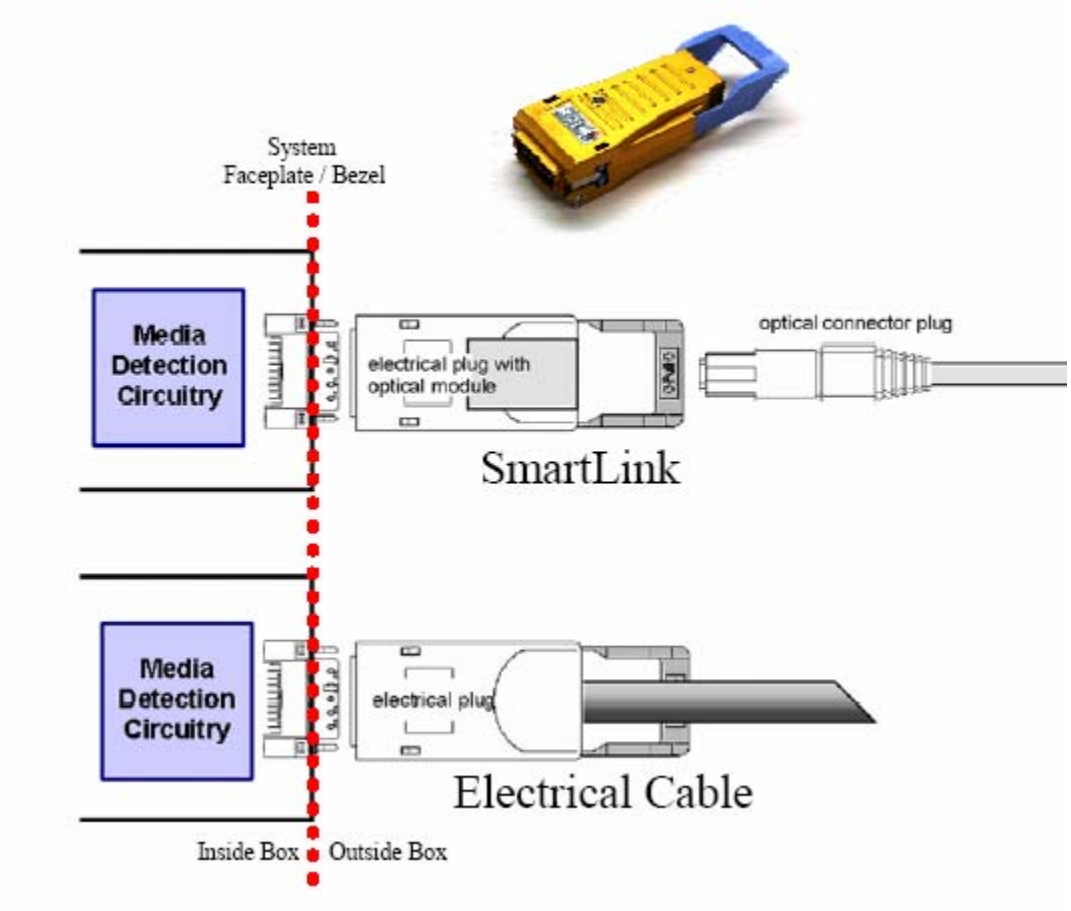
- InfiniBand Technology Terms:
  - Node: **Device, usually a server, which can access the IB fabric.**
  - Partition: **A group of IB ports that are allowed to communicate with one another. Ports can be members of multiple partitions. Ports can be either full members of the partition (allowed to talk to all other member ports) or limited members of the partition (allowed to talk only to full members of the partition).**
  - Path (Route): **The set of links, switch ports, and host ports a packet traverses from source to destination.**
  - Subnet: **Set of switches and channel adapters interconnected by links, and managed by a common Subnet Manager.**
  - Host Channel Adapter: **A device which terminates an IB link, and transports it to another media (e.g. PCI-X, PCI-e) and exposes its functionality to the Operating System or an application through a verbs software layer (VAPI).**

Terms and Components (continued)

- InfiniBand Technology Terms:
  - Subnet Manager: **Software which runs on the switch and manages & configures routes within the IB subnet fabric.**
  - Switch: **A device that routes packets from one link to another on the same subnet at line speed using a linear forwarding table.**
  - Verbs Application Programming Interface (VAPI): **A software layer which exposes the functionality of an InfiniBand channel adapter to an application or an Operating System (OS).**
  - HPC: **High-Performance Computing. Building a virtual super computer from a large cluster of commodity servers.**
  - Multi-Fabric Input/Output (MFIO): **The use of SFS products to interconnect InfiniBand and Ethernet and/or Fibre Channel networks. An InfiniBand deployment that includes at least one SFS I/O chassis and gateway module.**

InfiniBand Physical Layer

- Copper and Fibre interfaces are specified
- Copper (passive)
  - Up to 15m for 4x (SDR) connections
  - Up to 10m for 12x (SDR) connections
- Optical (active)
  - Availability via transceiver solution
  - Up to 150m
  - Long Haul possible, leverages DWDM infrastructure



InfiniBand Physical Layer (cont)

- Link is bonded 2.5 / 5 / 10Gbps (1x SDR/DDR/QDR) links

  Fiber is a ribbon cable

  Copper is a multi-conductor cable
- Each Link is 8b/10b encoded
- 4x (12x) Link is 4 (12) Physical Connections
- SAR provides a single data connection

Links	SDR	DDR	QDR
1X	2.5 Gb/s	5 Gb/s	10 Gb/s
4X	10	20	40
12X	30	60	120

InfiniBand System Architecture

- Connection Oriented Architecture
  - Central connection routing management (SM)
  - All communications based on send/receive queue pairs
- Two primary connection types
  - Reliable Connection
  - Unreliable Datagram
- Unused connection types
  - Unreliable Connection
  - Reliable Datagram
  - Raw Datagram

InfiniBand Connections

- Reliable Connection

  Host Channel Adapter based guaranteed delivery

  Uses HCA onboard memory (or system memory with PCI-E) for packet buffering

  Primarily used for RDMA communications

  Can use end-to-end flow control based on credits related to available receive buffers

- Unreliable Datagram

  Best effort forwarding

  Used for IP over IB communications

InfiniBand Subnet Manager (SM)

- All devices under the control of a single Master SM
- May have multiple slaves with replicated SM database state
- At system startup:
  - SM discovers topology – parallel sweep technology
  - All devices register with the SM
- Central Routing function
  - Shortest Path First Routing
  - Equal Paths Load Balanced with Static Round Robin distribution
  - Connection Failover

InfiniBand Subnet Manager (SM) cont.

- IB fabric communications are connection oriented
- Connections are managed by the Subnet Manager (SM)
- SM works in conjunction with the Subnet Administrative Agent (SA) on the host to provide end-to-end context information
- SM has total network view available at all times
- SM sets paths (routes) for all connections through the IB network.

InfiniBand Subnet Manager (SM) cont.

- Concept of Primary and Standby SM's

  Higher Priority Wins

  Election of lower GUID

  Non pre-emptive election process

- The IB spec does not specify how routing is setup in the fabric and every vendor has its own implementation.
- SM receives asynchronous notification of port or link failure, and re-establish connections and routes when possible

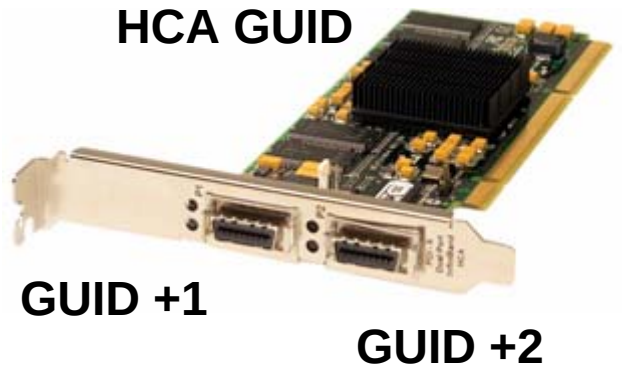
IB Addressing—Global Unique ID (GUID)

- Global Unique ID 64 bits in length
- Used to uniquely identify a port or port group
- HCA and each port has a GUID
- HCA GUID

  00:05:ad:00:00:01:02:03
- Port 1 GUID

  00:05:ad:00:00:01:02:04
- Port 2 GUID

  00:05:ad:00:00:01:02:05



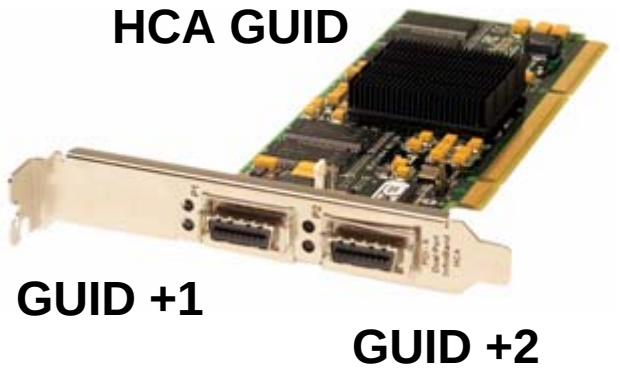
IB Addressing—Global ID (GID)

- Global ID 128 bits in length
- GUID plus Subnet prefix
- Used for host lookup on a subnet
- Used for inter-subnet IB routing (future)
- HCA and each port has a GID
- HCA GID

  fe:80:00:00:00:00:00:00:00:05:ad:00:00:01:02:03
- Port 1 GUID

  fe:80:00:00:00:00:00:00:00:05:ad:00:00:01:02:04
- Port 2 GUID

  fe:80:00:00:00:00:00:00:00:05:ad:00:00:01:02:05



IB Addressing—Local ID (LID)

- Local ID
- Assigned by SM to define a switchable endpoint in the network
- Subnet Local address
- Each port has a LID (assuming both are attached to the same IB Fabric)
- Port 1 LID

  00:04
- Port 2 LID

  00:05

IB Addressing—Queue Pair

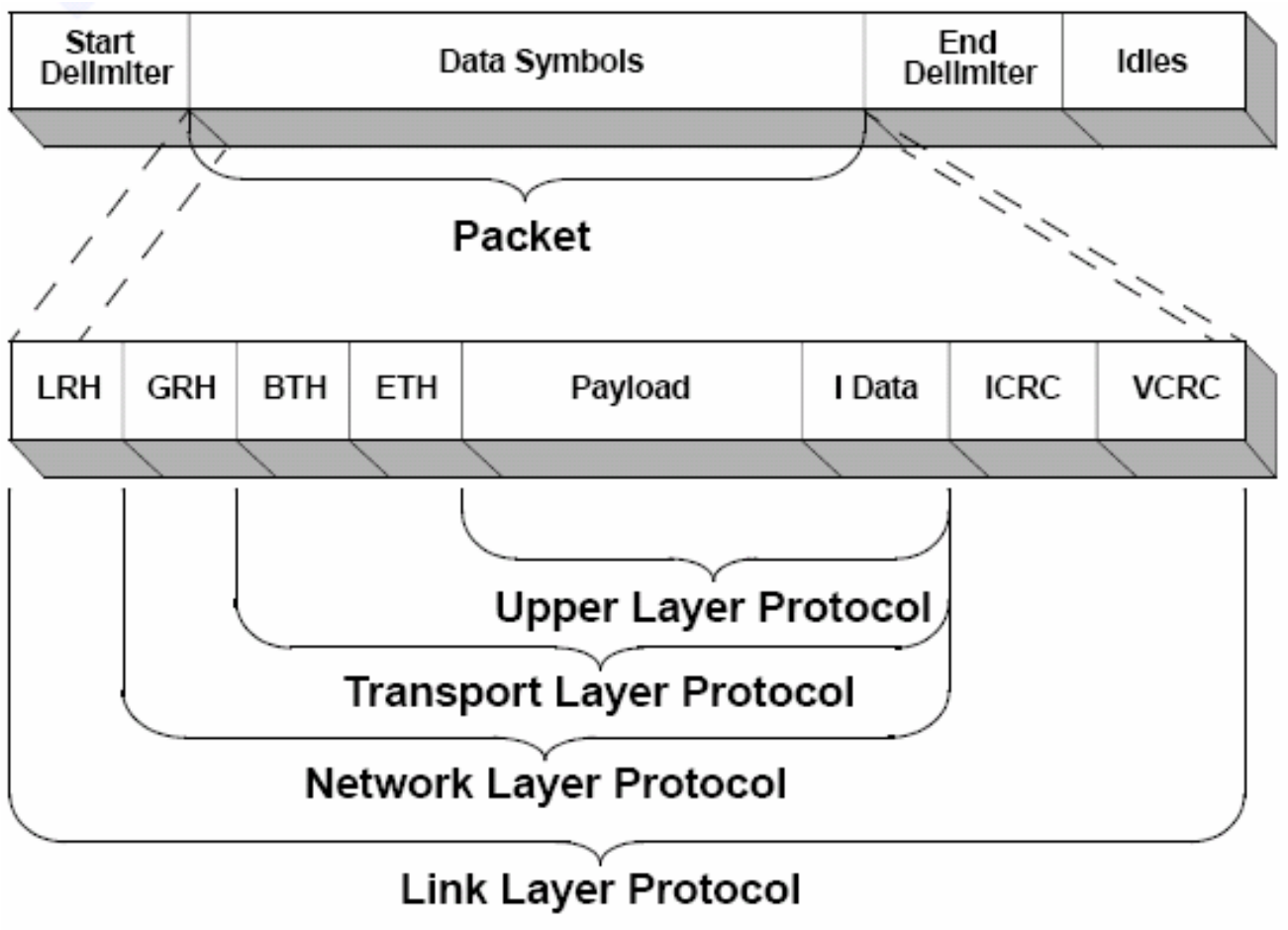
Queue Pair

- In conjunction with LID defines send/receive queues for End to End context
- Similar to a socket on an IP port
- Process address within the host

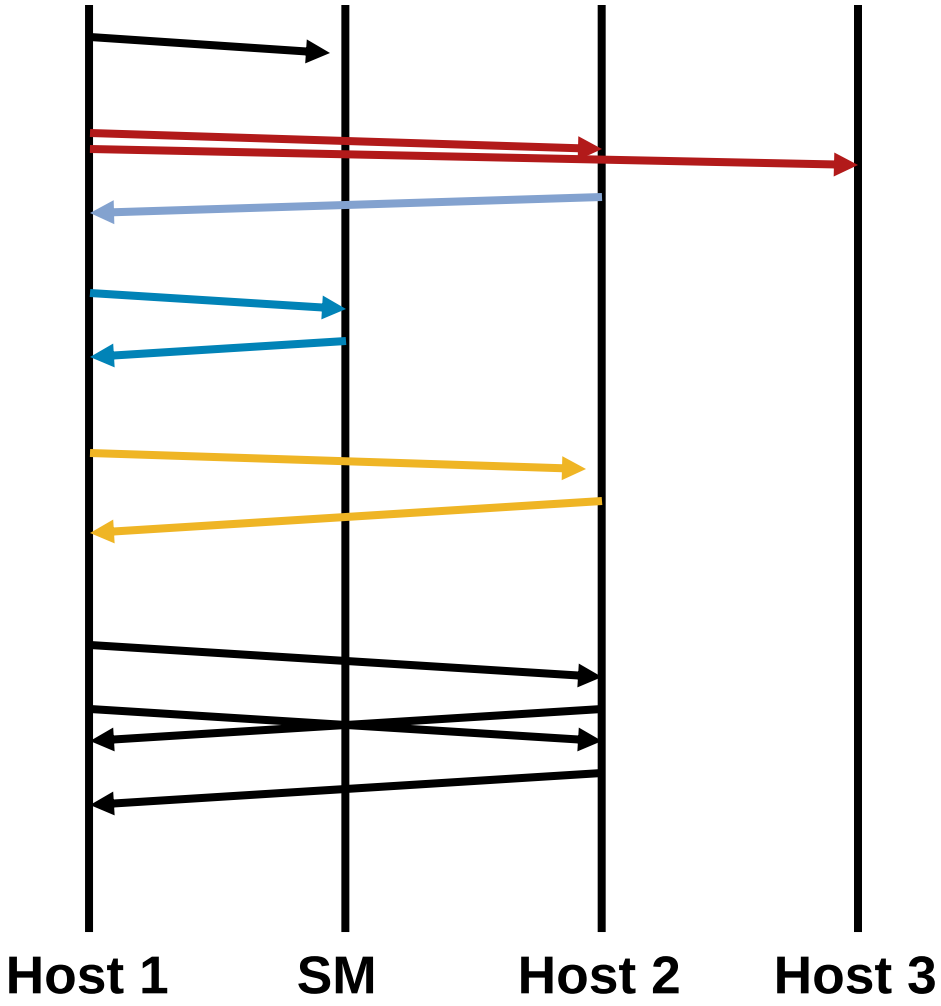
InfiniBand Routing

- Every packet has a LRH (Local Route Header) with a source and destination LID (assigned by the SM)
- LID's are 16 bit values assigned by the local subnet manager
- LID addressing is used for routing within a subnet
- Packets that traverse multiple subnets contain a global route header (GRH) that specifies the GID of the source and destination end node

IBA Data Packet Format



Address Resolution



Join Multicast Group

Send ARP on “Broadcast addr”

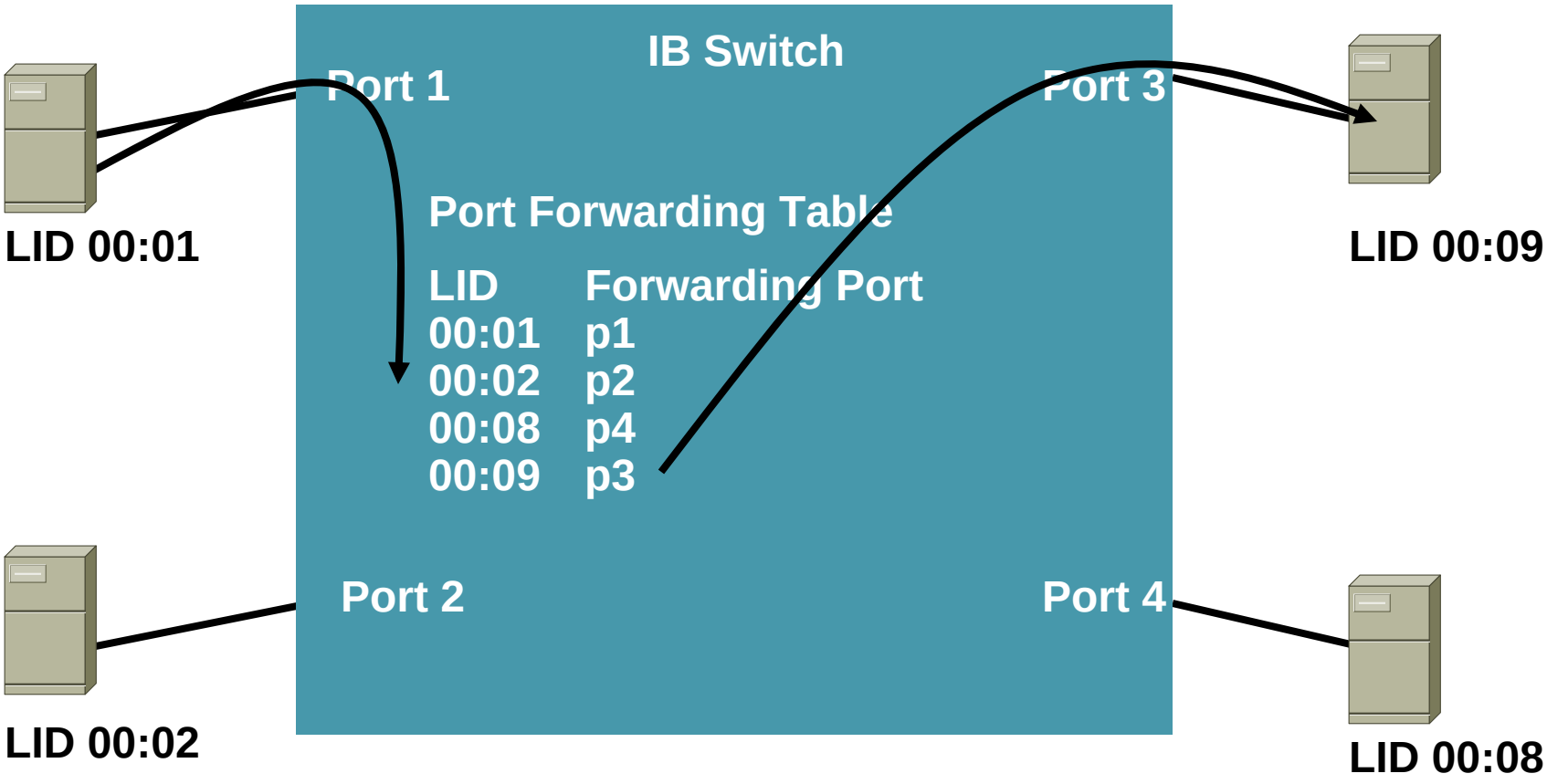
Receive remote GID via ARP

Ask SM for GID->LID mapping

Ask Host for Service info (QP)

Communicate

IB Local Communication



LID 00:01 talking to LID 00:09

InfiniBand and QoS

- IBA provides several mechanisms that permit a subnet manager to administer various quality of service guarantees for both connected and connectionless services.
- These mechanisms are

  Service Level

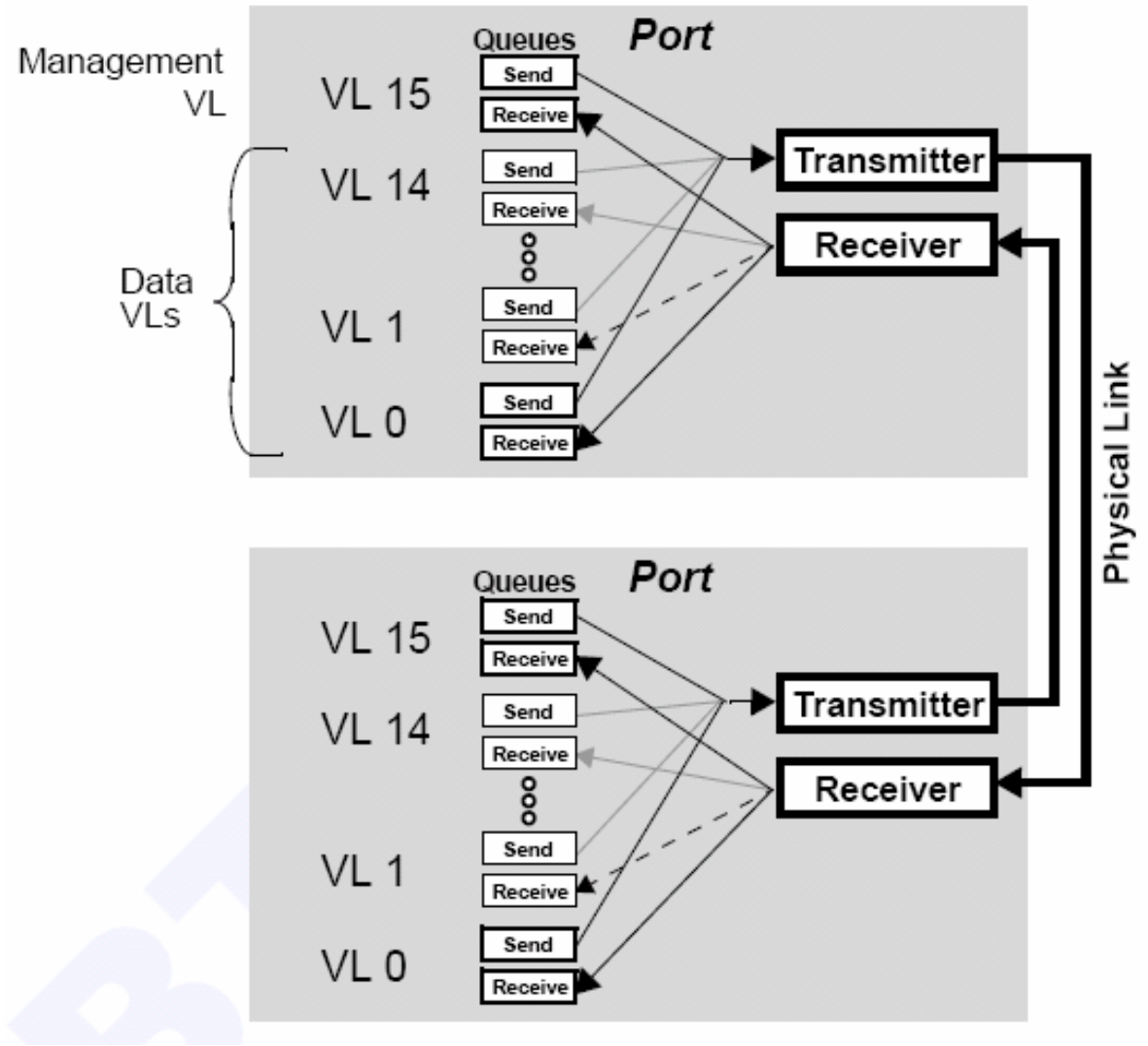
  Service Level to Virtual Lane Mapping

  Partitions

Virtual Lanes (VL)

- Virtual lanes (VL) provide a mechanism for creating multiple virtual links within a single physical link
- A virtual lane represents a set of transmit and receive buffers in a port
- All ports support VL15 which is reserved exclusively for subnet management.
- There are 15 other VLs (VL0 to VL14) called data VLs and all ports support at least one data VL (VL0)
- The default is to use VL0 until the SM determines the number of VLs that are supported by both ends of the link and programs the port's SL to VL mapping table.

Virtual Lanes (VL) cont.



Service Level (SL) and Virtual Lanes (VL)

- IBA defines a Service Level (SL) attribute
- IB packets operate at one of 16 service levels
- A packet's SL determines which VL is used on the next link
- Each port (switches, routers, endnodes) has a SL to VL mapping table that is configured by subnet management
- Packets addressed to QP0 are Subnet Management Packets (SMP) and exclusively use VL15 and their SL is ignored.
- VL15 (the management VL) is not a data VL

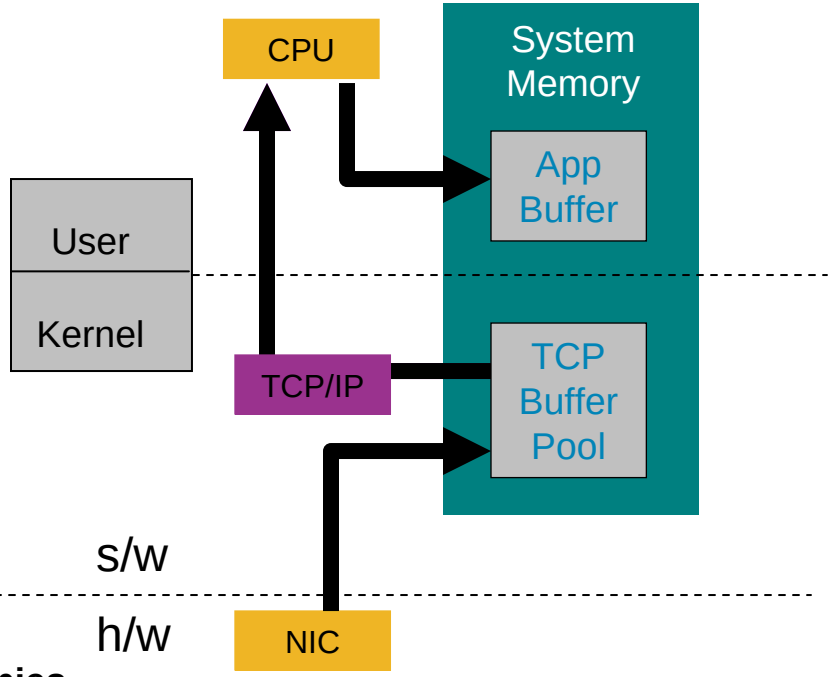
InfiniBand Partitions

- Partitioning enforces isolation among systems sharing an InfiniBand fabric
- Partitioning is not related to boundaries established by subnets, switches, or routers
- A partition describes a set of endnodes within the fabric that may communicate.
- Each port of an endnode is a member of at least one partition and may bea member of multiple partitions.
- A partition manager assigns partition keys (P_Keys) to each channel adapter port.
- Each P_Key represents a partition
- Each QP is assigned to a partition and uses that P_Key in all packets it sends and inspects the P_Key in all packets it receives
- Reception of an Invalid P_Key causes the packet to be discarded.

RDMA and Upper Layer Protocols

Server Overhead

Sources of Overhead in Server Networking	CPU Overhead
Transport Processing	40%
Intermediate Buffer Copying	20%
Application Context Switches	40%



Solutions for Overhead in Server Networking

Transport Offload Engine (TOE)

- Moves Transport processor cycles to the NIC
- Moves TCP/IP protocol stack buffer copies from system memory to the NIC memory

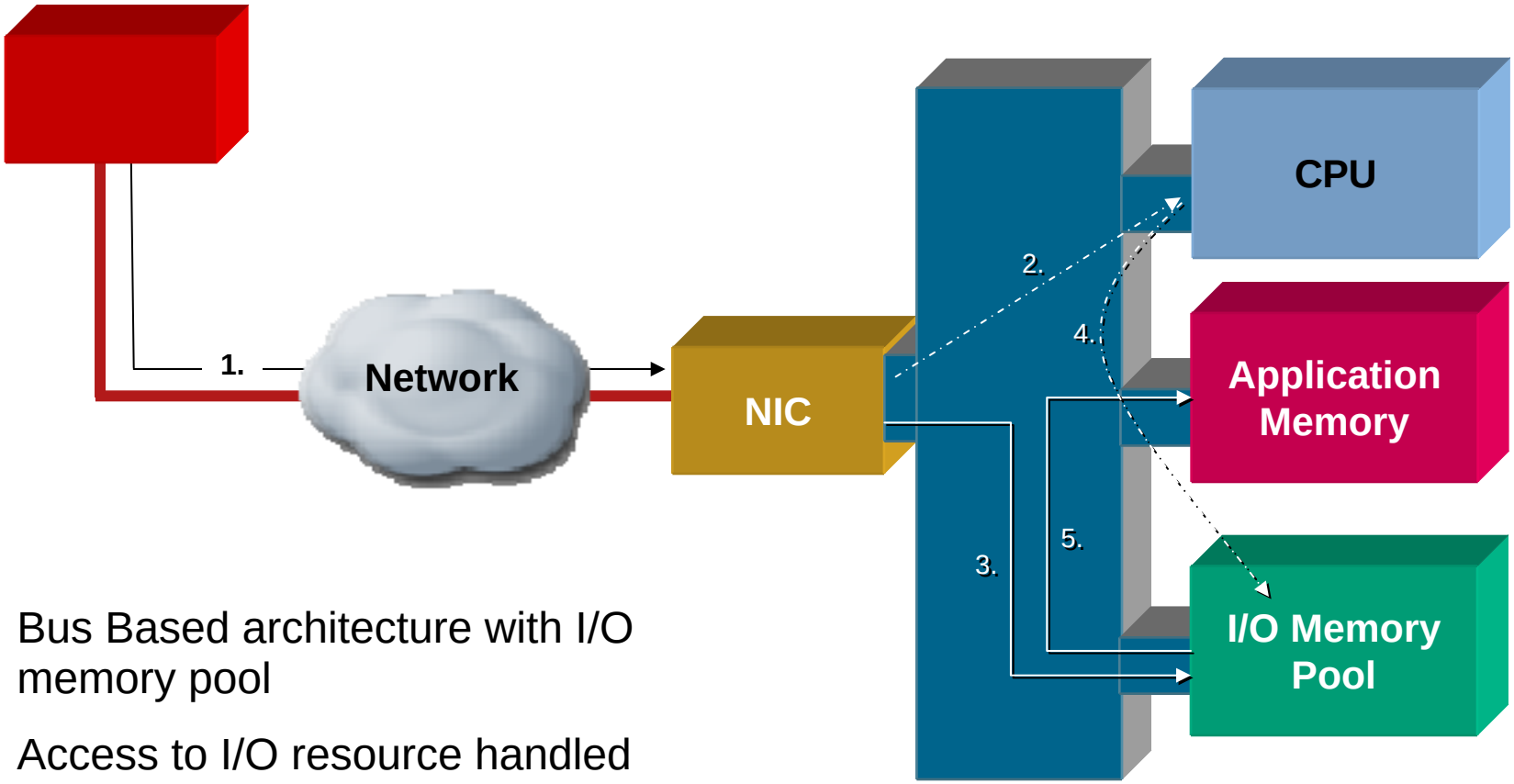
RDMA

- Eliminates intermediate and application buffer copies (memory bandwidth consumption)

Kernel Bypass – direct user-level access to hardware

- Dramatically reduces application context switches

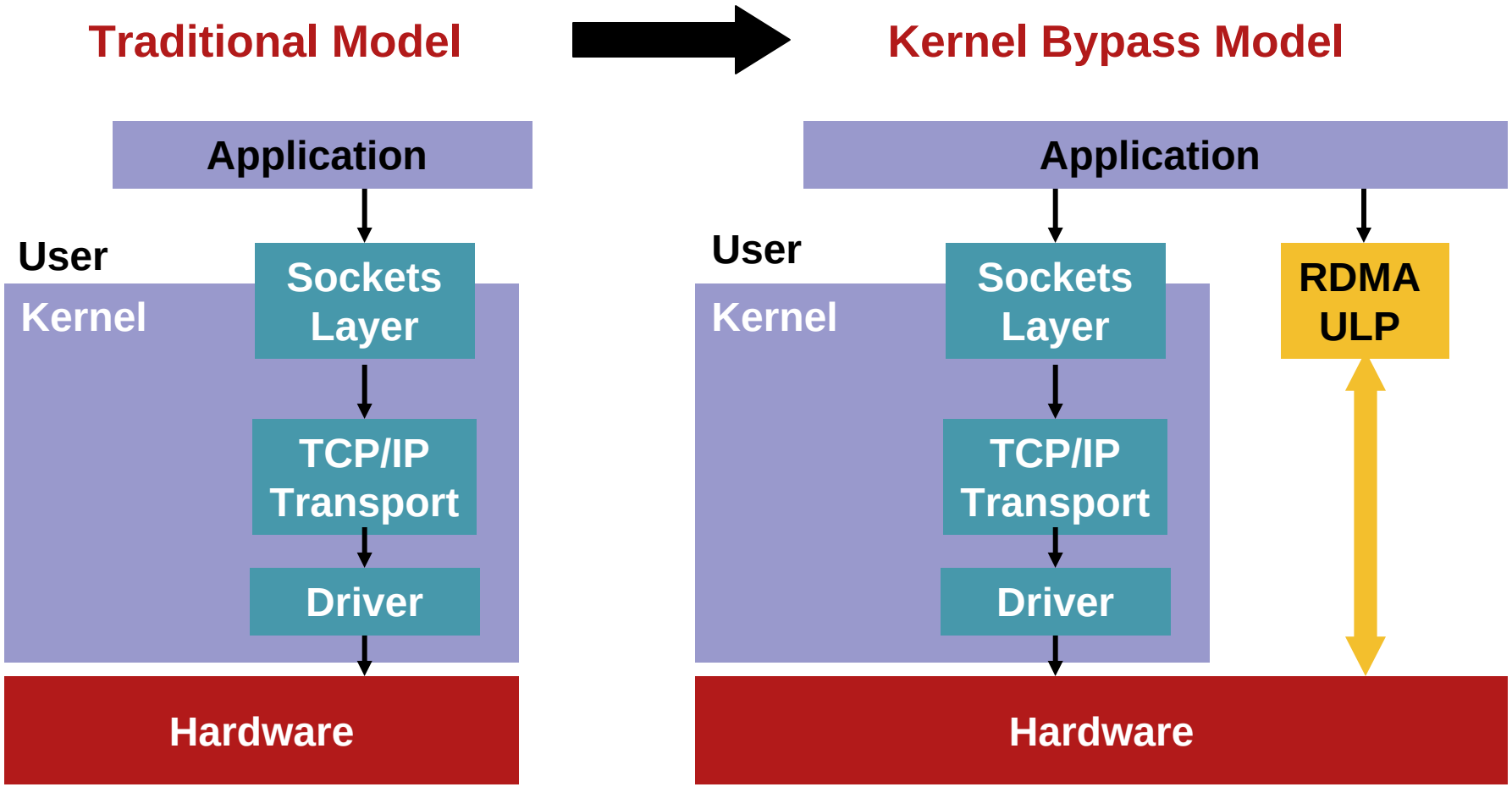
Traditional Server I/O architecture



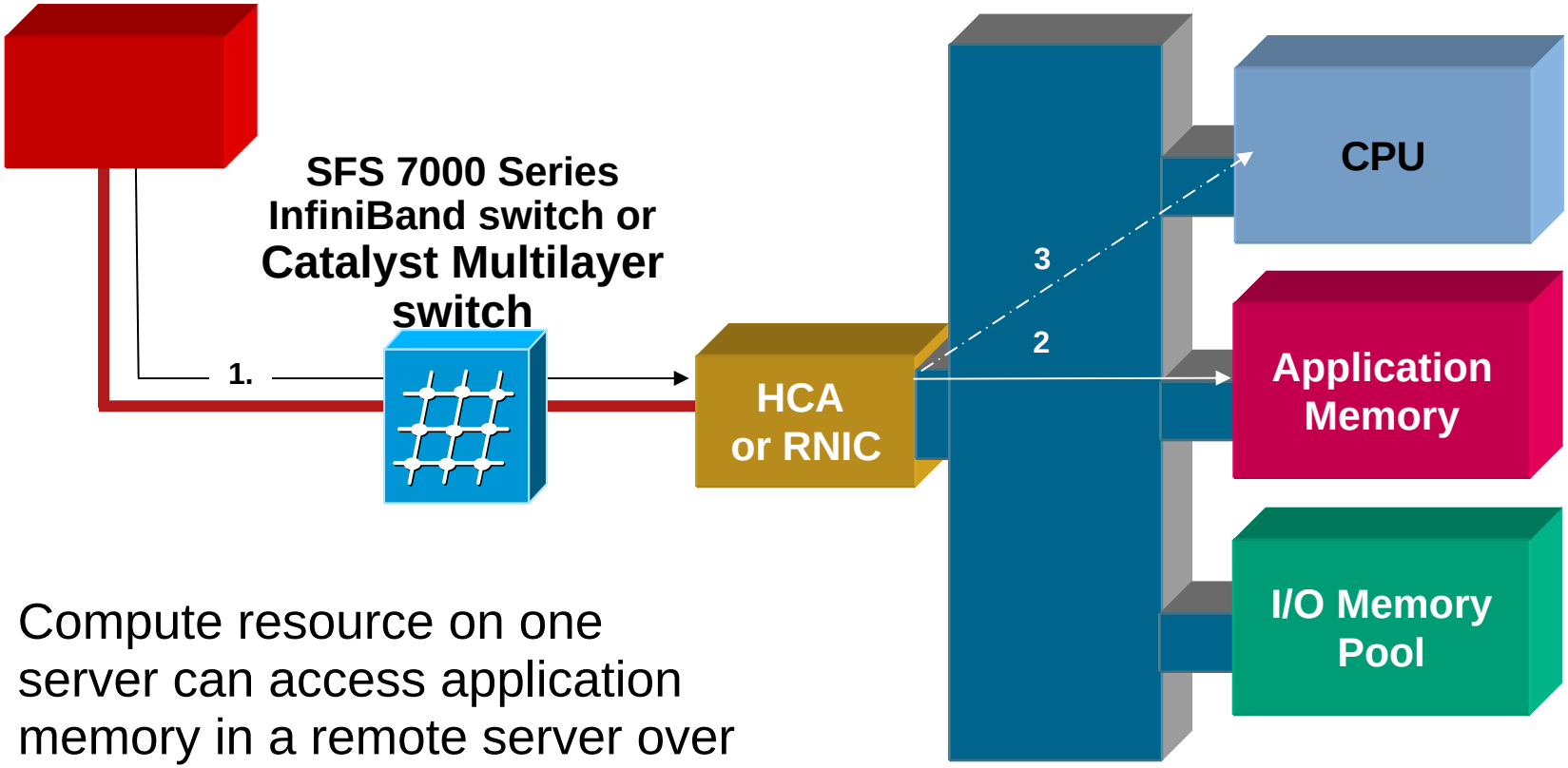
- Bus Based architecture with I/O memory pool
- Access to I/O resource handled by BIOS
- A data packet is typically copied across the bus three times

  CPU Interrupts, Bus bandwidth constrained, Memory bus constrained

Remote Direct Memory Access (RDMA) and Kernel Bypass



RDMA I/O Architecture

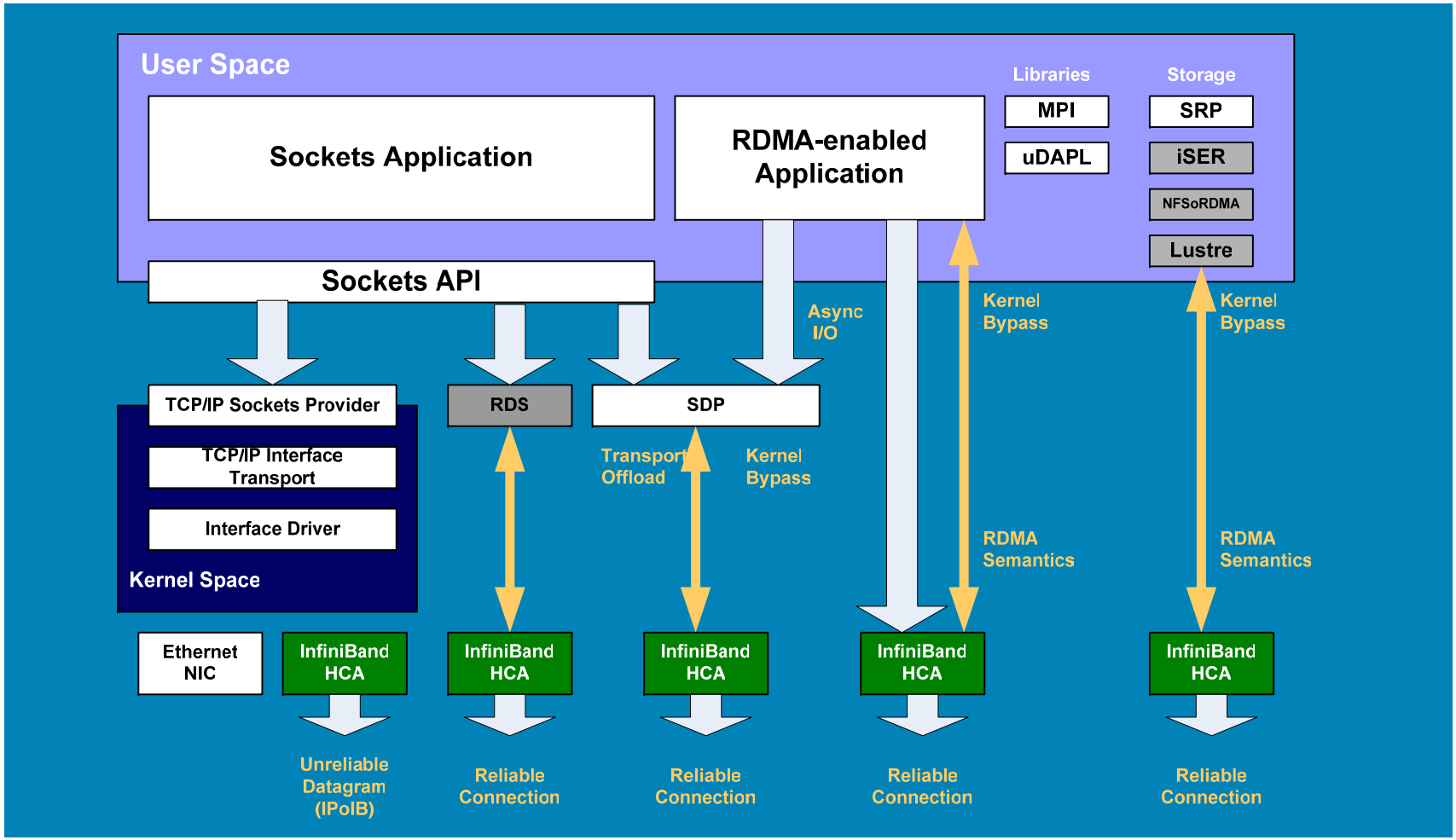


- Compute resource on one server can access application memory in a remote server over InfiniBand connection

InfiniBand Protocol Summary

	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.**
RDS (Reliable Data Sockets)	**RC communication with UDP**	**Used for IPC communication between cluster nodes for Oracle 10G RAC**
uDAPL (Direct Access Programming Library)	**Enables maximum advantage of RDMA flexible programming API.**	low level interface for application direct or kernel direct RDMAfunctions
MPI (Message Passing Interface)	**Low latency protocol used widely in HPC environments.**	**HPC applications.**

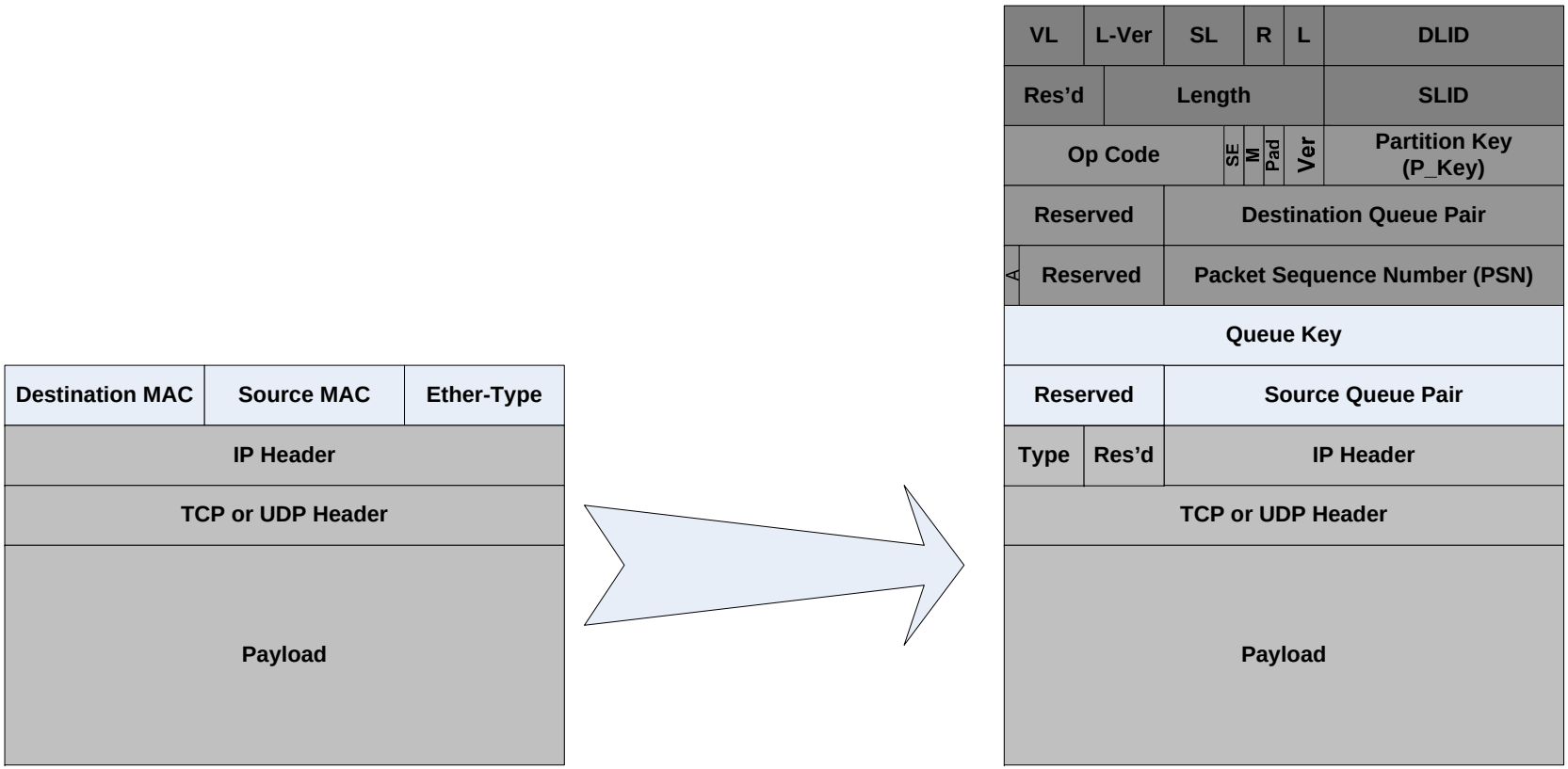
InfiniBand Upper Layer Protocols



IP over InfiniBand (IPoIB)

- IETF Standard (RFC 4931 / 4932)
- IP over InfiniBand provides the TCP / UDP socket interface for InfiniBand
  - Uses IB as a transport for IP
  - Support for IP Multicast over IB
  - No RDMA available in IPoIB
  - IPoIB is also used for address resolution for other protocols such as RDS, SDP, iSER
- Highest level of application compatibility no application change necessary
- Supported under Linux, Solaris, AIX, HP-UX, Windows
- SFS 3000 Ethernet Gateway may be used to bridge IPoIB traffic from InfiniBand to Ethernet

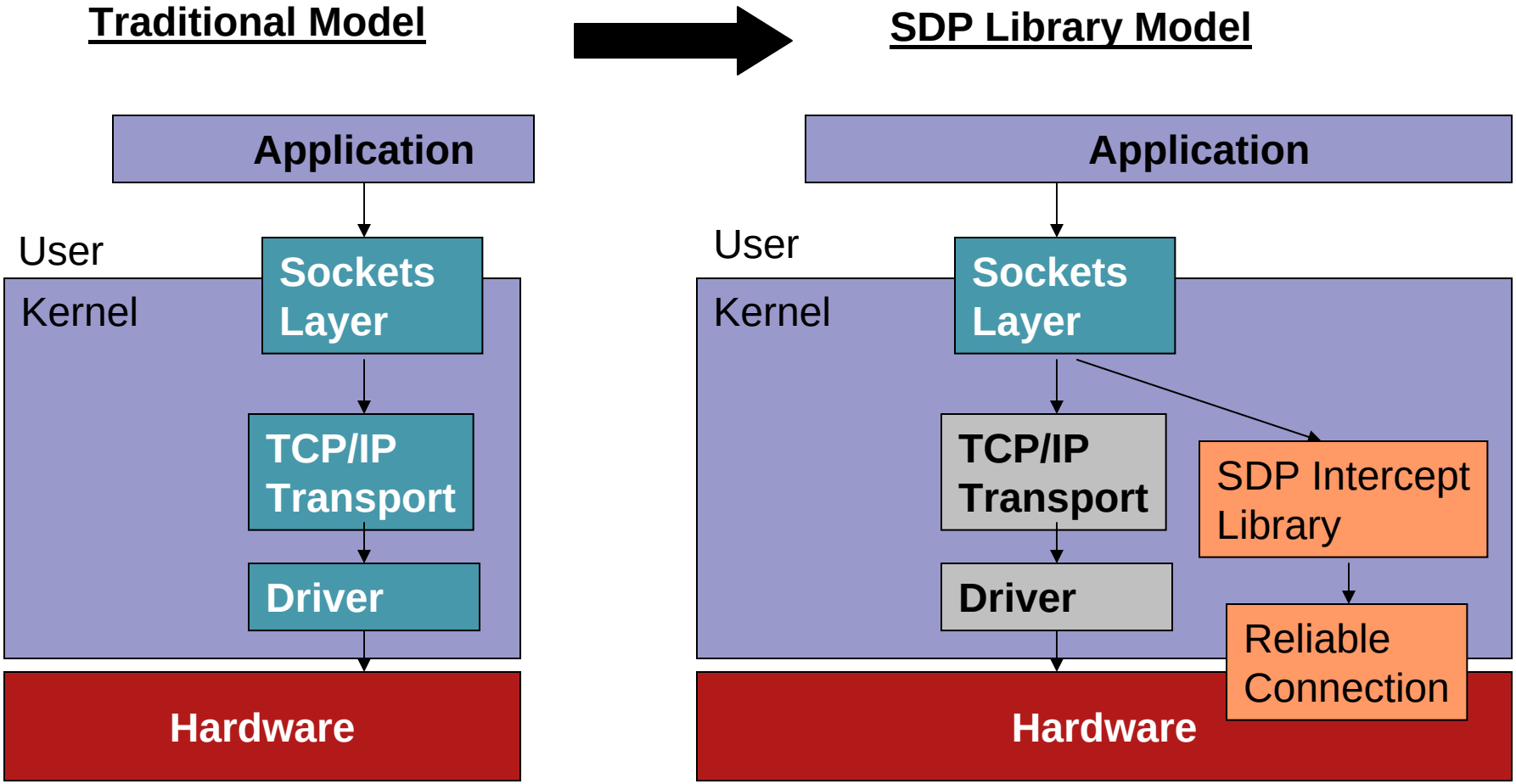
IPoIB Packet translation



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) cont



SCSI RDMA Protocol (SRP)

- Intended to run SCSI protocol to run over InfiniBand for SAN usage
  - T10 specification, similar to FCP
  - Transactions use RDMA for data movement from target to initiator
- Host drivers tie into standard SCSI/Disk interfaces in kernel/OS
- Linux, Windows, Solaris implementations
- Native SRP storage targets available today
- Not IB specific (no iWARP implementation yet)

Direct Access Provider Library (DAPL)

- Two variants: User DAPL (uDAPL)/Kernel DAPL (kDAPL)
- RDMA semantics API
- Provides low level interface for application direct or kernel direct RDMA functions (memory pinning, key exchange, etc.)
- DAT Collaborative makes open source implementation of uDAPL and kDAPL available: http://www.datcollaborative.org/

Message Passing Interface (MPI)

- MPI is the defacto standard API for parallel computing applications
- RDMA capabilities added via a set of patches to the base MPI code (MPICH, one of many available MPI libraries), initially developed at Ohio State University

  http://nowlab.cis.ohio-state.edu/projects/mpi-iba/

- Other MPI libraries now also include RDMA/IB capabilities such as Open-MPI

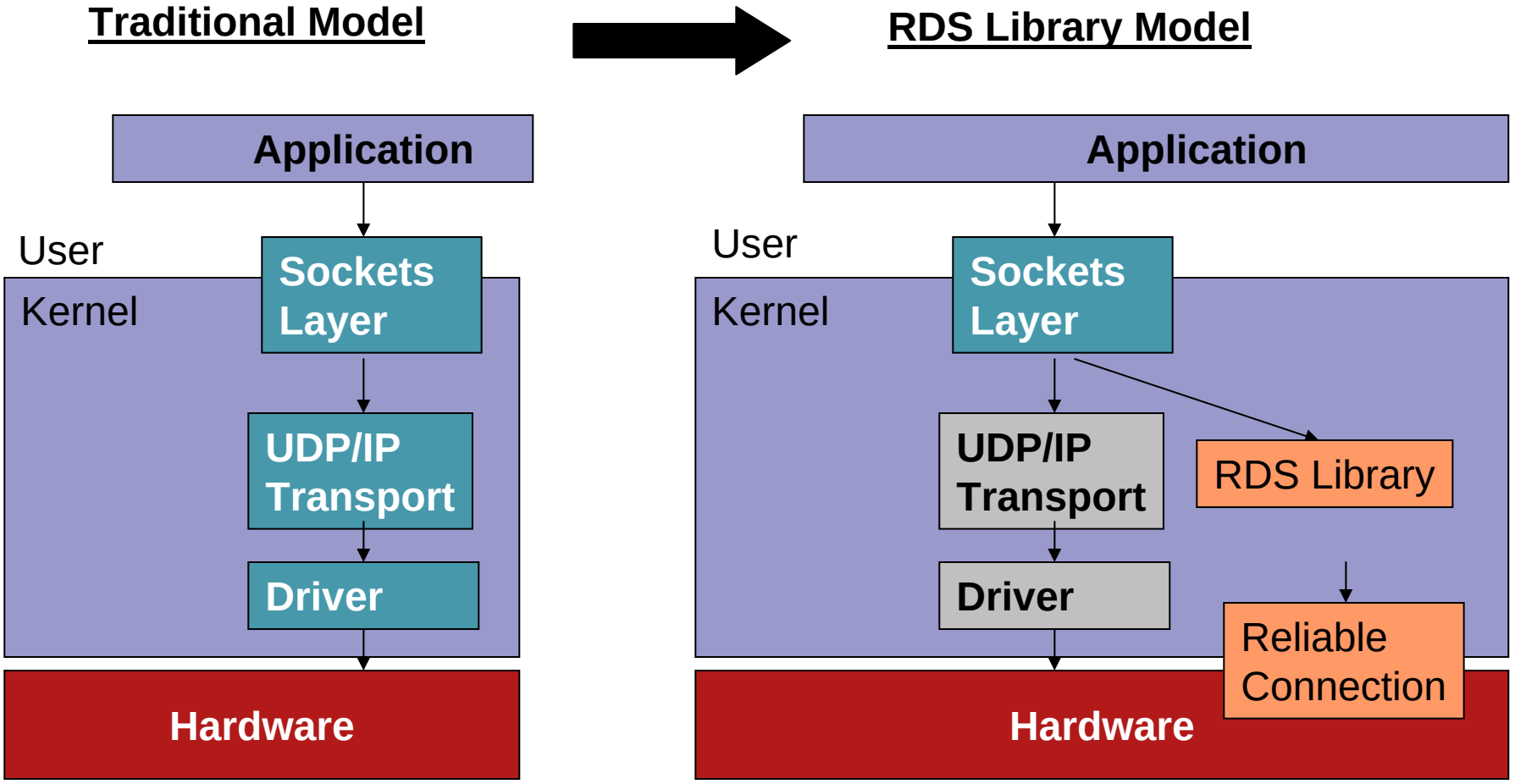
  http://www.open-mpi.org/

InfiniBand Performance

Measured Results

	Sockets API						MPI	
	TCP				SDP		MVAPICH	
	IP		IPoIB					
	Gigabit Ethernet	10 GE	SDR IB	DDR IB	SDR IB	DDR IB	SDR IB	DDR IB
Latency (us)	45.7	25.8	20.3	14.8	10	8.8	3.6	3.2
Bandwidth MB/s	118	1214	560	584	896	1033	960	1350
CPU	9%	25%	23%	26%	27%	28%	25%	25%

Reliable Datagram Sockets (RDS)



Reliable Datagram Sockets (RDS)

- Allows messages to be sent reliably to multiple destinations from a single socket
- User application can interface with RDS through a BSD style interface by specifying a separate socket family
- Minor code change / recompilation is required
- Runs socket based UDP/IP traffic with connection semantics
- RDS supports connectionless semantics at application level while using a RC QP InfiniBand connection between nodes.
- RDS is developed by Oracle for RAC accelerator to simplify their RAC implementation.

InfiniBand Hardware Overview

Cisco InfiniBand Hardware

New

IBM Bladecenter H
InfiniBand 4X
Switch & HCA

Cisco High
Performance
Subnet Manager

CiscoWorks - LAN
Management System

New

SFS-3012P

SFS-3012
Gateways

SFS-3012

SFS-7024P

SFS-7012P

New

4X DDR
PCI-Ex HCA

Device Fault
Manager

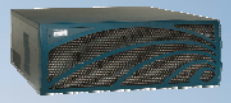
Resource Manager
Essentials

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

SFS-7000D
Family

New

SFS-7008P

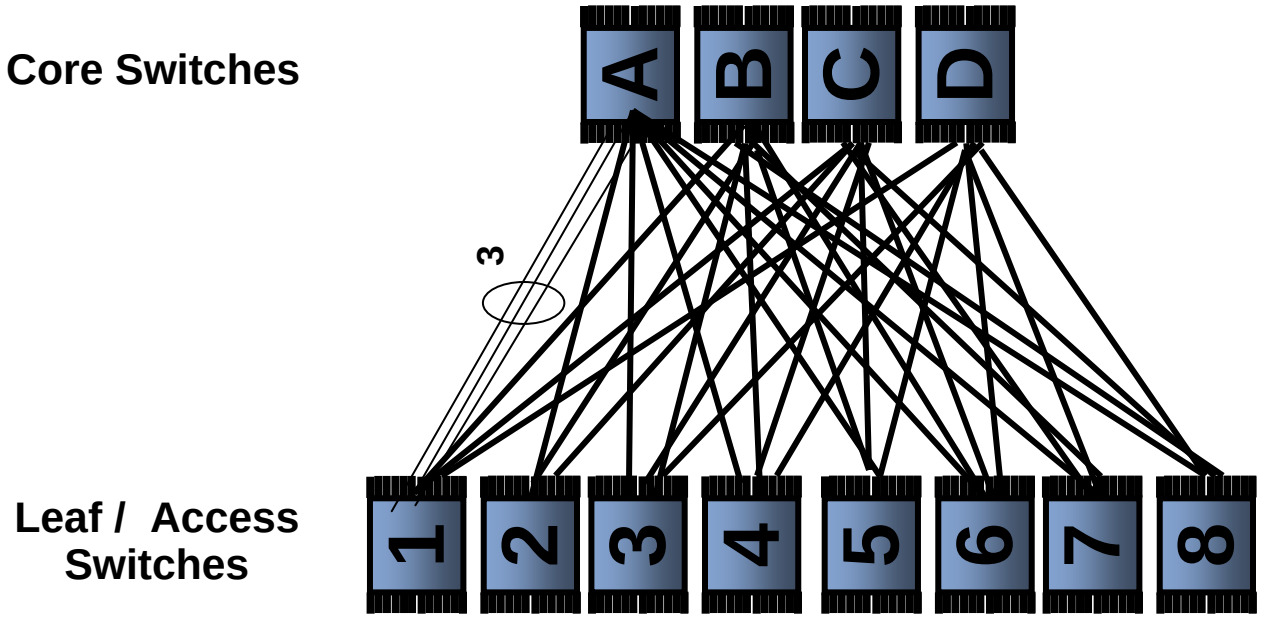


IBM Bladecenter
InfiniBand 1X
Switch & HCA

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

SFS-7000P

SFS 7008 (modular 96 Port) Internal Architecture



- The 24 port SFS 7000 is based on a single Anafa-II chip
- SFS 7008 96 port non-blocking switch needs 8 access switches
- 8 access switches have (8x12) 96 uplinks, hence we need four 24 port core switches
- Design Rules

  - Equal number of links from each access to core - 12 / 4 = 3 links to each core
  - Access switch must connect to each core switch

24 Port Switch on Chip (SOC) Features

- Switch on chip (SOC) ASIC architecture
- Single Data Rate InfiniBand offers 200ns latency
- Double Data Rate InfiniBand offers 140 ns latency
- Cut-through switch with sufficient memory
- InfiniBand is a credit-based network
- Full bi-section bandwidth available within switch chip
- QoS (SL-VL mappings)
- Port – span
- Subnet Management Agent on each chip

InfiniBand Host Channel Adapter (HCA)

- Network interface for IB attached Servers
- 2 IB Ports (Independent, Fail-over or Aggregate modes)
- Memory and Memory-Less versions
- PCIx and PCI-e
- Provides:

  Hardware Virtual/Physical Memory Mapping with Memory Protection

  Direct Memory Access (DMA) and Remote DMA transfer engine

  Reliable Packet forwarding capability

  Communication Queue Pair and Credit Management

  Async I/O Operations

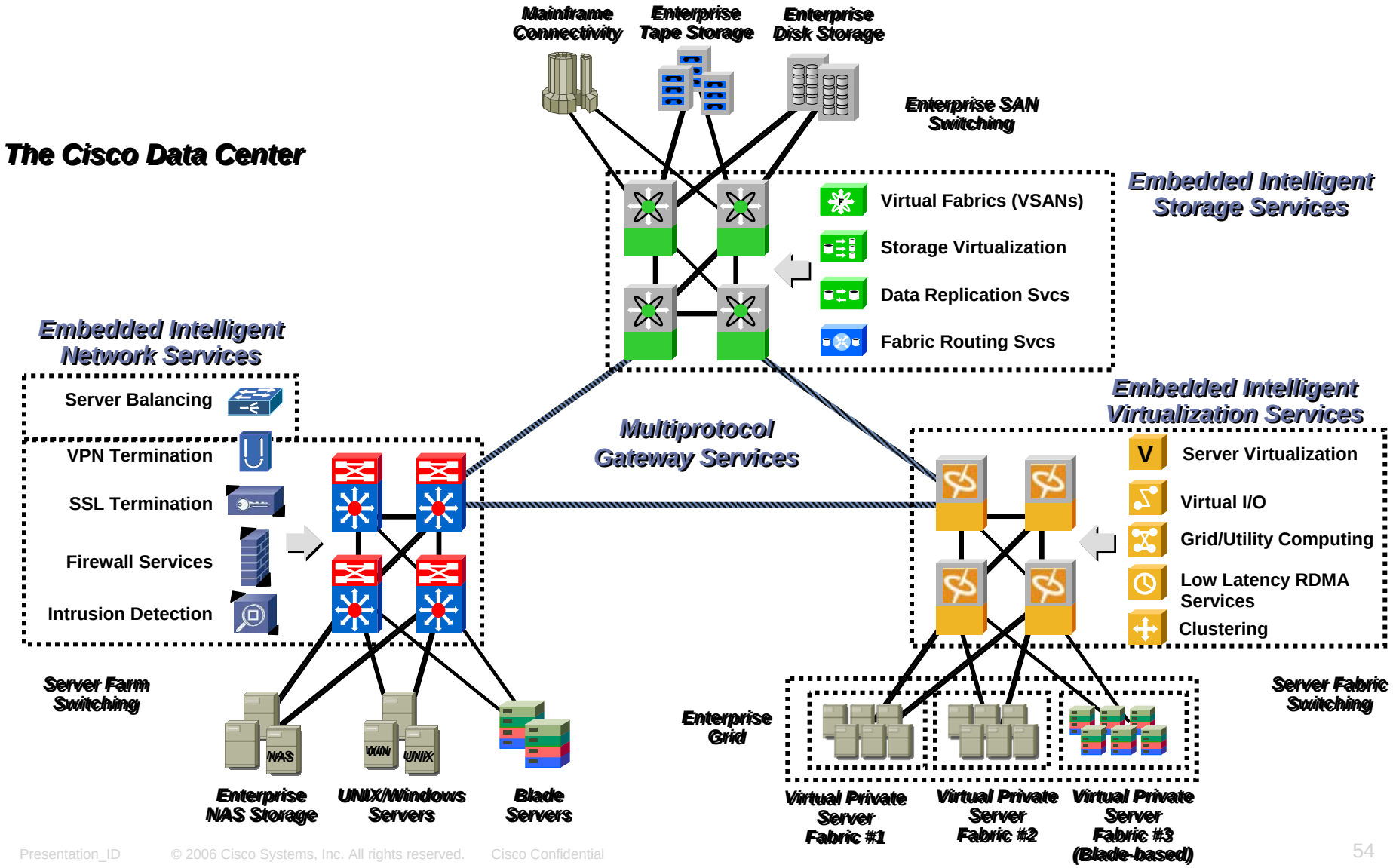
InfiniBand Gateway

- Technically a Target Channel Adapter (TCA)
- Similar to an HCA attached to an embedded device
- Usually doesn't require virtual memory manipulation and mapping
- Simplified HCA on a specialized device

  Examples, Ethernet to InfiniBand or Fibre Channel to InfiniBand packet forwarding engines

InfiniBand Applications

InfiniBand in the Data Center



InfiniBand Applications

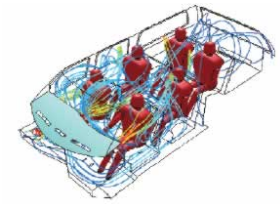
- High Performance Computing (HPC)
- High Speed Cluster Interconnect (Oracle RAC)
- Financial Services Industry
- I/O Virtualization (Multi Fabric I/O)

High Performance Computing

Enterprise HPC Application Areas

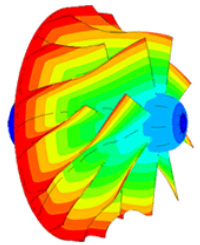
Computational Fluid Dynamics

- Fluent
- Star-CD
- Exa PowerFlow
- Vectis



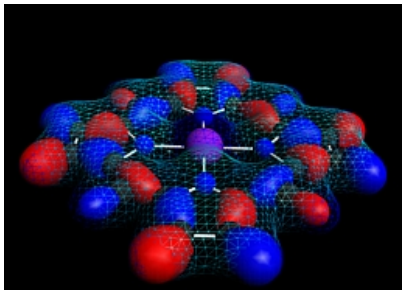
Finite Element Analysis

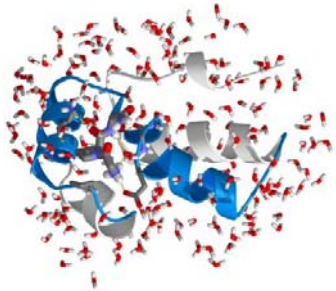
- Abaqus
- Ansys
- LS-Dyna (3-D FEA)
- RADIOSS
- NASTRAN
- PAM-Crash



Computational Chemistry Material Sciences:

- Accelrys
- ADF
- GAMESS
- CHARMm
- Namd
- NWChem
- GROMACS



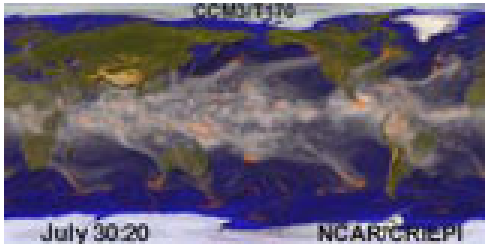


BioInformatics

- BLAST
- FASTA
- ClustaW
- EMBOSS

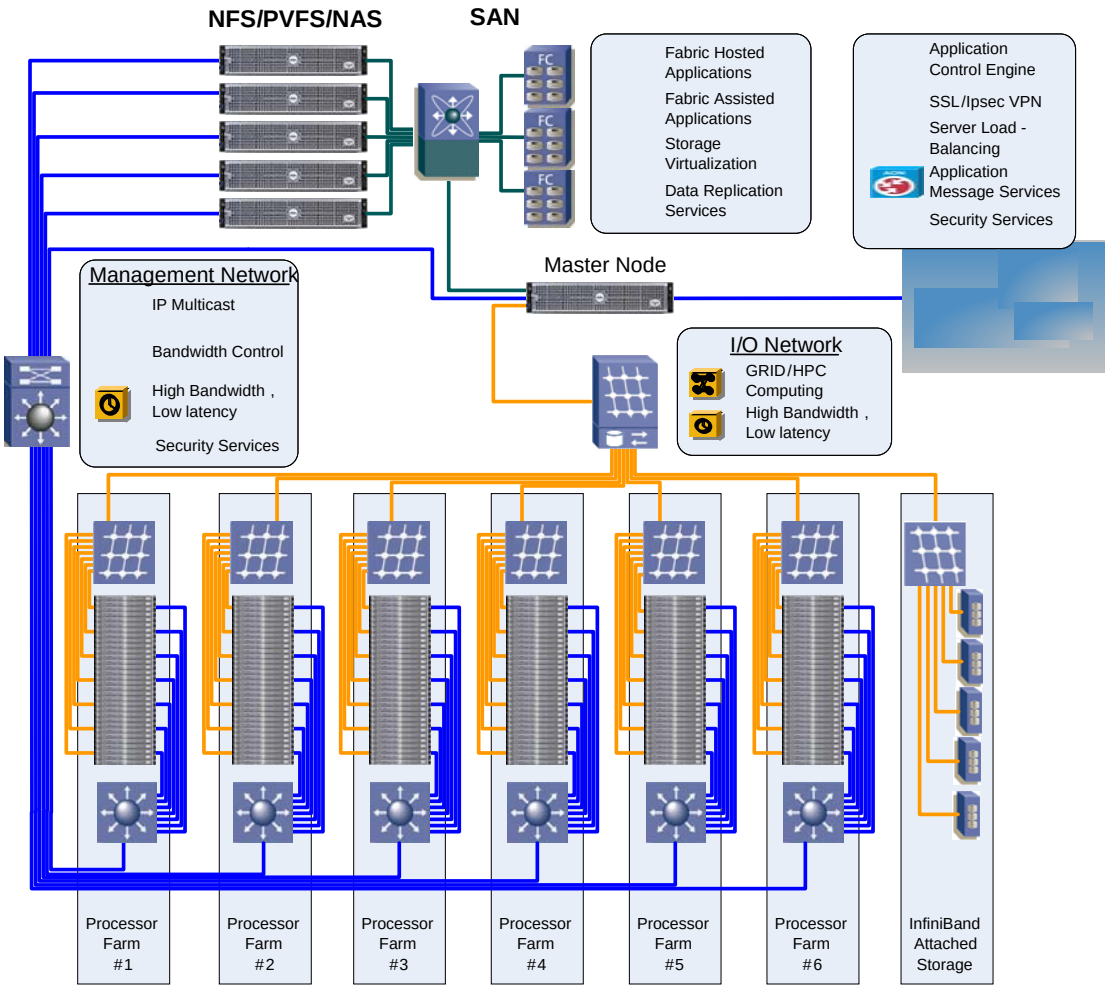
Climate & Weather Simulation

- MM5
- WRF
- CCM3



High-Performance Computing Solution

- **Interprocessor Communication (IPC) Network**
  - Low latency, high bandwidth on standard open source MPI over Infiniband network
  - Catalyst switching for TCP-based applications, can benefit from policing, QoS and multicast
- **Management and I/O Network**
  - Used for job scheduling, network monitoring
  - TCP or UDP based – benefits from Quality of Service and Multicast
  - Netflow reporting, NSF/SSO for high availability
- **Storage Network**
  - NAS or iSCSI over Ethernet fabric
  - IB-attached storage for lower storage overhead
  - FibreChannel storage with data replication, integrated applications



HPC InfiniBand Network Design

- Decide or determine the number of nodes in the fabric
- Determine if you will use a single switch or a “CLOS-style” network design

  Almost all IB networks should be CLOS-based designs
- Decide on a blocking factor based on application throughput needs (all data rates full-duplex)

  Non-blocking = 8Gb/s true data rate throughout fabric

  2:1 Oversubscribed = 50% blocking = 4Gb/s true data rate worst-case

  3:1 Oversubscribed = 67% blocking = 2.67 Gb/s true data rate worst-case
- Determine which types of switches you will use for your edge, core, or single switch needs
- Use the Cisco Clustomizer to determine the number of HCAs, switches, and cables in your design
- Determine your floor plan, rack layout, and cable lengths

HPC InfiniBand Network Design cont.

Non-Blocking	**8Gb/s app data**
50% Blocking	**4Gb/s app data**
67% Blocking	**2.67Gb/s app data**

"app data" = actual data rate with IB overhead removed

- Direct trade off between budget and performance
- Blocking fabrics require many fewer switches and cables

  But not a linear decrease: 50% blocking fabrics require more than 50% of the switches and cables
- Latency is unaffected by blocking factor, as long as throughput never exceeds true data rate

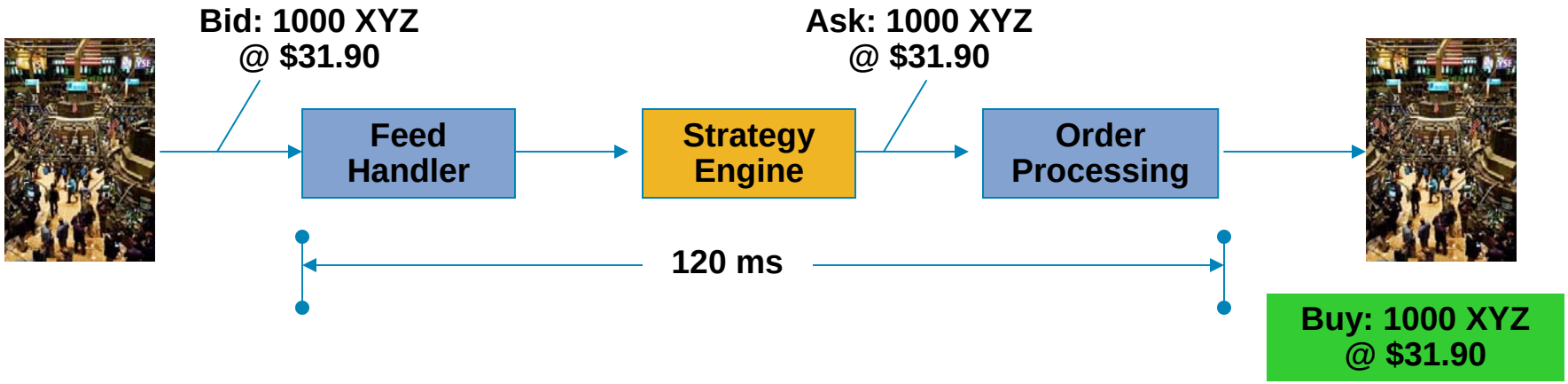
  If throughput exceeds 8Gb/s on a non-blocking fabric, latency increases just the same
- 50% blocking most popular config today

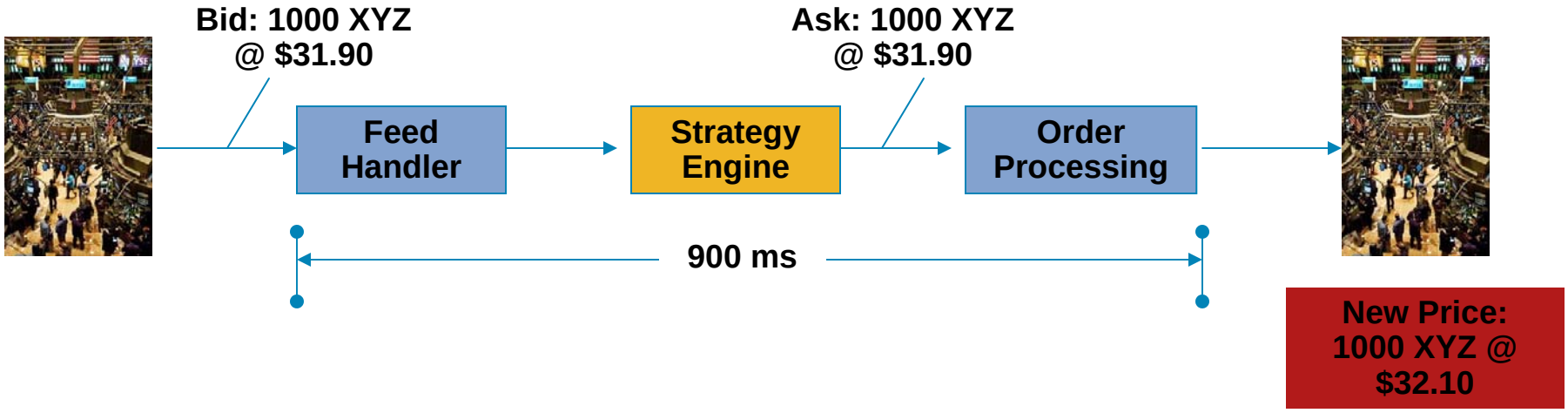
HPC InfiniBand Network Design cont.
DDR Advantages and Limitations

- DDR does not offer significant latency advantages over SDR
- DDR 30% more expensive for a best case 30% bandwidth gain
  - Higher bandwidth for large message sizes
  - DDR ROI dependent on application's data traffic pattern
  - Does your application primarily depend on large message transfers?
- DDR is distance limited to 8 meters
  - Currently investigating active cable technology to extend DDR cable reach
- DDR Possible for Trunking Between Edge and Core Layers
  - Reduces costs of core ports and cables
  - Speed Transition Increases Latency with "Store and Forward Penalty" (translates SDR to DDR) – up to 1usec max

Financial Services Industry

Impact of Latency on Proprietary Trades Desk





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

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

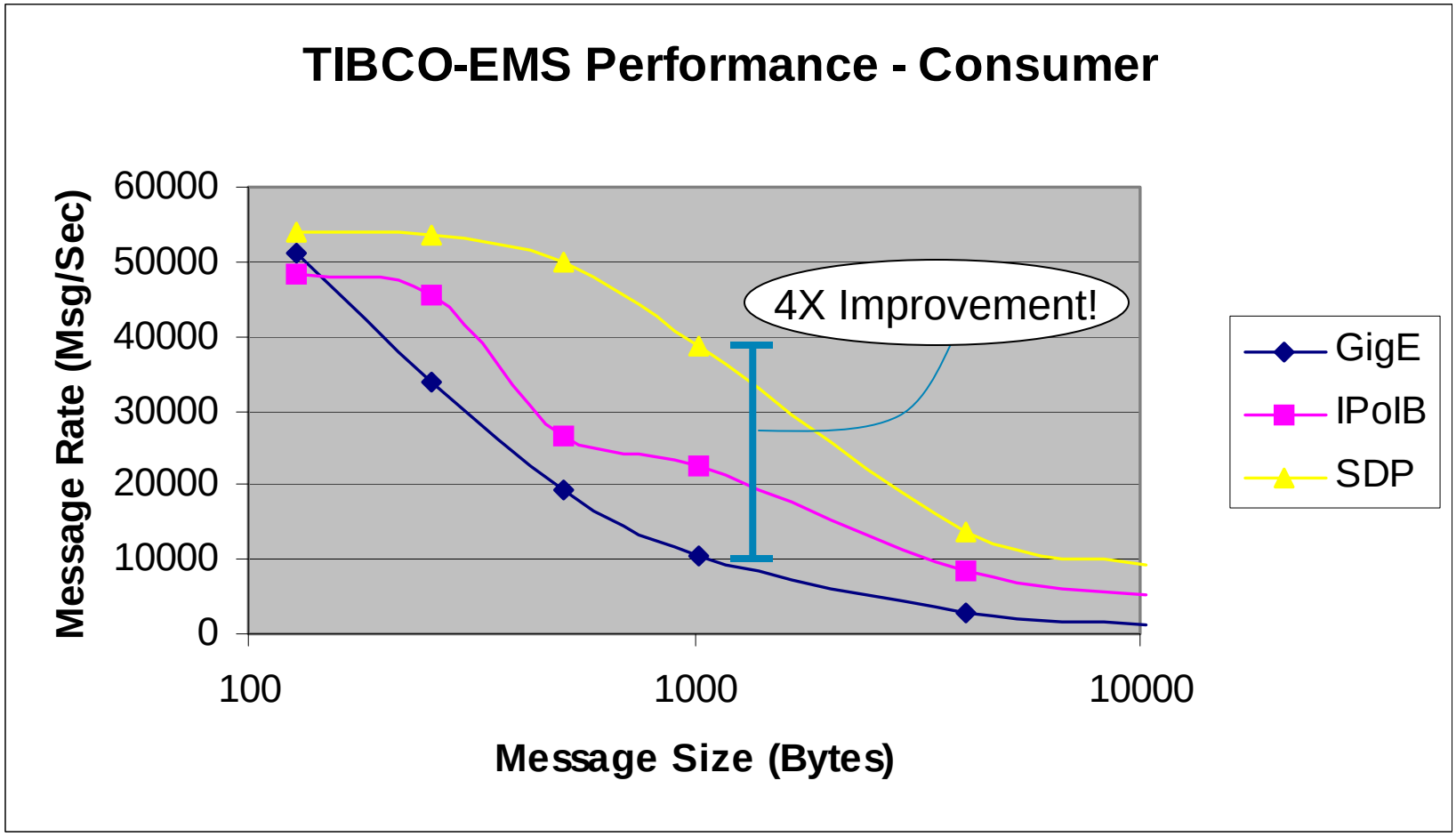
  High Throughput – From IB rates of 10Gbps and 20Gbps

  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

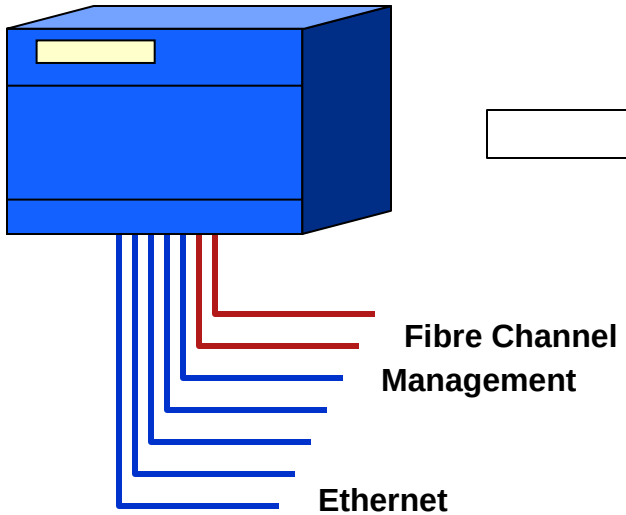
TIBCO-EMS Benchmarks (Consumer)



Multi-Fabric I/O

Multi Fabric I/O (I/O Virtualization)

Traditional Datacenter I/O

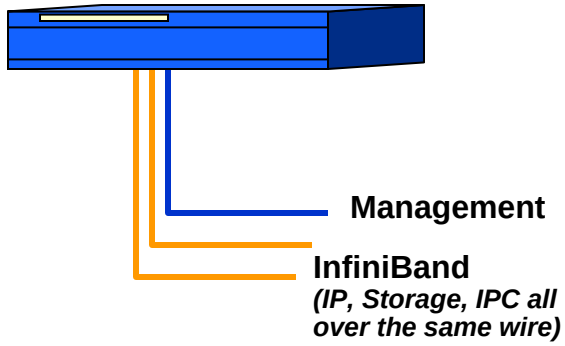


Ethernet (2-5 connections)
- **Web/Client**
- **Backup/Restore**
- **Management**
- **DMZ**
- **Database**

Fibre Channel (0-4connections)
- **up to four for I/O intensive apps like Oracle**

Proprietary cluster interconnects

Unified Fabric I/O



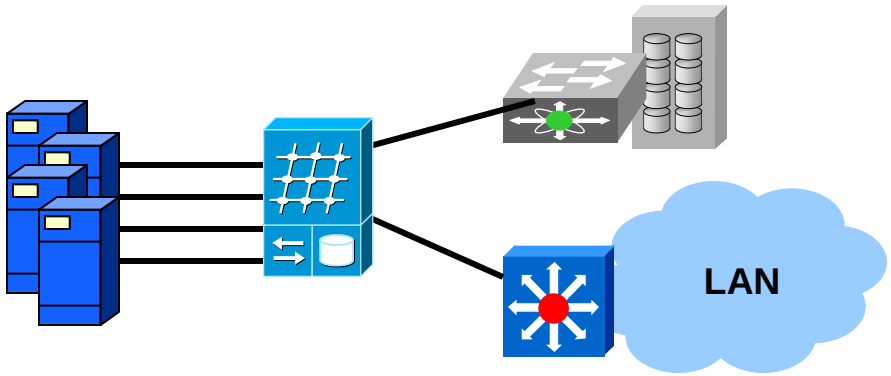
Two InfiniBand connections
- **IP, FC and IPC traffic all run over one or two 10Gbps IB pipes**
- **Data Center shares IP and FC access through central I/O gateways**

One Ethernet for management

Physical vs. Logical View

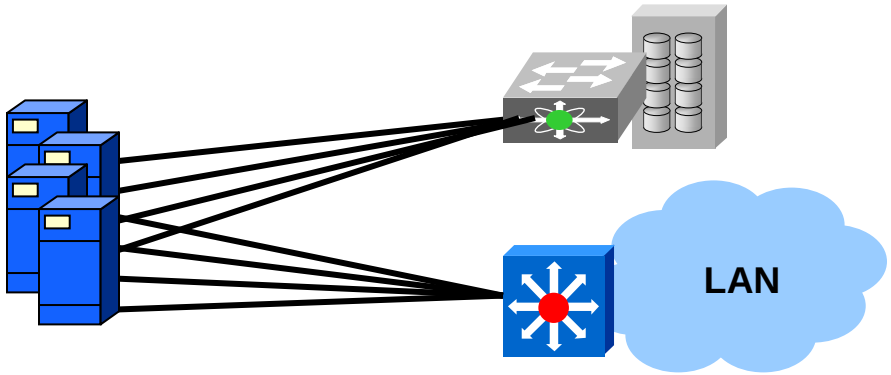
Physical View

- 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



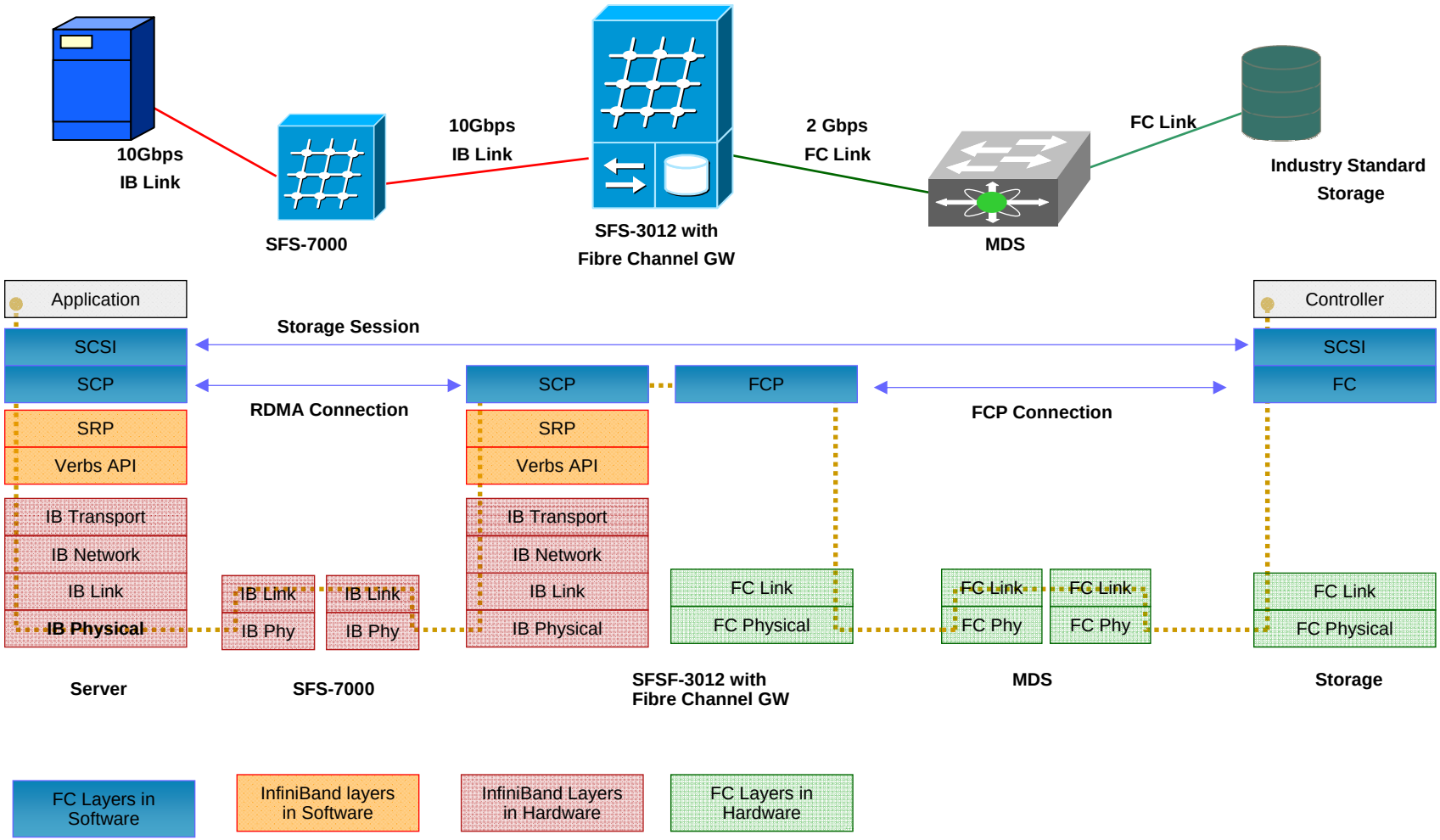
InfiniBand-to-Ethernet Gateway Features

- IP-Only protocols (no SDP)
- 802.1Q VLAN support
- One VLAN is mapped to one IB partition
- Up to 32 VLAN's per gateway
- Link aggregation
- IPv4 multicast support
- Loop protection
- Ethernet jumbo frames up to 9k
- IP fragmentation
  - Up to 2044 bytes IPoIB MTU
  - Ethernet frames larger than 2044 bytes are fragmented
  - Up to 9k Ethernet MTU
- High availability

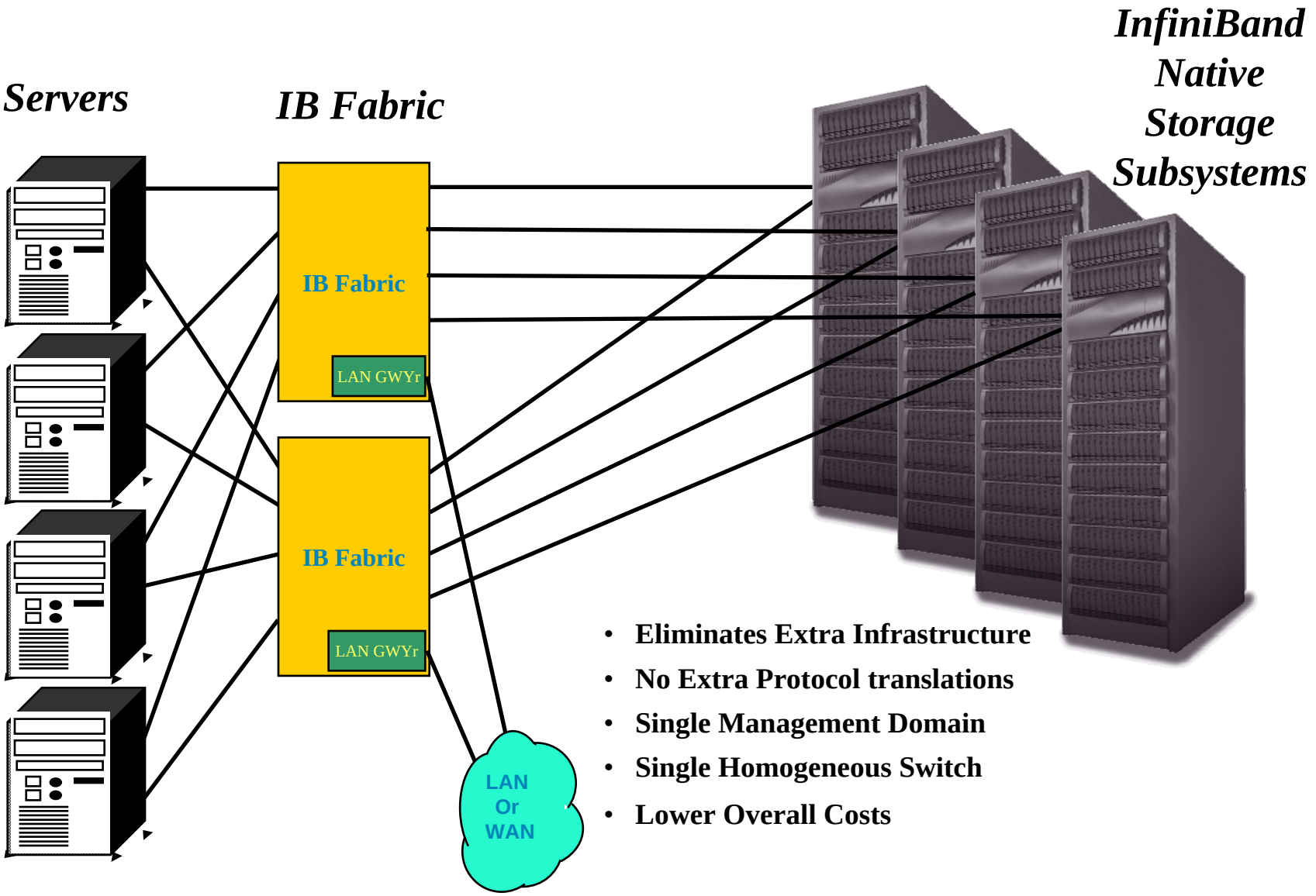
InfiniBand-to-Fibrechannel Gateway Features

- **Storage Gateway presents either:**
  - **Fabric Attached Loops**
  - **E-Port**
- **SCSI RDMA (SRP) Driver installs on host as normal SCSI driver. Defined by ANSI T10 standards.**
- **Each IB/SRP Initiator is assigned:**
  - **1 FC WWNN and**
  - **Multiple WWPNs**
- **Unique WWNs allow normal zoning to work as usual.**
- **Storage-based load balancing works as usual.**
- **Enhanced multipathing and I/O consolidation**

SCSI over InfiniBand (SRP) Gateway



Native SCSI over InfiniBand

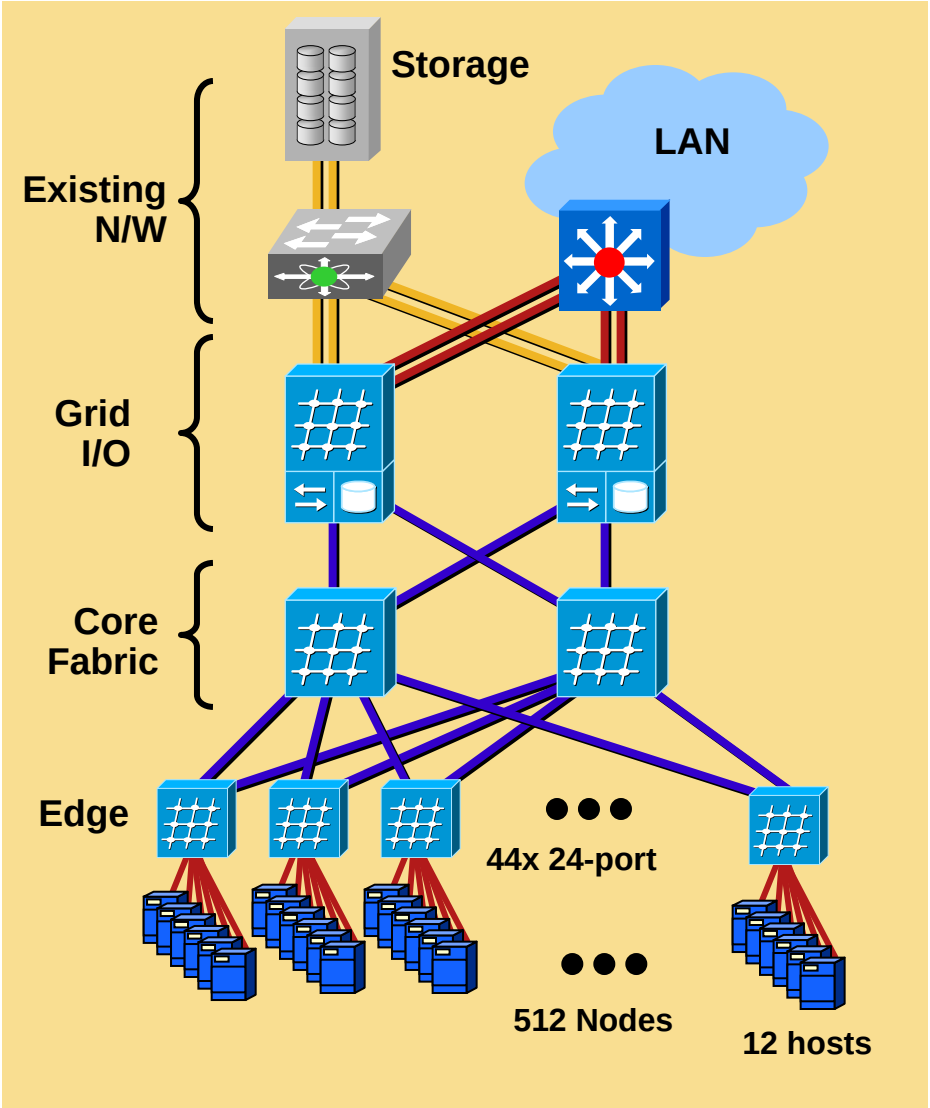


- **Eliminates Extra Infrastructure**
- **No Extra Protocol translations**
- **Single Management Domain**
- **Single Homogeneous Switch**
- **Lower Overall Costs**

Case Studies

Case Study: Large Enterprise Customer
Enterprise Grid Computing

- Application:
  - Replace proprietary platforms with standards-based components
  - Build scalable “on-demand” compute grid for financial applications
- Benefits:
  - 30-50% Application Performance Improvement
  - Standards-based solution for on-demand computing
  - Environment that scales using 500-node building blocks
  - Centralized shared I/O pool for flexibly allocating SAN/IP bandwidth



Meet the Experts
Data Centre

- Victor Moreno
  Technical Leader

Recommended Reading

BRKDCT -3009

- Continue your Networkers learning experience with further reading from Cisco Press.
- Visit the on-site Cisco company store, where the full range of Cisco Press books is available for you to browse.

Q and A