INFINIBAND TECHNOLOGY FOR HIGH PERFORMANCE COMPUTING AND SERVER VIRTUALIZATION

用InfiniBand技术实现高性能计算和服务器虚拟化

Jerry LIU

poweredbycisco. networkers 2005

Agenda 议程

- **InfiniBand Hardware Overview (IB硬件概况)**
- **InfiniBand System Overview (IB系统概况)**
- **RDMA and Upper Layer Protocols (RDMA和上层协议)**
- **High Performance Computing Architectures** （高性能计算体系结构）
- **HPC Building-blocks**（**HPC** 构成块）
- **I/O Virtualization**（**I/O** 虚拟化）
- **Server Virtualization**（服务器虚拟化）

InfiniBand Hardware Overview
硬件概观

什么是 IB What Is InfiniBand?

- **InfiniBand is a high speed – low latency technology used to interconnect servers, storage and networks within the datacenter (IB 是高带宽低延迟的互联技术)**
  在数据中心里将服务器，存储和网络连接在一起
- **Standards Based – InfiniBand Trade Association**

  http://www.infinibandta.org 基于工业标准
- **Scalable Interconnect:** 可扩展的互联技术

  1X = 2.5Gb/s

  4X = 10Gb/s

  12X = 30Gb/s

InfiniBand Physics 物理特性

- **Copper and Fiber interfaces are specified**
- **(使用铜缆和光纤传送数据)**
- **Copper 铜线电缆**
  - **Up to 15m* for 4x connections 10G最长15m**
  - **Up to 10m for 12x connections 30G最长10m**
- **Optical 光缆**
  - **Initial availability via dongle solution 最早使用dongle**
  - **Up to 300m with current silicon 目前芯片到300m**
  - **Long Haul possible, but not with current silicon**
  - 不使用目前芯片可以支持更长距离

*** 20m in certain circumstances**

InfiniBand Physics物理性质

- **Link is bonded 2.5Gbps (1x) links 以2.5G为一个links**
- **（一个链接以2，5Gbps 为单位，）**

  Fiber is a ribbon cable 光纤电缆

  Copper is a multi-conductor cable 铜缆为 多心电缆

- **Each Link is 8b/10b encoded 每个Link做8b/10b 编码**

  4x Link is 4 2.5Gbps Physical Connections

  4倍速的连结是 4 个 2.5Gbps 的物理连接

  Each connection is 2Gbps data 每一个连接2Gbps数据

  SAR provides a single 8Gbps data connection (4x)

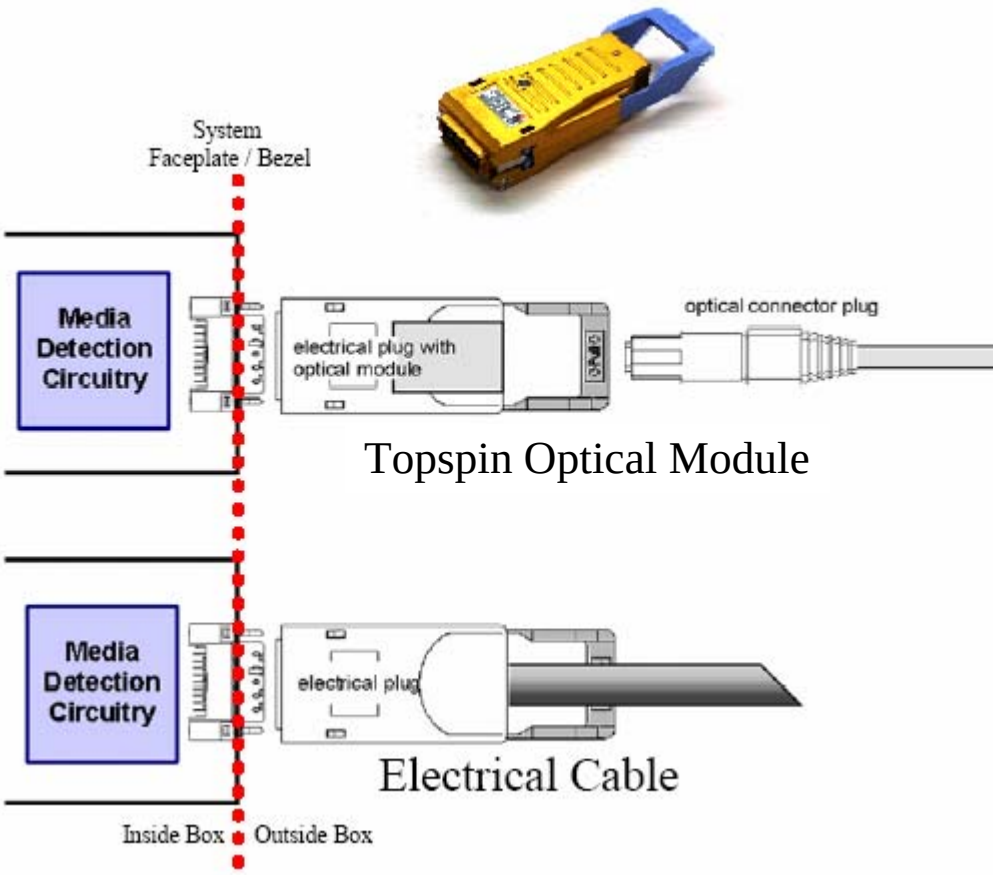
  24 Gbps (12x) SRP提供单一的8Gbps 或者24Gbps 数据连接

Pluggable Optics Module 光纤插接模块

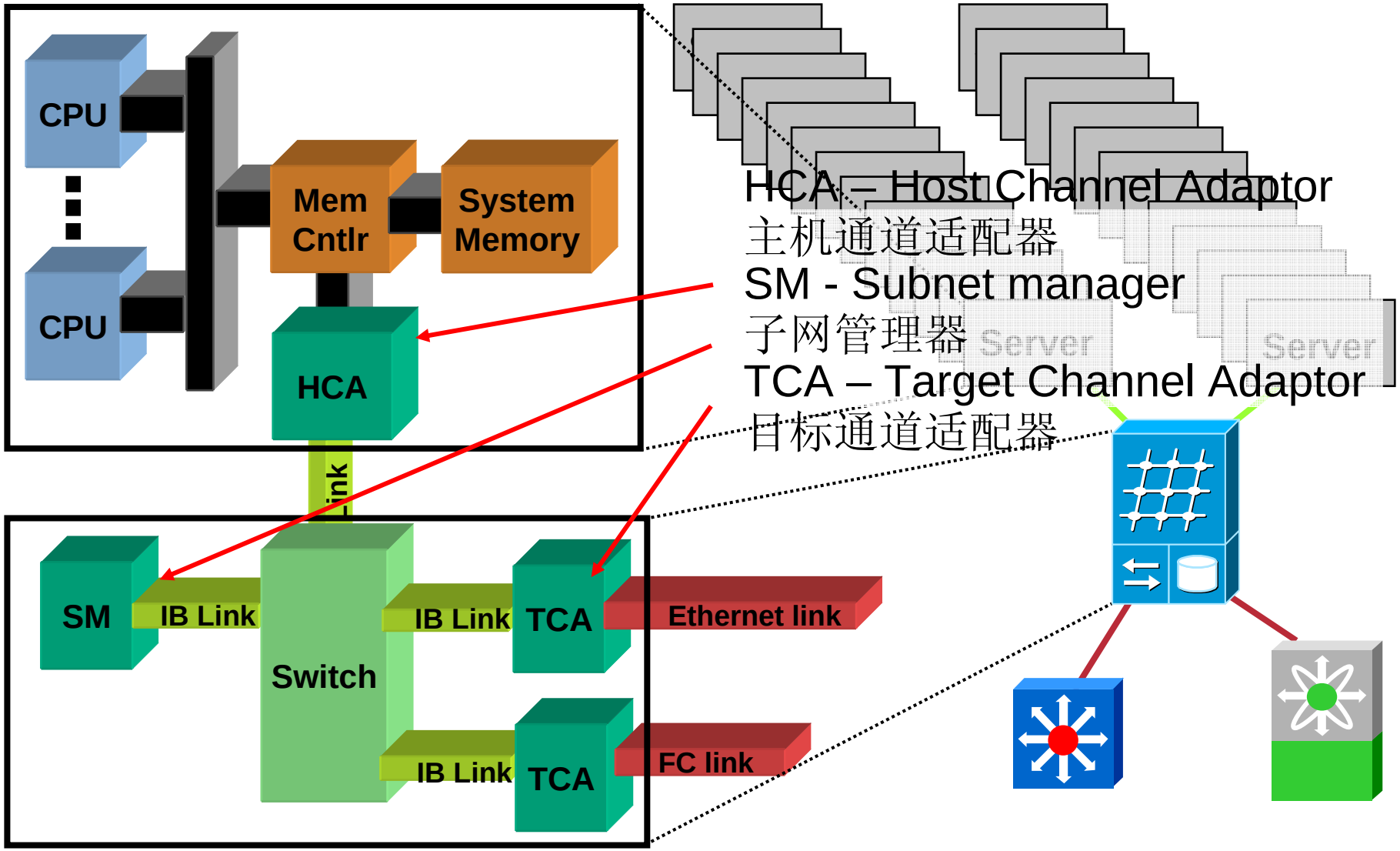
Transforms Powered Copper Ports to Optical Ports

铜缆口转换成光缆口

- **Coverts a copper port to an optical port on a port by port basis**
- **Extends port to port reach to 150m - 300m with fibre ribbon cables**
- 使用光缆可以延长口到口的距离达到**150m- 300m**



InfiniBand Nomenclature 专业术语



InfiniBand Switch Hardware 交换机硬件

- **Hardware switch devices is a cut-through memory switch** 交换机硬件是一个直接连接内存的交换机
- **Full-duplex, non-blocking 24 port tag forwarding switch** 全双工，无阻塞**24**口为单元的交换机
- **Tags are system Local ID, provided to all network endpoints by the Master Subnet Manager on system startup** 系统启动后，通过系统的主子网管理器来提供整个系统的节点管理，表记系统的本地**ID**（标识）

InfiniBand Host Channel Adapter 主机通道适配器

- **Network interface for IB attached Servers**
- （**IB**连接的服务器的网络接口）
- **Provides hardware Virtual/Physical memory mapping, Direct Memory Access (DMA), and memory protection**
- （提供虚拟**/**物理内存的映像的硬件，内存直接访问和内存保护）
- **Provides RDMA (Remote DMA) data transfer engine and reliable packet forwarding capabilities**
- （提供远程**DMA**来做数据传送和可靠的数据包提升传送能力）

InfiniBand Gateway 网关

- **Technically a Target Channel Adapter**
- （目标通道适配器的技术）
- **Similar to an HCA attached to an embedded device**
- （类似与**HCA**直接连接一个嵌入设备）
- **Usually doesn’t require virtual memory manipulation and mapping** 通常不需要虚拟内存操作和映像
- **Simplified HCA on a specialized device**
- （专有设备上的简单的**HCA**）

  Examples, Ethernet to InfiniBand or Fibre Channel to InfiniBand packet forwarding engines例如，以太网到**IB**， 光线通道到**IB** 数据包的传送引擎

InfiniBand System Overview
InfiniBand 系统概观

InfiniBand System Architecture 系统体系结构

- **Connection Oriented Architecture**连接导向架构
  - **Central connection routing management (SM)**
  - 中央连接路有管理（子网管理器）
  - **All communications based on send/receive queue pairs**
  - 所有的通讯基于一对发送/接受队列
- **Two primary connection types**两个主要类型的连接
  - **Reliable Connection** 可靠 **Unreliable Datagram**不可靠
- **Unused connection types**不使用的连接类型
  - **Unreliable ConnectionReliable Datagram Raw Datagram**

InfiniBand Connections 连接

- **Reliable Connection** 可靠的连接
  - **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** 不可靠的数据包
  - **Best effort forwarding**
  - **Used for IP over IB communications**

InfiniBand Subnet Manager 子网管理

- **IB Fabric is called an InfiniBand Subnet**
- **（IB网络叫做InfiniBand 子网）**

  All devices under the control of a single Master Subnet Manger (SM)　所有的设备在在一个SM下控制

  May have multiple slaves with replicated SM database state

  可能会有多个从属 **SM** 可以复制**SM**数据库的状态

- **At system startup, all devices register with the SM** 系统启动时，**SM** 上注册所有的设备

  Central Routing function 中央路由功能

  Shortest Path First Routing 最短路由

  Equal Paths Load balanced with static round robin distribution

  Connection endpoint lookup 连接重点地查找

Clusters 2.0 Subnet Manager: Fabric Sweep Performance 网络启动性能

Number of Hosts	Time
32	< 1 sec
64	< 1 sec
128	2 sec
256	4 sec
512	22 sec
1,024*	35-40 sec
2,048*	1-1:30 min**
4,096*	5-7 min**

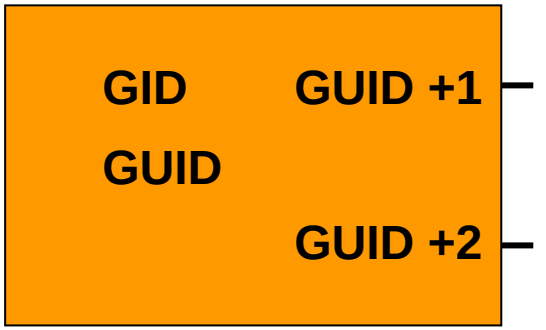
** Requires HPC Subnet Manager for this performance*

*** Estimated based on simulation*

- Assumes InfiniSwitch-III based two tier topology
- Embedded SM can handle up to 1,024 nodes

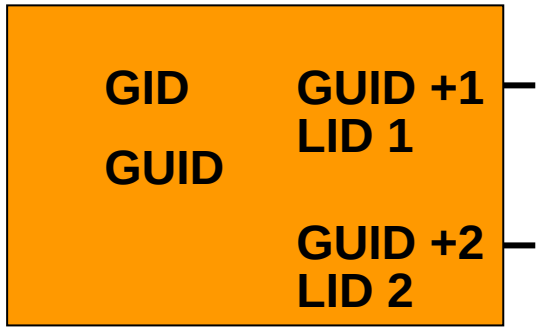
IB Addressing 地址

- **Three addresses: GUID, GID, LID**
- **GUID**
  - **Global Unique ID 64 bits in length**
  - **Used to uniquely identify a port or port group**
  - **HCA and each port has a GUID**
  - **(e.g 00:05:ad:00:00:01:02:03)**
- **GID**
  - **GUID plus Subnet prefix GUID 加子网前缀**
  - **Used for host lookup on a subnet 在子网里查找主机**
  - **Used for inter-subnet IB routing (future) 用以在子网内部路由**
  - **(e.g. fe:80:00:00:00:00:00:00:00:05:ad:00:00:01:02:03)**



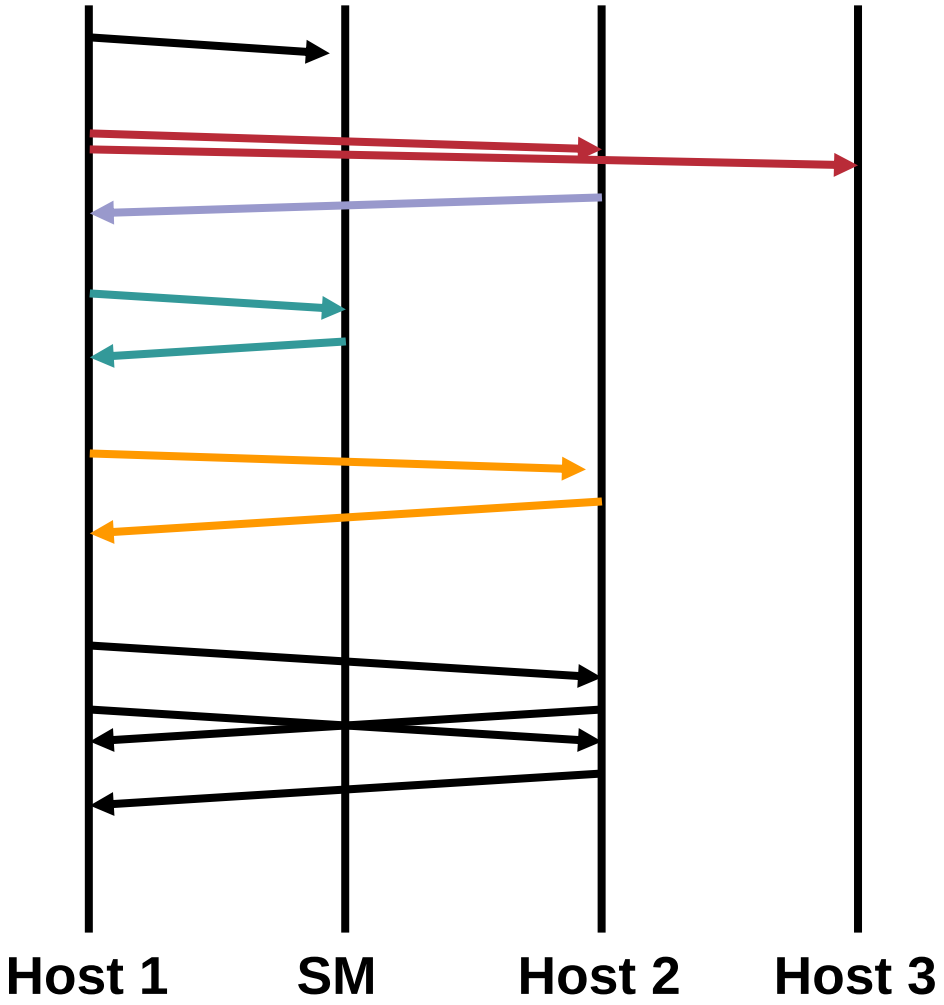
IB Addressing 地址

- **LID**
  - **Local ID 本地ID**
  - **Assigned by SM to define a switchable endpoint in the network**
  - **Subnet Local address**



- **Queue Pair 一对队列**
  - **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**

Address Resolution 寻址



Join Multicast Group

Send ARP on “Broadcast addr”

Receive remote GID via ARP

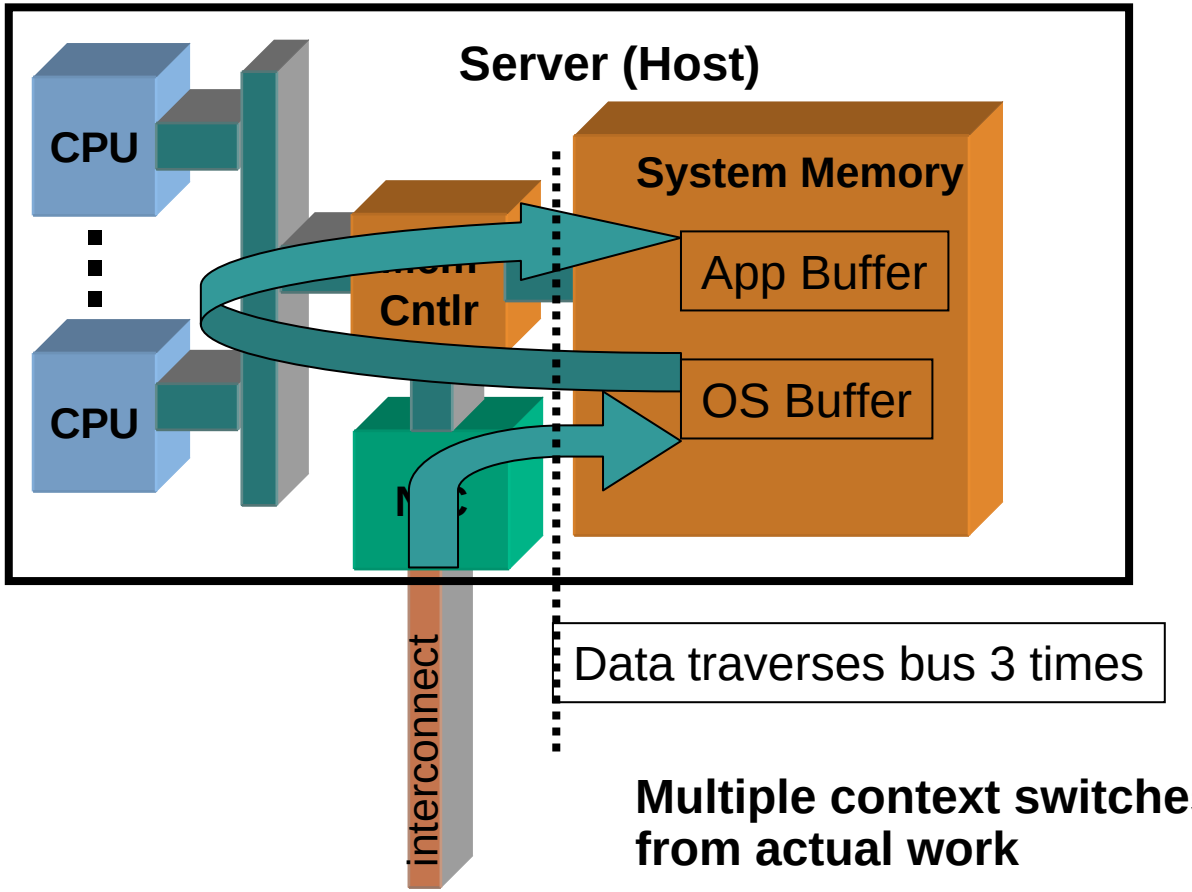
Ask SM for GID->LID mapping

Ask Host for Service info (QP)

Communicate

RDMA and Upper Layer Protocols
RDMA 和上层协议

Current NIC Architecture 目前NIC架构

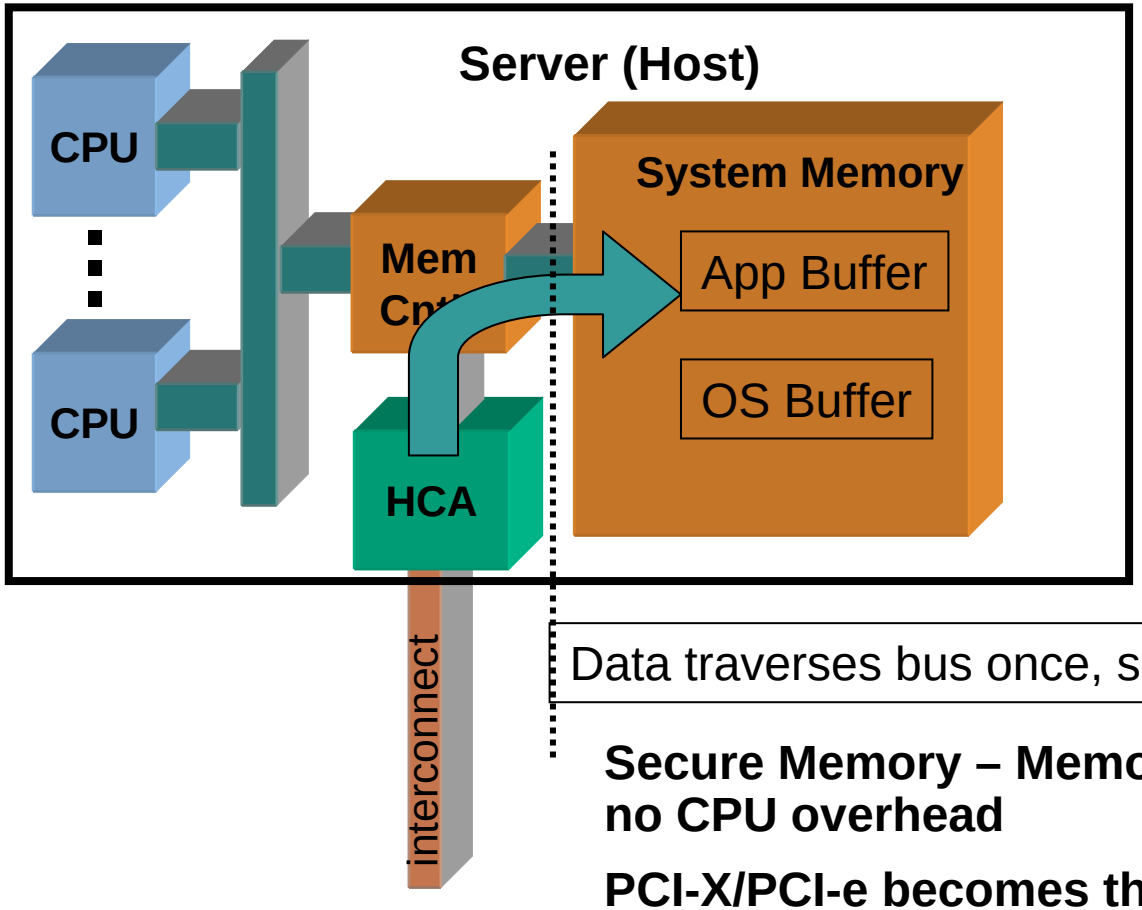


Multiple context switches robs CPU cycles from actual work

Memory bandwidth and per packet interrupts limit max throughput

OS manages end-to-end communications path

With RDMA and OS Bypass 使用RDMA旁路OS



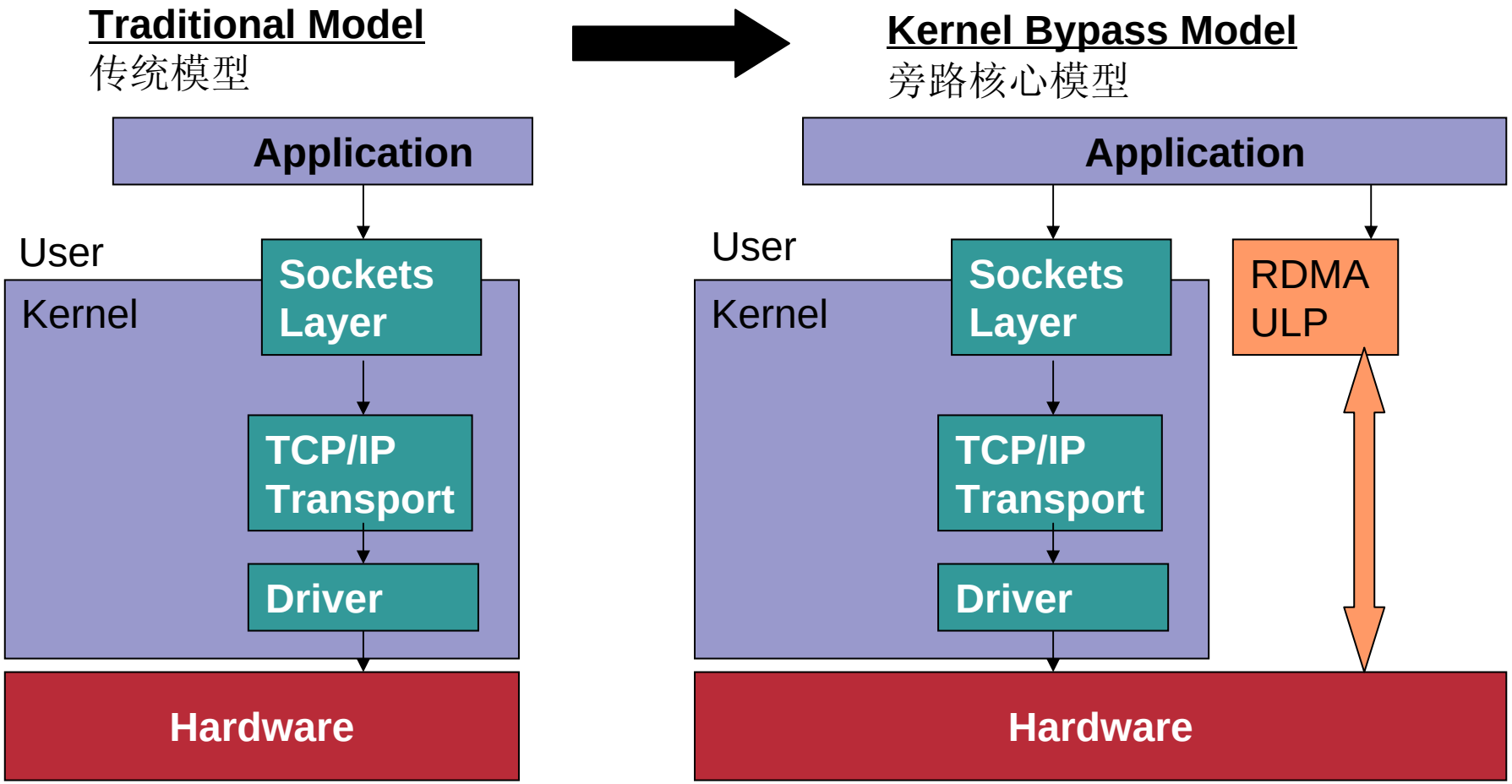
Data traverses bus once, saving CPU and memory cycles

Secure Memory – Memory transfers with no CPU overhead

PCI-X/PCI-e becomes the bottleneck for network data transmission

HCA manages remote data transmission

Kernel Bypass 旁路核心



Upper Layer Protocols 上层协议

- **Variety of software protocols to handle high speed communication over RDMA**
- （各种软件协议通过**RDMA**实现高速传送）
- **Protocols include** 协议包括

  IP-over-InfiniBand – IETF http://www.ietf.org/internet-drafts/draft-ietf-ipoib-ip-over-infiniband-09.txt

  SDP – InfiniBand Trade Association http://infinibandta.org

  SRP – ANSI T10 http://www.t10.org/ftp/t10/drafts/srp/srp-r16a.pdf

  DAPL – DAT Collaborative http://www.datcollaborative.org

  MPI – MPI Forum http://www.mpi-forum.org

IP over InfiniBand

- **IETF draft specification**
- **Leverages InfiniBand Multicast for broadcast requirements (ARP)**
- **Supports TCP, UDP, IP Multicast**

Sockets Direct Protocol (Sockets Direct协议)

- **STREAM Sockets over InfiniBand Reliable Connections**
- **TCP offload function for IB attached devices**
- **Can be used by TCP application without re-building the application**
- **Asynchronous I/O model also available with true RDMA forwarding – requires application re-write**

SCSI RDMA Protocol 协议

- **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!)**

Direct Access Provider Library 直接存取库

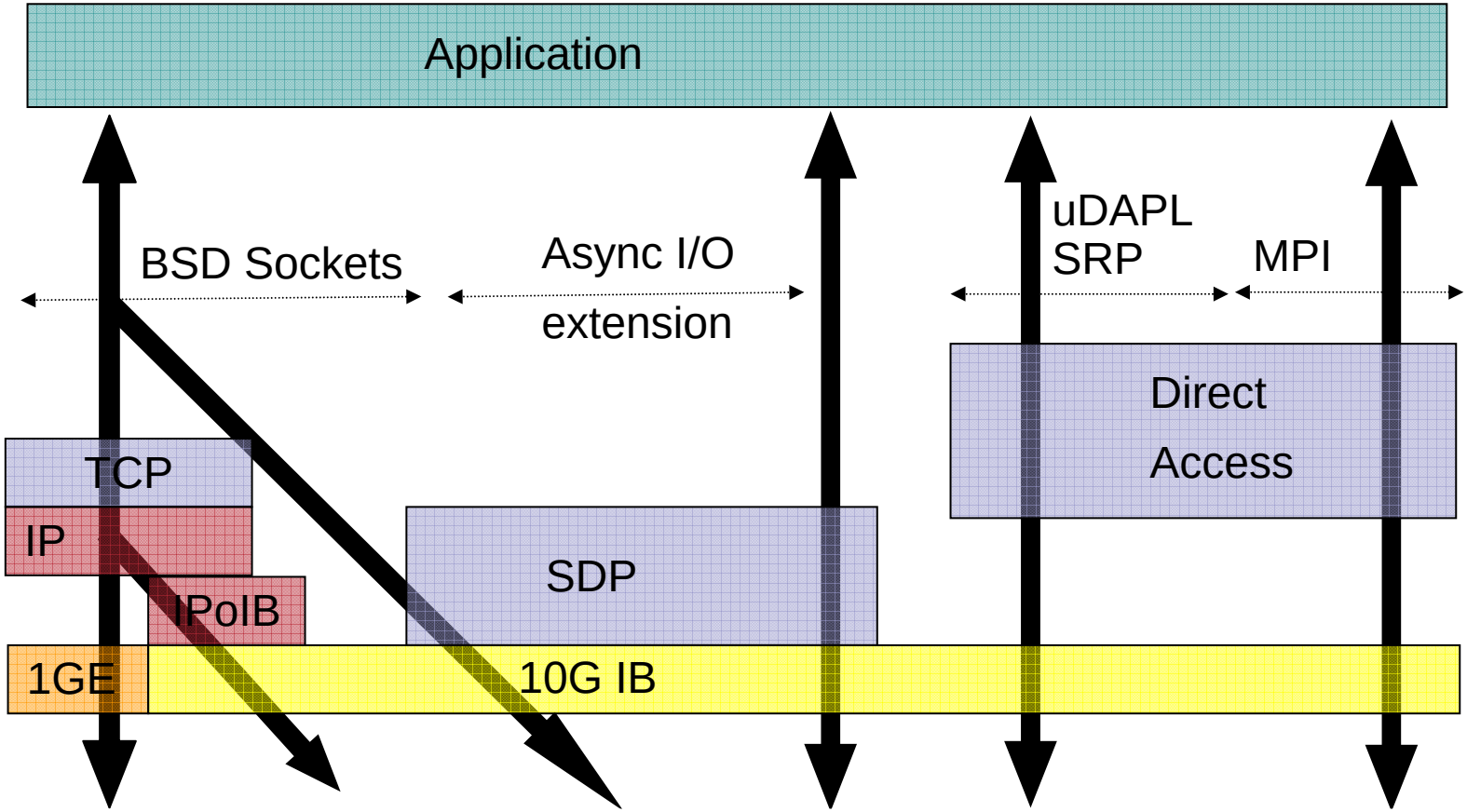
- **Two variants: User DAPL (uDAPL)/Kernel DAPL (kDAPL)**
- **RDMA semantics API**
- **Provides low level interface for application direct or kernel direct RDMA functions (memory pinning, key exchange, etc.)**

Message Passing Interface 消息传递接口

- **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/

InfiniBand Performance
InfiniBand 性能 *Measured Results*



Throughput	1 Gb/s	4.1Gb/s	4.5 Gb/s	7.9Gb/s	8Gb/s	8Gb/s
Latency	40-60 usec	30 usec	18 usec	18 usec	8 usec	3.5 usec

IB Glossary 术语

- **IB – InfiniBand Architecture (not InfinityBand)**
- **HCA – Host Channel Adapter (NIC)**
- **RDMA – Remote Direct Memory Access**
- **SM – Subnet Manager (management process)**
- **SRP – SCSI RDMA Protocol**
- **SDP – Sockets Direct Protocol**
- **TCA – Target channel Adapter (gateway)**

High Performance Computing
高性能计算

High Performance Computing Applications

- **Parallel processing applications** 并行处理的应用
  - **Closely coupled** 紧耦合
    - **Finite Element Analysis (Crash Simulation)**有限元分析
    - **Fluid Dynamics (Injection Molding)** 流体力学
  - **Loosely coupled** 松耦合
    - **Dataset searches (Terabyte->Petabyte datasets)**
    - **Monte-Carlo simulation (10,000s of repetitions)**

High Performance Computing Networks

- **Two Standards Based Technologies** 两种牛基于标准的技术
  - **Gigabit Ethernet/10 Gigabit Ethernet**
  - **InfiniBand**
- **Multiple Uses** 广泛的应用
  - **HPC interconnect HPC**内部连接
  - **Storage traffic** 数据传送和存储
  - **Load/Unload data movement** 加载/卸载数据环境
  - **Application/Systems management** 开发/系统管理

Network Types 网络类型

- **Network style is guided by application** 依赖于应用
  - **Closely coupled applications** 紧耦合应用
    - **Latency is a problem** 延迟的问题
    - **Throughput is key to resolving latency issues**
  - **Loosely coupled applications** 松耦合应用
    - **Load/unload of datasets can be a key bottleneck**
    - **Low latency for rapid response is critical**

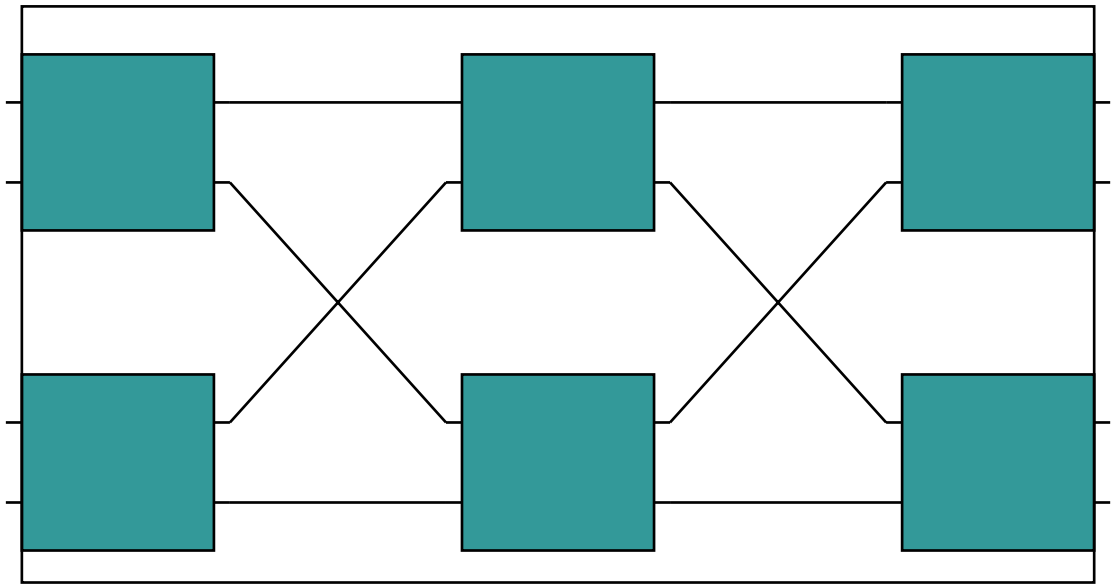
Network Architectures 网络架构

- **Multiple architectures have been used in the past**
  - **Hypercube** 立方体
  - **Mesh** 网状
  - **Taurus**
  - **Fat tree** 胖树
  - **Star** 星形
- **Today there is primarily one architecture**
  - **Fat tree**

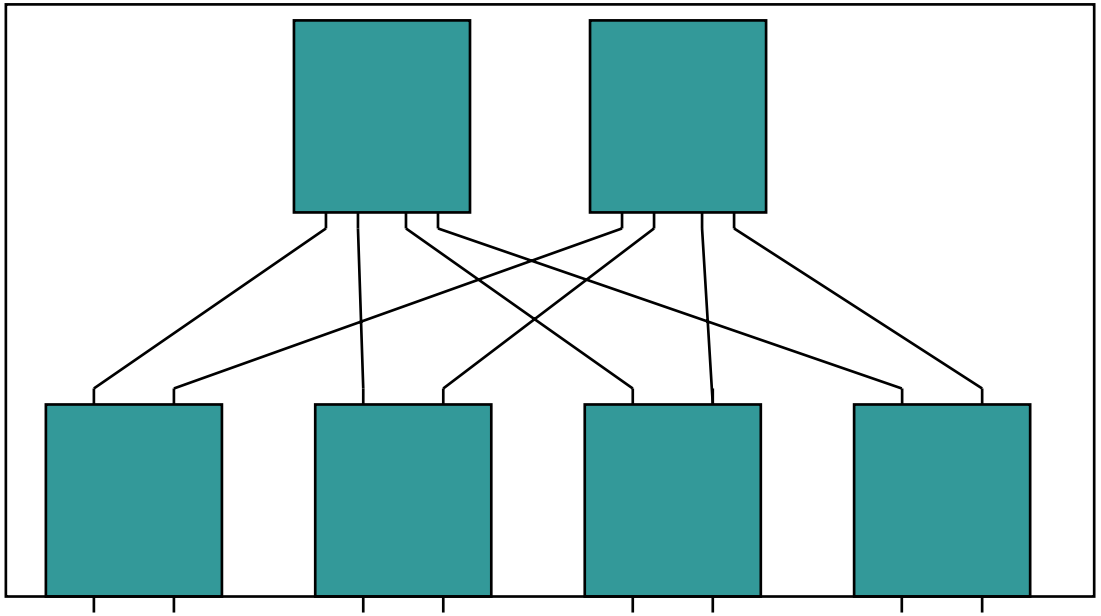
Fat-tree/CLOS architecture 胖树架构

- **Based on a non-blocking network architecture**
- **Usually based on an equivalent sized non-blocking building block switch**
- **Sometimes combined with a star architecture to provide a hybrid network**
  - **Small pools of non-blocking performance combined to provide a larger cost-effective fabric**

Building a Fat Tree 胖树的构成



Building a Fat Tree



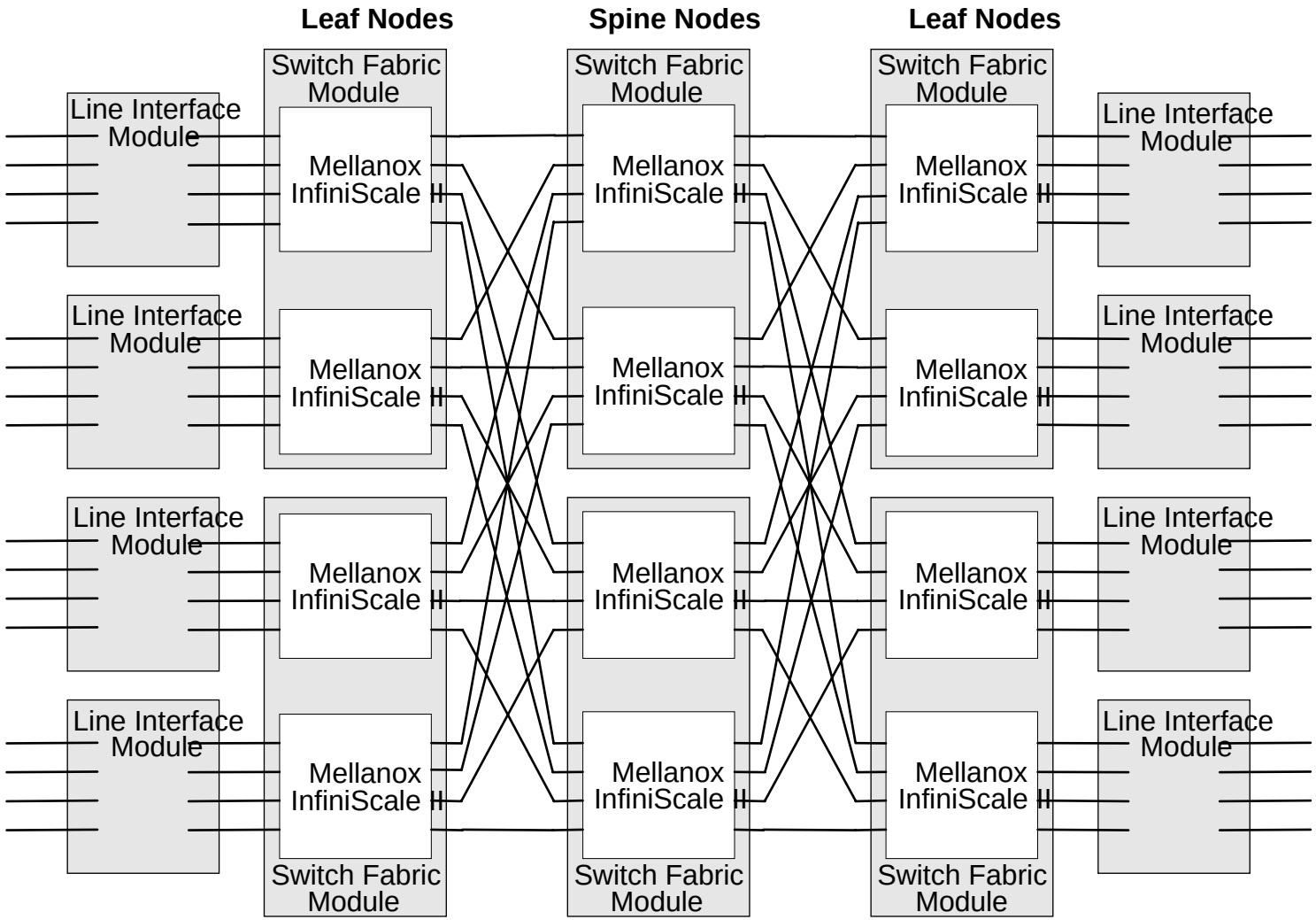
Core
核心
Spine

Leaf
边缘
Edge

Large switch architecture 大型交换机架构

- **Most IB based monolithic switches are based on the Fat Tree architecture**
- **Cisco SFS 7008 is an example of one of these systems**
- **Based on a 24 port non-blocking core switch component**
- **Total of 12 switches make up the entire system**

Cisco SFS 7008 Internal Architecture内部架构

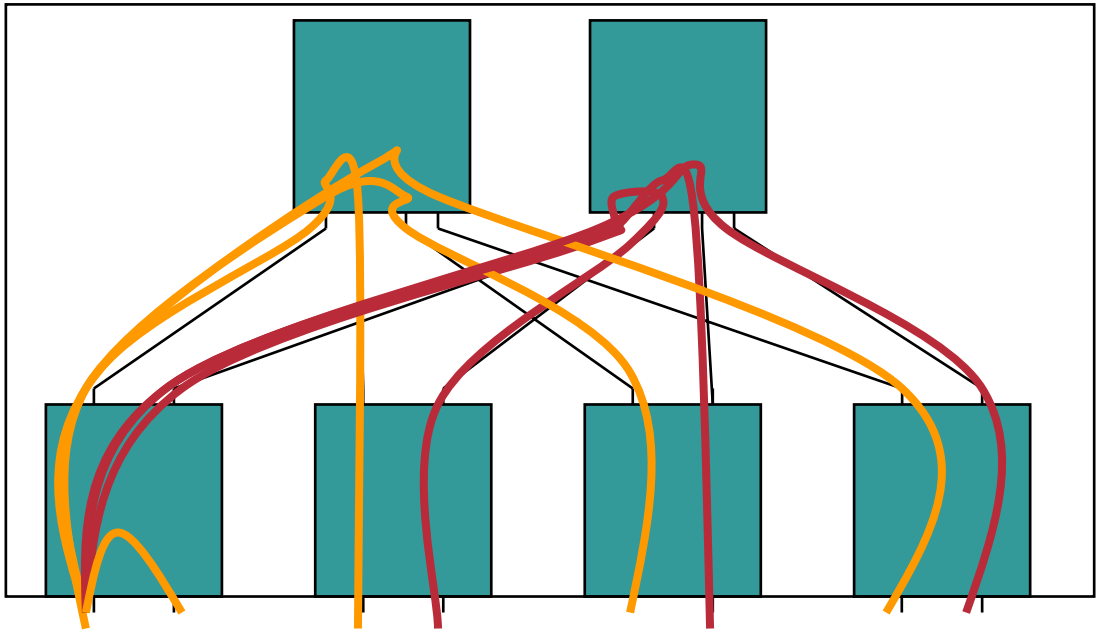


Note: all links are 12x

IB Routing for HPC　路由

- **Subnet Manager provides global route engine for entire IB Fabric**子网管理提供整个**IB** 网的整个路有引擎
- **Shortest Path First routing** 短路径优先
- **Round Robin load balancing** 负载均衡
- **Static routes** 静态路由

IB Routing for HPC



HPC Storage Problem 存储的问题

- **As the job grows so does the storage**
  - **Clusters growing to 1000 nodes or more** 集群超过**1000**节点
  - **Storage growing into the PetaByte range** 存储超过**PB**数据
- **The storage problem**
  - **Load/Unload a large dataset** 加载/卸载大量数据
  - **Get shared access to large datasets on the fly**
  - 随时共享存取大量的数据
  - **Performance is an issue: Need multi-GigaByte/s throughput**
  - 性能问题：

Current Solutions 当前的解决方案

- **NFS** 老的标准，不能很好的扩展
  - **Old standby: Doesn’t scale well (single server)**
- **Current commercial cluster file systems**
- （当前商用集群文件系统）
  - **Designed for multiple reader/writer situations**
  - **Don’t scale beyond 10s of nodes**
  - **Don’t necessarily manage the throughput problem**
  - **Need to build out separate Fibre Channel fabric**

HPC Storage Solutions 存储解决方案

- **Next Generation Cluster File-systems**
- （下一代集群文件系统）

  Based on new file-systems or modification of old (nfs or iSCSI) 基于新的文件系统或者对老的进行修改

  Split the data across multiple file service hosts

  将数据存取从多个服务器上分开

  Either act as a RAIF (Redundant Array of Inexpensive File-servers), or allow clients to access any one file server, while everyone shares the same back-end storage devices

  当共享的同一后端数据设备访问时，或是一个**RAIF**或允许客户端访问一个服务器，

InfiniBand Storage Acceleration 存储加速

- **Further accelerate storage access**
- （将来加速数据的存取）

  IB based forwarding over SDP or DAPL can provide CPU offload and increased throughput 基于IB

  Low latency improves overall throughput and performance

  低延迟

  High bandwidth improves load/unload times dramatically

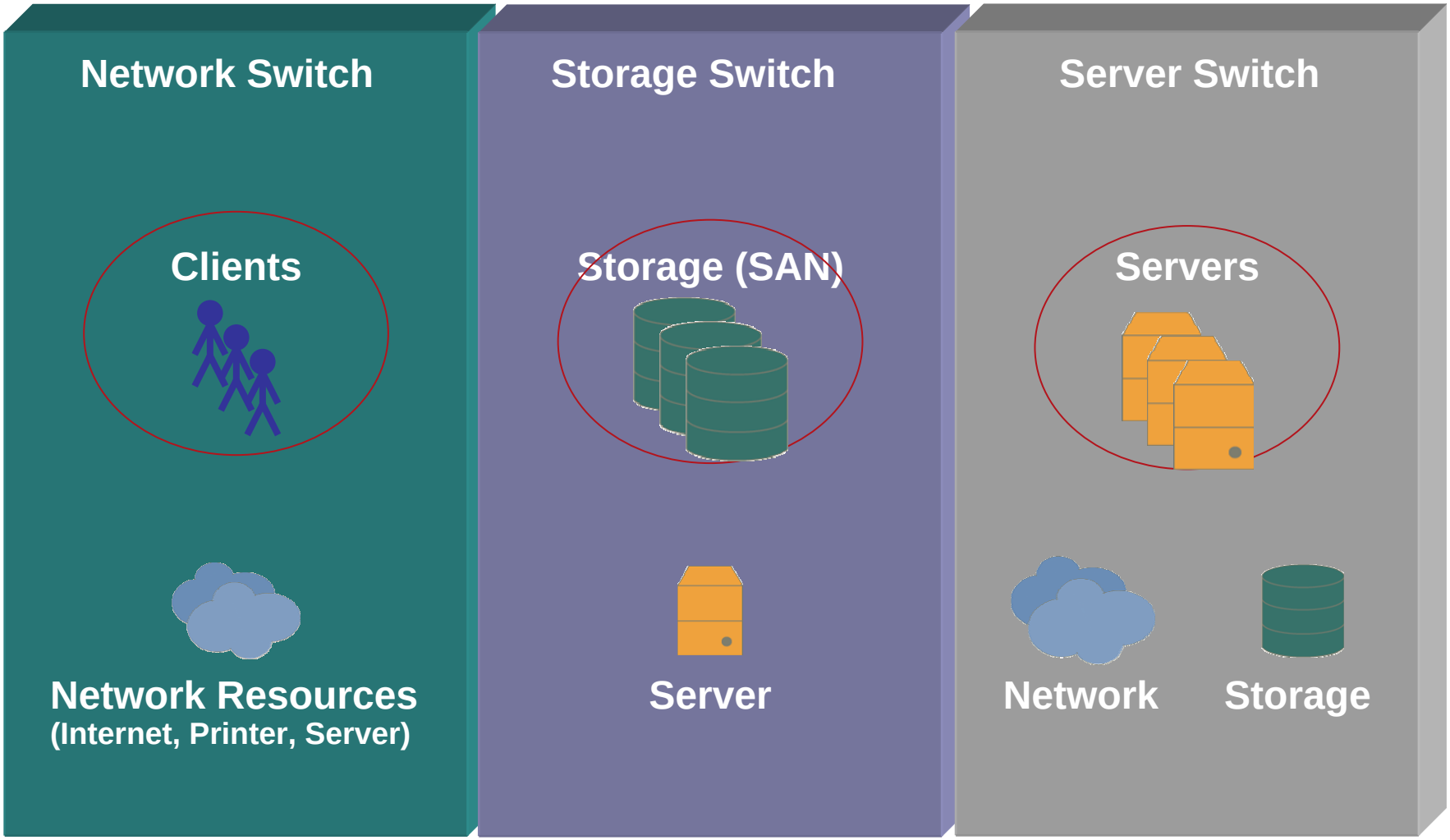
  高带宽

HPC Glossary 术语

- **Fat-tree – Non-blocking switch architecture (aka CLOS) 胖树 无阻塞Bisectional bandwidth – the total system bandwidth across the middle of the network 带宽**
- **Non-blocking – full host bandwidth all-2-all communication 胖树**

A New Infrastructure Category
一种全新的基础架构

A New Category of Data Center Infrastructure-
The Server Fabric Switch 一种全新的数据中心基础架构



What Makes The Server Fabric Switch Different?
服务器交换机有何不同?

High Performance Server-to-Server Interconnect

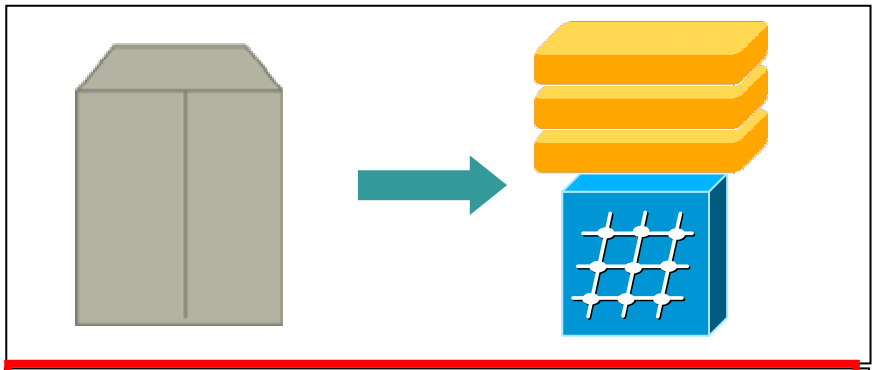
Virtualization (I/O, Storage, and CPU)

Policy-Based Dynamic Resource Mapping

Performance and Control

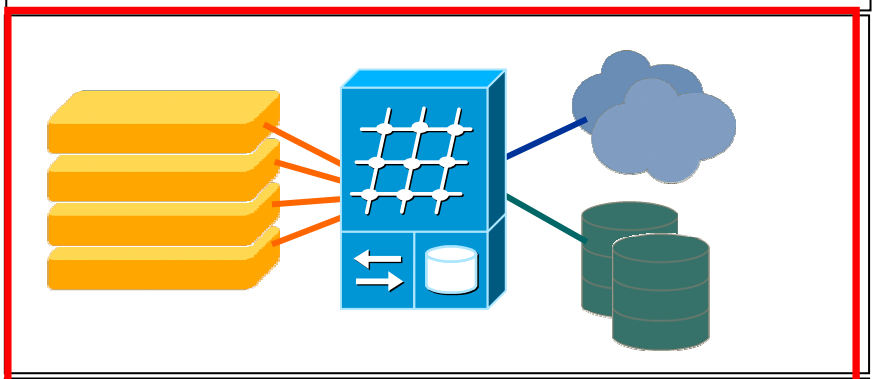
Server Fabric Switch Applications
Why Performance and Control?服务器网络交换机应用

Server Clustering



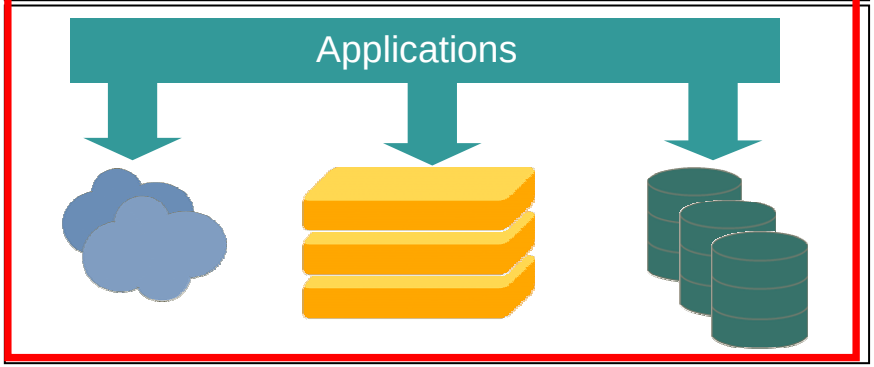
–High Performance Computing (HPC)
–"Enterprise-Class" HPC
–Database Scalability

I/O Virtualization



–I/O Consolidation
–I/O Aggregation
–Server Consolidation

Utility or Grid Computing



–Application Provisioning
–Server Re-purposing
–Server Migration

I/O Virtualization I/O 虚拟化

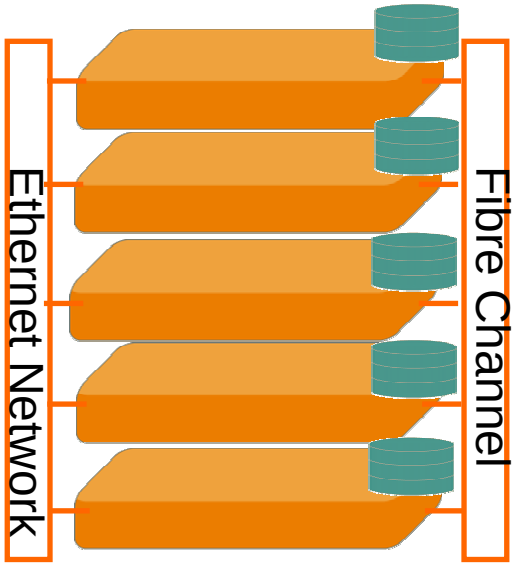
The Evolution of I/O Virtualization I/O虚拟化革命

SMP

Pro: Single managed entity, fast backplane

Con: Expensive, Proprietary server + backplane

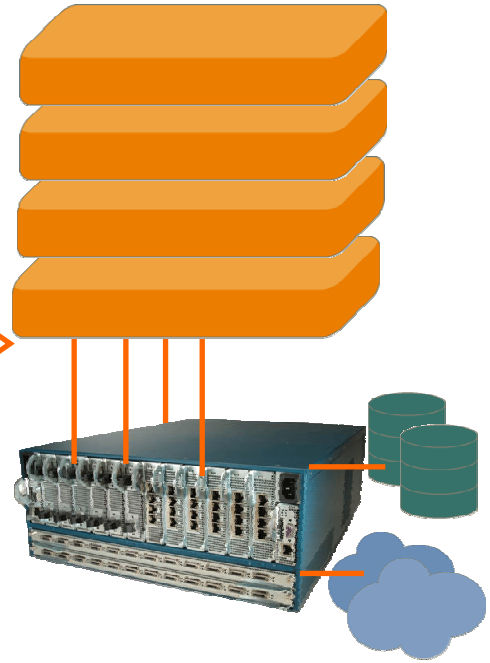
Dis-aggregation



Pro: Standard servers, inexpensive

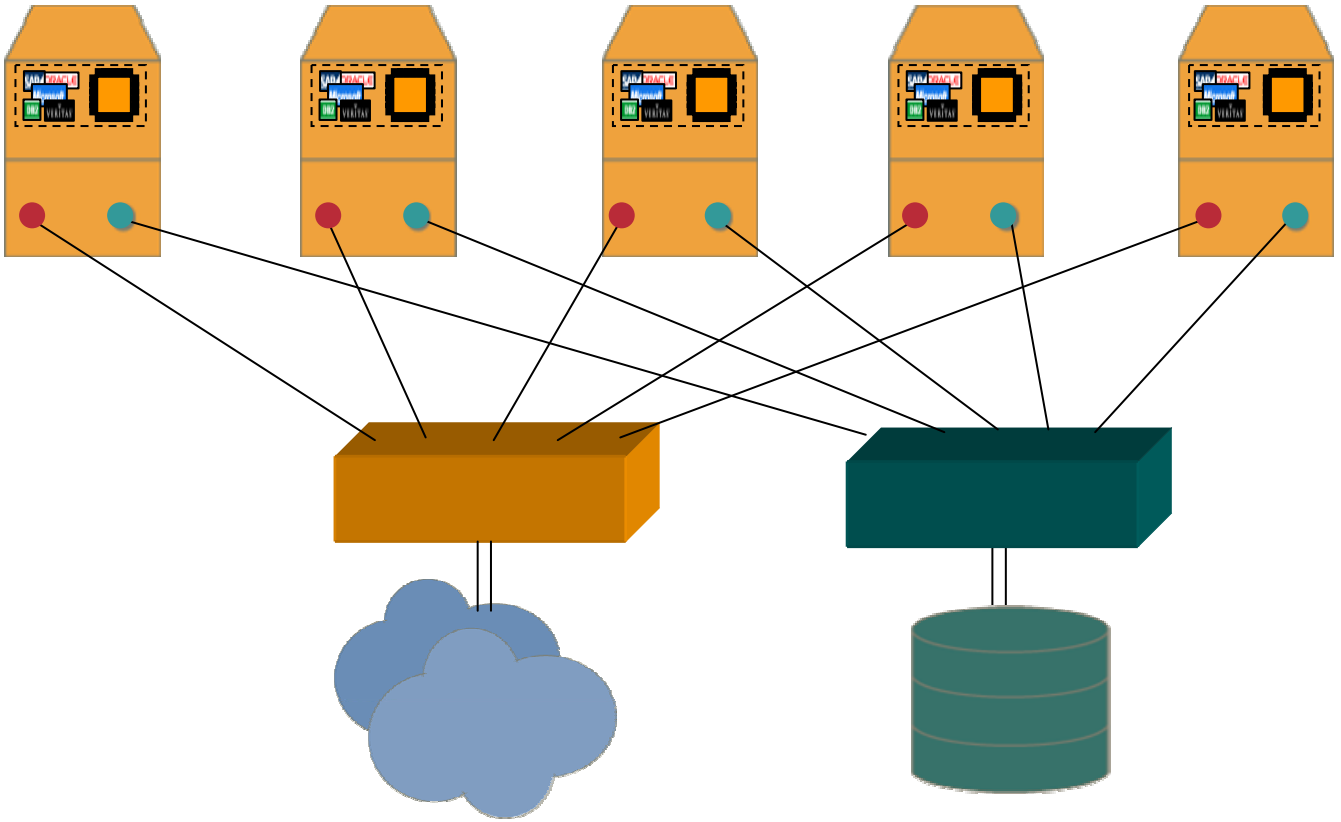
Con: Lots of managed components, low-performing interconnect

Virtualization

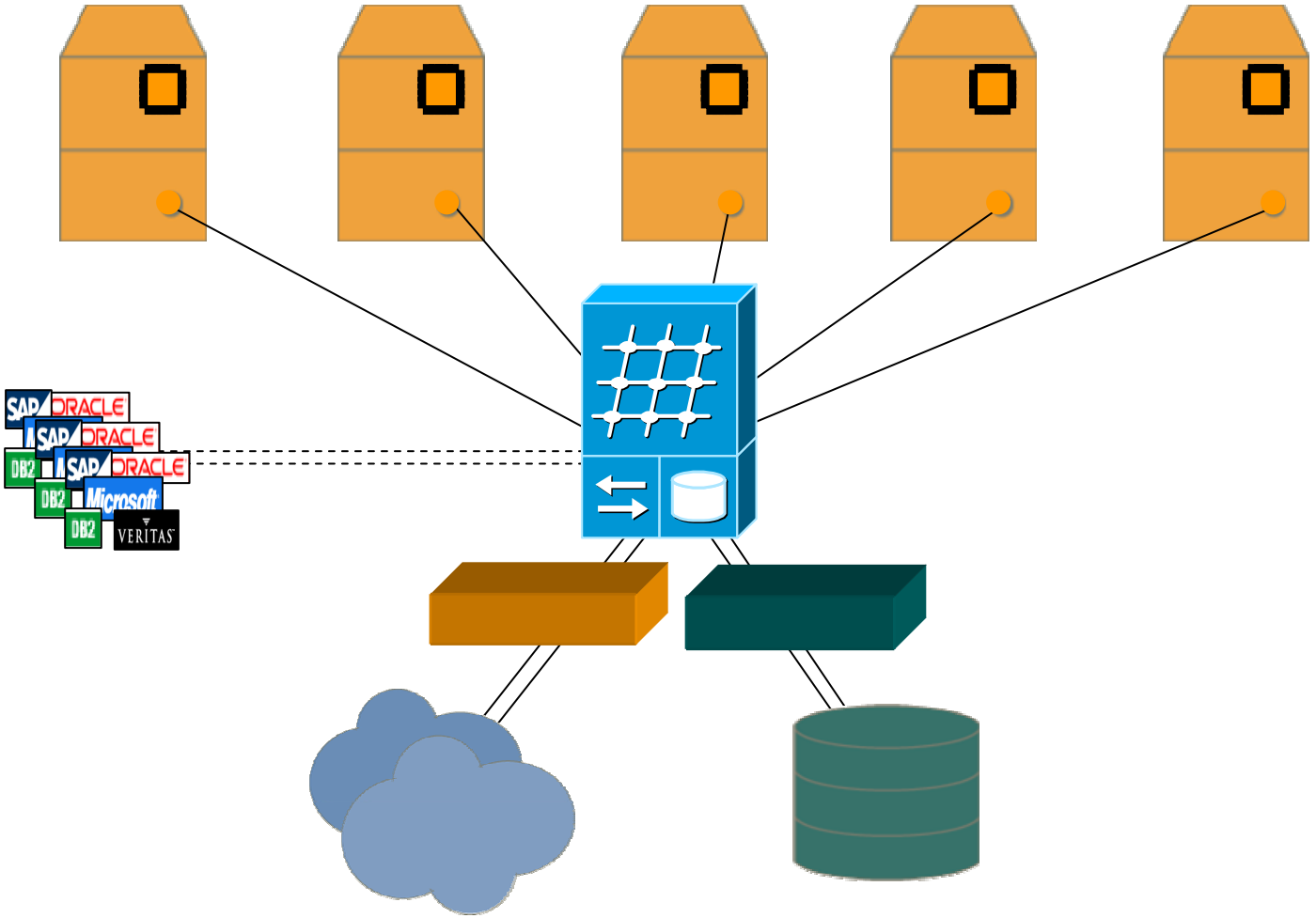


Pro: Reduced # of managed components, virtual I/O, *fast* standards backplane

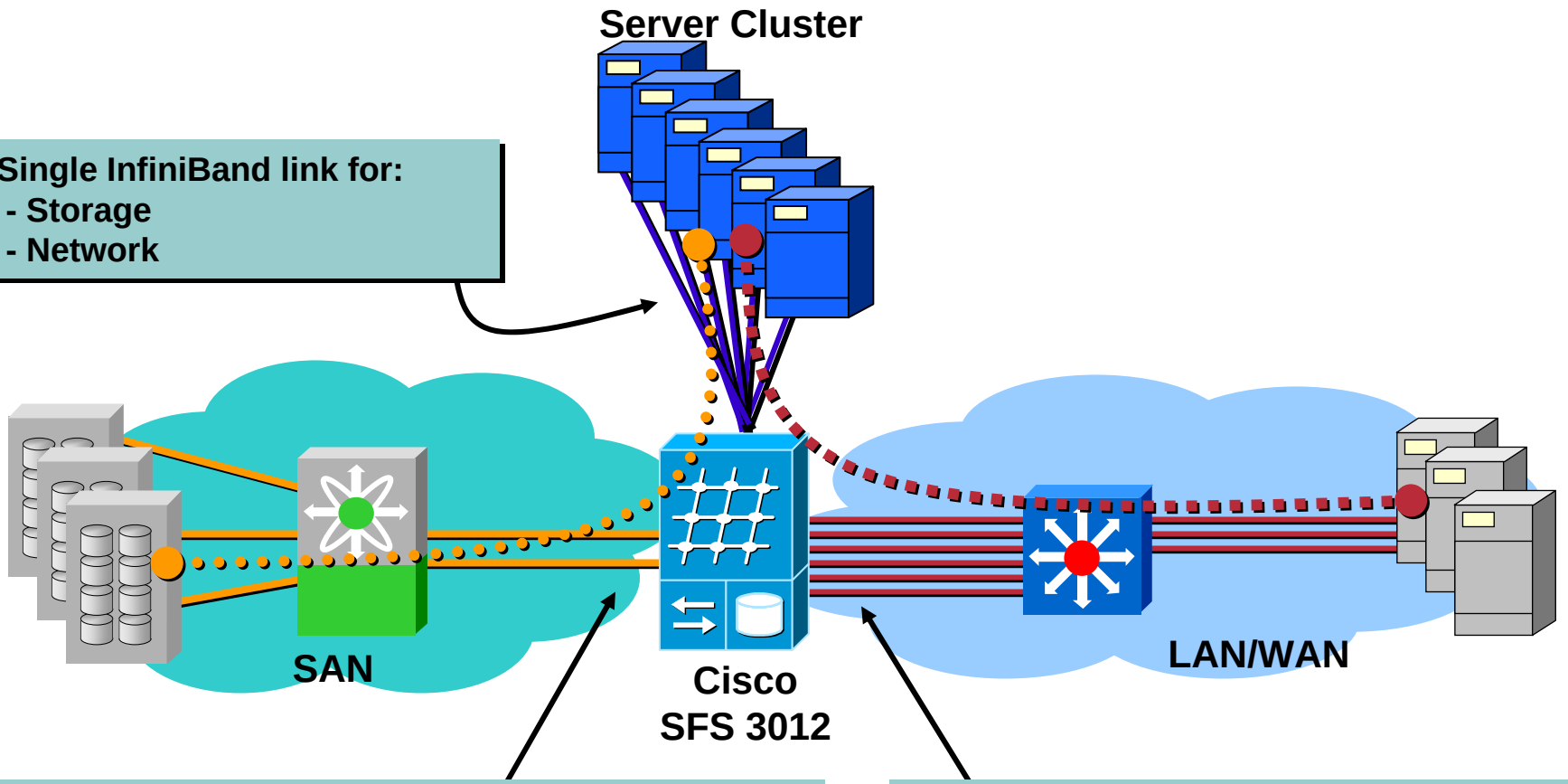
Evolution of the Data Center 数据中心演变
Network and Storage Virtualization 网络和存储虚拟化



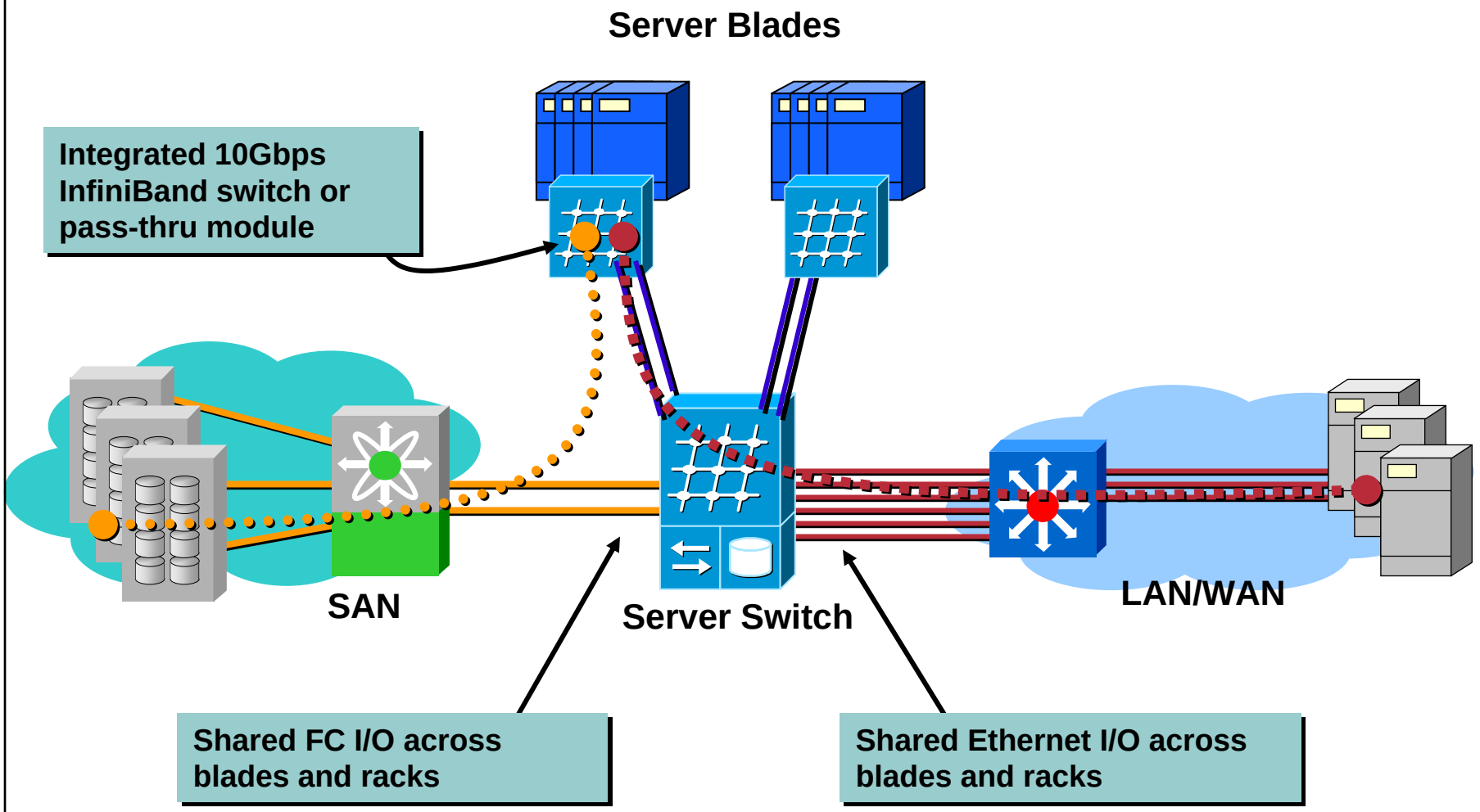
Evolution of the Data Center
Server Virtualization - The Server Switch
服务器交换机 –服务器虚拟化



Virtual I/O for Network and Storage
Unified "wire-once" fabric　为网络和存储的虚拟化I/O

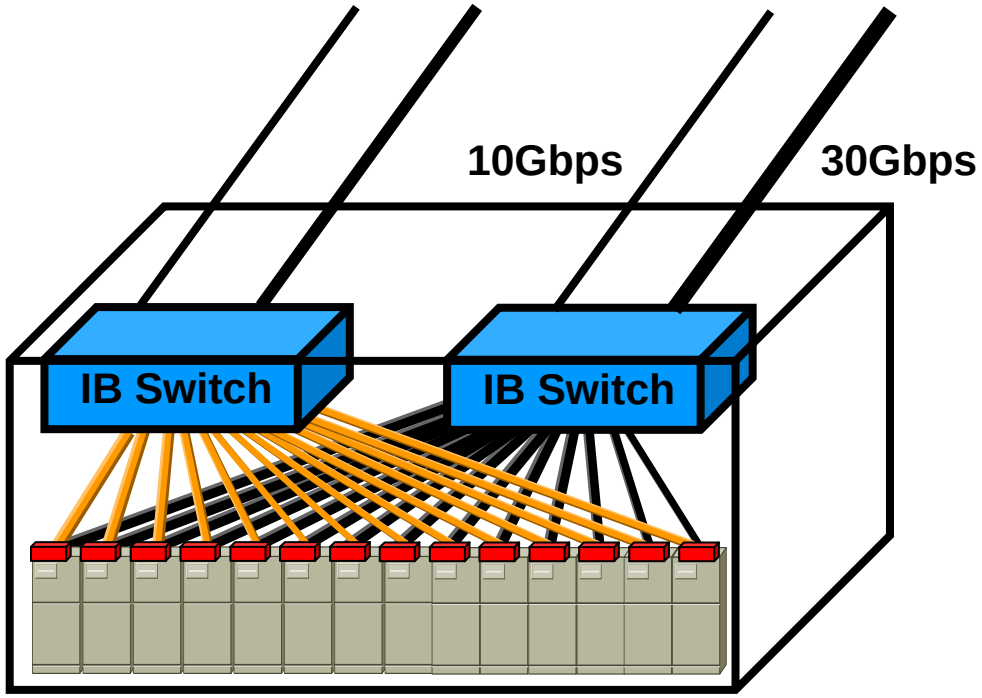


Virtual I/O for Blade Servers
Eliminating I/O Bottlenecks 刀片服务器的I/O虚拟化



Integrated InfiniBand for Blade Servers
Create “wire-once” fabric 刀片服务器集成 InfiniBand

- Integrated 10Gbps InfiniBand switches provide unified “wire-once” fabric
- Optimize density, cooling, space, and cable management.
- Virtual I/O provides shared Ethernet and Fibre Channel ports across blades and racks
- Option of integrated InfiniBand switch (ex: IBM BC) or pass-thru module (ex: Dell 1855)



Blade Chassis with InfiniBand Switches

HCA

Virtual I/O: How it Works
虚拟I/O 是如何工作的

Transparent Topology Architecture 透明的拓扑架构

- **IP Communications IP 通讯**

  Inside Fabric: IP over InfiniBand (IPoIB) enables transparent communications for any IP-based applications.

  Outside Fabric: InfiniBand-to-Ethernet Gateways provide transparent access to existing IP Ethernet infrastructure.

- **Fibre Channel Storage FC 存储**

  Inside Fabric: SCSI RDMA Protocol (SRP) enables SCSI over InfiniBand. 网内

  Outside Fabric: InfiniBand-to-Fibre Channel Gateways provide transparent access to existing SAN infrastructure. 网外

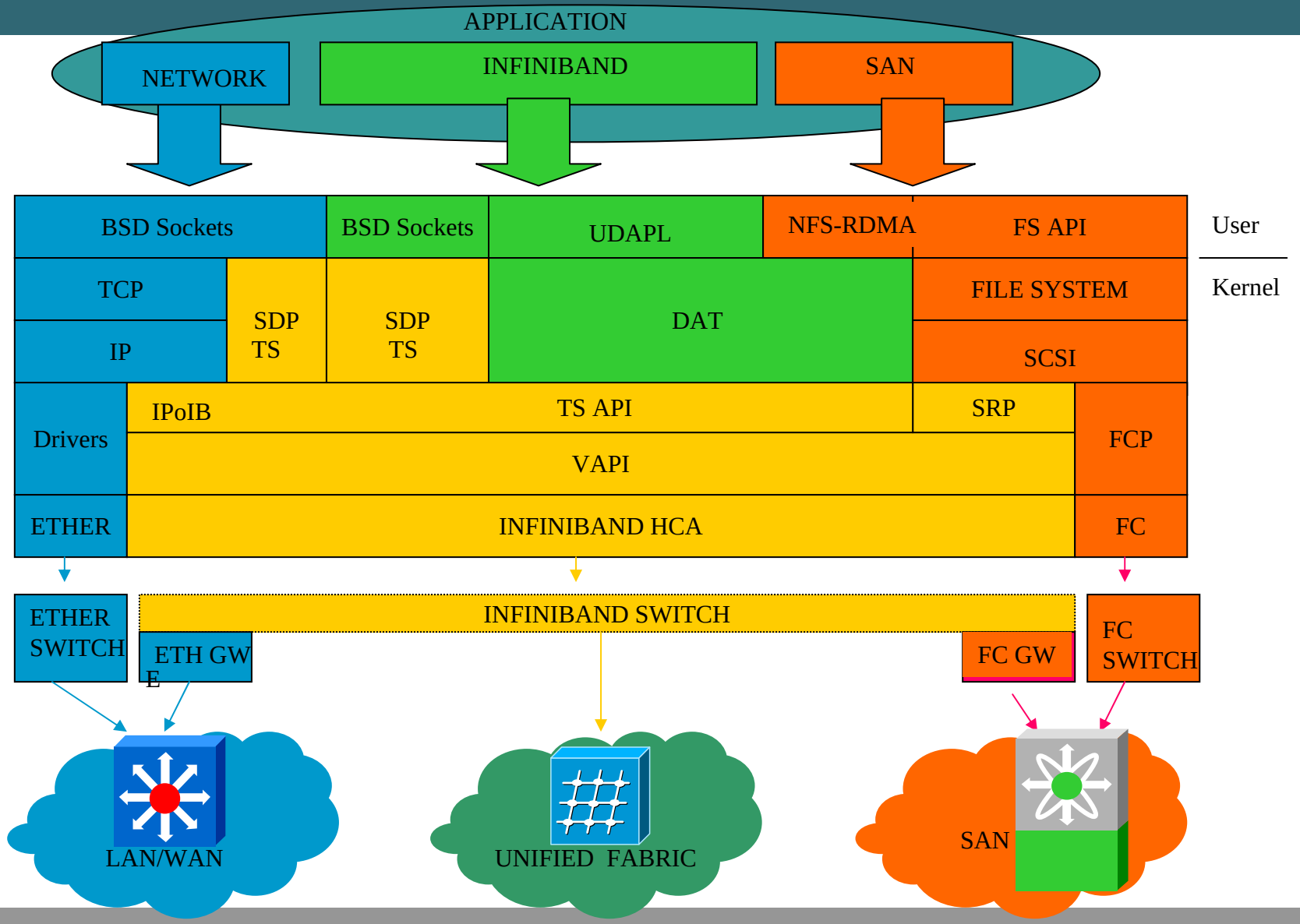
- **Cluster Communications 集群通讯**

  Inside Fabric: Remote DMA protocols for transparent performance enhancements.

InfiniBand Protocol Summary IB协议摘要

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 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.

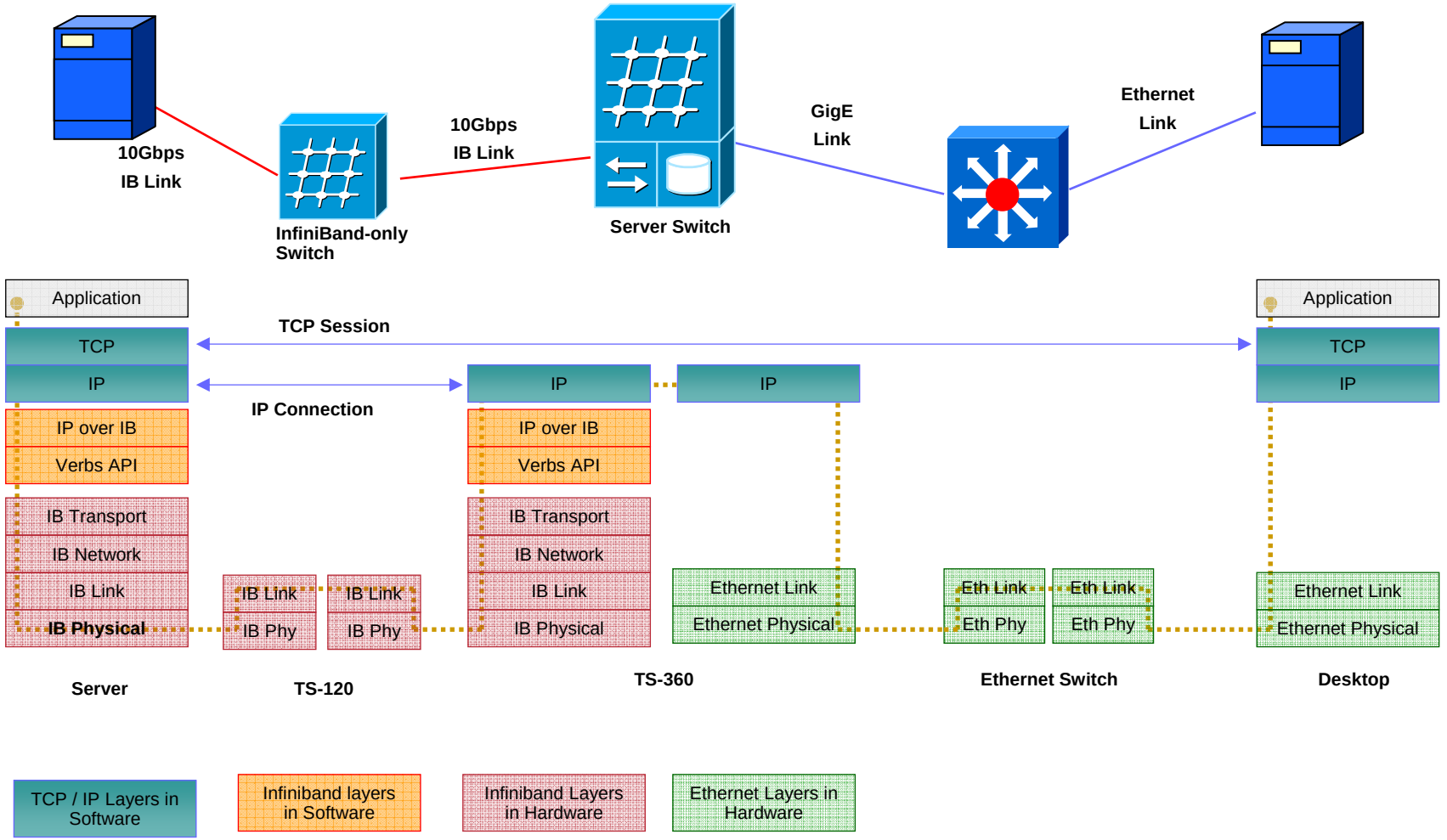
The InfiniBand Driver Architecture 驱动架构



IP over InfiniBand

- **Transmission of IP over Infiniband** 通过InfiniBand 传送IP
  - **Use IB as a link layer for IP** **Use Infiniband UD transport mode**
  - **Define data link and link layer address**
  - **Encapsulation for ARP, IPv4 and IPv6**
  - **Address resolution** **Transport IP multicast over IB**
- **Provides highest level of application compatibility.**
- 提供最上层应用兼容
- **Applications do not need to be re-written or re-compiled**
- 应用无需重新编写和编译
- **Standard IP utilities and applications work as usual:**
- 标准的IP程序和以前一样工作
  - **Ifconfig, ping, telnet, File sharing (NFS, CIFS); Login access (ssh, telent, etc); Cluster heartbeat**
  - **DHCP over IB** **IP over InfiniBand MIB**

How IP over InfiniBand works 如何工作



* Notes: Uses standard Berkeley TCP/IP libraries

Sockets Direct Protocol

- **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)**

InfiniBand-to-Ethernet Gateway Overview

- **Ensures seamless integration with IP-based applications.**
- **Act like L2 bridge between IB and Ethernet**
- **Bridge group is the main forwarding entity**
- **Bridge group has two bridge ports Ethernet and IPoIB**
- **Bridge group bridges one VLAN to one IB partition**
- **Ethernet bridge port can be tagged or untagged**
- **Ethernet bridge port can aggregate up to 6 ports**

InfiniBand-to-Ethernet Gateway Features功能

- IP-Only protocols
- 802.1Q VLAN support
- Link aggregation
- IPv4 multicast support
- Loop protection
- Ethernet jumbo frames up to 9k
- IP fragmentation
- High availability

VLAN Support　支持VLAN

- Standard 802.1Q VLAN support
- Static port based VLAN’s
- One VLAN is mapped to one IB partition
- Up to 32 VLAN’s per gateway
- Full range of VLAN ID’s
- Tagged and untagged ports

Link Aggregation 连接聚合

- Standard 802.3ad link aggregation
- Static link aggregation group configuration
- One link assigned to one or multiple bridge groups
- Up to 6 link aggregation group per gateway
- Seven different frame distribution types
- Link aggregation group can carry up to 32 VLAN’s
- Link aggregation group can not span multiple gateway

Multicast Support

- InfiniBand switches support true IB Multicast in Hardware
- InfiniBand-to-Ethernet gateways support multicast in hardware.
- IB Switches use two types of Forwarding Tables:
  - Linear Forwarding Table (1 to 1 - Message In/Out)
  - Multicast Forwarding Table (1 to Many - Message In/Out)
- IB Partitions can be used to Segregate Traffic Domains
- Hardware Multicast Support means:
  - No Host Overhead for sending Multicast Messages
  - No Appreciable Latency between 1st Message & Last Message
  - No Superfluous Network Traffic
  - Multiple IB Switches in a Fabric Effectively Creates a Parallelized Multicast Delivery Mechanism (Scales Very Large, Very Fast)

Jumbo Frames / Fragmentation

- **Up to 9k Ethernet MTU**
- **Up to 2044 bytes IPoIB MTU**
- **Ethernet frames larger than 2044 bytes are fragmented**
- **No fragmentation for IB frames**
- **Used mainly by UD based protocols**

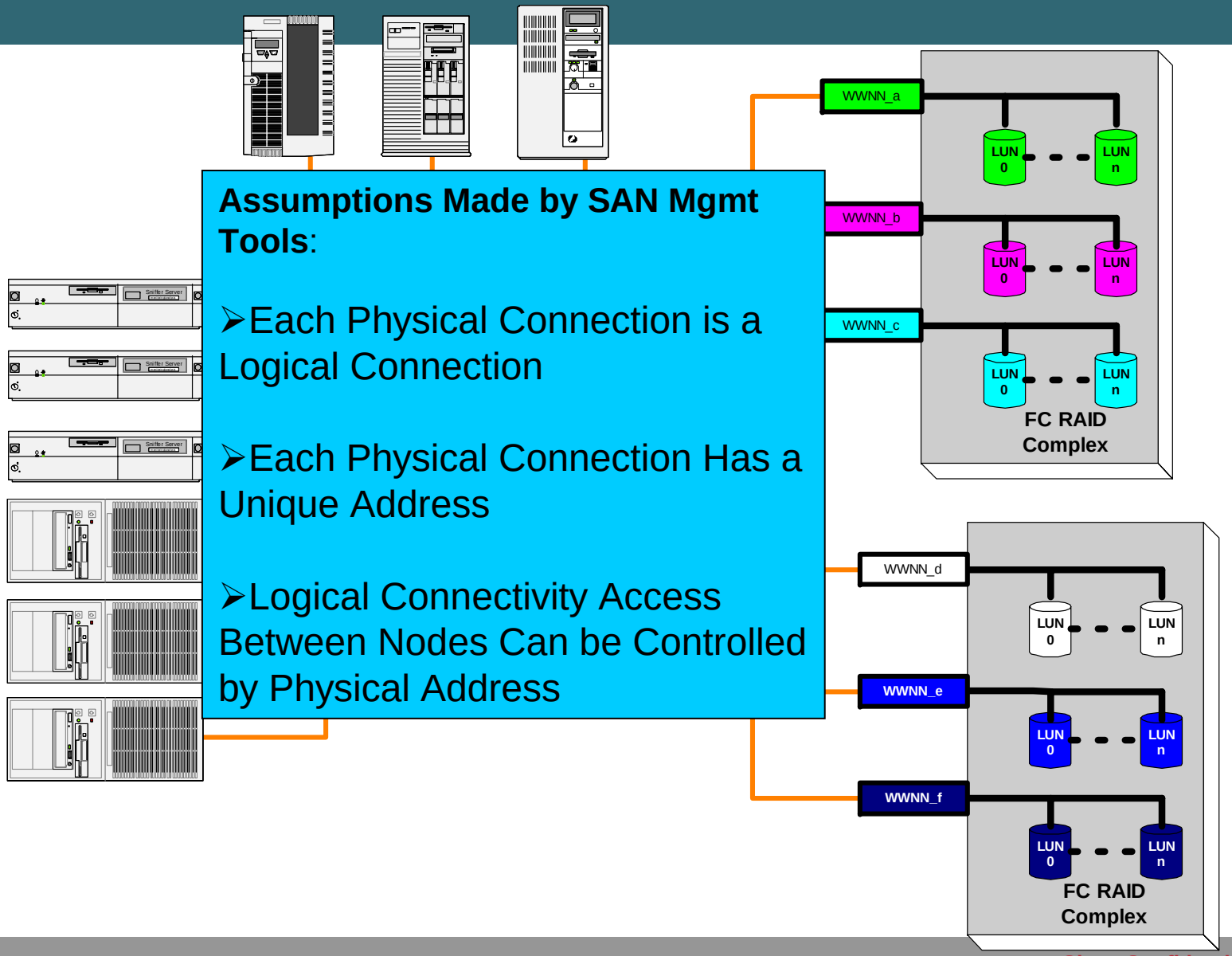
High Availability

- Bridge group based redundancy
- Bridge group member of a redundancy group
- One redundancy group cover one VLAN
- Active – passive and active – active modes
- Automatic fail-over and fail-back
- Uses gratuitous ARP to redirect traffic
- Redundancy group can span multiple chassis
- Proprietary redundancy protocols for address distribution and bridge group election

InfiniBand-to-Fibre Channel Gateway

- **Ensures seamless integration with important SAN tools.**
  - **Fabric-based Zoning**
  - **LUN-based access controls**
  - **Storage and host-based HA and load balancing tools**
- **Creates SAN network addresses on InfiniBand.**
  - **SAN Management Tools must “see” each node.**
  - **Creates “talk-through” mode with virtual WWNNs per server.**
- **Enables SAN Interoperability with InfiniBand.**
  - **Appears as virtual NL-Port, N-Port, E-Port.**
  - **Proven interoperability with Cisco MDS, Brocade, McData, Qlogic, EMC, IBM, Hitachi, and more.**

Typical SAN Today

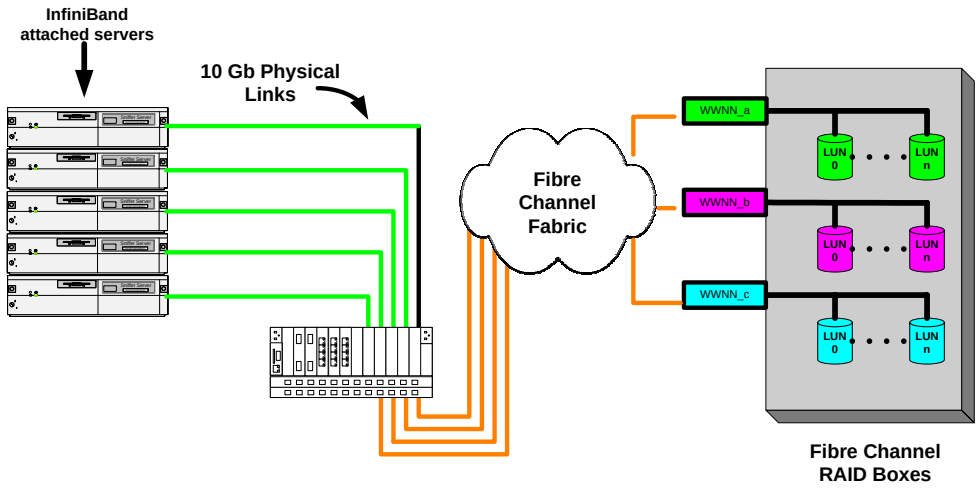


Assumptions Made by SAN Mgmt Tools:

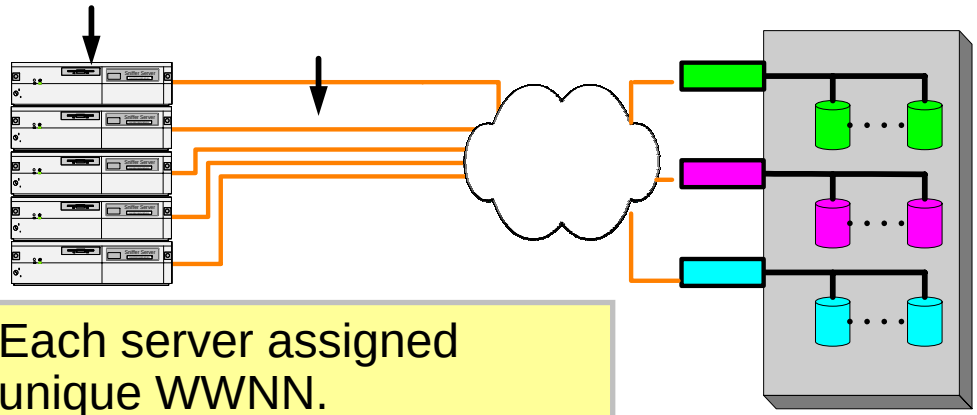
- Each Physical Connection is a Logical Connection
- Each Physical Connection Has a Unique Address
- Logical Connectivity Access Between Nodes Can be Controlled by Physical Address

Physical vs. Logical View 物理Vs. 逻辑视图

Physical View



Logical View



- Each server assigned unique WWNN.
- Appears as direct-attached N_Port.

Topology Transparency: How it Works

- **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**

Server Virtualization
服务器虚拟化

Three Categories of Server Virtualization
服务器虚拟化三种类型

- **Virtual Machine: Splits a servers into independent virtual servers.　虚拟机**

  VMWare*, *XEN*, *MSFT

  Main value is higher server utilization.

- **Virtual SMP: Combines servers together into a single managed powered entity.　虚拟SMP**

  Virtual Iron*, *Qlusters

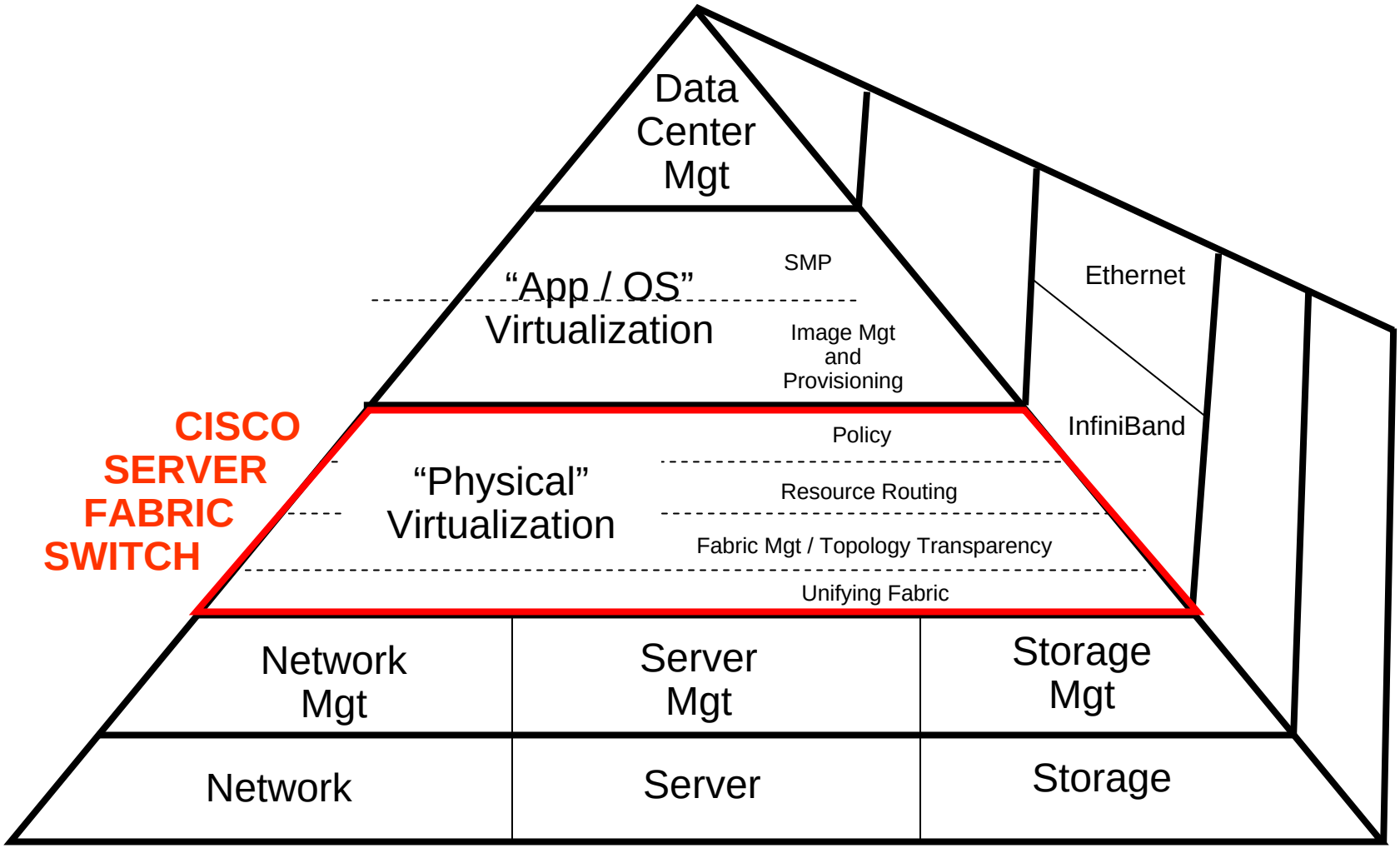
  Main value is scaling mission critical apps on commodity HW.

- **Physical Server Virtualization: Makes servers stateless by moving server identity into the network, including storage and I/O subsystem.　物理服务器虚拟化**

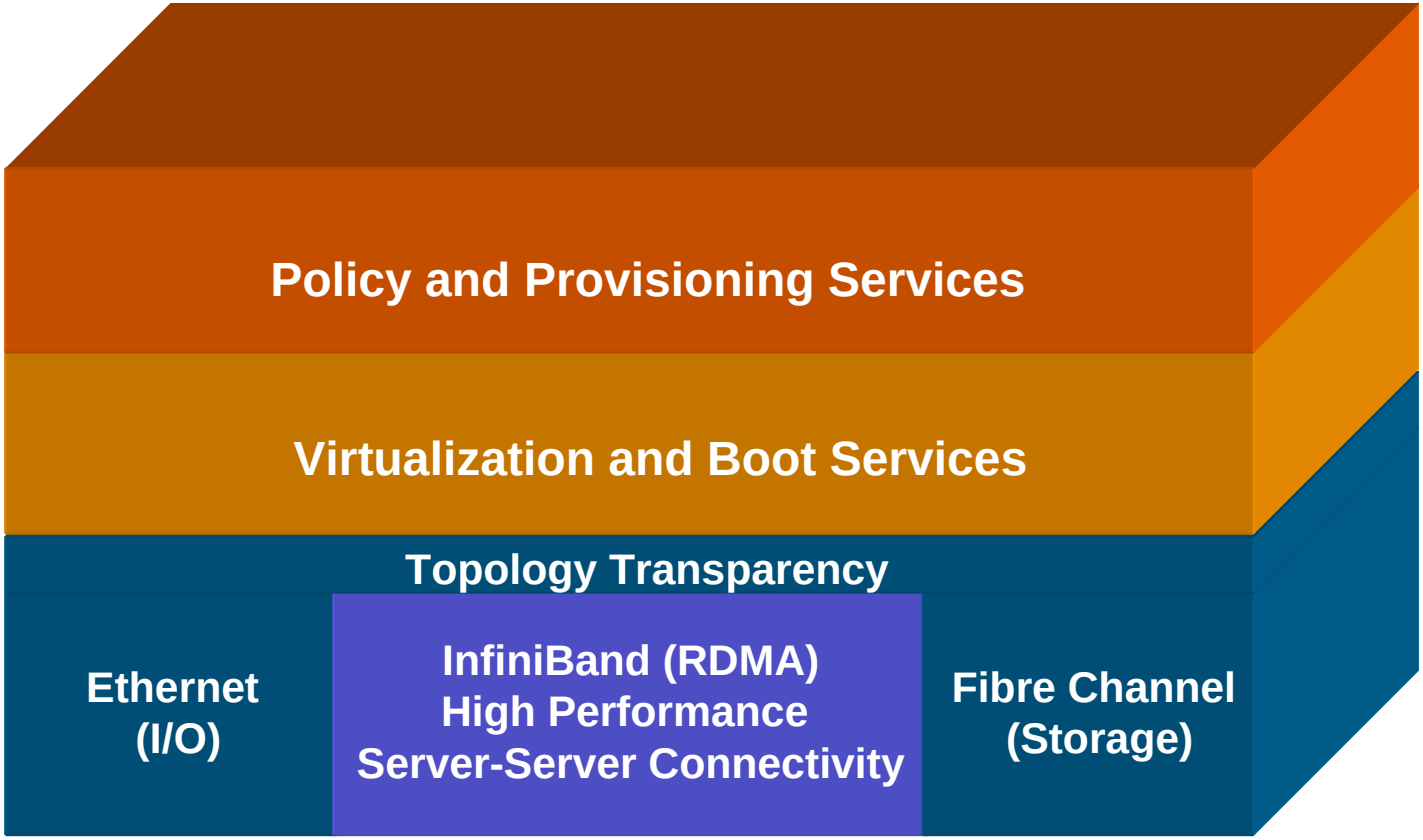
  Cisco VFrame™*, *Egenera

  Main value is making infrastructure change easier in heterogeneous environment.

Virtualization Ecosystem



VFrame Server Virtualization Framework
Building Blocks



VFrame™

- **Software suite that makes the Server Switch** *programmable*
- **Three main components**

VFrame™ Embedded System Logic

Policy ingestion, interpretation, and enforcement at the server switch

VFrame™ APIs (and SDK)

Allows 3rd party (End-user Customers, Software Partners, System Vendor OEMs) management and provision tools to program and manage the server switch fabric Software Partners

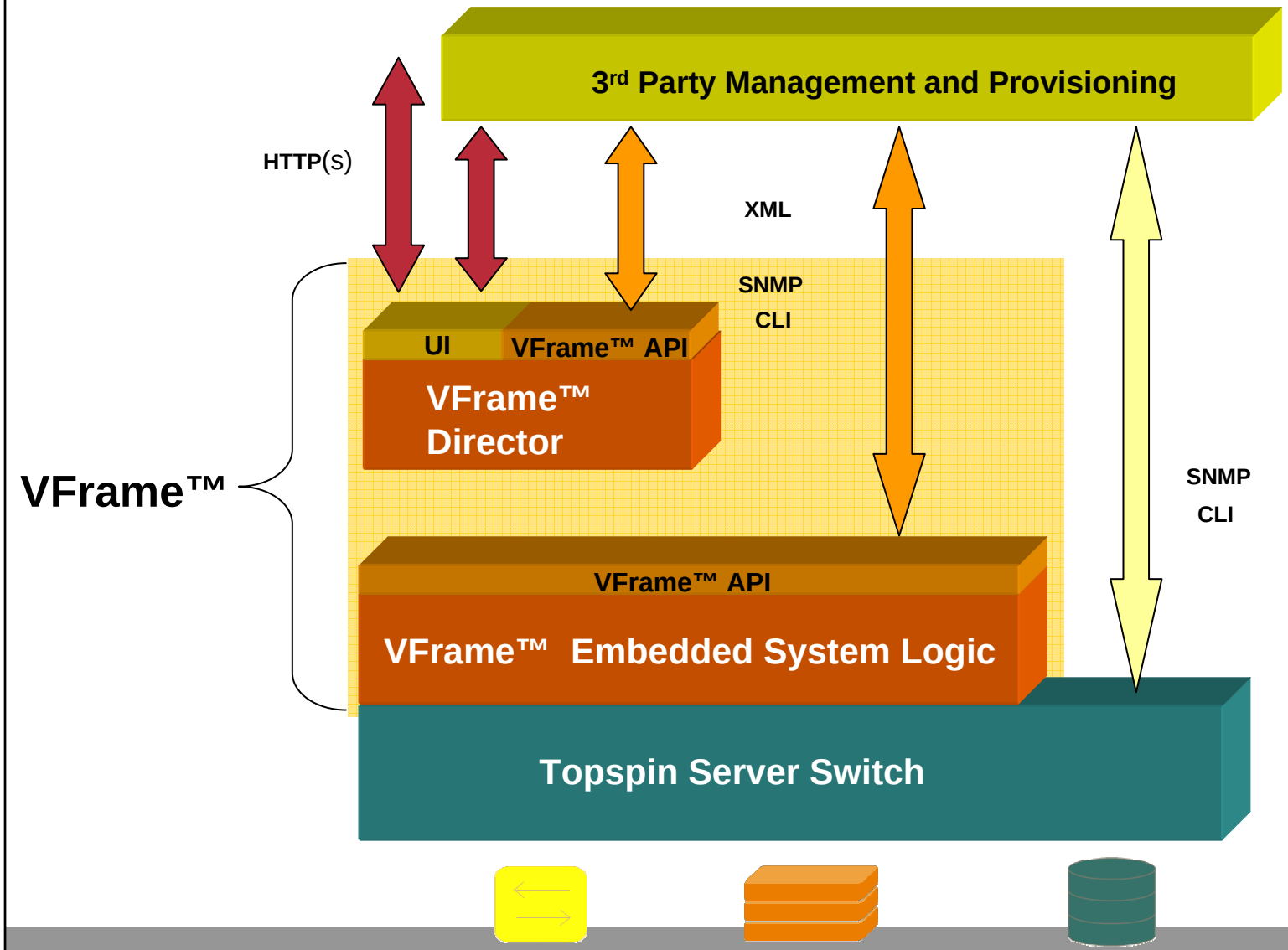
VFrame™ Director

Software package disseminates policies to server switch fabric

Central policy enforcement provides better system wide decision making and conflict arbitration

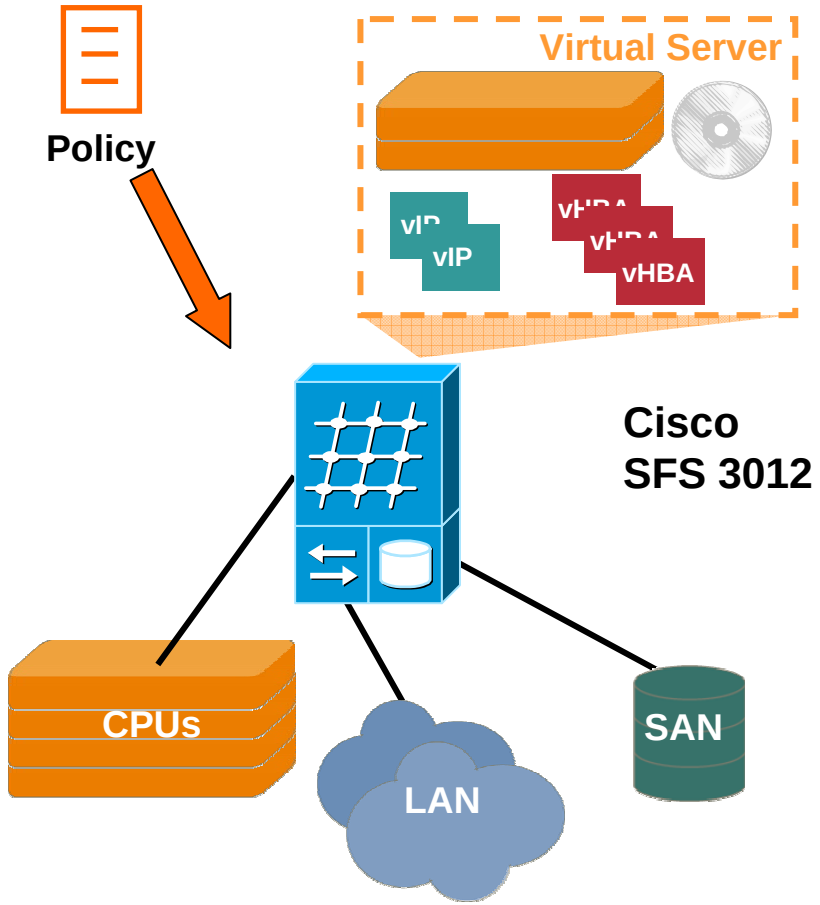
Can be installed on any server in the network

VFrame™ Architecture 架构



Programmability
可编程的*VFrame™*

1) Server Switch receives policy from VFrame™ Director or 3rd party software.
2) Based on policy, Server Switch assembles the virtual server
   - Selects server(s) that meet minimum criteria (e.g. CPU, memory)
   - Boot server(s) over the network with appropriate app/os image
   - Creates virtual IPs in servers and maps to VLANs for client access.
   - Creates virtual HBAs in servers and maps to Zones, LUNs, and WWNNs for storage access



How it Works 如何工作
Policy Definition

A Virtual Server combines:

Everything but the physical hardware. Ex:

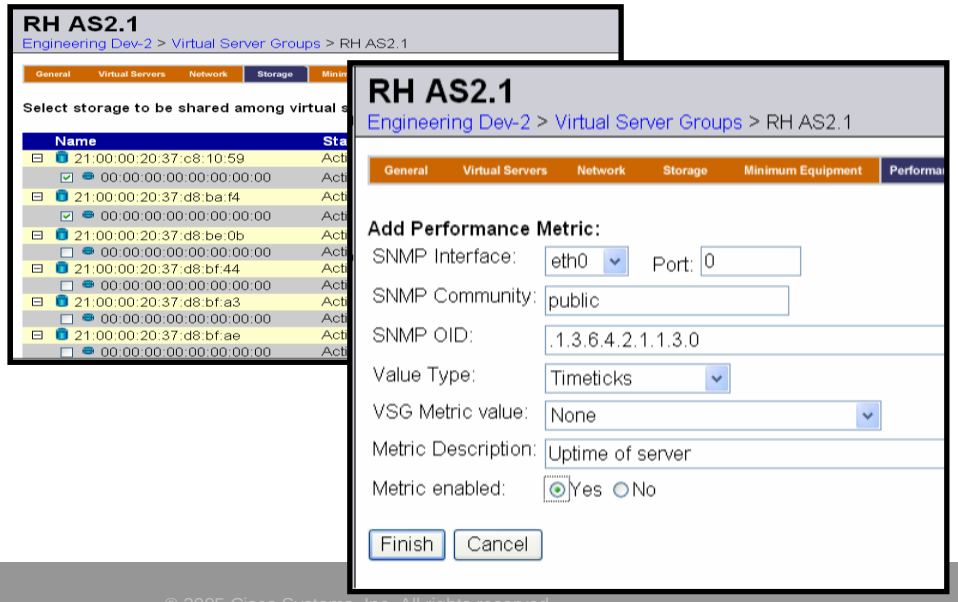
- Network Interfaces
- SAN WWNs
- Server Customization scripts

A Virtual Server Group combines:

- One or more Virtual Servers
- Shared Storage
- VLAN / SAN Zoning
- Performance Monitors
- **Policies**

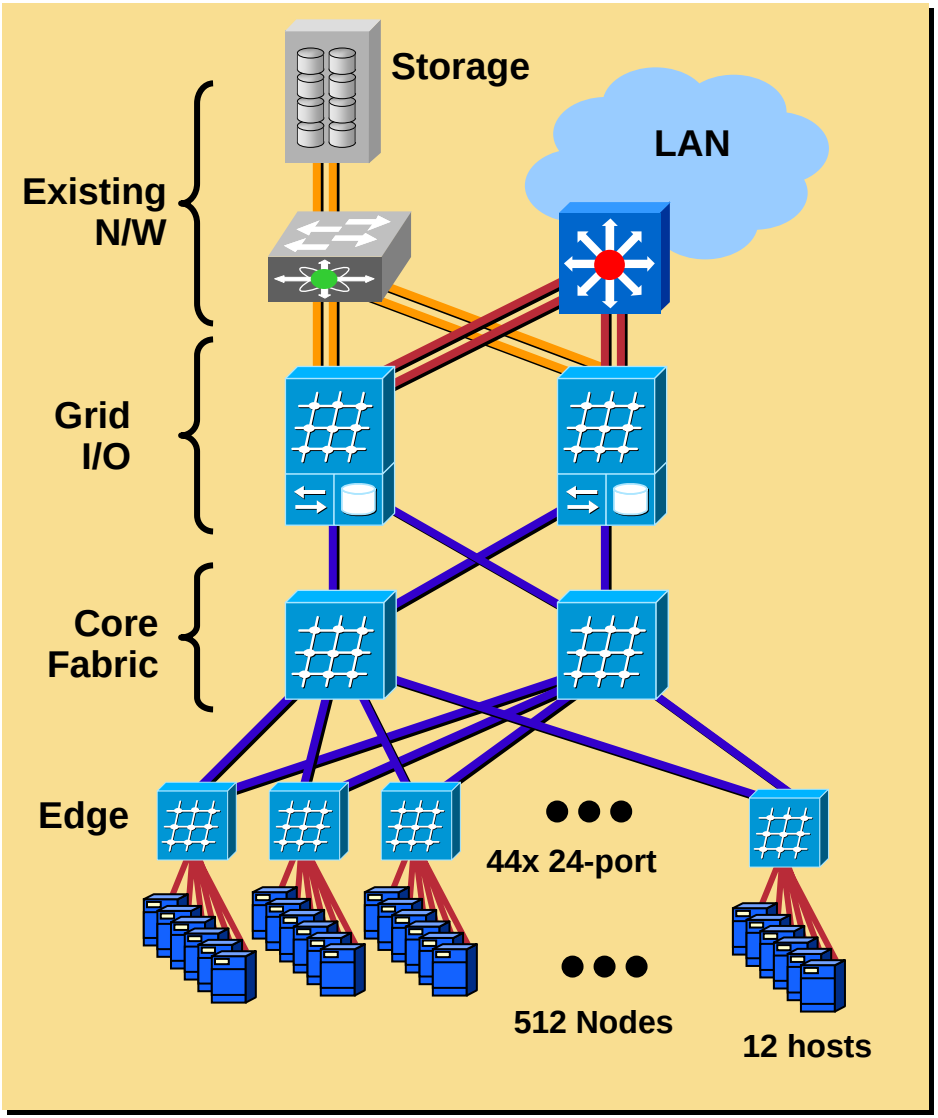
Policies Consist of:

- One or more **Trigger(s)**
  - Component Failure, Performance Metric, Scheduled Event, Custom Script
- One or more **Action(s)**
  - Add/Remove/Change Server or Group
  - Failover Server
  - Email Notification
  - Custom Script



Case Study: Large Wall Street Bank
Enterprise Grid Computing

- **Application:**
  - Replace proprietary platforms with standards-based components
  - Build scalable “on-demand” compute grid for financial applications
- **Benefits:**
  - 20X Price/Performance Improvement over four years
  - 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



Case Study: Utility Computing Service
Wall Street $ per CPU hosted grid

- **Application:**

  Build scalable “on-demand” compute service for enterprise customers (license $/CPU)

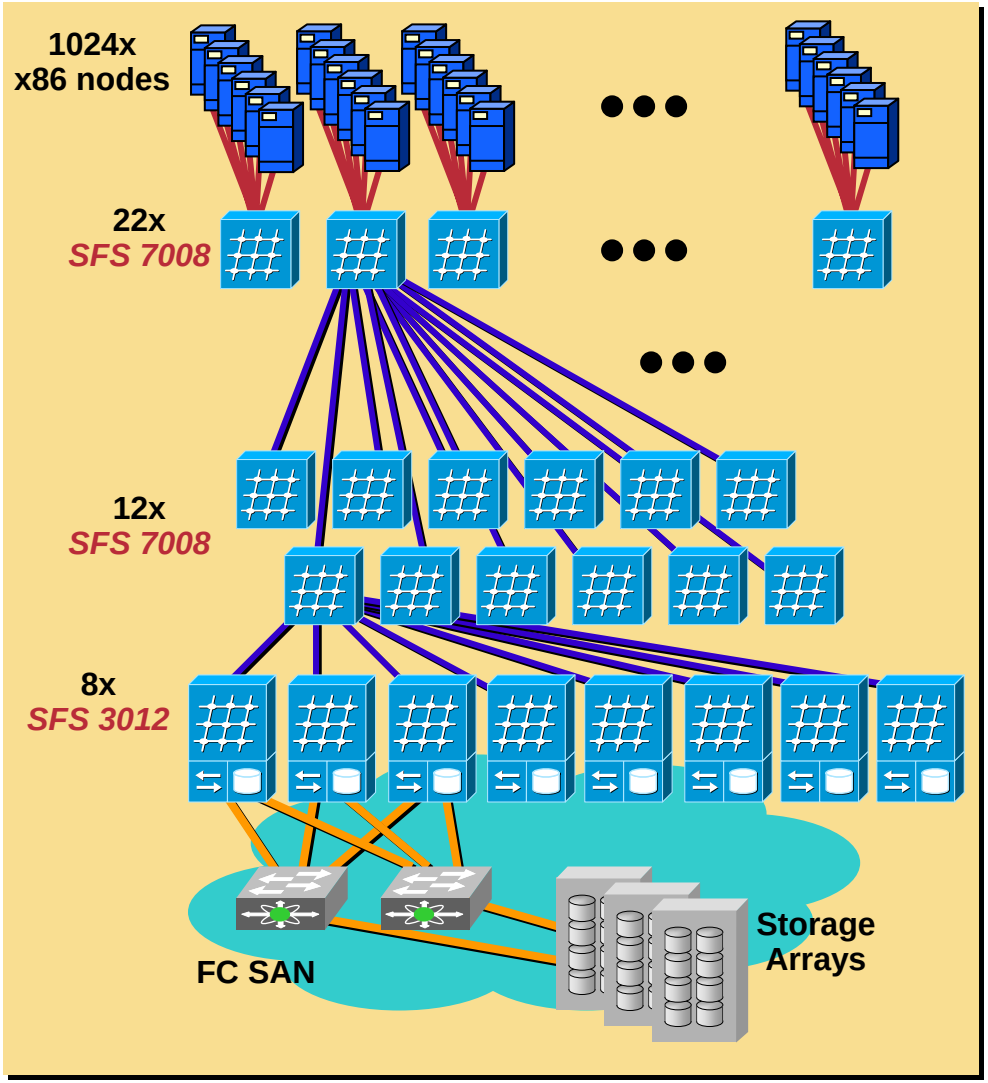
  Key initiatives around Financial Services and Energy verticals

- **Benefits:**

  Ability to outsource computing services to many customers with common infrastructure

  Flexibility to assign I/O resources to any server on-demand without recabling.

  Change between Linux, Windows, and Solaris in seconds.



Additional Content

- **Cisco Press –**

http://www.cisco.com/go/datacenter

- **Cisco Press –**

http://ciscopress.com/datacenterfundamentals

Q and A