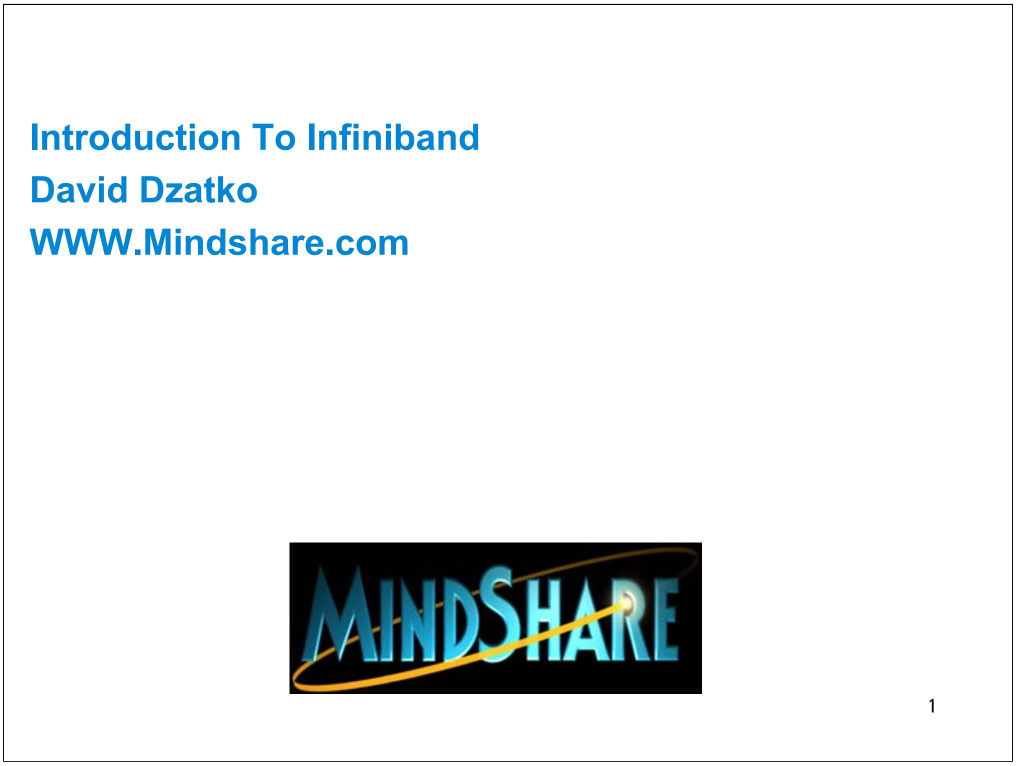
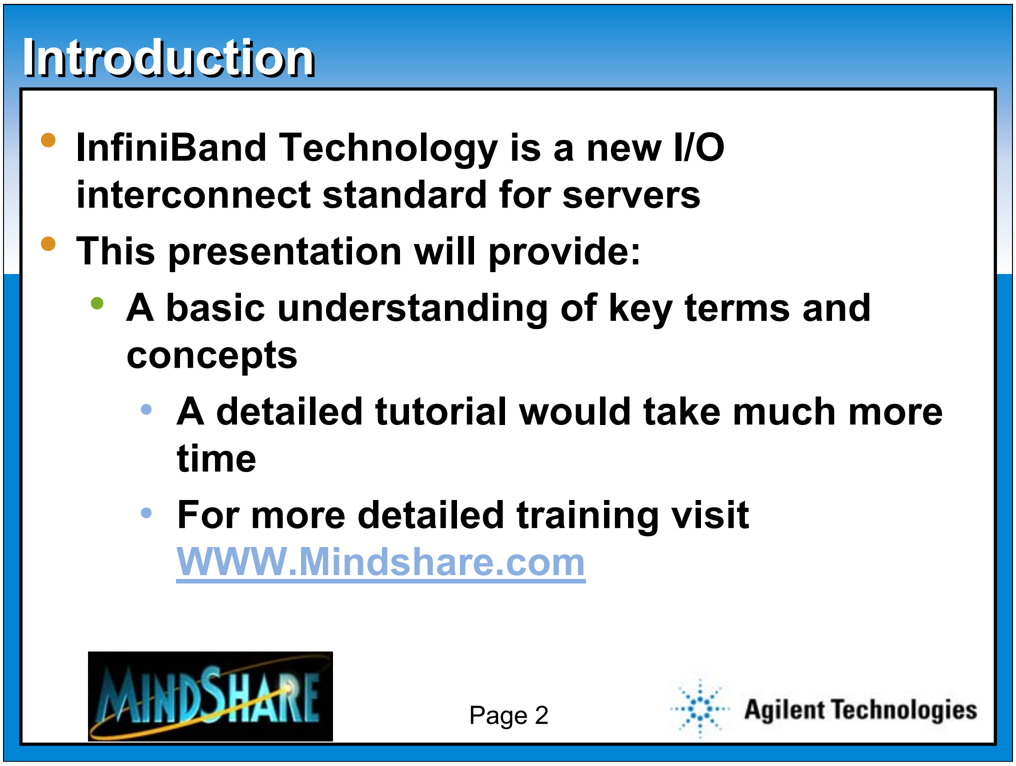


Introduction To Infiniband

David Dzatko

WWW.Mindshare.com

Infiniband System Architecture class from Mindshare.

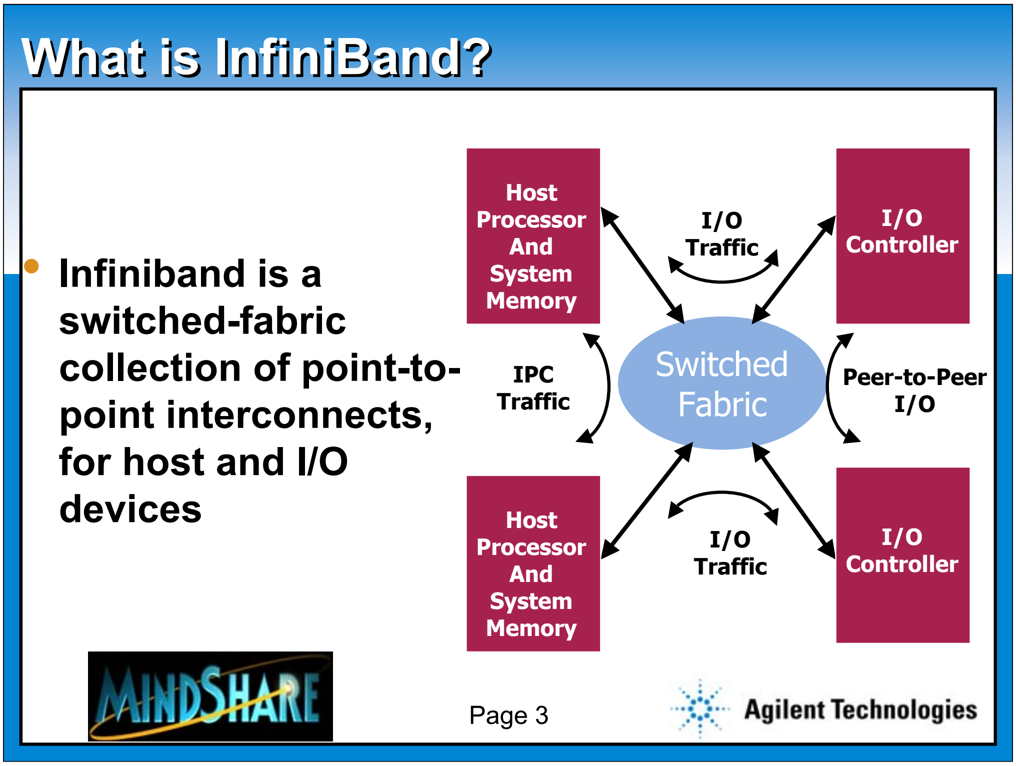


Introduction

- InfiniBand Technology is a new I/O interconnect standard for servers
- This presentation will provide:
  - A basic understanding of key terms and concepts
    - A detailed tutorial would take much more time
    - For more detailed training visit WWW.Mindshare.com

Infiniband Technology is a new initiative to bring a powerful I/O architecture to the server computer industry. The specification is now available for adoption by system and peripheral manufacturers. This presentation has two main goals:

1. To introduce the key features and benefits of Infiniband.
2. To provide a brief introduction to some elements of Infiniband is provided. This discussion focuses on the components that make up the Infinband architecture and the message passing communication model.



What is InfiniBand?

- Infiniband is a switched-fabric collection of point-to-point interconnects, for host and I/O devices

Host Processor And System Memory

I/O Traffic

I/O Controller

IPC Traffic

Switched Fabric

Peer-to-Peer I/O

Host Processor And System Memory

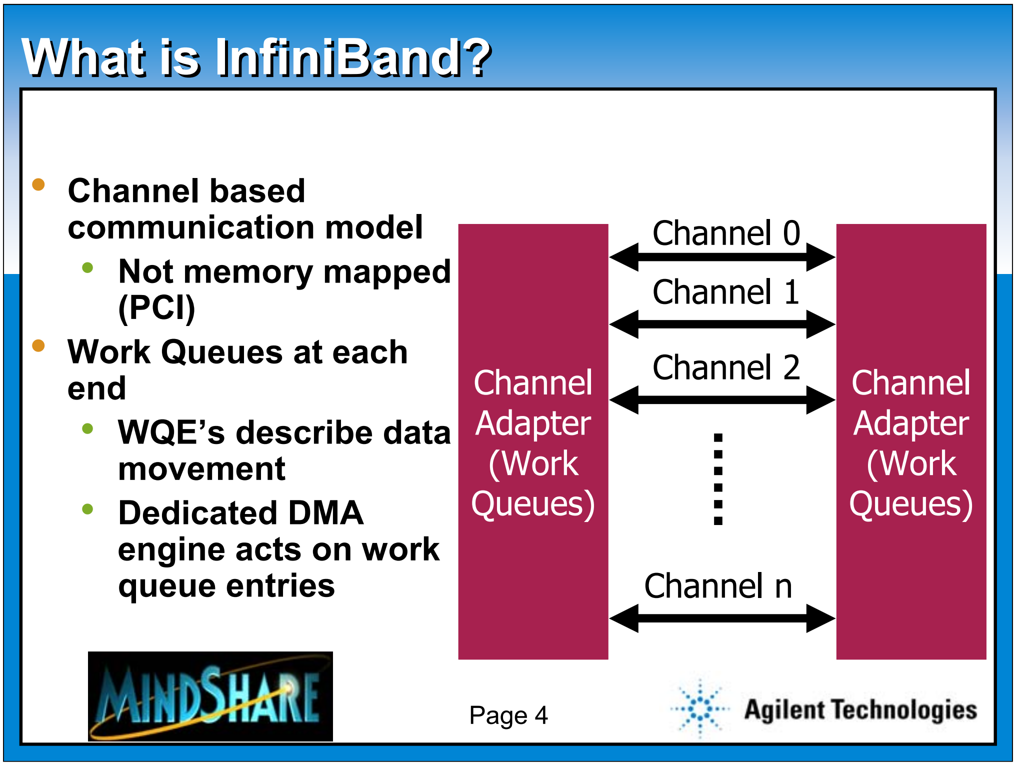
I/O Traffic

I/O Controller

MINDSHARE

Agilent Technologies

Infiniband is a switched fabric, meaning that a network of multiple switches and point-to-point links will interconnect two devices together. Point-to-point links will allow good signal integrity at higher frequencies. The devices attached will include servers, networks and storage arrays. Infiniband will also be used to interconnect parallel clusters of processors for the purpose of inter-process communication (IPC).



What is InfiniBand?

- Channel based communication model
  - Not memory mapped (PCI)
- Work Queues at each end
  - WQE's describe data movement
  - Dedicated DMA engine acts on work queue entries

Channel Adapter (Work Queues)

Channel 0

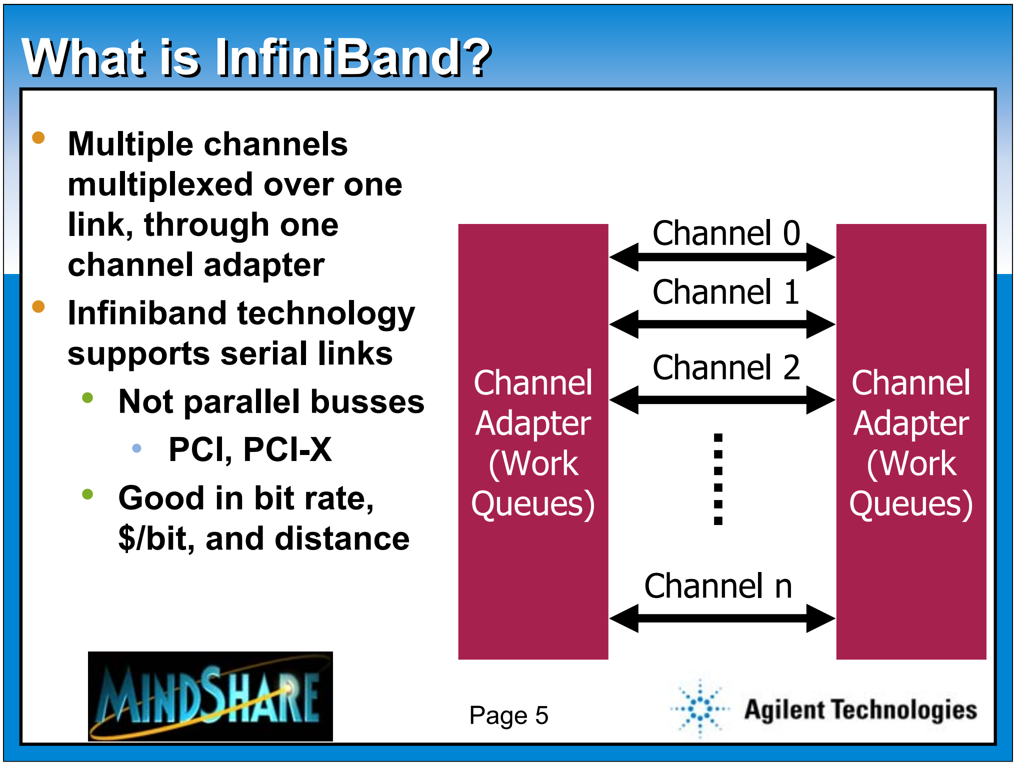
Channel 1

Channel 2

Channel n

Channel Adapter (Work Queues)

Channel based communication implies that a logical connection or path between two address spaces is established, allowing messages to flow through it. At each end of the channel there are Work Queue pairs (Send/Receive). The establishment of the channel associates the two Work Queue pairs. Work queue elements (WQE's) describe the type of data movement that occurs across the channel. Also at each end, there is a dedicated DMA engine that acts on the work queue elements.



What is InfiniBand?

- Multiple channels multiplexed over one link, through one channel adapter
- Infiniband technology supports serial links
  - Not parallel busses
    - PCI, PCI-X
  - Good in bit rate, $/bit, and distance

Channel Adapter (Work Queues)

Channel 0

Channel 1

Channel 2

Channel n

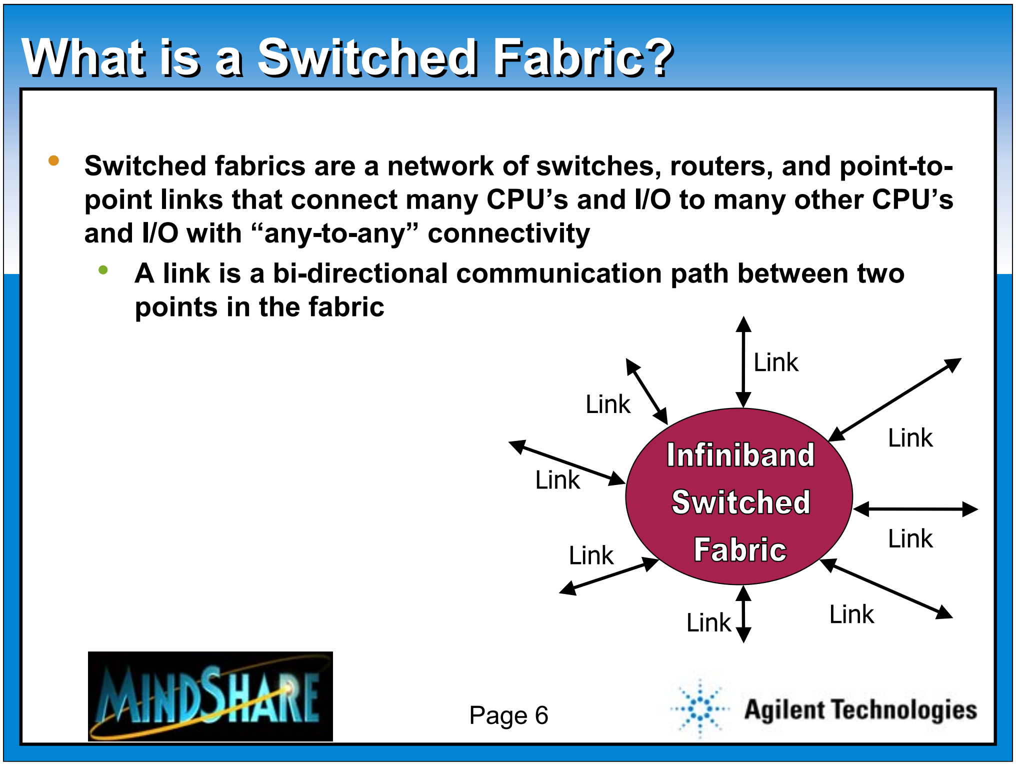
Channel Adapter (Work Queues)

MINDSHARE

Agilent Technologies

Multiple channels are time-multiplexed or scheduled across a single serial link. This is different compared to the memory mapped I/O and system memory accesses found on a PCI based machine, which move data using a Load/Store model.

Infiniband's serial links will allow for a high bit rates, low cost per bit, and great spacing between devices.



What is a Switched Fabric?

- Switched fabrics are a network of switches, routers, and point-to-point links that connect many CPU's and I/O to many other CPU's and I/O with "any-to-any" connectivity
  - A link is a bi-directional communication path between two points in the fabric

Infiniband Switched Fabric

Link

Link

Link

Link

Link

Link

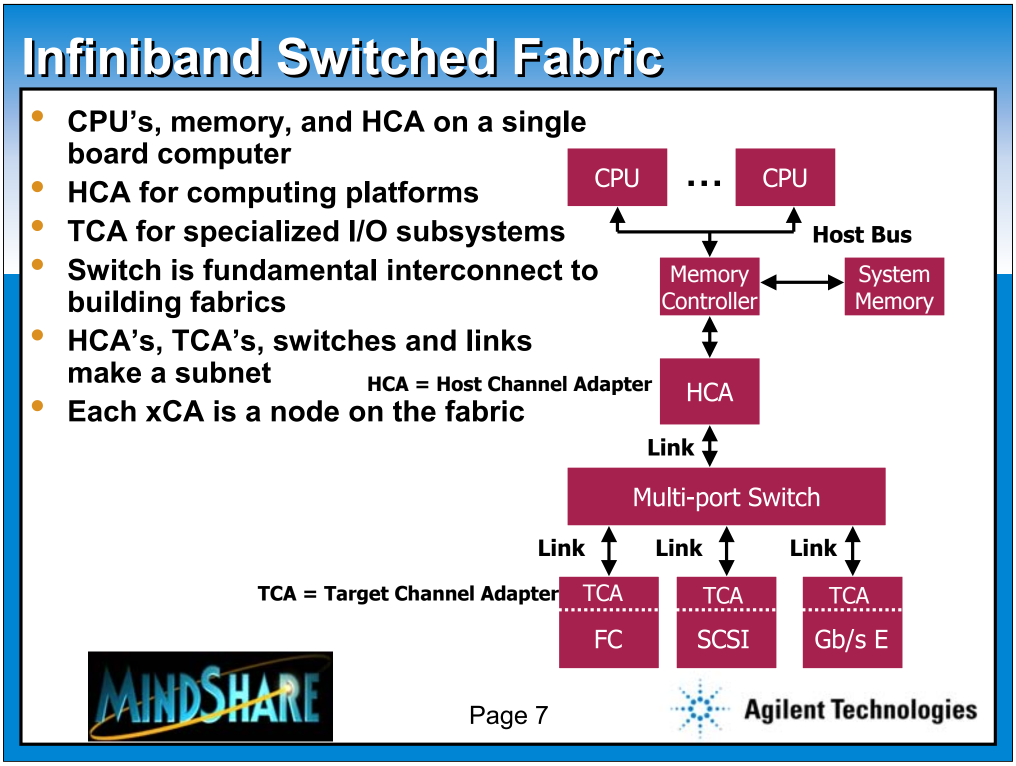
Link

Link

MINDSHARE

Agilent Technologies

Switch fabrics provide inherent scalability. Redundancy can also be easily enabled. Peer-to-peer communication is more natural and easier to implement. Flexible topologies can be easily created for tuning system performance.



Infiniband Switched Fabric

- CPU's, memory, and HCA on a single board computer
- HCA for computing platforms
- TCA for specialized I/O subsystems
- Switch is fundamental interconnect to building fabrics
- HCA's, TCA's, switches and links make a subnet
- Each xCA is a node on the fabric

CPU

. . .

CPU

Host Bus

Memory Controller

System Memory

HCA = Host Channel Adapter

HCA

Link

Multi-port Switch

Link

Link

Link

TCA = Target Channel Adapter

TCA

FC

TCA

SCSI

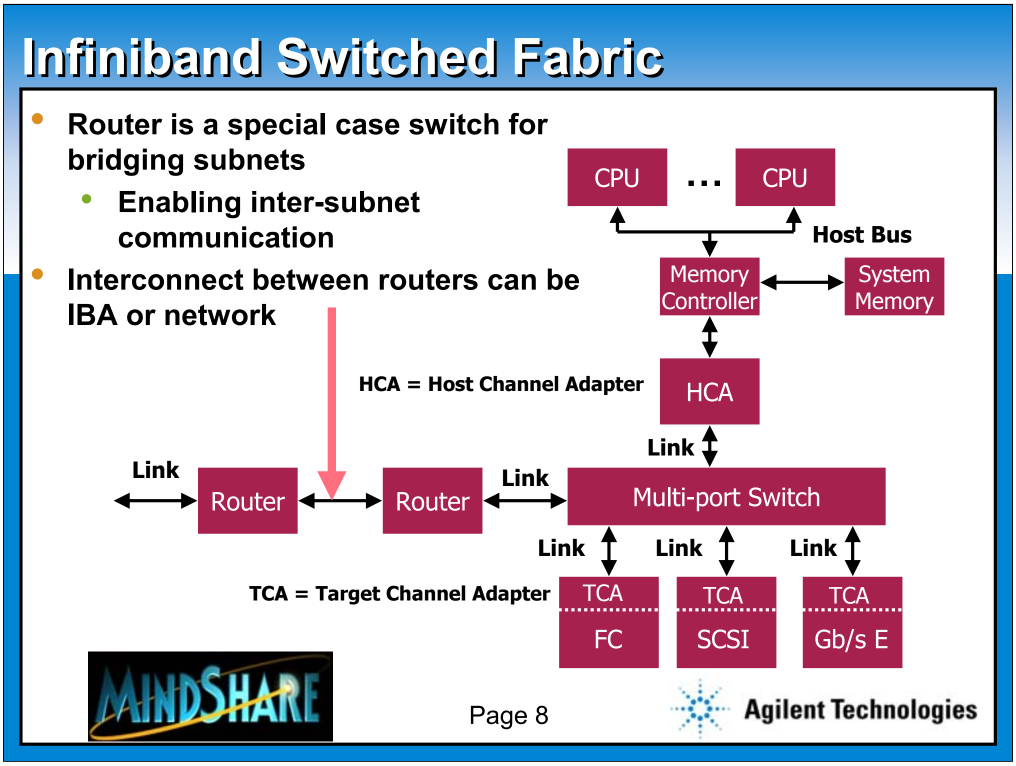
TCA

Gb/s E

MINDSHARE

Agilent Technologies

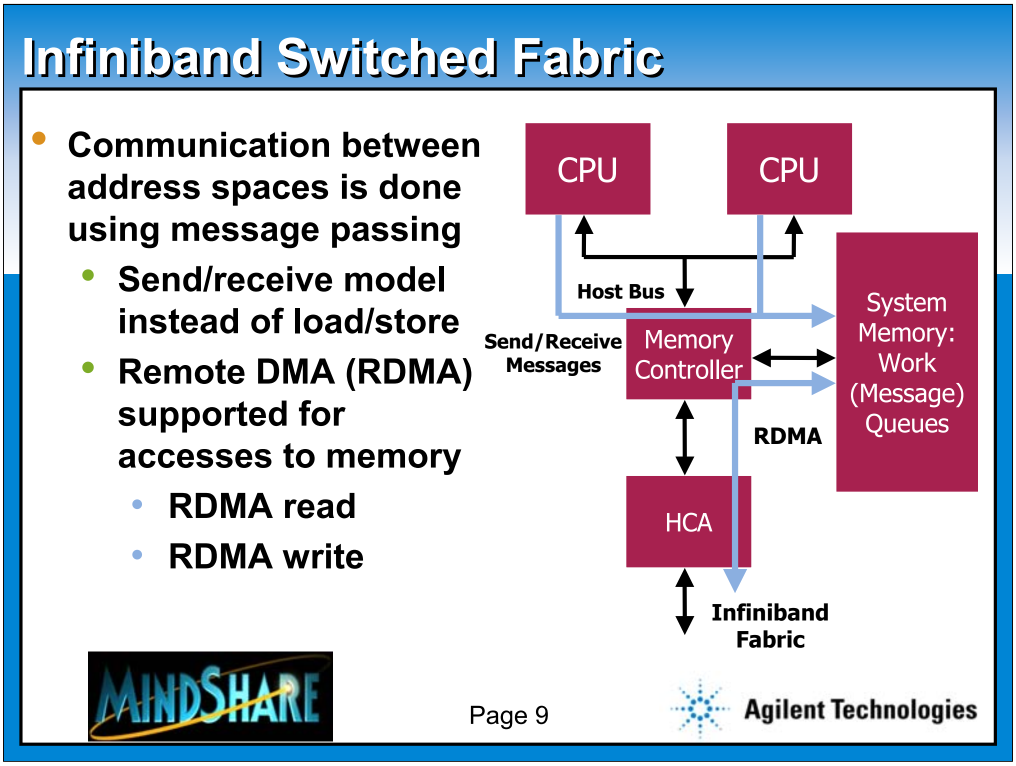
The Host Channel Adapter will provide the interface between host processors and system memory to the Infiniband fabric. The HCA is intended to be part of the chipset; meaning that IBA will be a first order network, with direct access to system memory. The processors, HCA, memory controller and memory could be part of a single board computer attached to a implementation-specific back plane through a standard module connector. Target channel adapters interface specialized I/O controllers to the Infiniband fabric. A switch is a fundamental component for creating fabrics by connecting HCA's and TCA's together. Note that you do not need a switch for fabrics. A multi-ported HCA could connect directly to a number of TCA's via individual links. Together, HCA's TCA's, switches and the interconnecting links make up an Infiniband subnet. Each TCA and HCA is a node on the fabric.



Infiniband Switched Fabric

- Router is a special case switch for bridging subnets
  - Enabling inter-subnet communication
- Interconnect between routers can be IBA or network

The Host Channel Adapter will provide the interface between host processors and system memory to the Infiniband fabric. The HCA is intended to be part of the chipset; meaning that IBA will be a first order network, with direct access to system memory. The processors, HCA, memory controller and memory could be part of a single board computer attached to a implementation-specific back plane through a standard module connector. Target channel adapters interface specialized I/O controllers to the Infiniband fabric. A switch is a fundamental component for creating fabrics by connecting HCA’s and TCA’s together. Note that you do not need a switch for fabrics. A multi-ported HCA could connect directly to a number of TCA’s via individual links. Together, HCA’s TCA’s, switches and the interconnecting links make up an Infiniband subnet. Each TCA and HCA is a node on the fabric.

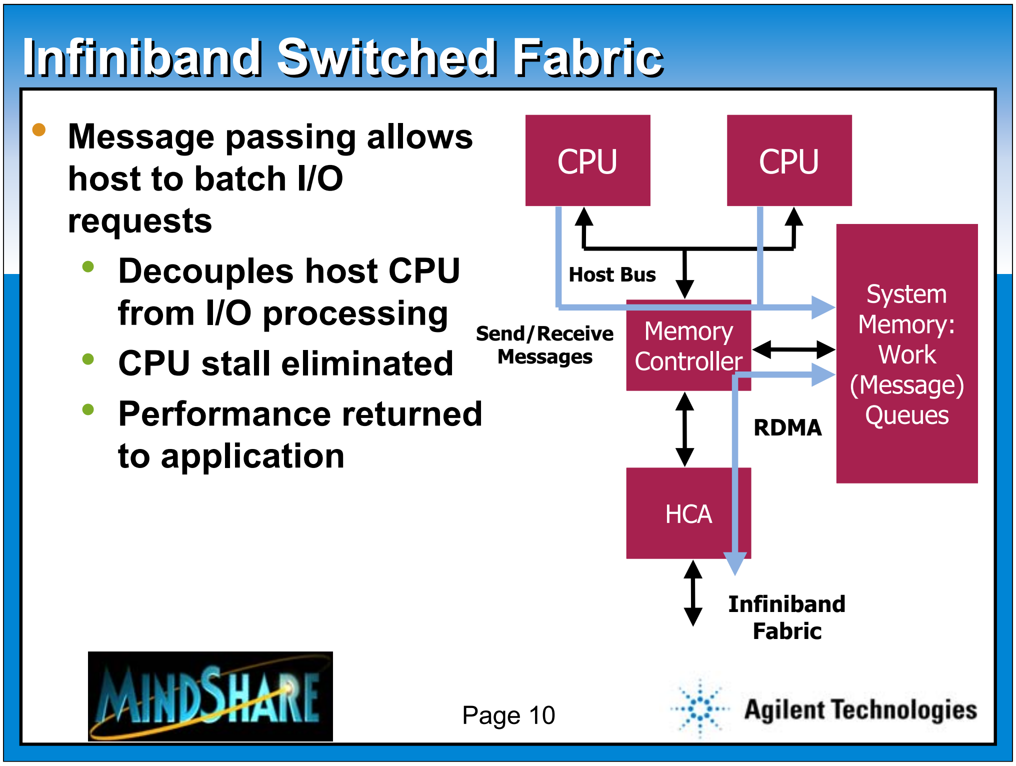


Infiniband Switched Fabric

- Communication between address spaces is done using message passing
  - Send/receive model instead of load/store
  - Remote DMA (RDMA) supported for accesses to memory
    - RDMA read
    - RDMA write

Agilent Technologies

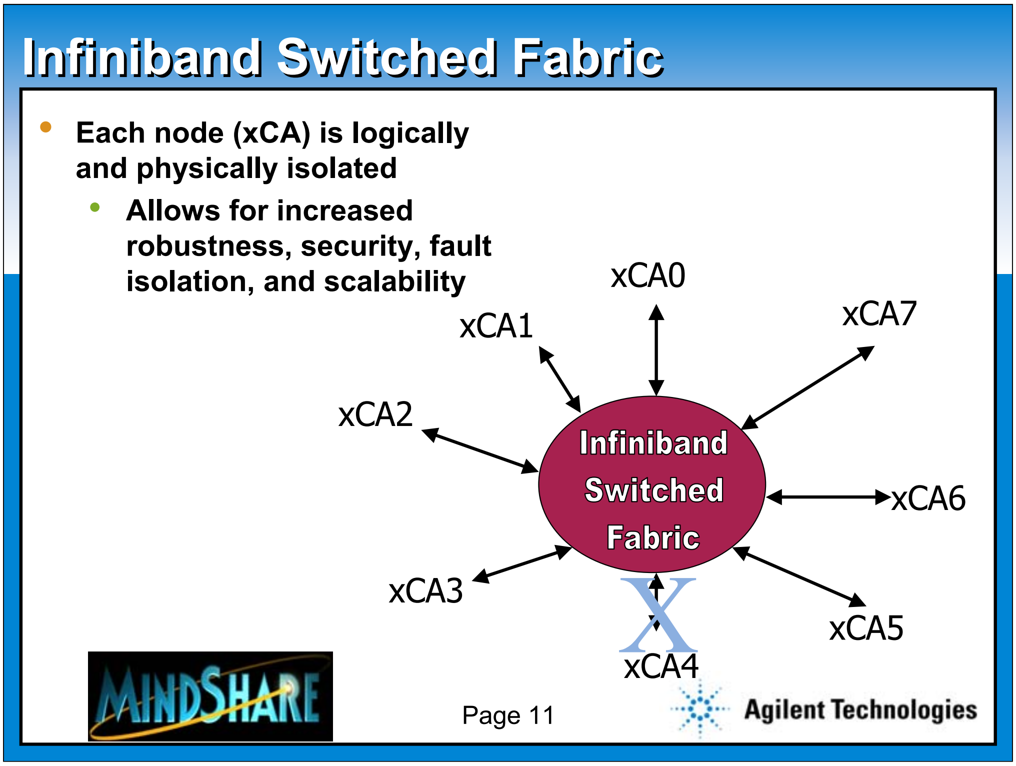
The Host Channel Adapter (HCA) is an interface between the Infiniband and the memory controller of a host server. Data travels between hosts and I/O devices through the passing of messages. Send and receive messages involve the cooperation of both end nodes. The initiator posts to the send queue a data structure describing the data to be sent. The target of the transfer posts to a receive queue a data structure describing where the received data is to be placed. RDMA messages involve the host or I/O device transferring data without the intervention of software on the other end. RDMA read involves a message to the send queue to describe the length and source of the data to be read and the local address space to put the data. RDMA write involves a message to the send queue describing the data to be written and where the data is to be written. Typically the send/receive method is used for control information, such as requests to transfer data and replies to data transfers. RDMA is for transferring large blocks of data. As an example, disk traffic and streaming network data would transfer using RDMA. RDMA does involve protection mechanisms for accesses to memory. The memory addresses associated with RDMA are virtual addresses, not physical addresses.



Infiniband Switched Fabric

- Message passing allows host to batch I/O requests
  - Decouples host CPU from I/O processing
  - CPU stall eliminated
  - Performance returned to application

With the implementation of work queues, I/O operations can be batched in system memory by a host processor. Once this queuing is complete, the processor is then free to perform other tasks not associated from the I/O operation. Since the processor does not have to interact with the I/O directly, the processor has not slowed to the speed of the I/O bus. With the I/O bus stall eliminated, that CPU bandwidth can be returned to the currently running application.



Infiniband Switched Fabric

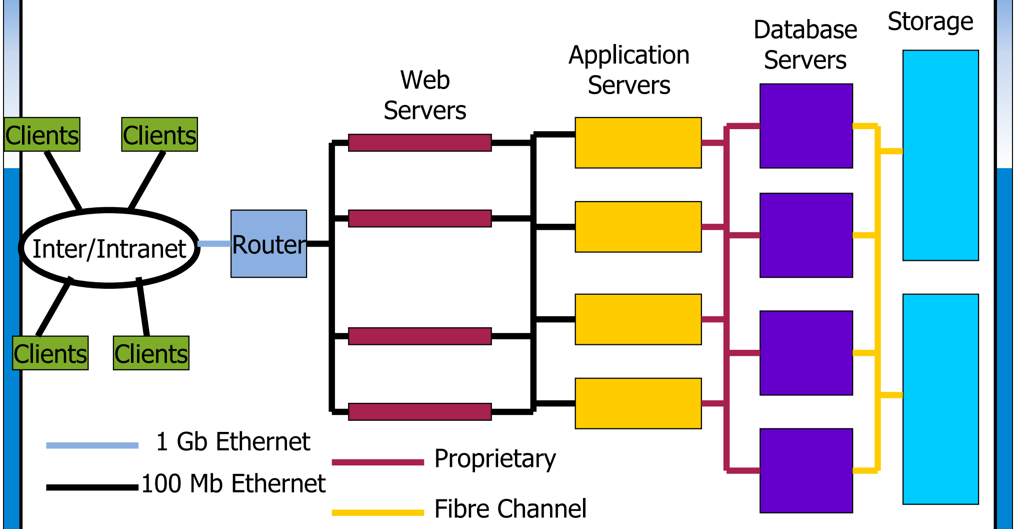
- Each node (xCA) is logically and physically isolated
  - Allows for increased robustness, security, fault isolation, and scalability

MINDSHARE

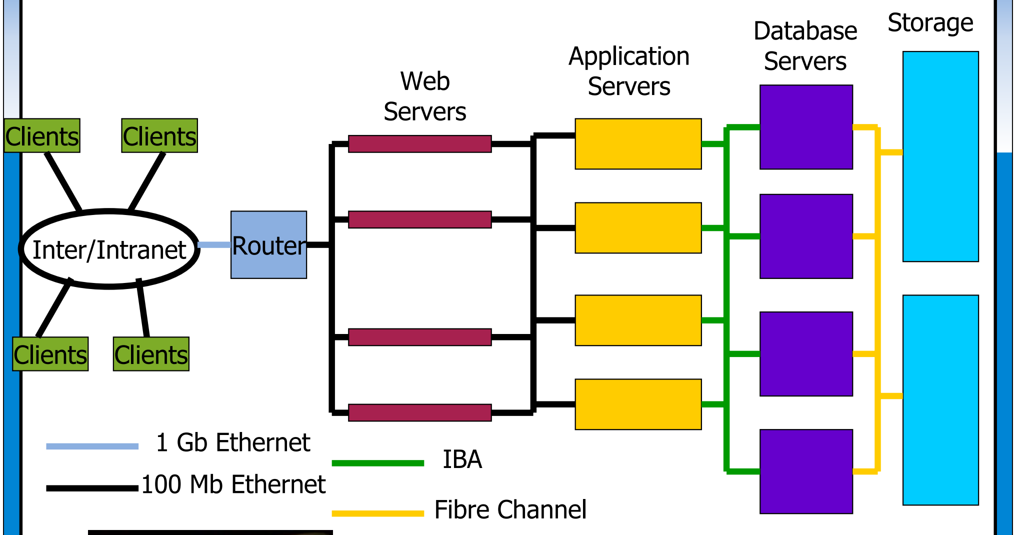
Agilent Technologies

xCA is a generic notation for host channel adapter (HCA) or target channel adapter (TCA). A TCA is an interface between an I/O controller (e.g. SCSI host bus adapter, Ethernet controller, FC switch or device, etc.) and the Infiniband fabric. Each node is a separate addressable space not mapped into system memory. Since the data transfer between two nodes does not depend on any other end node, if a node is lost due to failure, the network is still intact. The failing node can easily be detached and replaced. Robustness includes easy configuration and run-time tuning. Easy scaling is achieved by just adding a new fabric to a switch port of an existing fabric.

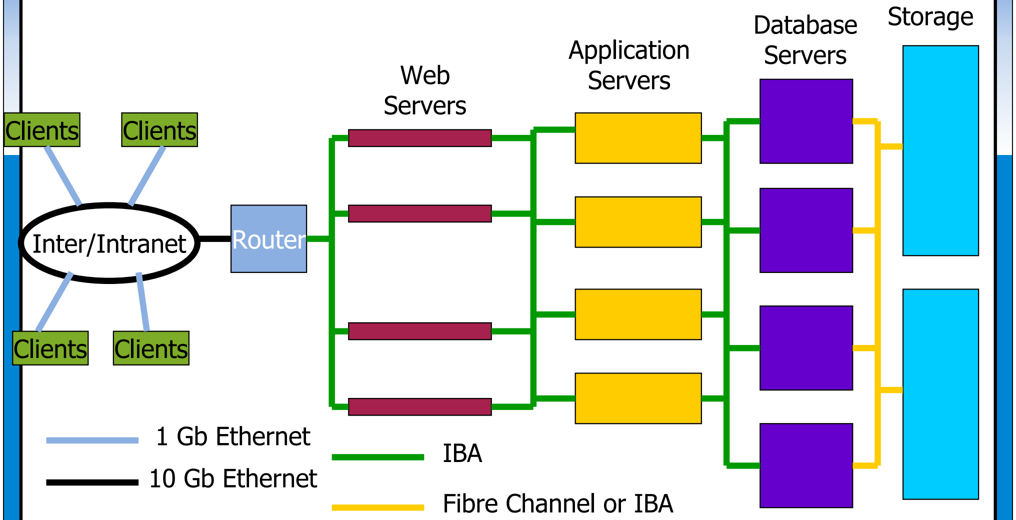
3 Tiered Data Center Today

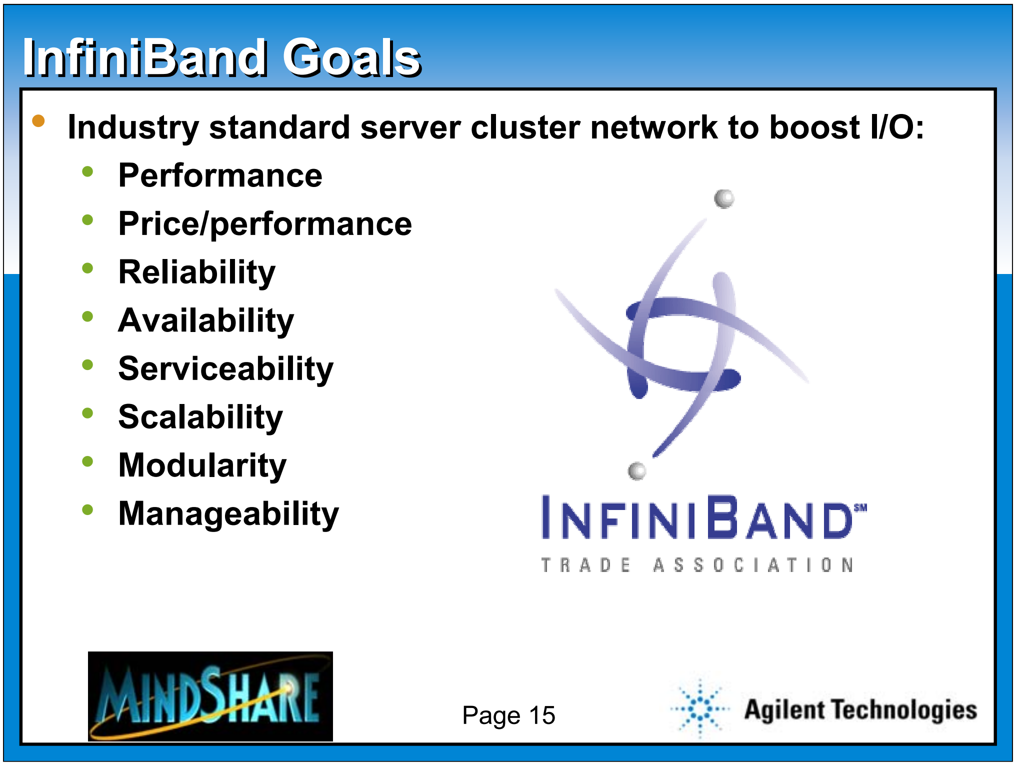


Early Adoption



Possible Future Deployment



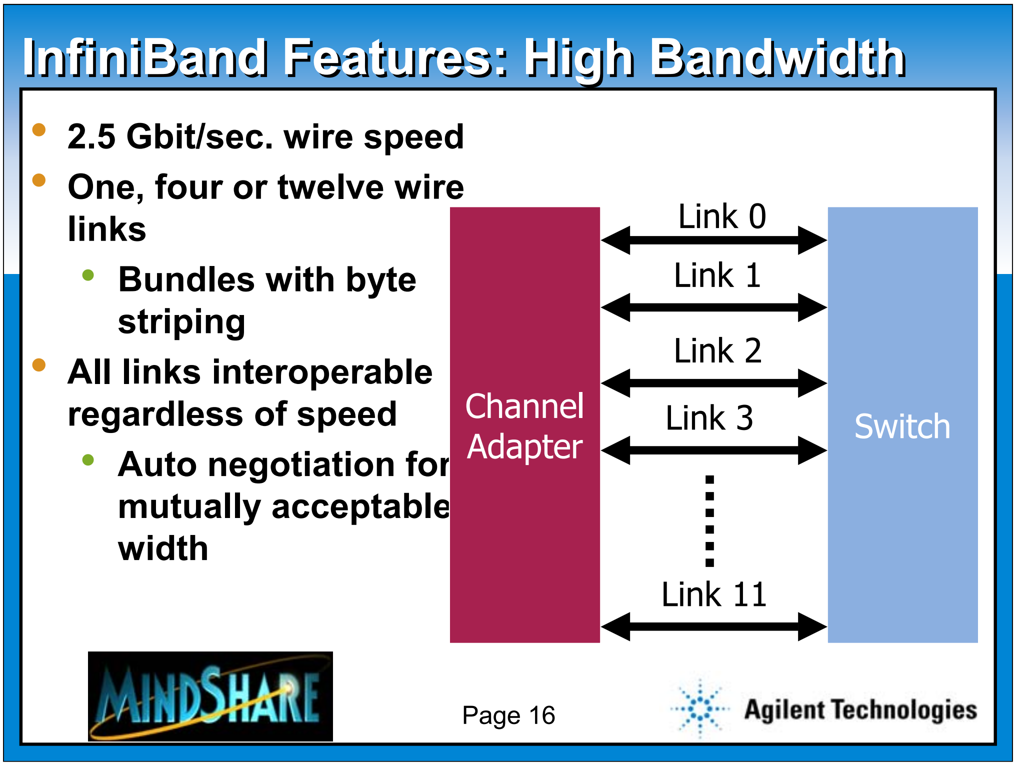


InfiniBand Goals

- Industry standard server cluster network to boost I/O:
  - Performance
  - Price/performance
  - Reliability
  - Availability
  - Serviceability
  - Scalability
  - Modularity
  - Manageability

INFINIBAND℠
TRADE ASSOCIATION

IBA surveyed IT managers looking for there requirements for next generation I/O. The goals of the Infiniband specification is to provide the industry a standard architecture for interfacing servers and I/O in such a way that improves the areas that IT managers have identified as important.

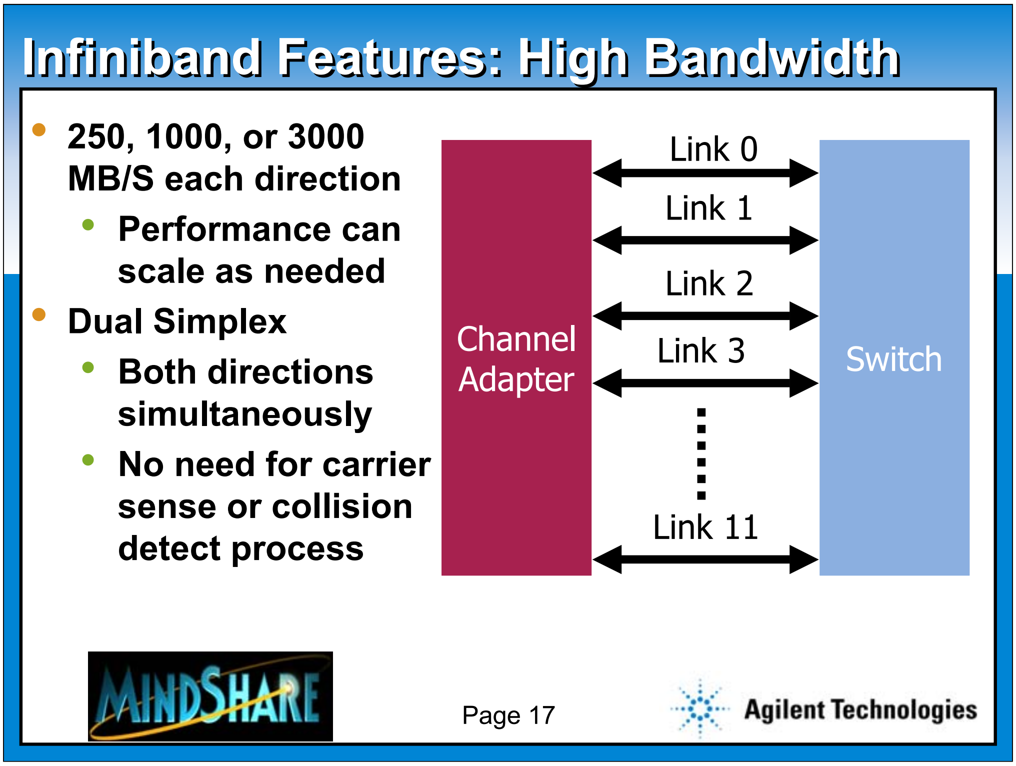


InfiniBand Features: High Bandwidth

- 2.5 Gbit/sec. wire speed
- One, four or twelve wire links
  - Bundles with byte striping
- All links interoperable regardless of speed
  - Auto negotiation for mutually acceptable width

Agilent Technologies

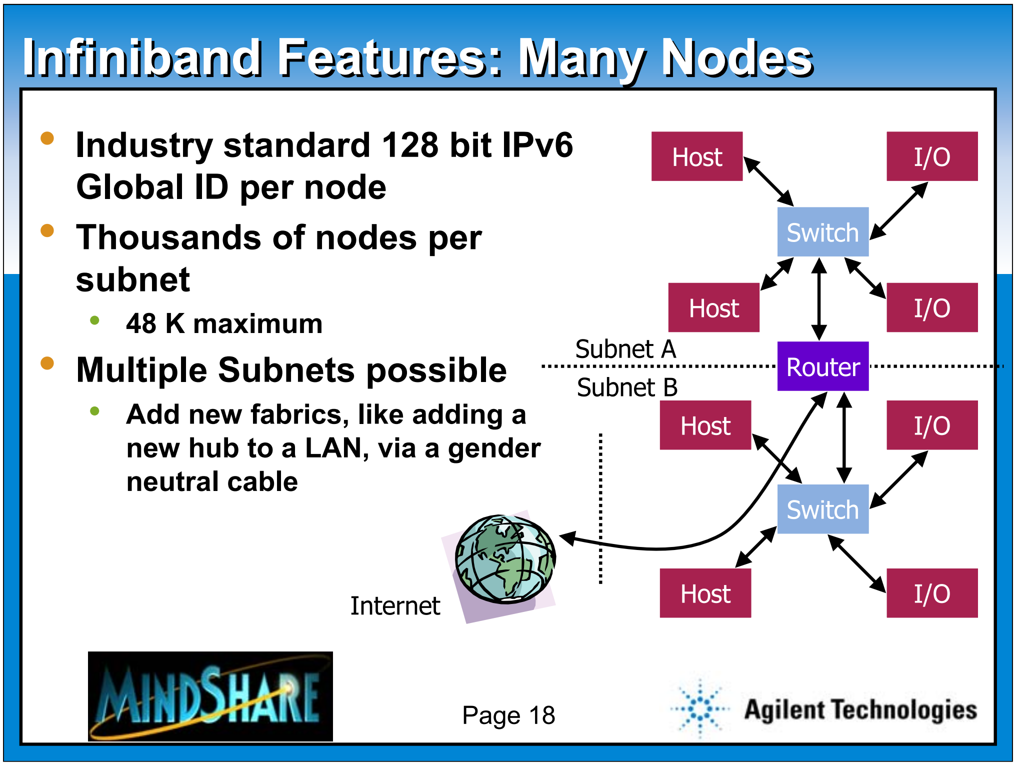
Each link will provide an aggregate 5 Gbit/second data transfer rate. The specification currently supports a 2.5 Gbit/second bit rate in both directions. If higher performance is required, links can be bundled to provide a 4X or 12X speed-up, relative to a single link, resulting in very high bandwidth. It is anticipated that hosts will be the initial implementers of the wider link widths. All packets are 4 bytes long or a multiple of four. Intermixing varying widths can occur on a single fabric. If multiple links are available, bytes will be striped along the wider bundles. If two nodes are dissimilar widths, the specification supports an auto negotiation algorithm between the two nodes to settle on a mutually acceptable width. The lowest common denominator is used in this scheme.



Infiniband Features: High Bandwidth

- 250, 1000, or 3000 MB/S each direction
  - Performance can scale as needed
- Dual Simplex
  - Both directions simultaneously
  - No need for carrier sense or collision detect process

Initially, the wire speed will be 2.5 Gbit/second. That rate can increase in the future, without dramatically changing the protocol, as higher data transfer rates are required. An upgrade capability is built into the specification. An auto-negotiation algorithm between links of dissimilar speeds is in place. A dedicated pair of wires in each cable will carry data in a single direction. Therefore, there’s no need to arbitrate for use of a link to send data in a given direction.



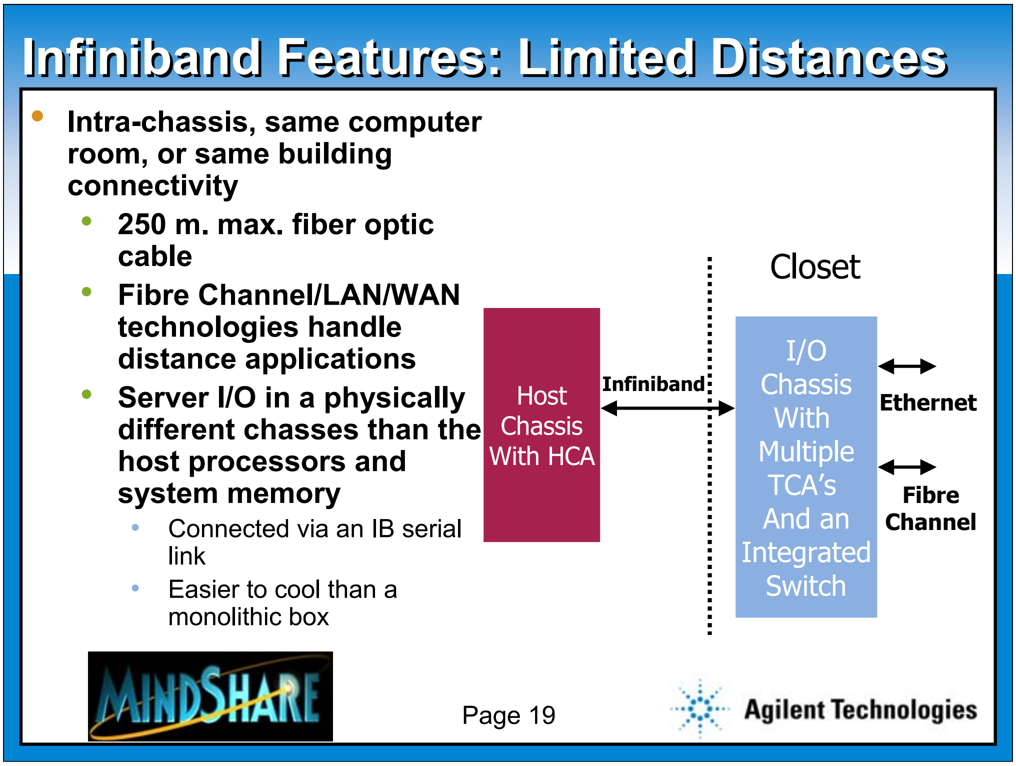
Infiniband Features: Many Nodes

- Industry standard 128 bit IPv6 Global ID per node
- Thousands of nodes per subnet
  - 48 K maximum
- Multiple Subnets possible
  - Add new fabrics, like adding a new hub to a LAN, via a gender neutral cable

MINDSHARE

Agilent Technologies

With Infiniband, it will be easy to dynamically grow a cluster without effecting existing hosts and I/O subsystems. This evolutionary, “pay as you grow” strategy of growth is very cost effective. A subnet can be added to an existing subnet by attaching two switches to a router as shown. The hardware and software will cooperate to integrate the new subnet into the system. The mechanical design of the cable will allow this connection to be easy.



Infiniband Features: Limited Distances

- Intra-chassis, same computer room, or same building connectivity
  - 250 m. max. fiber optic cable
  - Fibre Channel/LAN/WAN technologies handle distance applications
  - Server I/O in a physically different chasses than the host processors and system memory
    - Connected via an IB serial link
    - Easier to cool than a monolithic box

Closet

Host Chassis With HCA

Infiniband

I/O Chassis With Multiple TCA's And an Integrated Switch

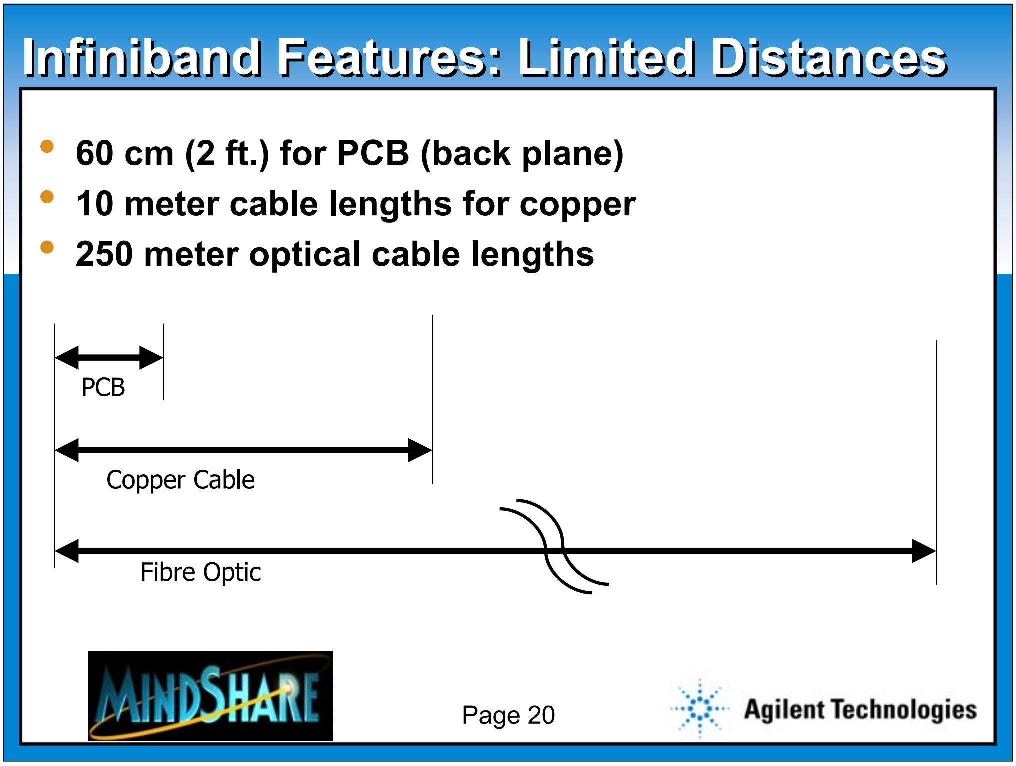
Ethernet

Fibre Channel

MINDSHARE

Agilent Technologies

The intent of the Infiniband is not to replace existing LAN/WAN and Storage Area Networks for distance connectivity. Instead, Infiniband has a relatively close connectivity focus, limited by the maximum cable lengths. Intra-chassis connectivity will allow board-to-board connectivity via a back plane. Same room connectivity through copper cables. Same building via fiber optics. Limited distance focus will allow the wire speeds to be relative high. Although the connections of Infiniband are close together, relative to other networks, the technology will allow servers and I/O to be within separate chasses, probably in the same room, connected through a serial link. Noisy I/O (disks) can be placed in the wire closet. The result is much more flexible systems compared to servers with a fixed number of I/O slots and CPU's, memory and I/O cards all in the same, monolithic box. Separate chassis will be easier to cool as well. The I/O chassis with the integrated switch is not too different in concept to a PCI-based expansion chassis.

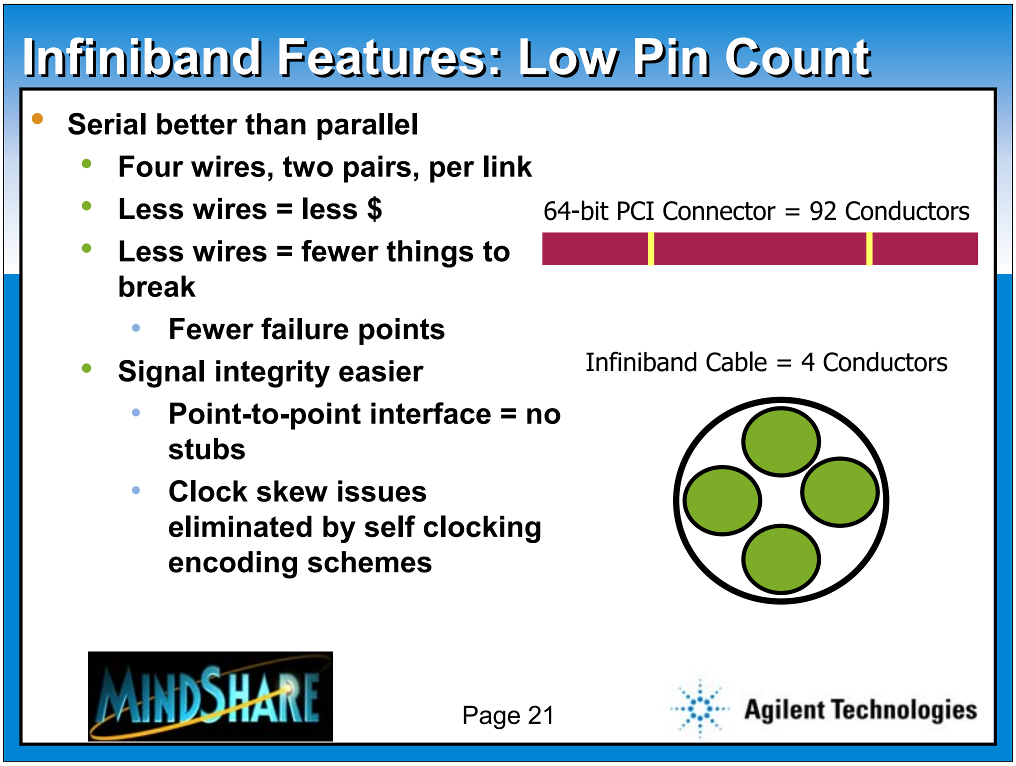


Infiniband Features: Limited Distances

- 60 cm (2 ft.) for PCB (back plane)
- 10 meter cable lengths for copper
- 250 meter optical cable lengths

PCB
Copper Cable
Fibre Optic

Internal system board traces for embedded I/O and local clustering are limited to 2 feet. Back plane connectors and modules for intra-chassis connections are defined in the spec. Copper cables are for medium distance, external connections within the computer room. Fiber Optic cables are for longer distances, like within a data center. Fibre Channel and LAN/WAN technologies will handle the distance applications.



Infiniband Features: Low Pin Count

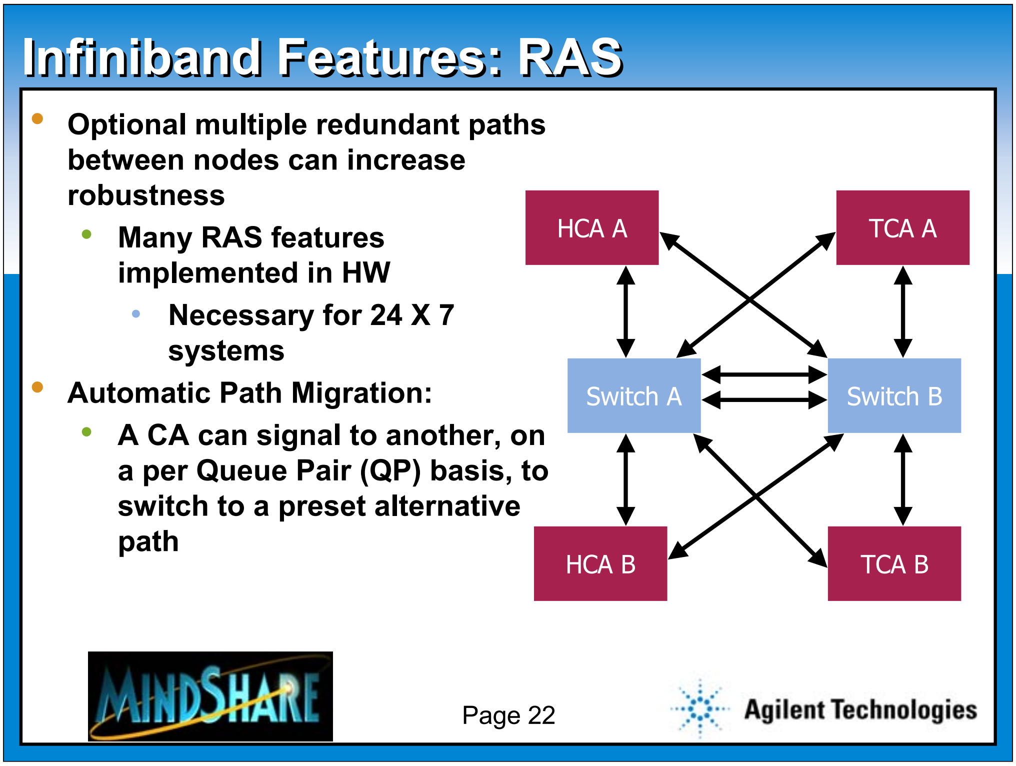
- Serial better than parallel
  - Four wires, two pairs, per link
  - Less wires = less $
  - Less wires = fewer things to break
    - Fewer failure points
  - Signal integrity easier
    - Point-to-point interface = no stubs
    - Clock skew issues eliminated by self clocking encoding schemes

64-bit PCI Connector = 92 Conductors

Infiniband Cable = 4 Conductors

There is a trend in the I/O industry of going to faster and narrower buses and away from wider, parallel buses. For low and medium speed peripherals, USB and IEEE 1394 Firewire serial buses have been widely implemented. Interconnect costs are reduced as fewer wires need to be bundled in a cable, laid down on a printed circuit board or within a connector. With few wires to break, there is a reduction in the possible failures that can occur with serial versus parallel connections.

Infiniband is a point-to-point interconnect. Signal integrity issues can be closely controlled. Traces can be stub-less; eliminating a source of secondary reflections that negatively affect signal integrity. The data encoding mechanism of IBA allows the clock to be embedded in transmission and recovered in reception. Therefore, clock skew is not an issue.



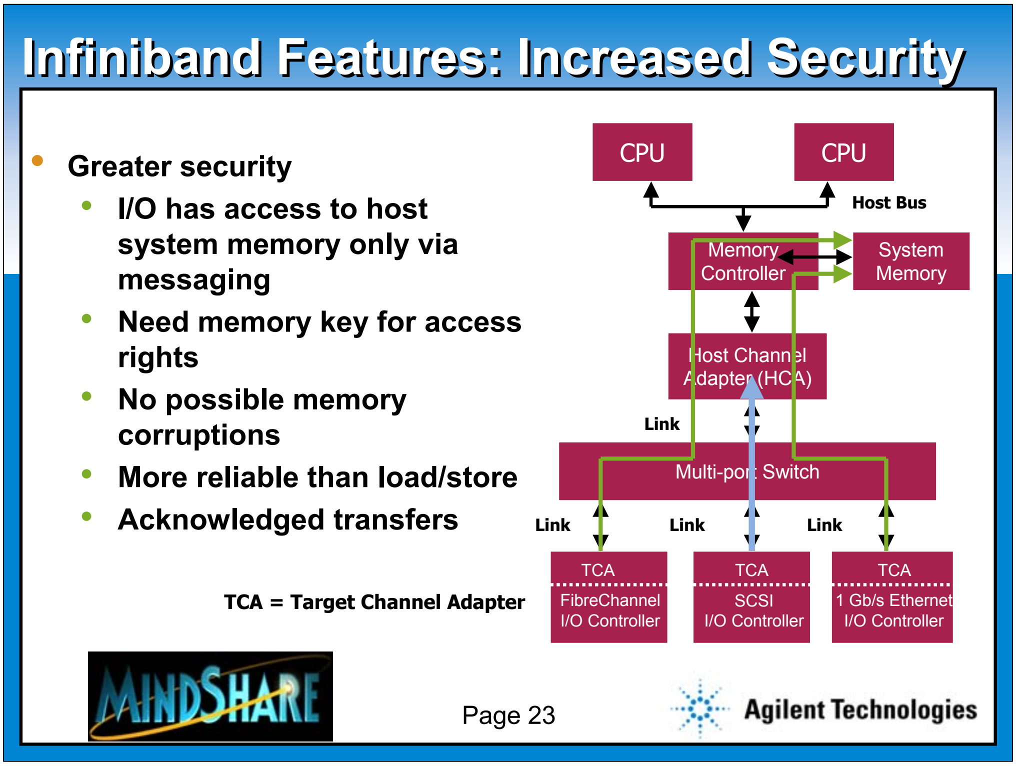
Infiniband Features: RAS

- Optional multiple redundant paths between nodes can increase robustness
    - Many RAS features implemented in HW
        - Necessary for 24 X 7 systems
- Automatic Path Migration:
    - A CA can signal to another, on a per Queue Pair (QP) basis, to switch to a preset alternative path

MINDSHARE

Agilent Technologies

Devices may be multi-ported for performance or enhanced reliability. In the above network, every node has two links of communication, such that if one has failed, data transfers can continue without downtime. Hardware and software cooperate to manage the alternative routing through the available links.



Infiniband Features: Increased Security

- Greater security
  - I/O has access to host system memory only via messaging
  - Need memory key for access rights
  - No possible memory corruptions
  - More reliable than load/store
  - Acknowledged transfers

TCA = Target Channel Adapter

CPU

CPU

Host Bus

Memory Controller

System Memory

Host Channel Adapter (HCA)

Link

Multi-port Switch

Link

Link

Link

TCA

FibreChannel I/O Controller

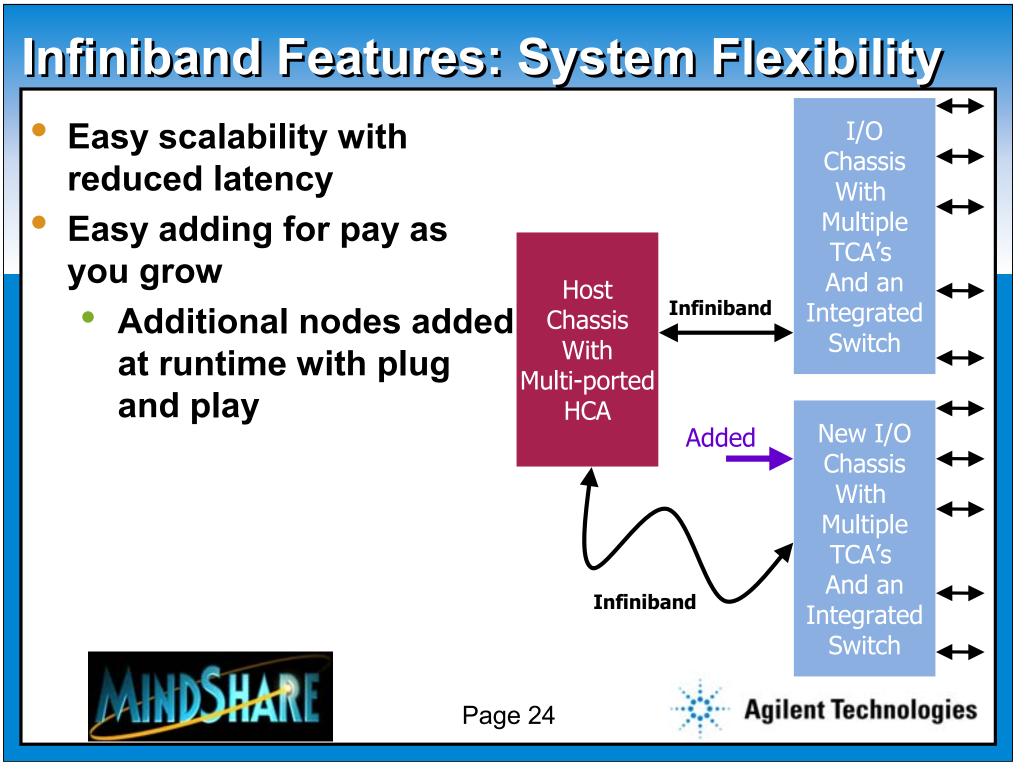
TCA

SCSI I/O Controller

TCA

1 Gb/s Ethernet I/O Controller

The memory management scheme of Infiniband will provide control over all accesses to system memory using virtual, contiguous ranges. Through hardware enforced protection in the HCA, I/O will be prevented from disturbing spaces not allocated for it. As part of the addressing scheme, a key will need to be provided by the remote DMA device for access into system memory. Message passing will be more reliable than the load/store model and shared memory in use today because currently the host bridge HW has no way to enforce access protection. Reliable data transfers uses acknowledges to guarantee uncorrupted delivery.



Infiniband Features: System Flexibility

- Easy scalability with reduced latency
- Easy adding for pay as you grow
  - Additional nodes added at runtime with plug and play

Host Chassis With Multi-ported HCA

Infiniband

I/O Chassis With Multiple TCA's And an Integrated Switch

Added

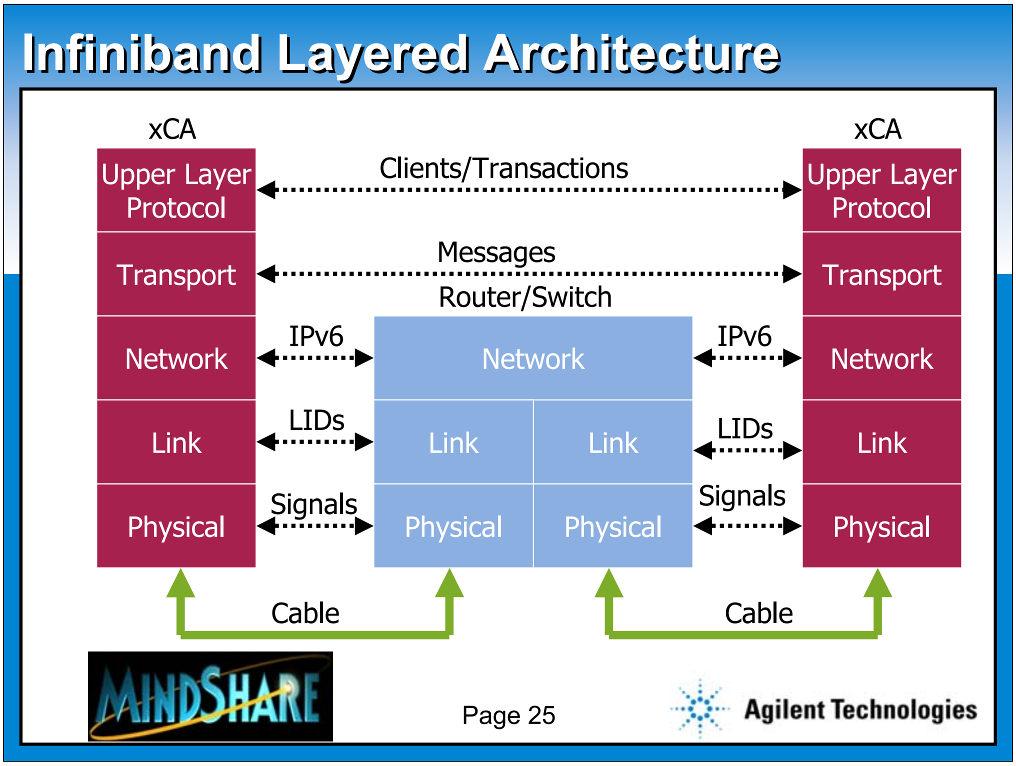
New I/O Chassis With Multiple TCA's And an Integrated Switch

Infiniband

MINDSHARE

Agilent Technologies

Infiniband allows for system expansion outside of the box to occur seamlessly. New I/O connects closer in this topology to system memory. This will reduce the latency experienced by the additional I/O devices when accessing system memory. This topology allows for more mileage out of the initial investment in the host chassis. Customers avoid the need to buy excess capacity up-front in anticipation of future growth. Instead, they buy what they need up-front and add capacity without impacting operations or installed systems with run-time plug and play.



Infiniband Layered Architecture

xCA

Upper Layer Protocol

Transport

Network

Link

Physical

Clients/Transactions

Messages

Router/Switch

IPv6

Network

LIDs

Link

Link

Signals

Physical

Physical

IPv6

LIDs

Signals

xCA

Upper Layer Protocol

Transport

Network

Link

Physical

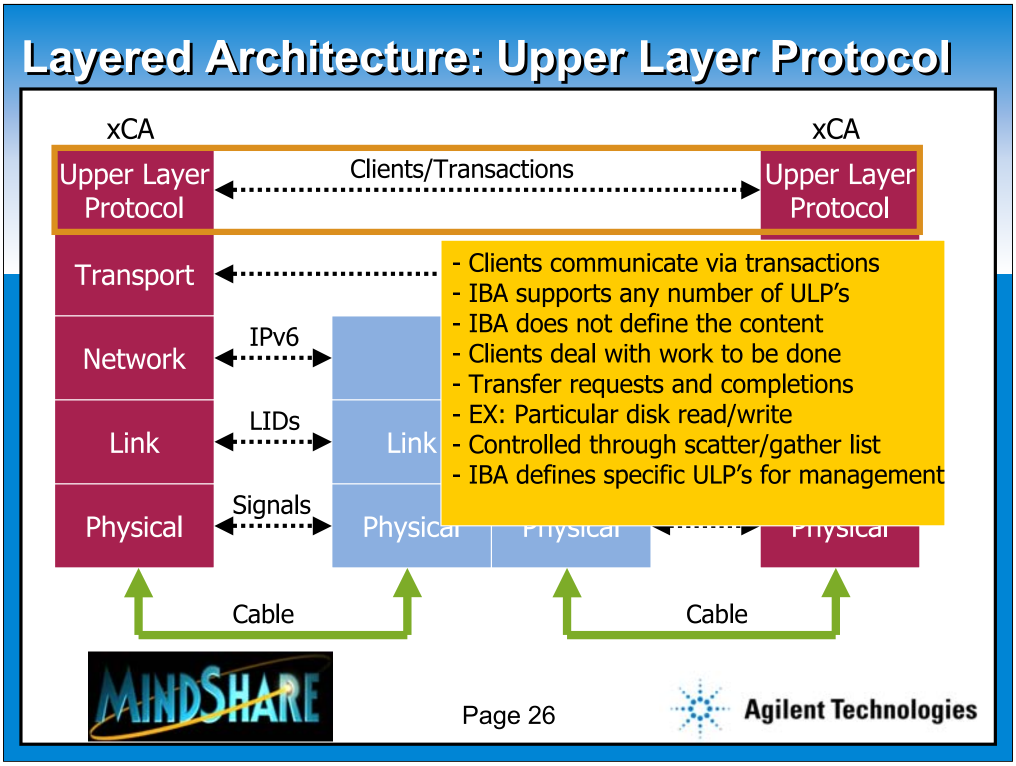
Cable

Cable

MINDSHARE

Agilent Technologies

IBA can be explained as a series of layers. Each layer is dependent on the layer below it. Each layer provides a service to the layer above it.



Layered Architecture: Upper Layer Protocol

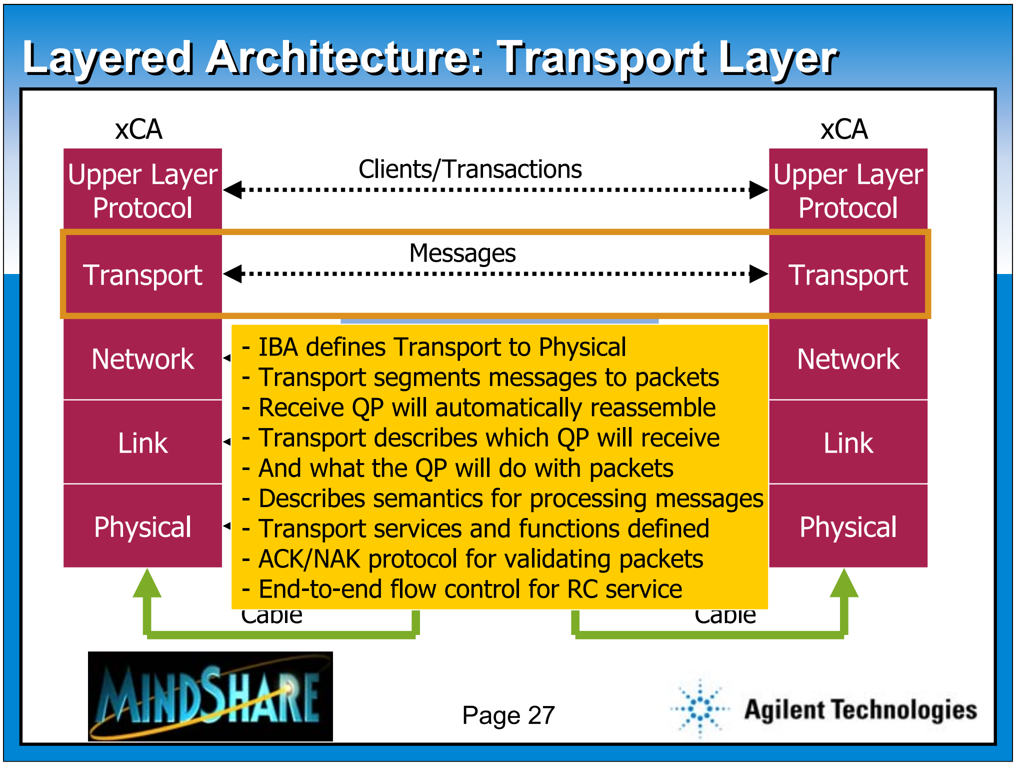
- Clients communicate via transactions
- IBA supports any number of ULP’s
- IBA does not define the content
- Clients deal with work to be done
- Transfer requests and completions
- EX: Particular disk read/write
- Controlled through scatter/gather list
- IBA defines specific ULP’s for management

MINDSHARE

Agilent Technologies

Upper Layer Protocol Layer-Consumers: Clients communicate through Transactions. IBA supports any number of higher level protocols used by client processes. IB doesn’t care what these are (IB does not specify content of messages). Clients are concerned about when did the data get there, is it good or not, is the transfer complete. Clients deal with work to be done: data transfer requests and completions. Example: A storage subsystem may be concerned about a particular disk read or particular disk write (controlled via a scatter/gather list).

IBA does support two specific ULP’s for management purposes: Subnet Management and Subnet Services. More on these later.



Layered Architecture: Transport Layer

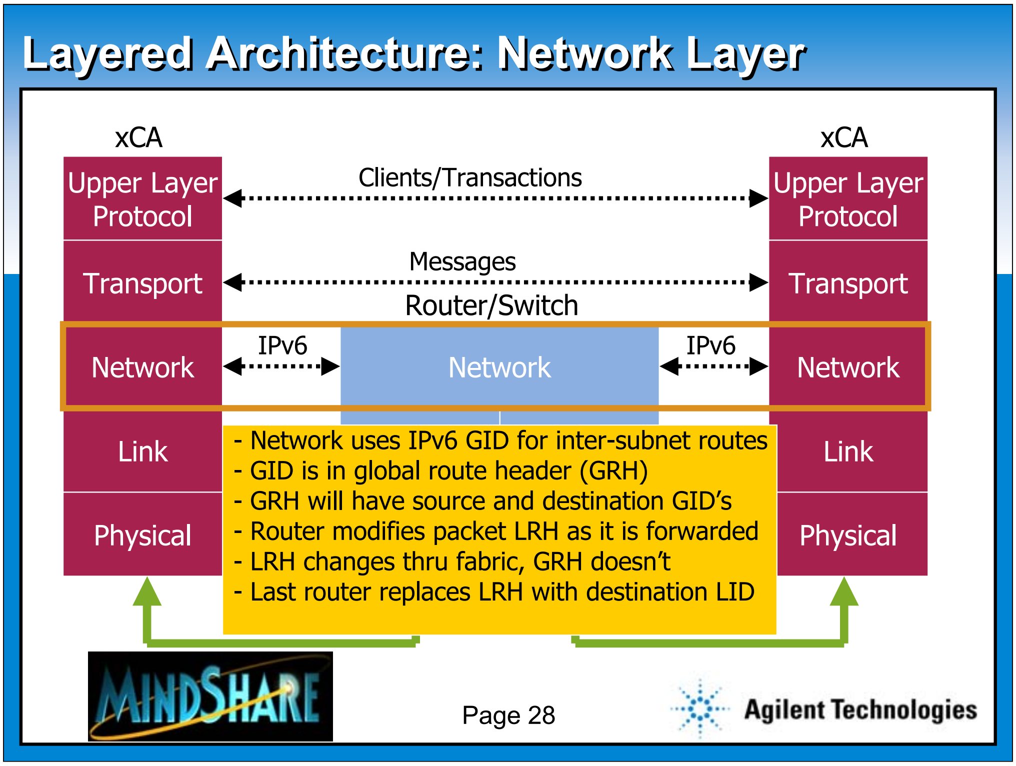
- IBA defines Transport to Physical
- Transport segments messages to packets
- Receive QP will automatically reassemble
- Transport describes which QP will receive
- And what the QP will do with packets
- Describes semantics for processing messages
- Transport services and functions defined
- ACK/NAK protocol for validating packets
- End-to-end flow control for RC service

MINDSHARE

Agilent Technologies

IBA defines the transport layer down to the physical layer. The Transport layer segments messages (units of work) into one or more packets. There may be one or more messages per transaction. When the data payload of the message is greater than the Path Maximum Transfer Unit (PMTU), transport will segment into multiple packets. Receiver QP’s perform automatic reassembly of packets into messages.

The Transport Layer also describes which Queue Pair will receive the packets and what the QP will do with those packets. Messages make sense to consumers. IBA provides transport services which are the semantics for processing messages. Examples include: Connected, Unconnected, Datagram, Multicast, and Raw Datagram. Transport functions describe how the queues behave when processing messages. Examples include Send/Receive, RDMA, and Atomics. Transport will implement the ACK/NAK protocol for validating that packets were or were not received correctly. End-to-end message flow control is performed for reliable connection service.



Layered Architecture: Network Layer

xCA

Upper Layer Protocol

Transport

Network

Link

Physical

Clients/Transactions

Messages

Router/Switch

IPv6

Network

IPv6

xCA

Upper Layer Protocol

Transport

Network

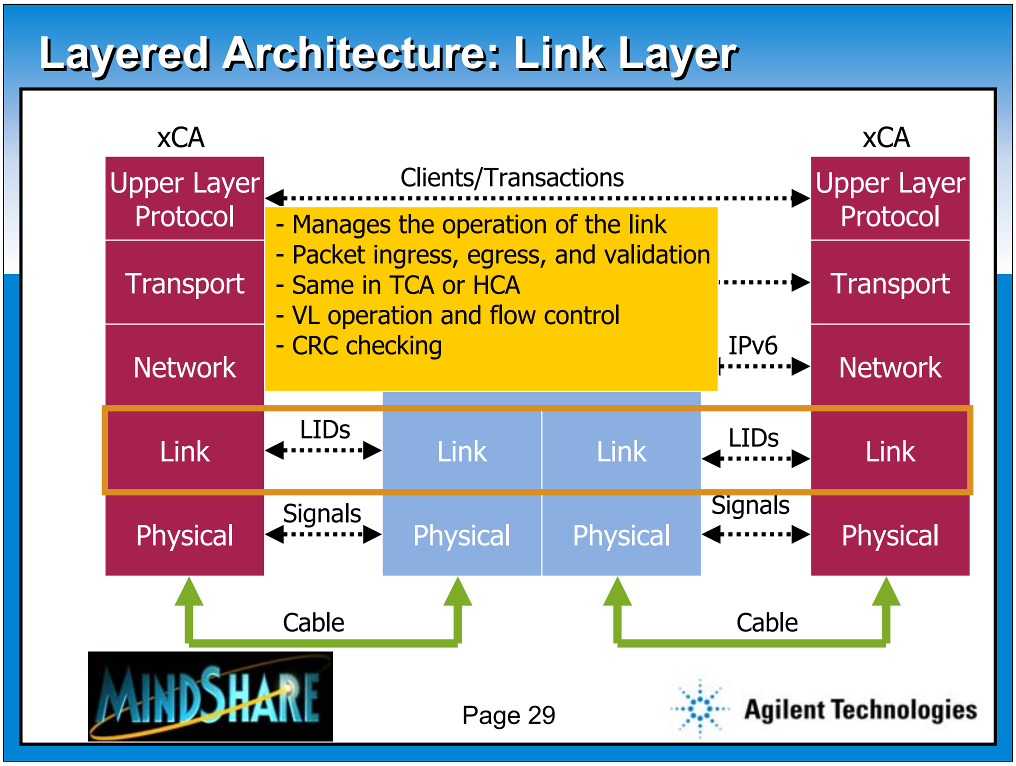
Link

Physical

- Network uses IPv6 GID for inter-subnet routes
- GID is in global route header (GRH)
- GRH will have source and destination GID's
- Router modifies packet LRH as it is forwarded
- LRH changes thru fabric, GRH doesn't
- Last router replaces LRH with destination LID

Agilent Technologies

The Network Layer uses IPv6 GID/LID addressing for inter-subnet routing purposes. A Global Route Header (GRH) will be present in any packet going between subnets. The GRH have an IPv6 address of the source and destination ports of the packet. Routers will forward the packet based on the GRH and a forwarding database. As a router forwards a packet between subnets, the LRH (Local Route Header) will be modified within the GRH, but the source/destination addresses do not change. This maintains the integrity of the end-to-end transport. The last router replaces the LRH with the LID of the destination node.



Layered Architecture: Link Layer

xCA

xCA

Upper Layer Protocol

Transport

Network

Link

Physical

Clients/Transactions

- Manages the operation of the link
- Packet ingress, egress, and validation
- Same in TCA or HCA
- VL operation and flow control
- CRC checking

IPv6

LIDs

LIDs

Signals

Signals

Link

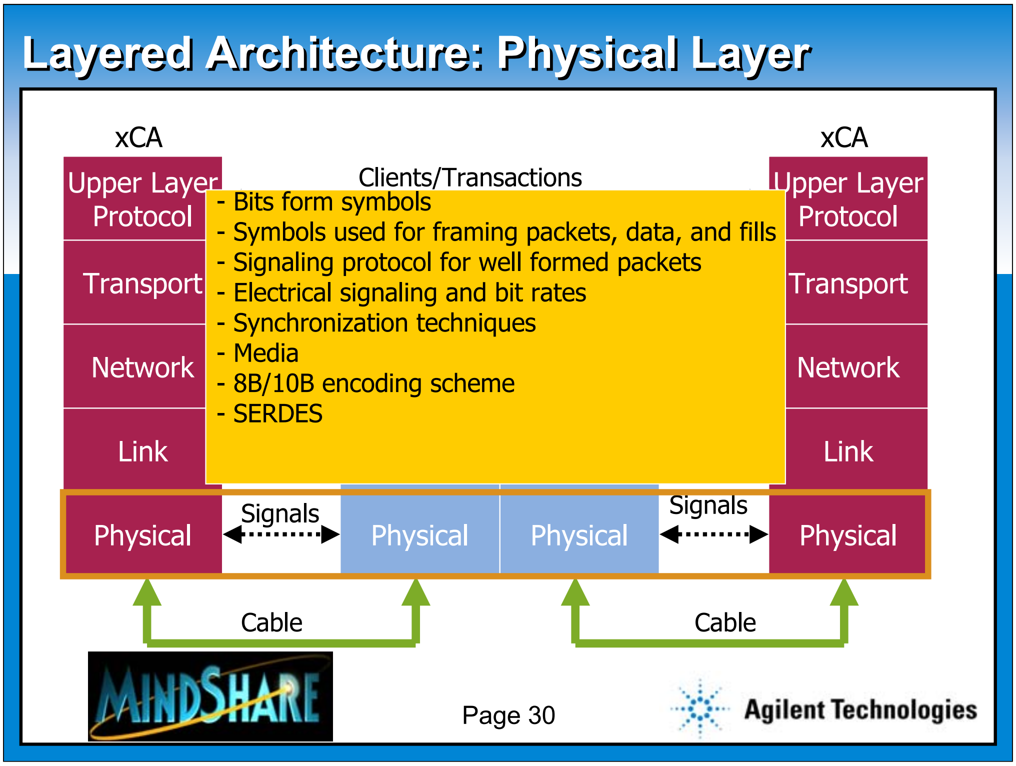
Physical

Cable

Cable

The Link Layer manages the operation of the link. Link layer protocol defines the formats of the packets. Packets are routed to the destination within the subnet via LIDs assigned by the subnet manager. The packet format contains a packet's Service Level (SL) and VL. Link protocol defines an Invariant CRC (ICRC) which does not change as the packet traverses the fabric. A Variant CRC (VCRC) covers all the fields of the packet. These CRC's allow packet errors to be detected. Link protocol detects errors and has standard error recovery. Link layer checks on data correctness, packet ordering and packet drops in the face of transient errors and network congestion.

The link layer also defines packet flow control techniques based on credits. The receiver on each link sends credits to the transmitter on the other end of the link. Credits indicate how many packets can be accepted on a particular VL. Packets will not be sent without credits indicating room in the receiver.



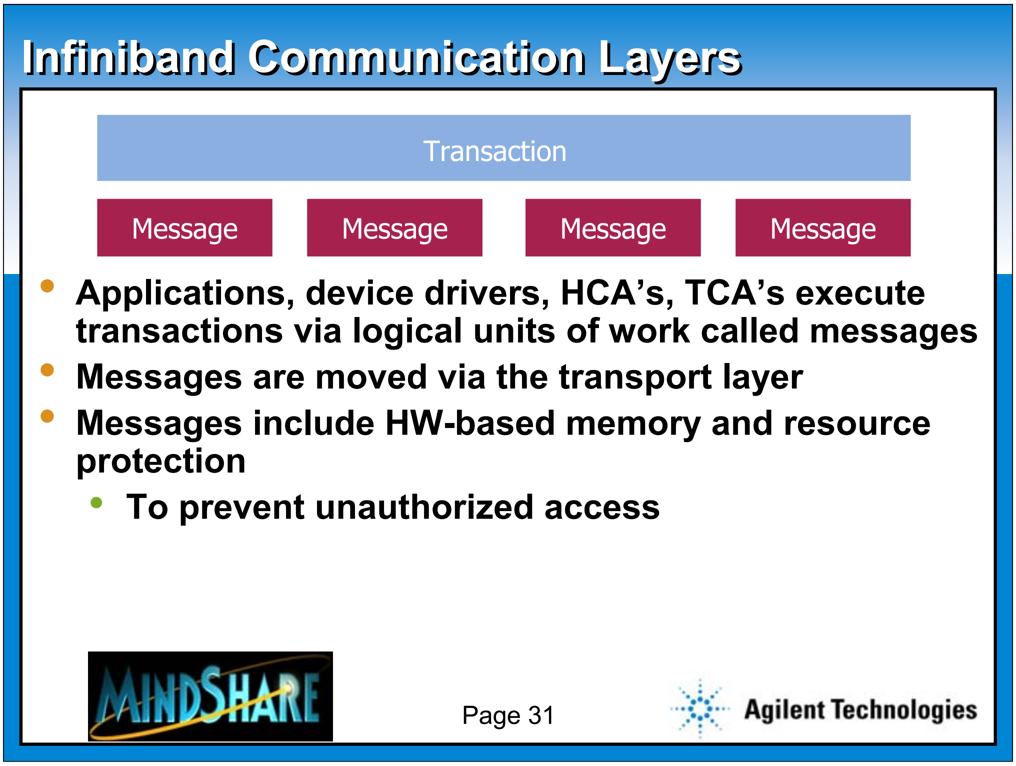
Layered Architecture: Physical Layer

- Bits form symbols
- Symbols used for framing packets, data, and fills
- Signaling protocol for well formed packets
- Electrical signaling and bit rates
- Synchronization techniques
- Media
- 8B/10B encoding scheme
- SERDES

MINDSHARE

Agilent Technologies

The Physical Layer defines how bits on the wire form symbols. Symbols are used to form packets. Symbols are used for framing (Start of packet and end of packet), data (packet content) and fill between packets (Idles). Definitions exist for the signaling protocol for properly formed packets and properly aligned framing symbols. Some errors are defined by the physical layer. Electrical signaling, bit rates, synchronization techniques, media and connectors are also defined by the the physical layer. 8B/10B encoding scheme.



Infiniband Communication Layers

Transaction

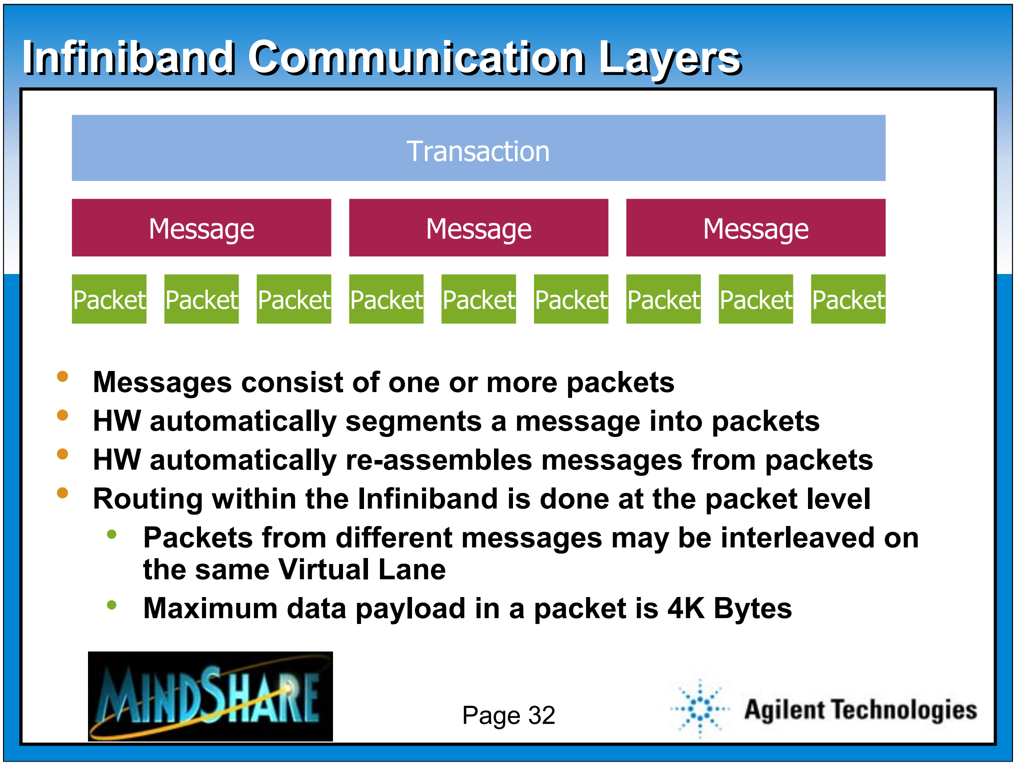
Message | Message | Message | Message

- Applications, device drivers, HCA’s, TCA’s execute transactions via logical units of work called messages
- Messages are moved via the transport layer
- Messages include HW-based memory and resource protection
  - To prevent unauthorized access

MINDSHARE

Agilent Technologies

At the transaction level, a device driver in kernel mode may have a scatter gather list that needs execution to perform an I/O operation. The device driver breaks the transaction into messages moved by the IBA transport layer. The messages will be represented by WQE’s. Each message can be up to 2 GB in size of data. Messages will include memory and resource protection to prevent a driver from unauthorized access to I/O.



Infiniband Communication Layers

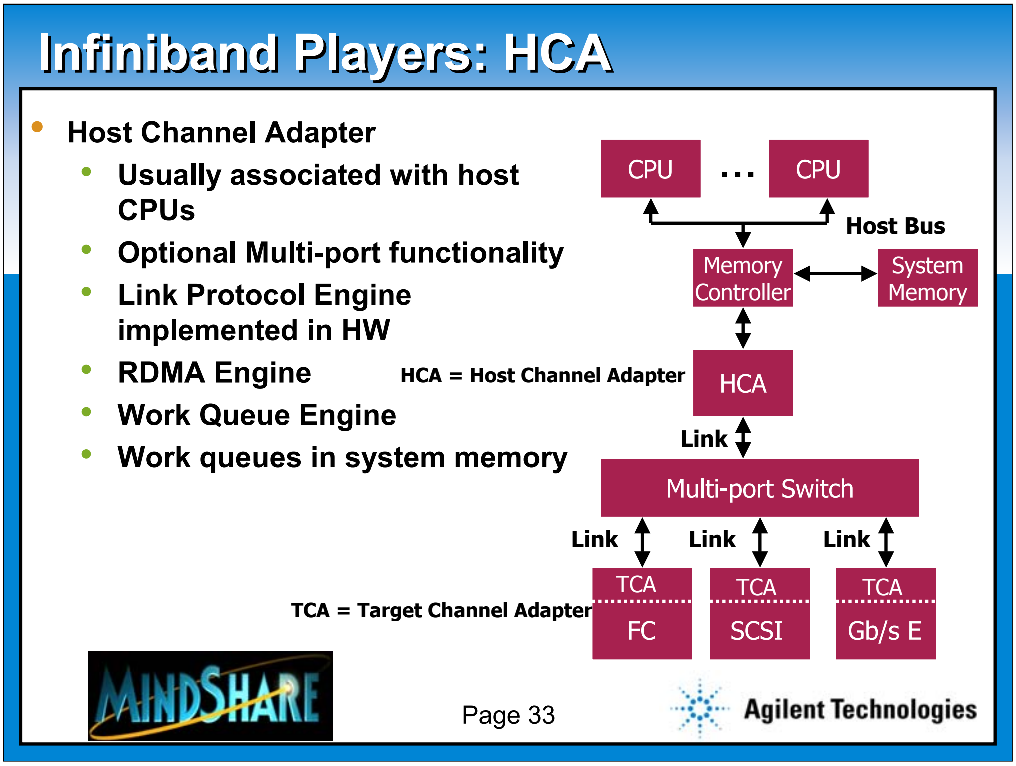
Transaction

Message | Message | Message

Packet Packet Packet Packet Packet Packet Packet Packet Packet

- **Messages consist of one or more packets**
- **HW automatically segments a message into packets**
- **HW automatically re-assembles messages from packets**
- **Routing within the Infiniband is done at the packet level**
  - **Packets from different messages may be interleaved on the same Virtual Lane**
  - **Maximum data payload in a packet is 4K Bytes**

Messages consist of one or more packets. Packets are the IBA routable unit of transfer. Data in packets can be as little as 256 B to 4 KB per. Longer packets will be more efficient, yielding higher BW. However, smaller packets may result in shorter latencies with less head of live blocking. As a long packet is routed, other packets wait longer to be routed. IBA supports auto-negotiated MTU's of 256 B, 512 B, 1024 B, 2048 B, and 4096 B. Automatic segmentation and re-assembly is the process of breaking data to be transferred into smaller parts and then re-assembling the data at the receiving end to reconstruct the original data.



Infiniband Players: HCA

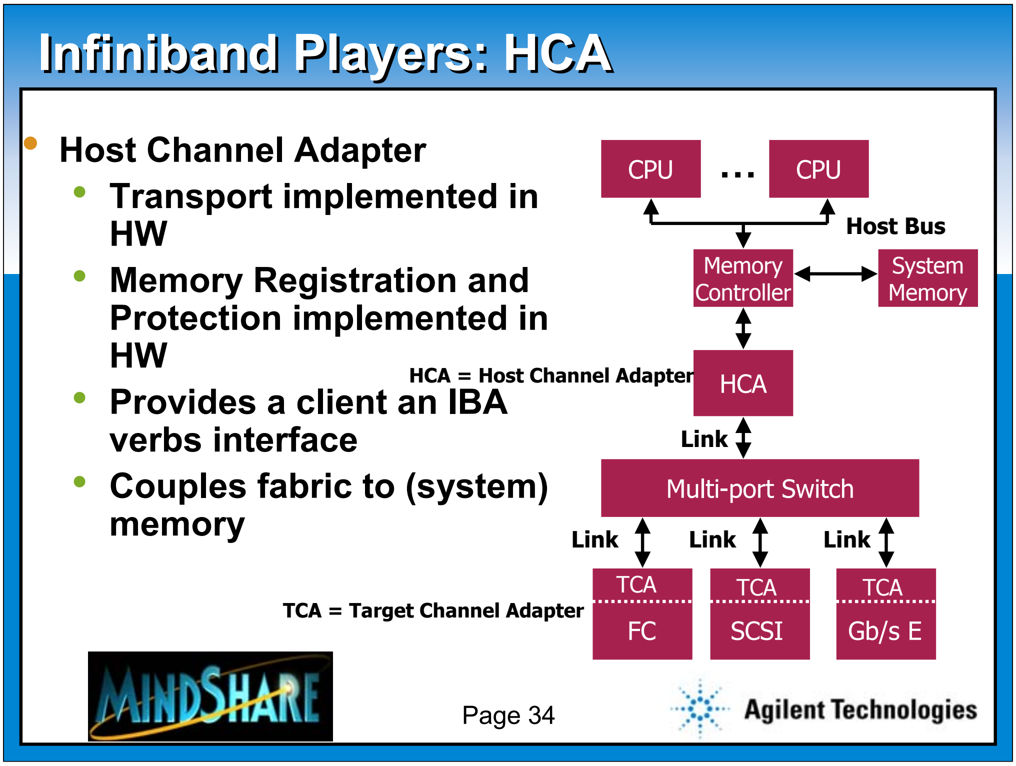
- Host Channel Adapter
  - Usually associated with host CPUs
  - Optional Multi-port functionality
  - Link Protocol Engine implemented in HW
  - RDMA Engine
  - Work Queue Engine
  - Work queues in system memory

HCA = Host Channel Adapter

TCA = Target Channel Adapter

A host channel adapter (HCA) provides a connection between the system memory controller and Infiniband switched fabric. The HCA may be single or multi-ported. Multi-ported HCA's could be used for enhanced reliability (redundancy), performance (bundling), and/or connectivity (multiple attachment points). The HCA provides support for the Link Protocol Engine in hardware, meaning no SW intervention is required for data movement. This engine understands the low-level, packet communication model. Link protocol engine detects errors and has standard error recovery. The HW checks on data correctness, packet ordering and packet drops in the face of transient errors and network congestion.

An RDMA engine will allow bulk data transfers to occur to or from system memory, independent of host CPU interaction. The work queue engine will handle the posted work requests and create the work completion queue entries. The work queues themselves are located in system memory. The HCA manages the transport layer in hardware. This means that the HCA understands the message types and the packet-to-message translations. The HCA converts a virtual address provided by the RDMA to a physical memory location in system memory. Access rights are also checked by the HCA hardware for any transfer to or from system memory. Key to an HCA is its support of the verbs interface to client processes.

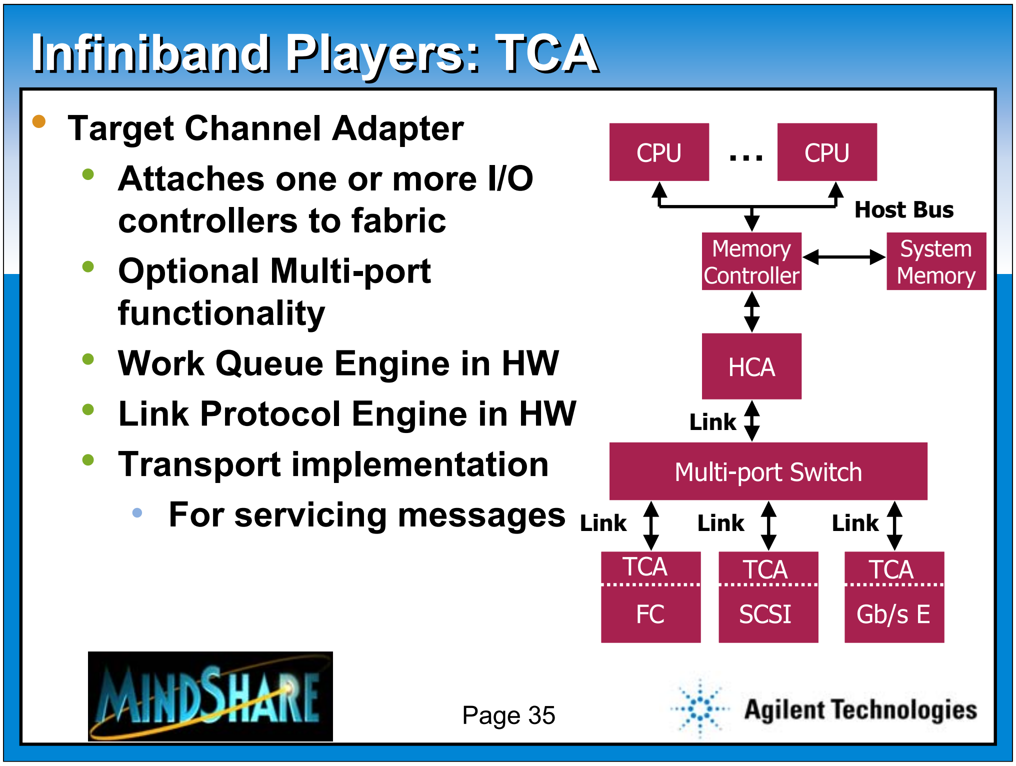


Infiniband Players: HCA

- Host Channel Adapter
  - Transport implemented in HW
  - Memory Registration and Protection implemented in HW
  - Provides a client an IBA verbs interface
  - Couples fabric to (system) memory

A host channel adapter (HCA) provides a connection between the system memory controller and Infiniband switched fabric. The HCA may be single or multi-ported. Multi-ported HCA’s could be used for enhanced reliability (redundancy), performance (bundling), and/or connectivity (multiple attachment points). The HCA provides support for the Link Protocol Engine in hardware, meaning no SW intervention is required for data movement. This engine understands the low-level, packet communication model. Link protocol engine detects errors and has standard error recovery. The HW checks on data correctness, packet ordering and packet drops in the face of transient errors and network congestion.

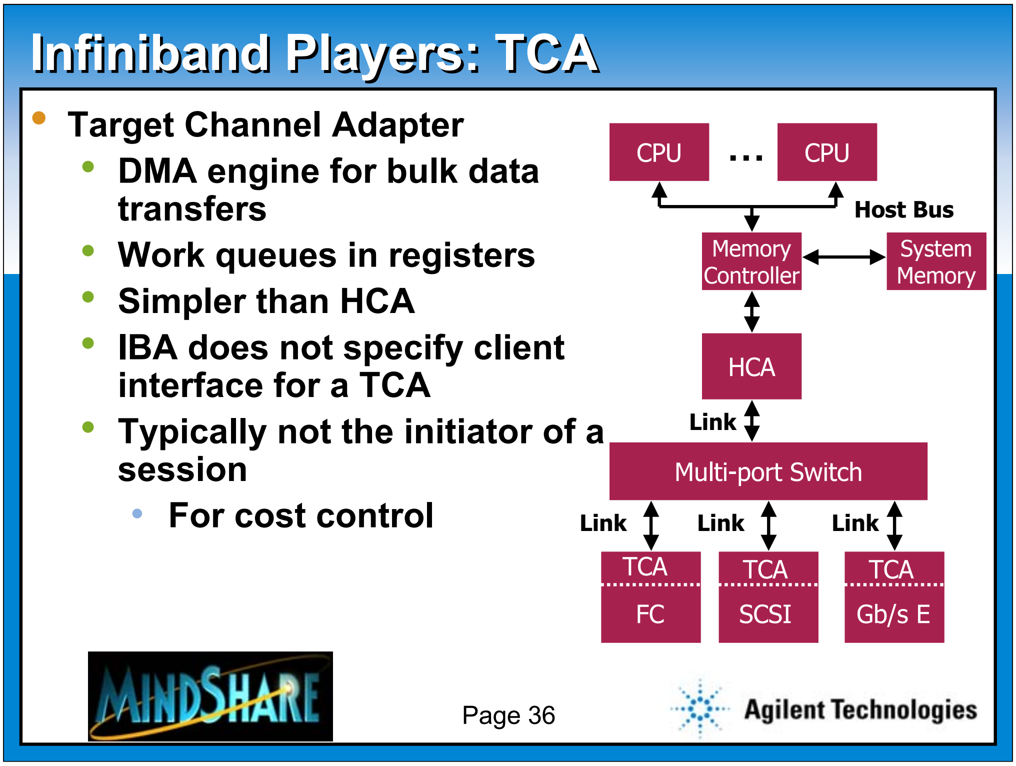
An RDMA engine will allow bulk data transfers to occur to or from system memory, independent of host CPU interaction. The work queue engine will handle the posted work requests and create the work completion queue entries. The work queues themselves are located in system memory. The HCA manages the transport layer in hardware. This means that the HCA understands the message types and the packet-to-message translations. The HCA converts a virtual address provided by the RDMA to a physical memory location in system memory. Access rights are also checked by the HCA hardware for any transfer to or from system memory. Key to an HCA is its support of the verbs interface to client processes.



Infiniband Players: TCA

- Target Channel Adapter
  - Attaches one or more I/O controllers to fabric
  - Optional Multi-port functionality
  - Work Queue Engine in HW
  - Link Protocol Engine in HW
  - Transport implementation
    - For servicing messages

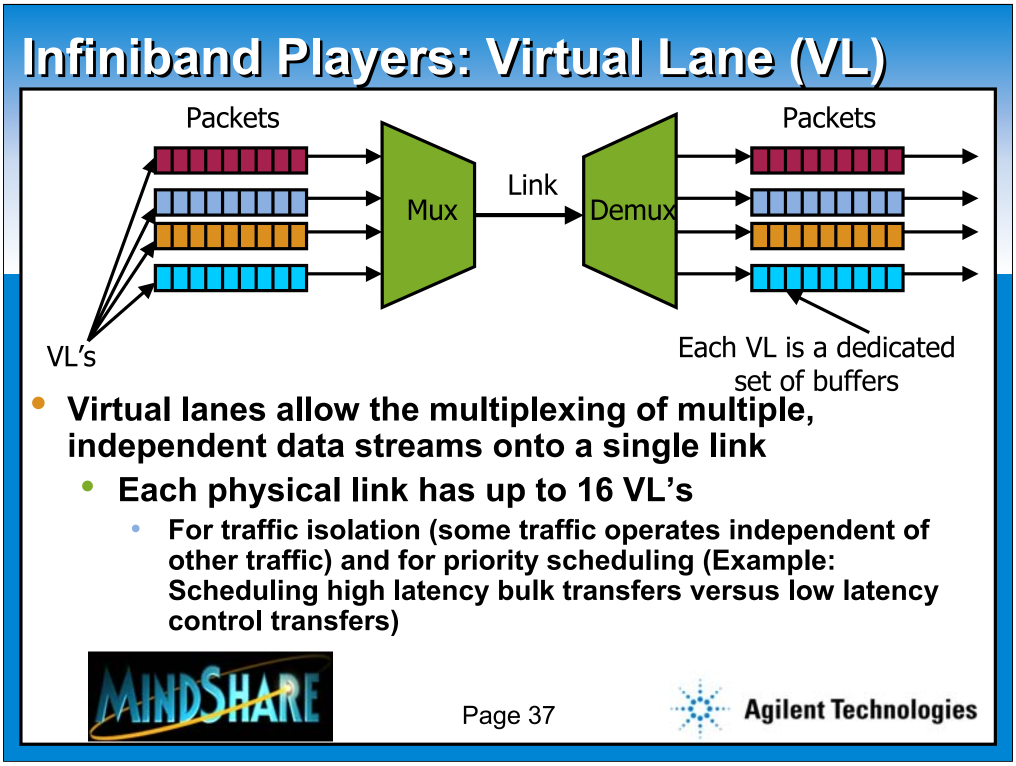
A target channel adapter (TCA) is similar to a PCI bus interface for an I/O controller, but much more sophisticated. The TCA provides the interface between the specific I/O controller and the Infiniband serial cable fabric. Protocols implemented by the I/O controller could be standard network or storage, or proprietary. A TCA may provide multiple ports for redundancy or performance. Analogous to a PCI multi-function device, a single TCA may be shared by multiple I/O controllers. Like an HCA, the TCA provides the Work Queue, Link Protocol, and DMA engines in hardware. A transport implementation is also provided by the TCA to allow the TCA to interpret messages. The work queues will be implemented in hardware registers or fifo's as opposed to implementing them in local memory. Typically, a TCA will be simpler in design relative to an HCA. The TCA only needs to implement the minimum set of features required to effectively attach the controller to the fabric. The TCA will also natively support peer-to-peer transfers in hardware.



Infiniband Players: TCA

- Target Channel Adapter
  - DMA engine for bulk data transfers
  - Work queues in registers
  - Simpler than HCA
  - IBA does not specify client interface for a TCA
  - Typically not the initiator of a session
    - For cost control

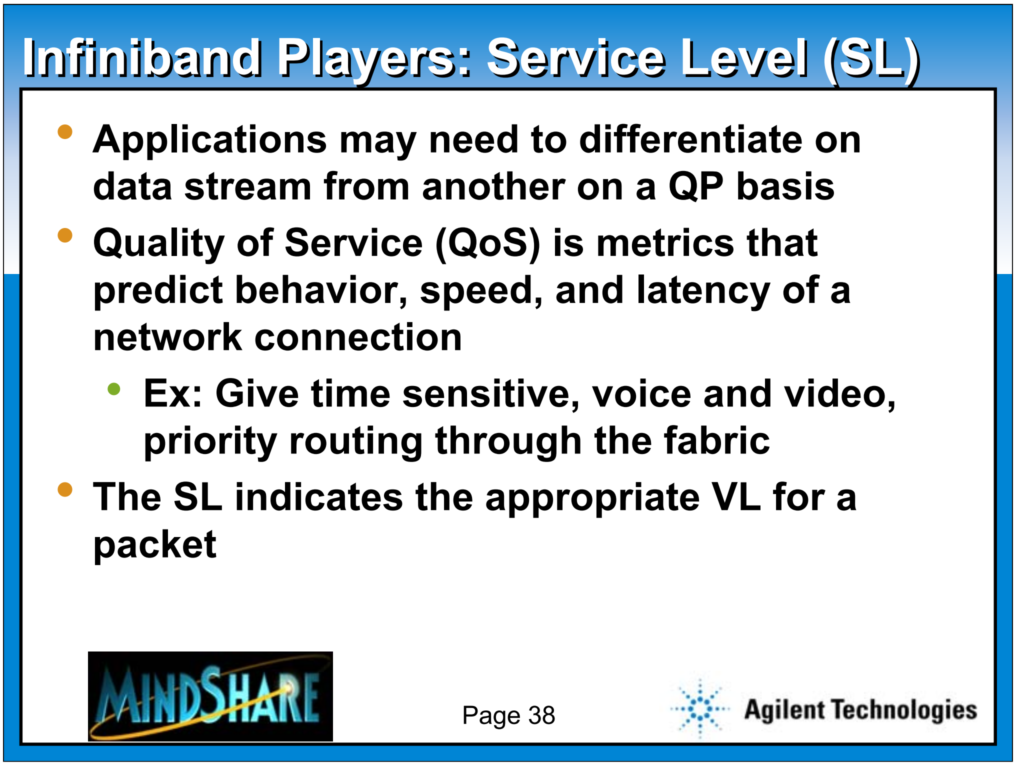
A target channel adapter (TCA) is similar to a PCI bus interface for an I/O controller, but much more sophisticated. The TCA provides the interface between the specific I/O controller and the Infiniband serial cable fabric. Protocols implemented by the I/O controller could be standard network or storage, or proprietary. A TCA may provide multiple ports for redundancy or performance. Analogous to a PCI multi-function device, a single TCA may be shared by multiple I/O controllers. Like an HCA, the TCA provides the Work Queue, Link Protocol, and DMA engines in hardware. A transport implementation is also provided by the TCA to allow the TCA to interpret messages. The work queues will be implemented in hardware registers or fifo's as opposed to implementing them in local memory. Typically, a TCA will be simpler in design relative to an HCA. The TCA only needs to implement the minimum set of features required to effectively attach the controller to the fabric. The TCA will also natively support peer-to-peer transfers in hardware.



Infiniband Players: Virtual Lane (VL)

- Virtual lanes allow the multiplexing of multiple, independent data streams onto a single link
  - Each physical link has up to 16 VL's
    - For traffic isolation (some traffic operates independent of other traffic) and for priority scheduling (Example: Scheduling high latency bulk transfers versus low latency control transfers)

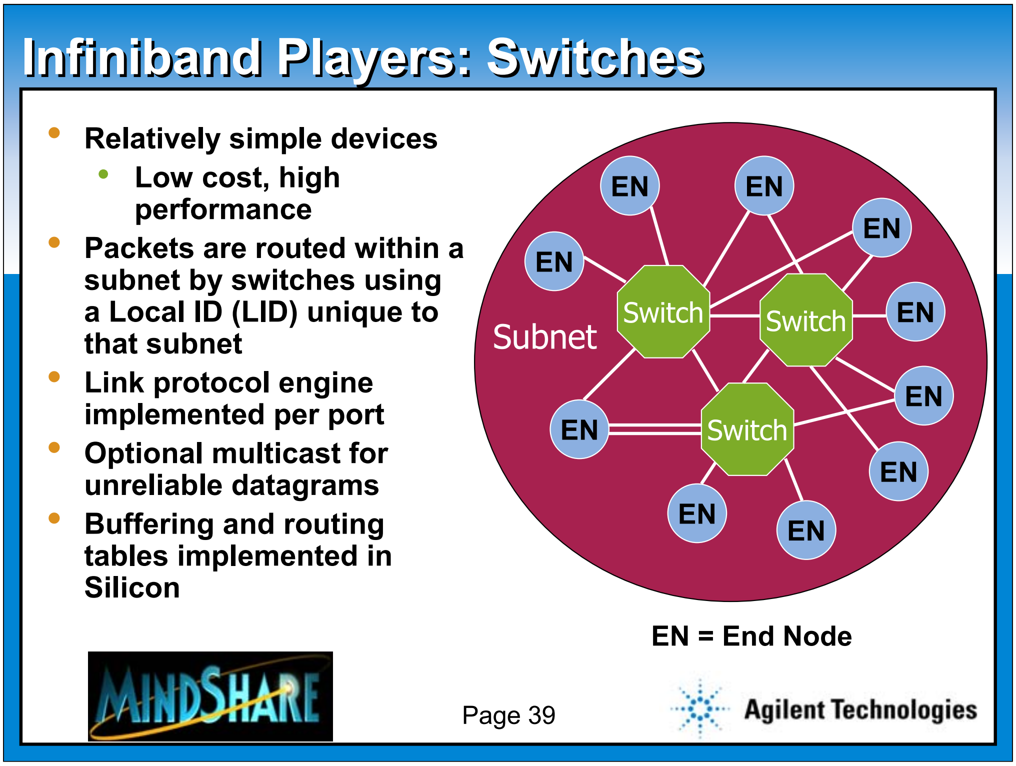
Virtual Lanes are a method of providing independent data streams on the same physical link. Multiple "virtual" links or lanes are multiplexed onto a single link. There are up to 16 possible VL's per link. The VL's that are actually used by a port are configured by the subnet manager (SM) and depends on a packet's service level (SL). A port for a link has an SL to VL mapping table programmed by the SM. Flow control is maintained by the port separately for each VL, as each VL has a dedicated set of data buffers. If there is heavy traffic on a particular VL, traffic is not blocked assigned to another VL.



Infiniband Players: Service Level (SL)

- Applications may need to differentiate on data stream from another on a QP basis
- Quality of Service (QoS) is metrics that predict behavior, speed, and latency of a network connection
  - Ex: Give time sensitive, voice and video, priority routing through the fabric
- The SL indicates the appropriate VL for a packet

SL's allow the implementation of differentiated services. Quality of service is metrics that predict the behavior, speed and latency of a network connection. Quality of service example: Give time-sensitive, voice and video data priority routing through the fabric. The service level (SL) indicates the appropriate VL for a packet.



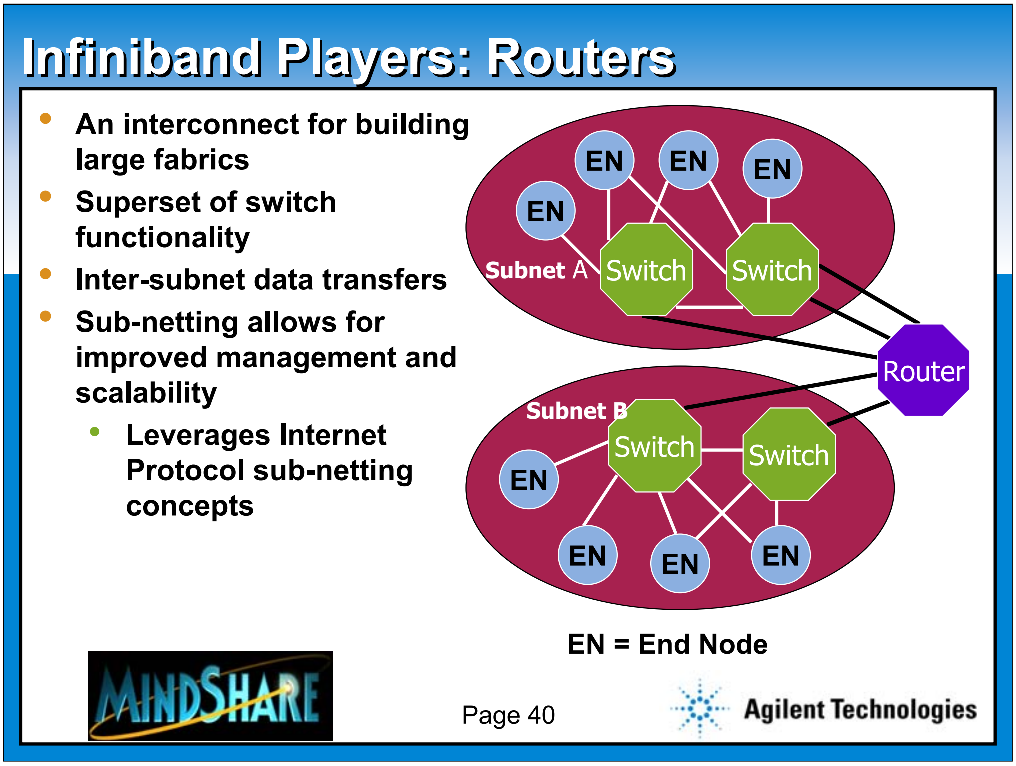
Infiniband Players: Switches

- Relatively simple devices
  - Low cost, high performance
- Packets are routed within a subnet by switches using a Local ID (LID) unique to that subnet
- Link protocol engine implemented per port
- Optional multicast for unreliable datagrams
- Buffering and routing tables implemented in Silicon

EN
EN
EN
EN
Subnet
Switch
Switch
EN
EN
Switch
EN
EN
EN
EN

EN = End Node

The architecture of an Infiniband switch is intended to yield high performance, low cost design. Thus, the fabrics created with these switches will be cost effective. A switch routes packets from one link to another within the same subnet using the DLID in the Local Route Header.

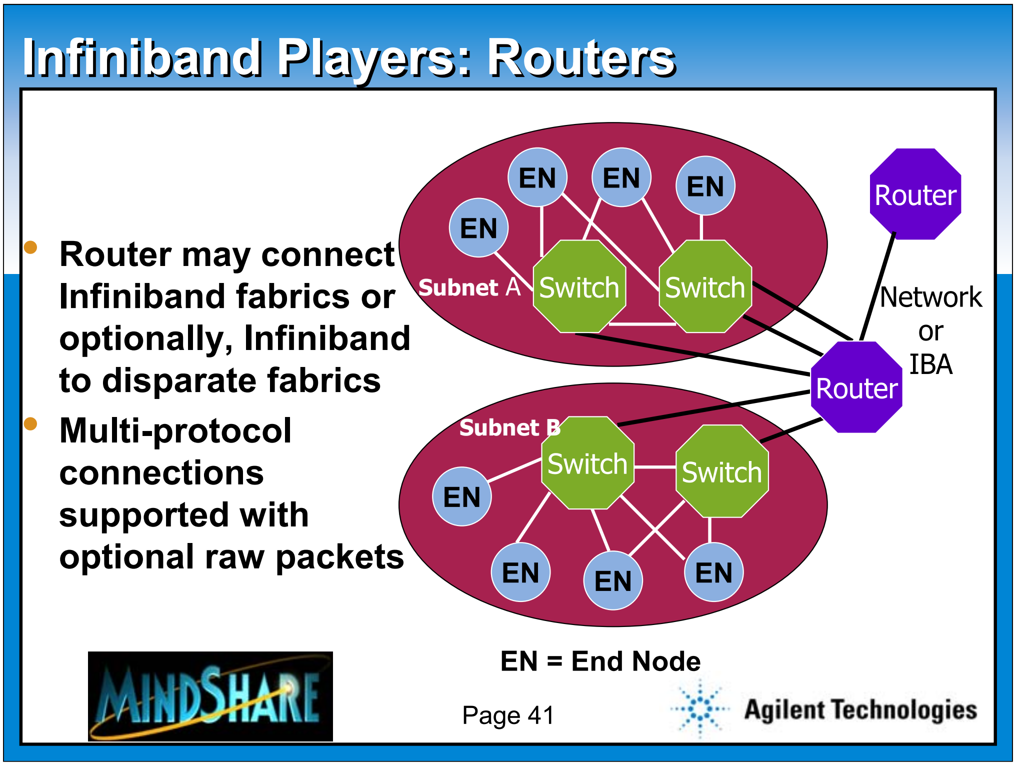


Infiniband Players: Routers

- An interconnect for building large fabrics
- Superset of switch functionality
- Inter-subnet data transfers
- Sub-netting allows for improved management and scalability
  - Leverages Internet Protocol sub-netting concepts

EN = End Node

A router is a special case switch for joining subnets. Administration domain is at the subnet. Routers bridge subnets, but also provide subnet isolation.



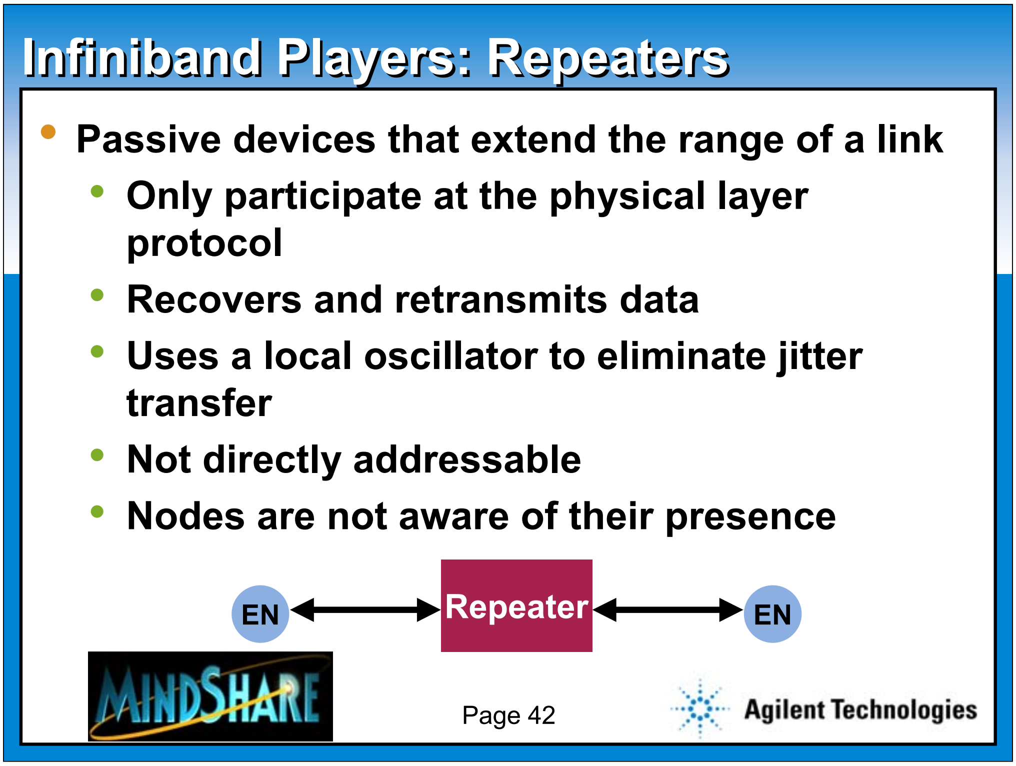
Infiniband Players: Routers

- Router may connect Infiniband fabrics or optionally, Infiniband to disparate fabrics
- Multi-protocol connections supported with optional raw packets

EN = End Node

Agilent Technologies

A router may optionally provide IBA fabric to disparate fabric routing. This is a powerful concept for global internetworking. Raw packet formats supported for this particular application.

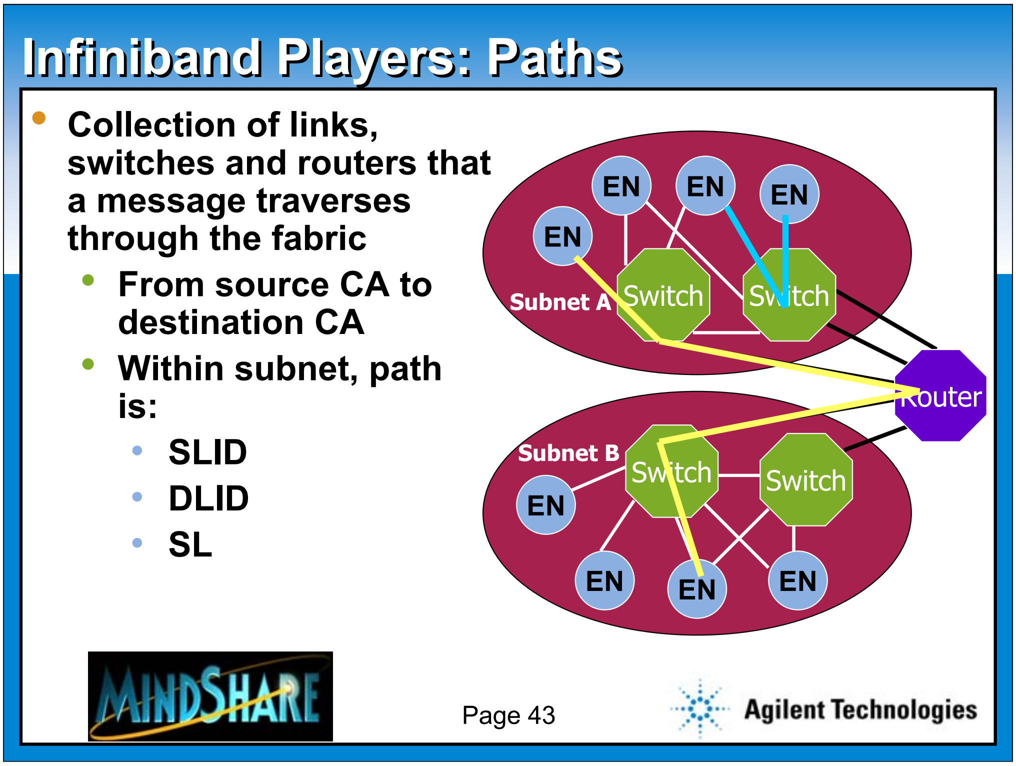


Infiniband Players: Repeaters

- Passive devices that extend the range of a link
  - Only participate at the physical layer protocol
  - Recovers and retransmits data
  - Uses a local oscillator to eliminate jitter transfer
  - Not directly addressable
  - Nodes are not aware of their presence

EN
Repeater
EN

Repeaters are transparent devices that extend the range of the link. Transparent in the sense that repeaters only participate at the physical layer of the protocol. A repeater is not directly addressable. Nodes of the link are not aware of their presence.



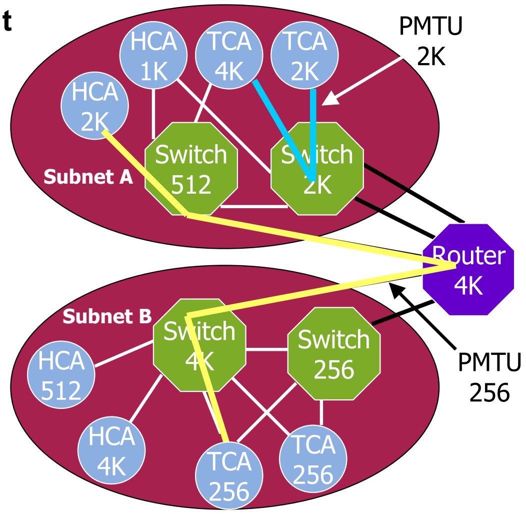
Infiniband Players: Paths

- Collection of links, switches and routers that a message traverses through the fabric
  - From source CA to destination CA
  - Within subnet, path is:
    - SLID
    - DLID
    - SL

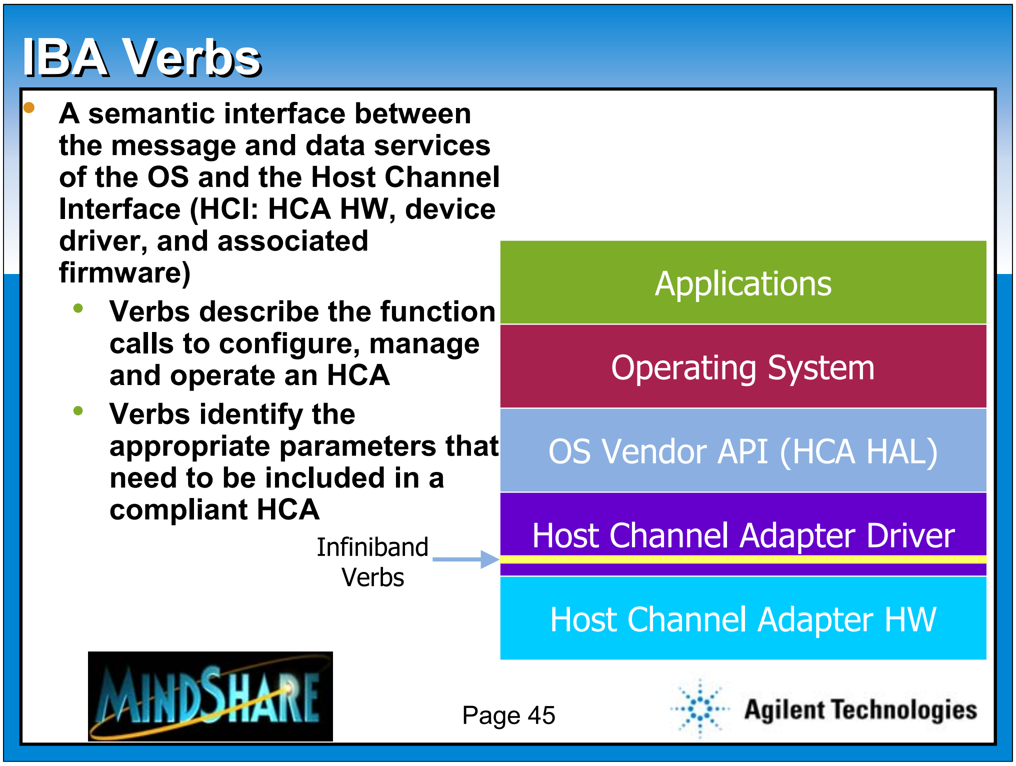
A message traverses through the fabric on a particular path. The path is the collection of links, switches and routes used to deliver the message from the source channel adapter to end channel adapter. To a switch, a LID represents the path through the switch, connecting one port to another. To a router, the path through a router is represented by the subnet prefix portion of the GID.

Infiniband Players: Path MTU

- Maximum Transmission Unit (MTU)
  - Maximum data payload through a port
  - 256, 512, 1K, 2K, and 4K Bytes
  - MTU does not include additional 126 Bytes of required header buffering
  - PMTU is minimum MTU through all ports on a given path
    - Not restricted to a single subnet



Page 44



IBA Verbs

- A semantic interface between the message and data services of the OS and the Host Channel Interface (HCI: HCA HW, device driver, and associated firmware)
  - Verbs describe the function calls to configure, manage and operate an HCA
  - Verbs identify the appropriate parameters that need to be included in a compliant HCA

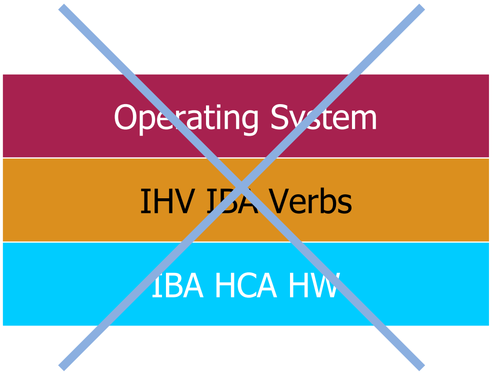
Applications
Operating System
OS Vendor API (HCA HAL)
Host Channel Adapter Driver
Infiniband Verbs
Host Channel Adapter HW

MINDSHARE

Agilent Technologies

IBA Verbs

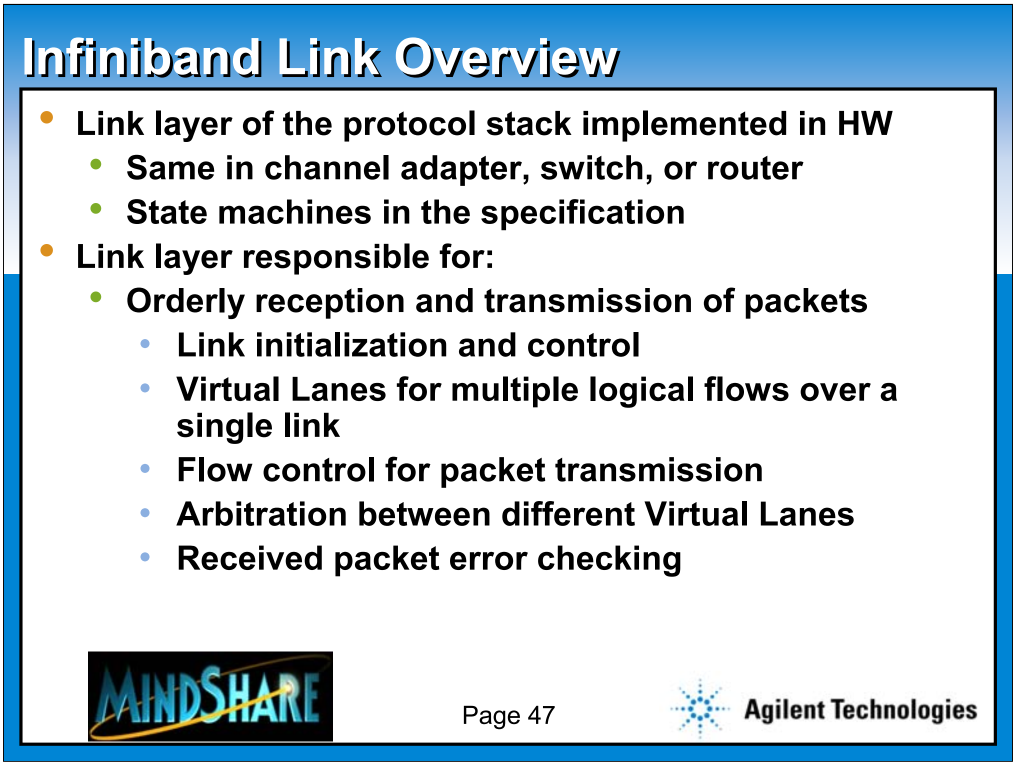
- An abstract description of the functionality of an HCA
  - Semantic behavior of the HW
- NOT AN API
  - No registers defined
- May influence API's
  - Verbs are intended to inspire OSV API's at the equivalent semantic level
    - Not required
    - OSV's may expose all or some of the HCA functionality via the API as necessary (TBD)



Operating System

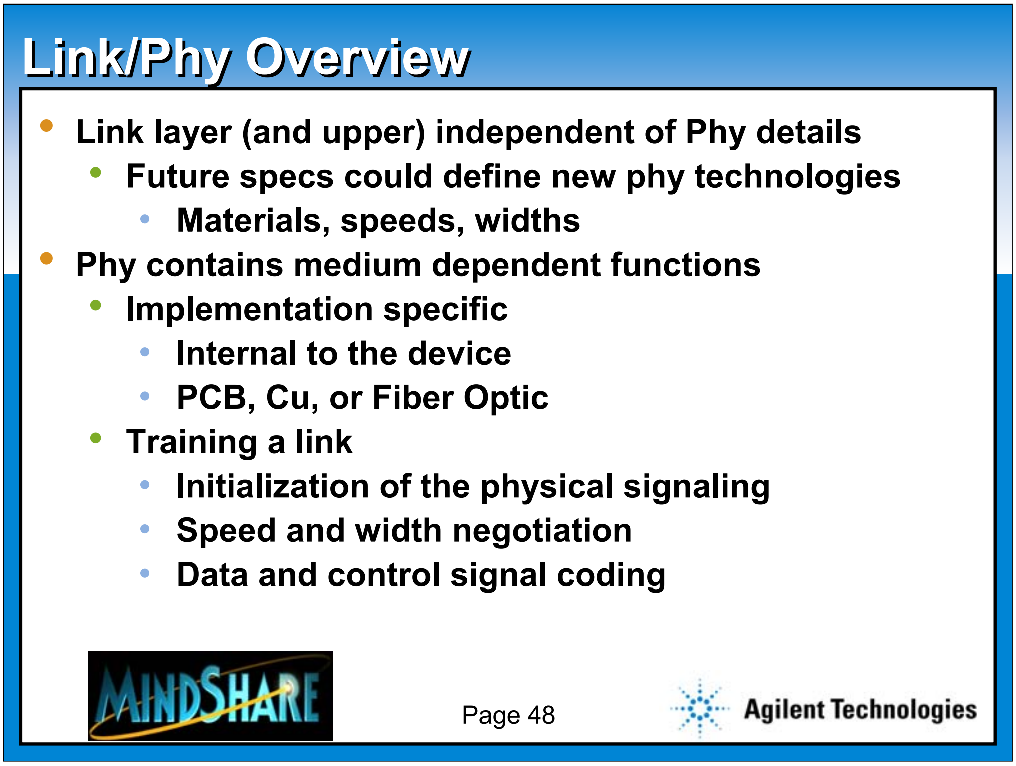
IHV IBA Verbs

IBA HCA HW



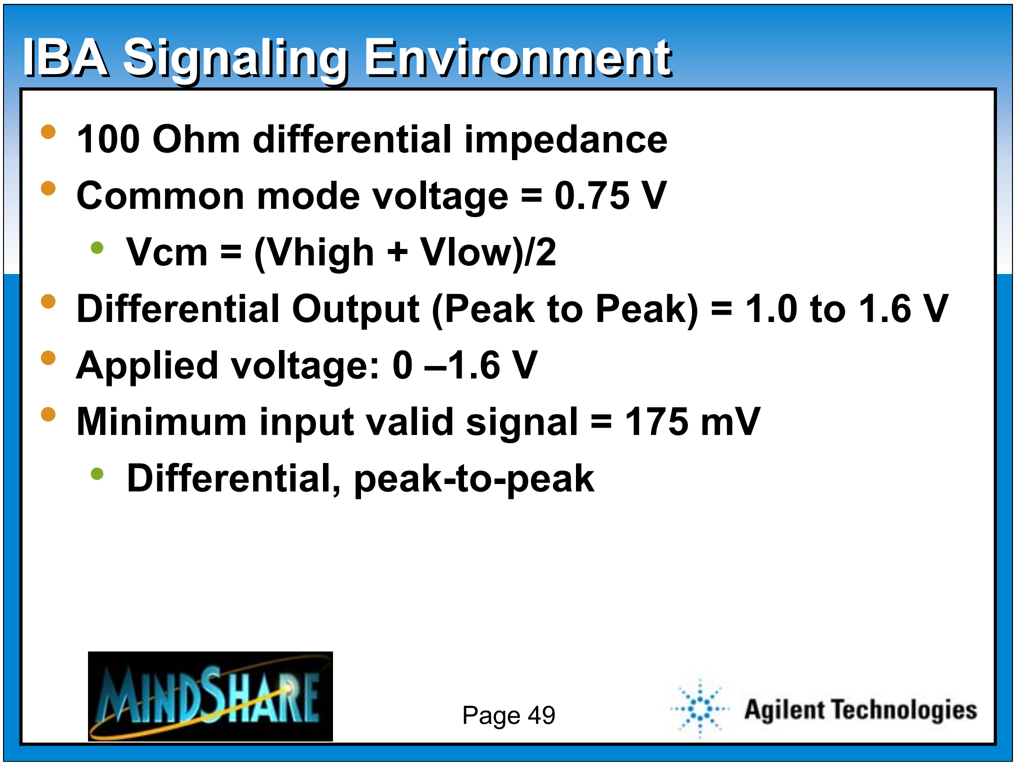
Infiniband Link Overview

- Link layer of the protocol stack implemented in HW
  - Same in channel adapter, switch, or router
  - State machines in the specification
- Link layer responsible for:
  - Orderly reception and transmission of packets
    - Link initialization and control
    - Virtual Lanes for multiple logical flows over a single link
    - Flow control for packet transmission
    - Arbitration between different Virtual Lanes
    - Received packet error checking



Link/Phy Overview

- Link layer (and upper) independent of Phy details
  - Future specs could define new phy technologies
    - Materials, speeds, widths
- Phy contains medium dependent functions
  - Implementation specific
    - Internal to the device
    - PCB, Cu, or Fiber Optic
  - Training a link
    - Initialization of the physical signaling
    - Speed and width negotiation
    - Data and control signal coding



IBA Signaling Environment

- 100 Ohm differential impedance
- Common mode voltage = 0.75 V
  - Vcm = (Vhigh + Vlow)/2
- Differential Output (Peak to Peak) = 1.0 to 1.6 V
- Applied voltage: 0 –1.6 V
- Minimum input valid signal = 175 mV
  - Differential, peak-to-peak

MINDSHARE

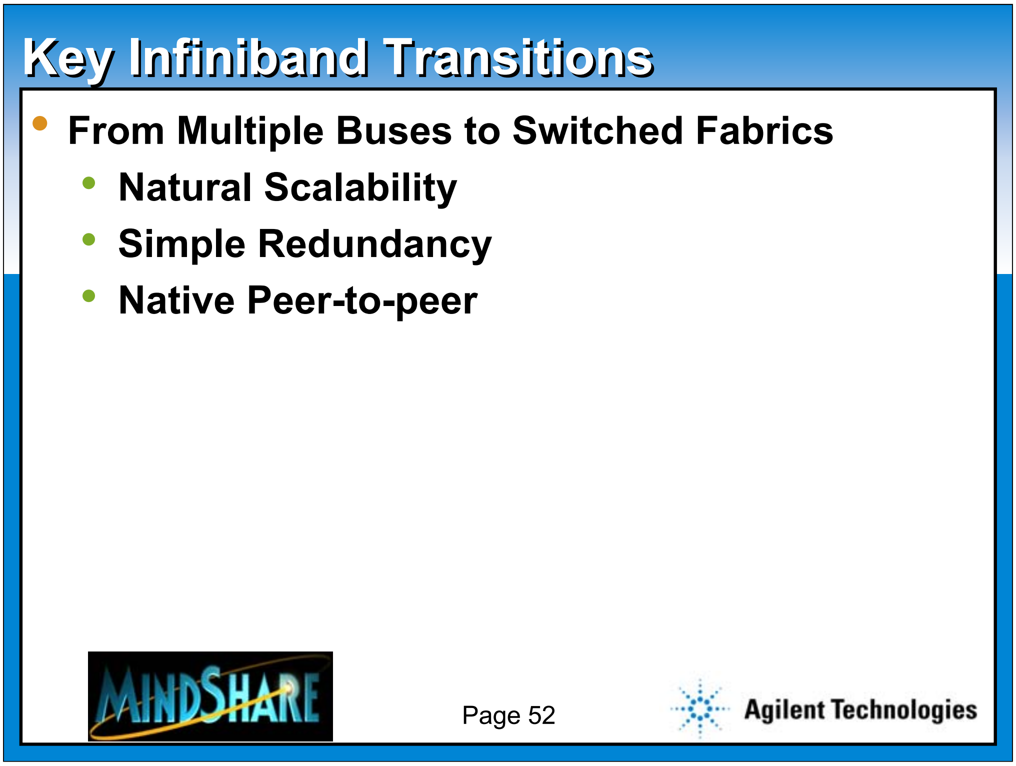
Agilent Technologies

Key Infiniband Transitions

- From Load/Store to Send/Receive
  - Data transfer requests queued in work queues
  - Dedicated Work Queue Engine to work on queues

Key Infiniband Transitions

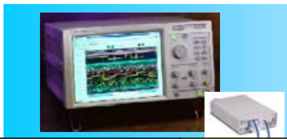
- From Shared Bus to Point-to-point Links
  - Simplifies management
    - Configuration
    - Error Recovery
  - Best in:
    - Price/performance
    - Performance
    - Distance
    - RAS
    - Electricals
    - Mechanicals



Key Infiniband Transitions

- From Multiple Buses to Switched Fabrics
  - Natural Scalability
  - Simple Redundancy
  - Native Peer-to-peer

Tools For InfiniBand Validation



16700 Logic Analyzer
cross-bus analysis for root
cause analysis

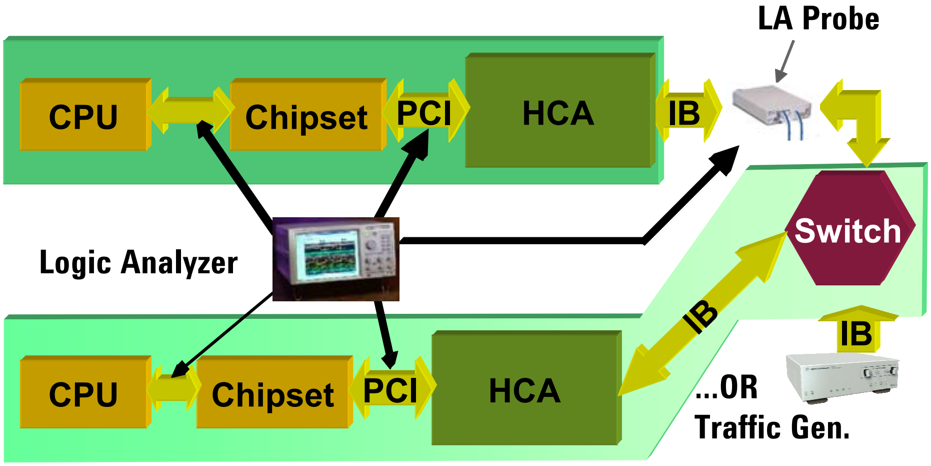
Port Physical	Port Link Level	Network	Transport	Appl.

DCA 86100 / TDR

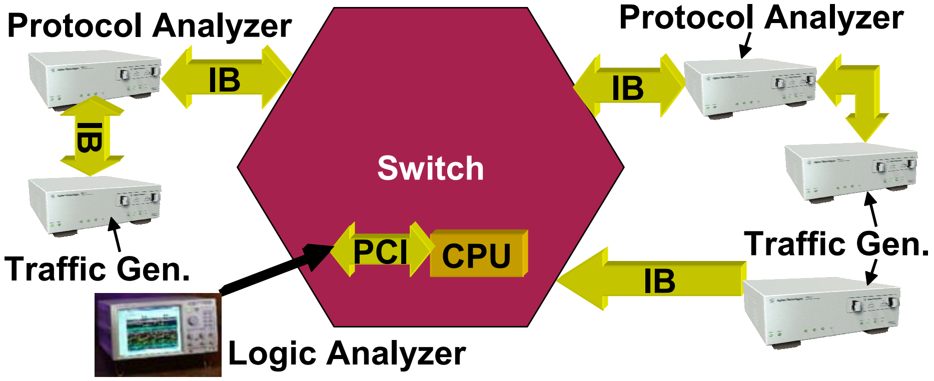
81250 / PHY stimulus
SERDES test, pBERT,

E2950 Series
- Analyzer
- Traffic Generator

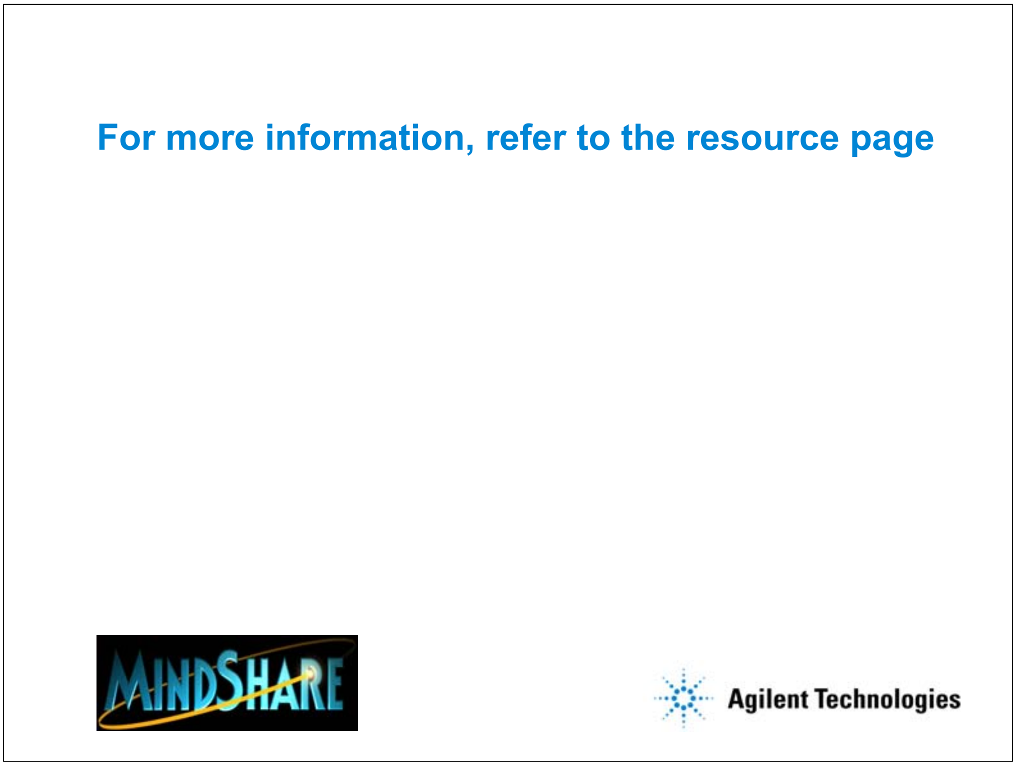
System Level Validation Platform



Fabric Validation Platform



Protocol Analysis Suits Fabric Validation



For more information, refer to the resource page

Infiniband System Architecture class from Mindshare.

Next Infiniband NetSeminar:

August 1

Focus will be on Infiniband physical layer characterization using differential TDR and vector network analysis

MINDSHARE

Agilent Technologies

Infiniband System Architecture class from Mindshare.

THANK YOU!

Infiniband System Architecture class from Mindshare.