



InfiniBridge™: An Integrated InfiniBand Switch and Channel Adapter

Chris Eddington

Director of Technical Marketing

Mellanox Technologies

chrise@mellanox.com

Mellanox Technologies, Inc.



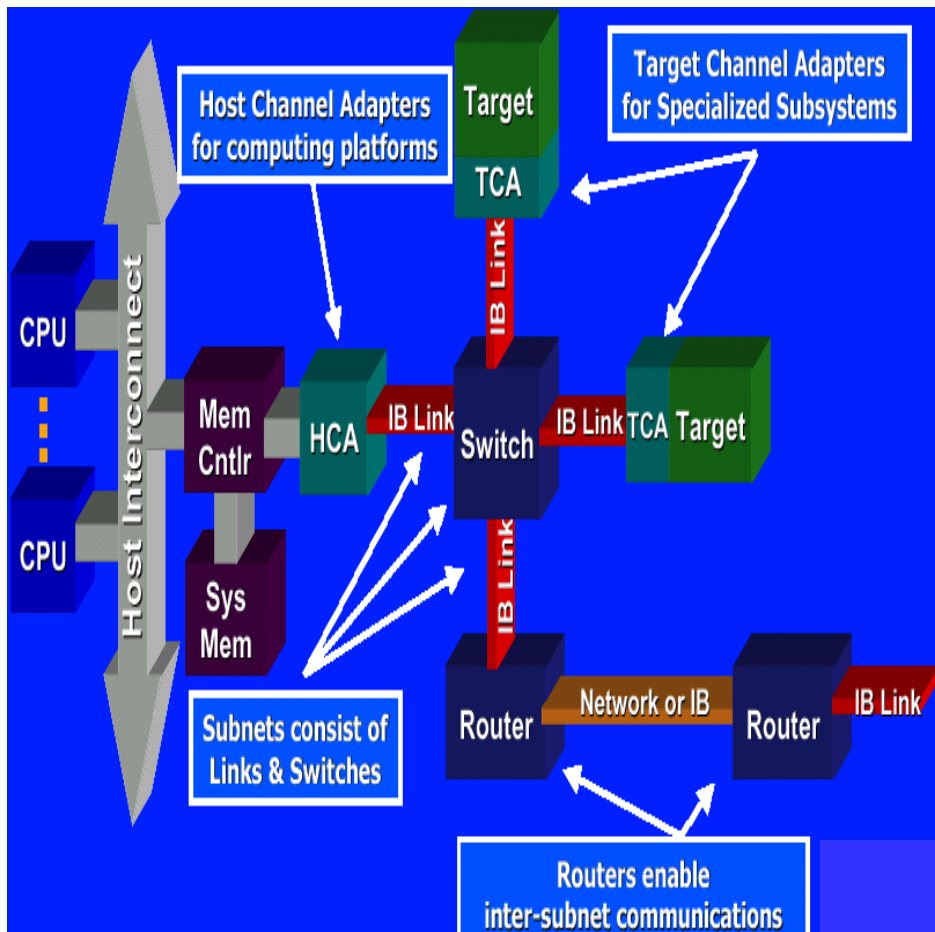


Agenda

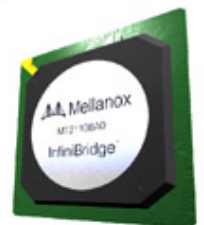
- **InfiniBand Overview**
- **Virtual Lanes and Virtual Fabrics**
- **Network Stack and Reliable Connections**
- **Virtual Interface Architecture**
- **InfiniBridge™ Transport Protocol Engines**
- **InfiniPCI Technology**
- **Summary**



InfiniBand Switch Fabric

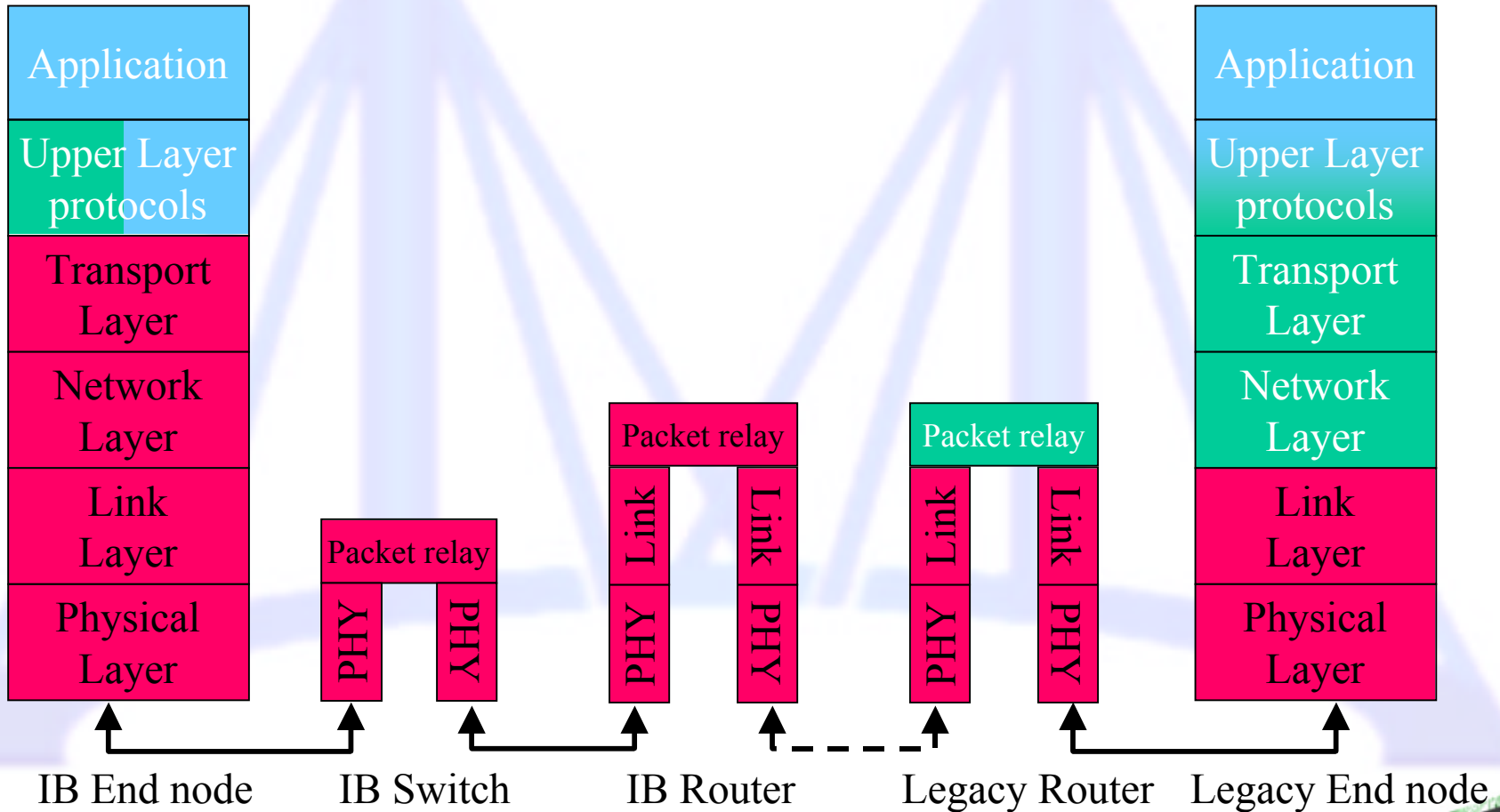
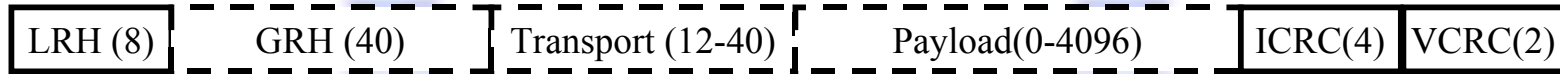
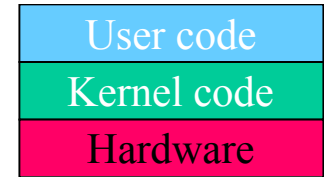


- HCA (Host Channel Adaptor)
 - Connects a CPU to the InfiniBand Fabric
- TCA (Target Channel Adaptor)
 - Connects I/O controllers such as Ethernet, SCSI, Fibre Channel to InfiniBand
- Switches:
 - Basic building block of InfiniBand Subnets
- Routers:
 - Connect IB subnets
 - Connect IB to SAN / LAN / WAN





Network stack





Packet Format

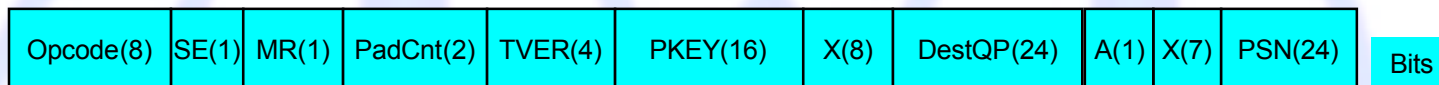
General IB Request Packet Structure



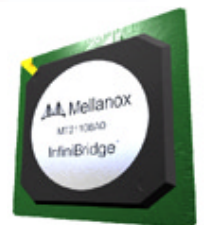
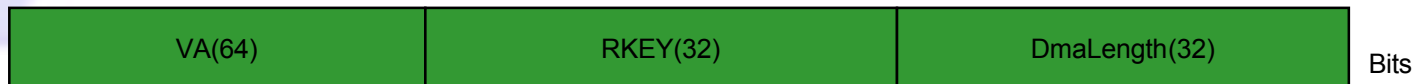
LRH – Local Route Header (8 Bytes)



BTH – Base Transport Header (12 Bytes)

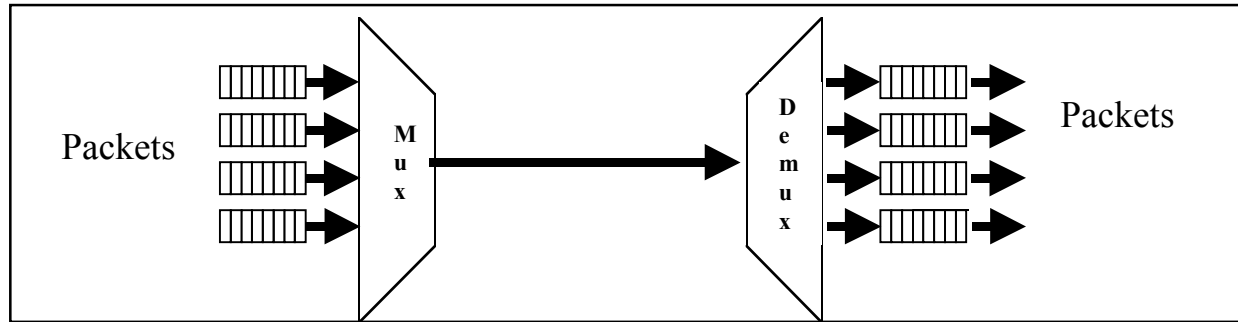


RETH – RDMA Extended Transport Header (16 Bytes)



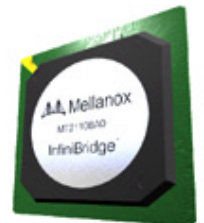
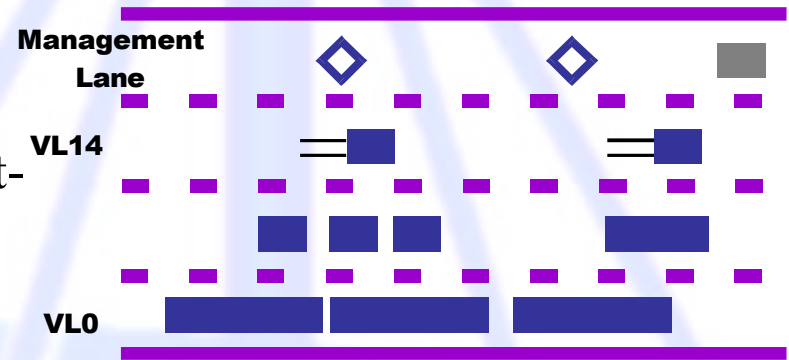


Virtual Lanes



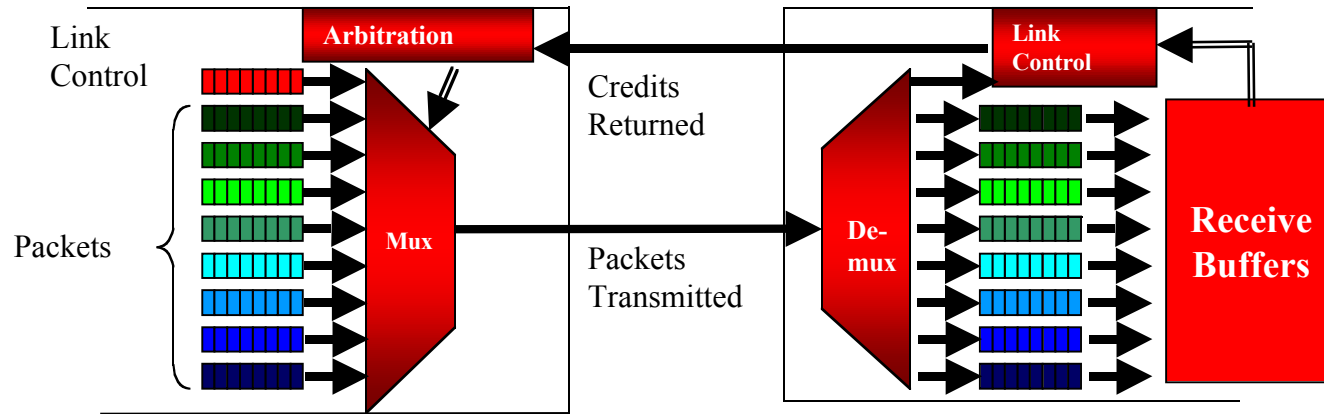
● Multiplex independent data streams onto a single physical link:

- Dedicated management lane
- Differentiated services on a packet-boundary basis
- Alleviates head-of-line blocking
- Allow VL-based load balancing across multiple paths

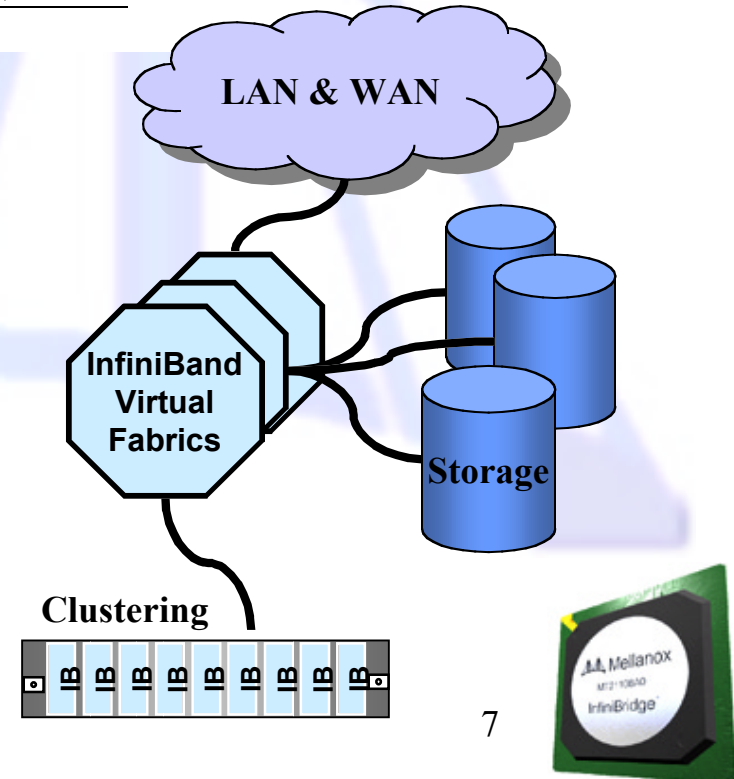




Link Level Flow Control



- Credit-based link-level flow control
 - Link Receivers grant packet receive buffer space credits per VL
- Separate flow control per VL enables Virtual Fabrics
 - Multiple protocols on a unified physical network
 - Congestion and latency on one VL does not impact traffic with guaranteed QoS on another VL

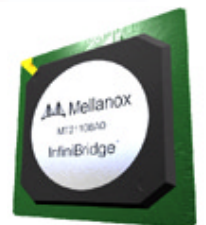




Reliable Transport

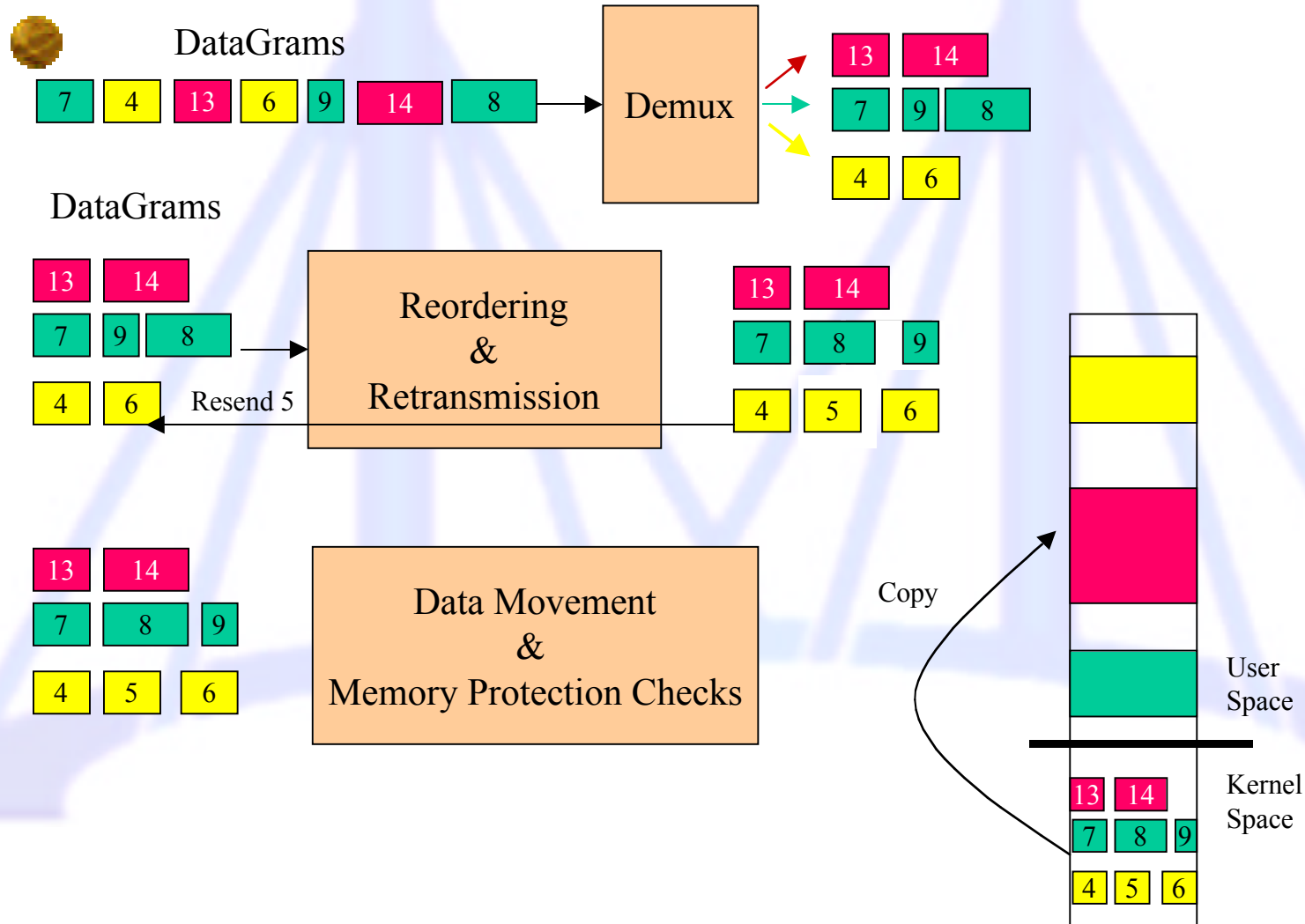
● What makes a Reliable Connection?

- **Reliability (Acknowledgement)**
 - Packets must be in-order
 - No missing packets
 - Flow Control – prevents end point buffer overflow
- **Connections**
 - End to end associations between user space processes (called sockets in TCP/IP)
 - Requires de-multiplexing of datagrams
- **Putting the data where it needs to be**
 - Message copying from kernel to user space

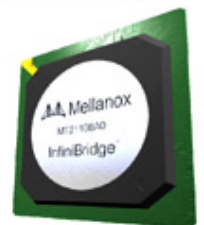
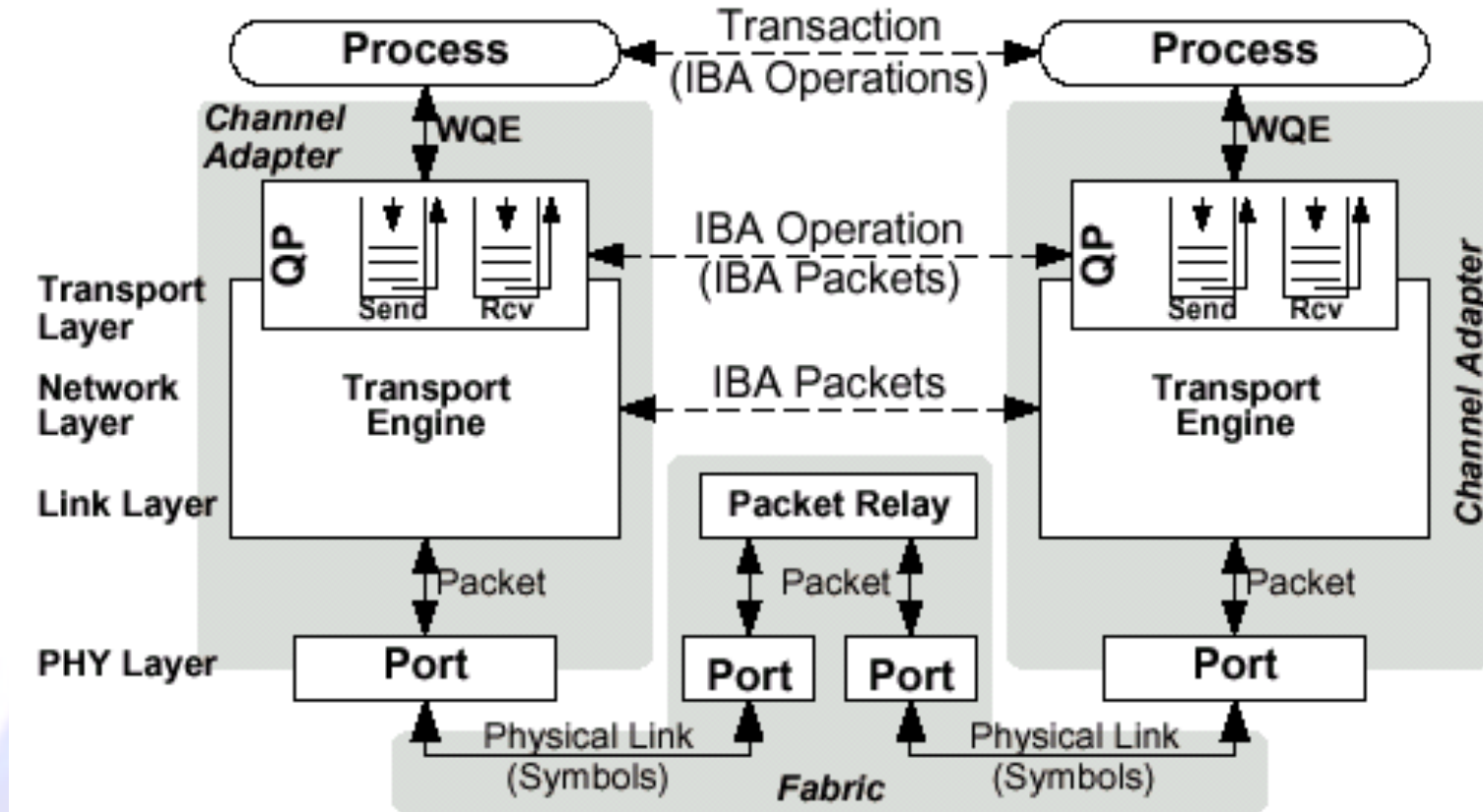




Reliable Transport



InfiniBand and Virtual Interface





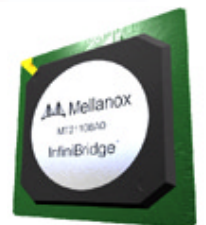
InfiniBridge™ Features

● Mellanox InfiniBridge™ MT21108

- Integrated channel adapter and switch

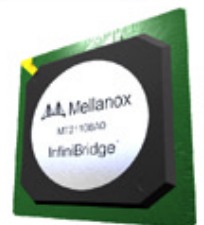
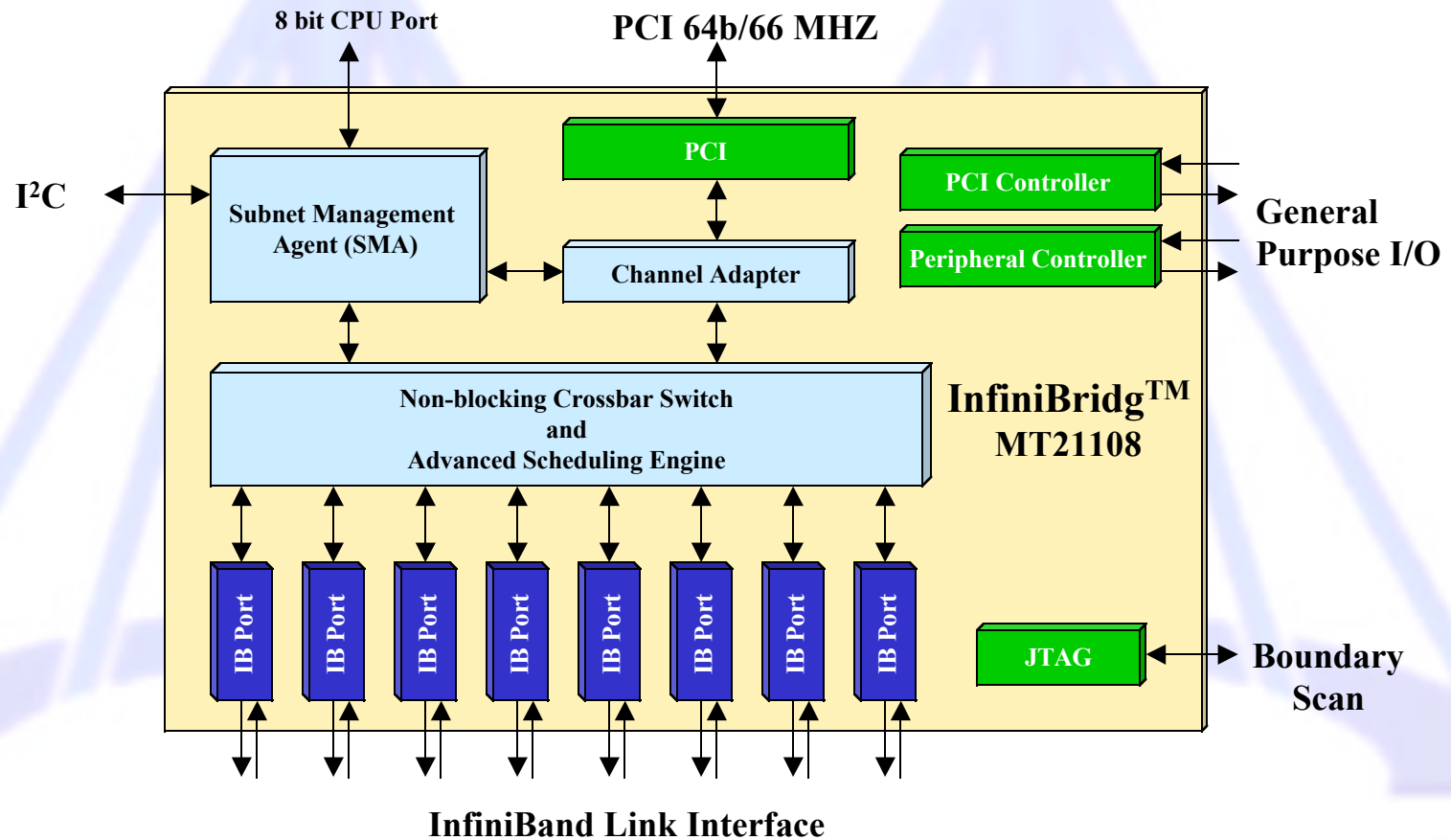
● Key Features:

- Supports both 1X (2.5Gb/s) and 4X (10Gb/s) InfiniBand Links
- Hardware Transport Protocol Engines deliver reliable in-order connection
- Multiple Virtual Lanes plus a Dedicated Management Lane
- Multicast Support
- Maximum Transfer Unit (MTU) up to 2K/4K bytes
- Greater than 100 Gb/s Internal Bandwidth
- InfiniPCI™: Transparent PCI-to-PCI Bridge





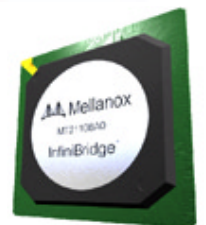
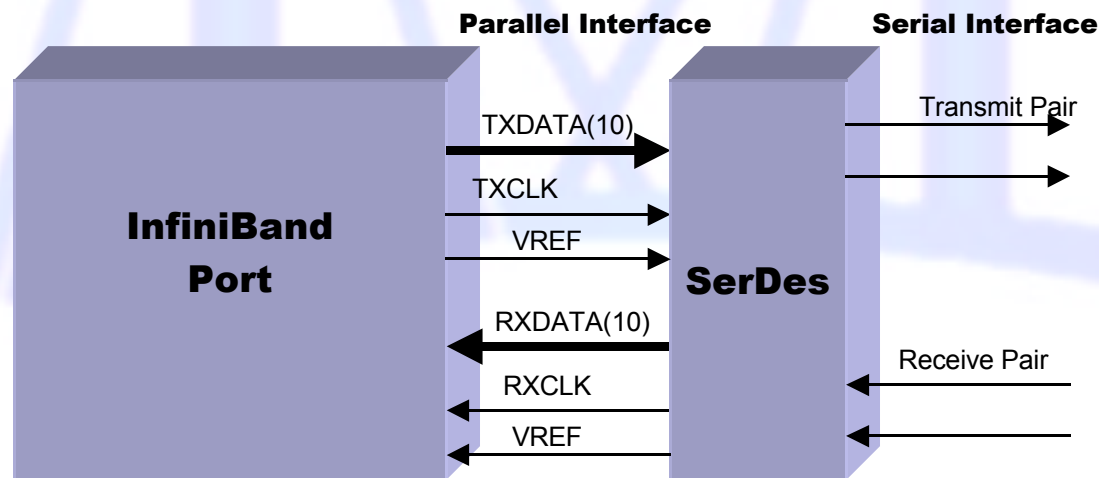
InfiniBridge™ High Level Block Diagram





InfiniBand Port Logic

- Only serial interface defined by InfiniBand TA
 - SerDes use Parallel interface interface to ASIC
 - Point to point, 125MHz, source synchronous, DDR
 - SSTL2
 - 10 pins + clock and reference voltage in each direction

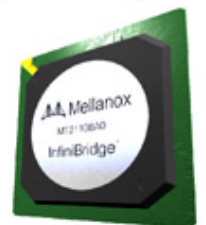




InfiniBand Switch

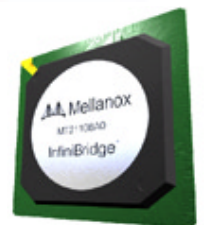
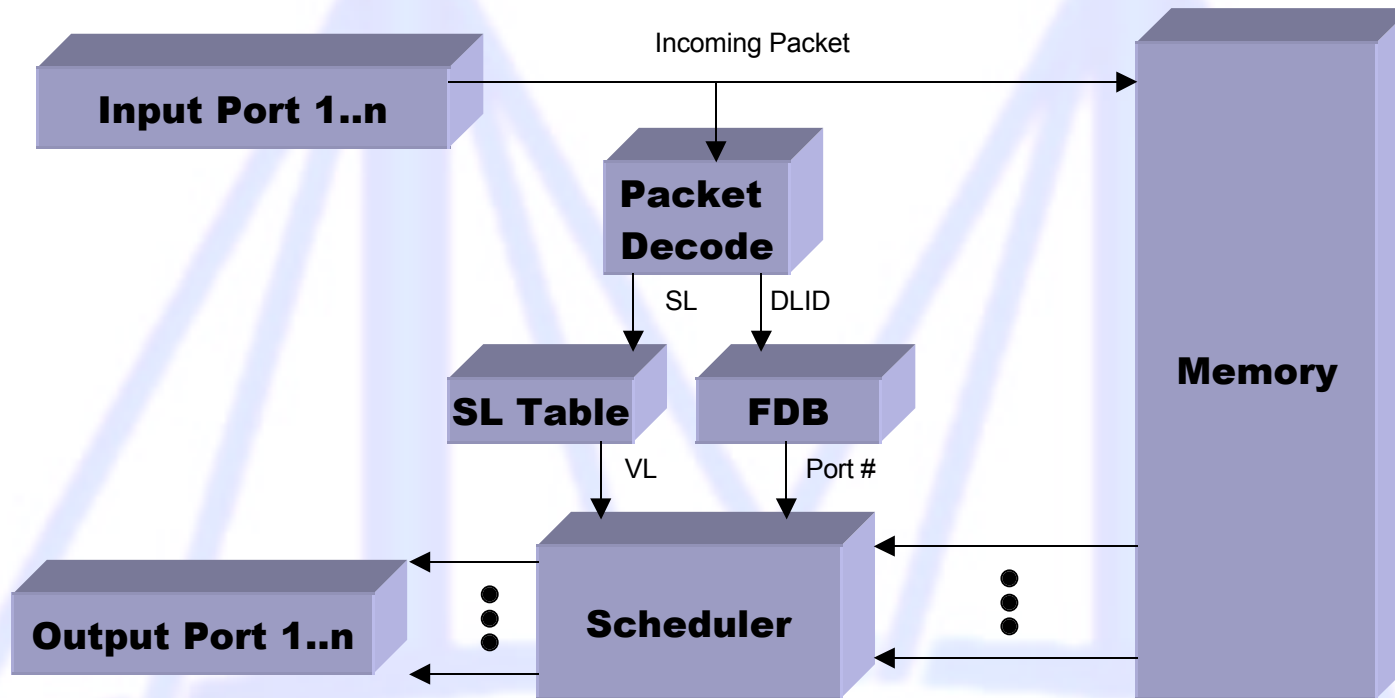
● Layer 2 Forwarding

- Decode Incoming Packet Header (LRH) to get DLID and SL
- Lookup destination port in FDB (Forwarding Database)
- Lookup VL from SL
- Output scheduler decides priority based on VL and integrity checks



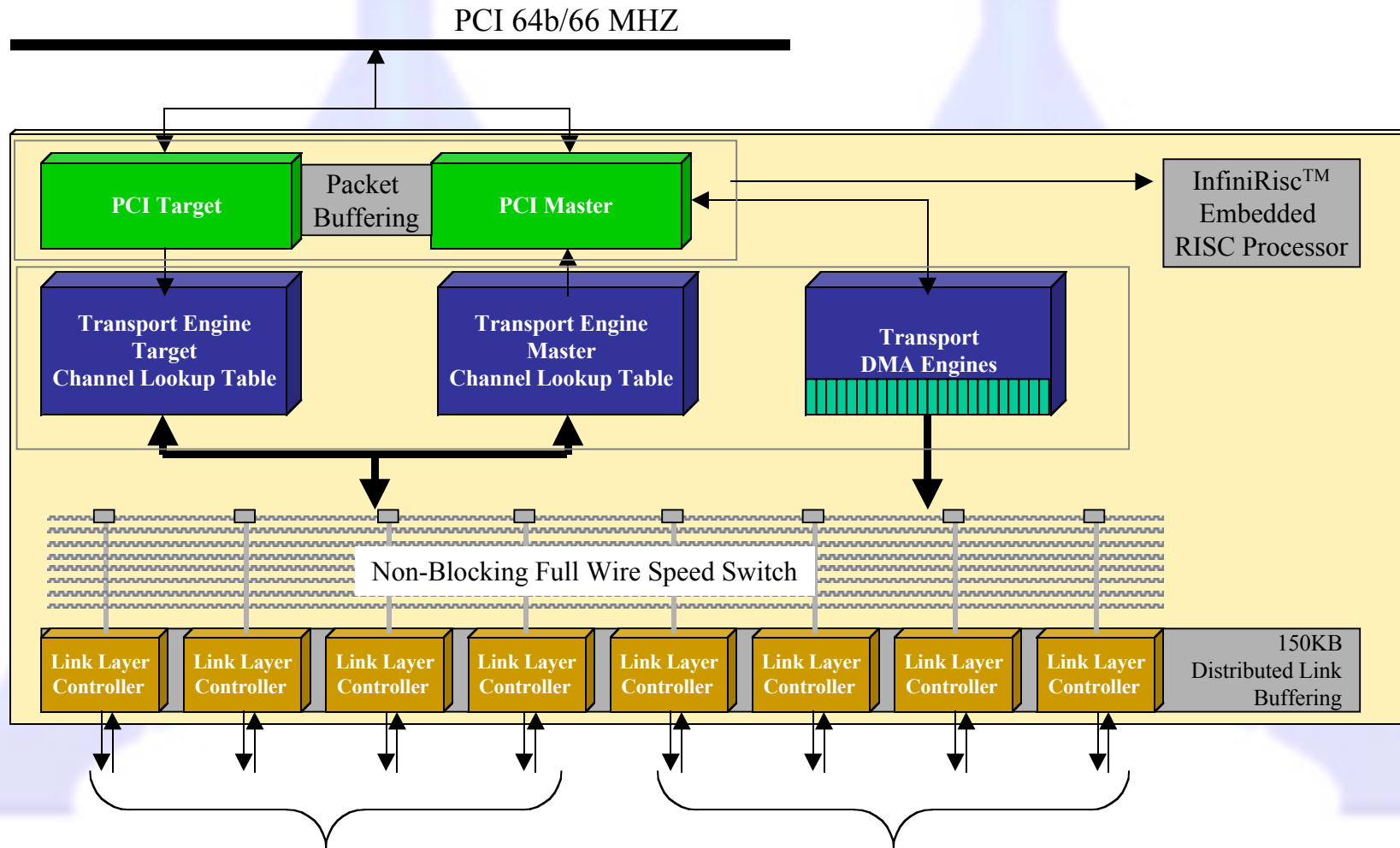


InfiniBand Switch (cont.)

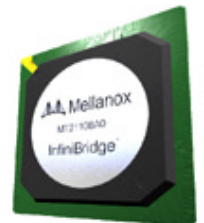




InfiniBridge™ Transport Engine Block Diagram



Four 1X links may be optionally bonded together to form a 4X (10Gb/s) Link

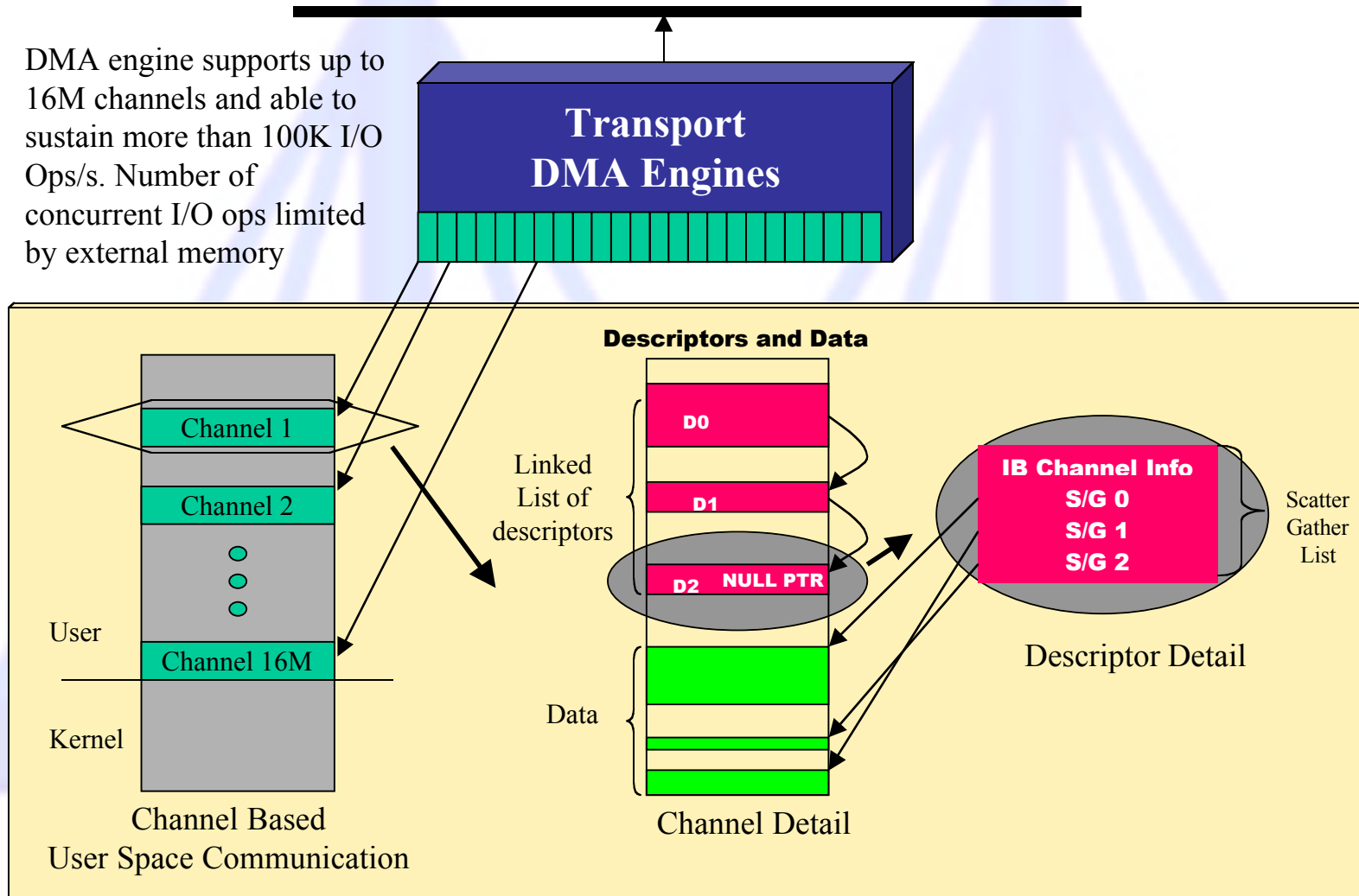




InfiniBridge™ Transport Protocol Engine

DMA engine supports up to 16M channels and able to sustain more than 100K I/O Ops/s. Number of concurrent I/O ops limited by external memory

PCI 64b/66 MHZ



System Memory

Mellanox Technologies, Inc.

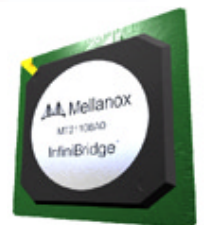




InfiniPCI™ Technology

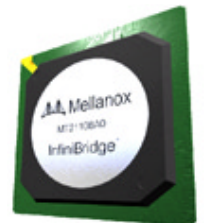
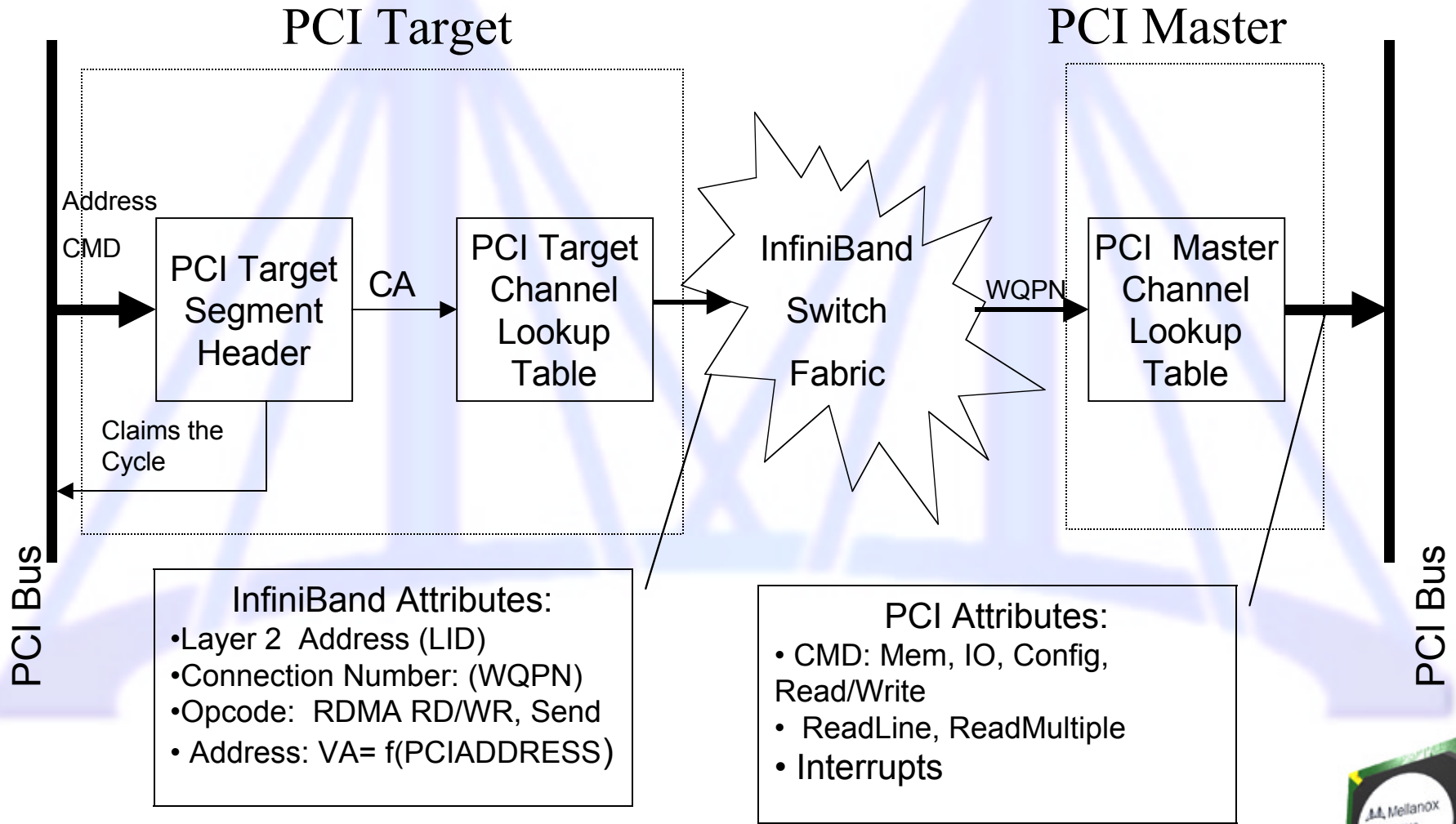
● InfiniPCI™ Technology

- Transparent PCI to PCI Bridging over standard InfiniBand Fabrics
- Functions with existing OS, BIOS, PCI software and hardware without modifications
- Use PCI semantics to create multi-segment backplanes, fully switched chassis, and multi-chassis fabrics



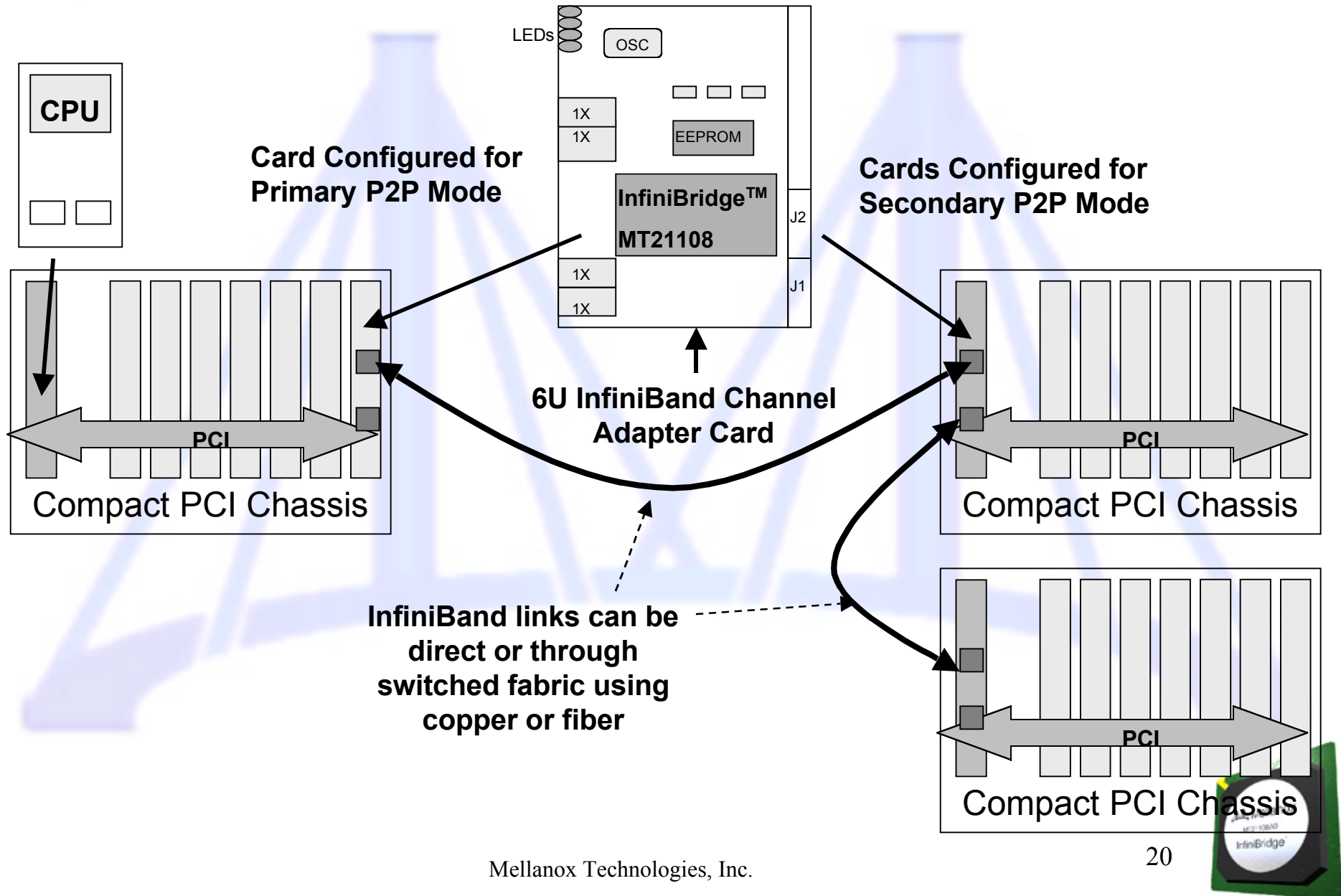


InfiniPCI™ System View



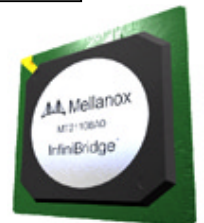
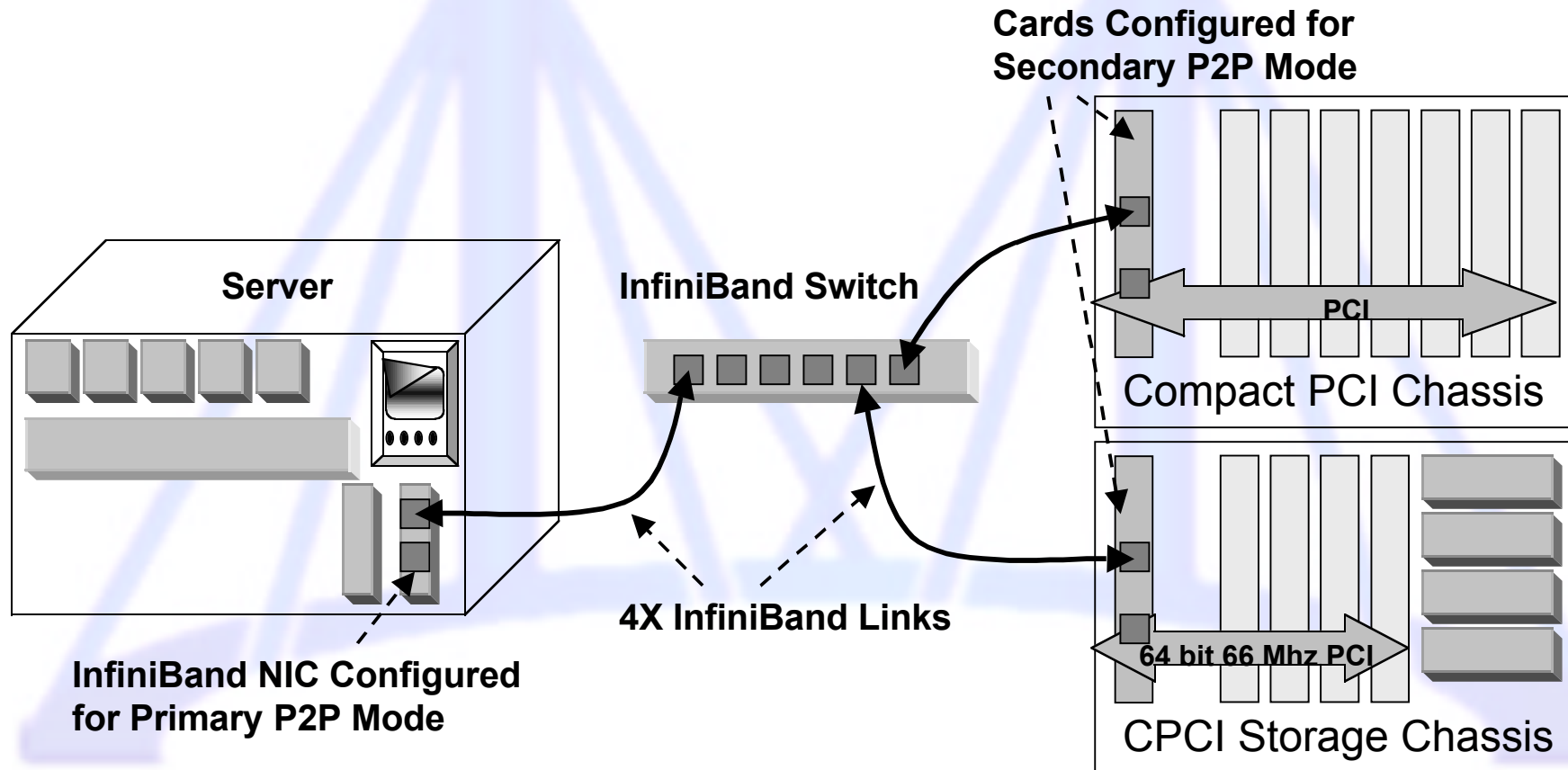


Chassis-to-Chassis Interconnect





Remote I/O Application





Summary

● **InfiniBridge™ Architecture**

- **Integrated 45 Gb/s non blocking switch and channel adapter**
- **Reliable transport in hardware**
- **Transport Protocol Engines support up to 16M connections with concurrency**
- **InfiniRISC™ embedded RISC processor**

● **Virtual fabrics enable multi-protocol networks**

● **InfiniPCI™ technology implements transparent PCI bridging**

