The Case for PCI Express® on the Backplane

Miguel Rodriguez

Sr. Product Marketing Engineer

PLX Technology, Inc.

Disclaimer

All opinions, judgments, recommendations, etc. that are presented herein are the opinions of the presenter of the material and do not necessarily reflect the opinions of the PCI-SIG®.

The information in this presentation refers to a specification still in the development process. All material is subject to change before the specification is released.

Overview

- Today's backplane requirements are more demanding than before
- De-facto backplane technology, GbE, can no longer keep up with growing demands
- Other technologies, 10GbE, IB and PCIe, are in line as the backplane technology successor
- PCIe has evolved from supporting a simple IO model to having the necessary features for becoming the ideal backplane interconnect

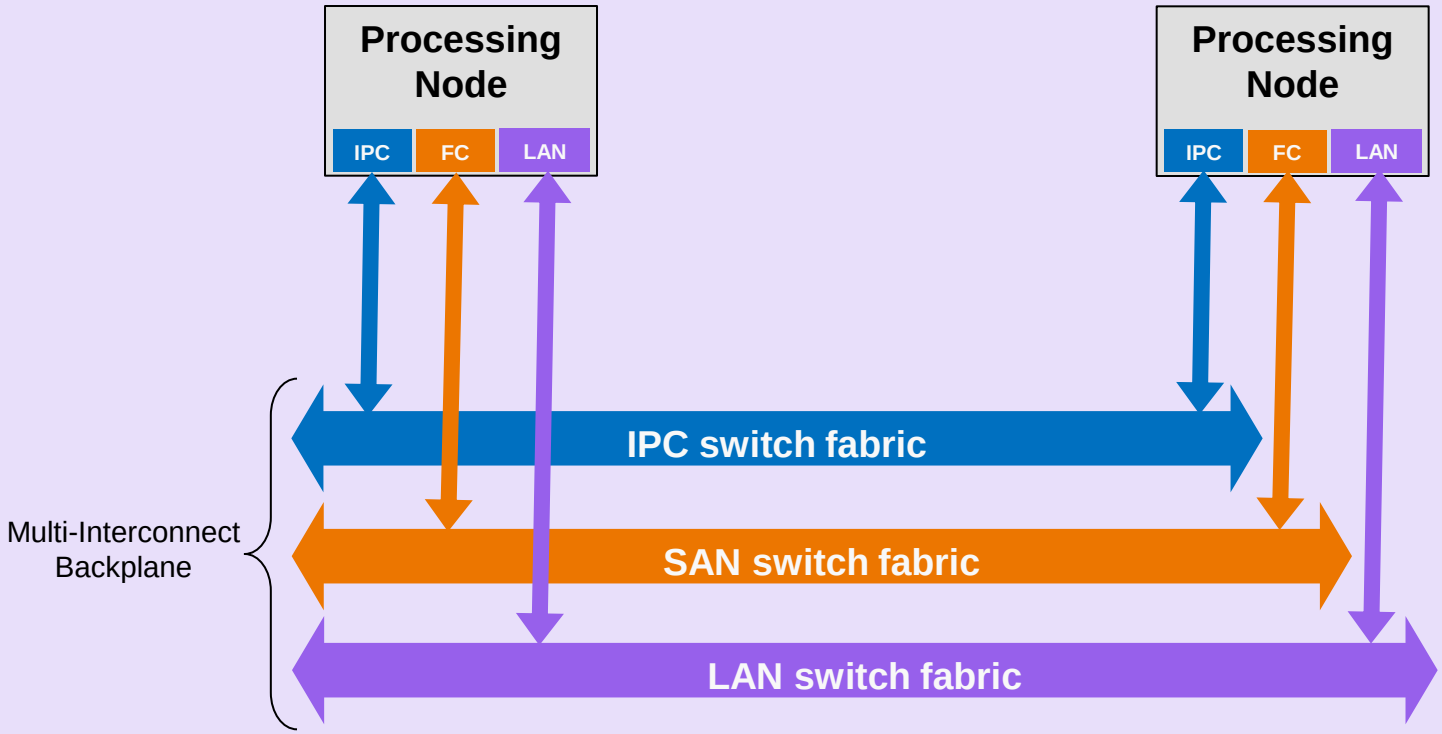
Backplane Requirements

- Need to support Inter-processor communication (IPC)
  - ✓ MPI and others
- Access to an external local area network (LAN)
  - ✓ Popular interconnect used now is gigabit Ethernet
- Access to an external storage area network (SAN)
  - ✓ Such as Fibre Channel, SAS, etc.

Backplane Requirements

- Needs support for a high performance fabric
  - ✓ High throughput: 10Gbps and over
  - ✓ Low latency: single digit microseconds
- Reliable physical interface
  - ✓ For deploying blade servers
    - Over 20 inches of FR4
  - ✓ Cabling between rack mounted servers
    - 5 meter long cables on average

Traditional Backplane



IPC 10GbE or Infiniband Adapter

FC Fibre Channel Adapter

LAN Gigabit Adapter

Backplane Technologies

- Three candidates are ready to make the claim for a unified backplane
  - ✓ PCI Express
  - ✓ 10 Gigabit Ethernet
  - ✓ Infiniband
- Each provide a set of capabilities which address the backplane requirements

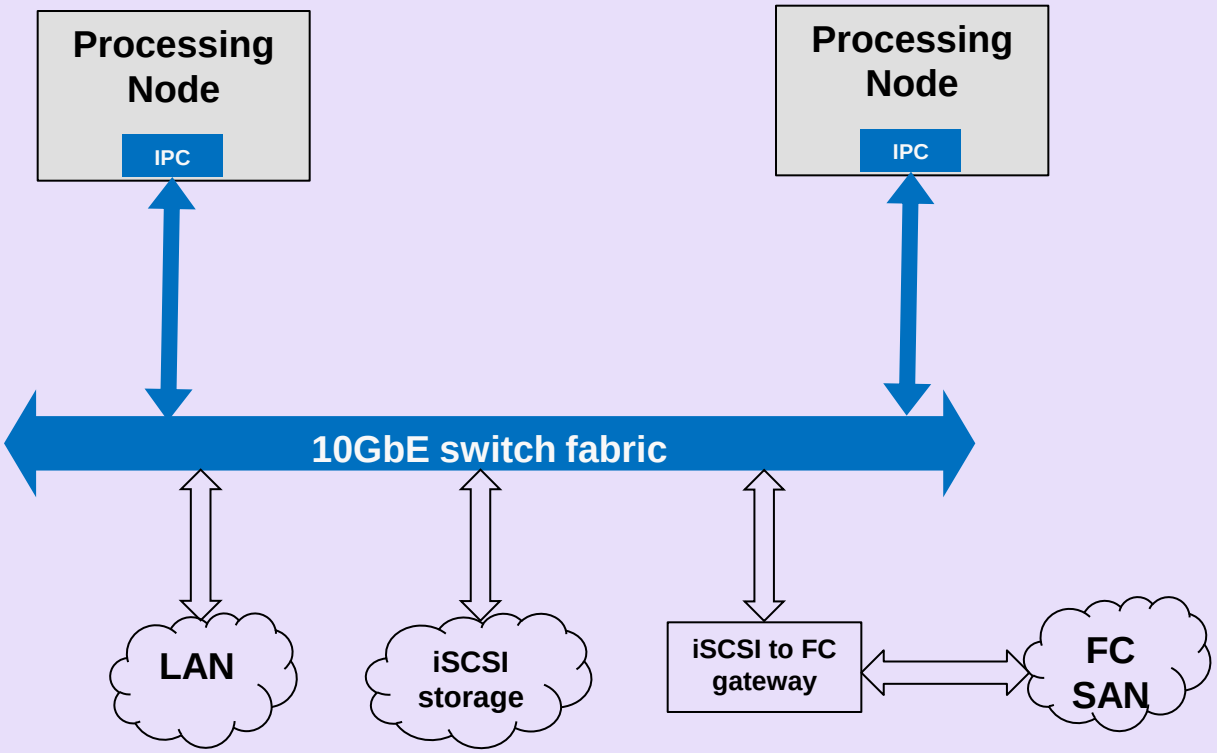
10GbE on the Backplane

- Ethernet is accepted and currently used for both LAN and IPC
  - ✓ Gigabit for LAN
  - ✓ 10Gigabit for IPC
- However, Ethernet is not ideal for storage traffic
  - ✓ It has a tendency to drop packets in the wake of congestion (lossy)
  - ✓ Upper layer protocols (TCP/IP) are used to improve reliability at the expense of inherent overhead

10GbE on the Backplane

- From storage perspective, iSCSI layer is built atop TCP/IP
  - ✓ Provides a more reliable mechanism over ethernet
- Complexities in iSCSI protocol require dedicated iSCSI adapter at server
  - ✓ Adapter handles actual storage communication
  - ✓ Supports iSCSI-based storage arrays only
  - ✓ Requires gateway device to other storage infrastructures
- Usage model of iSCSI device contradicts the definition of a unified IO interface

10GbE on the Backplane

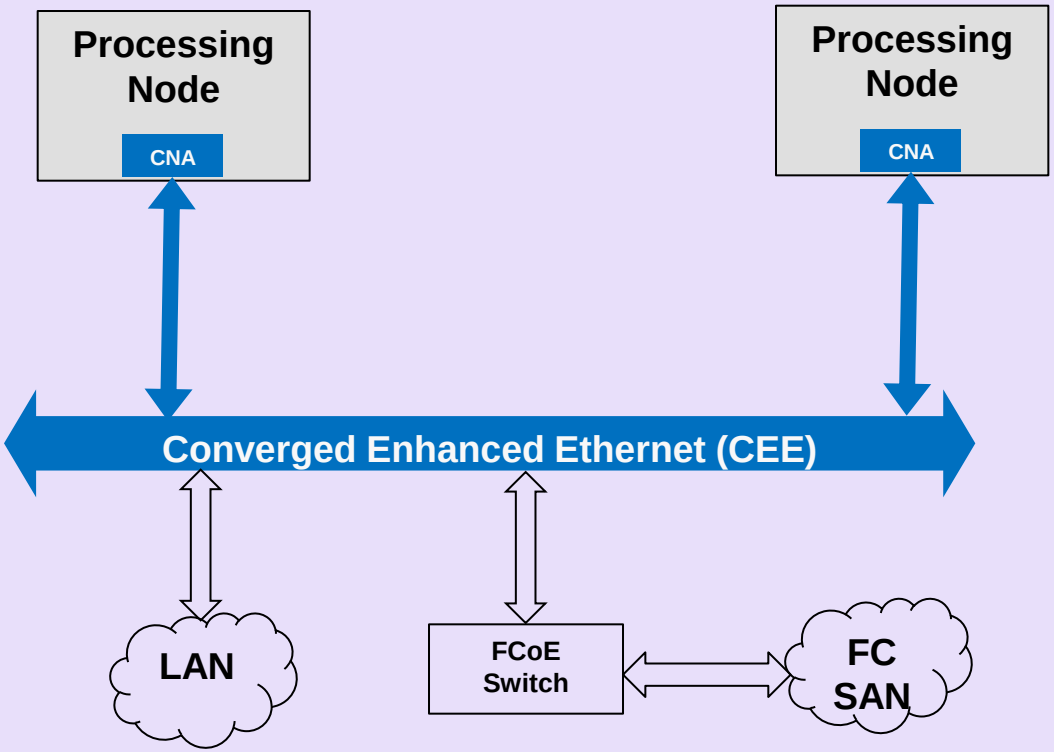


IPC 10GbE Adapter

10GbE on the Backplane

- Converged Enhanced Ethernet (CEE)
  - ✓ Recent improvement efforts to Ethernet to improve reliability (prevent packet drops)
  - ✓ Allows the FC protocol to run directly over Ethernet
    - Fibre Channel over Ethernet (FCoE)
    - Full benefits of FCoE not realized unless it is offloaded to hardware
  - ✓ Supports both functions simultaneously (FCoE and 10GbE)
- Both models (CEE and FCoE) still in early stages
  - ✓ Not widely deployed as most infrastructure is FC based
  - ✓ Requires new type of FCoE switch capable of handling both FCoE and TCP/IP
- Uses Converged Network Adapters (CNA) to create a unified fabric
  - ✓ Supports both functions simultaneously (FCoE and 10GbE)

10GbE on the Backplane

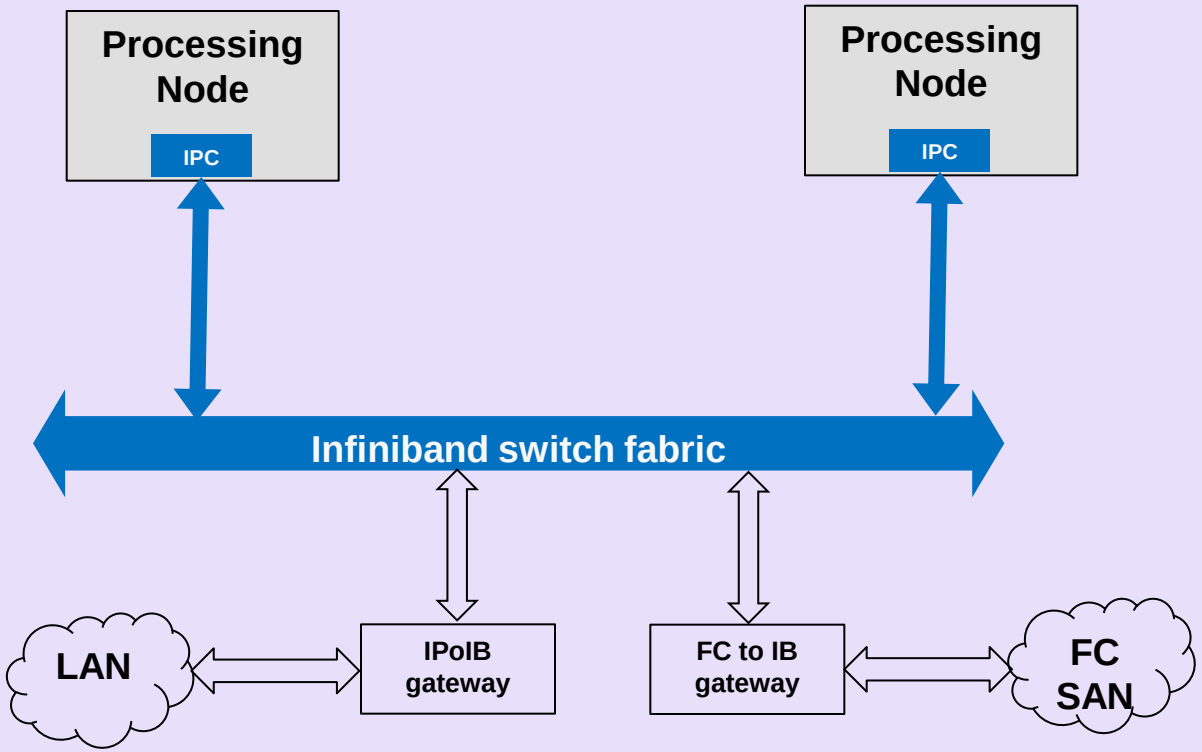


CNA 10GbE CEE Adapter

Infiniband on the Backplane

- High throughput and low latency technology
- Ideal for IPC and commonly deployed in High Performance Computing (HPC)
- LAN connectivity requires TCP/IP over IB (IPoIB) gateway
- SAN connectivity requires Fibre Channel (FC) to Infiniband (IB) gateway
- Additional overhead and cost to connect to LAN and SAN

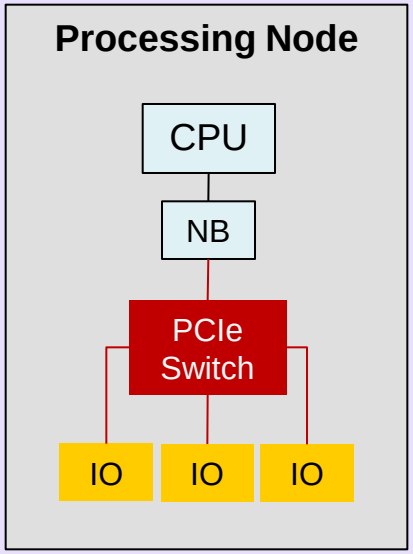
Infiniband on the Backplane



IPC Infiniband Adapter

Traditional PCIe System

- Parent/Child model
  - ✓ One CPU to many IO endpoints
  - ✓ CPU manages all the IO in a system
  - ✓ Single address domain

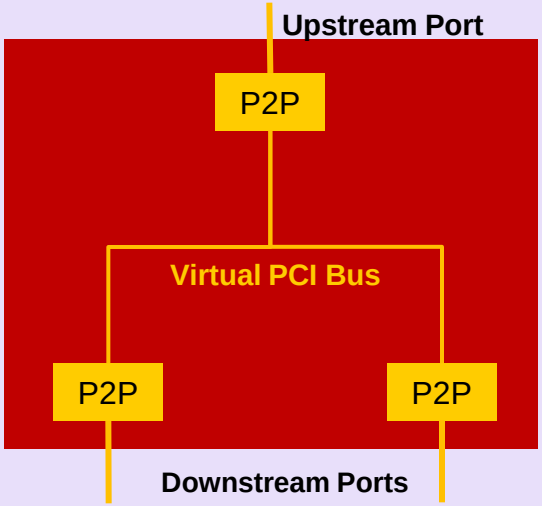


PCI Express on the Backplane

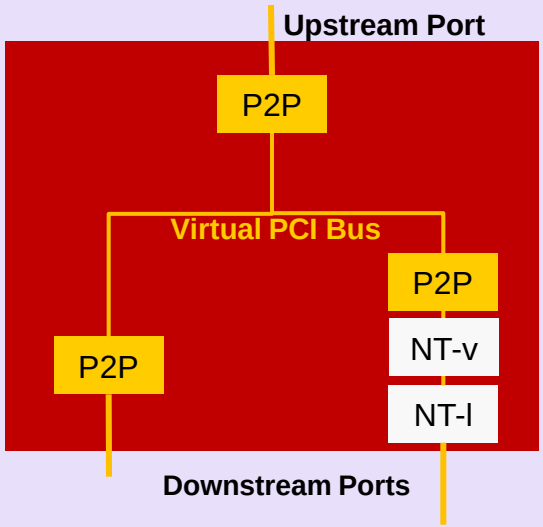
- No longer a CPU to IO interconnect only
  - ✓ Supports more advanced models
- Implements advanced mechanisms to enable IPC and LAN directly on the PCIe fabric
  - ✓ Via address translation, doorbell and scratchpad mechanisms
- Supports Shared IO model
  - ✓ Only one fabric to manage
- Supported natively on the CPU
  - ✓ No need to bridge to other protocols for IPC

Non-Transparent Bridging

- Vendor specific feature built on top of PCIe
  - ✓ Implemented as an endpoint pair inside a PCIe switch
    - NT-v: endpoint found by the host on the upstream port
    - NT-l: endpoint found by the host on the downstream port
  - ✓ Presents a Type 0 configuration header to software
    - Configuration transaction terminated at NT endpoint



Software View of a Three Port Transparent PCIe Switch

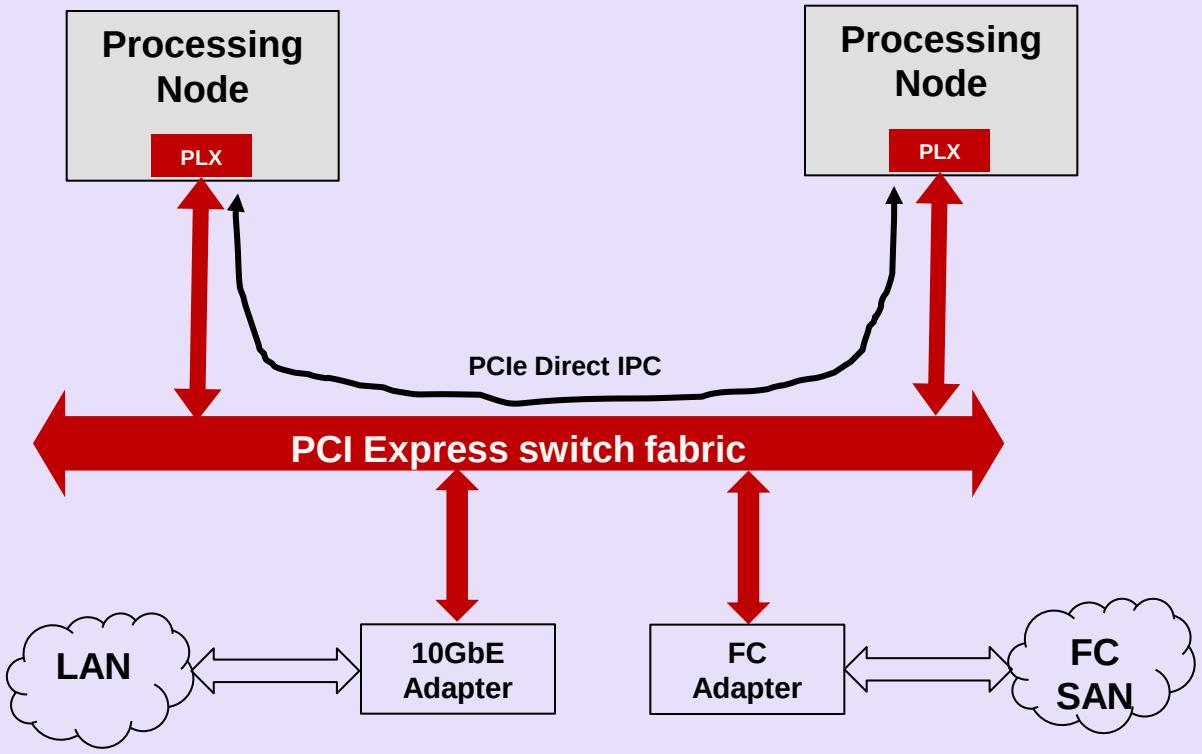


Software View of aThree Port Transparent PCIe Switch with NT

PCI Express on the Backplane

- Non-Transparent support enables clustering
  - ✓ Implements address translation mechanisms to enable address isolation between hosts
  - ✓ Utilizes doorbell and scratchpad registers for generating interrupts between hosts
  - ✓ Connects to an external PCIe switch fabric to which other nodes and IO are also connected to
- NT hardware serves as a software platform
  - ✓ NT driver can expose industry standard APIs
    - TCP/IP, Sockets, etc.

PCI Express on the Backplane



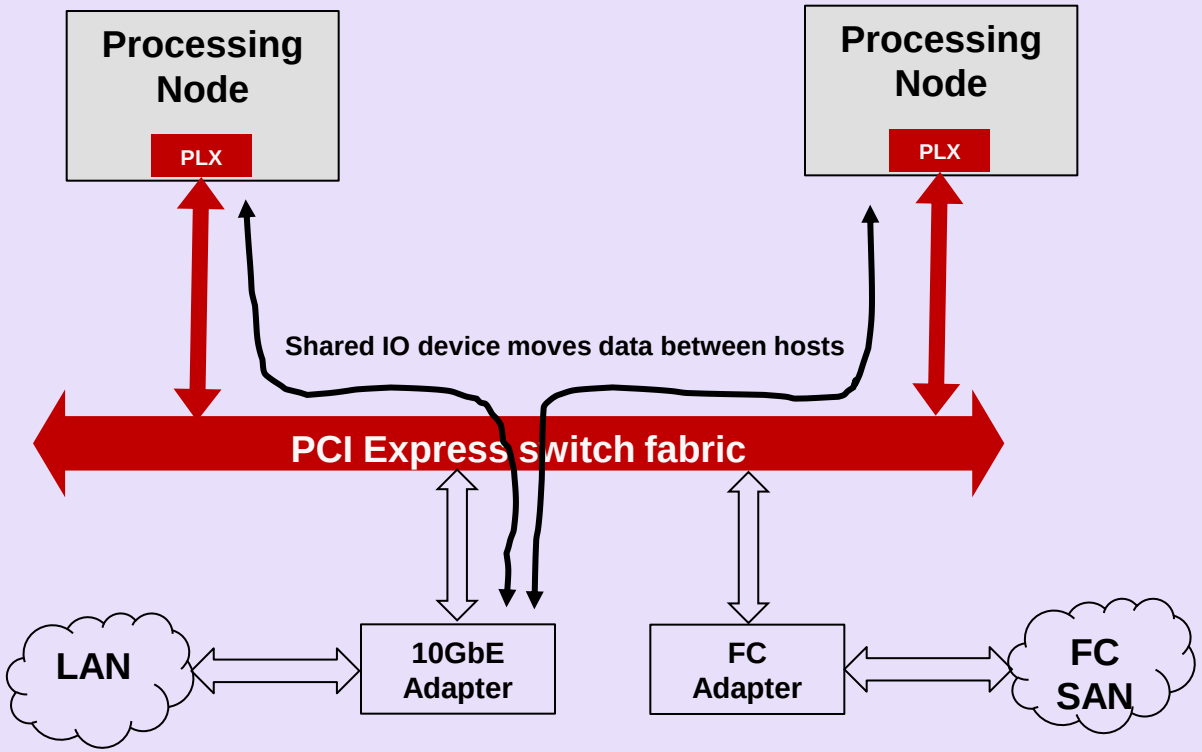
PLX PCIe NT Adapter

Note: PCIe NT functionality can be part of the PCI Express switch fabric.

PCI Express on the Backplane

- Shared IO model using Non-Transparent port
  - ✓ NT port and device driver provide the requisite support for IO sharing
  - ✓ Does not require MR-IOV; uses SR-IOV endpoints instead
  - ✓ Shared adapter appears private to each processing node
  - ✓ Avoids the over-provisioning of the IO resources
- Communication to other nodes
  - ✓ Logically, similar to that of using a dedicated IPC adapter
  - ✓ Physically, adapter resides remotely

PCI Express on the Backplane



PLX PCIe NT Adapter

Performance

- Low latency and high throughput
  - ✓ Makes for a good performance foundation
- Additionally, interconnect must support an efficient interface
  - ✓ In order to take advantage of hardware capabilities
- Hardware capabilities and application usage model for
  - ✓ 10GbE
  - ✓ Infiniband
  - ✓ PCI Express

Hardware Capabilities

- 10GbE offers evolutionary scalability with minimal disruption
  - ✓ Higher throughput (gigabit to 10 gigabit)
- 10GbE fails in two important areas
  - ✓ Latency
    - Lossy interconnect can drop packets → unpredictable latencies
    - CEE enhancements may alleviate congestion; improvement on latency is not yet clear
    - Current latency offered by Infiniband and PCIe < 2us
  - ✓ Jitter

Hardware Capabilities

- Infiniband
  - ✓ Throughput
    - Can sustain high data rates (40Gbps QDR; 80Gbps EDR)
  - ✓ However, throughput limited to PCIe interface capability
    - PCIe 2.0 x8 cannot sustain an IB dual-port interface at 80Gbps EDR
  - ✓ Offers low latency interface
  - ✓ Serves as an PCIe to IB bridge

Hardware Capabilities

- PCI Express
  - ✓ High throughput with PCIe 3.0
    - 8GT/s per lane → x16 = 128GT/s
  - ✓ Low latency
    - PCIe switch latency is less 130ns
    - End-to-end latency is less than 1us typically
  - ✓ PCIe native on the processor/chipset
    - No need to terminate inside the box
    - Can extend outside the system allowing user to take advantage of the latency and bandwidth potential
    - Allows direct read/write of remote memory

Application Usage Model

- An application running on a server can transfer data to/from remote server in one of two ways
- Messaging model
  - ✓ Application writes to a common buffer on remote node
  - ✓ Data is locally copied from the common buffer to the target application buffer
  - ✓ Predominantly used by Ethernet
  - ✓ Also supported by Infiniband

Application Usage Model

- Remote Direct Memory Access (RDMA)
  - ✓ Enables an application to directly access data on the memory of a remote server
  - ✓ Eliminates data copies → low latency; high bandwidth
  - ✓ Natively supported by Infiniband
  - ✓ Supported by Ethernet via iWarp
    - RDMA at the TCP/IP level
    - Complete protocol needs to be offloaded to hardware to gain any meaningful performance

Application Usage Model

- Direct Addressing Mode
  - ✓ Unique mechanism supported by PCIe only
  - ✓ Provides capability for local server to access data in remote memory (remote server)
  - ✓ Non-Transparent approach allows translation mechanism for remote memory access
  - ✓ From sender's perspective, memory operation targets its own address space
  - ✓ From receiver's perspective, memory operation originates within its own address space
- In the context of IPC communication
  - ✓ Direct address mode is considered more efficient

Summary

- PCIe provides an attractive value proposition
  - ✓ Extremely affordable with exceptional performance
  - ✓ Integration in mainstream processor chips
- Key Features that distinguish PCIe
  - ✓ High performance in terms of bandwidth and latency
  - ✓ Support for IPC using Non-Transparent shared memory
  - ✓ IO virtualizations in multi-host environment

Thank you for attending the
PCI-SIG Developers Conference 2010.

For more information please go to
www.pcisig.com