PCI SIG™

PCI Express™ Basics & Applications in Communication Systems

Akber Kazmi

PLX Technology

Agenda

- ➡ PCI Express Overview, Components & Architecture
- PCI Express Protocol Layers
- Needs of Communication Systems & PCIe
- PCI Express in Communication Systems
- Summary

PCI Express
High Level Overview

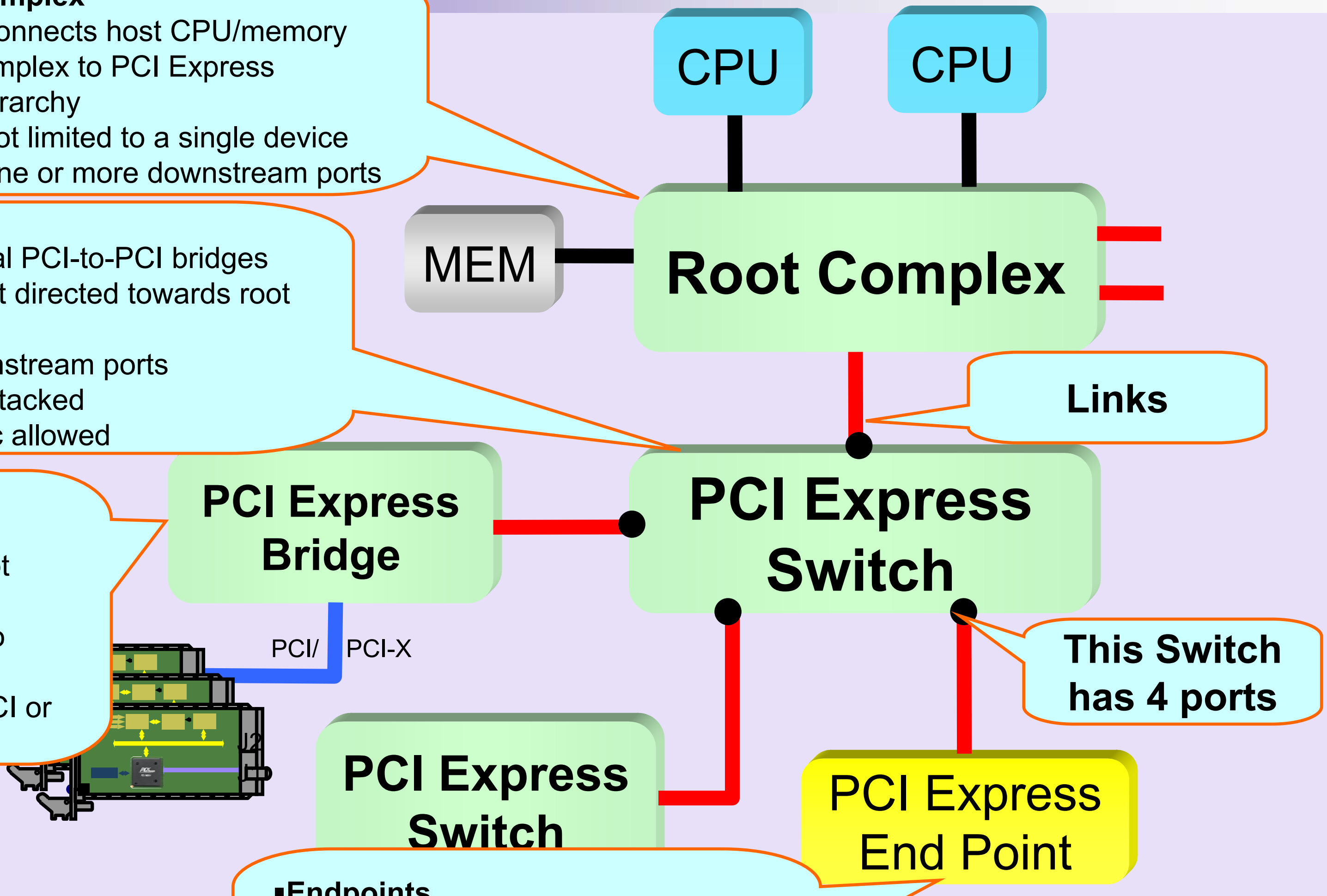
- Chip/chip and fabric interconnect technology
- High speed serial, packet based
- Fully open and standardized
- Complete compatibility with PCI & PCI-X
- Cost driver: PCs/Graphics (economies of scale)
- Advanced features: QoS, Flow Control, data error detection
- Applicable to wide variety of applications
  - ✓ Servers, Storage, Communications, embedded
- Extensive industry support

PCI Express Features/Benefits

PCI Express Features	Benefits
• PCI transparency	• Smooth migration, SW re-use, simple validation
• TC/VC mechanism	• QoS & isochrony
• High bandwidth	• Peak traffic loads, support high throughput apps.
• Flow control	• Buffer size flexibility, cost flexibility
• Reliable link layer	• No dropped packets, simplified SW, high availability
• Robust link layer	• Maintain communication for HA or diagnosis
• E-CRC	• End-to-end data integrity
• Error reporting, fault isolation	• System management, serviceability, availability
• Hot-plug	• Optimize density, support cold spares
• Power management	• Reduced power consumption and emissions
• High Speed Serial	• Reduced cost, pin count, PCB layers & area

PCI Express can be used in many market segments

Typical PCI Express System



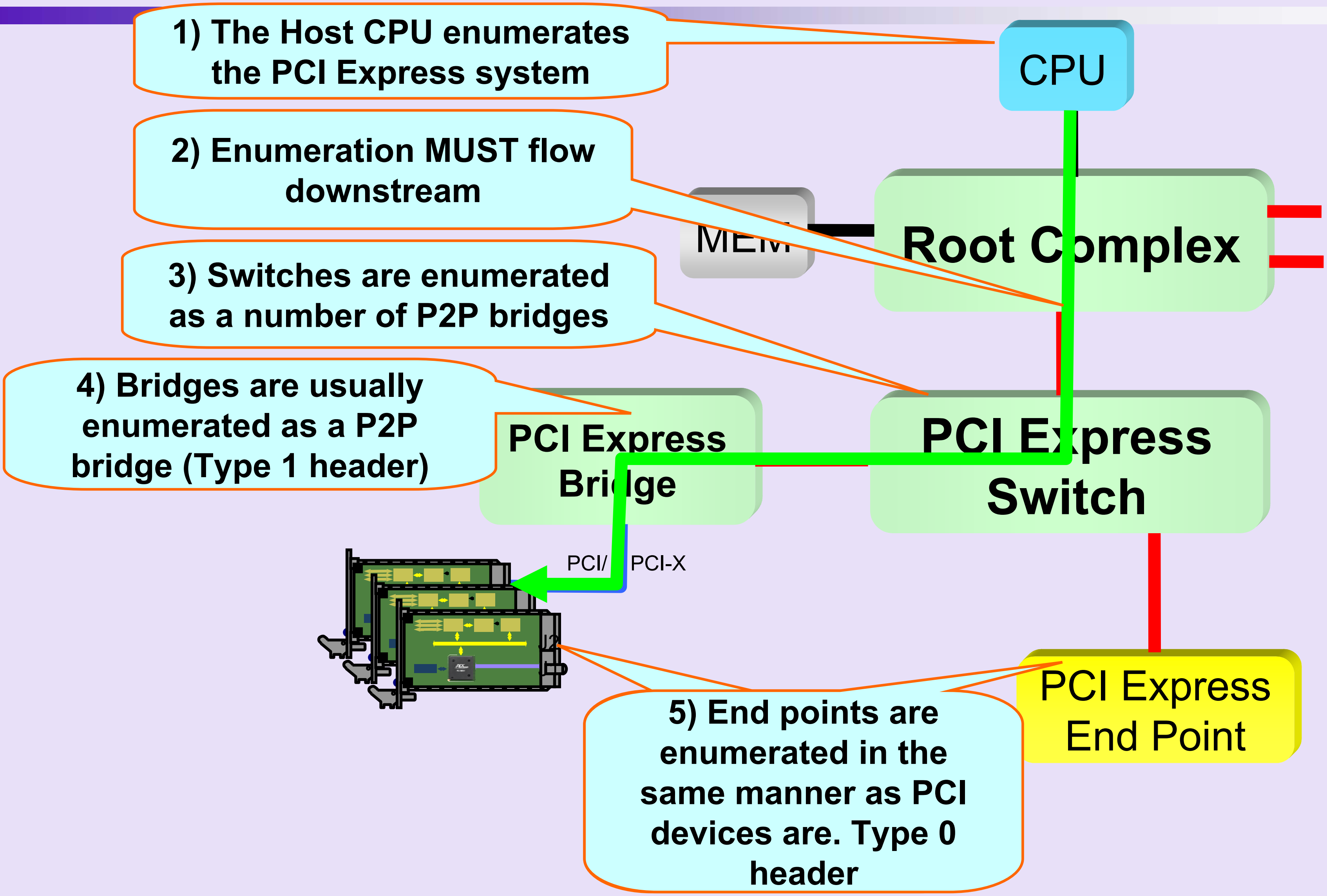
- **Root Complex**
  - Connects host CPU/memory complex to PCI Express hierarchy
  - Not limited to a single device
  - One or more downstream ports

- **Switch**
  - Assembly of logical PCI-to-PCI bridges
  - One upstream port directed towards root complex
  - One or more downstream ports
  - Switches can be stacked
  - Peer to peer traffic allowed

- **Bridge**
  - One upstream port directed towards root complex
  - One downstream to other devices
    - Example: PCI or PCI-X bus

- **Endpoints**
  - Native PCI Express Endpoints
  - Examples:
    USB, InfiniBand, E’net
    FibreChannel, etc.

Configuration



1) The Host CPU enumerates the PCI Express system

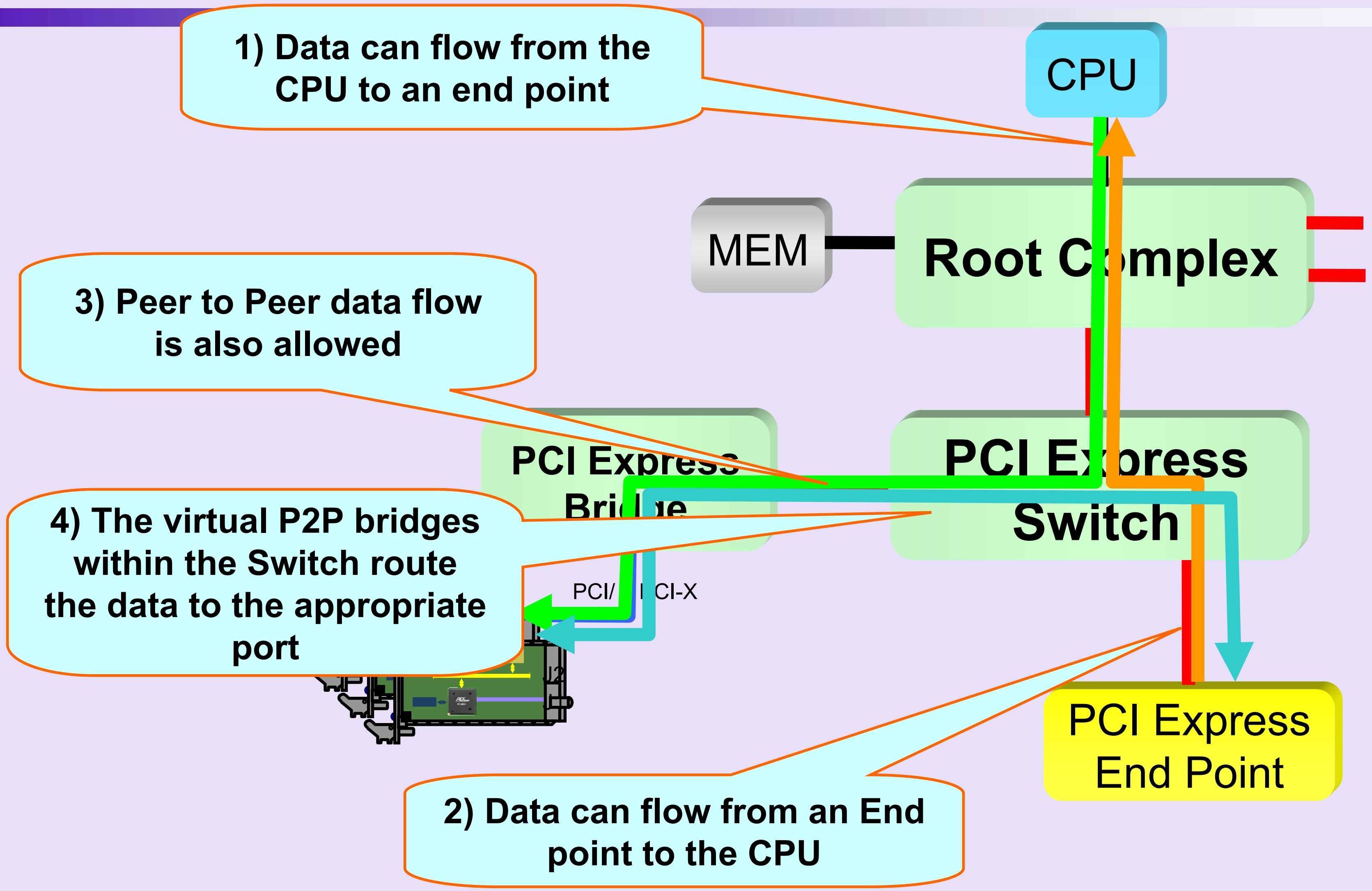
2) Enumeration MUST flow downstream

3) Switches are enumerated as a number of P2P bridges

4) Bridges are usually enumerated as a P2P bridge (Type 1 header)

5) End points are enumerated in the same manner as PCI devices are. Type 0 header

Data Flow



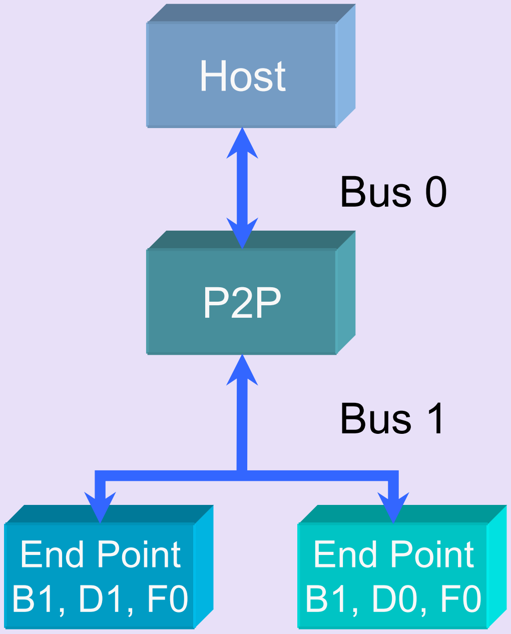
1) Data can flow from the CPU to an end point

2) Data can flow from an End point to the CPU

3) Peer to Peer data flow is also allowed

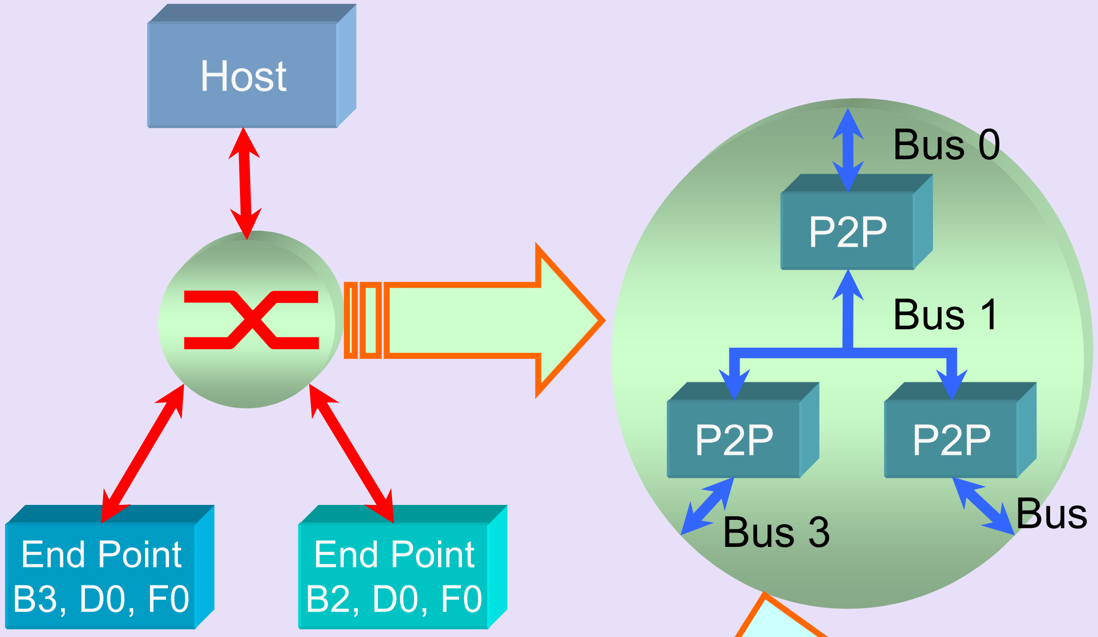
4) The virtual P2P bridges within the Switch route the data to the appropriate port

PCI Express – Software Model



PCI System

Where: B=bus,
D=device,
F=function



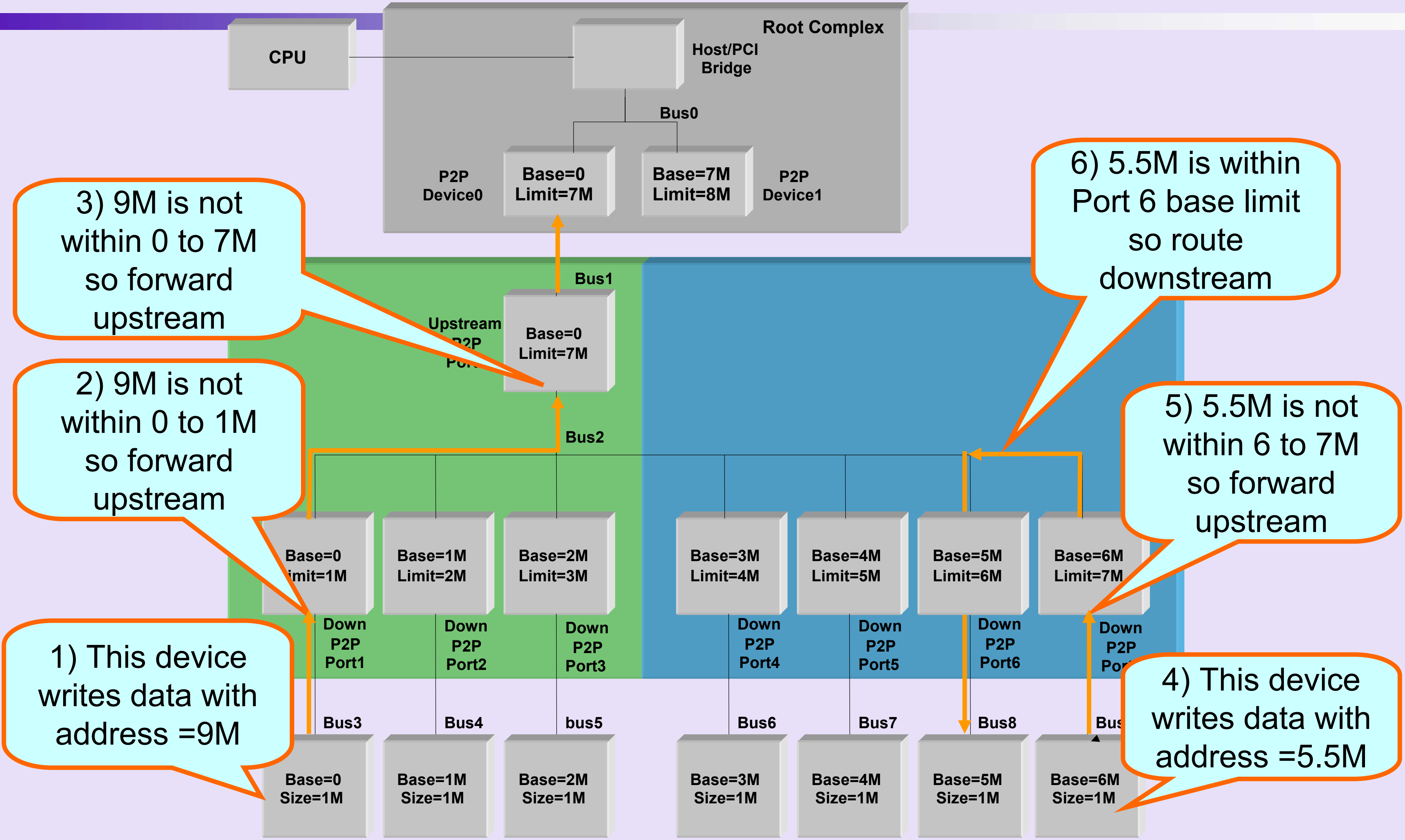
Equivalent PCI Express System

A switch looks like a collection of P2P bridges. Bus 1 is a virtual PCI bus

Data Routing

- PCI Compatible Routing Methods
  - ✓ Address Routing
    - – Memory and I/O read/write
    - – Optional for messaging
  - ✓ ID Routing
    - – Configuration read write
    - – Completions
    - – Optional for messaging
- PCI Express only routing methods
  - ✓ Implicit Routing
    - – Messaging
      - packets are routed based on a sub-field in the packet header.
      - Implicitly routed messages eliminates most of the sideband signals for interrupts, error handling, and power management.

Address Routing Examples



1) This device writes data with address =9M

2) 9M is not within 0 to 1M so forward upstream

3) 9M is not within 0 to 7M so forward upstream

4) This device writes data with address =5.5M

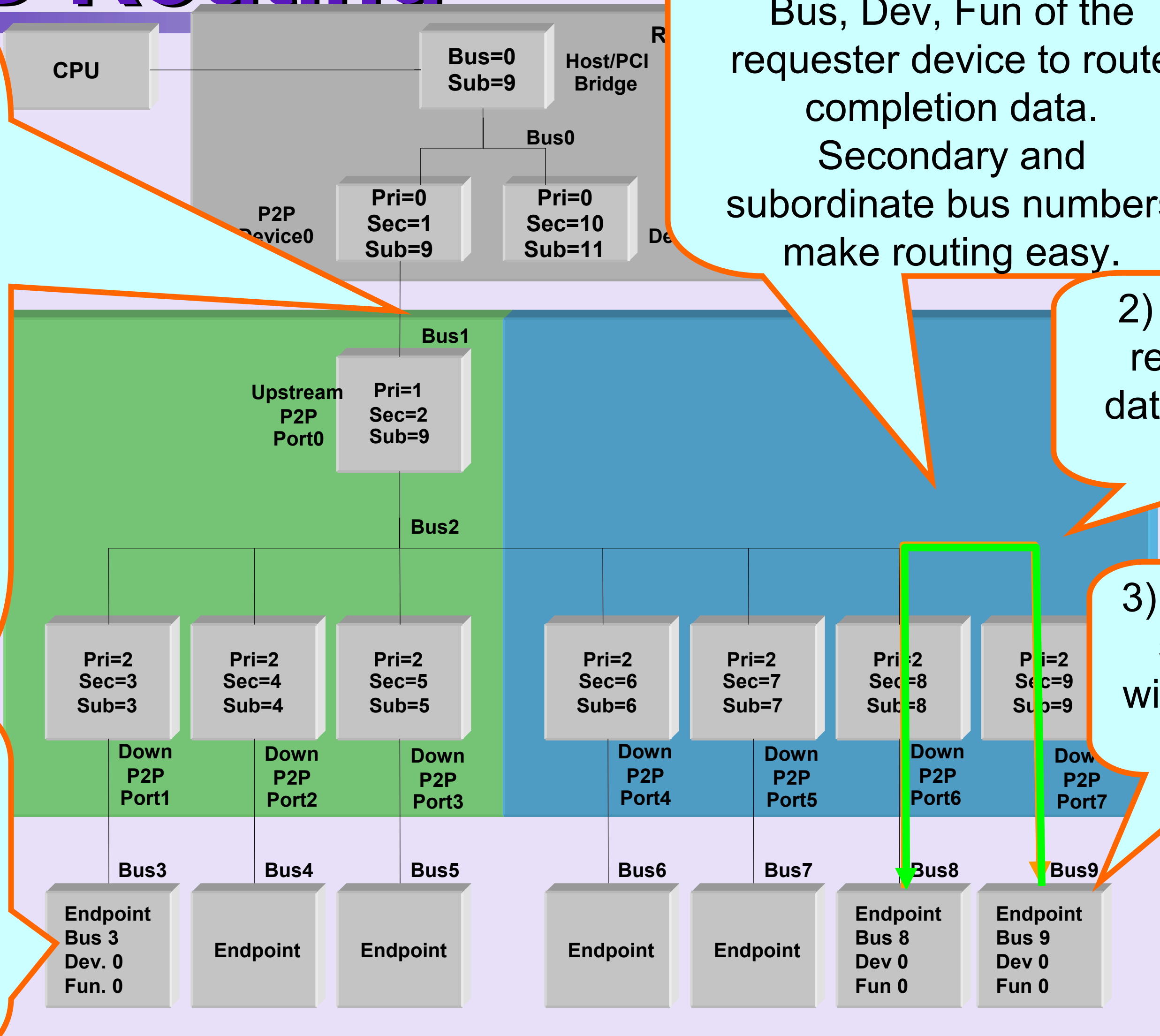
5) 5.5M is not within 6 to 7M so forward upstream

6) 5.5M is within Port 6 base limit so route downstream

ID Routing

Type 1 configuration accesses are converted to Type 0 accesses at the destination bus. E.g. a Type 1 access to a device with bus number 1 is converted to a Type 0 access here

Configuration and completions accesses use Bus, Device, Function numbers.

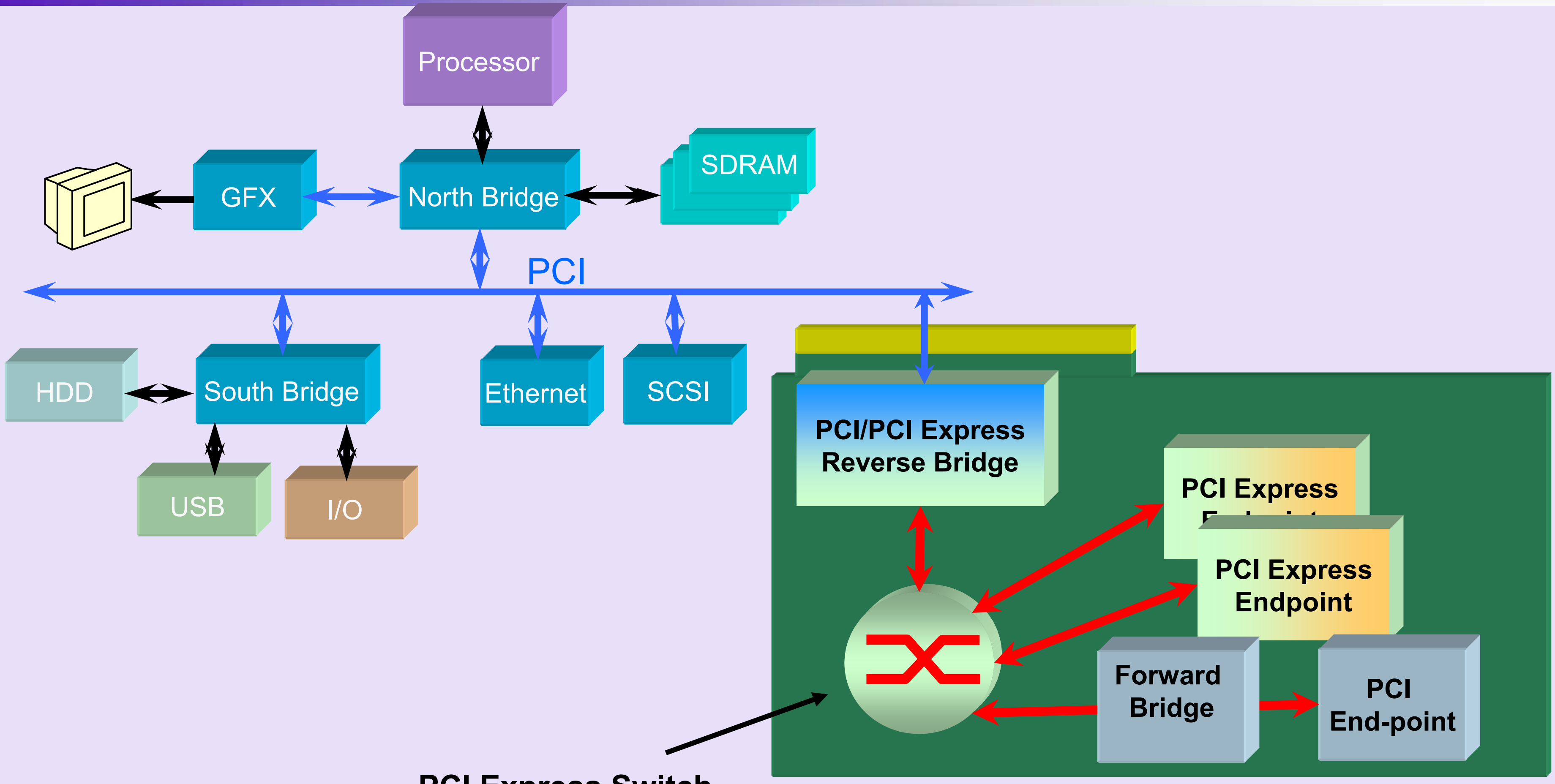


1) Completions use the Bus, Dev, Fun of the requester device to route completion data. Secondary and subordinate bus numbers make routing easy.

2) Device 8,0,0 requests read data from device 9,0,0

3) Device 9,0,0 sends data with Requester ID of 8,0,0

Reverse and Forward Bridging

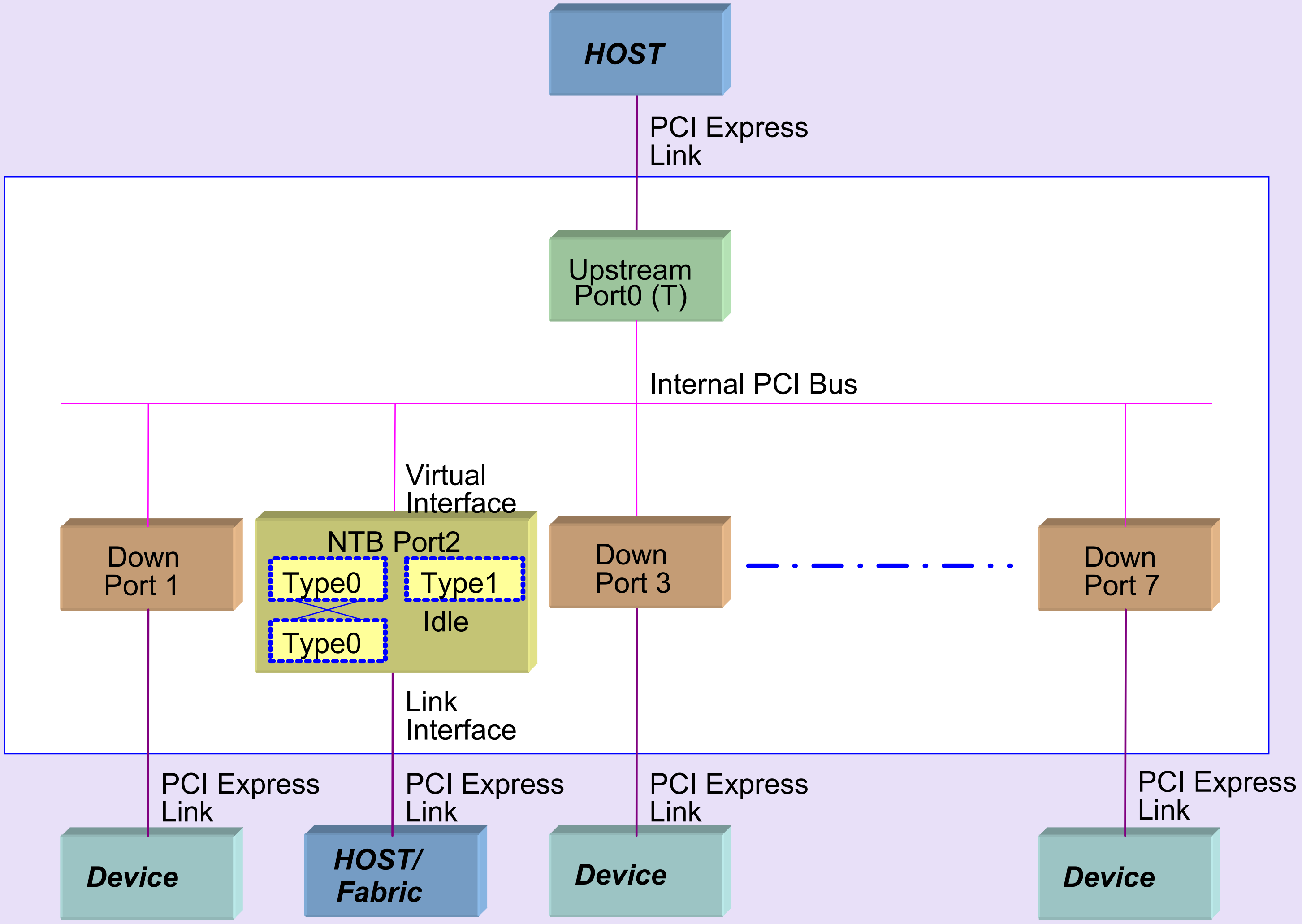


PCI Express Switch

Non-Transparent Bridge

- Provides isolation of host memory domains
- Presents the whole Sub-system as a Type0 Endpoint to Host
- Enables Inter-domain communication through address translation and Requester ID translation
- Provides Door-Bell and Scratch PAD register mechanism for host communication

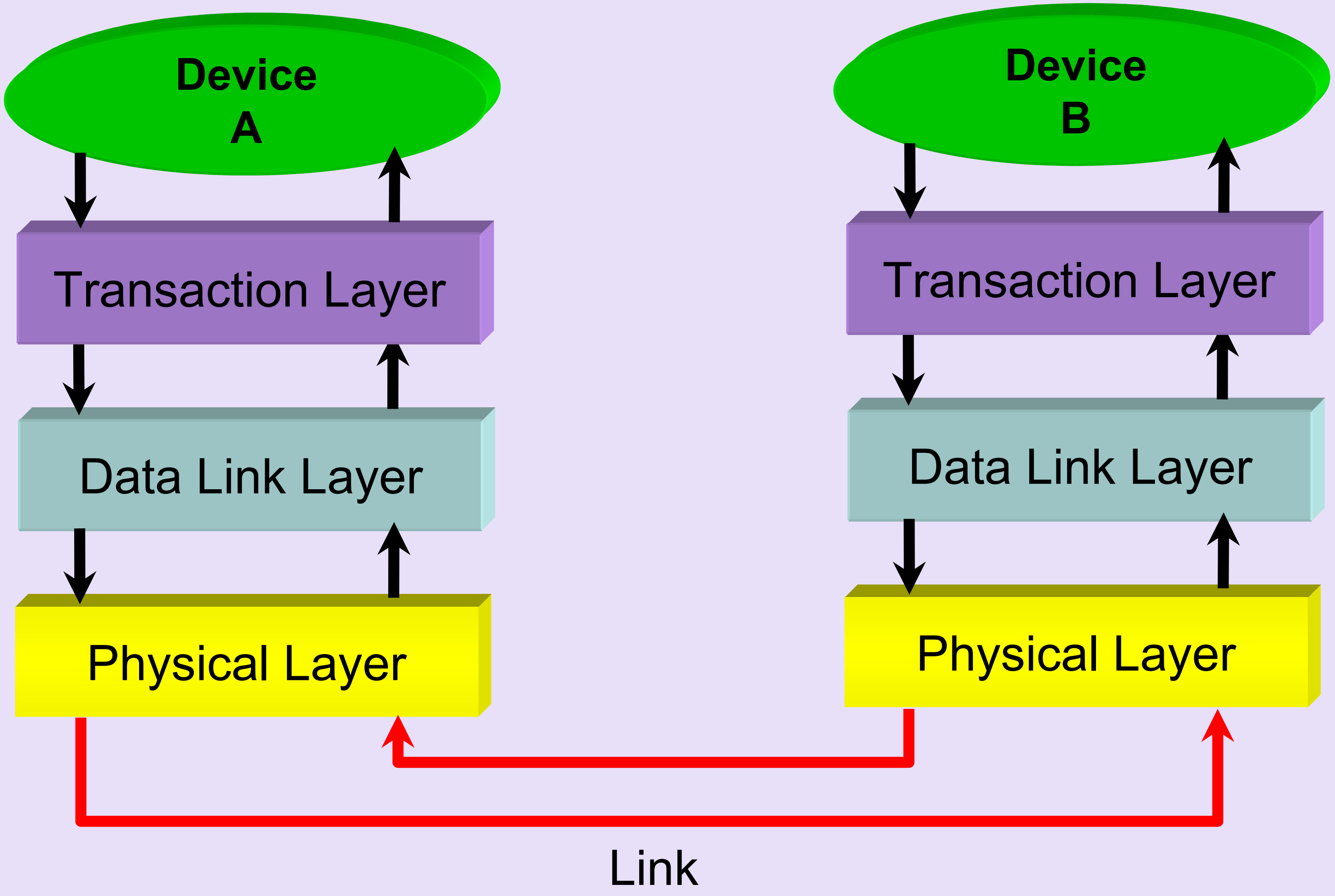
Non-Transparent Bridging



Agenda

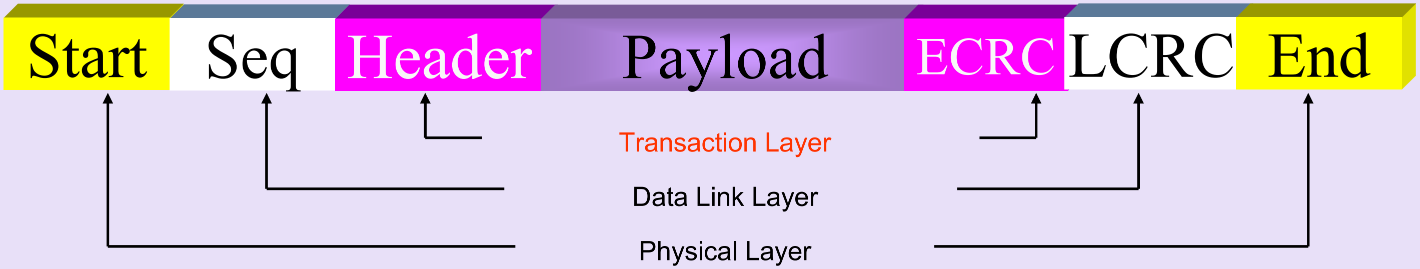
- PCI Express Overview, Components & Architecture
- ➡ PCI Express Protocol Layers
- Needs of Communication Systems & PCIe
- PCI Express in Communication Systems
- Summary

Protocol Stack



Transaction Layer

- Upper layer of PCI Express protocol
- Responsible for;
  - ✓ Storing negotiated and programmed configuration information
  - ✓ Managing link flow control
  - ✓ Enforcing ordering and Quality of Service
  - ✓ Power management control/status
  - ✓ Transaction Layer Packet processing
  - ✓ Assembly, disassembly, high-level error checking



Transaction Layer
- Implementation

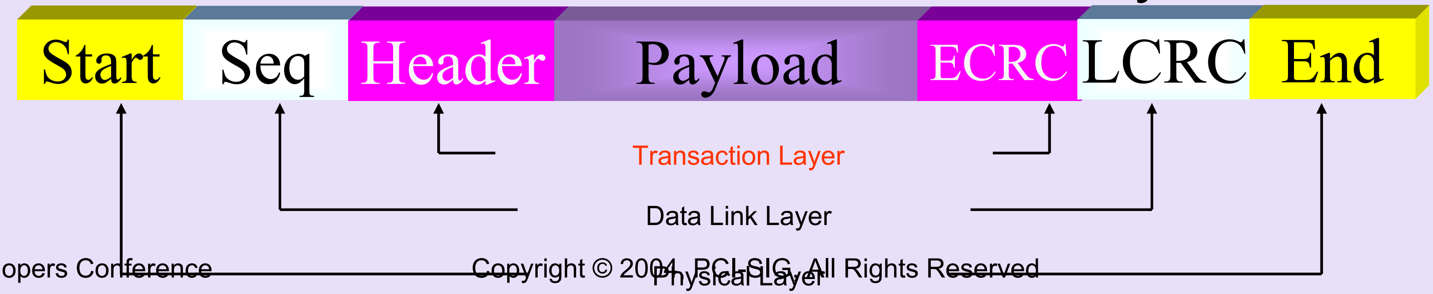
- Packet Header for Address Routing is either 12 or 16 bytes

<table>
<tr><td></td><td colspan="8">0</td><td colspan="8">1</td><td colspan="8">2</td><td colspan="8">3</td></tr>
<tr><td></td><td>7</td><td>6</td><td>5</td><td>4</td><td>3</td><td>2</td><td>1</td><td>0</td><td>7</td><td>6</td><td>5</td><td>4</td><td>3</td><td>2</td><td>1</td><td>0</td><td>7</td><td>6</td><td>5</td><td>4</td><td>3</td><td>2</td><td>1</td><td>0</td><td>7</td><td>6</td><td>5</td><td>4</td><td>3</td><td>2</td><td>1</td><td>0</td></tr>
<tr><td>Byte 0</td><td>R</td><td colspan="2">Fmt</td><td colspan="5">Type</td><td>R</td><td colspan="3">TC</td><td colspan="4">R</td><td>T
D</td><td>E
P</td><td colspan="2">Attr</td><td colspan="2">R</td><td colspan="10">Length</td></tr>
<tr><td>Byte 4</td><td colspan="32">Request Specific</td></tr>
<tr><td>Byte 8</td><td colspan="32">Address [63:32]</td></tr>
<tr><td>Byte 12</td><td colspan="30">Address [31:2]</td><td colspan="2">R</td></tr>
</table>

<table>
<tr><td></td><td colspan="8">0</td><td colspan="8">1</td><td colspan="8">2</td><td colspan="8">3</td></tr>
<tr><td></td><td>7</td><td>6</td><td>5</td><td>4</td><td>3</td><td>2</td><td>1</td><td>0</td><td>7</td><td>6</td><td>5</td><td>4</td><td>3</td><td>2</td><td>1</td><td>0</td><td>7</td><td>6</td><td>5</td><td>4</td><td>3</td><td>2</td><td>1</td><td>0</td><td>7</td><td>6</td><td>5</td><td>4</td><td>3</td><td>2</td><td>1</td><td>0</td></tr>
<tr><td>Byte 0</td><td>R</td><td colspan="2">Fmt</td><td colspan="5">Type</td><td>R</td><td colspan="3">TC</td><td colspan="4">R</td><td>T
D</td><td>E
P</td><td colspan="2">Attr</td><td colspan="2">R</td><td colspan="10">Length</td></tr>
<tr><td>Byte 4</td><td colspan="32">Request Specific</td></tr>
<tr><td>Byte 8</td><td colspan="30">Address [31:2]</td><td colspan="2">R</td></tr>
</table>

Data Link Layer Packets

- Data Link Layer Functions
  - ✓ Integrity of Transaction layer packet (TLPs)
    - – Link-level error detection and re-transmission of bad TLP's
  - ✓ Tracking state of link and passing link status to upper layers
  - ✓ Conveying power management state info.
  - ✓ Initialization and updates of credit based flow control
- Classes of DLLPs
  - ✓ Transaction Layer Packet acknowledgements (Ack/Nak)
  - ✓ Power management
  - ✓ Flow Control (Flow Control packets)
  - ✓ Vendor specific DLLP
- Create and terminate DLLPs for Link layer info



DLL and TL Interaction

Transaction Layer originates header, data and digest, checks flow control credits and forwards to DLL.

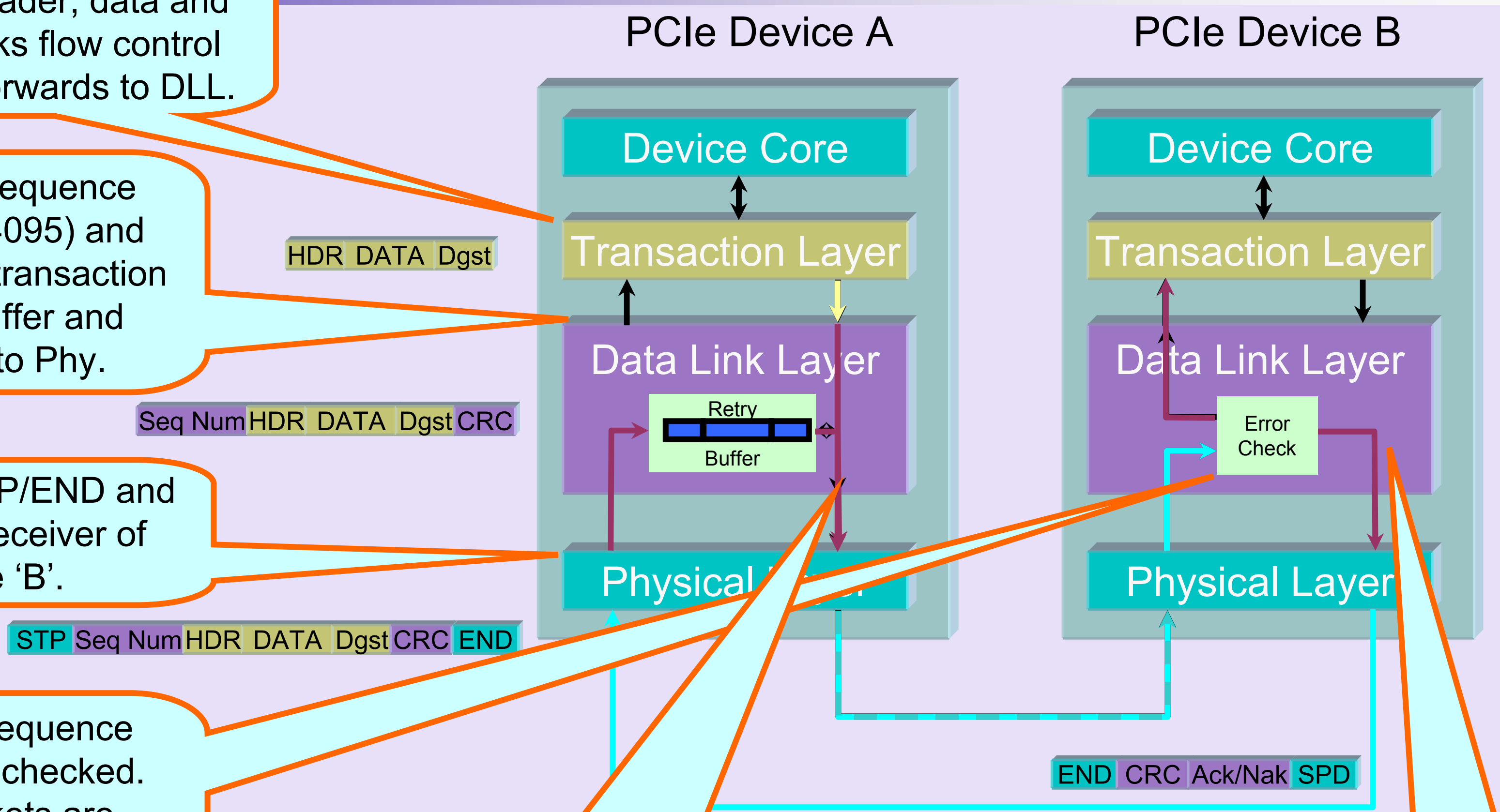
DLL adds sequence number (0-4095) and CRC, stores transaction in Retry buffer and forwards to Phy.

Phy adds STP/END and sends to Receiver of device ‘B’.

CRC and sequence number are checked. Valid packets are forwarded to Transaction Layer

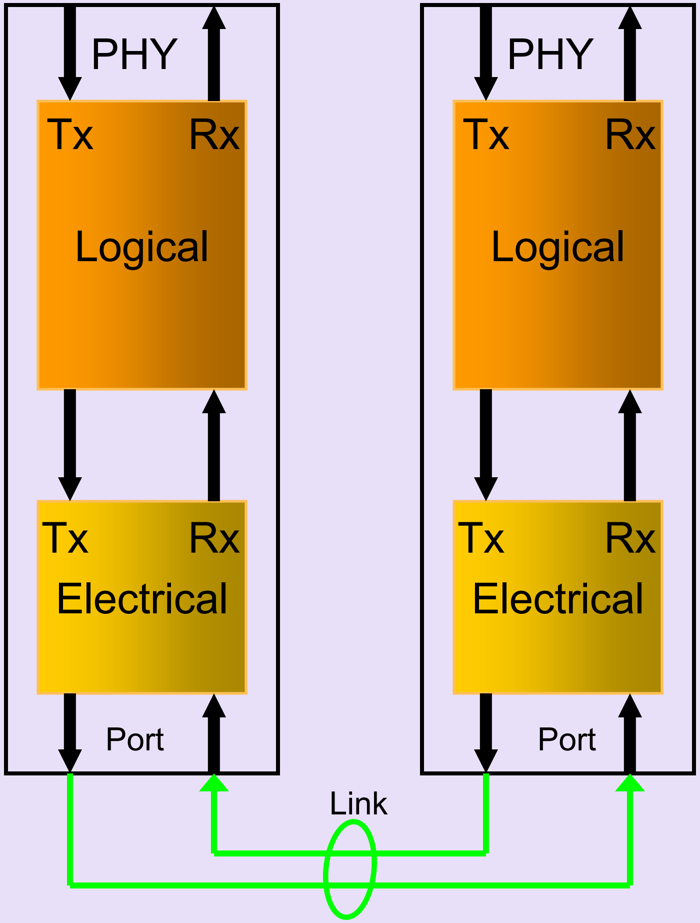
‘A’ checks if an ACK. TLP’s with sequence number <= current one are removed from buffer. If a NAK then all unacknowledged TLP’s are resent

A NAK is sent for bad & an ACK is sent for good TLP’s



Physical Layer Function

- Provides the physical connection between devices
- Logical Functions
  - ✓ Link training and status
  - ✓ Packet framing, Data striping/Data assembly
  - ✓ Data scramble, 8B/10B encode/decode
  - ✓ Symbol lock
- Electrical Functions
  - ✓ Receiver detect
  - ✓ Receive clock recovery
  - ✓ Bit lock, Serialization/Deserialization
  - ✓ LVDS signaling



Agenda

- PCI Express Overview, Components & Architecture
- PCI Express Protocol Layers
- ➡ Needs of Communication Systems & PCIe
- PCI Express in Communication Systems
- Summary

The Challenge

In general, too many interconnects

- Goals
  - ✓Minimize the number of interconnects
    - – Reality: there will always be multiple interconnects
  - ✓Technically suitable and economically viable
    - – Relieve the need to create proprietary technologies
    - – Provide broad based industry acceptance & economies of scale
  - ✓Interoperable multi-sourced switches, bridges & end-points

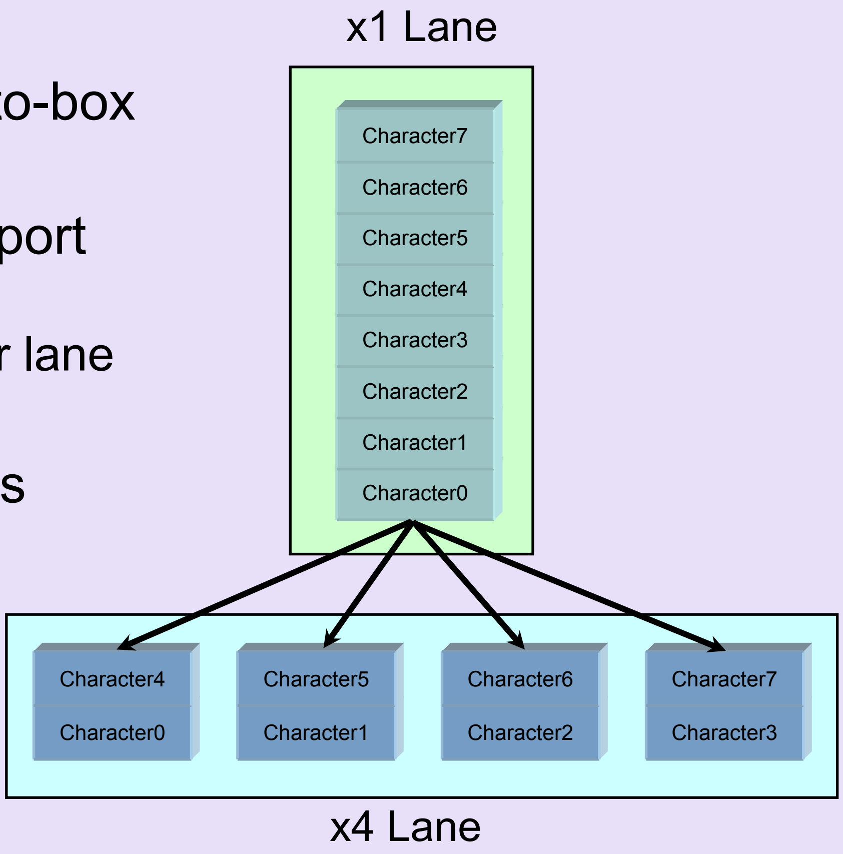
High Speed Serial Interface with Economies of Scale

Functional Needs

- Connectivity, Bandwidth and Scalability
- Data Integrity and Reliability
- Serviceability and Availability
- Quality of Service

Connectivity, Bandwidth & Scalability

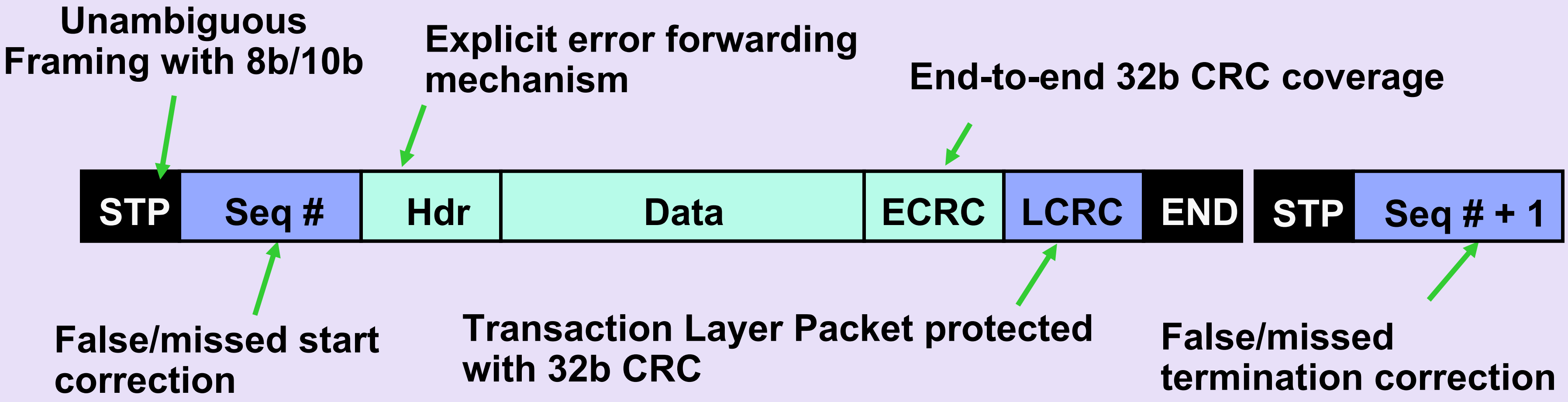
- Chip-to-chip, board-to-board, box-to-box
  - ✓ Cable spec in development
- Combining multiple lanes in wider port (x1, x4, x8, x16, x32)
  - ✓ Current spec supports 2.5GB/s per lane
  - ✓ Gen-2 in definition
- Byte striping used for multiple lanes
- No sideband signals
  - ✓ 8b/10b encoding used



x4 Byte Striping

Data Integrity Support

- Data Link Layer Mechanisms (Link/Local):
  - ✓ TLPs protected using 32bit CRC
  - ✓ DLLPs protected using 16bit CRC
  - ✓ TLP error recovery through Data Link-level retry
  - ✓ Supplemental coverage through 8b/10b
  - ✓ Loss of packets detected using Sequence Numbers
- Transaction Layer Mechanisms (End-to-End):
  - ✓ Optional coverage using 32bit CRC
  - ✓ Data Poisoning capability



Unambiguous Framing with 8b/10b

Explicit error forwarding mechanism

End-to-end 32b CRC coverage

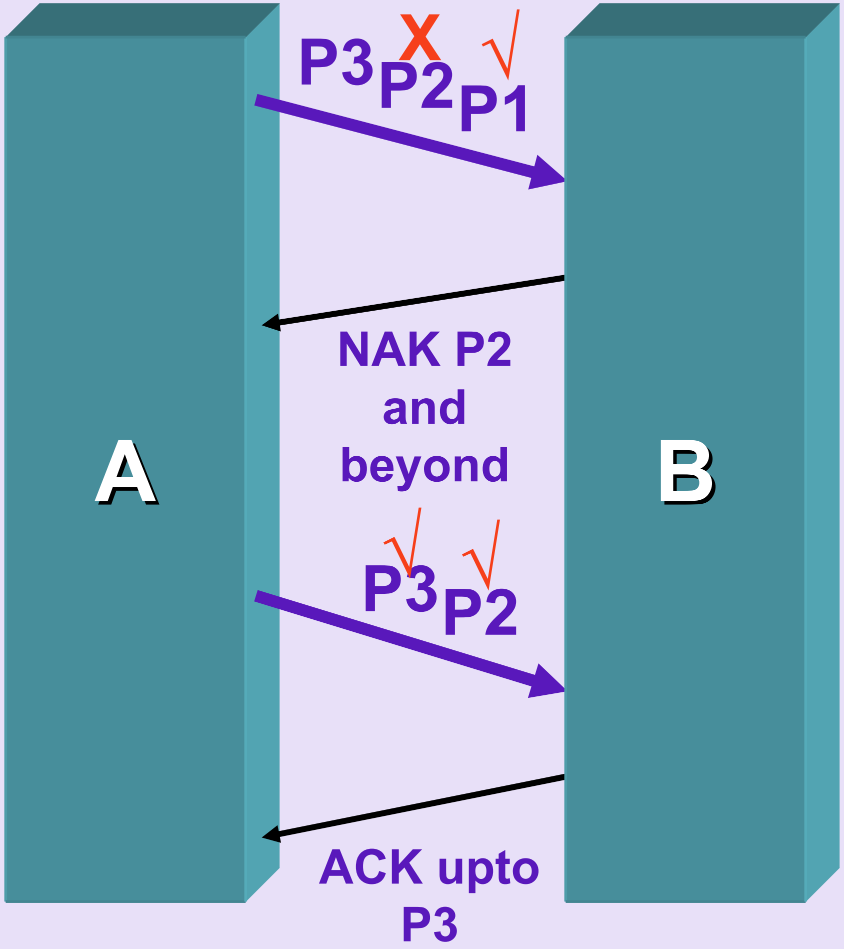
False/missed start correction

Transaction Layer Packet protected with 32b CRC

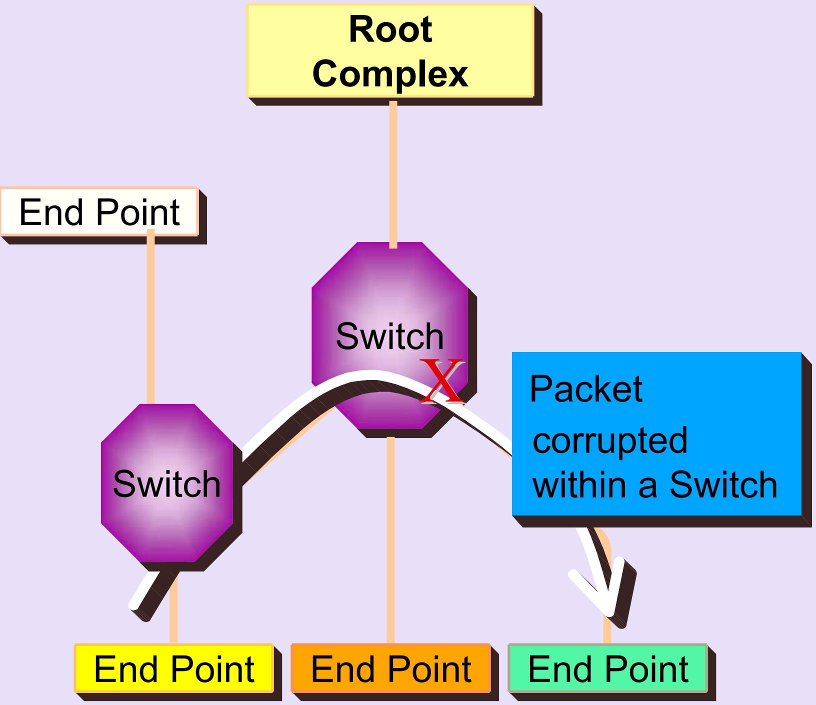
False/missed termination correction

Link Data Integrity – Retry Example

1. Three TLPs sent from A to B
2. Packet 2 corrupted
3. B detects corruption and issues Nak DLLP
4. A resends Packet 2 and following Packet
5. B acknowledges successful receipt of Packets



End-to-End Data Integrity - ECRC



- Component internal errors are critical
  - ✓ Header errors → TLP misrouting
  - ✓ Data corruption → application and system failure
- End-to-end data integrity using ECRC
  - ✓ Protecting from system-wide errors
  - ✓ Enabling upper layers error recovery
- ECRC basics:
  - ✓ Optional Capability – additional 32bit field (part of TLP)
  - ✓ Generated by the source component – applies to all invariant TLP fields
  - ✓ Switches must pass ECRC unchanged
  - ✓ Checked in the destination component – resulting behavior is device specific

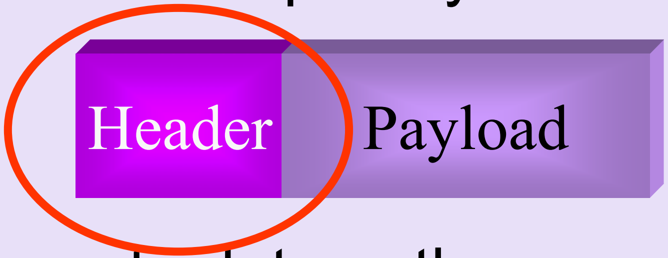
PCI Express Hot Plug

- PCI Hot Plug enables add or remove of PCI add-in device without interrupting normal system operation or requiring a power down/system reset
  - ✓ Root ports and downstream ports of switches are the hot pluggable ports in a PCI Express hierarchy
  - ✓ Elements of the Standard hot plug usage model derived from SHPC
  - ✓ Hot plug registers are integral part of the PCI Express registers
    - Do not require a separate set of memory mapped registers like PCI SHPC
  - ✓ Native hot plug solution is specific to PCI Express
    - SHPC continues to be the mechanism for parallel bus PCI implementations

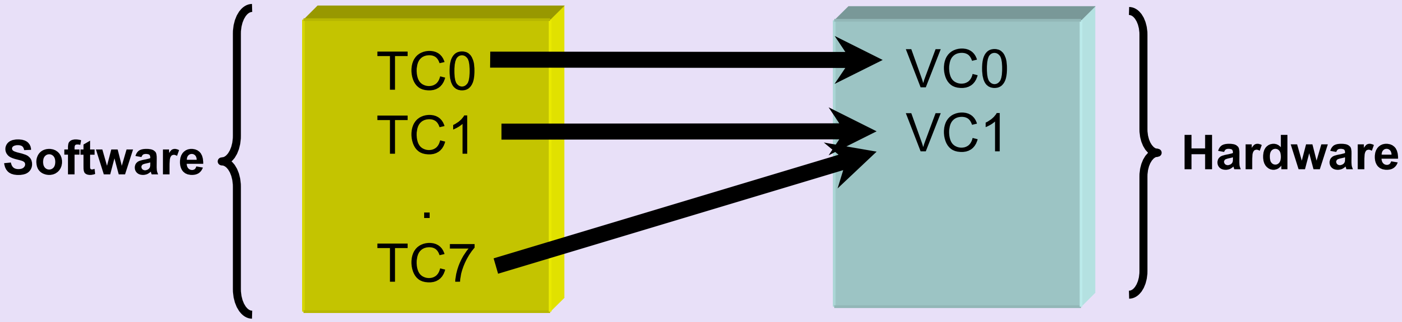
PCI Express Enables Hot Plug Capability for the Mainstream

Quality of Service

- Traffic Classes (TC)
  - ✓ Software-controlled method to add traffic priority
  - ✓ Part of HEADER field in a TLP



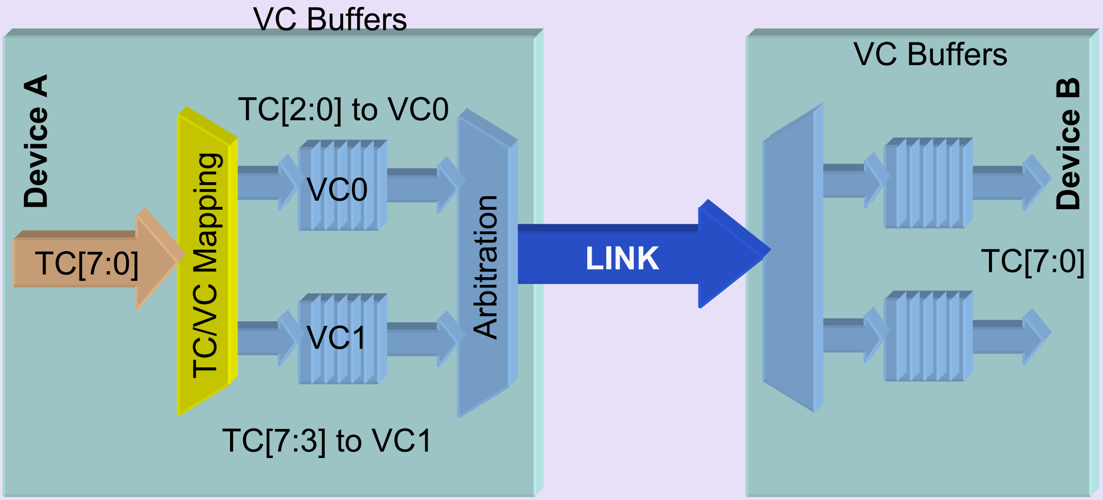
- Virtual Channels (VC)
  - ✓ Hardware method to provide separate data paths
  - ✓ Part of queue structure in switches and bridges
  - ✓ Hardware may have fewer than 8 VCs



VCs and TCs

QoS though VC’s and TC’s

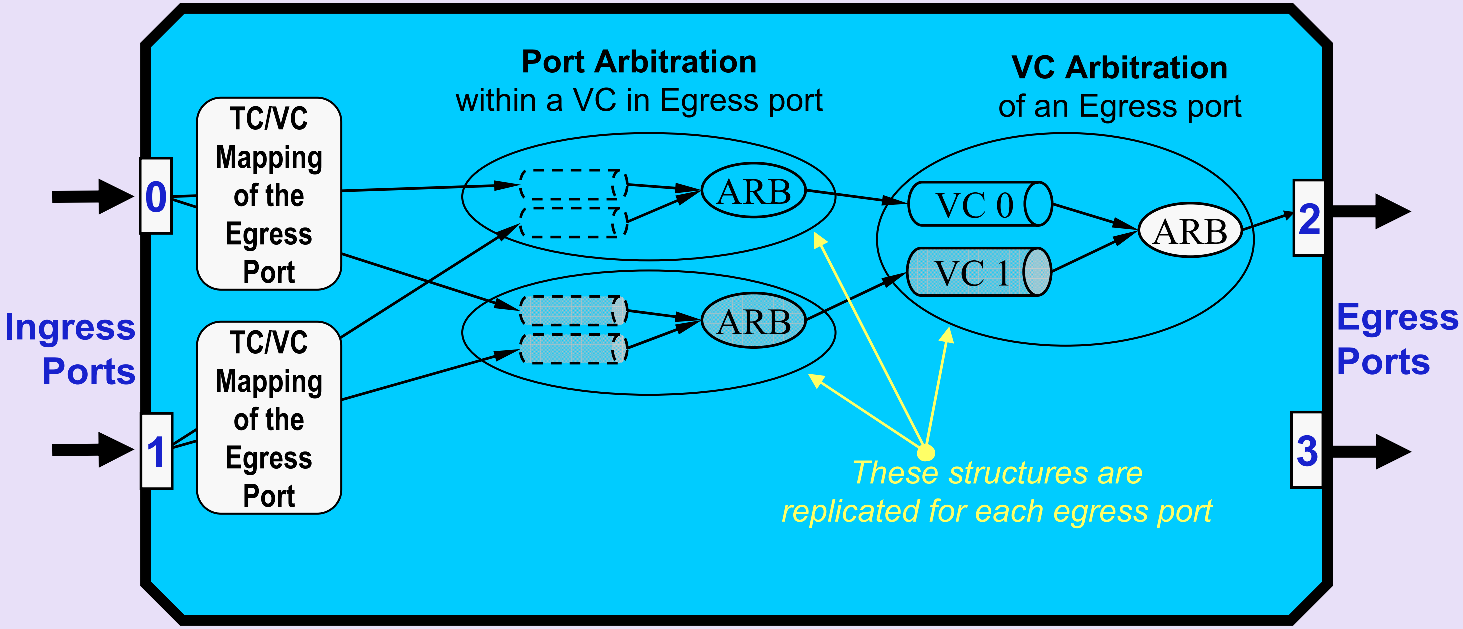
- ✓ Software decides what TC a packet should use
- ✓ VC’s allow multiple independent logical data flows over the link
- ✓ TC’s are mapped into VC’s
- ✓ Multiple TC’s may be mapped into one VC
- ✓ TC/VC mappings can be configured per port
- ✓ Ingress and egress payload credits are programmable per VC, port and transaction type



Arbitration

- ✓ TC’s are routed through switches with different priorities based on arbitration policy
  - Switches use Port arbitration and VC arbitration
  - TC mapping, Port and VC arbitration schemes can be configured on a per port basis – stored in PCI Express Extended Capability set.
  - Arbitration schemes include;
    - Hardware Fixed
    - Weighted Round Robin (32)
    - Weighted Round Robin (64)
    - Weighted Round Robin (128)
    - Weighted Round Robin (256)
    - Timed weighted (128)
  - Arbitration schemes are set up in VC Arbitration Tables and Port Arbitration Tables

Port & VC Arbitration


Port Arbitration
within a VC in Egress port
VC Arbitration
of an Egress port
TC/VC Mapping of the Egress Port
TC/VC Mapping of the Egress Port
0
1
2
3
Ingress Ports
Egress Ports
ARB
ARB
ARB
VC 0
VC 1
These structures are replicated for each egress port

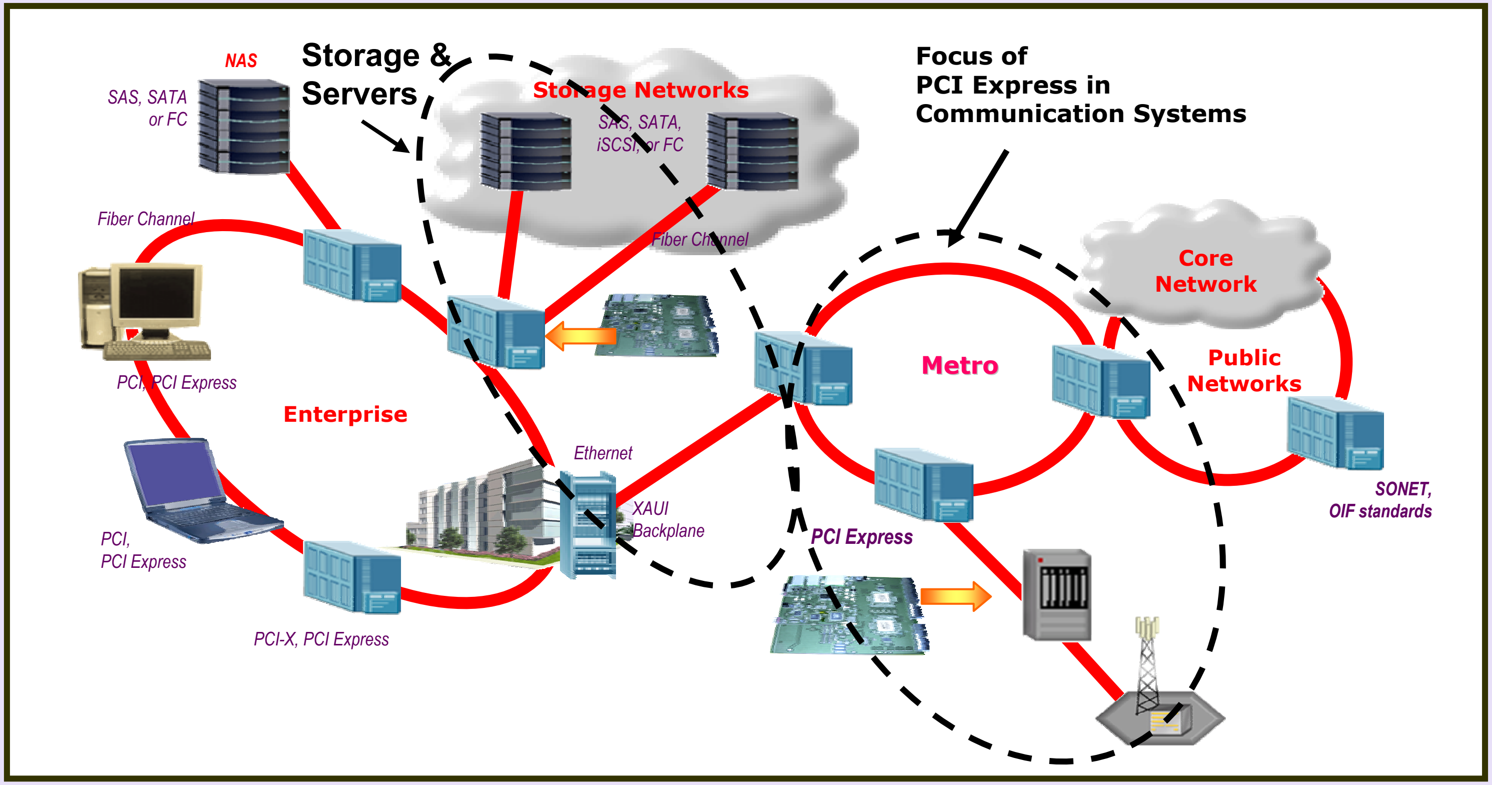
- Port Arbitration:
  - ✓ Traffic targeting same VC/Egress Port
  - ✓ Fixed Round-Robin (RR), programmable Weighted RR, programmable Time-based WRR
- VC Arbitration:
  - ✓ Traffic from different VC competing for the Link
  - ✓ Fixed priority, RR, programmable WRR

Agenda

- PCI Express Overview, Components & Architecture
- PCI Express Protocol Layers
- Needs of Communication Systems & PCIe
- PCI Express in Communication Systems
- Summary

PCI Express in Communications

- PCI Express meets the interconnect needs of the communications industry
- Suited for Metro, Edge, Mobile and Storage network equipment

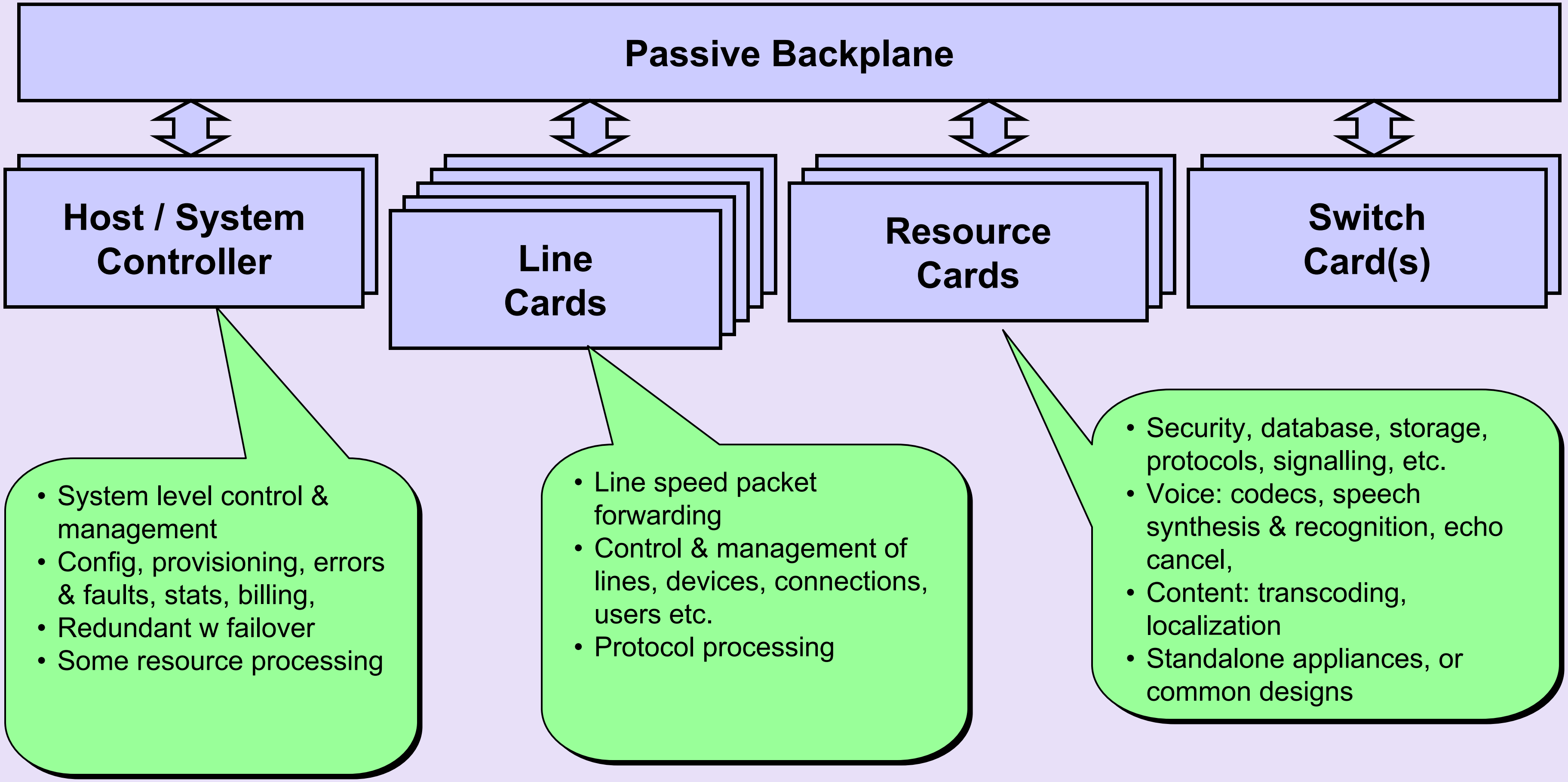


Single Host Interconnect

- PCI Express best suited as a local interconnect of single-host systems.
  - Connects the host with the I/O subsystems
  - Subsystems may be on same board, or separate I/O cards
  - Serves the needs of both control and data traffic
- Supports single board, mezzanine and bladed systems
- Communications needs of
  - ✓ Peer-to-peer transfers are supported thru switching
  - ✓ Multi-host can be supported with non-transparent bridge implementation (same as PCI)

Reliable Link layer with Flow Control

Chassis-Based System



Passive Backplane

Host / System Controller

- System level control & management
- Config, provisioning, errors & faults, stats, billing,
- Redundant w failover
- Some resource processing

Line Cards

- Line speed packet forwarding
- Control & management of lines, devices, connections, users etc.
- Protocol processing

Resource Cards

- Security, database, storage, protocols, signalling, etc.
- Voice: codecs, speech synthesis & recognition, echo cancel,
- Content: transcoding, localization
- Standalone appliances, or common designs

Switch Card(s)

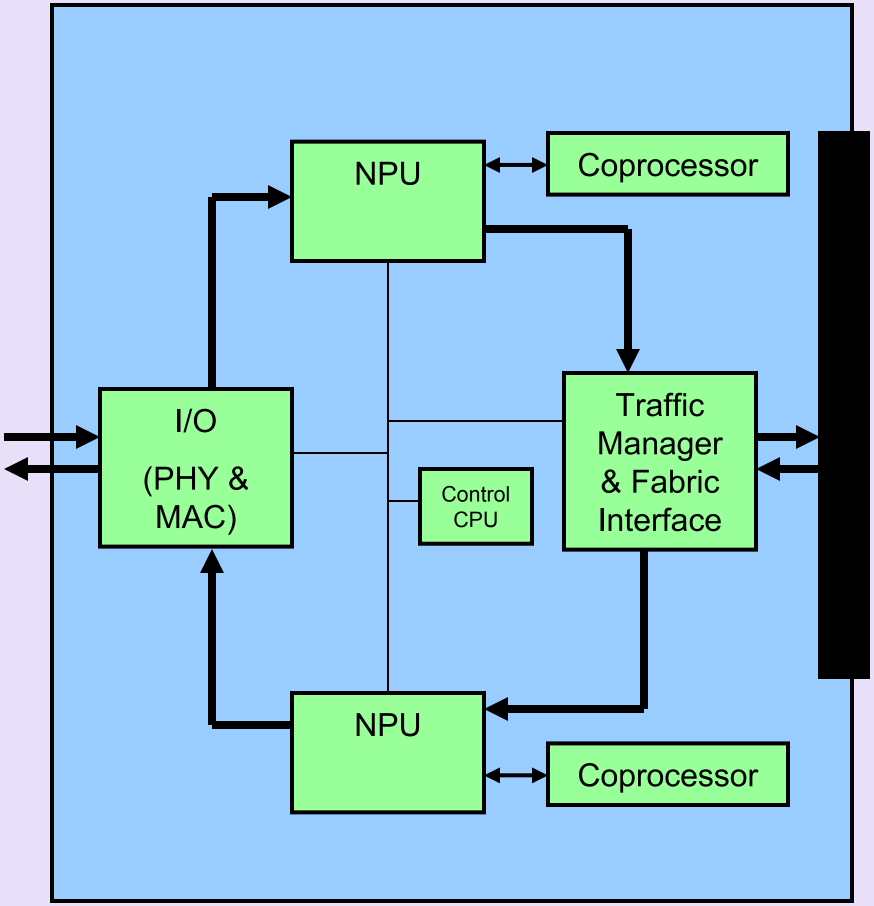
PCI Express Backplanes

- Analogous situation to PCI
- Single host + I/O cards
- Dual redundant hosting requires non-transparent bridging
  - ✓ Non-transparent function may be embedded in switch ports
- Distributed processing moves to system fabric
  - ✓ Issues are scalability, system management, etc.
  - ✓ Replace a shared bus with switch fabric
  - ✓ May integrate host controller on the Switch Fabric blade

Line Card Architecture (now)

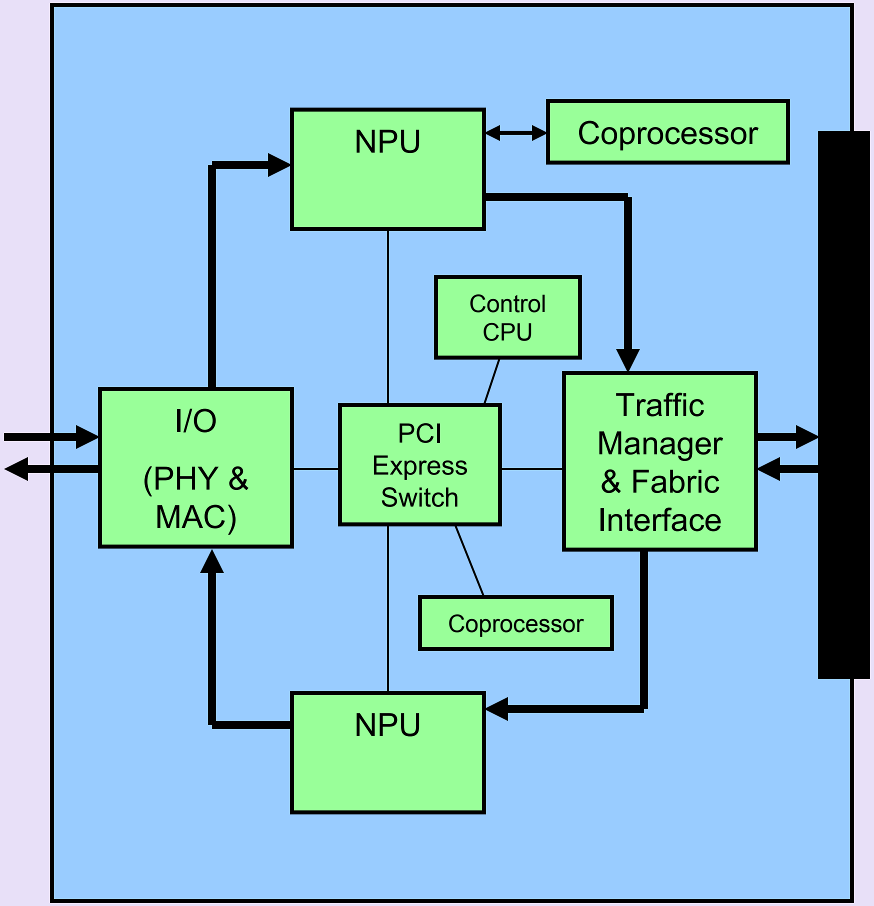
Current implementations:

- ✓ Fixed Configurations
- ✓ Chips connected in discrete daisy chain fashion
- ✓ Optimized for particular applications
- ✓ Devices must pass/process traffic destined for another device

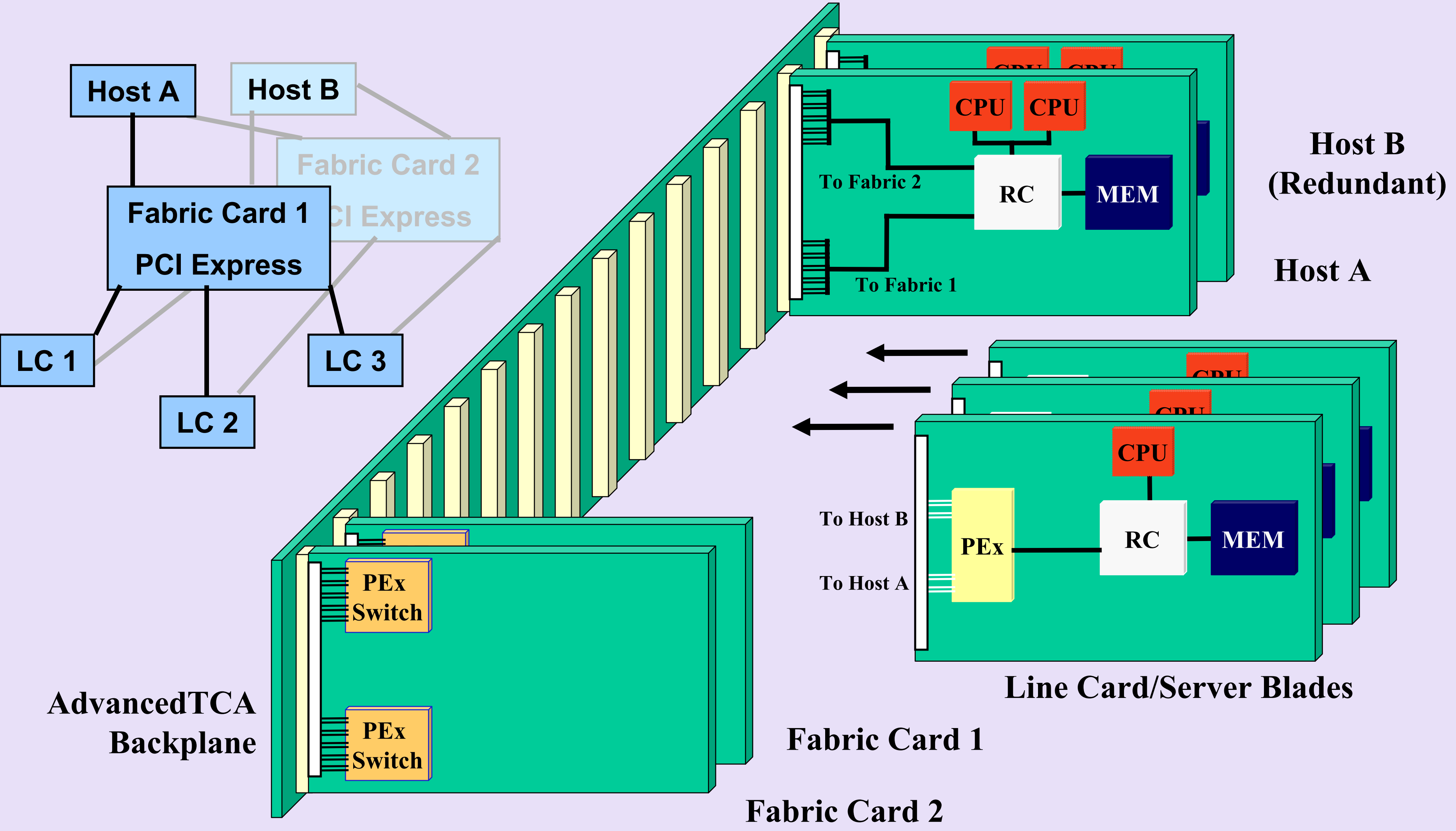


Line Card Architecture (future)

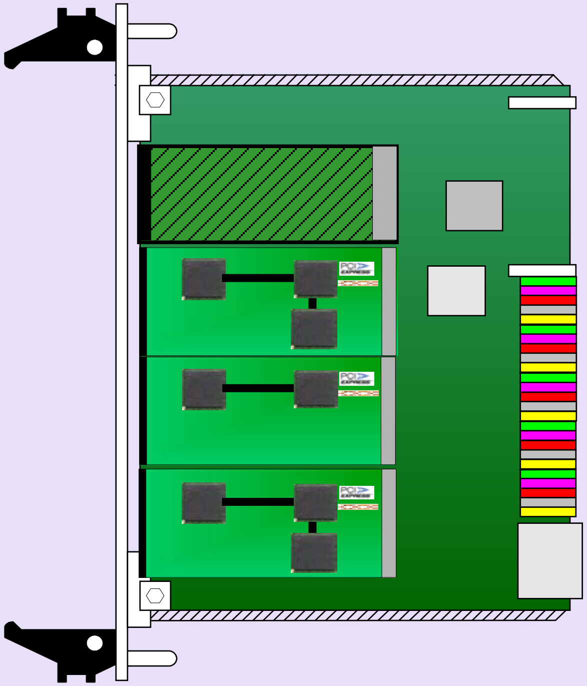
- PCI Express Switch based architecture
  - ✓ more flexible
  - ✓ scalable
  - ✓ reusable architecture
  - ✓ fewer traces ->cheaper boards
  - ✓ no multi-drop issues



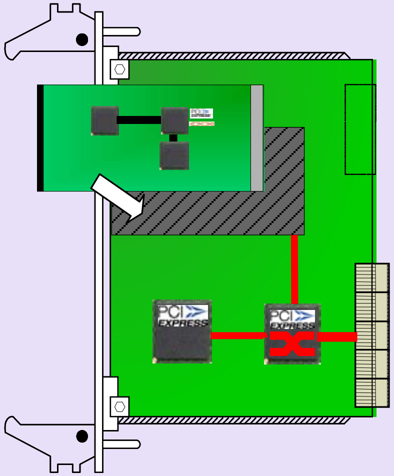
PCI Express in ATCA (3.4)



I/O Mezzanine Form Factors

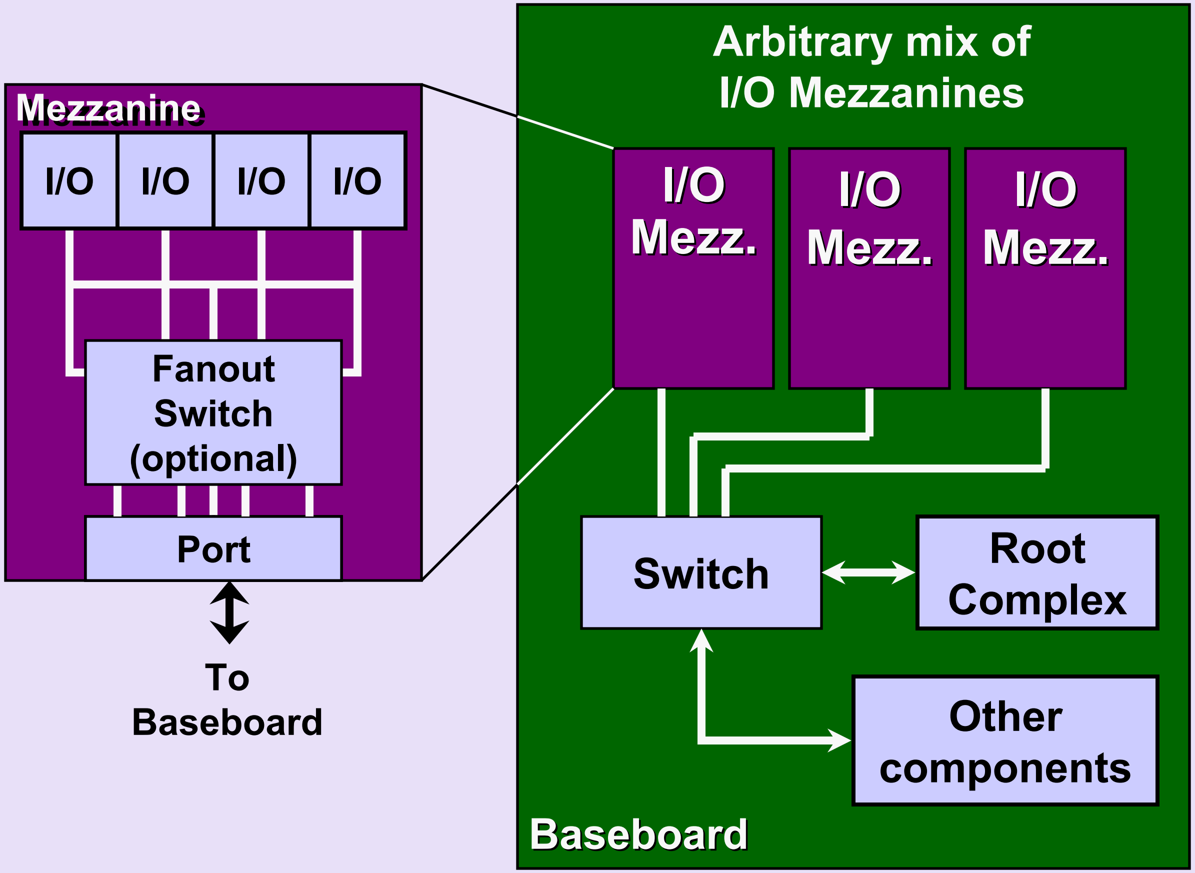


AdvancedTCA
AMC Mezzanine Card



PICMG Express
XMC Mezzanine Card

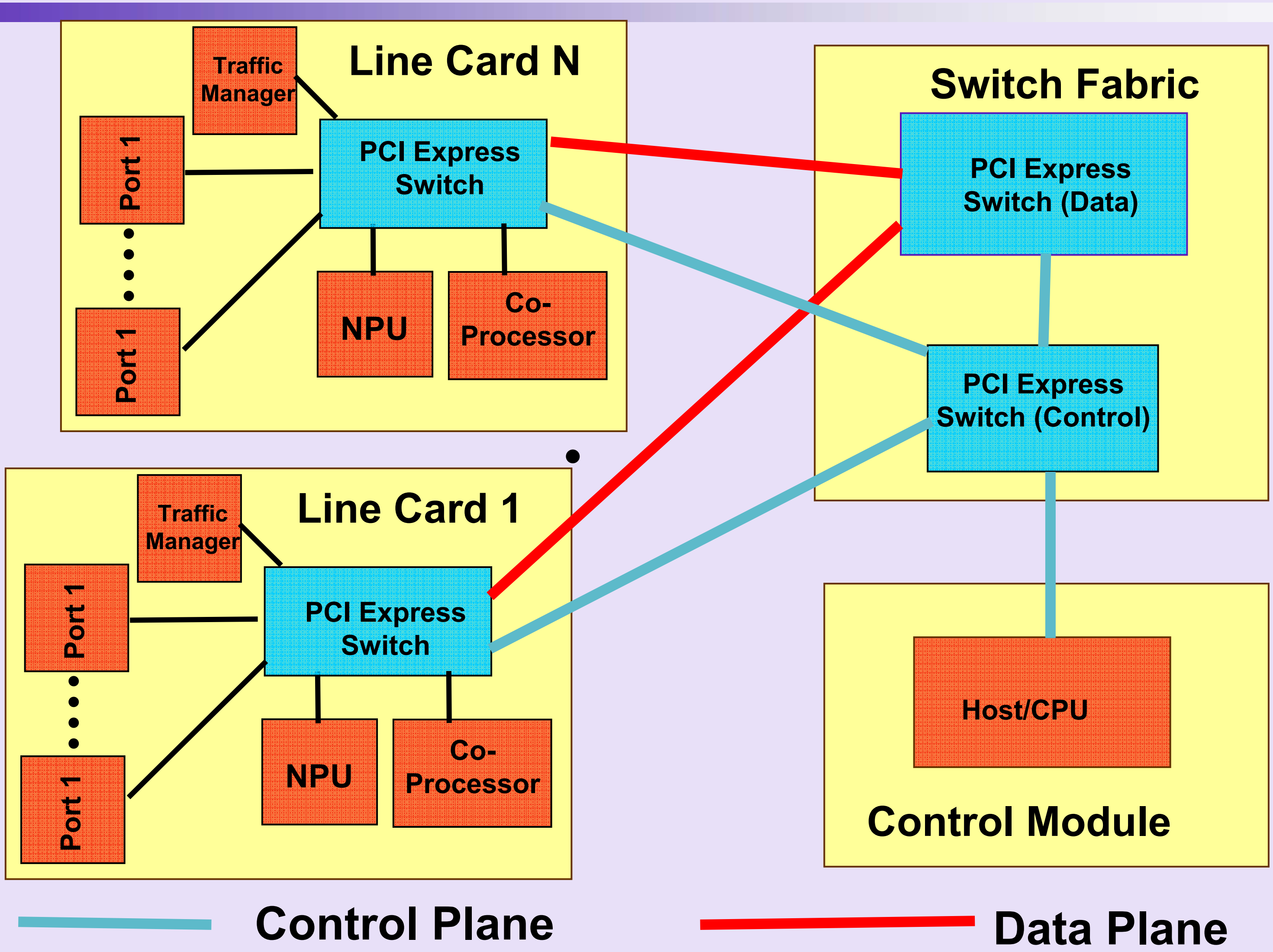
I/O Mezzanine Cards



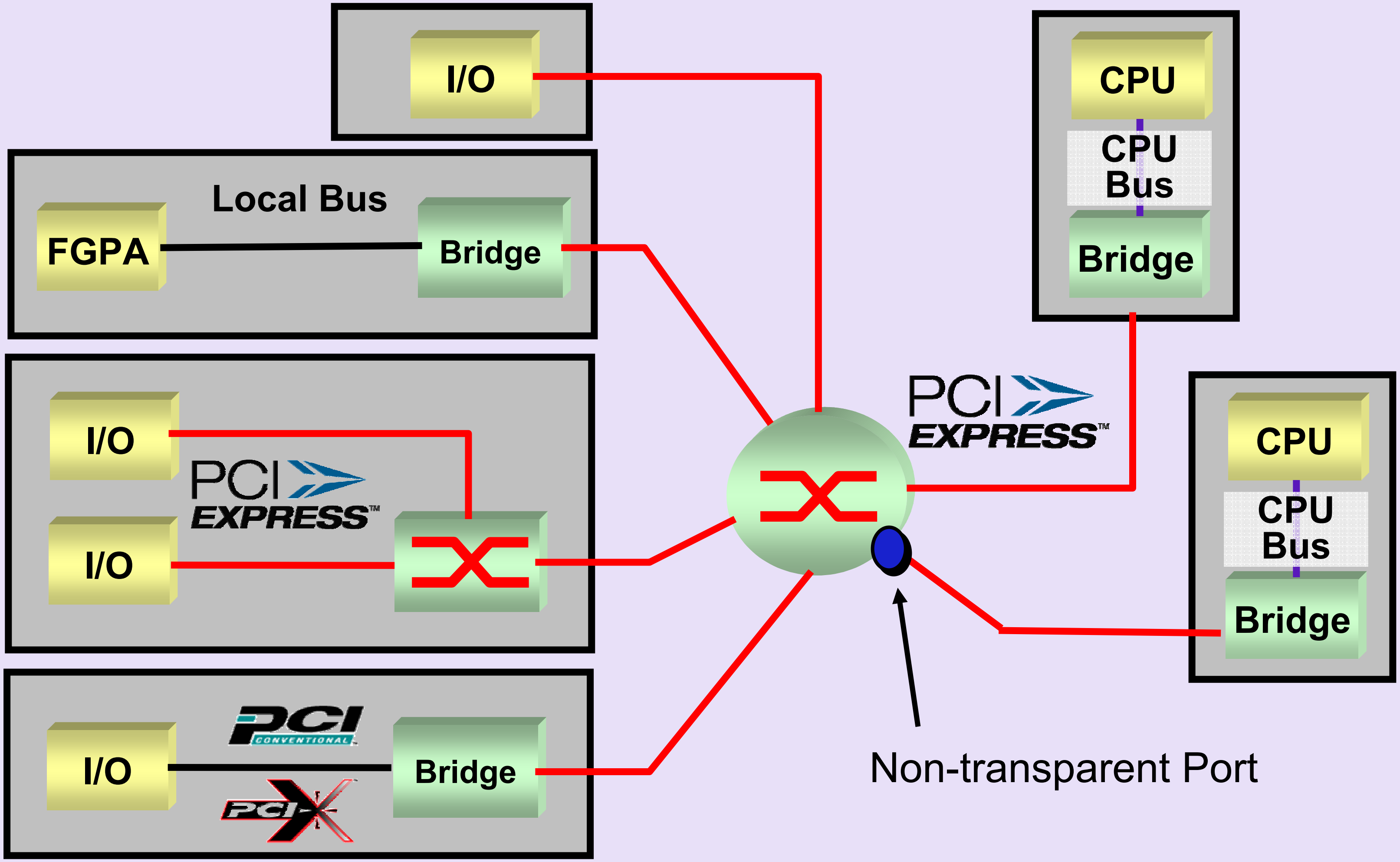
- Follows I/O device migration to PCI Express
- Supports multiple I/O mezzanines
- Host CPU (root complex) could be on the baseboard or on a mezzanine card
- Processor mezzanine interconnect is electrically similar to a mini backplane

* Other brands and names may be claimed as the property of others.

Low/Mid Range Systems



Control Plane (Switch)



Non-transparent Port

Agenda

- PCI Express Overview, Components & Architecture
- PCI Express Protocol Layers
- Needs of Communication Systems & PCIe
- PCI Express in Communication Systems
- ➡ Summary

Summary

- Mature Specification (1.0a)
- High speed serial interconnect technology
- Packet based layered protocol
- Full compatibility with PCI based software
- Data integrity at link and transaction layers
- Flow control for optimum bandwidth/buffer usage
- Hot plug and power management for RAS
- Traffic Classes and Virtual Connections for quality of service (QoS) support
- Valuable features for communication systems design
- Serves control plane and low/mid range data plane
- Leverage and re-use existing PCI software

Thank you for attending the
PCI-SIG Developers Conference 2004.

For more information please go to
www.pcisig.com

PCI
SIG™