SNIA | Data, Storage & Networking

Everything You Wanted to Know About PCIe But Were Too Proud to Ask

Live Webinar
July 29, 2025
8:00 am PT / 11:00 am ET

Martin Chao
Broadcom

Tim Lustig
NVIDIA

Today's Presenters

Tim Lustig
Director, Corporate Ethernet Marketing
NVIDIA
Moderator

Martin Chao
Principal R&D Application Engineer
Broadcom

The SNIA Community

200
industry leading organizations

2,000
active contributing members

50,000
IT end users & storage pros worldwide

SNIA | DSN A SNIA Community

SNIA | Data, Storage & Networking

What We Do

Drive the awareness and adoption of a broad set of technologies, including:

- Storage Protocols (Block, File, Object)
- Traditional and software-defined storage
- Disaggregated, virtualized and hyperconverged
- AI, including storage and networking considerations
- Edge implementation opportunities and factors
- Storage and networking security
- Acceleration and offloads
- Programming frameworks
- Sustainability

How We Do It

By delivering:

- Expert webinars and podcasts
- White papers
- Articles in trade journals
- Blogs
- Social Media
- Presentations at industry events

Logistics

- The slides are available under the attachments tab at the bottom of your console.
- Questions are welcome!
- Please rate the session and provide feedback!
- Want more sessions like this or other topics, let us know!
  - JOIN US! We meet on Thursday mornings at 11:00 AM eastern.
  - Email dsn-chair@snia.com if you have questions.

SNIA Legal Notice

- The material contained in this presentation is copyrighted by SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Agenda

- PCI Express Overview
- PCIe Enumeration / AER / DPC / Hot-Plug
- Non-Transparent Bridge
- SR-IOV
- PCIe Switch in AI Application
- New in PCIe Generation 6

PCI Express Overview

Introduction to PCI Express (PCIe)

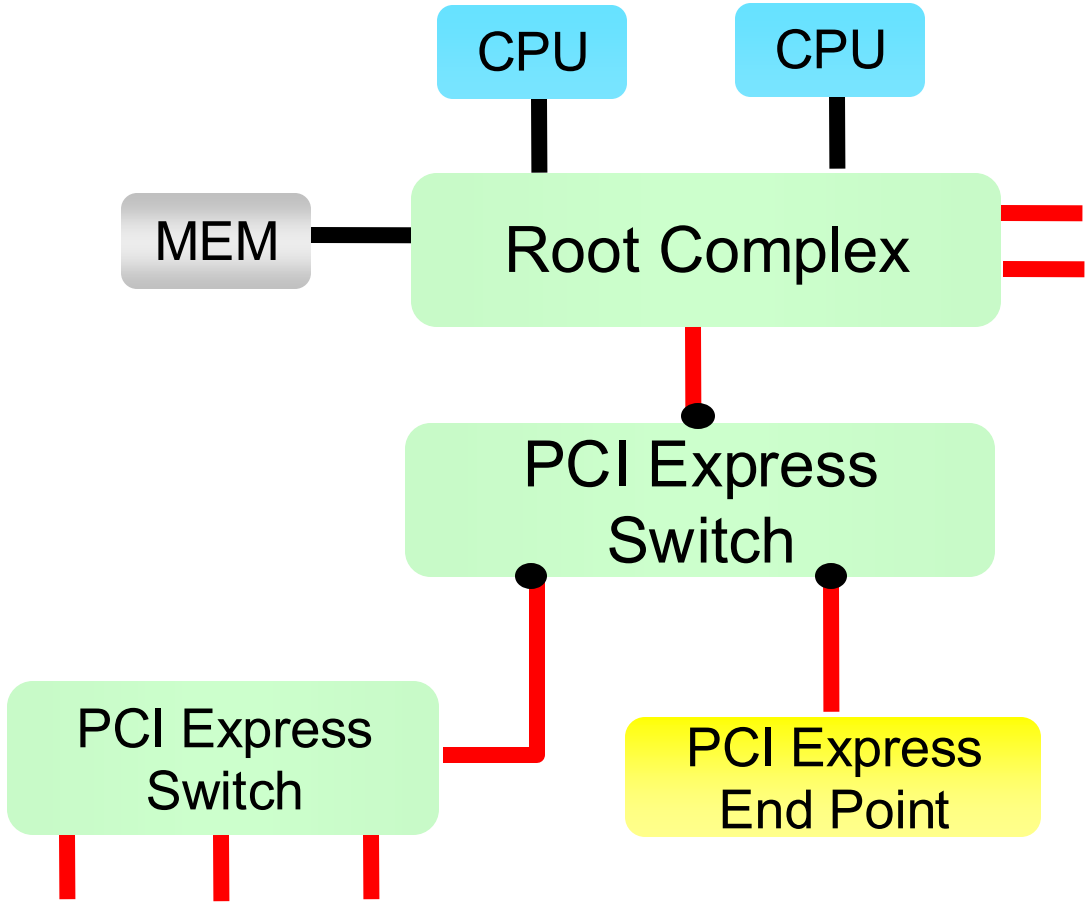
- Peripheral Computer Interconnect - standardized interface for motherboard components like graphics, memory and storage.
- PCIe brought a major shift from parallel bus model (PCI or PCI-X) to serial bus
  - Similar to serial interfaces like Serial Attached SCSI (SAS), SATA, InfiniBand or Fibre Channel
- PCI compatible software model
  - Same address spaces as PCI
  - Supports existing legacy configuration space model
  - PCI Express add-in cards co-exist with PCI cards

PCI Express Generations

- Evolved several generations so far (Gen 1 ->2 -> 3,…. 6)
- Bandwidth doubled for every generation
- Encoding method changed from Gen2 to Gen3 and then from Gen5 to Gen6
  - Gen1 and Gen2 use 8b/10b encoding
  - Gen3 to Gen5 use 128/130 encoding (Add 2 bits as sync header for each 128-bit)
  - Gen6 uses 1:1 ratio by 256B FLIT

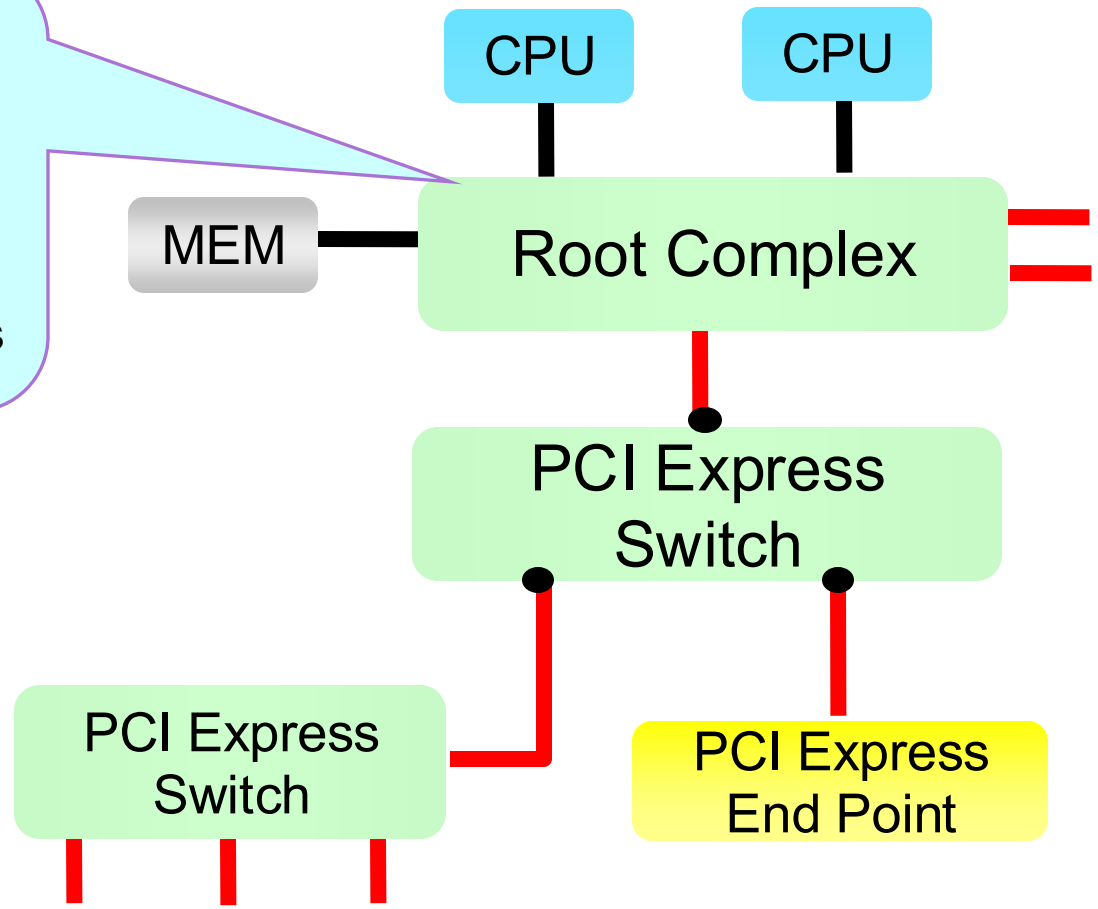
Generation	Speed	Encoding
Gen1	2.5GT/s	8b/10b
Gen2	5.0GT/s	8b/10b
Gen3	8GT/s	128/130b
Gen4	16GT/s	128/130b
Gen5	32GT/s	128/130b
Gen6	64GT/s	256B FLIT (1:1 ratio)

Typical PCI Express System



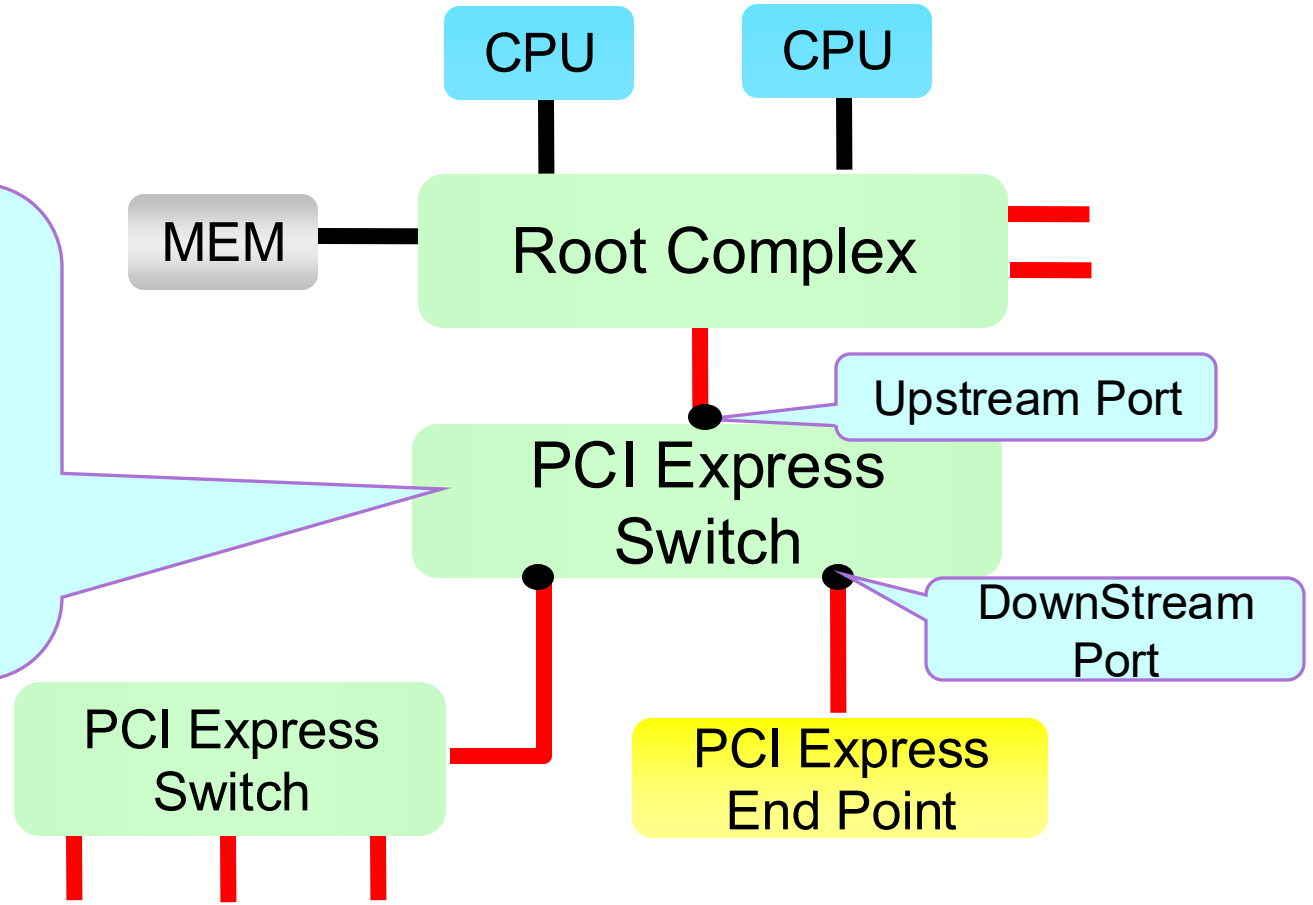
Typical PCI Express System

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

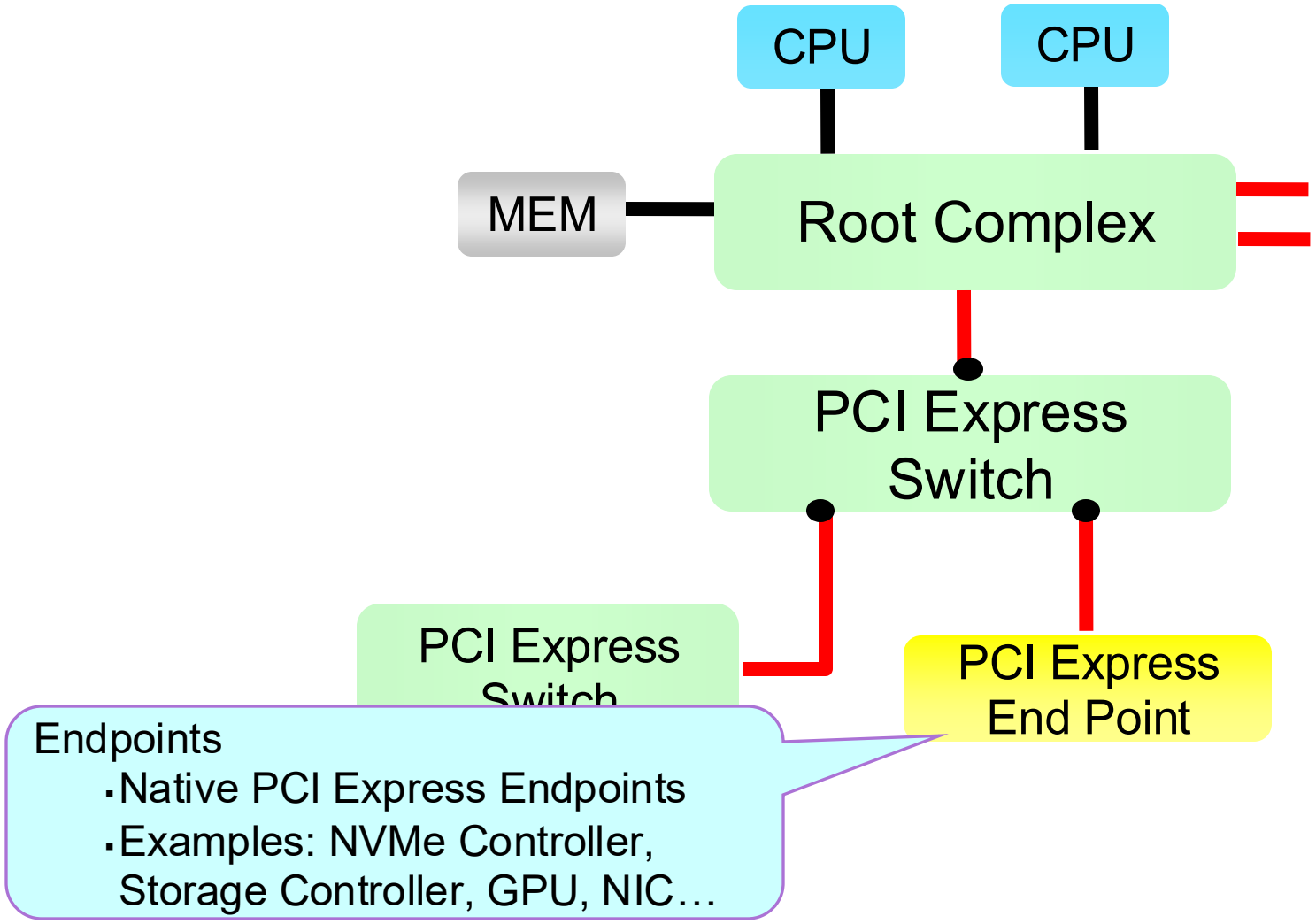


Typical PCI Express System

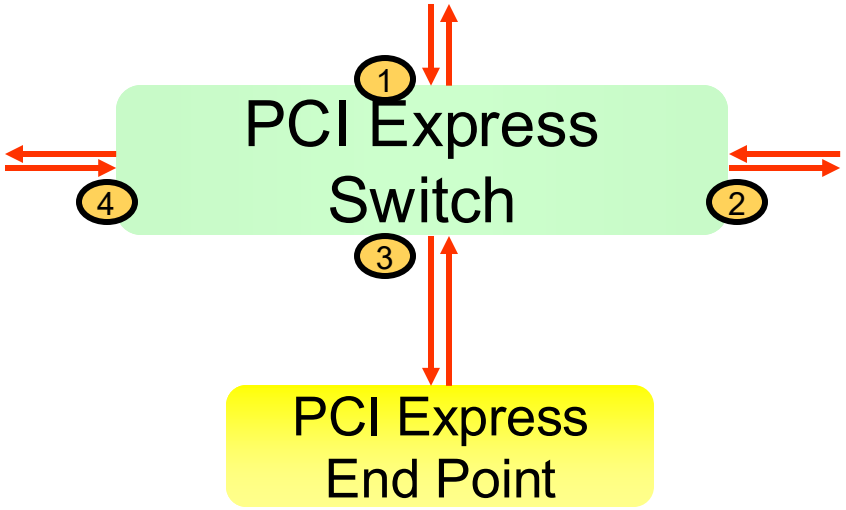
Switch
- Assembly of logical PCI-to-PCI bridges
- One upstream port directed towards root complex
- One or more downstream ports
- Switches can be stacked
- Peer to peer traffic allowed

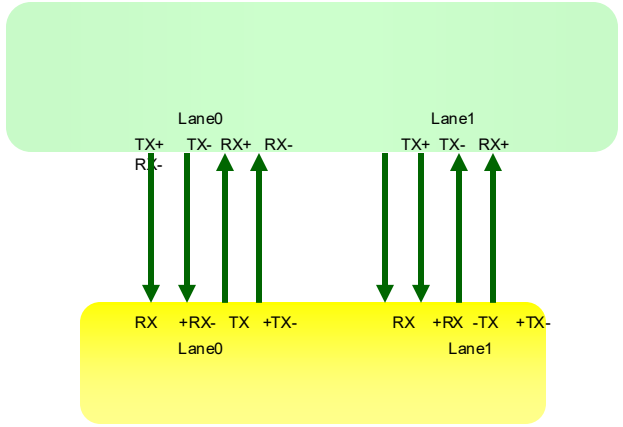


Typical PCI Express System



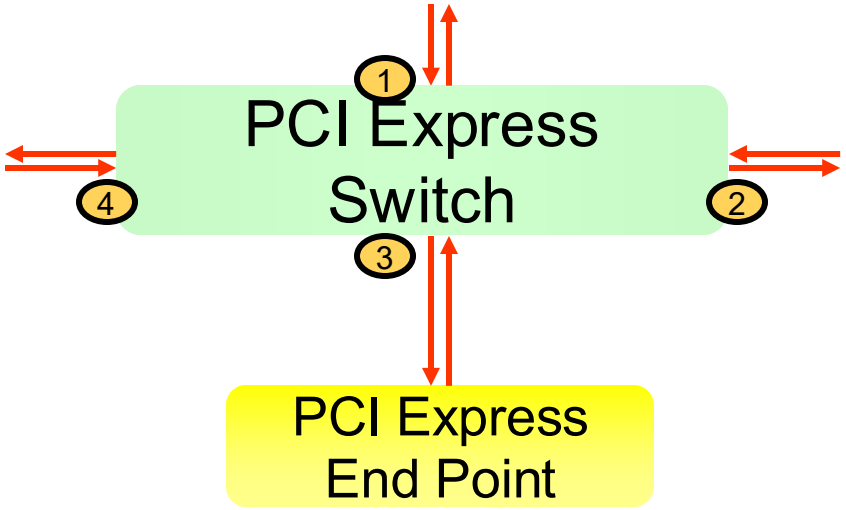
Terminology – Ports, Links, Lanes

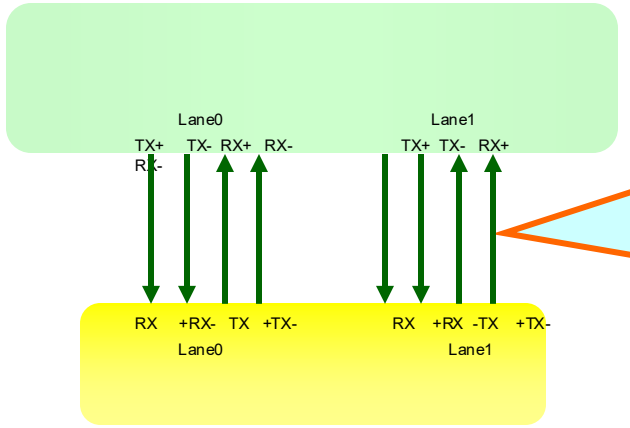




SNIA | DSN A SNIA Community

Terminology – Ports, Links, Lanes



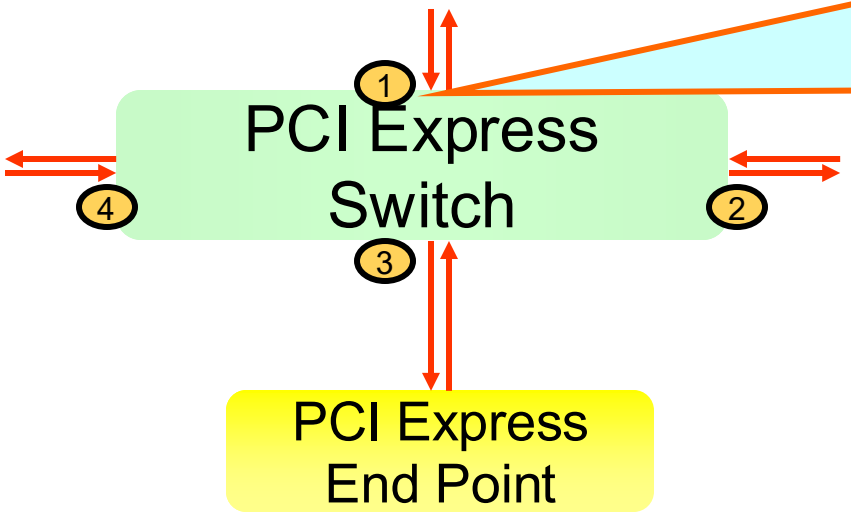


Lane

A set of differential signal pairs, one pair for transmission and one pair for reception. This shows 2 lanes i.e. a x2 link

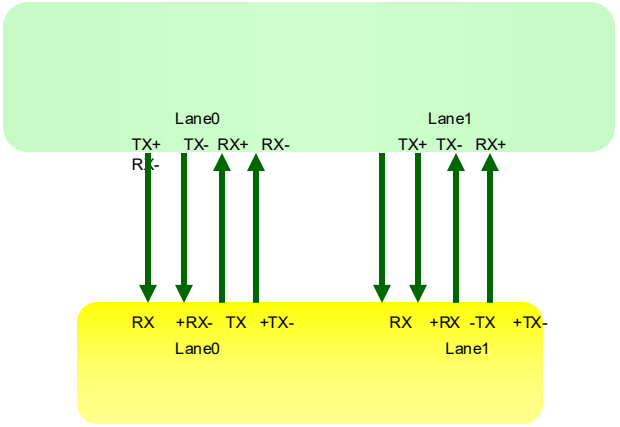
SNIA DSN A SNIA Community

Terminology – Ports, Links, Lanes

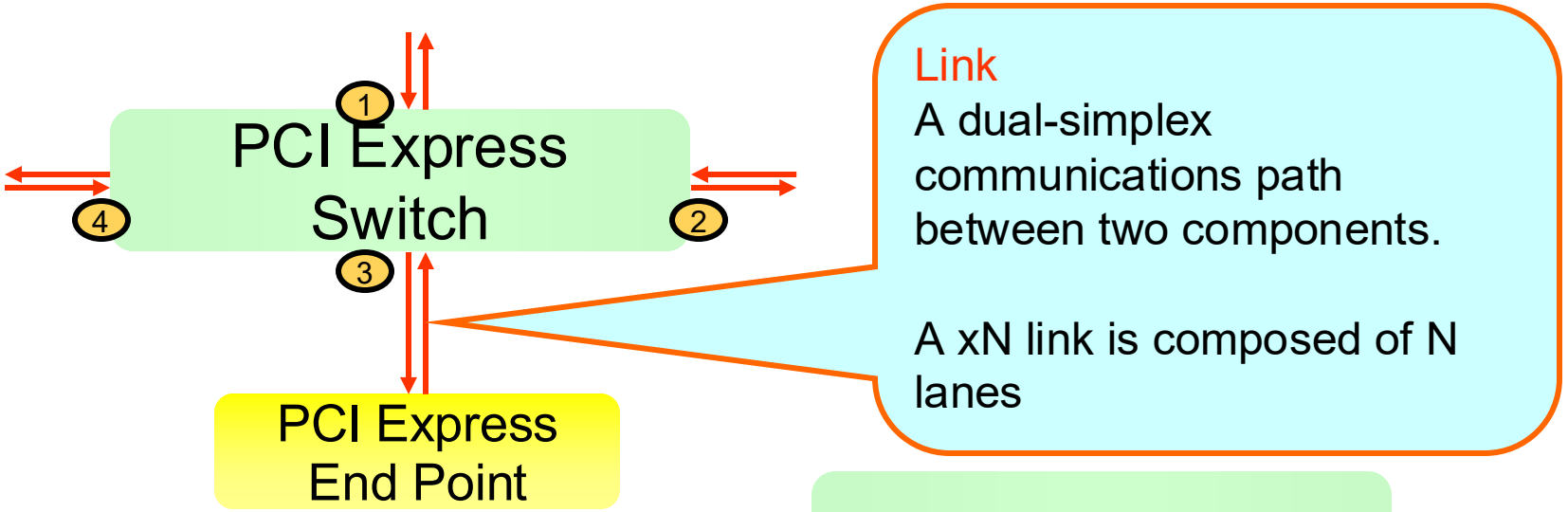


Port

A group of transmitters and receivers located on the same chip that define a link. Typically Switches are multi-port (this switch has 4 ports) end points only have 1 port. Each port is one end of the physical connection between two devices.



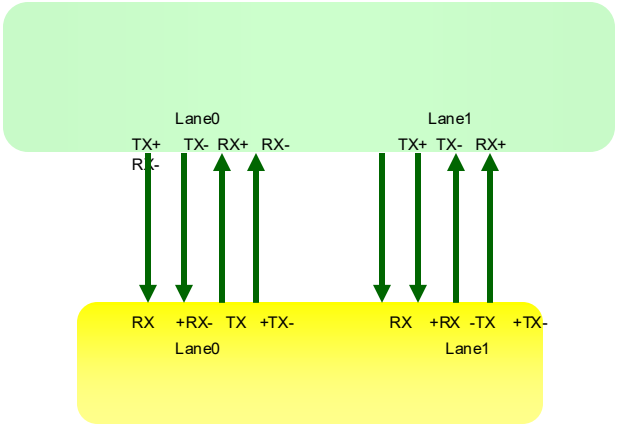
Terminology – Ports, Links, Lanes



Link

A dual-simplex communications path between two components.

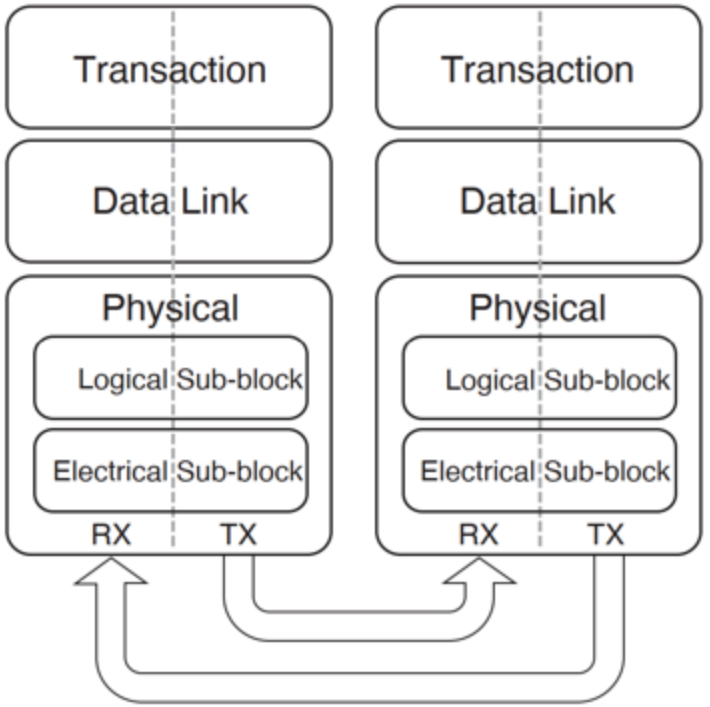
A xN link is composed of N lanes



SNIA | DSN A SNIA Community

PCI Express Layering

- Transaction layer creates the packet with data and header
  - Transaction Layer Packet processing
  - Storing negotiated and programmed configuration information
  - Assembly, disassembly, high-level error checking
  - Memory Read/Write, Configuration Read/Write, Messages, I/O Read/Write
- Data Link Layer
  - Adds Sequence number at the head of TLP and LCRC (Link CRC) for signal integrity at the tail
  - Initialization and updates of credit based flow control mechanism
  - Transaction Layer Packet acknowledgements (Ack/Nak)
- Physical Layer
  - Add Start and End symbol at the head and tail
  - Speed, link width, lane reversal and polarity auto negotiated during link training at initialization

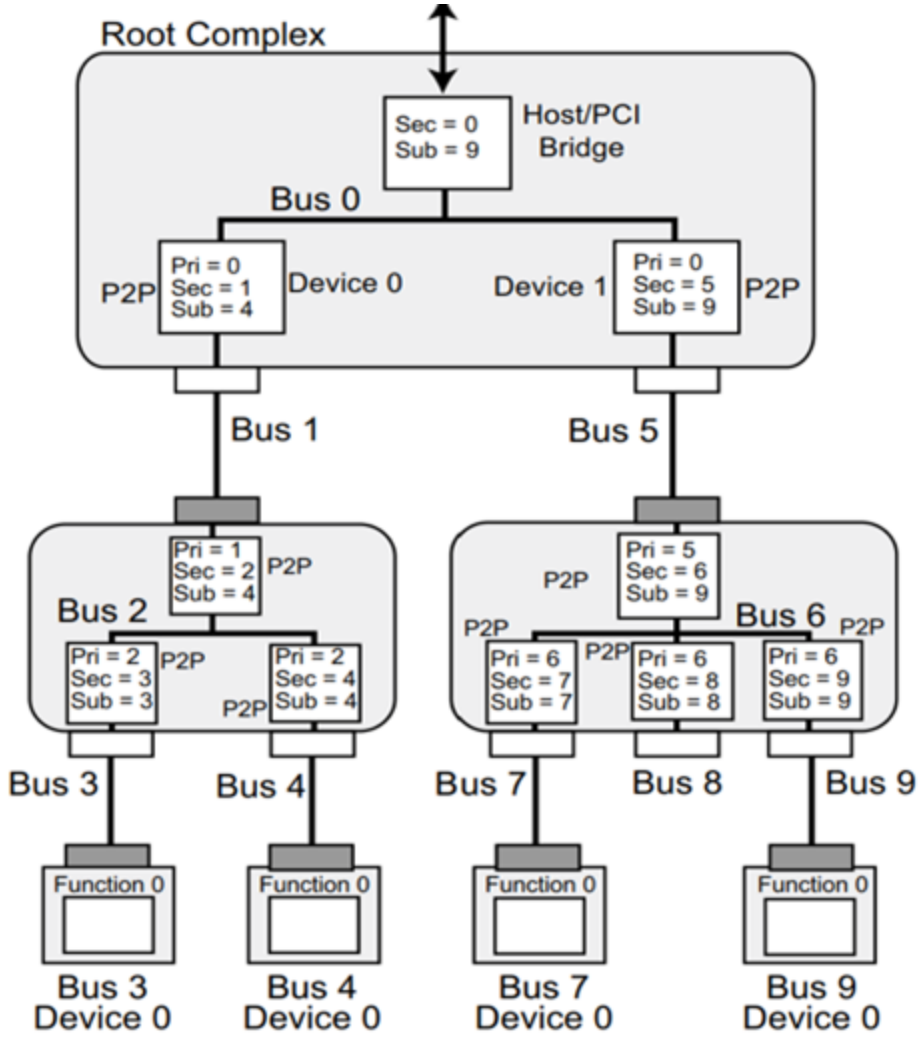


SNIA | DSN
A SNIA Community

PCIe Enumeration

Bus-Device-Function ID Assignment

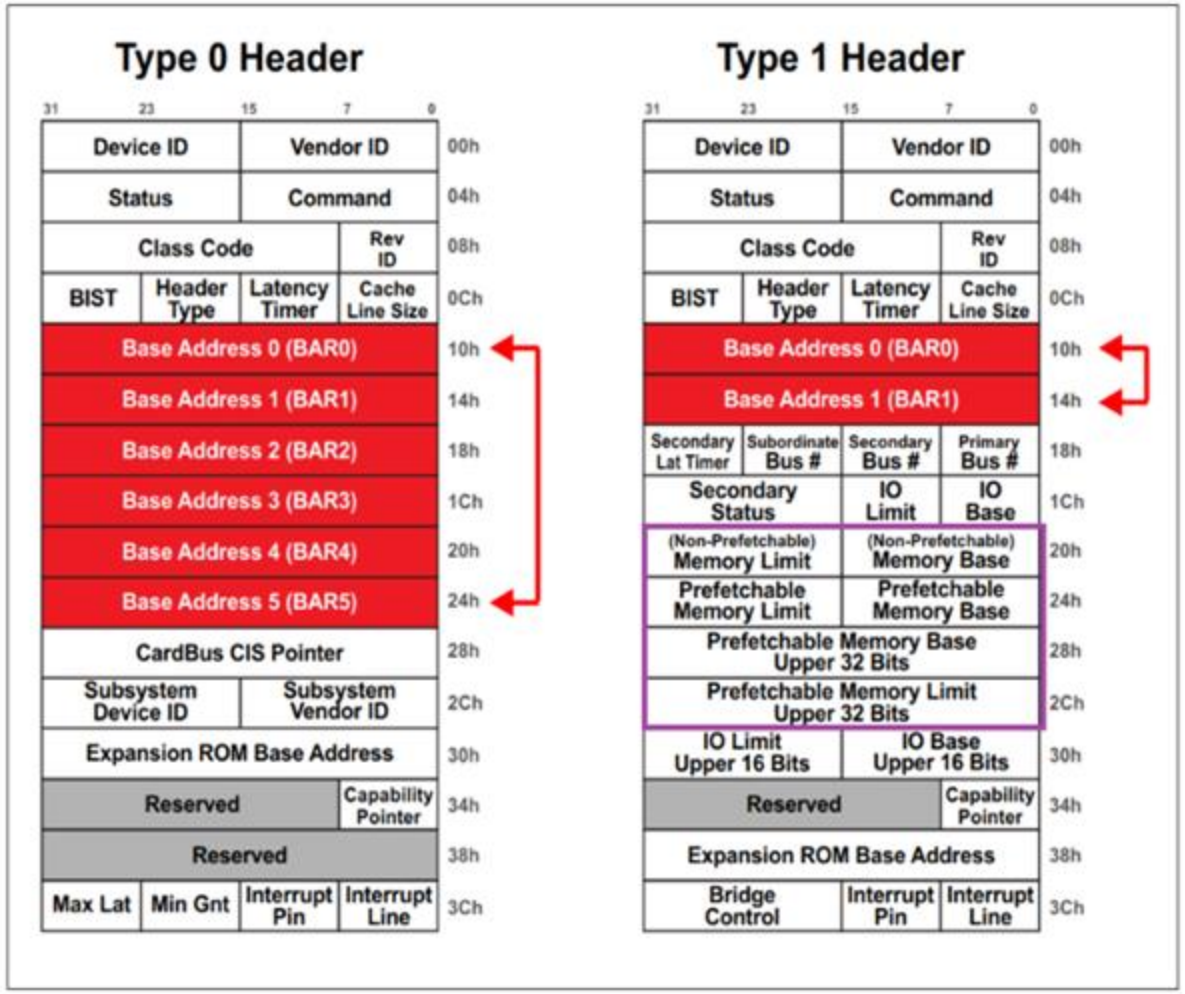
- Depth First Search
- BDF ID: 8 bits Bus#, 5 bits Device#, 3 bits Function#
  - Primary Bus Number
    - Contains the bus number to which the upstream side of a bridge is connected.
  - Secondary Bus Number
    - Contains the bus number to which the downstream side of a bridge is connected.
  - Subordinate Bus Number
    - Contains the deepest bus number on the downstream side of a bridge.
- Gen6 introduced as extension to the BDF



SNIA DSN A SNIA Community

Base Address Assignment

- Config Space Header
  - Endpoint: Type 0
  - Bridge: Type 1
- Host assigns the memory mapped address for EP and Bridge owned.
- Host updates the Prefetchable/Non-Prefetchable memory BASE/LIMIT of the bridge to cover the range of BAR address for all the downstream devices below it
- Used for Memory TLP address routing.



SNIA DSN A SNIA Community

Advanced Error Reporting

Error Reporting Messages

- Error signaling messages generated by the device detecting the error
- Three types of error message;
  - ERR_COR
  - ERR_NONFATAL
  - ERR_FATAL
- Error messages always routed to the Root Complex
- Translation to platform specific events by the Root Complex

Correctable Errors

- Hardware is responsible for recovery and no software or application involved
- May degrade system performance by going down to narrower width or lower speed
- No loss of information
  - Same data is re-sent by hardware
- Logging of these errors is possible
  - Error details are logged in AER (Advanced Error Reporting)
  - ERR_CORR message sent to Root Complex
- Typical example is a TLP retry event or recovery on the link
  - TLP retry event – Initiator receives a NAK from connected device
  - When the link detects too many errors, a recovery is initiated

Uncorrectable Errors

- Errors which causes data loss
  - May require device reset or re-enumeration by the system
  - Error handling specific to platform
  - Similar to SERR# in PCI/PCI-X

- Error severity can be specified as Fatal or Non-Fatal
  - Fatal error causes ERR_FATAL message sent to RC
  - Non-Fatal error causes ERR_NONFATAL message sent to RC

- A typical example is Completion Time Out (CTO)
  - A non-posted TLP initiator does not receive the corresponding completion
  - Can cause a system crash if the initiator is the CPU

Downstream Port Containment

Downstream Port Containment (DPC)

- DPC is the automatic disabling of the Link at a DSP (Downstream Port) following an uncorrectable error
- Convert an uncorrectable error detected at a DSP or an uncorrectable error message received at DSP to a system interrupt
- Prevents the potential spread of data corruption from propagating to other parts in the system
- Improves PCIe error containment and allows software to recover from async removal events
  - A surprise hot removal action can cause uncorrectable errors on the link

SNIA | DSN
A SNIA Community

Hot-Plug

Standard Hot-Plug

- Orderly Removal / Addition
- Downstream ports need special hardware hot-plug controller to support it
- Physical Elements listed for supporting hot-plug. Defined in given form factors and may be optional.

Standard Hot-Plug

Table 6-7 Elements of Hot-Plug §

Element	Purpose
Indicators	Show the power and attention state of the slot
Manually-operated Retention Latch (MRL)	Holds adapter in place
MRL Sensor	Allows the Port and system software to detect the MRL being opened
Electromechanical Interlock	Prevents removal of adapter from slot
Attention Button	Allows user to request hot-plug operations
Software User Interface	Allows user to request hot-plug operations
Slot Numbering	Provides visual identification of slots
Power Controller	Software-controlled electronic component or components that control power to a slot or adapter and monitor that power for fault conditions
Out-of-band Presence Detect	Method of determining physical presence of an adapter in a slot that does not rely on the Physical Layer

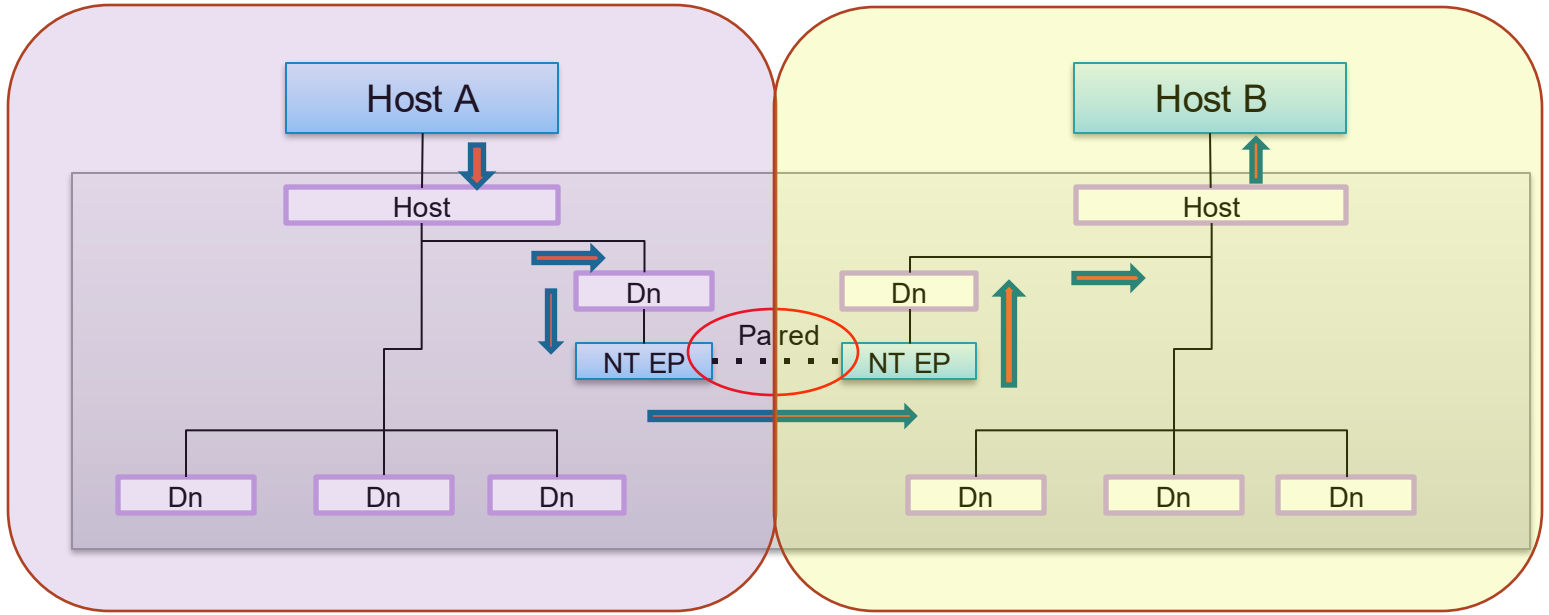
Async Removal

- Removing an add-in card without prior notification, previously called surprise hot-plug, is now called Async Removal.
- Hotplug Surprise Support bit
  - In downstream port slot capability register
  - If this bit is set in this device, surprise down error on downstream port will be prevented during device hot removing.
- PCIe 5.0 spec recommends handling Async Removal as a Downstream Port Containment (DPC) event
  - Instead of setting Hotplug Surprise Support bit

Non-Transparent Bridge

Non-Transparent Bridging (NTB)

- NTB implementation is vendor specific and is not defined in the spec
- Hosts in two separate PCIe hierarchies can talk with each other
  - Standard PCIe hierarchies are independent
  - Accessing remote host memory via NT EP BAR address mapping



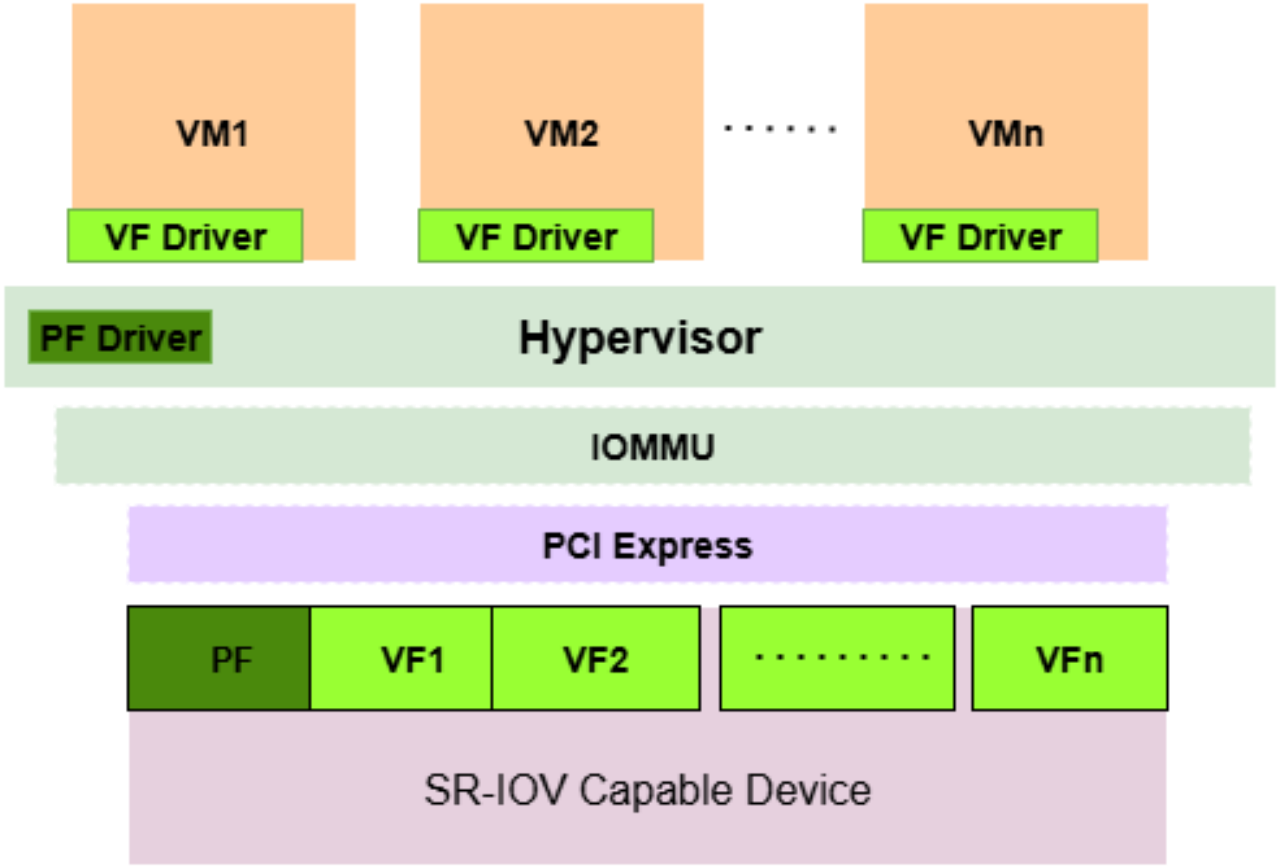
SNIA DSN A SNIA Community

SNIA | DSN
A SNIA Community

SR-IOV

Single Root IO Virtualization

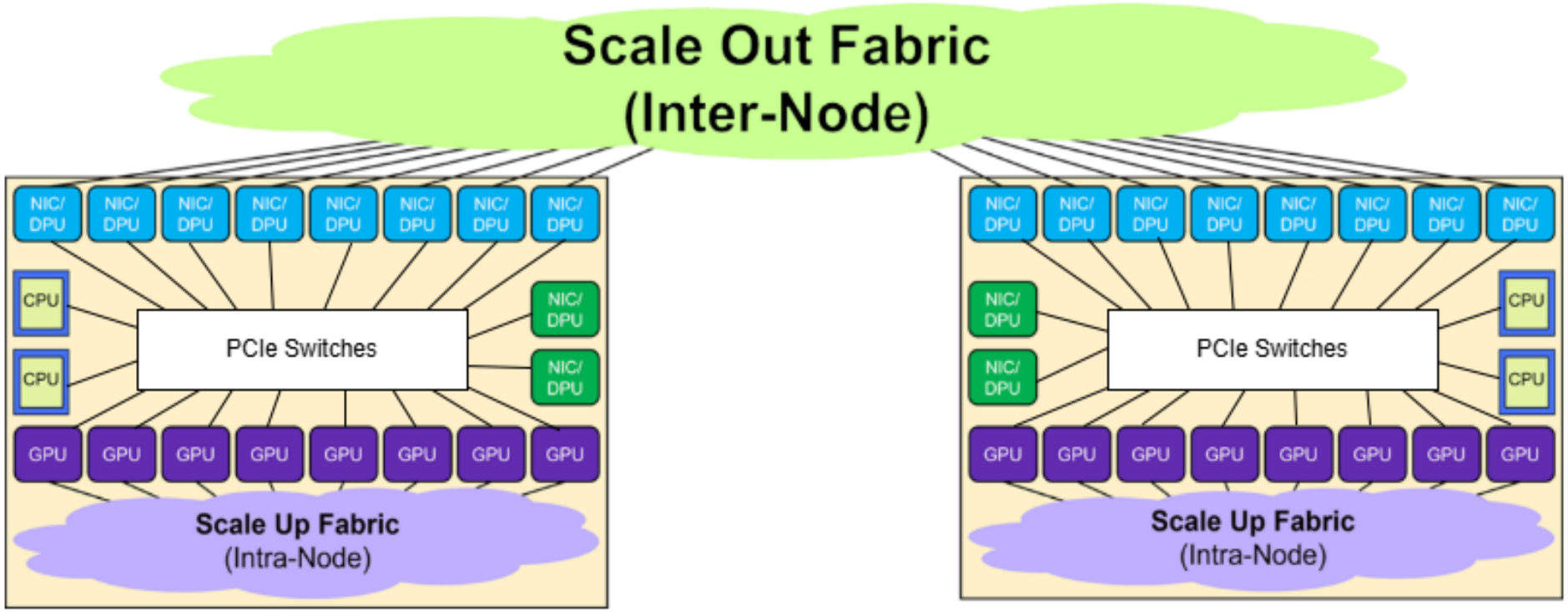
- SR-IOV devices contains physical functions(PF) and virtual functions(VF) associated with single Root Complex.
- Hypervisor configures the VFs for VMs via full-functional PF.
- Once configured by the hypervisor, a VM can directly access its assigned VF through PCIe.
- Performance improvement for VM due to direct VF access.



PCIe Switch in AI Application

PCIe Switch in AI Application

- Enabling High-Speed, Low-Latency Interconnection between xPUs, NICs, and so on.
- Peer-to-Peer Communication (P2P)
- Flexible Topologies for AI Clusters, with Fabric Link support
- Resource Pooling / Dynamic Assignment



SNIA DSN A SNIA Community

SNIA | DSN
A SNIA Community

PCIe Gen6 - What's New

PAM4 vs NRZ

- NRZ (Non-Return-to-Zero)
  - PCIe Gen1 – Gen5 from 2.5GT/s to 32GT/s
  - 2 states (1 bit) per transition
  - BER: 10^-12
- PAM4 (Amplitude Modulation 4-level)
  - PCIe Gen6 64GT/s
  - 4 states (2 bits) per transition
  - BER: 10^-6

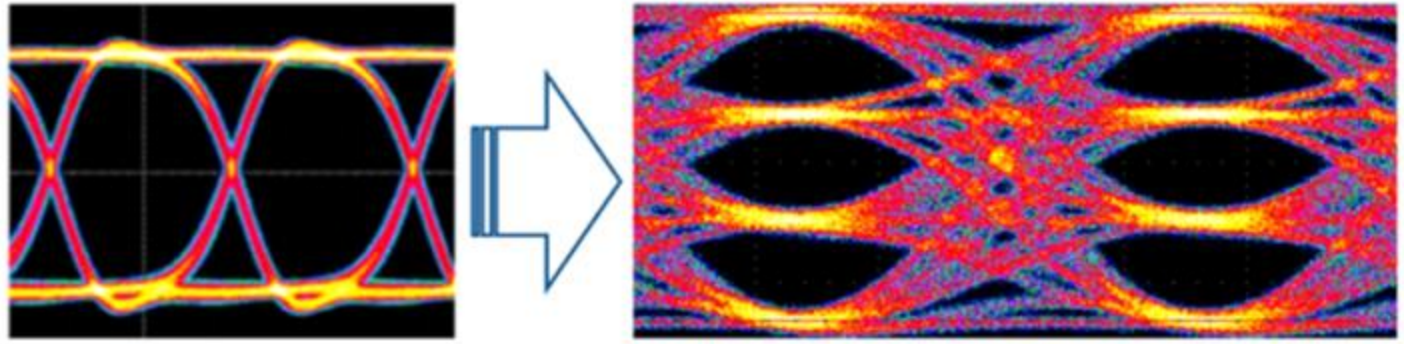
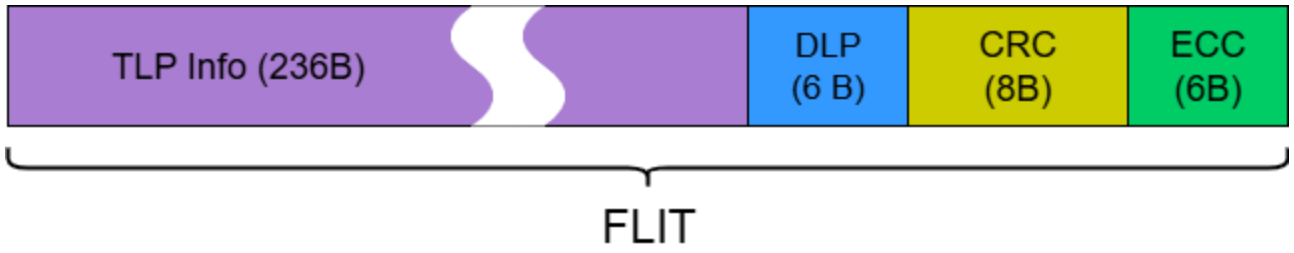


Image 1: NRZ Eye vs PAM4 Eyes

FLIT Mode (FM)

- PCIe Gen6 has added a new mode of operation called Flit Mode
  - If FM is enabled, it is enabled for all speeds
    - Once enabled, it remains enabled until a Reset or link down
  - FM is required when operating at Gen6 speed
  - Flit mode can be disabled by Link Control 3 register, but it limits the link speed to Gen5
- In FM, the basic unit of transfer is no longer packets, but is a fixed-sized (256B) Flit.
  - One or more TLPs can be held within a Flit, and a single TLP may span across multiple Flits (based on Max Payload Size and start location within the Flit)
  - The ACK/NAK protocol and retry mechanisms are on a Flit level and not a TLP level



Replay Overview in Flit Mode

- Only the TLPs in Flit Data is replayed, not the DLP port
- PCIe Gen6 introduces a new replay scheme, called Selective Replay
  - Old replay scheme is called Standard Replay
- Standard Replay
  - Receiver Nak's a Flit with a particular sequence number and the transmitter replays all Flits after the sequence number
- Selective Replay
  - Receiver Nak's a Flit with a particular sequence number and the transmitter only replays that particular Flit
  - Receiver needs to keep following valid Flits in its receive buffer before it receives the replay Flit for keeping the ordering rule

Shared Flow Control Mechanism

- Prior to Gen6, Dedicated Flow Control (FC) resources are implemented for each Virtual Channel(VC). Sometimes the utilization efficiency is not consistent.
- Shared FC can reduce the cost of implementing multiple VCs because the set of FC resources used by all implemented VCs can now be shared.
  - A larger common pool of shared resources
  - The Transmitter can still indicate the use of dedicated credit for a specific TLP
  - Shared FC is used only in FLIT mode

L0p

- New power saving state in PCIe Gen6 – Allow two connected devices to shut down some lanes in the link such as downgrade a x16 link to x8 by shutting off the upper 8 lanes without going through Recovery
- Supporting L0p is negotiated between two devices in Configuration phase during link training
- Transit to L0p is handshake via Link Management DLLP between the two devices
- Allow up-size request or down-size request.
- Only support in FLIT mode.

Q&A

After this Webinar

- Please rate this webinar and provide us with your feedback
- This webinar and a copy of the slides are available at the SNIA Educational Library snia.org/educational-library
- A Q&A from this webinar, including answers to questions we couldn't get to today, will be posted on our blog at sniablog.org
- Follow us on X/Twitter @SNIA

SNIA | DSN
A SNIA Community

Thank You