PCIe Protocol

Presented by: Mitesh Khadgi

Date: 13/05/2025

Agenda

- History of PCI Express (PCIe)
- PCI and PCIe Generation-wise enhancements
- PCIe slots
  - Typical Example Cards for PCIe x1 slots
  - Typical Example Cards for PCIe x4 slots
  - Typical Example Cards for PCIe x8 slots
  - Typical Example Cards for PCIe x16 slots
- Pin Breakdown for PCIe x1
- Key Features of PCIe
- Additional Features
- Layers in PCIe
  - Transaction Layer
  - Data Link Layer
  - Physical Layer
- PCIe Version Improvements Table
- PAM4 vs NRZ (Used in PCIe 1.0-5.0)
- NRZ vs PAM4 Encoding
- FLIT Mode (Flow Control Unit)
- FEC (Forward Error Correction)
- Encoding types used in PCIe
- Combined Use in PCIe 6.0/7.0
- Performance and Bandwidth
- Configuration Space
- PCIe Topology
  (Root complex position in PCIe topology)
  - Root Complex
  - End Point (EP)
  - PCIe Bridge
  - LTSSM (Link Training and Status State Machine)
  - LTSSM States
  - PCIe Switch
- PCIe Protocol Stack
- Data Transmission in PCIe
- Every 3 Years Projection
- Factors to Consider when choosing PCIe Lane Counts
- PCIe size comparison
- Examples
  - Intel Processor and Chipset PCIe Lane Configurations
  - AMD Processor and Chipset PCIe Lane Configurations
- What will Replace PCIE?
- Q & A

History of PCI Express (PCIe)

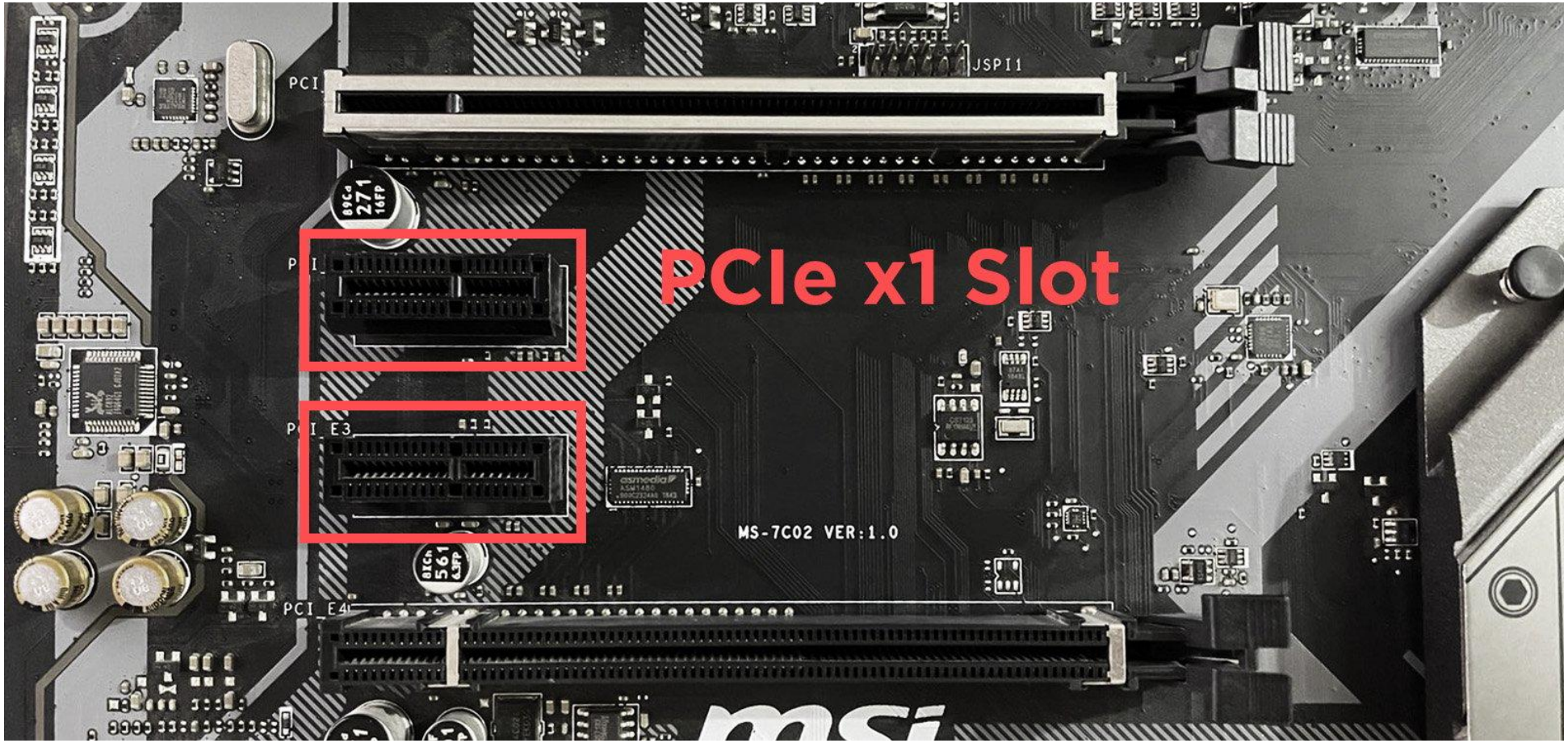
- Before PCIe, systems used:
  - ISA (Industry Standard Architecture) – Introduced in the 1980s, slow and parallel.
  - PCI (1992) – PCI (Peripheral Component Interconnect), Parallel bus with shared bandwidth; the dominant interface in the 1990s.
  - AGP (1997) – Advanced/Accelerated Graphics Port, used for video cards, but limited to graphics.
- Understand Architectural Trade-offs
  - Each PCIe generation introduced new encoding, signaling, and protocol features to overcome physical and practical limitations:
  - PCI (parallel bus) -> had shared bandwidth, signal integrity issues, and poor scalability
  - PCIe (serial point-to-point) -> solved these by introducing dedicated lanes and high-speed serial links
- PCI Express (PCIe) is a high-speed serial computer expansion bus standard designed to replace older bus standards like PCI (Peripheral Component Interconnect), PCI-X, and AGP. It was developed by the PCI-SIG (PCI Special Interest Group), a consortium of tech companies including Intel, IBM, and Dell.
- Knowing this explains:
  - Why PCIe adopted packetized communication?
  - Why transitions like NRZ -> PAM4 or 8b/10b -> 128b/130b occurred?
  - PCI introduced INTA#, INTB#, INTC#, INTD# collectively referred to as INTx

PCI and PCIe Generation-wise enhancements

Generation	Year	Encoding	Signaling	Data Rate (x1)	Notable Advance
PCI	1992	N/A	Parallel, 33–66 MHz	133 MB/s (32-bit @ 33 MHz)	Industry-standard desktop interface
PCI-X 1.0	1998	N/A	Parallel, up to 133 MHz	~1.06 GB/s (64-bit)	Higher bandwidth for servers
PCI-X 2.0	2003	N/A	Parallel, up to 533 MHz	Up to ~4.3 GB/s	Doubled clock rates, ECC support
PCIe 1.0	2003	8b/10b	**NRZ (Non-Return-to-Zero)**	250 MB/s	Switched to serial, point-to-point lanes
PCIe 2.0	2007	8b/10b	NRZ	500 MB/s	Doubled throughput
PCIe 3.0	2010	128b/130b	NRZ	~985 MB/s (1 GB/s)	More efficient encoding (less overhead)
PCIe 4.0	2017	128b/130b	NRZ	~1.97 GB/s (2 GB/s)	Higher signaling rate for GPUs/SSDs
PCIe 5.0	2019	128b/130b	NRZ	~3.94 GB/s (4 GB/s)	Supports AI/ML and fast storage needs
PCIe 6.0	2022	**FLIT (Flow Control Units)** over **PAM4**	**PAM4 (4-level Pulse Amplitude Modulation)**	~7.88 GB/s (8 GB/s)	Major shift to PAM4, low-latency FLIT mode
PCIe 7.0	2025 (exp.)	FLIT (PAM4)	PAM4	~15.75 GB/s (16 GB/s)	Doubles PCIe 6.0, targeted at HPC, AI, 800G+ networking

Typical Example Cards for PCIe x1 slots

Port/Hub Expansions
Sound Cards
Network Cards
Capture Cards

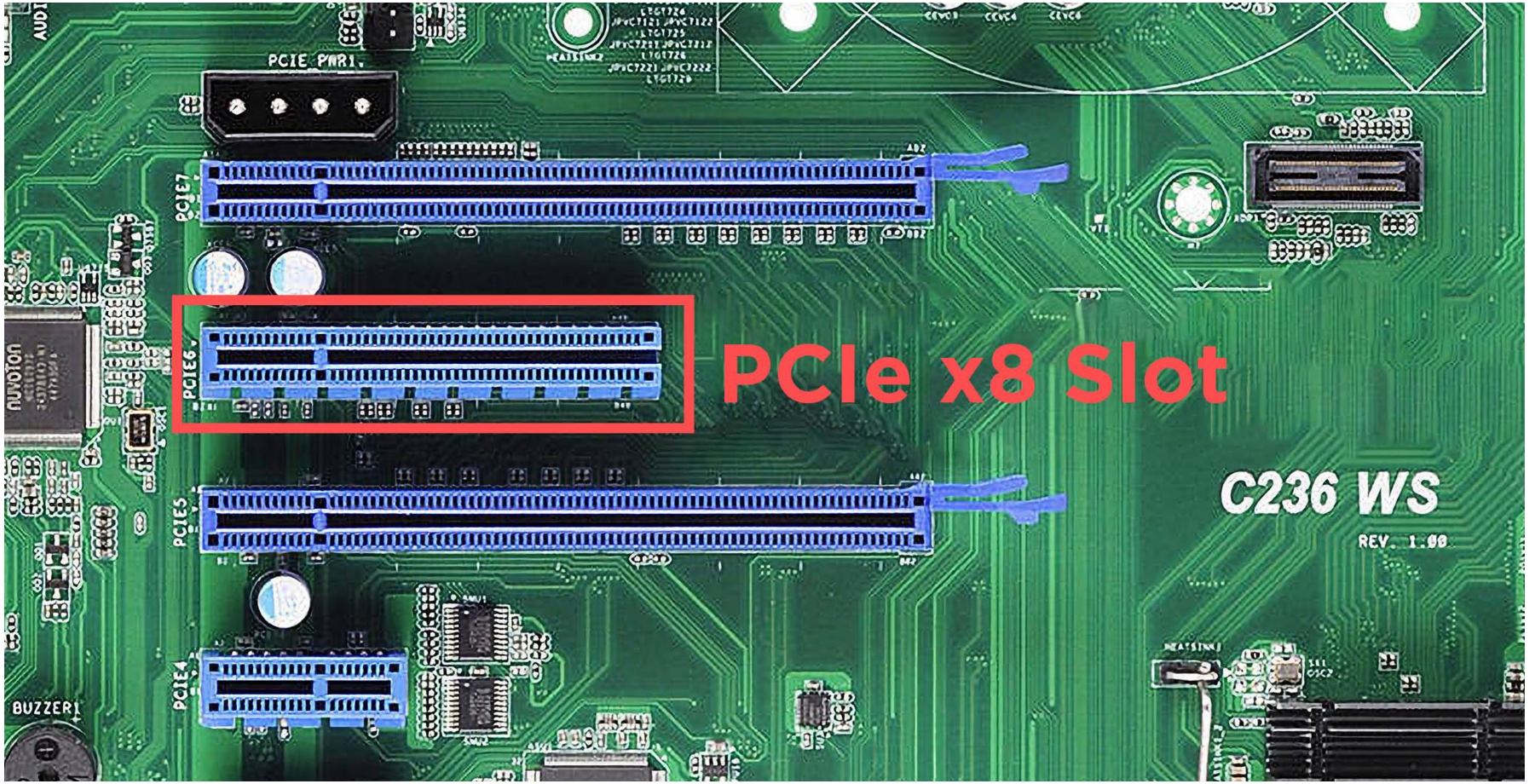


Typical Example Cards for PCIe x4 slots

Port/Hub Expansions
High-Bandwidth Network Cards
NAS Storage
RAID Controller Cards
Capture Cards
M.2 and NVMe Adapters

Typical Example Cards for PCIe x8 slots

Higher-Bandwidth Implementations of other Expansion Cards
Multi-Slot NVMe Adapters
Low-End Graphics Cards that are actually made for PCIe x8 slot lengths (something like an AMD RX 560)

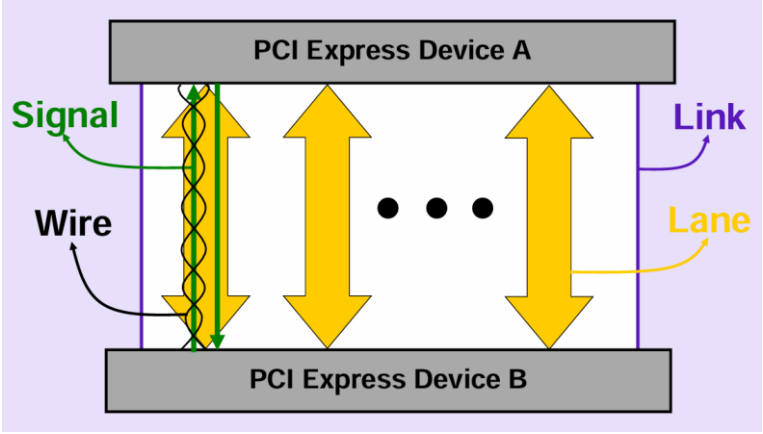


Typical Example Cards for PCIe x16 slots

Graphics Cards, in general
Enthusiast or Server-Grade Expansion Cards (Network and Storage most common)

Pin Breakdown for PCIe x1

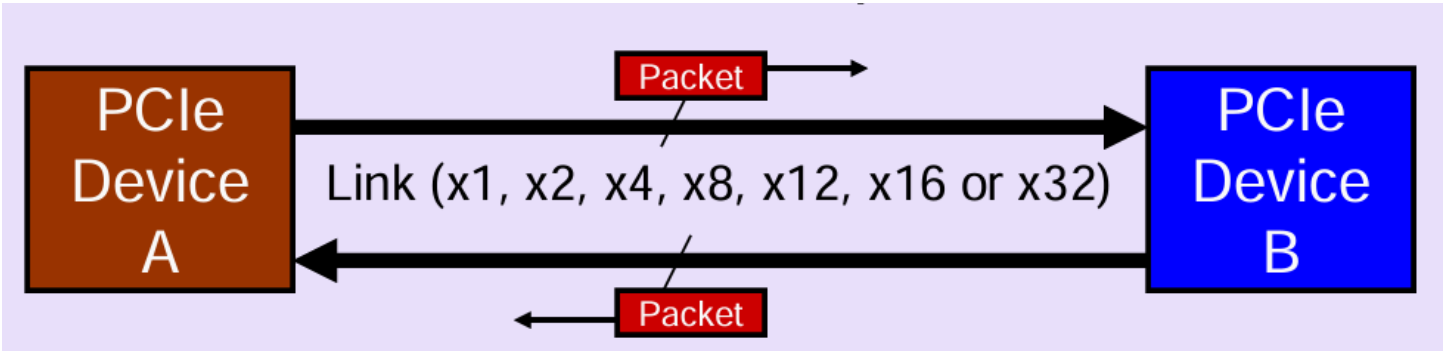
Pin Group	Typical Count	Description
Power (3.3V)	2	Device power supply
Ground (GND)	8–10	Signal reference and noise shielding
Transmit (TX±)	1 pair	Transmit differential signal
Receive (RX±)	1 pair	Receive differential signal
Control	4–6	PERST#, WAKE#, CLKREQ#, etc.
Reserved	Few	For future expansion



PCIe Terminology (Link and Lane)

Key Features of PCIe

- PCIe (Peripheral Component Interconnect Express) is a high-speed, scalable interface used for connecting various peripherals to the CPU.
- Key features include
  - **high bandwidth** (up to 128 GT/s in PCIe 7.0)
  - **low latency** for efficient data transfer
  - **scalable lanes** (x1, x2, x4, x8, x12, x16/x32)

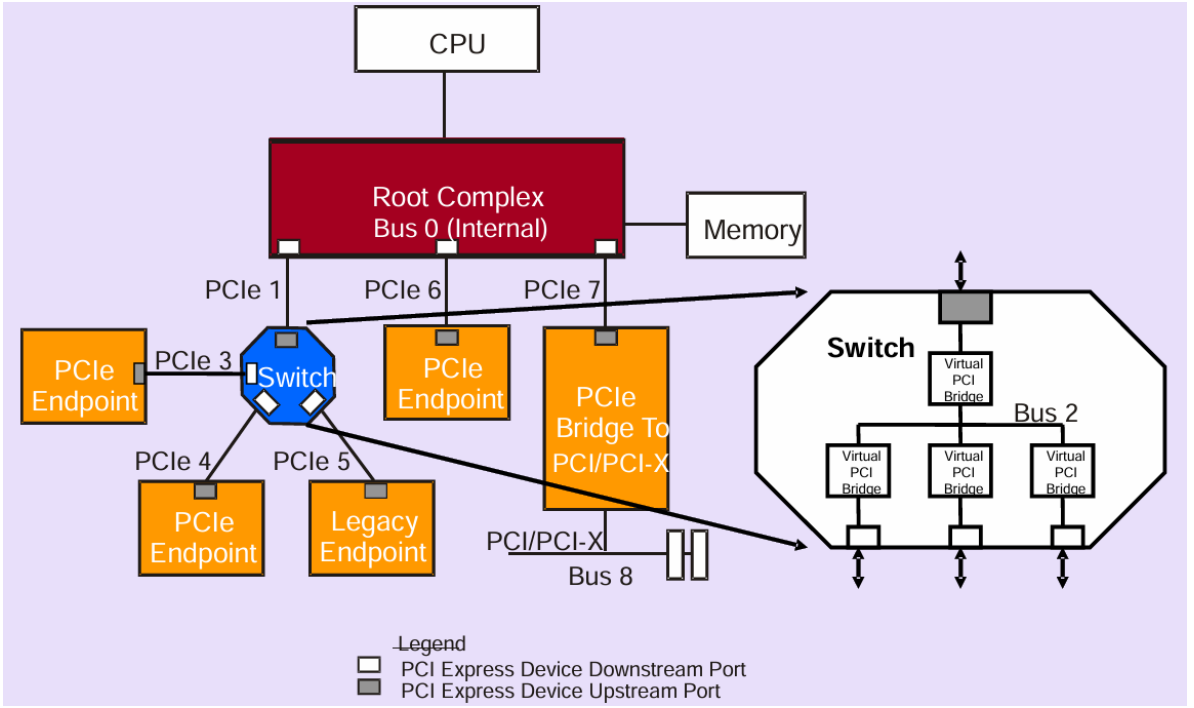


  - supports **backward compatibility**, **error correction** (FEC, CRC), and **power efficiency** (Active State Power Management).
  - enables **flexible device support** (e.g., GPUs, SSDs) and **dynamic link equalization** for signal optimization.
  - supports **hot-plug** and **virtualization** (SR-IOV), making it ideal for data-intensive tasks, AI workloads, and high-performance systems.
- NRZ (Non-Return-to-Zero): Traditional binary signaling (2 voltage levels).
- PAM4 (Pulse-Amplitude Modulation 4): Uses 4 voltage levels to encode 2 bits per symbol—doubles data rate without increasing frequency.
- FLIT: Packetized data transmission used in PCIe 6.0+, improves latency and error handling.

Additional Features

- Data Integrity and Error Handling
  - Link-level “LCRC”
  - Link-level “ACK/NAK”
  - End-to-end “ECRC”
- Credit-based Flow Control
  - No retry as in PCI
- MSI/MSI-X style interrupt handling
  - Also supports legacy PCI interrupt handling in-band
- Advanced power management
  - Active State PM
  - PCI compatible PM
- PCI Express system will boot “PCI” OS
- PCI Express supports “PCI” device drivers

PCIe Topology

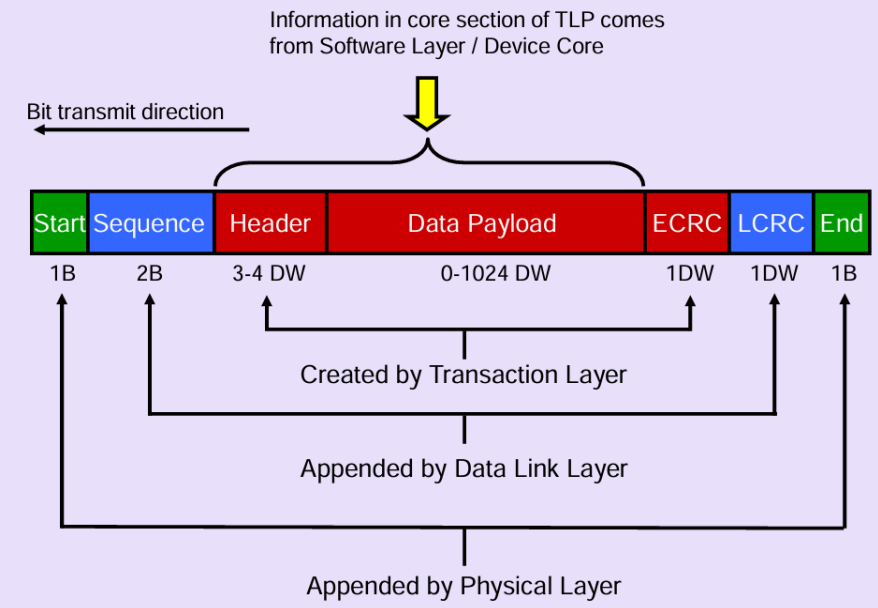


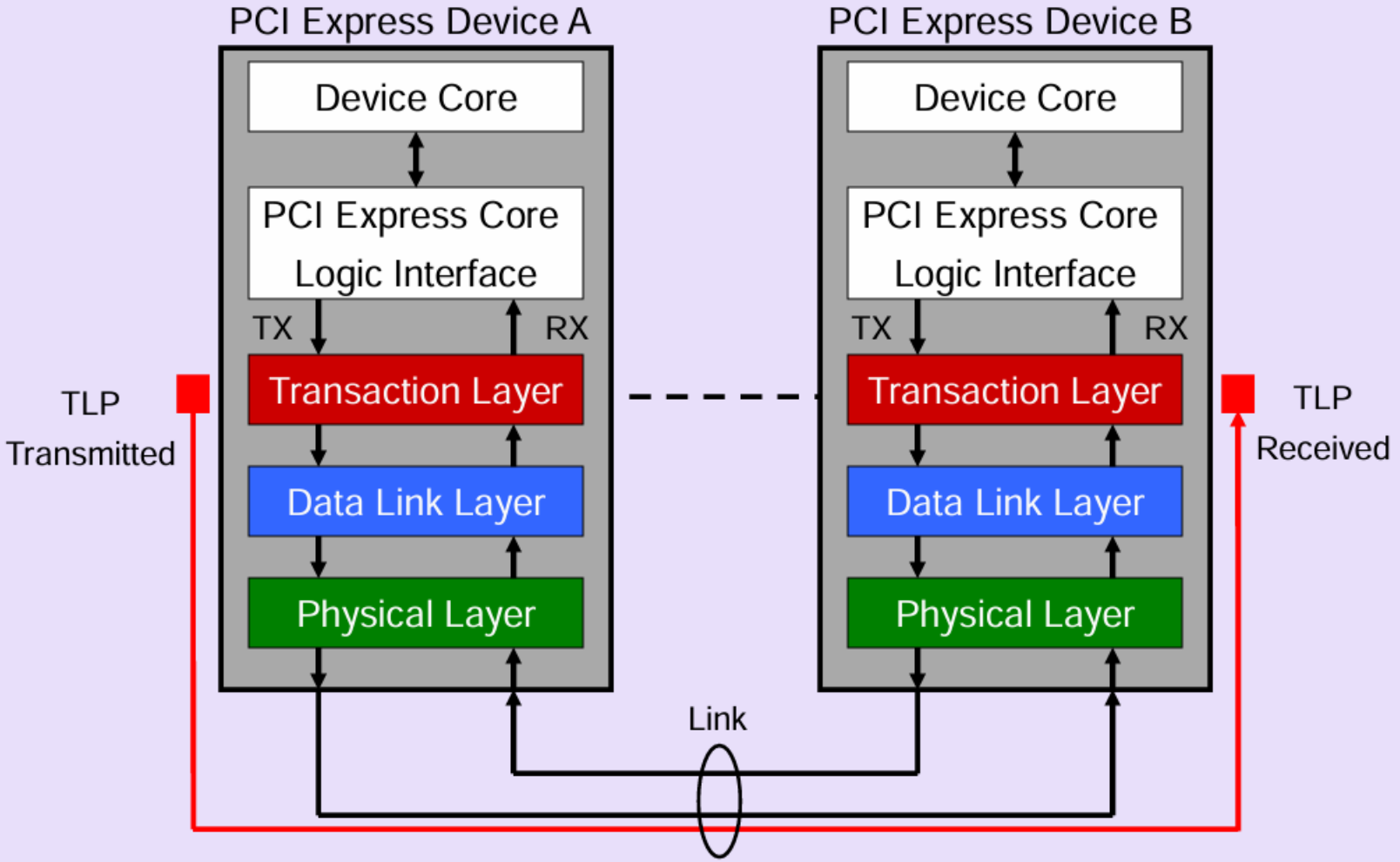
Layers in PCIe

- PCI Express (PCIe) is designed using a layered architecture to ensure flexibility, scalability, and high-speed data transfer. It follows a protocol stack model, similar in concept to the OSI model in networking, and consists of the following three main layers:
  - Transaction Layer
  - Data Link Layer
  - Physical Layer

Transaction Layer

Section	Size	Notes
Header	3 DWs (12 bytes)	Includes type, address, etc.
Payload	Variable	Contains write data
ECRC (Opt.)	4 bytes	Optional checksum

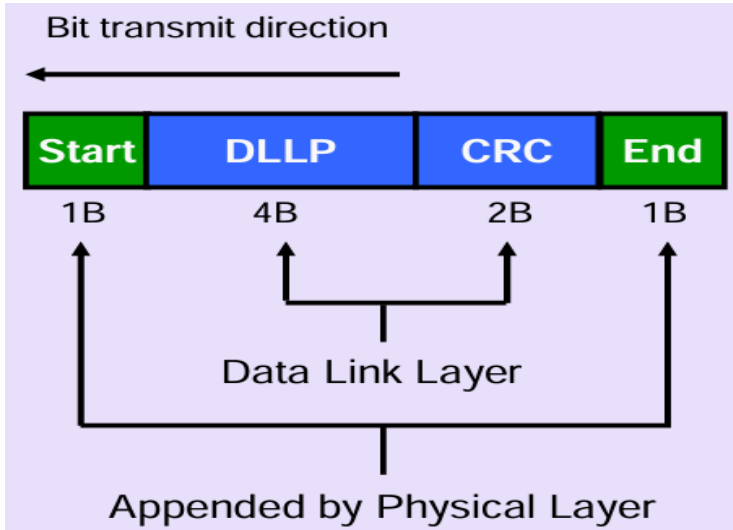


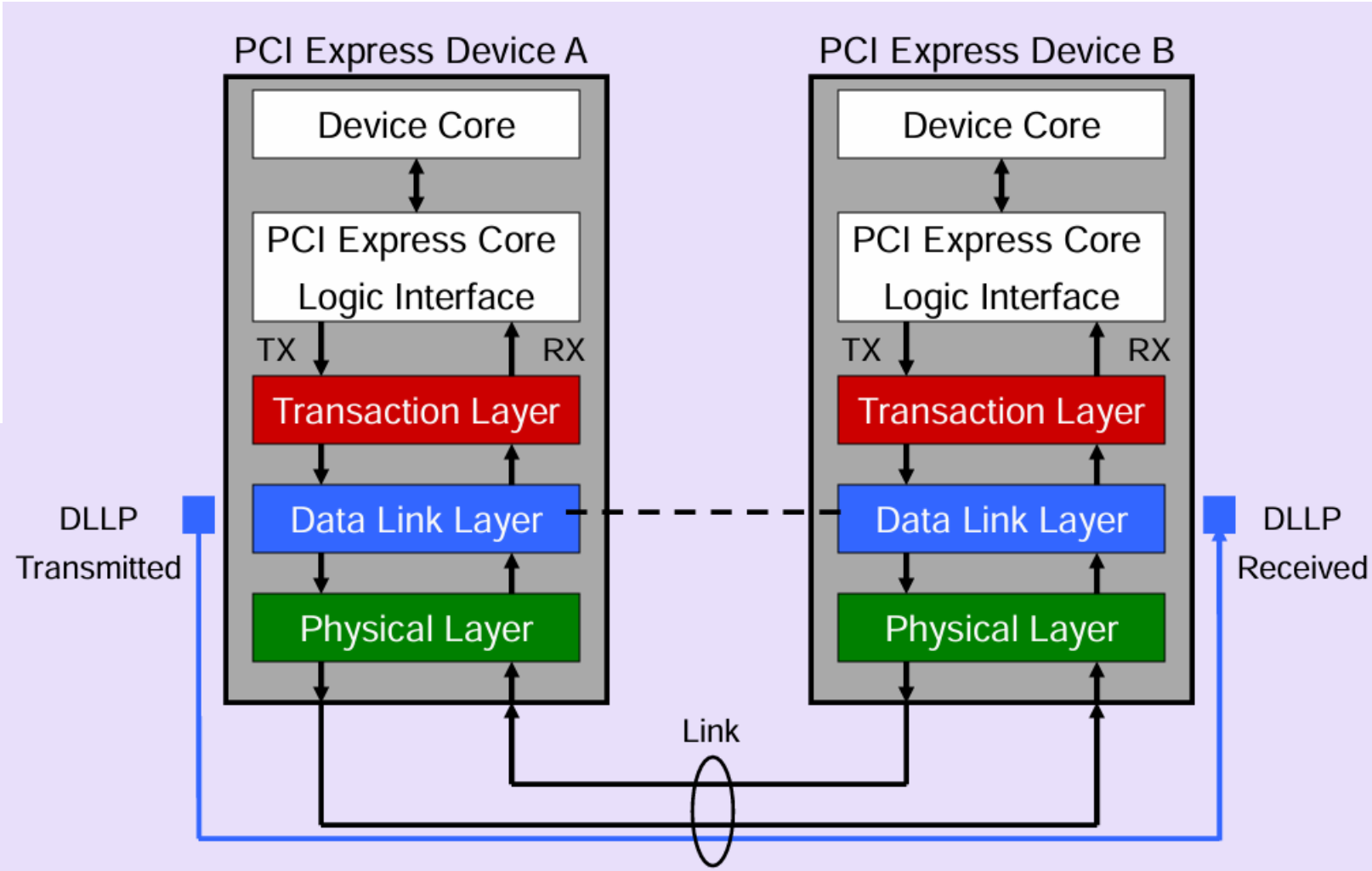


Data Linik Layer

A DLLP is always exactly **6 bytes (48 bits)** long and consists of:

Field	Size (bits)	Description
Type	8 bits	Identifies the type of DLLP (ACK, NAK, Flow Control, PM, etc.).
Data	32 bits	Payload specific to the DLLP type (e.g., credit info, ACK ID).
CRC (Checksum)	8 bits	Protects the DLLP from transmission errors (CRC-8).

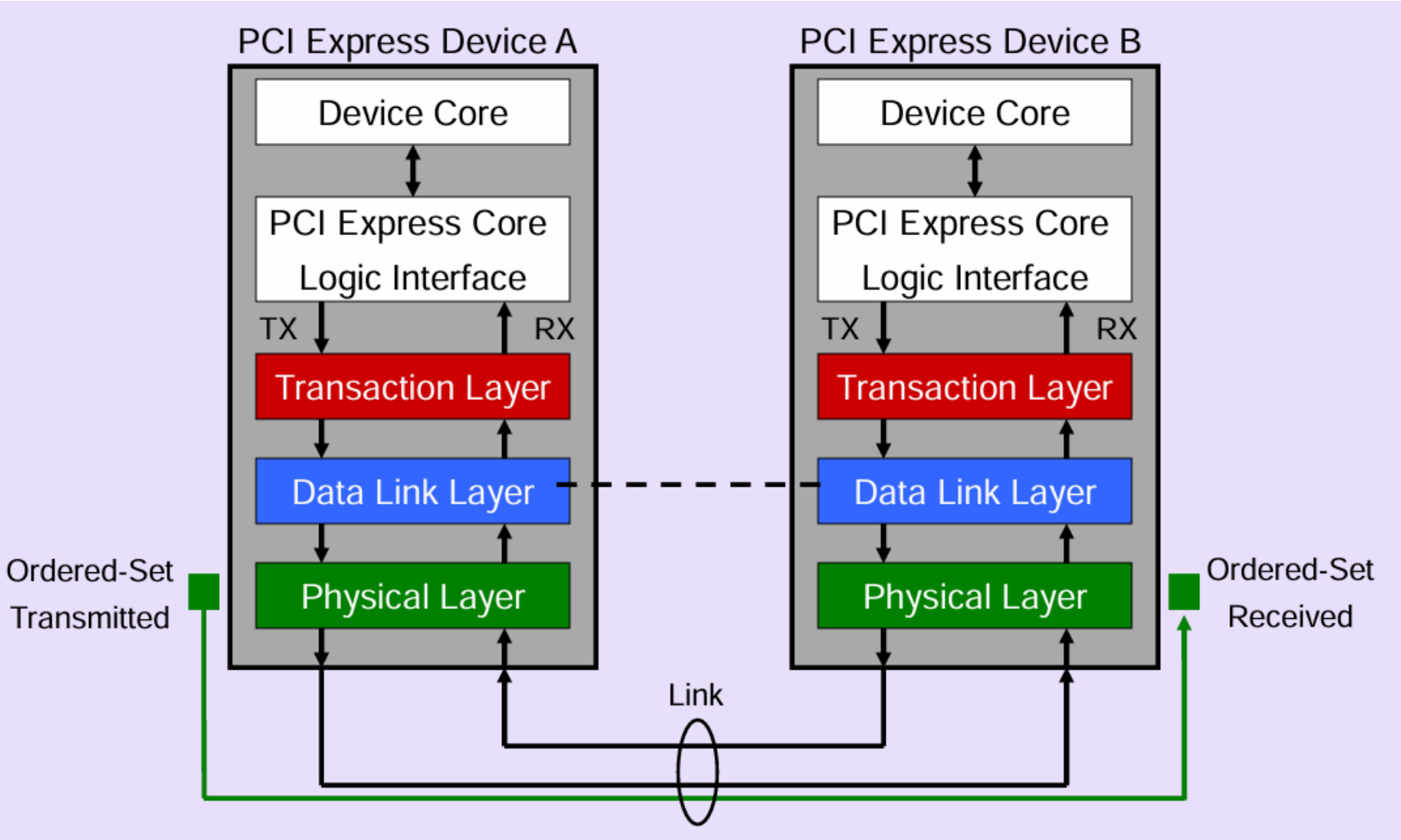




Physical Layer

Evolution of Physical Layer Signaling:

PCIe Version	Encoding	Signaling	Bandwidth per Lane (1 direction)
1.0 / 2.0	8b/10b	NRZ	250 / 500 MB/s
3.0 / 4.0 / 5.0	128b/130b	NRZ	1 / 2 / 4 GB/s
6.0	128b/130b	PAM4	8 GB/s
7.0 (planned)	TBD	PAM4 or better	16 GB/s



PCIe Version Improvements Table

Feature	PCIe 1.0	PCIe 2.0	PCIe 3.0	PCIe 4.0	PCIe 5.0	PCIe 6.0	PCIe 7.0
Release Year	2003	2007	2010	2017	2019	2022	2025 (planned)
Max Raw Line Rate (GT/s)	2.5	5.0	8.0	16.0	32.0	64.0	128.0
Max x1 Bandwidth (GB/s)	0.25	0.5	1.0	2.0	4.0	~7.45	~15.4
Max x16 Bandwidth (GB/s)	4.0	8.0	~15.75	~31.5	~63.0	~119–128	~245–256
Duplex x16 Bandwidth (GB/s)	8.0	16.0	~31.5	~63.0	~126	~238–256	~490–512
Signaling	NRZ	NRZ	NRZ	NRZ	NRZ	PAM4	PAM4
Encoding	8b/10b	8b/10b	128b/130b	128b/130b	128b/130b	FLIT + FEC	FLIT + FEC
Efficiency	80%	80%	~98.5%	~98.5%	~98.5%	~93-95% (with FEC)	~93-95% (with FEC)
Error Correction (FEC)	None	None	None	None	None	FEC + CRC + Replay	Improved FEC
Use of FLITs	No	No	No	No	No	256B fixed-length	Optimized
Power Efficiency	Moderate	Moderate	Improved	Higher draw	High draw	Target: ~1.2 pJ/bit	Target: ≤1 pJ/bit
Compatibility	✅	✅	✅	✅	✅	via fallback	via fallback
Backward Compatibility	N/A	✅	✅	✅	✅	✅	✅
Major Features Added	Basic I/O	Higher speeds	128b/130b encoding	Double speeds	Double speeds	PAM4, FLIT, FEC	128 GT/s, Memory Coherence (CXL)

PAM4 vs NRZ (Used in PCIe 1.0–5.0)

Feature	NRZ	PAM4
Bits per symbol	1	2
Voltage levels	2 (High/Low)	4 (00, 01, 10, 11)
Signal-to-noise ratio	Higher	Lower
Complexity	Lower	Higher (requires FEC & equalization)
Used in PCIe versions	1.0 to 5.0	6.0 and 7.0

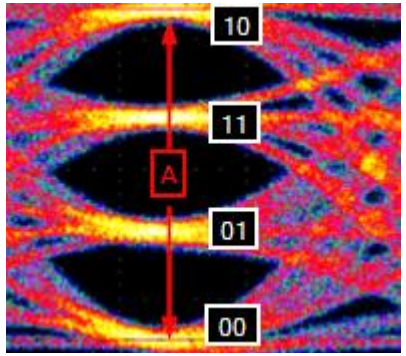
PAM4 is a multi-level signaling technique that uses 4 distinct voltage levels to represent 2 bits per symbol (00, 01, 10, 11). This effectively doubles the data rate without doubling the clock frequency or needing more lanes. By contrast, previous PCIe generations used NRZ (Non-Return-to-Zero), which only used 2 voltage levels (1 bit per symbol).

NRZ vs PAM4 Encoding

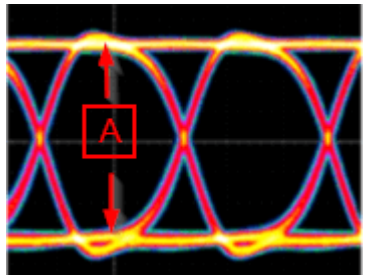
Pulse Amplitude Modulation 4-level (PAM4) is a multilevel signal modulation format used to transmit signal. Each signal level can represent 2 bits of logic information.

For the PAM 4 eye, Image 3, we can see the three eyes formed using four voltage levels (00, 01, 11, 10). The eye height is also important here, and A covers the height of all three eyes. Larger eye openings equal a better quality signal.

Diving a little bit further, we can also look at NRZ and PAM4 graphically. This is shown by PAM4 Eye Image where in PAM4 each clock cycle can be sent four different voltages of 0, 1, 2, or 3 in which correspond to 00, 01, 11, and 10 respectively. Over "Word 1" for NRZ vs PAM4 **you can see with PAM4 the signal of 101100100100 was transmitted in half the time.**

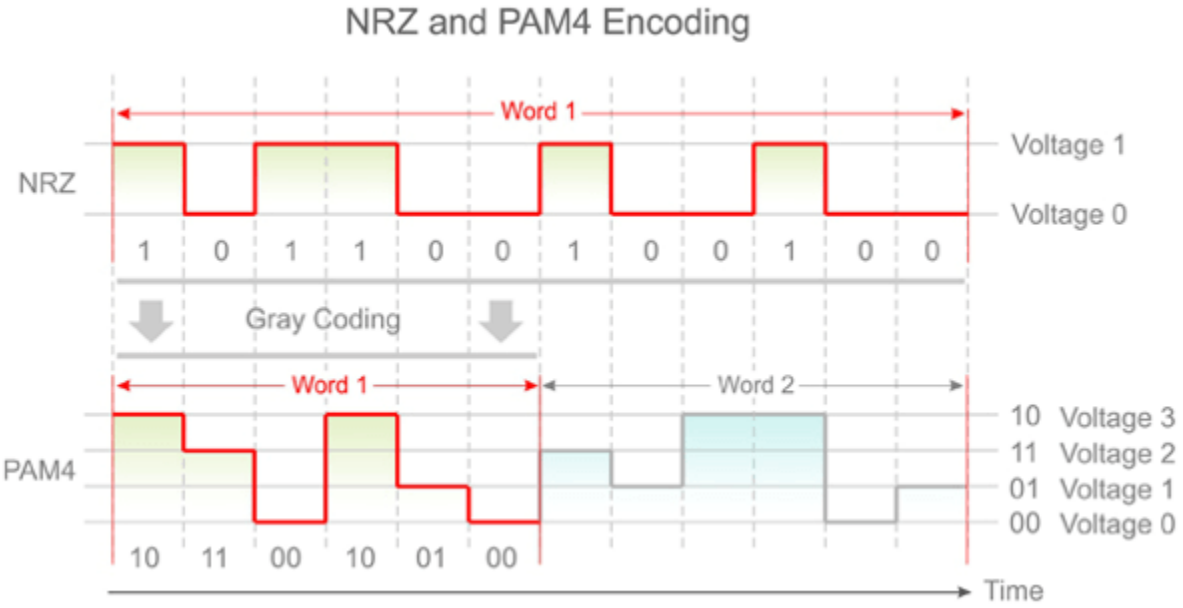


PAM4 Eye



NRZ Eye

Non-Return-to-Zero (NRZ), also called Pulse Amplitude Modulation 2-level, is a binary code using low and high signal levels to represent the 1/0 information of a digital logic signal. NRZ can only transmit 1 bit, i.e. a 0 or 1, of information per signal symbol period.
To better understand this idea we can look at an NRZ Eye and an NRZ Encoding graph, Image 4. For the NRZ Eye, we can look at NRZ Eye Image where the top horizontal line would represent a 1 and the bottom line would represent a 0. For the eye height "A" you want to see as large of an opening as you can get. The larger the eye equals a better quality signal.



FLIT Mode (Flow Control Unit)

- What is FLIT?
  - FLIT stands for Flow Control Unit, a method introduced in PCIe 6.0 to structure data transmission. Instead of a continuous byte stream like in PCIe 1.0–5.0, FLIT mode breaks data into fixed-size packets.
  - FLIT size in PCIe 6.0: 256 bytes
  - Each FLIT includes:
    - Header: control information
    - Payload: actual data
    - CRC: error detection checksum
- Why FLIT Mode?
  - Necessary to support PAM4, which is more prone to errors
  - Improves error recovery, latency predictability, and data integrity
  - Enables low-latency FEC decoding and packet retry mechanisms
- Additionally, FLIT encoding eliminates DLLP (Data Link Layer Packets) overhead and 128B/130B encoding from earlier PCIe specifications, leading to much greater TLP (Transaction Layer Packet) efficiency, particularly for large transactions.

FEC (Forward Error Correction)

- FEC is an error control technique where redundant bits are added to data to detect and correct errors without needing retransmission.
  - In PCIe 6.0/7.0, FEC is mandatory due to PAM4's higher error rate.
  - Works in real-time, inline with transmission.
- In this method, the sender sends a redundant error-correcting code along with the data frame. The receiver performs necessary checks based upon the additional redundant bits. If it finds that the data is free from errors, it executes error-correcting code that generates the actual frame. It then removes the redundant bits before passing the message to the upper layers.
  - Transmitter encodes the FLIT with redundant error-correcting data.
  - Receiver uses FEC logic to detect and correct errors.
  - CRC is used to verify if a FLIT has too many errors (uncorrectable).
  - If CRC fails -> data is dropped or retried via Replay mechanisms.
- Latency-Optimized FEC in PCIe is designed to correct small errors quickly, rather than focusing on heavy-duty correction.

Encoding types used in PCIe

- In PCIe (Peripheral Component Interconnect Express), **8b/10b** encoding is a method where each 8-bit data word is converted into a 10-bit symbol for transmission. This technique helps to ensure a DC-balanced signal, limit run lengths of consecutive 1s or 0s, and enable clock recovery at the receiver. While 8b/10b was used in early PCIe generations, newer versions like PCIe Gen3 and beyond have shifted to alternative encoding schemes like 128b/130b or 64b/66b, which offer higher efficiency and better bandwidth.
- In PCI Express (PCIe), **128b/130b** encoding is used for PCIe 3.0 through 5.0, and it involves adding two synchronization bits to each 128-bit data transfer, resulting in 130-bit blocks. This encoding scheme offers a high data rate with a small overhead (about 1.54%). The two sync bits are used for data block (01b) or ordered set (10b) identification.
- In PCIe 6.0, **PAM4 (Pulse Amplitude Modulation with 4 levels) encoding and FLIT (Flow Control Unit)** mode work together to achieve higher bandwidth. PAM4 uses four voltage levels to transmit data, effectively doubling the data rate compared to previous generations. FLIT mode, which uses 256-byte fixed-size packets, is crucial for error correction and data management with PAM4.
- In PCI Express (PCIe), the **64b/66b** encoding is primarily used in 10 Gbit/s Ethernet and is not the standard encoding for PCIe itself. PCIe uses 128b/130b encoding. The 64b/66b encoding adds two bits of overhead for every 64 bits of data, which is then used to transmit the 66-bit block. This overhead is primarily used to achieve DC balance and sufficient data transitions for clock recovery at the receiver.

Combined Use in PCIe 6.0/7.0

Component	Purpose	Benefit
PAM4	Doubles data rate per lane	More bandwidth with same clock
FLIT	Packetized structure	Predictability, easier error detection
FEC	Error correction	Makes PAM4 viable
CRC	Error detection	Ensures corrupted data is identified
Replay	(If CRC fails) resend FLIT	Maintains reliability under error

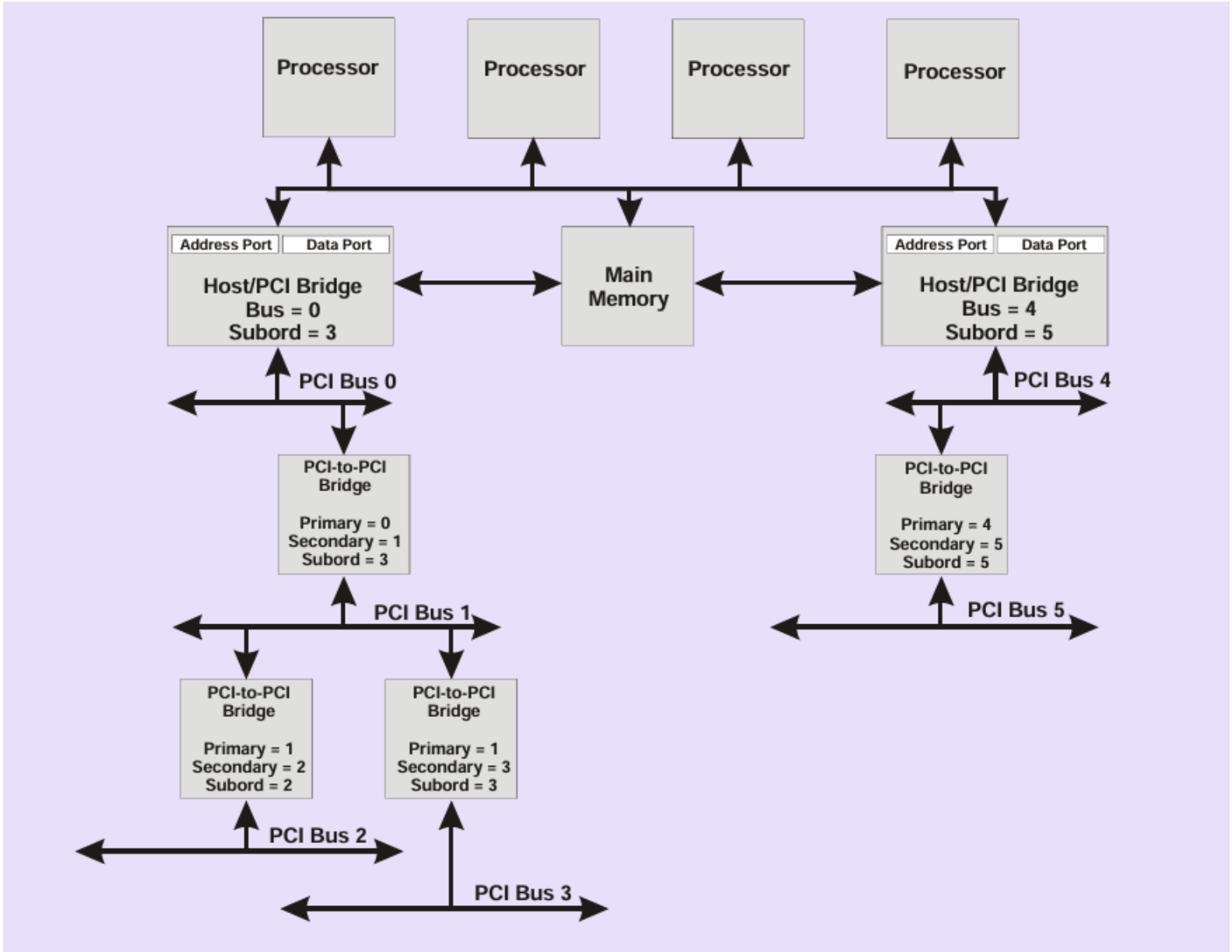
PCIe Version	Signaling	Encoding	FLIT	FEC	Reason
PCIe 5.0	NRZ	128b/130b	❌	❌	Signal quality manageable
PCIe 6.0+	PAM4	FLIT	✅	✅	Needed for PAM4 reliability

Performance and Bandwidth

Interconnect	Max Raw Bandwidth (x16 or equiv.)	Encoding	Data Rate per Lane	Latency Focus
PCIe 7.0	512 GB/s (bi-dir)	PAM4 + FEC	128 GT/s	Low
CXL 3.0	Same as PCIe 6.0/7.0	PAM4 + FEC	64/128 GT/s	Ultra-low
USB4	40 Gbps (v2: 80 Gbps bi-dir)	NRZ + FEC	20/40 Gbps	Medium
Ethernet	400 Gbps (and beyond)	PAM4 + strong FEC	100–400+ GT/s	Moderate-high

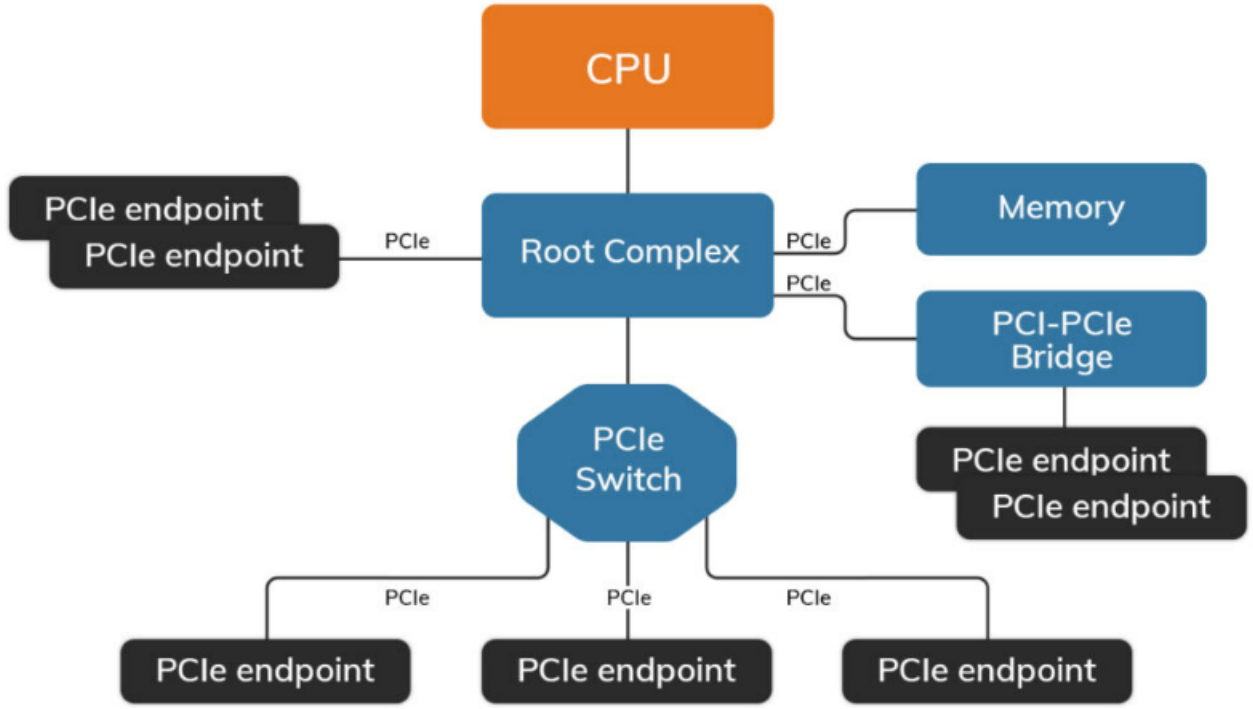
NOTE: PCIe 6.0/7.0 and CXL share the same base - FLIT + FEC - but CXL adds **coherent cache/memory protocol layers**.

Configuration Space



PCIe Topology
(Root complex position in PCIe topology)

- Root Complex
- End Point (EP)
- PCIe Bridge
- LTSSM
- PCIe Switch



Root Complex

- It is the “root” of the PCI inverted tree topology and acts on behalf of the CPU to communicate with the rest of the devices. It connects the system CPU to the PCIe topology. It initiates configuration requests as the requester. Figure shows the position of the root complex in PCIe topology.
- The main controller that connects the CPU and system memory to the PCIe fabric.
- Acts as the host in a PCIe system.
- Functions:
  - Initiates PCIe transactions (e.g., reads/writes)
  - Enumerates and configures PCIe devices during boot
  - Manages memory mapping and interrupt routing
- Example: In a PC, the CPU or chipset integrates the root complex to communicate with devices like GPUs or SSDs.

End Point (EP)

- A PCIe device that responds to the root complex.
- Cannot initiate PCIe bus enumeration or configuration.
- According to the PCIe specification in the PCIe topology there can be 256 buses, 32 devices on each bus, and 8 functions in each device. An endpoint can support a maximum of up to 8 functions and every function has its own separate configuration space. A function in an endpoint can be a separate individual entity where it has its functionality. PCIe-based NVM and PCIe-based SSDs are two end-point devices on a computer system.
- Examples:
  - GPUs
  - NVMe SSDs
  - Network Cards
  - Sound Cards
- NOTE: All data flows to/from End Points through the PCIe hierarchy, often via switches or bridges.

PCIe Bridge

- A device that connects different PCIe domains or buses, often to legacy systems.
- Also used to expand connectivity or segment PCIe topologies.
- PCIe PCI bridge: They are adapters that allow PCI devices to connect to PCIe slots in systems by doing protocol conversions from PCI specification to PCIe 1x specification. The master sends requests with the necessary parameters to the PCIe bridge. PCIe bridge converts requests into point-to-point transfers on the requested lane in the interface.
- Types:
  - PCIe-to-PCI bridge: Connects modern systems to older PCI devices.
  - PCIe-to-PCIe bridge: Allows hierarchy expansion — forms part of PCIe switches.

LTSSM
(Link Training and Status State Machine)

- LTSSM is an abbreviation of link training and the status state machine which manages PCIe devices. It is the main state machine control that detects, Polls, Configures, Recovers, Resets, and Disables the devices at the right times during operation.
- A finite state machine that manages the link bring-up and maintenance process in PCIe.
- NOTE:
  - LTSSM ensures the link is stable and operational before and during data transmission.
  - Crucial for hot-plug, power management, and error recovery.

LTSSM States

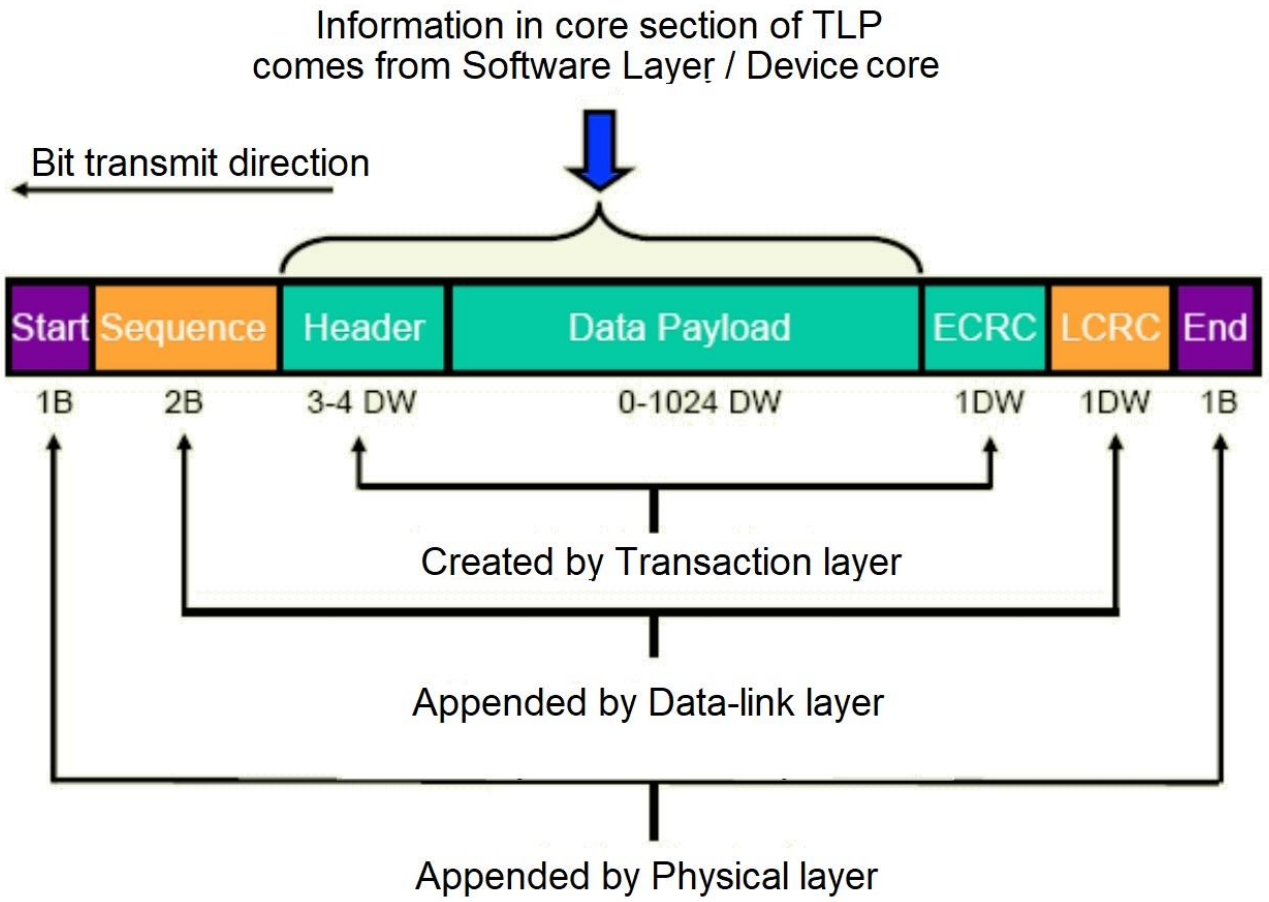
State	Purpose
Detect	Detect electrical presence of link partner
Polling	Exchange of basic signaling patterns
Configuration	Set link width, speed, capabilities
L0	Active data transmission state
L1/L2	Low-power link states
Recovery	Handle link errors or speed changes
Disabled	Link is down

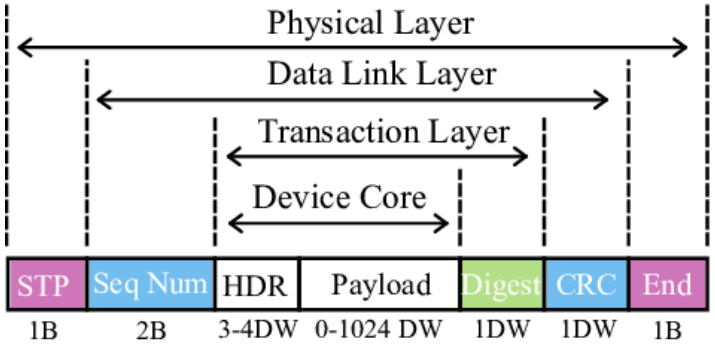
PCIe Switch

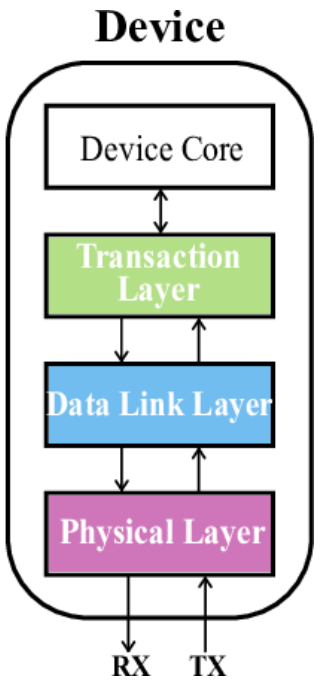
- A multi-port PCIe device that allows multiple End Points to connect to a single Root Complex.
- Functions like a network switch, but for PCIe traffic.
- Features:
  - Each port acts as a PCIe Bridge
  - Supports non-blocking switching, QoS, and traffic prioritization
  - Often used in:
    - Servers (multiple SSDs or GPUs)
    - AI accelerators
    - Expansion chassis
- Example: A PCIe switch might allow one CPU to control 8 NVMe drives via one x16 slot, splitting it into eight x2 links.

PCIe Protocol Stack

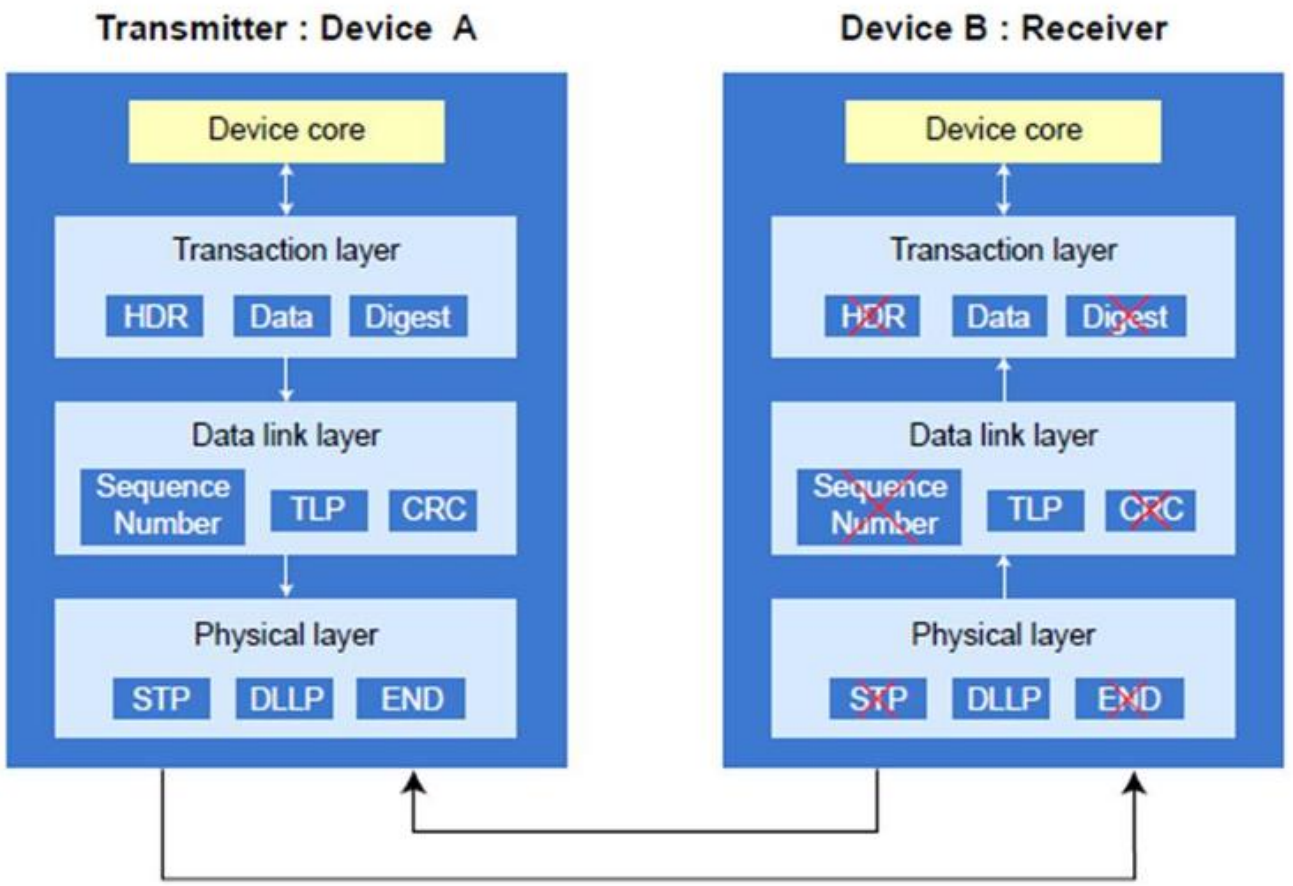
TLP	Type
Memory Read	Non-Posted
Memory Write	Posted
I/O Read	Non-Posted
I/O Write	Non-Posted
Configuration Read	Non-Posted
Configuration Write	Non-Posted
Message	Posted
Completion	Completion





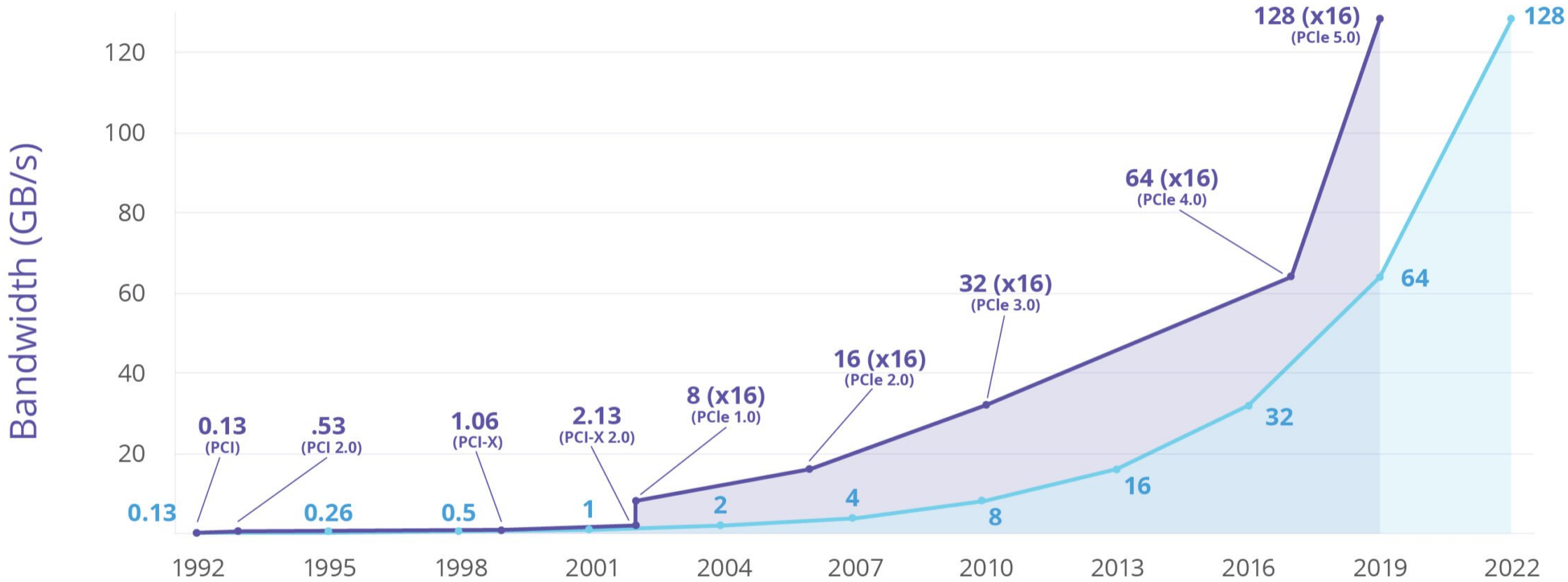


Data Transmission in PCIe



EVERY 3 YEARS

I/0 Bandwidth Doubles



PCI BANDWIDTH 1992-2019

Actual Bandwidth (GB/S)

I/O Bandwidth Doubles Every Three Years

Factors to Consider When Choosing PCIe Lane Counts

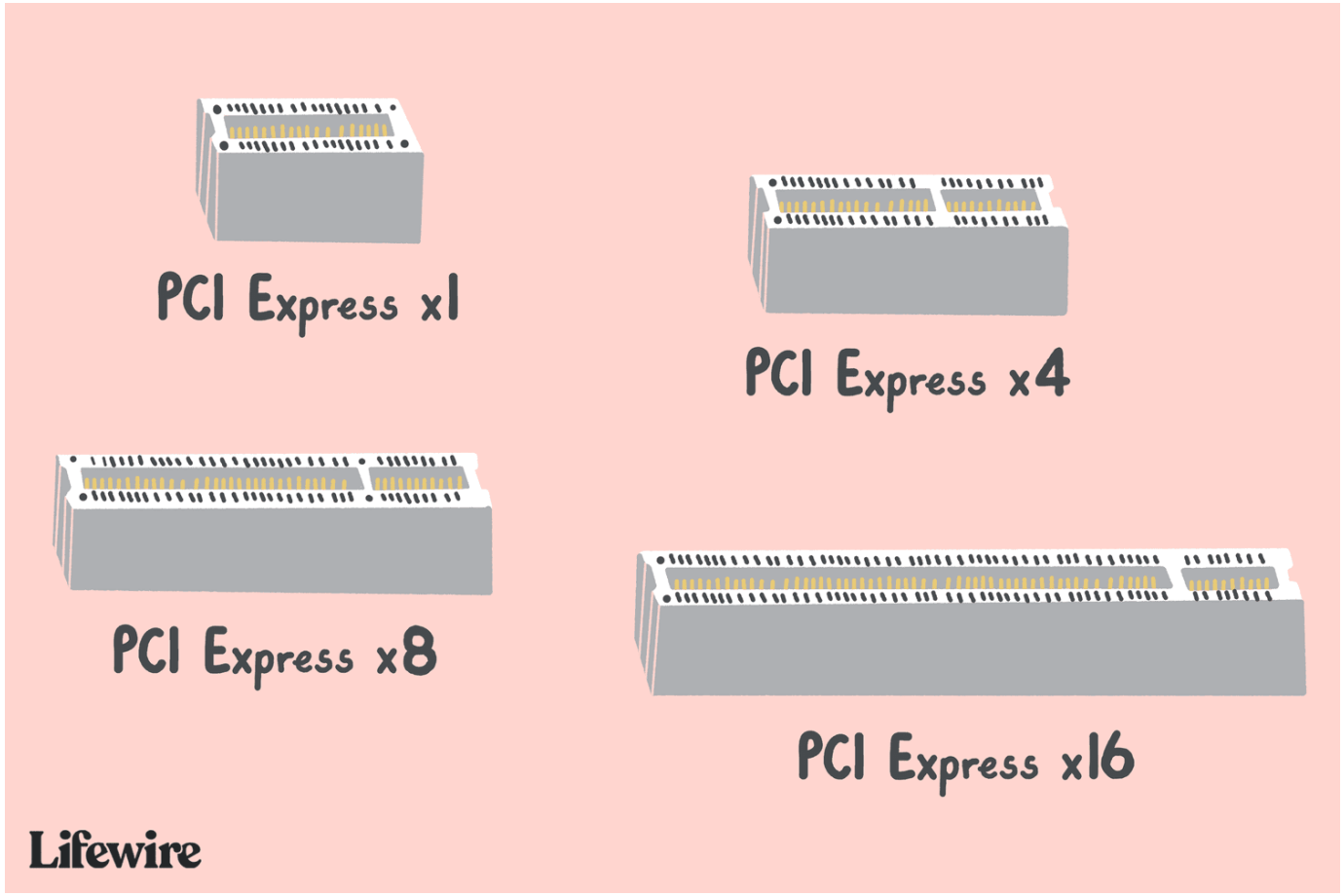
Device Bandwidth Requirements

- Each PCIe generation and lane count offers a specific **bandwidth**:

PCIe Gen	Bandwidth per Lane (x1)	x4	x8	x16
3.0	~1 GB/s	~4 GB/s	~8 GB/s	~16 GB/s
4.0	~2 GB/s	~8 GB/s	~16 GB/s	~32 GB/s
5.0	~4 GB/s	~16 GB/s	~32 GB/s	~64 GB/s
6.0	~8 GB/s	~32 GB/s	~64 GB/s	~128 GB/s

Step	Decision Criteria
1	Identify the device type and speed
2	Check how many lanes your CPU/motherboard offers
3	Match the device’s required lanes to available slots
4	Consider bifurcation or adapters if needed
5	Leave room for future expandability

PCIe size comparison



PCI Express Size Comparison Table		
Width	**Number of Pins**	**Length**
PCI Express x1	18	25 mm
PCI Express x4	32	39 mm
PCI Express x8	49	56 mm
PCI Express x16	82	89 mm

Intel Processor and Chipset PCIe Lane Configurations

- Intel 13th Gen (Raptor Lake-S) Desktop Processors
    - Processor PCIe Lanes:
        - 16 PCIe 5.0 lanes
        - 4 PCIe 4.0 lanes
        - 8 DMI 4.0 lanes (Direct Media Interface to the chipset)
    - Chipset PCIe Lanes:
        - 8 PCIe 4.0 lanes
    - Example Processor:
        - Intel Core i9-13900K
            - 16 PCIe 5.0 lanes
            - 4 PCIe 4.0 lanes
            - 8 DMI 4.0 lanes

AMD Processor and Chipset PCIe Lane Configurations

- AMD Ryzen 7000 Series (Zen 4) Desktop Processors
  - Processor PCIe Lanes:
    - 16 PCIe 5.0 lanes
    - 4 PCIe 5.0 lanes reserved for chipset connection
  - Chipset PCIe Lanes:
    - 24 PCIe 5.0 lanes
    - 4 PCIe 4.0 lanes
  - Example Processor:
    - AMD Ryzen 9 7950X
    - 16 PCIe 5.0 lanes
    - 4 PCIe 5.0 lanes reserved for chipset connection

What Will Replace PCIe?

- Video game developers are always looking to design ever more realistic games. They can only do that if they can pass more data from their game programs into your VR headset or computer screen; faster interfaces are required for that to happen.
- Because of this, PCI Express won't continue to reign supreme, resting on its laurels. PCI Express 3.0 is amazingly fast, but the world wants faster.
- PCI Express 5.0, ratified and released in 2019, supports a bandwidth of 31.504 GB/s per lane (3938 MB/s), twice what's offered by PCIe 4.0.
- The technology industry has many other non-PCIe interface standards, but since they would require significant hardware changes, PCIe looks to remain the leader for some time to come.

Q & A

Thank You! ☺