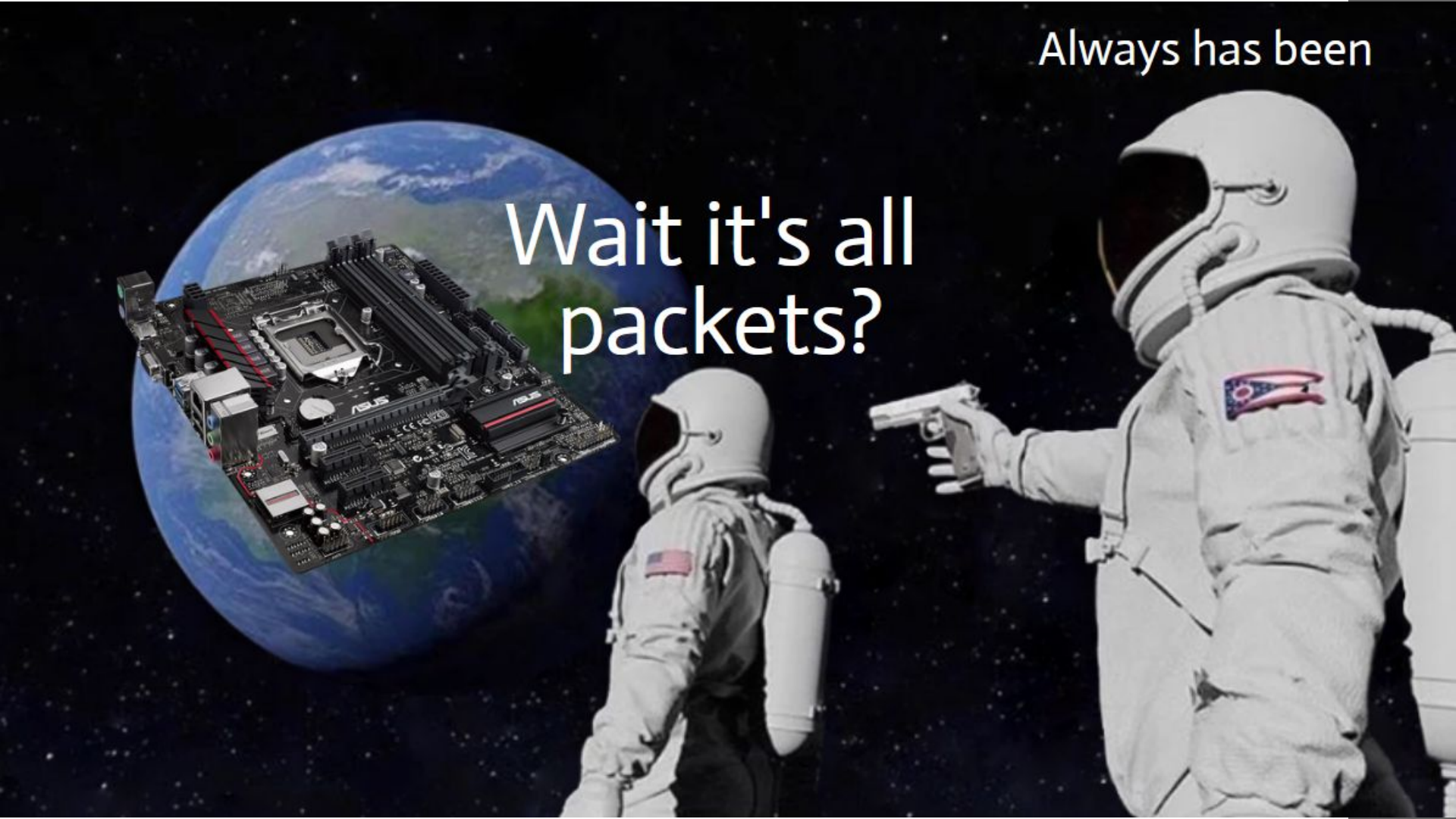
CPE 470 - PCI Express

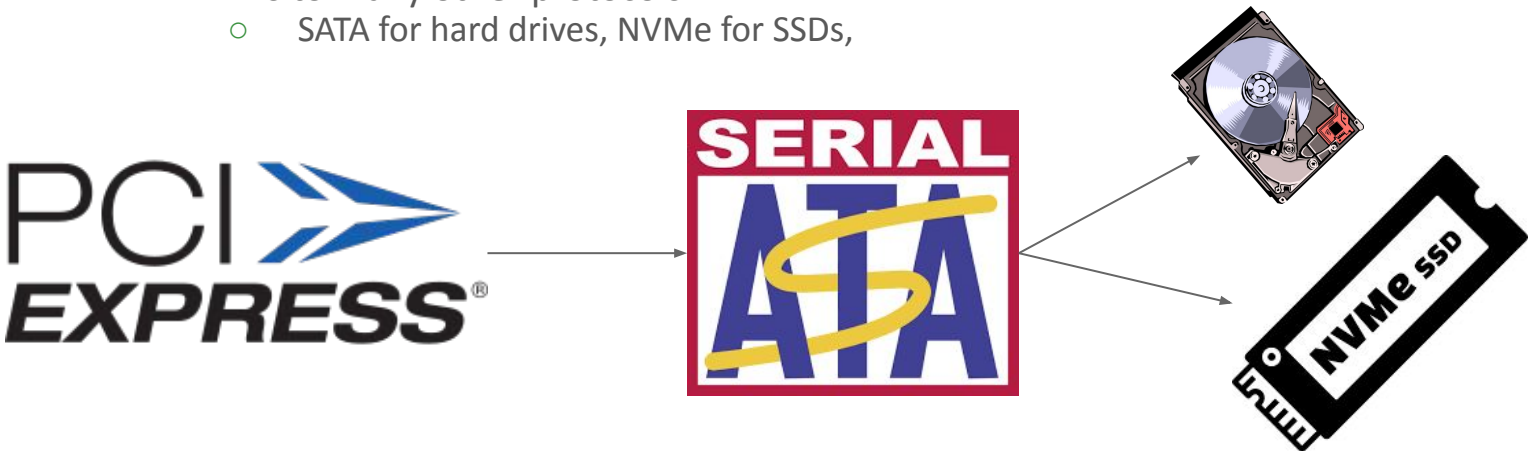

Always has been
Wait it's all packets?

Why PCIe?

Glossary

PCIe: Peripheral Component Interconnect Express

- Industry Standard
  - Primary interconnect used by PCs, graphics cards, servers, data centers
- Fast
  - Maxes out at ~**64 GBps**
- Fault Tolerance
  - Unlike previous interconnect protocols (like SPI), PCIe has inbuilt fault tolerance, and deals with
- Links to many other protocols
  - SATA for hard drives, NVMe for SSDs,



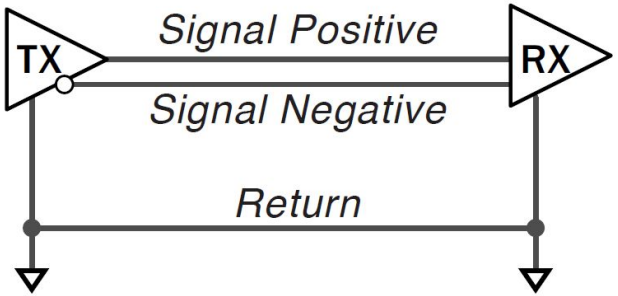
But First…

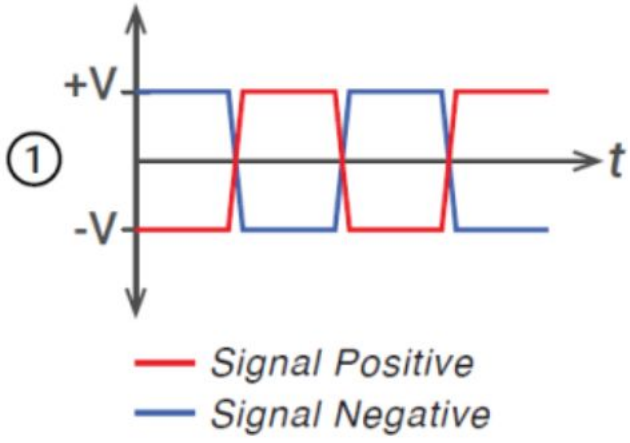
Differential Signals

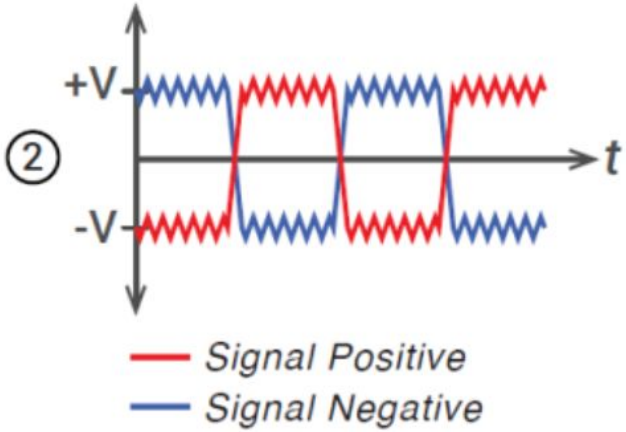
Glossary

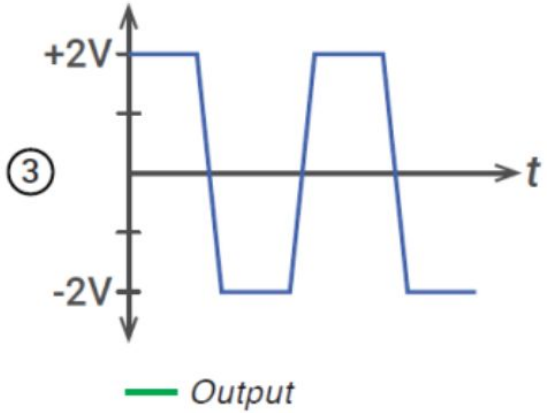
Differential Pair: two wires carrying opposite polarities of a differential signal

- At higher speeds, signal noise becomes a dominating consideration
  - Solution: Use double the wires
- **Differential Pair** encodes signal and its inverse
  - Noise equally impacts each wire
  - The difference offers twice the signal strength, and higher immunity to noise





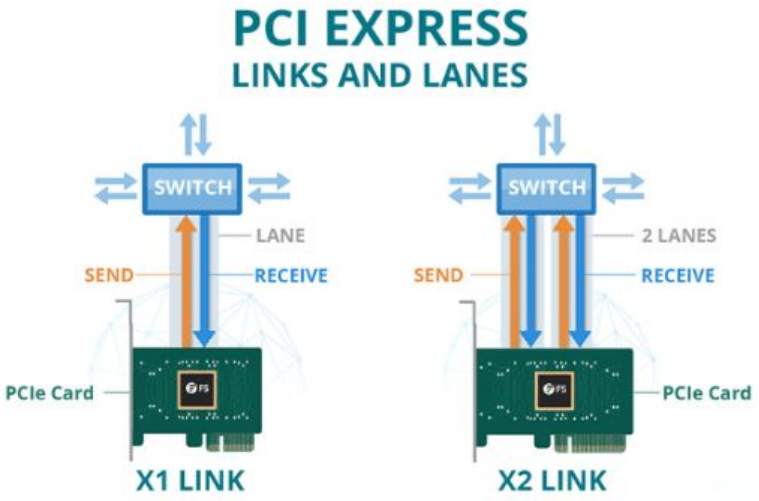




PCIe - Physical Layer

PCIe **Link** between devices:

- Interconnect composed of **Lanes**
- **Lane** is composed of two differential pairs
  - One Transmit Pair, One Receive Pair
  - 4 wires total
- Interconnect characterized by number of Lanes:
  - Powers of 2:
    - x1, x2, x4, x8, x16
    - Where x16 is 16 lanes, 64 wires
- What is missing?
  - Clock?
  - How can each **lane reach 4GB/s?** (PCIe Gen 5)

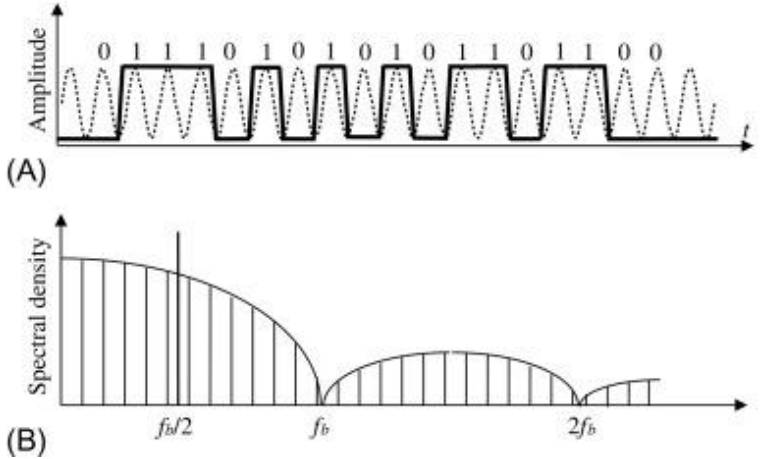


Clock Recovery and Encoding

Glossary

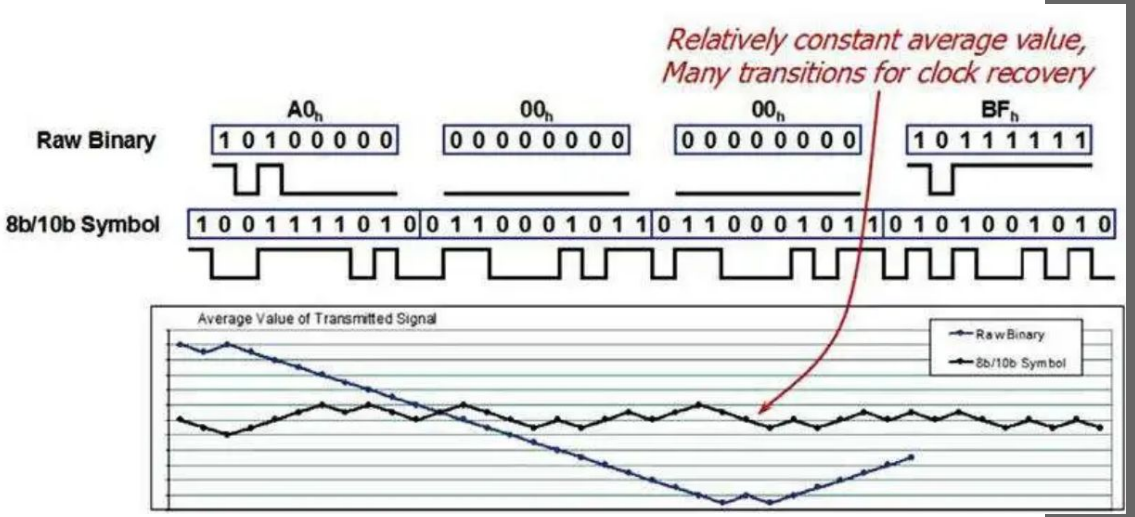
Clock Recovery: extract clock from serial data stream

- Clock Synchronization across systems is hard
  - PCB routing a 4 GHz clock would be nightmarish
- How can we transmit data without a common clock?
  - Encode the clock in the data!
- Use **clock recovery** to extract clock from switching in the data
  - Problem: How can you encode the clock in a static signal, such as sending all 0 or 1?
  - Solution: Encoding Schemes



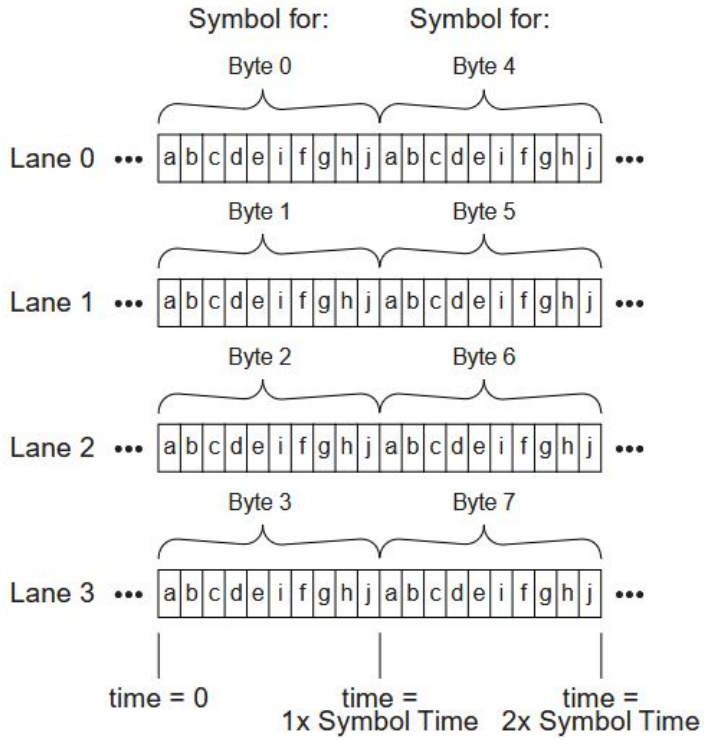
Encoding Schemes

- Encoding Schemes
  - Use more bits to guarantee switching
- PCIe uses **8b/10b**:
  - Use 10 bits to represent 8 bits
  - 8’b0 becomes 10’b1001110100
  - Lose 25% of bandwidth
- PCIe 3 moved to 128b/130b
  - Lose under 2% of bandwidth



PCIe - Byte Ordering

- PCIe differs from SPI in its byte ordering
  - In SPI, QSPI, etc, one byte gets distributed across all data lines
  - In PCIe, each byte stays in their lane!
    - Different bytes on each lane

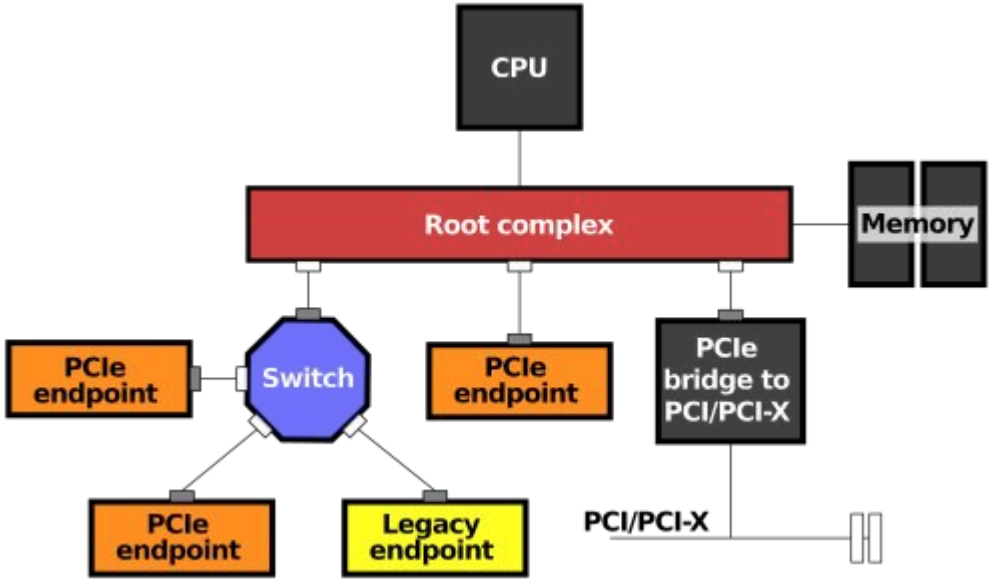


PCIe Topology

Glossary

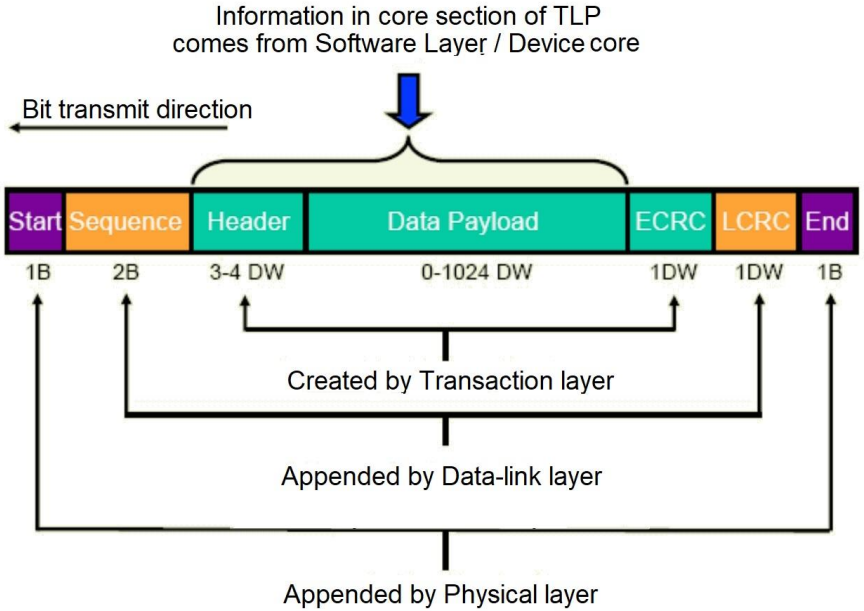
Root Complex: connects CPU and its memory to one or more PCIe ports

- Point-to-Point Connections
  - One device per Link
- All links reach CPU through the **Root Complex**
- Use a **Switch** to connect multiple endpoints together
  - Can act as a bottleneck if trying to talk to multiple endpoints



PCIe Packet Structure

- Starts with primary **transaction layer** header and data
  - Often memory operation with data to transfer
- Wrapped in a **data link layer**
  - Uniquely identifies the packet
- Appended with start and end sequence at the **physical layer**



PCIe Transaction Layer

Glossary

TLP: Transaction Layer Packet

- This layer controls what data is being transmitted
- Memory reads/writes is most common
  - Many non-memory devices are mapped as memory
- IO interfaces specifically with IO ports, which were an x86 standard
  - Tends to be outdated, kept for legacy reasons

Address Space	Transaction Types	Basic Usage
Memory	Read Write	Transfer data to/from a memory-mapped location
I/O	Read Write	Transfer data to/from an I/O-mapped location
Configuration	Read Write	Device Function configuration/setup
Message	Baseline (including Vendor–defined)	From event signaling mechanism to general purpose messaging

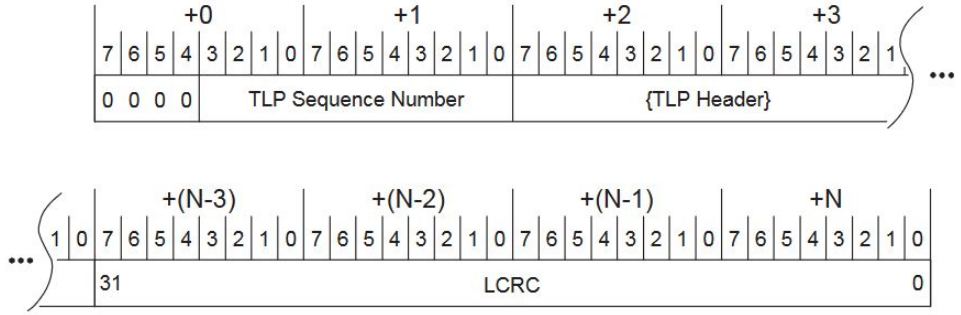
PCIe Data Link Layer

Glossary

DLLP: Data Link Layer Packets

CRC: Cyclic Redundancy Check

- Devices initialize their connection over the Data Link Layer
  - Use **DLLP**s to set up link, power, and flow control
  - PCIe is hot swappable → need a way to discover new devices when plugged in
- Uniquely identify each packet with a sequence number
  - Helps to acknowledge or indicate errors with packets later on
- Create a **CRC** for error detection
  - If any bits are lost in transit, CRC should show this and lead packet to be invalidated



PCIe Flow Control

- Flow Control is used to make sure TLPs are only sent when there is available room for them
  - TLPs are stored in buffers
  - If there was no space left in a buffer, a TLP would be dropped → failure
- PCIe implements flow control using **Credits**
  - Each link gives its partner a set number of credits
  - As transmitter sends packets, it consumes credits
  - Once out of credits, has to wait for receiver to finish processing and give it new credits

SATA

- Interface with hard drives and SSDs!
- Similar to one lane of PCIE in some ways:
  - Set of 2 differential signals
  - Clock recovery
  - 8b/10b encoding
- Protocol specific to storage devices
- Often adapted from PCIe
  - “Chipset” on a motherboard adapts from PCIe to interfaces like Sata

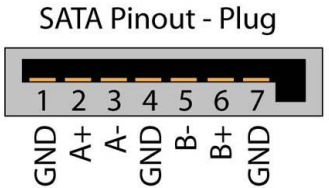
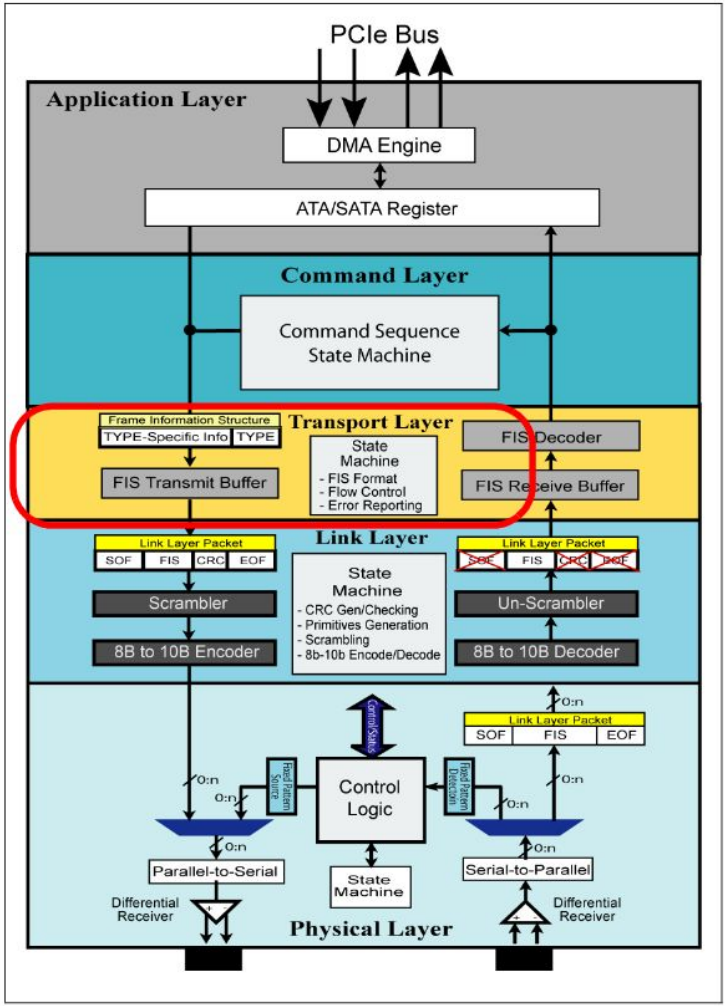
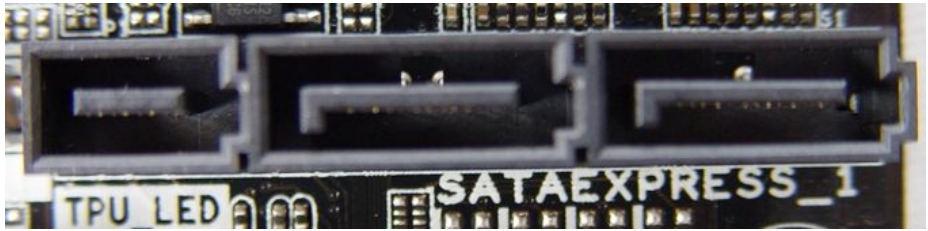


Figure 4-2: Transport Layer



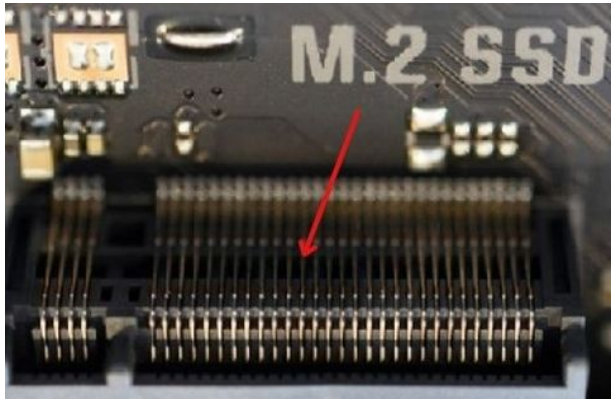
NVMe: for SSDs

- SSDs tend to have their own form factors based on existing protocols
- Need higher speed than SATA alone
  - Combine SATA with PCIe or just use PCIe



SATA Express

- 2 SATA Ports
- 2 PCIe Lanes



M.2

- Up to 4 PCIe Lanes

References

- https://www.synopsys.com/blogs/chip-design/pcie-gen1-speed-basics.html
- https://sparxeng.com/blog/hardware/mastering-differential-signals
- https://en.wikipedia.org/wiki/PCI_Express