WOLF Advanced Technology

PCIE SWITCHES AND LANE-RATE CONVERSION

– WHITEPAPER

Written by: Shari Beck
Sr. Product Manager

WOLF® ADVANCED TECHNOLOGY

INTRODUCTION:

This paper explains how a PCIe switch can receive data on one port using one PCIe lane rate and transmit it over a different port using a different lane rate.

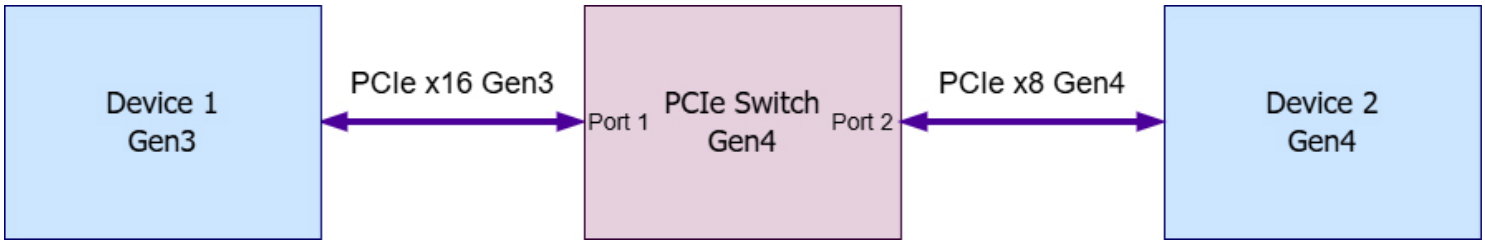

Figure 1: Example of a typical use case where lane-rate conversion is used

PCIe BASICS

PCIe was designed for point-to-point data transfer. Data moves over physical connections, or links. Each link consists of lanes, where each lane supports bi-directional data transfer using two differential pairs, one for transmit and one for receive. The number of lanes per link can be x1, x2, x4, x8, or x16 physical connections, and the number of lanes is referred to as the lane width or link width. Each lane in the connection carries data packets in parallel with the other lanes. Each PCIe lane can transmit and receive data simultaneously at full bandwidth. However, shared internal resources (such as buffers, queues, and flow control mechanisms) normally make bi-directional traffic less efficient than uni-directional traffic.

PCIe LINK TRAINING

At power-up or reset, PCIe links undergo training to establish a valid connection between devices. Once links are established, system software (such as firmware or the operating system) enumerates downstream devices. During link training, the link width (number of lanes) and the fastest supported speed are negotiated based on the capabilities of both ends. The speed negotiation typically starts at Gen1 and progresses to the highest generation supported by both devices. If link conditions change, the PCIe link will attempt to retrain automatically.

PCIe DATA PACKETS

PCIe uses structured data packets, as shown in the following diagram.

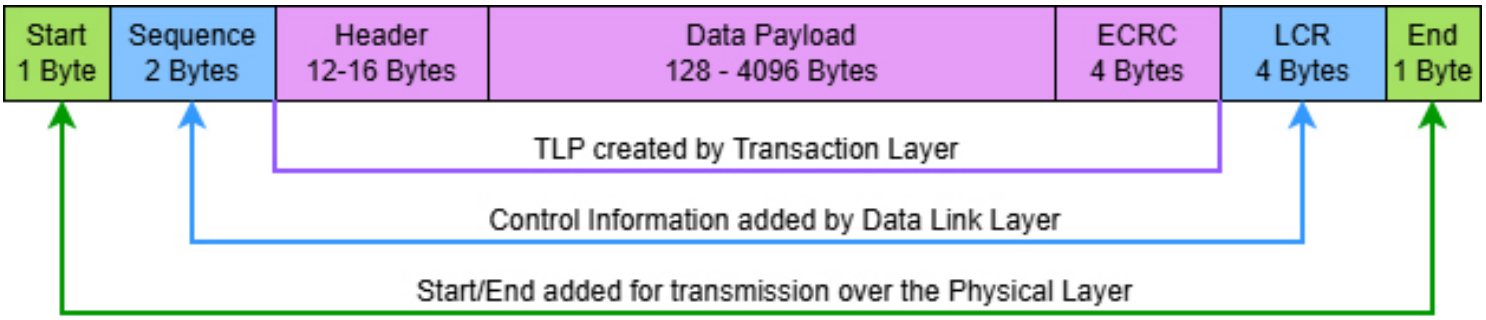

Figure 2: PCIe packet structure

TRANSACTION LAYER PACKET (TLP):

A TLP is constructed by the device that will be transmitting the packet. The TLP includes the data payload that uses one of several allowed payload sizes ranging from 128 to 4096 bytes. Each device has a Maximum Payload Size Supported (MPSS). During enumeration a Linux or Windows OS will walk through all devices to the root, determine the smallest MPSS of the devices in the available paths, and then set the Maximum Payload Size (MPS) of each device so that devices can only generate packets that are within the maximum size supported.

The TLP includes Payload data, headers, and additional small overhead data. Using smaller TLPs can provide lower latency. Using larger TLPs can provide greater efficiency as the payload data will be a larger percentage of the total packet size.

The size of the payload data is often specific to the application. For example, a GPU device will typically use a MPSS of 256 bytes or 512 bytes. GPUs perform many parallel computations, where smaller, more frequent transfers can be handled efficiently, and the resulting reduction in PCIe efficiency has minimal impact on overall system performance.

DATA LINK LAYER (DLL):

The Data Link Layer (DLL) provides the information required to detect missing packets or errors and to ask for retransmission if needed. It also manages flow control, negotiating between the receiver and sender so that no packets are lost due to buffer overflow. For a PCIe transmission the flow control is based on a credit-based flow control mechanism, where the receiver determines how many credits (buffer capacity) it has and the transmitter consumes credits when it is sending packets, and must stop sending packets when it runs out of credits. For a PCIe switch credits are managed per link, not per switch.

The DLL is generated for each hop by the transmitting device. For example, if the data path is Root Complex to Switch to Endpoint, then the root complex will add the DLL for the hop from the Root Complex to the Switch. When the Switch receives the packet it will remove the DLL created by the Root Complex and create a new DLL for the next hop from the Switch to the Endpoint.

PHYSICAL LAYER:

The Physical Layer determines how the data will be encoded into electrical signals and how those signals will be sent between devices. Physical Layer signaling is generated hop-by-hop. Among other things, it manages link training which establishes the PCIe generation (Gen) being used. Then depending on the Gen determined it encodes the data into signals that are used by that Gen. PCIe Gen1 and Gen2 use 8b/10b encoding, meaning a byte (8 bits) is encoded into a 10 bit-symbol for error reduction and improved signal integrity. This requires a 20% overhead. Starting with PCIe Gen3, the 8b/10b encoding was replaced with 128b/130b encoding, a high efficiency line coding scheme that reduced the overhead to ~1.54%. With 128b/130b encoding data is grouped into 128-bit blocks with a 2-bit synchronization header for a 130-bit frame. This improves bandwidth efficiency since less overhead data is transmitted compared to 8b/10b encoding.

During link training the Physical Layer negotiates the number of lanes (link width) to use to send data between the connected devices. In multi-lane situations, the physical layer also performs the lane striping, also known as byte striping. During lane striping the physical layer distributes one outgoing data stream across multiple lanes of a wide link (x4, x8, x16), which enables a higher aggregate throughput by using multiple lanes in parallel. The physical layer transmitter interleaves ("stripes") consecutive bytes across the available lanes in round-robin order. The receiver then reassembles ("un-stripes") the bytes back into their original order. Striping enables a wide line to behave as one fast logical link instead of as multiple independent links.

PCIe TRANSFER RATE

PCIe transfer rate is measured in Giga Transfers per second (GT/s). Usually transfer speeds are specified in Gigabits per second (Gbits/s or Gbps). The usable throughput per lane is the transfer speed less the overhead required by the encoding method. For example: Given that PCIe Gen4 runs at a transfer rate of 16.0 GT/s per lane per direction with 128b/130b encoding, the usable throughput line-rate is:

Per lane, per direction: 16GT/s x (128/130) = 15.7538 Gb/s ≈ 1.969 GB/s

PCIe Gen	Encoding Scheme used	Transfer Rate (GT/s) per lane & direction	Throughput (GB/s) per direction			
			x1	x4	x8	x16
1.0	8b/10b	2.5	0.25	1	2	4
2.0	8b/10b	5	0.50	2	4	8
3.0	128b/130b	8	0.9846	3.938	7.877	15.754
4.0	128b/130b	16	1.969	7.877	15.754	31.508
5.0	128b/130b	32	3.94	15.754	31.508	63.01

Table 1: PCIe Generational Performance, Gen1 to Gen5

PCIe SWITCH BASICS

A PCIe switch has a defined number of ports that can be used to connect to devices, and maximum number of lanes that can be configured by each port. The switch will have a maximum PCIe Gen speed, which defines the maximum bandwidth supported. The switch uses internal buffering when receiving and redirecting packets.

Each of the ports on the PCIe switch will independently undergo PCIe link training with the connected device during power-up or reset. Different ports can operate at different bandwidths and use different numbers of lanes.

LANE-RATE CONVERSION

A feature of modern PCIe switches (Gen3 or newer) is lane rate conversion. This is the ability of the PCIe switch to change the rate of the data being sent, provided sufficient buffering and downstream bandwidth are available to transmit/receive the data.

For example, one device is connected to the switch by a PCIe Gen 3 ×16 link, while a second connected device has an established PCIe Gen4 ×8 link. Table 1 shows that both links can theoretically handle the same bandwidth (15.754 GB/s), because the PCIe Gen 3 link has twice the number of lanes of the PCIe Gen4 link, while each Gen4 lane can transmit data twice as fast as a Gen3 lane.

When the switch receives a packet, it removes the transmitter's Data Link Layer (DLL) information and Physical Layer information. This data is re-generated for every hop. The DLL and Physical Layer for the next hop will be generated using the new control information required for the outgoing port, including the encoding and striping. The switch then wraps the TLP which includes the payload data with the new control information and transmits the packet. That is, the content remains the same while the link-level encapsulation is regenerated every hop.

The TLP packet size does not change during lane-rate conversion. PCIe switches can't operate as a "packet re-segmenter", that is, a PCIe switch cannot split TLPs into smaller TLPs to accommodate a downstream device. However, the Maximum Payload Size Supported (MPSS) that was determined during enumeration will guarantee that no device in the path will generate a TLP larger than any device in that reachable path can support.

In summary, the PCIe switch will support lane-rate conversion without any extra manual configuration requirements from the user since it is a built-in feature of modern PCIe switches.

WOLF®
ADVANCED TECHNOLOGY