DMA Videodata Transmission over PCI Express

Thomas Zerrer

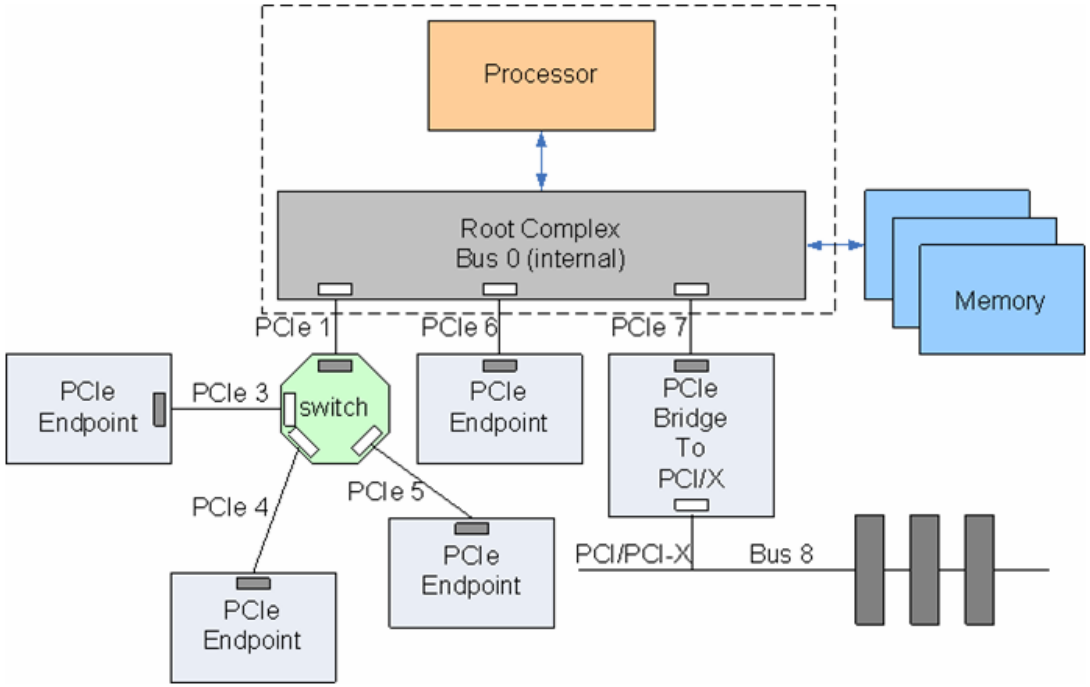
History:

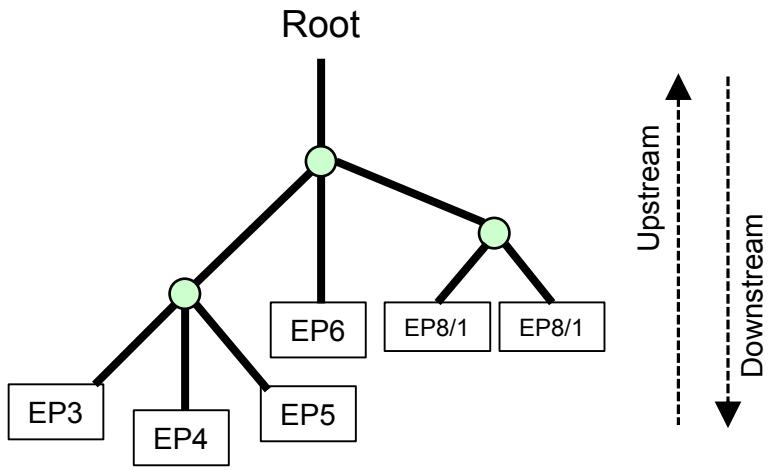
- founded in 2005
- Office in Hildrizhausen, near Stuttgart
- Team of 4 Engineers
- Customers from the Industrial, Test & Measurement and Automotive branches
- Xilinx Alliance Partner

Objectives:

- Understand the structure and basic terms of a PCI Express System
- Estimate the data throughput at a given Linkwidth / Linkspeed
- Advantages of DMA Datatransfer over Memory Read Requests
- typical FPGA architectures for transmitting DMA Data

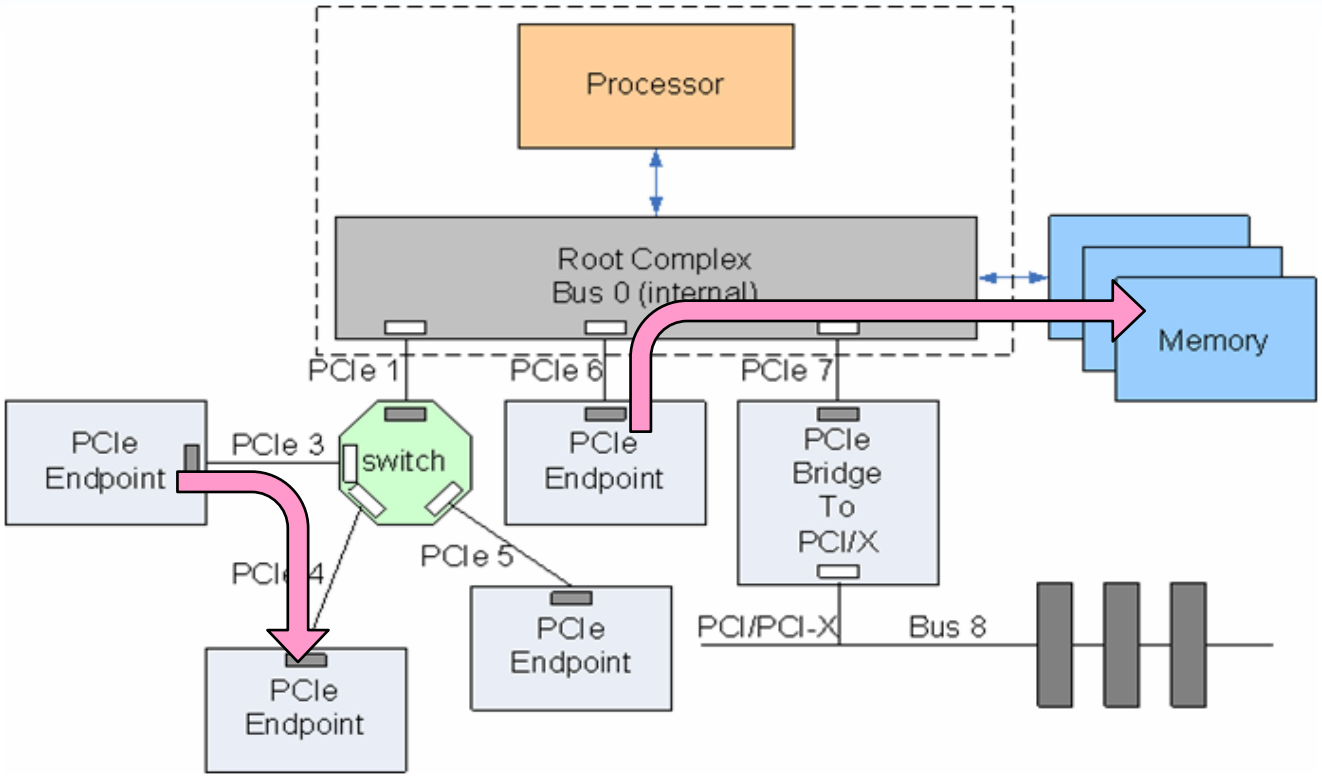
Typical PCI-Express Architecture





Features:

- PCI Express is always a point to point connection
- Switches allow the access of several endpoints from one PCIe connection (PCIe1)
- Every endpoint can exchange data with the Root Complex **or** other PCIe endpoints

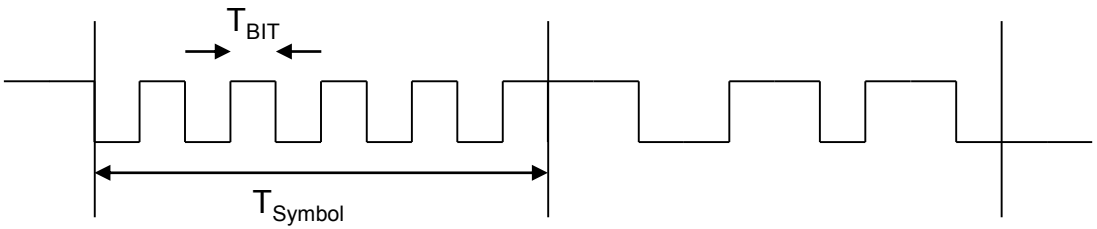


Data Transmission

Bus Mastering Features:

- Every endpoint can be a Bus Master and write or request data
- No processor involvement for such direct accesses
- Data Transfer can be to the Rootcomplex (e.g. Memory) or to any PCIe endpoint

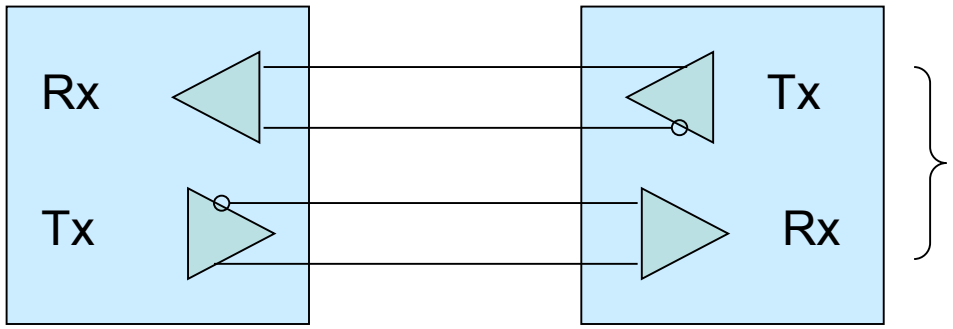
PCI-Express is a serial Highspeedlink:



10 Bits = 1 Symbol for Gen 1 & 2,
~8 Bits = 1 Symbol for Gen 3-5

T_{BIT}:		Release Date
Gen 1 = 400 ps	2.5 GBit/s	2003
Gen 2 = 200 ps	5.0 GBit/s	2006
Gen 3 = 125 ps	8.0 GBit/s	2010
Gen 4 = 62.5 ps	16.0 Gbit/s	2017
Gen 5 = 31.25 ps	32.0 Gbit/s	2019

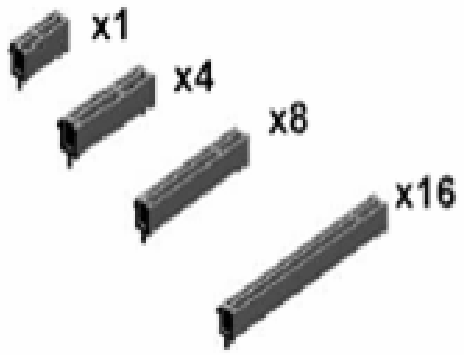
PCI-Express is differential (LVDS) and full duplex:



One Lane
consists of 4 physical
signal traces

The Bitclock is embedded in the data and must be recovered with a Clock Data Recovery circuit

Link-Width		x1	x2	X4	x8	x16
Theoretical throughput in MByte / s* per direction	Gen 1 : 2.5 Gbit/s	250	500	1.000	2.000	4.000
	Gen 2 : 5.0 Gbit/s	500	1.000	2.000	4.000	8.000
	Gen 3 : 8.0 Gbit/s	1.000	2.000	4.000	8.000	16.000
	Gen 4 : 16.0 Gbit/s	2.000	4.000	8.000	16.000	32.000



Motherboard Connectors

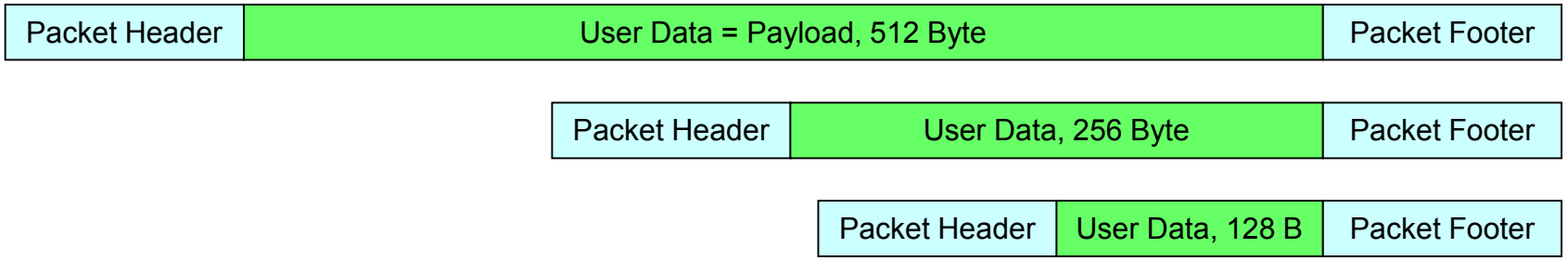
* 1 MByte = 10^6 Byte

In order to achieve more data throughput than one lane provides, several lanes can be grouped together, forming higher link widths. This is called a multilane link.

Caution: The actual throughput is less than the given values above, because of protocol overhead and other decreasing factors !

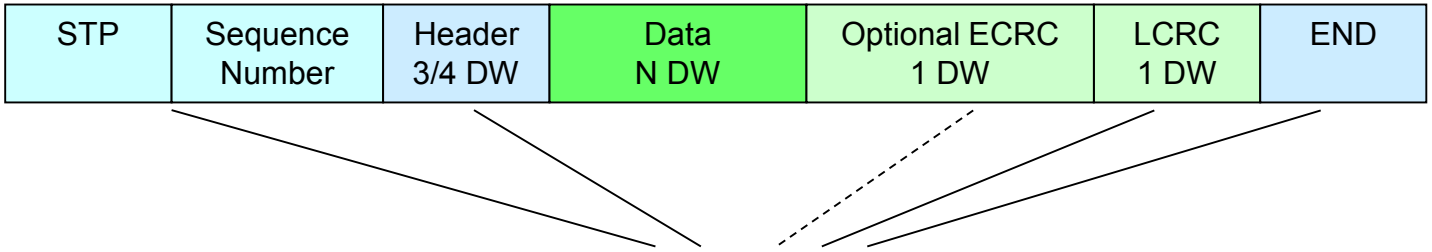
Understanding the maximum payload size (MPS) :

Example Data Packets:



- MPS defines the maximum amount of user data (= payload) contained in a PCI-Express data packet (TLP).
- The higher this value is, the less is the protocol overhead, since Packet header and Packet Footer remain the same.
- The actual MPS value is negotiated during link training between the endpoint and the link partner and remains fixed until powerdown.
- The Spec defines MPS values of 128, 256, 512, 1024, 2048 or 4096 Bytes.
- Important : You have to select the right CPU (Host, Rootport) carefully. MPS > 512 are rare !

TLP Packet:



Packet Overhead = 5 DW for 32 Bit Addressing and 6 DW for >4 Gbyte Addressing
(no ECRC/Digest assumed)

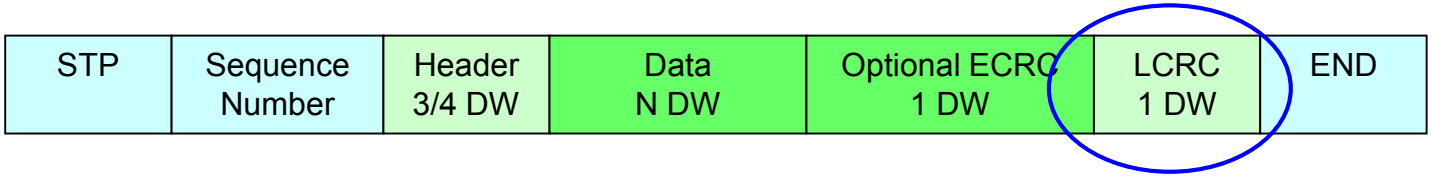
MPS* in Bytes	5 DW Overhead	5 DW Loss in %	Max Throughput (Gen 1, x1, 5 DW)**	Max Throughput (Gen 1, x1, 6 DW)**
128	20 / (128+20)	13,5 %	216,3	210,5
256	20 / (256+20)	7,2 %	232	228,5
512	20 / (512+20)	3,8 %	240,5	238,75
1024	20 / (1024+20)	1,9 %	245,3	244,25
4096	20 / (4096+20)	0,5 %	248,8	248,75

* Maximum Payload Size

** in MByte/s, 1 MB = 10^6 Byte

DW = Double Word = 32 Bit

Bad Signal Integrity causes packet replays



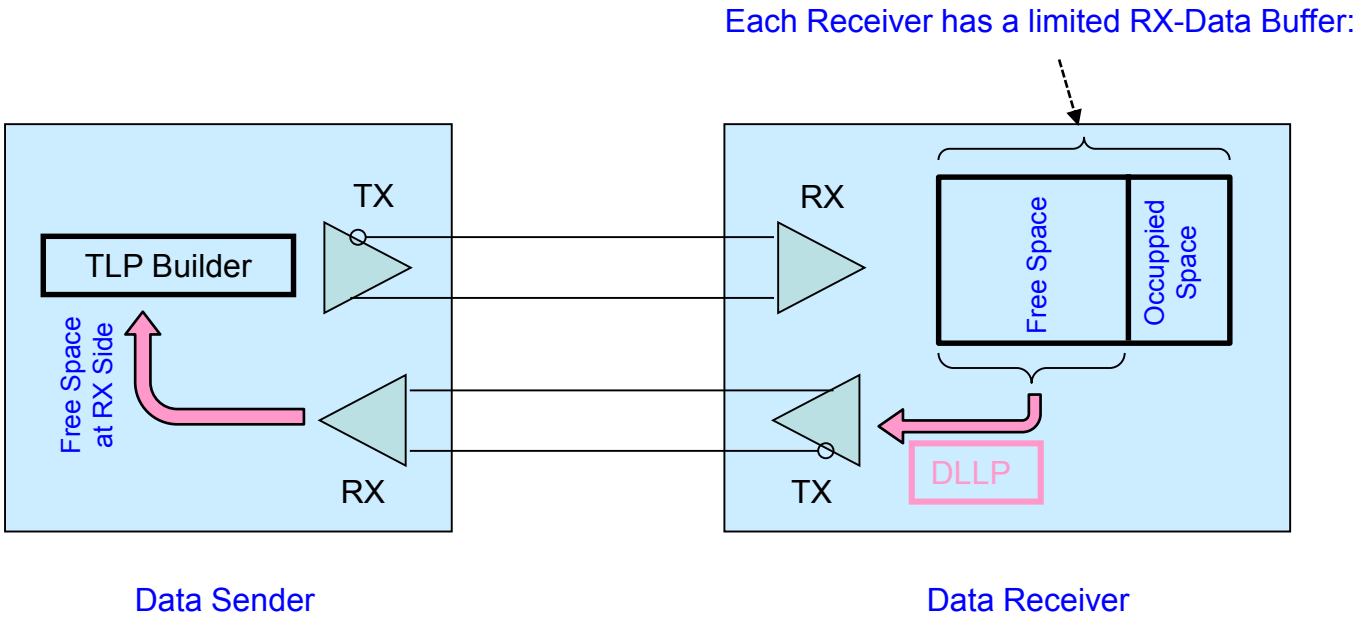
- Every PCIe Data Packet (TLP) contains a CRC checksum (PCIe term “LCRC”) in order to ensure the data integrity
- If this TLP is lost or contains a wrong CRC, the sender is informed to resend the TLP
- If the endpoint has to replay the packets very often the effective DMA Performance decreases.

Actions:

- The Transceiver Parameter of the endpoint have to optimized according to the effective trace lengths in order to ensure the best quality of the received signal.
- The PCB Layout has to obey high speed rules, typically found at the vendors website
- The host System has to be carefully selected in order to ensure good signal quality

Note : The amount of packet retries can be measured with Smartlogic’s DMA Performance Demodesign

Understanding Flow Control



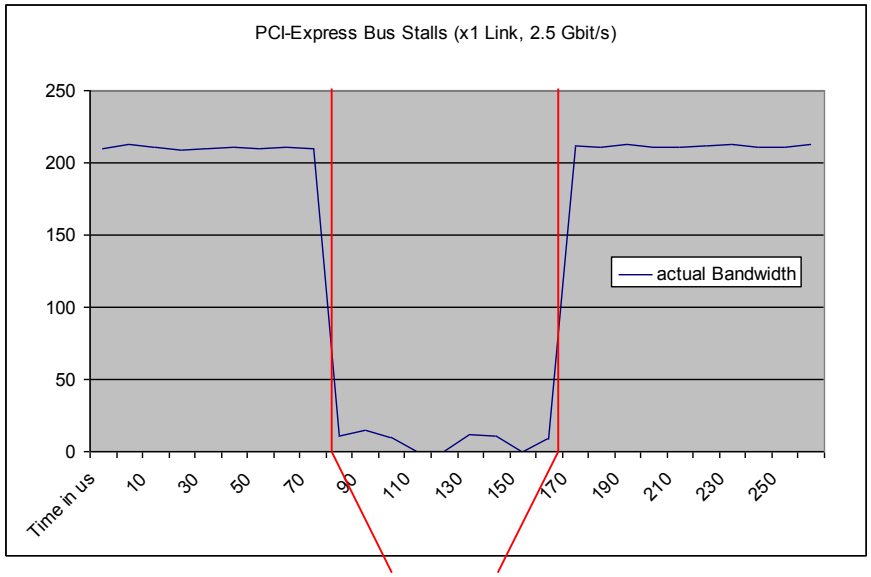
- Flow Control is used to limit the amount of transmitted data in order to prevent Fifo overflows at the receiver side.
- The Free Space in the RX FIFO is transmitted periodically to the Link Partner in form of a Data Link Layer Packet (DLLP)

Flow Control Rules

- Flow Control is used to limit the amount of transmitted data in order to prevent Fifo overflows at the receiver side.
- Therefore a transmitter is only permitted to send data, if it has enough flow control credits from the link partner.
- If the link partner does not advertise credits, the transmitter is not allowed to send data.
- Since the PCI-Express Protocol defines different categories of incoming data (i.e. completions of read requests, Write requests, etc) different receiver buffers exist, which have their own dedicated credit pools. So it is possible, that an endpoint might be allowed to send completions but is not allowed to send DMA data requests.
- The names of the different credit categories are: posted header, posted data, non-posted header, non-posted data, completion header, completion data.

Slow Flow Control updates from the Host decrease DMA Performance

DMA Average
Performance is good

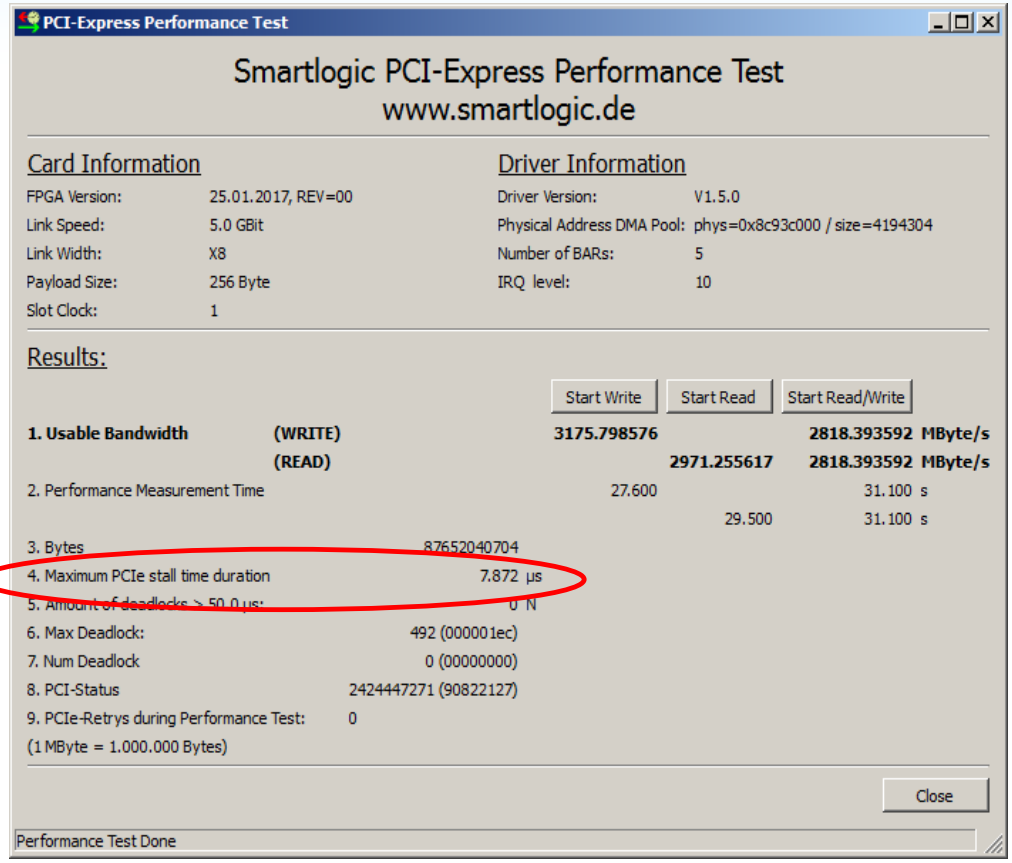


PCI-Express Bus Stall

Result:

PCIe Bus Stalls are Host dependent. They directly impact the design (i.e. FIFO depths, etc)
Good systems show Stall-Times < 10 us. Bad systems show stall-times up to 140 us !

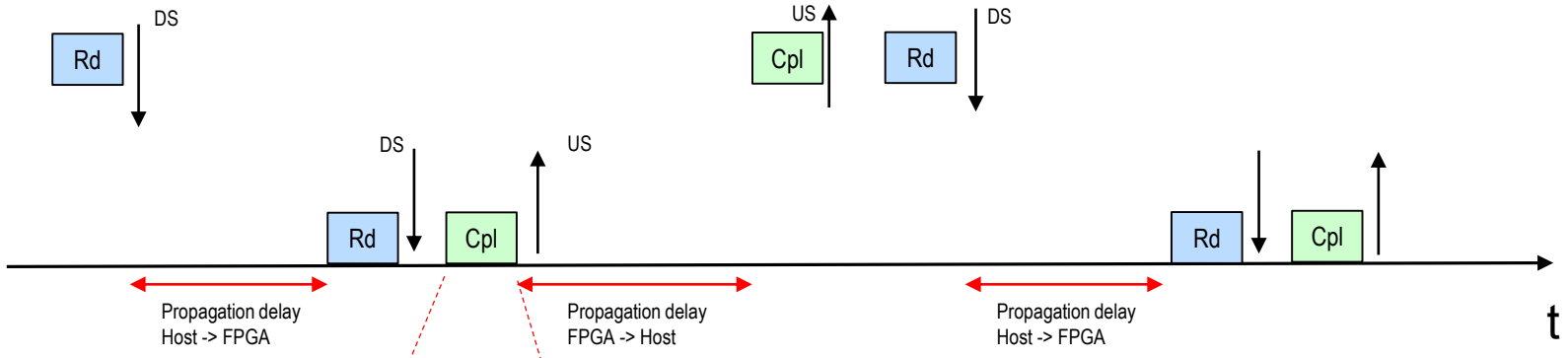
Note : PCIe Bus stall time can be measured with Smartlogic's Performance Demo Design.



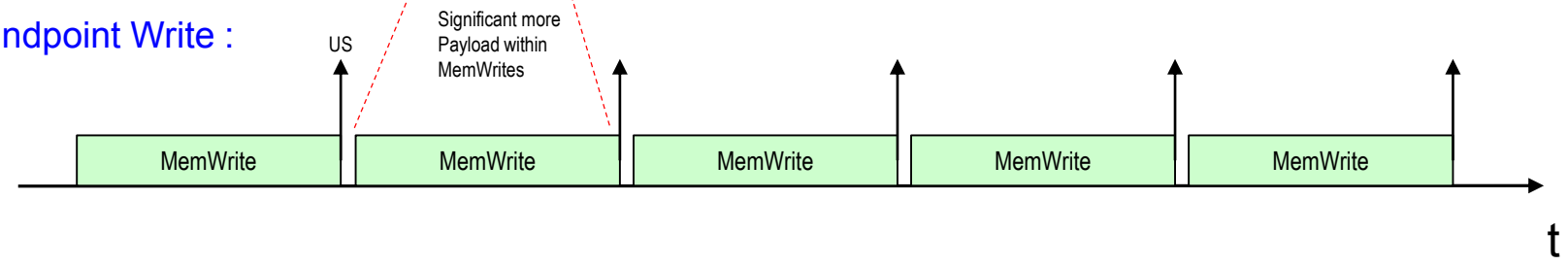
Features:

- Performance Measurement for Read & Write Directions to measure Host Parameters like Throughput, MPS, CRC Errors, Stall Time
- Available for Evaluation with Xilinx Demoboards (Bitstream, GUI, Driver)
AC701, KC705, KC105, VCU108 and others (see smartlogic Webpage)

CPU Read:



Endpoint Write :



- Read Requests are very inefficient compared with maximum packed MemWrite Requests:
  First Reason is the Propagation Delay between Rootport and Endpoint
  Second Reason is, that the Read Requests never request more than 4-8 DWs
- For Maximum PCIe throughput, Memwrite Requests are necessary

Link Speed / Width	Type	4 DW Read MBytes / s	2 DW Read MBytes / s	1 DW Read MBytes / s
G1 / X1	Typical Slot	6,3	3,15	1,6
G1 / X4	Typical Slot	9,1	4,55	2,3
G2 / X4	Typical Slot	10,4	5,2	2,6
G1 / X1	Graphics Slot	8,9	4,5	2,25
G2 / X4	Graphics Slot	21,7	10,9	5,5
G2 / X8	Graphics Slot	27,3	13,65	6,9
G3 / X8	Graphics Slot	43,7	21,85	11

1 MByte = 10^6 Byte

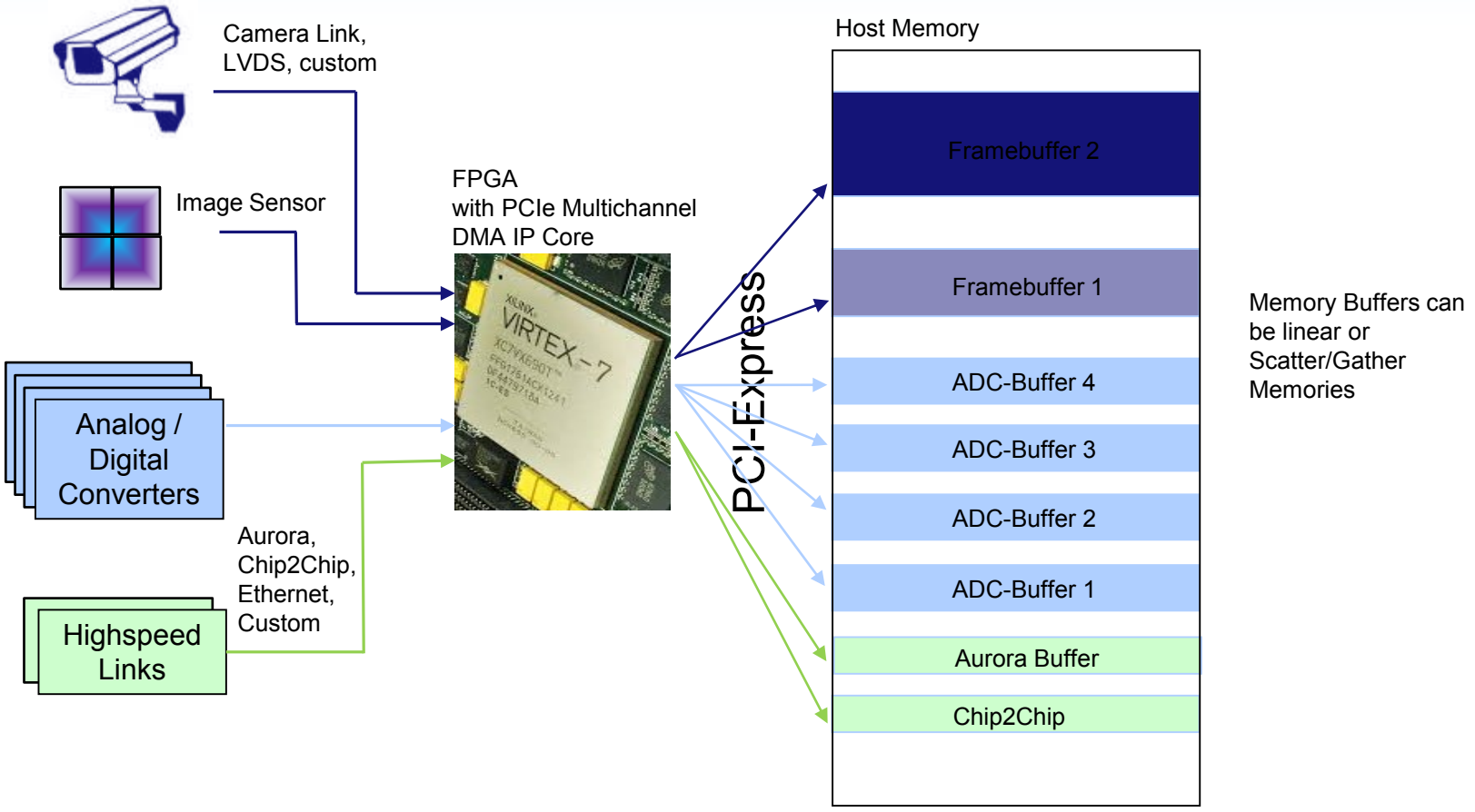
Facts:

- PCI Express Read Performance is very poor. Conventional PCI even had a better Read Performance
- The Read Performance depends on the number of switches, the access type and the used OS

The table above provides measured values on a Dell PC System with Win 7-64 Bit. The values may vary on other systems.

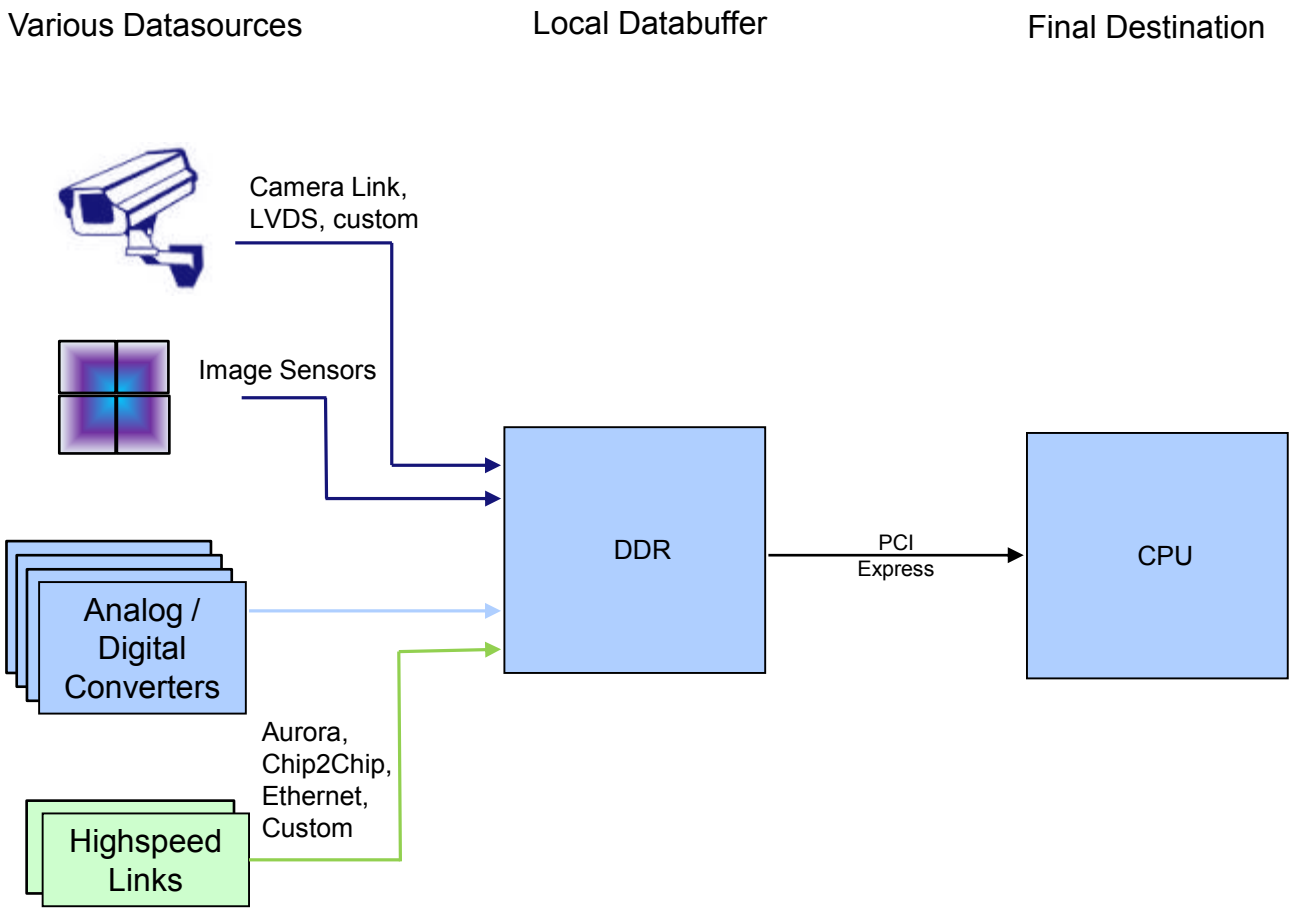
Streaming different Datasources to the Host

SMARTLOGIC

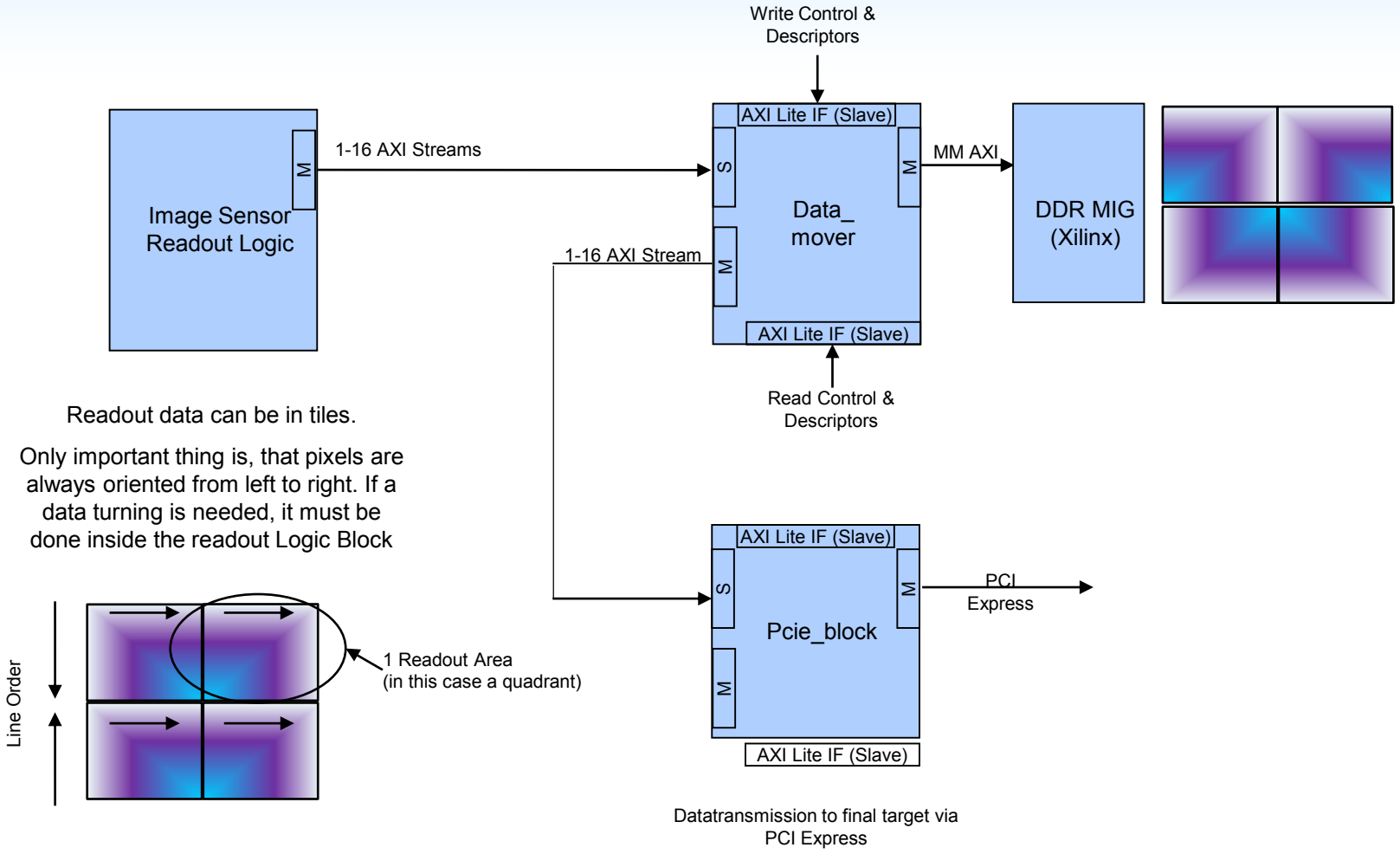


- Up to 16 Datasources possible
- Each Datasource is stored in a separate Memory Buffer

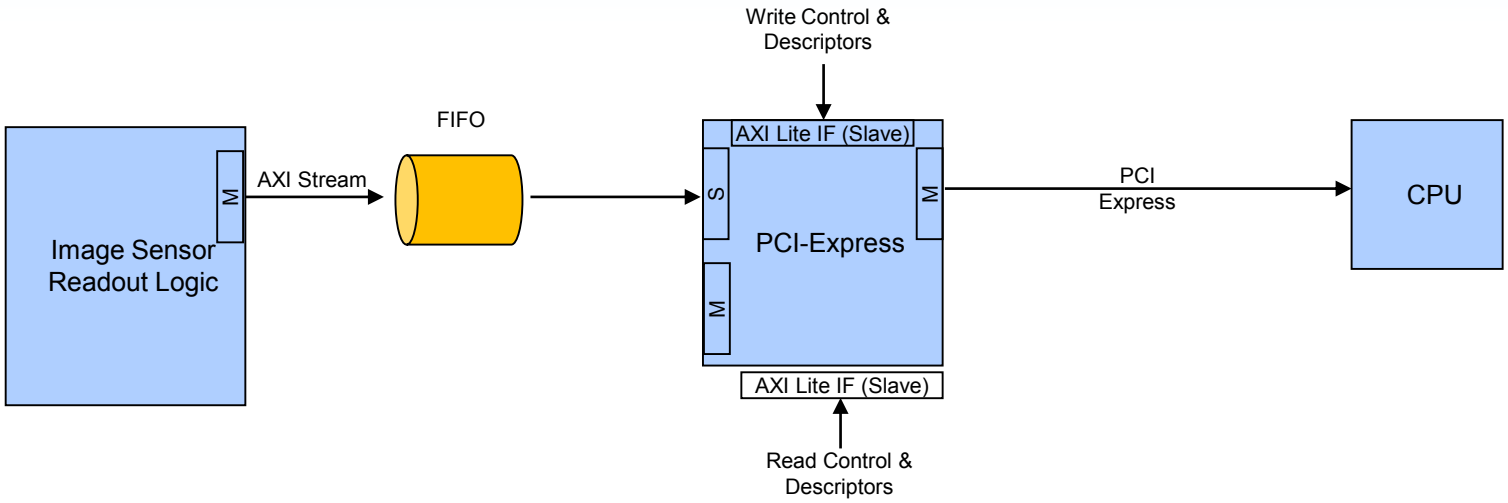
- Core cares for complete addressmanagment, User only supplies data via AXI Stream Interfaces
- The destination can be the Host Memory (as shown) or an other PCIe endpoint Device



Two basic IP Blocks

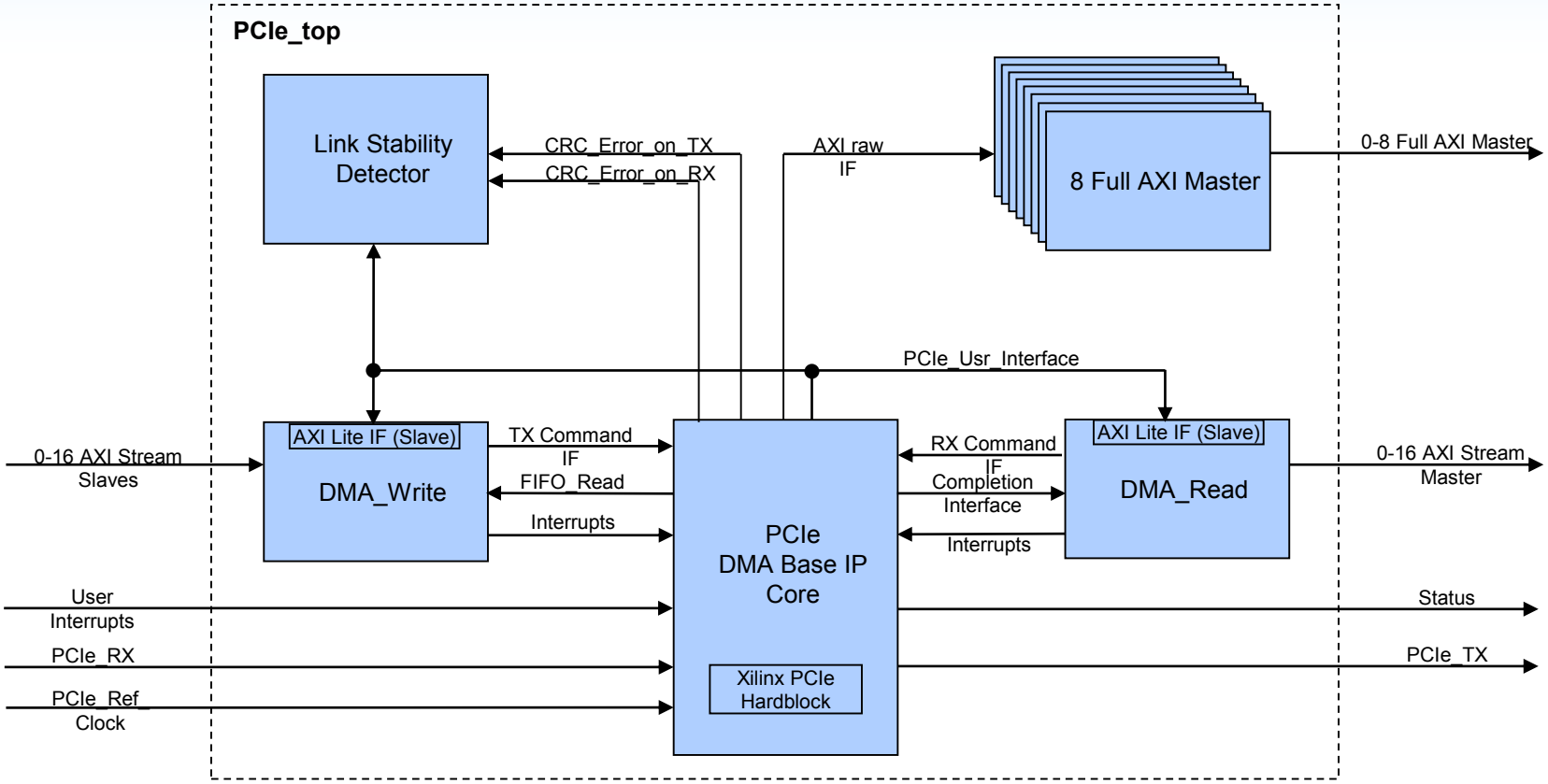


- Manage Multiple AXI Streams to AXI Memory mapped (Datamover)
- Provide suitable addressing schemes
- Transmit multiple AXI Streams over PCI Express (PCIe Block)



It is possible to transmit the DMA Data without a DDR Framebuffer, but make sure that :

- SRAM based FIFOs must be provided
- PCI Express Stall Times must be deterministic to calculate FIFO depths
- Required throughput should not exceed 80% of net Link Bandwidth



Highlights:

- Each AXI Stream interface is configurable regarding data width and has ist own clock input.
- Unused AXI Master / Slaves do not use Logic resources.
- Link Stability detector for production testing (i.e. soldering problems) for PCB Layout validation