

Practical introduction to PCI Express with FPGAs

Michal HUSEJKO, John EVANS
michal.husejko@cern.ch
IT-PES-ES

Agenda

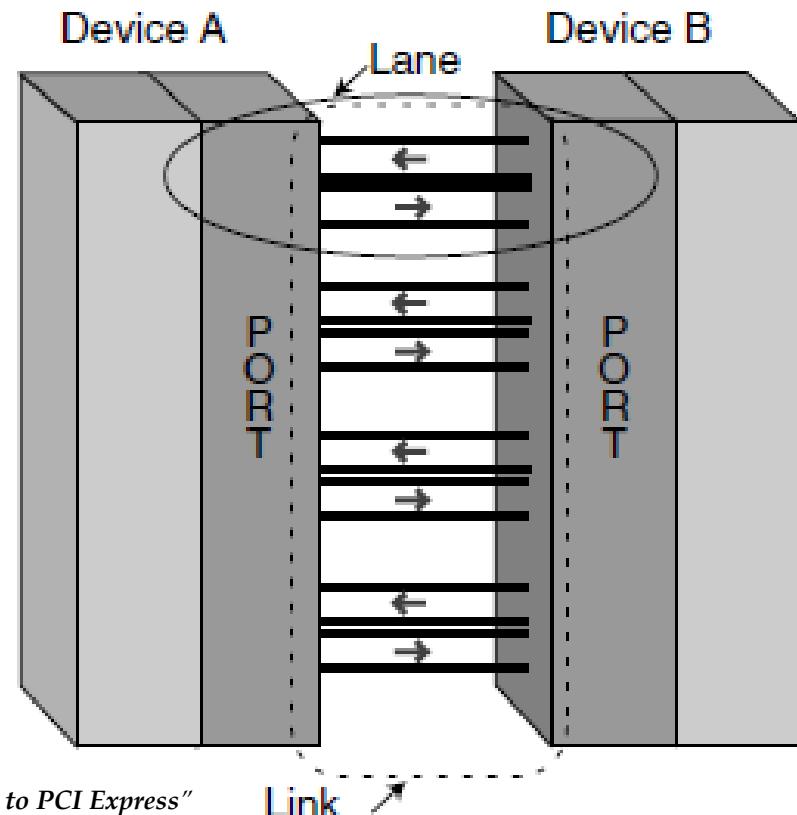
- What is PCIe ?
 - System Level View
 - PCIe data transfer protocol
- PCIe system architecture
- PCIe with FPGAs
 - Hard IP with Altera/Xilinx FPGAs
 - Soft IP (PLDA)
 - External PCIe PHY (Gennum)

System Level View

- Interconnection
- Top-down tree hierarchy
- PCI/PCIe configuration space
- Protocol

Interconnection

- Serial interconnection
- Dual uni-directional
- Lane, Link, Port
- Scalable
 - Gen1 2.5/ Gen2 5.0/ Gen3 8.0 GT/s
 - Number of lanes in FPGAs: x1, x2, x4, x8
- Gen1/2 8b10b
- Gen3 128b/130b



Tree hierarchy

- Top-down tree hierarchy with single host
- 3 types of devices: Root Complex, Endpoint, Switch
- Point-to-point connection between devices without sideband signalling
- 2 types of ports: downstream/upstream
- Configuration space

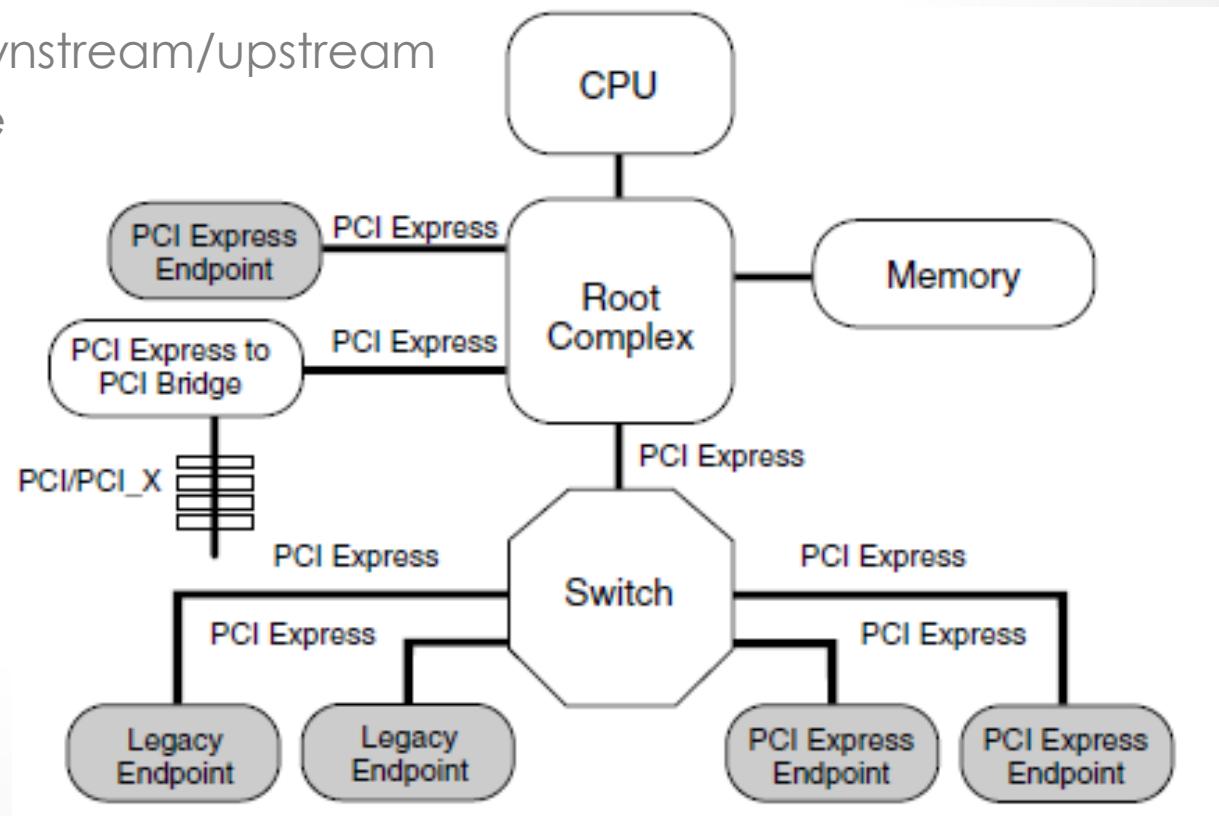


Image taken from "Introduction to PCI Express"

PCIe Configuration space

- Similar to PCI conf space – binary compatible for first 256 bytes
- Defines device(system) capabilities
- Clearly identifies device in the system
 - Device ID
 - Vendor ID
 - Function ID
 - All above
- and defines memory space allocated to device.

PCIe transfer protocol

- Transaction categories
- Protocol
- Implementation of the protocol

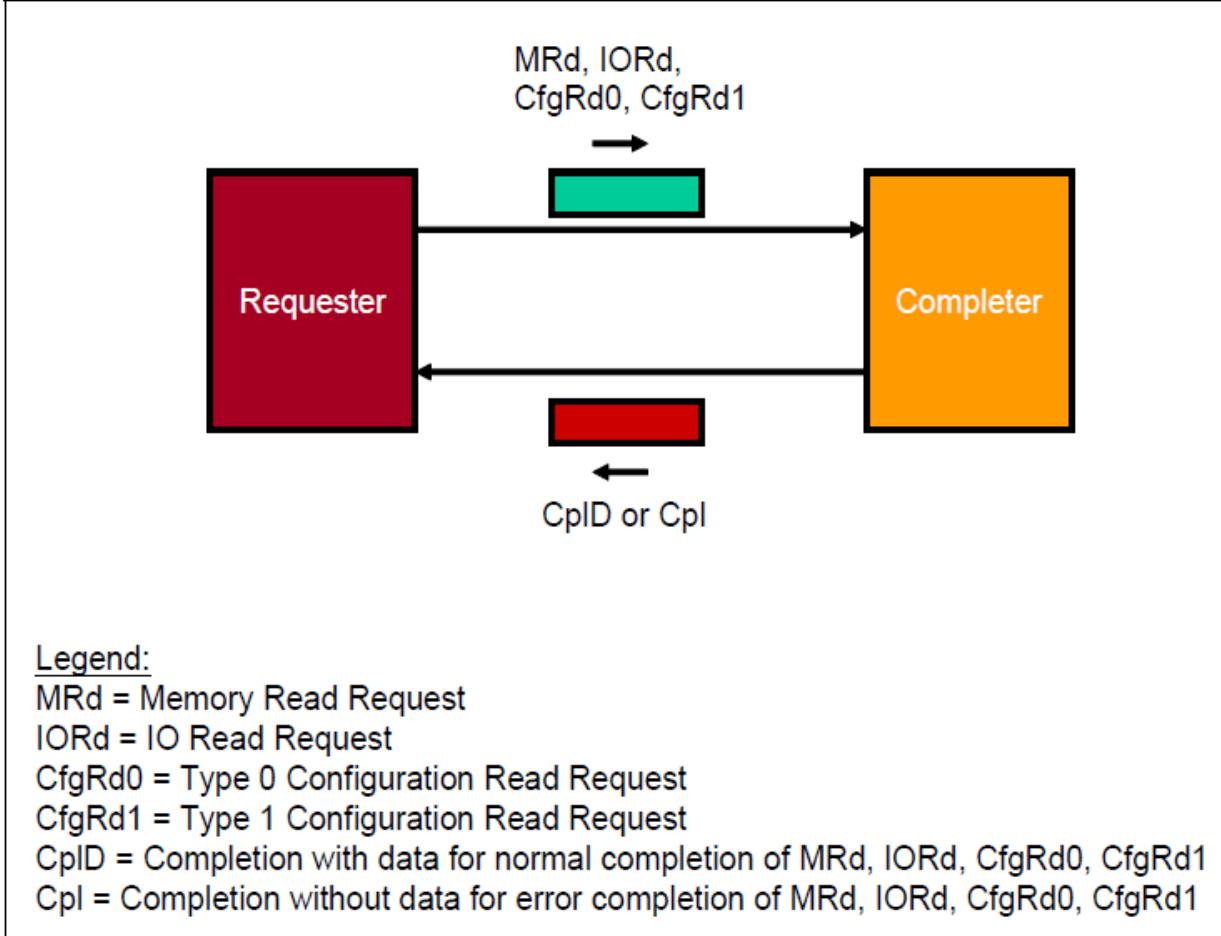
Transaction categories

- Configuration – move downstream
- Memory – address based routing
- IO – address based routing
- Message – ID based routing

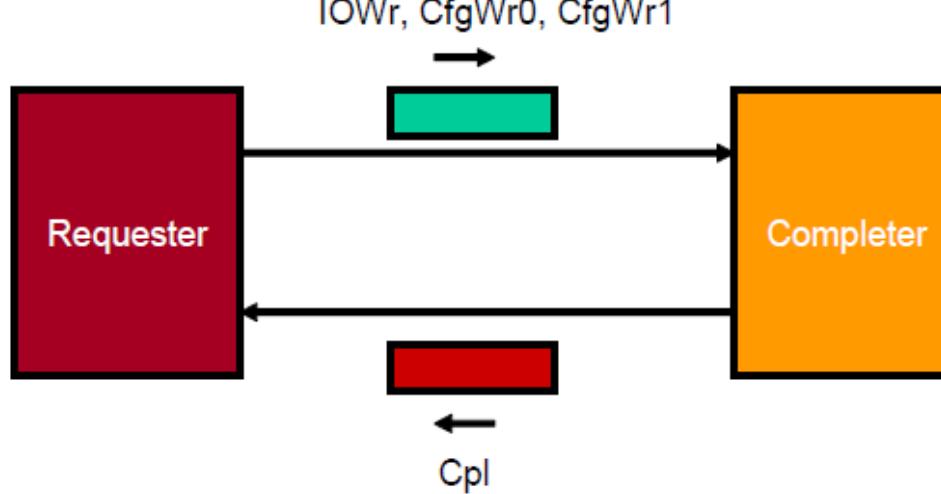
Transaction Types

Transaction Type	Non-Posted or Posted
Memory Read	Non-Posted
Memory Write	Posted
Memory Read Lock	Non-Posted
IO Read	Non-Posted
IO Write	Non-Posted
Configuration Read (Type 0 and Type 1)	Non-Posted
Configuration Write (Type 0 and Type 1)	Non-Posted
Message	Posted

Non-posted read transactions



Non-Posted write transactions



Legend:

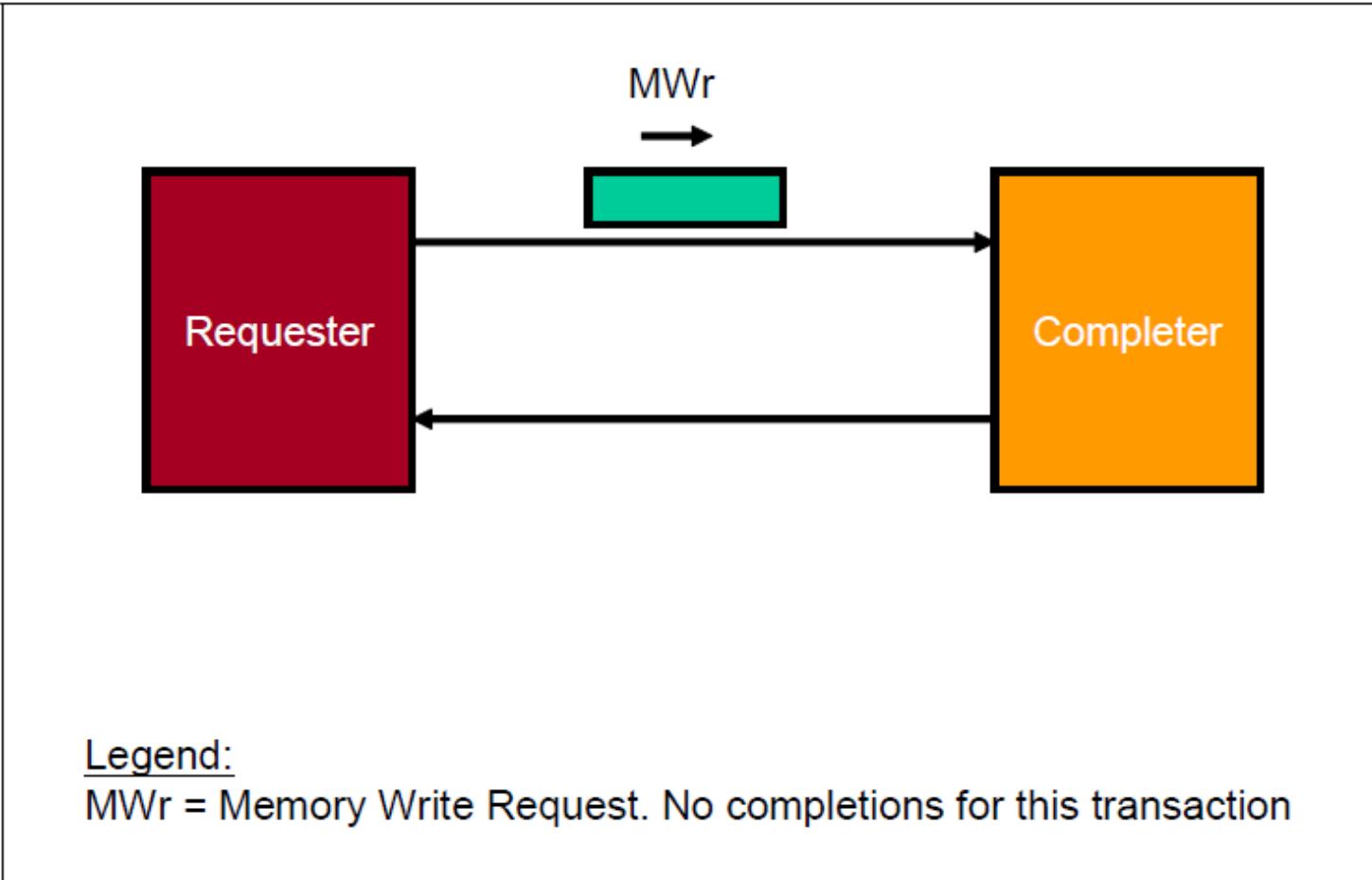
IOWr = IO Write Request

CfgWr0 = Type 0 Configuration Write Request

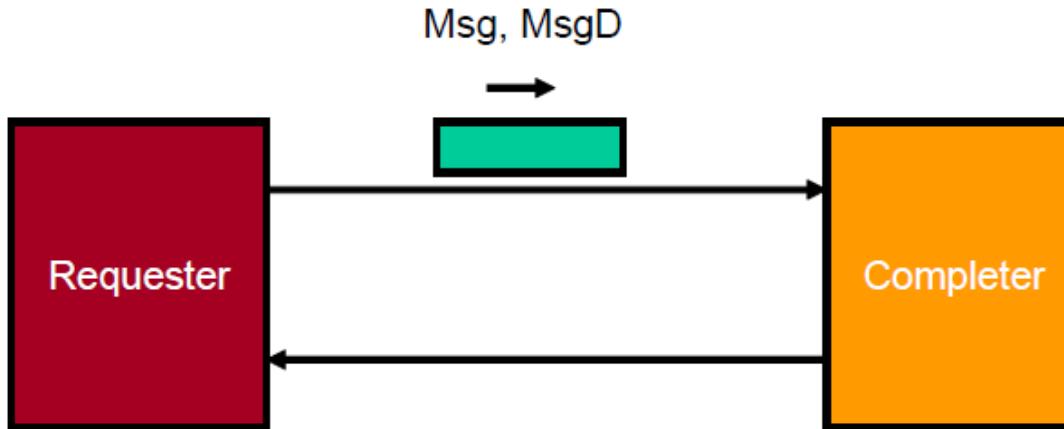
CfgWr1 = Type 1 Configuration Write Request

Cpl = Completion without data for normal or error completion of IOWr, CfgWr0, CfgWr1

Posted Memory Write transactions



Posted Message transactions

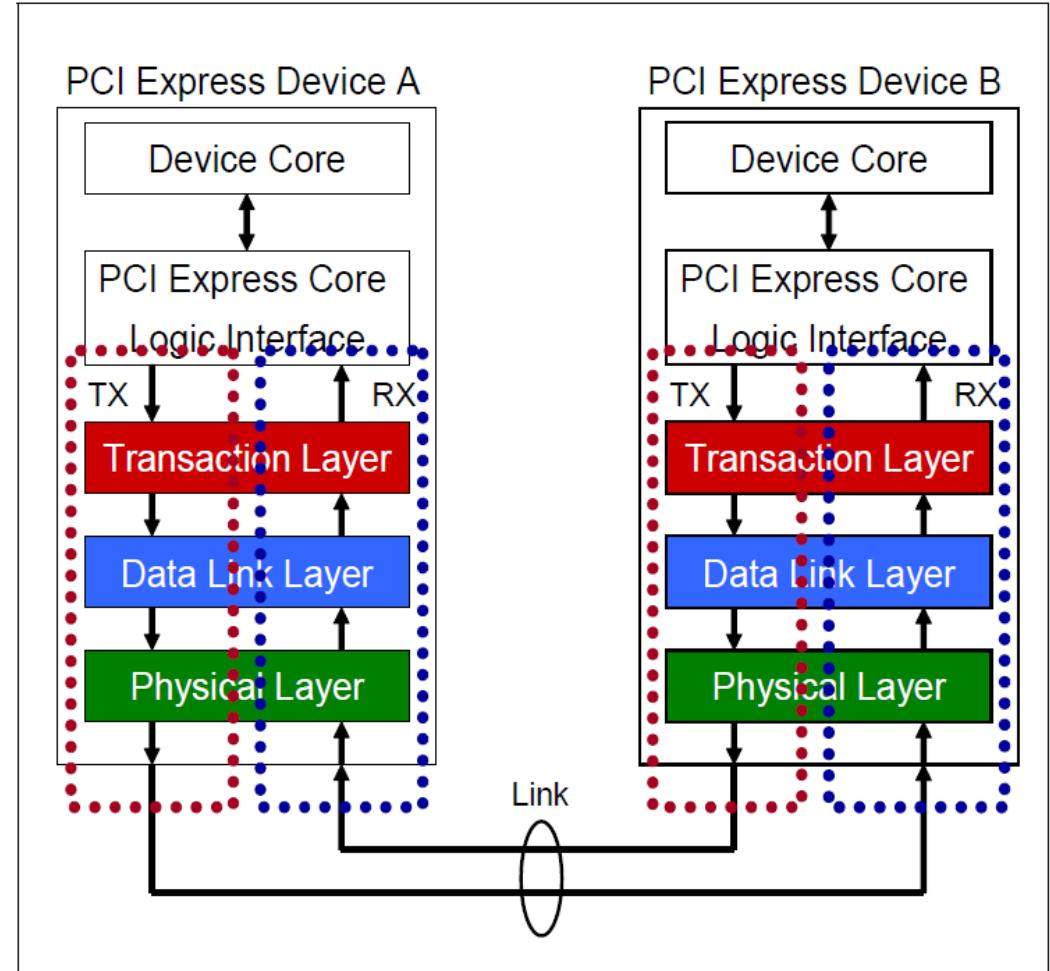


Legend:

Msg = Message Request without data
MsgD = Message Request with data

PCIe Device Layers

- 3 layer protocol
- Each layer split into TX and RX parts
- Ensures reliable data transmission between devices



Physical Layer

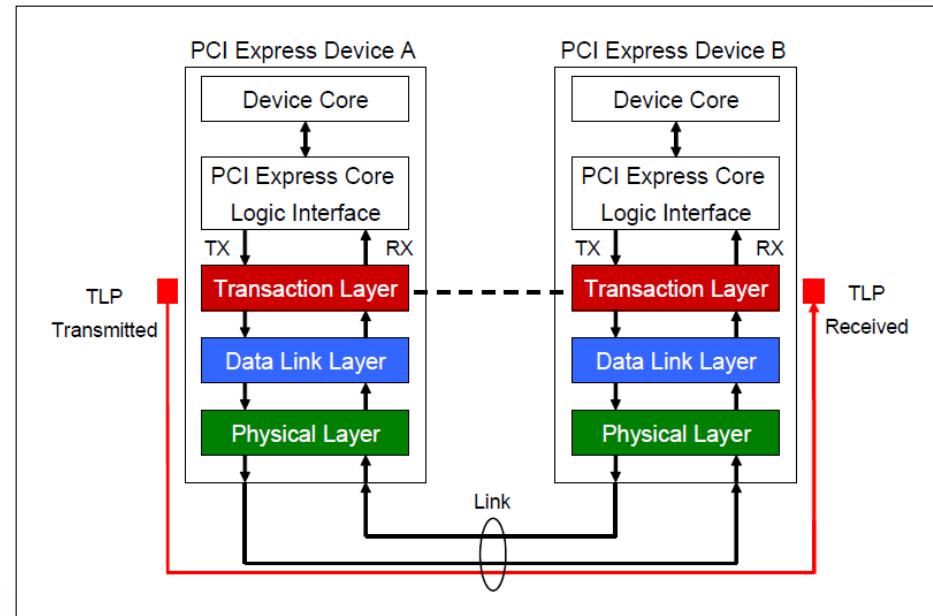
- Contains all the necessary digital and analog circuits
- Link initialization and training
 - Link width
 - Link data rate
 - Lane reversal
 - Polarity inversion
 - Bit lock per lane
 - Symbol lock per lane
 - **Lane-to-lane deskew**

Data Link layer

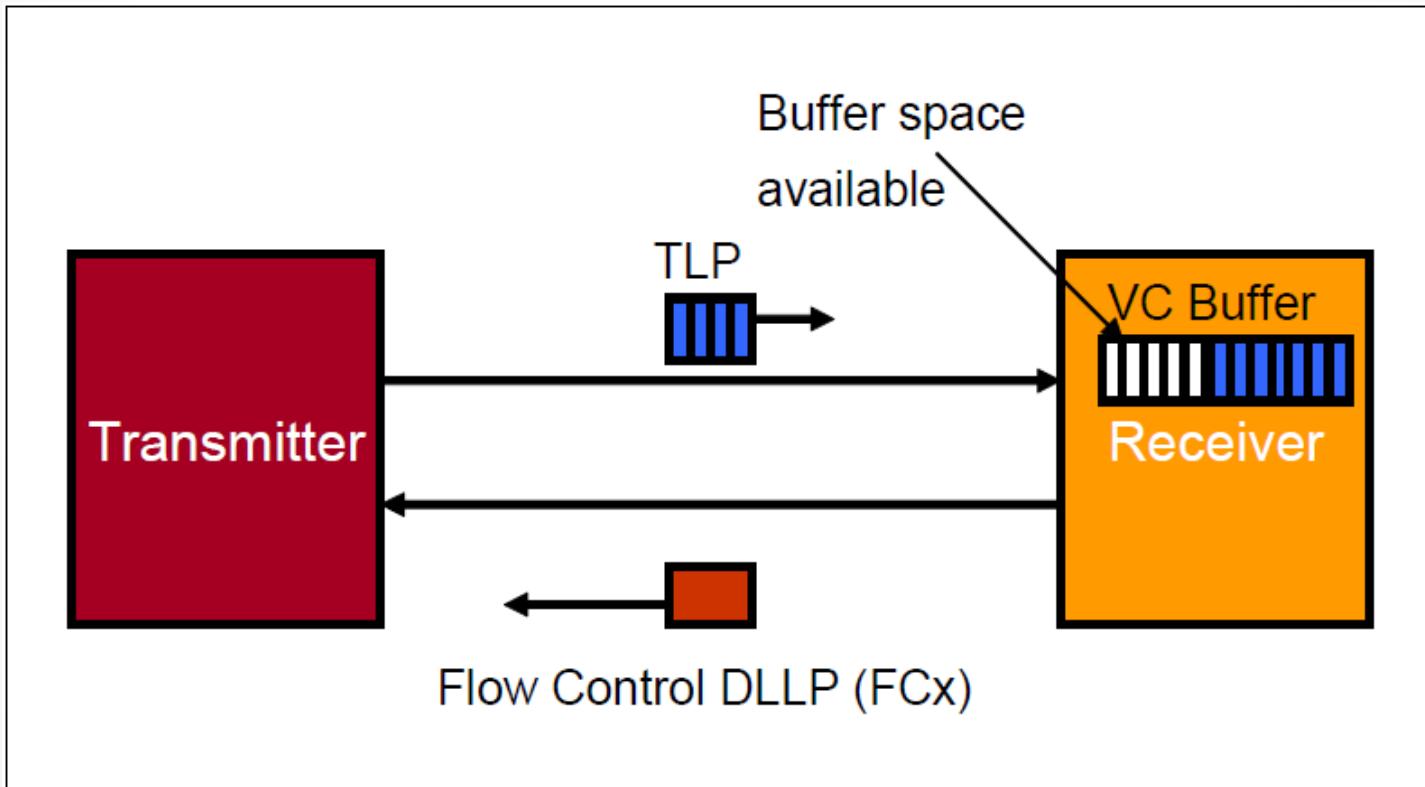
- Reliable transport of TLPs from one device to another across the link
- It's done by using DLL packets:
 - TLP acknowledgement
 - Flow control
 - Power Management

Transaction layer

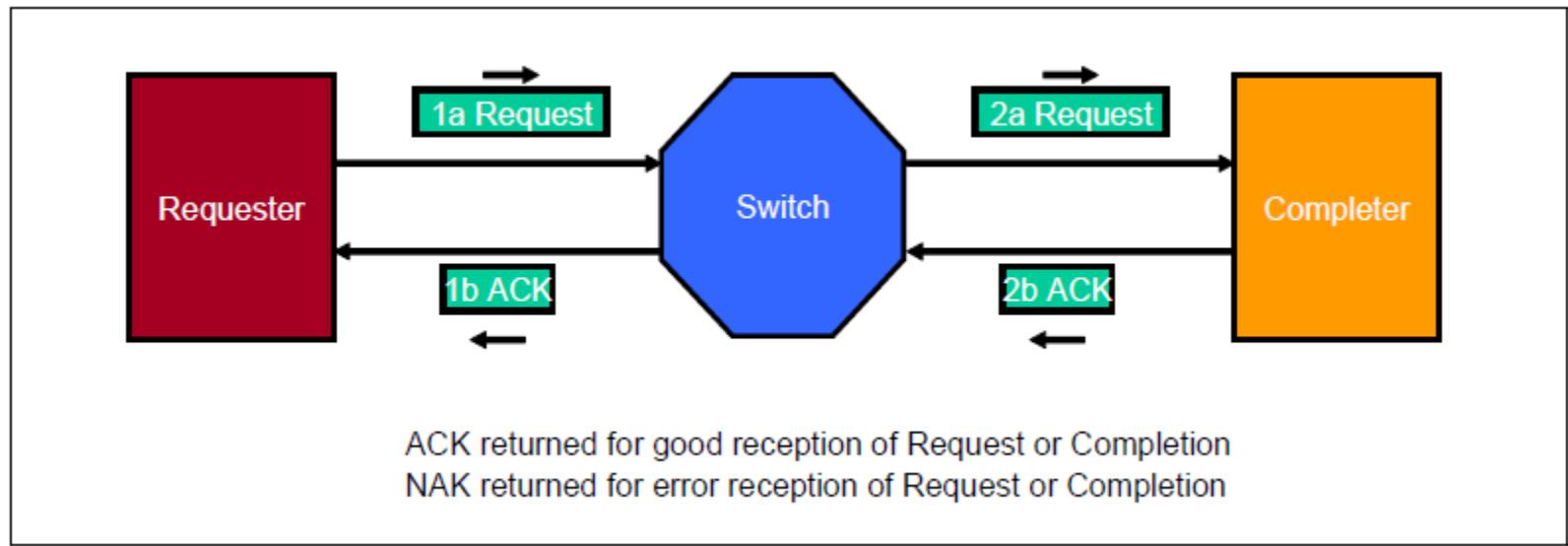
- It turns user application data or completion data into PCIe transaction – TLP
- Header + Payload + ECRC
- used in FPGAs IPs



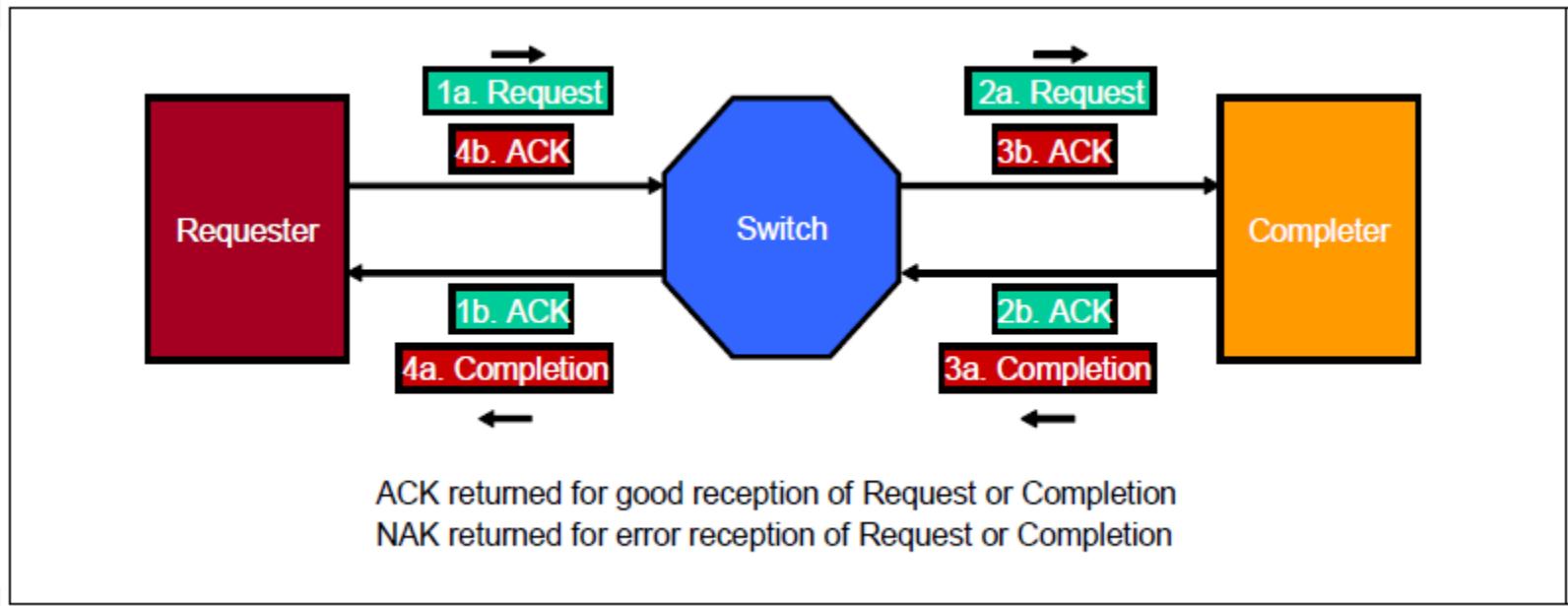
Flow control



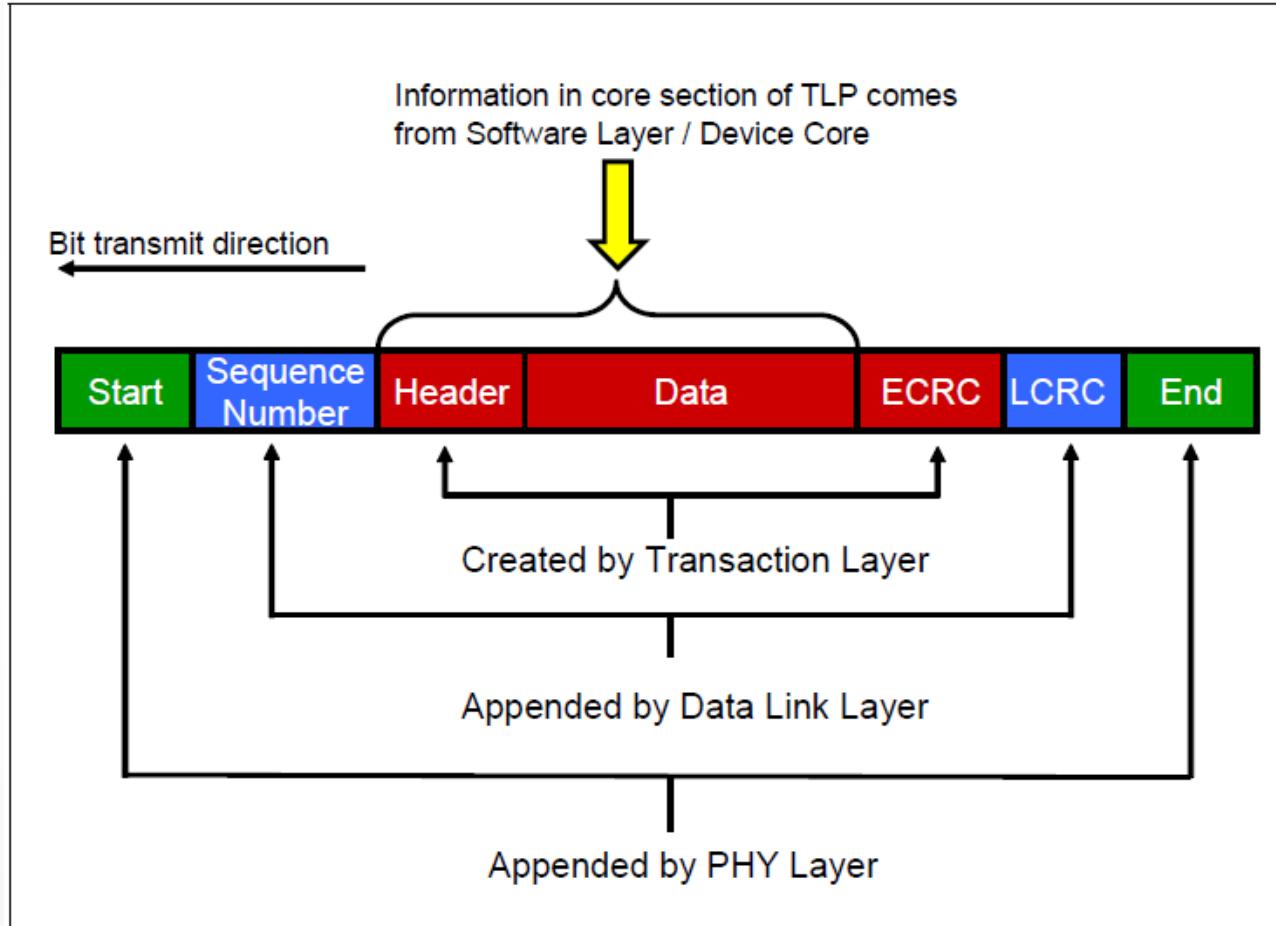
Flow control – posted transaction

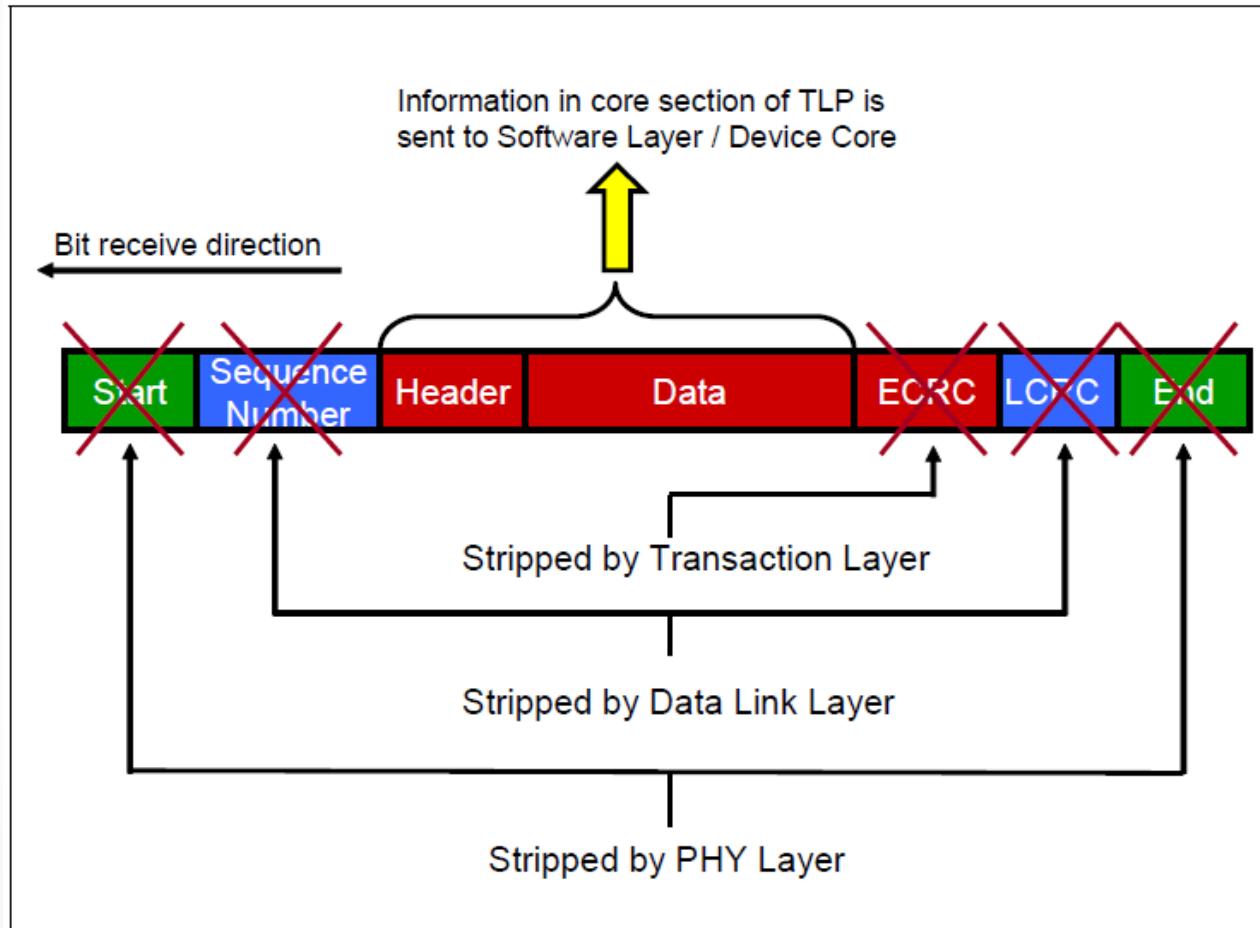


Flow control – non-posted transaction



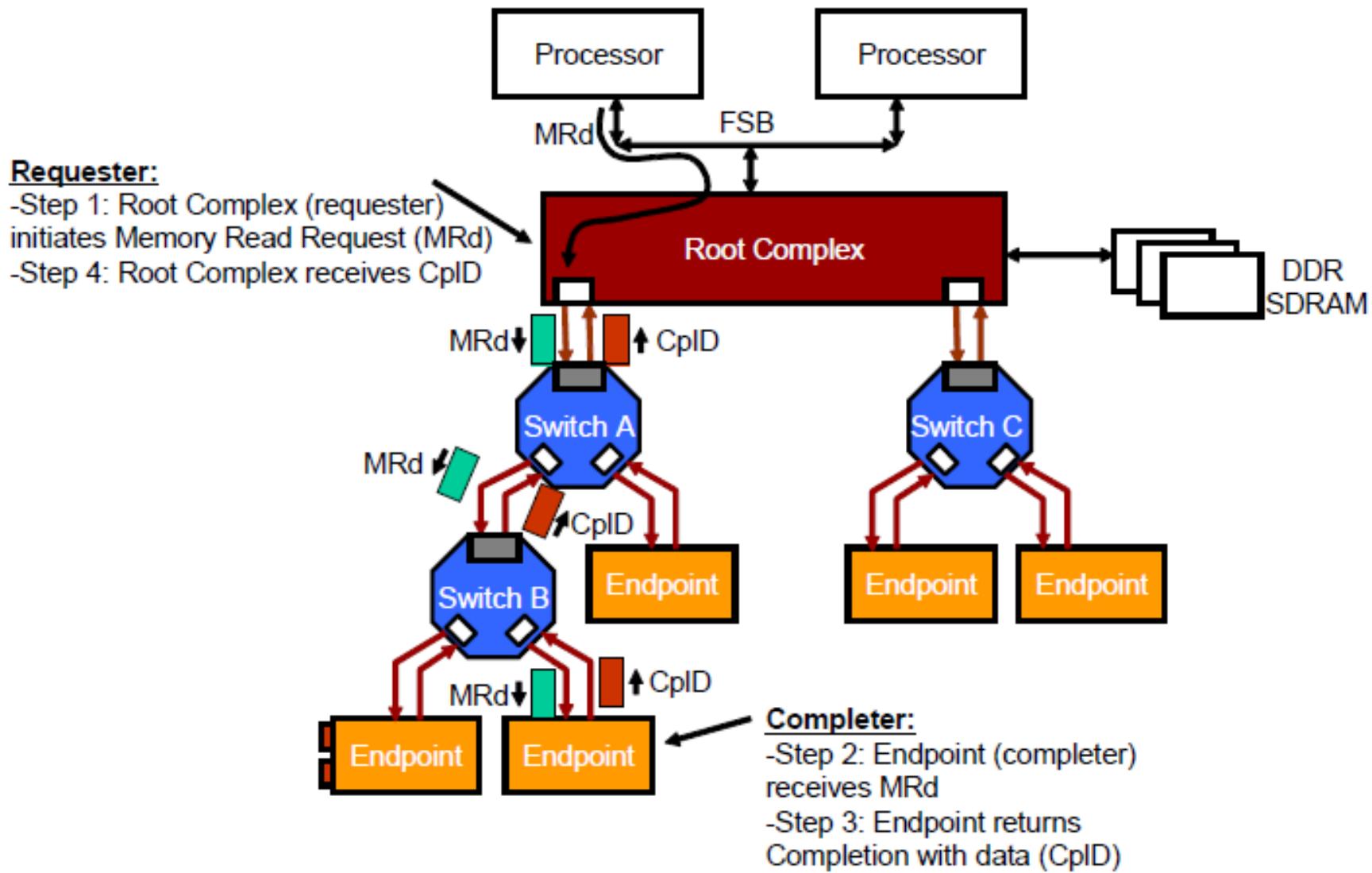
Building transaction



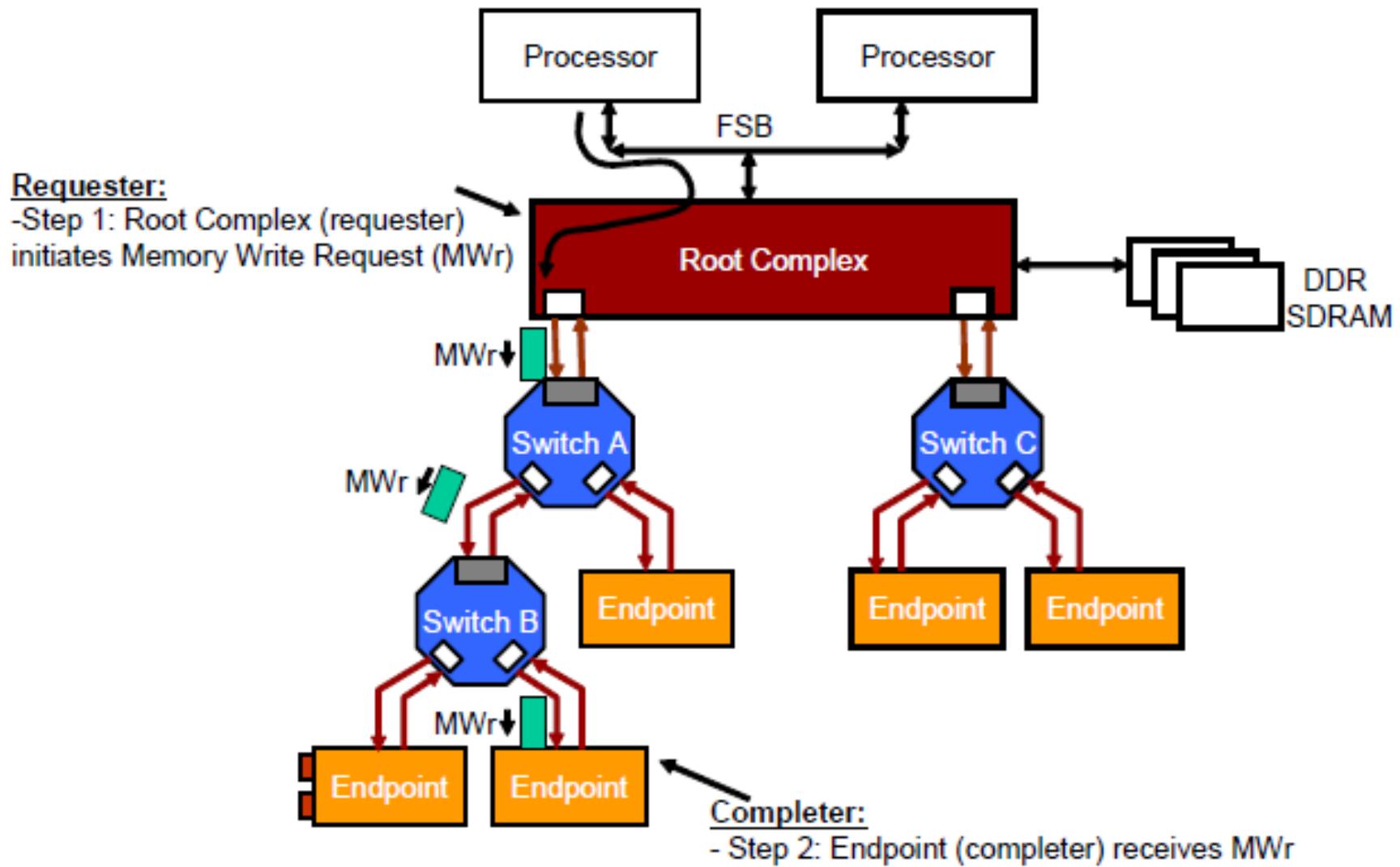


Example

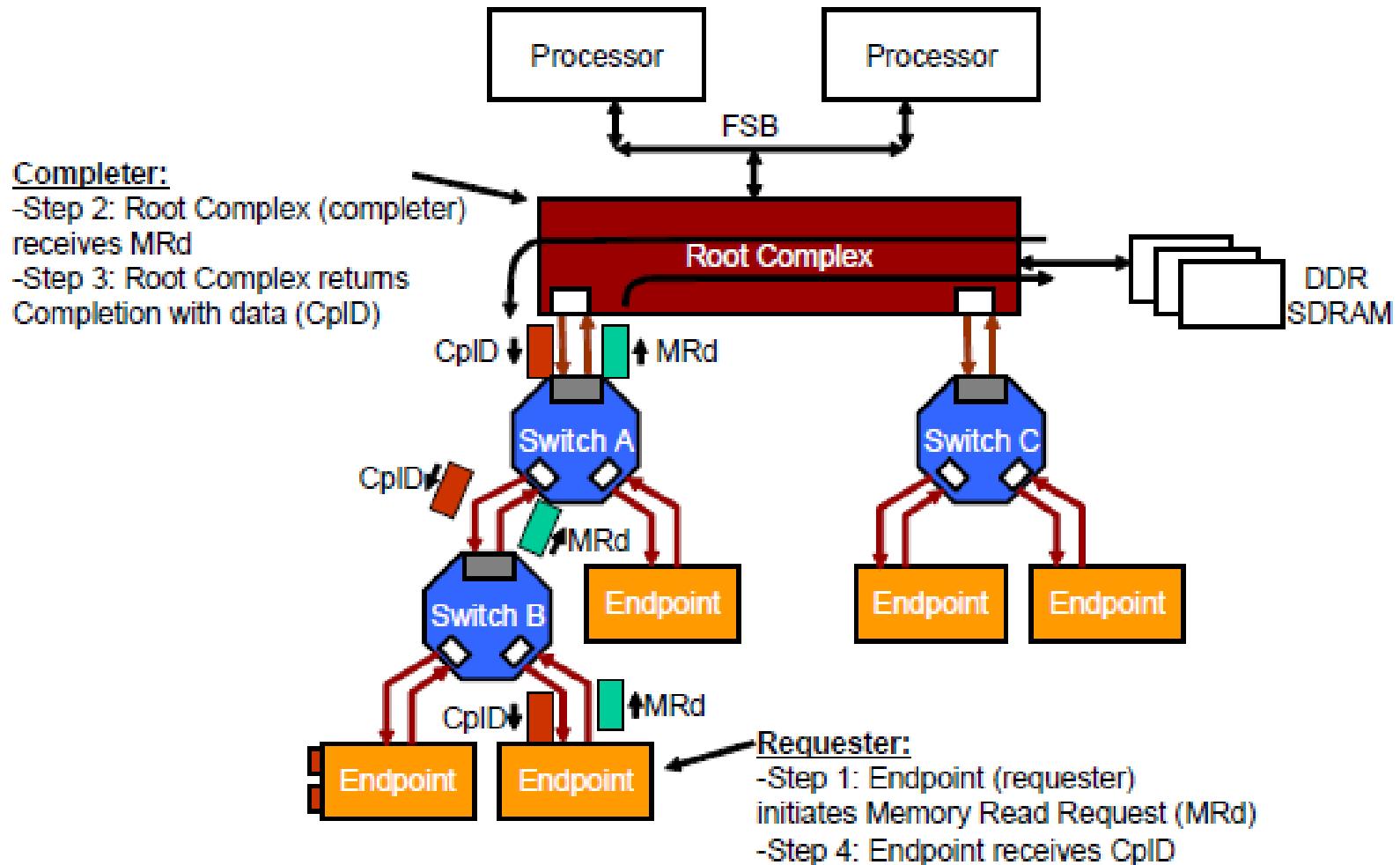
CPU MRd targeting an Endpoint



CPU MWr targeting Endpoint



Endpoint MRd targeting system memory

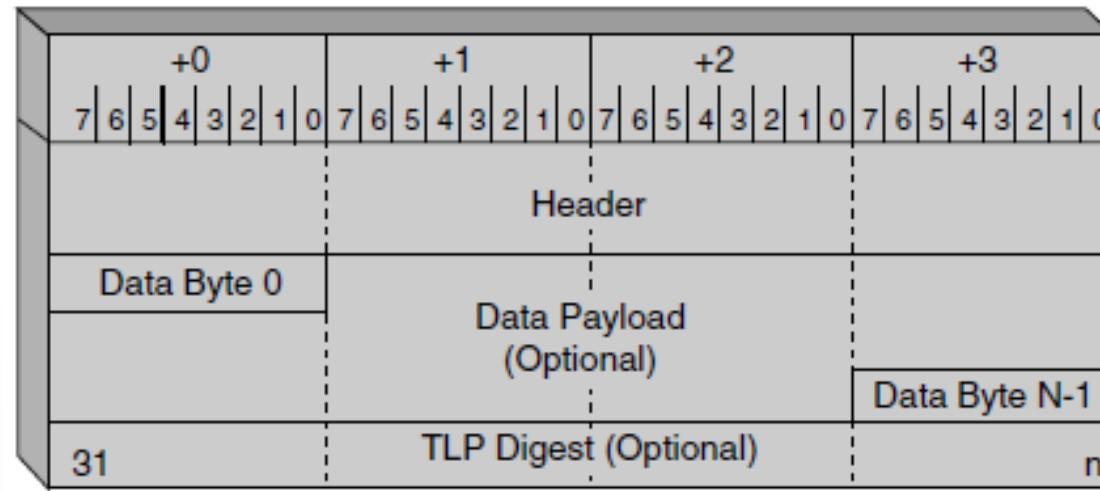


Packet constraints

- Maximum Payload Size (MPS)
 - default 128 Bytes
 - least denominator of all devices in the tree
- Maximum Read Request Size (MRRS)
 - Defined by RC
- Maximum Payload/ Read req. size 4 kB
 - defined by spec
 - No 4kB boundary crossing allowed
- Example: Intel x58 : MPS=256B, MRRS=512B

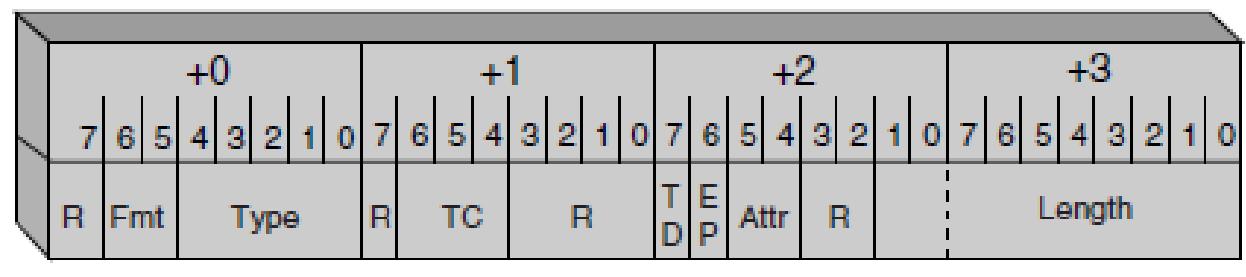
HEADER description

- Little endian
- 3DW or 4DW (Double Word – 4 bytes)



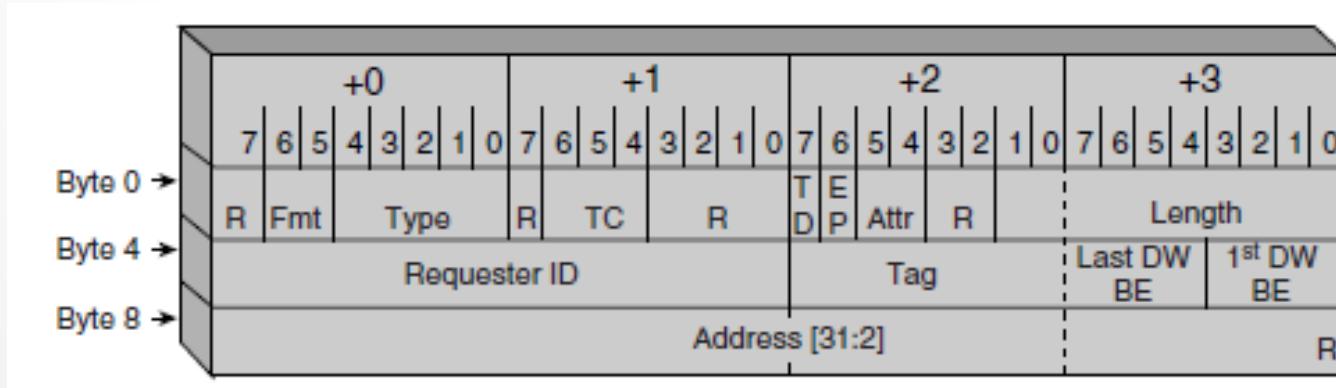
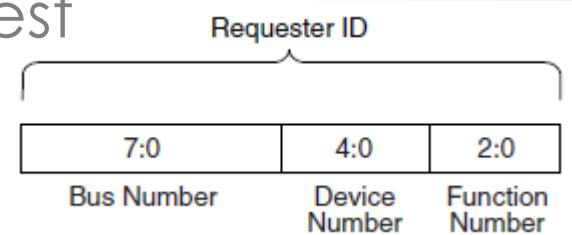
HEADER – base part

- Fmt – size of the header, is there payload ?
- Length – in DW
- EP – Poisoned
- TC – Traffic class
- TD – TLP digest – ECRC field
- Attr – status (success, aborted)



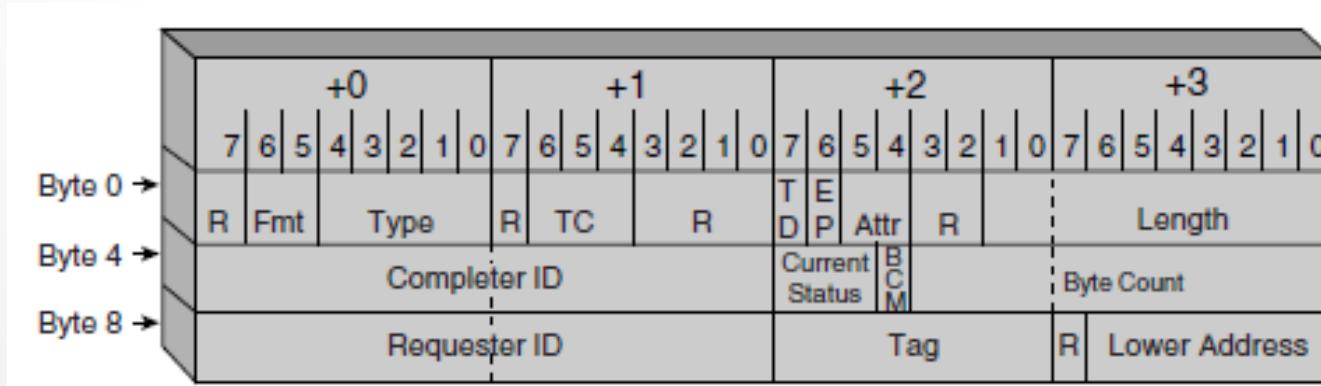
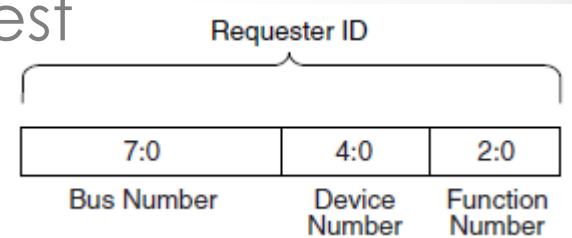
HEADER Memory Request

- TAG - Number of outstanding request
- Requester ID



HEADER Completion

- TAG - Number of outstanding request
- Requester ID

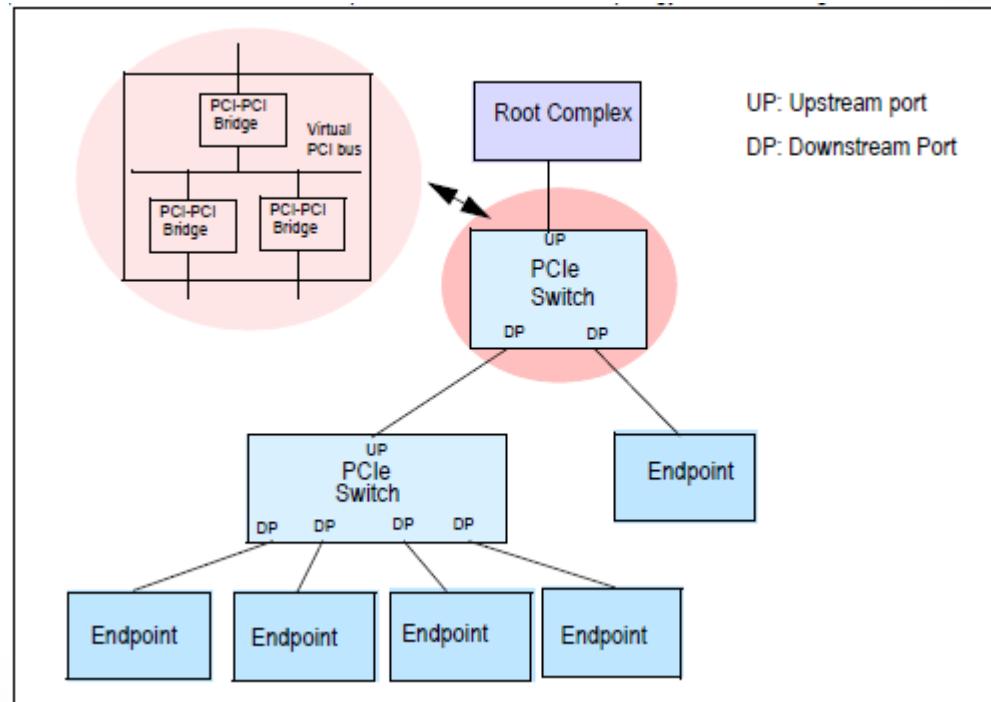


PCIe System Architecture

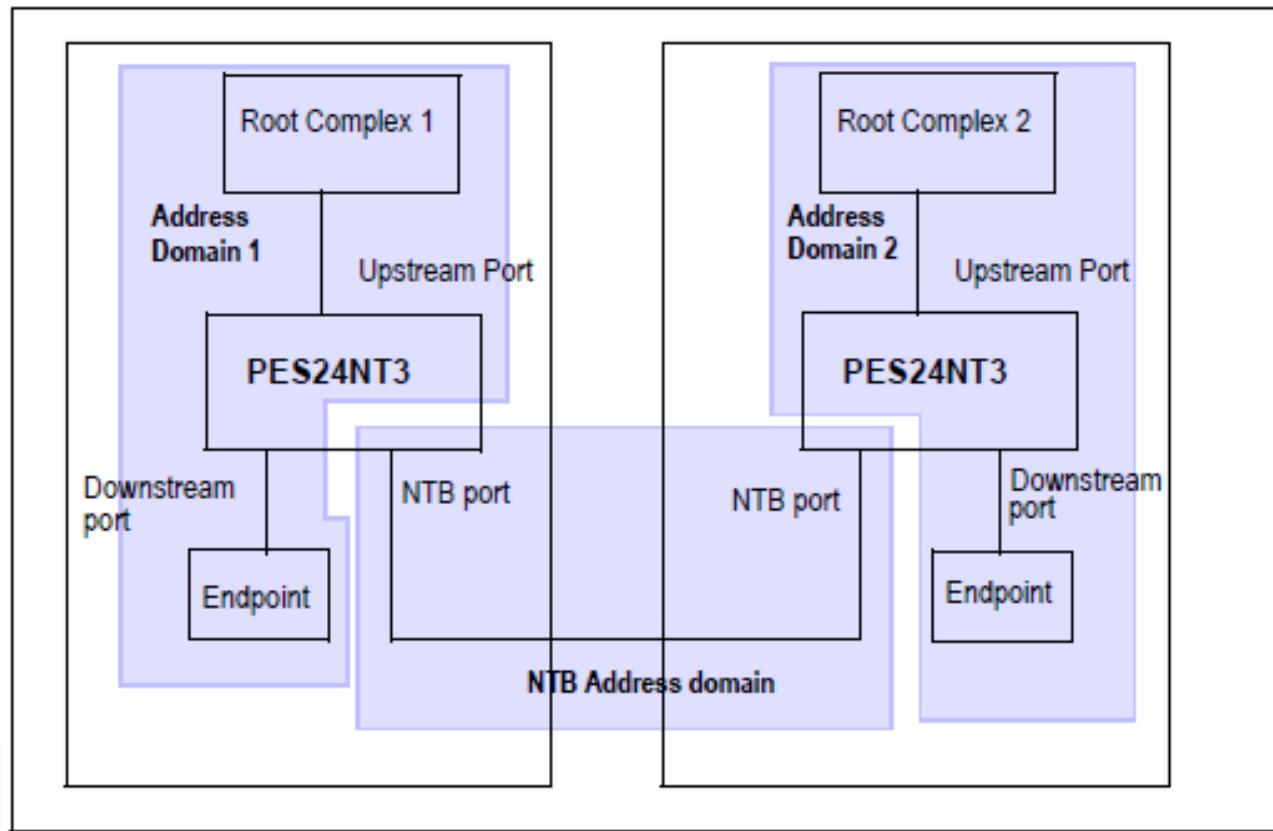
- Switches
 - Extend interconnection possibilities
 - DMA
 - Performance improvement functions
 - Non Transparent Bridging
- Extending distance
 - Bus re-drivers
 - Copper and optical cables

PCIe switches

- Non Transparent Bridging (NTB)
- Virtual Partitioning
- Multicasting
- DMA
- Failover



NTB + Virtual Partitioning



Cabling

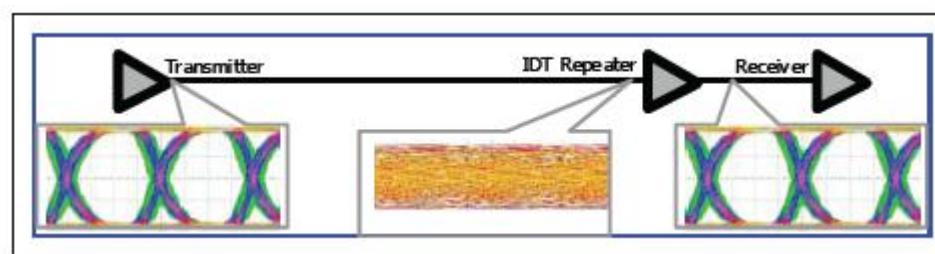
- Copper cables
- Optical cables
- Cable re-drivers(repeaters)



Image taken from www.ioxos.ch



<http://www.alpenio.com/products/pciex4.html>



www.idt.com

PCIe with FPGAs

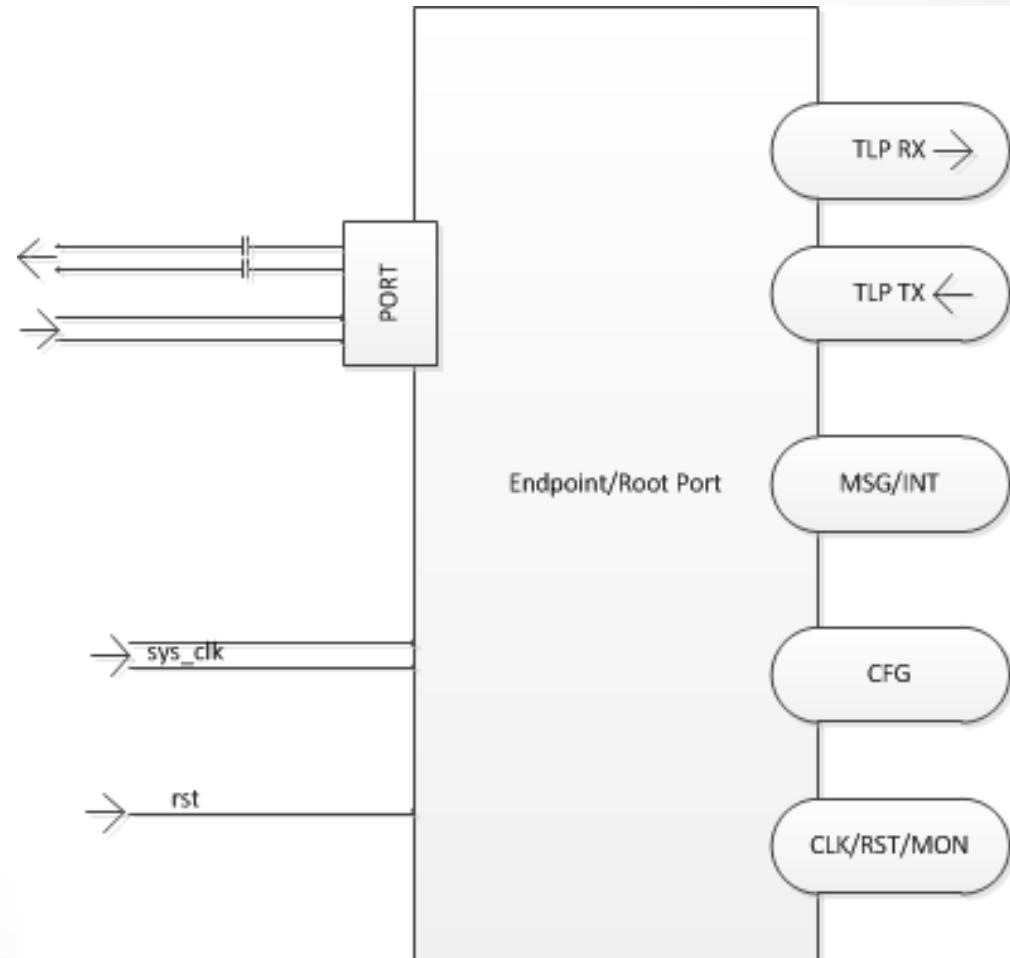
- Technology overview:
 - Hard IP – Altera and Xilinx
 - Soft IP – PLDA
 - External PHY – Gennum PCIe to local bus bridge
- Vendor documents – app notes, ref designs, Linux/Win device drivers
- Simulation – Endpoint/Root port

Xilinx Hard IP solution

- User backend protocol same for all devices
 - Spartan – 6
 - Virtex – 5
 - Virtex – 6
 - Virtex – 7
- Xilinx Local Link (LL) Protocol and ARM AXI
- For new designs: use AXI
- Most of the Xilinx PCIe app notes uses LL

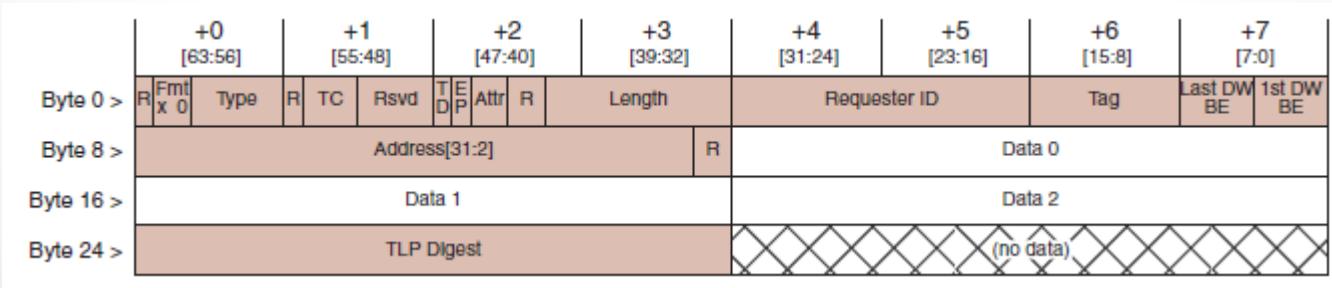
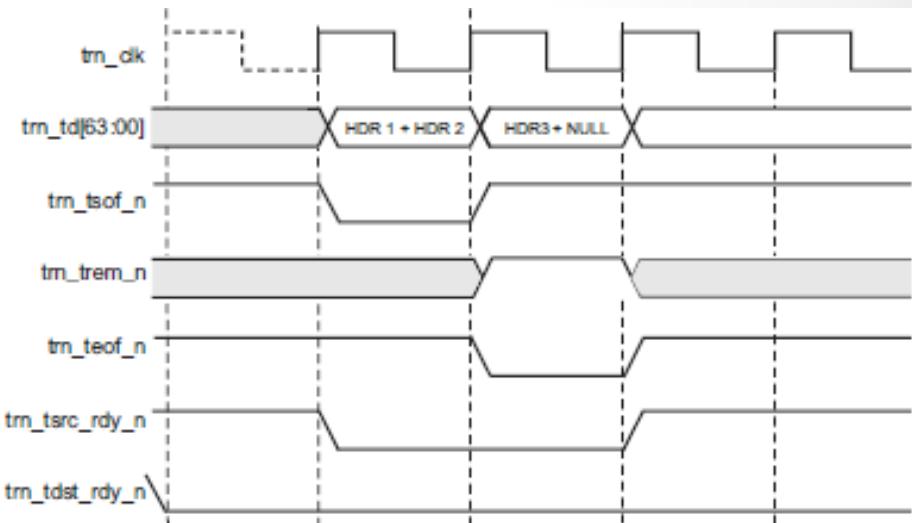
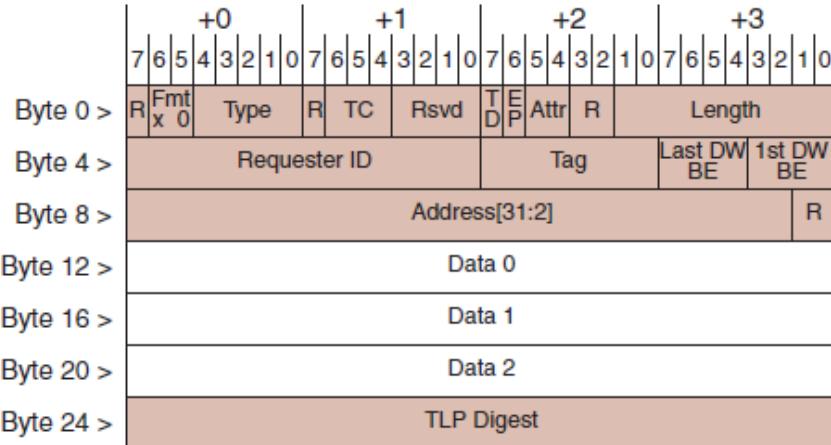
Xilinx Hard IP interface

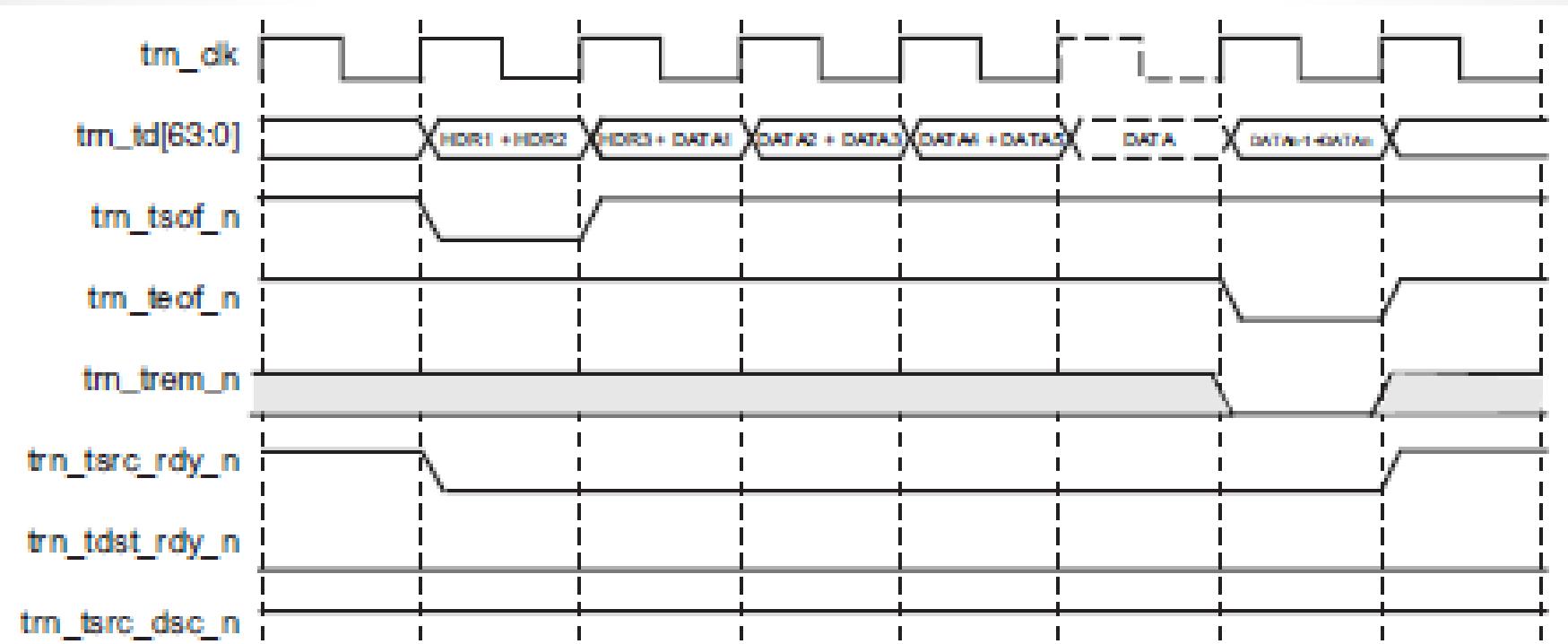
- External world: gt, clk, rst – (example x1 needs 7 wires)
- CLK/RST/Monitoring
- TLP TX if
- TLP RX if
- CFG if
- MSG/INT if



PCIe LL protocol

- TLP packets are mapped on 32/64/128 bit TRN buses

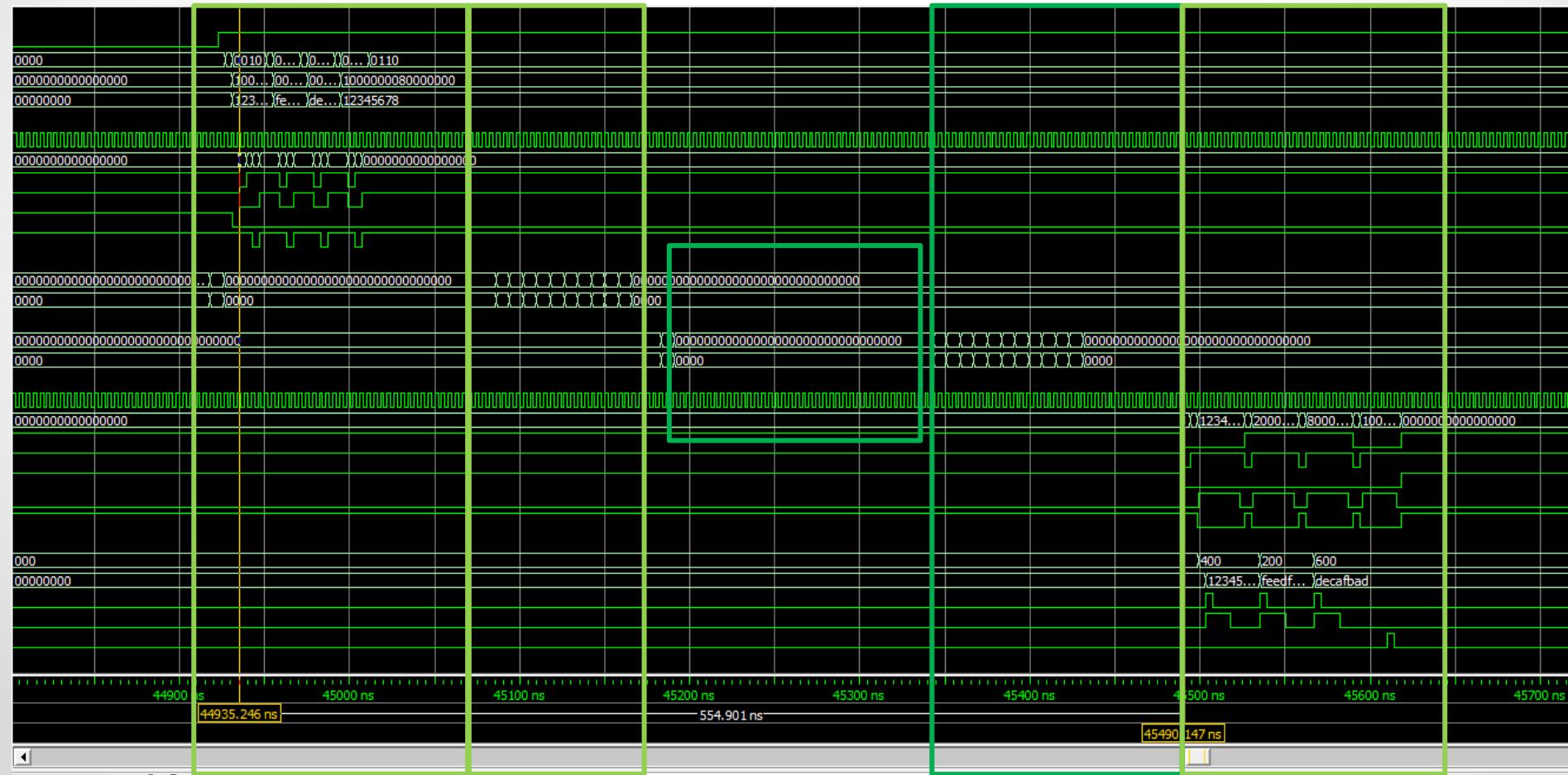




Xilinx simulation

RP <-> EP

- Gen1, x8, Scrambling disabled in CORE Gen

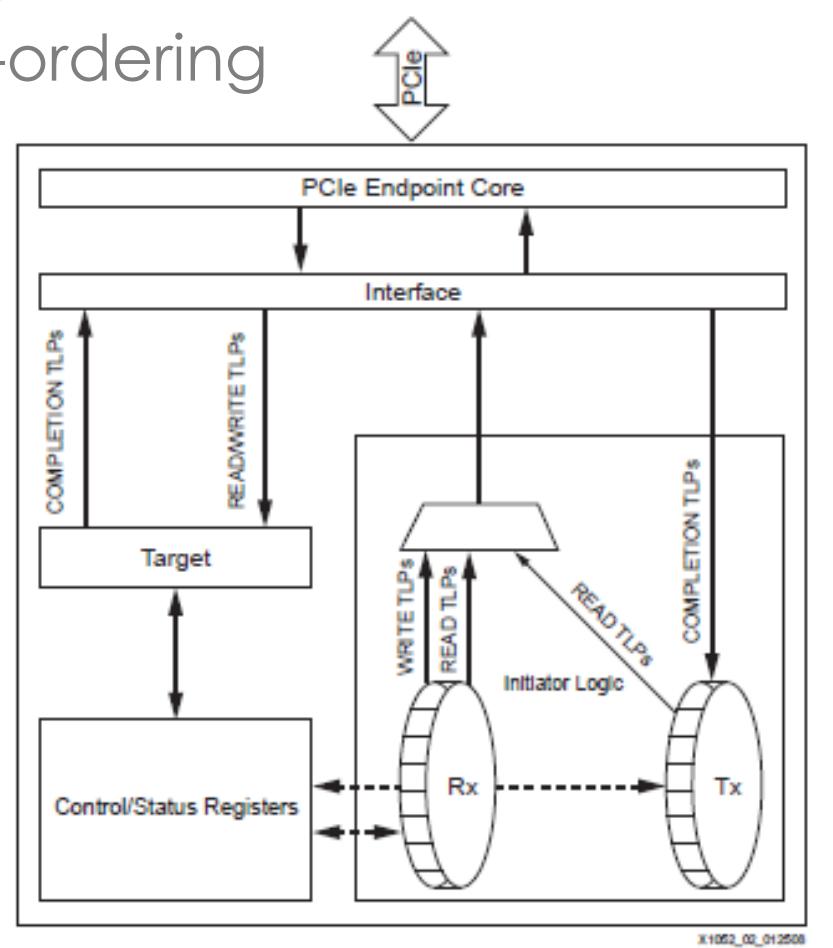


How to design with Xilinx PCIe Hard IP

- Application notes
- Reference designs
- CORE Gen Programmable IO (PIO)
hardware/simulation examples

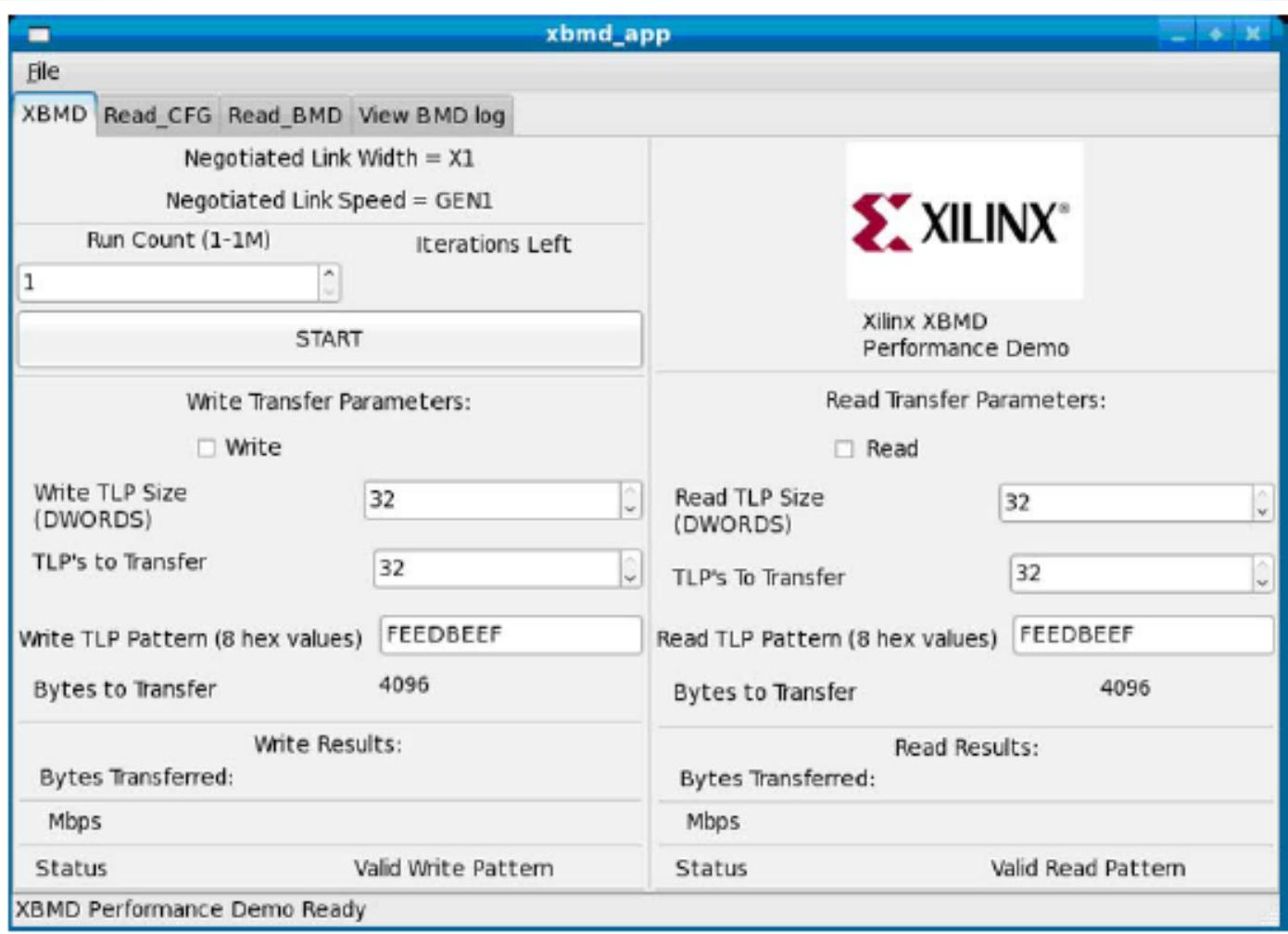
XAPP 1052

- Block DMA in Streaming mode
- No CplID transaction re-ordering



XAPP 1052

- GUI for Win(VisualBasic)
- GUI for Linux (Glade)
- Driver for Win/Linux

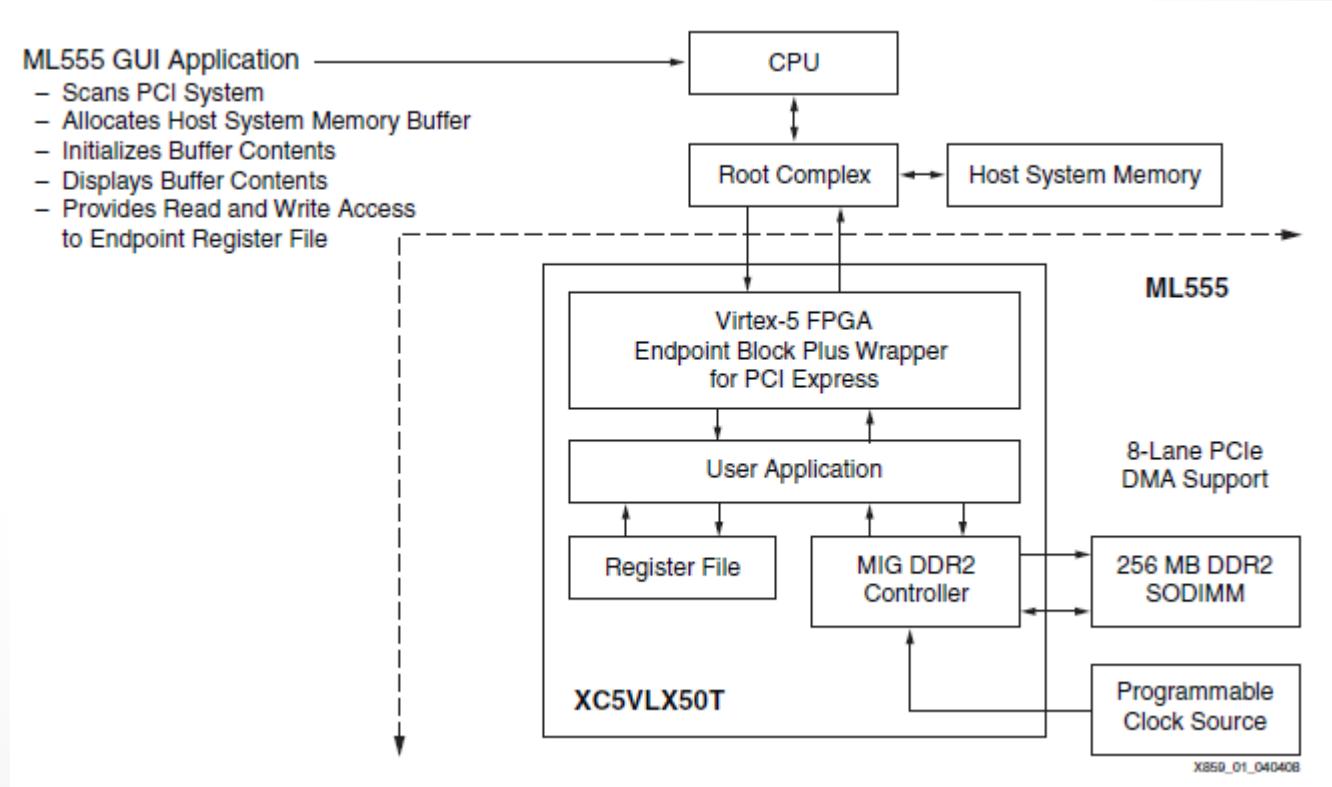


XAPP1052 – performance

- Intel Nehalem 5540 platform
- Fedora 14, 2.35. PAE kernel
- Gen1, x4, PCIe LeCroy analyser
- DMA config
 - Host configures (MWr) DMA engine – around 370 ns between 1DW writes
 - Host checks DMA status: MRd (1DW) to CplD (1DW) response time – around 40 ns
- DMA operation:
 - DMA MRd(1st) -> CplD response time around 2.76 µs
 - DMA MRd(8th) -> CplD response time around 3.82 µs
 - DMA MWr -> around 750-800 MB/s (Gen1,

XAPP 859

- Block DMA: Host <-> DDR2
- Jungs Win device driver
- C# GUI



Xilinx DMA Initiator Design Platform for XAPP859

[Run Demo](#)[Run Read DMA](#)[Run Write DMA](#)[FullDuplex DMA](#)[NL555 Activity Log](#)[Exit](#)

Read DMA Setup

Transfer Size (bytes)

- 128 256 512 1K 2K 4K 8K
 16K 32K 64K 128K 256K 512K 1M

Number of Transfers

- 1 25 50 75 100

Write DMA Setup

Transfer Size (bytes)

- 128 256 512 1K 2K 4K 8K
 16K 32K 64K 128K 256K 512K 1M

Number of Transfers

- 1 25 50 75 100

Host Memory Buffer

Base Address: 00100000

Starting at
Offset Below:

Print 1K DWORD

[Fill Buffer](#) 0x Incrementing
PatternOffset Address:
(Max = 0xFFFF)

0000

Buffer Offsets

Read DMA

Host PC Source:

0

NL555 DDR2 Dest:

0

Write DMA

NL555 DDR2 Source:

0

Host PC Dest:

0

PCIe Config Space

Max Read Request Size = 512 bytes

Max Payload Size = 128 bytes

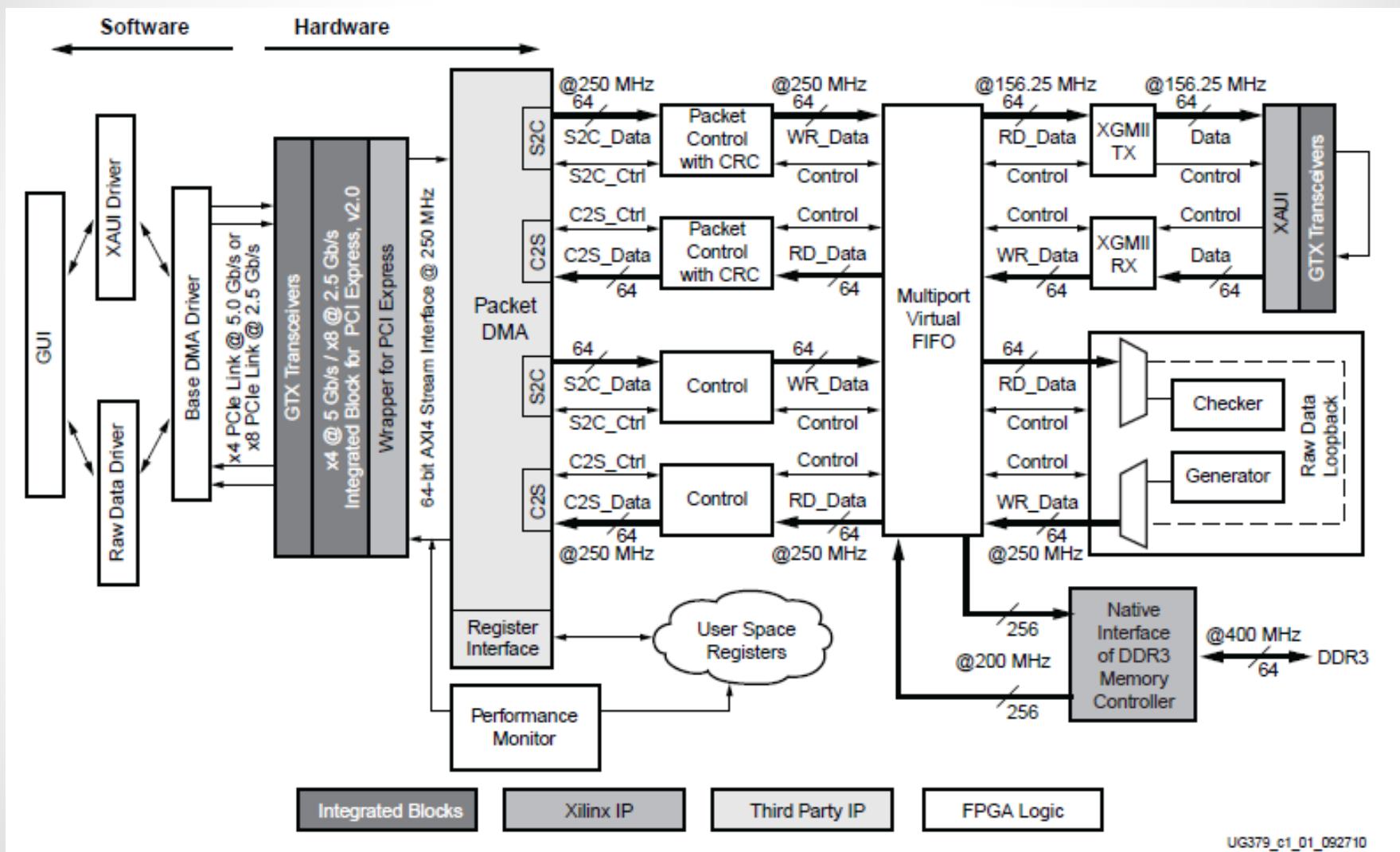
RCB = 64 bytes

Link Width = 8 Lanes

[Compare Buffer](#)[Display RegFile](#)[Reset To Defaults](#)[Clear](#)[Print](#)

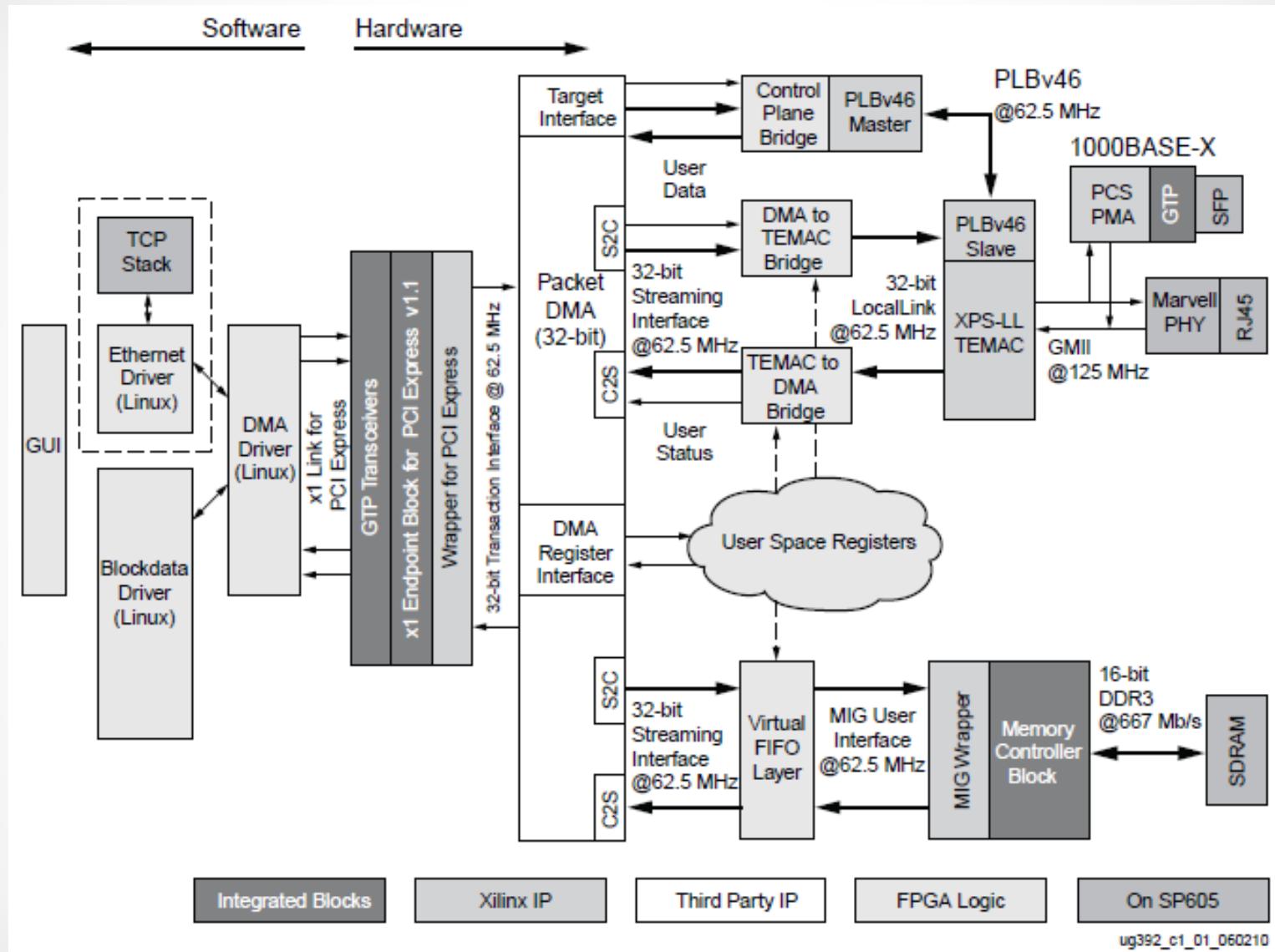
Xilinx V6 Connectivity Kit

- PCIe to XAUI
- PCIe to parallel loopback
- VirtualFIFO based on DDR3 (MIG, SODIMM)
- Northwest Logic User Backend IP – Packet (SG)
DMA



Xilinx S6 Connectivity Kit

- PCIe to 1 Gb Eth
- PCIe to parallel loopback
- VirtualFIFO based on DDR3 (MIG, Component)
- Northwest Logic User Backend – Packet (SG) DMA

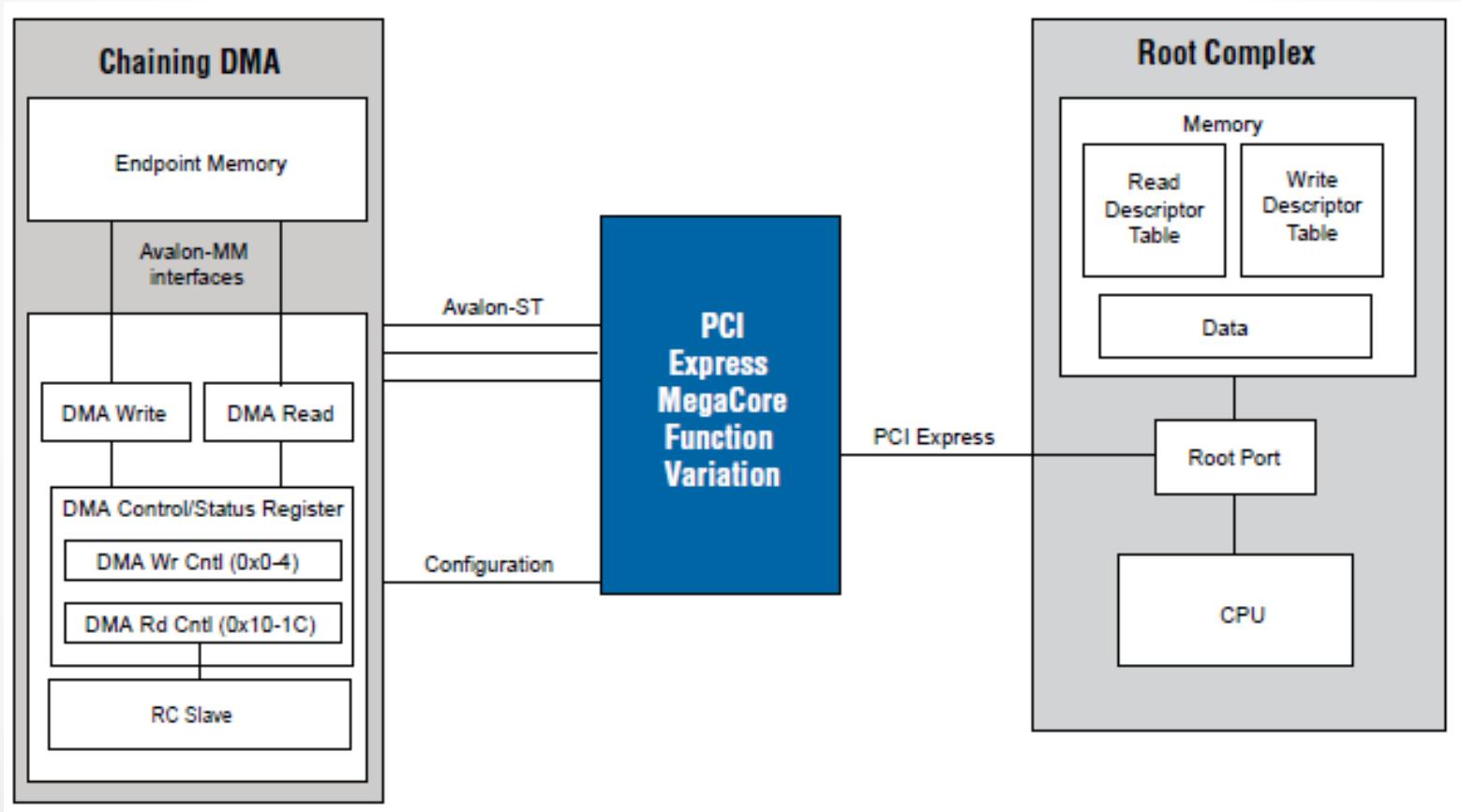


Altera Hard IP solution

- Target devices:
 - Cyclone IV GX
 - Arria I/II GX
 - Stratix II/IV GX
- Similar to Xilinx in terms of user interface – TLP over Avalon ST or User application with Avalon MM
 - ST – streaming mode, for high performance designs
 - MM – memory mapped, for SOPC builder, lower performance
- CvPCIe – FPGA reconfiguration over PCIe
 - I/O and PCIe programmed faster than the rest of the core

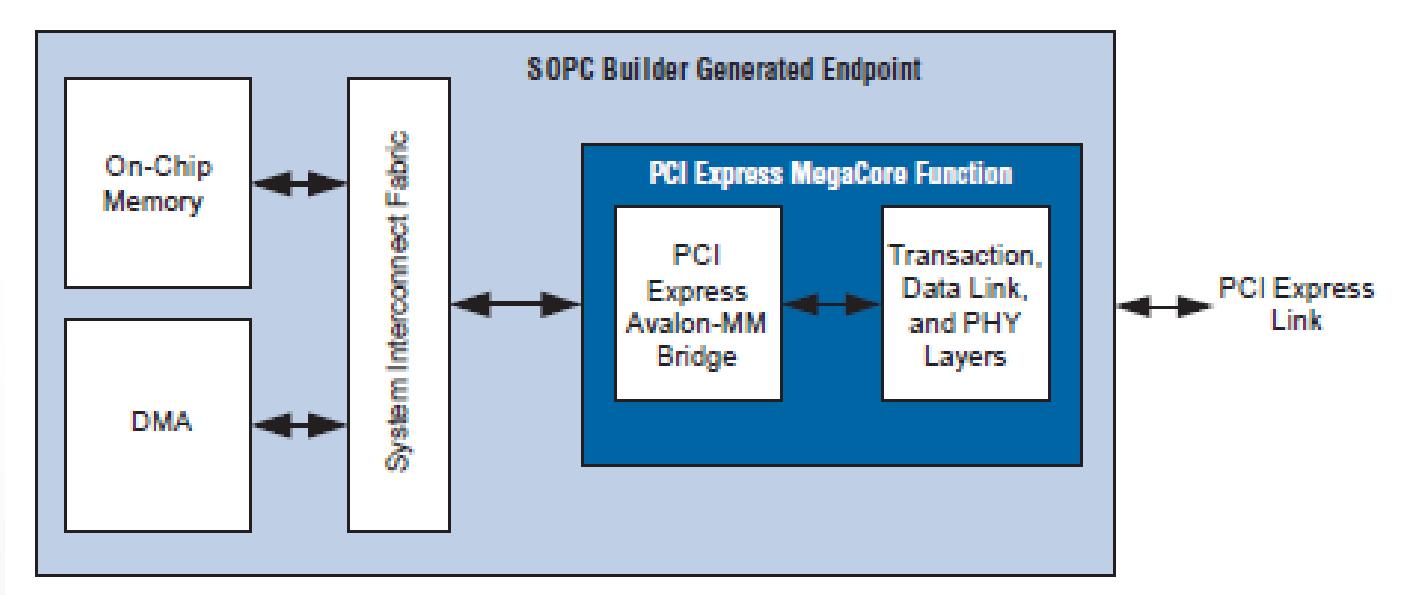
Altera Megacore Reference Designs

- Endpoint Reference Design
 - PCIe High Performance Reference Design (AN456) – Chained DMA, uses internal RAM, binary win driver
 - PCIe to External Memory Reference Design (AN431) – Chained DMA, uses DDR2/DDR3, binary win driver
- Root Port Reference Design
- SOPC PIO
- Chained DMA documentation
 - also Linux device driver available
- BFM documentation
 - Extensive simulation with Bus Functional Models

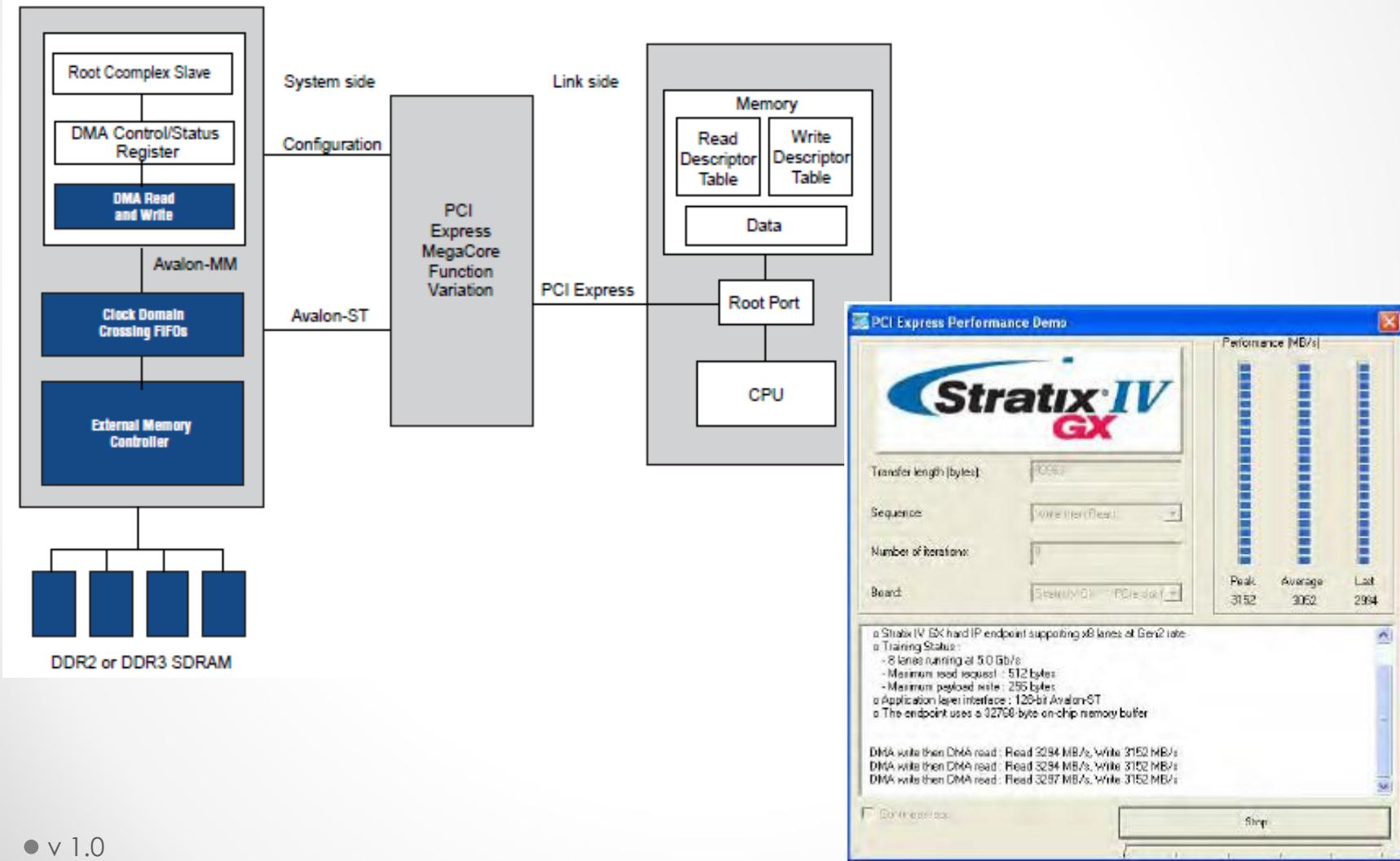


SOPC Based Design

- SOPC Builder Based
- Gen 1, x4
- DMA
- Sim and HW



AN431 – PCIe to DDR3



PLDA PCIe IPs

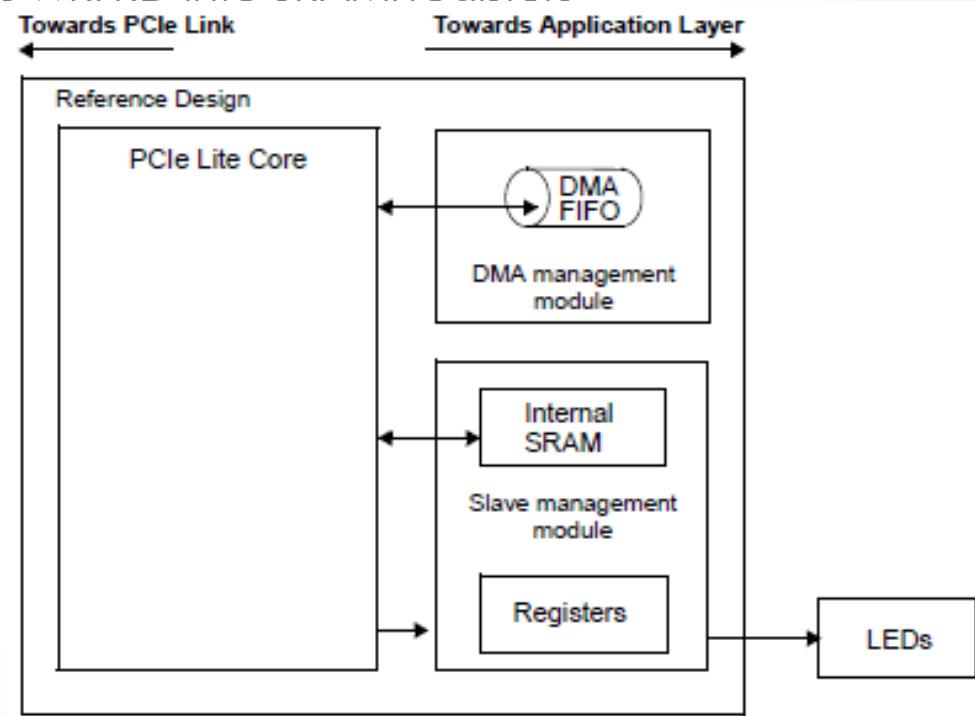
- XpressLite
 - currently available at CERN
 - Soft IP, Gen1 Endpoint only, x1/x2/x4
 - Stratix GX, Stratix II GX, and Arria GX support
 - No S4GX, C4GX and A2GX Hard IP support
- EZDMA2 Altera/Xilinx
 - Support Hard IP inside Altera: Cyclone IV GX, Arria II GX, and Stratix IV GX
 - Hard IP inside Xilinx: Virtex-5/6, Spartan-6
 - Same user/DMA interface as XpressLite
- XpressRich – rich version
 - Are you rich ?
- Northwest Logic ?

PLDA XpressLite

- Stratix GX, Stratix II GX, and Arria GX support only
 - No S4GX, C4GX and A2GX Hard IP support
- Generated with JAVA GUI: Windows/Linux
- Synthesis: single VHDL/Verilog encrypted file
- ModelSim: pre-compiled lib (Win/Linux)
- Ncsim: protected lib (Linux)
- Testbench: RP emulation
- Device drivers, API, tools (C++ source available)

PLDA XpressLite

- Maximum 8 DMA channels with Scatter Gather
- Reference design:
 - PCIe Lite – Endpoint only
 - Single DMA engine – C2S(WR) + S2C(RD)
 - Single target module – accepts WR/RD into SRAM/registers

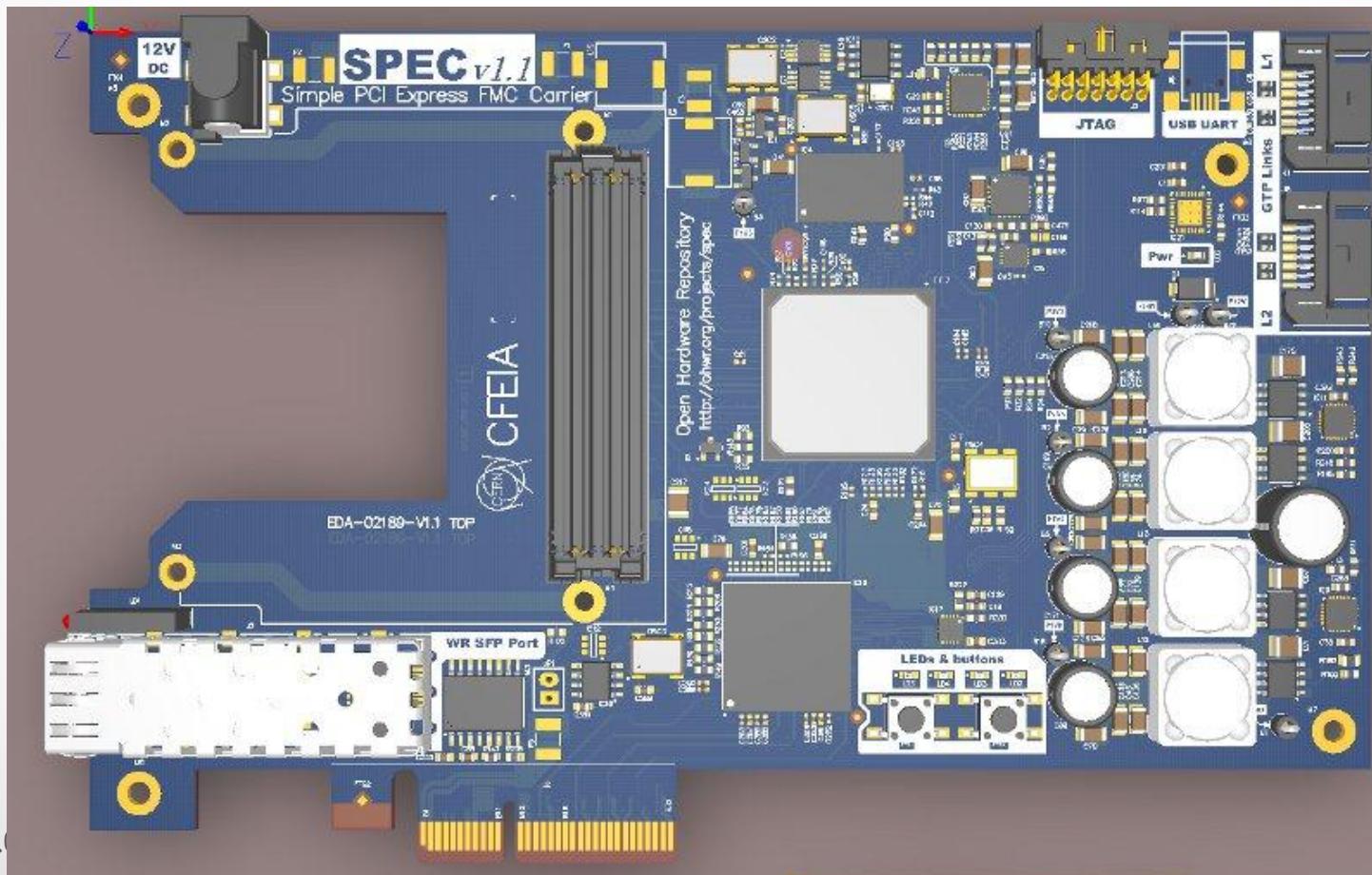


External PCIe chips - Gennum

- TLP interface with simple framing signalling
- FPGA serial programming
 - FPGA can be reprogrammed without affecting PCIe link
- GPIO interface/Interrupts
- IP (with DMA) provided for Altera and Xilinx
- Device drivers and Software DK provided
- Already used at CERN:
 - Open source IP for Xilinx device developed by CERN group
 - Wishbone
 - SG DMA
 - device driver
 - More info www.ohwr.org

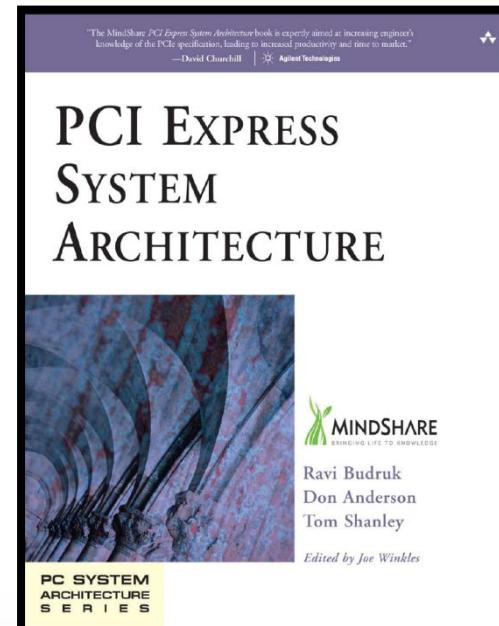
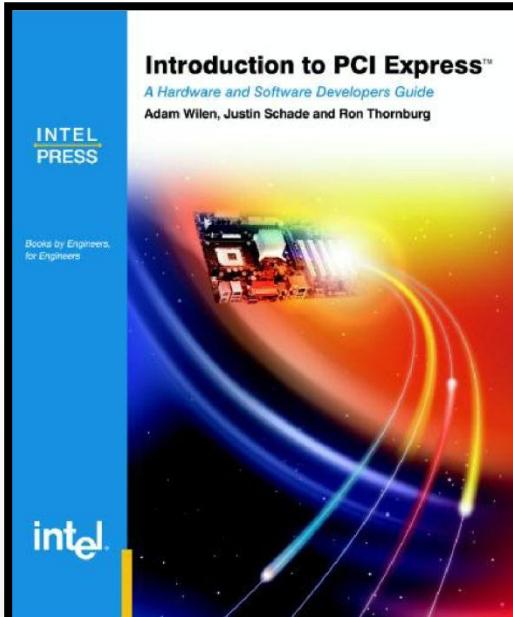
Gennum PHY + Spartan6

- <http://www.ohwr.org/projects/spec/wiki>
- Open source IP, SG DMA, device driver



More information

- Books:
 - Introduction to PCI Express – CERN Library (hardcopy)
 - PCI Express standards – CERN Library – CDS.CERN.CH
 - PCI Express System Architecture – mindshare.com (ebook+ hardcopy)



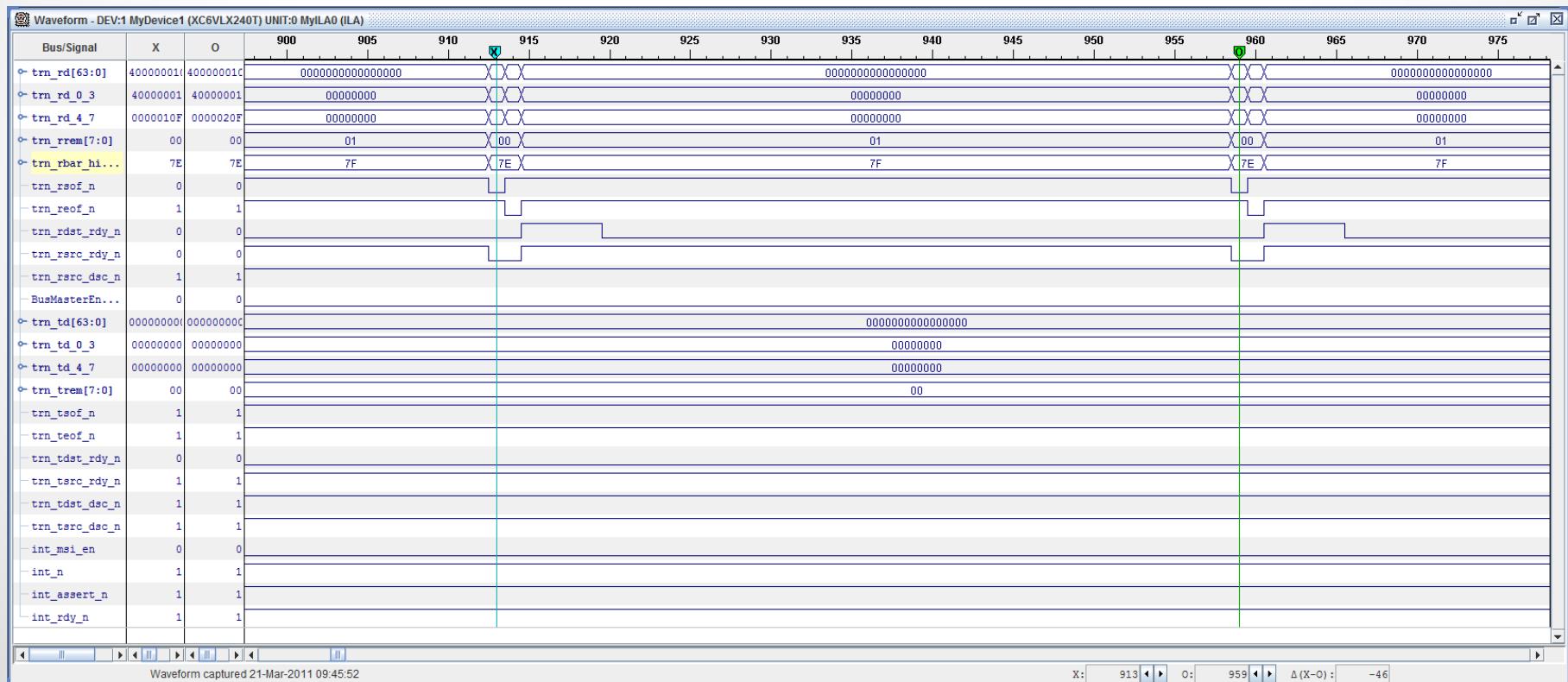
eda.support@cern.ch

- PCIe demos available on request
- IDT PCIe Switch dev. kit. coming soon
- Evaluating EZDMA2 for Xilinx.

Extras

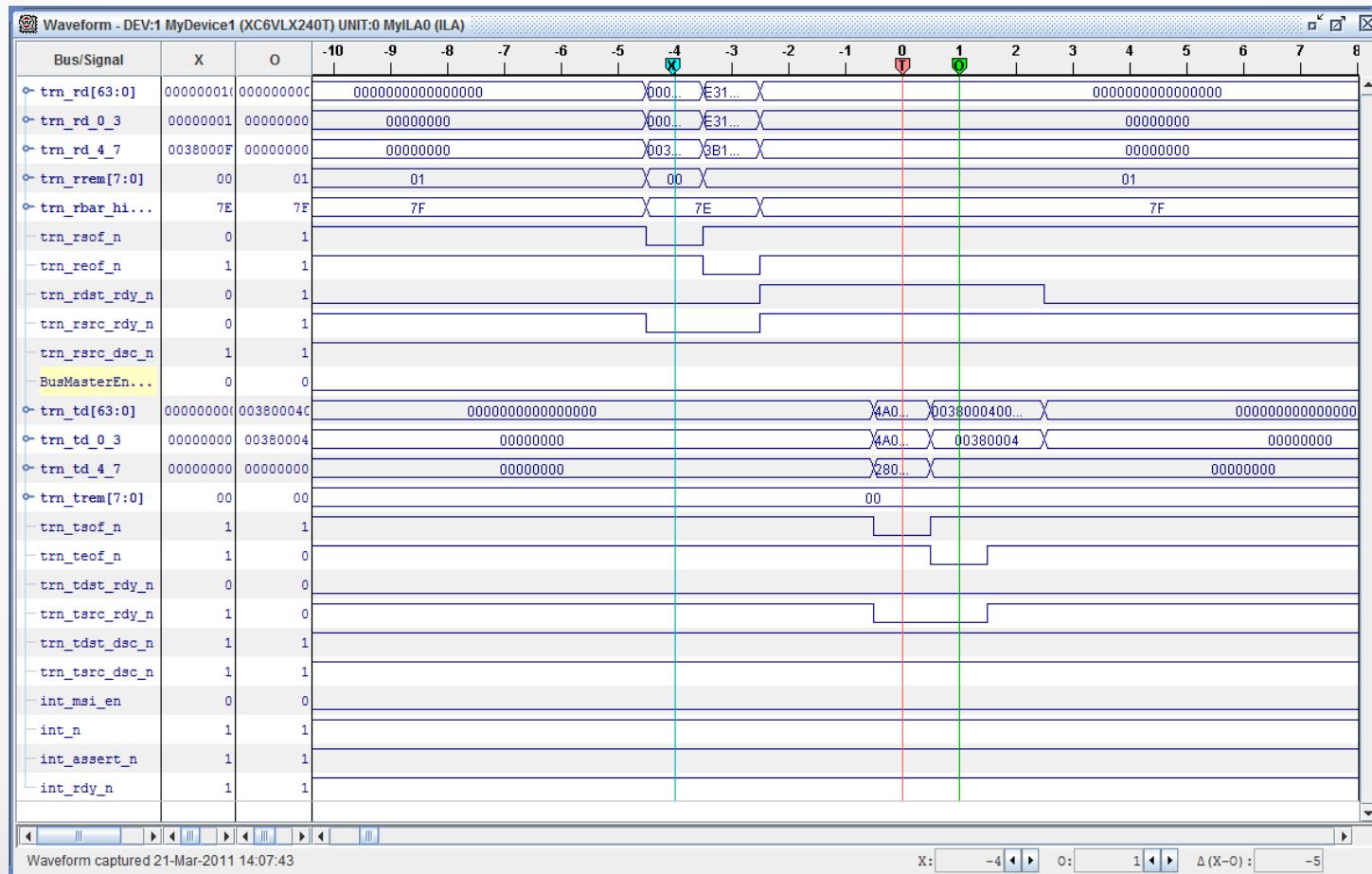
XAPP1052 DMA Config WR

- Host configures (MWr) DMA engine – around 370 ns between 1DW writes



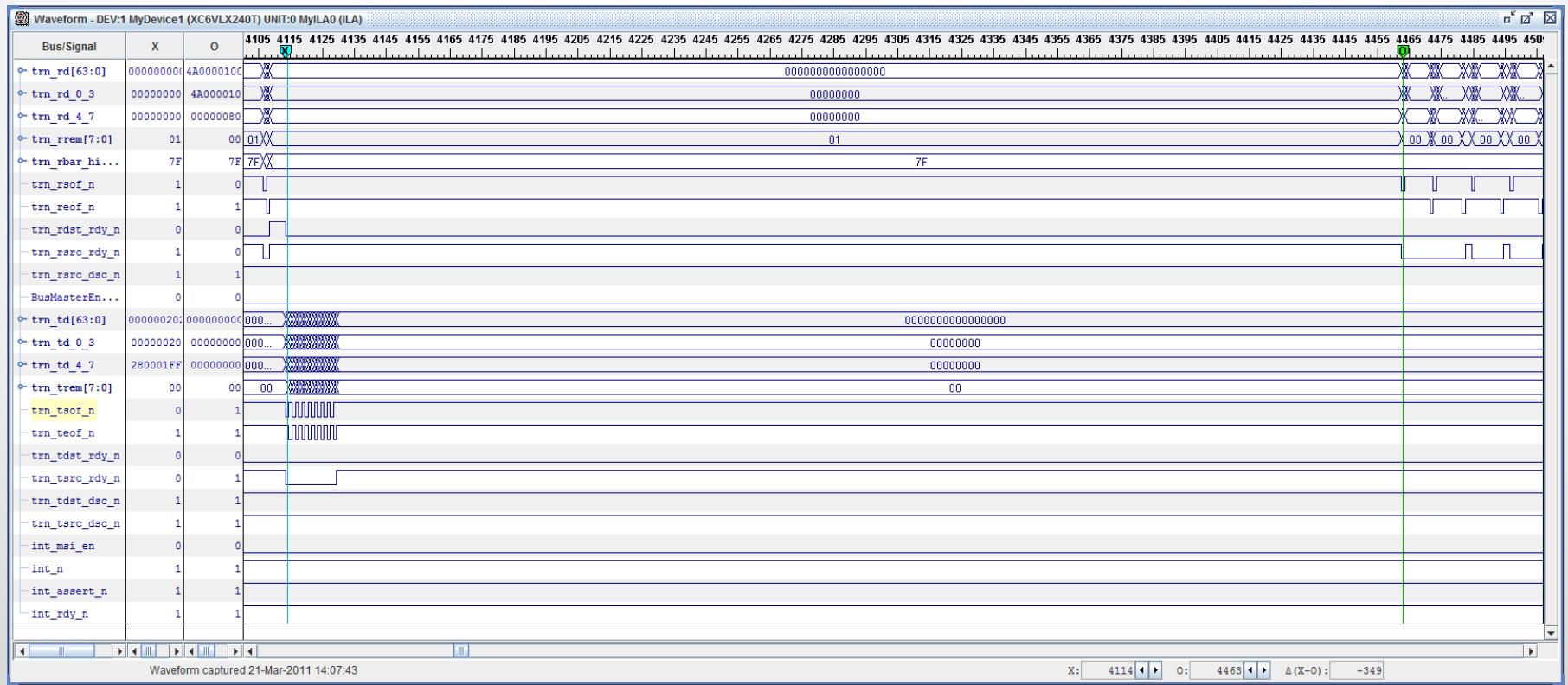
XAPP 1052 DMA Config RD

- MRd (1DW) to CplD (1DW) – around 40 ns



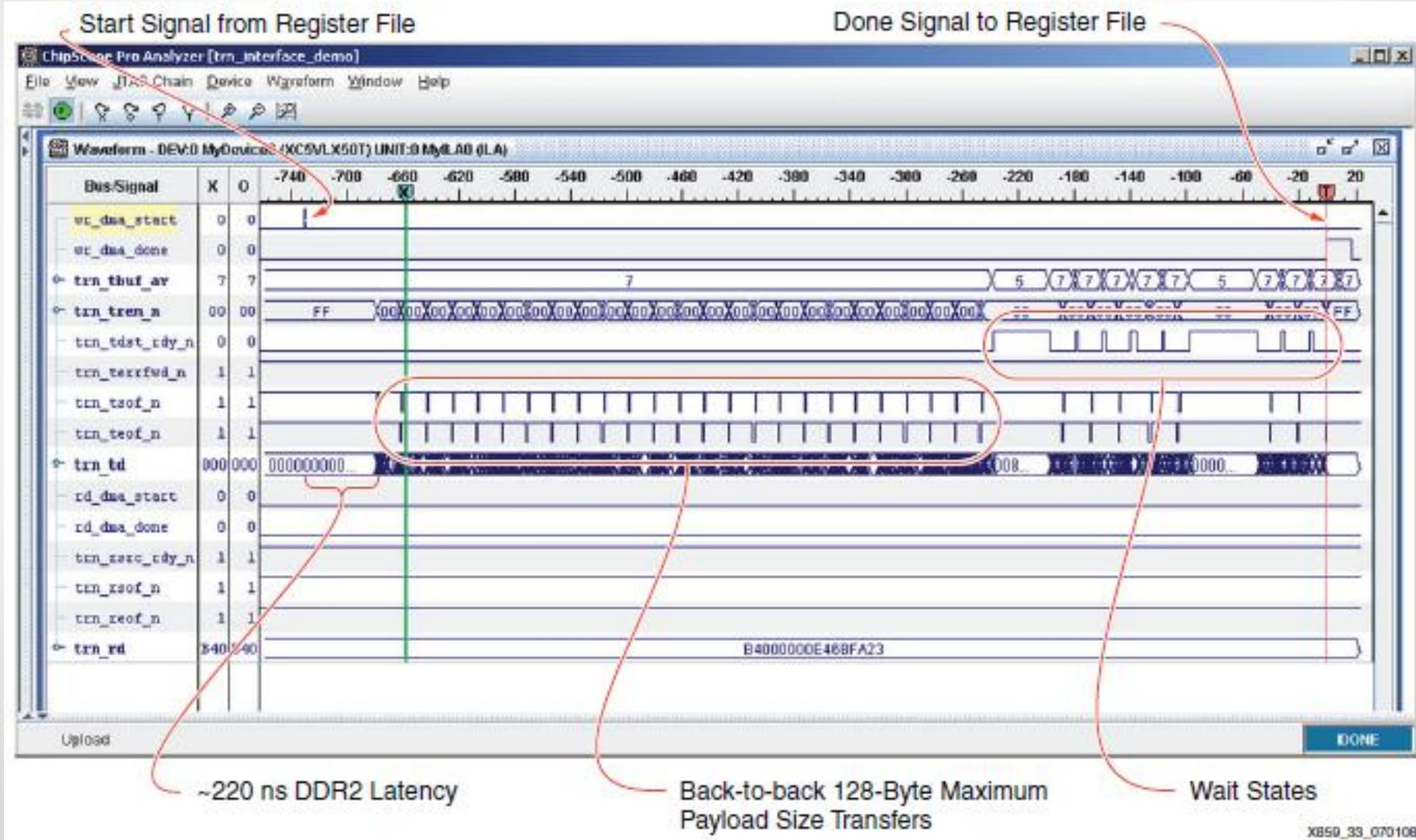
MRd to System Memory

- Intel Nehalem 5540 platform
- MRd(1st) -> CplD response time around 2.76 μ s
- MRd(8th) -> CplD response time around 3.82 μ s

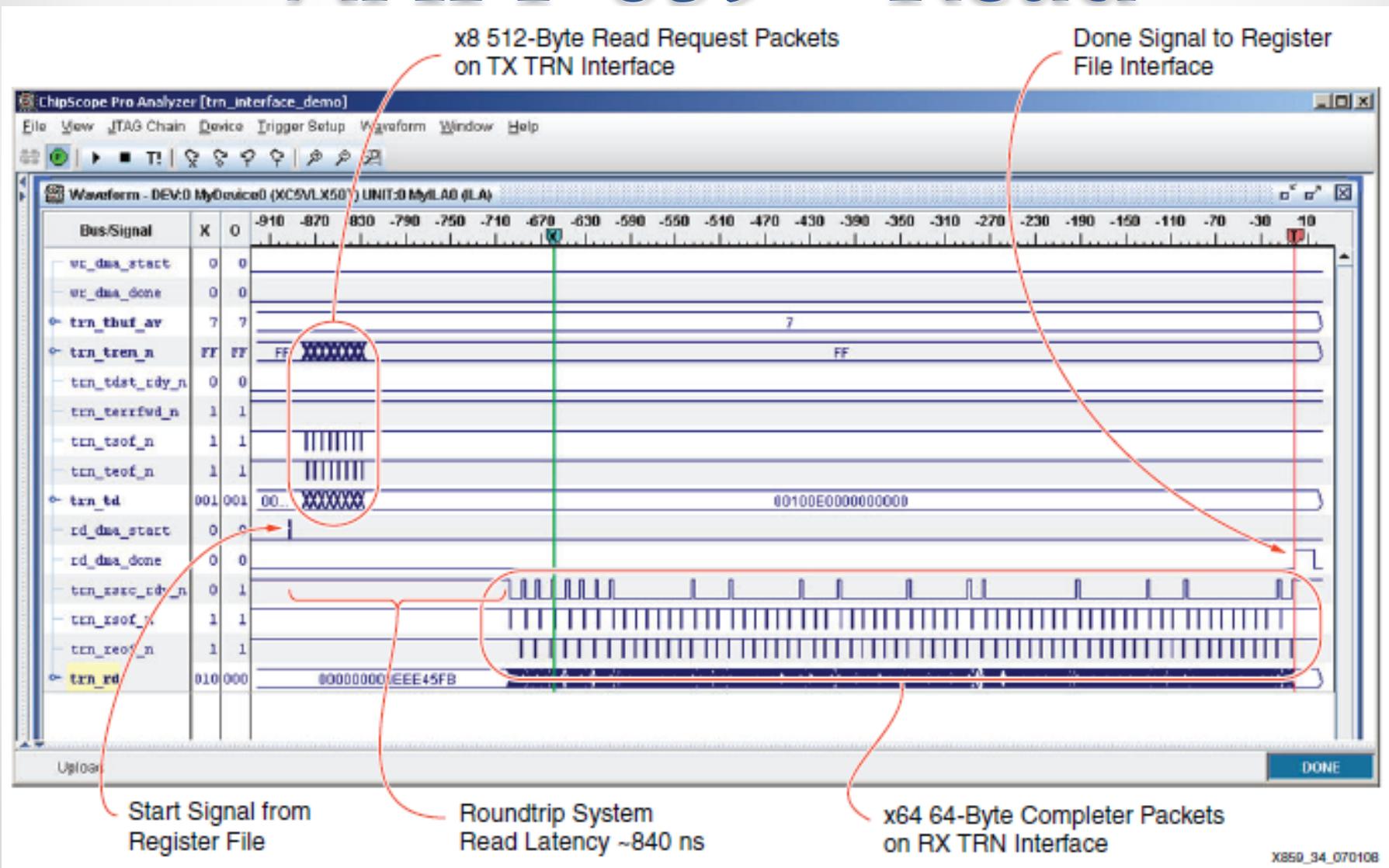


Link Tra 0	R→ x4	2.5 TLP 2083	Mem	MWr(32)	Length 10:00000	RequesterID 000:00:0	Tag 3	Address E3100000	1st BE 1111	Last BE 0000	Data 1 dword	VC ID 0	Explicit ACK Packet #2639	Metrics 2	# Packets 2	Time Delta 104.000 ns	Time Stamp - 0000 . 000 000 104 s
Link Tra 1	R→ x4	2.5 TLP 2084	Mem	MWr(32)	Length 10:00000	RequesterID 000:00:0	Tag 0	Address E3100000	1st BE 1111	Last BE 0000	Data 1 dword	VC ID 0	Explicit ACK Packet #2641	Metrics 2	# Packets 2	Time Delta 448.000 ns	Time Stamp 0000 . 000 000 000 s
Split Tra 0	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 000:07:0	CompleterID 040:00:0	Tag 0	TC 0	VC ID 0	Address E3100004	Status SC	Data 1 dword	Metrics 2	# LinkTras 7.336 µs	Time Delta 0000 . 000 000 448 s	Time Stamp		
Link Tra 4	R→ x4	2.5 TLP 2086	Mem	MWr(32)	Length 10:00000	RequesterID 000:00:0	Tag 1	Address E3100010	1st BE 1111	Last BE 0000	Data 1 dword	VC ID 0	Explicit ACK Packet #2655	Metrics 2	# Packets 2	Time Delta 368.000 ns	Time Stamp 0000 . 000 007 784 s
Link Tra 5	R→ x4	2.5 TLP 2087	Mem	MWr(32)	Length 10:00000	RequesterID 000:00:0	Tag 2	Address E310000C	1st BE 1111	Last BE 0000	Data 1 dword	VC ID 0	Explicit ACK Packet #2658	Metrics 2	# Packets 2	Time Delta 400.000 ns	Time Stamp 0000 . 000 008 152 s
Link Tra 6	R→ x4	2.5 TLP 2088	Mem	MWr(32)	Length 10:00000	RequesterID 000:00:0	Tag 3	Address E3100024	1st BE 1111	Last BE 0000	Data 1 dword	VC ID 0	Explicit ACK Packet #2661	Metrics 2	# Packets 2	Time Delta 368.000 ns	Time Stamp 0000 . 000 008 552 s
Link Tra 7	R→ x4	2.5 TLP 2089	Mem	MWr(32)	Length 10:00000	RequesterID 000:00:0	Tag 0	Address E3100020	1st BE 1111	Last BE 0000	Data 1 dword	VC ID 0	Explicit ACK Packet #2665	Metrics 2	# Packets 2	Time Delta 336.000 ns	Time Stamp 0000 . 000 008 920 s
Link Tra 8	R→ x4	2.5 TLP 2090	Mem	MWr(32)	Length 10:00000	RequesterID 000:00:0	Tag 1	Address E3100014	1st BE 1111	Last BE 0000	Data 1 dword	VC ID 0	Explicit ACK Packet #2668	Metrics 2	# Packets 2	Time Delta 368.000 ns	Time Stamp 0000 . 000 009 256 s
Link Tra 9	R→ x4	2.5 TLP 2091	Mem	MWr(32)	Length 10:00000	RequesterID 000:00:0	Tag 2	Address E3100018	1st BE 1111	Last BE 0000	Data 1 dword	VC ID 0	Explicit ACK Packet #2671	Metrics 2	# Packets 2	Time Delta 368.000 ns	Time Stamp 0000 . 000 009 624 s
Split Tra 1	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 000:07:0	CompleterID 040:00:0	Tag 0	TC 0	VC ID 0	Address E3100010	Status SC	Data 1 dword	Metrics 2	# LinkTras 1.912 µs	Time Delta 0000 . 000 009 992 s	Time Stamp		
Split Tra 2	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 000:07:0	CompleterID 040:00:0	Tag 0	TC 0	VC ID 0	Address E310000C	Status SC	Data 1 dword	Metrics 2	# LinkTras 1.696 µs	Time Delta 0000 . 000 011 904 s	Time Stamp		
Split Tra 3	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 000:07:0	CompleterID 040:00:0	Tag 0	TC 0	VC ID 0	Address E3100024	Status SC	Data 1 dword	Metrics 2	# LinkTras 1.704 µs	Time Delta 0000 . 000 013 600 s	Time Stamp		
Split Tra 4	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 000:07:0	CompleterID 040:00:0	Tag 0	TC 0	VC ID 0	Address E3100020	Status SC	Data 1 dword	Metrics 2	# LinkTras 1.904 µs	Time Delta 0000 . 000 015 304 s	Time Stamp		
Split Tra 5	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 000:07:0	CompleterID 040:00:0	Tag 0	TC 0	VC ID 0	Address E3100014	Status SC	Data 1 dword	Metrics 2	# LinkTras 1.936 µs	Time Delta 0000 . 000 017 208 s	Time Stamp		
Split Tra 6	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 000:07:0	CompleterID 040:00:0	Tag 0	TC 0	VC ID 0	Address E3100018	Status SC	Data 1 dword	Metrics 2	# LinkTras 1.800 µs	Time Delta 0000 . 000 019 144 s	Time Stamp		
Link Tra 22	R→ x4	2.5 TLP 2098	Mem	MWr(32)	Length 10:00000	RequesterID 000:00:0	Tag 3	Address E3100048	1st BE 1111	Last BE 0000	Data 1 dword	VC ID 0	Explicit ACK Packet #2712	Metrics 2	# Packets 2	Time Delta 400.000 ns	Time Stamp 0000 . 000 020 944 s
Link Tra 23	R→ x4	2.5 TLP 2099	Mem	MWr(32)	Length 10:00000	RequesterID 000:00:0	Tag 0	Address E3100004	1st BE 1111	Last BE 0000	Data 1 dword	VC ID 0	Explicit ACK Packet #2714	Metrics 2	# Packets 2	Time Delta 868.000 ns	Time Stamp 0000 . 000 021 344 s
Split Tra 7	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 040:00:0	CompleterID 000:00:0	Tag 1	TC 0	VC ID 0	Address FFC00000	Status SC	Data 32 dwords	Metrics 3	# LinkTras 36.000 ns	Time Delta 0000 . 000 022 212 s	Time Stamp		
Split Tra 8	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 040:00:0	CompleterID 000:00:0	Tag 2	TC 0	VC ID 0	Address FFC00080	Status SC	Data 32 dwords	Metrics 3	# LinkTras 44.000 ns	Time Delta 0000 . 000 022 248 s	Time Stamp		
Split Tra 9	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 040:00:0	CompleterID 000:00:0	Tag 3	TC 0	VC ID 0	Address FFC00100	Status SC	Data 32 dwords	Metrics 2	# LinkTras 36.000 ns	Time Delta 0000 . 000 022 292 s	Time Stamp		
Split Tra 10	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 040:00:0	CompleterID 000:00:0	Tag 4	TC 0	VC ID 0	Address FFC00180	Status SC	Data 32 dwords	Metrics 2	# LinkTras 44.000 ns	Time Delta 0000 . 000 022 328 s	Time Stamp		
Split Tra 11	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 040:00:0	CompleterID 000:00:0	Tag 5	TC 0	VC ID 0	Address FFC00200	Status SC	Data 32 dwords	Metrics 2	# LinkTras 36.000 ns	Time Delta 0000 . 000 022 372 s	Time Stamp		
Split Tra 12	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 040:00:0	CompleterID 000:00:0	Tag 6	TC 0	VC ID 0	Address FFC00280	Status SC	Data 32 dwords	Metrics 2	# LinkTras 44.000 ns	Time Delta 0000 . 000 022 408 s	Time Stamp		
Split Tra 13	R→ x4	2.5 Mem 00:00000	MRd(32)	RequesterID 040:00:0	CompleterID 000:00:0	Tag 7	TC 0	VC ID 0	Address FFC00300	Status SC	Data 32 dwords	Metrics 2	# LinkTras 36.000 ns	Time Delta 0000 . 000 022 452 s	Time Stamp		

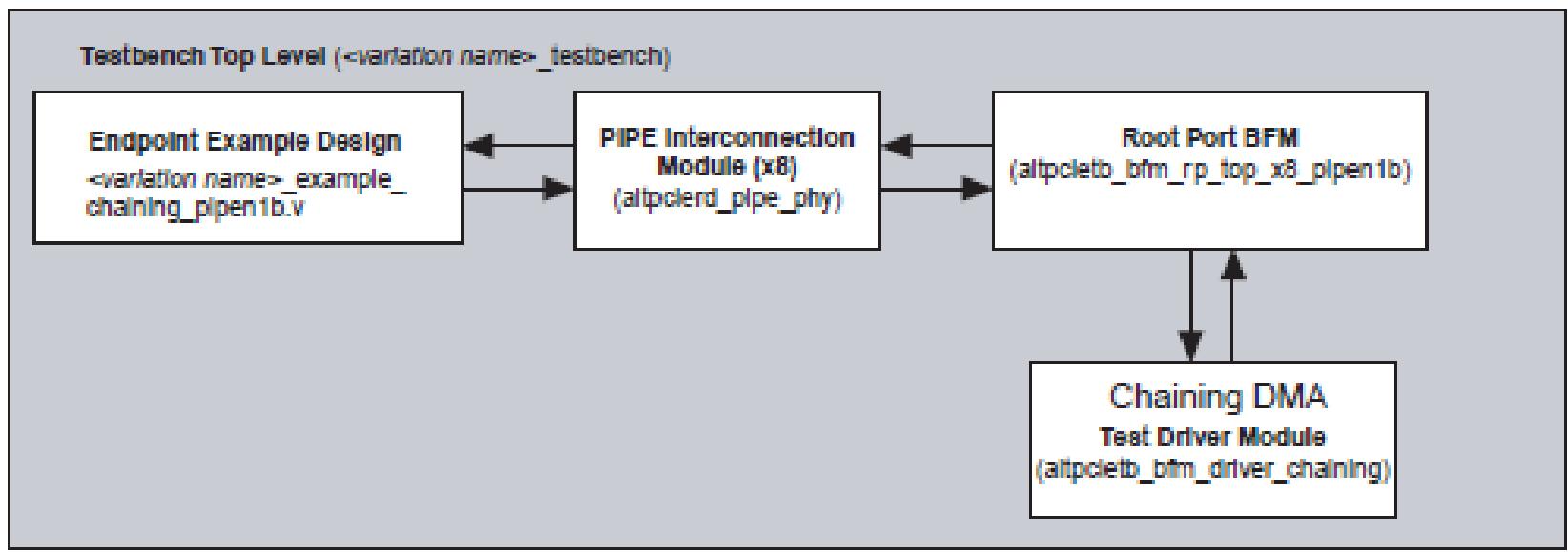
XAPP 859 – Write



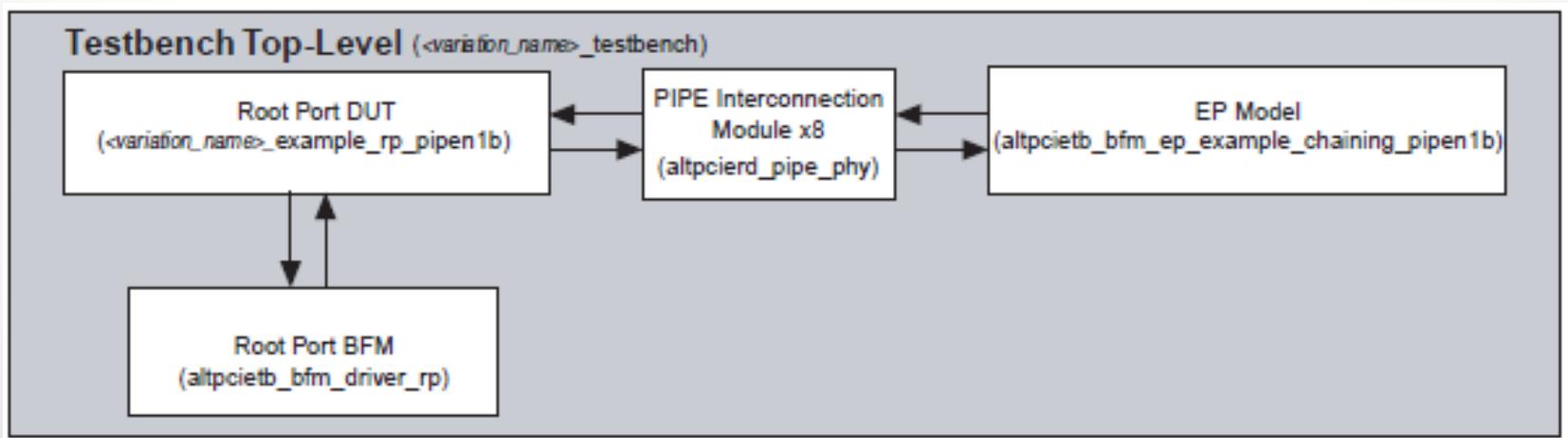
XAPP 859 – Read



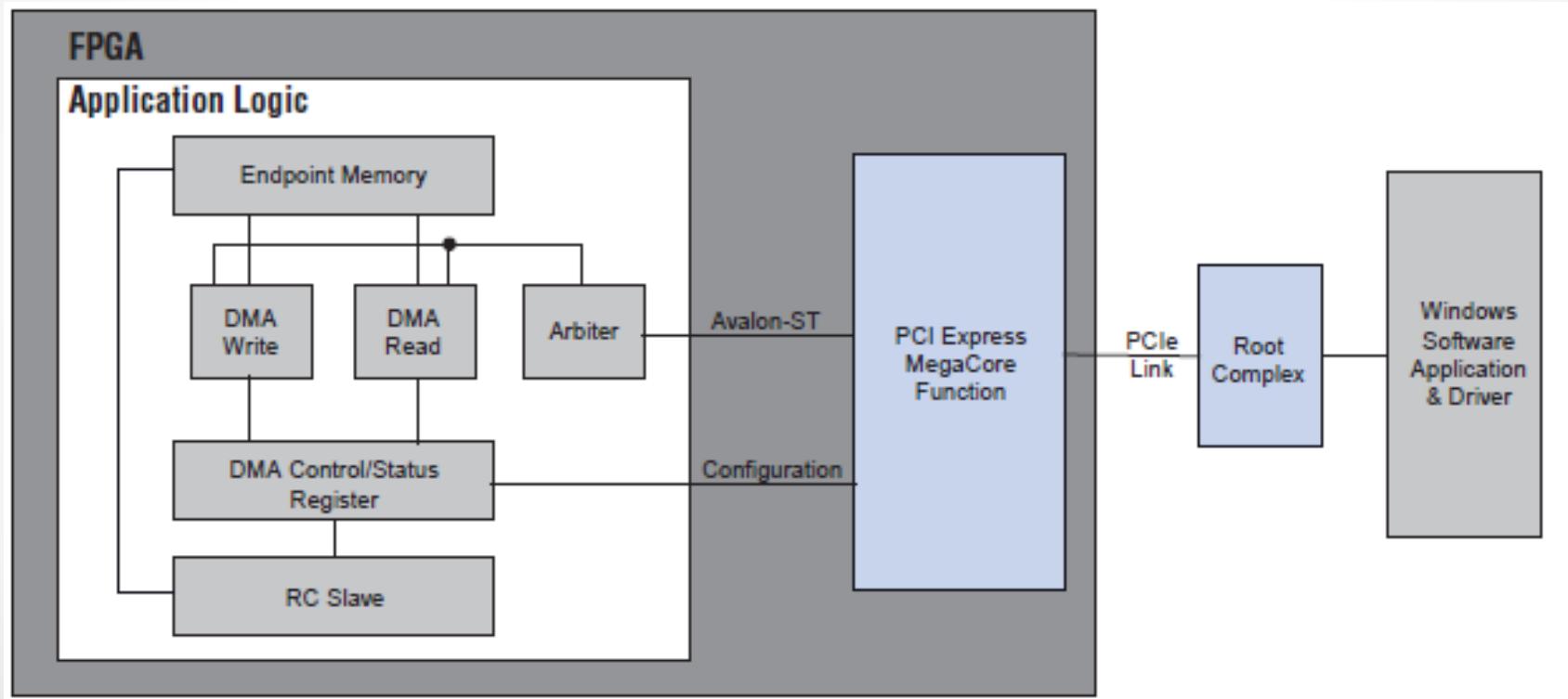
Endpoint TB

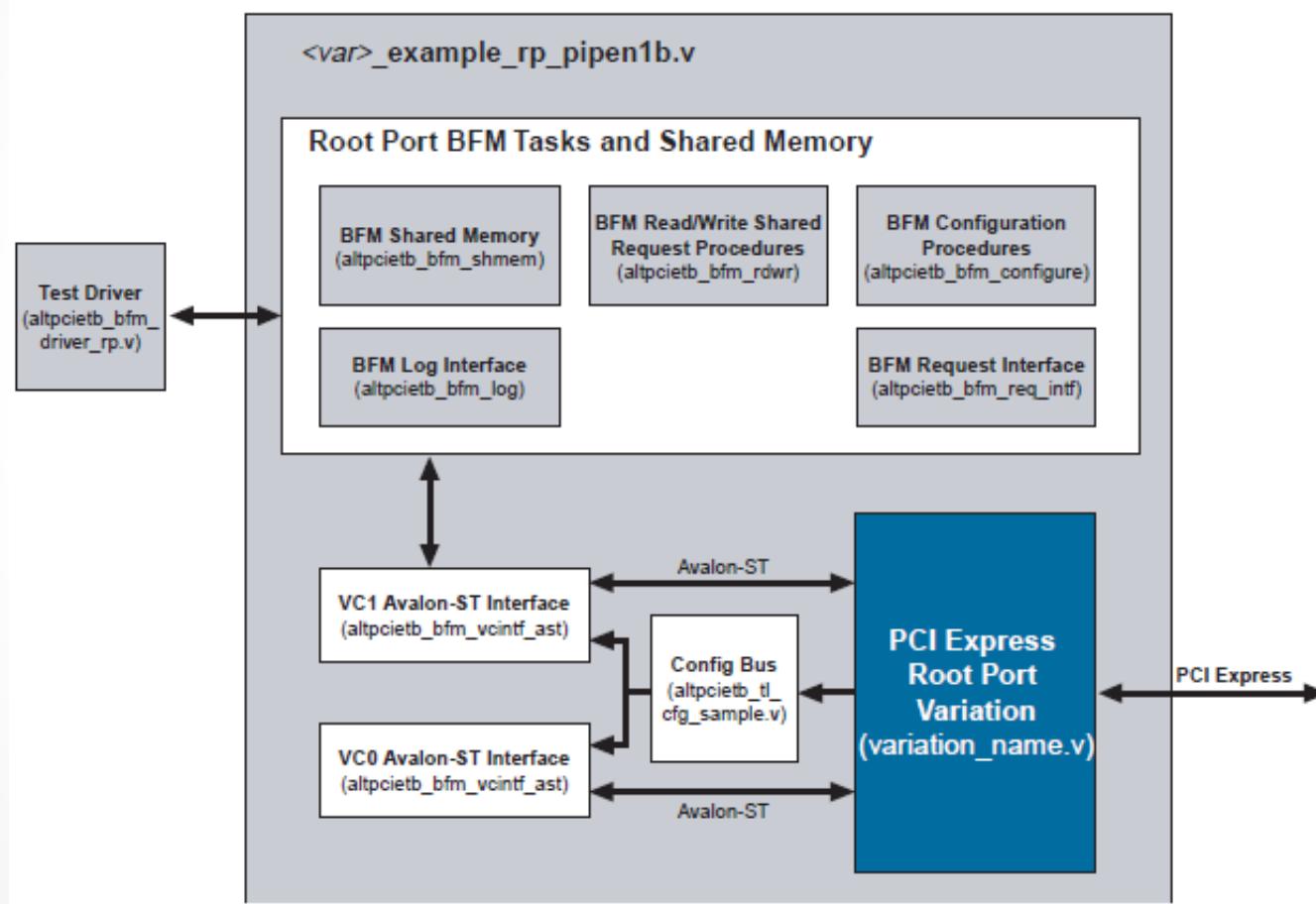


Root Port TB



AN456 – Chained DMA





Endianness

- 0x12345678
- Big-Endian stores the MSB at the lowest memory address. Little-Endian stores the LSB at the lowest memory address. The lowest memory address of multi-byte data is considered the starting address of the data. In Figure 1, the 32-bit hex value 0x12345678 is stored in memory as follows for each Endian-architecture. The lowest memory address is represented in the leftmost position, Byte 00.
- <http://en.wikipedia.org/wiki/Endianness>

Endian Order	Byte 00	Byte 01	Byte 02	Byte 03
Big Endian	12	34	56	78 (LSB)
Little Endian	78 (LSB)	56	34	12