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
Features:

- PCI Express is always a point to point connection
- Switches allow the access of several endpoints from one PCIe connection (PCle1)
- Every endpoint can exchange data with the Root Complex or other PCIe endpoints
PCI-Express : Bus Mastering

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: Physical Layer

PCI-Express is a serial Highspeedlink:

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

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

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

$T_{\text{BIT}}$

<table>
<thead>
<tr>
<th>$T_{\text{Symbol}}$</th>
<th>$T_{\text{BIT}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$T_{\text{BIT}}$</th>
<th>Release Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen 1 = 400 ps</td>
<td>2.5 GBit/s</td>
</tr>
<tr>
<td>Gen 2 = 200 ps</td>
<td>5.0 GBit/s</td>
</tr>
<tr>
<td>Gen 3 = 125 ps</td>
<td>8.0 GBit/s</td>
</tr>
<tr>
<td>Gen 4 = 62.5 ps</td>
<td>16.0 Gbit/s</td>
</tr>
<tr>
<td>Gen 5 = 31.25 ps</td>
<td>32.0 Gbit/s</td>
</tr>
</tbody>
</table>

One Lane consists of 4 physical signal traces
### PCI Express: Multilane Links

**Theoretical throughput in MByte / s**

<table>
<thead>
<tr>
<th>Link-Width</th>
<th>x1</th>
<th>x2</th>
<th>x4</th>
<th>x8</th>
<th>x16</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theoretical throughput in MByte / s</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gen 1: 2.5 Gbit/s</td>
<td>250</td>
<td>500</td>
<td>1.000</td>
<td>2.000</td>
<td>4.000</td>
</tr>
<tr>
<td>Gen 2: 5.0 Gbit/s</td>
<td>500</td>
<td>1.000</td>
<td>2.000</td>
<td>4.000</td>
<td>8.000</td>
</tr>
<tr>
<td>Gen 3: 8.0 Gbit/s</td>
<td>1.000</td>
<td>2.000</td>
<td>4.000</td>
<td>8.000</td>
<td>16.000</td>
</tr>
<tr>
<td>Gen 4: 16.0 Gbit/s</td>
<td>2.000</td>
<td>4.000</td>
<td>8.000</td>
<td>16.000</td>
<td>32.000</td>
</tr>
</tbody>
</table>

*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:

<table>
<thead>
<tr>
<th>MPS* in Bytes</th>
<th>5 DW Overhead</th>
<th>5 DW Loss in %</th>
<th>Max Throughput (Gen 1, x1, 5 DW)**</th>
<th>Max Throughput (Gen 1, x1, 6 DW)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>20 / (128+20)</td>
<td>13,5 %</td>
<td>216.3</td>
<td>210,5</td>
</tr>
<tr>
<td>256</td>
<td>20 / (256+20)</td>
<td>7,2 %</td>
<td>232</td>
<td>228,5</td>
</tr>
<tr>
<td>512</td>
<td>20 / (512+20)</td>
<td>3,8 %</td>
<td>240,5</td>
<td>238,75</td>
</tr>
<tr>
<td>1024</td>
<td>20 / (1024+20)</td>
<td>1,9 %</td>
<td>245,3</td>
<td>244,25</td>
</tr>
<tr>
<td>4096</td>
<td>20 / (4096+20)</td>
<td>0,5 %</td>
<td>248,8</td>
<td>248,75</td>
</tr>
</tbody>
</table>

* 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

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)
PCI Express Read Performance vs Write Performance

CPU Read:

- 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

Endpoint Write:

- For Maximum PCIe throughput, Memwrite Requests are necessary
# PCI-Express Read Performance

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

1 MByte = 10⁶ 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

- Up to 16 Datasources possible
- Each Datasource is stored in a separate Memory Buffer
- Core cares for complete address management, User only supplies data via AXI Stream Interfaces
- The destination can be the Host Memory (as shown) or an other PCIe endpoint Device

Memory Buffers can be linear or Scatter/Gather Memories

Image Sensor with PCIe Multichannel DMA IP Core

Camera Link, LVDS, custom

Analog / Digital Converters

Aurora, Chip2Chip, Ethernet, Custom

Host Memory

Framebuffer 2

Framebuffer 1

ADC-Buffer 4

ADC-Buffer 3

ADC-Buffer 2

ADC-Buffer 1

Aurora Buffer

Chip2Chip

Highspeed Links

Copyright Smartlogic 2005-2019, All Rights reserved. Confidential
Basic Framebuffer Architecture

Various Datasources

- Image Sensors
  - Camera Link, LVDS, custom

Analog / Digital Converters

- Highspeed Links
  - Aurora, Chip2Chip, Ethernet, Custom

Local Databuffer

- DDR
  - PCI Express

Final Destination

- CPU
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)

Readout data can be in tiles. Only important thing is, that pixels are always oriented from left to right. If a data turning is needed, it must be done inside the readout Logic 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 its 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.