



Debugging PCIe® Link & Transaction Layer Issues

Roland Scherzinger
Technical Marketing Expert
Agilent Technologies



Agenda

- Introduction
- Debugging PCI Express[®] Protocol bugs
- Is it a protocol problem?
- What are the symptoms?
- What should I look for?
- How do I find what I am looking for?

What are the Symptoms?

- The system “Blue Screens”
- The performance is not as expected
- Errors are logged
- Drivers do not load properly
- The link is not established

Sorry, Your computer has crashed

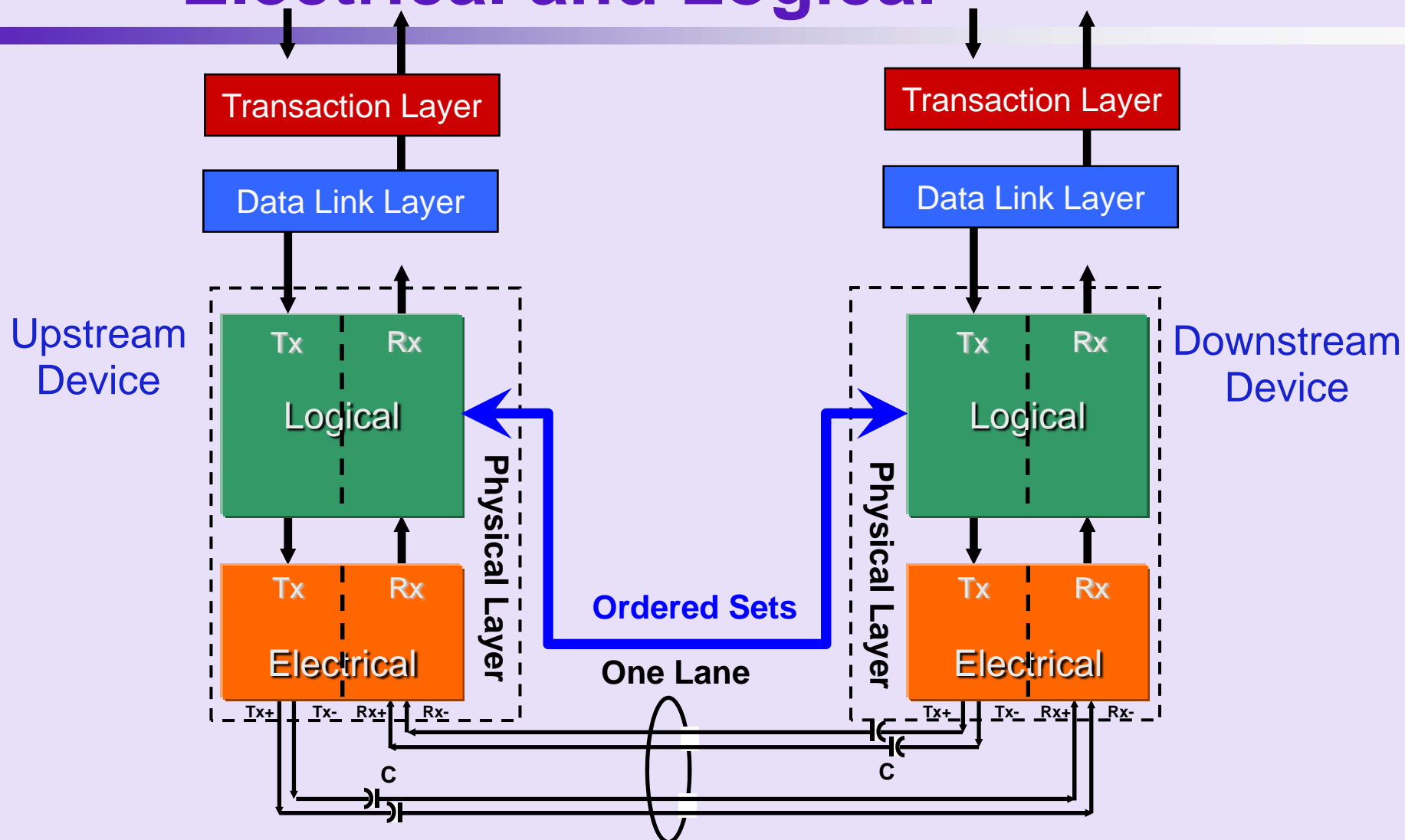
I am not going to tell you what is wrong

Please use your protocol analyzer to check it out.....

Is it a Physical Layer Error?

- Some errors at the physical layer may show when using a protocol analyzer.
- Errors such as disparity, symbol or 8B/10B encoding errors may show in the idle data when the link is in L0 or may happen in a TLP or DLLP
- The PCIe[®] Specification allows for X number of errors
- Many errors are recoverable
- How do I determine if it is a Physical Layer error?

Physical Layer Elements – Electrical and Logical

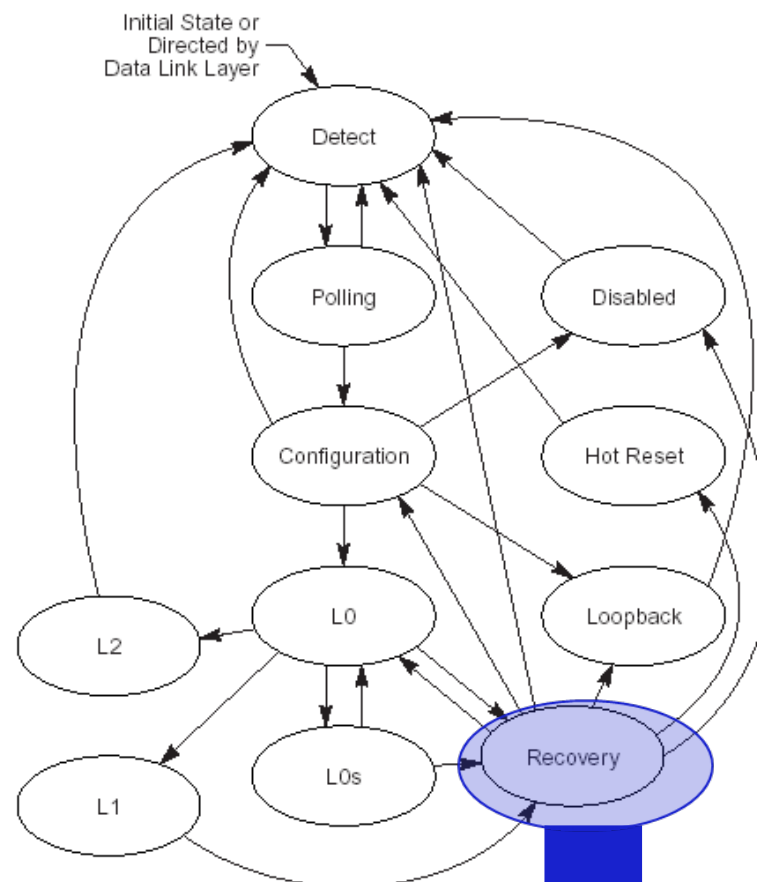
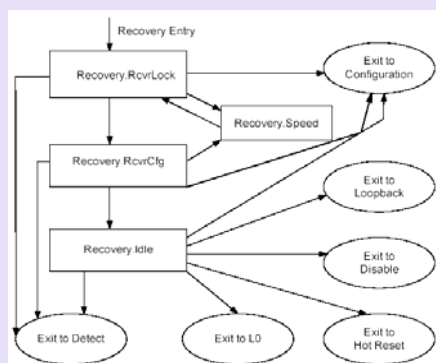


Symptoms of a Physical Layer Error

- At the Link and Transaction layers, physical layer errors may result in NAKs being sent if a CRC is corrupted
- Performance may not be as expected
- Link may enter recovery state from L0
- Does the link recover properly?
- If many disparity or symbol errors are observed on the protocol analyzer, it may be worth checking the electrical quality of the link using an oscilloscope

How does Link Training Work?

- Each of the devices implements a so called LTSSM that controls the link training.
- LTSSM stands for “Link Training and Status State Machine”
- Link training starts in state Detect.
- An active link that can transport transaction layer packets is in state “L0”.

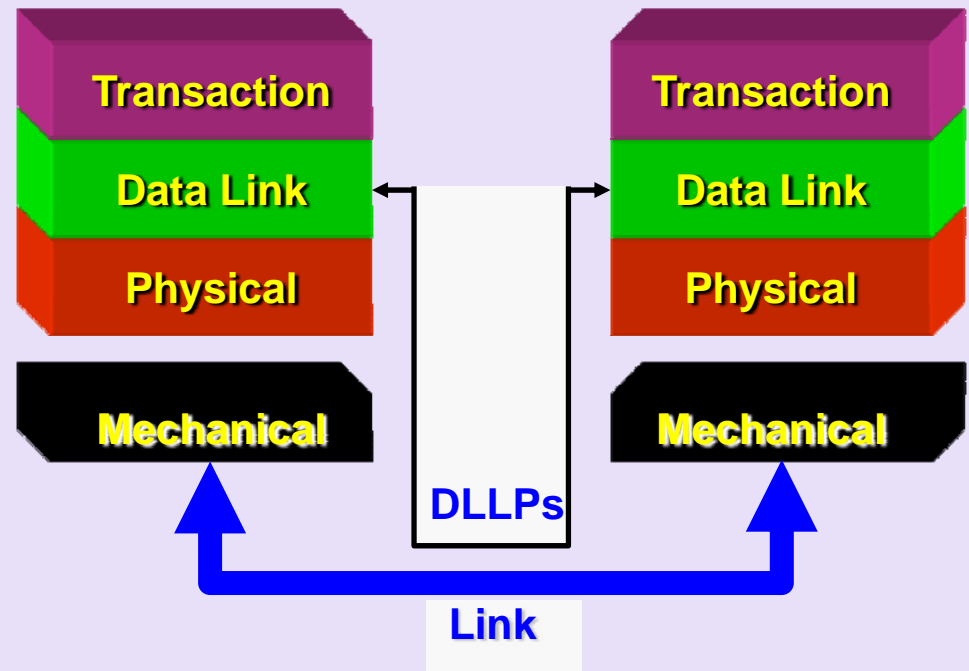


Link Layer Errors

- The Data Link Layer is responsible for ensuring that packets are transmitted and received successfully.
- This includes the ACK/NAK protocol and also Flow control protocol
- Many protocol errors exhibit these mechanisms

Data Link Layer Responsibilities

- Primary role is to assure the integrity of TLPs moving across the link.
- Link initialization and power management
- Tracking of link state and passing messages and status between Transaction and Physical layers.
- Exchanges traffic with neighboring port's DLL using DLLPs.
- DLLPs start and end at the Data Link Layer of each device.
- DLLPs and TLPs are interleaved on the link.



Link Layer Error example

- An endpoint is sending continuous DMA traffic to the host and it is observed that only 2 requests are sent back to back followed by a 5ms delay
- The performance of the device is not as expected
- What should one look for on a protocol analyzer?

Example of Bad LCRC Behavior

- This example shows the NAK instance and the responses

System Protocol Tester: Session: 2 - C:\Documents and Settings\gorgetty\Desktop\traces for mike\BadCRC.xml

File Edit View Capture Listing Window Help

Port Overview

Port	Name	Records	File
101/1	101/1:Upstream 101/1:Downstream	14095	C:\Documents and Settings\gorgetty\Desktop\traces for...

101 to 102 = 0 ns

Packet Viewer-1

Record No	Timestamp	Rel. Timest...	101/1:Upstream: Type	101/1:Downstream: ...	Sequence Nu...	Tag	Length	Address, Register Nu...	Data, DataFC	LCRC, Message Code	Completion Status	Requester ID, Com...
3296	24.23222801...	428 ns	Ack		6 C4							
3298	24.23222868...	672 ns	Completion with data		6 D1	03	0 01		Data=00 00 11 10	LCRC=10 31 11 8F	SC (Successful Co...	Requester ID=00 00 Completer ID=04 00
3300	24.23222918...	492 ns		Ack	6 D1							
3449	24.23249784...	268.664 us		Config Read Type 0	6 C5	09	0 01	Register Number=00		LCRC=AC 48 F1 97		Requester ID=00 00 Completer ID=04 00
3450	24.23249828...	444 ns	Nak		6 C4							
3452	24.23249873...	448 ns	Message		6 D2	1F				LCRC=45 B0 4A 47 Message Code=ERR_COR		Requester ID=04 00
3453	24.23249874...	4 ns		Config Read Type 0	6 C5	09	0 01	Register Number=00		LCRC=AC 48 F1 17		Requester ID=00 00 Completer ID=04 00
3455	24.23249918...	444 ns	Ack		6 C5							
3456	24.23249922...	44 ns		Ack	6 D2							
3459	24.23249986...	636 ns	Completion with data		6 D3	09	0 01		Data=86 80 5E 10	LCRC=1C 23 50 A4	SC (Successful Co...	Requester ID=00 00 Completer ID=04 00
3460	24.23250035...	492 ns		Ack	6 D3							

Stopped Offline

Link Layer Error Example-continued

- It is possible that all of the packets show correctly on the analyzer display, assuming there are no physical layer problems
- It may be possible to use filtering to show the delays caused by the flow control credit – showing the relative time between each of the requests will quickly show if there is a flow control bottleneck
- Flow control may cause problems in other areas, for example: A Fibre Channel HBA may require a high percentage of the bandwidth required on the PCI Express link and may show limited performance on the Fibre Channel Link
- See example on next slide

Example – Filter Only Memory Read

System Protocol Tester: Session: 2 - C:\Documents and Settings\gorgetty\My Documents\08 - Settings Files\02 - Trace Files\read_with_multiple_compl...

File Edit View Capture Listing Window Help

Port Overview

Port	Name	Records	File
101/1	101/1:Downstream 101/1:Upstream	35301	C:\Documents and Settings\gorgetty\My Documents\08 - ...

M1 to M2 = 0 ns

Packet Viewer-1

Record No	Timestamp	Rel. Timest...	101/1:Downstream: ...	101/1:Upstream: Type	Sequence Nu...	Tag	Length	Address, Register Nu...	Completion Status	Requester
2	8.896 us	0 ns		Memory Read	A 71	15	4 00	Address=FF 00 10 00		Requester
41	11.196 us	2.300 us		Memory Read	A 72	14	4 00	Address=FF 00 20 00		Requester
83	13.600 us	2.404 us		Memory Read	A 73	13	4 00	Address=FF 00 30 00		Requester
136	16.816 us	3.216 us		Memory Read	A 74	12	4 00	Address=FF 00 00 00		Requester
179	19.216 us	2.400 us		Memory Read	A 75	11	4 00	Address=FF 00 10 00		Requester
223	21.700 us	2.484 us		Memory Read	A 76	10	4 00	Address=FF 00 20 00		Requester
276	24.896 us	3.196 us		Memory Read	A 77	0F	4 00	Address=FF 00 30 00		Requester
320	27.380 us	2.484 us		Memory Read	A 78	0E	4 00	Address=FF 00 00 00		Requester
364	29.888 us	2.508 us		Memory Read	A 79	0D	4 00	Address=FF 00 10 00		Requester
420	33.236 us	3.348 us		Memory Read	A 7A	0C	4 00	Address=FF 00 20 00		Requester
462	35.624 us	2.388 us		Memory Read	A 7B	0B	4 00	Address=FF 00 30 00		Requester
501	37.936 us	2.312 us		Memory Read	A 7C	0A	4 00	Address=FF 00 00 00		Requester
544	40.340 us	2.404 us		Memory Read	A 7D	09	4 00	Address=FF 00 10 00		Requester
598	43.548 us	3.208 us		Memory Read	A 7E	08	4 00	Address=FF 00 20 00		Requester
641	46.056 us	2.508 us		Memory Read	A 7F	07	4 00	Address=FF 00 30 00		Requester
684	48.488 us	2.432 us		Memory Read	A 80	06	4 00	Address=FF 00 00 00		Requester
736	51.644 us	3.156 us		Memory Read	A 81	05	4 00	Address=FF 00 10 00		Requester
777	54.224 us	2.580 us		Memory Read	A 82	04	4 00	Address=FF 00 20 00		Requester
817	56.532 us	2.308 us		Memory Read	A 83	03	4 00	Address=FF 00 30 00		Requester
884	60.420 us	3.888 us		Memory Read	A 84	02	4 00	Address=FF 00 00 00		Requester
916	62.128 us	1.708 us		Memory Read	A 85	01	4 00	Address=FF 00 10 00		Requester
958	64.712 us	2.584 us		Memory Read	A 86	00	4 00	Address=FF 00 20 00		Requester

Filter Complete Stopped Offline

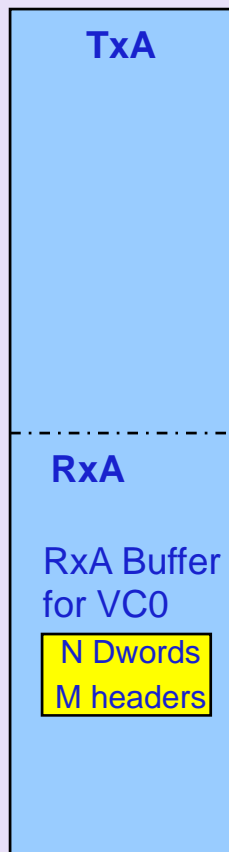
Flow Control

How it works

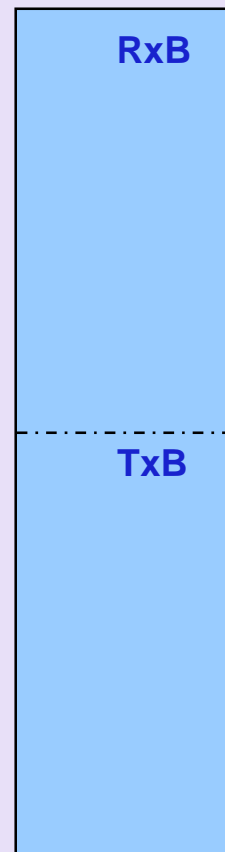
Example:

**Device A flow control
(Device B follows
same procedure)**

Device A



Device B



**Device B may only send
a packet if CREDITS_
CONSUMED is lower
than the CREDIT_LIMIT.
Credit limit is determined
by incoming flow control
updates**

Flow Control

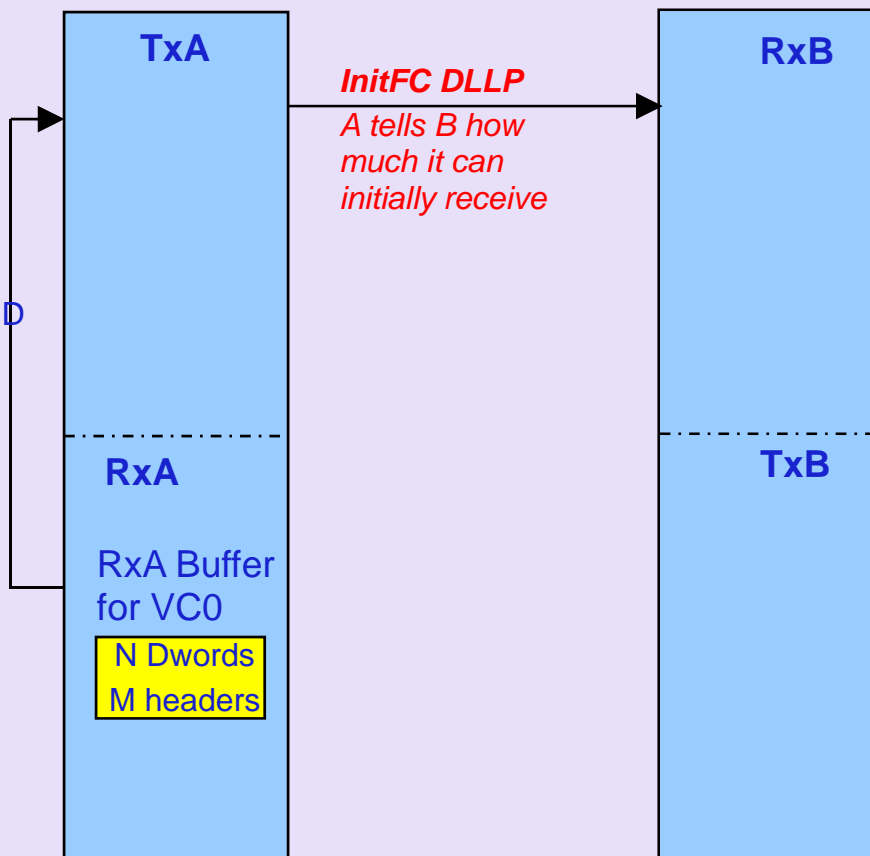
How it works

Example:

Device A flow control
(Device B follows
same procedure)

Device A

Device B



Device B may only send a packet if CREDITS_CONSUMED is lower than the CREDIT_LIMIT. Credit limit is determined by incoming flow control updates

Flow Control

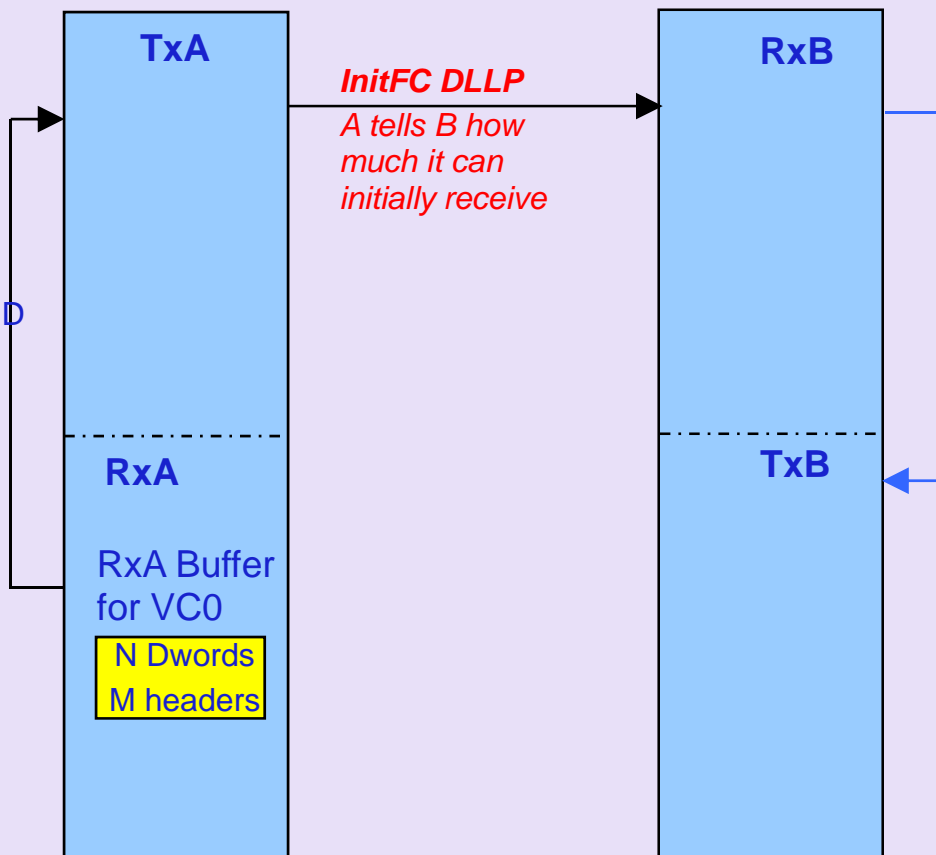
How it works

Example:

Device A flow control
(Device B follows same procedure)

Device A

Device B



1. CREDITS_ALLOCATED

RxA Initial buffer size – sent as InitFC

Device B may only send a packet if CREDITS_CONSUMED is lower than the CREDIT_LIMIT. Credit limit is determined by incoming flow control updates

2. TxB – to determine when it can send TLPs, it tracks CREDITS_CONSUMED = 0
CREDIT_LIMIT = # from A

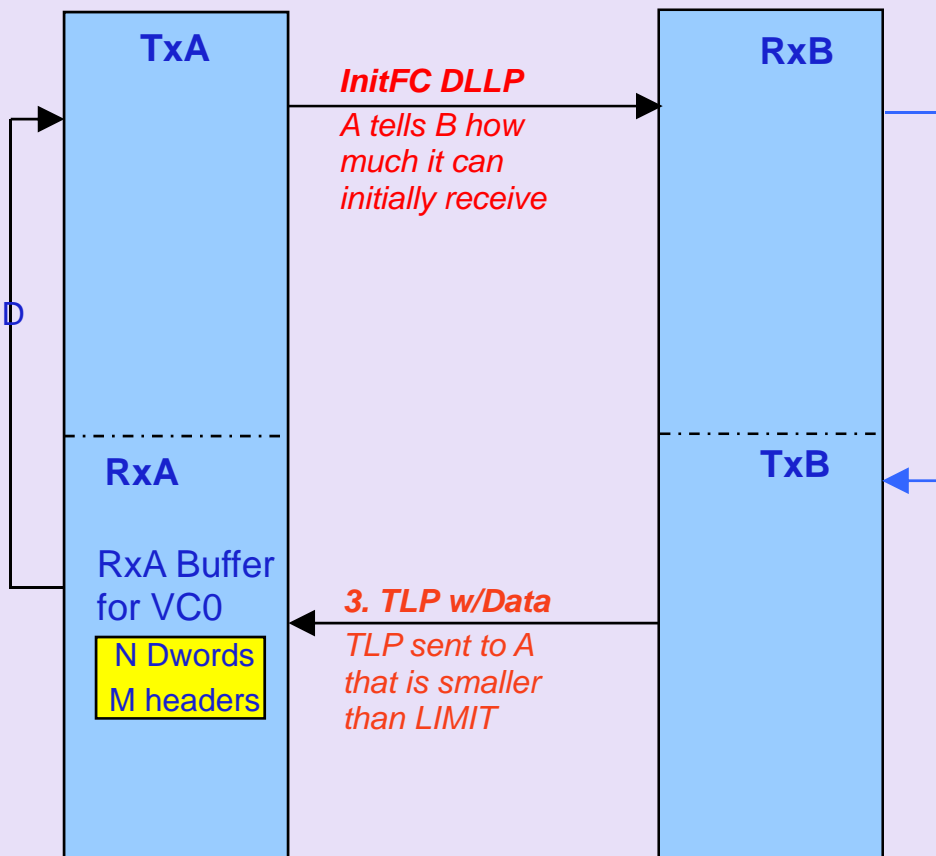
Flow Control

How it works

Example:
Device A flow control
(Device B follows same procedure)

Device A

Device B



Device B may only send a packet if CREDITS_CONSUMED is lower than the CREDIT_LIMIT. Credit limit is determined by incoming flow control updates

2. TxB – to determine when it can send TLPs, it tracks CREDITS_CONSUMED = N CREDIT_LIMIT = # from A

↓
CREDITS_CONSUMED
increased according to FC rules

Flow Control

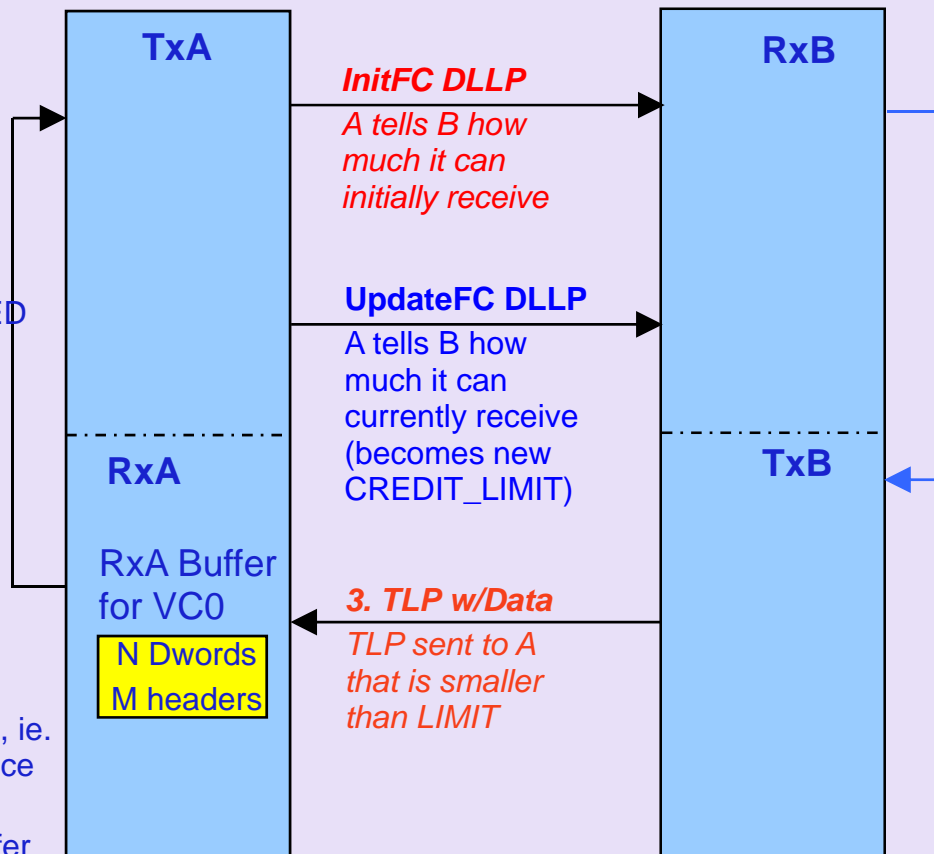
How it works

Example:

Device A flow control
(Device B follows same procedure)

Device A

Device B



Transaction Layer Errors

- Protocol Analyzers are extremely important for debugging transaction layer problems. This is the only way that the software or driver developers can see exactly what is happening on the bus
- It can be the timing or sequence of events which causes a system hang or failure
- It may also be only in obscure cases that problems show up
- Often, the trigger condition is not known, only the symptoms can be triggered on
- This is important when considering how to find a problem – first clearly define the symptoms

Example – Transaction Layer Error

- What are the symptoms?
 - ✓ Does the Link enter recovery state?
 - ✓ Does the system hang?
 - ✓ What happens when the system hangs?
 - ✓ Is the link still active even although the system hangs?
- Usually the events prior to the hang are most interesting
- Careful setting of the trigger position will ensure the correct information is captured.

Example Continued

- Does the Error happen at both 2.5G and 5G?
- Is the behavior the same at all link widths?
- Is the problem related to the system or the card or both?
- Does the device involve a switch? Does a downstream event relate to a problem on the upstream side?
- Is the problem intermittent?
- Does the problem happen each time after power up?

Performance Considerations

- What should the performance of a link be?
- What is the theoretical maximum possible?
- What is the actual throughput being achieved?

Overhead Considerations

- There are many factors to take into account when calculating the theoretical maximum bandwidth
- Some of these are:
 - ✓ Flow control considerations
 - ✓ Logical idle
 - ✓ TLP headers
 - ✓ ACK/NAK latency

Quick Calculation

- Work out the theoretical maximum of a x1 link at 2.5G:
 - ✓ 1 byte = 10 bits on the wire
 - ✓ Data rate is 2.5Gb/s
 - ✓ Approximately equal to 250MB/s
 - ✓ This 250MB/s would be the bandwidth of the link in each direction assuming all bits are counted.
 - ✓ Now add in TLP header information
 - ✓ Example: a Memory Write request requires at least 3 DWords of header and a minimum of 1 DWord of payload = 4 DWords

Performance Calculation

- Add the STP, END, Sequence Number and LCRC fields adds another 2 DWords so a minimum of 6 DWords is used to transfer 1 DWord of Data.
- Assuming these were back to back with no Flow Control limitation that would mean the maximum throughput with a data payload size of 1DW would be 41.66MB/s

Increase the Payload Size

- Increasing the payload size would increase the throughput drastically, take 16 DW payload length for example:
- Overhead is 5 DW, so if the payload length is 16 DW (64 Bytes), theoretically the maximum bandwidth for transferring data could be 171.875MB/s assuming the packets are sent back to back

Using Max Payload Size

- Using the maximum allowable payload length of 4KB (1024 DW), then the maximum data bandwidth, without flow control consideration would be 249.875MB/s.
- In real life, this would never happen because each end of the link has a physical buffer size which limits the maximum number of headers and payload, typical payload sizes being somewhere around 64 or 128 Bytes.
- The physical buffer size also determines how many packets can be received at one time, if as few as 2 credits are available, then a delay of 1us happens before the next 2 packets can be sent (after a flow control update) The performance decreases drastically.

Performance Considerations

- $1\mu\text{s} = 1000\text{ ns} = 250\text{ symbol (10b Bytes) times.}$
- Now we have 334 bytes (including idle and flow control update times) for each 64 bytes of payload transferred
- $64/334 * 250\text{MB/s} = 47.90\text{MB/s}$ assuming $1\mu\text{s}$ for the flow control delay
- Now it is clear that there are several factors which will determine the maximum throughput that can be achieved on a PCI Express link.

How do I measure it?

- Protocol analyzers have different methods of working out performance:
 - ✓ One time capture
 - With a one time capture, the time length where the performance can be measured will be relatively short given finite trace buffer size
 - This method with careful filtering and storage qualification can provide more detail
 - ✓ Real time capture
 - For a high level overview, real time capture uses counters which provide less information but show a much larger time range

Configuration Space Errors

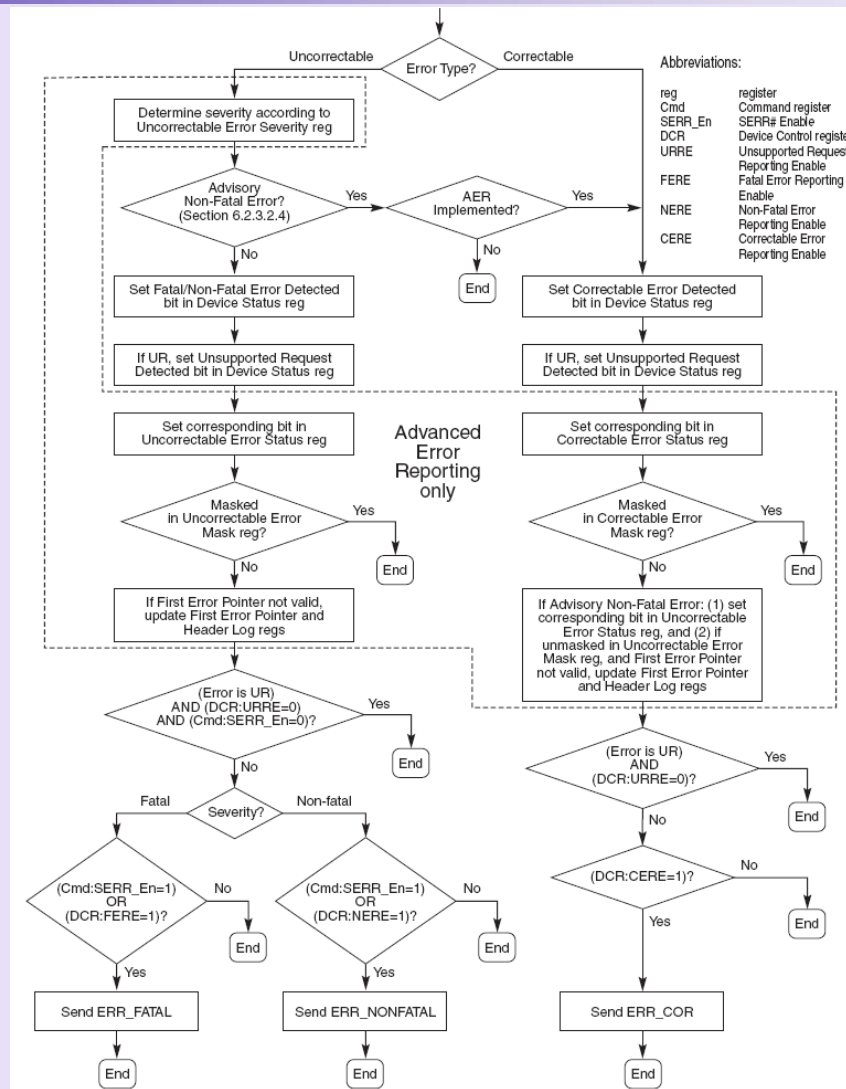
- PCI Express has a large configuration space which contains many different registers
- Do the registers contain the correct values?
- Do the registers behave in the correct manner?

Type 0 Configuration Space Header

31				0				Byte Offset
Device ID		Vendor ID		00h				
Status		Command		04h				
Class Code			Revision ID	08h				
BIST	Header Type	Master Latency Timer	Cache Line Size	0Ch				
Base Address Registers				10h				
				14h				
				18h				
				1Ch				
				20h				
				24h				
Cardbus CIS Pointer				28h				
Subsystem ID		Subsystem Vendor ID		2Ch				
Expansion ROM Base Address				30h				
Reserved			Capabilities Pointer	34h				
Reserved				38h				
Max_Lat	Min_Gnt	Interrupt Pin	Interrupt Line	3Ch				

OM14316

Advanced Error Reporting (AER) Test Challenges



Summary

- When debugging protocol issues a protocol analyzer can be a very powerful tool
- Determine if it is a physical, data link or transaction layer error
- Determine the symptoms
 - ✓ What can I trigger on?
- How do the symptoms relate to the problem
- Re-test to check for resolution

Contact Information

- Roland Scherzinger
- Email: roland_scherzinger@agilent.com
- ++49 7031 464 1066

Thank you for attending the
PCI-SIG Developers Conference
Europe 2009

For more information please go to
www.pcisig.com