Deadlock Avoidance in PCI Express® Based Architectures

Isaac Livny

Applications Engineer

LSI Corporation

Agenda

- Deadlock issues in bus architectures
- Deadlocks in PCI Express® architectures
- Deadlocks at common resource level
  - ✓ Example: ASPM L1 re-entry delay
- Deadlock at RC/chipset level
  - ✓ Example: PME fence mechanism hang-up
- Deadlock at device level – EP or RC
  - ✓ Example 1: Incompliant NAK
  - ✓ Example 2: CLKREQ# readiness
- Deadlock at PCIe® link level
  - ✓ Example: Credit leakage

Agenda

- Deadlock issues in bus architectures
- Deadlocks in PCI Express architectures
- Deadlocks at common resource level
  - ✓ Example: ASPM L1 re-entry delay
- Deadlock at RC/chipset level
  - ✓ Example: PME fence mechanism hang-up
- Deadlock at device level – EP or RC
  - ✓ Example 1: Incompliant NAK
  - ✓ Example 2: CLKREQ# readiness
- Deadlock at PCIe link level
  - ✓ Example: Credit leakage

General

- Deadlock avoidance crucial for every bus architecture
- Occur when multiple agents mutually depend on response from each other or common resource
  - ✓ Access to a resource cannot complete until other agent has completed its access to the same resource
- Conventional PCI back-off schemes
  - ✓ Disconnect
  - ✓ Retry
- PCI Express replaces retry and disconnect termination schemes with split transaction protocols.
  - ✓ Complexity of protocols can create more deadlock pitfalls

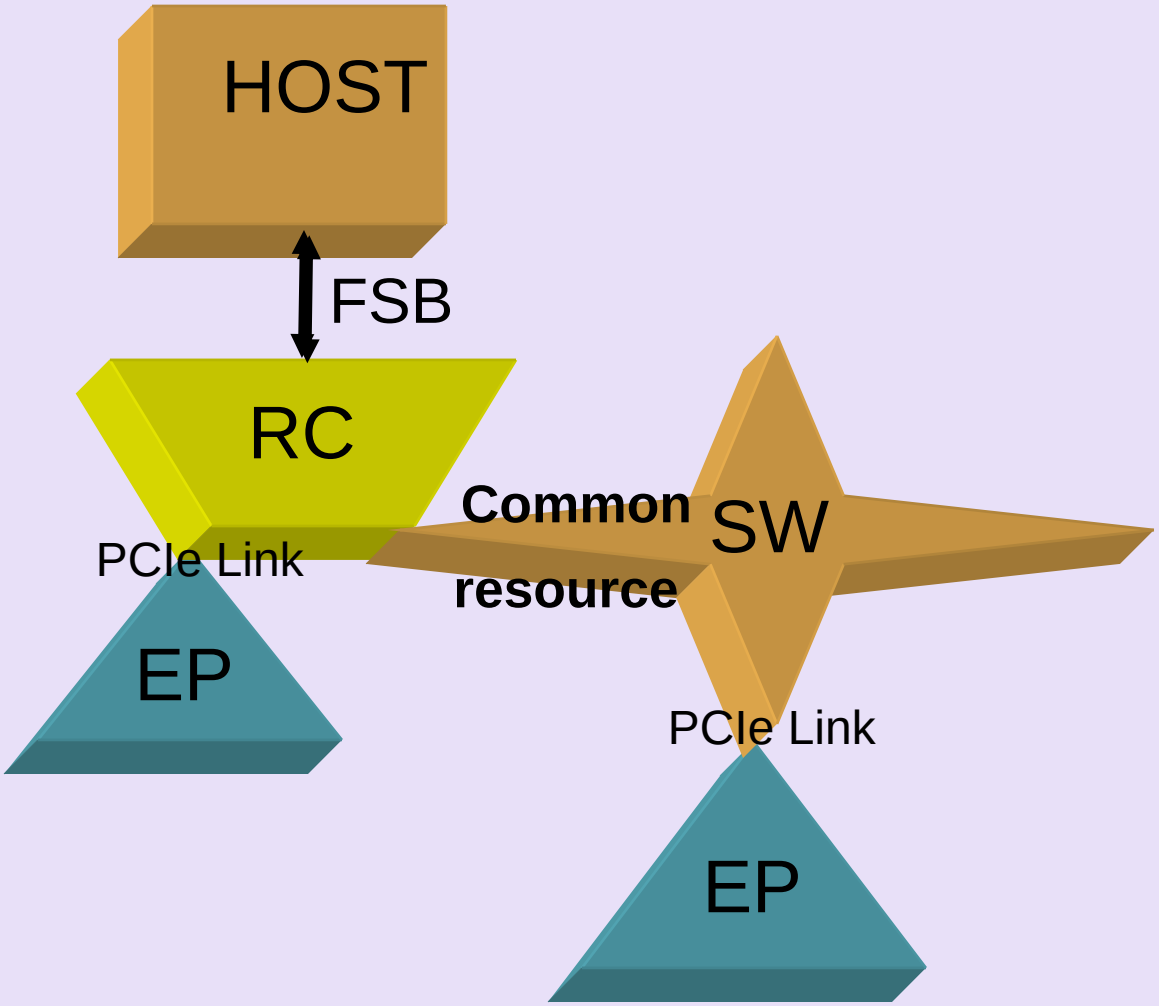
Agenda

- Deadlock issues in bus architectures
- Deadlocks in PCI Express architectures
- Deadlocks at common resource level
  - ✓ Example: ASPM L1 re-entry delay
- Deadlock at RC/chipset level
  - ✓ Example: PME fence mechanism hang-up
- Deadlock at device level – EP or RC
  - ✓ Example 1: Incompliant NAK
  - ✓ Example 2: CLKREQ# readiness
- Deadlock at PCIe link level
  - ✓ Example: Credit leakage

Deadlock scenarios

- 4 types of deadlock scenarios
  - characterized by where they can occur in the PCI Express fabric
  - ✓ Within an endpoint
  - ✓ Associated with a particular PCI Express link
  - ✓ Associated with a resource common to a number of PCI express devices or within a PCI express switch
  - ✓ At the Root Complex level, within the chipset or at the system level

PCI Express Fabric



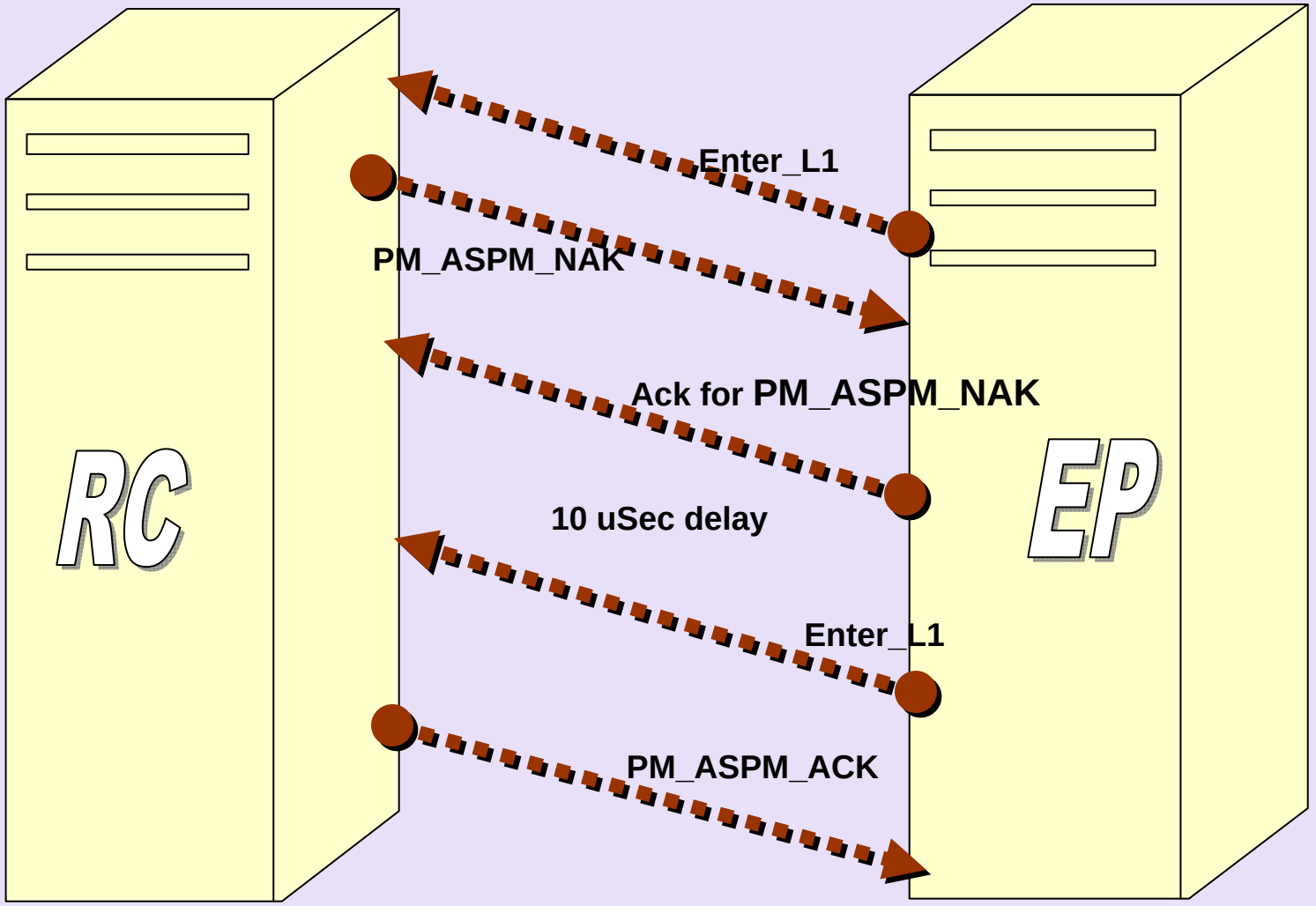
Deadlocks and avoidance

- PCIe protocols affected
  - ✓ ASPM L1 entry and exit protocols
  - ✓ PME fence mechanism
  - ✓ Error reporting schemes
  - ✓ Credit handling for special cases
  - ✓ Clock power management readiness
- Deadlock avoidance schemes
  - ✓ At the device level
  - ✓ Root Complex level
  - ✓ System level
  - ✓ Possible specification enhancements

Agenda

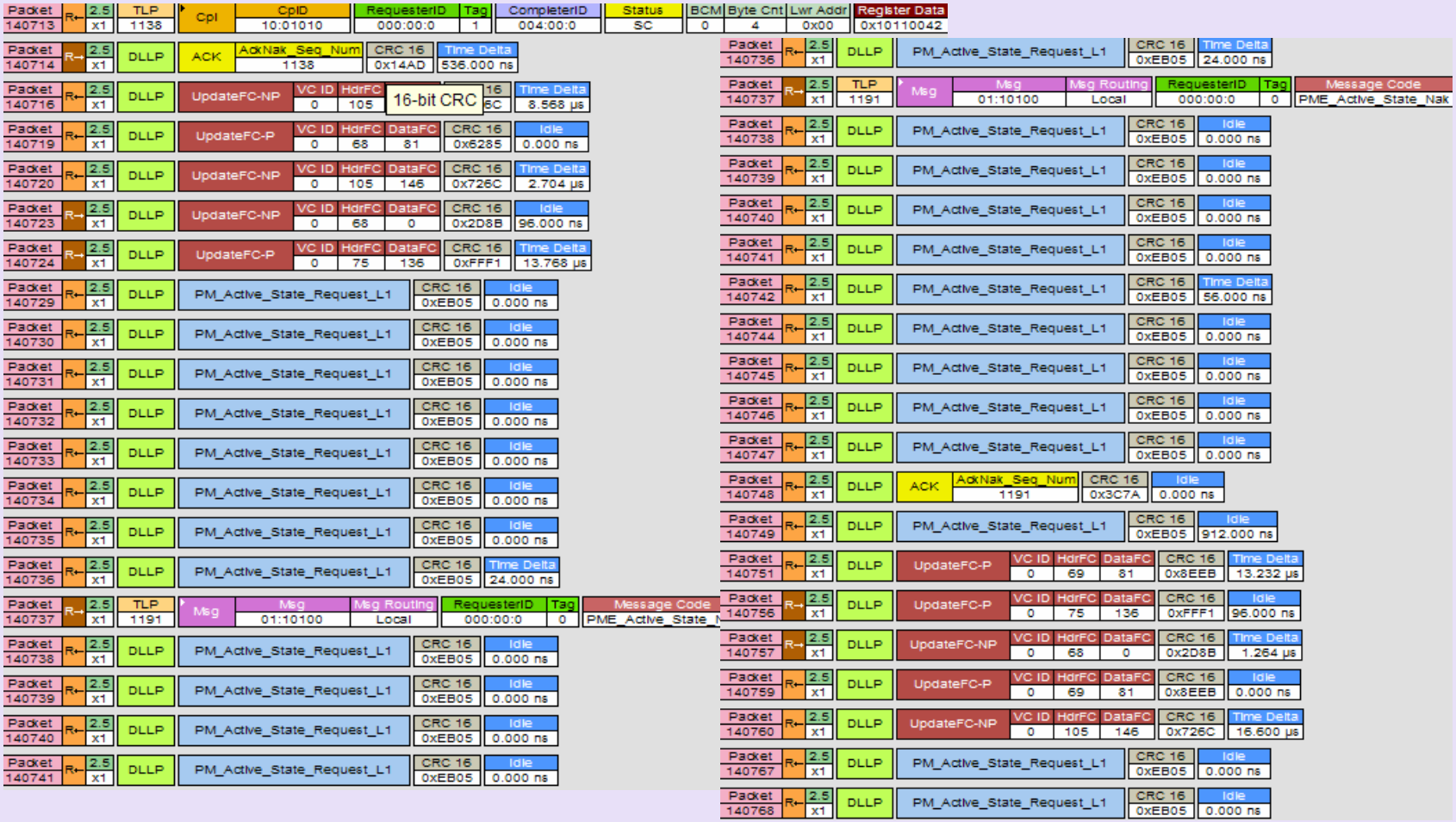
- Deadlock issues in bus architectures
- Deadlocks in PCI Express architectures
- Deadlocks at common resource level
  - ✓ Example: ASPM L1 re-entry delay
- Deadlock at RC/chipset level
  - ✓ Example: PME fence mechanism hang-up
- Deadlock at device level – EP or RC
  - ✓ Example 1: Incompliant NAK
  - ✓ Example 2: CLKREQ# readiness
- Deadlock at PCIe link level
  - ✓ Example: Credit leakage

ASPM L1 Re-entry delay



Deadlock example at common resource level

ASPM L1 re-entry delay deadlock



L1 Re-entry delay pitfall

- ASPM L1 enabled at endpoint, disabled at root port
  - ✓ PCIe 1.1 Requires correct software sequencing
  - ✓ Used to force PM_Active_State_NACK
- Traffic recess – EP requests L1, Root port NAKs
- EP Data Link Layer Acks the NACK message
- EP required ASPM L1 re-entry delay > 10 usec
- If re-entry is violated – RC interprets as old request
- RC resources congestion blocks additional traffic
- Deadlock situation arises.
- If RC low on resources deadlock can also happen when re-entry delay is within spec
- RC should flush L1 requests not to consume resources

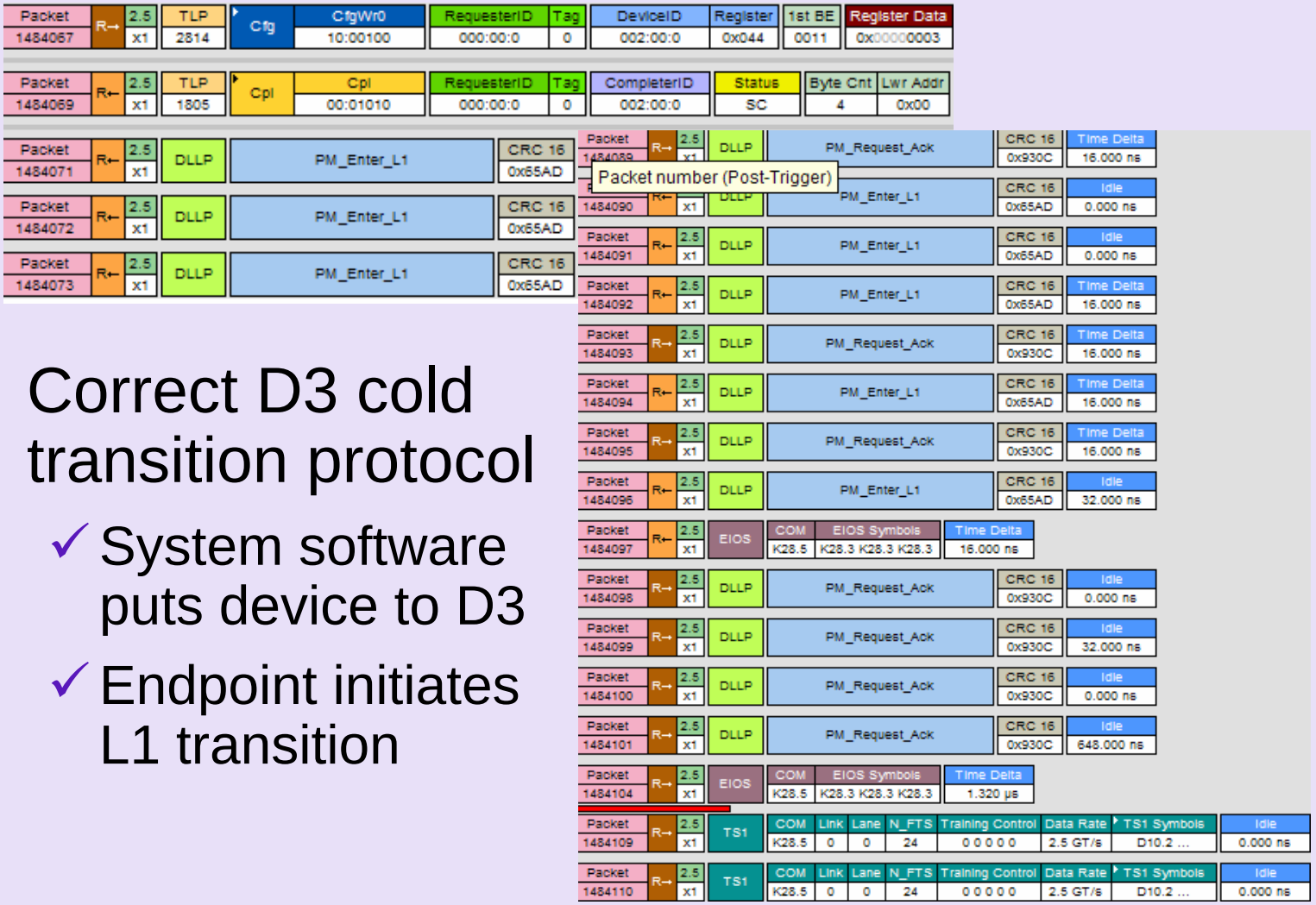
Agenda

- Deadlock issues in bus architectures
- Deadlocks in PCI Express architectures
- Deadlocks at common resource level
  - ✓ Example: ASPM L1 re-entry delay
- Deadlock at RC/chipset level
  - ✓ Example: PME fence mechanism hang-up
- Deadlock at device level – EP or RC
  - ✓ Example 1: Incompliant NAK
  - ✓ Example 2: CLKREQ# readiness
- Deadlock at PCIe link level
  - ✓ Example: Credit leakage

D3 Cold Transition Protocol

- System software puts device to D3
- Endpoint initiates L1 transition
- L1 exit, Link transitions back to L0
- System software/RC initiate PME fence mechanism
- Endpoint Acks PME Shutdown message
- Transition to L2/L3 ready
- L23 entry message ack'ed
- Link goes to electrical idle

D3 Cold Transition Protocol



- Correct D3 cold transition protocol
  - ✓ System software puts device to D3
  - ✓ Endpoint initiates L1 transition

D3 Cold Transition Protocol

- L1 exit, Link back to L0
- PME Fence mechanism
- L2/L3 ready transition

D3 Cold Transition Protocol

- L23 entry message ack'ed
- Link goes to electrical idle

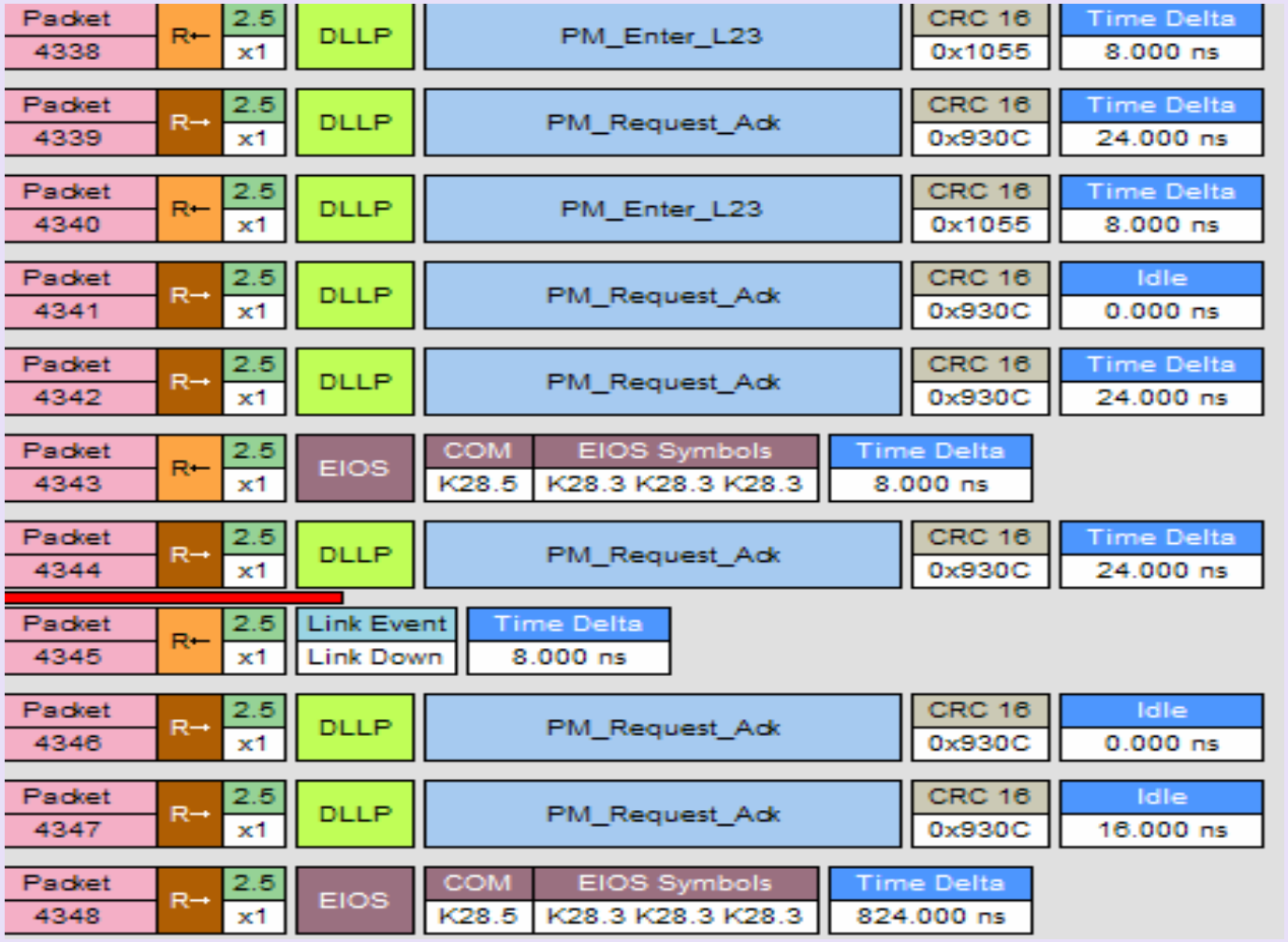
Power Sequencing Deadlock Scenario

- Power on
- Wait for the BIOS to kick in
- Toggle main supply/RESET
  - ✓ Typical for mobile systems
- Windows starts
- Disable the device in device manager, or go to Hibernate
- Machine hangs

Observed Sequence

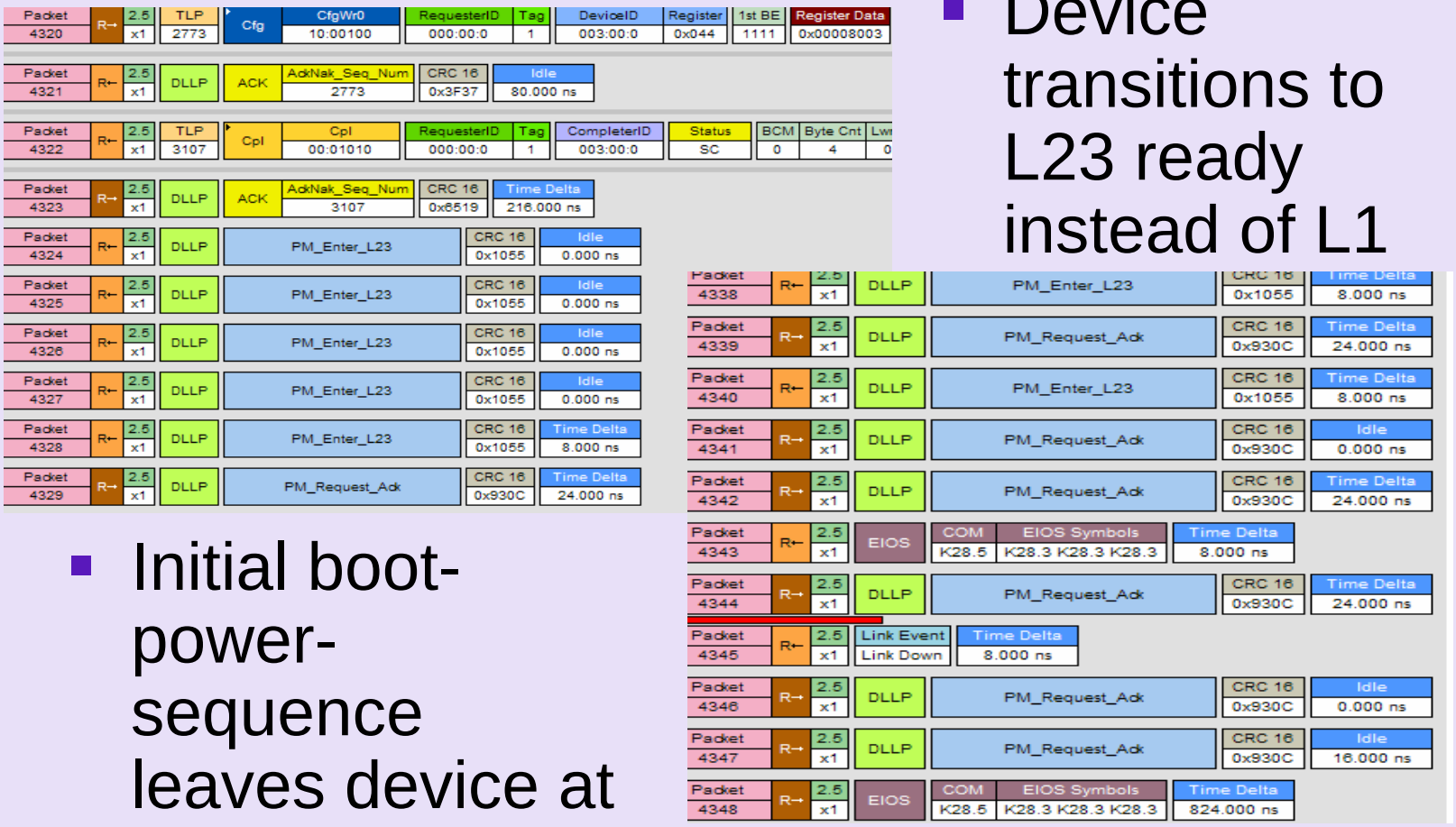
Packet												
4320	R→	2.5 x1	TLP 2773	Cfg	CfgWr0 10:00100	RequesterID 000:00:0	Tag 1	DeviceID 003:00:0	Register 0x044	1st BE 1111	Register Data 0x00008003	
4321	R←	2.5 x1	DLLP	ACK	AckNak_Seq_Num 2773	CRC 16 0x3F37	Idle 80.000 ns					
4322	R←	2.5 x1	TLP 3107	Cpl	Cpl 00:01010	RequesterID 000:00:0	Tag 1	CompleterID 003:00:0	Status SC	BCM 0	Byte Cnt 4	Lw
4323	R→	2.5 x1	DLLP	ACK	AckNak_Seq_Num 3107	CRC 16 0x6519	Time Delta 216.000 ns					
4324	R←	2.5 x1	DLLP	PM_Enter_L23		CRC 16 0x1055	Idle 0.000 ns					
4325	R←	2.5 x1	DLLP	PM_Enter_L23		CRC 16 0x1055	Idle 0.000 ns					
4326	R←	2.5 x1	DLLP	PM_Enter_L23		CRC 16 0x1055	Idle 0.000 ns					
4327	R←	2.5 x1	DLLP	PM_Enter_L23		CRC 16 0x1055	Idle 0.000 ns					
4328	R←	2.5 x1	DLLP	PM_Enter_L23		CRC 16 0x1055	Time Delta 8.000 ns					
4329	R→	2.5 x1	DLLP	PM_Request_Ack		CRC 16 0x930C	Time Delta 24.000 ns					

Observed Sequence



Packet						
4338	R←	2.5 x1	DLLP	PM_Enter_L23	CRC 16 0x1055	Time Delta 8.000 ns
4339	R→	2.5 x1	DLLP	PM_Request_Ack	CRC 16 0x930C	Time Delta 24.000 ns
4340	R←	2.5 x1	DLLP	PM_Enter_L23	CRC 16 0x1055	Time Delta 8.000 ns
4341	R→	2.5 x1	DLLP	PM_Request_Ack	CRC 16 0x930C	Idle 0.000 ns
4342	R→	2.5 x1	DLLP	PM_Request_Ack	CRC 16 0x930C	Time Delta 24.000 ns
4343	R←	2.5 x1	EIOS	COM K28.5	EIOS Symbols K28.3 K28.3 K28.3	Time Delta 8.000 ns
4344	R→	2.5 x1	DLLP	PM_Request_Ack	CRC 16 0x930C	Time Delta 24.000 ns
4345	R←	2.5 x1	Link Event Link Down	Time Delta 8.000 ns		
4346	R→	2.5 x1	DLLP	PM_Request_Ack	CRC 16 0x930C	Idle 0.000 ns
4347	R→	2.5 x1	DLLP	PM_Request_Ack	CRC 16 0x930C	Idle 16.000 ns
4348	R→	2.5 x1	EIOS	COM K28.5	EIOS Symbols K28.3 K28.3 K28.3	Time Delta 824.000 ns

PME Fence Mechanism Deadlock Summary



- Device transitions to L23 ready instead of L1
- Initial boot-power-sequence leaves device at D3-cold ready state

Power Sequencing Hangup Analysis

- Host tries to communicate while device in D3 cold
- RC blocks transactions since it has transitioned the link to L23
- Machine hangs on no completions coming back from the device
- Power down state machine should not transition to L23 before receiving PME_Turn_Off message and fully executing the PME fence mechanism

Agenda

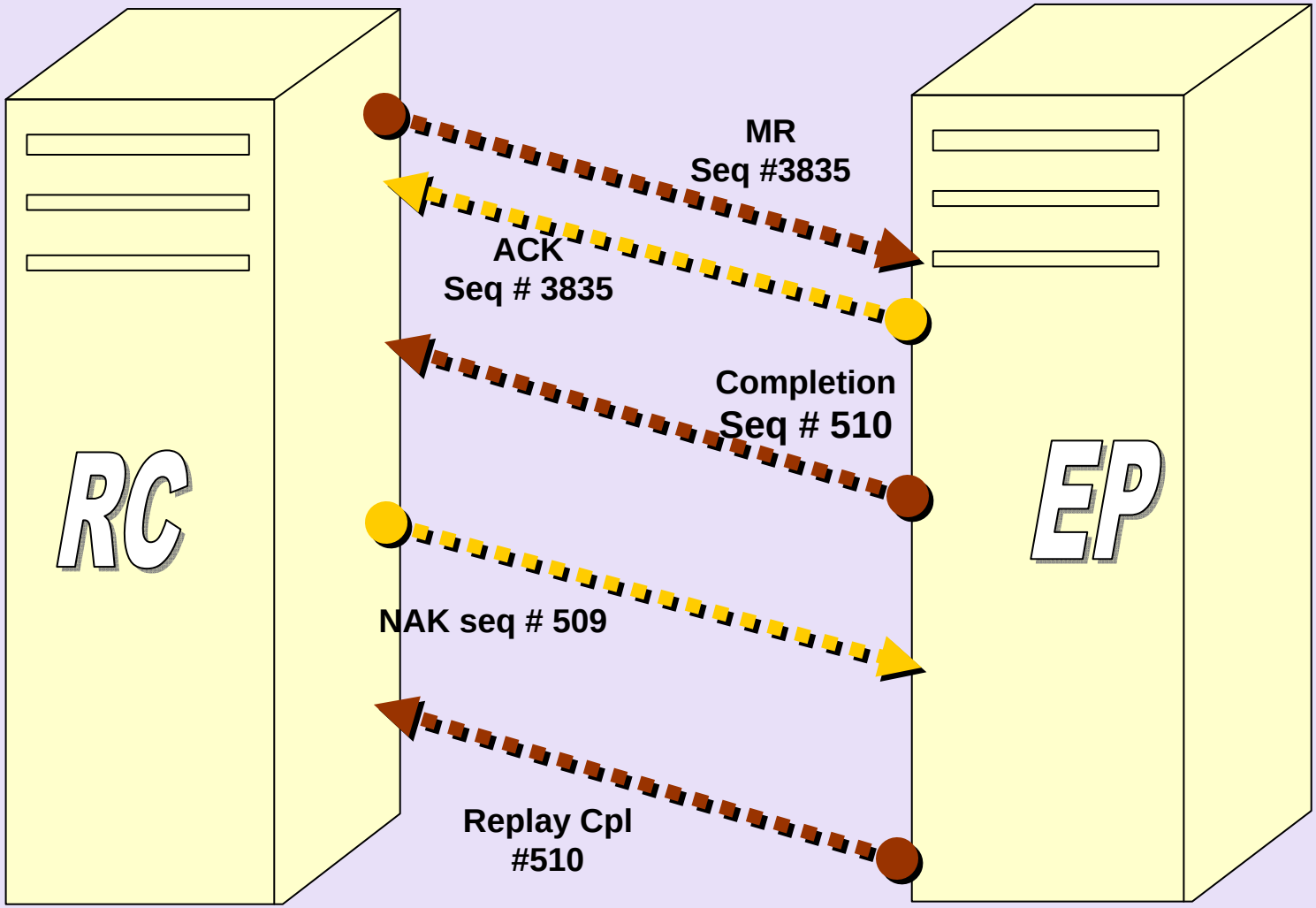
- Deadlock issues in bus architectures
- Deadlocks in PCI Express architectures
- Deadlocks at common resource level
  - ✓ Example: ASPM L1 re-entry delay
- Deadlock at RC/chipset level
  - ✓ Example: PME fence mechanism hang-up
- Deadlock at device level – EP or RC
  - ✓ Example 1: Incompliant NAK
  - ✓ Example 2: CLKREQ# readiness
- Deadlock at PCIe link level
  - ✓ Example: Credit leakage

Deadlock at the Device Level: Handling of Illegitimate NAK

- Compliant NAK
  - ✓ TLP # 510 Nak'd by NAK 509
  - ✓ TLP #510 is replayed

Packet																				
93203	R→	2.5 x1	TLP	Mem	MRd(32)	TC	TD	EP	Attributes	Length	RequesterID	Tag	Address	1st BE	Last BE	LCRC	Time Delta			
			3835		00:00000	0	0	0	01	1	000:00:0	1	FEAFF180	1111	0000	0xBD332CCA	272.000 ns			
Packet																				
93204	R→	2.5 x1	DLLP	ACK	AckNak_Seq_Num	CRC 16	Idle													
					3835	0x1139	56.000 ns													
Packet																				
93205	R→	2.5 x1	TLP	Cpl	CplD	TC	TD	EP	Attributes	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data		
			510		10:01010	0	0	0	01	1	000:00:0	1	002:00:0	SC	0	4	0x00	11040000		
Packet																				
93207	R→	2.5 x1	TLP	Cpl	CplD	TC	TD	EP	Attributes	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data		
			510		10:01010	0	0	0	01	1	000:00:0	1	002:00:0	SC	0	4	0x00	11040000		
Packet																				
93210	R→	2.5 x1	TLP	Cpl	CplD	TC	TD	EP	Attributes	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data		
			510		10:01010	0	0	0	01	1	000:00:0	1	002:00:0	SC	0	4	0x00	11040000		
Packet																				
93211	R→	2.5 x1	DLLP	NAK	AckNak_Seq_Num	CRC 16	Time Delta													
					509	0xCAC4	280.000 ns													
Packet																				
93214	R→	2.5 x1	TLP	Cpl	CplD	TC	TD	EP	Attributes	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data		
			510		10:01010	0	0	0	01	1	000:00:0	1	002:00:0	SC	0	4	0x00	11040000		
Packet																				
93215	R→	2.5 x1	DLLP	ACK	AckNak_Seq_Num	CRC 16	Idle													
					510	0xC28F	480.000 ns													
Packet																				
93216	R→	2.5 x1	DLLP	ACK	AckNak_Seq_Num	CRC 16	Idle													
					510	0xC28F	1.312 µs													

Compliant NAK



Compliant NAK

- ✓ NAK of TLP #866 Nak's TLPs # 867, #868
- ✓ TLP #867, 868 are replayed
  - Note: TLP# 869 sent before NAK processed

Packet															
3213	R←	2.5 x1	TLP 866	Mem	MWr(32) 10:00000	RequesterID 002:00:0	Tag 0	Address 06037880	1st BE 1111	Last BE 1111	Data 32 dwords	LCRC 0xC872881E	Time Delta 952.000 ns		
3214	R←	2.5 x1	DLLP	ACK	AckNak_Seq_Num 866	CRC 16 0x36A9	Time Delta 376.000 ns								
3215	R←	2.5 x1	TLP 867	Mem	MWr(32) 10:00000	RequesterID 002:00:0	Tag 0	Address 06037900	1st BE 1111	Last BE 1111	Data 32 dwords	LCRC 0x59E9112C	Idle 720.000 ns		
3216	R←	2.5 x1	TLP 868	Mem	MWr(32) 10:00000	RequesterID 002:00:0	Tag 0	Address 06037980	1st BE 1111	Last BE 1111	Data 32 dwords	LCRC 0x6F85E849	Time Delta 1.128 µs		
3217	R←	2.5 x1	DLLP	NAK	AckNak_Seq_Num 866	CRC 16 0xDDCE	Time Delta 168.000 ns								
3218	R←	2.5 x1	TLP 869	Mem	MWr(32) 10:00000	RequesterID 002:00:0	Tag 0	Address 06037A00	1st BE 1111	Last BE 1111	Data 32 dwords	LCRC 0x46BE0908	Idle 0.000 ns		
3219	R←	2.5 x1	TLP 867	Mem	MWr(32) 10:00000	RequesterID 002:00:0	Tag 0	Address 06037900	1st BE 1111	Last BE 1111	Data 32 dwords	LCRC 0x59E9112C	Idle 0.000 ns		
3220	R←	2.5 x1	TLP 868	Mem	MWr(32) 10:00000	RequesterID 002:00:0	Tag 0	Address 06037980	1st BE 1111	Last BE 1111	Data 32 dwords	LCRC 0x6F85E849	Idle 24.000 ns		

Incompliant NAK

Packet			TLP		MRd(32)	RequesterID	Tag	Address	1st BE	Last BE	LCRC	Time Delta
75124	R→	2.5 x1	576	Mem	00:00000	000:00:0	1	FEAFF180	1111	0000	0x0BD503A0	272.000 ns

Packet			DLLP		AckNak_Seq_Num	CRC 16	Idle
75125	R←	2.5 x1	DLLP	ACK	576	0xC655	56.000 ns

Packet			TLP		CplD	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	LCRC	Time Delta
75126	R←	2.5 x1	3852	Cpl	10:01010	000:00:0	1	002:00:0	SC	0	4	0x00	1 dword	0x99A353D5	472.000 ns

Packet			DLLP		AckNak_Seq_Num	CRC 16	Idle
75127	R→	2.5 x1	DLLP	ACK	3852	0xC912	224.000 ns

Packet			TLP		MWr(32)	RequesterID	Tag	Address	1st BE	Last BE	Data	LCRC	Time Delta
75128	R→	2.5 x1	577	Mem	10:00000	000:00:0	0	FEAFF180	1111	0000	1 dword	0x1AAAC0DA	288.000 ns

Packet			DLLP		AckNak_Seq_Num	CRC 16	Time Delta
75129	R←	2.5 x1	DLLP	ACK	577	0x674E	2.337 ms

Packet			DLLP		AckNak_Seq_Num	CRC 16	Time Delta
75130	R→	2.5 x1	DLLP	NAK	3852	0x2275	2.764 ms

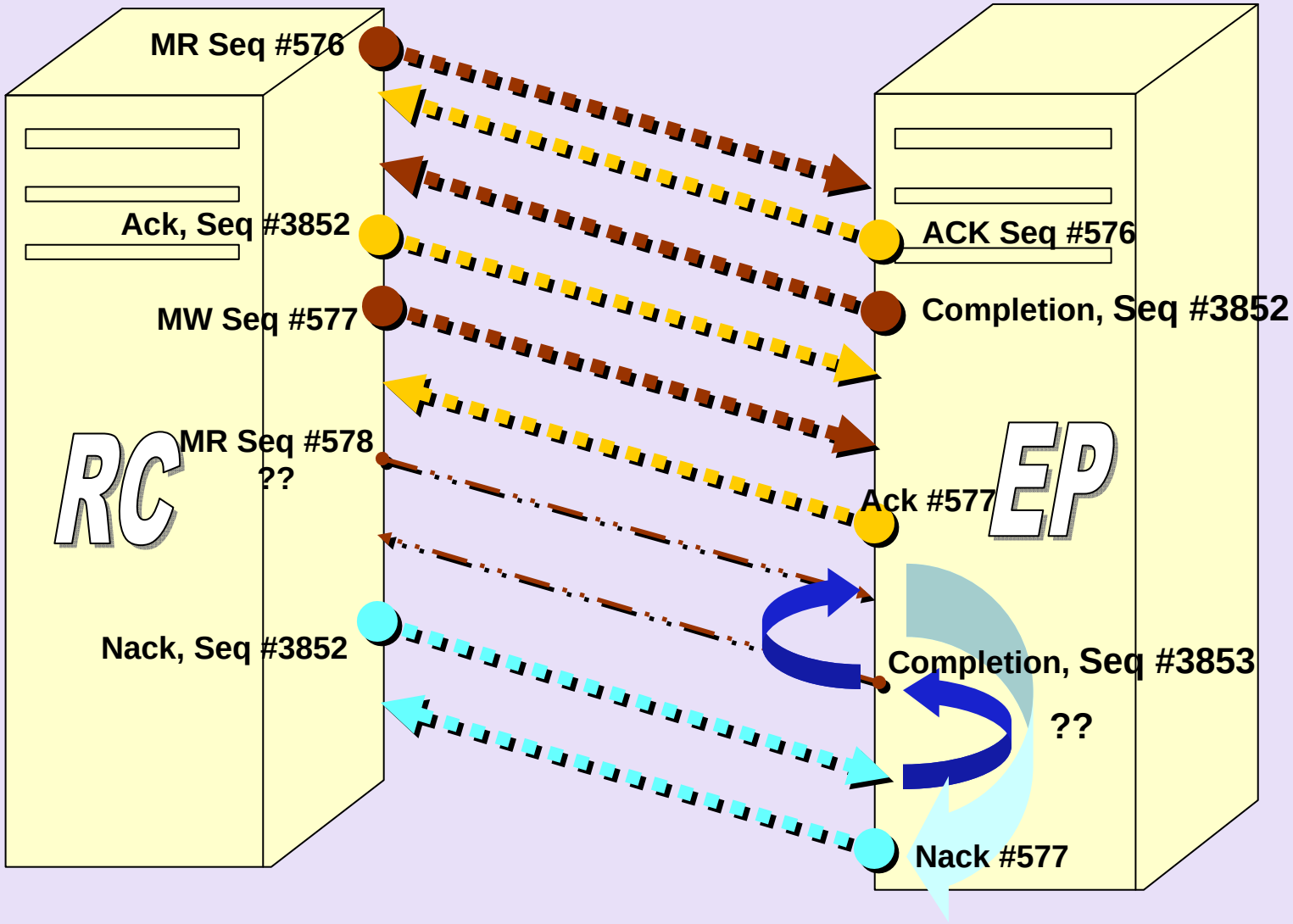
Packet			DLLP		AckNak_Seq_Num	CRC 16	Time Delta
75131	R←	2.5 x1	DLLP	NAK	577	0x8C29	345.602 ms

Packet			TLP		MRd(32)	RequesterID	Tag	Address	1st BE	Last BE	LCRC	Time Delta
75132	R→	2.5 x1	578	Mem	00:00000	000:00:0	0	FEAFF180	1111	0000	0xE5B3030B	256.000 ns

Incompliant NAK

- Undelivered TLP #3853 Nak'd by NAK #3852
  - ✓ NAK applies to completion for TLP never sent?
- EP interprets NAK #3853 as for bad completion it never sent
- EP alerts RC by NAK #577 that if it expects completion #3853 it should replay Request #578
- RC hangs for 345 msec processing NAK #577
- RC finally plays request #578
- Repeated incompliant NAKs lead to deadlock

Incompliant NAK deadlock



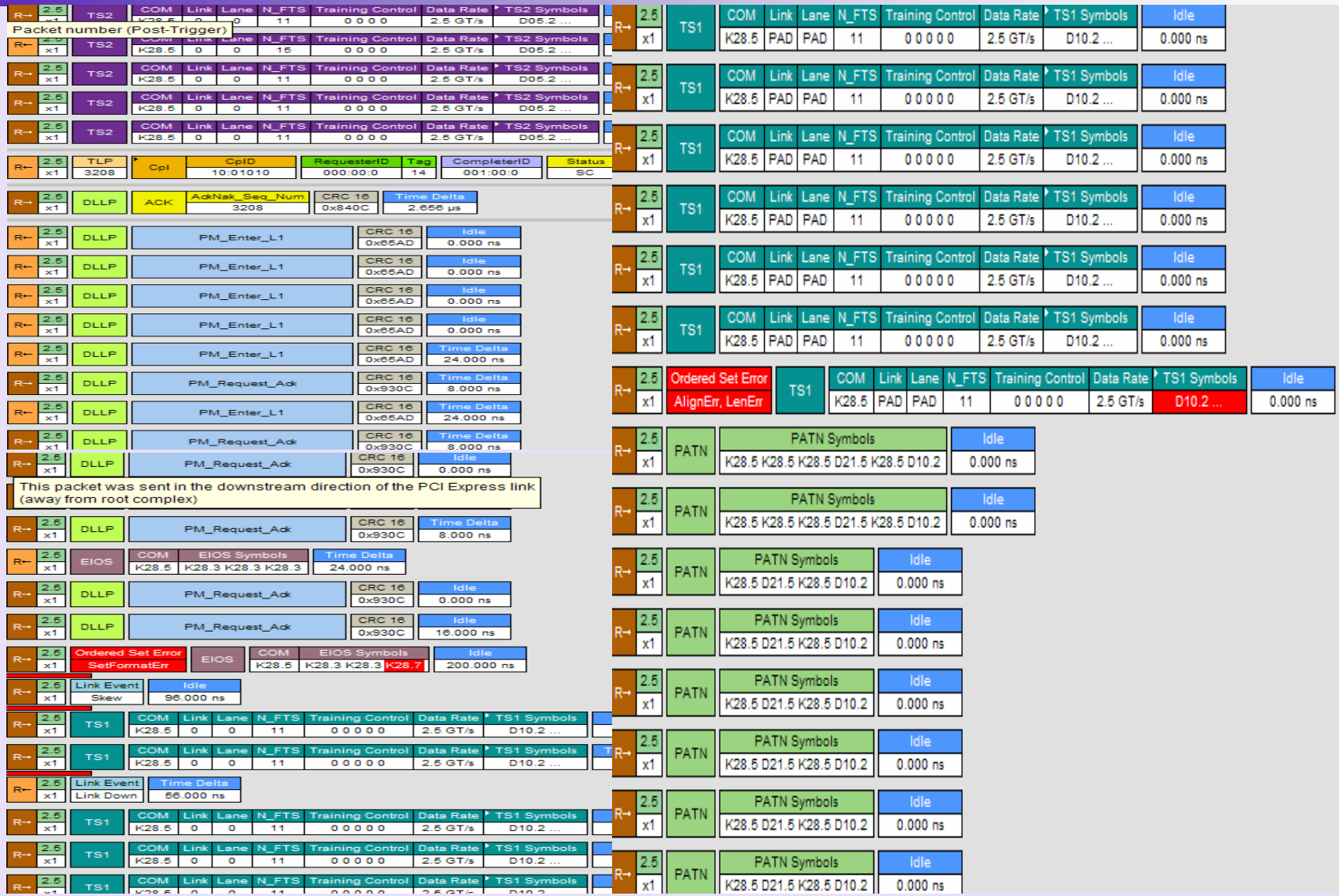
Agenda

- Deadlock issues in bus architectures
- Deadlocks in PCI Express architectures
- Deadlocks at common resource level
  - ✓ Example: ASPM L1 re-entry delay
- Deadlock at RC/chipset level
  - ✓ Example: PME fence mechanism hang-up
- Deadlock at device level – EP or RC
  - ✓ Example 1: Incompliant NAK
  - ✓ Example 2: CLKREQ# readiness
- Deadlock at PCIe link level
  - ✓ Example: Credit leakage

ASPM L1 and Clock Power Management

- L1 exit latency does not account for CLKREQ# enabled, used in miniExpress cards to shut down REFCLK# and further power savings
- Upon deassertion of CLKREQ# EP shuts down PLL in preparation for the loss of REFCLK#
- Device does not account for its own increased PLL recovery time
- System timeout due to long L1 recovery
- PCIe 2.0 requires components to account for additional latency due to loss of REFCLK#
- Both EP and RC need to account for additional latency due to clock power management
- Use backup slow clock to replace bit clock when CLKREQ# de-asserted

ASPM L1 and Clock Power Management



Agenda

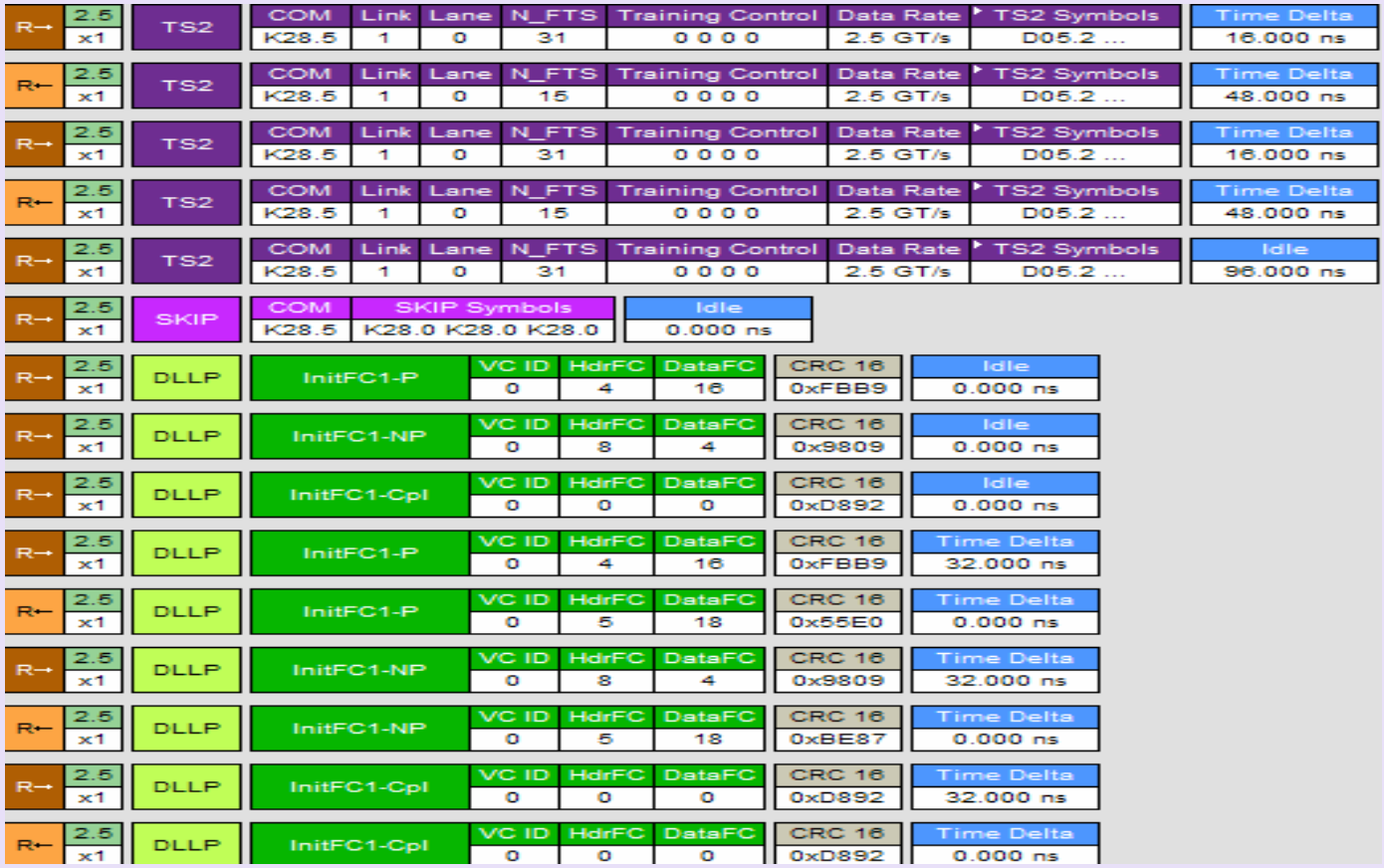
- Deadlock issues in bus architectures
- Deadlocks in PCI Express architectures
- Deadlocks at common resource level
  - ✓ Example: ASPM L1 re-entry delay
- Deadlock at RC/chipset level
  - ✓ Example: PME fence mechanism hang-up
- Deadlock at device level – EP or RC
  - ✓ Example 1: Incompliant NAK
  - ✓ Example 2: CLKREQ# readiness
- Deadlock at PCIe link level
  - ✓ Example: Credit leakage

Deadlock Associated with Particular PCIe Link

- Credit starvation deadlocks
  - ✓Malformed TLP going upstream
  - ✓Downstream device accounts for credit
  - ✓Upstream device does not account for credit
  - ✓Additional malformed TLPs lead to credit starvation

Malformed TLP Deadlock

- **Downstream FC Init for posted transactions is 16, or 16 x 16 = 256 bytes**



R→	2.5 x1	TS2	COM	Link	Lane	N_FTS	Training Control	Data Rate	TS2 Symbols	Time Delta	
			K28.5	1	0	31	0 0 0 0	2.5 GT/s	D05.2 ...	16.000 ns	
R←	2.5 x1	TS2	COM	Link	Lane	N_FTS	Training Control	Data Rate	TS2 Symbols	Time Delta	
			K28.5	1	0	15	0 0 0 0	2.5 GT/s	D05.2 ...	48.000 ns	
R→	2.5 x1	TS2	COM	Link	Lane	N_FTS	Training Control	Data Rate	TS2 Symbols	Time Delta	
			K28.5	1	0	31	0 0 0 0	2.5 GT/s	D05.2 ...	16.000 ns	
R←	2.5 x1	TS2	COM	Link	Lane	N_FTS	Training Control	Data Rate	TS2 Symbols	Time Delta	
			K28.5	1	0	15	0 0 0 0	2.5 GT/s	D05.2 ...	48.000 ns	
R→	2.5 x1	TS2	COM	Link	Lane	N_FTS	Training Control	Data Rate	TS2 Symbols	Idle	
			K28.5	1	0	31	0 0 0 0	2.5 GT/s	D05.2 ...	96.000 ns	
R→	2.5 x1	SKIP	COM	SKIP Symbols			Idle				
			K28.5	K28.0 K28.0 K28.0			0.000 ns				
R→	2.5 x1	DLLP	InitFC1-P	VC ID	HdrFC	DataFC	CRC 16	Idle			
				0	4	16	0xFBB9	0.000 ns			
R→	2.5 x1	DLLP	InitFC1-NP	VC ID	HdrFC	DataFC	CRC 16	Idle			
				0	8	4	0x9809	0.000 ns			
R→	2.5 x1	DLLP	InitFC1-Cpl	VC ID	HdrFC	DataFC	CRC 16	Idle			
				0	0	0	0xD892	0.000 ns			
R→	2.5 x1	DLLP	InitFC1-P	VC ID	HdrFC	DataFC	CRC 16	Time Delta			
				0	4	16	0xFBB9	32.000 ns			
R←	2.5 x1	DLLP	InitFC1-P	VC ID	HdrFC	DataFC	CRC 16	Time Delta			
				0	5	18	0x55E0	0.000 ns			
R→	2.5 x1	DLLP	InitFC1-NP	VC ID	HdrFC	DataFC	CRC 16	Time Delta			
				0	8	4	0x9809	32.000 ns			
R←	2.5 x1	DLLP	InitFC1-NP	VC ID	HdrFC	DataFC	CRC 16	Time Delta			
				0	5	18	0xBE87	0.000 ns			
R→	2.5 x1	DLLP	InitFC1-Cpl	VC ID	HdrFC	DataFC	CRC 16	Time Delta			
				0	0	0	0xD892	32.000 ns			
R←	2.5 x1	DLLP	InitFC1-Cpl	VC ID	HdrFC	DataFC	CRC 16	Time Delta			
				0	0	0	0xD892	0.000 ns			

Malformed TLP Deadlock

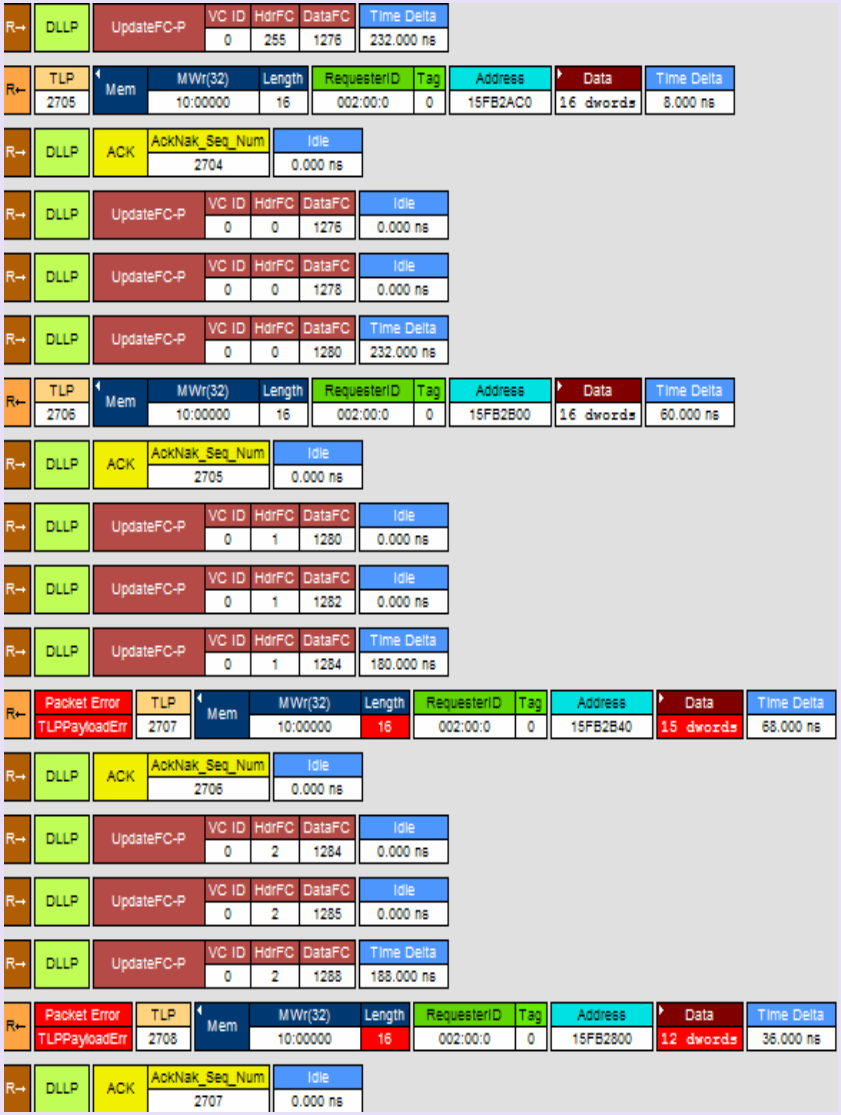
- RC advertises buffer space for a request for 4 x 16 = 64 posted data bytes
- EP sends MW request for 16 x 4 = 64 bytes
- RC consumes 64 bytes and advertises release of buffer space
- EP sends additional memory write request
- RC receives a malformed TLP
- RC does not consume, does not update
- Result is credit leakage

Malformed TLP Deadlock

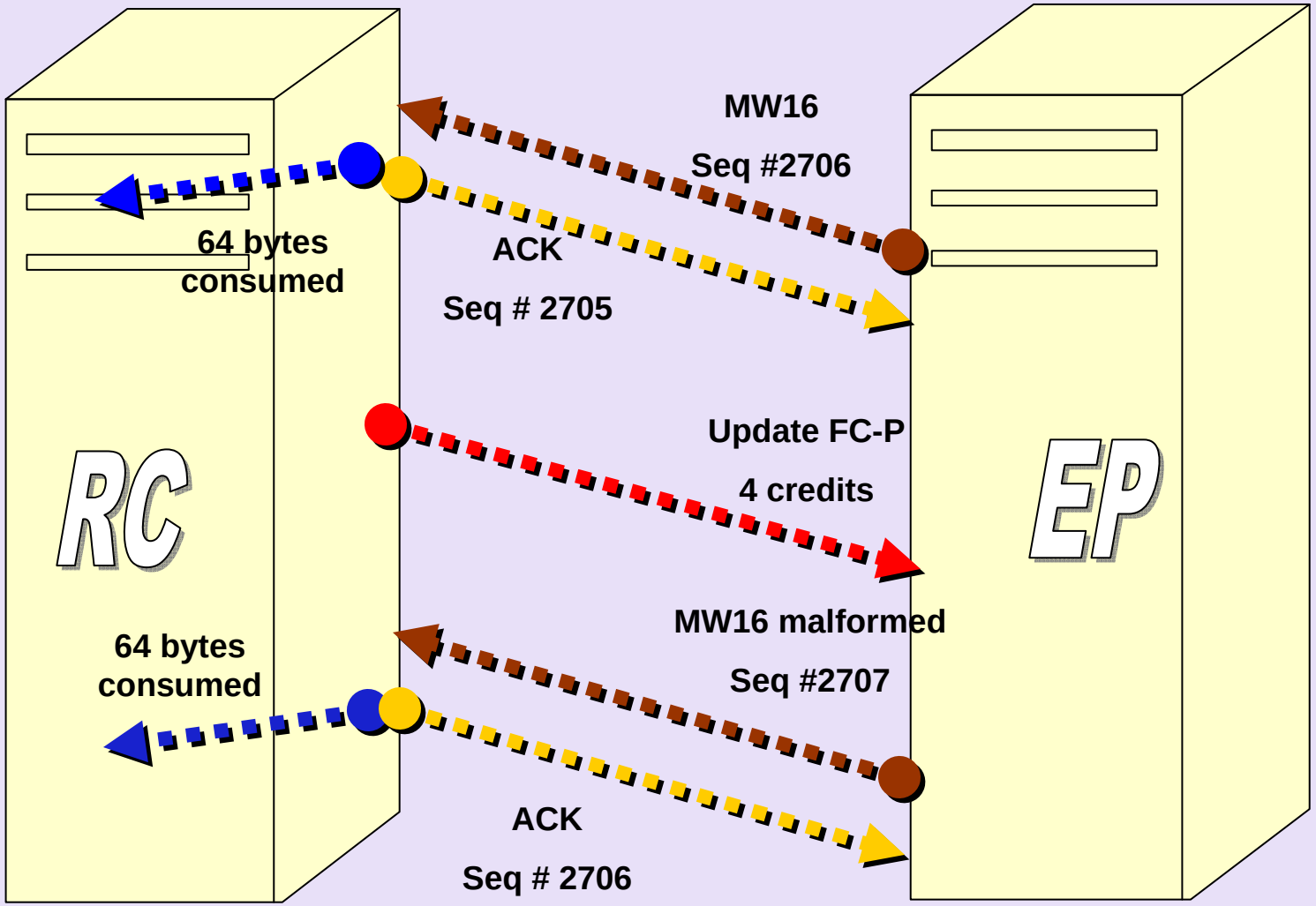
- Memory Write TLPs require free resource of 16 x 4 = 64 bytes

  = 4 credits

- TLP 2705 is sent, but only TLP 2704 is ack'ed
- 4 credits are consumed and FC updated
- EP lags 4 credits behind RC
- Each malformed TLP adds additional lag of 4 credits

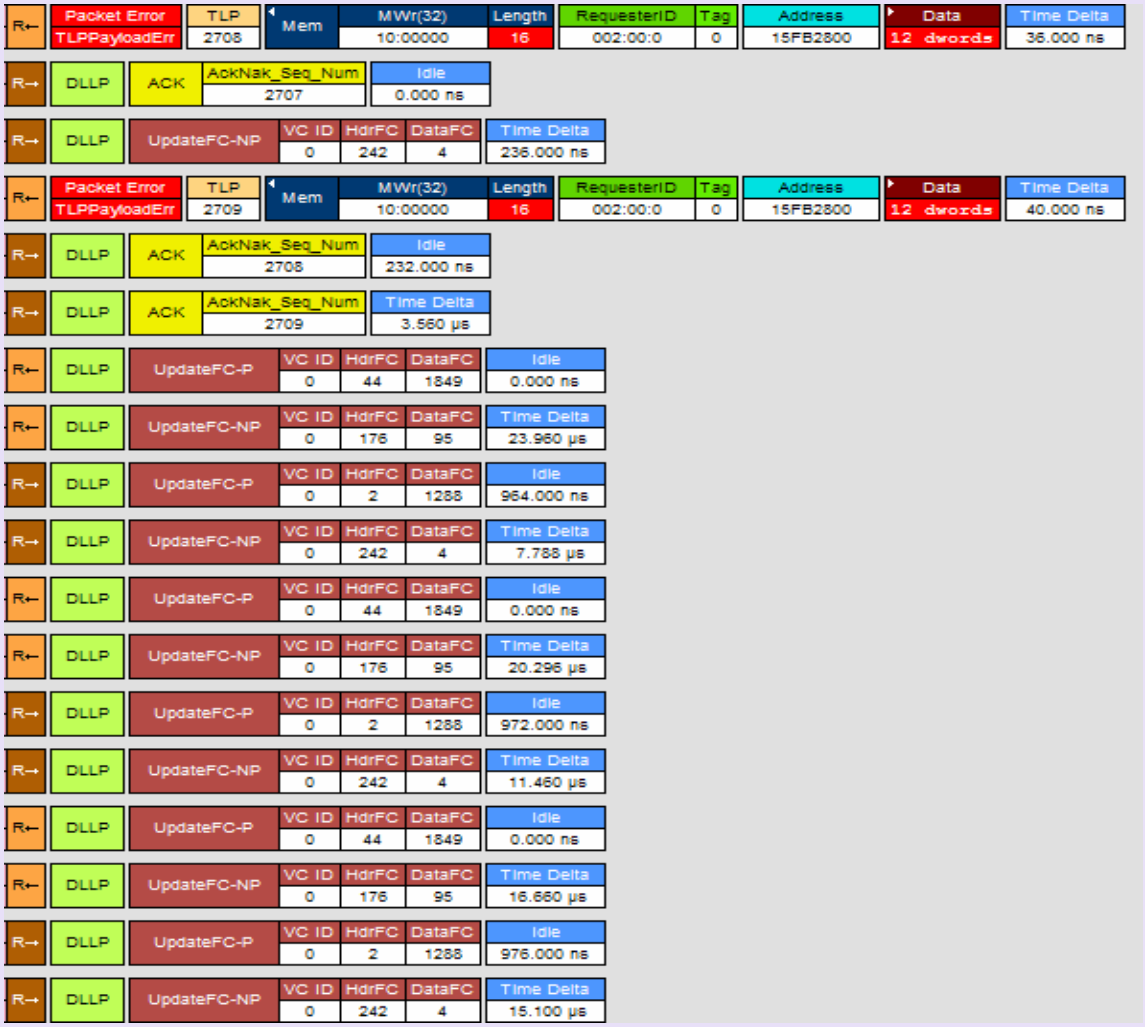


Malformed TLP Deadlock

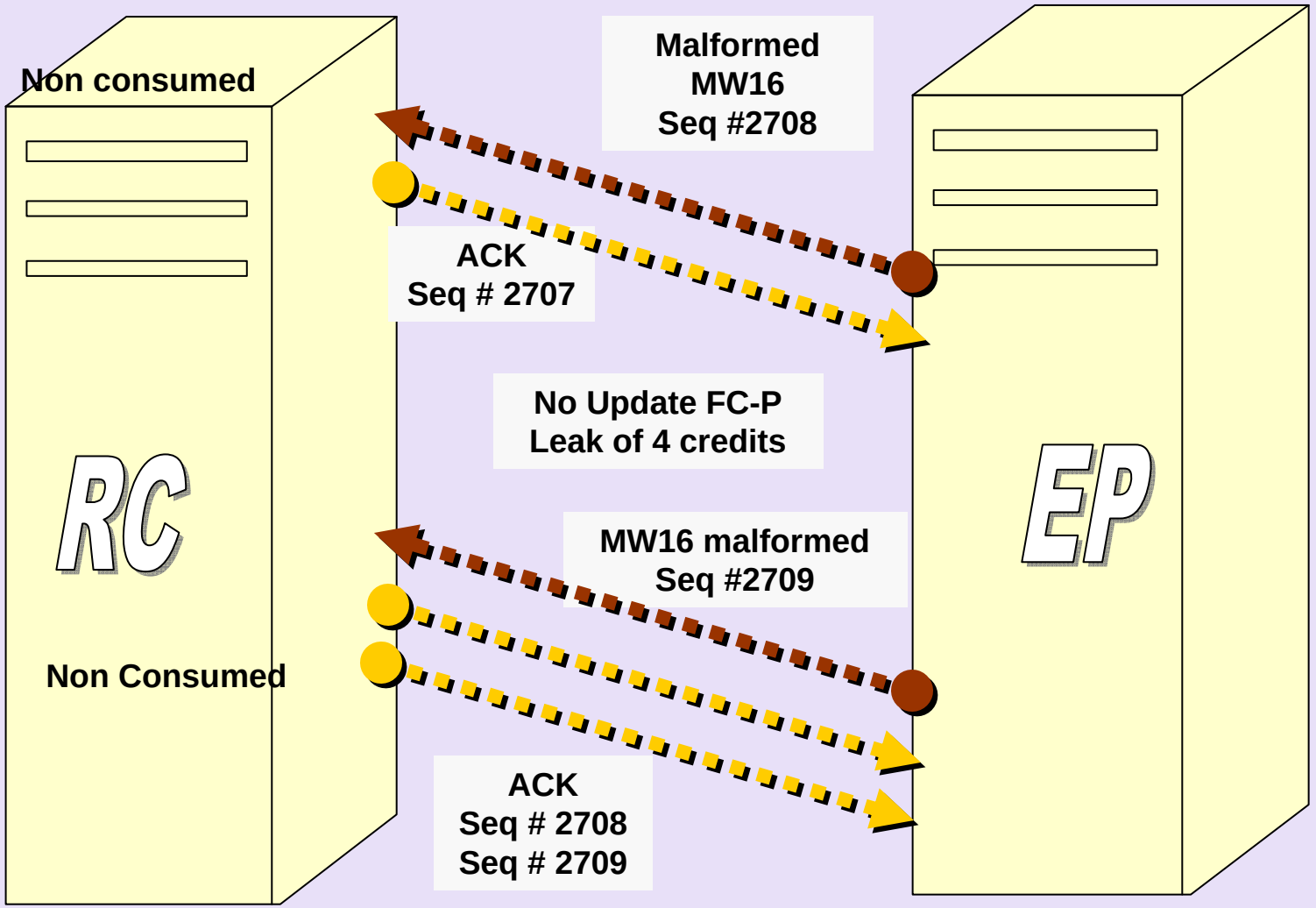


Malformed TLP Deadlock

- 3 malformed TLPs create a lag of 12 credits
- Leads to EP credit starvation and blocks further traffic
- Deadlock scenario resulting from credit starvation



Malformed TLP Deadlock

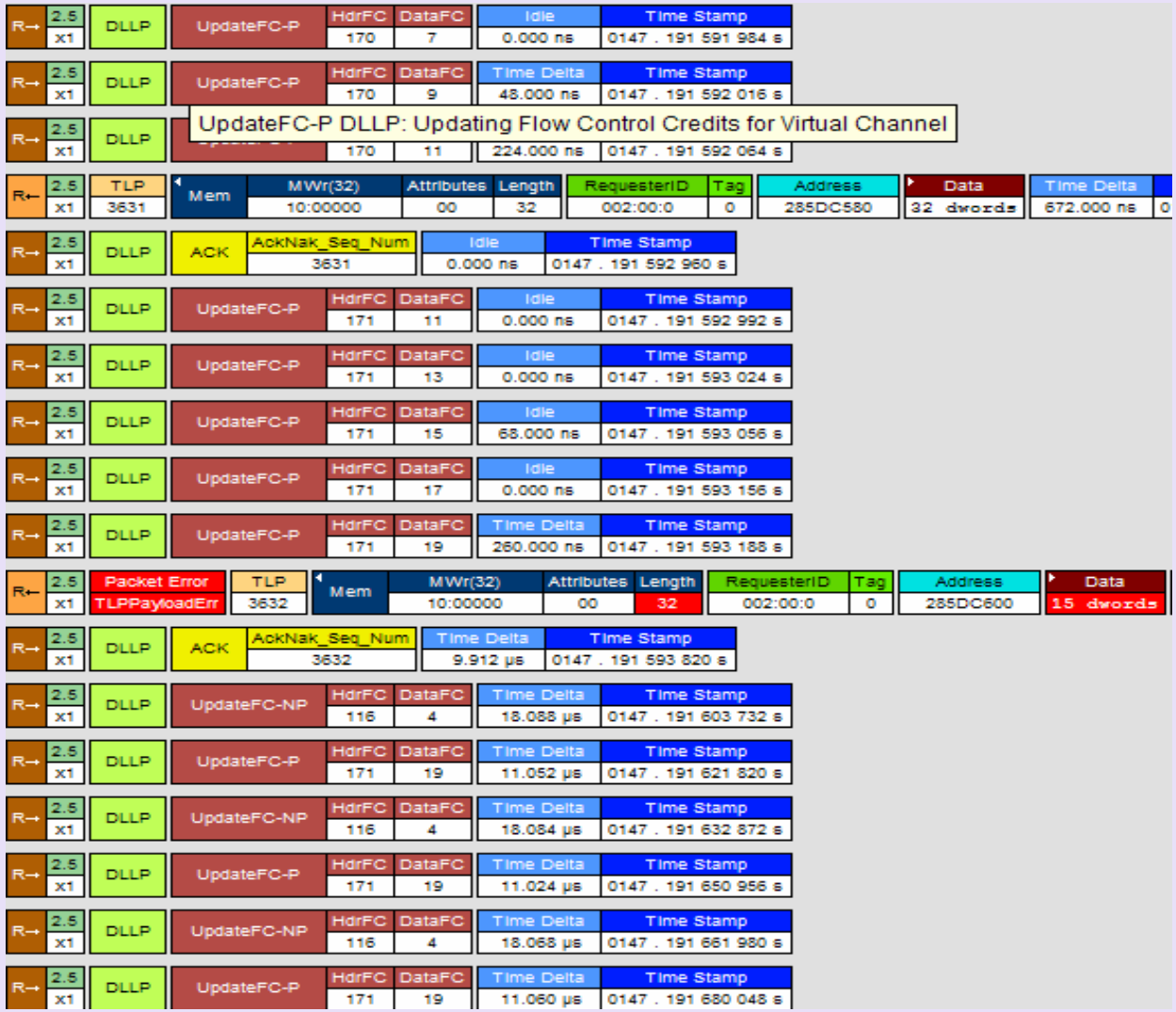


Malformed TLP Deadlock

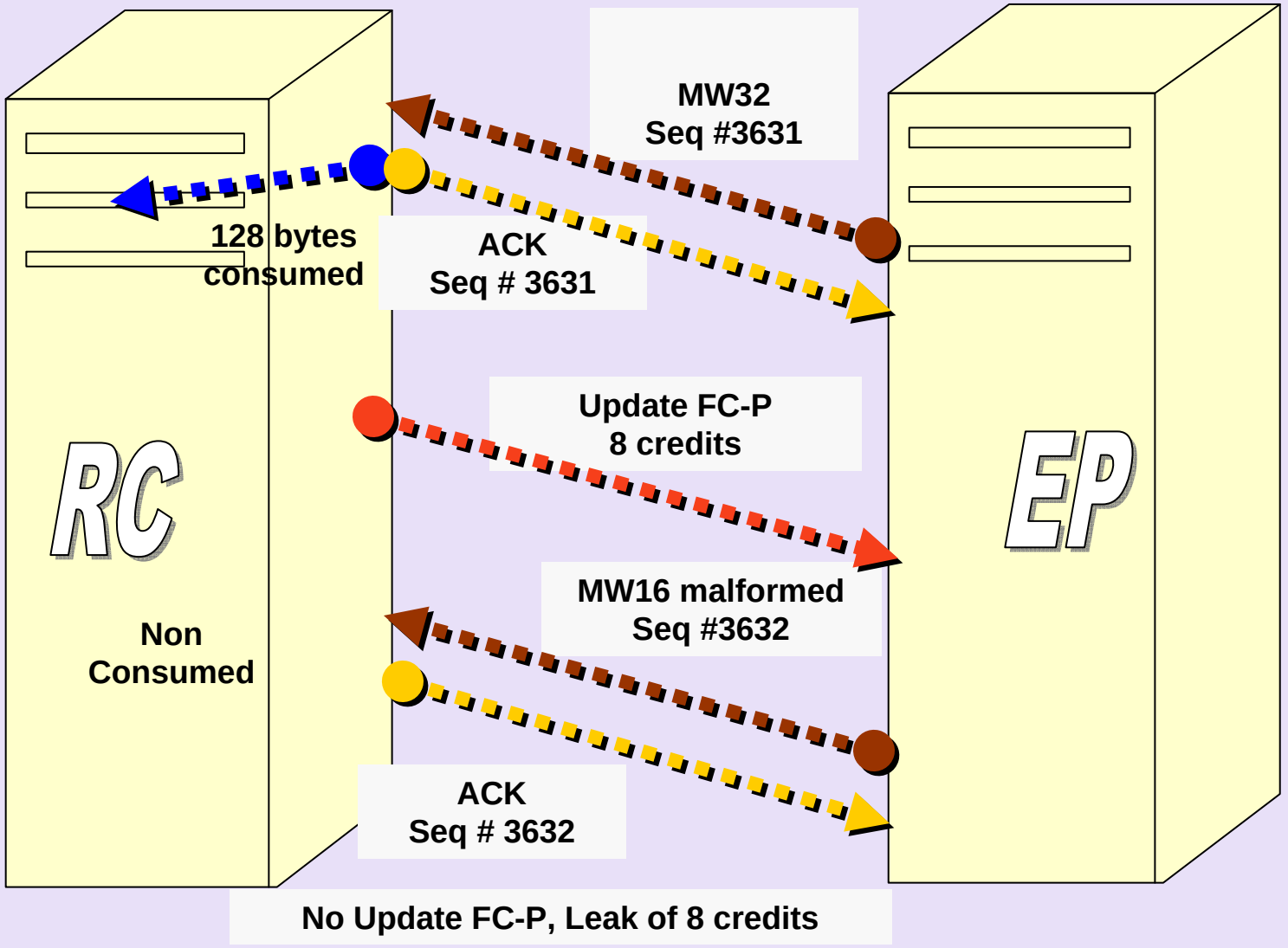
- TLP of length 32 requires

  32 * 4 / 16

  = 8 credits
- No Updates
- Credit starvation will occur with only 2 malformed TLPs



Malformed TLP Deadlock



Malformed TLP Deadlock

- Error reporting
  - ✓ Malformed TLP Error reporting message required for downstream TLP, meaningless for upstream TLP
- Specification enhancement recommendation
  - ✓ Downstream FC Init or link retrain should be mandatory following malformed TLP, or other scenarios leading to credit starvation
  - ✓ ECR proposal under study

Thank you for attending the PCI-SIG Developers Conference Europe 2007.

For more information please go to www.pcisig.com

Deadlock Avoidance in PCI Express Based Architectures

Isaac Livny

Applications Engineer

LSI Corporation