Flash Memory Summit

Challenges of Testing PCIe Gen. 4 SSDs *and Beyond*

Justin Treon

Advantest

Agenda

- MPT3000 SSD Tester
- Implementing PCIe Gen. 4 Early
- Specification Issues
- Testing Issues
- Test Compliance Issues
- PCIe Gen. 5 and 6
- SAS-4 and SAS-5

Flash Memory Summit

Our PCIe Gen. 4 Product

- The MPT3000 is a multi-protocol SSD tester
  - FPGA based implementation allows for the protocol switches

Risking Early Development

- Hardware developed for the 0.7 physical layer of the PCIe Gen. 4 specification
- By using the FPGA we were able to test before the Gen. 4 devices became available

Testing the Link

- No devices to test the link available

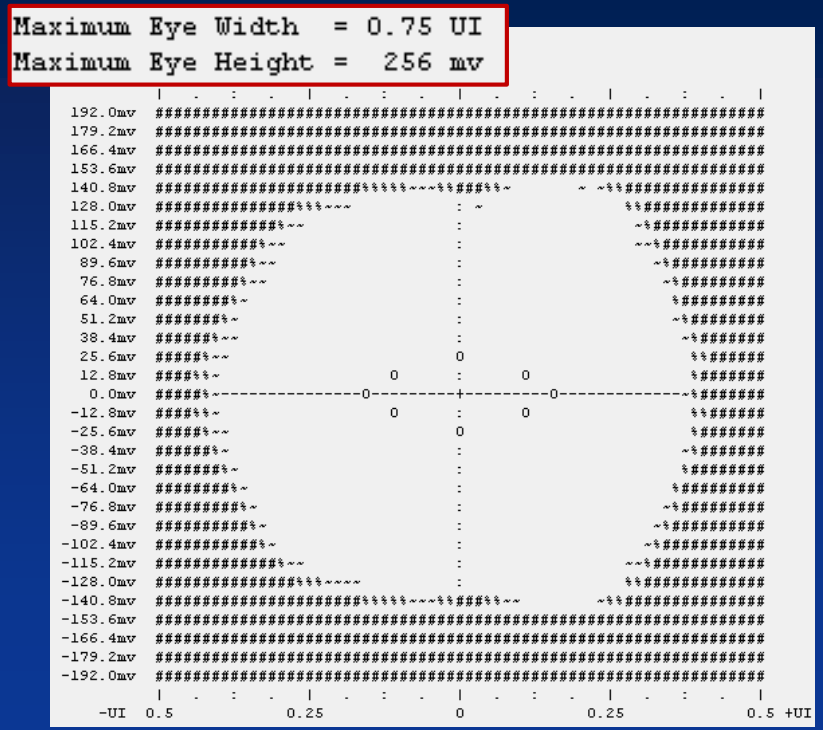
  … so we made our own

  - To be an early adopter you need to build you own test equipment
- Test board has many debug features to test the link

Flash Memory Summit

Eye Diagram

- Eye Diagram tool added to debug the link
  - Similar to Rx Margin (a.k.a. Lane Margining)
- Adding eye diagram to SSD controllers is suggested

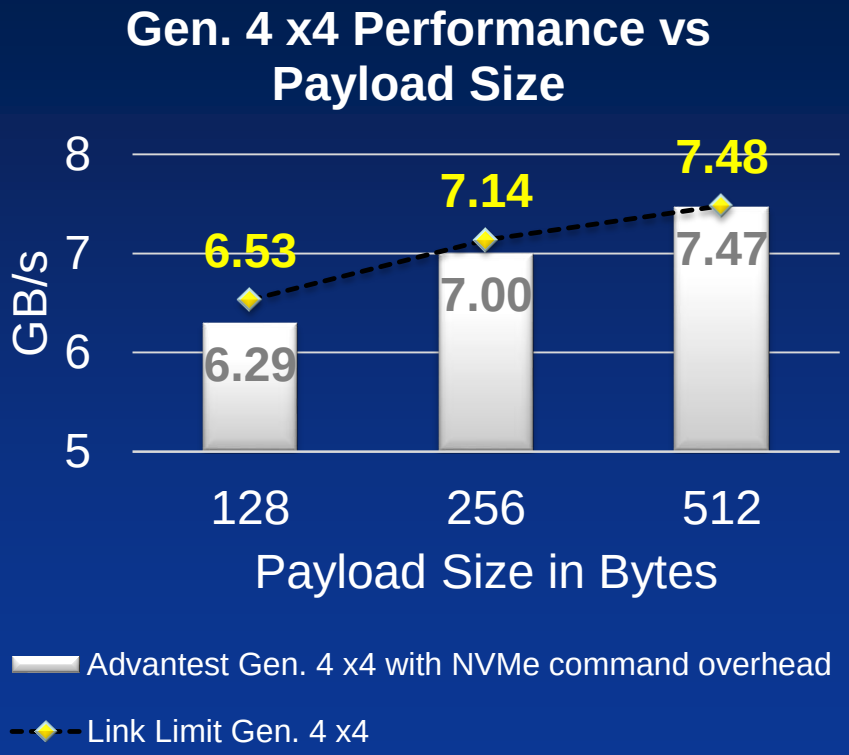


Changes to the Specification

- The optional lane margining feature was made mandatory at the last moment … *surprise*
- Causing implementation mismatch requiring rearchitecting
  - Budget time and resources for specification changes

Testing the Performance

- No SSD available to test with
  … so we made our own
- DMA transfers simulated the NVMe command
  - Setting data engine speed
  - Determining performance with NVMe command overhead



PCI SIG Compliance

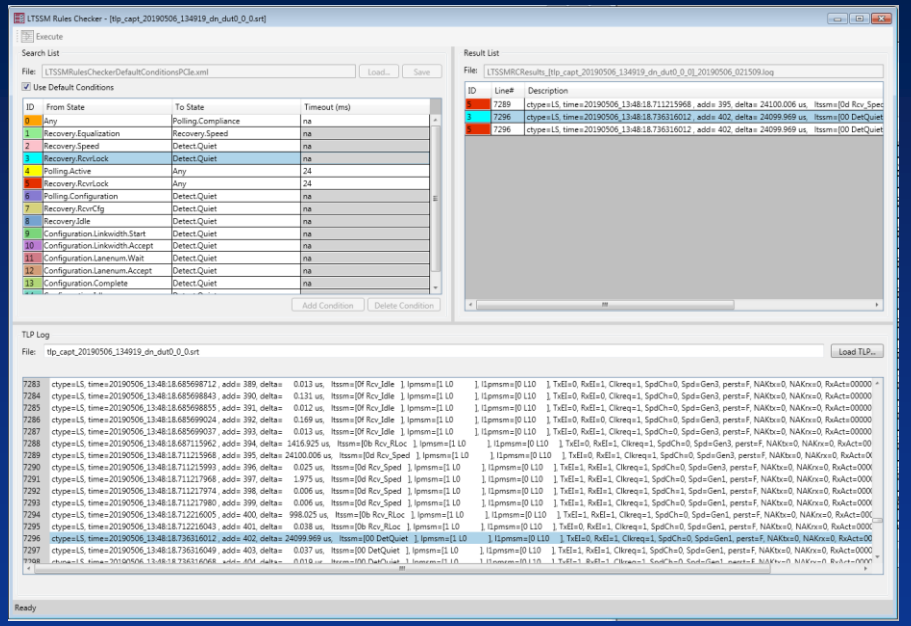
- Test systems were not ready
  - False positives for pass and fail
  - Engineering resources used to correct tests
- Test specification as of last plug fest
  - Not possible to obtain PCI SIG certification

Hurry Up and Wait

- PCIe Gen. 4 market intercept pushed out from original expectations
- Extra time allowed for feature development and stability improvements

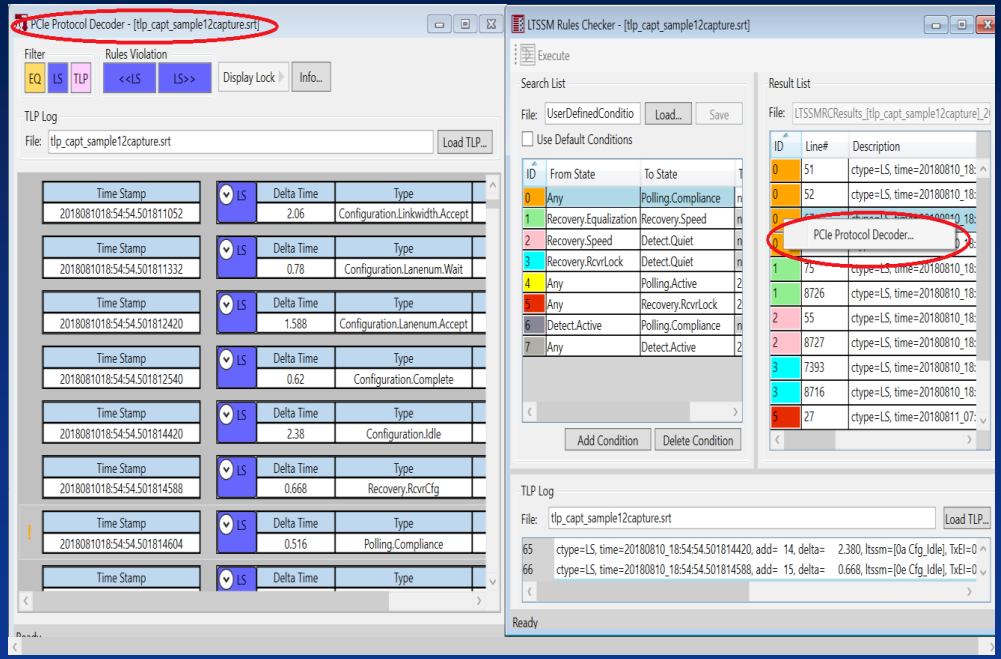
TCDT (Traffic Capture and Debug Tool)

- Traffic Capture and Debug Tool
  - Suite of user friendly graphical tools for parsing and debugging captures
  - LTSSM Rules Checker
    - Tools allow users to parse and analyze logs in a quick and user friendly manner



TCDT (2)

- TLP Capture
  - Packet capture at the TLP layer and above
- PCIe Protocol Decoder
  - Protocol Decoder allows the user to graphically analyze the traffic



Flash Memory Summit

TLP Capture Example (1)

- PCIe link framing
  - Test Log

```
Minimize Voltage and disconnect!

************************************
PCIe Link Status :
LinkUp L0        = true
LTSSM            = 0x10, LTSSM_L0
LinkWidth        = x4
LinkSpeed        = 8.0G
ActiveLanes      = 00001111 (0x0f)
ValidLanes       = 00001111 (0x0f)
LinkUpCount      = 1
LinkRetrainCnt   = 22
**************************
Sun Jan 27 23:54:47 2019 
Error in connecting and n

Power up PCIe levels! ERROR in setting the PCIe Power Level.
```

link retrain count 22 > 20.

Example of link framing failure

  - Log from TLP Capture

```
delta=473746.780 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.005 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.005 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.005 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.945 us,   ltssm=[0e Rcv_RCfg   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.240 us,   ltssm=[0f Rcv_Idle   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.340 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=122430.750 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.005 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.005 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.005 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.940 us,   ltssm=[0e Rcv_RCfg   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.245 us,   ltssm=[0f Rcv_Idle   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.335 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta= 38072.200 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.005 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.005 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.005 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.970 us,   ltssm=[0e Rcv_RCfg   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.240 us,   ltssm=[0f Rcv_Idle   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.340 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta= 38072.180 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.005 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
delta=     0.005 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0          ], l1pmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Gen3, perst=F, NAKtx=0, NAKrx=0, RxAct=111
```

```
l0recovery=[ framing ]
l0recovery=[ directed ]
```

```
l0recovery=[ framing ]
l0recovery=[ directed ]
```

```
l0recovery=[ framing ]
l0recovery=[ directed ]
```

```
l0recovery=[ framing ]
l0recovery=[ directed ]
```

TLP Capture Example (2)

- No block device not ready
  - Test Log

```
Minimize Voltage and disconnect!

**************************************
PCIe Link Status :
LinkUp L0       = true
LTSSM           = 0x10, LTSSM_L0
LinkWidth       = x4
LinkSpeed       = 8.0G
ActiveLanes     = 00001111 (0x0f)
ValidLanes      = 00001111 (0x0f)
LinkUpCount     = 1
LinkRetrainCnt  = 3
Sun Mar 24 01:55:04 2019 ERROR : Block Device is not present.
find_protocol_fpga_virtual_ep: ---> bus (number=14 primary=12 s
nvme_wait_ready: Error: device not ready after 20500
nvme_probe: nvme_configure_admin_queue failed, result
```

  - Log from TLP

```
Block Device is not present.
```

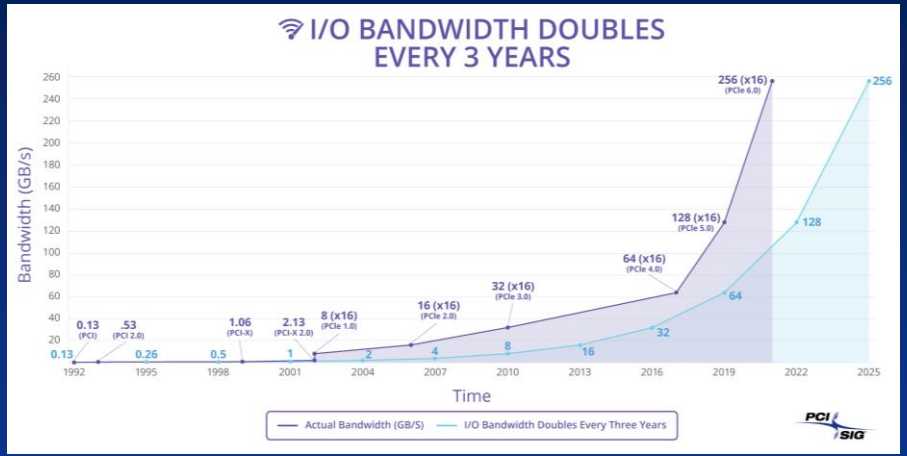
```
20190324_01:54:12.277536572 [ 142]: -->DN1[ Mem32 RdReq ] 'h00000001 'h0000000f 'hca01001c ..........  .......... (3:11,0,1.12) LEN=4    CPU
20190324_01:54:12.277537080 [ 138]: <--DN1[ Cpl w/ data ] 'h4a000001 'h0e000004 'h0000001c 0x00000000 ..........  .......... (4:11,0,0.02) LEN=4    CPU

[... many lines reporting the same status deleted ...]

                                                                                                    ....

20190324_01:54:32.585139616 [ 337]: -->DN1[ Mem32 RdReq ] 'h00000001 'h0000000f 'hca01001c ..........  .......... (3:11,0,1.12) LEN=4    CPU
20190324_01:54:32.585140124 [ 333]: <--DN1[ Cpl w/ data ] 'h4a000001 'h0e000004 'h0000001c 0x00000000 ..........  .......... (4:11,0,0.02) LEN=4    CPU
20190324_01:54:32.689161116 [ 338]: -->DN1[ Mem32 RdReq ] 'h00000001 'h0000000f 'hca01001c ..........  .......... (3:11,0,1.12) LEN=4    CPU
20190324_01:54:32.689161616 [ 334]: <--DN1[ Cpl w/ data ] 'h4a000001 'h0e000004 'h0000001c 0x00000000 ..........  .......... (4:11,0,0.02) LEN=4    CPU
```

    - NVMe controller status register CSTS.RDY reports not ready

Example of device ready failure

PCIe Gen. 5 and Beyond

- PCIe Gen. 5 is almost the same as Gen. 4
  - Early adopters will need to make their own tools
- Gen. 6 specification in 2021
  - Significant changes
    - 64 GT/s
    - Pulse Amplitude Modulation
    - Forward Error Correction
- PCI SIG on three year cycle, Gen. 7 in 2024



SAS-4 and SAS-5

- There may be a SAS-4 market … *maybe*
  - SAS-4 demand is limited to companies that do not want to update/upgrade databases
  - SAS optimized databases do not make good use of potential performance gains
- Will there even be a SAS-5?

Closing

- Making your own tools is crucial as an early implementor
- Starting early allows the product to mature, but is costly *… and frustrating*
- Beware supply chain issues
  - High speed components may be in short supply
  - Resin for high speed PCBs will be costly and is in short supply

Questions?

?

Backup

TLP Capture Example (3)

- PCIe link rcvd_ts
- Test Log

```
======================Test: Display Link Status, Wed Mar
ActiveLanes      = 00001111 (0x0f)
ValidLanes       = 00001111 (0x0f)
TxElecIdle       = 11110000 (0xf0)
TxDetectRx       = 00001111 (0x0f)
RxElecIdle       = 11110000 (0xf0)
RxPolarity       = 00000000 (0x00)
LTSSM            = 10000    (0x10)  L0
Link Width       =   x4     Changed
Link Speed       =   5.0G   Changed
Link Up          = 1
ReTrain          = 0
LinkUp  Count    = 1
ReTrain Count    = 65535
Time since last ReTrain: 1.1e-05 seconds
```

Example of link training failure

- Log from TLP Capture

```
delta=     0.000 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.000 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.016 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.016 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     1.332 us,   ltssm=[0e Rcv_RCfg   ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.448 us,   ltssm=[0f Rcv_Idle   ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.504 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=    14.004 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111, 10recovery=[ rcvd_ts ]
delta=     0.016 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.016 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     1.340 us,   ltssm=[0e Rcv_RCfg   ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.448 us,   ltssm=[0f Rcv_Idle   ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.496 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=    14.012 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111, 10recovery=[ rcvd_ts ]
delta=     0.016 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.016 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     1.332 us,   ltssm=[0e Rcv_RCfg   ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.448 us,   ltssm=[0f Rcv_Idle   ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.512 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=    13.996 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111, 10recovery=[ rcvd_ts ]
delta=     0.016 us,   ltssm=[10 --L0--     ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
delta=     0.016 us,   ltssm=[0b Rcv_RLoc   ], lpmsm=[1 L0         ], llpmsm=[0 L10     ], TxEI=0, RxEI=0, Clkreq=0, SpdCh=0, Spd=Gen2, perst=F, NAKtx=0, NAKrx=0, RxAct=00001111
```

```
10recovery=[ rcvd_ts ]
10recovery=[ rcvd_ts ]
```

PCIe Gen. 4 Changes

1. **Speed change to 16GT/s**
2. **Equalization updates for 4.0 (8 GT/s to 16 GT/s)**
3. TSx OS changes
4. 16 GT/s EIEOS
5. SKP OS changes (CTRL SKP)
6. Polling.Compliance update
7. **10-Bit Tag**
   - Allowing for greater token count
8. Data Link Feature Exchange
9. Flow Control Scaling
10. **Rx Margining (a.k.a. Lane Margining)**
    - **Host cabling feature**
    - Pushed through as a requirement by Intel
11. Retimer
12. **Configuration space register updates**