PCI Express Performance Measurements

Application Note 1565

Overview

The serial point-to-point PCI Express technology supports up to 4 GB/s bandwidth per direction. Depending on the link width, the bandwidth scales from 250 MB/s to 4 GB/s per direction. But, this high theoretical bandwidth does not guarantee the overall performance will be optimal. Performance always depends on the efficiency of both devices on a PCI Express link. Parameters like payload size, flow control credit availability and different latencies strongly influence the overall result.

It's not an easy job to predict the actual performance of a new device. The numerous input factors make it very difficult to find a precise estimate of the real-live performance.

A first performance estimate

What is the maximum throughput, one can get for read completions on a x1 PCI Express link under the following conditions?

1. The requester is able to accept completion packets at maximum rate (ideal requester).

2. The completer is able to send completion packets at maximum rate.

3. The completer splits the completion at each 64 byte Read Completion Boundary.

    1. 240 MB/s
    2. 210 MB/s
    3. 190 MB/s
    4. 170 MB/s

The result is 190 MB/s. Is this surprising? Is it lower or higher than expected? Why is the maximum only 190 MB/s?

Definition of Performance Parameters

The most interesting performance parameter is the link **Throughput**, the actual amount of bytes being transferred in one second.

Also interesting is the information how the link is being **utilized**. How does the link usage time compare to total link time?

$$\text{Utilization} = \frac{\text{LinkActiveSymbols}}{\text{TotalSymbols}}$$

Finally, the link **efficiency** builds a ratio of number of payload symbols divided by the amount symbols while the link is active. In other words, efficiency is evaluated with the equation:

$$\text{Efficiency} = \frac{\text{PayLoadSymbols}}{\text{LinkActiveSymbols}}$$

where

LinkActiveSymbols =
OverheadSymbols + PayLoadSymbols

This parameter tells how many symbols would be transferred if the complete link time was used.

The actual throughput is calculated with the formula below.

Throughput =
MaximumThroughput * Utilization * Efficiency

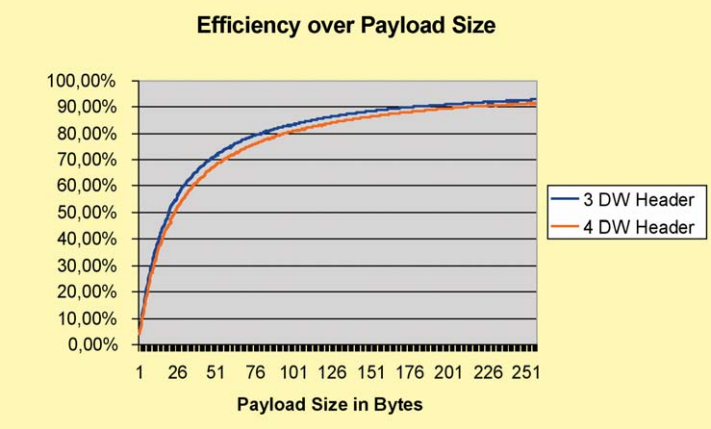

Figure 1. Efficiency over payload size

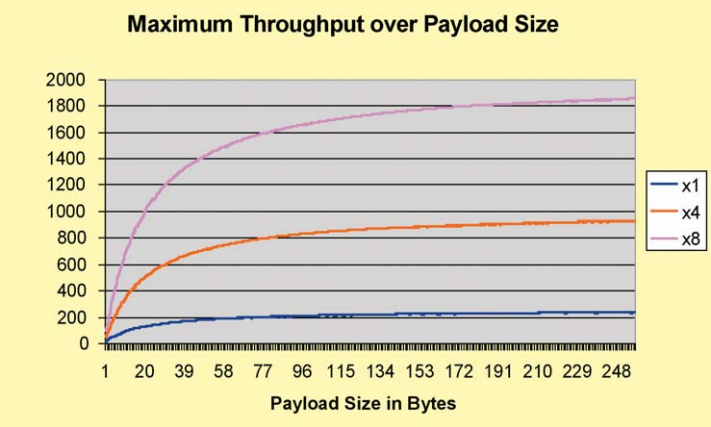

Figure 2. Maximum throughput over payload size

Latencies

The **request to completion** latency (Figure 3) determines the responsiveness of the system. One can distinguish here between first DWORD latency and last DWORD latency. The values here may differ, depending on actual load condition of the backend. Missing credits for completion headers or data may also influence these numbers.

TLP to flow control (FC) update
When the transaction layer packet (TLP) is received, sequence number and CRC checking takes place. If there's no error it will be put into the receive buffer. Then the TLP will be given to the transaction layer. When the transaction layer finally has accepted the TLP, the buffer spot will be freed again, and the transmitter will send a flow control update to the link partner.

TLP to flow control update (Figure 4) is the time between the end of a TLP and the flow control update data link layer packet (DLLP) that returns the credits that were used by the originating TLP.

Flow control update to TLP
When a flow control update is received, CRC checking takes place. Then it is forwarded to the transaction layer. If this flow control update results in additional posted, non-posted or completion credits, a TLP that was waiting for credits will be forwarded from the transaction layer to the data link layer (if replay buffer space is available). The data link layer will add the framing and finally transmit the TLP.

The flow control update to TLP latency (Figure 5) is the time it takes from receiving a FC update DLLP until a TLP that was waiting for credits is transmitted.

The **flow control update latency** (Figure 6) is the sum of TLP to FC Update plus FC Update to TLP plus the DLLP length.

Figure 3. Request to completion latency

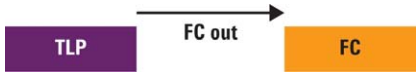

Figure 4. TLP to flow control update latency

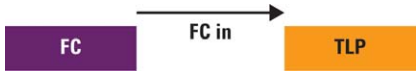

Figure 5. Flow control update to TLP latency

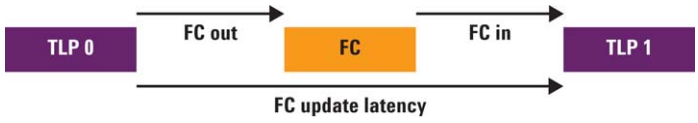

Figure 6. Flow control update latency

Latencies (continued)

Buffer sizes for each virtual channel need to be sufficient so that big flow control update latency does not cause lower utilization and throughput. It's really important to realize this latency is determined by both sides of the link. Buffer sizes need to be chosen so the performance requirements of the device are met at maximum link width and big flow control update latency.

In other words: If a device received credits for 8 headers and 1024 bytes it needs to wait after it has used up all the credits if the flow control update for the first TLP did not come in on time.

For example in a by one link, if the flow control update latency is 1 µs, a device needs sufficient header and payload credits for sending 250 bytes. If the initially advertised credits from the link partner are lower, the device can not achieve full line rate as it has to eventually wait for additional credits. Utilization will drop in this case.

The device, with a by 8 link, needs credits for sending 2 KB if the flow control update latency is 1 µs.

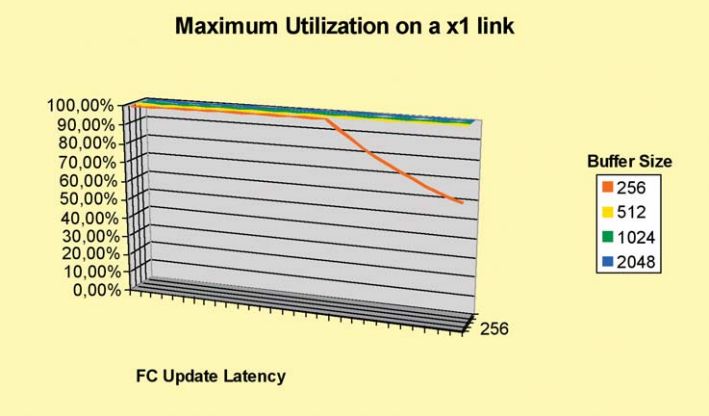

Figure 7. Maximum utilization over FC update latency (x1)

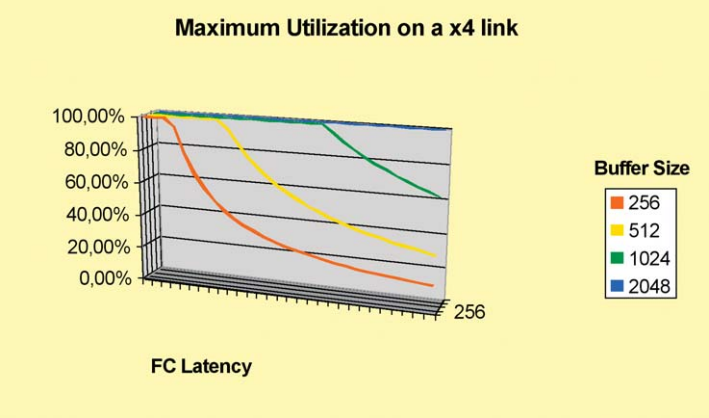

Figure 8. Maximum utilization over FC update latency (x4)

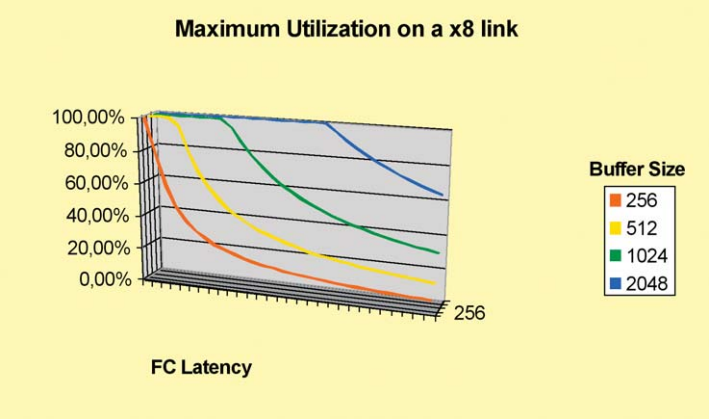

Figure 9. Maximum utilization over FC update latency (x8)

TLP to ACK/NAK latency

When a TLP is received by the data link layer it will check that TLP for framing and CRC errors. Depending on the result of that test it will schedule either an acknowledge (ACK) or a not acknowledge (NAK) data link layer packet (Figure 10).

The TLP to ACK/NAK latency is the average time between the end of the TLP and the ACK or NAK DLLP for the appropriate TLP.

Figure 10. TLP to ACK/NAK latency

Figure 11. ACK to buffer free

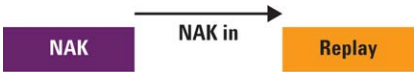

Figure 12. NAK to replay

ACK to buffer free and NAK to replay

Receiving an ACK or NAK data link layer packet uses up some time. Also, the action that is either buffer free or replay needs some time (Figure 11).

The time from reception of the ACK DLLP until the receive buffer is freed is the ACK in latency.

The time from reception of the NAK DLLP until the TLP is replayed is the NAK in latency (Figure 12).

The NAK in latency can be measured whereas the ACK in latency can not be measured since buffer free does not result in an observable event on the link.

Nevertheless, a device may need to wait for replay buffer space depending on ACK latency, link speed and replay buffer size (Figure 13).

Figure 13. ACK/NAK latency

Test methods and setup

Setup 1

The performance measurements were taken with an Agilent Technologies E2960 protocol analyzer. It was set up to measure throughput, efficiency and utilization on the link between Device 1 and Device 2.

It's only possible to measure the actual results with the protocol analyzer. This setup (Figure 14) does not allow measuring the maximum capabilities of Device 1 or Device 2.

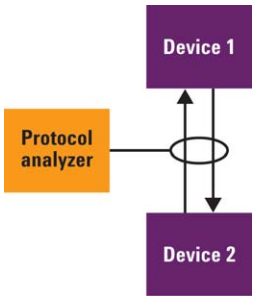

Figure 14. Setup for measuring actual performance

Setup 2

In order to measure the maximum capabilities of a device, an ideal link partner is required. An ideal stimulus is a device that does not influence the performance parameters of the device under test.

An Agilent E2960A Protocol Exerciser and Analyzer setup was used for this task. The exerciser is stimulating the system with ideal traffic. The protocol analyzer measures the actual performance numbers in this setup (Figure 15).

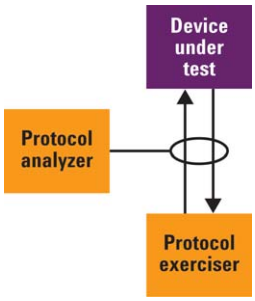

Figure 15. Setup for measuring maximum capabilities

Measurement Results

Actual device performance

The Figure 16 bitmap shows throughput, utilization and efficiency on a x1 link. The result was:

Direction throughput	Upstream 7 MB/s	Downstream 7 MB/s
Utilization	5%	10%
Efficiency	55%	30%

The upstream direction was more efficient. Therefore, utilization of the upstream direction was half as big as the downstream direction.

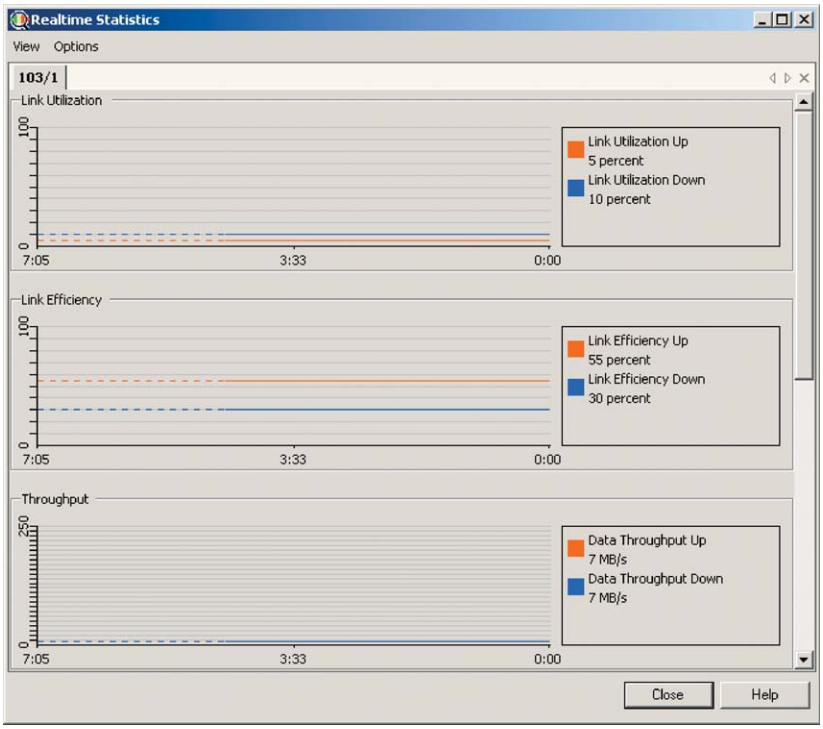

Figure 16. Actual performance

Maximum completion throughput

Now the exerciser was used in order to send infinite read requests to the device under test (Figure 17). The receiver of the exerciser was configured to show infinite credits for completions. This way it's possible to measure the maximum completion throughput the device under test is able to drive.

The result was 180 MB/s at 99% utilization and 75% efficiency. The low efficiency was due to the average payload size of 64 bytes.

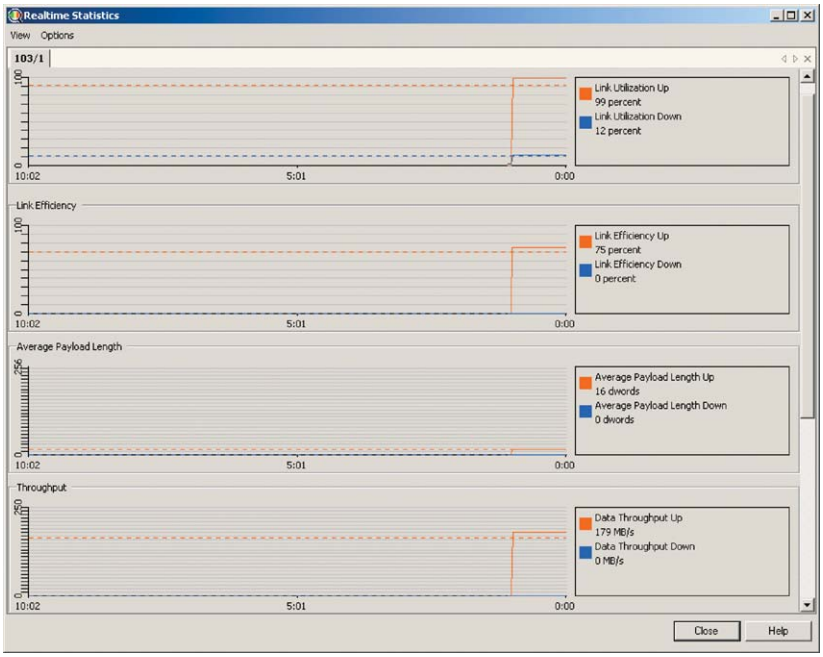

Figure 17. Maximum completion throughput

Measurement Results (continued)

TLP to FC update latency

Here (Figure 18) the exerciser was programmed to send a memory write request to the DUT. The protocol analyzer was used to measure the time between that TLP and the next flow control update. The result was 624 ns. Since the TLP duration was 240 ns, the real TLP to FC update latency for posted writes on this device was 384 ns.

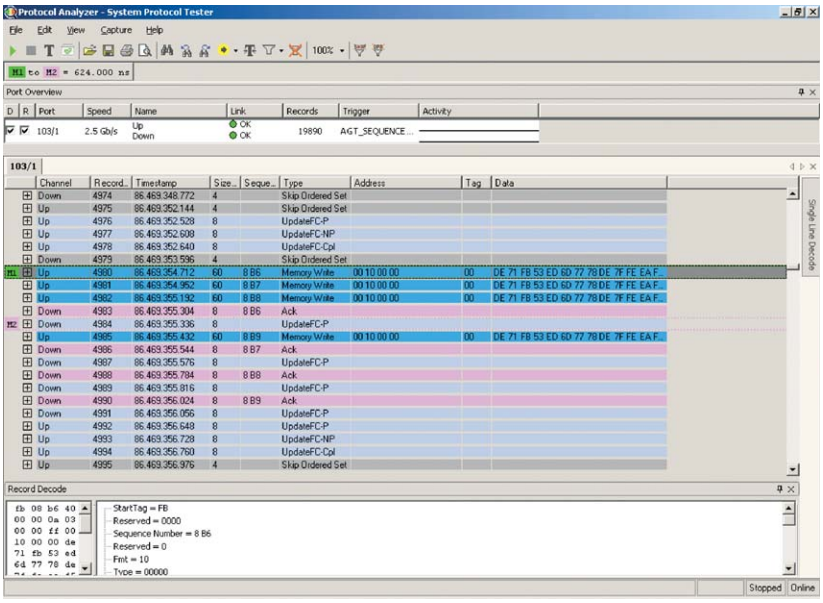

Figure 18. TLP to FC update latency

FC update to TLP latency

Finally, the exerciser was programmed to show very limited completion credits for the device under test so that it was forced to wait for flow control updates (Figure 19). This way it's possible to measure the FC update to TLP latency by measuring the time between a flow control update (completion) packet and the next completion TLP. The device under test showed a latency of 432 ns.

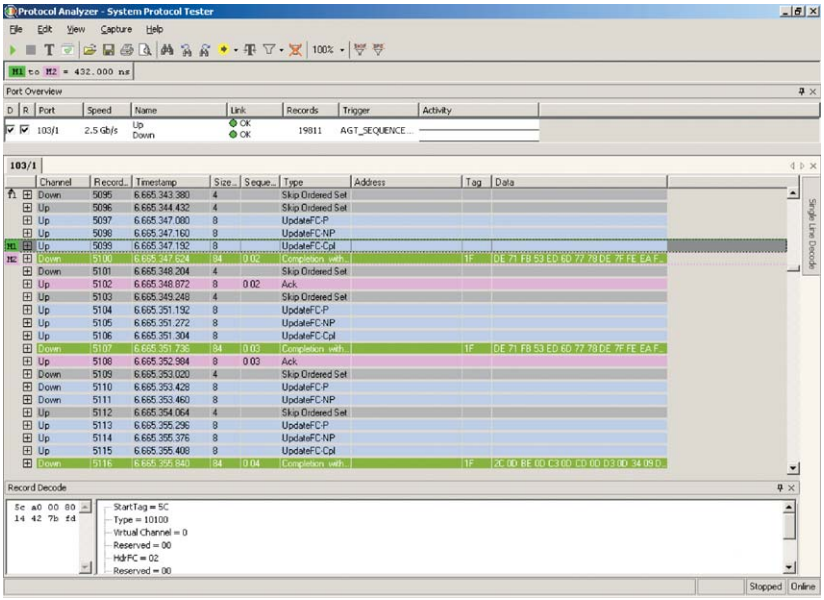

Figure 19. FC update to TLP latency

How to avoid performance surprises

As it has been shown, the performance on a PCIe link depends on the characteristics of both devices on the link. In order to make sure performance requirements are met, it's a good idea to anticipate the device at the other side of the link has high latencies. Here are some suggestions on meeting performance requirements:

- Make sure the device is sending packets with maximum payload size.
- Avoid unnecessary DLLP's.
- Minimize the flow control and ACK/NAK latencies of the device.
- Supply sufficient buffer size for each virtual channel and the reply buffer so that big flow control and ACK/NAK latencies at the other side of the link do not hurt.

Summary

- PCI Express parameters such as TLP size, availability of flow control credits and latencies have a strong influence on the overall performance.
- Both sides of a link are influencing the overall performance.
- For corner case measurements an ideal stimulus is required.

Agilent Open

www.agilent.com/find/open
Agilent Open simplifies the process of connecting and programming test systems to help engineers design, validate and manufacture electronic products. Agilent offers open connectivity for a broad range of system-ready instruments, open industry software, PC-standard I/O and global support, which are combined to more easily integrate test system development.

Agilent Email Updates

www.agilent.com/find/emailupdates
Get the latest information on the products and applications you select.

Agilent Direct

www.agilent.com/find/agilentdirect
Quickly choose and use your test equipment solutions with confidence.

For more information on Agilent Technologies' products, applications or services, please contact your local Agilent office. The complete list is available at:

www.agilent.com/find/contactus

Phone or Fax
United States:
(tel) 800 829 4444
(fax) 800 829 4433

Canada:
(tel) 877 894 4414
(fax) 800 746 4866

China:
(tel) 800 810 0189
(fax) 800 820 2816

Europe:
(tel) 31 20 547 2111

Japan:
(tel) (81) 426 56 7832
(fax) (81) 426 56 7840

Korea:
(tel) (080) 769 0800
(fax) (080) 769 0900

Latin America:
(tel) (305) 269 7500

Taiwan:
(tel) 0800 047 866
(fax) 0800 286 331

Other Asia Pacific Countries:
(tel) (65) 6375 8100
(fax) (65) 6755 0042
Email: tm_ap@agilent.com
Contacts revised: 05/27/05

Printed in USA,
September 15, 2006
5989-4076EN

Agilent Technologies