# Choosing PCI Express Packet Payload Size

January 15, 2007 Version 1.0

## 1 Introduction

In evaluating system interconnects, much is made of protocol efficiency – the ratio of payload to header and other overhead symbols on the wire. In PCI Express (PCIe), the maximum payload is a system wide constant set to the least common denominator of device support in the system. Designers choose the value of maximum payload to support based upon cost/performance tradeoffs, the needs of their application, and their expectation of market needs. The protocol engine then automatically segments longer transfers into packets equal to or smaller than the maximum supported in the path of the packet. This white paper gives guidelines to device designers based upon consideration of protocol efficiency and market requirements.

## 1.1 Market Segmentation of Maximum Payload Support

We observe distinct market segmentation in the support for various maximum payload values.

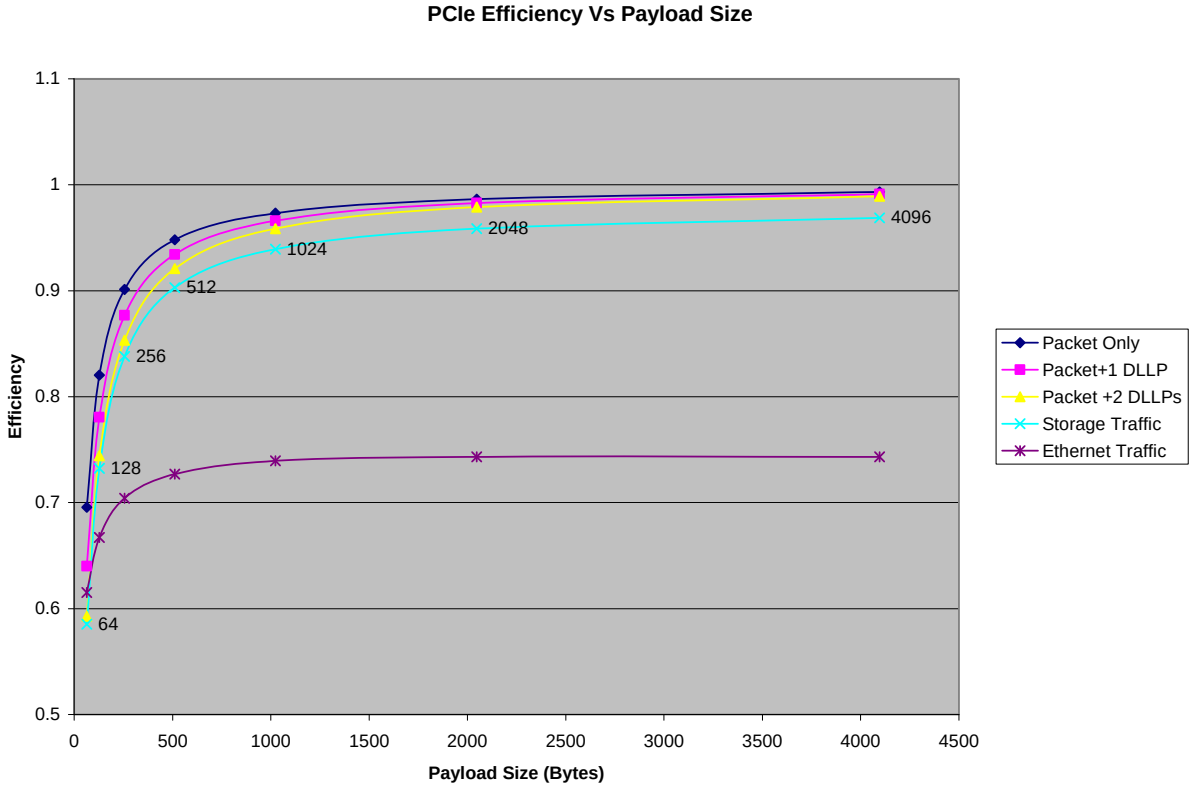
Intel desktop chipsets support at most a 64-byte maximum payload while Intel server chipsets support at most a 128-byte maximum payload. The primary reason for this is to match the cache line size for snooping on the front side bus. A secondary reason may be that the memory controller itself is optimized around handling cache line sizes. Finally, the buffer memory required is roughly proportional to the maximum payload size; supporting longer packets raises device cost. The majority of the market is well served with a maximum payload of 256 bytes or less.

Chipsets produced by vendors other than Intel have supported a higher value; 512 bytes is the commonly known maximum payload value for a server North Bridge. As will be shown later, this value provides higher throughput for storage and network traffic and in fact seems to mark the point of diminishing returns, except for a specialized storage infrastructure.

The storage infrastructure is optimized to transfer long files segmented into disk sector size payloads of 2K- or 4 K-bytes. Storage ASSPs typically support payloads of up to 4 K-bytes and in most cases exhibit less than optimal performance with shorter packet lengths due to internal architectural tradeoffs and lack of large packet-size support in other components in the path of transfer. This is the case for storage “boxes” as opposed to storage HBAs that are limited by the North Bridge’s maximum payload capability. Most storage OEMs have adopted to short payload restrictions but a small percentage of systems using proprietary North Bridge function through custom ASICs have been consistent in requiring support for longer packet lengths.

## 1.2 PCIe Protocol Efficiency vs Payload Size

The figure below shows PCIe protocol efficiency as a function of payload length, assuming a 16-byte header and the use of ECRC (4 bytes). The top line shows the efficiency of a single packet considered alone. Even the relatively small maximum payload length supported by typical server chipsets (256 bytes on average) shows greater than 80 percent efficiency when a single packet is considered in isolation. However, the data link layer protocol requires an ACK packet and a flow-control packet for every two (roughly speaking) transaction layer packets. The middle two lines of figure below show the efficiency taking the ACK and flow-control overhead into account. The figure also clearly illustrates the improvements (or lack thereof) in efficiency with payload sizes over 512 bytes.




A typical PCIe application also includes overhead for DMA descriptor read and occasional transfer complete interrupts. Storage traffic is composed primarily of full-sector transfers and thus require a descriptor read and an interrupt only every 4 K-bytes. Ethernet traffic, on the other hand is based, on a two pronged mix of short and long traffic, with an average payload size estimated at 600 bytes.

Furthermore, the Ethernet traffic mix includes many short (64 bytes or less) packets whose presence pulls the Ethernet efficiency well below that of the storage mix and does so almost independently of the maximum payload length. To show this effect, the Ethernet traffic was modeled as a mix of 64-byte and 1.5 K-byte packets in proportions to achieve a 600-byte average packet length. We assumed that for Ethernet, a buffer architecture and synchronization strategy are used so that a descriptor read plus an interrupt aren't required for short packets; the severity of the short packet problem mandates such a solution.

Storage applications show higher overall efficiency than does Ethernet because their DMA and interrupt overhead are amortized over longer data blocks and storage uses a very high proportion of long packets. Of course, when Ethernet is used for storage, as in iSCSI, the traffic mix will take on the characteristics of storage traffic and support for longer payloads for Ethernet will be beneficial.

The reader should also be aware that even when claiming support for an extended maximum payload, a North Bridge **may** still supply read completion packets with only 64 or 128 bytes of payload. Thus, the higher efficiency might be enjoyed only when writing memory.

## 1.3 Buffer Size Impact

PCIe flow control is modeled upon an input buffer. In order to achieve full wire speed, a device must state sufficient credits to mask the delay of the flow-control credit update latency loop. Since, in the worst case, an FcUpdate or ACK DLLP may have to wait for a maximal length packet to pass by, the size of the input buffer required to achieve full wire speed becomes strongly dependent upon the maximum payload size. If a designer implements support for a longer packet length without increasing the buffer memory size appropriately, then throughput will suffer when the longer packet length is used. The throughput loss through credit starvation may well exceed that gained due to the higher efficiency of the long packets, especially when many ports share the common buffer pool inside the device.

In a switch, a fixed amount of buffer memory is available. PLX Technology switches are configurable as to the number of ports. When the PEX 8548, for example, is configured to a lesser number of ports, memory is reallocated from the idle ports and can be used to support a higher maximum payload value without the cost penalty of a larger memory buffer. PLX switches are specified with sufficient buffer memory for most common port configurations and maximum payload values supported without compromise to throughput.

## 1.4 Summary

PCIe devices support different maximum payload sizes. Those cost-optimized for computer I/O – the majority of devices – regrettably limit their support to the 64-, 128- or 512-byte limit enforced by chipsets. Performance oriented devices for storage and networking should support a maximum payload of at least 512 bytes but will benefit from this only when used with North Bridges that have the same or greater capability. ASSPs for the storage market can benefit from a maximum payload capability of up to 4 K-bytes, but the improvement over 512 bytes is small. PLX makes a range of switches in which the maximum payload capability adapts inversely to the number of ports configured. Thus we are able to provide the extended maximum payload capability required by storage applications without penalizing cost-sensitive applications that require more ports and to provide uncompromised throughput in all configurations.