PCI Express 3 was announced in August, 2007. Like its predecessor, it sought to double performance but unlike its predecessor it did not do so by simply doubling the bit rate. Why?

Conventional limitations

Like processors in the past, sooner or later interconnects like PCI Express are going to hit a barrier imposed by physics. The physical dimensions of a transmission line are limited by dielectric loss (amongst other things), which is a function of frequency. PCI Express can only afford to double in frequency perhaps once more before expensive materials like Teflon are required in the manufacture of printed circuit boards. Similarly, the other way of adding bandwidth – more lanes – is limited by package pin count.

Power consumption in electronic circuits is also function of frequency. As well as green considerations, heat dissipation imposes limits on performance, and power consumption with attendant air conditioning demands has become a major headache for operators of data centres.

So how can photonics help?

Photonics

Advantages

By comparison, optical solutions can multiplex well over 100 wavelengths (lanes) over a single fibre, each of which can comfortably be modulated at 10s of gigahertz or faster. As well as being economic with physical transmission paths, this also means a single switching element can switch every lane in a particular fibre at once instead of one switch per lane as in electronic switches.
Perhaps the most radical promise is that photonics offers the prospect of decoupling performance from power consumption. It turns out that the reason for this is both a strength and a weakness of photonic technology. Electronic systems consume power when a transistor switches state because the signal is regenerated from the power supply, where in photonics, a light source is modulated without storing and regenerating the signal and power demands are determined by the physical characteristics of the technology.

No storage means less processing delay and obviates the need for buffer management, with the potential result of lower latency, higher efficiency throughput.

**Disadvantages**

No storage also means no buffering, which means traditional store-and-forward techniques are impractical in photonic switches. The way photonic systems are built must, therefore, be carefully rethought.

Although optical buffers are not impossible, Rod Tucker of Melbourne University demonstrated in 2005 their infeasibility. He analysed optical network routers in terms of power consumption and physical packaging, and concluded that for the foreseeable future, electronic switches beat optical switches on both counts. It seems likely that OEO (Optical-Electrical-Optical) conversion is there to stay for those applications.

An architecture which eschews optical buffering has the opportunity to benefit from avoiding the latency which buffers introduce. Local interconnect networks are small enough for it to be feasible to centralise scheduling for the entire network, making it possible to use edge buffering (before the first EO conversion) to avoid contention and therefore the need for optical buffering in the switch altogether.

The lack of storage also imposes the requirement that all agents be synchronous. That means that the host interface of a lowly USB hub would have to operate at the same speed as a high performance GPU. This might be mitigated by the use of multifunction devices, subordinate, low-speed buses, or perhaps slower modulation at the expense of switch efficiency.

**Time slots and Optical Burst Switching**

An agent cannot transmit a packet unless the switch is configured to deliver it to the intended destination. A scheduling mechanism is therefore required. Most switching fabrics would usually provide ingress buffering, and sometimes buffering for packets in transit. Where optical buffering is unavailable, the buffers which already exist in each PCI Express agent can be used to provide edge buffering. With suitable scheduling – and because global state can be known to the scheduler – no other buffering should be required.

The easiest way to schedule time in an optical switch fabric is to divide time into slots (separated by guard bands to allow the switching elements to settle). Practicality dictates that these time slots be fixed size in length.
Superficially, this looks a little bit like Optical Burst Switching, a technique in which packets are stored at the edge of a network and transmitted in batches when the channel becomes available. OBS is useful in fabrics which have slow switching times (such as MEMS, which has a switching time in the order of milliseconds), but can make latency much worse – something which we think should be avoided. The difference is that in OBS, switch efficiency depends on the slot period being much greater than the packet period.

Research

Latency

PCI Express buses are barely recognisable as being related to the standard which was introduced in 1993. It is at best a distant cousin to PCI/-X. Both support a range of bus widths and clock frequencies.

One significant innovation introduced by PCI-X was the option to split request and completion into different bus transactions. Split transactions were helpful because they eased timing compliance for slower devices, at the cost of potentially unbounded completion latency. My experience has been that after about 50 ms, a host will lock up – during which time, the processor which initiated the transaction is halted.

Unlike PCI-X, PCI Express is compelled to split all transactions that require any sort of response because it is not a bus in the conventional sense. It is much closer to a packet oriented message passing system; indeed, PCI Express 1 and 2 both use the same coding and physical transceivers as gigabit Ethernet.

The current draft of PCI Express revision 3 abandons the 8b10b coding used by both previous revisions (and gigabit Ethernet) in favour of 128b130b. While considerably reducing coding overhead, this change increases marshalling latency by a factor of 16 for the benefit of only twice the throughput.

Does latency matter?

"Money can buy bandwidth, but reducing latency involves bribing God."

– attributed variously

The instinctive answer tends to be yes, and David Patterson wrote in 2004 on how bandwidth has grown so much faster than latency. At the very least, PCI Express exemplifies this and arguably was built in such a way that latency suffered. To begin my research, I proposed a set of experiments to investigate the relevance of latency in PCI bus performance, and to evaluate the effects of time slots on latency.
The first experiment involves directly measuring latency in PCI, PCI-X and PCI Express buses. The second involves using a benchmark (such as SpecSFS) to investigate the effect on performance of varying latency using a programmable ‘delay line’ which I shall build for the purpose. A modified version of the delay line can be used to characterise the impact of slot allocations on bus performance.

Together, I hope for these to make the case that any optical local interconnect must keep low latency at the forefront of its design, especially given the enormously greater bandwidth which optical systems offer.

**Practical implementations**

The technology does exist to build an optical switch from discrete switching components, and I hope to build a model of the switch to demonstrate the feasibility of the optical PCI interconnect.

In the first instance, the model will probably be built from electronic components, but under the same restrictions under which optical switches operate.

Arrays of switching elements are quite feasible, and there have been exciting developments in recent years in building waveguides in silicon so the degree of integration required to make such a technology viable commercially may be just around the corner.

**References**


PCI-SIG®, “PCI Express® Base Specification Revision 3.0 (draft 0.3)”, 22 May, 2008