Optical Interconnects in Future HPC Systems

Steve Scott
Cray Inc.
sscott@cray.com

Abstract: This talk will describe the primary interconnection network challenges as we attempt to build exascale computers over the coming decade. We discuss the role that optics will play in these systems, and key attributes for signaling technology from a system builder's perspective.
OCIS codes: (060.4250) Networks; (060.4258) Network topology

1. Introduction

A continuation of historical HPC system scaling trends would have us achieving a sustained Exaflop on the High Performance Linpack (HPL) benchmark by around 2019 [12]. That target presents significant design challenges in many areas, including the interconnection network. This paper provides a system-builder's perspective on these network-related challenges, with a specific emphasis on the impact of optics on future network designs. An expanded version of this paper was presented at the 2009 Hot Interconnects conference [9].

Optical signaling technology will play a key role in future HPC network design. By breaking the tight relationship between cable length, cost, and signaling speed, optical signaling opens the door to network topologies with much longer links than are feasible with electrical signaling. Cost-effective optics will thus enable a new class of interconnects that use high-radix network topologies to significantly improve performance while reducing cost.

2. Design goals and challenges

Large-scale HPC systems contain up to tens of thousands of compute nodes in hundreds of compute cabinets, with spans of up to 40-50m from corner to corner. The interconnects need to support both message-passing and global address space workloads. Thus, performance on both bulk data transfer and short packets is important. Traffic can be highly irregular and time-varying, so packet-level adaptive routing is important and fast routing decisions are required (virtual circuit setup is not practical).

Network performance is evaluated primarily on the sustained bandwidth and latency experienced by real workloads. Systems currently under design require on the order of 10 GB/s per node of network injection bandwidth, with hardware network latencies in the hundreds of ns for large systems. To maximize system price-performance, both performance and price-performance of the network matter. Thus, network design basically involves hitting some absolute performance goals while minimizing cost, and while constrained by the set of available signaling and packaging technologies. Secondary measures such as reliability, diagnosability, configurability, serviceability and scalability also play a role.

Network bandwidth can be measured for both point-to-point and global/irregular traffic. Both are important for HPC workloads. At Cray, we tend to favor global bandwidth as a metric (peak bandwidth for all-to-all or uniform random communication). Even for jobs with "nearest neighbor" communication amongst logical nodes, the virtual-to-physical node mapping can cause contention on network links, and job scheduling tends to lead to physical fragmentation of node sets over time in order to maximize machine utilization. Many jobs also perform long-distance or irregular communication amongst logical nodes. Thus we cannot rely on physical locality within networks, and global bandwidth matters.

Unfortunately (for network designers), computational capability continues to grow at a substantially faster pace than network bandwidth. The Cray T3E, for example, used 375 Mb/s network signaling in 1996, while QDR Infiniband used 10 Gbps transmission 14 years later in 2010, an average increase of 26% per year. Meanwhile, processors grew from around 1GF to 100GF over that same time period, an average increase of around 39% per year. The coming advent of high-flop-count accelerators will only exacerbate this trend.

The design of exascale systems presents several key challenges, including power, resiliency, application concurrency, and programming complexity. The greatest of these is power. A feasible first Exaflop system near the end of the decade might contain around 100,000 processor sockets of 10 TF each. To achieve 0.01 B/flop of global bandwidth (about half that of a current 200 GF x86 node with a x4 QDR Infiniband fat-tree network) would require 100 GB/s of bandwidth per node, or 10 PB/s of global bandwidth.

A widely-assumed upper end for the power of our largest HPC systems is around 20MW. Assuming we were willing to spend up to 5MW just on network transmission (likely over-generous), that would limit us to (5MW/80Pbit/s) ≅ 60 pJ/bit to move data across the global network. Given that current signaling technology can

easily use 10-20 pJ/bit across a single link, and the average distance in a 100K-node 3D torus is over 30 hops, we can see that advances in both network topology (to reduce diameter) and signaling technology will be needed to achieve viable networks for exascale computers.

Cost and physical packaging of networks is also going to become an increasing problem as the growth of signaling rates continues to lag that of computational rates. We are also starting to face significant limitations on reach of electrical signals as signaling rates exceed 10 Gbps. This will be discussed further in the next section.

3. HIGH-RADIX NETWORKS AND THE IMPACT ON LINK LENGTH

Network designs often involve trade-offs between link widths, link lengths and network diameter. By narrowing the link width, a router can have more ports (that is, a higher radix), and reduce the diameter of the network (number of hops a packet must traverse to cross the network, either average or worst case). For example, in *k*-ary *n*-cube networks, the average network diameter for a network with $N = k^n$ nodes is proportional to $n(\sqrt[n]{N})$, which shrinks as the dimensionality, n, is increased. Moreover, as the network diameter shrinks, the total number of wires necessary to achieve a given global bandwidth shrinks proportionately, which can reduce network cost. Conversely, given a fixed number of wires per node, a higher-radix network can achieve higher bisection bandwidth.

Several attractive network topologies can be created with high-radix routers, including the folded Clos [3] (a.k.a. fat-tree [8]), **flattened butterfly** (or *k*-ary *n*-fly) [4], and **dragonfly** [6]. These have very low diameter compared to 3D tori and global bandwidth that scales linearly with system size. The flattened butterfly and dragonfly *require* high-radix routers because of the large number of ports needed per router. Interestingly, though they can provide scalable global bandwidth, they are direct networks (no external router stages required as the system grows).

Despite the apparent advantages of high-radix routers, many HPC commercial networks have been built with low-radix routers [11][1][7][13]. Two significant reasons for this are serialization latency over narrow links [5], and the negative impact of longer physical cable lengths on cost and signaling rate.

While network packet sizes have remained roughly constant over time (e.g.: a packet might carry one 64B payload), the bandwidth of network routers has increased by over two orders of magnitude in the past two decades. For this reason, packet serialization latency for narrow links has become relatively insignificant. Meanwhile, new router architectures have enabled high-radix routers to be designed [5][10].

Though exact details vary, higher radix networks generally require longer cable lengths. As will be discussed in Section 4, this significantly reduces the achievable electrical signaling rate. Electrical cables can also be quite bulky, making cable mats for high-radix networks physically challenging, and potentially limiting network bandwidth due to physical space for routing cables. The cost of electrical cables is also highly correlated with length, with wire costs that scale linearly with cable length and a relatively modest connector cost. Packaging overheads related to cable construction and connector back-shells can also make extremely narrow cables inefficient.

Optical signaling rates (and to a lesser degree, costs) are insensitive to cable lengths, and optical cable bulk is quite low. Thus, optical signaling has the potential to largely eliminate the negative impacts of physical cable length, which, along with the reduction in link serialization latency, could make high-radix networks very attractive.

There has of course been a long-standing interest in using optics in multiprocessor interconnects (the *Massively Parallel Processing Using Optical Interconnections* conference was started in 1994), but it has not been cost effective over most of this period. Optics was first used for long-haul networks, then local area networks, then cluster interconnects. It is now planned for use within next-generation multiprocessor interconnects from both IBM [2] and Cray. Many different optical network topologies have been proposed, and many metrics of value have been forth, of varied importance, in our opinion, to the design of practical HPC systems. The next section discusses which metrics we believe are important in evaluating optical and electrical signaling technology.

4. Optical Signaling Technology Metrics

The most important metric for optical signaling technology is simply the *cost per unit of bandwidth* ($/Gbps). This must be measured over some given physical path or distance: between chips on board, between boards across a backplane, between adjacent cabinets, over a 5m cable, 10m, etc.

$/Gbps almost always grows with distance and is highly related to packaging hierarchy. PCB routing is considerably less expensive than cables, and each additional connector also adds cost. As some point, as distance is increased, the signaling rate can no longer be sustained, and either more expensive materials (PCB, connectors and/or cables) or repeaters must be used, or the signaling rate must be dropped.

Though electrical signaling is currently much less expensive than optical signaling over PCBs and short cables (due to the high cost of the optical transceivers), its costs rise more steeply per meter than with optical signaling, and transmission line losses at high frequencies require the periodic use of repeaters for longer electrical links. We estimate the current price-performance cross-over point to be somewhere between 5 and 10 meters.

The interplay between topology, link length, signaling rate and cost lead to interesting trade-offs in the *cost per unit of global bandwidth* ($/GBW). Accurate calculations of $/GBW are exceedingly complicated, involving many degrees of freedom with respect to target performance levels, topology, scale, system density, packaging options, PCB/connector/cable materials choice, configurability, etc. Our own analysis has indicated that $/GBW can be minimized for future large systems using optical link technology and high-radix networks such as the flattened butterfly and dragonfly. A large reduction in $/Gbps in optical signaling would significantly strengthen this conclusion.

The second most important metric for optical signaling technology is the *power per unit of bandwidth* (mW/Gbps or pJ/bit). System power is very important today and getting more so every year, and the network can be a large fraction of total system power. The total power for an optical link must include the electrical power used to communicate to/from the transceiver. Careful co-design of the electrical and optical circuits should be used to minimize this total link power. *Power per global bandwidth* (W/GBW) can be used to compare the power efficiencies of two topologies or to compare an electrical interconnect to an optical interconnect.

Other metrics are less important, but still of significant interest to system builders. *Bandwidth density* can play a large role in a system design. This can be the total bandwidth out of a package, or the achievable bandwidth off the edge of a board (Gbps/inch). A transition to optics integrated onto the chip package, perhaps in conjunction with wave division multiplexing, could lead to a significant increase in achievable bandwidth off the chip, which could be transformational to system design, but only if the cost and power were affordable.

There are a number of mechanical/packaging metrics that are moderately important, including weight (*Kg/GBW*), signaling density (*Gbps/ m^2*), and bend radius (*cm*). Component reliability is also important, and adding optical transceivers to a link will generally increase failure (FIT) rates.

There are several metrics that are *not* important for system builders. Bandwidth per fiber is not important in and of itself unless it reduces overall cost/Gbps. Bit error rate certainly can't be ignored, but adding expense to improve bit error rates is generally not productive. Even with BERs in excess of 1e-9 (much higher than typical optical links), a CRC-protected channel with hardware retransmission can provide extremely high reliability with less than 1% bandwidth overhead from re-transmissions.

The ability to broadcast an optical signal to multiple listeners is not needed; the occasional tree-based broadcast can be performed in hardware or software over conventional networks with point-to-point links. Likewise, the ability to perform optical routing of incoming optical data (an all-optical-network) is not needed. Our view is that optics is attractive as a transmission medium, not for performing logic.

5. Summary

After several decades in which electrical signaling has out-performed optics on key metrics of value for system interconnects, optical signaling is now looking attractive for the longer links in next-generation systems. The key metrics that the optical industry should focus on improving are *$/Gbps* and *pJ/bit*.

The next major disruption point will be when optical signaling can be used directly off the processor and router packages. This has the potential to substantially increase the available bandwidth in the system, so long as cost and power can be kept in check.

Continued improvements in optical signaling relative to electrical signaling will strengthen the arguments for high-radix, low-diameter networks, and give the system designer greater flexibility to build networks that optimize overall system price-performance.

6. References

[1] N. Adiga, *et a*l., An Overview of the BlueGene/L Supercomputer, *SC'02*, November 2002.

[2] B. Arimilli, et al., "The PERCS High-Performance Interconnect," *Hot Interconnects 18*, August 2010.

[3] C. Clos, A Study of Non-Blocking Switching Networks, *The Bell System Technical Journal*, 32(2): 406-424, March 1953.

[4] J. Kim, W. J. Dally, and D. Abts, Flattened Butterfly: A Cost-Efficient Topology for High-Radix Networks, *ISCA '07*, June 2007.

[5] J. Kim, W.J. Dally, B. Towles, and A.K. Gupta, Microarchitecture of a high-radix router, *ISCA '05*, pp 420-431, June 2005.

[6] J. Kim, W.J. Dally, S. Scott, D. Abts, Technology-Driven, Highly-Scalable Dragonfly Topology, *ISCA '06*, June 2006.

[7] J. Laudon and D. Lenoski, The SGI Origin: A ccNUMA Highly Scalable Server, *ISCA '97*, pp 241-251, June 1997.

[8] C. Leiserson, Fat-trees: Universal networks for hardware efficient supercomputing, *IEEE Transactions on Computers*, October 1985.

[9] M. Parker and S. Scott, "The Impact of Optics on HPC System Interconnects," *Hot Interconnects 17*, August 2009.

[10] S. Scott, D. Abts, J. Kim, and W.J. Dally, The BlackWidow High-Radix Clos Network, *ISCA '06*, June 2006.

[11] S. Scott and G. Thorson, The Cray T3E Network: Adaptive Routing in a High Performance 3D Torus, *Hot Interconnects 4*, August, 1996.

[12] Top500 Supercomputer Sites, http://www.top500.org.

[13] J. S. Vetter, et al., Early Evaluation of the Cray XT3, *Proc. 20th IPDPS*, April 2006.