The Impact of Optics on HPC System Interconnects

Mike Parker and Steve Scott
map@cray.com, sscott@cray.com

Cray Inc.

Index Terms—interconnection network, high-radix router, network topology, optical interconnect

1. INTRODUCTION

Optical signaling has long been used for telecommunications, where its low-loss signaling capability is needed and the relatively high termination costs can be amortized over long distances. Until recently, Cray has not found it advantageous to use optics in its multiprocessor interconnects. With recent reductions in optical costs and increases in signaling rates, however, the situation has changed, and Cray is currently developing a hybrid electrical/optical interconnect for our "Cascade" system, which will be shipping in 2012.

In this position paper, Cray was asked to answer the question "*Will cost-effective optics fundamentally change the landscape of networking?*" The short answer is yes. By breaking the tight relationship between cable length, cost, and signaling speed, optical signaling technology 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.

Section 2 of this paper discusses the design targets for HPC system interconnects, setting the context in which optical technology will be considered. Section 3 discusses the impact of network topology and reviews the case for high-radix networks, which create the need for longer physical links in the network. Sections 4 through 6 then present a variety of important – and not-so-important – metrics by which optical signaling technology should be judged. Finally, Section 7 presents a summary and conclusions.

2. NETWORK DESIGN GOALS

At Cray, we design systems with hundreds to tens of thousands of compute nodes, connected via a custom interconnect. The systems range in physical size from a single cabinet up to a few hundred cabinets, perhaps as large as 40-50 meters from corner to corner.

Future 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. Bandwidth demands will of course continue to increase over time with increases in compute node processing power.

Both performance and price-performance 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.

The traffic pattern must be considered when evaluating network performance. Some network topologies are preferable for nearest-neighbor communication, while others are preferable for global or irregular communication. At Cray, we tend to favor global bandwidth[1] as a metric. While many applications do perform logical nearest-neighbor communication, most applications don't take on the complexity of understanding their logical-to-physical node mapping and the machine's physical topology, and optimizing communication to reduce link contention. Also, in practice, to maximize system utilization job schedulers tend to grab whatever compute nodes are available when launching jobs, such that the sets of physical processors on which jobs run become physically fragmented over time. Therefore even jobs that are performing logical nearest-neighbor communication tend to have communication better characterized as semi-random permutation patterns. Adaptive routing can be used to smooth out temporal non-uniformities and hot spots, and make link utilization more uniform. By attempting to schedule jobs on sets of physically proximate processors, average communication distance can be reduced, but we still cannot count on physical nearest-neighbor communication. And of course many applications also perform long-distance or irregular communication amongst logical nodes. So, global bandwidth matters. Point-to-point bandwidth between two nodes is also important because it can limit performance of all data transfers even at light global loads.

Clearly network topology has a significant impact on sustained global bandwidth. The next section discusses the interplay between topology, link length and signaling speed.

3. HIGH-RADIX NETWORK TOPOLOGIES

Given a physical design constraint, such as the number of pins available for a router node, the choice of network topology often involves a trade-off between link width 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

[1] We use *global bandwidth* rather than *bisection bandwidth*, because it is agnostic to network topology. Global bandwidth is the peak bandwidth for all-to-all or uniform random communication, whereas bisection bandwidth is the peak bandwidth crossing a minimum bisection cut of the machine. For some topologies (such as a mesh), injected all-to-all traffic crosses the bisection 0.5 times on average, whereas for others (such as a fat-tree), injected all-to-all traffic crosses the bisection once on average. Thus, bisection doesn't give a consistent measure of the bandwidth available for all-to-all or global communication.

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. In a binary hypercube, average network diameter is just $\frac{1}{2}\log_2(N)$. Moreover, as the average network diameter shrinks, the total number of wires necessary to achieve a given global bandwidth shrinks proportionately, which can reduce network cost.

Figure 1 illustrates the advantage of increasing the radix of a network. A group of 16 nodes each have *L* lanes of network pins with bandwidth *B* per lane per direction. If connected as a 1D torus with link width of *L*/2 lanes, the bisection bandwidth of the network is 2*BL* and the average diameter is 4 hops. If connected as a 2D torus (in this case, a hypercube) with link width *L*/4 lanes, the bisection bandwidth is doubled and the average network diameter is halved.

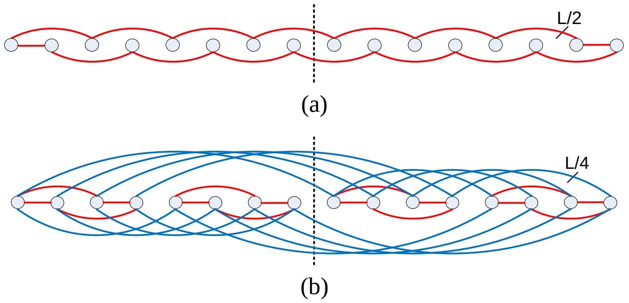

Figure 1. Advantages of Higher Radix

Consider a set of 16 nodes, each with L lanes (one differential pair in and out) of pin bandwidth signaling at B bits/s/lane/dir.

(a) 1D torus with link width L/2 lanes. Bisection bandwidth = 2BL. Average distance = 4 hops. (Max = 8 hops.)

(b) 2D torus (hypercube) with link width L/4 lanes. Bisection bandwidth = 4BL. Average distance = 2 hops. (Max = 4 hops.)

There are several attractive network topologies that can be created with high radix routers, including:

- The folded Clos network[3] (a.k.a. fat-tree[8]) provides global bandwidth that scales linearly with node count, yet can also be easily *tapered* to reduce cost. In its full instantiation, it can route any permutation conflict free, and with proper support, can balance load across all links even for deterministically-routed traffic[9]. It has low network diameter compared to a torus or a hypercube, and has many redundant paths, allowing for resiliency in the face of faults. While a folded Clos can be built with low-radix routers, a high-radix folded Clos has lower network diameter and requires fewer network stages.
- The flattened butterfly (or *k*-ary *n*-fly) network[4] is similar to an *n*-stage *k*-ary butterfly with the *n* stages collapsed into a single router. It can also be thought of as a *k*-ary *n*-cube torus, where the *k* nodes along each dimension are connected by an all-to-all instead of a ring. The flattened butterfly can *only* be created with high-radix routers, since each node requires *(k-1)n* links. It also requires the use of adaptive routing, because non-uniform traffic patterns can cause up to a factor of *k* greater contention than uniform traffic when routed deterministically. Under uniform loading, the flattened butterfly causes only half the wire utilization of a folded Clos network.
- The dragonfly network[6] is a variation on the flattened butterfly that uses extra (inexpensive) local hops to reduce the number of (expensive) global hops. Local *groups* are internally connected as a flattened butterfly, and treated as a very-high-radix router to create a *single*, global all-to-all stage amongst groups. As with the flattened butterfly, the dragonfly network requires the use of adaptive routing and high radix routers. Both the flattened butterfly and dragonfly networks provide the scalable global bandwidth and very low network diameter of a high-radix folded Clos without requiring external router stages (that is, they are *direct* networks).

Given the advantages of higher radix networks, it might seem surprising that many commercial networks have been designed with low-radix routers[1][7][10]. There are two primary reasons why high-radix networks have not been used historically: high link serialization latency and the impact of physical cable length on cost and signaling rates.

The link serialization problem was described in detail by Kim, *et al*[5]. The latency to send a packet across a network in a pipelined fashion can be broken into two components: the time to route the head through the network (which is proportional to network diameter times per-hop delay), and the time for the tail to catch up to the head (packet serialization latency across a link, which is inversely proportional to link bandwidth). With low signaling rates, the link serialization latency can be quite high, resulting in a significant latency penalty for narrow links.

However, over the past couple decades, router pin bandwidth has increased by close to a factor of 10 every five years[5]. Meanwhile, the size of individual network packets has remained roughly constant, carrying perhaps a single cacheline of data, or even just a command or acknowledgement. Link serialization has now ceased to become a significant factor in network latency. An 80-byte packet takes only 64ns to be serialized over a single bit lane at 10 Gbps. This change in underlying technology makes high-radix routers more attractive, and increases the optimal dimensionality of interconnection networks.

Physical signaling and packaging considerations have also historically conspired against very high radix networks. Though the exact details vary with physical layout and topology, 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. This can be at least partially overcome by cable aggregation, however, in which multiple narrow links connected to different routers are carried in the same physical cable.

Low-cost optical cables have the potential to largely eliminate the penalties for longer cable lengths. In conjunction with the increases in router pin bandwidth that have eliminated the link serialization penalty, this can usher in a new generation of networks built on high-radix topologies.

Optical signaling technology has of course been around for a long time, and has been proposed for use in multiprocessor interconnects for over 15 years. (The *Massively Parallel Processing Using Optical Interconnections* conference series was started in 1994.) Proposals over the years have included free-space optics, holographic optics, optical switching, wave-division multiplexing, and various architectures that take advantage of the ability to have multiple transmitters and receivers on a single optical channel. Many metrics of value have been put forth, of varied importance, in our opinion, to the design of practical HPC systems. The next few sections discuss which metrics we believe are important in evaluating optical and electrical signaling technology.

4. COST-PERFORMANCE

The most important metric in evaluating a signaling technology is the *cost per unit of bandwidth* ($/Gbps). Of course, to be relevant, this cost-performance 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. In addition to price-performance, HPC systems may have some absolute performance requirement they must hit, particularly with respect to the package pin bottlenecks. In order to use all available bandwidth coming off a router chip, for example, reducing the signaling rate may not be an option.

At present, electrical interconnects offer superior price-performance on board, across a backplane within a chassis, and even over short (few meter) cables including several inches of PCB foil at the ends. Above this distance, however, electrical signaling is starting to hit transmission limits as signaling rates move beyond 10 Gbps.

Electrical interconnects suffer more from transmission line losses (often termed dB loss) at high frequencies than optical interconnects. As a result, for a given data-rate, repeaters are required periodically in long electrical cables to re-drive the signals. Alternately, lower data-rates can be used per electrical signal pair to reduce the impact of the transmission line loss over a desired distance.

For a fixed data rate, the cost of an electrical interconnect increases linearly as the length of the cable increases. At points where the length of the cable exceeds the length that can be reliably driven at that data rate, a stair-step occurs in the cost as electrical repeaters are inserted in the line.

For a cable with no repeaters, transmission losses increase as the square of the distance. Thus, once the maximum allowable losses occur, signaling rate must be reduced as the square of the distance, resulting in a quadratic increase in $/Gbps. While optical signals are also attenuated over the length of an optical fiber, the distances that can be driven without repeaters is more than sufficient for even large scale supercomputers.

Thus, it is useful to compare the cost of transmitting a unit of bandwidth a particular distance. Figure 2 compares the cost cost/Gbps of an electrical cable with repeaters and an optical cable as cable length is increased. The costs are normalized to the cost of a zero-length electrical cable. For reference, a zero-length electrical cable (*i.e.:* connectors only) costs the equivalent of about 8-9 meters of copper cable.

The cost of the optical cable is dominated by the endpoints, which may cost 1.5-2x the cost per Gbps of the electrical cable endpoint. The cost of the optical cable grows slowly with distance compared to the electrical cable, with the electrical cable per meter per Gbps cost growing at 2-2.5x the cost of the optical cable. The electrical cable (here assuming 16 Gbps signaling) requires a repeater every 6 meters. The first two repeaters are required at shorter distances to compensate for losses from PCB foil and connectors at the endpoints. Here the repeaters cost about the equivalent of 2 meters of electrical cable.

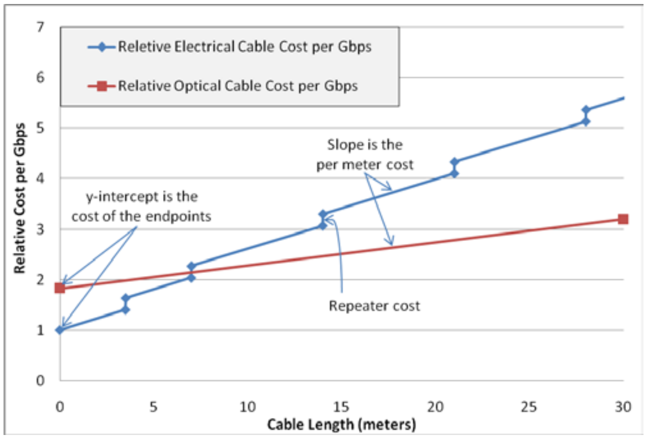

Figure 2. Electrical and Optical Cable Cost/Gbps vs. Length

Your mileage may vary with respect to the precise numbers, but the qualitative trends are clear. We currently estimate the price-performance cross-over point to be somewhere near seven meters. A significant reduction in optical link prices would of course lower the optical graph, and thus reduce the length of the electrical/optical cost-performance cross-over.

The superior cost-performance of optical links at 10m+ lengths, coupled with their relative insensitivity to link length (and lower cable bulk), opens the doors to a wealth of topology choices, as discussed in Section 3. 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 (really!), 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. Further reductions in the cost of optical links (relative to copper cables) would strengthen that conclusion, as would further increases in signaling rates that reduced the distance over which electrical communication could be performed without repeaters.

5. PACKAGING AND OTHER CONSIDERATIONS

HPC interconnect design is greatly impacted by mechanical packaging, which gives rise to several other important metrics for a candidate signaling technology. As the number of signal pins on IC packages continues to rise, and more and more computational power is placed on a board, there is ever increasing pressure on signaling bandwidth at all levels of the packaging hierarchy.

At the system level, we are primarily interested in cable management. The bandwidth demands of large HPC systems translate to a large number of electrical or optical cables that need to be routed both within and out of the cabinet. Connections to neighboring cabinets can often be routed horizontally between the cabinets. Cables connecting to more distant cabinets generally connect out the top or bottom of the cabinet. Cray typically routes them over the top of the machine rather than routing them under the floor tiles.

Jaguar, the Cray XT5 system at Oak Ridge National Labs, has over 3000 miles of interconnect network wires. Management of such a large bulk of wiring can be a significant challenge. This is important for getting cables to fit within the available space and for making sure the system is maintainable and expandable. Here metrics such as cable volume, cross-sectional area, and bend radius are all important. The volume occupied by a set of cables is useful for reasoning about the infrastructure required to support global cabling. The cross-sectional area of a set of cables is useful for understanding the minimum area required to escape a cabinet or route a set of cables down a row of cabinets.

For comparison purposes, these cable bulk metrics can be normalized to a particular bandwidth (*e.g.*: $m^2/Gbps$). Optical cables tend to be significantly less bulky than their electrical counterparts for a given bandwidth. In electrical cabling, this disadvantage is made worse when considering cabling with periodic active repeaters. Repeaters also raise a number of issues related to packaging, serviceability, and supplying power to the repeaters.

The bend radius of cables has an impact on the amount of wasted volume required within the back of a cabinet needed to escape the cables away from the system backpanels and to escape the cables out of the cabinet. Small signal-pair count electrical cables generally have a fairly small bend radius, whereas electrical cables with more signal pairs (often the size of garden hoses) are not nearly as flexible. Optical fiber is much smaller than the corresponding copper wire pairs and can be physically bent in a fairly tight radius. However, there are restrictions on how tight one should bend an optical fiber to minimize optical attenuation. For the medium signal-pair counts found in HPC interconnection networks, optical interconnects have an advantage in this space.

Though of lesser concern, the weight of HPC systems, including the interconnect, can be a factor in some HPC applications. Again, an advantage of optical links is that they weigh much less per unit of bandwidth then their copper counterparts.

At the board level, signals need to be routed off a router die, through a chip package, and to the edge of a circuit board where they can connect to other parts of the system. Pin bandwidth, or *Gbps/pin*, is useful for understanding the number of pins required to get bandwidth off the router chip. Higher pin counts can significantly increase both the packaged part cost and board cost, due to the additional layers required in the PCB to escape the signals. Thus, when high bandwidth is desired, it is generally advantageous to signal at the highest rates feasible in the silicon technology being used.

At the edge of the board, signals must route through an optical or electrical connector. Here the key metrics for signaling technology are *Gbps/inch* or *signals/inch* at some given data rate. A lower signal density can limit bandwidth off the board, or result in larger boards (or lower computational density!) to achieve the desired bandwidth. Though both this metric and pin bandwidth are indirectly reflected in cost per bandwidth, it is more convenient (especially earlier in the design process) to deal with these bandwidth density metrics directly

In general, higher data rates (within reasonable limits) lead to lower costs per Gbps at most lengths for system interconnects. They also result in better cost at the board and chip level. As previously mentioned, bringing optics on the board and directly to the chip isn't cost effective in current systems. Thus getting the signals over to optical transceiver chips or over to the board edge is done electrically.

Neither optical nor electrical interconnects appear to have a clear advantage in terms of bandwidth density off the board. While some optical couplers may consume less area on a backpanel per unit of bandwidth than an electrical equivalent, there is usually additional PCB area consumed for the optical transceivers. As optics straight to the router package becomes cost effective, there will likely be a significant advantage here for optics.

Optical interconnects come in two broad categories: active optical cables (AOC), where cables connect electrically on both ends and have an optical transceiver built in to the two cable ends, and fiber with transceivers directly on the router board. Because there is not a significant advantage to bringing signals off the backpanel optically versus electrically, it makes sense to consider other advantages to active optical cables. One minor advantage of AOCs is that the laser is never exposed. While there is some safety advantage to this approach, it is not necessarily a compelling reason to use AOCs in HPC interconnects, where qualified and trained individuals perform maintenance of the machine. One more significant advantage of AOCs, however, is that optical and electrical cables can potentially be interchanged as the system size or configuration is varied. Shorter connections can use less expensive copper cables, and longer connections can use optical cables, where they are more cost effective.

As the radix of the network increases, there is an increase in the number of independent connections to a particular printed circuit board within a cabinet. Each of those independent connectors has an overhead (a connector back-shell) associated with it. Thus, there is a cost or overhead (in signals per inch and Gbps/inch) associated with high-radix topologies. While there is some cost to be paid for this, it is generally cheaper than the alternative of routing significantly more signals and/or lower global bandwidth that accompanies lower radix topologies. This can be seen by calculating the global bandwidth the topology is capable of attaining per inch of local board space.

Interconnect power is also an important factor in HPC systems. 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. *Joules/bit* at a given distance or similar metrics can also be used to compare the energy efficiency of two individual cables. When comparing optical to electrical power, it is important to count the power all the way back to the endpoints (the routers).

The addition of optical transceivers and electrical repeaters has a negative impact on system reliability (MTBF). Each of these parts has a failure rate (FIT rate) associated with it. For shorter connections, passive electrical connections have a lower FIT rate, as they have no active components that go bad over time. Optical cables only have transceivers at either end of the optical fiber. Thus they end up having a reliability advantage over longer length electrical cables where several repeaters are required.

6. NOT SO IMPORTANT CONSIDERATIONS

Several metrics often discussed for optical signaling have little relevance to HPC interconnects. Chief among these is bandwidth per fiber. For the majority of cables in a system, the cost of the fiber is quite small compared to the cost of the transceivers, and optical cable bulk is generally not a problem. Thus, using wave division multiplexing (WDM) to push up the total bandwidth per fiber is not that helpful unless it lowers the cost of the transceivers. Similarly, there is no benefit to being able to transmit more than around 50 meters for HPC interconnects.

There has also been recent work on snoop detection and other security mechanisms associated with optical cables. While these technologies may be important for long haul fiber optics, where the fiber may not be fully physically secure, HPC systems tend to be isolated to a machine room where access can be controlled.

While bit error rate (BER) can’t be ignored in HPC interconnects, 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 could be useful for certain traffic patterns, but is not needed in general; 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. Electrical switching capabilities are keeping up with data rates, and do not add significant switching latency compared to time of flight in large networks. Electrical switching furthermore allows flow control, adaptive routing, configurability and other performance and administrative features. Our view is that optics are attractive as a transmission medium, not for performing logic.

7. SUMMARY AND CONCLUSIONS

After several decades in which electrical networks have out-performed optical networks on key metrics of value, optical signaling technology is poised to exceed the bandwidth/$ of electrical signaling technology for long network links (greater than a few meters). Along with high router pin bandwidth, this starts making high-radix network topologies look attractive. As the price-performance cross-over length continues to shrink, it will make high-radix networks quite compelling. There will be no reason not to exploit the low latency and scalable global bandwidth of topologies such as the high-radix folded Clos, flattened butterfly and dragonfly.

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, both to local memory[2] and between nodes of the system.

Optical wave-division multiplexing (WDM) has limited benefit in current systems, because the bandwidth off the chips is constrained by the electrical signaling rate and the number of package pins. All of those electrical signals must be routed to the optical transceivers, and it matters little whether they are sent on individual fibers or merged onto a smaller number of fibers using WDM (as mentioned above, it would matter only insofar as it affected the price of the optical link). Once the conversion to optics happens on-package where the number of available electrical signals is much greater, however, then WDM may allow for greater total bandwidth on-to and off-of the package. Signaling power may also be reduced by reducing the distance that the electrical signals have to drive.

We don’t believe that the transition to on-package optical signaling will result in another change to network topologies. It will simply provide a large bump in achievable network bandwidth. We also don’t see a coming need for optical switching. Electrical switching performance should continue to improve along with computational performance over the coming generations of silicon technology. Optical signaling will simply provide a superior mechanism for moving bits between chips, as evaluated by the metrics discussed earlier in this paper.

8. REFERENCES

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

[2] C. Batten, *et a*l., Building Manycore Processor-to-DRAM Networks with Monolithic Silicon Photonics, *Hot Interconnects 16*, August 2008.

[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* pages 126–137, 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*, C-34(10):892-901, October 1985.

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

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