Networking the National Leadership Computing Facility

Steven Carter, *Oak Ridge National Laboratory*

ABSTRACT: *Today's supercomputers produce and consume as much data as several thousand normal Internet users and place a special burden on the network. Special attention must be paid to provide an appropriate local- and wide-area infrastructure to properly support and efficiently share these resources. The most important metric to determine success is application–to-application performance. For this reason, it is important to understand the characteristics of each architecture and how they affect network performance. The National Center for Computational Sciences has made many improvements to the Cray X1's Cray Network Server enabling it to achieve high-throughput transfers in the local- and wide-area.*

KEYWORDS: *X1, XD1, XT3, Network Performance*

1 Introduction

The National Leadership Computing Facility (NLCF) is seeking supercomputers of unprecedented size and power. These computers produce and consume as much data as several thousand normal users and place a special burden on the network beyond that which standard enterprise networks can bear. Since a supercomputer runs scientific applications, network performance cannot be measured by the number of bits that flow from one network port to another in any given second. The measure of performance must be the degree to which the network enables an application to produce good scientific results. The network should be seen as extending into the supercomputer to the very system call the application makes to transfer data.

For this reason, NLCF is making a substantial commitment to improving application to application performance to enable its supercomputers to produce the best science possible. This commitment includes significant local- and wide-area network improvements as well as a concerted effort to optimize applications to take best advantage of each supercomputer's I/O characteristics.

The National Center for Computational Sciences (NCCS) has recently started a major upgrade of its local area network, allowing it to support 10Gb/s interfaces throughout the network and accommodate multiple 10Gb/s transfers over its backbone. In addition, Oak Ridge National Laboratory is deploying an optical network with a total wide area capacity close to 1Tb/s. This year, 6 new 10Gb/s connections will be added to increase the NLCF's ability to share its computing resources with distant researchers.

In concert with the upgrade in network hardware, the NCCS is making a significant commitment in optimizing the operating systems and software to best utilize the available local and wide area bandwidth. Last year, NCCS integrated Net100 technology into the Cray Network Subsystem (CNS) allowing for faster wide area transfers. As part of this work, the TCP assist functionality was re-written and later adopted by Cray for use in the latest versions of the CNS. Currently, researchers are able to get 700-800Mb/s throughput locally and 400-500 Mb/s throughput wide area for a single file transfer through a single CNS on a regular basis. Work is underway within NCCS to increase single CNS throughput to 5Gb/s.

2 NCCS Network Infrastructure

2.1 *Local-Area Connectivity*

The NCCS network core consists of three Cisco 6500 class Ethernet switches with Supervisor 720 blades (Figure 1). One 6509 with dual Supervisor 720s acts as a 10Gb/s aggregation switch, allowing for a very high bisection bandwidth among the 10Gb/s connected hosts. The remaining 6509 and 6513 are interconnected with multiple 10Gb/s interfaces and house Firewall Services Modules (FWSM) and Content Switching Modules (CSM). Although capable of housing 10Gb/s interfaces, these switches serve predominantly as 1Gb/s aggregation points.

At speeds above 1Gb/s, switch backplane bandwidths become the least limiting factor. There are currently no firewalls that support 10Gb/s streams. Multiple vendors make firewalls that support aggregates of 10Gb/s or more, but single streams are still limited to 1Gb/s. In order to accommodate 10Gb/s streams, a hybrid approach was taken to intrusion prevention. Cisco Access Control Lists (ACLs) are used to prevent unwanted access to hosts that require streams greater than 1Gb/s. Remaining hosts are protected with the Cisco FWSMs. Fewer features are available with ACLs than are available with a true firewall, but they provide a good mix of performance and security.

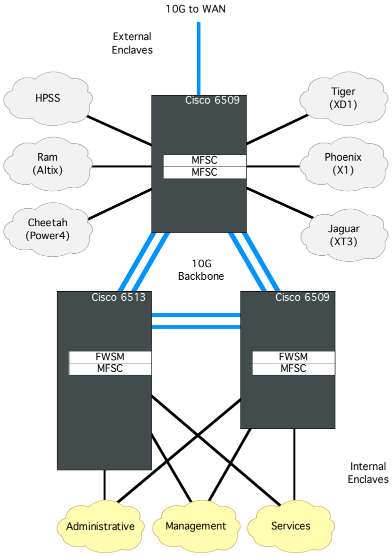


Figure 1: NCCS Internal LAN

To further enhance security, hosts with similar security requirements are collected together into Enclaves. Each enclave is protected by a mini-firewall from each other enclave. In the case of major externally reachable resources such as supercomputers and HPSS, each is given its own Enclave. All Enclaves that contain resources with non-administrative access are collectively known as External Enclaves and considered inherently hostile. Other resources such NFS, LDAP, and DNS servers that have administrative only access are placed in Internal Enclaves. There is a one-way relationship between Internal and External enclaves. Services can be provided from Internal to External Enclaves, but a host in an External Enclave cannot log into a host in an Internal Enclave. If a resource in an External Enclave is compromised, it provides little extra access to other Internal and External resources.

2.2 *Wide-Area Connectivity*

In the summer of 2005, ORNL will acquire three new production 10Gb/s connections: one to ESNet, one to Internet2, and one to the ETF. Access to these circuits will be provided by a Juniper T640 router. Although the NCCS will have only a single 10Gb/s connection to this router, it has sufficient capacity to add more interfaces as needed.

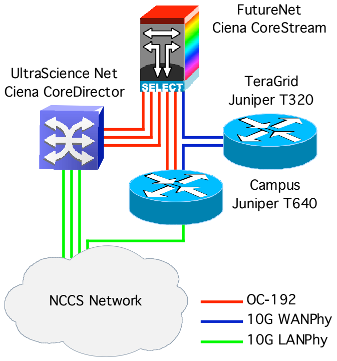


Figure 2: NCCS WAN Connectivity

In addition to the three production links, three 10Gb/s research circuits will also come online in the same timeframe: two to the UltraScience Net, and one to the Cheetah Network. The UltraScience Net is a network research project lead by ORNL to develop circuit-switched (as opposed to packet-switched) techniques in support of next generation data-transfer requirements. UltraScience Net's backbone consists of 2 x OC-192s providing dedicated channels up to 20 Gb/s between its four main switching hubs: Sunnyvale, California, Seattle, Washington, Chicago, Illinois, and Atlanta, Georgia. The sites connected to these hubs include Stanford Linear Accelerator Center (SLAC), Pacific Northwest National Laboratory (PNNL), Fermi National Accelerator Laboratory (FNAL), and the Oak Ridge National Laboratory (ORNL) and, of course, the National Leadership Computing Facility (NLCF). It is expected that Argonne National Laboratory (ANL), and the National Energy Research Supercomputer Center (NERSC) will also connect to the Chicago and Sunnyvale hubs, respectively, in the near future. The combination of these circuits makes the NLCF accessible by any researcher located anyplace on any of the main national research networks, namely ESnet, Internet2, the TeraGrid, and National Lambda Rail.

The Cheetah Network is a similar network funded by the National Science Foundation providing a single OC-192 backbone to NCSU and CUNY. The UltraScience and Cheetah Networks peer at ORNL giving the NLCF the unique ability to conduct coast-to-coast experiments (Figure 3).

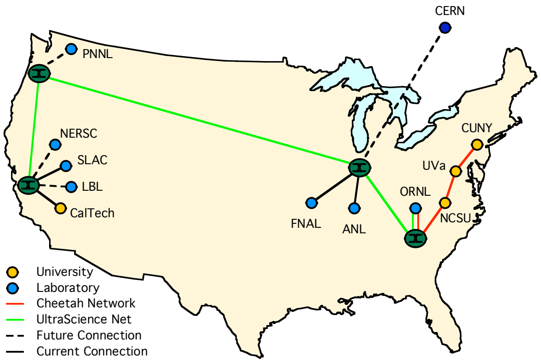


Figure 3: DOE UltraScience + NSF Cheetah

3 Architectures

When the network is viewed as extending from socket call to socket call, it is necessary to have an accurate view of the actions within the operating system and the data flow through the hardware. Each supercomputer embodies slightly different concepts that make it behave in different ways that achieve varying levels performance. Perhaps no line of supercomputers illustrates these differences more vividly than Cray's three most recent supercomputers: the X1, the XD1, and the XT3.

3.1 *X1*

The Cray X1 is a vector processor machine with shared memory and a high-speed memory interconnect presented as a single system image. The X1 contains node 4 processor node boards that come in two types: application node and OS node. The application nodes run the computations while the OS nodes run the operating system. Nodes communicate to neighbors directly in pairs, and to distant nodes via routers.

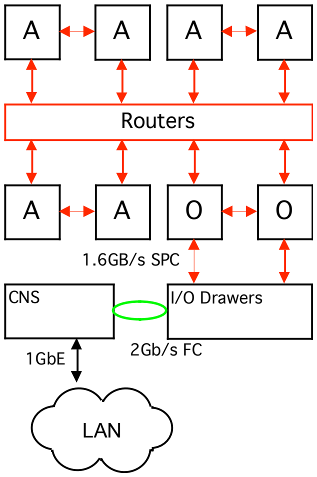


Figure 4: X1 Overview

The OS nodes contain SPC connections to PCI-X bridges in the I/O drawers. Fiber Channel HBAs are plugged into the PCI-X slots to interface the X1 to disks and networking. IP connectivity is provided via IP-over-Fiber Channel (IP/FC) to a Cray Network Server (CNS).

The CNS allows the X1 to communicate from its native fiber channel interfaces to either Ethernet or HIPPI. In the case of UDP datagrams, the CNS merely routes and fragments the 64K IP/FC datagrams into the size required by the local network. For TCP, the CNS intercepts connections and redirects them to the TCP Assist Daemon (tcp_assistd). Tcp_assistd terminates the incoming TCP stream and opens another TCP stream to the original destination. The daemon then inserts a rule into IP Tables to masquerade the new connection to make it appear to come from the first originator. This daemon transfers data between the two sockets in order to agglomerate the smaller frames arriving on the Ethernet interface into larger frames to be sent on the Fiber Channel interface or vise-versa.

When a system call is made on an application node, the system call is checked to see if it is a migrating system call. If it is, the system call is rescheduled on one of the OS Nodes. Since the X1 is a single system image, the thread reschedule is a matter of stopping the thread and placing it on an OS processor's CPU work queue. When the system call is completed, a check is made to see if the system call thread was migrated. If it was, then the thread is rescheduled back to last user PE.

Since Unicos/mp is a single system image, the thread can be scheduled on any OS node regardless of whether it has a connection to the FC HBA used by the IP/FC interface. However, there will be a performance difference depending on the location of the FC HBA due to the latency in the X1 mesh. For that reason most I/O connections are made close to the OS nodes.

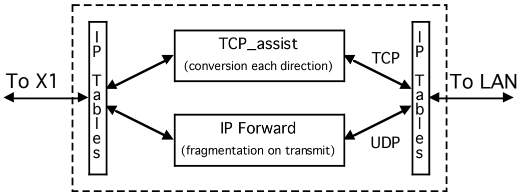


Figure 5: CNS Internal Diagram

There are many obstacles in getting good network performance from the X1:

- The vector nature of the X1 means sub-optimal performance for scalar-dominated network operations.

- The lack of network off-load functionality inherent in using IP-over-FC exacerbates the scalar problem.
- The thread migration required when opening sockets from application nodes degrades performance significantly.
- Socket calls on OS nodes without a directly attached interface can have up to a 20% penalty in performance.
- It is hard to predict when a socket call will be placed on an OS node without a directly attached interface.

3.2 *XD1*

An XD1 chassis contains compute nodes with two Opteron processors connected via HypterTransport. Each node contains one or two RapidArray processors for communications over the RapidArray interconnect (Figure 6).

The PCI-X expansion card connects to nodes 5 and 6 via each node's HyperTransport. As seen from the rear, the left two PCI-X slots are managed by node 6 (slots 1 and 2). The right two PCI-X slots are managed by node 5 (slots 3 and 4). Nodes 5 and 6 run the Ethernet drivers for the GigE cards and automatically create the interfaces for the cards they manage. The drivers initialize the card, do housekeeping, and load and unload Ethernet frames from the GigE cards.

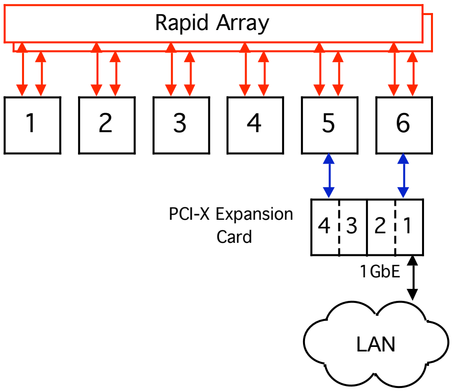


Figure 6: XD1 Overview

Nodes 1-4 do not have direct hardware access to the physical interfaces in the PCI-X slots. A layer 2 forwarding arrangement is set up to provide a node with access to a physical interface managed by nodes 5 or 6. Through Active Manager, nodes 5 or 6 can be configured with a MAC address and an IP address for a node without a physical interface. In addition, Active Manager adds the virtual interface to the node for which the MAC and IP addresses were added. Data sent to the virtual interface gets encapsulated in TCP/IP and Ethernet headers in a conventional manner; however, the Ethernet frames are handed to the RapidArray driver instead of a NIC. The RapidArray driver then forwards the Ethernet frames across the RapidArray to the appropriate node, which then sends it out a physical interface

The physical interfaces on nodes 5 and 6 run in promiscuous mode to watch for Ethernet frames destined for its own local physical interfaces or for remote virtual interfaces for which it is configured. When it sees a packet destined to one of its addresses, it is passes that packet to the RapidArray driver to be forwarded to the appropriate node. No Network Address Translation (NAT) is performed and all forwarded is done at layer 2 in the drivers to minimize overhead.

Of the three Architectures, the XD1 perhaps poses the least trouble with network I/O. All of its nodes run the same operating system. The only real problem is the lack of physical Ethernet interfaces on each node. Although the overhead is kept to a minimum, there is still a performance difference caused by the switch latency and a slight overhead on the nodes with the physical interfaces. In order to get optimal performance, care must be taken to perform network transfers on a node with a physical Ethernet interface.

3.3 *XT3*

The Cray XT3 consists of compute nodes and service nodes connected in a mesh (Figure 7). Each compute nodes runs a scaled down microkernel (Catamount) that does not support socket calls. The application nodes run a version of SUSE Linux modified to support its particular flavor of functionality: I/O, network, and login. I/O nodes have directly connected disks, network nodes have 10Gb/s Ethernet interfaces, and login nodes have 1Gb/s Ethernet interfaces.

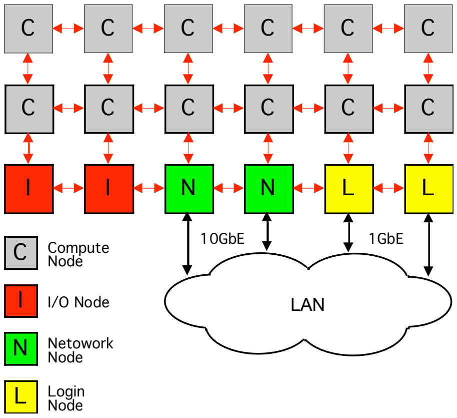


Figure 7: XT3 Overview

Network and login nodes with physical network interfaces have a conventional TCP/IP and Ethernet driver stack. I/O, Network, and Login nodes can be interconnected with IP-over-Portals interfaces. IP-over-Portals allows data to be encapsulated into TCP/IP headers and send via portals over the mesh. A network or login node can then route those packets out its physical interface. In this manner a user on a login node can send packets out the 10Gb/s Ethernet interface on a Network Node instead of through the 1Gb/s Ethernet on the Login Node.

Although not as complex as the X1, the XT3 does present many challenges to applications sending data or visualization streams over the network. First and foremost, the XT3 does not have the capability to perform network I/O from an application node. Users are supposed to send all network I/O from the login node. The XT3 can add a large number of network nodes, giving it significant network I/O capability. However, it is unclear how interfaces spread across so many nodes can be used in an effective manner by users on the Login Node.

4 Current work

Work began to characterize and improve the X1's performance soon after its arrival in the spring of 2003. As part of the characterization process, UDP data was sent from the X1 to CNS with varying packet sizes and the resulting throughput measured (Figure 8). The slope of the line between 1K and 31K packet sizes show that the X1 can send packets at roughly 3800 packets per second regardless of the packet size. While the performance is good for large packets, it is quite poor for small packets.

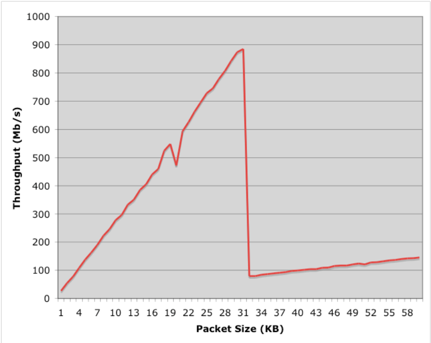


Figure 8: X1 UDP Throughput vs. Packet Size

Also evident in the data was a UDP "knee" just past 31k making it difficult to fully utilize the 2Gb/s FC connection between the X1 and the CNS. After this was pointed out to Cray, they were able to correct the behavior giving the X1 a significant performance increase.

The wide-area performance of the X1 was particularly bad as well. Early tests showed a top throughput of 15Mb/s to the Pittsburg Supercomputing Center, far below the available bandwidth of the link. In order to find the reason for the poor performance, Net100 [Net100] technology was integrated into the CNS to monitor and tune the TCP parameters while making wide area transfers. Immediately obvious was the fact that the TCP buffers used by the stock tcp_assistd were inadequate to match the Bandwidth Delay Product (BDP) of the wide-area path. The BDP is simply the bandwidth of the link multiplied by the delay between the time a packet is sent and that same packet is received. If the TCP window is smaller than the BDP, there will be unused bandwidth on the link. Since the source was not available for tcp_assistd, it was re-written from scratch. Although not the only way to adjust the buffers, it offered visibility into each component of the CNS for future work. Merely increasing the TCP read and write buffered increased wide area throughput by up to 400%.

4.1 *SuperCNS Version 1*

In order to get beyond speeds of 1Gb/s, work began on building a "SuperCNS". The first attempt involved the standard CNS hardware with 2 dual channel Emulex Fiber Channel Host Bus Adapters (HBAs) to connect the CNS to the X1 (Figure 9).

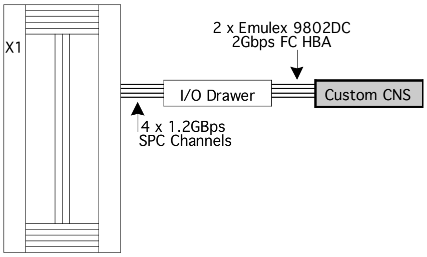


Figure 9: First CNS Test Setup

CNS version 1.2, which includes a 2.4 kernel, was installed on the CNS hardware. In cooperation with Cray engineers, the bonding functionality of Unicos/mp was used in round-robin mode to load-balance packets across each of the interfaces. Both the proprietary Emulex drivers that came with the CNS and the open source Emulex drivers were tested.

As seen in Figure 10 and Figure 11, there is a significant difference between the open source and the proprietary Emulex drivers in both performance and system resource utilization. The proprietary driver's throughput scaled much better with the number of bonded channels than did the open source drivers. One reason for this difference might be then level of resource utilization attained by each of the drivers. The proprietary drivers were able to achieve much higher system utilization and therefore higher throughput.

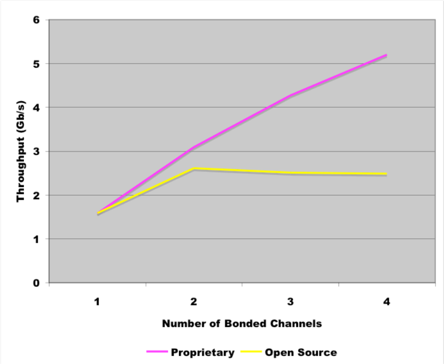


Figure 10: Throughput vs. Number of Bonded Channels for the Emulex Open Source and Proprietary Drivers

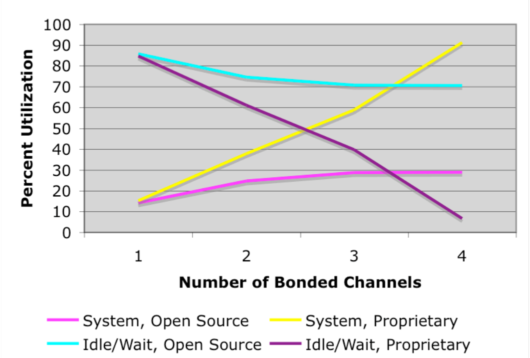


Figure 11: System Utilization vs. Number of Bonded Channels for the Open Source and Proprietary Emulex Drivers.

4.2 *SuperCNS version 1.2*

For the next attempt at a "SuperCNS", the stock CNS hardware was replaced with a dual Opteron host containing two pairs of 133Mhz, 64-bit PCI-X slots on two separate PCI-X bridges. The same two Emulex 9802DCs were used, but a Chelsio T110 10Gb/s Ethernet NIC was added. Suse 9.2 with a stock 2.6.6 kernel was installed with the version 2.1.0a-BETA of the Chelsio drivers and version 2.10 of the Emulex open source drivers. Version 2.10 of the Emulex drivers include various bug fixes and modifications to work with the 2.6 kernel. A similar host with a Chelsio T110 was used as a data sink (Figure 12).

The extra level of performance from the dual Opteron, coupled with a 10Gb/s NIC with TCP offload engine were meant to allow the new SuperCNS to accommodate high-speed streams in each direction. Although version 2.10 of the Emulex drivers offered little performance improvements, the added host capacity did allow the host to pass streams in each direction. In fact, 2Gb/s file transfers were performed on 2GB files from the X1 to the sink host.

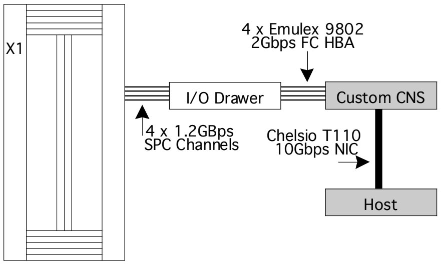
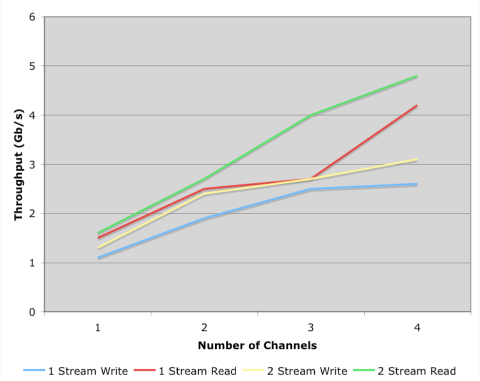


Figure 12: SuperCNS version 1.2 Test Setup

In order to attain the maximum throughput available from the SuperCNS, channel bonding was disabled, and parallel streams sent through the individual channels. In this configuration, memory-to-memory transfers reached 3Gb/s for writes and 4.8Gb/s for reads.



In addition to the non-bonded testing, a file transfer was done over bonded channels using bbcp yielding 1.8Gb/s for a 2 GB file. This rate was likely limited by the speed of the disks on the data sync.

5 Future Work

In the near future, the SuperCNS will be put into production with the Terascale Supernova Initiative (TSI) over the Cheetah Network. As part of this work, tcp_assistd will be modified to work with UDP to test the use of protocols designed for use over dedicated channels.

In an attempt to get more efficient resource utilization, version 8 of the Emulex drivers will be tested. This might involve porting the existing IP-over-FC functionality to the version 8 drivers.

Finally, similar work will be done to characterize and improve the network performance of the Cray XD1 and Cray XT3.

6 Summary

With relatively little effort, the NCCS has been able to significantly increase the network performance of the Cray X1 in a way that directly benefits the researchers that use it. It has also been shown that the latest SuperCNS hardware has more bandwidth available to be further taped for data transfers. Future work of this type will be necessary to efficiently share the NLCF's resources to make them effective as a national resource.

7 References

[Net100] *Net100*. http://www.net100.org/.

[X1Overview] *Cray X1 System Overview Version 2.4*, March 2004.

[XD1Overview] *Cray XD1 System Overview Release 1.1*.

[XT3Overview] Cray XT3 System Overview Version 1.0, February 2005.

8 Acknowledgment

This work was sponsored by the U.S. Department of Energy's Office of Advanced Scientific Computing Research and performed at the Oak Ridge National Laboratory, managed by UT-Battelle, LLC under contract number DE-AC05-00OR22725. This work is partially sponsored by the Laboratory Directed Research and Development Project at ORNL.