Networking the National Leadership Computing Facility

Steven Carter

Network Lead, NCCS

Oak Ridge National Laboratory

scarter@ornl.gov

Outline

- Introduction
- NCCS Network Infrastructure
- Cray Architecture Overview
- NCCS Enhancements
- Future Work
- Summary

Introduction

- Big Machine. Needs to be shared.
- The goal is to enable science:

Area	Present-2008	2008-Beyond	Remarks
High Energy Physics	100 Gb/s	1 Tb/s	High Throughput
Climate	160-200 Gb/s	n Tb/s	High Throughput
SNS Nanoscience	1 Gb/s	n Tb/s	Remote Control & High Throughput

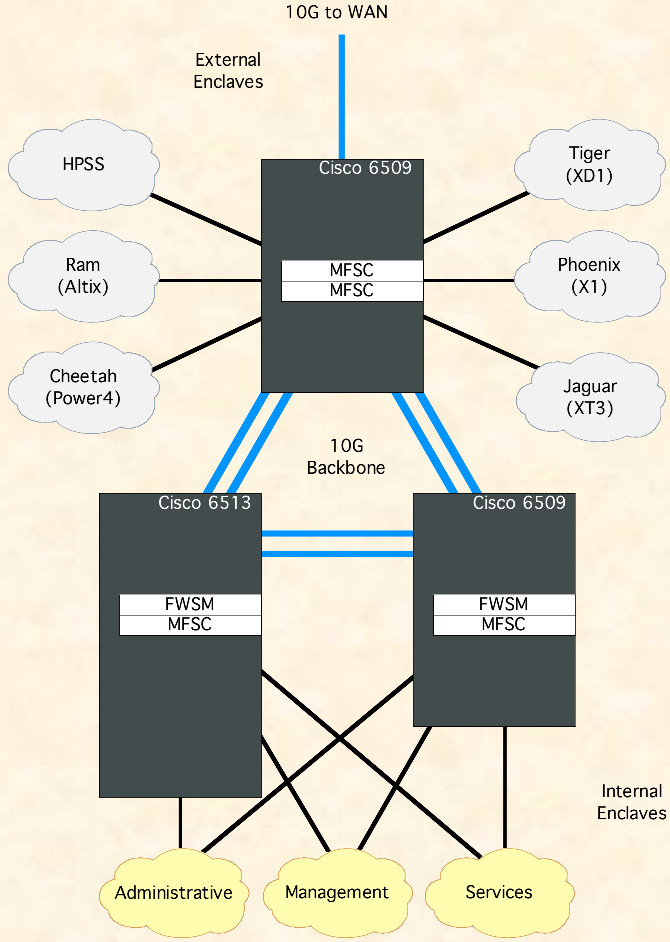
- Most of our work has been specifically aimed at make applications more productive (e.g. TSI).

Introduction (Cont.)

- Many problems trying to get to 10G:
  - Stock TCP has problems at high speeds.
    - Many possible solutions: TCP variants, UDP protocols, L2 protocols, etc.
  - 10G cards are a burden on the system.
    - This is getting better: TOE, less memory movement.
  - Many system components cannot handle 10G rates (e.g. PCI-X too slow).
    - PCI-Express on the way out.
  - Many disk subsystems cannot handle 10G rates.
    - Still a big problem.
  - Security (Firewalls, IDS) has not caught up.
    - Some 10G solutions: MeteNetworks, Endace.

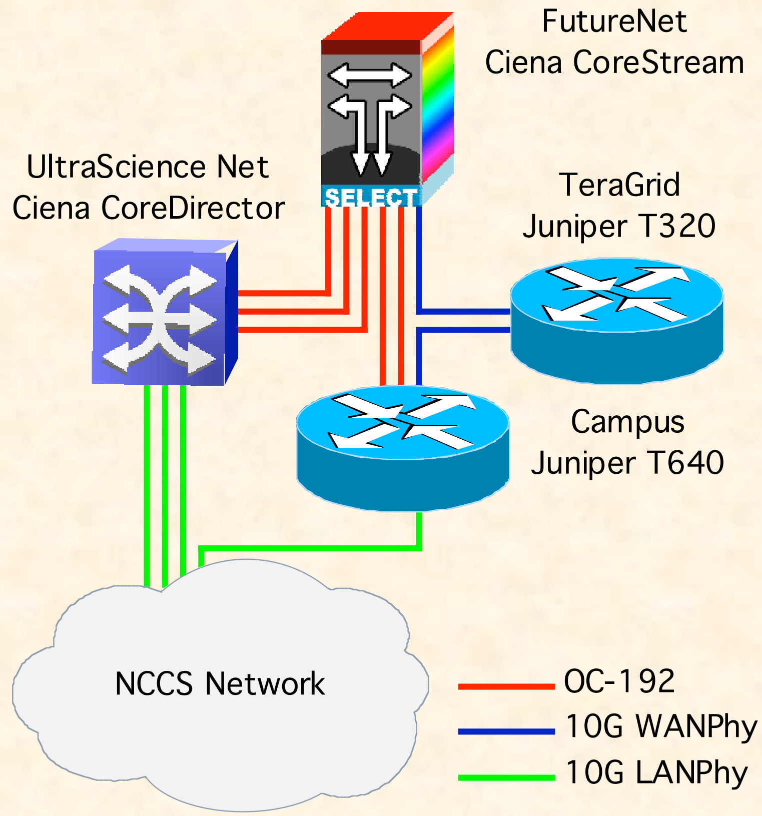
NCCS Network Infrastructure (LAN)

- The NCCS is making a substantial commitment in local- and wide-area networking to enable the NLCF machines to produce good science.
- Local-Area Network:
  - 3 x Cisco 6500 series switches.
  - 10G aggregation switch.
  - 10G Backbone.
  - Hybrid firewall/ACLs.
  - 10G interface to wide-area.
- Try to reach a good compromise between security and performance.

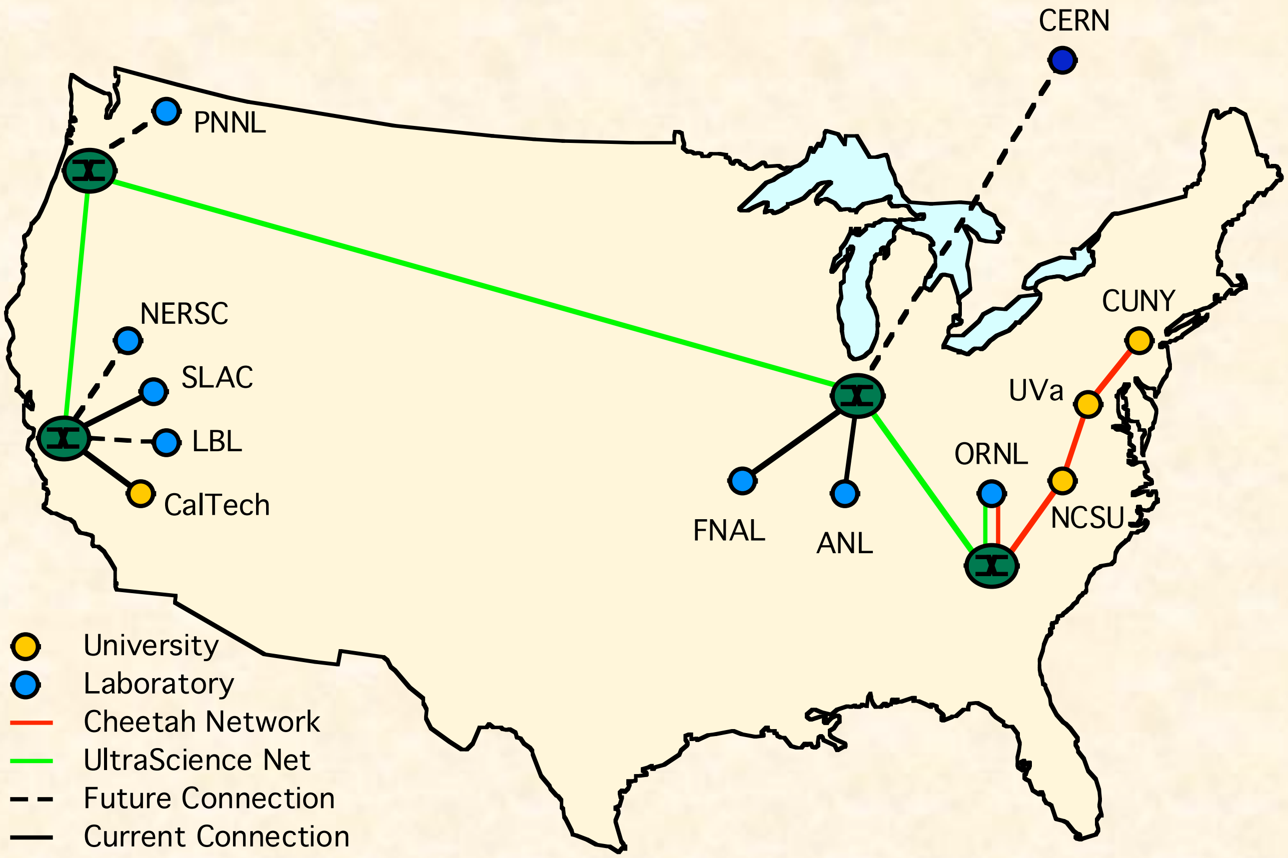


NCCS Network Infrastructure (WAN)

- **Wide-Area Network:**
  - **ORNL Connector:**
    - **1Tb/s Total Capacity**
    - **Will provide 10G circuits to ESNet, Internet 2, Teragrid, UltraScience Net, Cheetah Network.**
  - **DOE UltraScience Net, NSF Cheetah Net:**
    - **Developing technology to enable application controlled dedicated circuits (among other things).**
    - **UltraScience and Cheetah will peer at ORNL giving coast to coast access.**



DOE UltraScience + NSF Cheetah

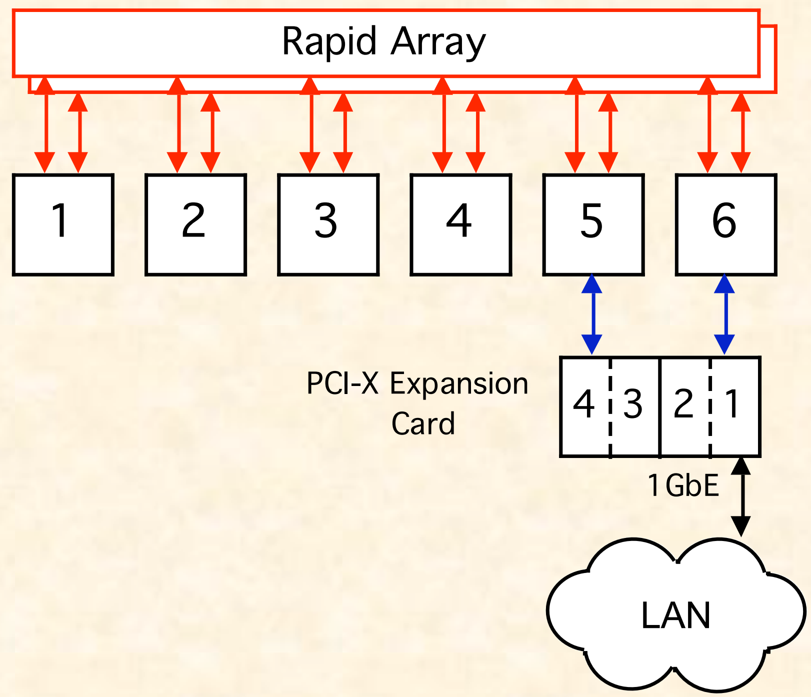


Cray Architecture Overview

- Each machine has slightly different ways of performing network I/O. Some are better than others, but you have to know what you are dealing with to use it effectively.

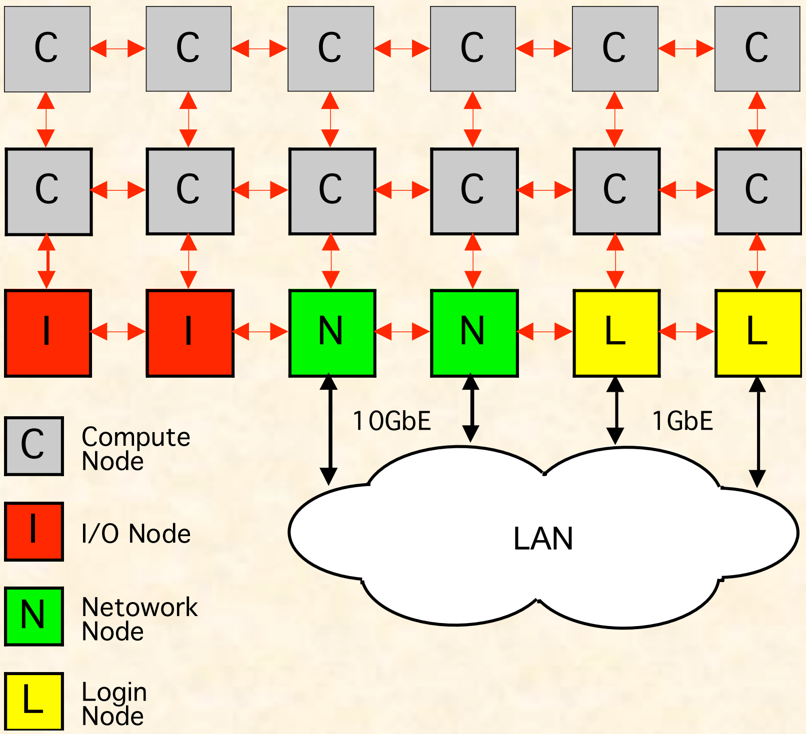
Cray XD1

- Nodes 5 & 6 own the physical interfaces.
- Nodes 5 & 6 present the MAC & IP addresses for the nodes with virtual interfaces.
- Bridging is done over the RapidArray.
- Penalty for using off-node interfaces.
- Some overhead on nodes 5 & 6.



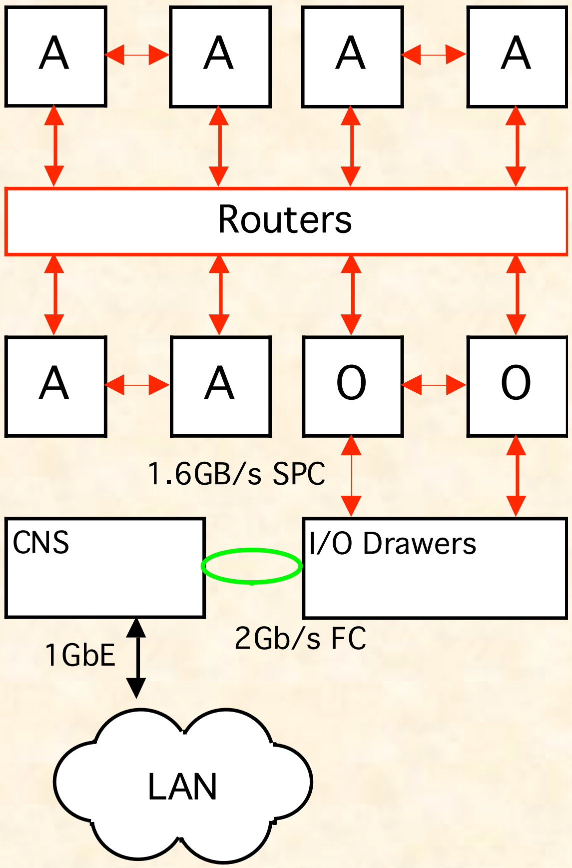
Cray XT3

- Login nodes have 1G interfaces.
- Network nodes have 10G interfaces.
- Application nodes cannot open sockets.
- IP-over-Portals network amongst the nodes.
- Penalty for using off-node interfaces.
- Not obvious how to use all of the interfaces effectively.



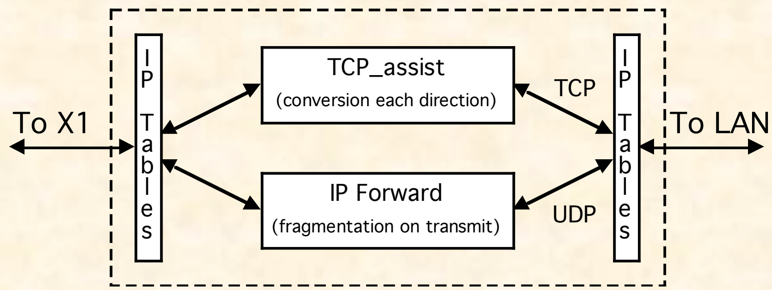
Cray X1

- Socket calls made on App Nodes are suspended and migrated to OS Nodes.
- ~500 system calls/sec from App Node. ~5000 system calls/sec from OS Node.
- OS Nodes connect to I/O drawers via SPC.
- I/O drawers have PCI-X bridges with Fiber Channel cards.
- CNS connects via Fiber Channel.



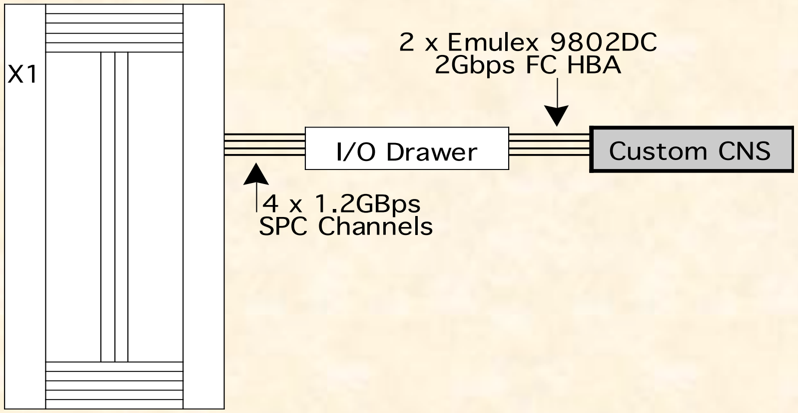
Cray X1: CNS

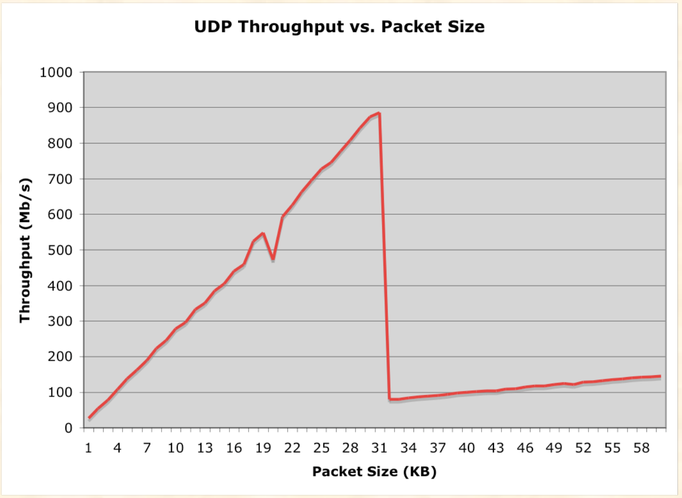
- Communicates with X1 via IP-over-FC w/ ~64k frames.
- Iptables diverts TCP streams to tcp_assistd and masquerades incoming and outgoing packets.
- tcp_assist terminates socket and opens another to the destination.
- Reading from one socket and writing to the other, tcp_assist either fragments or coalesces packets to match the MTU.
- tcp_assist forwards TCP packets faster then ip_forward.
- UDP packet are handled my the stock ip_forward functionality of the system.



CNS Testing, Round 1

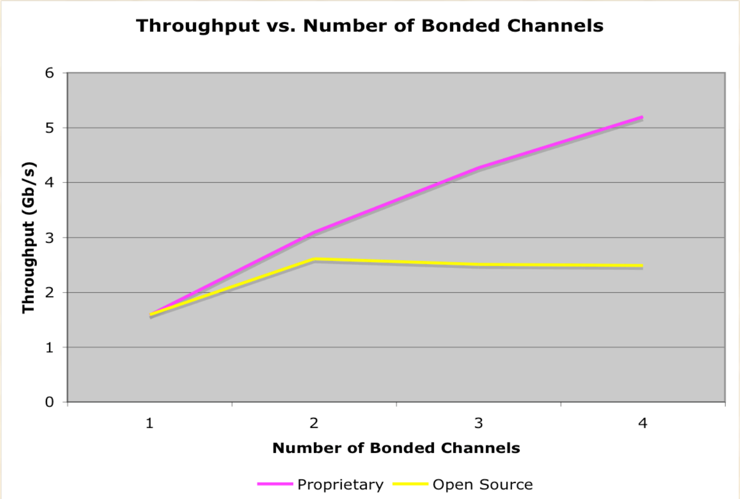
- Hardware:
  - Standard CNS w/2 x Emulex 9802DC
- Software:
  - CNS 1.2 base
  - 2.4 kernel.
- UDP "Knee"
- Net100 modifications:
  - Re-wrote tcp_assistd (ships in CNS 1.4 and above… sorry for any problems).
  - WAN performance increase by 400%.
- Bonded interface testing:
  - 1-4 Interfaces.
  - Proprietary and Open source Emulex drivers tested.

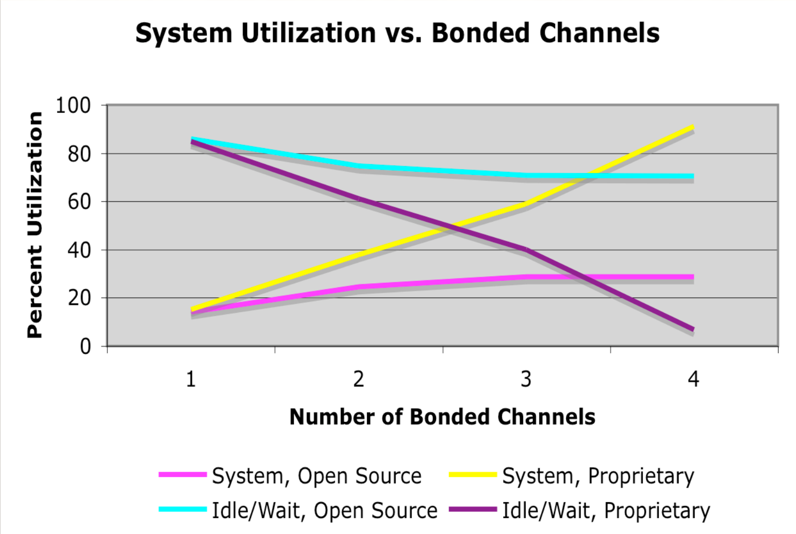




Round 1 Results

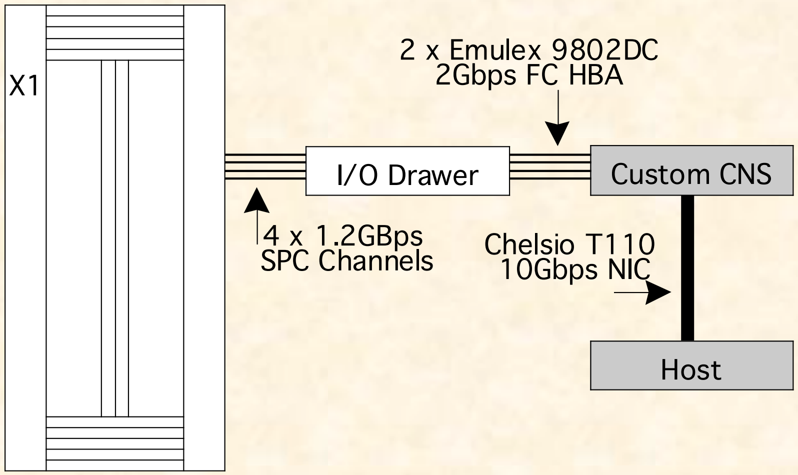
- Big difference between Open Source and Proprietary Drivers.
- Proprietary Drivers:
  - Higher throughput.
  - Higher system utilization.
- Open Source Drivers:
  - Lower throughput.
  - Lower system utilization.
- Conclusion:
  - System utilization too high to use effectively without 10G NIC with good offload.





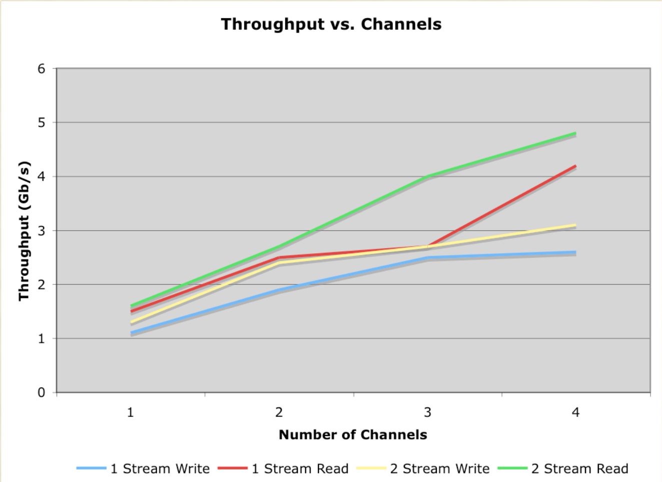
CNS Testing, Round 2

- Hardware:
  - 2.2GHz Dual Opteron
  - 2 x Emulex 9802DCs
  - 1 x Chelsio T110 10G NIC.
- Software:
  - Suse 9.2 base.
  - 2.6.6 kernel.
  - ORNL's tcp_assistd.
  - Emulex 2.10 drivers.
- Tested both bonded interfaces and non-bonded interfaces.



Round 2 Results

- Throughput of 5Gb/s (10Gb/s aggregate).
- 1.8 Gb/s file transfer using bbcp over bonded channels.
- Even unbonded channels can be exploited for single transfers (e.g. GridFTP).



Future Work

- Put the SuperCNS into production with the test networks (i.e. UltraScience Net, Cheetah) for TSI.
- Port IP-over-FC Functionality to Emulex's latest open source drivers (v8).
- Same type of work with the XD1, and XT3.

Summary

- It will be challenging to meet the networking needs stated by the various science areas.
- A holistic approach needs to be taken achieve these goals (i.e. local- and wide-area, host tuning/design, application modifications).
- NCCS's current work has paid dividends in enabling scientists to do their work (TSI can now transfer files in hours instead of days).

References

- **[Net100]** ***Net100*. http://www.net100.org/.**
- **[X1Overview]** ***Cray X1 System Overview Version 2.4*.**
- **[XD1Overview]** ***Cray XD1 System Overview Release 1.1*.**
- **[XT3Overview]** ***Cray XT3 System Overview Version 1.0*.**

Acknowledgments

- 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.

- The submitted manuscript has been authored by a contractor of the U.S. Government under Contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of the contributions, or allow other to do so, for U.S. Government purposes.

The End

- Comments? Questions? Criticisms?