Cray Networking on Product Line Systems

Jay Blakeborough

May 15, 2003

A Bit of History

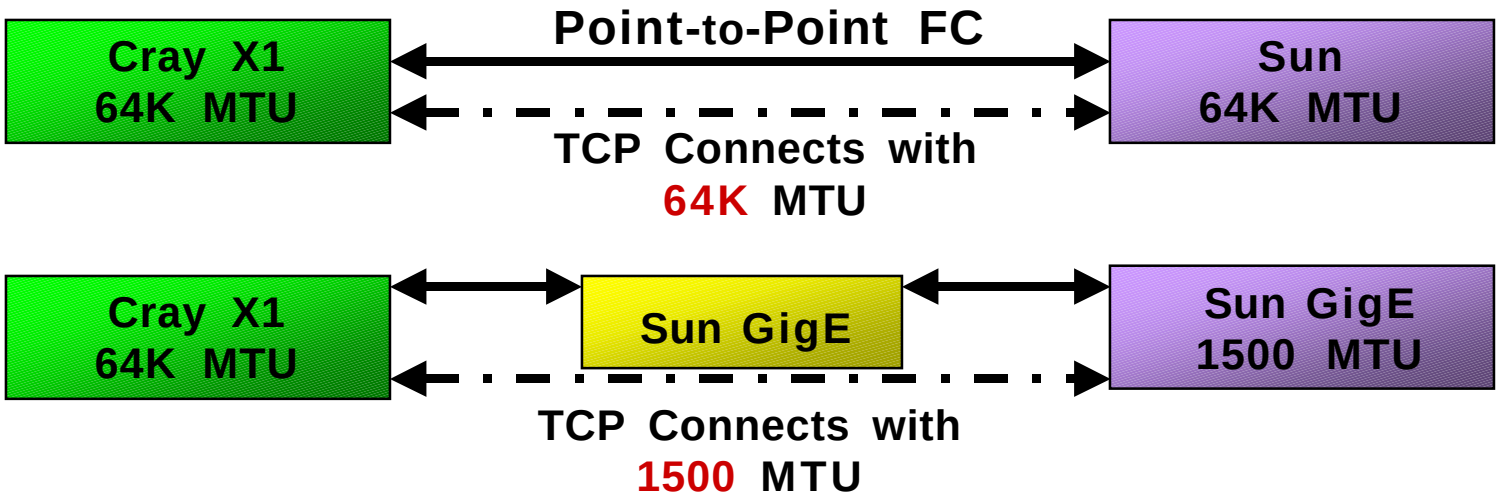
- Traditional Cray PVP platforms performed well on large MTU interfaces (e.g. 600+ Mb/s on 800 Mb/s HiPPI with 64K-byte MTU)
- Same platforms performed ***poorly*** on small packets (e.g. 30 Mb/s on 100 Mb/s Ethernet with 1500-byte MTU)
- Cray L7R released in late 2001 and provided 90 Mb/s on 10/100 Ethernet and 350 Mb/s on Gigabit Ethernet (GigE)

Cray X1 Beta Networking Plan

- Utilize IP over Fibre Channel to a Sun server and bridge to a 1500-byte MTU GigE network (Sun server used as pseudo-Cray X1 for testing)
- Performance was much less than expected, even with large write sizes (11-15 Mb/s)
- UNICOS networking déjà vu
- Sun IP over Fibre Channel performance was sub-optimal (~40 MB/s with 64K MTU)

Cray X1 Networking

- Began Evaluation of Options to Increase Performance in 1/2002:
  - Tune networking parameters on the Sun and the Pseudo Cray X1



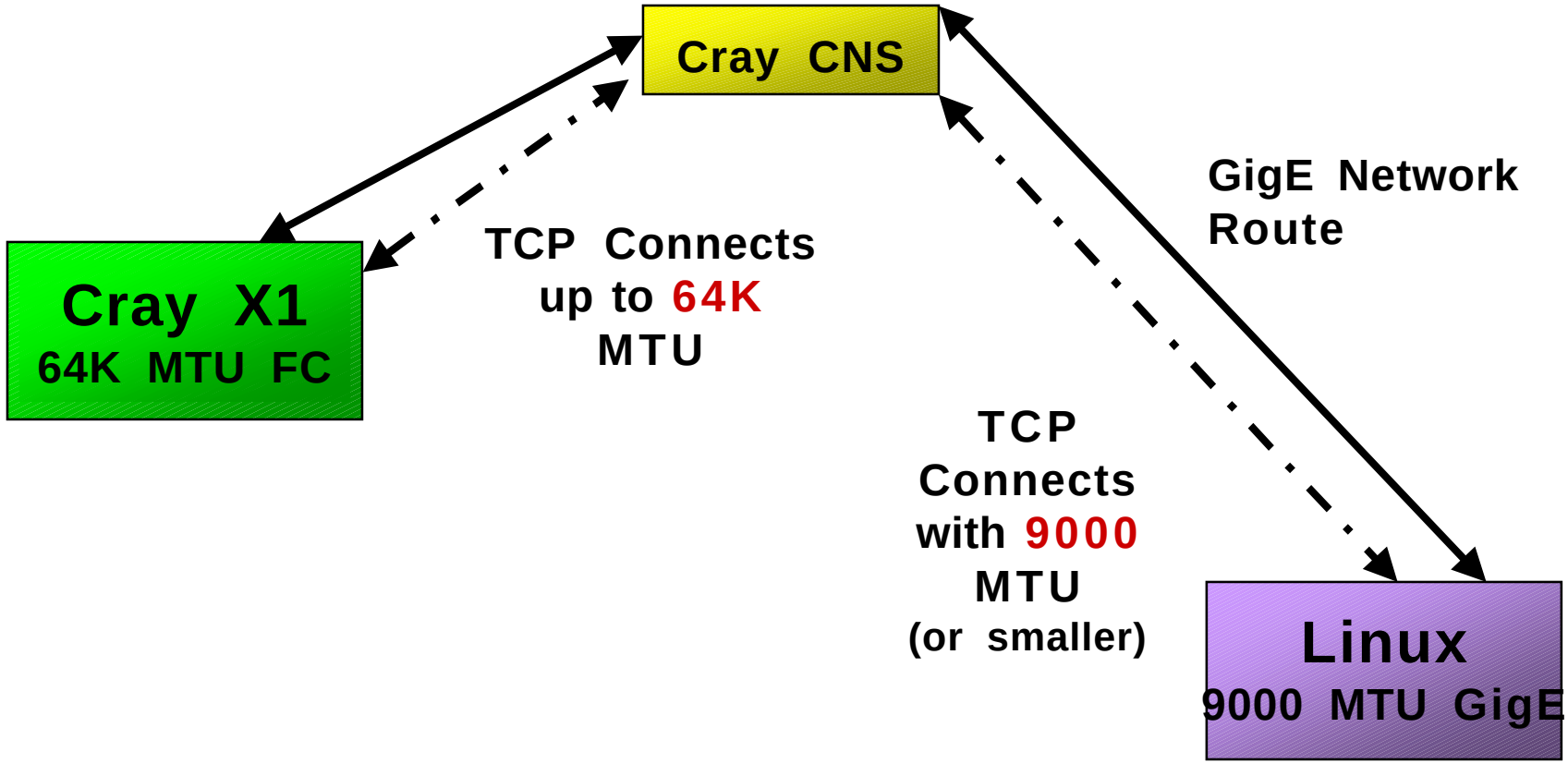
Cray X1 Networking

- Additional Options Considered:
  - GigE Off-load NIC directly attached to the Cray X1 PCI-X bus
    - Cray X1 supports only PCI-X (not PCI)
    - No PCI-X GigE off-load NICs available at the time
    - No HiPPI connectivity
  - Utilize/Improve Cray L7R Technology
    - Fibre Channel experiments with commodity hardware showed promise
    - Good experiences at sites with Cray L7R routers

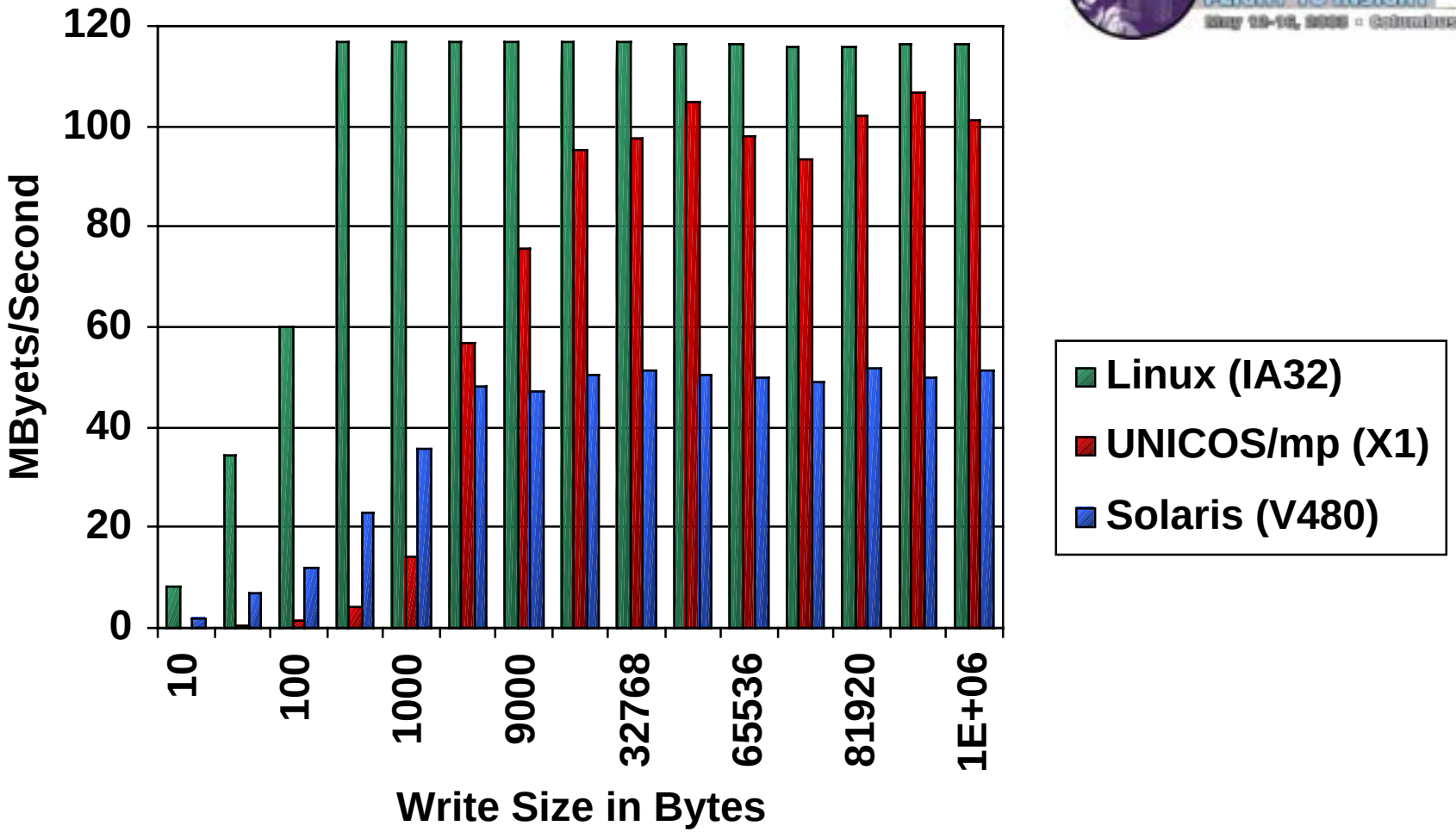
Cray Network Subsystem (CNS)

- Chose to create enhanced version of the Cray L7R, called the CNS
- Using new commodity hardware platform
- Fibre Channel HBA running IP-over-FC to Cray X1
- Gigabit Ethernet (Copper and Fiber)
- HiPPI Available
- CNS 1.0 Hardware and Software released in 12/2002

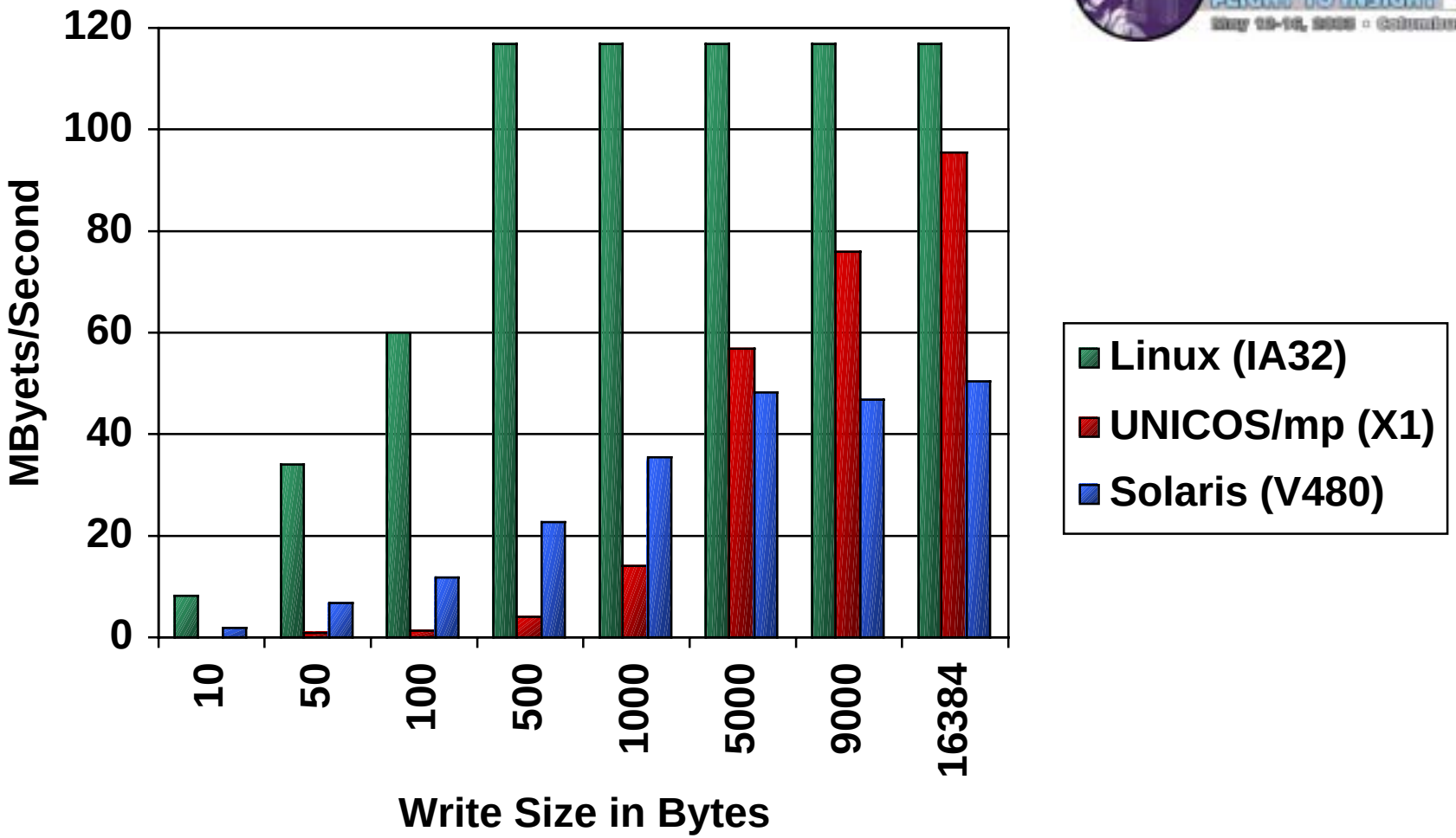
CNS Concept


Cray CNS
Cray X1
64K MTU FC
TCP Connects
up to 64K
MTU
GigE Network
Route
TCP
Connects
with 9000
MTU
(or smaller)
Linux
9000 MTU GigE

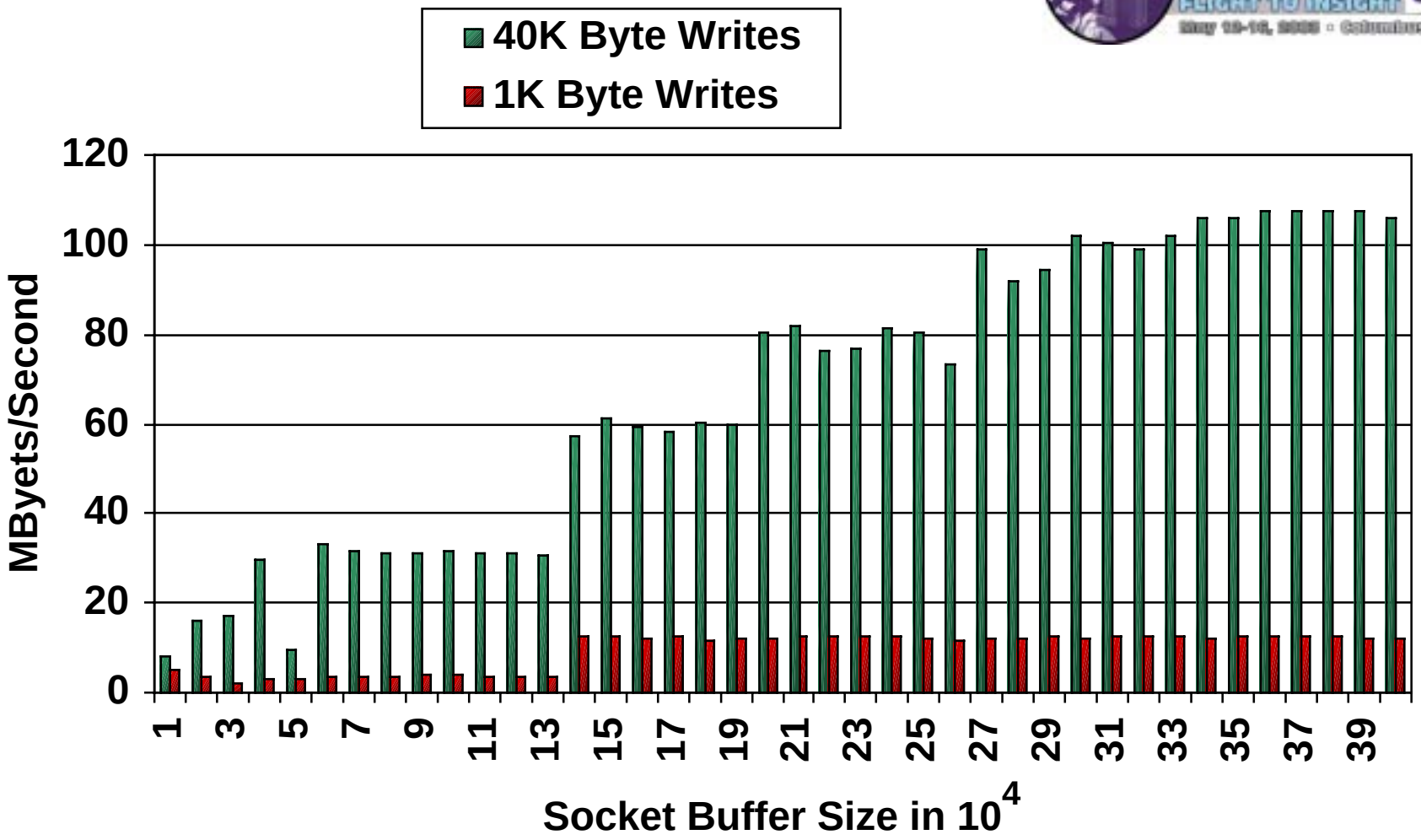
GigE Performace to Linux (1500 MTU)



Focus on Smaller Write Sizes



Effects of Socket Buffer Size



CNS Plans

- CNS 1.1 Release Planned for end of June 2003
  - Driver updates and fixes
  - Improved Installation and Configuration
  - Updates will be easier to install
- Future CNS Release - Not yet scheduled
  - Multiple Fibre Channel connections to the Cray X1 mainframe for resiliency and potential performance improvements
  - Multiple GigE customer network connections

What's the Networking Buzz?

- TCP Off-Load Engines (TOE)
  - Checksum and interrupt hold-off
  - Transmit segmentation
  - Full "fast-path" off-load
  - Evaluating
- Trunking/Bonding
  - Research and evaluation for resiliency and performance
  - Short-term focus is on CNS-to-X1 communication
  - Longer-term focus on customer network connections

What's the Story on 10 GigE?

- No planned commitments for current Product Line Systems (Cray X1 and X1e)
- Other Cray Inc. projects are investigating
- A "fast-path" TOE will likely be required to achieve good performance on a variety of systems
- Plans for Copper-based 10 GigE equipment are just being discussed - may be limited to short distances with dual-cable requirements
- 10 GigE full-bandwidth NICs and switches will be very expensive for a while
- *10 GigE is not yet a mature networking technology*

Looking Forward...

Cray Inc. Product Line Networking Vision:

We will utilize current mature networking technologies to provide industry-standard, single-stream networking performance to our customers. We will design and implement methods to provide system aggregate network bandwidth of at least 8 times the single-stream performance.