Application Sensitivity to Link and Injection Bandwidth on a Cray XT4 System

Cray User Group Conference
Helsinki, Finland
May 8, 2008

Kevin Pedretti, Brian Barrett, Scott Hemmert, and Courtenay Vaughan
Sandia National Laboratories

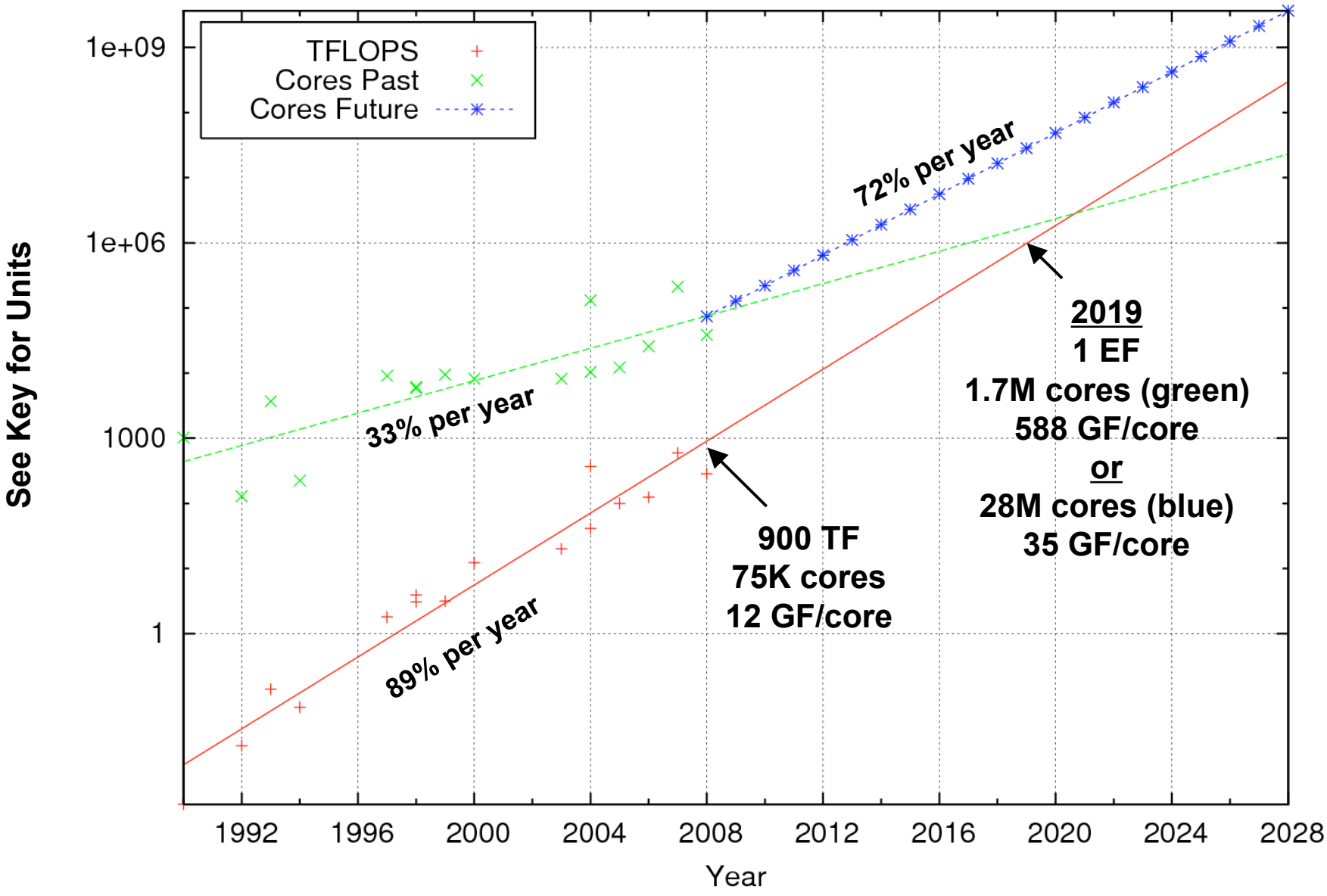
ktpedre@sandia.gov

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Outline

- **Introduction**
- **Link Bandwidth Detuning**
- **Injection Bandwidth Detuning**
- **Application Testing on Red Storm**
- **Future Work and Conclusions**

Challenges: Exponentially Increasing Parallelism, Decreasing Balance Due to Power Constraints



Motivation

- Many challenges ahead on path to Exa-FLOPS
  - Institute of Advanced Architecture (IAA) starting up
  - Need tools to evaluate machine balance trade-offs

- Hmm… wouldn't it be great if we could vary link bandwidth, injection bandwidth, latency, and message rate independently from one another, on a real system like Red Storm?
  - Perform larger and longer experiments than possible with simulation
  - Validate application simulations and models
  - Guide future decisions

Approaches

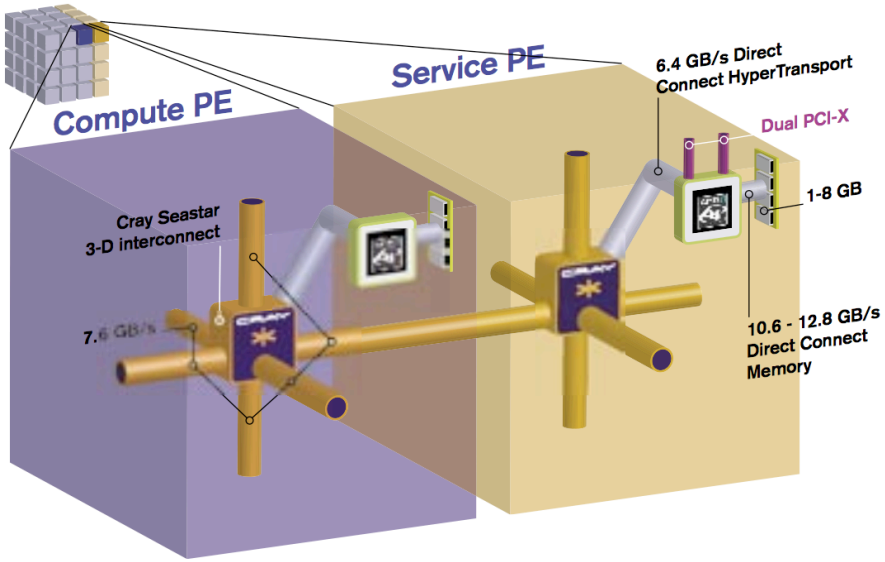
- Application modeling
  - Develop mathematical models describing applications
  - Sandia: http://www.sandia.gov/PMAT/
- Simulators
  - Model hardware with software, FPGA, etc.
  - Speed generally decreases with fidelity
  - Sandia: SST – Structural Simulation Toolkit
    - Processor, memory, and network models (inc. SeaStar)
- Execution driven simulators
  - Run application on real system, virtual time tracked by centralized network scheduler/simulator (possibly SST)
  - Sandia: Seshat
- Empirical experiments
  - This talk
  - Related to MPI detuning work on ASCI Red (Ron Brightwell)

Outline

- Introduction
- **Link Bandwidth Detuning**
- Injection Bandwidth Detuning
- Application Testing on Red Storm
- Future Work and Conclusions

From the XT4 Brochure…

Cray XT4 Scalable Architecture



Interconnect Reliability Features

Each link on the chip runs a reliability protocol that supports Cyclic Redundancy Check (CRC) and automatic retransmission in hardware. In the presence of a bad connection, a link can be configured to run in a degraded mode while still providing connectivity.

The Cray SeaStar2 chip provides a service port that bridges between the separate management network and the Cray SeaStar2 local bus. This service port allows the management system to access all registers and memory in the system and facilitates booting, maintenance, and system monitoring.

What is this "degraded mode" and how is it enabled?

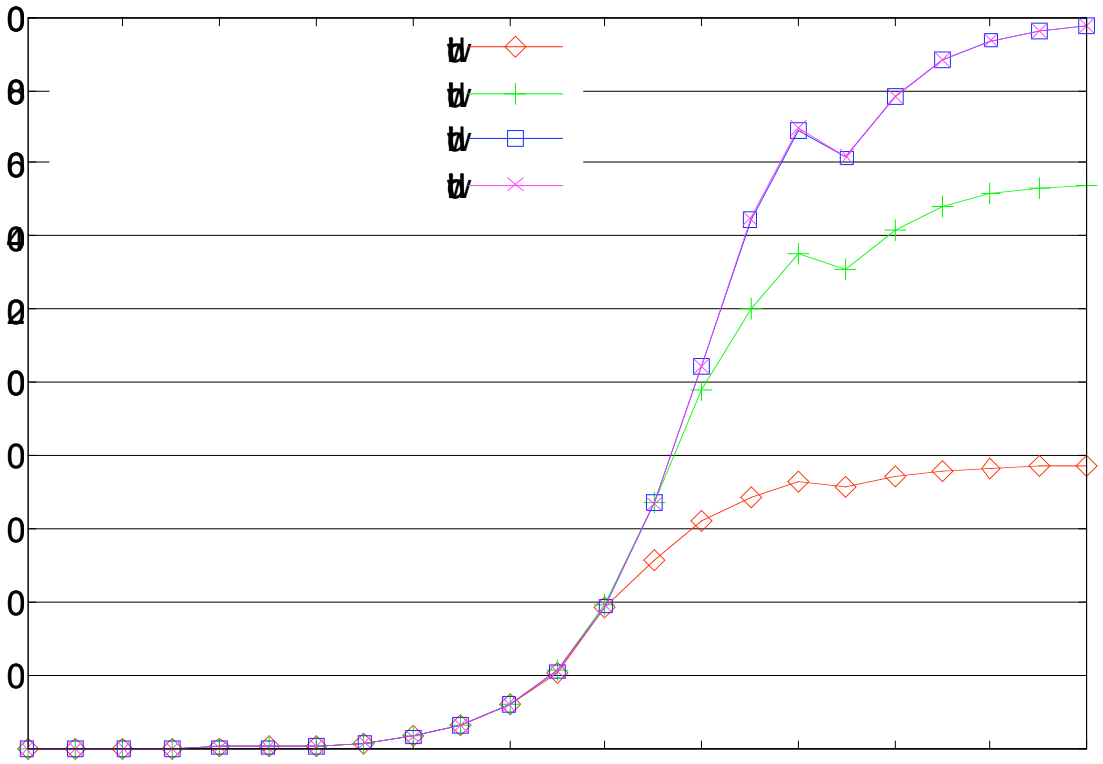
Sandia National Laboratories

Degraded Link Mode

- Network links consist of many parallel wires
- What if one of the wires/drivers goes bad?
  Options:
  - Fix or disable link and reboot (answer today)
  - Reroute on-the-fly, avoiding bad link
  - Disable faulty wire(s), distribute traffic over remaining wires => degraded mode

- Degraded mode can be enabled on a per link-type basis via the rm.ini file (/opt/cray/etc/rm.ini)

Proof of Concept on XT4 Development Cage

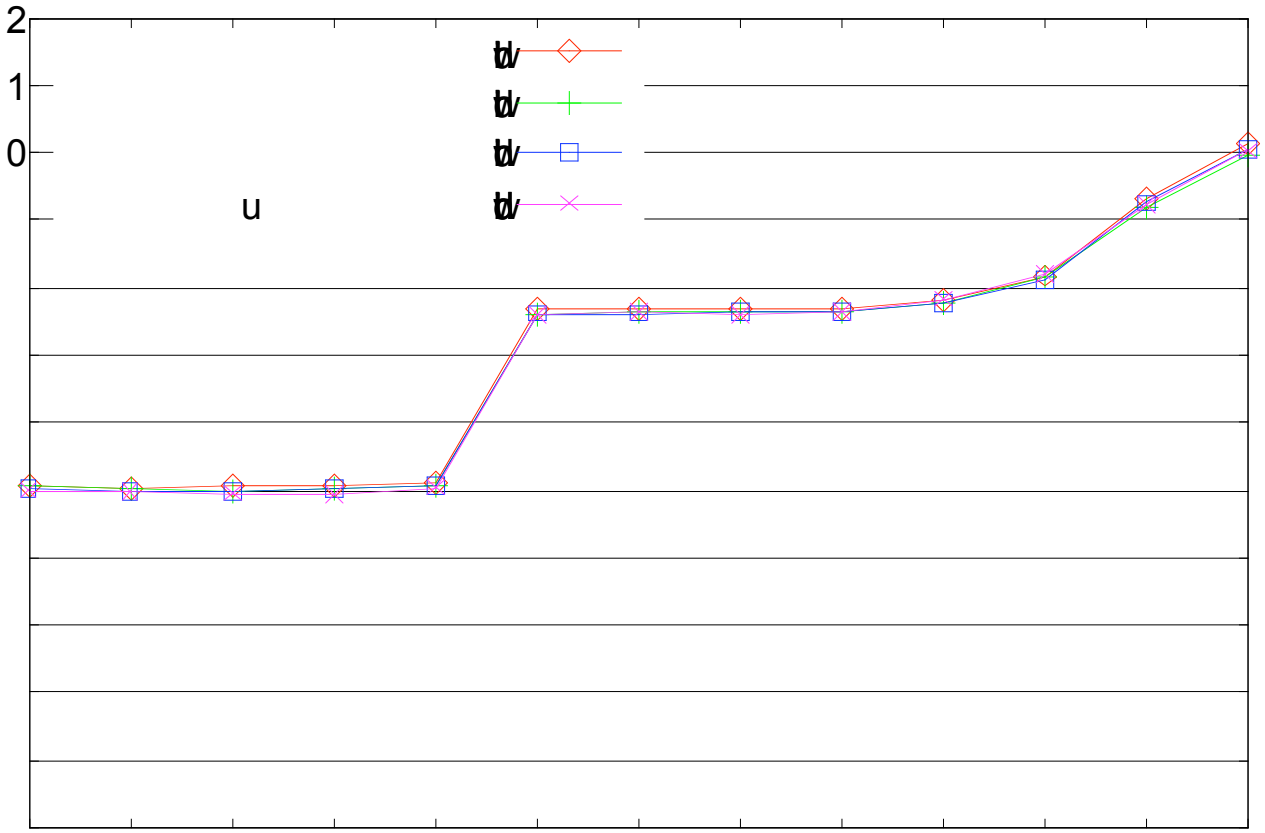


Point-to-point bandwidth is clearly throttled by host injection rate.

No difference between ¾ and Full link bandwidth configurations.

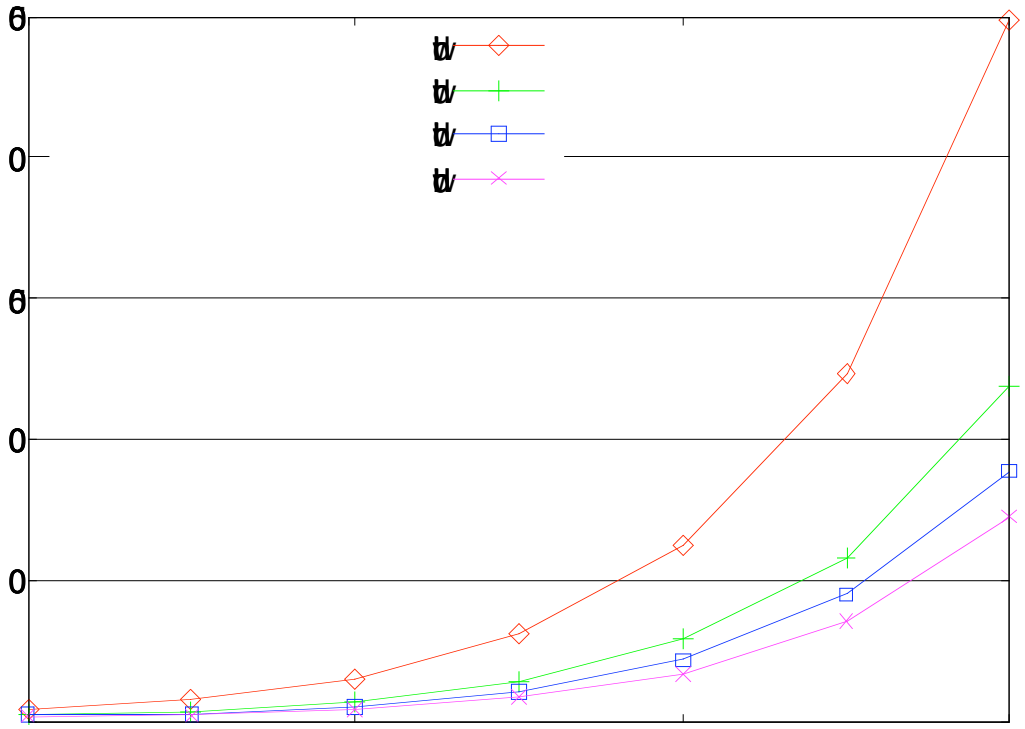
Full link bandwidth is approx. 773 MB/s * 4 = 3092 MB/s in each direction.

Latency Remains Unchanged



Sandia National Laboratories

16-node MPI Alltoall



Difference <= 1% compared to full link bandwidth for message sizes:

- ¼ BW: <= 2 KB
- ½ BW: <= 4 KB
- ¾ BW: <= 8 KB

At 4 MB msg size, compared to full link bandwidth:

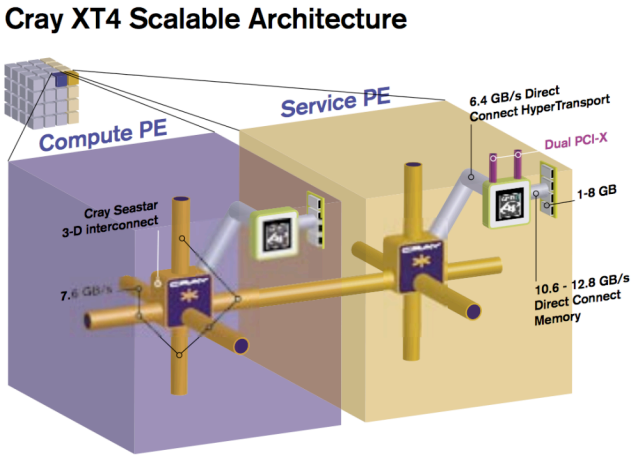
- ¼ BW: 3.42x worse
- ½ BW: 1.63x worse
- ¾ BW: 1.22x worse

Outline

- Introduction
- Link Bandwidth Detuning
- **Injection Bandwidth Detuning**
- Application Testing on Red Storm
- Future Work and Conclusions

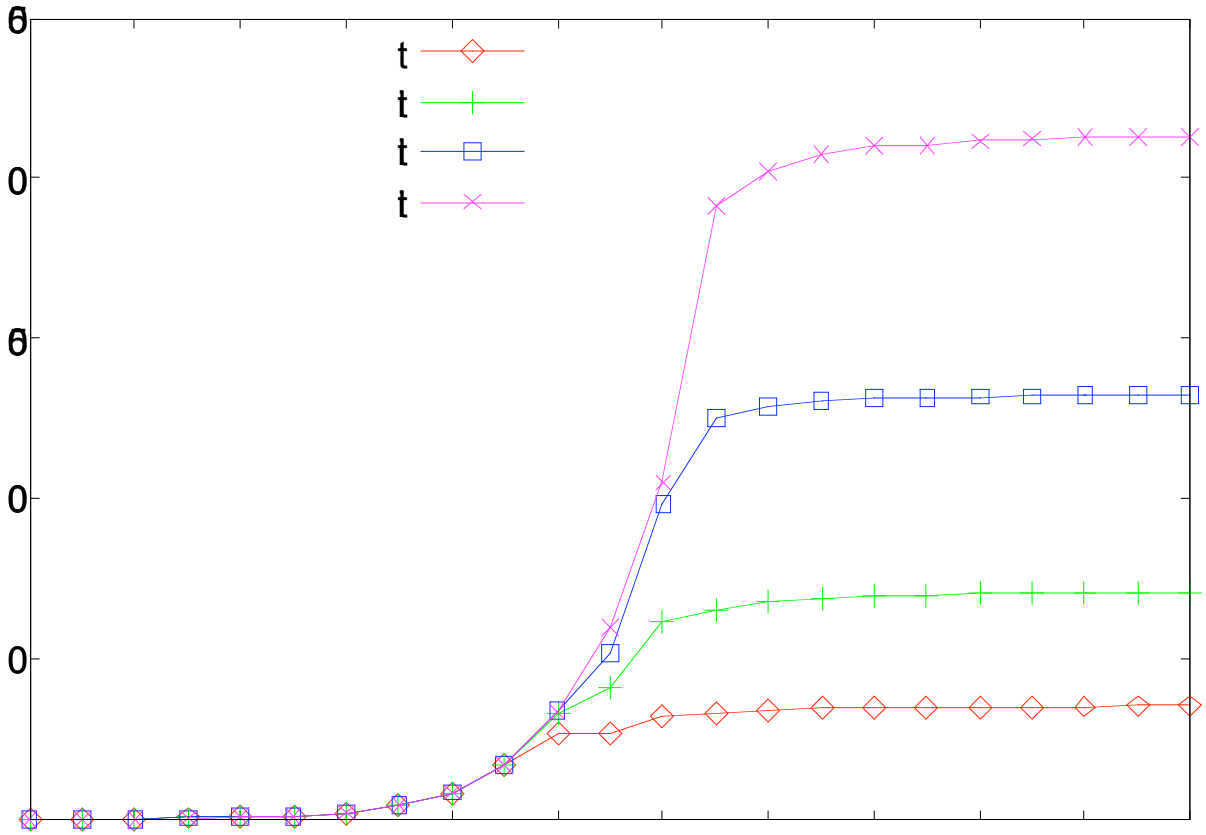
HyperTransport Detuning

- HyperTransport link between Opteron and SeaStar setup at boot by Coldstart on Cray XT, BIOS on standard PCs
- Anyone may query HT widths and frequencies supported by SeaStar via the PCI config space:
  - 8 or 16 bits wide
  - 200, 400, or 800 MHz
- HT link config currently hard-coded in Coldstart
- Ran into HT watchdog timeouts with 200-MHz, 8-bit config (400 MB/s), easy fix via xtmemio



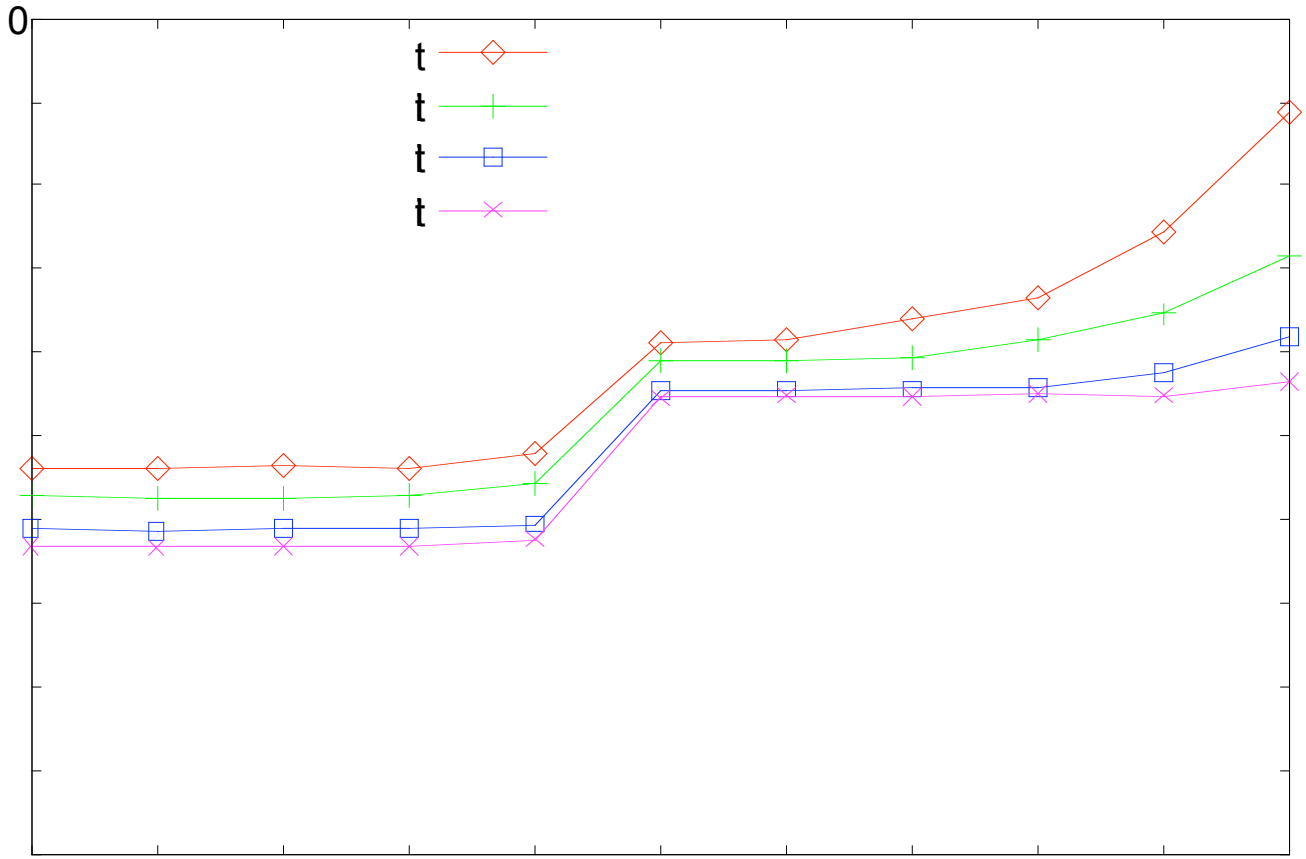
Sandia National Laboratories

HyperTransport Link Detuning: Effect on Bandwidth



Sandia National Laboratories

HyperTransport Link Detuning: Effect on Latency



Sandia National Laboratories

Outline

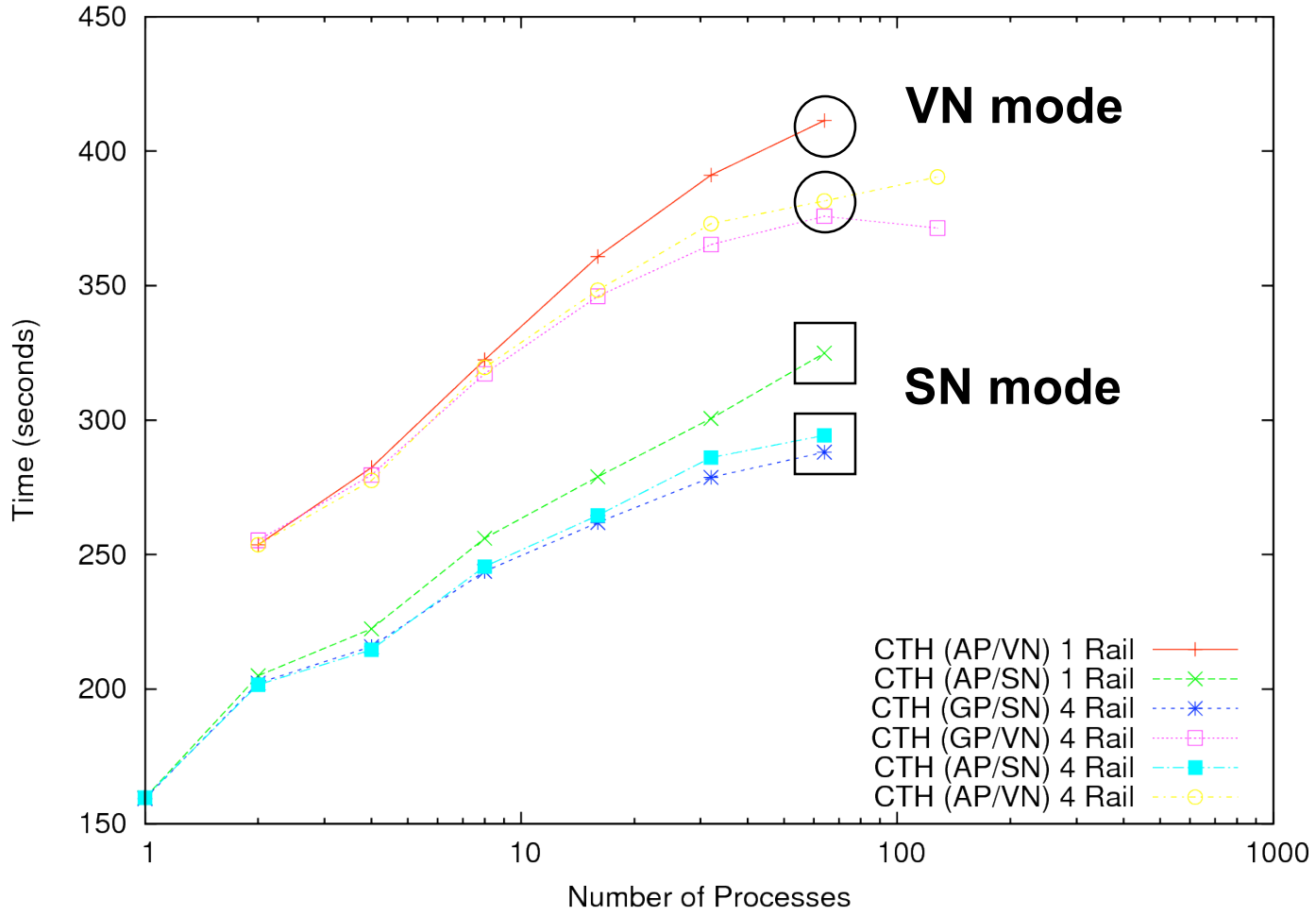
- Introduction
- Link Bandwidth Detuning
- Injection Bandwidth Detuning
- **Application Testing on Red Storm**
- Future Work and Conclusions

Single Cabinet Testing, 80-nodes

- Used to build confidence before full Red Storm testing

- Applications tested:
  - CTH
    - Shock physics
    - Weak scaling, non-AMR
  - HPCCG
    - Sparse conjugate gradient solver mini-application
    - Strong and weak scaling
  - LAMMPS
    - Molecular Dynamics
    - Strong and weak scaling
  - SAGE
    - Hydro-dynamics
    - Strong scaling

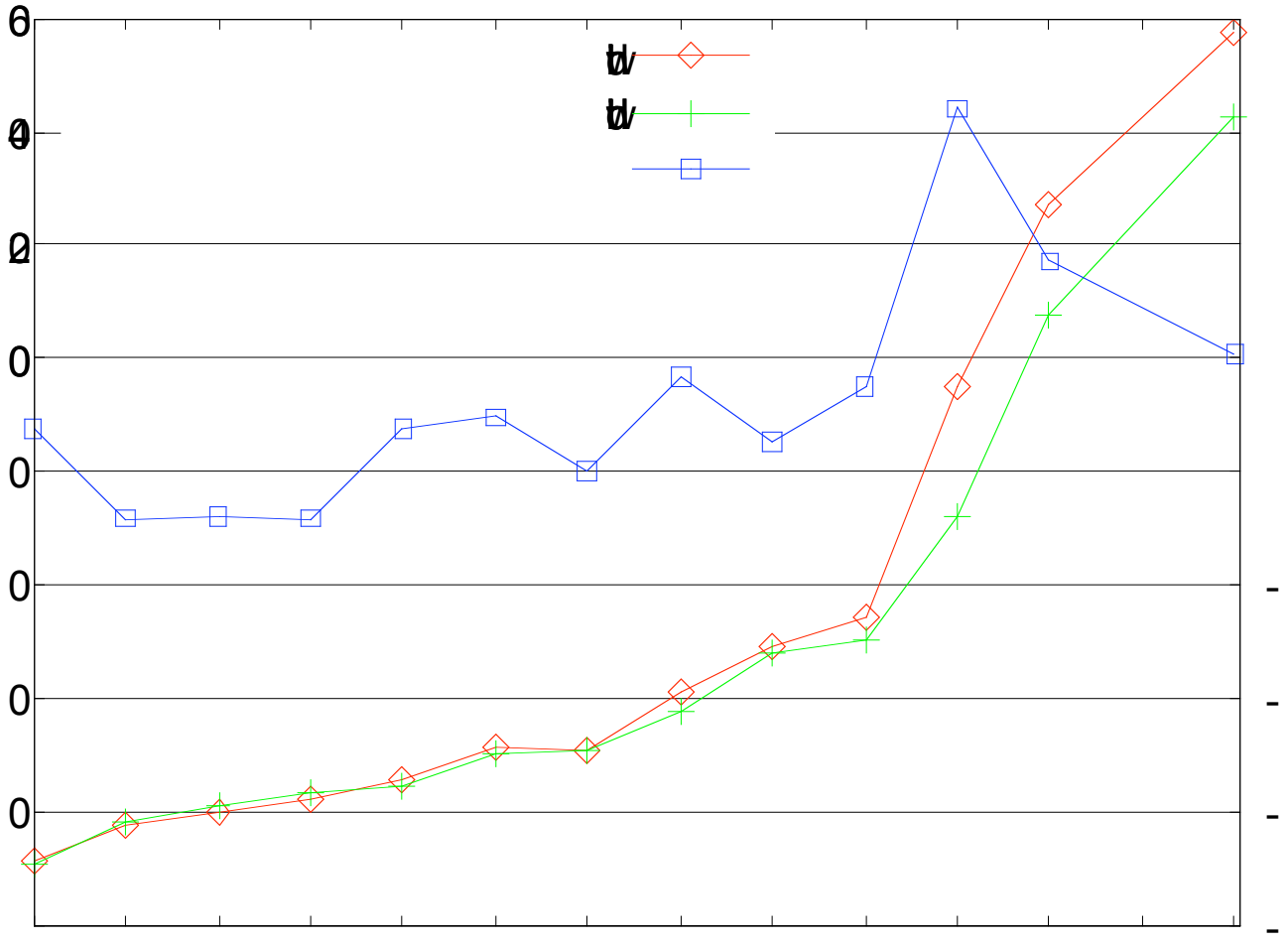
Only Application with Significant Difference was CTH



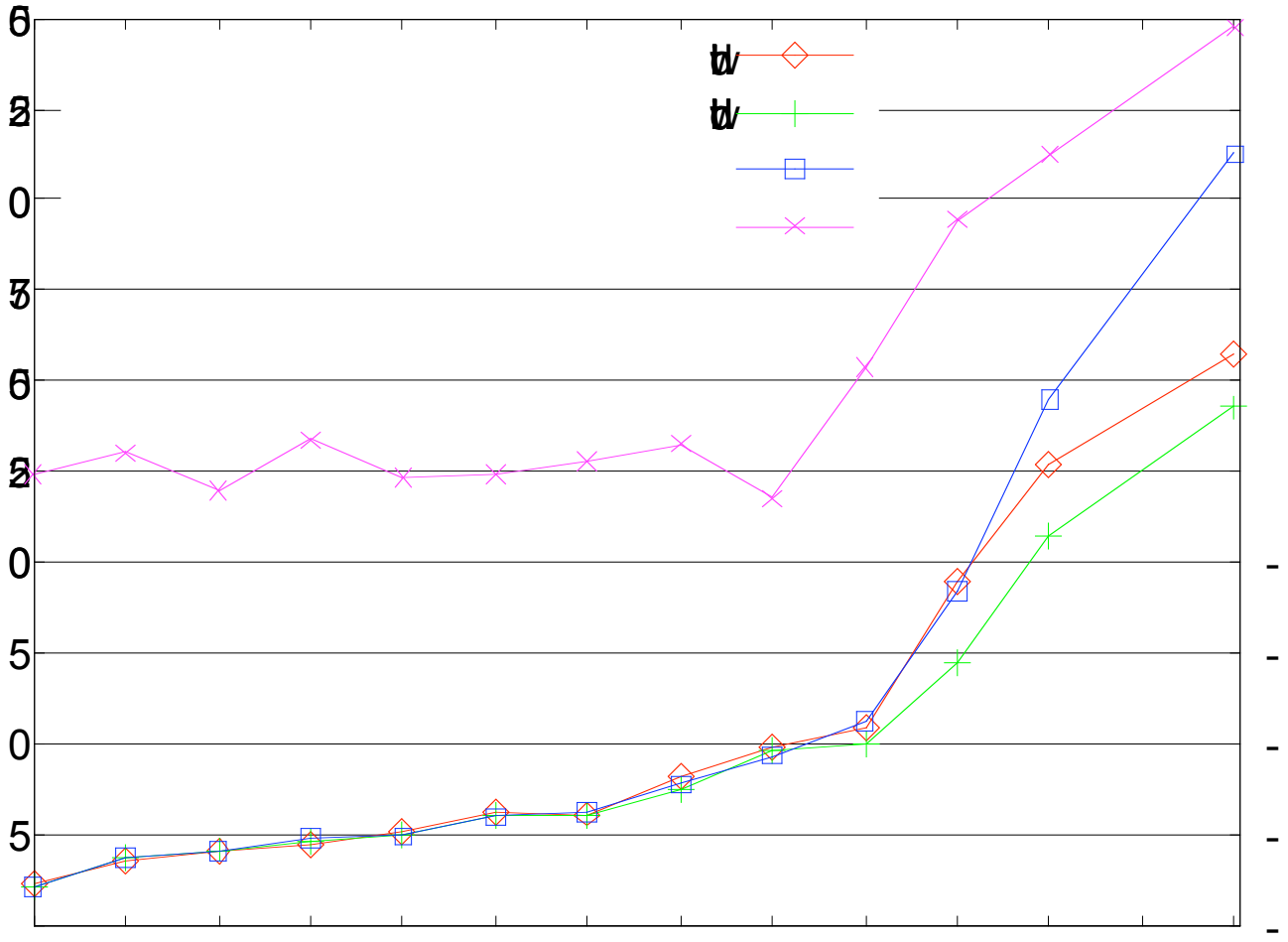
Sandia National Laboratories

Red Storm Jumbo Testing in Degraded Mode

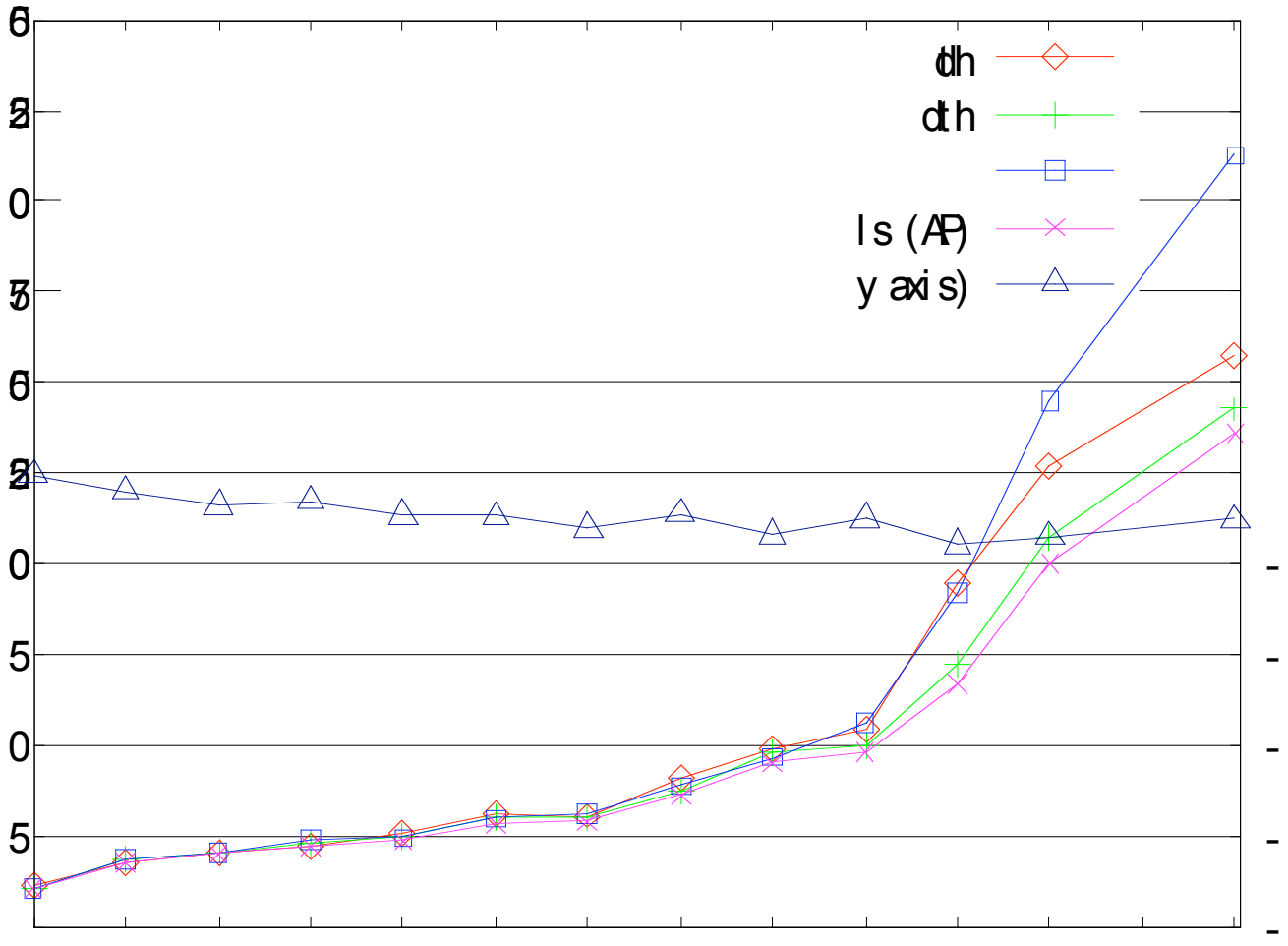
- Somehow convinced management that Red Storm would not be harmed (Thanks! ☺)
  - Successful cabinet testing
  - Simple configuration file change (rm.ini) + reboot
- Testing performed April 22-24, 2008
  - First three days used for Quad-Core Catamount testing and comparison with CNL (Courtenay Vaughan's talk on Wednesday)
  - One 8-hour window for degraded link bandwidth testing
  - Tested CTH and Partisn codes
- Caveats
  - Non-identical node layouts (MOAB vs. Interactive)
  - Only enough time for one trial at each data point



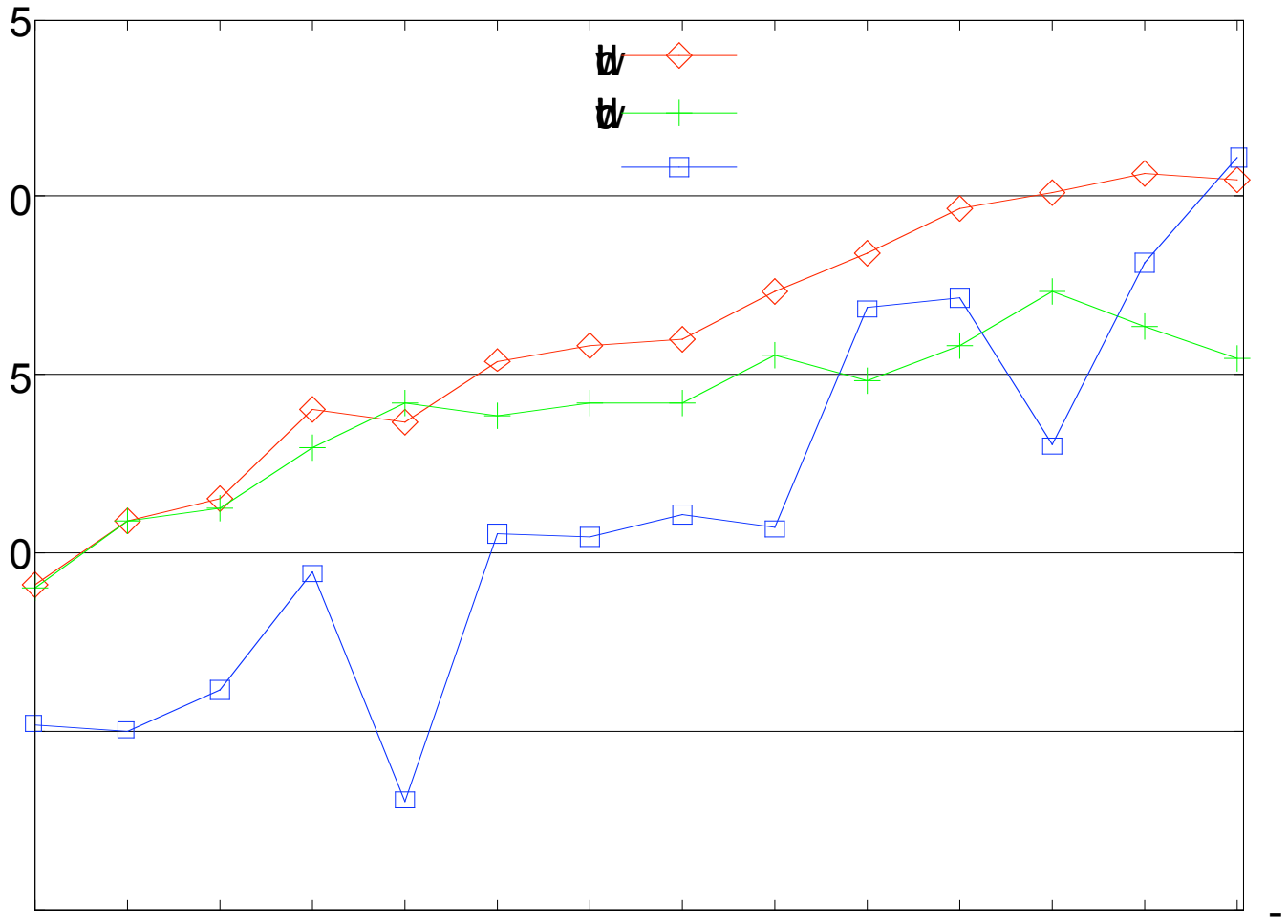
At 8192 nodes, ¼ bandwidth config is 10.3% worse than full bandwidth.
Standard Partisn test problem setup to stress latency, ~50 MB per process.



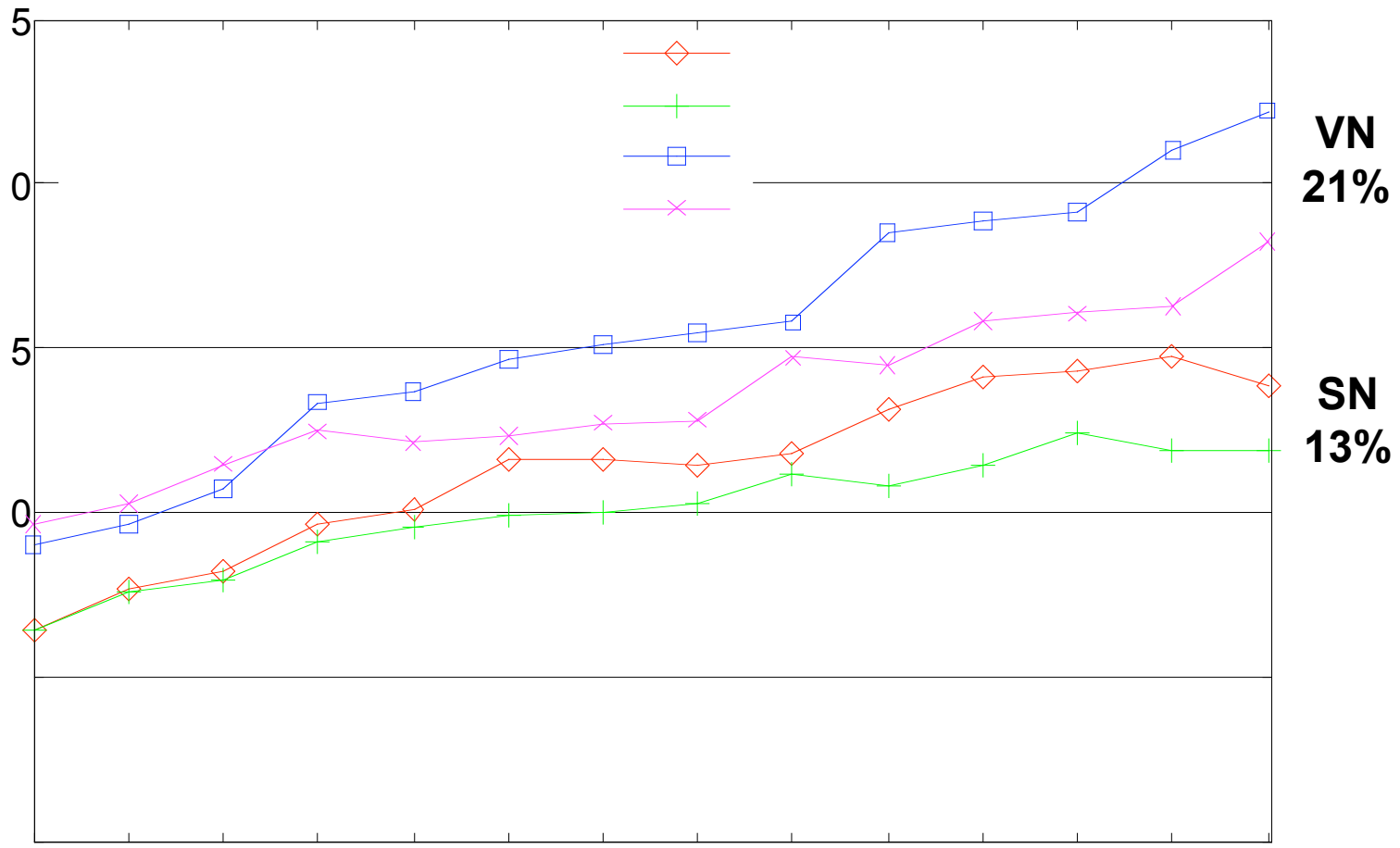
At 8192 nodes, CNL (2.0.44) is 49% worse than Catamount on this problem. Doesn't appear to be a bandwidth issue.



Accelerated Portals (AP) has ~30% lower latency than Generic Portals (GP), but only improves Partisn performance 1-8%.

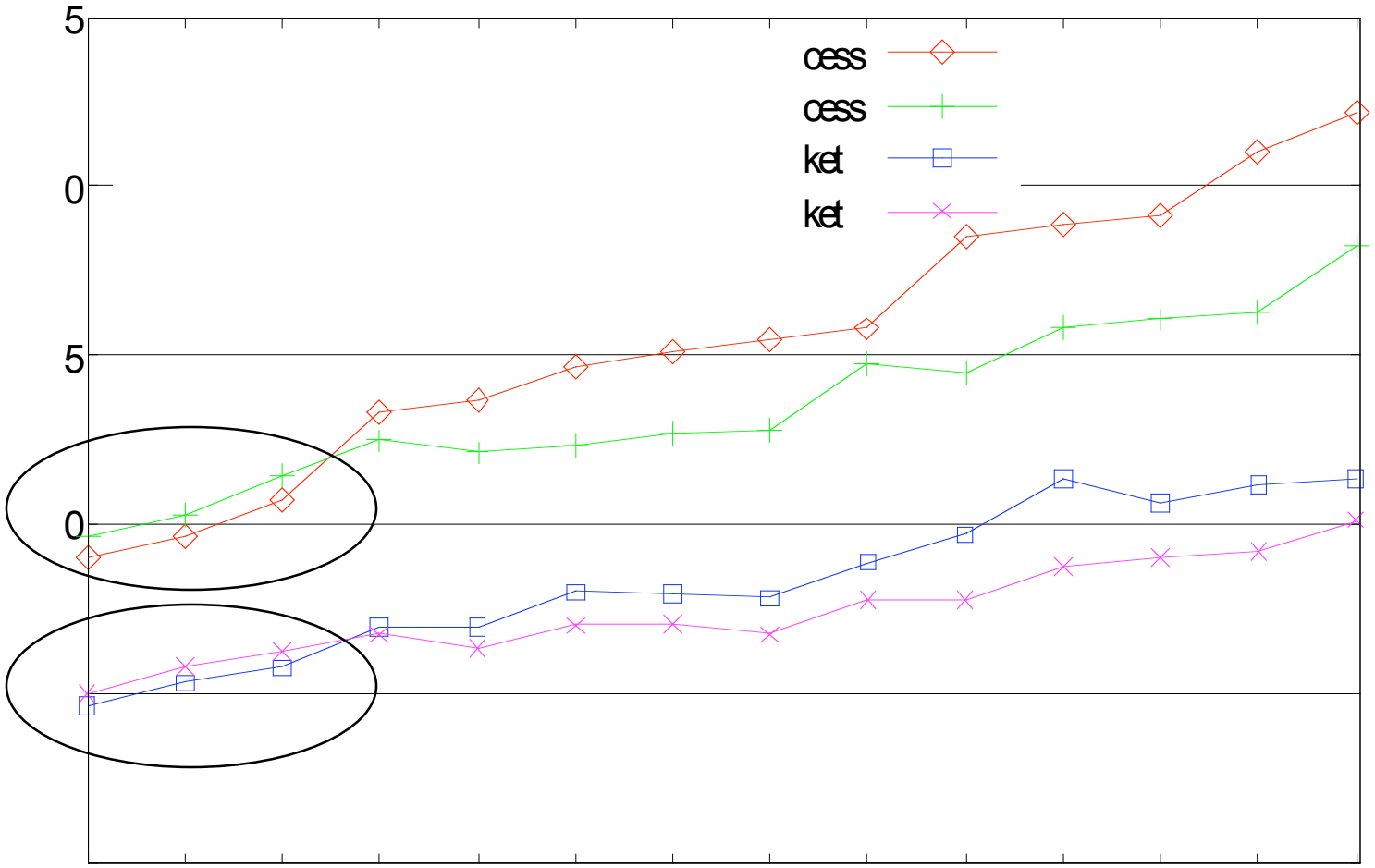


At 8192 nodes, ¼ bandwidth config is 32.6% worse than full bandwidth.
Many ~2.5 MB nearest-neighbor messages for this problem.



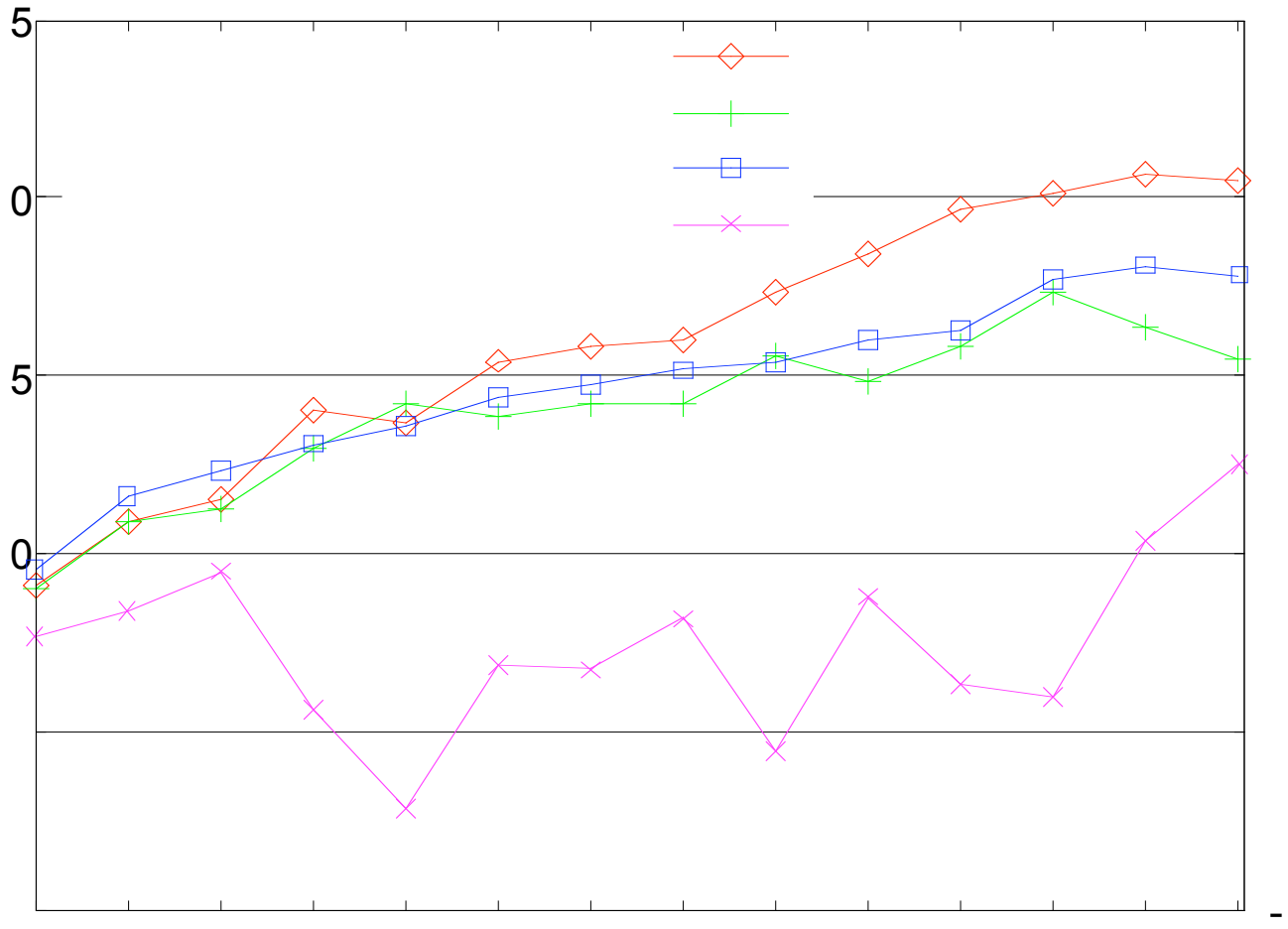
At 8192 nodes, ¼ bandwidth config is 13% worse than full bandwidth for SN mode, 21% for VN mode.

Many ~2 MB nearest-neighbor messages for this problem.



[

Unexplained performance boost in degraded mode for <= 4 nodes (one board) in VN-mode. SN-mode behaves as expected.



More jagged Catamount curve thought to be caused by MOAB, which preferentially allocates 2 GB nodes on Red Storm. CNL tested using interactive mode aprun.

Future Work

- Independent control of message latency
  - Leverage earlier work on Portals event timestamps
  - MPI library modifications
  - Null-entries at head of match list
- CPU frequency and memory detuning

- Application testing using checkpoints from production runs
  - Real problems
  - Run a few timesteps rather than entire problem

Conclusions

- It is possible to independently control link bandwidth and injection bandwidth on Cray XT4 systems
- Application testing on full Red Storm system booted into degraded ¼ link bandwidth mode was successful
  - Partisn: 10.3% worse at 8192 nodes
  - CTH: 13 - 36.2% worse 8192 nodes
- Useful platform for large-scale machine balance experiments

Acknowledgements

- Courtenay Vaughan – Red Storm testing
- Kurt Ferreira – single cabinet testing
- Sue Kelly and Bob Ballance – approving and allocating Red Storm system time
- Cray – consultation