Cray XT3 Experience so far

Horizon Grows Bigger

Richard Alexander
May 2006
raa@cscs.ch

Support Team

Dominik Ulmer

Hussein Harake

Richard Alexander

Davide Tacchella

Neil Stringfellow

Francesco Benvenuto

Claudio Redaelli

\+ *Cray Support*

Recognition

- Mario Mattia
- Paolo Palazzi
- Mario Marchi
- Roberto Anseloni
- Kevin Stelljes
- Don Mengel
- Dave Wallace
- Jim Harrell
- John Metzner
- Steve Johnson

I am sure there are many more I am unaware of!

Short History

- Cray turn over on 8 July 2005
- Initially marked by great instability
- Dec 2005 acceptance
- “production” late Jan 2006

Adventures in Reliability

- Reliability has been a step-wise function
- Three major steps
  - Aug/Sep 05 Seastar 1.5v to 1.6v
    - losing multiple nodes per hour
  - Feb/Mar 06 Vector 18 & 14 fixes - 1.3.17
    - losing multiple nodes per day
  - Today 1.3.21 - losing multiple nodes per week

What is an XT3?

- Massively parallel Opteron-based UP
- Catamount job launcher on compute
  - One executable
  - No sockets, no fork, no shared memory
- Suse Linux on “service nodes”
  - I/O nodes, login nodes, system db, boot
  - “yod or pbs-mom” nodes

Cray XT3 Use Model: Fit the work to the hardware

- **Palu** - 1100 compute nodes: Batch only
- **Gele** - 84 compute nodes: New users
  - compile, debug, scale, interactive nodes
- **Fred** - 56 nodes: Test environment

Current Palu configuration

- 4 login nodes with DNS rotary name
- 2 yod/mom nodes
- Scratch for general users
  - 4 Lustre servers (1 MDS / 15 OSTs)
- Scratch for PSI users
  - 3 Lustre servers (1 MDS / 11 OSTs)
- each OST = 600MB

Palu Load Characteristics

- Current usage:
  - >90+% node utilization
  - Jobs using 64-256 nodes are “typical”
  - Max nodes / job is 768 now going to 1024 out of 1600

- Heavily Oversubscribed
- To be upgraded with 600 more processors (6 racks) and Engenio 6998 dual FC disk controller.

Two groups telling us they have done science they couldn’t do before.

Open issues and next steps

- High speed network not 100% stable
- High-speed file system Lustre young and immature
- Bugs lead to intermittent node failures

BIG DEAL

- Major trouble where subsystems collide: PBSpro, Lustre and CPA

- DDN disk arrays have been problematic!

High Speed Network

- When the high speed network is sick => Nothing Works
- Stability improved VASTLY over time
- Stability still not satisfactory; Cray analyzing problems
- Most errors affect/abort single jobs
- today: reboot the whole system.

File system Lustre

- Genuine parallel file system
- Performance varies from Great to Bad

- Very young and immature feature set
- When Lustre is unhealthy (~once or twice a month) the system must be rebooted

- Real Lustre Errors versus HSN problems!
  Difficult to differentiate

PBSPro batch system

- Standard PBSPro can **not** implement desired management policies (Priority to large jobs, Back filling, Reservations)
- **These are NOT exotic requirements!**

- Cray has delivered a “special” version of PBS with a TCL scheduler - uses standard TCL script
- CSCS is very pleased with the outcome of this collaboration!

Help Cray out of the scheduler business!

Problems du Jour: Intermittent failures

Extremely difficult to diagnose

- Job start failures - one of our highest priority bugs
  - *Extremely difficult to diagnose*
- Lustre performance - pathetic
  - Example of lack of maturity “fragmented filesystem”

- Currently nodes stay down until next machine reboot
- Single node reboot available with 1.4
- We need to run diagnostics on one node while the system is up!

Near Term Future

- Single node reboot
- We need your help for PBSPro scheduler
- Dual core service nodes with upgrade of palu system
- Test Linux on compute node - fred

Summary

- System into production and available for development work
- System gaining maturity.
- Main current open issues in
  - Parallel file system maturity
  - Node failures
  - 1.4 stability ?????

  More Interruptions that we would like!

CSCS non-standard Areas

- Dual GigE links for boot & sdb nodes
- PBSPro TCL-based scheduler
- Yod/Mom nodes separated from Login nodes

- Use model is unique, we think.