NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

PDSF at NERSC
Site Report
HEPiX April 2010

Jay Srinivasan

(w/contributions from I. Sakrejda, C. Whitney, and B. Draney)

jay@nersc.gov

(Presented by Sandy Philpott, JLAB)

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Outline

- PDSF Structure
- Compute Node H/W
- Storage H/W
- Networking and Data Transfer
- PDSF Grid Infrastructure
- Other NERSC News

PDSF Structure

- PDSF structure has not changed.
- We still have a total of 2.5 FTEs working on PDSF
- All nodes are now 64-bit systems (AMD and Intel)
  - 1st gen Opteron, Barcelona
  - Xeon 5400, Nehalem
- New purchases
  - Compute nodes, quad-core Nehalems
  - Storage FC/SATA, Dell MD3K
  - Networking – 10GE switch Dell 8024

Compute Node H/W

- New compute nodes
  - Quad-core, dual-socket, Intel ONLY
  - Nodes have 3GB RAM per core
  - HEP-SPEC06 of ~92 (8core), ~115 (16core)
  - For now, running 8 tasks per node, investigating ways of running more (memory limitations of jobs prevent using 16 slots)
- Started using 2U-Twin nodes (Supermicro) for admin nodes.
- Evaluating Dell 6100C (also a 2U-Twin)

Storage H/W

- Continuing with a mix of F/C-SATA storage and SAS-SATA storage.
- Evaluating some “integrated” storage, i.e. server, RAID controller and disks all in one chassis for non-GPFS XRootD servers.

Networking and Data Transfer

- Evaluating the new 10GE switch from Dell (PC 8024F)
  - Uses SFP+ connectors
- Continue with Dell 6200 switches for top-of-the-rack (GigE to hosts and 10GE uplink)
- Providing a "Data Transfer Node" service to users.
  - High bandwidth connection to the outside world – 10 Gigabit connection to NERSC backbone and internal NERSC
  - High bandwidth connection to local PDSF storage
  - Network stack tuned for large-scale transfers

PDSF Grid Infrastructure

- One gatekeeper OSG 1.2.6
  - Reporting as NERSC-PDSF
- Two storage elements OSG 1.2.6 + Bestman 2.2.1.3.10.
  - SRM servers NERSC-PDSF and NERSC-DTN
  - Help with testing of Bestman for the LBNL dev. team
  - Bestman buffering on a disk local to the server (2TB cache) data transfers from the cluster to archival storage. This allows jobs to finish without waiting on the batch nodes for the transfers to complete.
- Prototype Alice VO box feasibility test conducted in December (Alien submits Condor-G jobs to OSG gatekeepers)
  - Passed scalability tests
  - Alice vo box being deployed - testing under way

Dirac GPU Cluster

- 48 nodes
  - Dual 2.4GHz Nehalem 5500 processors
  - 24GB memory
  - 1 TB drive
  - Dual PCIe x16 slots
    - ConnectX2 QDR
    - Tesla GPU soon to be Tesla 'Fermi' GPU
- Resource available from the Carver (IBM iDataPlex) cluster

Tesla GPUs

- C2050
- 448 CUDA cores
- ECC Memory
- 3 GB memory
  - 6 GB for C2070
- 1.5 GHz memory speed
- 515 Gflops peak double precision floating point

- C1060
- 240 CUDA cores
- 4 GB memory
- 800 MHz memory speed
- 78 Gflops peak double precision floating point

IPv6 testing at NERSC

- NERSC has started testing the use of IPv6
- We have brought up an IPv6 peering with ESnet and are testing and modifying our security systems.
  - We are still looking for a scientific customer use of IPv6
  - We will create a roadmap to deliver IPv6 according to need
  - First likely requirement will be a cluster that needs more than 8k public addresses.