SLAC National Accelerator Laboratory Site Report

A National Lab in Transition

Randy Melen, Deputy CIO
Computing Division, Operations Directorate
SLAC National Accelerator Laboratory

April 20, 2010

SLAC is changing …

- SLAC is changing from a single purpose High Energy Physics lab …
- To a multi-purpose lab predominantly doing Photon Science.
- While particle physics and astrophysics will continue, there are many changes.

Computing Management Changes

- Chief Information Office Donald Lemma
  - Hired from industry, started summer 2009
- Head of Infrastructure & Operations, Norm Ringgold
  - Hired from government and industry, started December 2009

Computing Management Changes

- Enterprise Architect (new), Imre Kabai
  - Hired from industry, started December 2009
- Head of Enterprise Apps (new), Yuhua Liu
  - Hired from industry, started March 2010
- Head of Scientific Computing(new), open position

Current Services & Capacity for Scientific Computing

- Two-floor computer room in Building 50 originally designed for IBM mainframes and peripherals
- Retrofitted with Computer Room Air Conditioners using chilled water from main cooling tower
- Using a basic hot aisle-cold aisle air flow rack arrangement
- Power from 2 substations: Substation 7 (behind Building 50) and Substation 8 (4'th floor); SLAC has plenty of power, the challenge is distribution

Current Services & Capacity for Scientific Computing (cont.)

- Added 2 Sun modular data centers with independent closed-loop chiller
- Contain ~500 Linux nodes with network and console support

Current Services & Capacity for Scientific Computing (cont.)

- Recently placed 192 Dell R410 2.93GHz Nehalem nodes with network and console support on SLAC's network but housed in Stanford's Forsythe Hall several miles away
- On SLAC's network, console-managed remotely, part of the full LSF cluster, very successful

Current Services & Capacity for Scientific Computing (cont.)

- 5072 cores available to the general queues
  - 4472 are RHEL5
  - 600 are RHEL4 (~13%)
  - RHEL3 is gone

- Expect to be off RHEL4 by end of summer

Current Services & Capacity for Scientific Computing (cont.)

- Batch scientific computing uses Platform's Load Sharing Facility (LSF) to manage almost all scientific batch computing as part of a "supercluster" of about 1,465 RedHat Linux nodes with about 8,200 cores or job slots (with a few Windows batch servers too)
- Typical workload is in excess of 6,500 jobs
- Most batch work is "embarrassingly parallel", i.e., independent serial jobs running for multiple hours

Current Services & Capacity for Scientific Computing (cont.)

- Seven subclusters (about 250 nodes) run as part of the LSF-managed supercluster
  - Infiniband or other low latency connections for MPI-based parallel work, one with Lustre filesystem (3 subclusters)
  - GPU hardware for rendering or many-core parallel computing (1 subcluster)
  - Some dedicated to particular groups or tasks (3 subclusters)

Current Services & Capacity for Scientific Computing (cont.)

- GPU computing example: KIPAC orange cluster
  - 4 SuperMicro servers with the current Nvidia Tesla GPUs
  - Each server has two 256-core GPUs + 2 Intel Nehalem CPUs
  - Cuda is the main API and OpenCL is available
  - Connected with IB to a ~41TB eight target Lustre 1.8.2 system

Current Services & Capacity for Scientific Computing (cont.)

- 1,200 nodes shared between BaBar, FGST, ATLAS, and other users based on an allocation with a fair-share algorithm
- Pre-emption for FGST satellite data processing pipeline every few hours
- ATLAS processing with Open Science Grid-LSF interface

Current Services & Capacity for Scientific Computing (cont.)

- Disk storage for scientific computing is mostly direct attach and using NFS, Lustre, Scalla/xrootd, and AFS filesystems
- About 3.5PB raw disk
- About 10 TB of AFS space across 28 AFS servers running AFS 1.4.11 (testing 1.4.12)

Current Services & Capacity for Scientific Computing (cont.)

- Robotic tape storage moving from 6 older Sun/STK silos to 2 new 1TB cartridge SL8500 silos
- Current tape storage is ~ 6.7PB
- New silos capable of 12PB, and using HPSS for data management

Current Services & Capacity for Scientific Computing (cont.)

- HPSS 7.1.1 running completely on RH Linux
- ~20,000 tapes (~15,000 9x40 tapes, 4,300 T10KB tapes)
- Two new silos with 16 “handbots” 8 LSMs each with 2 handbots)

Current Services & Capacity for Scientific Computing (cont.)

- 34 drives for HPSS, 4 for Windows Netbackup, 4 for UNIX TSM, 2 for AFS backup
- TSM + AFS → TiBS in the future
- 12 HPSS tape mover servers with 30 TB of disk cache

Staffing

- No operators but 4 “technical coordinators” who do physical installs, manage “break/fix” with vendor, 7x24 pager response to client “urgent” messages
- No Help Desk, no “user services” or technical writers
- Responsible for business computing & scientific computing servers and storage

Staffing (cont.)

- Lab-wide Network Administration: 5 FTE
- Windows team: 7 FTE (1 open)
- Unix team: 10 FTE
- Storage team: 5 FTE (1 Windows, 4 UNIX)
- Email team: 2 FTE
- Database Admin team: 5 FTE
- All of the above include working supervisors

Current Limits

- Reliability: Old infrastructure in need of improvements
- Space: Limited raised floor space – 15,000 sq. ft.
  - About 1000 sq. ft. available for new racks
  - About 1000 sq. ft. used, needing earthquake retrofit
  - About 1000 sq. ft. used for old tape silos, available in 2011
- Power: At 1.5MW current use, above the 60% best practices limit of 1.3MW, close to the 80% safe limit of 1.7MW
- Cooling: Approaching the cooling limits based on 8” water pipes in the building

Addressing These Limits

- Reliability
  - Investing approx. $3.43M (over 2years) in physical infrastructure upgrades for reliability & safety and some expansion for the short term (< 12 months)
  - Replace old unreliable UPS systems with a single high efficiency UPS; replace old transformers with new; modernize Substation 8
  - Diesel generator for phones, network, data center cooling, core servers and storage

Addressing These Limits (cont.)

- Power
  - Bringing in another ~1MW power to a new power room and (~270kW of that to second floor)
- Space
  - Use ~1000 sq. ft. of 1st floor computer room by retiring 6 older Sun/STK tape silos by end of CY2010
  - Technology investigation of 100 kW racks

Addressing These Limits (cont.)

- Investigating cooling possibilities
  - Run warmer computer rooms
  - Better isolation of hot and cold air
  - Add an independent external closed-loop chiller with more water-cooled cabinets

Future Expansion Possibilities

- Expand into 1st floor of existing computer building with hot aisle containment pods, efficient UPS and external closed-loop chiller
- Retrofit an existing building on site, possibly partnering with Stanford University
- Use 7500 sq. ft. of new possible “signature” building for data center annex

Questions?