LANL Computing Environment for PSAAP Partners

Robert Cunningham
rtc@lanl.gov
HPC Systems Group (HPC-3)
July 2011
LANL Resources Available To Alliance Users

• Mapache is new, has a Lobo-like allocation
  – Linux (TOSS) cluster, Moab scheduler, shared /scratchN
  – 4,736 Xeons with IB, 50.4 TF, ? procs/job

• Conejo: companion to Mapache – small ASC allocation

• Lobo is current workhorse – TLCC platform
  – Linux (TOSS), Moab scheduler, shared /scratchN
  – 4,352 compute CPUs; max 2,144 procs/job

• Small Roadrunner hybrid platform available: Cerrillos

• Big future for Turquoise (Mustang, etc.) but not big ASC
LANL Mapache Cluster

- SGI XE1300 Series with quad-core Intel Xeon 5550
- 75% allocation for ASC, primarily for PSAAP
- Architecture:
  - 4,352 x Intel Xeon@ 2.66 GHz cores on 592 compute nodes = 8 cores/node
  - Mellanox InfiniBand interconnect
  - 14.2 TeraBytes RAM = 24GB/node
  - Theoretical peak of ~50.4 teraflops
LANL Lobo Cluster

- Standard DoE Lab cookie-cutter cluster from Appro, Intl.
- 75% allocation for ASC, primarily for PSAAP
- 2 Connected Units (CUs) combined include:
  - 4,352 x AMD Opteron @ 2.2 GHz cores on compute nodes
  - Voltaire InfiniBand interconnect
  - 8.7 TeraBytes RAM
  - Theoretical peak of ~17.4 teraflops
LANL Cerrillos Cluster

- Hybrid architecture: Opteron+Cell, from IBM.
- 25% allocation for ASC, primarily for PSAAP
- 2 Connected Units (CUs) combined include:
  - 1,440 x AMD Opteron @ 1.8 GHz cores on compute nodes
  - 1,440 Cell Broadband Engines
  - Voltaire InfiniBand interconnect
  - 11.8 TeraBytes RAM
  - Theoretical peak of ~152 teraflops
LANL TLCC2 Cluster Plans

- No order placed, so this is a prediction only, for one ASC cluster in Turquoise network: “moonlight.lanl.gov”
- Hybrid architecture: Opteron+GPGPU, from Appro
- Primarily for PSAAP, but other ASC will use it
- Expected in ~December
- 308 compute nodes, ie. 2 SUs
  - Intel Sandy Bridge, dual proc node
  - Nvidia M2090
  - Qlogic InfiniBand interconnect
  - Theoretical peak of ~1.6 TF/node
LANL HPC Environment

- Obtain an account, acquire cryptocard (foreign nationals start early!)
- Access HPC platforms in Open Collaborative (Turquoise) Network via firewall/gateway: `ssh wtrw.lanl.gov`
- After connecting, `ssh` into front-end: lo-fe[1|2], mp-fe1, ce-fe[1|2]
- Moab + [Slurm | Torque] schedules nodes, batch or interactive
  - `msub` `scriptname` - or - `msub -I`
  - `checkjob, showstate, showq, mjobctl`
  - `mpirun` to run parallel jobs across nodes
  - Fairshare scheduling to deliver pre-determined allocation
- Unique security model
  - no connecting (`ssh`) between platforms
  - No kerberos tickets; keep cryptocard handy
  - `scp` or `sftp` using File Transfer Agents (FTAs), turq-fta1.lanl.gov and turq-fta2.lanl.gov
LANL HPC Environment

• Usage model
  – Front-end: text editing, job script set-up, pathname arrangements, compilations, linking, and job launching
  – Compute nodes: run applications
  – I/O nodes: out to local disk, no direct user access
  – Possible intra-network File Transfer Agent (FTA) arrangement in future

• Modulefiles to establish compilers, libraries, tools in $PATH, $LD_LIBRARY_PATH

• Compilers: Intel, PGI, Pathscale. MPI: OpenMPI, Mvapich

• Math libraries provided by compiler vendors + ATLAS
LANL Turquoise HPC Storage

- Tiny home directories, not shared between clusters (security)
- Larger NFS-based “workspace” in `/usr/projects/proj_name`
- Big parallel, globally-accessible filesystem: `/scratchN`
  - Cross-mounted to all HPC nodes in Turquoise
  - ~800 TB total space
  - Fast for parallel I/O, slower than NFS for serial transfers
  - No automated back-ups
  - Purged weekly, 30-days or older!
- Archival storage, offline, available via GPFS
- File transfers – an everpresent problem on the brink of a solution
Parallel Scalable Back Bone (PaScalBB)

- Relieve the compute nodes
- Multiple clusters sharing large, global namespace parallel I/O subcluster
  - Includes Cerrillos/Lobo/Coyote
- Network is combination of HPC Interconnect + commodity networking bridge
- Panasas is storage vendor
- I/O through a set of fileserver nodes over IB; nodes serve as interconnect<->GigE routers.
LANL Turquoise File Transfers

- File Transfers between Turquoise and the wild are \textit{S L O W}, we are addressing this now; may need your help to test.
- Throttled by gateway/firewall and security: \textasciitilde1 MB/s!
  - Packet reordering
  - Sniffing
  - Only \texttt{scp} allowed today: encryption
  - Panasas filesystem (/scratchN) – slow serial
  - All data routed through tiny \texttt{wtrw}, twice
- Solution currently in testing: double-hop through a security enclave
  - GPFS-based way station
  - Parallel transfers using \texttt{bbcp}
  - Orders of magnitude faster: 10s MB/s up to low 100s
- Turquoise holds the future for unclassified work, big changes ahead
Turquoise High Performance File Transfer Service

Clusters

WTRW

SSH Existing

LANL Authentication

Firewall

GPFS

New File Transfer

New Perf Sonar

LANL Yellow

Collaborators via the Internet

GPFS

Turquoise HPC

Clusters

WTRW

SSH Existing

LANL Authentication

Firewall

GPFS

New File Transfer

New Perf Sonar

LANL Yellow

Collaborators via the Internet
LANL Turquoise Tools

- TAU (Tuning and Analysis Utils) -- profiling and tracing toolkit
  [http://www.cs.uoregon.edu/research/tau/home.php](http://www.cs.uoregon.edu/research/tau/home.php)

- STAT (Stack Trace Analysis Tool) from LLNL
  [https://computing.llnl.gov/code/STAT/](https://computing.llnl.gov/code/STAT/)

- Boost – C++ Utility Libraries

  - sampling experiments
  - callstack analysis
  - hardware performance counters
  - MPI profiling and tracing
  - I/O profiling and tracing
  - floating-point exception analysis
LANL Turquoise Tools

- Javelina -- code coverage tool that uses dynamic instrumentation
  http://javelina-cc.sourceforge.net/
- Valgrind -- instrumentation framework for dynamic analysis,
  http://valgrind.org/
  - memory errors
  - cache and branch prediction profiler
  - thread error detection
  - heap profiling
- memP – parallel heap profiling library
  http://memp.sourceforge.net/
- mpiP – Lightweight, Scalable MPI profiling
  http://mpip.sourceforge.net/
LANL Debugging

• gdb (Gnu debugger) – comes with distro
  http://www.gnu.org/software/gdb/

• Totalview – interactive medium-scale parallel debugger
  http://www.roguewave.com/company/overview/totalview-tech-acquisition.aspx
  – parallel independent process views
  – ThreadSpotter
  – MemoryScape
  – ReplayEngine
  – GUI or command-line (TCL)
3 Useful LANL Web Sites for Users

- User Docs
  http://bear.lanl.gov/drupal/?q=tlcc_home
  (or) https://bear.lanl.gov/drupal
  (or call 505-665-4444 option 3, consult@lanl.gov)

- Calendar
  http://ccp.lanl.gov

- HPC Training
  http://int.lanl.gov/projects/asci/training
HPC Accounts

Don’t forget Photo Op!
Questions?