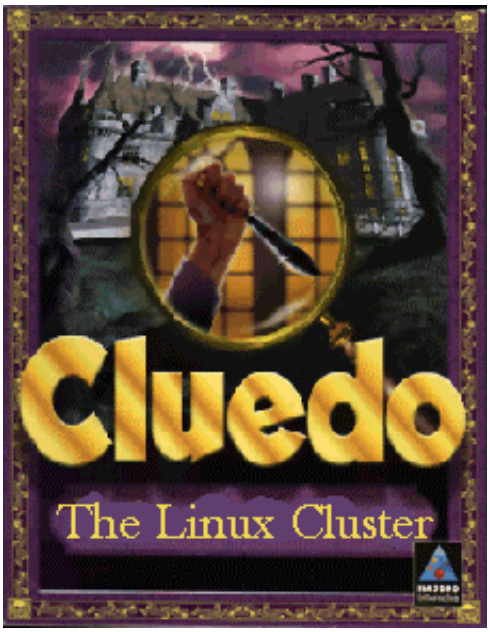
Linux Clusters and D0 Software: The Clued0 Experience



Dugan O'Neil MSU
D0RACE - December 6, 2001

Outline

- What/Why is CLuED0?
- Earthquakes, Floods and Redhat 7.1
- D0 Software on CLuED0
- PBS Batch on CLuED0
- SAM/CLuED0 Theory
- Conclusions

What is ClueD0

- Linux cluster at D0 providing centralized services and management for desktop machines.
- `http://www-clued0.fnal.gov/clued0`

  includes user info, admin howtos, etc.
- 145 machines, 45 institutes, 375(ish) users
- Operating system upgraded to RedHat 7.1 in May 2001 (testing since Beta releases).
- Institutes own/buy their machines. Diverse hardware, capabilities.
- Use LDAP + CLuMP (cluster management software written by Roger Moore) to control system configuration.
- Use PBS + MAUI scehduler for batch
- Current D0 software releases supported by release managers
- Software firewall on all nodes but one "gateway"

Earthquakes, Floods and RedHat 7.1

- RedHat 7.1 was not as stable as we would have liked. Mainly problems in 2.4 linux kernel series.
- Since we made it to 2.4.5XFS our problems have mainly been solved. Some issues for high impact applications (eg. a full D0 build, heavily loaded server, etc.).
- Basic advice: don't bother running with stock RedHat 7.1 kernel especially if you need to NFS mount disks. Upgrade immediately to (at least) 2.4.10.
- For large disks you want a journalling filesystem (prevent long reboots). We started using ReiserFS....disaster. XFS (from SGI) is much better. Haven't tried ext3 yet but we'd like to soon.

D0 Software on CLuED0

- This is really not an "offsite" situation
- The release managers do the RedHat 7.1 builds and put them on a central machine (d02ka). We simply mount the disk from this machine.
- Offsite you will need to download frozen releases to a central machine and mount disks from clients. This is more work.
- We do have UPS/UPD installed. Fermilab has quite a few docs describing the setup, etc. Roger Moore packaged them as rpms (easier to install) available from

  `http://www-clued0.fnal.gov/~oneil/rpms`

- MSU used to have this setup and maintained at MSU.....but then Dylan Casey left. It is still there, but am not sure if it is maintained.

Security Stuff

- Kerberos clients come with RedHat 7.1 by default. We used standard kerberos, not fermi-kerberos. This works fine.
- We run an "iptables" software firewall on all machines (save one) which blocks all ports from incoming connections. Not required, but nice.
- To connect to fermilab in principal you cannot allow connections except by ssh and kerberos means.

PBS

- `http://www.openpbs.org/`
- Free, easier to configure/maintain than NQS.
- Supported on many platforms
- MAUI scheduler used by supercomputing centers around the world. Very flexible fair-share algorithms.
- Batch time allocated by contribution to the cluster. Easy to share cluster resources among diverse institute makeup. At remote sites might allow easy sharing among groups, departments, etc.

SAM/CLuED0 Theory

- We have not had much luck with SAM so far, but the future is brighter
- One server machine with access to d0mino cache + enstore tape. Client's transfers are all made from our server's (large) cache.
- So far we only have the server and 5 desktops in a test configuration. Heidi's success on the farms last week gives us reason to hope for more.

Conclusions

- ClueD0 is a rapidly growing multi-institute cluster of linux machines
- Not all experience we've gained is helpful offsite.
- Recommendations:
  - Don't buy expensive quad or 8 processor machines. Not worth the money/performance, not as well tested/supported by linux.
  - RedHat 7.1 is fine with a recent kernel ($\geq$2.4.10)
  - Mixing NFS with other OS's is not as nice as pure Linux
  - PBS is a nice, solid batch system
  - LDAP/CLuMP is a nice way to manage diverse machines....a little cutting edge
  - It doesn't really matter if you use fermi kerberos or standard kerberos, you decide.

Conclusions

- How can we help?
  - If you send questions to clued0-admin@fnal.gov we will answer
  - If you send us people to be admins for 6 months we will send them back well trained, well fed and happy.
  - If you set up a clued0-like cluster somewhere we ask that you document the whole thing....we haven't.....