

L1 Trigger Xserve Farm Test Stand Installation

Jim Kowalkowski

1 Introduction

1.1 Purpose and Goal

The purpose of this report is to record the steps it took to set up the Apple Xserve cluster at Fermilab. The report will also include problems encountered during setup and how they were resolved.

This cluster will be used as the development platform and test stand for the level-1 trigger software infrastructure and algorithm. We will use it to do throughput tests and also to evaluate potential operating systems. The first candidate will be Mac OS-X.

1.2 Useful Resources

- <http://ctserver.uchicago.edu/Cerberus-setup/xserve-cluster-config.pdf>
- http://mccluster.epfl.ch/Building_McCluster.pdf
- <http://developer.apple.com/server/>
- <http://www.apple.com/support/>
- <http://www.afp548.com/articles/system/headlessg5.html/>
- <http://docs.info.apple.com/article.html?artnum=107912>
 - Getting Started
 - Command-Line Administration
 - Network Services
 - Open Directory
 - File Services
 - User Management

2 Materials

All the hardware is installed in FCC2. This includes

- 15 Xserve cluster nodes (11g5xx where xx in [1,15])
- 1 head node (bfoot)
- 1 24 port Infiniband switch
- 16 Infiniband HCAs
- 16 Infiniband cables
- 42U rack
- 3 APC 20amp power strips with twist lock plugs

The software we ordered with the system includes:

- 15 copies of OS-X server 10 client
- 1 copy of OS-X unlimited client

- 2 copies of Apple Remote Desktop (ARD)
- Infiniband drivers for OS-X from Voltaire

The systems came with the OS preinstalled, without the licensed validated with a key. The validation is need before the nodes can be used. The key comes in the OS-X software box.

3 Terminology and Tools

Here is a list of technologies and tools that Apple makes use of that you might want to know about.

- Apple Remote Desktop (ARD): A tool for executing commands, doing software update, and doing software installations across all cluster nodes at once. This is an important tool. It also allows use of VNC to connect into a system from a remote location.
- NetInfo database: This database is used to house all the information that is normally kept in /etc files such as hosts, passwd, and fstab. Editing the etc files has no effect; the values must be edited in this database with command line tools or with the NetInfo Manager editing tool.
- DHCP: Gives out IP addresses to cluster nodes. This facility also supplied the location of the global resources such as Open Directory.
- DNS: This is used to resolve names in the private subset. We are looking for another solution so this tool does not need to be present.
- StartupItems: Most Unix system put information about things to start at boot time in /etc/rc.d. This OS has a directory /System/Library/StartupItems that contains a set of scripts that run when the system starts.
- WatchDog: This facility serves the same purpose of init (inittab) in a standard Unix system.
- Open Directory: An LDAP server that hold information common to all nodes, such as user accounts, mount points, network names, host names, and boot information. This is a newer version of the NIS or yellow pages service.

GUI-based tools

- Server Admin: start, stop, and configure system services here
- Work Group Admin: create users and export file systems here. Use the Inspector from here.
- NetInfo manager: Edit the system configuration information with this tool, such as hosts, groups, fstab, etc.
- Inspector: A tool for editing similar information as NetInfo except that it works also with Open Directory. You must enable this in the preferences of the Work Group Administration tool.
- Directory Access: A tool for specifying how the system will look for files and configurations.
- Server Assistant: Tool used to configure remote nodes without being logged into an account
- System Preferences - Startup Disk: Set the disk that will be for booting.

Here is a list of useful commands that help with administration.

- `sa_srchr`: check for nodes that need to be configured on this subnet.
- `hdiutil/installer`: `hdiutil` mounts a `dmg` file as though it were a disk. `Dmg` files will usually include a `pkg`. The installer command installs the software. Run it like this (usually):
 - `hdiutil attach xx.dmg`
 - `installer -pkg Desktop/xx.pkg -target /`
- `kickstart`: This tool is located in `/System/Library/CoreServices/RemoteManagement/ARDAgent.app/Contents/Resources/`. It is used for starting, stopping, and configuring ARD from the command line.
- `niutil/nidump/niload`: Use these tools to edit and manipulate the NetInfo database from the command line.
- `ldapsearch/ldapadd/lpaddelete/ldapmodify/ldif`: Use these commands to manipulate the Open Directory database from the command line.
- `Serversetup/serveradmin`: These command line tools are used to administer the machine. They may not be in the path. Find them in `/System/Library/ServerSetup`.

Files that are of interest to you include the following:

- `sesolv.conf`: contains nameserver information
- `sshd_config`: rules about logging into the machine via `ssh`
- `hostconfig`: This file contains the hostname (as opposed to the computer name), and a YES or NO for each of the services that are supposed to be running or not running on the system at boot time.
- `/var/db`: The NetInfo database information is kept here. Some system state files are located here also.
- `watchdog.conf`: the `inittab` equivalent.
- `hosts.equiv`: Used by `rsh` to determine what users are allowed in without password.
- `hosts.allow`: rules about machine networks or services are allowed into the machine.

4 Procedure

In the BTeV document database is a file called `setup.tar.gz`. This file has a README and a set of scripts and configuration files that were used to set up the cluster. Many of the things mentioned below are already coded in these scripts and just need to be run at the right time.

I'm assuming that the head node is on both the public and a private LAN. The head node gets set up first, since it is the main point of administration. Fill out the CD network node forms to get an IP address and also to get Kerberos host/ftp principles. Do not install them yet. Plug the machine into the public and private networks. In this setup, the *public network is in en0 and the private is in en1*. We chose *btev.net* as the private network name.

4.1 Inventory File

First create a file with the following columns: node_name, serial_number, ethernet_address, ip_address, slot_in_rack, license_key, status. Assign the node_names. We are in a private LAN, so set the ip_addresses now also. Read the serial numbers and ethernet addresses from the back of the machines. Fill in the slot number as you install the servers. At this time, assign the license keys to servers. It is very important to get this table correct, especially if something goes wrong. The file used for this setup is in the BTeV document database and it is called "serialnumbers". We will verify most of the numbers next.

4.2 Head Node Setup – bfoot

Turn the node on and answer all the simple questions it asks. When you get to the network setup screens, activate only en0 and en1.

Here are the values for en0:

| | |
|--------------|-----------------|
| Address mode | Manual |
| IP address | 131.225.198.5 |
| Netmask | 255.255.255.0 |
| DNS server | 192.168.200.1 |
| Router | 131.225.198.200 |
| Domain | Btev.net |

Here are the values for en1:

| | |
|--------------|---------------|
| Address mode | Manual |
| IP address | 192.168.200.1 |
| Netmask | 255.255.255.0 |
| DNS server | 192.168.200.1 |
| Router | Btev.net |

Make sure to select "Standalone Server" when asked about server type.

Copy the krb5.conf and sshd_config files from the setup.tar.gz file to the /etc directory on the head node. Install the host keys using

```
kadmin -r FNAL.GOV -p host/bfoot.fnal.gov@FNAL.GOV -w <key_from_email> -q  
"ktadd host/bfoot.fnal.gov@FNAL.GOV"
```

You should now be able to access the machine via Kerberos from a remote machine.

4.3 NetInfo Modifications

DHCP will first give out fixed IP addresses for ethernet addresses that it recognizes. The standard place for this information is the NetInfo machines directory. Included in the document database with this document are files called machines_others and networks. These files are in the NetInfo database raw text

file format. The DHCP section of the Network Services Administration Guide explains the options allowed in each of the sections. The machines_other file and networks file was created by running

- nidump -r /machines / > machines_other
- nidump -r /networks / > networks
- vi machines_other
- vi networks

Add one section for each machine; include its IP address, ethernet address, and node name. Add a new network called btev.net. When done, to the following:

- niload -r /machines / < machines_other
- niload -r /networks / < networks

4.4 Bfoot DHCP setup

On bfoot, go to the Finder->Applications->Server->ServerAdmin screen (it is also on the bottom tool bar). Go to DHCP. Activate DHCP for en1, and then double click on it for configuration.

| | |
|---------------|----------------------------------|
| Address range | 192.168.200.20 to 192.168.200.50 |
| Netmask | 255.255.255.0 |
| Router | |
| DNS server | 192.168.200.1 |
| Domain | btev.net |

Do not set up the LDAP options at this time. Start the DHCP service now.

4.5 DNS setup

Near the DHCP setup screen is the DNS screen - go there. Setting this up is covered in the document from the University of Chicago (also in the document database) and will not be repeated here. Be sure to use our head node, network name, and client names instead of the ones in the document. Remember to use the replicate button near the + and - buttons for adding records and zones (the overlapping squares). Add a btev.net zone with bfoot.btev.net as the SOA. Add PTR records to the reverse mapping zone 200.168.192.in-addr-darpa for each of the nodes. Add one address (A) record for each of the nodes. Add a forward zone with addresses 131.225.8.120 and 131.225.17.150. Start the service and check the log to make sure that everything is working. Reboot bfoot to make sure everything is configured and starts properly.

Test that the DNS is working properly by pinging a known server in the fnal.gov domain. Be sure to include the entire domain name e.g. bone.fnal.gov.

This facility will need to be shut down as soon as possible - as soon as an alternate method is discovered.

4.6 Verify node addressing

The cluster nodes can now be turned on. Watch the DHCP log for the incoming requests, you will see the ethernet addresses and then see the DHCP supplied IP address from the NetInfo machines directory – the numbers must match what is in your inventory file. If any do not match, correct all the tables and reboot the node in question. I turned one machine on at a time to make sure that the slots matched the names and the ethernet addresses.

4.7 Verify Default Login

Test the default login on each of the cluster nodes. Use ssh to login to root on each node. If the DNS is set up properly, you should be able to use their node names. The password is the first 8 characters of the serial number. Correct any errors in serial number recording (there are bound to be a few) in the inventory file.

4.8 Use Server Assistant

Start up the server assistant. Answer the simple questions. When you get to the server list page, enter all the cluster nodes and their passwords (first 8 characters of the serial numbers). When you get to the serial number page, enter all of them. The scroll bar does not appear and this will cause you grief. Double click on the first serial number and *then* use the tab key to move to the entry that is off the screen. I used the cut and paste feature to get them right (from the inventory file). The system will tell you if any of the keys are invalid. It will take a while for all the cluster nodes to be configured.

The assistant will prompt you for an administration account name, we use “btev” as the short login and “BTeV Admin” for the longer name.

Make sure to check the “apple remote desktop” service box so this feature starts when the nodes reboot.

The server assistant named all the machines improperly. This will need to be repaired later on.

4.9 Apple Remote Desktop

Install the Apple Remote Desktop (ARD) package and get the software update from Apple. Start ARD and restrict the IP search range to the cluster nodes. Select all of them and drag them to the master list. You will be asked to supply an admin login – use the btev account for this.

4.10 Node Configuration

Look at the README in the setup.tar.gz file to understand a bit more about what is going on. Complete the setup of the head node by installing:

- OS-X update (from Apple – use softwareupdate)
- X11 from Apple’s web site
- Fink
- Xemacs from fink

- Gsl from fink
- Cmake from fink

Almost everything in this list is under the “Manage” menu on the top bar. Do the following:

- Upgrade ARD on the clients to 2.x.
- Rename the computer to the correct names
- Copy the files in the setup.tar.gz bundle over to each copy (one step process)
- Run the script to load machines, mounts, and network tables in NetInfo

Download the OS-X server upgrade from Apple and use ARD to install it on all the cluster nodes at once.

Finally, install the Infiniband package on all the cluster nodes using ARD. At the end, install this package on the head node.

4.11 Test

Go to a couple nodes using the barm and btev accounts and make sure that you can:

- Log into other nodes by name using rsh
- Go to directory /home
- Go to directory /local
- Go to directory /sw

5 Problems Encountered

5.1 Auto Server Setup

I believe the proper way to set up machines is to use the auto server setup feature. All you need to do is to create a directory in /Volumes/*/ called “Auto Server Setup” and place a properly named plist file in it with all the sections filled in. See the plist files in setup.tar.gz for examples. I attempted to use “Auto Server Setup” and failed. The machine never reappeared on the network. I had to reload the operating system on the system that I tried this out on.

For recovery, I used the technique of loading the drive into the head node and then installing on it from there. After setup, the machine will be assigned that new drive for booting. Hold down the “C” key as the machine is rebooting and a menu will appear asking what drive you want to boot from - select the original drive. When the machine boots, go to the preferences screen and select the startup disk (near the bottom). Pick the original disk to make the change permanent.

5.2 Server Assistant Crash

The Server Assistant application continually crashed and the remote nodes could not be configured. The OS had to be reloaded to resolve the problem.

5.3 Server Assistant Peculiarities

The Server Assistant did not create a scroll bar for a large number of servers when entering the serial numbers.

The Server Assistant set all the computer names in the batch to the same name. I had to write a script to repair the names (contained in setup.tar.gz).

5.4 Switch Misconfiguration

None of the machines saw the DHCP server when we first started them up. It turned out that the ports were not properly configured on the switch and the DHCP server was working properly.

6 What Should Still Be Done

Conversion over to Open Directory is not complete. After this is done, the accounts can be centrally maintained.

The DNS need to be shut down. This may involve configuring it as a caching server and making adjustments to /etc/resolv.conf and the named database area.

7 Other Useful Things

Go to the web and find OSXvnc. Install the package somewhere on bfoot. Using this tool, you can use UltraVNC from a Windows machine to use bfoot remotely. The command line tools for OSXvnc are located inside the package. Use them as follows:

- `./storepasswd anything_you_want ~/mypassword`
- `./OSXvnc-server -rfbport 20001 -rfbauth ~/mypassword`

Go to the windows machine and start UltraVNC. For the session, enter "bfoot.fnal.gov:20001". Enter your password at the prompt and you should be up and running.

8 Infiniband Setup

Install the Voltaire Mac OS-X Infiniband package. You must get this from their web site after getting a customer login from them. Install it on all the cluster nodes and the head node. Reboot all nodes after installation.

The switch setup was easy – I just followed the user's guide. The included console serial cable is a null modem cable, so hook it directly to a laptop serial port. Activate the ethernet port according to the instructions. Afterwards, you can go to the web site hosted on the switch and look around. Check that the client links are up and running properly.

For the remaining tests, always run from the barm account. Of course you will need to be root to modify files in /etc.

8.1 Test an Infiniband link

Go to machine l1g502. Run the utility "vping -d". This is the ping daemon mode and the Infiniband lds will print out for you to use. Logon to l1g501 in another

window and type “vping 0x... 0x...” where ... is the DLIP and QP parameters printed on the console of l1g502. You should see a successful completion status.

8.2 Check connectivity

The default configuration from installation of the Voltaire package modifies /etc/profile so that PATH points to the right spots. From l1g501, run “rsh l1g502 ‘echo \$PATH’” to see if the rsh command is working properly and that the path is set up properly to point at /usr/mellanox/bin, /usr/voltaire/bin, and /usr/voltaire/mpi/bin.

If rsh does not work, modify /etc/xinet.d/shell, setting “disable=no” on all the cluster nodes. On the head node, set “disable=no” and add “bind=192.168.200.1” so that rsh is active only on the private network. Edit /etc/host.equiv and add one line of “+ barm”. Restart the xinetd daemon by sending a hup signal i.e. “kill – HUP xxx” where xxx is the inetd process ID. Check that the shell service is connected properly to only the btev.net network using the netstat utility.

If the PATH printed is not correct, edit ~barm/.bashrc and add the paths listed above to the PATH variable.

Check that l1g501 can connect to bfoot using rsh by running “rsh bfoot ls”. Check that bfoot can contact l1g501 in the same way.

8.3 Voltaire MPI

Copy the examples from /usr/Voltaire/mpi/examples to ~barm/mpiexamples. Go to the copy directory and type “make mpi_bandwidth”. Login to l1g501 and go to the mpiexamples directory. Run “mpirun_rsh –np 2 l1g501 l1g502 ./mpi_bandwidth 10 10000”. This command should run successfully.

There is a problem with the current configuration of bfoot (the head node) that prevents mpirun_rsh from running properly. This program uses the name bfoot.fnal.gov instead of just bfoot to configure the MPI job. The MPI system wants to send data using rsh and the domain fnal.gov is not valid for rsh. I am still working on a fix for this problem.

The environment variable for MPI called VIADEV_RENDEZVOUS_THRESHOLD is set to 1800 by default. This is very low and we need it set to something like 11000. Performance suffers with the value this low. You cannot change this value without breaking the system – no MPI programs work properly at values greater than 1800.

8.4 NetPIPE

Get the NetPIPE benchmark program from the web. Untar it, go to its directory and type “make ib”, “make tcp”, and “make mpi”. This will create test executables named NPib, NPtcp, and NPmpi. You need two windows open to two different machines to run the ib and tcp programs. Start the ib test by running “NPib” on one machine (l1g501) and running “NPib –h l1g501” on the other (l1g502). After the test completes, copy the np.out file to a meaningful name. Start the tcp test in

the same manner and copy the np.out file again. Run the NPmpi test from I1g501 using “mpirun -np 2 I1g501 I1g502 NPmpi”. Save the result again.

8.5 MVAICH

You can download MVAICH from Ohio State University. Edit the mvapich.make.macosx script and correct the values in there. Change the INSTALL_PATH and make sure the configure line contains the following options:

```
./configure --without-f77 --without-f90 --with-device=vapi --with-arch=macosx \  
-prefix=$INSTALL_PATH --without-romio
```

This implementation will probably be much faster than the Voltaire supplied one, even though the Voltaire one is based on an earlier version of MVAICH. Run this using “mpirun_rsh -rsh -np 2 I1g501 I1g502 exec_name”. Make sure the paths in /etc/profile and ~barm/bashrc are pointing to the mvapich directory and not the Voltaire one.

8.6 LAM MPI

The distribution of LAM MPI contains an Infiniband device. You can download, compile, and try it out. I got it to work just fine using the tcp device, but not with the ib device.

9 Conclusion

The ARD tool is very useful and should be started and used as soon as possible during the setup. The ability to restart servers, run a command on all the nodes, and copy files to all nodes are very useful features

There is a bunch of legacy stuff and quirks that made things somewhat difficult. I also caused several problems during installation that caused me much grief. Following the steps in this document will reduce the number of errors and shorten the recovery period.

The cluster is set up in a minimal configuration. Ideally, Open Directory would be running and the cluster nodes would locate this facility through the DHCP records (it is made to do this). The cluster nodes would use Open Directory to get user account data and mounted file system information so that these things would be centrally located on the head node.

Right now each cluster node has a barm and btev account (locally administered) and a local hosts database. The head node advertises /usr/local (as /local on cluster nodes), /sw (for fink distributed products), and /home (for shared user accounts and data transfers. Each cluster node has a local /data directory for storing temporary result calculations. A DHCP server resides on bfoot and gives out fixed addresses to cluster nodes based on their MAC addresses when they boot. A DNS is running on the head note temporarily until we can figure out how to allow the head node to contact the outside world and resolve cluster names at the same time (the DNS solution solves this problem).

The Infiniband software includes MPI for running programs on the cluster. Further instructions will be created for using this facility. A batch system may also be added to this cluster in the near future.

Draft