Scaling Into Tomorrow

Nicholas P. Cardo
Lawrence Berkeley National Laboratory
National Energy Research Scientific Computing Center
cardo@nersc.gov

ABSTRACT: *NERSC's recent addition of a 102 Cabinet Cray XT4™ system greatly enhances the computational capability of the center. An overview of the systems balanced configuration will be described along with early performance measurements and system metrics. In addition, the challenges of facility preparation will be discussed. This paper will take into account all from preparation into production.*

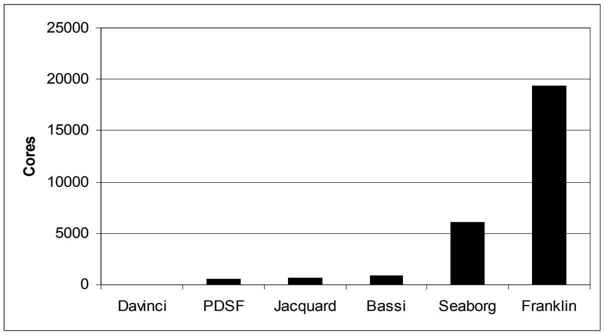
KEYWORDS: XT4, Install, Configuration, Scalability

1 Introduction

NERSC continues to provide world class computational resources serving numerous scientific disciplines. There are currently five main production systems at NERSC:

System	Compute Cores	Type
PDSF	550	mixed
Davinci	32	IA-64
Seaborg	6,080	Power 3
Jacquard	712	Opteron
Bassi	888	Power 5

NERSC's newest system, Franklin, will take scalability to new levels by providing 19,376 computational cores.



2 The NERSC 5 System(s)

There are many components governing the purchase of a large scale system. While the obvious is the computer itself, other components include storage, networking and a dedicated test system.

2.1 Meet Silence

The NERSC 5 system includes a dedicated test system that is completely isolated from the main production system. This allows for specialized testing to occur without the fear of interruptions to the main production system. The system consists of a single Cray XT4™ cabinet with two full chassis. The configuration is designed to mimic the configuration the main production system, although not at scale.

The system is named for Silence Dogood, the widow alias used by young Benjamin Franklin to submit letters to the Boston Current.

The system consists of 5 IO Modules and 11 Compute Processor Modules. This yields 10 Service Node and 44 Compute Nodes.

There are 2 login nodes that are part of a Domain Name Service (DNS) round-robin. Each node has a dual port GigaBit Ethernet adapter providing external connectivity.

There are 2 dedicated network nodes, each configured with a single port 10 GigaBit Etherenet Adapater. These two nodes provide connectivity into NERSC's internal high speed network infrastructure as well as external high speed networking.

The remaining 6 service nodes provide the Lustre MetaData Server (MDS) and Object Storage Server (OSS) functionalities.

2.2 Meet Franklin

The main production NERSC 5 system takes its name from Benjamin Franklin, who is considered by many to be *America's First Scientist*. The computer system will perform computational analysis in many scientific disciplines, just as Benjamin Franklin, the man. The name is truly honouring of the man as well as the machine.

2.3 XT4 Cabinets

Franklin consists of 102 Cray XT4™ cabinets connected via a 3-Dimensional Torus high speed switch. The system consists of a total of 9,740 dual-core Opteron nodes for both computational work as well as system services.

2.4 Storage

System storage is provided by five DataDirect Networks (DDN) S2A9550 High Performance RAID Storage Systems couplets for a scratch filesystem. Each couple has 32 Tiers and utilizes 300GB FC Drives.

An additional DDN S2A9550 couplet with 16 Tiers of 300GB FC drives is used to provide a home filesystem.

2.5 Fun Facts

Sometimes the awe factor of a system can be observed by simply looking at the components that make up the system. While a 102 cabinet system may not sound large, the details exemplify the systems size.

- 9,740 Total Nodes
- 38,960 Memory DIMMs
- 39138 GBs of Memory
- 3,366 Interconnect Cables
- 9,133.8 Meters of Interconnect Cables
  29,968 Feet (5.7 Miles)
- 96 Wires per Interconnect Cable
- 2,876,927.8 feet of Interconnect Wire
  544.9 Miles
- 1,628 300GB Disk Drives

3 Facility Preparation

Having purchased a new system, the dubious task of installing it was at hand. There are several unique requirements for installing a Cray XT4™ system at NERSC.

3.1 Seismic Isolation

All computer systems installed at NERSC have conformed to seismic standards. The Cray XT4™ is no exception. However, new advances in technologies have shown the use of seismic isolation platforms to be the preferred choice.

Picture a set of large ball bearings sandwiched between to steel platforms. This solution allows the computer to remain relatively still while the ground beneath it shifts. Previous methods used rigid bracing to tie the cabinets to the concrete subfloor. The problem with the previous solution is that the cabinets could get shaken to pieces resulting in significant damage.

The entire Cray XT4™ sits atop an ISO-Base™ by Worksafe Technologies[1]. The ISO-Base™ solution provides for 10 inches of free floating in all directions. This seismic isolation platform allows the entire Cray XT4™ to float freely above the shifting ground below. The system will remain relatively still because it is isolated from the shaking ground below.

The same applies to the storage devices. All storage cabinets also sit atop an ISO-Base™ in order to seismically isolate them. Special consideration was needed for the cabling of the storage units. While the Cray XT4™ has overhead troughs for routing the SeaStar cables, the DDN cabinets utilize underfloor cabling. The hole through the floor tile needed to be located in a way that would not cause cables to be sheered during platform movement. Furthermore, sufficient cable slack was needed to allow for 10 inches of free float for the platform.

3.2 Floor Space

The Cray XT4™ system is configured with a 3-Dimensional Torus for its high speed interconnect. This is accomplished by a unique cabling system that provides six connections in 3 axes. The X dimension is cabled within a row. The Y dimension spans rows. While the Z dimension spans chassis within a cabinet.

In order to accomplish this unique cabling system, cabinets must be laid out in a precise pattern. All cabinets within a row must be set physically next to each other. All rows must be a specific distance apart. This leaves no room for such devices as air handlers or power distribution units.

The 102 cabinets were to be installed in 6 rows with 17 cabinets per row. However, a structural support column happened to align itself in such a way that it would be impossible to open the back of one cabinet. To further complicate things, the seismic isolation platform needed to have sufficient space available in order freely float. The solution was to insert a blank cabinet that would just have cables running through it. While this solved the problem with cabinet access it did not solve the need to freely float. Cray engineers solved this by manufacturing a short cabinet. The ISO-Base™ under the cabinet was also shortened.

3.3 Power

The electrical requirements for the system presented additional challenges. The Cray XT4™ individual cabinet has a power requirement of 15 to 22.5 KW yielding a system requirement of 1530 to 2295 KW[2].

To put this into perspective, according to data from the Energy Information Administration[3], a single Cray XT4™ cabinet would equal the power consumption of approximately 16 United States homes.

To accommodate this demand, two new transformers were installed to deliver approximately 1.1MW of electricity each. While this sounds simple to accomplish, the power feed from the electric company required accommodating. Power cables were fed from the street, through the basement and into the parking lot. At this point, a trenched duct bank was installed to complete the cable runs through the parking lot to the transformers.

This major electrical installation presented the opportunity to install a small Uninterruptible Power Supply (UPS). The intent is to redundantly connect all storage devices and critical systems to the UPS, providing a safety margin against power failures.

As part of the system hardware checkout, Cray engineers ran CPU and memory tests which stress the components. Any component that did not survive the trip to California could be identified and replaced. This also proved to be a useful test of the new electrical power system. One of the transformers tripped a 1 MW breaker, powering off half the system. The root cause turned out to be an incorrect setting on a programmable breaker.

3.4 Cooling

The Cray XT4™ is uniquely designed to pull air directly from under the raised floor and exhausting out the top of the cabinet. A large blower in the base of the cabinet pulls cold air from under the cabinet and forces it through the components within the cabinet at nearly 3000 cubic feet per minute[4].

The challenge is to deliver sufficient cold airflow without starving the other equipment in the computer room.

The first part of the overall solution was to clear out the underfloor space by removing or relocating all cables and piping. The biggest challenge was the relocation of the chilled water pipes for the required air conditioning.

The second part of the solution was to install twenty two 40 ton air handlers encompassing three sides of the Cray XT4™ perimeter. The underfloor airflow created such pressure that a tile was blown out. Actually, a ceiling tile was blown out of position. A floor tile was pulled for some underfloor work which resulted in the ceiling tile, 14 feet above, being blown out of position.

The DDN 9550s also required specific cooling. It was found that the controllers were consistently running hot resulting in task exception failures. To mitigate this, the cabinets were analysed and plan set forth to drop the cabinet temperature. It was found that the heat generated from the drives was not being exhausted quickly enough resulting in a noticeable temperature increase in the upper part of the cabinet. This imposed extra externally generated heat on the controllers that sit on top of the drives.

In order to drop the temperature in the cabinet, a series of changes were made. Air flow in front of the cabinets was increased by alternating standard 25% air flow perforated tiles with high flow 65% air flow grated tiles. Temperature readings in front of the cabinet were now sufficiently cold enough to supply the cabinet, eliminating this as a possibility. Air flow was increased into the cabinet by removing the air flow pillow in the raised floor cable hole. To increase the exhaust rate, two plastic inserts were removed from the tops of the cabinets.

The results of these changes improved air delivery by supplying more air. An increase in hot air exhaust from the cabinet was achieved by providing two new exhaust ports in the top of the cabinet as well as increasing cold air flow into the cabinet from the underfloor. These simple changes resulted in a 10° C decrease in internal temperatures.

4 System Configuration

There are several major components of a Cray XT4™ system from hardware specifications to software configurations. Each plays an important role in providing an integrated balanced system.

4.1 3-D Torus

The number of total cabinets and cabinets per row determines the SeaStar2 torus cabling configuration. Franklin consists of 6 rows of 17 cabinets each for a total of 102 cabinets. This results in a Class 3 torus topology identified by the following equation:

C_r x (4 * R) x 24

Where:

C_r = Cabinets per Row

R = Number of Rows

For Franklin, this results in a 3 Dimensional torus configuration of:

17 x (4 * 6) x 24 or

17 x 24 x 24

The X dimension of the Torus connects cabinets within a row. The Y dimension connects cabinets between rows. The Z dimension connects the cages within a cabinet.

4.2 Compute Partition

There are two types of nodes in the system. The first type is a compute node which makes up the compute partition.

There are four compute nodes per compute module. There are eight modules per chassis and three chassis per cabinet. This gives a total of 96 compute nodes per cabinet.

Each compute node has one 2.6GHz AMD dual core Opteron processor and four 1GB memory DIMMS for a total of 4 GBs per node of memory.

The system has 9,688 compute nodes for a total of 19,376 computational cores. Total aggregate computational memory comes to 38,752 GBs or about 37.8 TBs.

UNICOS/lc represents the complete software package for the system. The software component running on the compute nodes is the Cray Catamount Microkernel.

4.3 Service Partition

The second type of node is a service node. The collection of all service nodes make up the system's service partition.

Service nodes are limited to two nodes per module. Two nodes are replaced with I/O bays for interface cards. Given two nodes per module, 8 modules per chassis and 3 chassis's per cabinet, a total of 48 service nodes are possible per cabinet.

Service nodes have personalities associated with them that identify their purpose in the system. Each Service Node runs SUSE Linux™, included in UNICOS/lc. There are four distinct node personalities for providing login, I/O, network and system services.

Login nodes on Franklin are each configured with 8 GBs of memory and one dual-port Gigabit Ethernet adapter. Franklin has 16 login nodes.

Network nodes on Franklin are each configured with 8 GBs of memory and one 10 Gigabit Ethernet adapter. Franklin has 4 network nodes providing high speed networking to internal and external networks.

There are 4 *System* nodes on Franklin. System nodes are each configured with 8 GBs of memory and provide the necessary services to boot and run the system. These include a boot and system database node (SDB).

Franklin has 28 *I/O* nodes for serving up the two Lustre filesystems. Each node is configured with 8 GBs of memory and two 4 Gb Fiber Channel Adapters.

4.4 Redundant Configuration

Franklin has been configured with characteristics of a highly available system. This is accomplished by the elimination of single points of failure for survivability as well as recoverability. It is just as important to be capable of quickly returning the system to production service as it is to continue to run with component failures.

As previously mentioned Franklin has four *System* nodes but only two system services were identified. Both the boot and SDB nodes have alternate nodes identified that be used to replace the primary nodes.

There are four *Network* nodes which provide two connections to an internal 10Gb network and two connections to external 10Gb networks. This provides redundant paths out of the system which are controlled by simple routing rules.

The service partition is actually split into four cabinets with two cabinets in row 0 and two cabinets in row 5. The boot and SDB nodes are located in cabinet 0 in rows 0 and 5 while all the remaining service nodes are split between cabinet1 in rows 0 and 5.

There are two System Management Workstations (SMWs) that provide a recoverable solution to a catastrophic failure of the SMW.

Although many system services do not currently have automatic failover solutions, the system is configured to support these as they become available.

4.5 Networking

Franklin's primary access point is through the login nodes. Routing over ippo0 provides access through the network nodes to the 10Gb networks. Normal interactive keystrokes will utilize the 1 Gb interface in each of the login nodes.

Franklin's network nodes are split between two cabinets. All traffic from service nodes within a single cabinet will utilize the network nodes in that same cabinet. This not only splits the network load between multiple network nodes but also reduces traffic over the Torus by minimizing hops between the nodes.

With so much network connectivity, managing configurations in the shared root environment is overly complicated as each node would require it's own customizations. The Red Storm Team at Sandia National Laboratory developed a way to use a single network file that contains all the network information for all the nodes. The startup scripts in the default class were modified to perform lookups in this file. However, this was insufficient to NERSC's configuration needs. The scripts were further enhanced to increase the capability of flexible routing topologies while providing the basic networking infrastructure.

The entire network setup is now contained in a single directory. Some basic changes in the startup scripts were needed. The base network startup scripts are found in `/etc/sysconfig/network`. The `ifcfg-eth0` and `ifcfg-ippo0` now point to a custom scripts in the network directory. Additionally, the script `scripts/set-route-arp` also points to a custom script in the network directory.

Three data files now exist in the network directory. These are:

`arp-ippo0`	`ippo0` arp table
`hosts-external`	external interfaces
`routes-external` …	additional routing

The `arp-ippo0` file contains a mapping of internet address to hardware mac address for a secondary address on the `ippo0` device. This information is used to preload the arp table on all service nodes.

All definitions of externally visible interfaces are maintained in the hosts-external file. The file contains nid, location, nodename, device, address, netmask and MTU if necessary. The two `ifcfg` scripts will lookup the nid in this file and set the interface appropriately.

All additional routing is maintained in the routes-external file. The file contains nid, location, nodename, device, destination address, gateway address, netmasks, MTU and ipforwarding. The `set-route-arp` script will first preload the arp table with the `arp-ippo0` file and then process appropriate entries in the `routes-external` file.

The end result is that all network files are specialized to the default class rather than on a per node basis. The configuration can be easily updated and provides complete flexibility to provide the necessary networking configurations.

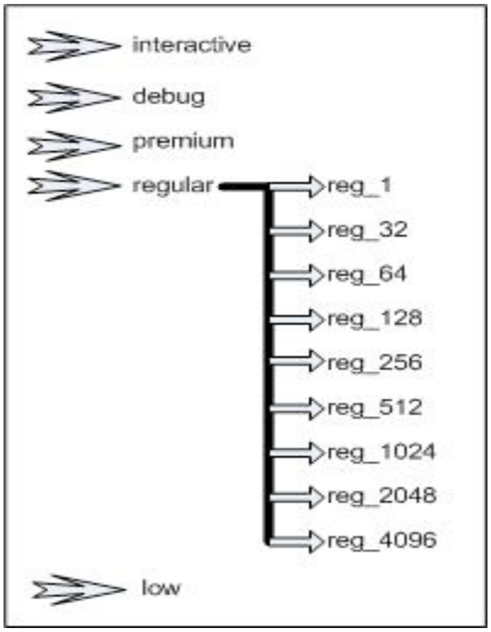
4.6 The Batch System

All nodes within Franklin are set to "batch", there are no "interactive" nodes. The inability to regulate the use of "interactive" designated compute nodes made this functionality undesirable in a large scale production operation.

The batch system is comprised of two components; a resource manager and a job scheduler. Franklin utilizes the TORQUE Resource Manager and Moab scheduler from Cluster Resources Inc[5]. While TORQUE manages the batch jobs, Moab initiates them. Although, having two separate components complicates things because each has their own configuration to maintain, the overall solution has increased functionality and flexibility.

The basic queue structure provides a simple to use interface for the users while maximizing the scheduling possibilities. The basic principle is that jobs are submitted to a class of service rather than specific queues. This allows the queue structure to evolve over time without impacting job scripts. There are five production classes of service: interactive, debug, premium, regular and low. These classes of service could serve as execution queues or routing queues to a series of finer grain execution queues based on job requirements.

The following diagram shows the flow of jobs through the production queue structure.



The ability to run `yod` from a login session is replaced by `qsub -I -q interactive`. However, without special handling, this part of the workload could be starved for resources. The solution comes from utilizing advanced features of the Moab scheduler. Moab supports the capability of queue bound standing reservations for predefined timeframes. A number of nodes are automatically reserved daily during primetime hours specifically for jobs submitted to debug or interactive classes of service.

4.7 Lustre

Three are two Lustre filesystems on Franklin: home and scratch.

The home filesystem consists of ten 2 TB Object Storage Targets (OSTs) on five Object Storage Servers (OSSs) and one Meta Data Target (MDT) on one Meta Data Server (MDS). The filesystem settings are:

```
stripecount: 2

stripesize: 1048576

blocksize :4096
```

The scratch filesystem consists of one hundred fifty eight 2 TB Object Storage Targets (OSTs) on twenty Object Storage Servers (OSSs) and one Meta Data Target (MDT) on one Meta Data Server (MDS). The filesystem settings are:

```
stripecount: 4

stripesize: 1048576

blocksize :4096
```

4.8 cron

Running the `cron` daemon is a necessity but a challenge in the shared root environment. The initial setup runs the daemon on one of the login nodes. However, crontabs can be submitted from any of the 16 login nodes. Access is controlled by the standard `allow` and `deny` files.

A second `cron` daemon runs on the boot node for administrative services. Crontabs can only be submitted on the boot node for this daemon.

4.9 LDAP

NERSC utilizes a centralized account management system which feeds a master LDAP server. A series of replicas have been strategically located to disperse the load of authentication lookups. Franklin has no local user accounts and relies on one of these replicas for primary account authentication lookups.

To improve lookup performance the Name Service Caching Daemon (nscd) is run on each of the login nodes. The initial time-to-live settings are 60 seconds. The impacts will be closely monitored as demand on the system increases to see if any adjustment is required.

5 In Closing

The Cray XT4™ is a large and complicated system. Harnessing the vast capabilities of such a system creates many challenges. As such, initial configurations may evolve to better suit the demands on the system. With over 2000 anticipated user accounts and such a large computational resource, many new challenges are yet to be discovered.

6 Acknowledgments

This work was supported by the Director, Office of Science, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract number DE-AC02-05CH11231.

This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy.

The process of designing, configuring, installing and testing a computer system of this scale requires the expertise of many people. The author wishes to acknowledge the accomplishments of the NERSC 5

procurement team as well as the NERSC 5 implementation team.

7 About the Author

Nicholas P. Cardo is the Project Lead and Lead System Administrator of Franklin. He is a senior member of the Computational Systems Group at NERSC. He can be reached at Lawrence Berkeley National Laboratory, National Energy Research Scientific Computing Center, 1 Cyclotron Rd, bldg 943r0256, Berkeley, CA 94720 USA, E-mail: cardo@nersc.gov.

8 References

1. Worksafe Technologies Inc., www.worksafetech.com.
2. Cray Inc. Cray XT4 Data Sheet, www.cray.com/downloads/Cray_XT4_Datasheet.pdf.
3. Energy Information Admistration, www.eia.doe.gov.
4. 4 Cray Inc. Cray XT4 Data Sheet, www.cray.com/downloads/Cray_XT4_Datasheet.pdf.
5. Cluster Resources Inc, www.clusterresources.com