University of California, Riverside

Computing and Communications

Computational Clusters @ UCR

March 28, 2008

1 Introduction

A recent survey showed that more than half of higher education institutions have seen an increase in the use of high-performance computing, high-performance networking and data storage and are predicting an even greater demand in the coming years.[1] This rise in demand coincided with the shrinking financial resources available to research faculty has created a unique challenge. PIs are turning more to collaborative, team-based research methods that draw upon expertise and pooled resources regardless of geographic location.

Cyberinfrastructure (CI) has emerged in recent years as a facilitator of this new research paradigm at universities everywhere and in particular the University of California. Here at UC Riverside as part of a larger CI strategy, Computing and Communications has developed a comprehensive three-year computational cluster plan.

2 Campus Datacenter Co-Location Facility

Given the shortage on premium datacenter space on campus, this plan makes heavy use of the campus datacenter in the Statistics and Computing building in a specifically designated area known as the Co-Location Facility.

Clusters that are housed here will have the benefit of a direct connection to not only the campus high speed 10 Gigabit backbone but will have a connection to an extremely fast 10 Gigabit link to Cenic/Internet2 making collaboration with other universities involving extremely large datasets much more feasible.

Other advantages of a co-located cluster are the power, cooling, and FTE provided by Computing and Communications that wouldn't be available to a cluster housed in a researcher's laboratory.

3 Three Cluster Models

The proposed plan includes support for three different cluster models that are housed in the Co-Location Facility. These include departmentally maintained clusters, dedicated clusters and a shared collaborative cluster.

3.1 Departmentally Maintained Clusters

These systems are built to particular PI / lab / center specifications and managed by PI funded staff, but housed within C&C's data center with C&C staff management / mentoring / backup provided to the departmental administrator. This type of cluster is meant for researchers who have computing needs that fall outside of the campus cluster standards described in the next section.

[1] *IT Engagement in Research: A Baseline Study*, EDUCAUSE Center for Applied Research, 2006

3.2 Centrally Hosted Dedicated Clusters

These systems are built to campus standards so that multiple systems servicing the needs of many researchers can be deployed and maintained by a relatively small number of systems administrators. These campus standard computational clusters will be maintained by Computing and Communications and housed in the campus Co-Location Facility.

Technical staff within Computing and Communications will assist all researchers who wish to be part of this program with purchasing the best solution for their needs.

3.2.1 Intended Users

These clusters will be designed to meet the needs of research labs, research centers, and PIs with relatively substantial compute requirements. The nodes are readily available high-power 64-bit systems, meant for applications that can easily be distributed, rather than a highly parallelized cluster system with a fast (and expensive interconnect).

3.2.2 Financial Support Model

PIs fund these Standardized/Dedicated Clusters via their own funding sources, including initial complements, extramural funds, etc. PIs are also responsible for maintaining these clusters.

3.2.3 Campus Cluster Standards

In order to maximize the efficiency of the support staff dedicated to managing these clusters a set of hardware standards have been established. Researchers that chose to house their cluster in the Co-Location Facility and have it managed by Computing and Communications must follow these standards.

<u>Compute Nodes</u>

- 1U Rack Mounted Servers
- Dual AMD Opteron 2214 Dual Core Processors
- At least 8 GB of RAM per node
- At least 1 80-160GB Hard Drive
- High quality hardware that is as green as possible
- Dual Gigabit Ethernet ports
- DDR Infiniband interconnect HCA with cable (*Optional)*
- IPMI
- 3 year warranty

<u>Master Node</u>

- 1U Rack Mounted Server
- Dual AMD Opteron 2214 Dual Core Processors
- At least 8 GB of RAM

- Dual Gigabit Ethernet ports
- 3 year warranty

Interconnect (Network)
- Gigabit Ethernet switch
- Infiniband fast interconnect switch with embedded subnet manager (*Optional)*

Storage
- No specific standard but PI must consult with Computing and Communications prior to purchase to ensure that the support staff is prepared

Scheduling/Administration Software
- Perceus for cluster OS provisioning
- Sun Grid Engine is highly recommended but any number of schedulers/resource managers are supported as long as they support interoperability with the UC/UCR Grid.

Operating System
- CentOS or ScientificLinux

These standards will be reviewed from time to time to ensure that clusters purchased as part of this program will be of the high quality and best performance available.

3.2.4 Software Installed

A standard set of software packages will be installed on every cluster that is part of this program. Researchers may choose to alter this list as they see fit but are responsible for purchasing any non open source software package that isn't part of the initial install.

Compilers
gcc (C)
g++ (C++)
g77 (Fortran 77)
gfortran (Fortran 95)
gdb (Debugger)
gnat (Ada 95)

Other Development Tools/Numerical Libraries
Sun JDK
MPICH
ARPACK
ATLAS
Basic Linear Algebra Subprograms (BLAS)
FFTW
GNU Scientific Library (GLS)
HDF(4)

HDF5
Intel Math Kernel Library (MKL)
LAPACK
netCDF
OpenMotif
ScaLAPACK

Databases
MySQL
PostGreSQL

Science and Mathematics
Octave

Plotting and Graphing
gnuplot
grace
Metis
ParMetis

Visualization and Modeling
OpenDX
ParaView
POVray
RasMol

3.2.5 Equipment Obsolescence

To maintain efficient use of all resources dedicated to this program it is important that all hardware housed in the Co-Location Facility be up to date. After a period of three years after any cluster has been purchased and housed in the Co-Location Facility it will be evaluated based on condition of the hardware and the cost of continued maintenance. If it is deemed to be still cost effective to continue to house the hardware it will be retained and evaluated annually after. If it is no longer feasible to continue to maintain the hardware, Computing and Communications will work with the equipment owners to either turn over administration of the cluster to their own staff or assist them with disposing of the obsolete hardware.

3.3 Collaborative Computation Cluster

UCR's collaborative cluster provides a shared system as a computing resource for campus researchers with limited financial resources.

3.3.1 Collaborative Cluster Components

The 69 compute nodes available in the cluster are divided into three logical sub clusters. The first 40 nodes are part of a general purpose cluster that is made available to researchers and sponsored graduate and undergraduate students who do not have access to any other cluster on campus to run computations that take a very short time to run. The next 24 nodes are part of a base shared cluster system available only to researchers who contribute nodes to expand the cluster. All nodes in this first set of 64 are part of an extremely fast InfiniBand interconnect. Jobs run on these nodes may take advantage of the low latency communication and high bandwidth that this interconnect offers. The remaining 5 nodes in the cluster are part of what is known as the Application Cluster and are intended for jobs that cannot take advantage of the Infiniband network.

3.3.2 Intended Users

The collaborative cluster is ideal for researchers with higher end compute needs but without the financial resources to obtain the infrastructure necessary for a fast interconnect (an InfiniBand interconnect switch can add more than $50,000 to the overall cost of a cluster).

Of this group two types are targeted: researchers who need occasional use of computational resources and those who regularly use high end computation in their research. The former group of researchers would be able to make use of the 40 node sub cluster to run the occasional short job. The latter group of researchers would have available to them the 24 nodes of the base shared cluster in additional to their own nodes they add to the cluster and for a longer amount of time.

3.3.3 Financial Support Model

The initial complement of 69 compute nodes, master node, InfiniBand interconnect and storage server is provided by Computing and Communications (see below for details) in the campus data center. The master node houses user accounts and is the node on which jobs are launched and applications created and/or compiled. Faculty researchers who make regular use of the cluster are encouraged to participate in building the cluster by using funding sources at their disposal (such as grants and initial complements) to purchase additional compute nodes.

3.3.4 Hardware

The hardware specification for the collaborative cluster is as follows.

<u>64 Compute Nodes</u>

- 1U Rack Mounted Servers
- Dual AMD Opteron 2214 HE Dual Core Processors
- 8 GB of RAM per node
- 1 80GB Hard Drive

- High quality hardware that is extremely green
- Dual Gigabit Ethernet ports
- DDR Infiniband interconnect HCA

Master Node
- 1U Rack Mounted Server
- Dual AMD Opteron 2214 HE Dual Core Processors
- 8 GB of RAM
- Dual Gigabit Ethernet ports
- 1.5 TB internal storage
- DVD-RW drive

Interconnect (Network)
- Gigabit Ethernet switch
- QLogic Infiniband fast interconnect switch

Storage
- Sun X4500 Storage Server (48TB raw capacity)
- Backup X4500 Storage Server (offsite, 48TB raw capacity)

As noted above, special care as been made to ensure that this cluster is as green as possible. This includes using lower voltage processors and more efficient power supplies.

Researchers who add nodes to the cluster must follow the campus cluster standards in section 3.2.3. To make the cluster as homogeneous as possible however, the InfiniBand HCA is required.

3.3.5 Software

The following software will be available to researchers who use any of the nodes that have been provided by Computing and Communications. Any nodes that have been added by researchers will have a software suite available that is similar to those with dedicated clusters (see 3.2.4 above). Researcher requests to add additional software will be evaluated on a case by case basis.

Compilers
ifort (Fortran 77/95)
icc (C/C++)
gcc (C)
g++ (C++)
g77 (Fortran 77)
gfortran (Fortran 95)
gdb (Debugger)
gnat (Ada 95)

Other Development Tools/Numerical Libraries

Sun JDK
MPICH
ARPACK
ATLAS
Basic Linear Algebra Subprograms (BLAS)
FFTW
GNU Scientific Library (GLS)
HDF(4)
HDF5
Intel Math Kernel Library (MKL)
LAPACK
netCDF
OpenMotif
ScaLAPACK

Databases
MySQL
PostGreSQL

Chemistry
Gaussian
Q-Chem

Science and Mathematics
Mathematica
Maple
Matlab
Distributed Matlab or Star-P
Octave

Plotting and Graphing
gnuplot
grace
Metis
ParMetis

Statistics
R
SAS*

Social Sciences
Stata
SPSS*

Visualization and Modeling
IDL

IDL Analyst
OpenDX
ParaView
POVray
RasMol
TecPlot

3.3.6 Cluster Access

Access to the cluster may be done through either the UCR Grid Portal or through SSH to the head node.

3.3.7 Queues

To ensure the maximum use of the available cycles on the cluster several queues will be established.

General Purpose Component

On the general purpose 40 node component a 24 hour queue will be in effect. This means that any job that is submitted to this queue will be limited to a run time of 24 hours. There is no guarantee of when the job will start and is based on the number of waiting jobs in the queue and the number of cores being requested. Jobs that require more cores will be pushed back in the queue.

Shared Component

Researchers who have contributed nodes will have access to a research group queue that has a time constraint of 14 days. This means that if a researcher has contributed 8 cores to the cluster a job may be submitted that uses up to 8 cores and may run for a maximum 14 days.

These researchers also have access to a 24 hour surplus cycle queue that takes advantage of unused cycles on the base set of 24 nodes provided by Computing and Communications as well as any unused cores on nodes that have been contributed by other researchers.

3.3.8 Storage

All users of the cluster of any either the general purpose component or the shared component will have access to the shared Sun X4500 storage server.

Users of the general purpose component will have a 2 GB home directory quota available to them. This storage space is reserved for data specific to the use of the cluster and is not intended for long term storage.

Researchers who contribute nodes to the cluster will have 1TB for free for distribution to members of their research group and may purchase additional storage in 1TB increments for $2000 each. This cost will not only cover the cost of the disk but of the maintenance and regular backups that will be done.

3.3.9 Governance

An advisory panel of one representative of each research group that has contributed will be established to recommend policy changes on the collaborative cluster to the Associate Vice Chancellor of Computing and Communications. This group will also review requests for abnormal uses of the cluster such as a request to use all available nodes for a special research project.

4 Grid Computing

An additional important component of UCR's Cyberinfrastructure strategy is to easily make available storage and compute resources outside their own dedicated cluster by the embracing of grid computing.

Other institutions within UC, notably UCLA, UCSB and UCI have already taken steps to form a UC wide grid. This enables researchers at any of those campuses to submit compute jobs that can run on any cluster that is part of the grid at any campus. UCR seeks to join the UC Grid as well as form a local grid of shared resources to be known as the UCR Grid.

4.1 UCR Grid

The UCR Grid is essentially a subset of the UC Grid. The collaborative cluster will be the first member of this grid. All cluster owners who participate in the other two cluster models will also be invited to take part in it as well.

Cluster owners may participate in the grid by adding a grid appliance node provided by Computing and Communication and making a certain number of cycles available to pool users of the grid.

Grid users will not only have access to resources on other clusters in the UCR Grid but will also have the ability to directly access clusters outside of the UCR Grid such clusters that are on the TeraGrid.

Jobs are submitted to the UCR Grid through the UCR Grid Portal. Users of the portal will have access to several tools including the Data Manager service which gives users a single interface to handle data management across all the UCR grid clusters a user has access to.

4.2 UC Grid

The UC Grid and associated UC Grid portal is very similar to the UCR Grid other than it gives users access to resources at UCLA, UCSB and UCI.

5 Conclusions and Future Direction

Computing and Communications is extremely committed to this new Cyberinfrastructure strategy with the establishment of these three cluster programs. In the future it is hoped that other services may be offered to researchers such as:

- Seminars on parallel programming
- Seminars on common research software use
- Consultations on parallel algorithm optimization
- Internal grants for cluster related research
- Establishment of cluster user groups

Through an investment in the relationship between Computing and Communications and the campus research community this campus can be a model for collaborative research throughout the UC.