HPC at Rice University

Using Technology for the Sober Fearless Pursuit of Truth, Beauty, and Righteousness.

Dr. B. Kim Andrews
kimba@rice.edu

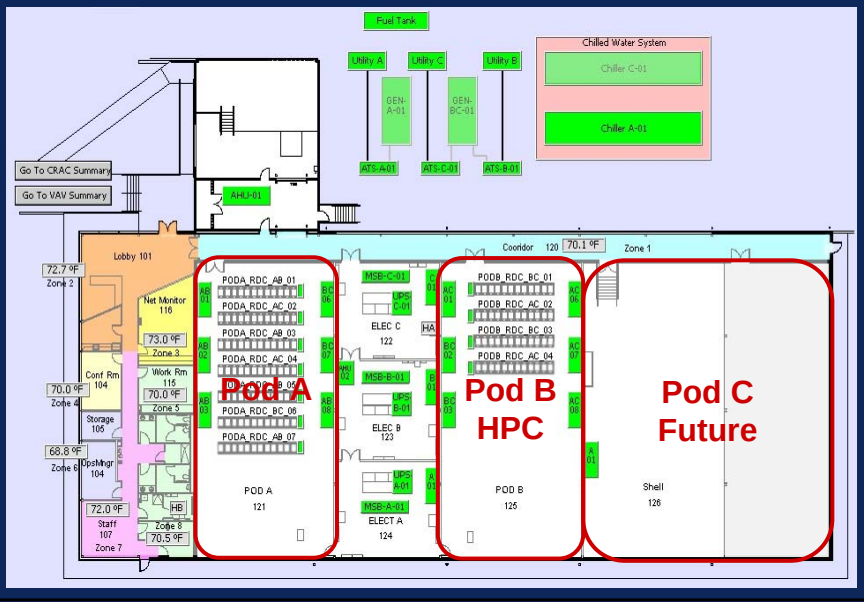
Manager of Research Computing

Information Technology

11/14/08

The Primary Datacenter

Data Center Overview



Physical Building Attributes

Physical Space	Size	Maximum Rack Capacity
Pod A	4,000 s.f.	96 x 30 RU = 2,880 RU
Electrical A, B, C	3,000 s.f.	-
Pod B	4,000 s.f.	72 x 30 RU = 2,160 RU
Pod C (shell space)	4,000 s.f.	TBD
Electrical D, E, F (shell space)	3,000 s.f.	-

Facility Capacities (day 1 and max)

Infrastructure Component	Redundancy	Current Capacity	Maximum Design Capacity
Electrical service	True A/B	2 mW useable 6 mW total	4 mW useable 6 mW total
UPS	True A/B	3 @ 625 kVA (500 kW) 1875 kVA (1500 kW) total	6 @ 625 kVA / 500 kW 3750 kVA (3000 kW) total
Chillers	N+1 = 2N	350 tons	1,000 tons
Pod CRACs	N + 1	12 @ 30 tons	28 @ 30 tons
Generators	True A/B	2 @ 2 mW	3 @ 2 mW
Pod A Racks	-	72 server, 18 network	96 server @ 5 kW ea. 18 network @ 7 kW ea.
Pod B Racks	-	36 server, 9 network	72 server @ 15 kW ea. 18 network @ 7 kW ea.
Electrical Circuits (typical)	2N	2 three-phase, 90 amp outlets per rack (6 circuits / 6 poles)	4 three-phase, 90 amp outlets per rack (12 circuits / 6 poles)

Cooling and Power Density

Pod	CRACs (30 ton)	Redundancy	Total Tons	Racks (12 per row)	kW per Row	Kw per Rack
A	2	N+1 (2N)	30	24	52	4.375
A	6	N+1	150	72	87.3	7.319
B	**6**	**N+2**	**120**	**36**	**140**	**11.667**

Pod	CRACs (30 ton)	Redundancy	Total Tons	Racks (12 per row)	kW per Row	Kw per Rack
A/B	**12**	**N+3**	**270**	**108**	**105**	**8.75**

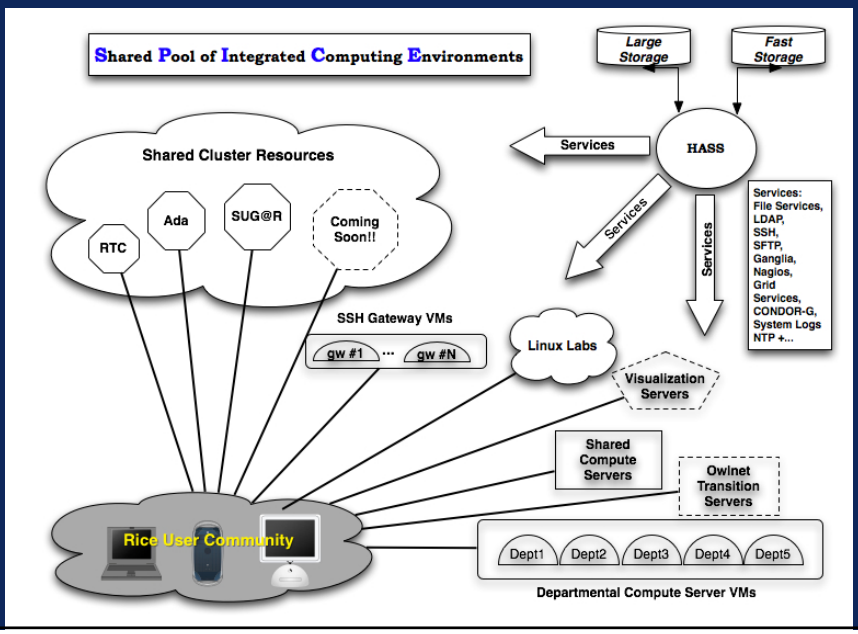
Design Intent Matrix

Design Element	Within Design Intent	Slightly Above Design Intent	Moderately Above Design Intent	Well Beyond Design Intent
kW per rack	8 or less	8 - 12	12 - 15	15 - 22
Racks per row / Type	12 / CPI	10 / CPI	10 / Special	10 / Very Special
Airflow Base = (Passive rack, front-to-back, hot-isle / cold-isle)	Base	Base + Some Active (rack door fans)	Base + More Active (by rack, baffling, plenum, or liquid)	Base + More Active (by rack, baffling, plenum, or liquid)
Redundancy	N+1, N+2	N+1, N+2	Potentially compromised	Potentially compromised
Cost increase	None long as capacity is available	Minimal as long as capacity is available	Potentially extensive	Likely very extensive
Flexibility	Very flexible system placement	System may need to be split into more racks than intended or expected	System placement will be completely dependent on power, cooling, and airflow considerations	System placement will be completely dependent on power, cooling, and airflow considerations
Other Considerations			Liquid cooling exchangers may complicate placement	May require Pod C Infrastructure Expansion

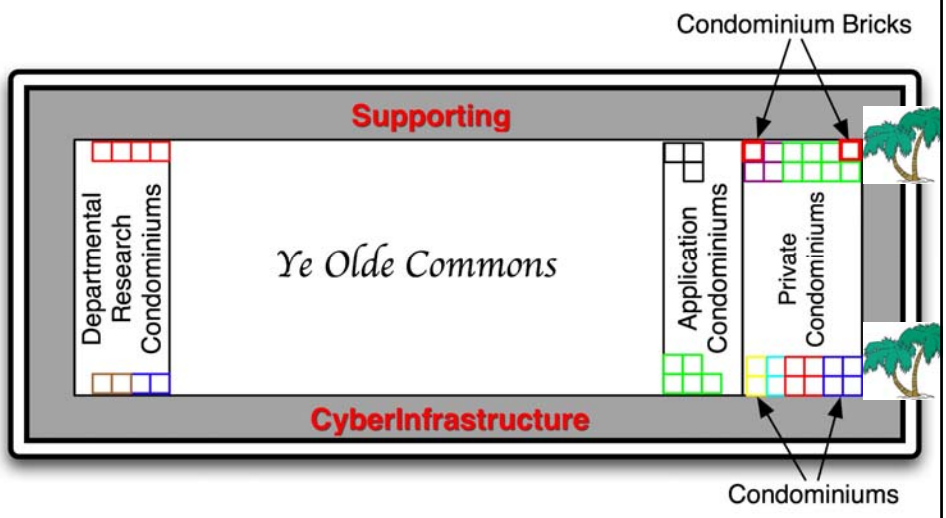
Hardware - Overview

- Variety of platforms, both old and new
- Both High-capability and High-capacity
- We use it so we can't lose it.
- Moderate resources with high utilization for very large, diverse user base
- No Leasing - The University funds infrastructure servers, but grants provide shared and dedicated compute nodes in fits and squirts
- Condominium Computing is the path forward

Campus Overview



The SUG@R Condominium Resort



Hardware - SUG@R

- Purchased in April 2008
- Full production in June 2008
- Intel Xeon processors
- SunFire x4150
- 134 compute nodes
- 1072 processing cores
- 2 login nodes
- 2 management nodes
- 10 TeraFLOPS peak

Hardware - SUG@R

- Two quad-core processors per node
- 16GB RAM per node
- All nodes connected via Gigabit ethernet.
- No fast interconnect
- Does not support MPI between nodes!
- 500GB storage (NFS) for home directories
- 1TB storage (NFS) for shared group directories
- 20 TB fast scratch storage on Panasas (PanFS) filesystem

Hardware - SUG@R

- LOM Network
- 4 GigE network links for public access
- Ideal for jobs requiring 8 processors (one node) or less
- Maximum run time is 24 hours per job - 32/64 CPU limit

Hardware - Ada

- MRI Grant: September 2004
- Production: February 2005
- AMD64 Opteron
- 158 compute nodes
- 632 processor cores
- 4 login nodes
- 4 file server nodes
- 2 management nodes
- 3 TeraFLOP peak

Hardware - Ada

- All nodes connected via RapidArray (Cray proprietary interconnect based on Infiniband, two 2GBytes/sec links per chassis)
- LOM Network for remote management
- 6 GigE network links for public access
- Ideal for parallel jobs that require a large (>128) number of processors
- Maximum run time is 8 hours per job - no limit on CPUs

Hardware - Ada

- Fibre Channel Storage
  - /users (NFS) and /projects (NFS), 5TB each
    - NFS (not for high performance job I/O)
    - 42 disks (Hitachi SATA 7200 RPM)
  - /lustre (fast scratch, 5TB)
    - Lustre Filesystem (large block job I/O)
    - 56 disks (Hitachi FC 10000 RPM)
    - Data is not persistent (two week expiration)

Hardware - RTC

- MRI Grant: September 2002
- Production: June 2003
- Intel Itanium 2 (900 MHz McKinley)
- 1.3GHz Madison nodes added 2007
- 134 compute nodes
- 286 processors
- 2 or 4 processors per node
- 3 login nodes
- 4 file server nodes
- 2 management nodes
- 1 TeraFLOP peak performance

Hardware - RTC

- LOM network for remote management
- All nodes connected via Gigabit Ethernet
- 65 nodes connected via Myrinet (2Gbits/sec for MPI traffic)
- 4 GigE network links for public access
- Ideal for 64 bit applications requiring large amounts of memory
- 700GB storage (NFS) for home directories
- 2.5TB storage (NFS) for shared group directories
- 5TB fast scratch (PVFS) for job I/O
- Maximum run time is 2 weeks per job

Hardware - Campus

Dedicated Clusters:

- 8 clusters ranging from a 128 CPU cluster in bioengineering to a 4 CPU cluster in physics.

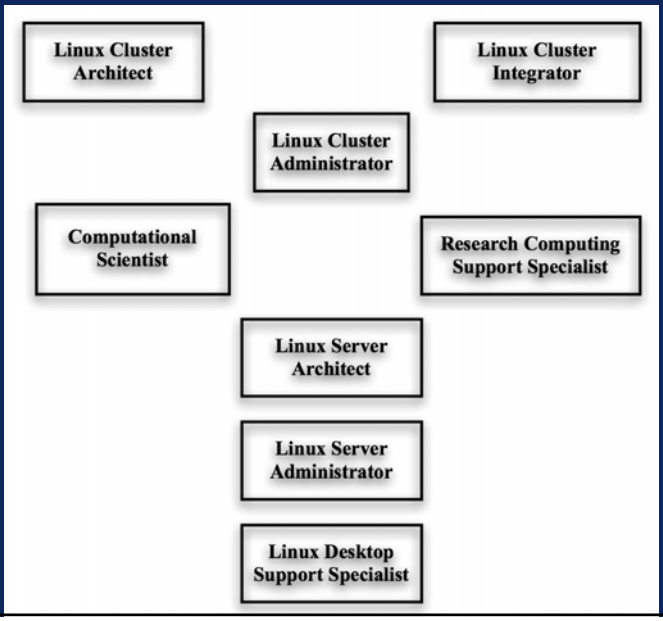
Dedicated Servers:

- More than 113 servers ranging from interconnected groups in Bonner Nuclear Lab and Electrical & Computer Engineering to stand-alone servers in use across campus.

Operations - OS Support

- Red Hat Enterprise Linux 5 on all Beowulf clusters
- SUSE is deployed only on the Cray
- Linux B & E Practices for campus
- Red Hat kickstart server for campus
- Red Hat Enterprise Linux standard image
- RHN Proxy / Satellite (for automated patch installation)

Technologists



Operations - Typical Work

- User account management
- Debugging of batch scripts
- Software installations and upgrades
- Software compilation assistance
- Job utilization monitoring, profiling and performance analysis
- Debugging compute node failures
- Data management (filesystem management)
- Capacity planning (CPU, networking and storage)
- Special projects and allocations / Export Control
- Provide problem solutions to more than 500 users (RT, FAQs, Workshops)
- Monitoring and tuning of job schedulers
- Documentation (FAQs)
- Routine maintenance (OS patches, hardware fix)

Operations - Infrastructure Software

- OpenLDAP server for authentication
- Storage (primary and disaster recovery)
- Web sites (rcsg.rice.edu)
- Account management system (PHP/MySQL)
- Flexlm License Server (software licenses)
- Nagios – system availability
- Ganglia - system utilization
- HAAS –High-availability system for the above resources (Red Hat Advanced Platform)
- Wiki for internal documentation

Operations - Access

- Software firewall (Linux iptables) blocks most network traffic to management/login nodes.
- Essential traffic (i.e. SSH, DNS, NTP, LDAP) are allowed.
- No access to compute nodes from campus network. Nodes are in a private network.
- No access to login nodes from off campus. Must use SSH Gateway (login gateway).
- Access to compute nodes is via job scheduler only!
- Login via SSH only.
- Operating system security updates applied as needed or 2-3 times per year.

Applications

- Huge variety of Linux-based applications
  - HomeGrown
  - Commercial
  - ExperiMental
  - Mental
- Jobs span requests that range from 1 to HMCIG? CPUs and runs from 10 mins to 10 mos
- Hardest balance is CS versus Gaussian
- Use faculty profiling tools (e.g. HPC Toolkit), H/W optimized libraries, RACT experiments as input
- Computational Scientist on staff
- Members of TeraGrid Champions program

Applications

- Amber – a set of molecular mechanical force fields for the simulation of biomolecules (which are in the public domain, and are used in a variety of simulation programs); and a package of molecular simulation programs which includes source code and demos
- Comsol – The COMSOL Multiphysics® simulation environment facilitates all steps in the modeling process —defining your geometry, specifying your physics, meshing, solving and then post-processing your results.
- CPLEX - ILOG CPLEX delivers high performance with robust, flexible optimizers for solving linear, mixed-integer and quadratic programming problems in mission-critical resource allocation applications.
- PETSc - a suite of data structures and routines for the scalable (parallel) solution of scientific applications modelled by partial differential equations. It employs the MPI standard for all message-passing communication.

Applications

- Matlab – many thousands of batch jobs submitted by single user or a few long running interactive graphical jobs.
- SAS – business intelligence application for analytics, data manipulation, and reporting.
- Mathematica – both interactive graphical and batch jobs.
- R – statistical computing (linear and non-linear modelling) and graphical techniques.
- Gaussian – used by chemists, chemical engineers, biochemists, physicists and others. Starting from the basic laws of quantum mechanics, Gaussian predicts the energies, molecular structures, and vibrational frequencies of molecular systems, along with numerous molecular properties derived from these basic computation types.
- NAMD – parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems

Applications -☐ Top Disciplines

- Biochemistry
- Physics
- Bioengineering
- Mechanical Engineering
- Earth Sciences
- Computational and Applied Mathematics
- Computer Science
- Chemistry
- Statistics

Applications - Teragrid

- Campus Champions
- Faculty or Staff member applies for DAC
  - Liaison between researchers and TeraGrid staff
  - Receives 30,000 SU (CPU hours)
  - Can add user accounts for researchers
  - Used for testing and training before faculty member applies for his/her own allocation
  - Provides access to TeraGrid and input to its staff
  - Quick access to TeraGrid
  - Access to Workshops
  - Free registration to annual conference

SC 08
November 15 – 21, 2008
Austin, Texas

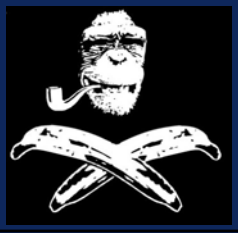
Campus Champions BoF: Bringing HPC to Your Campus
Wednesday, November, 19 at noon.

User Issues

- High-volume, low-CPU jobs
- QoS versus inefficient scalability
- Number of active jobs
- Fair Share across partitions & systems
- Optimization motivation for apps
- Compute-on-Demand requirements
- Disk Hogs
- Contiguous runtime requirements

The Research Computing Support Group at Rice University

The RCSG



"Putting the arRRRRRrrrrrrr back into
R*esearch*
C*omputing*
S*upport…"*

Thank You for Feigning Polite Interest!

Thank You for Your Interest!