

USC HPCC
Campus
Research Support
I2 Workshop 2006
Jim Pepin

HPCC

- ◆ Provide common facilities and services for a large cross section of the university that requires leading edge computational and networking resources.
- ◆ Leverage USC central resources with externally funded projects.
- ◆ Campus ISD support for base facilities
- ◆ Not just usual suspects play

HPCC

- ◆ Faculty advisory group
- ◆ Allocation committee
 - Large Disk/cpu
 - Favors larger scale problems.
 - Favors interdisciplinary
 - 5 users over 500k hrs
 - Routinely setup 512 node runs
 - 256 are standard queue
 - Show pbsstop
- ◆ Large memory jobs are also common
- ◆ Viz/Data initiatives
- ◆ 13.8TF cluster

HPCC

◆ Highlights

- Leverage across university.
- ISD is catalyst for inter-disciplinary work
 - ◆ USC strategic plan stresses this
- 3,000,000+ node hours in last year
- 8 'condo' users 600+ nodes
- 70TB 'condo' disk
- Networking
 - ◆ Los Nettos
 - ◆ Pacwave
 - ◆ Dark fiber/waves
- Shared staff

Current HPCC Resources

- ◆ High Performance Computing Resources
 - Linux Cluster (1830 nodes/5384cpus, 2Gb/sec Myrinet), many file head nodes
 - ◆ ~100TB shared disk, 18GB - 40GB local disk per node.
 - ◆ Ranks in top few for academic clusters last time. (stay tuned)
 - Myrinet switch is 1952 total nodes possible now.
 - Adding nodes funded by USC research groups.
 - Sun Core Servers (E15k shared memory)
 - ◆ 72 processors, 288GB memory, 30TB shared disk
 - Mass Storage Facilities (QFS)
 - ◆ 10,000 tape capacity
 - ◆ 1.1PB on tape

Building a Big Cluster

- Large cluster represent unique challenges
 - ◆ Power
 - ◆ A/c
 - Air flow
 - ◆ Hot spots
 - ◆ Volume
 - What happens when a/c fails
 - ◆ Wiring
 - Density
 - Testing
 - Power cabling.
 - Blocking cooling

Why It's Hard

- Large cluster represent unique challenges
 - ◆ Software installation
 - Non-homogenous cluster
 - ◆ Built over time
 - ◆ Different vendors
 - ◆ Different hardware base configurations
 - ◆ Merging new 'chunks' is complicated
 - High speed network (Myrinet)
 - ◆ New spine(s)
 - Gb ethernet
 - ◆ New ports
 - How to do this in running cluster
 - ◆ VERY carefully
 - ◆ Pre-position cables/switches
 - ◆ Lots of labor in short time

Configuration

- Types of processors
 - ◆ Sun V60 (3.0Ghz, 500Mhz, 2GB) (some 4GB)
 - ◆ Dell. 3.2 (3.2Ghz, 800Mhz, 2GB)
 - ◆ Sun V20z (opterons) (2.2Ghz, 2GB)
 - ◆ Sun V20z (opterons) (2.0Ghz, 4GB dual core, dual node)
 - ◆ Sun 4100 (opterons) (2.0Ghz, 4GB dual core, dual node)
 - ◆ Able to mix arch.
- Goal is to increase in cores/processors per node
 - ◆ 16 cores in next 12-24 months
 - ◆ All nodes 64 bit in year
- Interesting power trade-offs with opterons vrs xeon
- Power pc options are out there.

Configuration

- Networks
 - ◆ Myrinet
 - ◆ 7 spine pairs on one chassis 128 ports each 896 total potential
 - New 4/1 cards 4 fiber on each port
 - Reduces clutter
 - Room for 3 more spine pairs
 - ◆ 15 128 port edge switches
 - 96 host ports
 - 32 spine ports
 - 3/1 host/spine ratio
 - ◆ 2 256 port edge switches
 - 256 host ports
 - 128 spine ports
 - 2/1 host/edge ratio
 - ◆ Ethernet for each node
 - File I/O
 - Many back-end file servers (15k, 440s,240s,v20zs,v40zs (qfs fses, nfs access))
 - ◆ Console concentrators for all nodes (management)

Hot Stuff

- Power loads at full song
 - ◆ 480v transformers
 - Another 80KW on 208v (runs building lights and suppl a/c (lieberts) and one main a/c)
 - ◆ We have surpassed power available on our generator
 - Special setup to allow cluster to be off generator

Transformer	Before	During
T1	273kw	360kw
T2	390kw	750kw

Real Life Cluster in Datacenter

- ◆ Slide show.
 - Heat density
 - Power
 - ◆ Lets melt
 - ◆ Web page
 - Building/logistics

New Data Center

- ◆ Online in fall
- ◆ Some blueprints