

Building/
Benchmarking
7TF+ Cluster
Jim Pepin
USC/ISD-HPCC

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.

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
- 6 'condo' users 295 nodes
- 50TB 'condo' disk
- 5 users over 200,000 node hours

Current HPCC Resources

- ◆ High Performance Computing Resources
 - Linux Cluster (1726nodes/3552cpus, 2Gb/sec Myrinet)
 - ◆ ~80TB shared disk, 18GB - 40GB local disk per node.
 - ◆ Ranks in top 10 for academic clusters.
 - Myrinet switch is 1426 total nodes.
 - Adding nodes funded by USC research groups.
 - Sun Core Servers (E15k shared memory)
 - ◆ 72 processors, 288GB memory, 30TB shared disk
 - Mass Storage Facilities (QFS)
 - ◆ 18,000 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

Building a Big Cluster

- Large cluster represent unique challenges
 - ◆ Benchmarking issues
 - Runs of entire cluster takes long time (on order of hour)
 - Do smaller runs.
 - REALLY important to find bad hardware
 - ◆ During complete runs power is almost 2x increase over idle.
 - ◆ When using different speed nodes benchmark is at speed of slowest processor in set.
 - ◆ Opteron are not a help in benchmark in our case.
 - 2.2 Ghz opteron is 'faster' for user jobs but slower than 2.8Ghz xeon on linpack.
 - ◆ Tried various ways to finesse this.
 - ◆ Results in table slides.

Upgrade Process

- Build Sequence
 - ◆ Before downtime
 - First step was to run small size before 'down time'
 - ◆ Wireless from Washington.
 - Scheduled 12 hr cable party during power work in building.
 - ◆ Put in 3 new Myrinet spines
 - Did run on new 356 nodes during next week, used a few existing 335s and v60s, v20s to get mixed results
 - ◆ Also provided burn in period for new xeons, checked cabling etc.

Benchmark Process

- Scheduled for 1 week
 - ◆ Did not plan to take that time (too much pressure)
 - ◆ Took system down Friday night
 - 3-4 hrs of final cable moves and some processor swaps
 - ◆ First run was 6.8
 - Fixed 3.2 came back online
 - ◆ Noticed we could up N a bit
 - ◆ Each run at this size was about 1 hr.
 - ◆ Got 7.291TF at NB=80
 - ◆ Ran some NB > 80
 - Incremental TFs were not going to produce better
 - ◆ “Deal” with staff was if I got 7+ we were done
 - ◆ Cluster back available Saturday.
- What the numbers look like

Configuration

- Types of processors
 - ◆ IBM x335s (2.8Ghz, 400Mhz, 1GB)
 - ◆ Sun V60 (2.8Ghz, 500Mhz, 1GB)
 - ◆ Sun V60 (3.0Ghz, 500Mhz, 2GB) (some 4GB)
 - ◆ Dell. 3.2s (3.2Ghz, 800Mhz, 2GB)
 - ◆ Sun V20s (opterons) (2.2Ghz, 2GB)
- Benchmark build-up
 - ◆ 2x2, 4x4, 16x16, 26x26, final 44x60
 - Pure speeds (2,8s, 3.0s, 3.2s)
 - Used 16x16 test chunk of cluster
 - ◆ 4x4 to narrow down bad nodes/cables
 - 26x26 to test new addition before full downtime

The Numbers

- Results at various speeds

- ◆ 4x4, n 40000 (~700MB/proc)

- ◆ 4x4 n25000(~400MB/proc)

Types	GF	%peak
3.0Ghz	65.9	68.6%
3.2Ghz	70.9	69.2%
3.2+2.8	63.8	71.3%
3,2+3.0+2.8	63.8	71.3%
3.2+3.0	67.1	69.8%
V20(2.2)	49.6	69.5%
3.2+2.2	42.6	60.0%

Types	GF	%peak
3.0Ghz	61.1	66.0%
3.2Ghz	67.2	63.6%
3.2+2.8	56.5	63.0%
3,2+3.0+2.8	-	-
3.2+3.0	63.3	66.0%

The Numbers

- Results at various speeds

- ◆ 8x8 n 70000 ~700MB

26x26 n 235000 ~700MB

Types	GF	%peak
3.0Ghz	-	-
3.2Ghz	268.2	65%
3.2+2.8	243.4	67%
3,2+3.0+2.8	-	-
3.2+3.0	-	-

Types	TF	%peak
3.0Ghz	-	-
3.2Ghz	2325	59.,7%
3.2+2.8	2213	58.3%
3,2+3.0+2.8	-	-
3.2+3.0	-	-

Final Answer

- Final Runs
 - ◆ 44x60, NB 80, 1320 nodes 2640 cpus

N	TF	Percent/peak
325000	6.855	46.3%
340000	7.291	49.3%

Hot Stuff

- Power loads
 - ◆ 480v transformers
 - Another 50KW on 208v (runs lights and suppl a/c (lieberts))

Transformer	Before	During
T1	273kw	360kw
T2	354kw	610kw
Non a/c (UPSes)		655kw(67%)

Pretty Pictures

- Build-it slide show