LQCD Facilities at Jefferson Lab

Chip Watson

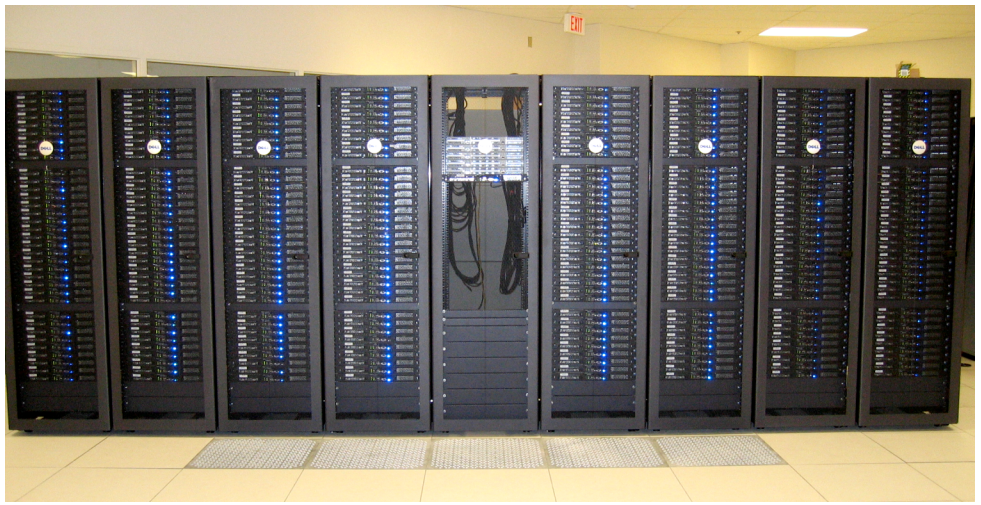
March 23, 2007

Existing Clusters

3g 2003 gigE mesh
2.66 GHz P4, 256 MB / node
½ decommissioned,
now just 128 nodes
no allocation this next year

4g 2004 gigE mesh
2.8 GHz P4, 512 MB/node
384 nodes, 3 sets of 128
start to decommission in 2008

6n 2006 infiniband
3.0 GHz Pentium-D
1 GB/node
280+ nodes



FY 2007 Cluster

Goals:

Formal goal: 2.9 Teraflop/s sustained <asqtad,dwf>,
& deploy by June 30

Science goal: get the most capacity that $1.4M can buy,
& deploy as fast as possible

Best Value RFP Process

- Explicitly described a cluster of ~800 processors with infiniband fabric (to guide vendors toward good solution)
- Allowed for ANYTHING
  (specifically wanted to all room for a BG/L proposal to compete)
- Included anisotropic clover as one of three benchmarks
- Chose local volumes corresponding to anticipated real jobs
  (not artificial "best performance" numbers)

Proposals

Single node performance, showing breadth of potential solutions:

action: local vol:	asqtad 12^4	clover 12x6x6x32	dwf 28x8x8x32	bandwidth per core	<as,cl,dwf> $/MF
1 dual core Xeon 2.66	3530	2487	4800	1500	$0.51
1 quad core Xeon 2.33	2540	3520	4400	715 - 900	$0.78
2 dual core Xeon 2.66	4630	4800	7491	1325	$0.56
2 quad core Xeon 233	4200	12000	5600	400	$0.51
2 dual core AMD 2.6	4900	4040	6560	1750	$0.49
amd w/ 1 GB dimms (not 512)	5140	4400	6950	1975	

Winning Proposal

Vendor: Koi ("whitebox", same supplier as Kaon at FNAL)

Node:

- dual cpu, dual core AMD 2218, 2.6 GHz
- DDR (20g) infiniband, 18:6 leaf switch oversubscription
- ASUS KFN4 motherboard; IPMI 2.0

Interesting Option:

Upgrade to quad core Opterons at steeply discounted price

- doubles number of cores
- doubles SSE issue rate / cycle (to match Intel)
- 2 MB L3 shared cache, + ½MB L2/core (effectively doubles cache)
- same power envelope (2.1 GHz vs 2.6 GHz)

Projecting Quad Core Performance

Reasoning:

- dual - dual core Xeons get the most flop/s per MB/sec of memory bandwidth (streams triad) – i.e. Xeons have enough peak flop/s to consume bandwidth
- Raw flops of quads will be 3x faster than Opteron duals

  (2x cores, 2x issue rate, ¾ clock speed)
- With this increase, Opteron will also be memory bound, like the Xeons
- Opterons have 50% - 60% more memory bandwidth than Xeons
- Therefore, quad cores should have ~ 50% performance boost
  20% cost, and delay of ~3 months on just 20% of funds (clear win)

In addition:

- Hyper transport bus & Opteron cache protocols are better at multi-threading, yielding additional architectural advantages

Modifying the proposal

- Preliminary decision to hold back 20% of the funds to be able to do the quad core upgrades

  (no real impact, since continuing resolution kept those funds from Jlab anyway)

  - increase memory / node to 4 GBytes
    (higher density, better performance for some reason)
  - decrease node count to 400
- In May, evaluate quad core chips in one of our nodes, verify that we get price/performance boost
- Order quads OR order additional 20% nodes
- Quad chips might take 3 months to receive (high initial demand); install as late as September

7n Timeline

April:

machine installation (mid month for racks, later for long cables)

May:

friendly user mode on 400 dual-duals

separate PBS server & queue, for 64 bit O/S

2 of the nodes used as 64 bit interactive & build/test
with 8 GB / node

June:

production on 400 dual-duals

July:

convert 6n to 64 bit; decommission old interactive nodes

September:

rolling outages to upgrade to quads

Optimistic Result: 2.9 TFlop/s deployed in FY 2007

File Server Upgrade

Currently have 5 servers from 1 to 4 years old;

total of 15 TBytes, but reliability decreasing.

Disk Allocation is currently by project; each server = 1-3 projects

(avoids need for disk management software & manpower)
But: some projects need more than 5 TBytes (largest server)

Next step: 3x disk capacity increase to match 3x computing performance increase.

Option 1: **many small** terabyte servers + software (like dCache)

Option 2: **a few larger, faster** servers

a single project still lives on single server (easy "flat" namespace management); servers cost more, but much less manpower expense; **single stream performance advantage** (good for bursty load)

Mid Range File Servers

Goals:

- >200 MB/sec streaming single file into head node of job
  (avoid need for parallel file system, re-writing aplications)
- Feed data in via infiniband fabric, avoiding completely the cost of a gigE fabric
  - saved $20K by using less expensive fast ethernet
  - achieve bandwidth goals (impossible via gigE)

Presently Evaluating

- Sun's Thumper - Sun Fire X4500: zfs, PCI-X infiniband HCA
  - 18 TBytes / box (could use one per large project)
  - 550 MB / sec disk to memory!!!
- Agami's AIS6119
  - 12 gigE links; no direct IB connectivity
  - IB gateway via 4 trunked gigE connections (router, or use one node)
- Others being considered

Infrastructure Upgrade Summary

- **Disk Cache**
  - 15 TB going to **45+ TB**
- **Wide Area Networking**
  - Upgraded this past year to **10g**
- **Local Area Networking**
  - Bandwidth: file server to silo going to **10g**
- **Power**
  - Over next 2 years add 1 Megawatt UPS
  - Over next 5 years add equivalent cooling capacity

QUESTIONS ?