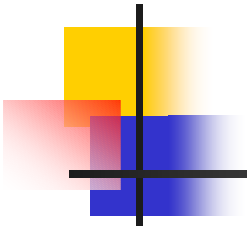
Lattice QCD Computing Project, and Fermilab Status

Don Holmgren
All-Hands Meeting
Fermilab
April 6-7, 2006



Outline

- Lattice QCD Computing Project
  - Scope
  - Goals and Milestones
  - Personnel/Budgets
- Fermilab Status
  - Hardware
  - Statistics
  - I/O
  - Computer Security
  - User Support

Lattice QCD Project - Scope

- Four years: Oct 1, 2005 → Sept 30, 2009
- Operation (administration, hardware maintenance, site management) of:
  - US QCDOC
  - SciDAC Clusters
    - FNAL: QCD, Pion clusters
    - JLab: 2M, 3G, 4G clusters
  - New Systems
    - FNAL Kaon
    - JLab 6N

Lattice QCD Project - Scope

- Purchase and deploy new systems
  - 2006: JLab 6N, FNAL Kaon
  - 2007 and beyond: at most one new system a year
- Not in scope:
  - Software development
  - Scientific software support

Lattice QCD Project - Mechanics

Funded via OMB Exhibit 300 process:

- Science case – how does proposed project fit agency strategic goals and President's Management Agenda?
- Business case – compare this investment to alternatives, determine best ROI, NPV
- Baseline budget and schedule
- Performance against baseline budget and schedule
- Project Management
- Acquisition Strategy
- Risk Management
- Performance Goals
- Computer Security

Lattice QCD Project - Mechanics

The OMB Exhibit 300 calendar:

- Submitting to DOE now for FY08, which starts Oct 1, 2007
- DOE grades and ranks submissions to determine investment portfolio
- DOE submits to OMB in September
- “Passback” to agencies in Nov/Dec – agencies modify submissions that are at risk
- Incorporated in President’s Budget (February)

Lattice QCD Project – Goals and Milestones

Milestones for each year:

- "Delivered Tflops-yrs"
  - Defined as available capacity expressed as average of DWF and asqtad inverter performance
  - "1 year" = 8000 hours
- "Deployed Tflops"
  - Defined as incremental capacity brought online, expressed as average of DWF and asqtad inverter performance
- External review of progress and next year's planned deployment by end of June

Lattice QCD Project – Goals and Milestones

Year	Tflops-yrs Delivered	Tflops-yrs Deployed
FY2006	6.2 (QCDOC = 4.2)	2.0
FY2007	9	3.1
FY2008	12	4.2
FY2009	15	3.0

Lattice QCD Project - Budget

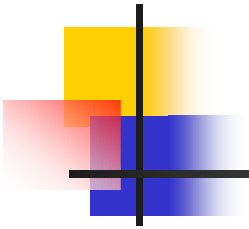
Funded operations at BNL, JLab, and FNAL in 2006 (fractions of a person):

Project + Base/SciDAC	sysadmin / technician	Scientific software & user support	site management
BNL	0.75	0.5	0.25
FNAL	1.75	0.5	0.25
JLab	0.65	0.5	0.25

Sysadmin/tech totals increase to 3.00 in 2007, 3.25 in 2008

Lattice QCD Project - Budget

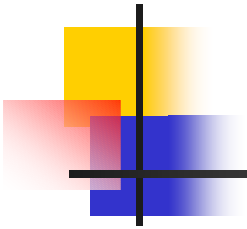
Year	Personnel	Equipment
FY2006	$650K	$1,850K
FY2007	$885K	$1,615K
FY2008	$955K	$1,545K
FY2009	$1030K	$670K



Lattice QCD Computing Project

Points of contact at the labs:

- Project Manager – Don Holmgren, FNAL
- Assoc. Project Mgr – Bakul Banerjee, FNAL
- BNL Site Mgr – Tom Schlagel
- FNAL Site Mgr – Amitoj Singh
- JLab Site Mgr – Chip Watson
- Metafacility Operations Mgr – Bálint Joó, JLab



Fermilab Status - Hardware

Current clusters:

- "QCD"
  - 127 nodes, 2.8 GHz Pentium 4, 1 GB memory
  - Myrinet (128th connection is to head node)
  - Online since June 2004
  - Performance (64 node runs):
    - DWF: 1400 Mflops/node
      Ls=16, average of 32x8x8x8 and 32x8x8x12
    - Asqtad: 1017 Mflops/node
      14^4 local lattice/node
    - Total capacity: ~ 150 Gflops

Fermilab Status - Hardware

Current clusters (cont'd):

- "Pion"
  - 520 nodes, 3.2 GHz Pentium 640, 1 GB memory
  - Infiniband
  - Full cluster online in December 2005
  - Performance (64 node runs):
    - DWF: 1729 Mflops/node
      Ls=16, average of 32x8x8x8 and 32x8x8x12
    - Asqtad: 1594 Mflops/node
      14^4 local lattice/node
    - Total capacity: ~ 860 Gflops

Fermilab Status - Hardware

Next cluster: “Kaon”

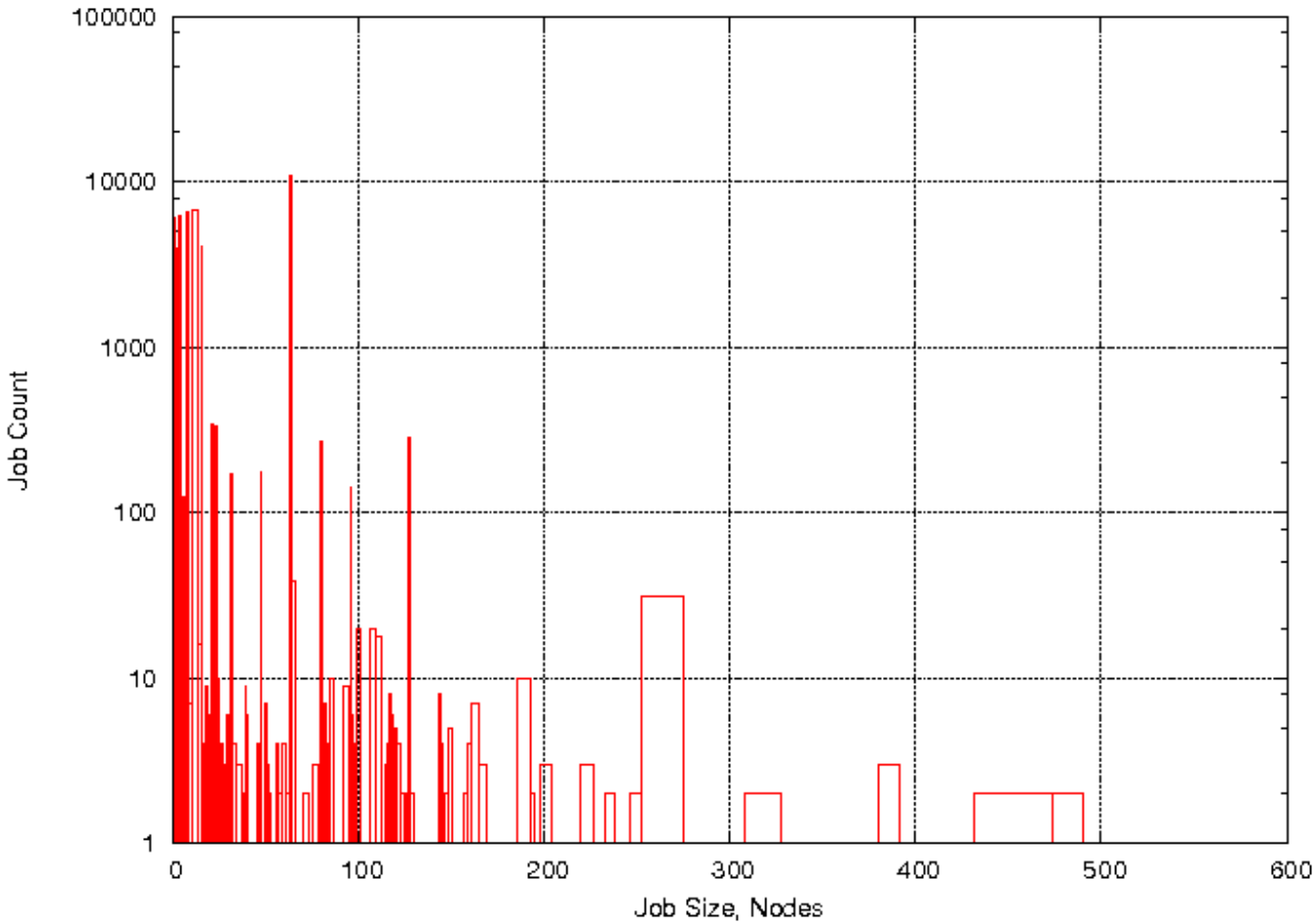
- Request for Proposals (RFP) released April 3
- Vendors must bid:
  - Dual processor, Intel (“Dempsey” dual core processors plus fully buffered DIMMs) or AMD (Opteron 270 or faster dual core processors)
  - Infiniband interconnect
- Bids due April 28
- Award May 17
- Deliveries in July, “Friendly Users” in August
- ~ 450 nodes, ~ 4200 Mflops/node, ~ 1.9 Tflops total

Fermilab Status - Statistics

- Since April 1, 2005:
  - Users submitting jobs:
    29 LQCD, 5 administrators
  - 104,000 jobs (90,333 multi-node)
  - 3.57 million node-hours (not including ~ 0.5 million node-hours on Pion, June-Sept, for configuration generation)
- Since start of project (Oct 1, 2005):
  - 3020 Tflops-hrs delivered (0.4 Tflops-yrs)
  - 92.8% uptime (excluding December)

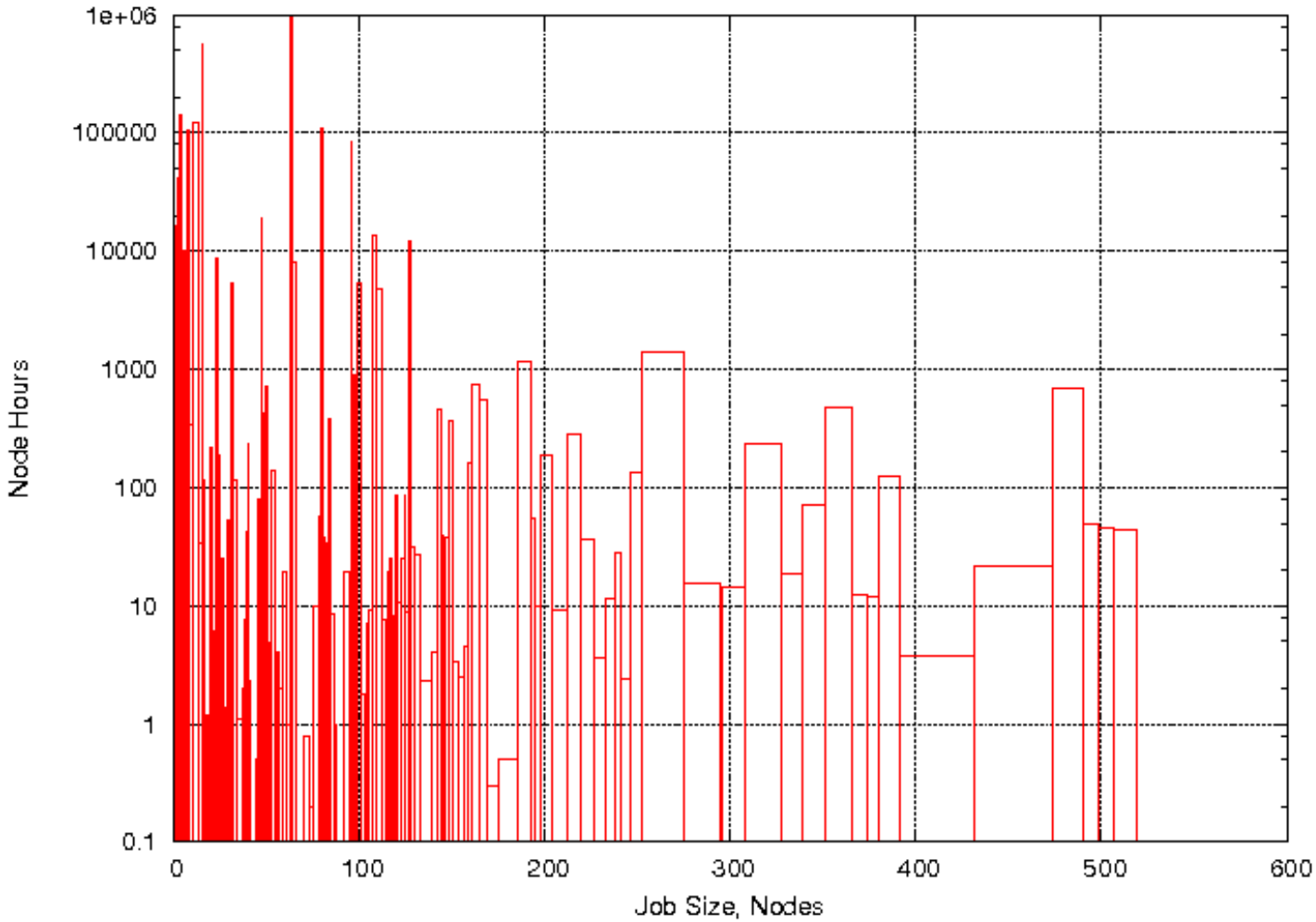
Fermilab Status - Statistics

Job distribution by node count:



Fermilab Status - Statistics

Job distribution by total node hours:



Fermilab Status – Mass Storage

"Enstore"

- Robotic, network-attached tape drives
- Files are copied using "encp src dest"
- 15 MB/sec transfer rate per stream
- Currently using ~110 Tbytes of storage
- Cost:
  - $375 for 1 Tbyte of tape (5 tapes)
  - $1000 per tape slot in the robot
  - 1 Tbyte of tape + slots = ~ 0.5 dual CPU node

Fermilab Status – Mass Storage

"Public" dCache (/pnfs/lqcd/)

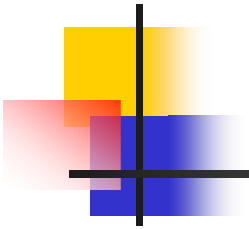
- Disk layer in front of Enstore tape drives
- All files written end up on tape ASAP
- Files are copied using "dccp src dest"
  - Pipes allowed
  - Also, direct I/O allowed (posix/ansi)
- On writing, hides latency for tape mounting and movement
- Can "prefetch" files from tape to disk in advance

Fermilab Status – Local Storage

Disk RAID arrays attached to head node

- /data/raid*x*, *x* = 1-6, total ~ 7 Tbytes
- Also, /project (visible from worker nodes)
- Data files must be copied by user jobs via rcp to/from head node
- Performance is limited:
  - By network throughput to/from head node
  - By load on head node
- Cost: $2200 for 1 Tbyte (RAID'd)
  - 1 Tbyte = ~ 0.8 dual CPU node

Fermilab Status – Local Storage



"Volatile" dCache (/pnfs/volatile/)

- Consists of multiple disk arrays attached to "pool nodes" connected to Infiniband network
- No connection to tape storage
- Provides large "flat" filesystem
  - Users don't have to keep track of /data/raid*x* paths
- Provides high aggregate read/write rates when multiple jobs are accessing system
- Supports file copies (via dccp) and direct I/O (via libdcap: posix/ansi style calls)
- About 10 Tbyte available

Fermilab Status – I/O

User requirements needed for budgets and planning:

- Local disk storage (dCache and /data/raid*x*)
  - Proportional to machine allocations
- Temporary tape storage (and duration)
- Archival tape storage

Fermilab Status - Security

- Kerberos
  - Strong authentication (instead of ssh)
  - Use kerberos clients or cryptocards
  - Linux, Windows, Mac support
- Transferring files
  - Tunnel scripts – provide “one hop” transfers to/from BNL and JLab
  - See web pages for examples
- Annual training/testing
  - Required at all three sites

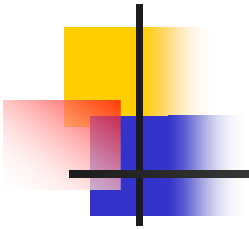
Fermilab Status – User Support

- Mailing lists
  - Lqcd-admin@fnal.gov
  - Lqcd-users@fnal.gov
- Transition to “help” tickets by summer
  - Submit via web form or e-mail
  - Admin’s and users will be able to track responses
- Level of support
  - 10 x 5, plus best effort off-hours
  - Increasing automation (remote reboots)

Fermilab Status – User Support

Fermilab points of contact:

- Don Holmgren, djholm@fnal.gov
- Amitoj Singh, amitoj@fnal.gov
- Kurt Ruthmansdorfer, kurt@fnal.gov
- Jim Simone, simone@fnal.gov
- Jim Kowalkowski, jbk@fnal.gov
- Paul Mackenzie, pbm@fnal.gov



Questions?