Cluster Development at Fermilab

Don Holmgren

All-Hands Meeting

Jefferson Lab

June 1-2, 2005

Outline

- Status of current clusters
- New cluster details
  - Processors
  - Infiniband
  - Performance
- Cluster architecture (I/O)
- The coming I/O crunch

Current Clusters (1)

- QCD80
  - Running since March 2001
  - Dual 700 MHz Pentium III
  - In June 2004, moved off of Myrinet and onto a gigE switch
  - 67 nodes still alive
  - Now used for automated perturbation theory

Current Clusters (2)

- nqcd
  - Running since July 2002
  - 48 dual 2.0 GHz Xeons, originally Myrinet
  - 32 nodes now running as prototype Infiniband cluster (since July 2004)
  - Now used for Infiniband testing, plus some production work (accelerator simulation - another SciDAC project)

Cluster Status (3)

- W
  - Running since January 2003
  - 128 dual 2.4 GHz Xeons (400 MHz FSB)
  - Myrinet
  - One of our two main production clusters
  - To be retired October 2005 to vacate computer room for construction work

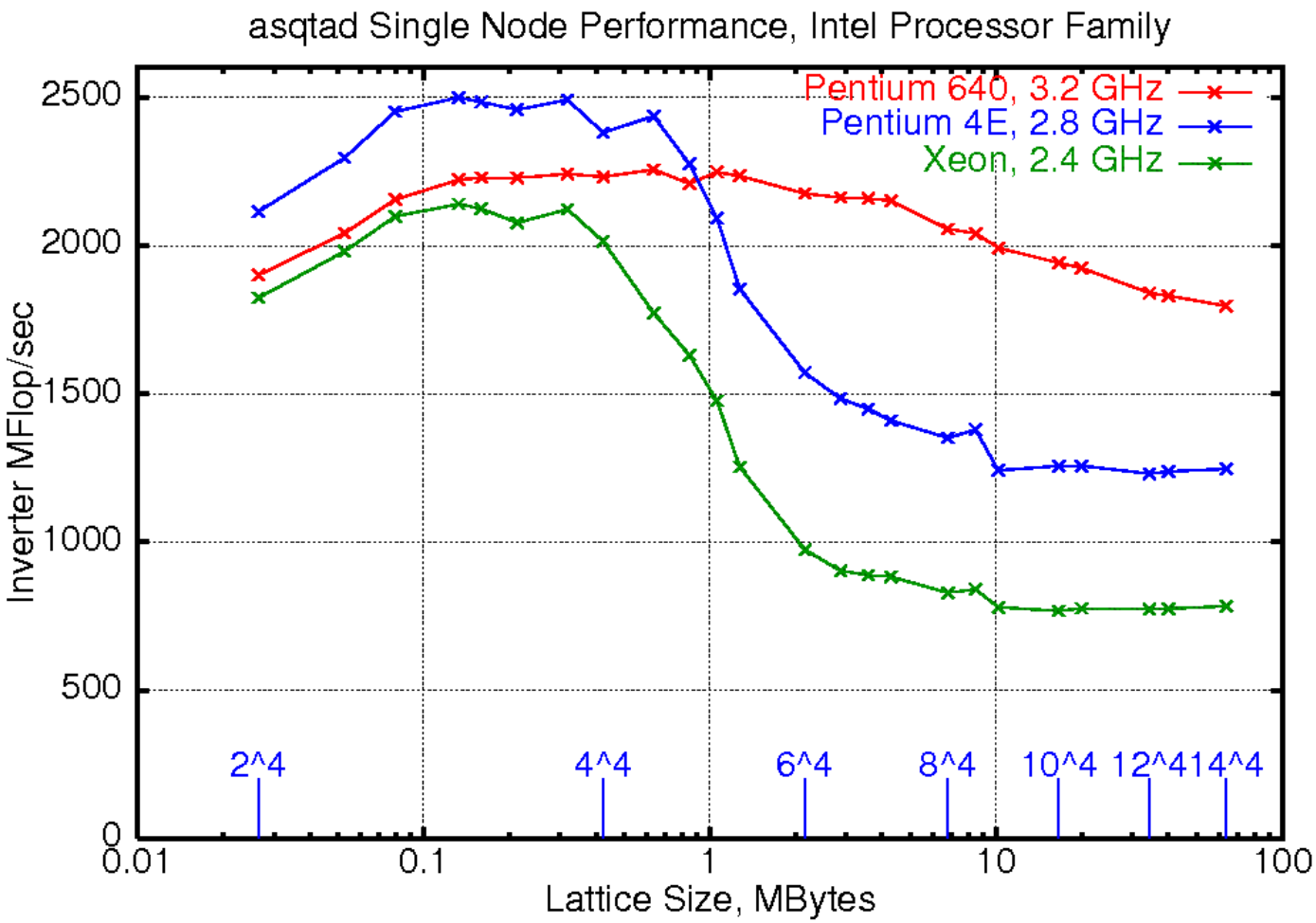
Cluster Status (4)

- qcd
  - Running since June 2004
  - 128 single 2.8 GHz Pentium 4 “Prescott”, 800 MHz FSB
  - Reused Myrinet from qcd80, nqcd
  - PCI-X, but only 290 MB/sec bidirectional bw
  - Our second main production cluster
  - $900 each
    - ~ 1.05 Gflop/sec asqtad (14^4)
    - 1.2-1.3 Gflop/sec DWF

Cluster Status (5)

- pion
  - Being integrated now – production mid-June
  - 260 single 3.2 GHz Pentium 640, 800 MHz FSB
  - Infiniband using PCI-E (8X slot), 1300 MB/sec bidirectional bw
  - $1000/node + $890/node for Infiniband
    - ~ 1.4 Gflop/sec asqtad (14^4)
    - ~ 2.0 Gflop/sec DWF
  - Expand to 520 cpus by end of September

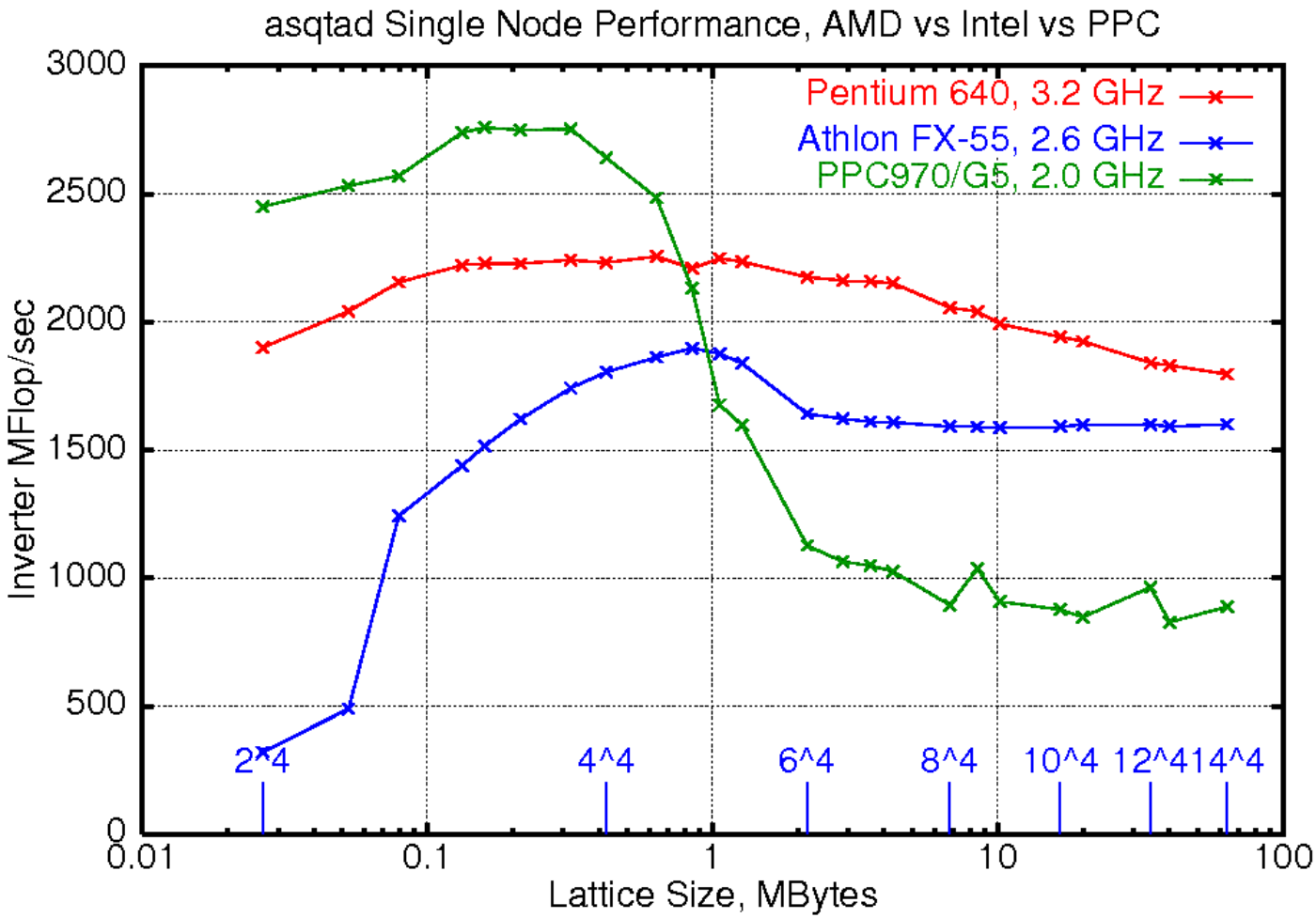
Details – Processors (1)



Processors on W, qcd, and pion – cache size and memory bus:

- W: 0.5 MB 400 MHz FSB
- qcd: 1.0 MB 800 MHz FSB
- pion: 2.0 MB 800 MHz FSB

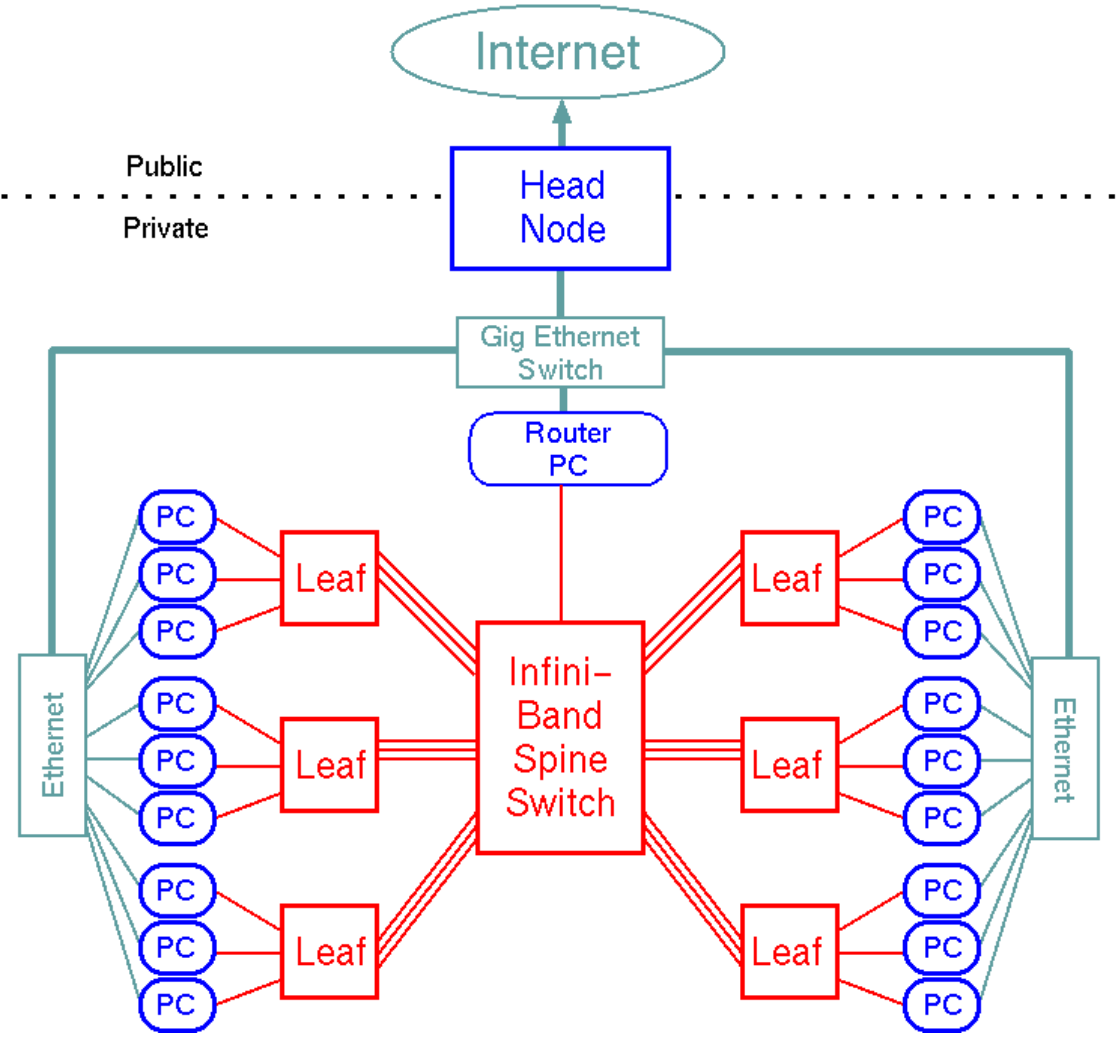
Details – Processors (2)



Processor Alternatives:

- PPC970/G5 1066 MHz FSB (split bus)
- AMD FX-55 is the fastest Opteron
- Pentium 640 best price / performance for pion cluster

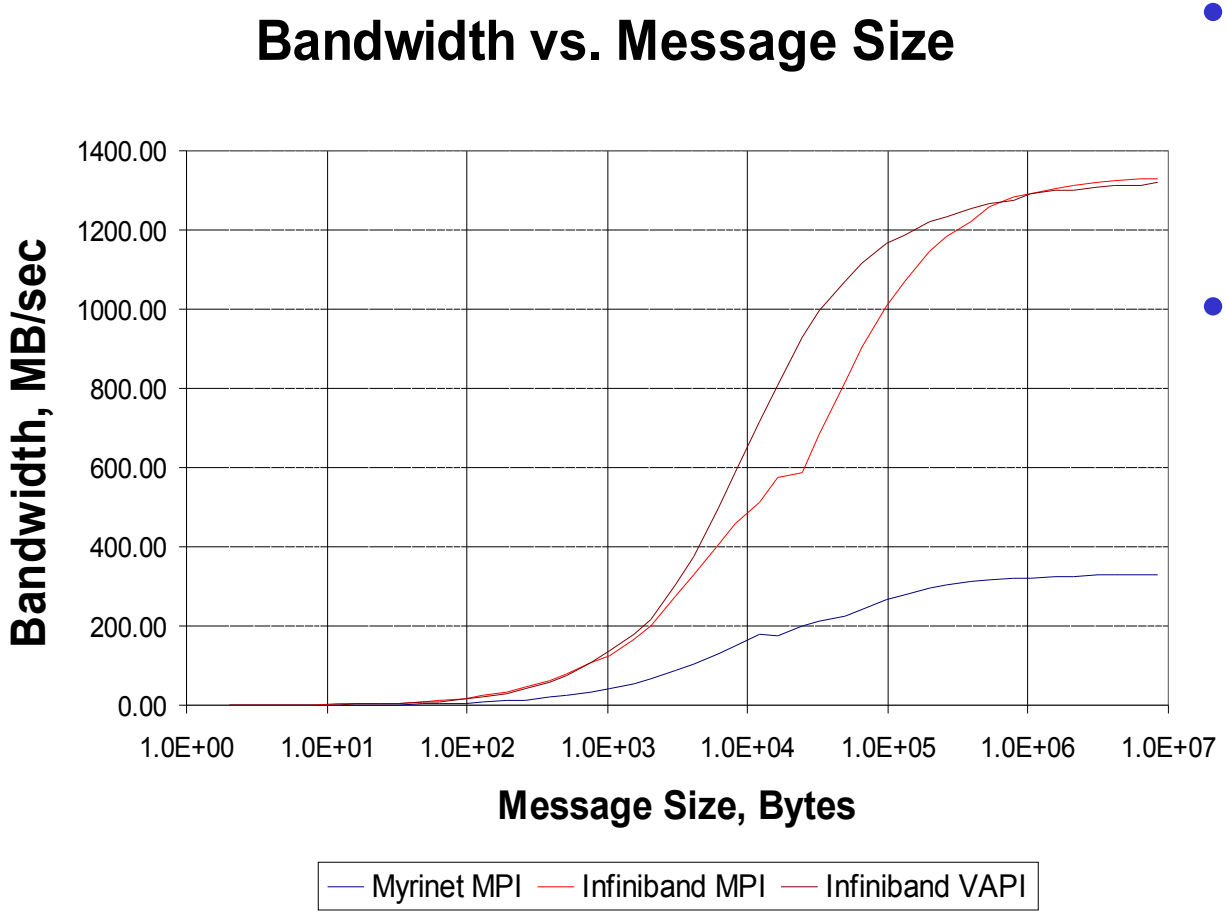
Details – Infiniband (1)



Infiniband cluster schematic

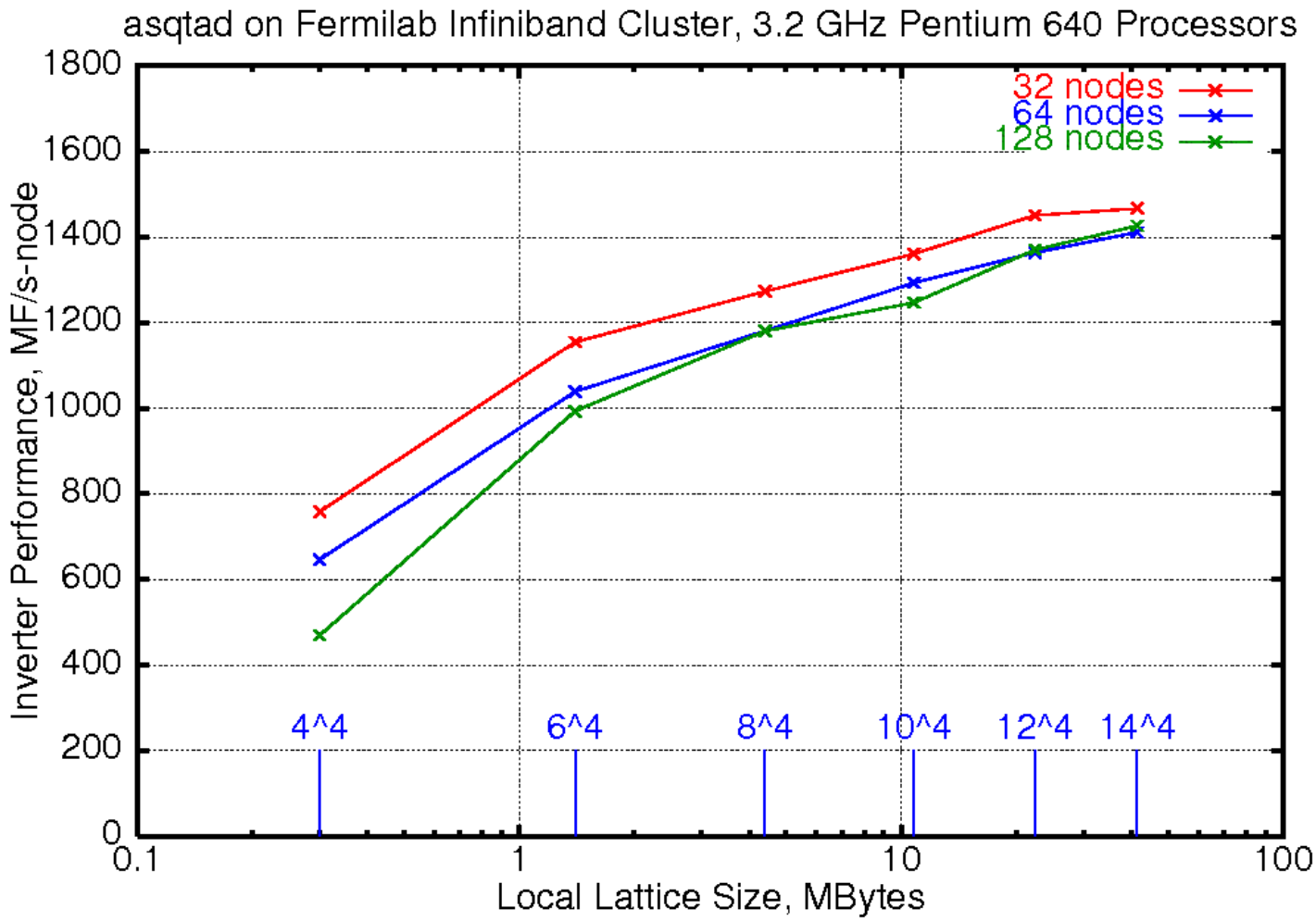
- "Oversubscription" – ratio of nodes:uplinks on leaf switches
- Increase oversubscription to lower switch costs
- LQCD codes are very tolerant to oversubscription
- pion uses 2:1 now, will likely go to 4:1 or 5:1 during expansion

Details – Infiniband (2)



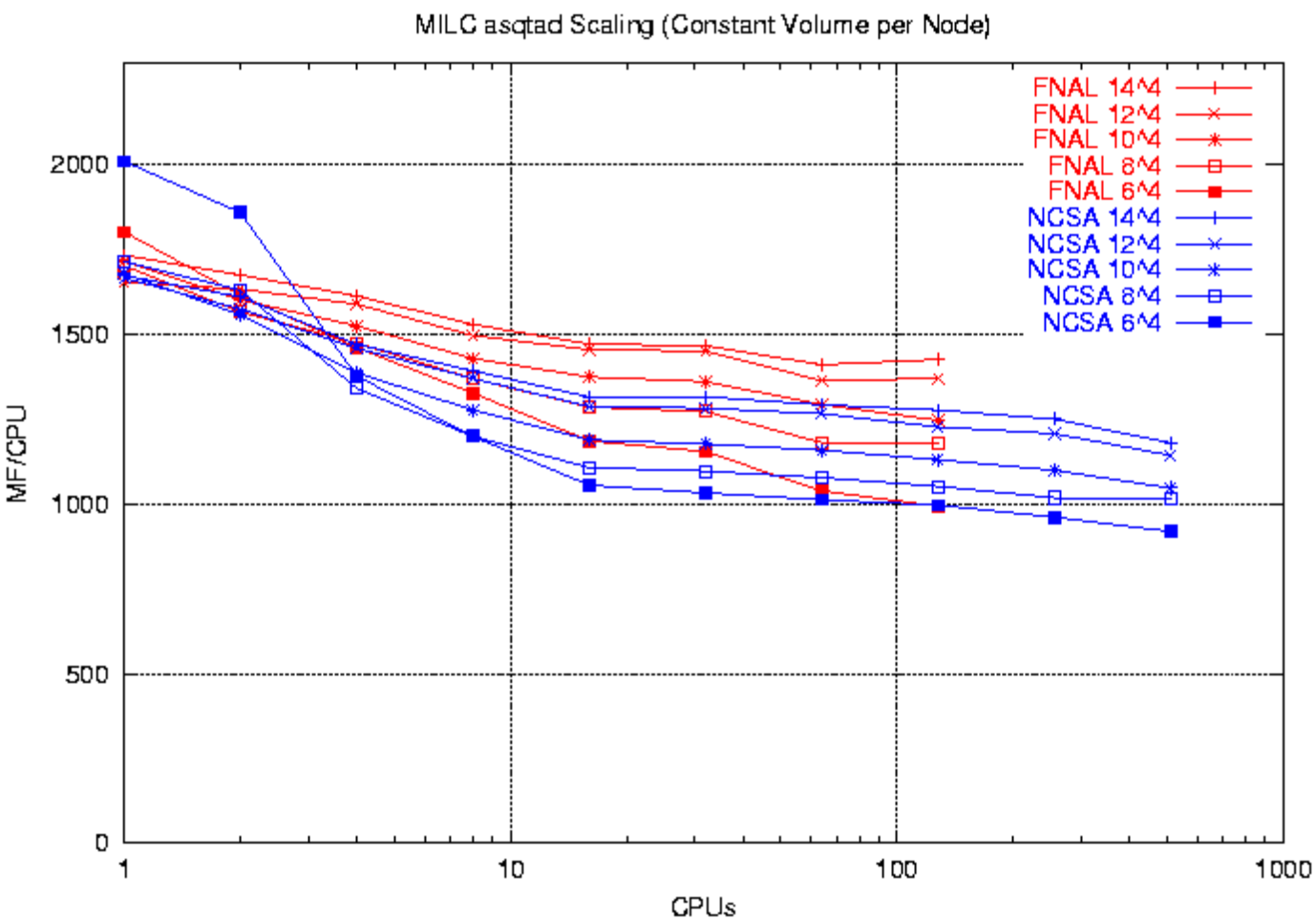
- Netpipe bidirectional bw data
  - Myrinet: W cluster
  - Infiniband: pion
- Native QMP implementation (over VAPI instead of MPI) would boost application performance

Details – Performance (1) - asqtad



Weak scaling curves on new cluster

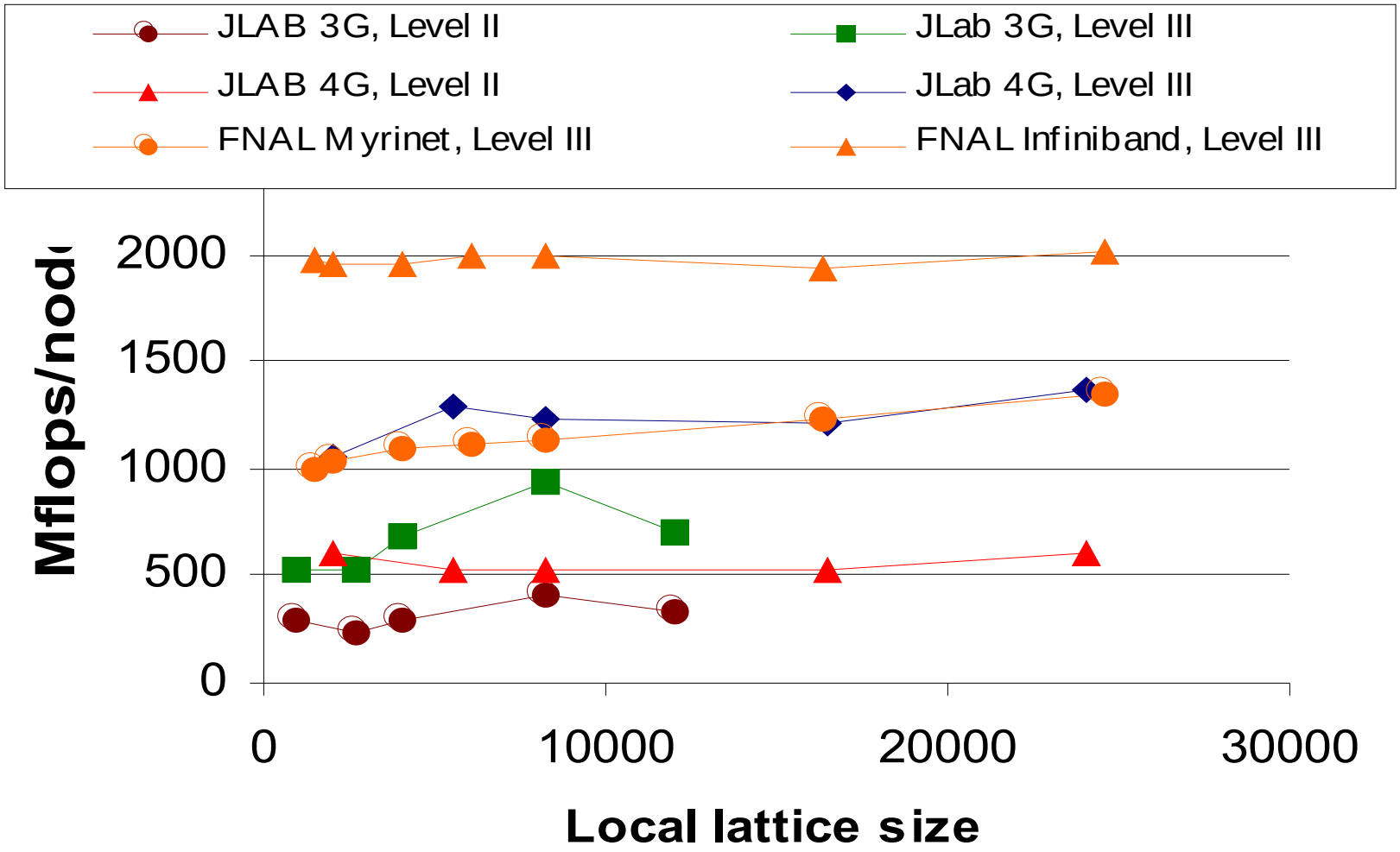
Details – Performance (2) - asqtad



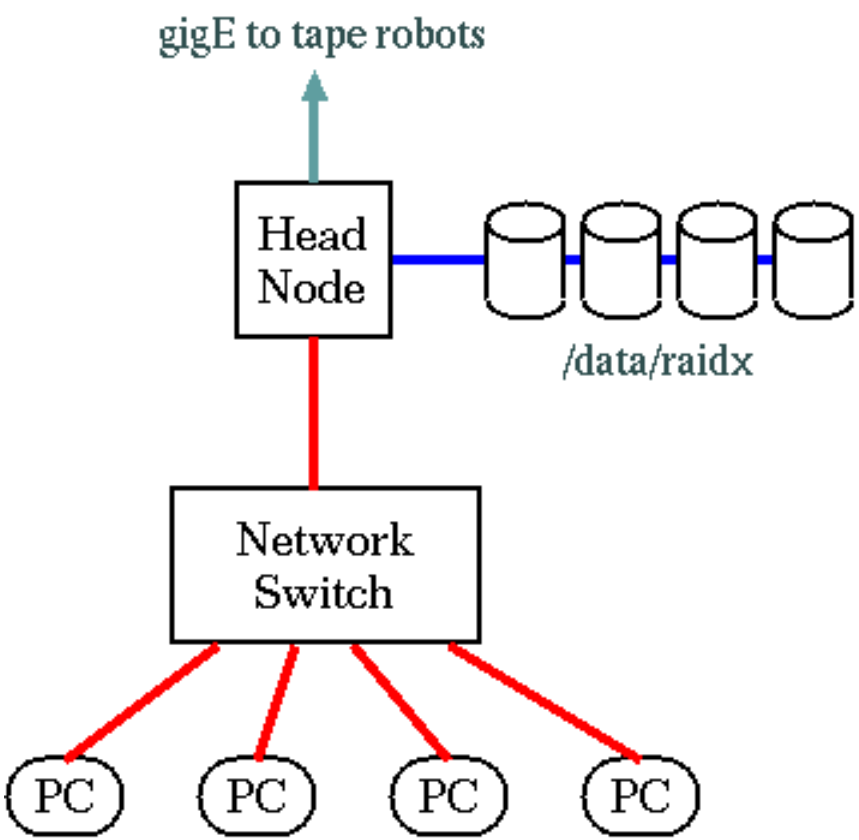
Weak scaling on NCSA "T2" and FNAL "pion"

- T2: 3.6 GHz Dual Xeon, Infiniband on PCI-X, standard MILC V6
- pion: MILC using QDP

Details – Performance (3) - DWF

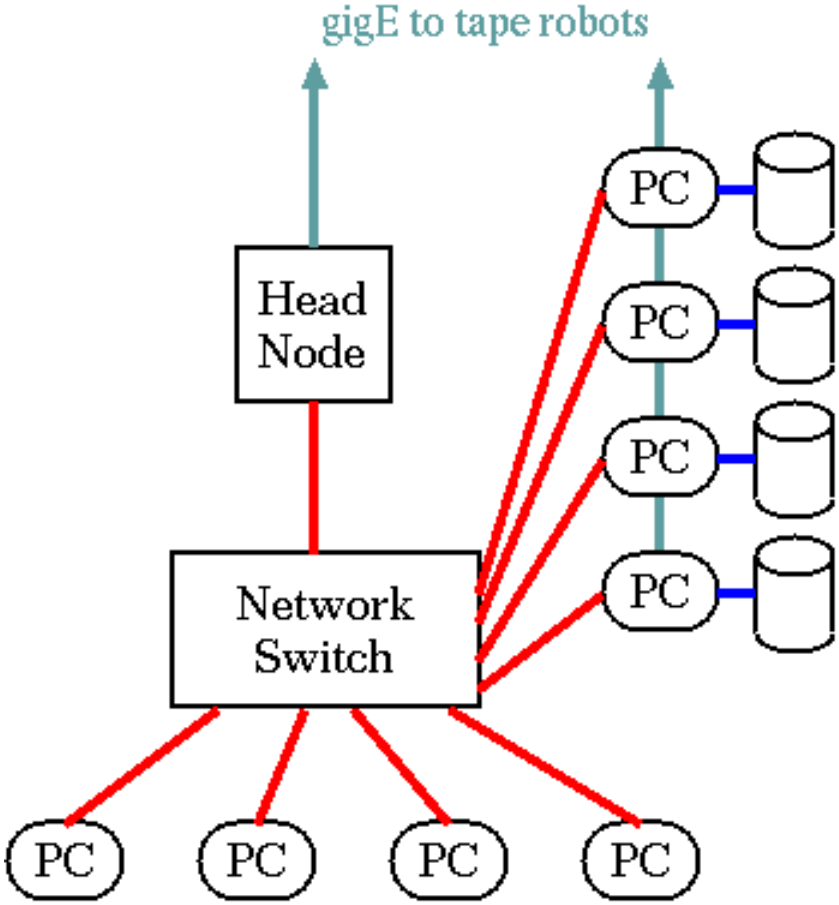


Cluster Architecture – I/O (1)



- Current architecture:
  - Raid storage on head node: /data/raid1, /data/raid2, etc.
  - Files moved from head node to tape silo (encp)
  - User's jobs stage data to scratch disk on worker nodes
- Problems:
  - Users keep track of /data/raidx
  - Users have to manage storage
  - Data rate limited by performance of head node
  - Disk thrashing – use fcp to throttle

Cluster Architecture – I/O (2)

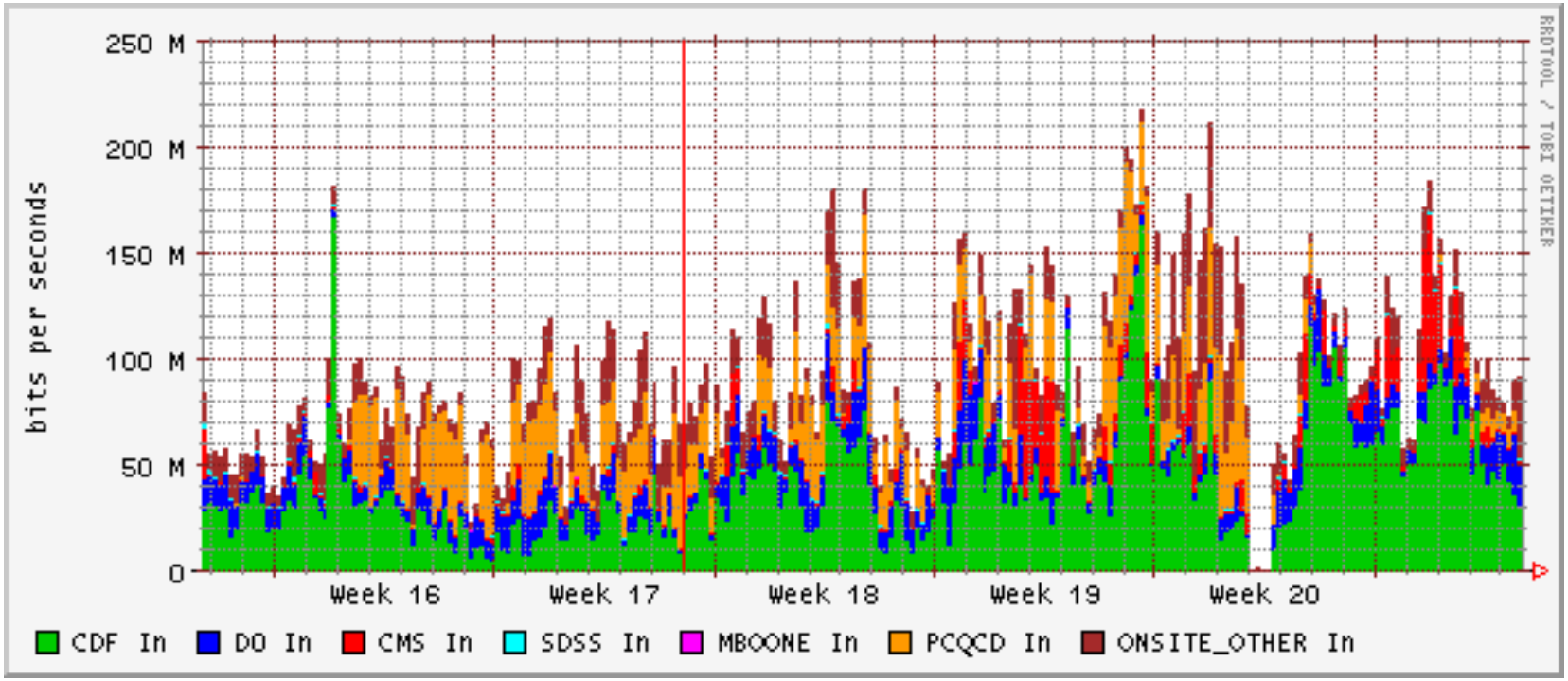


- New architecture:
  - Data storage on dCache nodes
  - Add capacity, bw by adding spindles and/or buses
  - Throttling on reads
  - Load balancing on writes
  - Flat directory space (/pnfs/lqcd)
- User access to files:
  - Shell copy (dccp) to worker scratch disks
  - User binary access via popen(), pread(), pwrite() [#ifdefs]
  - User binary access via open(), read(), write() [$LD_PRELOAD]

Cluster Architecture Questions

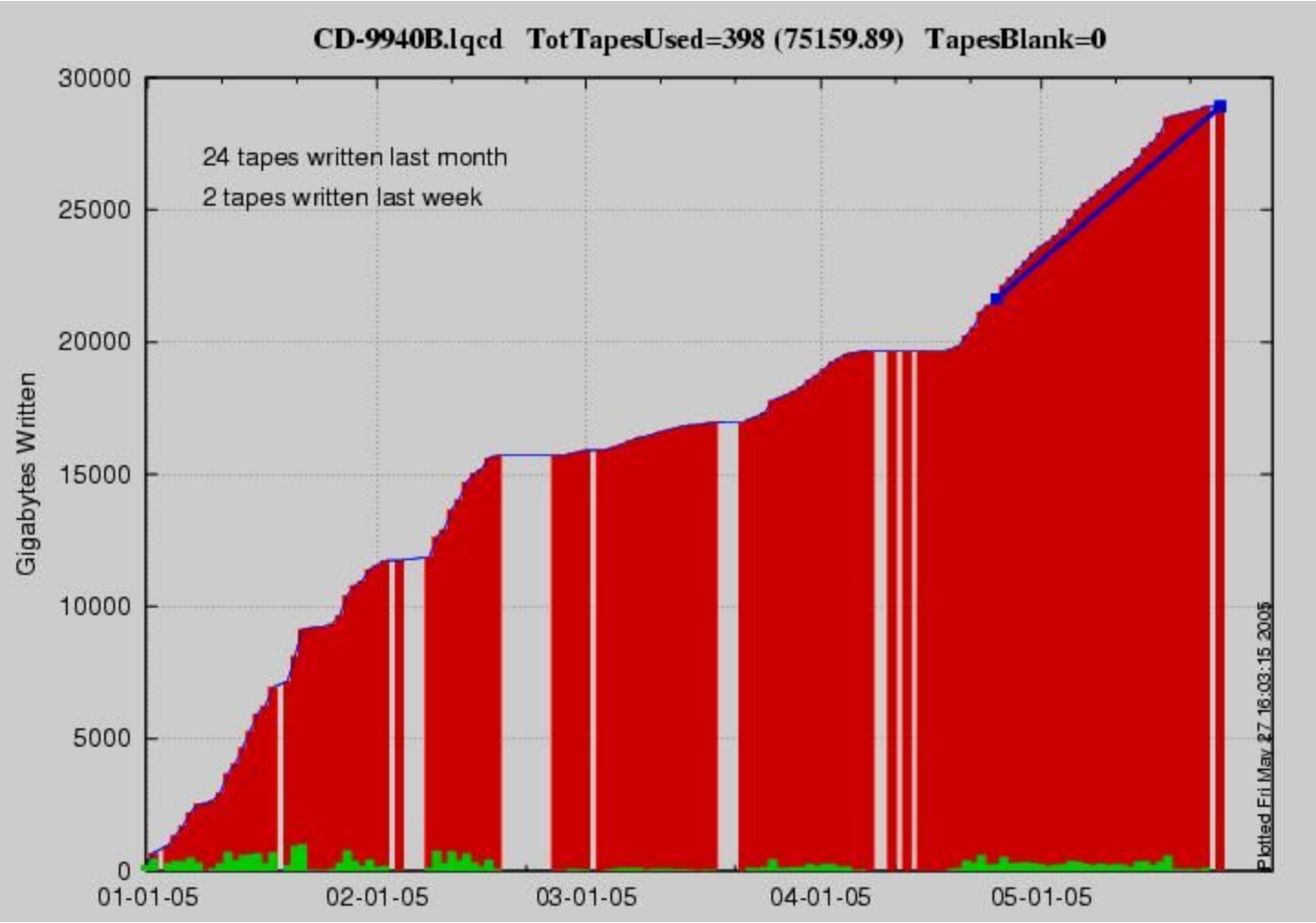
- Now:
  - Two clusters, W and qcd
  - Users optionally steer jobs via batch commands
  - Clusters are binary compatible
- After Infiniband cluster (pion) comes online
  - Do users want to preserve binary compatibility?
    - If so, would use VMI from NCSA
  - Do users want 64-bit operating system?
    - > 2 Gbyte file access without all the #defines

The Coming I/O Crunch (1)



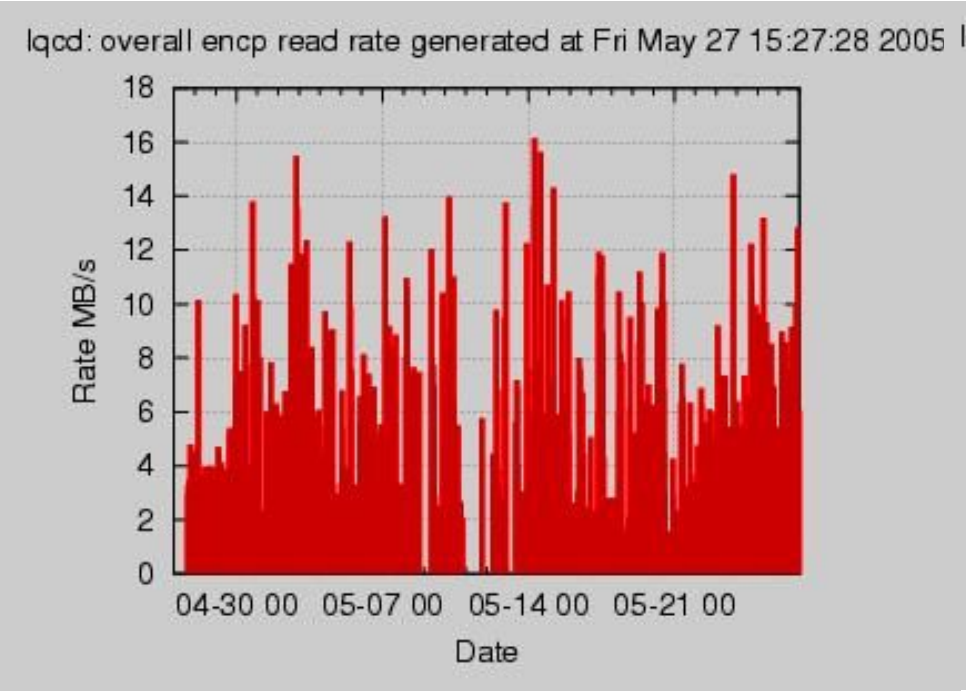
LQCD already represents a noticed fraction of incoming data into Fermilab

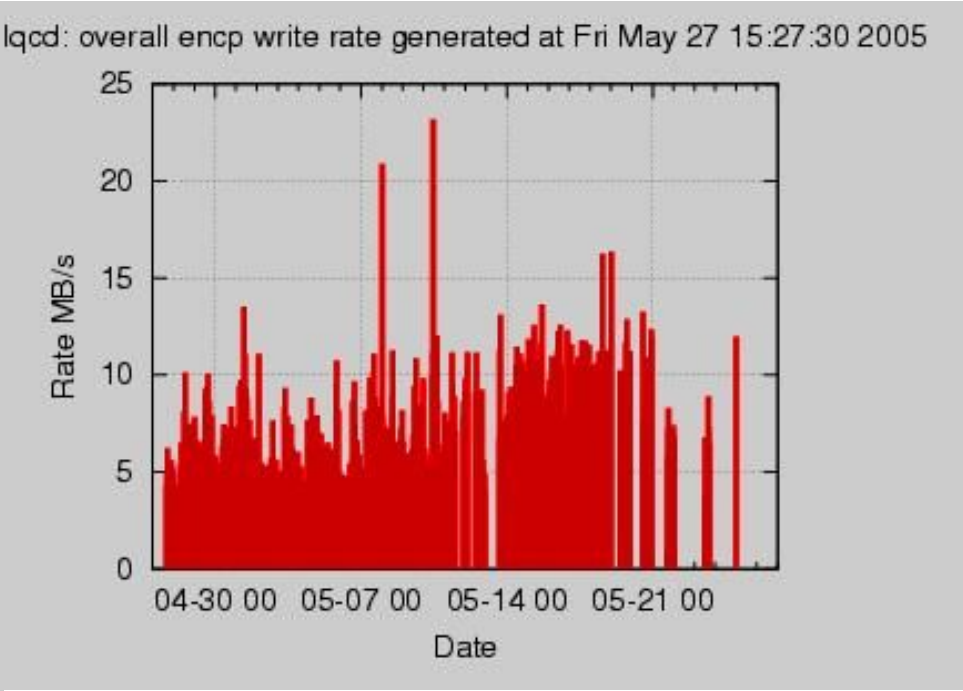
The Coming I/O Crunch (2)



Propagator storage is already consuming 10's of Tbytes of tape silo capacity

The Coming I/O Crunch (3)





- Analysis is already demanding 10+ MB/sec sustained I/O rates to robots for both reading and writing (simultaneously)

The Coming I/O Crunch (4)

- These storage requirements were presented last week to the DOE reviewers:
  - Heavy-light analysis (S. Gottlieb): 594 TB
  - DWF analysis (G. Fleming): 100 TB
- Should propagators be stored to tape?
  - Assume yes, since current workflow uses several independent jobs
  - Duration of storage? Permanent? 1 Yr? 2 Yr?
  - Need to budget: $200/Tbyte now, $100/Tbyte in 2006 when LTO2 drives are available

The Coming I/O Crunch (5)

- What I/O bandwidths are necessary?
  - 1 copy in 1 year of 600 Tbytes → 18.7 MB/sec sustained for 365 days, 24 hours/day!
  - Need at least 2X this (write once, read once)
  - Need multiple dedicated tape drives
  - Need to plan for site-local networks (multiple gigE required)
- What files need to move between sites?
  - Configurations: assume O(5 Tbytes)/year QCDOC ←→ FNAL, similar for JLab (and UKQCD?)
  - Interlab network bandwidth requirements?

Questions?