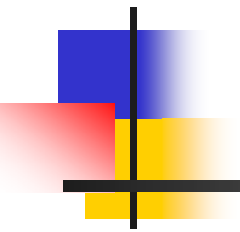


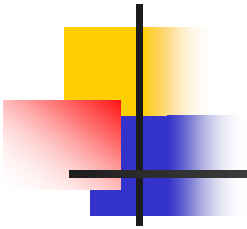
Fermilab Status

Don Holmgren

USQCD All-Hands Meeting

Fermilab

May 14, 2009



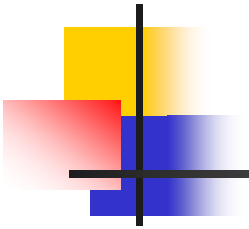
Outline

- Current Hardware
- FY10/FY11 Deployment
- Storage/Filesystems
- Statistics
- User Authentication
- User Support

Hardware – Current Clusters

Name	CPU	Nodes	Cores	Network	DWF	Asqtad	Online
QCD	Single 2.8 GHz Pentium 4	127	127	Myrinet 2000	1400 MFlops per Node	1017 MFlops per Node	June 2004 0.15 TFlops
Pion	Single 3.2 GHz Pentium 640	518	518	Infiniband Single Data Rate	1728 MFlops per Node	1594 MFlops per Node	June 2005 / Dec 2005 0.86 TFlops
Kaon	Dual 2.0 GHz Opteron 240 (Dual Core)	600	2400	Infiniband Double Data Rate	4696 MFlops per Node	3832 MFlops per Node	Oct 2006 2.56 TFlops
J/ψ	Dual 2.1 GHz Opteron 2352 (Quad Core)	856	6848	Infiniband Double Data Rate	10061 MFlops per Node	9563 MFlops per Node	Jan 2009 / Apr 2009 8.40 TFlops

Time on QCD will not be allocated this year, but the cluster will be available.



Hardware

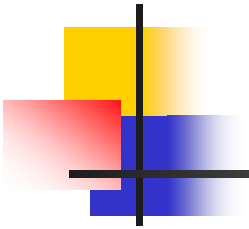
- Pion/Kaon
  - Run 64-bit version of Scientific Linux 4.x
  - Access via *kaon1.fnal.gov*
  - Run same binaries on both clusters
- JPsi
  - Runs 64-bit version of Scientific Linux 4.x
  - Access via *jpsi1.fnal.gov*
  - Binary compatible with Pion / Kaon

Hardware

- QCD
  - Runs 32-bit version of Scientific Linux 4.1, so large file support (files > 2.0 Gbytes in size) requires the usual *#define*'s
  - Access via *lqcd.fnal.gov*
  - Not binary compatible with Pion / Kaon / Jpsi
  - Will be decommissioned sometime in 2010

Hardware – GPUs

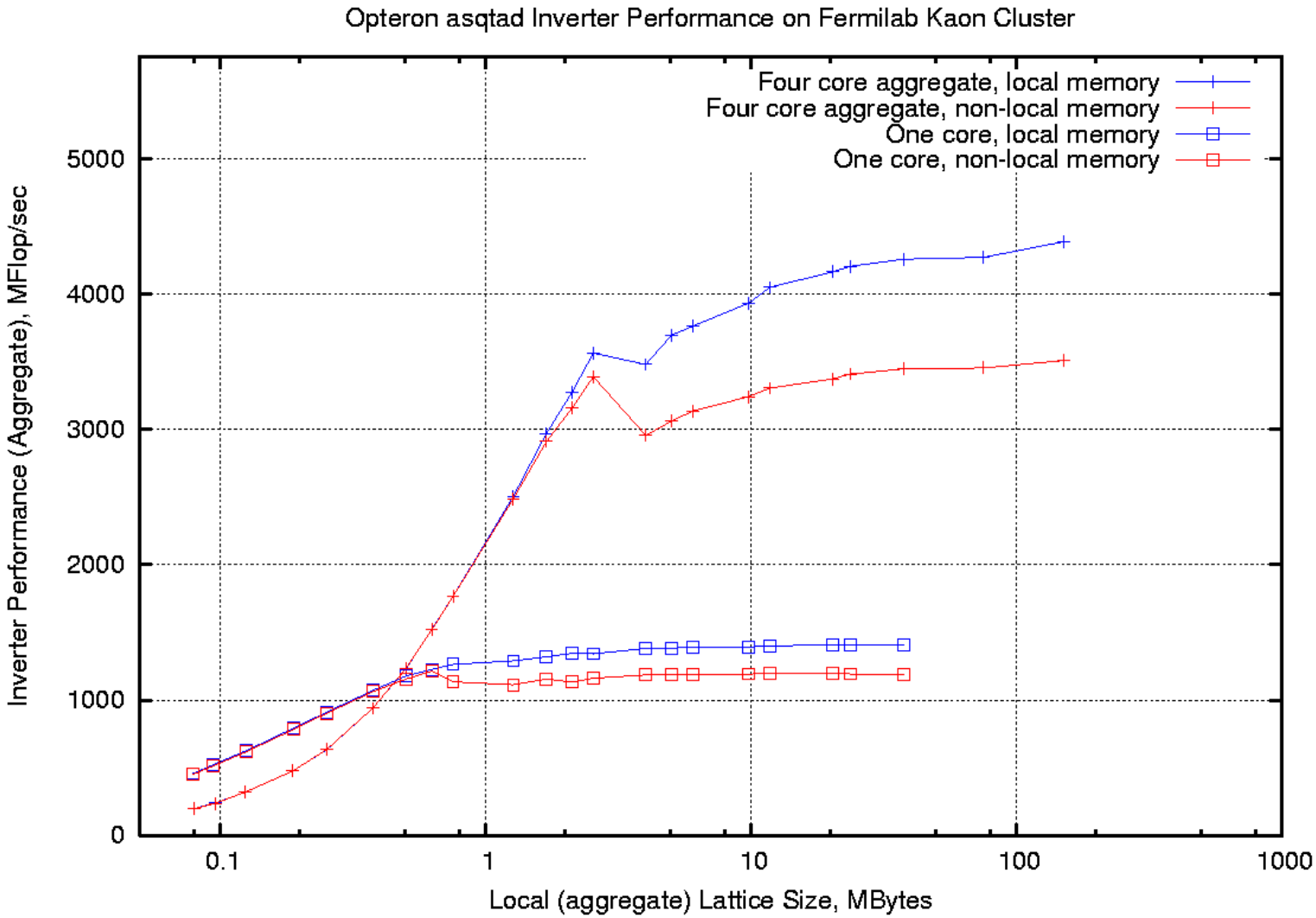
- Four Nvidia Tesla S1070 systems are available for CUDA programming and production
  - Each S1070 has 4 GPUs in 2 banks of 2
  - Each bank of 2 GPUs is attached to one dual Opteron node, accessed via the JPsi batch system
    - Nodes are “gpu01” through “gpu08”
    - Access via queue “gpu” (***`qsub -q gpu -l nodes=1 -I -A yourproject`***)
    - Parallel codes using multiple banks can use two or more nodes with MPI (or QMP) over Infiniband
  - Send mail to lqcd-admin@fnal.gov to request access



Numa Effects

- For new users (and a reminder to existing users), please be aware that Kaon and JPsi are NUMA (non-uniform memory access machines)
- In order to achieve the best performance it is important to lock processes to cores and utilize local memory
- The MPI launchers provided on Kaon and JPsi (**`mpirun_rsh`**) will correctly do this for you
- You can use **`numactl`** to manually lock processes and memory – we're happy to give advice

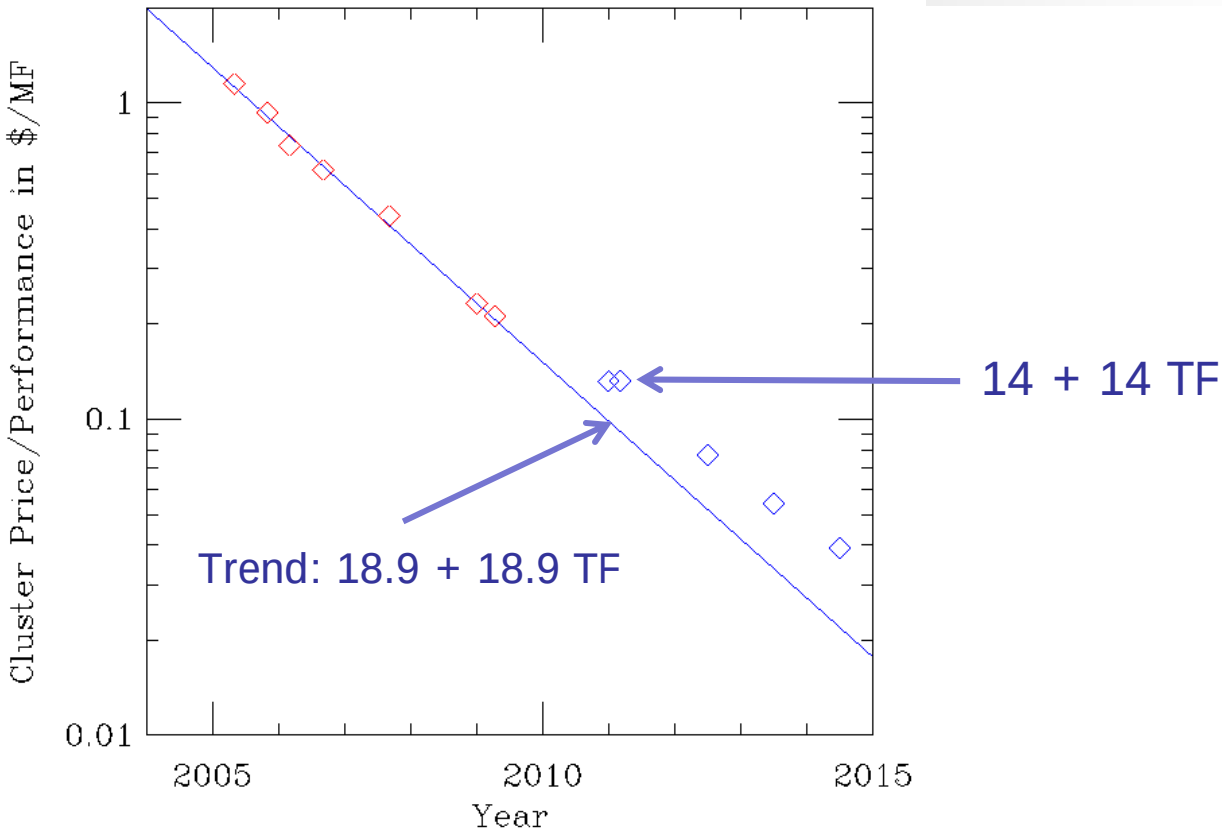
NUMA Effects



FY10/FY11 Deployment

- The LQCD-ext project plans currently call for a combined FY10/FY11 deployment at Fermilab
- Probable configuration:
  - Intel-based ("Nehalem" or "Westmere") dual-socket quad-core or hex-core, or AMD Opteron hex-core
  - QDR Infiniband
  - Either a close duplicate of the JLab ARRA machine or the next generation
- Conservative performance estimate for OMB-300: 28 TF

FY10/FY11 Cost and Performance Basis



Cluster	Price per Node	Performance/Node, MF	Price/Performance
Pion #1	$1910	1660	$1.15/MF
Pion #2	$1554	1660	$0.94/MF
6n	$1785	2430	$0.74/MF
Kaon	$2617	4260	$0.61/MF
7n	$3320	7550	$0.44/MF
J/Psi #1	$2274	9810	$0.23/MF
J/Psi #2	$2082	9810	$0.21/MF

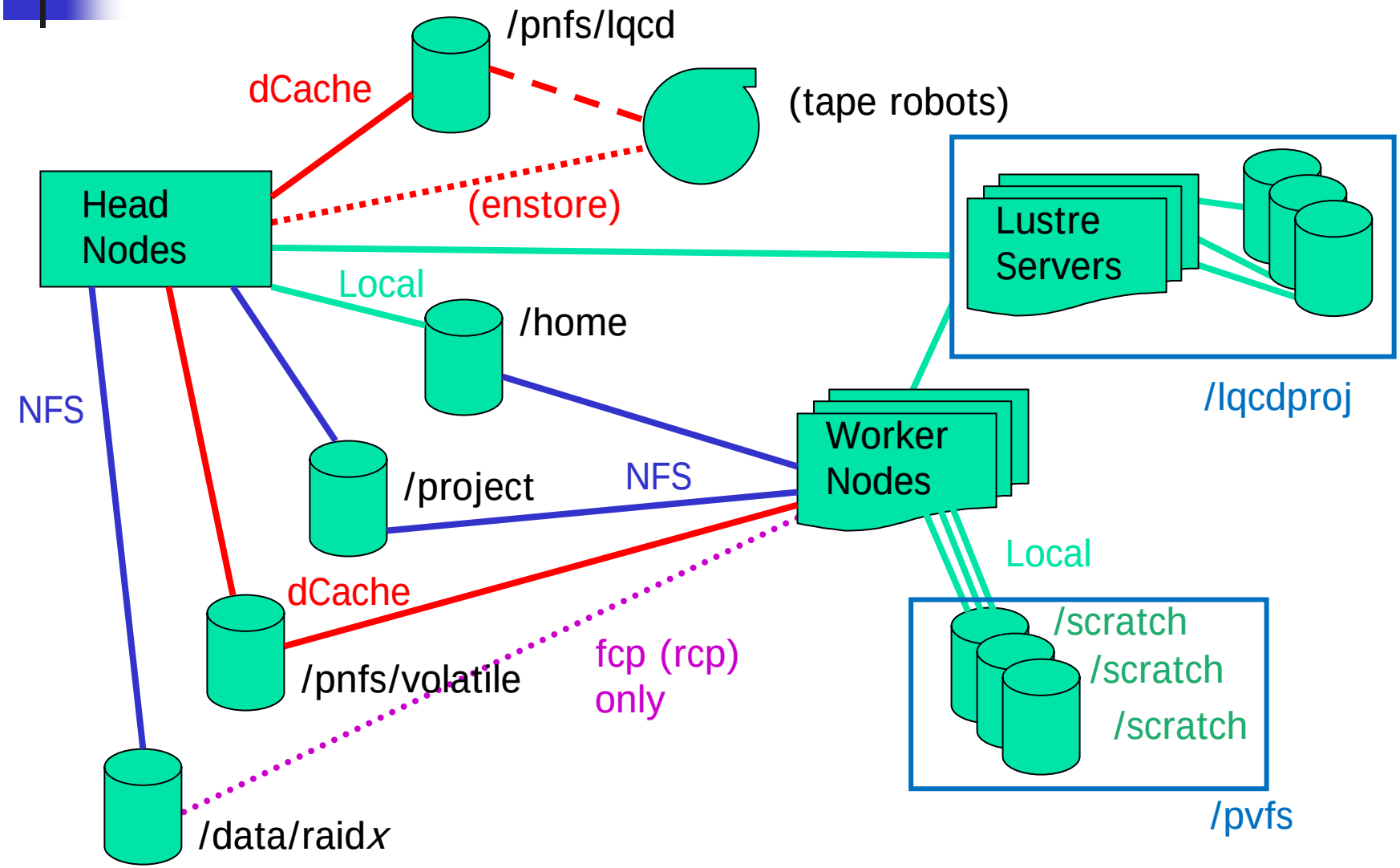
Performance of Current x86 Processors

Cluster	Processor	DWF Performance per Node	Clover Performance per Node	Asqtad Performance per Node
7n	1.9 GHz Dual CPU Quad Core Opteron	8800 MFlops	5148 MFlops	6300 MFlops
J/Psi	2.1 GHz Dual CPU Quad Core Opteron	10061 MFlops	7423 MFlops	9563 MFlops
Shanghai	*2.4 GHz Dual CPU Quad Core Opteron*	*12530 MFlops*	*Not measured*	*10370 MFlops*
Nehalem 1066 MHz FSB	*2.26 GHz Dual CPU Quad Core Xeon*	*22200 MFlops*	*12460 MFlops*	*15940 MFlops*
Nehalem 1333 MHz FSB	*2.93 GHz Dual CPU Quad Core Xeon*	*27720 MFlops*	*15260 MFlops*	*19390 MFlops*

- 7n and J/Psi performance figures are from 128-process parallel runs (90% scaling from single to 16-nodes)

- *Shanghai and Nehalem performance figures are estimated from single node performance using 90% and 80% scaling factors, respectively*

Storage


/pnfs/lqcd
dCache
(tape robots)
Head Nodes
(enstore)
Lustre Servers
Local
/home
/lqcdproj
NFS
Worker Nodes
/project
NFS
Local
dCache
/scratch
/scratch
/scratch
/pnfs/volatile
fcp (rcp) only
/pvfs
/data/raidx

Properties of Filesystems

Name	Type	Visibiilty	Integrity	I/O Restrictions
/home	NFS	Global within cluster (qcd, pion/kaon, jpsi)	Backed up nightly	Limited data rate
/project	NFS	Global	Backed up nightly	Limited data rate
/scratch	Local disk	Each worker has own	Erased at beginning of each job	High scalable data rate
/pvfs	Set of local disks	Each worker of a job can see	Optionally created at beginning of a job, destroyed at the end	High scalable data rate, large size
/data/raidx	NFS	Head nodes only	RAID hardware but not backed up	Limited rate, use fcp to access
/pnfs/volatile	dCache	Global	Not backed up, oldest files deleted on demand	Scalable rate, no appends
/pnfs/lqcd	Enstore / dCache	Head nodes only	Data are on tape	No appends
/lqcdproj	Lustre	Global	RAID hardware but not backed up	None (POSIX) Scalable rate

Statistics

- Since April 1, 2008:
  - Users submitting jobs:
    62 USQCD, 6 administrators or other
  - 1,390,428 jobs (1,221,629 multi-node)
  - 10.6M node-hours = 17.6M 6n-node-hours

User Authentication

- Kerberos
  - Use Kerberos clients (ssh, rsh, telnet, ftp) or cryptocards
  - Linux, Windows, Mac support
  - Clients are much easier than cryptocards
- Kerberos for Windows
  - See our web pages for kerberos-lite
  - I highly recommend using Cygwin with kerberos-lite
- Kerboros for OS/X
  - See http://www.fnal.gov/orgs/macusers/osx/
  - The “OpenSSH Client Only 3.x Downgrade Packages” links will give you ssh’s that will work to access our head nodes

User Support

- Web Pages
  - http://www.usqcd.org/fnal/
- Mailing lists
  - lqcd-admin@fnal.gov
  - lqcd-users@fnal.gov
- Trouble tickets
  - Please send all help requests to lqcd-admin@fnal.gov
  - Fermilab is transitioning to a new help-desk system; sorry, but new accounts will require a few extra days compared to the past until the kinks are worked out of the system
  - Once the help-desk system is working smoothly, we will encourage users to use it instead of e-mail for help requests (likely many months away)

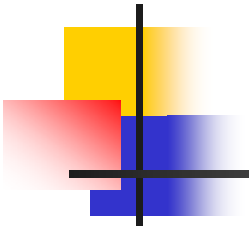
User Support

- Level of support
  - 10 x 5, plus best effort off-hours
- Backups
  - /home, /project are backed up nightly from kaon1, jpsi1, and lqcd; restores are available for up to 12 months
  - /data/raidx, /pnfs/volatile, /lqcdproj are not backed up – users are responsible for data integrity
- Quotas: **`quota -l`** to check disk, **`lquota`** (on lqcd.fnal.gov) to check account usage

User Support

Fermilab points of contact:

- Best choice: lqcd-admin@fnal.gov
- Don Holmgren, djholm@fnal.gov
- Amitoj Singh, amitoj@fnal.gov
- Kurt Ruthmansdorfer, kurt@fnal.gov
- Nirmal Seenu, nirmal@fnal.gov
- Jim Simone, simone@fnal.gov
- Ken Schumacher, kschu@fnal.gov
- Rick van Conant, vanconant@fnal.gov
- Bob Forster, forster@fnal.gov
- Paul Mackenzie, mackenzie@fnal.gov



Backup Slides

Mass Storage

"Enstore"

- Robotic, network-attached tape drives
- Files are copied using "encp src dest"
- > 40 MB/sec transfer rate per stream
  - Currently limited to ~ 120 MB/sec total across clusters
- Currently using ~220 Tbytes of storage
  - An increase of 60 Tbytes since last year

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

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
- Provides high aggregate read/write rates when multiple jobs are accessing multiple files on different pools
- Supports file copies (via dccp) and direct I/O (via libdcap: posix/ansi style calls)
- ~ 40 Tbytes available – another 40 Tbytes can be added
- No appends. Any synchronization between nodes in a job (MPI collectives) may lead to deadlocks.

Local Storage

Disk RAID arrays attached to head node

- /data/raid*x*, *x* = 1-9, total ~ 10 Tbytes
- Also, /project (visible from worker nodes)
- Data files should be copied by user jobs via fcp (like rcp) to/from server node
- Performance is limited:
  - By network throughput to/from server node
  - By load on server node

Local Storage

/scratch

- Each worker node has a local disk (30 GB on QCD and Pion, 90 GB on Kaon
- 30-40 Mbyte/sec sustained rate per node
- Cleaned at the beginning of each job
- Suitable for QIO “multifile” operations

Local Storage

/pvfs

- On a multinode job, the individual /scratch partitions can optionally be combined into a larger single filesystem mounted at /pvfs and visible to all of the nodes in that job
- This is useful when /scratch on the head node of a job is not sufficient in size
- To request /pvfs creation, add "-v PVFS=" to your qsub command
- /pvfs is destroyed at the end of your job

FCP instead of RCP

- In your job scripts, please use fcp instead of rcp or rsync
  - fcp is a throttled (by queuing) form of rcp
  - By restricting the number of fcp's in flight, we avoid head thrashing on disks and increase aggregate throughput
  - Syntax:

```
fcp [host]:src.file [host]:dest.file
fcp –c rcp –r –p src_dir [host]:dest
fcp –c rsync –a src host:dest
```

Moving Files Between USQCD Sites

- Avoiding doing double scp's:
  - ssh Tunnel scripts – provide "one hop" transfers to/from BNL and JLab
  - See web pages for examples