MINOS Computing Infrastructure

Experiment specifics

- Neutrino Oscillations/disappearance
- Beam – CC, NC, NuE, NuTau
- Cosmic – Charge ratio, Atmospheric

- Number of users
  - About 120 active collaborators
  - About 100 using Fermilab facilities, over 75 of these are active.

Experiment schedule

	Pre-2009	2009	2010	2011	2012	2013	2014
Planning							
Constructio							
Commissio							
Data taking	X	X	X	X	MNV	MNV	MNV
Data	X	X	X	X	X	X	X

Data

- How much data/year?
  - Test beam – 3 GB, 23K files caldet_data
  - Pedestal and calibration - negligible
  - Normal data – 1.5 TB/year, 20K files
    see http://www-numi/computing/dh/CFL/CFLSUM
  - Normal data after quality filtering? same
- How large are the major data streams ?
  - Raw / Reco / Summary
    3.8 / 6.0 / 1.2 Tbytes Near
    2.6 / 6.0 / 1.2 TBytes Far
  - MC / Reco / Summary
    10. / 52. / 5.6 Tbytes daikon_04-cedar_phy_bh

Central FNAL systems

- Uses:
  - Software development and debugging
  - Reconstruction and data filtering
  - Calibration and alignment
  - User data analysis
  - No MC Generation
- Minos Cluster - 27 old 2-core SLF 4.4
  - plan to replace with six new 8 core hosts
- FNALU -
  - access to SunOS for web servers
  - FNALU batch – has AFS. But only 26 mostly slow slots. We mostly ignore it
- FermiGrid
  - GPFarm allocation 400, 100 guaranteed, 64 with AFS (retiring for Parrot)
  - Have used rest of FermiGrid effecitvely, up to 1000, when availablel

Central FNAL systems

- Storage used
  - Enstore – 301 Tbytes
  - Dcache
    - 8 TB DAQ read/write
    - 14 TB Minos read (ntuples)
    - 20 TB Public read (m
  - BlueArc
    - 8 TB /minos/scratch (apps)
    - 50 TB /minos/data, being doubled now

Data flow

	Pre-2009	2009	2010	2011	2012	2013	2014
Raw Data,	6	7.5	9	10.5			
Processed Data, TB	220	50	50	50			
User data,	20	20	30	30			
Simulated data, TB	50						

Please enter incremental quantities

CPU needs

	Pre-2009	2009	2010	2011	2012	2013	2014
Running	?	?	?	?			
Reconstruc	200	200	300	400			
Calibration	small	small	small	small			
Skimming	100	100	100	100			
Analysis	400	500	600	600			
Simulation	small	small	small	small			

Please use CPU-years on a current machine
e.g. # events * time per event in sec * 3 x10^7 * reprocessing factor

Operating systems

- Minos Cluster – SLF 4.4
  - Stuck until kcron/aklog problems fixed at SLF 4.7 and 5
- FermiGrid – SLF >= 4.4
- Offsite – many including MacOS.
  - They roll their own, no central support

Data storage and tracking

- SAM Data File Catalog
- Remote access to data

  ftp from Dcache server

  xrootd with server on Minos Cluster node(s)

Remote systems

- Monte Carlo generation at

  Caltech, Minnesota, Rutherford, Tufts, William & Mary
- All shared with or borrowed from other users in various ways. Sometimes through cooling constraints !
- There is no organized sharing of remote systems for user analysis.
- We will soon add TACC/Austin for MC generation and, for the first time, remote reconstruction.

Data distribution to remote sites

Where, what, quantity, speed, method

- UMN copies all raw data, ftp from dcache
- TACC copying all neardet_data, ftp from dcache

  2 to 3 Mbytes per second, this is adequate,
- RAL/Caltech have copied ntuples, as above.

  Driven by local file list, from CFL or SAM listings
- MC data import

  scp -c blowfish from many production sites into BlueArc

  Typically 1 Mbyte/second (due to latency), good enough

  mcimport cron job validates, srmcp's to Dcache, catalogs

Grid

- Grid usage
  - Production reconstruction on Gpfarm, up to 800 jobs
  - User analysis via glideinWMS, up to 1000, goal of 5000 peak
  - User analysis on Minos Cluster via local Condor pool, about 40
- Grid tools
  - Using SRM locally, per CD recommendation
- Use of glideinWMS
  - User analysis jobs on GPFarm and FermiGrid
  - Farm processing is moving to this now.
- General grid resources
  - Certified code at TACC/Austin. Plan to run MC, and reco there

Databases

- Oracle for SAM (the only option)
- Mysql for CRL and all other activity
- SAM is under 16 GB. Mysql is about 70 GB.
- Access rate –
  - SAM is small,
  - Mysql runs with 250 connection limit, should go to 1500
- Replication
  - Many remote and laptop users, via Nick West's dbmauto
  - Local farm processing, for isolation
- Archives
  - Oracle daily; Mysql monthly with binlogs for incremental

Conditions

- Conditions and calibrations
  - Stored in and accessed from mysql
  - Framework jobs use root's tSQL API

Code management

- Code repository
  - CVS, using the Fermilab CDserver/browser
- Build system
  - SoftRelTools
- Distribution system
  - Local builds
  - Parrot gives access to software via HTML

Standard packages

- What standard packages are used:
  - clhep
  - dcap
  - encp
  - geant / geant4
  - genie
  - lhapdf
  - neugen
  - pythia6
  - Root - bleeding edge used in development
  - stdhep

What worked really well?

- SAM as the Data File Catalog, including the Web services interface, and the browser for getting file lists.
- Mysql has been virtually maintenance free. Handing it over to the DBA's has been a lot more work than maintaining it.
- File (ntuple) concatenation has reduced the number of files to be handled over 20-fold, a big winner.
- Offsite MC generation has worked very smoothly, after we learned to deal with the import process in an automated way (error detection, error recover, etc.)
- glideinWMS has delivered a huge increase in analysis capacity, using FermiGrid opportunistically.

What would you not do again?

- SRM has severe limitations (reliability, speed, latency, diagnostics, installation)
- We are moving user mysql access to a replica, to avoid overloading the primary database.
- Would build code in Bluearc, due to capacity limits of AFS, and to aid running on FermiGrid .
- Would put login areas in BlueArc, not AFS, if permitted.
- Would plan earlier for file concatenation and splitting.
- Will probably move from ESNET phone bridge to EVO

What should you know ?

- Enstore overheads – 200 MB and 5 seconds per file
- Dcache directory overhead .1 second/file
- Tape lifetime is around 2000 mounts
- Dcache DOS outages a couple of times per year due to users writing 10's of thousands of files
- OSE security - malicious root user on worker nodes

  /grid/app mounted exec only on grid

  /grid/data mounted writeable, noexec on grid

Monitoring

http://www-numi.fnal.gov/computing/dh

- checklist via frame at left
- check glideinWMS activity at condor monitoring
- manually check 'predator' process scanning raw data

Minos status page

http://computing.fnal.gov/cgibin/cdsystemstatus/system_status.pl

What's up ?

- DCache server is moving to NFSv4 protocol soon
  - NFSv4 client is in latest Linux kernel
  - The end of dcap/dccp etc
- Disk costs less than tape
  - consider using Enstore disk movers
- Enstore may provide file aggregation – loose talk so far