

FY 2007 Minos Analysis Computing Resource Expansion

Guidance

In response to the Young Minos requests in DocDB 3254-v1, we used the remaining FY07 Minos computing budget to purchase computing capacity for analysis batch jobs, and enough NFS served disk to support that additional computing.

Analysis CPU

The Young Minos request was to triple/quadruple the dedicated 12 core/36 GHz being delivered by 6 Minos Cluster nodes at that time, a net of up to 150 GHz

<i>Analysis Computing Resources</i>	<i>Capacity in GHertz</i>
FNALU Batch	137
Minos Cluster (in LSF)	75
Minos Cluster (not LSF)	75
Minos Grid (new nodes)	170

We increased the amount of LSF in the Minos Cluster to 24 cores/75 GHz.

We have asked that condor queues be established on the entire Minos Cluster. This is being actively pursued, we will work with the support people daily.

We have purchased eight new hosts for Analysis computing. They are Dell PowerEdge 1950 systems, with 2 x Quad-Core Intel Xeon 2.66 Ghz, 16GB RAM, and 500GB SATAu HD. They should be available by the end of October.

The new Analysis nodes will be deployed as part of the FermiGrid compute pool, so that other groups can use them when we do not. We will do what is necessary to make these nodes useful for Minos analysis, including making /afs/fnl.gov available on these systems.

Jobs will be submitted via the Condor system already used by the farm and by other experiments. Steven Cavanaugh has started testing this for Minos, and has run Loon jobs. The addition of AFS will greatly simplify this process.

We are guaranteed to get as much capacity as we have purchased. More capacity beyond our purchase is available via the general FermiGrid. Use of the full FermiGrid may require the more general grid tools being tested by Steve and pioneered by Nick West.

Batch CPU

Here is a rough outline of the scale of the overall FermiGrid capacity, in this case just counting job slots, the information that is most readily available. Our present plan and hope is to have AFS available on the full GP_GRID.

You might figure about 2 GHz per slot, allowing for old nodes and overbooking. Information is from the plots at <http://fermigrid.fnal.gov/fermigrid-metrics.html>
FermiGrid - Production Clusters ([Globus Gatekeeper Service Monitor](#))

<i>Fermigrid resource</i>	<i>Job slots (perhaps 2 GHz)</i>
Minus GP_GRID allocation	300
GP_GRID	1900
CDF (CAF)	3800
D0 (part of CAB)	1100

Storage

<i>Storage Resources</i>	<i>Capacity in TBytes</i>
Old Minos Cluster /local/scratch	5.7
AFS MINOS_DATA	10.5
/minos/scratch	10
/minos/data	20
/grid/data/minos	0.5
/grid/app/minos	0.030

The new disk is a 32 TB SataBeast array served by a BlueArc NFS server. Usage allocations can be adjusted without interrupting service.

We plan to deploy the disk as follows :

10 TB – to replace the /local/scratch disks on the Cluster, allowing us to move more Cluster nodes to analysis, and to use some existing Cluster nodes for special servers.

20 TB – to supplement the existing 5 TB AFS physics work space

Once this disk is mounted and has proven to be performant and reliable, we will probably want to dedicate several of the present Minos Cluster interactive nodes to other roles (Analysis batch, xrootd, data import, etc.)

We will ask to have this mounted on the GP_GRID nodes, as well as the Minos Cluster and server systems.

The /grid/data/minos area is presently heavily used by the farm.

It is intended to be used for temporary user files.

We will make /grid/data/minos/users/* directories to keep things manageable.

We have started to look into the obvious management issues.

The /grid data/app/minos area is for centrally managed products and releases.

This would ideally just be rsynced from the existing AFS product and release areas.

There are some subtle problems with this, not quite trivial at present !

Other Experiments

For comparison, the Batch slot allocation for most other active experiments is uniform at 300. That includes KTev, MiniBoone, SDSS.

Lower priority projects get 225 (CDMS, HyperCP, MIPP, Auger, DES, ILC)

I'm told that MiniBoone does their Analysis batch computing via a condor system running on a collection of about 50 desktop systems, and depending on AFS for a uniform software environment. They have some tens of Terabytes of NFS shared disk deployed for user data, similar to our new BlueArc systems. They have purchased an 8 host (64 core) computing supplement identical to our new systems, but I have not yet tracked down where it will be deployed.

Other Issues

Well, life is neverquite so simple as just deploying hardware.

We have not pushed harder on Condor deployment because there are some non-trivial practical details that needed to be resolved.

The biggest problem with Condor has been that the recent versions in production use at Fermilab spontaneously rerun some large fractions (over 1/20) of the users jobs. The jobs are simply executed twice. There seem to be two or three known technical problems. The last of these is advertized to have been resolved in the lastest Condor released just this Tuesday, 25 September.

A Condor job submission, as supported at Fermilab, starts a thread of execution on the node from which you submitted the job. That process must stay around until your job has run. That produces some real scalability issues (imagine submitting a 10,000 job request from an already overloaded node.) We will likely need to provide one or more gatekeeper servers for submissions.

Our general infrastructure probably could not tolerate thousands of jobs running at once (database and I/O limitations.) We will need to learn to limit out peak computing. (What a great problem to have !)