

## FY12 Plan for **Intensity Frontier/General Computing**

Prepared by: Lee Lueking

Date: November 3, 2011

Relevant Strategic Plans - Strategic Plan for Intensity Frontier

### **Executive Summary**

The experiments in the Intensity Frontier area continue to mature both in their internal organization and their use of the computing services provided by the Computing Division. The goal of the Computing Division to provide common solutions for software and services is being realized and will continue to emerge in FY12. We now have operational experience with the existing job submission and monitoring tools. These, and planned improvements, will allow experiments to obtain the computing and storage resources they need in FY12, and enable the REX Department to provide monitoring and operational support.

The GPCF resources will continue to be the center for interactive login and batch submission as we expand the number of VM's and local batch slots to meet the increased needs. Grid batch slots provided by the General Purpose Grid cluster will continue to be the main source of computing power for IF and the need will increase. BlueArc disk continues to be the central data storage resource and we plan to grow this significantly. This central storage is immensely popular due the efficiency it brings to the processing task, and its cost appears to have been reduced significantly compared to last year. Archiving data to tape in ENSTORE will continue and the small file aggregation service that team is providing should prove to be very valuable.

There will be a year-long accelerator shutdown scheduled to begin in Feb 2012, although this may be delayed to May or June. Although this stops data taking, it is not anticipated to reduce the computing needs for most of the experiments. MINOS, MINERvA, MiniBooNE and NOvA will continue to analyze existing data and prepare for post-shutdown operation. Mu2e, g-2, LBNE, MicroBooNE and others will continue to ramp up their detector design and beam simulation studies. We have compiled trend lines for computing and storage for last year's consumption and we are using these, along with input from each experiment, to assess the usage levels needed for the coming year.

### **Tactical Plan Goals**

- Support for offline computing for data taking experiments in 2012.
- Implement and support tools and procedures for data handling.
- Integrate new experiments into Intensity Frontier computing model.
- Develop sound computing and analysis solutions for Intensity Frontier experiments.

- Provide both interactive and batch computing facilities including easy access to FermiGrid computing resources.
- Central disk storage resources needed by experiments.
- Establish easy to use and effective tape archival procedures for experiments.
- Provide application and operations support.
- Comprehensive solutions for database services.

### **Maintenance/Compliance Drivers**

1. Oracle Red Hat Enterprise Linux licenses for two SAM DB servers and one Minerva conditions DB server.
2. Oracle maintenance for the SAM DB instances.
3. Oracle host contracts for two SAM DB servers.
4. BlueArc maintenance
5. Enstore Drive and robot maintenance
6. Networking maintenance

### **Upgrades and Enhancement Drivers**

1. Replacements for 286 cores GP Grid computing nodes
2. 580 new GRID slots (CPU cores)
3. Approximately 200TB new Central Storage BlueArc disk
4. 15 new 4-core interactive VM's to be added to the GPCF
5. 750 x 800Gb tapes for ENSTORE robot storage
6. A small number of service VMs , for possible services like database, or file services. Probably on the order of 10-12 x 2 core VMs.

### **Strategic Drivers**

1. MINOS, MiniBooNE and MINERvA will continue running with their full detectors until the Spring 2012 shutdown.
2. NOvA will continue running its Near Detector on the Surface and putting in place elements of the far detector facility at Ash River, MN.
3. The other experiments continue to design their beam lines and detectors for future running periods and require significant GRID computing and storage resources.
4. Using GRID computing resources at remote sites will be encouraged. This may amount to 25% of the computing needs for the Intensity Frontier program, if needed.
5. The cost of central disk storage solutions seems to be decreasing and this may make it easier to supply the needed storage.
6. Software solutions for staging data provided through SAM or other applications are in progress.

7. The effect of shutting down Run II and the potential for reclaiming some of those resources is being explored.
8. In the event that we are unable to replace existing storage and CPU hardware, we will keep end-of-life DISK and CPU hardware and pay the implicit costs for increased maintenance, power and cooling.
9. JIRA issue tracking used by REX will be transitioning to Service NOW.

## **Activities and Work Definition**

### **Activity = Intensity Frontier/ General Computing**

- Activity type: Service
- Description: User support for Interactive and Batch submission cluster.
- Timescale: Ongoing
- Metrics: GPCF cluster usage and GRID usage

### **Activity = Intensity Frontier/ General Computing**

- Activity type: Service
- Description: Assist in establishing grid submission to remote sites
- Timescale: Ongoing
- Metrics: Processing fraction at remote sites

### **Activity = Intensity Frontier/ General Computing**

- Activity type: Service
- Description: Operational support for SAM catalog and data management systems
- Timescale: Ongoing
- Metrics: The use of the SAM catalog and other features by IF experiments.

### **Activity = Intensity Frontier/ General Computing**

- Activity type: Service
- Description: Set up and ongoing support for Control Room computers
- Timescale: Ongoing
- Metrics: Critical support for online data taking

### **Activity = Intensity Frontier/ General Computing**

- Activity type: Service
- Description: Assist in specifying, procuring and commissioning of new hardware
- Timescale: As needed, when needed
- Metrics: timely acquisition and installation of new hardware

## Detailed Tactical Plan Objectives and Priorities

The computing needs of the IF program for FY12 are based on estimates provided by each experiment, and trend lines of actual usage compiled over the last year. We continue to use MINOS as a baseline to compare against and assume that it represents a good model of usage for other experiments, when scaled appropriately. Although there is a year-long accelerator shutdown scheduled to begin in the spring of 2012, we do not anticipate this will significantly reduce the computing resource needs. The budget for Maintenance and Compliance Drivers are summarized in Tables 1, and that for Upgrades and Enhancement needs for each experiment are shown in Table 2. Estimated unit cost for DISK and CPU are based on FY11 purchases and preliminary quotes we have for FY12. Table 3 is a reference to show the total anticipated resources allocated to each experiment including the FY12 purchases.

Table 1. Details of Miscellaneous Intensity Frontier Budget items.

<b>Item</b>	<b>FY11 Total Cost (\$)</b>
Replacement computing for neutrino people	12000
Oracle redhat license	3x\$2040=6120
Oracle host contract	2x\$3000=6000
Oracle maintenance	15000
bluearc maintenance	9650
small items on procard and contingency	20000
furniture/books/training/licenses/computing accessories	15000
Domestic travel	12x\$2000=24000
Foreign travel	2x\$4000=8000
Enstore Drive Maintenance	2 x \$1750=3500
Enstore Robot Slots	300x\$40=12000
Enstore robot maintenance	1000
Networking maintenance	4000
<b>Total</b>	

Table 2. FY12 additional resources requested itemized by experiment. Numbers include upgrades and enhancement budget items. Items pre-purchased in FY11 are indicated by “\*” with the requested quantity in ().

Experiment	Disk (TB)	Tape (TB)	Int. Login (4 core VMs )	Local Batch (cores)	GRID (cores)	Servers
MINOS	0* (30)	100	0	0	0	0* (3)
NOvA	100	300	5	20	300	0
MINERvA	80	50	5	20	200	0* (3)
Mu2e	5	0	0	0	30	0
LBNE	10	0	4	20	50	0
MicroBooNE	23	50	0	20	0	0
ArgoNeuT	0	0	0	0	0	0
MiniBooNE	15	100	0	0	0	0
Muon g - 2	0* (5)	0	1	0	5	0
replace/added	580 (add)			40(repl)	300 (repl)	
Recovery from Run II					285	
<b>TOTAL</b>	780	600	15	125	300	0
<b>TOTAL Cost</b>	\$270k	\$21k	\$12k	\$14.5k	\$34.8k	\$0k

Table 3. Total resources expected to be allocated to each experiment by the end of FY12.

	Disk(TB)	Tape(TB)	Int. Login (4 core VMs)	Local Batch (slots)	GRID (slots)
<b>MiniBooNE</b>	140	380	0	0	200
<b>MINOS</b>	190	700	2	30	1000
<b>SciBooNE</b>	50	50	0	0	0
<b>MINERvA</b>	210	160	5	40	500
<b>Argoneut</b>	20	20	1	0	10
<b>NOvA</b>	160	490	10	40	600
<b>MicroBooNE</b>	30	500	2	20	60
<b>LBNE (all)</b>	30	6	5	20	200
<b>Mu2e</b>	10	0	2	20	50
<b>g minus 2</b>	10	0	2	5	40
<b>Totals</b>	850	2306	29	175	2660

Storage and computing resource usage in the last year are shown in the figures below. Experiments are provided separate central disk allocations for *applications* and *data*. The total used and allocated space for *applications* is shown in Figure 1 and is relatively flat at around 14 TB. The total space used and allocated for *data* is shown in Figure 2, and this indicates approximately linear growth from 200TB to 330TB over the 8 month period from March through October 2011. This normalized to the full year (1.5 \* 130) is

195TB, which is similar to the experiments' requests. Individual charts for data storage space used by each IF experiment can be found in Appendix 1.

Figure 1. BlueArc application areas disk storage used for all IF experiments. (The green line is space used, red line is space allocated.)

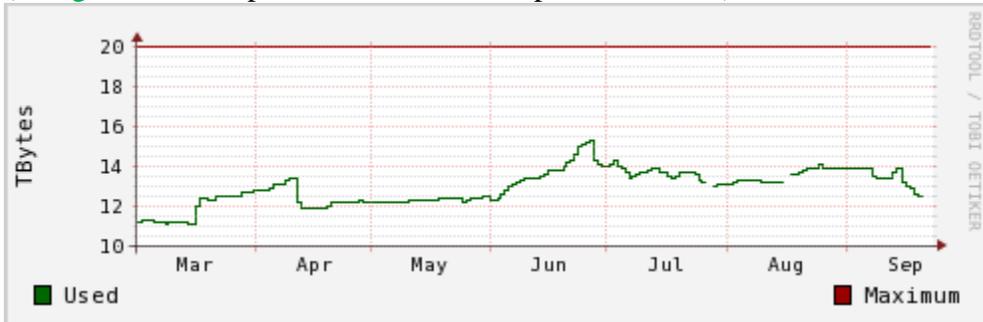
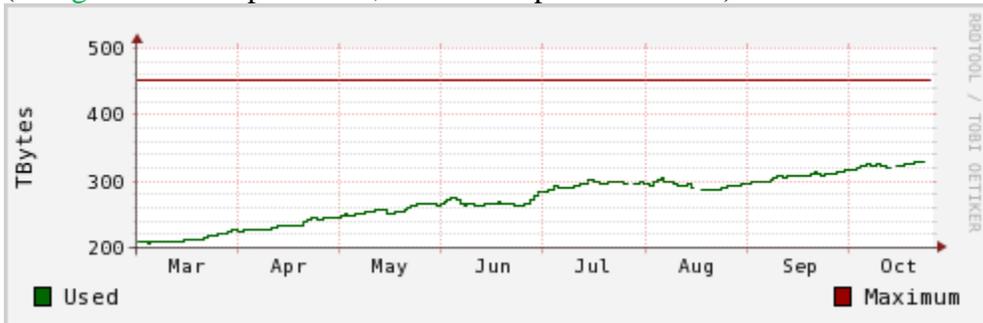


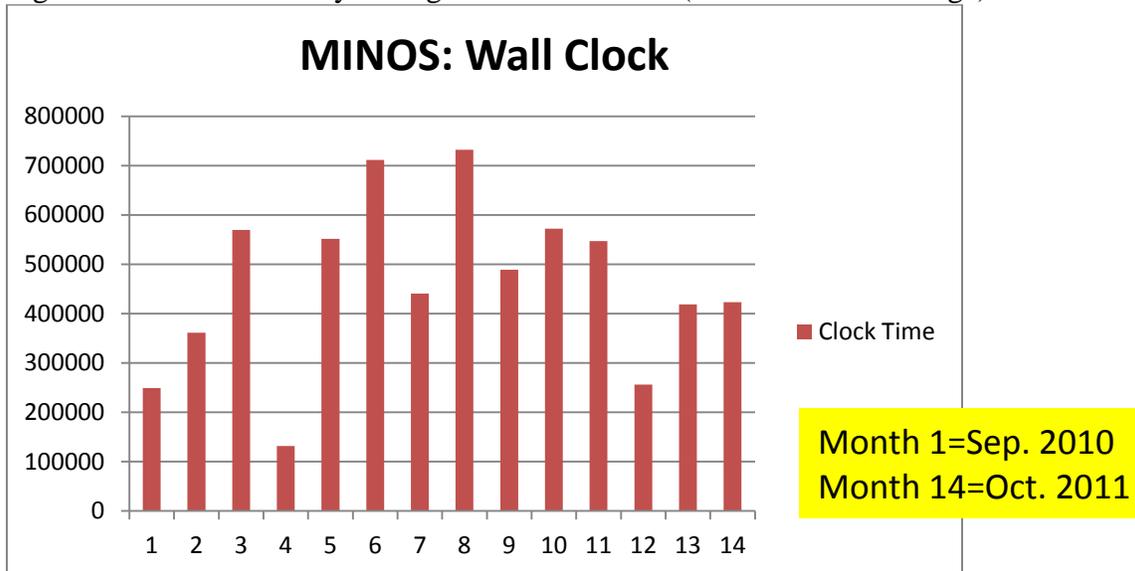
Figure 2. BlueArc data areas disk storage used for all IF experiments. (The green line is space used, red line is space allocated.)



A record of the GP Grid Cluster computing resources used by each experiment in the Fermilab VO is recorded in the Gratia database. This enables evaluation of trends in computing usage and can be used to understand future needs. Figure 3 shows the wall-clock time for MINOS processing per month over the period of September 2010 through October 2011. Wall clock time is a useful metric as it indicates total grid-slot occupation time, including staging, initialization, processing, and so on. There are approximately 750 hours per month, so the monthly average for MINOS represents about 614 slots filled during a typical month.

Similar charts for the heaviest GP Grid using experiments are included in Appendix 2. Many experiments have used fewer processing resources last year than we had estimated (e.g. Minerva, Nova), but some have used significantly more (e.g. Mu2e). We anticipate FY12 will see a significant increase in usage and we use MINOS as a “typical” experiment. It is expected that Minerva, Nova and others will meet or exceed this usage level in FY12.

Figure 3. MINOS Monthly average 461 kHrs/month (614 filled-slot average)



## MINOS

MINOS has 8 years of experience using and estimating their storage needs. Minos storage includes Dcache, BlueArc and tape. The current breakdown for dCache and BluArc is as follows:

RawDataWritePools - 15 TB, all raw data through 2012, using 13 now

MinosPrdReadPools - 20 TB current ntuples (i.e. ND full cosmics)

ReadPools - 14 TB shared with other experiments

WritePools - 6 TB shared cache, non-rawdata files

Bluearc

/grid/fermiapp - 1 TB software releases

/minos/app - 7 TB should shift most of this to /minos/data/users

/minos/data - 190 TB data working area, 30 TB more will be needed

Tape usage is estimated to be 100TB for FY12 and MINOS is currently using 634 TB net as of October 2011. Minos would like to see the deployment of Open Hardware disk solutions at around \$100/TByte, as Enstore movers and/or as PBDM backends to Bluearc. They will start to use the new Fermicloud ftp services where appropriate to set correct user file ownerships for grid jobs. They are evaluating and encourage the move to Chimera based DCache. This could be deployed as a PBDM backend to Bluearc.

MINOS computing is performed both on dedicated resources and shared GRID machines. Interactive login and local batch are provided by the Minos Cluster comprises 5 dual quad nodes that were purchased in 2010. AFS is used for user home areas and code is now in Bluearc /minos/app. The Condor local pool has 32 job slots, is heavily used most of the time and has been very successful. Grid resources on the GPfarm are used extensively with 1000 slot standard allocation where 100 slots are priority due to purchase. On FermiGrid MINOS usually gets close to 2000 cores opportunistically and this will continue. They have expanded the limit to 5000, and have observed 4000 in simultaneous use.

MINOS remote processing plans include TACC (Teragrid) and several other sites. About 1/2 the CPU for the dogwood3 reconstruction (full ND Cosmic sample) Monte Carlo simulation is 100% offsite. Sites included in the offsite effort include the following: William and Mary, RAL, Caltech, Tufts, U.T. Austin (TACC). Data transfer to Bluearc was via scp. The transition to the Bestman Gridftp gateway started in early FY12.

Several servers for DB, Web and special purpose are needed. The list includes mysql production/dev servers, and IF shared SAM production/dev db servers. Servers were purchased in FY11 to provide calibration database production, development and replication functionality. These will be shared with other experiments in the Intensity Frontier. Currently the Condor dedicated master nodes are minos25 and minos54. The data movement and monitoring node is minos27. The Condor management nodes will move to virtual systems in FY12.

Special CD manpower needs include continued operations and software support. MINOS employs the following software tools: UPS/UPD, SoftRelTools, ECL, DocDB, SAM, mysql (DBA support), Enstore, Dcache, Bluearc, condor, GlideinWMS, ROOT, IMAP with Thunderbird alpine and other IMAP clients. System Management of Control Room computers is moving from REX to FEF in FY12.

## **NOvA**

NOvA will have significant need for computing in FY12. Central disk and tape is required for large Monte Carlo samples being generated. Large data sets will be acquired during commissioning and even when the beam is not on, cosmic ray samples will continue to grow for calibration and background studies. Data rates will vary from 100GB to 1000GB/day. For tape, the estimate is 3 x disk spaces with the expectation to save data processed/simulations to tape 2 or 3 times a year to free disk space for new development work on the data/simulations. Grid processing is expected to grow to around 600 total slots. These will be used initially for Monte Carlo production and data processing. Significant computing could be done at sites beyond Fermilab if needed; specifically Harvard and SMU have resources for this.

Help from CD will be needed in the areas of data handling, job monitoring, user support, control room setup, some DAQ setup and database maintenance. Special CD manpower needs include help with the GENIE and GEANT4 simulation package and beam simulation. Also, support for the ART Framework is required.

## **MINERvA**

The experiment estimates they will need 80 TB/year of central disk for raw data, calibration, analysis samples. This is based on running experience with the full detector over the past year and a half. Their tape requirements will be at least 100 TB for dual backup of all their raw data and single copies of processed and MC data. Interactive logins will move to 5 or more VM's for pure interactive work and to 40 cores of GPN01 for fast batch testing. For 2012 Minerva requests 500 slots on the GP grid farm. Most database and web server needs are being met through shared machines so there is not a request for any new dedicated servers. However, Minerva does request that several of the existing nodes be retained for I/O intensive applications even after transition of interactive use to VM's.

In the area of manpower they will need help with Grid, Code build/ports, DB applications, SAM, neutrino simulations. 1FTE for coordination, production, accounts, web. Also support for ECL and supporting DB's and help with MINOS->MINERvA transition. They will need support for SLF, grid tools, ups/upd or other code management systems which allow shared use of remote systems.

They anticipate significant processing at Hampton and other Universities. At these sites they will perform most of their MC generation and some small scale analysis. They estimate data transfer needs of 10 GB/day to/from 10 sites.

## **MiniBooNE**

MiniBooNE is scheduled to run until the 2012 shutdown, however there are some caveats. The experiment may continue to run after the shutdown. Future analysis will include 1 year of active analysis after run ends and 2 more years of keeping data accessible.

In FY12 it is assumed MiniBooNE will use the FY11 allocation of 2M CPU-hours or less. The needs for data storage include the files needed for analysis currently kept in 3 Bluearc volumes and 8 mbdata0[1-8].fnal.gov servers. They currently have about 25 TB available on Bluearc and need about 20TB/year to accommodate data streams. Help from CD is requested for handling retired servers and they propose copying files from retired servers to mitigate risk from catastrophic failure on older machines, but leave services running.

## **LBNE**

Resource requests are combined for Beamline, Near Detector (ND), Water Cerenkov (WC), and Liquid Argon (LAr). The advantage of doing it this way is LBNE does not yet know how much they will need and sharing the request reduces the chance of resource underutilization. However, the risk is that there will be competition for resources among the LBNE sub-components. Currently they are not using all the resources they have available, but as their needs grow, they will need to use dedicated resources instead of borrowing them.

The storage needs for disk and tape include MonteCarlo, Code, Documentation, and User home directories. The Monte Carlo samples include raw and processed versions with varying amounts of “truth” information. MC samples will become obsolete rapidly and be replaced with newer, better ones, so we prefer disk to tape at this stage. Some archival tape storage of CD-1 and CD-2 MC sets for future reference is needed. In the future large amounts of raw data (mostly noise) produced by the online DAQ – filtered sets available in several places. Near Detector data is most convenient to handle at FNAL. Backed-up storage for code and documentation is also required.

The need for Interactive login and batch resources depends on how collaborators are used to doing their work. MINOS and NOvA have a large, shared interactive cluster with access to BlueArc disk and shared code. CDF relies more on users stripping subsamples of the data with micro-ntuples on desktops made on the batch farms (infrequent access to the entire data set, while LBNE will have much more frequent access to the entire data set). A fairly large pool of interactive cores connected to disk with the full data is requested.

On the GRID they will need to run simulation and reconstruction for WCD, LAr, ND, and beamline. The estimate of WCD CPU needed once we are running (3 100 KTon WC baseline):

- Data: 2 passes x263 CPU years per year of collected data for
- production (in 2010 CPU-years)
- Cosmic MC: 109 CPU years – can be re-used for each year of data
- Neutrino MC: 2x47xNrunning-years CPU years.
- 

They do not need all this in FY11 and FY12, but we will need a subset of it to design the detector and evolve our computing environment.

Servers for DB, Web and special purpose will be needed. Already they have multiple docdb's, web areas (wiki, BNL trac, svn repositories for code development and CDR drafts) and they may need a special purpose calibration DB server

## **g-2**

g-2 will attain CD-0 approval in early 2012. We have started writing the Conceptual Design Report (CDR) necessary for CD-1. A large number of simulations using both GEANT4 and MARS will be needed, requiring substantial disk and batch processing power.

## Assumptions and Risks:

There was significant disk (150TB) purchased in FY11 that was a “pre-buy” for FY12. This covers MINOS needs for FY12,13 and provided Minerva an additional 70TB “cushion” in FY11. It appears that disk prices for FY12 will be significantly lower than FY11 (\$350/TB vs. \$800/TB). This gives us significant flexibility to pre-buy a large portion of the FY13 need (FY13 total estimate is 320TB), provide additional storage for experiments, and possibly rely more heavily on use disk archiving as opposed to tape. Assumptions for Local Batch cores, we are retiring/repurposing IF nodes which are used exclusively by Minerva and doubling capacity of the Local Batch. We are assuming that GP Grid resources released from Run II may be sufficient to meet most of our needs. We will replace 300 end of life grid slots.

## Strategic Drivers

(red needs updating)

Table 4. Some of the strategic drivers for each experiment are listed here.

Experiment	Activity
MINOS	<ul style="list-style-type: none"> <li>• Official MINOS running completed, preparation for MINOS+</li> <li>• Ongoing simulation, processing and analysis</li> </ul>
NOvA	<ul style="list-style-type: none"> <li>• Near Detector on Surface data taking</li> <li>• Exercising all on-line and off-line systems.</li> <li>• Beam simulation and analysis</li> </ul>
MINERvA	<ul style="list-style-type: none"> <li>• Data taking with full detector using MINOS near detector as muon catcher</li> <li>• Development and processing for full Simulation</li> <li>• Data processing and analysis at full speed</li> </ul>
MiniBooNE	<ul style="list-style-type: none"> <li>• End of run at spring shutdown.</li> <li>• Ongoing use of existing storage and grid resources</li> </ul>
ArgoNeuT	<ul style="list-style-type: none"> <li>• Detector development and testing – upgrade of purification system, possibly move into new experimental location</li> <li>• Simulation – Will need lots of simulation for analysis of first run</li> </ul>
MicroBooNE	<ul style="list-style-type: none"> <li>• CD2 in 1st Quarter</li> <li>• Will need lots of simulation for setting baseline</li> </ul>
LBNE	<ul style="list-style-type: none"> <li>• Targeting September 1 2010 for submission of the CD-1 documentation.</li> <li>• CD-2 target is mid-2012</li> <li>• More detailed simulations of the proposed alternatives</li> <li>• Develop LBNE-specific reconstruction algorithms</li> </ul>
Mu2e	<ul style="list-style-type: none"> <li>• CD-1 review Q1 CY 2011</li> <li>• Detector development and testing: CRV, (Straws?). Work in AD and TD on beams and magnets.</li> </ul>

	<ul style="list-style-type: none"> <li>• Simulation: Continuous code development; a few months of production MC.</li> </ul>
Muon g-2	<ul style="list-style-type: none"> <li>• Working with BNL data</li> <li>• Preparing for initial approval</li> </ul>

## Software Projects

We are starting, or continuing, several software projects to support the Intensity Frontier offline program. These include tools for Local Batch/Grid Job Submission and monitoring jobs on the grid cluster. New features are also being built for SAM that provide a web-based interface and make the user interface more accessible, and data caching more usable for grid jobs. These will work within the ART and GAUDI frameworks to provide flexible file access. Work is ongoing to also employ Globus Online asynchronous and make these file transfer advantages available to the IF grid user community. As part of the Globus Online work, a third party transfer agent has been developed which solves the file ownership issue that users generating files on the grid have needed to deal with by hand in the past. Table (below) represents an attempt to estimate the manpower needed for these tasks.

Table 5. IF software projects and estimated manpower.

Project	Description	FY12 FTE
Jobsub enhancements	Continued improvements to the job submission tools for local and GRID jobs	0.5
Condor DAG usage	Documentation and scripts to enable experiments to use Condor DAGs in their job definition and submission.	0.2
IF/GRID Monitoring	Framework and infrastructure for monitoring GRID jobs and operations.	0.5
File Transfer System (FTS)	Automated file archival system using SAM to store data to ENSTORE tapes.	0.2
IF SAM	Web-based interface for SAM. Uses FUSE to provide familiar user interface to SAM catalog. Integrates with IF experiment frameworks. Enhancements to SAM needed to provide file caching in the IF environment.	0.5
IF Beam Data	High availability Booster and NuMI beam monitoring and recording application.	0.2
CVM-FS evaluation	Evaluation of CERN VM-FS for IF. This is the CERN Virtual Machine File system in use by some of the LHC experiments for software and conditions data distribution.	0.1
Total		2.2 FTE

## Staffing Requests from Experiments

The experiments have requested 6 FTE's for various needs in FY12. This is summarized in Table 6.

Table 5. Experiment requests for manpower.

<b>Experiment</b>	<b>Manpower Request (FTE)</b>
MINOS	1
NOvA	2
MINERvA	2
Mu2e	0
LBNE	0
MicroBooNE	1
ArgoNeuT	0
MiniBooNE	0
Muon g-2	0
<b>Total</b>	<b>6</b>

## References

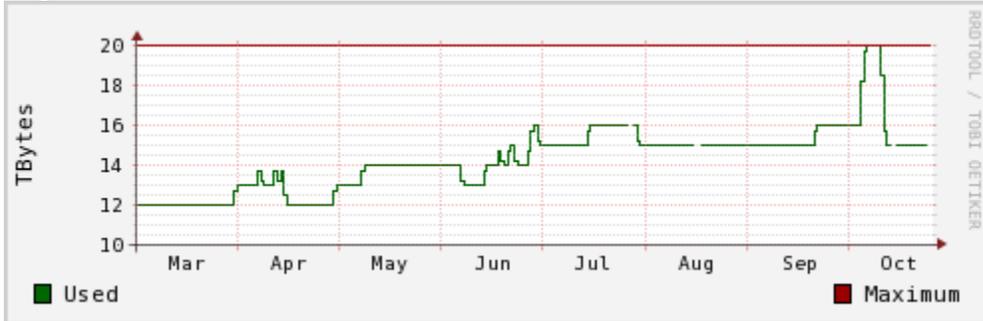
Last year's Tactical plan

Planning spreadsheet

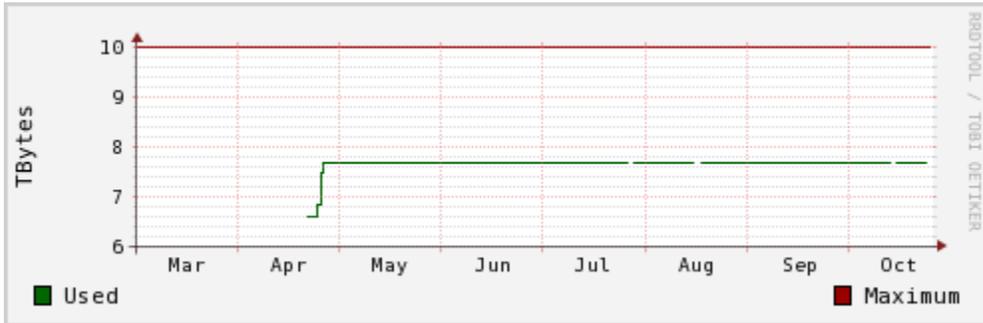
## Appendix 1. Trend lines of Central Storage usage

The charts below show the growth in Central Storage (BlueArc) for each experiment over the months March through October 2011.

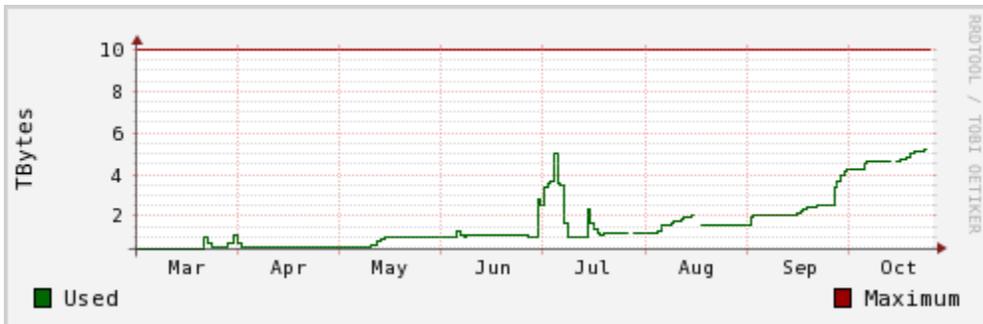
ArgoNeuT



Gm2



LBNE



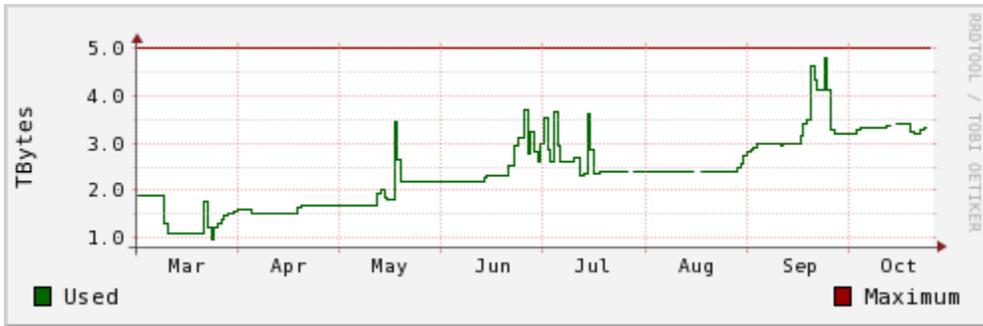
### Minerva



### MINOS



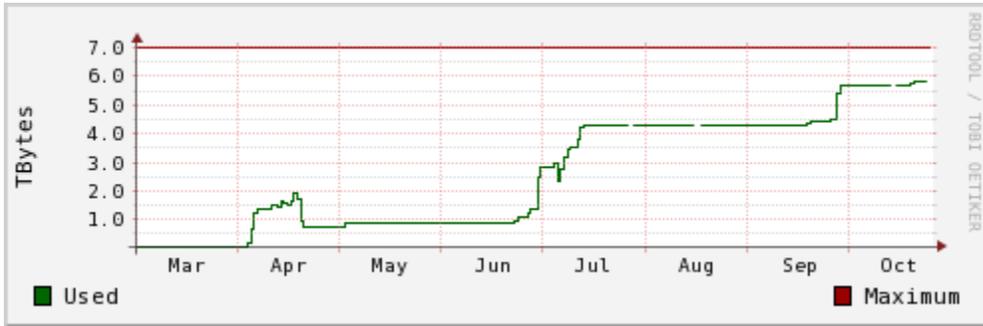
### Mu2e



### NOvA



### MicroBooNE

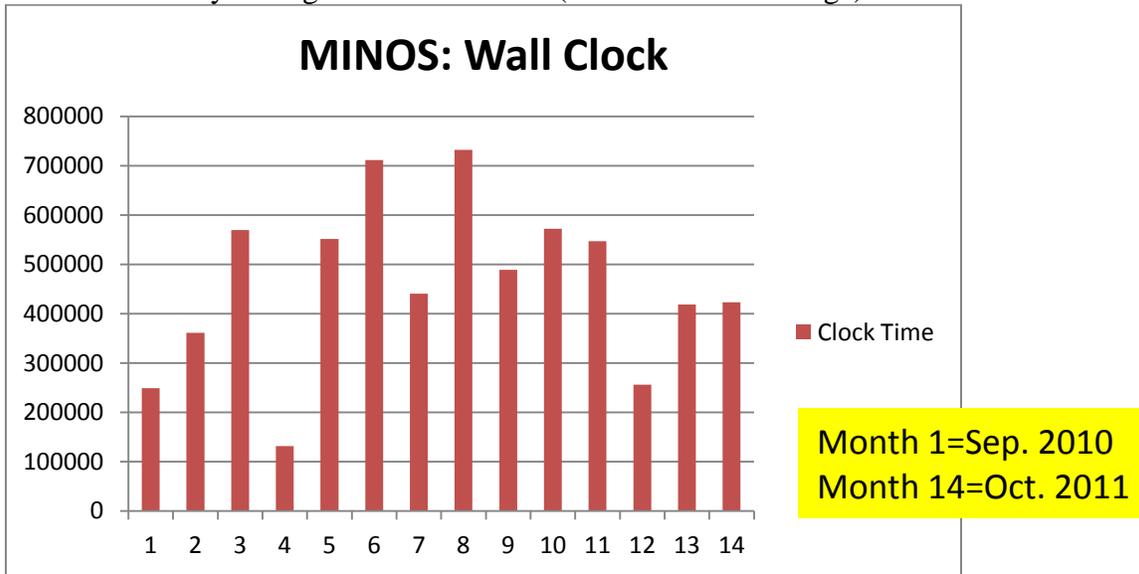


### MiniBooNE

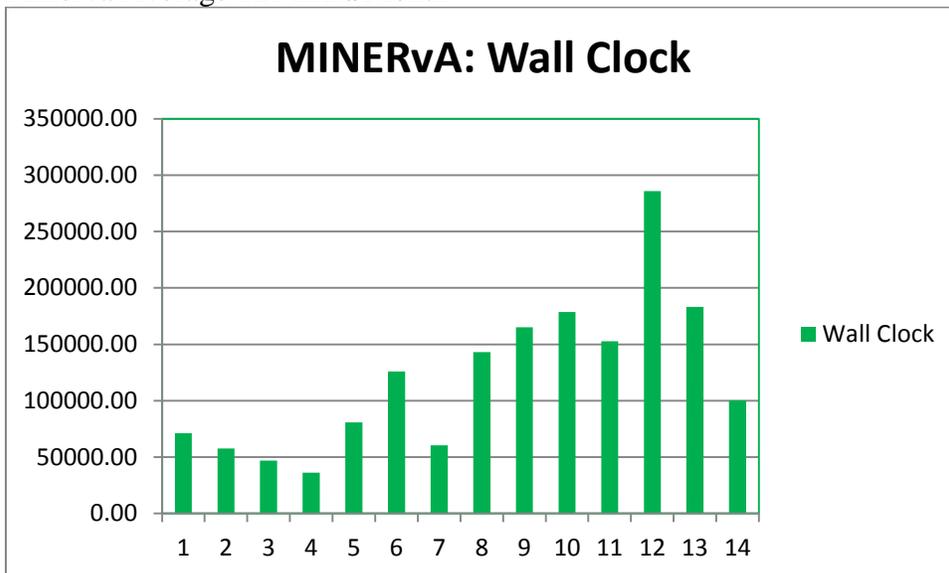


## Appendix 2. GP Grid Cluster usage for Major Experiments

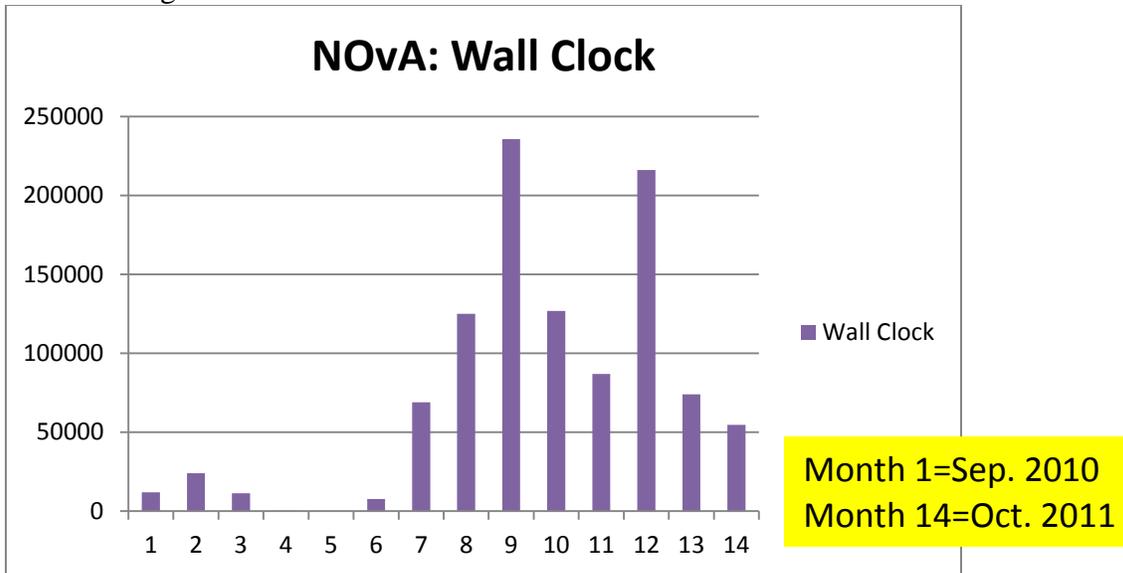
MINOS Monthly average 461 kHrs/month (614 filled-slot average)



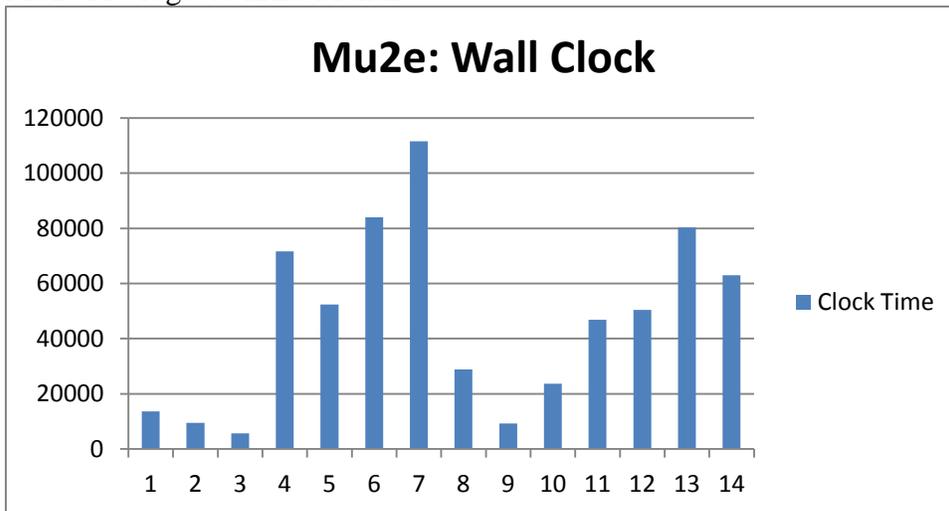
Minerva Average 121 kHrs/Month



Nova Average 75 kHrs/Month



Mu2e Average 46 kHrs/Month



LBNE 1 kHrs/Month

