

FermiGrid Bluearc Unmount Task Force Report

Avoiding direct User access to Bluearc Data on Fermigrid

Version 1.0 2015/04/29

CS Liaisons - please review this and comment.
Contact any Task Force member or kreymer@fnal.gov
or comment directly in
<https://cdcv.s.fnal.gov/redmine/issues/8157>

Charge

To write a proposal including straw man schedule for a staged unmounting of bluearc disks from the Fermigrid worker nodes.

The motivation is :

- to improve application portability to general OSG sites
- to reduce Bluearc server overloads affecting Fermigrid

The proposal needs to include a feasibility study of which file systems can be unmounted and the impact to both the user groups and service providers. (e.g., grid and cloud services, data management, network storage).

Additionally, it should include an analysis of the frequency of the problems over the last 6 months and the metrics we need to demonstrate that this approach would make offline computing significantly more robust.

Background

Bluearc is a proprietary, high performance NFS/CFS/FTP server used widely for core computing services at Fermilab. See performance details in the Appendix

There are four major storage areas, the last of which is our primary focus here.

Small projects not needing dedicated space work under the /grid directories. Larger projects have their own app and data volumes.

- /grid/fermiapp - 2 TB

 - shared software, and small client software

- /grid/data - 27 TB

 - data files for smaller clients

 - legacy files before /<project>/data areas existed

- /<project>/app - 41 TB

 - software

- /<project>/data (and similar) - 1600 TB

 - non-executable data files

App areas are typically accessed directly, and have not been a performance problem.

/grid/fermiapp supports smaller clients, and the small parts of large client software appropriate for this area, such as tools to establish a working environment for users.

A /grid/app area is writable from worker nodes, for use by external OSG users. While a technical risk, usage has been light, there are no known overloads due to /grid/app, and it is not discussed further in this report.

The data areas are intended for 'project' files, which are relatively volatile and are not archived to tape with the usual Fermilab dCache/Enstore system. Bluearc has quotas, and files are not removed without user intervention. Until the 2014 deployment of dCache scratch pools, Bluearc was the primary resource for Intensity Frontier clients' non-archived files. Many of these files are being moved to dCache now, initially concentrating on managed production files.

There remains a need for Bluearc style storage, with quota, persistence, and low overhead for cases needing many small files.

Bluearc files are served by 'head's. These can become overloaded when too many simultaneous unique transfers are attempted.

We have reduced the rate of overloads and their impact by placing data and app areas on separate heads, and by requiring use of a software layer called ifdhc, to regulate access to data files. Use of ifdhc or its predecessor cpn have been required since 2009.

See the performance documentation later in this report.

The only allowed direct access to data has been to Auxiliary highly shared files, for which alternate access methods are now available.

Proposal

DATA

The actions planned now for the Data areas are:

- move /grid/data to the data head (transparent)
- remount /*/data areas on hidden paths

/grid/data was not moved to the Data Bluearc head with other data areas in 2013. This should be done to further protect App areas.

While we learn to make effective use of newly deployed dCache scratch areas, and for use cases which are not a good fit for dCache, we continue to need Bluearc project areas for unmanaged user analysis files.

To improve portability to general OSG sites, and to avoid overloaded servers from unregulated client access, we require all access to Bluearc to go through the ifdhc layer. Any direct access (cd to a directory, read or write directly) should fail.

This has been the policy since 2009, enforced by manual intervention to cancel offending jobs. We have looked the other way for specific cases of highly shared Auxiliary data files, but now are testing alternatives for these use cases.

We will remove the risk of server overloads by removing the traditional /*/data type mount points on Fermigrid GPGrid worker nodes. This will be done project by project. New projects will not have /*/data mounts established.

Until the present FTP servers have appropriate Service Level Agreements established, and to minimize risk, we will initially move Bluearc data to and from Fermigrid via hidden NFS mount points. This lets us proceed immediately while evaluating longer term alternatives. Smaller clients can dismount completely, if their volume is low. This is all transparent to the users.

When appropriate FTP servers are deployed, the hidden NFS mounts can be transparently removed.

APP

No immediate action is planned for the app areas.

- slowdowns are rare
- OSG deployment provides a good incentive to use cvmfs

Strawman Schedule

- 2015-02-19 - ifdhc v1_7_2 current, supports hidden mounts
- now - stop mounting new projects data disks on Fermigrid workers
- now - move /grid/data from the Bluearc app to data head
- to schedule - unmount each project as authorized by Liaisons

Impact Statement

For processes using ifdhc, there will be no impact, and a net benefit due to reduced overloads. The hidden NFS mounts provide the same data rates and availability as now, but with protection from overloads. When supported ftp servers are deployed that are well matched to the workload, there be a further positive impact on user jobs.

Jobs not using ifdhc will fail, as desired. We think all of these use cases are handled by one of the Auxiliary file tools presently being tested. This includes appropriate use of CVMFS for files that behave like code (as small as typical shared libraries and all jobs read the same files), and Alien Cache CVMFS for cases that do not behave like code. We continue to explore other possibilities for these Aux data files including solutions based on xrootd.

Small clients not using Aux files can immediately have their data files unmounted, as there should be no impact.

Medium clients are nearly ready to move to hidden mount points, as they are well aware of these issues. Details will be determined by the Liaisons.

Large impact clients need to change their access models, as they are not presently using ifdhc. These are generally smaller volume clients.

IMPACT	PROJECTS
small	coupp, ds50, e906, gm2, ilc, lariat, lbne, mu2e, nusoft
medium	argoneut, genie, lar1nd, minerva, minos, nova, numix, uboone
large	accelerator, d0, cdms, des, mars*, miniboone, patriot (/grid/data)

APPENDIX

Problems and Performance

Bluearc performance is monitored in detail both internally and on clients.

We have logs, plots and alarms for

- Server open files
- Server loads and performance metrics
- Fermigrid client open files
- Client data rates for the major file systems and hosts
- Gridftp server availability

There are a variety of root causes of overloads. The classic symptom is that the Bluearc head gets very busy (high RBF values, RBF being an internal Bluearc metric, Running Bossock Fibers). At this point the head can detach from the network, causing NFS mounts to go stale or readonly.

This is usually due to user scripts on Fermigrid not making use of ifdhc layer to access Bluearc. Frequently this is due to use of old scripts that pre-date ifdhc.

- Accelerator Division / APC has a job structure that writes directly to /grid/app and /grid/data, they need to be onboarded to jobsub/ifdhc otherwise their jobs will break.
- CDMS has been flagged for heavy direct bluearc access, they need to be onboarded as well.
- Miniboone runs its own non-Fermigrid farm, and on Fermigrid without using jobsub, and does not use ifdhc.
- Some of the MARS groups (marslbne we think) do some direct bluearc writes.

Historical summary

DATE	SLOW RATE	COMMENT
pre 2009/08	5%	d0ora2 Oracle RMAN hardware errors d0 unregulated access
2009/11	0.01%	separation of Minos and d0 data
2011/06	1%	3 months during Minerva startup
2012/11	2%	1 month during mu2e and lbne startup
2013/03	.01% app 0.5% data	data/app heads were separated
2015/03	0.05% data	better since Oct 2014

Performance Metrics

The present metrics are linked to from
<https://cdcvs.fnal.gov/redmine/projects/fife/wiki/FGBMon>

The principal metric we will use to measure availability is the fraction of time under 5 MB/sec as summarized at
<https://cdcvs.fnal.gov/redmine/projects/ifdhcpn/wiki/PERFORMANCE#section-2>

Historically, the APP failure rate has been well under 1/10000, 0.01% , and we would like to see similar numbers for the DATA areas, presently slowing at around 5/10000, 0.05%

Storage System summary

Bluearc is a low latency robust file server, accessed via NFS, CIFS or FTP. Aggregate throughput to the present data areas is over 1/2 GByte/second. It has a limited ability to handle large scale parallel access, which we have regulated with ifdhc locks.

dCache is a moderate latency system with built in regulation of parallel file access. It supports both tape backed and scratch storage, with file aggregation for archival of small files. It has a high throughput, 1GByte/second per pool, and multiple pools, with the network setting the practical limit. Protocols include DCAP, NFS 4.1, WEBDAV and xrootd. File access is limited to file creation and reading, no appending or modification.

In February 2015 the experiment-specific FTP servers moved to 10 GBit hosts, which are a better match to the Bluearc servers than the former hosts, which had limited capacity and support.

Fermigrid Bluearc Mounts

This is a summary of GPGrid Bluearc mounts as of 2015 March 20

Project	app TB	data TB	IMPACT	AUX	comments
argoneut	2	35	M	F	
cdf			S		code is not in BA no data mounts on Fermigrid
coupp	1	12	S		
d0		1	L		prj_root on d0grid, clued0 78 TB Bluearc 466 TB non-Bluearc
des			S		sim = 24 TB orchestration = 100 TB des20/21 are not BA
ds50	1	25	S		
gm2	4	21			
ilc			L		accelerator = 1 TB ilc4c = 4 TB sid = 5 TB ild = 5 TB
lariat	2	8	S		
lbne	2	30	S		data2 = 30 TB
minerva	2	240	M	F	data2 = 50 TB data3 = 25 TB
miniboone		140	L		mounted on mbdata01-08
minos	7	236	M	F,L,T	
mu2e	1	70			data2 = 10 TB
nova	10	140	M	F,L	ana = 95 TB prod = 100 TB
nusoft	2	25	L		shared app hosts CWS web pages (seaqwest)
e906	1	3	S		
uboone	3	52	M	F	
lar1nd	1		M	F	
grid	2	27	L		app mount is /grid/fermiapp

IMPACT levels indicate the impact on the specific project(s)

Small - already using ifdhc, not using Aux files

Medium - already using ifdhc, need Aux file plan/test

Large - not using ifdhc

AUX highly shared files include

F - beam flux

L - library

T - template

Flux files model neutrino beam characteristics, and are used by all neutrino beam projects in their simulations.

Library files are used in analysis jobs to identify particles by matching to event data. Shared files can be up to 100 GB, sometimes loaded into memory for speed. Mainly used by Minos and NovA.

Template file collections are usually under a few GBytes, used in various ways by analysis jobs.

In some cases most efficient running requires shared cache on worker nodes. Existing direct access provides this via local worker memory cache and Bluearc head caching. But this does not work on OSG, and opens the door to overloads when non-shared files are accessed. We think AUX file tools presently under test will provide good performance and OSG readiness. See the Impact Statement above.

Working Documents

The work of the task force is primarily contained in the Redmine Wiki, <https://cdcvs.fnal.gov/redmine/projects/fife/wiki/FermiGridBlue>

The task force is focused mainly on preparing the Strawman plan, and providing pointers to supporting documentation.

Since there is a nearly complete overlap with FIFE membership, we have had short meetings immediately after FIFE weekly meetings as needed.

Redmine Issues track detailed technical discussions.

Under <https://cdcvs.fnal.gov/redmine/issues/>

- 7100 - Overall milestones

- 7121 - Move /grid/data to Data head

- 7122 - Locks based on Bluearc disk

- 7123 - Alternate data mountpoints

- 7124 - GridFTP scaling

- 8157 - Computing Sector Liaison comments

Task Force Membership

MEMBER	ORGANIZATION
Gerard Altayo	DCSO
Dmitry Litvintsev	DMS
Marc Mengel	SDPS/DMA
Andy Romero	ESO/SVS
Marco Slyz	SDCS/USDC
Steve Timm	ECF/VFP
Matt Tamsett	NOvA
Arthur Kreymer (chair)	SDPS/DMA