Data Management Requirements for Intensity Frontier Experiments (DRAFT)

CD/REX Department

v0.2

9 Nov 2010

Introduction

We outline here the Data Management requirements for the IF experiments. This includes file systems used by these projects, and the associated meta-data, planning and management tools.

The issue of file delivery to batch and Grid processes is discussed separately in the Data Handling Requirements document.

Storage Elements

We give a brief outline of storage systems presently in use or investigation.

Local files

Interactive and local worker nodes may have of order 200 GBytes per core of local disk storage.

Grid works nodes have less, under 50 GBytes per core.

These are needed for low latency, high rate, high reliability activities.

Local scratch areas are not archived.

AFS

Use of AFS is primarily historical. Minos has moved its software releases and data handling to Bluearc. User login areas are still in AFS for most IF user accounts, but this could change.

AFS is still the primary location for the CD Web Servers. Bluearc is also supported, and we are like to shift there soon. Perhaps to /grid/fermiapp/www/...

Access in authorized via AFS tokens.

Available anywhere on the network, but not on Fermigrid.

Need service level 24x7, due to login usage.

Only the login areas are archived.

NFS

The GPCF has about 20 TBytes of high performance scratch space, intended to replace local disk storage.

Available only withing GPCF.

Access is authorized via Unix UID/GID.

Need service level 8x7, consistent with GPCF support.

Local scratch areas are not archived.

Bluearc

The bulk of IF disk storage is in Bluearc. We have deployed 10 to 100 TBytes per experiment.

Bluearc is a proprietary high performance NFS server.

These systems are available throughout Fermilab, including most Fermigrid computing elements.

This has become the primary working space for IF experiments.

Software is installed in 'app' areas, mounted readonly and executeable on Fermigrid nodes.

Data is stored in 'data' aread, mounted writeable and non-executable on Fermigrid nodes.

Access is authorized by Unix UID/GID, and in the future via NFS 4.1 ACL's.

Scaling is controlled by requiring use of 'cpn' for Posix access, and via SRM for ftp access.

Bluearc files are available on the grid via Grid ftp.

Need service level 24x7.

The Bluearc areas are not archived.

DCache

DCache is needed as a read and write buffer between the user and the Enstore archives.

We need dedicated write pools for raw data, to provide a tapeless data path for analysis.

The DCache read pools are presently about 15 TB Public and 35 TB for Minos.

DCache is supported directly on the Femrilab site, and via kerberized and Grid ftp remotely.

Scaling is managed internally by the DCache servers.

Need service level 8x7

Enstore/tape

All permanent archives are via the Fermilab Enstore system, backed primarily by LTO-4 tape.
We duplicate critical files (raw data and important archives) in physically separate locations.

Enstore is needed only on select hosts, as most access is through DCache.

Need service level 8x7 (24x7 is being provided)

Investigations

We are participating in LCG storage studies, and the Fermilab Grid Storage Investigation. File systems being studied include AFS, GPFS, Hadoop, Lustre and xrootd .

Meta-data

File system

The great majority of file transactions are by individual analysis users. They work with what is provided by the file system : file names, paths and sizes.

SAM

Minos, Minerva and Nova are using SAM metadata for officially managed files.

All SAM users (CDF, D0, Minos, Minerva, Nova) use the same back-end Oracle schema.

Typical file meta-data include storage locations, dates, sizes, checksums, parentage, data streams/tiers/types/families, and application names and versions

Parameters can be added without making a schema change, but with caution. Each distinct parameter value creates a new row in the database. So parameters are good for things like a short list of strings, but not for floating point values. For example, Minos Monte Carlo data uses parameters to describe the Beam configuration, particle flavors, software release and vertex regions. Each of these have only a few possible values.

Luminosity

Historically, associating luminosity with files is a difficult task, handled differently by each experiment. There are basic hooks for this in SAM, but the data is handled differently by each experiment.

Data Quality

File listings

The Enstore system provides a daily complete file listing, needed for making global file scans. These listings are necessary to reduce the load on the backend databased from user listings

Minos creates frequent file usage summaries for its AFS and Bluearc systems, to discourage users from doing full file scans.

We need easy to use methods of generating relevant file listing. Examples include 'sam translate constraints' at the command and program level, and various Web data browsers, mostly using SAM metadata on the backend.

We need permanent named collections of files, such as the dataset snapshots used by SAM projects.

Monitoring

Performance

We need to actively monitor reliability and data transfer rates from each storage element. This is being done for Minos Bluearc systems. We should add this for the other systems.

Quotas

User quota are important for shared files systems written by individual users. This is both to prevent denial of service, and to track usage.

Summaries

It is important to track overall file system usage by user, project, and data category.

This is critical both for operations, and for capacity planning.