Journal of Physics: Conference Series

PAPER • OPEN ACCESS

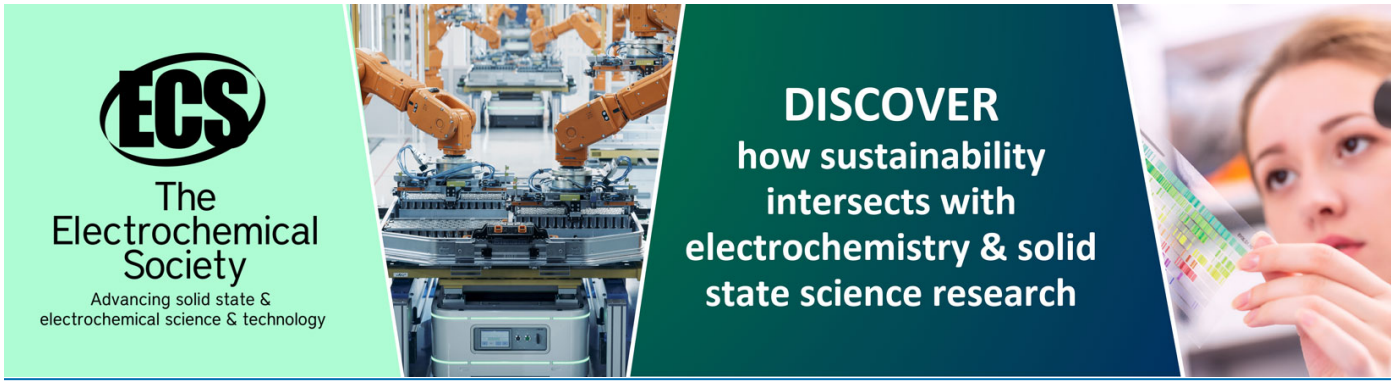
Archiving Scientific Data Outside of the Traditional HEP Domain, Using the Archive Facilities at Fermilab

To cite this article: A. Norman *et al* 2015 *J. Phys.: Conf. Ser.* **664** 042039

View the article online for updates and enhancements.

You may also like

- Mixed higgsino dark matter from a large SU(2) gaugino mass
Howard Baer, Azar Mustafayev, Heaya Summy et al.
- BEAMING NEUTRINOS AND ANTI-NEUTRINOS ACROSS THE EARTH TO DISENTANGLE NEUTRINO MIXING PARAMETERS
Daniele Fargion, Daniele D'Armiento, Paolo Desiati et al.
- Experiences with permanent magnets at the Fermilab recycler ring
James T Volk



This content was downloaded from IP address 18.191.240.243 on 27/04/2024 at 03:03

Journal of Physics: Conference Series **664** (2015) 042039 doi:10.1088/1742-6596/664/4/042039

Archiving Scientific Data Outside of the Traditional HEP Domain, Using the Archive Facilities at Fermilab

A. Norman[1], **M. Diesbug**[1], **M. Gheith**[1], **R. Illingworth**[1], **M. Mengel**[1]

[1]Fermi National Accelerator Laboratory, Bativa IL, USA

E-mail: anorman@fnal.gov

Abstract. Many experiments in the HEP and Astrophysics communities generate large extremely valuable datasets, which need to be efficiently cataloged and recorded to archival storage. These datasets, both new and legacy, are often structured in a manner that is not conducive to storage and cataloging with modern data handling systems and large file archive facilities. In this paper we discuss in detail how we have created a robust toolset and simple portal into the Fermilab archive facilities, which allows for scientific data to be quickly imported, organized and retrieved from the multi-petabyte facility.

In particular we discuss how the data from the Sudbury Neutrino Observatory (SNO) for the COUPP dark matter detector was aggregated, cataloged, archived and re-organized to permit it to be retrieved and analyzed using modern distributed computing resources both at Fermilab and on the Open Science Grid. We pay particular attention to the methods that were employed to uniquify the namespaces for the data, derive metadata for the over 460,000 image series taken by the COUP experiment and what was required to map that information into coherent datasets that could be stored and retrieved using the large scale archives systems.

We describe the data transfer and cataloging engines that are used for data importation and how these engines have been setup to import data from the data acquisition systems of ongoing experiments at non-Fermilab remote sites including the Laboratori Nazionali del Gran Sasso and the Ash River Laboratory in Orr, Minnesota. We also describe how large University computing sites around the world are using the system to store and retrieve large volumes of simulation and experiment data for physics analysis.

1. Overview

Large and small scale experiments in the HEP and Astrophysics communities are producing petascale datasets need to be archived, managed and retrieved processing and analysis. The organization of these datasets is as they are produced are typically dictated by technical constraints of the data acquisition hardware and by the model that the experiments have used to describe their data. These organizations can very widely in file size, structure, format and innate complexity and are often tied in a key manner to the underlying physics goals of the experiment. This native organization of the is often not well suited to efficient storage, searching and retrieval of the data. More over the interconnections that exist between different elements of the data may not be well suited to being preserved when interacting with or storing information in modern mass storage systems. These issues of complexity and organization, when storing or accessing


Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

data in a large storage facility can be a significant roadblock to individuals that are not experts in the organization and operation of large scale data management systems.

To address these issues of complexity and organization, there needs to be a general strategy for performing the mapping of both traditional high energy physics data as well as non-traditional HEP and astrophysics data into the storage systems and management layers that are used to access the storage. This strategy needs to applicable to a wide variety of experiments and must scale to meet their needs both in terms of data volumes, but also in terms of available manpower and automation.

To address the above issues we have developed one such model which is capable of performing the mapping into storage based on user defined rule sets and have produced a set of tools which facilitate storage, cataloging, and access to data in a large multi-petabyte data storage facility.

2. Data Models

Traditional HEP experiments, both collider based and fixed target bases, have long used a well defined "Run/Event" to define and organize their data. In this model, as shown in FIG. 1, relies on a hierarchical arrangement of information that originates at the top level with the base configuration and conditions that define a "Run" and then subsequent conditions which are able to divide a run logically into discrete subsets or intervals which can be considered "subruns". Below the subrun level, the information from the detector is traditionally organized as a set of discrete and independent "events" which correspond to some triggering condition that defined the readout and recording of the actual hit data from the detector.

This modular hierarchy maps well into a simple file based schema for recording the data, where each run or subrun maps into a file or set of files which can be logically associated by their parentage. This also allows for individual events to be stored sequentially within each run/subrun file and allows again for the simple association of the events through their owning parent. This also results in a reduction in complexity across the system, since in this data model the encapsulation of the events within the run or subrun objects removes the bulk of the accounting that would be required if the events were treated as independent storage objects.

In contrast non-traditional HEP and astrophysics data uses many different schemes to record their data that do not inherently lend themselves to the run and event model. In particular some experiments take data in the form of discrete digital image data, like those shown in FIG. 2, while other experiments collect similar information as a series of discrete but linked images, in much the same manner as a motion picture film. Others collect time series data which are represented as long continuos waveforms like the one shown in FIG. 3 for the Darkside experiment. Still others, like the NOνA experiment shown in FIG. 4 combine all of these to create a true continuos readout which can be visualized as a series of discrete images, but where each hit is actually a waveform readout.

The organization of these types of non-traditional readouts can sometimes be mapped into a very loose "run" model, but more often is related more closely to a generic "time window". The real problem of these types of data models are that instead of encapsulating the the data into large container, they instead result in a large number of individual objects (e.g. data images) that are stored separately but need to in some way be associated together as a logical collection and need to be similarly managed as a unified collection.

At the data acquisition level, for many of the experiments we encountered, this had been accomplished through the creation of a series of different files corresponding to the configuration data, condition and state data, readout summary data and then the individual digital image data organized in a complicated treed directory hierarchy. This object collection, as depicted in FIG. 5, would then be repeated over and over with additional meta information about the state that it was taking in encoded in the directory structure of the file system on which the data was being stored.

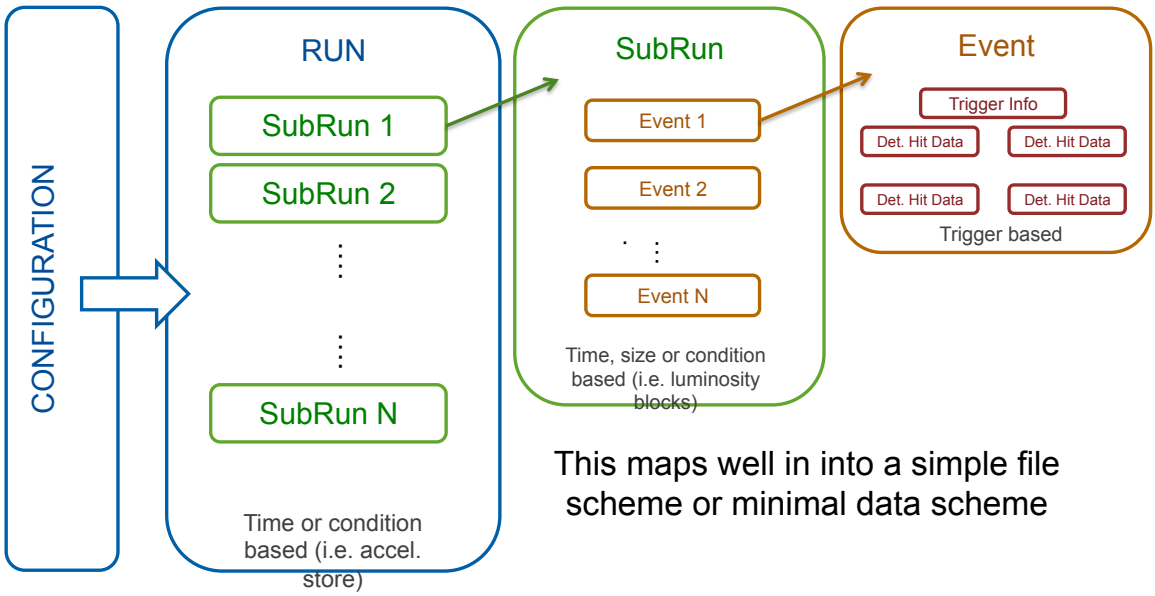


Figure 1. Schematic depiction of the Run/Event model used for the organization of data in traditional HEP experiments. The model relies on a hierarchical arrangement of information propagating from the base configuration down to the individual hit data.

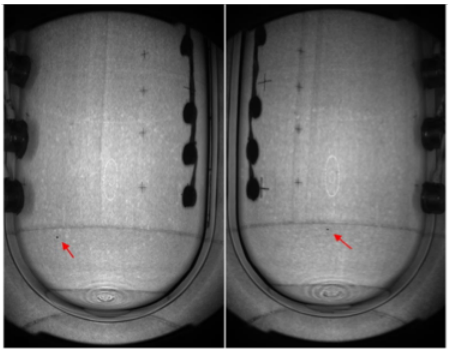


Figure 2. Digital image readout from the COUPP bubble chamber experiment

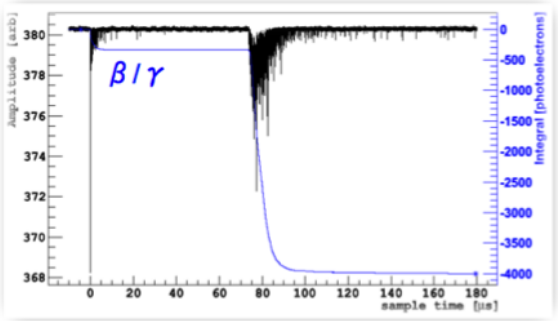


Figure 3. Continuos waveform readout of an event from the Darkside experiment.

The difficulty in mapping this type of data into a mass storage and data management system, is that all of the associations between the individual files need to be preserved for the data to be of value, but there can not be an implicit reliance on a file system implementation to carry these associations due to the need for the data to be portable between storage elements. The other difficulty in working with data organizations of these types are that the scale of the migrations and mappings that need to occur between the DAQ environment where the data was collected and the storage environment where it is to be archived, need to be scalable such that they can easily handle the importation of 400 thousand or more digital images or other large collections of individual objects. More over there is a need for an efficient organization of the data based on its characteristics which is completely agnostic to the actual underlying storage. As a result the model that we have adopted for working with these types of data sets is one of a cataloged, hierarchical, pseudo-object store.

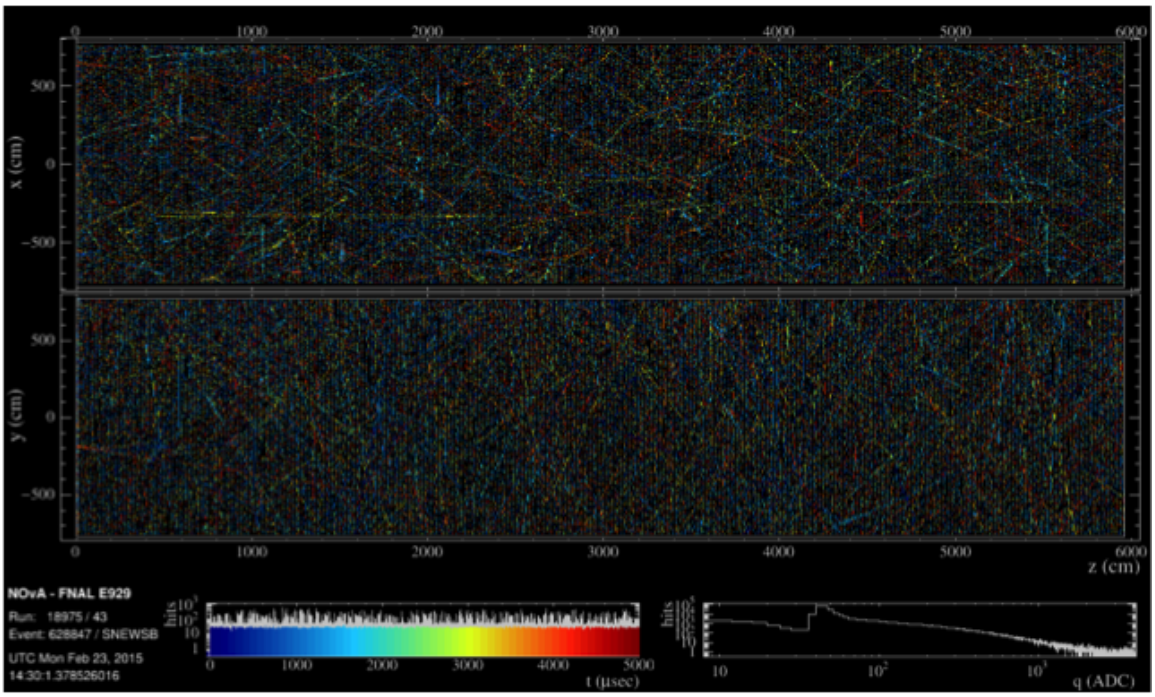


Figure 4. A 5 ms time window frame of the continuous readout of the NOνA detector. Each individual hit is a digitized waveform.

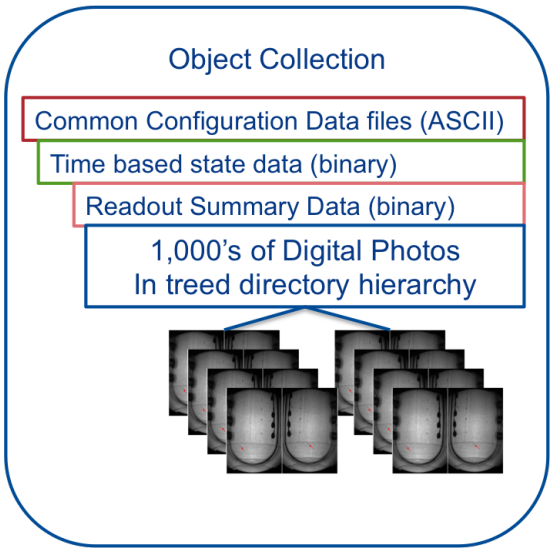


Figure 5. Schematic depiction of the collection of objects that would constitute the data that needed to be stored and associated for a non-traditional HEP experiment.

These types of storage and data management are difficult for normal physics who are more focused on the actual data taking and analysis to deal with effectively. As a result our goal in designing the portal to the Fermilab archive facilities[1] was to design:

(i) A simple method for data to enter the storage facility without knowledge of the details of the facility's operations

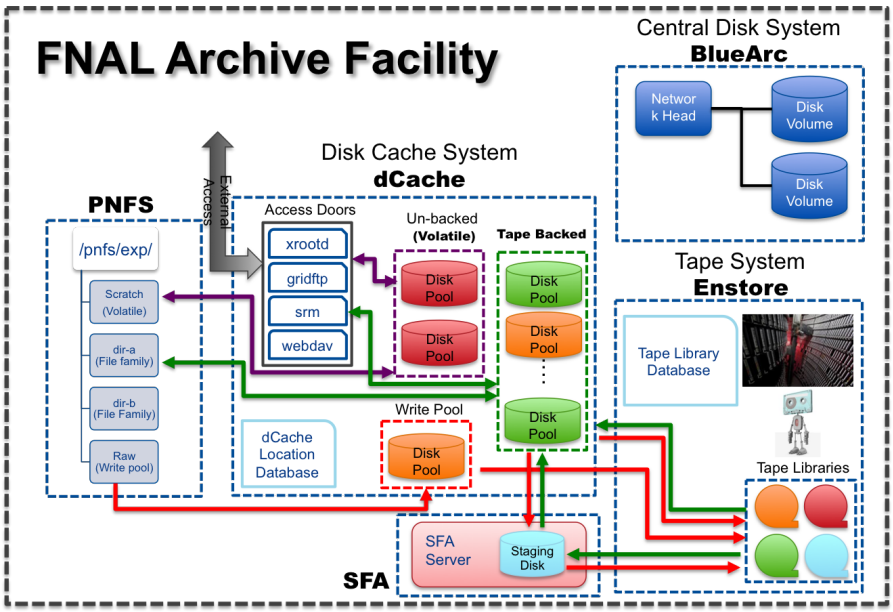


Figure 6. Schematic layout of portions of the Fermilab archive facility.

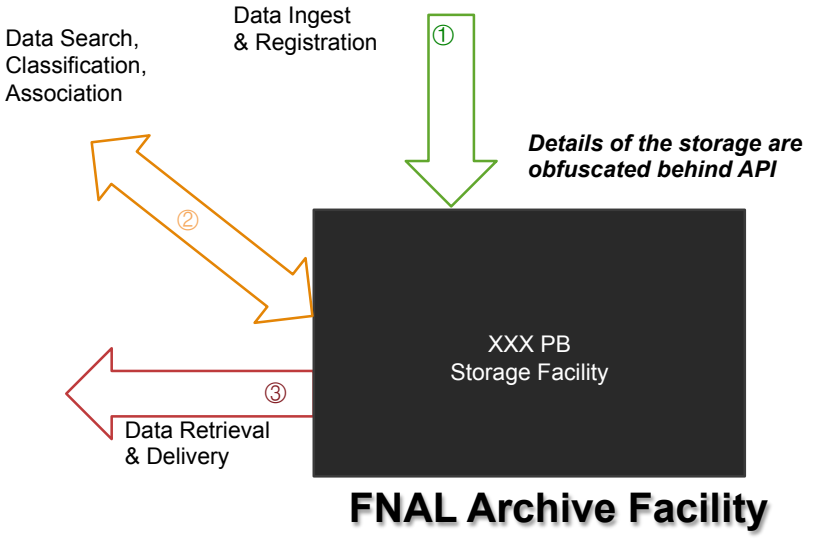


Figure 7. Abstraction of the Fermilab archive facility as a black box for the purposes of user interaction and data storage.

(ii) The ability to specify meta information which could be attached to the data to provide descriptions, associations and hierarchical relationships between data objects

(iii) The ability to locate/retrieve the data from the storage systems based on metadata instead of knowledge of the storage facilities

(iv) A mechanism for delivery of the data to a user specified location or locations which would reconstitute everything that was needed for analysis

Where each of these items could be automated, easy to use and would scale to the appropriate level to handle the data that was presented to us.

In our model we wanted to hide the details of the archive facility shown in FIG. 6 with the

abstraction and simplification of the facility as a black box with three generalize interaction paths which were described by a simple API as shown in FIG. 7.

3. Storage Tool Set

To implement the storage model described in section 2, a data handling toolset consisting of three main components was developed and then adapted to fit the needs of the experiment whose data needed to be archived into the Fermilab facilities.

The first component of the tools set, the Fermi File Transfer Service (F-FTS) was designed to provide the data injest path to the archive facility shown in FIG. 7. The F-FTS provides a simple interface for reliably transferring data between sites, registering data with a centralized data catalog and replica catalog and for injecting data into the storage facility with verification and validation of the successful archiving of the information. The F-FTS supports arbitrary data types and supports both standard and fully customizable metadata generation through a modular python based plugin mechanism. The system operates as an asynchronous client side daemon that is capable of watching specified areas on the client's file systems and then pushing new files to the archival storage systems.

The F-FTS presents experiments with a simple "dropbox" interface to the storage and file catalog systems. With this interface all the experimenters have to do is place a file (or set of files) in the dropbox location and a set of customizable rules engines examine the files and perform the proper association of metadata information and initiate the proper transfer protocols to transport the data from the client, into a location on the mass storage systems. The rules engines control the mapping of the data between the locations and permit complicated reorganization and aggregation of the information to allow for mapping of legacy data structures and data collections into formats and arrangements that are more suited to the storage destination.

In addition to load balancing/throttling the declaration and transfer of data between sites the F-FTS also provides transfer queues to deal with high latency WAN connections, and provides a multi-hop functionality to permit the chaining together of multiple F-FTS systems to navigate site boundaries, firewalls and other restrictions on networking that may be required to establish the full pathway between the data acquisition domain that the experiment is using and the storage domain where the archival systems are homed. The F-FTS also provides automated cleanup of the input areas after the files have been successfully transfered and verified. This permits F-FTS systems to run in an almost completely autonomous mode and service large ongoing and repeated data transfers.

The F-FTS was successfully used with the data from the COUPP experiment[2], which is a bubble chamber based dark matter search running at SNOLAB. The F-FTS was configured to recursively descend the COUPP data hierarchy to parse out and record all of the configuration information, collection date and time information, data subsets and individual events. These data were stored in a combination of text based files, binary data and bitmap images as well as as metadata stored in the file system paths of the DAQ systems where the data was collected. From this information a set of consistent metadata was defined for each file which allowed the associations inherent in the original directory structure to be reconstituted both through queries to the data catalog as well as during the data retrieval. The actual data files were then both aggregated by the F-FTS into higher level container objects and renamed to present a data elements that have a unique identifier across the namespace managed by the SAM data catalog. Using the above organization COUPP has stored approximately 460 thousand detector events taken on 1005 different days and in 17 different configurations, all of which are fully restorable to their original organization for analysis.

The F-FTS was also used to archive data from the Fermilab Holometer experiment[3]. For this experiment that data is a series of laser interferometer interference patterns that are acquired through a combination of National Instruments PXI controller cards running MS Windows and

Linux operating systems. In this configuration the data volumes on the acquisition machines were exported via NFS and CIFS shares to an offline machine where the F-FTS ran. This configuration allowed the F-FTS to provide archival storage to a DAQ system independent of OS compatibility issues. The F-FTS was configured with rule sets to permit it to properly the gravitational wave frame format (.gwf) and Hierarchical Data Format v5 (HDF5) format files that the data is acquired in. This configuration also allows for independent maintenance of the F-FTS system that does not interrupt or otherwise affect the holometer's DAQ and operations.

For the DarkSide liquid argon TPC based dark matter search [4] being performed at the Laboratori Nazionali del Gran Sasso, the F-FTS was used to transfer neutron veto data from Italy into the archive facility at Fermilab. The F-FTS was configured on one of the DarkSide DAQ machines and from there initiates a gridftp based transfer over the wide area network into the dCache based front end of the Fermilab tape archive. This F-FTS setup was used to transfer over 500 TB of data spread across approximately 100 thousand files between the two sites.

The NOνA experiment has likewise use the F-FTS system extensively in all aspects of operations from the DAQ systems[5] running at the far detector site in Ash River Minnesota, to multiple F-FTS instances handling the output of Monte Carlo simulations, large scale reconstruction and official analysis samples. NOνA has used this system to transfer and archive over 1.6 PB of data across over 12 million files. NOνA has used the system in the unique multi-stage transfer setup to navigate the pathway between the highly isolated far detector DAQ network and the archival storage systems and large NAS volumes at Fermilab. This configuration and the associated rule set was used to automate the creation of multiple replicas of each file through the gateway F-FTS and there by avoid saturating the limited network connections that exist out of the far detector site.

The second component of the data handling toolkit that was developed and adapted to needs of the experiments was the HTTP based API to the SAM data catalog system[6]. This system provides full access to the, storage system aware, SAM metadata and replica catalog. The SAM data handling system that has been described previously [], provides the fully queryable metadata and replica catalog that allows the experiments to map the proper associations between the different files in the data collection and then retrieve the actual location of the files in mass storage systems, without having to know the details of the actual underlying structure of the archive or cache systems that the data reside on.

The last component of the toolkit that was used to facilitate the use of non-tradition datasets that had been imported into the Fermilab archive facilities was the Intensity Frontier Data Handling Client (IFDH) which has been designed specifically to interact with the replica cata to find specific data elements and deliver them through the "last mile" of data movement between the storage facilities and the local storage that exists on the distributed systems that are used for data analysis.

The IFDH tool can move data between arbitrary storage elements between local disks, dedicated site specific cache systems and storage elements, NAS volumes, as well as large cache pools like the Fermilab dCache and the dCache fronted Fermilab tape libraries. IFDH also acts as a protocol abstraction layer and will select the "correct" protocol chain for transporting information between the different storage elements. It is module and by default includes support for gridftp, srm, dccp, AWS S3, local cp and dd operations as well as cross protocol bridging (i.e. transfer to proceed out of Amazon AWS's S3 storage system and into the gridftp doors of the Fermilab dCache pools.

4. Summary

The set of tools that was created to address the needs of modern HEP and non-HEP experiments archive their data using the Fermilab archival storage facilities has been successfully used to import large amounts of data from a while variety of experiments throughout the HEP and

astrophysics communities, ranging from dark matter searches, neutrino oscillations experiments and astrophysics datasets. The tools have been used to not simply store the data taken by the experiments, but organize it and map it into the Fermilab data management systems in a manner that is independent of the actual storage system backends.

These tools allow for the full reconstruction of the associations between data and configuration elements that are present in the experiment's original datasets and allow for retrieval and restoration of individual data elements as well as full datasets from the mass storage systems.

These tools have opened up the use of the Fermilab mass storage and archive systems to a wide variety of experiment both at Fermilab and around the world.

Acknowledgements

The author acknowledges support for this research was carried out by the Fermilab scientific and technical staff. Fermilab is Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy

References

[1] Bakken J, Berman E, Huang C H, Moibenko A, Petravick D and Zalokar M 2003 The Fermilab data storage infrastructure *Mass Storage Systems and Technologies, 2003. (MSST 2003). Proceedings. 20th IEEE/11th NASA Goddard Conference on* pp 101–104

[2] Bolte W, Collar J, Crisler M, Hall J, Krider J *et al.* 2006 *J.Phys.Conf.Ser.* **39** 126–128

[3] Kamai B, Chou A, Evans M, Glass H, Gustafson R, Hogan C, Lanza R, McCuller L, Meyer S, Richardson J *et al.* 2013 The Fermilab Holometer: Probing the Planck scale *American Astronomical Society Meeting Abstracts# 221* vol 221

[4] Baudis L *et al.* 2012 Darwin dark matter WIMP search with noble liquids *Journal of Physics: Conference Series* vol 375 (IOP Publishing) p 012028

[5] Zálešák J, Biery K, Guglielmo G, Habig A, Illingworth R, Kasahara S, Kwarciany R, Lu Q, Lukhanin G, Magill S *et al.* 2014 The NOνA far detector data acquisition system *Journal of Physics: Conference Series* vol 513 (IOP Publishing) p 012041

[6] Lyon A, Illingworth R, Mengel M and Norman A 2012 *J. Phys.: Conf. Ser.* **396** 032069