Analyzing how we do Analysis and Consume Data, Results from the SciDAC-Data Project

P. Ding[1], L. Aliaga[1], M. Mubarak[2], A. Tsaris[1], A. Norman[1], A. Lyon[1], R. Ross[2]

[1]Fermi National Accelerator Laboratory, Bativa IL, USA
[2]Argonne National Laboratory, Lemont IL, USA

E-mail: {dingpf,laliaga,atsaris,anorman,lyon}@fnal.gov {mubarak,rross}@mcs.anl.gov

Abstract. One of the main goals of the Dept. of Energy funded SciDAC-Data project is to analyze the more than 410,000 high energy physics datasets that have been collected, generated and defined over the past two decades by experiments using the Fermilab storage facilities. These datasets have been used as the input to over 5.6 million recorded analysis projects, for which detailed analytics have been gathered. The analytics and meta information for these datasets and analysis projects are being combined with knowledge of their part of the HEP analysis chains for major experiments to understand how modern computing and data delivery is being used.

We present the first results of this project, which examine in detail how the CDF, $D0$, NOvA, MINERvA and MicroBooNE experiments have organized, classified and consumed petascale datasets to produce their physics results. The results include analysis of the correlations in dataset/file overlap, data usage patterns, data popularity, dataset dependency and temporary dataset consumption. The results provide critical insight into how workflows and data delivery schemes can be combined with different caching strategies to more efficiently perform the work required to mine these large HEP data volumes and to understand the physics analysis requirements for the next generation of HEP computing facilities.

In particular we present a detailed analysis of the NOvA data organization and consumption model corresponding to their first and second oscillation results (2014-2016) and the first look at the analysis of the Tevatron Run II experiments. We present statistical distributions for the characterization of these data and data driven models describing their consumption.

1. Overview

Since the advent of the Tevatron Run II (in experiments such as CDF [1] and $D0$ [2]) and extending through the present day with the neutrino physics program (in experiments such as NOvA [3], MINERvA [4] and MicroBooNE [5]) at Fermilab, detailed information on the organization and use of physics data has been collected by the unified data catalog and data management systems used by these experiments and known as Sequential Access via Metadata (SAM) [6, 7]. The information contained in the SAM databases represents a rich computing ecosystem that describes the different ways that experiments have organized, generated and consumed data as well as how the supporting computing and data handling infrastructure have shaped these characteristics.

The SciDAC-Data project was started to examine this information in order to provide data-driven descriptions of High Energy Physics (HEP) workflows and data management [8]. These

descriptions are serving as the basis of models that are being developed to understand the scaling characteristics of current and future computing facilities. The data and analysis being presented have been converted into generalized distributions that are being used as the inputs to simulations of the Fermilab grid computing environment [9]. The data has also been released as a public dataset which is made available to the HEP and computing communities to aid in similar modeling and understanding of use cases in HEP [10].

2. SAM Data Handling System

The SAM data handling system was initially developed during Tevatron Run II [11] to serve as a unified data catalog/replica catalog that could provide related data files into "datasets" based on metadata selection criteria. The unification of the formal metadata based data catalog, with the location awareness of the replica catalog within SAM, and with a workflow level bookkeeping interface, allows the system to schedule and provide optimized delivery of files to production/analysis level tasks. In particular the system is able to optimize the restoration of data from slow archival media (tape) and manage or interact with disk caching layers in front of the slow archival systems.

Central to this functionality, SAM operates with the concept of a "dataset" defined as a logical selection criteria applied to the metadata associated with each individual file in the data catalog. This selection criteria is then evaluated at run/processing time against the current state of the data catalog to obtain a collection of zero or more files matching the criteria. This collection is referred to as a snapshot and represents the deterministic evaluation of the dataset at a specific point in time.

When an analysis is being performed against a given "dataset", the workflow management and bookkeeping portions of the SAM system perform this evaluation and then start a corresponding "project" that gives analysis clients work. Each client connects to the project and requests "the next file" in the dataset that is available for analysis. In this manner the system does not rely on pre-placement of data (data is instead delivered to the compute element that is to perform the work) and files are delivered in an order that is optimal and meets the requirements for the high latency storage systems (e.g. groups of files from the same physical tape are staged in together and then dispatched to the client compute nodes when ready). This system can fully leverage cache layers and performs predictive pre-staging or "look ahead" to keep the analysis pipelines near full capacity.

2.1. Dataset Organization and Metadata

The SAM data catalog contains a highly extensible schema to allow for the classification and organization of the experiment's data. The general organization of the metadata falls into two generalized categories A) core/physical characteristics and B) physics and user defined characteristics. The core metadata includes such parameters as the creation/start/end times of the files, file sizes, file format, data integrity checksums, data tier to which the data belongs, number of "events" contained in the file and other similar fields. The physics metadata in contrast contains values ranging from the state/configuration of the accelerator beam to information on the Monte Carlo generators used for simulation files. Figure 1 shows the number of custom metadata parameters currently in use by a selection of Fermilab hosted experiments that are actively using the SAM system for analysis or simulation (left) and the number of datasets that each experiment has defined based on these parameters (right).

For the NOvA experiment there are over 140 parameters in this general category that can be selected against when constructing a dataset, CDF has 214 parameters while MicroBooNE, MINOS, MINERvA and Mu2e have less than 100. The $D0$ experiment had over 700 physics parameters that were used by the experiment to describe their data. This relatively high number of parameters in $D0$ in comparison to other experiments reflects the fact that $D0$ integrated many

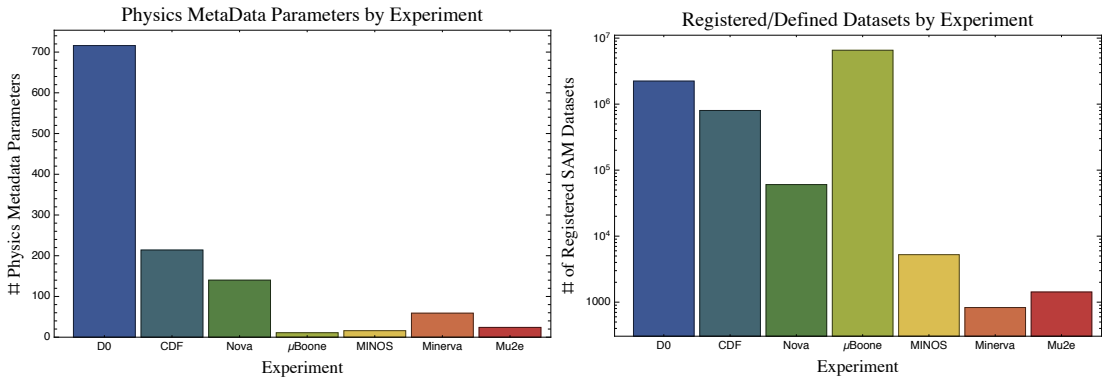

Figure 1. Number of metadata parameters (left) and total number of datasets (right) defined by Fermilab experiments from 2004-2016 (left).

of their MC parameters into the SAM database as metadata and used them for its bookkeeping. However, in practice, the experiments use less than 100 parameters to do their queries.

The number of datasets created by using these parameters ranges from hundreds to millions. A new experiment like MicroBooNE[1], has almost 7 million registered datasets, while long running experiments like CDF and $D0$ have around 800,000 and 10 million datasets, respectively[2]. MINOS and MINERvA, long running neutrino experiments [3], have 5,000 and 800 datasets, respectively. In contrast, a relatively new experiment like NOvA[4] has more than 60,000 datasets. On the other side, Mu2e has around 1,000 datasets [5]. These values show the various stages of SAM catalog and the differences in the treatment of the datasets across experiments. For instance, MINERvA likely makes few production-grade datasets while MicroBooNE has an open dataset creation policy.

The actual fraction of datasets used by each experiment under study is shown in Fig. 2 and it is calculated as the ratio between the number of datasets that have had one or more registered analysis project running against them over the total number of registered datasets for the experiment. Most of the experiments have more than 30% dataset usage fraction. Particularly, CDF and MINOS have high usage fraction (>70%). However, MicroBooNE has a much lower usage fraction ($\sim$ 10%). This reflects the fact that this experiment uses the dataset definition in a way of bookkeeping for different stages of their analysis workflow. Many of the datasets are not directly used for analysis, instead they are used to form higher levels datasets, which in turn are used by end users for analysis.

2.2. Analysis

After an overview of the data organization within each experiment under study, CDF, $D0$, NOvA, MINERvA and MicroBooNE experiments, we analyze how this data is consumed. Figure 3 (left) summarizes the variations in the number of files that datasets contain, as well as the owners of the datasets. Most of the datasets have less 1,000 files, dedicated specially for testing and short studies conducted by the analyzers. The long distribution tail corresponds to different analysis goals when organizing the data. Particularly, datasets with very high number of files relative to average for experiments are likely to be created to contain all data taken or simulated

[1] MicroBooNE was fully constructed and began data taking in July 2015.
[2] CDF started their operations in 1985 and $D0$ in 1992.
[3] MINOS was fully operational in 2005 while MINERvA began taking data in 2009.
[4] NOvA started taking data in their main detectors in 2013.
[5] Mu2e is expected to be commissioning in 2019.

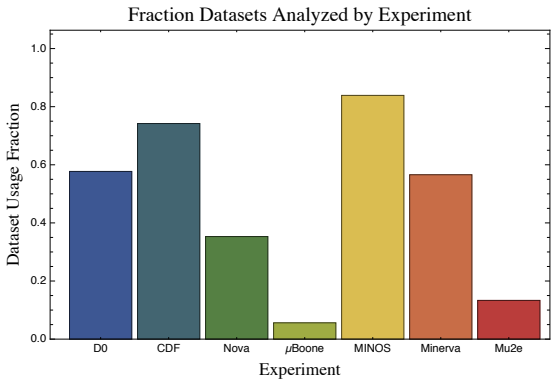

Figure 2. Usage fractions for datasets defined by experiments. The usage fraction is defined as the number of datasets against which there is at least one analysis project divided by the total number of datasets defined by the experiment.

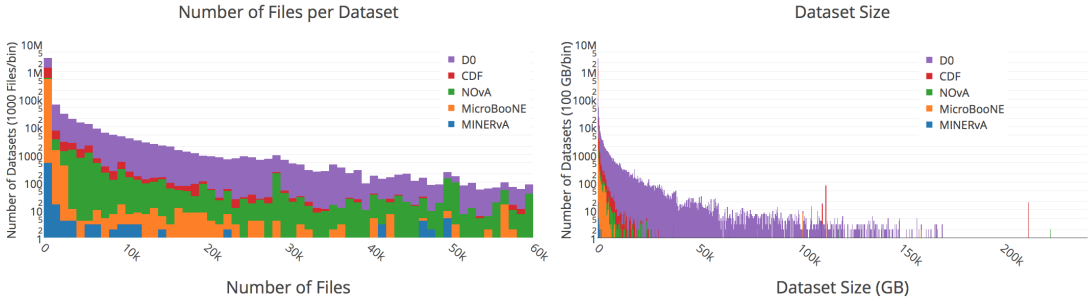

Figure 3. Distributions representing the organization of data within the CDF, *D*0, NOvA, MINERvA and MicroBooNE experiments. The distributions show the number of unique files that belong to each dataset that was defined by the experiments (left) and the total size of datasets defined and used by the experiments (right).

and used for final physics results to be released publicly.

Figure 3 (right) shows the dataset sizes in Gigabytes (GB) defined and used by the experiments. The distributions have a characteristic peak at a few GB and a long tail extended to hundreds of Terabytes (TB). This is related, approximately, to the distribution of number of files per dataset seen in Fig. 3 (left) and to the ways the different data tiers are generated. For example, datasets that correspond to raw data or reconstruction might contain larger amount of information and the resulting outputs, after applying complex algorithms to extract physics variables, go to smaller analysis files which usually have higher reusage comparing to raw and reconstruction datasets.

The dataset size distributions between experiments also varies due to the type of the experiments. While Intensity Frontier experiments (such as the neutrino experiments: NOvA, MINERvA and MicroBooNE) constantly increase their datasets' volume by adding more files for final results, collider experiments (such as CDF and *D*0) generally split their datasets in independent runs and physics channels in their study. Another reason for different dataset sizes within the same experiment can be the experiment lifetime. Typically HEP experiments run for years and there are always upgrades to increase the amount of data that they are taking. This effect will naturally result in larger datasets and the tail in the distribution for Intensity Frontier experiments will continue grow since those experiments are currently taking data.

Another way to look at how the experiments organized their data is looking at the dataset

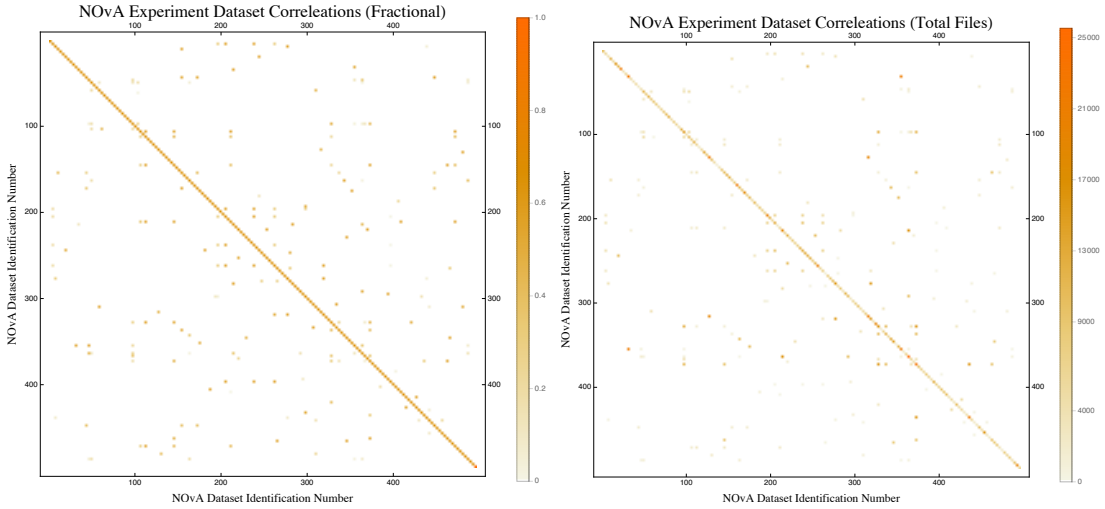

Figure 4. Correlation matrix representing the fractional (Left) and absolute (Right) overlap (in number of files) between 500 most popular datasets for the NOvA Experiment.

correlations. Depending on their definitions, some of the files may be shared between different datasets. Figure 4 shows the fractional (left) and absolute (right) overlap (in number of files) between datasets for the NOvA experiment (an identification number has been assigned to each dataset). We found that the correlation is high for this experiment: 21% of datasets overlap with at least one other dataset, in comparison with other experiments like $D0$ that has only 1.2% correlation. This difference can be explained by the fact that NOvA has fewer parameters than $D0$ and, as was mentioned before, collider experiments generally separate their data in independent samples.

The number of times that a dataset was processed by an analysis application is shown in Fig. 5. The high dataset reusage values in CDF and $D0$ might reflect the fact that those experiments are in their final stage of analysis rerunning against the datasets after optimizing each step of the data and simulation process. Conversely, experiments that are currently taking data are implementing new analysis methods or better reconstruction algorithms and they are likely creating new datasets constantly. This is related to the time between successive analysis projects being started on a dataset that can be seen in Fig. 6. The datasets are highly used in analysis project within a few months. We can also relate these distributions with the effective operational lifetime of the dataset (i.e. time until the dataset is no longer being actively analyzed) shown in Fig. 7.

Figure 8 shows the time series of observed user requested analysis projects of datasets per data tier for the NOvA experiment. The total data considered in this study spans over 1,620 days (from October 2011 to April 2016), however, the figure starts a few days before the beginning of the fully operational main detector (October 2013) [6].

The structure of this time series corresponds to the priority of the file production along NOvA history and illustrates, from the user perspective, the relation between the data tiers and the workflows from unpacking the raw data or MC generation up to the particle identification (PID) stage in terms of the time request for data and MC. The location and sequence of every peak matches different production campaigns. For example, a MC calibration effort (at 980 days) is followed by a peak in the MC reconstruction (at 1,050 days) which, in turn, is followed by a MC

[6] The first part of the distribution (not shown in the figure), from 0 to 600 days, corresponds to the detector prototype data that was used to study the performance of the detector systems.

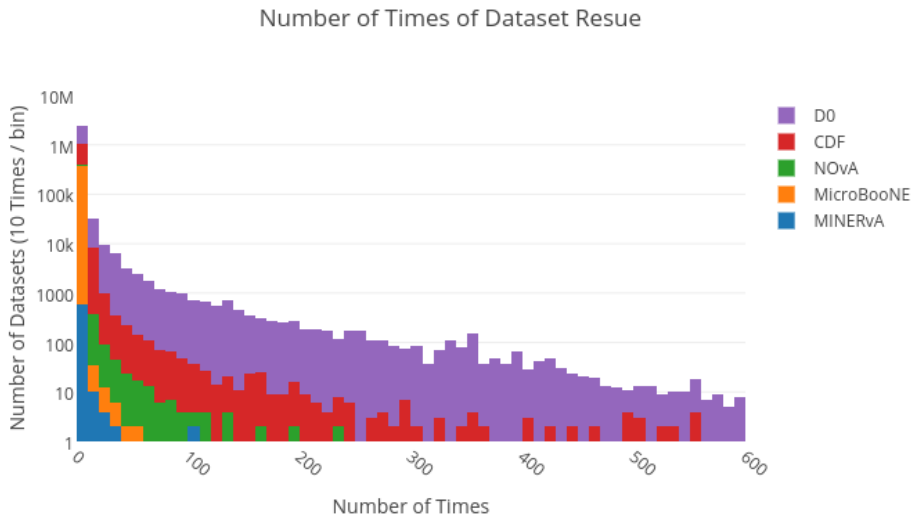

Figure 5. Distribution of the number of times that a given dataset was processed by an analysis application for the NOvA, MINERvA, MicroBooNE, CDF and *D*0 experiments.

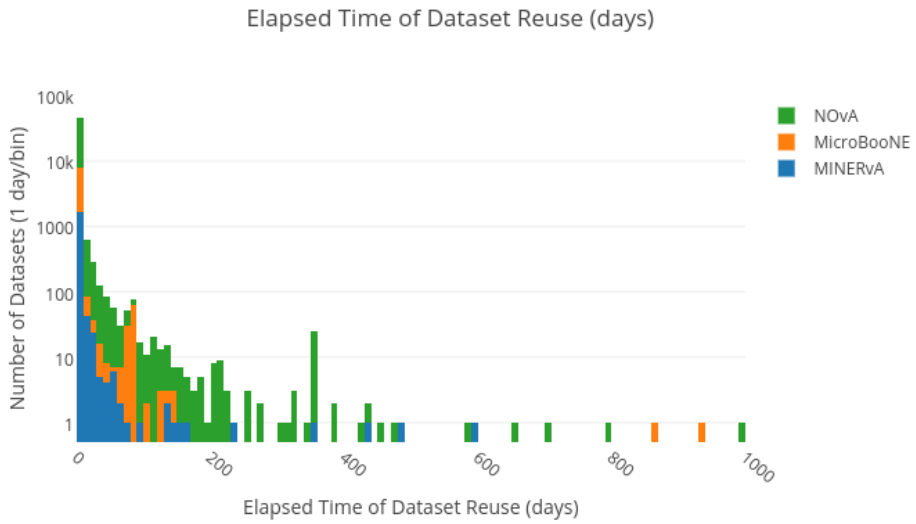

Figure 6. Distribution of the time (ΔT) between successive analysis projects being started on a dataset for the NOvA, MicroBooNE and MINERvA experiments. The distribution is related to the effective operational lifetime of the dataset (i.e. time until the dataset is no longer being actively analyzed).

particle PID stage (at 1,080 days).

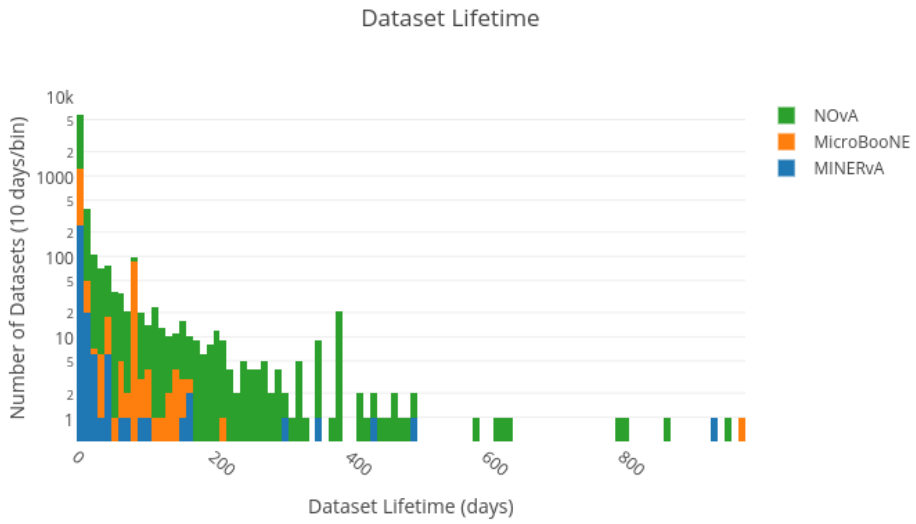

Figure 7. Dataset lifetime for NOvA, MicroBooNE and MINERvA experiments.

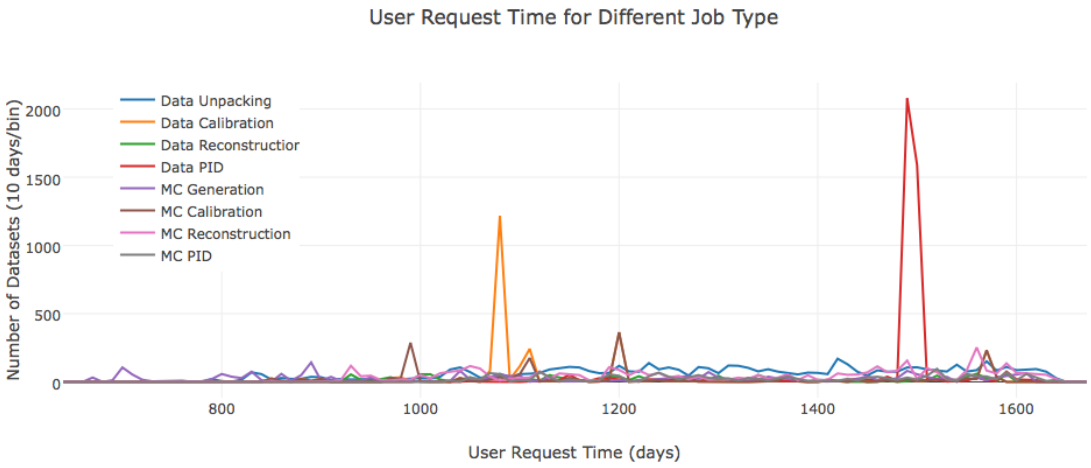

Figure 8. Time series of observed user requested analysis of datasets per data tier for the NOvA experiment: Data unpacking (also known as raw) for measured files or MC Generation for simulated Monte Carlo files, Calibration, Reconstruction and particle identification.

3. Summary

The first results of the SciDAC-Data project have been presented in detail. We analyzed the dataset organization, classification and consumption of the Tevatron Run II and the current neutrino experiments using the information from SAM data handling system. We have been particularly focused on data usage patterns as well as temporary dataset consumption and correlations.

The results of this analysis provide critical insight to improve the Fermilab grid computing environment and it has been used as an input for data driven simulations of the large HEP

computing consumption and infrastructure [9]. This study also helps to understand the requirements for the next generation of HEP computing facilities.

Acknowledgements
The author acknowledges support for this research that was carried out by Fermilab and Argonne National Laboratory. Fermilab is Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy. Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

References

[1] Fermilab The collider detector at Fermilab, url = https://www-cdf.fnal.gov
[2] Fermilab The DZero experiment, url = https://www-d0.fnal.gov
[3] Fermilab The NOvA neutrino experiment, url = https://www-nova.fnal.gov
[4] Fermilab The MINERvA experiment URL `http://minerva.fnal.gov`
[5] Fermilab The MicroBooNE experiment, url = http://www-microboone.fnal.gov
[6] Illingworth R 2014 A data handling system for modern and future Fermilab experiments *Journal of Physics: Conference Series* vol 513 (IOP Publishing) p 032045
[7] Lyon A, Illingworth R, Mengel M and Norman A 2012 *J. Phys.: Conf. Ser.* **396** 032069
[8] SciDAC The SciDAC project, url = http://www.scidac.gov
[9] Mubarak M, Ding P, Aliaga L, Tsaris A, Norman A, Lyon A and Ross R SciDAC-Data, A Project to Enabling Data Driven Modeling of Exascale Computing *(In this volume)*
[10] Fermilab SciDAC data project URL `http://scidac-data.fnal.gov`
[11] Terekhov I 2003 *Nucl. Instr. Meth. Phys. Res. A* **502** 402–406