Scientific Data Management Technology for NERSC

Kurt Stockinger
Lawrence Berkeley National Laboratory

January 2007

Abstract

In this article we summarize the main data management requirements based on a recent NERSC user survey. Next we provide some recommendations about how these requirements can be addressed using existing data management technology.

1 Introduction

The NERSC data management requirements are based on the following information:

- NERSC 2006 User Survey Results: Visualization and Data Analysis [7].
- Analytics User Survey: User Input on Analytics Needs [6].
- ERCAP 2007 Allocation Requests.

The analytics user survey covers the following application areas:

- High-Energy Physics (Accelerator Modeling, Detector Simulations)
- Astrophysics (Supernova Factory, Thermonuclear Supernovae, Cosmic Microwave Background Analysis)
- Fusion
- Molecular Dynameomics
- Material Sciences
- Climate Modeling

First we discuss the main data management requirements that we identified in the user surveys. Next we provide specific data management solutions to address these problems.

2 Main Results of User Surveys

Based on the user survey, the data management requirements can be categorized as follows:

- Data transfer
- Data access and file formats
- Meta-data access
- Query-driven visualization
- Comparative data analysis
- Fault-tolerant workflows

Let us now analyze these requirements in more detail.

2.1 Data Transfer

Many experiments expressed the need to transfer data from supercomputers to mass storage systems. Afterwards the data is transfered to data analysis systems. The amount of data is often in the orders of GB to TB and involves thousands of files. Handling this large number of files is non-trivial due to network failures, limited storage space, different API etc. The main requirement is to (1) manage large volumes of data across HPSS and NGF with the goal to minimize data transfers and (2) to ensure the data is present at the appropriate file system when requested by the analysis job.

2.2 Data Access and File Formats

Different experiments use different file formats to store and access their data. Some experiments use open-source database systems such as MySQL or PostgreSQL, while others use scientific data formats such as HDF, netCDF or FITS. Some experiments store the data in raw binary format and provide some meta-data information about how the data is stored. Due to the large number of different storage solutions and file formats, data often has to be converted for subsequent data analysis and visualization. However, this data conversion process is often very tedious and error-prone since a converter has to be written for basically every data format. The main requirement is to design community specific data formats.

In order to resolve certain small features in the physical domain, adaptive mesh refinement (AMR) techniques are used. A main requirement is to interactively analyze hierarchical data stored in AMR and to perform efficient data subsetting.

2.3 Meta-data Access

Scientific data formats such as HDF and netCDF store the data and meta-data information in a single file. Accessing the meta-data information is efficient if the data is on disk. However, often the data is stored on mass storage systems and thus accessing meta-data information is quite expensive. The main requirement is to separate the data and meta-data information and thus always keep the meta-data on disk. This approach allows fast searching of meta-data information on disk and subsequently only accessing the relevant data on the mass storage system.

2.4 Query-Driven Visualization

Many experiments require querying and analyzing large amounts of data that contain interesting data such as "temperature values < 100". Traditional data analysis and visualization tools would read the whole data set in order to find the regions of interest. This approach is not scalable to data sets that exceed the main memory limitations. The requirement is to use indexing and query mechanisms that quickly identify the regions of interest without the need to read the whole data set. Another requirement is to store the retrieved regions of interest for subsequent interactive analysis.

2.5 Comparative Data Analysis

Scientific data often consists of experimental and simulation data. One of the main requirements is to perform comparative analysis of both experimental and simulation data based on some statistical properties. This approach often requires some sort of machine learning techniques.

2.6 Fault-Tolerant Workflows

Typical experiments often handle complex data flows such as data generation, data transfers, data conversion, data analysis and visualization. Each of these steps might fail due to different errors such as network failures, sever failures or storage limitations. A main requirement is to build a workflow system that allows checkpointing of data flows and thus error recovery without the need to re-run the whole data flow.

3 Data Management Technology

In this section we give an overview of data management technology for solving some of the open problems reported by the experiments. Note that this list is very selective and is based on tools that are mainly used in production by DOE projects.

3.1 Data Transfer

- Storage Resource Manager (SRM) [13]: The SRM manages transfers and access of large files across different storage systems both on disk and tape. The API is standardized for various storage systems. An important feature of the SRM is space management and fault-tolerance.

  SRMs are widely used in the High-Energy Physics community for transferring large amounts of data between different labs and universities across the world. SRMs are also standardized by the Grid community.

- Storage Resource Broker (SRB) [12]: The SRB allows access to files stored an different storage systems and also provides meta-data information.

  SRBs are used in various Grid projects by many different application domains.

3.2 Data Access and File Formats

- Scientific data formats: HDF5, netCDF, FITS.

  These data formats provide high-performance I/O operations for data typically stored as multi-dimensional arrays.

  H5Part [2] is a simplified API on top of HDF5 to provide scalable I/O for MPI codes. Due to the standardized file format, H5Part allows transparent sharing of various data sets. H5Part was originally designed for the accelerator modeling community as a standardization effort on file formats. However, the methodologies are applicable to any scientific domain.

- Open-source database systems: MySQL, PostgreSQL.

  In scientific applications MySQL and PostgreSQL are mainly used for storing meta-data. The base data is often stored in scientific data formats that provide high-performance I/O operations.

- ROOT [9]: ROOT is an object-oriented database system with integrated data analysis and visualization technology.

  ROOT is mainly used in the High-Energy Physics community and currently manages among the largest data sets in the world.

3.3 Meta-data Access

- Metadata Catalog Service (MCS) [5]: MCS is a stand-alone metadata catalog service that associates application-specific descriptions with data files, tables, or objects.

  MCS is widely used in various Grid projects.

- SRB's MCat [4]: MCat is a meta data repository system to provide a mechanism for storing and querying system-level and domain-dependent meta data using a uniform interface.

MCat is based on SRB and is used in various Grid projects.

3.4 Query-Driven Visualization

Query-driven visualization of large multi-dimensional data sets is still in an early stage. Some promising research results can be found in [14, 1].

3.5 Comparative Data Analysis

Research in this field is still in an early stage. Typical techniques for comparative data analysis are independent component analysis. Some promising results have been achieved in the Sapphire Project [10].

3.6 Fault-Tolerant Workflows

- Kepler [3]: A scientific workflow system to access remote resources and services. Kepler is based on Ptolemy II [8].
- SCIRun [11]: Problem Solving Environment for modeling, simulation and visualization of scientific problems.

4 Conclusion

In this document we provided a brief overview of recommended data management technology based on various user surveys conducted in 2006. More information can be found on the NERSC Analytics Webpages: http://www.nersc.gov/nusers/analytics/.

References

[1] E. Wes Bethel, Scott Campbell, Eli Dart, Kurt Stockinger, and Kesheng Wu. Accelerating Network Traffic Analysis Using Query-Driven Visualization. In *IEEE Symposium on Visual Analytics Science and Technology*. IEEE Computer Society Press, 2006.

[2] H5Part. http://www-vis.lbl.gov/Research/AcceleratorSAPP/.

[3] Kepler. http://www.kepler-project.org/.

[4] Metadata Catalog (MCat). http://www.sdsc.edu/srb/index.php/MCAT.

[5] Metadata Catalog Service (MCS). http://www.globus.org/grid_software/data/mcs.php.

[6] NERSC Analytics User Survey: User Input on Analytics Needs. https://staph.nersc.gov/twiki/bin/view/Collaborations/Analytics.

[7] NERSC 2006 User Survey Results: Visualization and Data Analysis. https://www.nersc.gov/news/survey/2006/vis-results.php.

[8] Ptolemy II. http://ptolemy.berkeley.edu/ptolemyII/.

[9] ROOT. http://root.cern.ch.

[10] Sapphire. http://www.llnl.gov/CASC/sapphire/.

[11] SCIRun. http://software.sci.utah.edu/scirun.html.

[12] Storage Resource Broker (SRB). http://www.sdsc.edu/srb/index.php.

[13] Storage Resource Manageer (SRM). http://sdm.lbl.gov/srm-wg/.

[14] Kurt Stockinger, John Shalf, Kesheng Wu, and E. Wes Bethel. Query-Driven Visaulization of Large Data Sets. In *Proceedings of IEEE Visualization*, October 2005.