Storage Techniques for RHIC Accelerator Data*

J.T. Morris, S. Binello, T.S. Clifford, T D'Ottavio, R.C. Lee, C. Whalen
Brookhaven National Laboratory, Upton, NY, USA

Abstract

During Relativistic Heavy Ion Collider (RHIC) operations, approximately ten gigabytes of data are stored each day to document beam characteristics and performance of accelerator systems. A data organization model has been chosen that reflects the natural grouping of data by RHIC machine cycles. Software tools built on this model simplify navigation through the data. Migration of older data to inexpensive file storage systems provides a cost effective way of keeping many years of past data online for analysis. Specialized system administration procedures provide efficient and reliable tape backups. This paper describes the organization of RHIC accelerator data and the techniques developed to manage this data.

Introduction

The RHIC Control System shares basic system infrastructure, including database services and file servers, with the older Alternating Gradient Synchrotron (AGS) Control System. This Collider-Accelerator Department (C-AD) system infrastructure was expanded in anticipation of the needs of RHIC operation, which began in March of 2000. The size of RHIC, with about 20 times the number of control points in the AGS control system[1], was one factor that contributed to a greater demand for data storage. A change in operating model was even more significant. Tuning of the rapid cycling AGS (AGS cycle times range from 2 to 10 seconds) had relied primarily on immediate feedback. Efficient tuning of the much longer RHIC machine cycle depends on the capture of a great deal of beam and system diagnostic data for analysis. Detailed power supply diagnostic information is captured for each acceleration ramp. The RHIC *post mortem* system saves large amounts of power supply and loss monitor data to support analysis of faults. Extensive beam instrumentation data is saved for all phases of the RHIC machine cycle.

During early RHIC operations, the increased storage of data put a strain on the Control System infrastructure. The amount of disk space required exceeded expectations. In addition, the AGS model of data storage was found to be ill suited to the volume of data being saved for RHIC. In the AGS storage model, data was primarily separated by system. Archived machine settings were separated by distinct chronological running periods, typically on the order of several months. Other than this separation by "run", there was no systematic chronological separation of data in the file system or database. RHIC data accumulated quickly leading to UNIX directories with hundreds or thousands of data files. This presented problems for users who had to gather data for a single RHIC machine cycle from many different directory locations. System administration procedures, such as tape backups and disk space analyses, were slowed down by traversal of these large directories and file system performance was adversely affected. A new approach to data organization was needed.

Run and Fill Data Organization

In the operation of RHIC, a "fill" is defined as the time period encompassing one complete machine cycle. This includes the injection, acceleration, and storage of colliding beams. Successful RHIC fills typically last three or four hours. Early in planning of RHIC controls, a unique fill number was recognized as an important tag to be used in identifying and correlating data. Mechanisms were put in place to make a fill number publicly available and to increment the fill number before each RHIC machine cycle.

* Work performed under the auspices of the U.S. Department of Energy.

In November of 2000, the RHIC fill was chosen as the basis for a new method of organizing RHIC data. Grouping data by fill was a good match for both the way the files were written and the way they were typically retrieved. Directory organization by fill had already been successfully employed in the operation of CERN's Large Electron-Positron Collider [2]. The fill concept was integrated with the AGS run concept to produce a two level chronological separation of data. At the base of the RunData directory tree, a new directory is created at the start of each RHIC running period. During RHIC operations, the RHIC Sequencer [3] runs a "newfill" sequence at the start of each new RHIC machine cycle. This sequence increments the public fill number and creates a fresh fill directory in the RunData area for the current run. The sequence also creates a currentFill link within the RunData area that points to the newly created fill directory. The currentFill link makes it easy for applications to access files for an active fill.

The RunData storage area was introduced for the second year of RHIC operations, which began in May 2001 and ended in January 2002. Listed below is a partial view of the top level of the first year of the RunData directory structure. Data for fill numbers 178 to 2320 was stored during this time period. Note that separate run areas were created for gold (rhic_au_fy01) and polarized proton (rhic_pp_fy02) running.

```
RunData/rhic_au_fy01/00178
                    /00179
                    /00180
                    …
                    /01860
                    /01861

RunData/rhic_pp_fy02/01862
                    /01863
                    …
                    /02320
```

Storing Data in the RunData Area

Applications can use the RunData/currentFill link to write data files into subdirectories in the area for the current fill. Applications have the responsibility of ensuring that the necessary subdirectory structure is created within the fill directory before writing files. This is easily accomplished using Control System tools or Unix system utilities. In order to simplify data retrieval, most applications insert database records for files that are saved in the RunData area. The database records identify the file path, run name, fill number, and time period of each file saved. Additional descriptive information is included in database records as needed. Software tools are available in the Control System libraries to make it easy for any application to add file information to the database. All major RHIC systems store their data in the RunData area. This includes beam instrumentation systems, the *post mortem* system, the ramp diagnostics system, and the generic data logging system.

Retrieving Data from the RunData Area

Applications typically use information from the database to present lists of data for selection by a user. The most common tool for retrieving data is the LogView application. LogView allows data to be selected from a variety of sources and displayed together for the same fill or time period. Figure 1 shows the LogView data selection window. The top menus are used to select data items. The bottom menu is used to select the time period of interest. Figure 2 shows a typical LogView data display of main dipole current and beam current for a single RHIC fill. Standard file selection tools have been made available in Control System libraries to support data retrieval in applications. It should be noted that not all access to the RunData area is handled by Control System tools. Some users retrieve their data directly from the UNIX file system. Other users write custom programs that access the database directly to locate files. The RHIC run and fill data organization facilitates all methods of access.

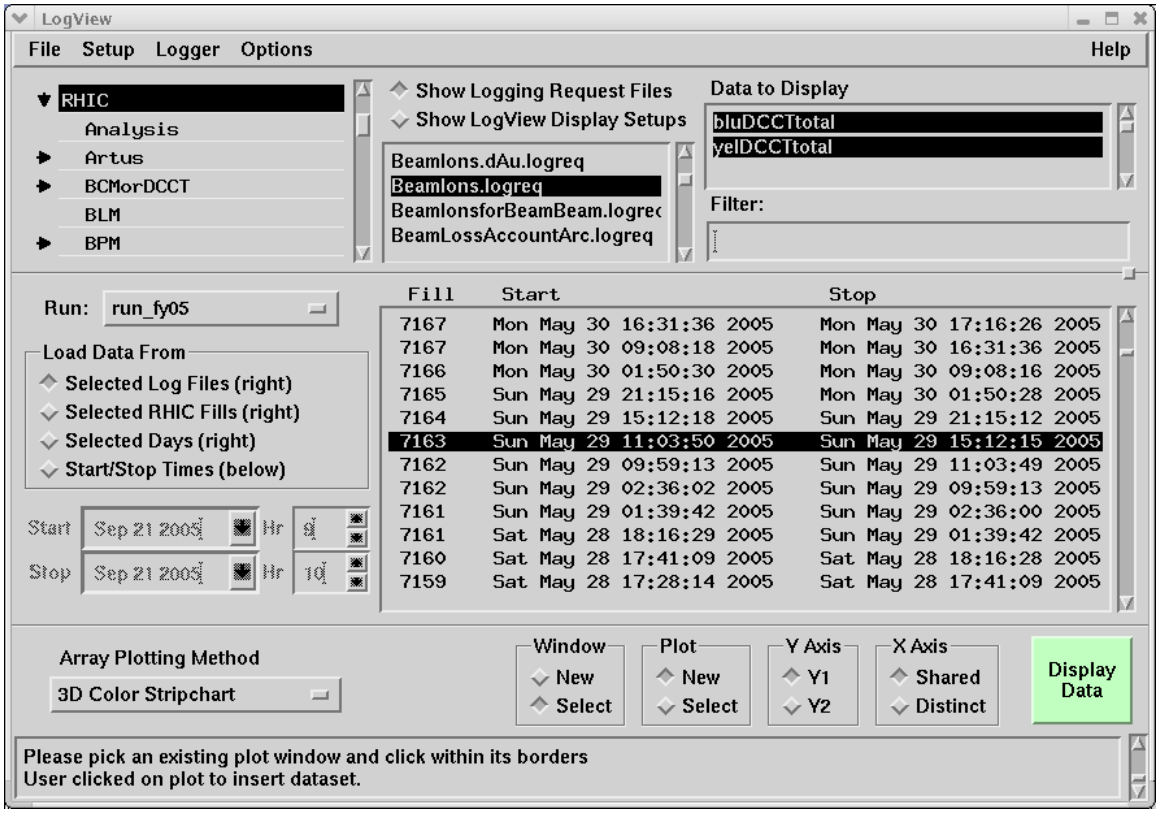

Figure 1: LogView Data Selection Window

Figure 2: LogView Display of Main Dipole Current and Beam Current for a Single Fill

Online Archive Servers

The problem of dealing with data from past runs was encountered early in RHIC operations. An average of 5 gigabytes(GB) of disk space were used each day during the rhic_au_fy01 and rhic_pp_fy02 runs, adding up to a total of 1.25 terabytes(TB) of data by the time these runs completed in January of 2002. The capacity of the C-AD operations file server at that time was only 700GB. Selected data from the completed runs was kept offline in a tape archive. It was restored on user request. The level of demand for the old data made it clear that it could not reasonably be maintained as an offline tape archive. The offline data was no longer critical to accelerator operations but it remained important for analysis purposes and as a reference for future operations.

The old data needed to be online but performance and reliability requirements were lower than that for other data. Write performance requirements were minimal since the data would only need to be written once when it was restored from tape to disk. Read performance requirements were not high since the data was not expected to be used in time critical accelerator operations. Reliability requirements were modest. Down times of even a day or two would be tolerable for most uses of the old data. Considering these reduced requirements, expansion of the primary operations file server was not considered necessary. Low cost storage alternatives were explored. Since data from past runs is separate in the run and fill data organization, using a separate storage solution did not present a problem.

The storage solution chosen in 2002 was a Linux workstation with room for 8 hard disk drives and a total capacity of 1.28TB. This system, considered an online alternative to tape archives, was referred to as an "online archive server." Like all other Control System file servers, it made files available using the Network File System protocol. One single online archive server provided enough space for one year of data with a price tag of about $5K. This has proven to be a cost effective way to provide convenient access to all data from past RHIC runs. A new online archive server has been added for each year of RHIC operations. This is an attractive approach since more recent data naturally ends up on newer, and presumably more reliable, systems.

Changes have been made in online archive system configuration since 2002. The amount of data stored during RHIC operation has continued to grow, averaging about 10GB a day during 2005 running. To keep pace with this growth, newer online archive servers have been purchased with 16 disk drives and 3TB of total disk space. In order to maximize available space and keep costs down, the earliest online archive servers did not have RAID protection for disk failures. This was considered an acceptable risk due to the relatively low reliability requirements. Experience has shown that it is worth investing in RAID. Several disk failures on online archive servers have been experienced each year. Restoring data from tape is a slow process and requires attention from system administrators. Data was sometimes unavailable for days. Since 2003, online archive servers have been purchased with RAID controllers and have been configured for RAID level 5. This has led to a higher level of data availability and a reduced burden on system administration staff. The price tag for the new online archive servers, with higher capacity and improved fault tolerance, remains under $8K. Some data from the older systems has been moved to available space on the newer servers.

Data Backup and Migration

Ordinary incremental backup procedures scan directory structures looking for files that are new or have been changed. Traversing directories of all the files saved during an entire RHIC run would put an unnecessary load on the file system. The run and fill data organization allowed specialized fill backup procedures to be developed. These specialized procedures take advantage of the fact that the directory tree for each fill in the RunData area is only written to for a limited period of time after it is created. All files in that fill directory structure are new. Shortly after the fill is completed, the automatic fill backup procedure marks the fill directory read-only to ensure that the contents of the fill directory will not be modified at a later time. The entire contents of that directory are then written to an online archive server and to backup tapes.

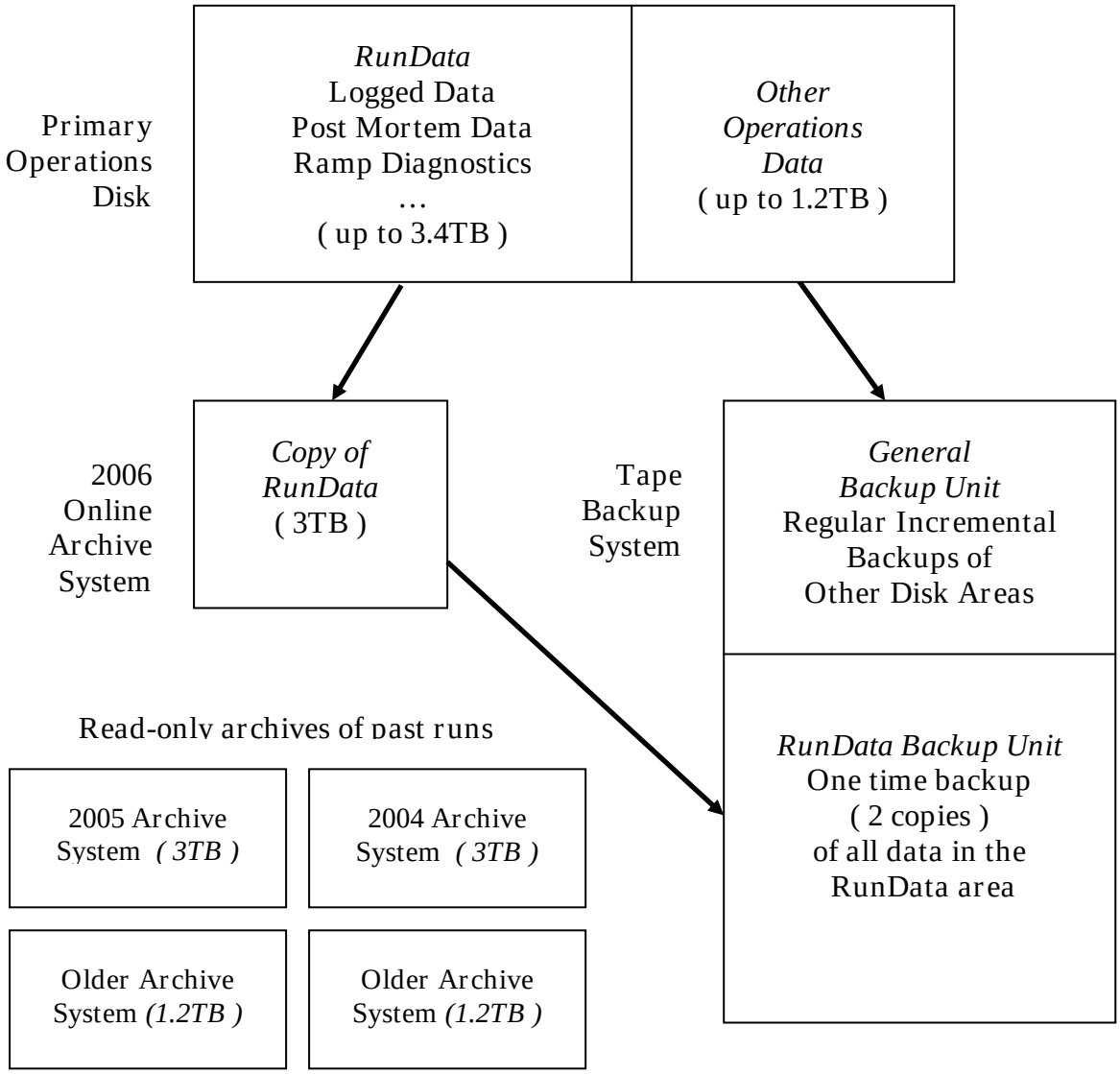

Figure 3. Data Backup and Migration Path for RHIC 2006 Run

Figure 3 shows the planned data backup and migration path for the RHIC 2006 run. The automated fill backup procedure first copies all the data for a fill to the designated online archive server for the current run. The same data is then copied from the online archive server to tape. At that point, an identical image of the directory structure for that fill exists on the primary operations file server, on the online archive server, and on tape. Before the start of a new running period, data from past runs is typically removed from the primary operations file server. The data remains available in the system on the online archive server. Links in the file system are changed so that the use of an online archive server is transparent to users. Note that the timing of the removal of data from the primary operations file server is flexible. During RHIC runs before 2004, the primary operations file server did not have enough space for all the data of the run in progress. Only data for the most recent fills was maintained on the operations file server. All other data was accessed from online archive servers. This level of flexibility was made possible by the chronological separation of data by fill.

Separation of data by fill simplifies other system administration tasks such as the monitoring of disk space use. It is a simple matter to track the amount of disk space used during each fill. System administration tools were developed to, in addition, track disk space use by different subdirectories within each fill. Areas of unexpected growth could be easily identified and investigated.

Conclusion

A revised data storage model for RHIC accelerator data has been in use for four years. The chronological separation of data by run and fill has yielded significant benefits. Access to data is simplified for users, whether they use Control System tools or direct directory access. The new data storage model has made it easier to administer Controls file systems. It has also facilitated the online storage of old data on low cost disk systems.

References

[1] D.S. Barton, et al., "The RHIC Control System," Nuclear Instruments and Methods in Physics Research A 499

[2] M. Lamont, CERN, private communication

[3] T. D'Ottavio, "Description of the RHIC Sequencer," Proc. ICALEPCS 2001