DATA CATEGORIZATION AND STORAGE STRATEGIES AT RHIC*

S.Binello†, K.Brown, T.D'Ottavio, J.Laster, R.Katz, J.Morris, J.Piacentino
Collider-Accelerator Department, BNL,Upton NY, USA

Abstract

This past year the Controls group within the Collider Accelerator Department at Brookhaven National Laboratory replaced the Network Attached Storage (NAS) system that is used to store software and data critical to the operation of the accelerators. The NAS also serves as the initial repository for all logged data. This purchase was used as an opportunity to categorize the data we store, and review and evaluate our storage strategies. This was done in the context of an existing policy that places no explicit limits on the amount of data that users can log, no limits on the amount of time that the data is retained at its original resolution, and that requires all logged data be available in real-time. This paper will describe how the data was categorized, and the various storage strategies used for each category.

INTRODUCTION

The NAS that had served as the main repository of data and software critical to running the accelerators, as well as the initial repository of logged data since 2008 was declared End Of Life by the vendor in 2014. As such, we needed to decide whether to continue using the system and obtain support from a third party, or simply replace the system. As this NAS had other limitations: insufficient storage, outdated small expensive 300GB drives and 1GB network interfaces, along with a hefty $15K annual maintenance fee for a mere 25TB of storage, the decision was made to replace the NAS.

DATA CATEGORIZATION

To facilitate and organize the analysis of requirements for the new system we categorized our data into three tiers: critical operational data, auxiliary operational data and historical/temporary data. Each tier was then assigned a level of required reliability.

Tier 1 – Critical Operational Data

This is data that is needed to run the accelerators. It consists of program executables, basic configuration information, Front End Computer (FEC) boot areas, archives of recent settings, and critical group home directories.

This data should always be available. A complete High Availability(HA) solution is required. That is, the storage device should have redundant components at all levels. In the event of multiple failures which makes this data unavailable, a backup with data as current as possible should be made available. A switch to the backup system should take less than 4 hours.

*Work performed under Contract Number DE-AC02-98CH0886 with the auspices of the US Department of Energy
† sev@bnl.gov

Tier 2 – Auxiliary Operational Data

This data may be important for some operational tasks, but is not critically needed for basic machine operation, for example, logged data from the current run, archived settings from past runs (and possibly archived settings from earlier in this run).

A total HA solution is desired, but is not absolutely necessary. Minimally, RAID should be used to protect from disk failures. If primary storage for this data is not a total HA solution, a backup version of the same data should be maintained on an alternate disk storage device. Data on the backup should be no more than three days old. A switch to the backup should take less than 4 hours.

Tier 3 – Historical or Temporary Data

This data is mostly composed of logged data from past runs, along with any data that may only need to be stored temporarily.

Some high availability features are desired. Minimally, RAID should be used to protect from disk failures. In the event of some other failure, restoration from tape would be necessary. Restoration of most types of data could be done in less than an hour but could take several days.

DATA STORAGE STRATEGIES

Once the data was classified attention focused no how best to support the requirements for each data tier.

Tier 1 – Storage For Critical Data

From the onset, it was understood that a High Availability (HA) NAS was the required solution for critical operational data. The focus was then on how to increase the availability of this critical data in the event of a failure of the primary NAS, and how to provide the most up-to date data possible. The Disaster Recovery (DR) system for the NAS that was to be replaced, was a low cost Linux storage server with internal SATA drives RAID'ed using 3-ware controllers (this is a configuration that is also use to support tier 3 data). Critical data was replicated from the NAS to the DR system using the Linux rsync utility. This pairing proved to have some flaws. The DR system was not identical to the NAS, and as the NAS was asked to store and perform more and more, over time it eventually outpaced the capabilities of the DR system. Additionally, we found that the rsyncs were also negatively impacting the performance of the primary NAS.

It was decided that in order to prevent this type of divergence in capabilities in the future, and in an attempt to limit the impact of data replication between the primary and DR unit, an identical NAS was needed for disaster

ISBN 978-3-95450-148-9

recovery. Currently, tier 1 data is replicated every 24 hours, but we expect to increase the frequency of replication. Optimized tools, provided by the NAS vendor, are used to efficiently replicate data between the primary NAS and the DR NAS.

Tier 2 – Support for Auxiliary Data

Several options were considered to store tier 2 data. As this data is not critical to running the accelerators, it does not require a complete HA NAS solution. As such, we needed to determine if we should continue storing tier 2 data alongside tier 1 on the NAS, or look for an alternate storage solution. Even though this data is not critical, a fault with the storage device could prevent the capture or availability of data that might still be of significant interest to users.

Two main reasons were identified to support the idea of a separate storage solution. One, was that access to this data might interfere with the read/write performance of tier 1 data. Another, was to provide a less expensive solution for tier 2 data. As such, two options were considered for tier 2 data. The first was to purchase an alternate less expensive NAS. The second, was to use an inexpensive Linux storage server configured with internal drives and protected with RAID controllers (i.e. our tier 3 storage server configuration). In this configuration there is no fail-over capability in the event of a RAID controller failure, nor in the event of a motherboard failure. To circumvent this limitation, we considered the possibility of configuring an alternate system to act as a backup. However, in the event of a failure the transition to this system would require down-time and data loss that we would not have to incur if we were to use a NAS .

Eventually, we resolved to store tier 2 and tier 1 on the same NAS. Financially it made sense to use the same NAS, as it only required the purchase of additional disks. And, even though it was not needed, it endowed tier 2 data with all the same HA capabilities as those provided for tier 1 data.

To address the concern that tier 2 data might negatively impact the reading/writing of tier1 data, we still wanted to separate these two tiers as much as possible. The initial thought was to direct tier 2 data to the DR NAS, while storing tier 1 data onto the primary NAS. In this configuration, each NAS acted as a primary for its designated data type while also acting as the DR system for its partner NAS. This was the preferred solution, however due to financial constraints the DR NAS did not have a redundant head (not a desirable configuration for a NAS acting as primary storage). As a result, tier 1 and tier 2 data were both stored on the primary NAS. To reduce contention between tier 1 and tier 2 data, the primary NAS was configured so that each tier was supported by its own NAS head, network connections, and dedicated disk drives.

Not only were the two tiers stored on separate disk drives, but different types of drives were used for each tier. As the amount of tier 2 data collected over the entire run is quite large (approximately 60TB compared to 6TB for tier 1), and we wanted to have the option of keeping an entire run's data on the NAS, it was decided to store tier 2 data on larger, cheaper, and slightly less reliable NL SATA drives. Tier 1 data was stored on 900GB SAS drives. This not only brought down the original cost of the NAS but also significantly reduced maintenance fees.

While the option exists to store a whole run's worth of tier 2 on the NAS, in fact the NAS is currently used as a temporary staging area for tier 2 data. Tier 2 data is initially captured and then stored for only a few weeks on the NAS, after which it is relocated to low-cost storage servers. This approach provides HA for the initial capture of tier 2, and during that period of time where it is most likely to be of interest to users.

Tier 3 – Support for Historical/Temporary Data

Tier 3 data is composed of older logged data from past runs (in effect tier 3 data is simply older tier 2 data), or large data files deemed not critical by users. As such it comprises the largest amount of stored data.

To provide real-time access to logged data from previous runs, we have had a strategy in place since 2001 where logged data is eventually moved from the NAS to inexpensive Linux storage servers. Data from previous runs is stored under dedicated directories for each run, and accessed through links on central directories that reside on the NAS.

When these storage servers fill up we simply purchase additional systems. We have found that over the years this approach allowed us to take advantage of the ever-increasing disk sizes and decreasing costs of SATA drives. Last year we purchased a 196TB system for $20K. When this strategy was initiated, the typical server stored about 5TB. Presently, storage for tier 3 data is provided by three 196TB, two 96TB and two 48TB systems.

Logged data on older servers, with less storage, is eventually consolidated onto newer servers. This may prove more cumbersome in the future as the amount of data to migrate increases.

Another option considered for tier 3 data was to store it on the the disaster recovery NAS. This NAS has the ability to support 4PB of data, and the ability to efficiently replicate data from the primary. However, there was concern about maintenance fees and cost of disk drives when compared to the generic Linux storage servers.

CONCLUSION

This latest NAS purchase provided us with an opportunity to categorize the data we store and to re-examine our data storage strategies. We considered various options, but in the end the solution was not all too different from the previous one. Tier 1 and 2 data is collected on the same HA NAS device, though tier 2 data is only stored there for a short time. Tier 3 data is stored on inexpensive generic Linux storage servers with limited HA capabilities. An almost identical NAS was purchased for disaster recovery, and is only used to backup tier 1 data and provides storage for tier 2 only in event the primary NAS fails. To be determined is to see how well this approach scales as we continue to store ever increasing amounts of data.

ISBN 978-3-95450-148-9