Mean PB To Failure – Initial results from a long-term study of disk storage patterns at the RACF

C Caramarcu, C Hollowell, T Rao, W Strecker-Kellogg, A Wong, A Zaytsev
Brookhaven National Laboratory, Department of Physics, PO Box 5000, Upton, NY 11973, USA

E-mail: tony@bnl.gov

Abstract. The RACF (RHIC-ATLAS Computing Facility) has operated a large, multi-purpose dedicated computing facility since the mid-1990’s, serving a worldwide, geographically diverse scientific community that is a major contributor to various HEPN projects. A central component of the RACF is the Linux-based worker node cluster that is used for both computing and data storage purposes. It currently has nearly 50,000 computing cores and over 23 PB of storage capacity distributed over 12,000+ (non-SSD) disk drives. The majority of the 12,000+ disk drives provide a cost-effective solution for dCache/XRootD-managed storage, and a key concern is the reliability of this solution over the lifetime of the hardware, particularly as the number of disk drives and the storage capacity of individual drives grow. We report initial results of a long-term study to measure lifetime PB read/written to disk drives in the worker node cluster. We discuss the historical disk drive mortality rate, disk drive manufacturers' published MPTF (Mean PB to Failure) data and how they are correlated to our results. The results help the RACF understand the productivity and reliability of its storage solutions and have implications for other highly-available storage systems (NFS, GPFS, CVMFS, etc) with large I/O requirements.

1. Introduction
The RHIC and ATLAS U.S. Tier 1 Computing Facility (RACF) at Brookhaven National Laboratory (BNL) supports the computational needs of the RHIC experiments, the U.S. collaborators in the ATLAS experiment and other activities undertaken by the Physics Department at BNL. It has operated continuously since the mid 1990’s, serving a geographically diverse, worldwide scientific community. The RACF is made up of a multi-silo robotic tape storage facility, high-availability disk storage system and an integrated Linux-based worker node cluster with over 3000 active Gigabit and 10 Gigabit-capable network ports distributed over several subnets. Currently, the combined tape and high-availability storage usage is approximately 53 PB. Additionally, the Linux-based worker node cluster provides approximately 24 PB of raw storage capacity spread over ~12,000 disk drives. Nearly 9,000 of these disk drives (~21 PB of raw storage capacity) are managed by dCache [1] and XRootD [2] as distributed storage for RHIC computing. The disk drives have varying capacities (0.5 to 4 TB), age (0-5 years) and brand (Seagate, Hitachi, Western Digital, etc.).

This paper reports initial results from a long-term analysis of worker node-based disk drive usage patterns at the RACF and has its origins in a Western Digital presentation at the HEPIX Fall 2013
meeting [3] on Mean PB To Failure (MPTF) as an alternative to the traditional Mean Time To Failure (MTTF) metric.

2. Methodology
To conduct this study, system I/O counters for Read/Write activity were recorded for disk drives on a daily basis and archived in a MySQL database. This project began in late September 2014 and has been gathering data continuously since then. For manageability, only data from a subset of the worker nodes are being recorded. This subset has approximately 1,200 disk drives (~10% of total) and was chosen as representative of the varying storage capacities, ages, brands and experiments supported at the RACF (PHENIX, STAR and USATLAS). Data trends are extrapolated over a full year and compared to manufacturer published specifications (where available). Finally, the results are correlated to expected and observed disk drive mortality.

3. Analysis

3.1 USATLAS
USATLAS worker nodes are storage-light systems, in which the available disk storage capacity is entirely devoted to scratch space. There is no distributed data storage at the worker node level. The worker nodes are used strictly for computing jobs. As a result very little I/O activity is observed to/from the worker-node. Of the subset discussed above that is dedicated exclusively to USATLAS, system I/O counters reveal a read rate of \(\approx 0.05\) TB/day per drive and a write rate of \(\approx 0.13\) TB/day per drive. Figures 1 and 2 are representative of the typical read/write rates on USATLAS worker nodes at the RACF. System I/O counters reset to zero when the worker node is rebooted.

![Figure 1. Read activity for an USATLAS worker node (acas1093)](image1)

![Figure 2. Write activity for an USATLAS node (acas1093)](image2)
3.2 PHENIX
PHENIX worker nodes are storage-heavy systems that use dCache for distributed storage management. PHENIX worker nodes are used for both computing and I/O intensive data movement activities. As a result, read and write rates are much higher than USATLAS (0.82 TB/day per drive and 0.75 TB/day per drive, respectively). Figures 3 and 4 are representative of typical read/write rates on PHENIX worker nodes at the RACF.

![Figure 3. Read activity for a PHENIX node (rcas2714)](image)

![Figure 4. Write activity for PHENIX node (rcas2714)](image)

3.2 STAR
Similar to PHENIX, STAR worker nodes are also storage-heavy systems, but it uses XRootD for distributed storage management and have a different data movement model. STAR uses the "write once, read many times" approach. Therefore, read and write rates are markedly different (1.56 TB/day per drive and 0.22 TB/day per drive, respectively). Figures 5 and 6 illustrate the read/write rates for typical STAR worker nodes at the RACF.
4. Published MPTF and Workloads at the RACF

In June 2013, Western Digital (WD) published a study on MTTF and MPTF describing a correlation between MPTF and disk drive failure rates [4]. As a result of the study, WD now publishes MTTF and also MPTF as the maximum workload (read + write) to achieve the stated MTTF for its storage products [5]. Seagate likewise publishes maximum workload to achieve stated MTTF for its products [6]. Among the disk drive inventory, the RACF uses several WD and Seagate disk drive models with a maximum workload of 550 TB/year (see references above). With these workloads, annual failure rates (AFR) are expected to be in the range of 0.44% to 0.73%, depending on manufacturer and disk drive model number. We define workload as:

\[ \text{Workload} = N \times \text{Read} + M \times \text{Write}, \]  \hspace{1cm} (1)

where N and M are weighted factors. For the initial study, we assume N=M=1.

The data from the subset of ~1200 disk drives allow us to calculate the activity (read and write) over the past ~6 months. We extrapolated linearly to a full year to estimate annual values. The results are summarized in Table 1 below.

![Read activity for STAR node (rcrs6229)](image1)

**Figure 5.** Read activity for STAR node (rcrs6229)

![Write activity for a STAR node (rcrs6229)](image2)

**Figure 6.** Write activity for a STAR node (rcrs6229)
Table 1. Disk drive activity extrapolated over a full year (in TB/year)

| Experiment | Read (TB/year) | Write (TB/year) | Workload (TB/year) |
|------------|----------------|-----------------|--------------------|
| USATLAS    | 17.2           | 45.8            | 63.0               |
| PHENIX     | 300.0          | 275.6           | 575.6              |
| STAR       | 567.8          | 81.4            | 649.2              |

5. Correlation with Disk Failure Rates
A preliminary survey of disk drive hardware failure was conducted to investigate correlation with disk drive workload. For this survey, we used only 2013 and 2014 data, and only disk drives with unrecoverable hardware failure (that is, disks are replaced in the server) are counted. The annual failure rate (AFR) is defined as a fraction of the total disk pool size in each cluster. AFR is calculated separately for each year (disk pool size was somewhat larger in 2014 than in 2013) and averaged. Results are summarized in Table 2 below.

Table 2. Disk drive annual failure rate

| Experiment | Annual Failure Rate (%) |
|------------|-------------------------|
| USATLAS    | 0.35                    |
| PHENIX     | 0.94                    |
| STAR       | 0.41                    |

Contributions from other factors (thermal conditions in the data centre, quality of electrical power to servers, etc.) have not been measured independently and are therefore folded into the AFR values listed above.

6. Summary
Some data movement patterns exceed published maximum workload by a non-negligible margin and should be a factor in observed AFR at the RACF. Early data suggests higher disk mortality rates with heavy write activity (as in PHENIX) when compared to other usage cases. Even so, AFR is acceptably low (under 1%), validating the worker node-based distributed storage model as a cost-effective solution.

The initial results reported here only cover ~186 days. A long-term study that covers the entire data collection, reconstruction and analysis cycle multiple times is ongoing to provide more precise results. This initial study also assumes read and write have the same impact on AFR (N=M=1). This may prove to be incorrect when more data become available.

7. Acknowledgements
The authors would like to gratefully acknowledge the useful discussions with Western Digital on the merits of the MPTF metric and its correlation to disk drive reliability.

8. References
[1] http://www.dcache.org.
[2] http://xrootd.org.
[3] Chattopadhyay A, “Why specify and monitor hard disk drive workloads”, HEPiX Fall 2013 (http://indico.cern.ch/event/247864/session/3/contribution/37).
[4] Tyndall G, “Why Specify Workload?” (http://www.wdc.com/wdproducts/library/other/2579-772003.pdf).
[5] See http://www.wdc.com/en/products/internal/enterprise/ for more details.
[6] See http://www.seagate.com/www-content/product-content/enterprise-hdd-fam/enterprise-capacity-3-5-hdd/constellation-es-4/en-us/docs/100740544d.pdf for full details.