PROCEEDINGS OF SCIENCE

Optimization of fast parallel operations with large disk arrays for the AMBER experiment

Martin Zemko,[a,b,*] **Dominik Ecker,**[e] **Vladimir Frolov,**[d] **Stephan Huber,**[e] **Vladimír Jarý,**[b] **Igor Konorov,**[e] **Josef Nový,**[b] **Benjamin Moritz Veit**[c] **and Miroslav Virius**[b]

[a] *Charles University in Prague, Czech Republic*
[b] *Czech Technical University in Prague, Czech Republic*
[c] *Johannes Gutenberg University of Mainz, Germany*
[d] *Joint Institute for Nuclear Research, Russia*
[e] *Technical University of Munich, Germany*

E-mail: martin.zemko@cern.ch, dominik.ecker@cern.ch, vladimir.frolov@cern.ch, stefan@klhuber.de, vladimir.jary@fjfi.cvut.cz, igor.konorov@cern.ch, novyjos1@fjfi.cvut.cz, b.veit@cern.ch, miroslav.virius@fjfi.cvut.cz

This contribution addresses the need for reliable and efficient data storage in the high-energy physics experiment called AMBER. The experiment generates sustained data rates of up to 10 GB/s, requiring optimization of data storage. The study investigates single-disk performance, including random and sequential disk operations, highlighting the impact of parallel access and disk geometry. A comparison with SSD drives reveals important differences. Various RAID configurations are assessed, considering their reliability, data rates, and capacity. Probability analysis is used to evaluate the RAID rebuilding procedure in the event of disk failure. In addition, an innovative approach of alternating disk access is proposed to ensure uninterrupted performance. Finally, the study identifies the most suitable RAID configuration for the AMBER experiment. The results of this study contribute to the design of high-performance storage solutions for data-intensive scientific experiments.

42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic

[*] Speaker

PoS(ICHEP2024)1044

1. Introduction

The AMBER streaming acquisition system relies on online data filtering performed at the CERN data center, with an estimated data rate of 10 GB/s from the full detector setup [1]. Local storage units are used to temporarily buffer data, ensuring continuous data collection during occasional upstream outages. These buffers can store up to three days of data, allowing independent operation.

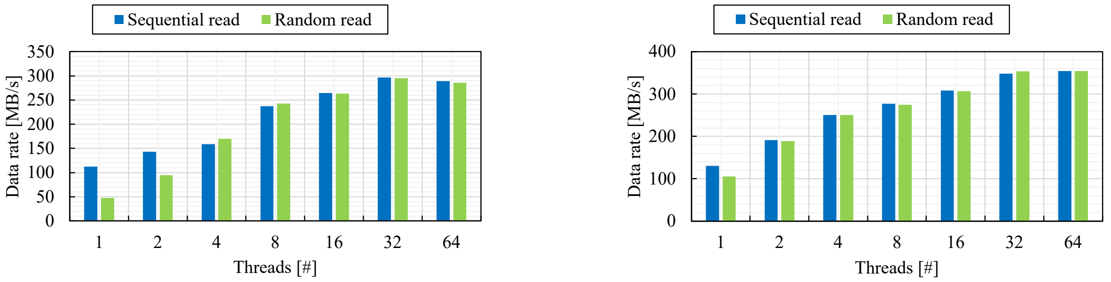
Each buffer must handle a sustained 1 GB/s data flow [2], tolerate a single disk failure, and have at least 95 % success rebuild rate. We employ redundant arrays of independent disks (RAID) to achieve these requirements, combining multiple disks for better performance and redundancy.

Our test setup includes four readout servers equipped with AMD 7313 CPU 3.0 GHz (16 cores, 32 threads) and 128 GB DDR4 RAM. Each server is connected to an external Promise VTrak J5800 storage chassis housing 24 Toshiba MG07ACA14TE (14 TB) disk drives. The chassis is connected to the host via a Broadcom MegaRAID SAS 9580-8i8e controller with 8 GB of DDR4 SDRAM.

2. Parallel access to single disk

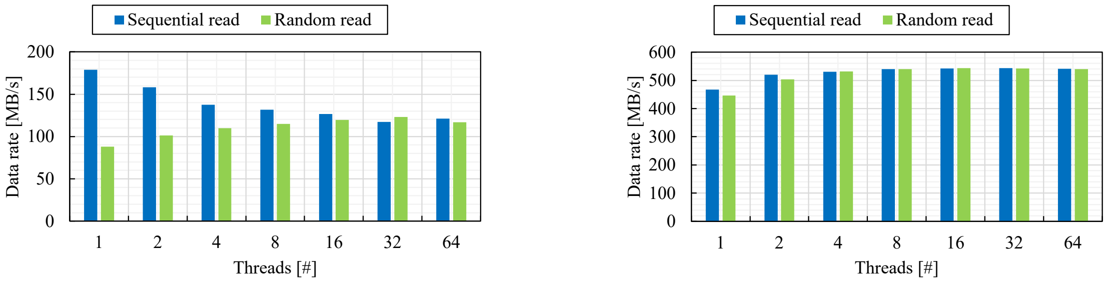
The AMBER readout system splits data into 1 GB chunks and processes them in parallel to maximize data throughput. We simulate this behavior by using multiple workers accessing the storage simultaneously. Parallel access can reduce performance, especially with HDDs that can handle only one request at a time, as the read head must switch rapidly between several positions.

In our first test shown in Figure 1 with 100,000 files of 4 KB each, both HDDs and SSDs performed better under parallel access, thanks to head trajectory optimization in HDDs and parallel stream support in SSDs. For larger files of 1 MB each, as shown in Figure 2, HDD performance dropped with multiple reading threads during sequential access but improved with random access. SSDs showed consistent performance across all cases, with better throughput under parallel access.



(a) Reading small files from HDD **(b)** Reading small files from SSD

Figure 1: Reading small data files from any storage type performs better with multiple reading threads.



(a) Reading large files from HDD **(b)** Reading small files from SSD

Figure 2: Performance of reading large files differs according to the disk type and access pattern.

3. Disk failure prediction

To increase the reliability of the storage, we developed a neural network to predict disk failures based on SMART (Self-Monitoring, Analysis, and Reporting Technology) metrics. Our model was trained on 2,000 drive entries compiled and published by Backblaze in 2016 [3].

For training, we used the TensorFlow framework with the Keras extension. The final neural network has nine inputs and one binary output. It consists of three hidden layers with 64, 32, and 16 neurons, each with a ReLU activation function. These layers are interleaved with normalization and dropout layers to prevent overfitting. The output layer has a single neuron with a sigmoid activation function. The final model provides excellent results, achieving an accuracy level of **94.95 %**, as shown in Figure 3. The ROC curve also indicates a good performance with AUC of **98.15 %**.

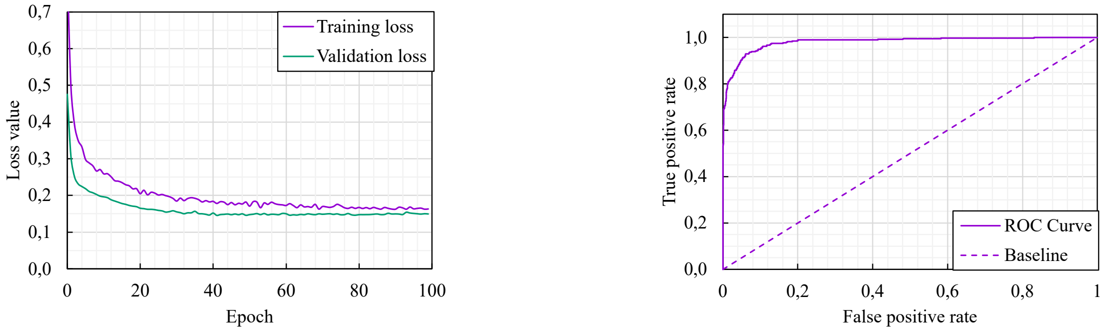


Figure 3: Trained model for predicting disk failures shows excellent performance and consistent results.

4. Redundant array of independent disks

We conducted multiple benchmarks focusing on the read and write throughput of nested RAID configurations with various span sizes. Every benchmark utilized 100 data files, each 1 GB in size, and was performed on the default readout computer setup. As illustrated in Figure 4, we observe that arrays with less than eight disks do not provide sufficient throughput. RAID 00 lacks the required redundancy, and RAID 10 provides only half the capacity. Moreover, RAID 50 performs better than RAID 10 and 60.

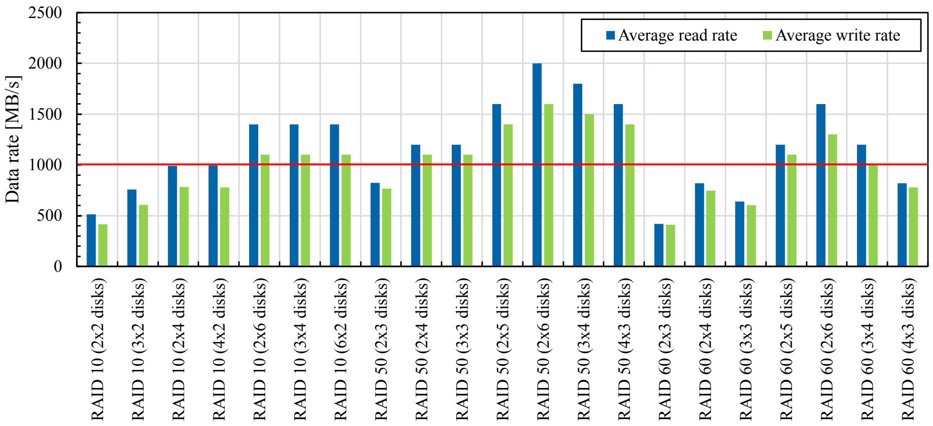


Figure 4: Performance of nested RAID arrays consisting of various span sizes.

4.1 RAID array rebuilding probability

If a disk fails, the array must be rebuilt from the remaining disks. Other disks may fail to read data during the rebuild due to non-recoverable read errors caused by data decay. To assess the likelihood of such an event, we derived a formula for the probability of a successful RAID array rebuild:

$$P_{\text{succ}} = \left(1 - \frac{E}{S}\right)^{C \cdot (N - N_{\text{P}})}$$

where:

- E denotes non-recoverable errors (bits),
- S represents the data size read for the given error rate (bits),
- C indicates single disk capacity (bits),
- N is the number of disks in RAID span,
- N_{P} is the number of parity disks.

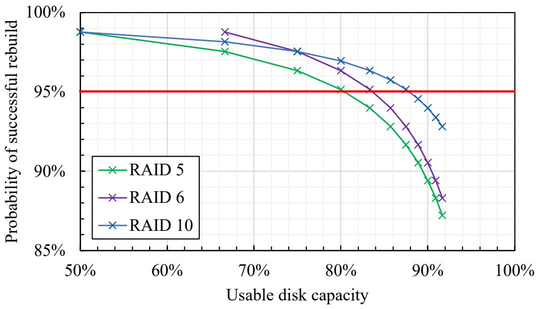


Figure 5: Probability of successful array rebuild for various RAID layouts.

Based on these calculations, as shown in Figure 5, we observe that the probability of a successful rebuild decreases for larger RAID arrays. Thus, we exclude all RAID 50 configurations with more than five disks and RAID 60 arrays with more than six disks.

5. Conclusion

We investigated HDD and SSD performance under various load conditions and thread configurations. Our findings reveal that all read patterns benefit from parallel access except for sequential reading of large files from HDDs. We also developed a robust neural network for predicting disk failures, achieving high accuracy using SMART data metrics. Our evaluation of RAID arrays shows that **RAID 50** (3 × 2 × 4 disks) array, as shown in Figure 6, offers the best balance between performance and redundancy. This configuration is the optimal data storage solution for the AMBER experiment, supporting sustained 1 GB/s data rate per readout server.

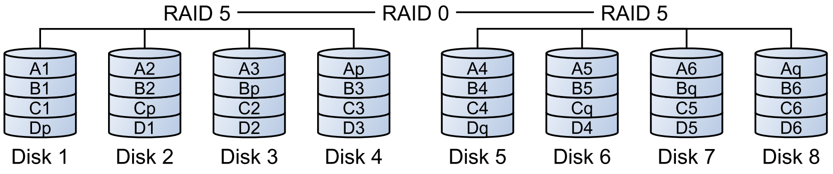


Figure 6: Optimal disk configuration is RAID 50 providing 75 % of the raw disk capacity.

References

[1] B. Adams, C. A. Aidala, and M. Alexeev et al. Proposal for Measurements at the M2 beam line of the CERN SPS. 2019.

[2] M. Zemko et al. *Triggerless data acquisition system for the AMBER experiment*. In Proceedings of 41st International Conference on High Energy Physics — PoS(ICHEP2022), Bologna, Italy: Sissa Medialab, 2022-11, p. 248. DOI: 10.22323/1.414.0248.

[3] E. Kim. *Hard Drive Test Data*. Kaggle data set, 2016. [CSV]. Available: `https://www.kaggle.com/datasets/backblaze/hard-drive-test-data`