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

Martin Zemko[1,2], Dominik Ecker[5], Vladimir Frolov[4], Stephan Huber[5], Vladimír Jarý[2], Igor Konorov[5], Josef Nový[2], Benjamin Moritz Veit[3], Miroslav Virius[2]

[1]Charles University in Prague, Czech Republic, [2]Czech Technical University in Prague, Czech Republic, [3]Johannes Gutenberg University of Mainz, Germany, [4]Joint Institute for Nuclear Research, Russia, [5]Technical University of Munich, Germany

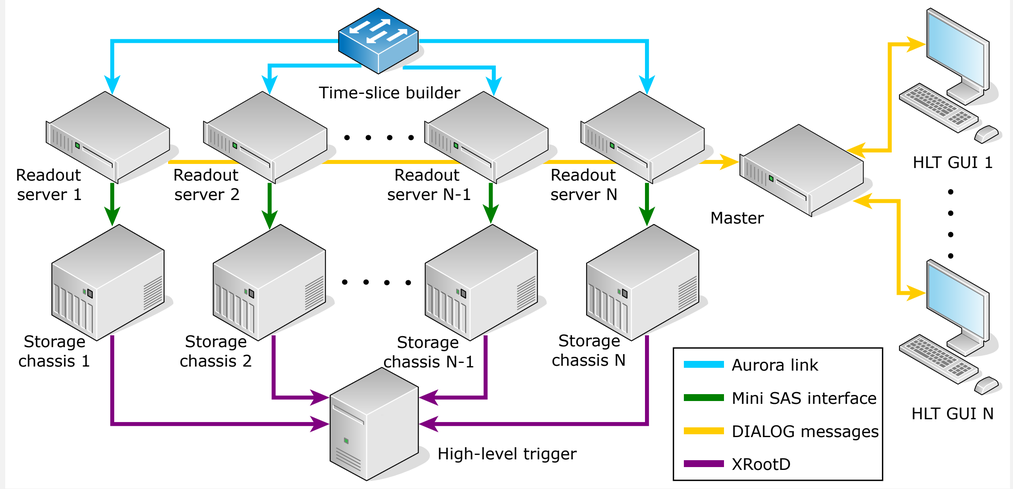
AMBER
Apparatus for Meson and Baryon Experimental Research

FACULTY OF MATHEMATICS AND PHYSICS Charles University

FACULTY OF NUCLEAR SCIENCES AND PHYSICAL ENGINEERING CTU IN PRAGUE

Introduction

- Particle detectors in high-energy physics produce high data rates that must be processed and stored efficiently.
- The AMBER experiment uses a streaming DAQ system with an average data rate of **10 GB/s** [1].
- Nested disk arrays can satisfy requirements for speed and redundancy.
- Our goal is to achieve a **1 GB/s** data rate per readout server [2].



Parallel access to single disk

- HDDs handle only a single request at a time, while SSDs manage several concurrent requests.
- HDD performance decreases with frequent head movements.
- The operating system can optimize the head trajectory in some cases.
- SSDs provide stable performance regardless of a file position.
- Parallelization is always beneficial for small files.

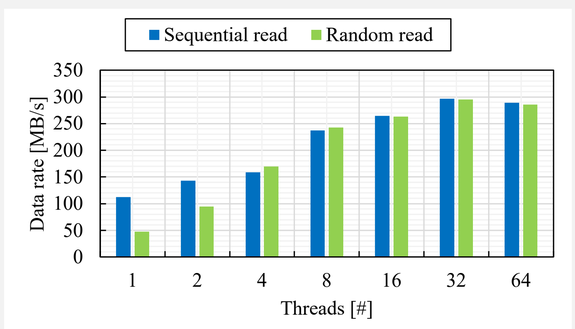


Figure 1: Reading small files from HDD

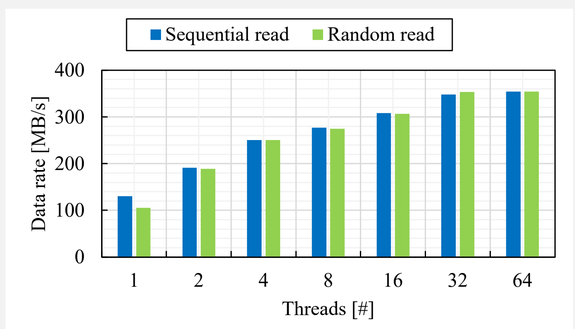


Figure 2: Reading small files from SSD

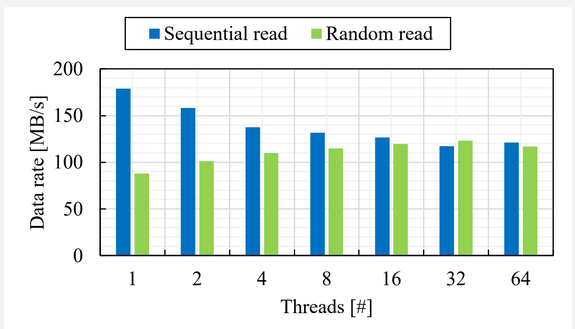


Figure 3: Reading large files from HDD

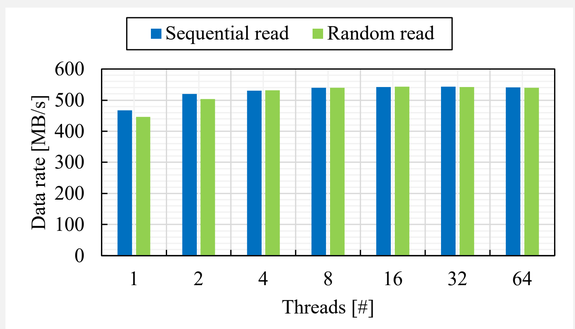
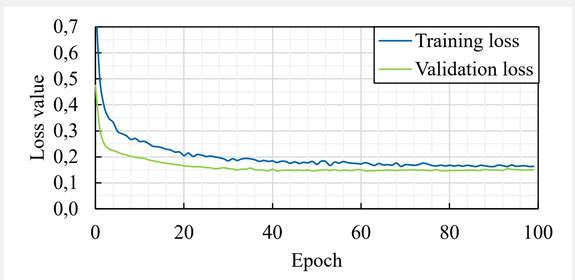


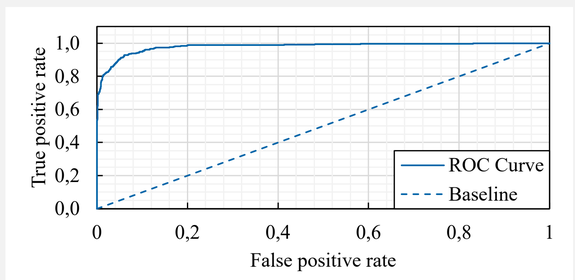
Figure 4: Reading large files from SDD

- Random reading of data files always benefits from more threads.
- For sequential reading of large files from HDD, a single thread is optimal.

Disk failure prediction

- We developed a neural network to predict disk failures based on SMART.
- The network recognizes failing disks in advance using 10 disk metrics.
- Our model was trained on 2,000 drives' metrics published by Backblaze.
- It achieves **94.95 % accuracy** on the validation data and **98.15 % ROC**.





Acknowledgements

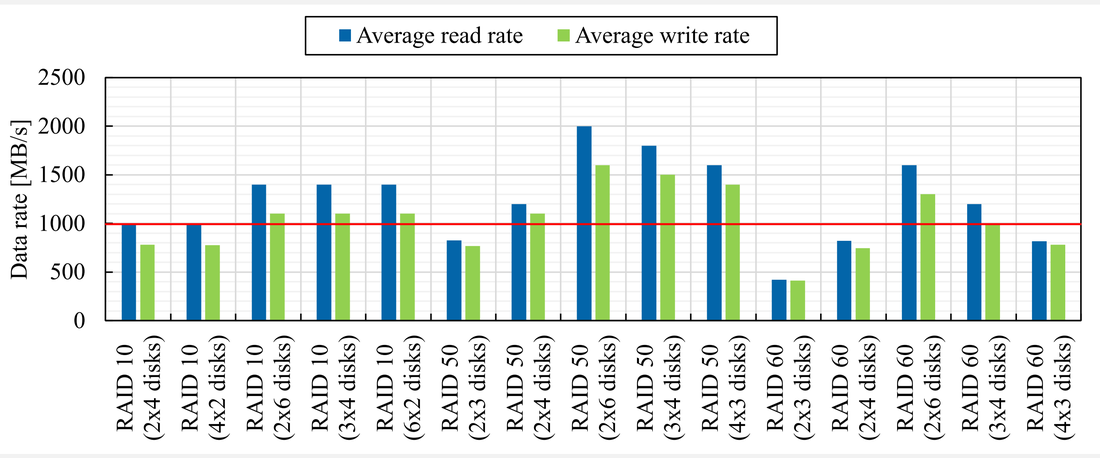
This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic (grant LM2023040), Charles University (grant PRIMUS/22/SCI/017), and the Grant Agency of the Czech Technical University in Prague (grant SGS23/190/OHK4/3T/14).

Test setup

- Test setup includes 4 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 with 24 Toshiba MG07ACA14TE (14 TB) disk drives.
- The chassis is connected via a Broadcom MegaRAID SAS 9580-8i8e.

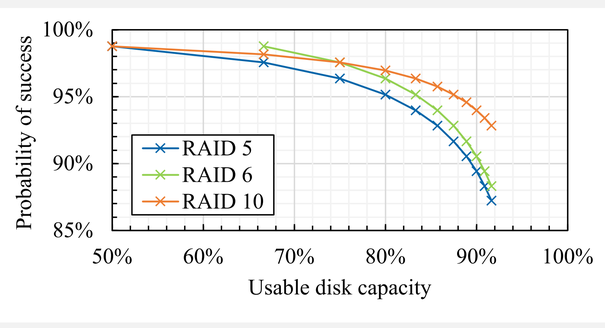
Performance of RAID arrays

- RAID 00 lacks redundancy, and RAID 60 offers low capacity.
- Arrays with less than 8 disks do not provide sufficient throughput.
- The optimal configuration is **RAID 50 with 2 spans of 4 disks**.
- Sustained throughput of 1 GB/s requires a nested RAID array.



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.
- The probability of a successful rebuild decreases for larger RAID arrays.

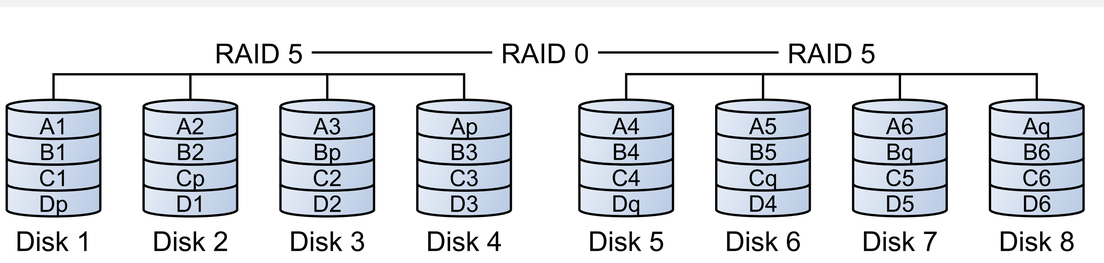


$$P_{succ} = \left(1 - \frac{E}{S}\right)^{C \cdot (N - N_P)}$$

E : non-recoverable errors
S : measured read size
C : disk capacity
N : total disks in array
N_P : parity disks

Conclusions

- 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 developed a robust neural network for predicting disk failures, achieving a high accuracy using SMART data metrics.
- Our evaluation of RAID arrays shows that **RAID 50 (3 x 2 x 4 disks)** array 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.



References

[1] M. Zemko, D. Ecker, and V. Frolov et al. Triggerless data acquisition system for the AMBER experiment. In *Proceedings of 41st International Conference on High Energy physics — PoS(ICHEP2022)*, page 248, Bologna, Italy, November 2022. Sissa Medialab.

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

[3] S. Huber, V. Frolov, and I. Konorov et al. Data Acquisition System for the COMPASS++/ AMBER Experiment. *IEEE Transactions on Nuclear Science*, 68(8):1891–1898, 2021.

Presented at

ICHEP 2024
July 17 – 24, 2024, Prague, Czech Republic