

IMPROVING AVAILABILITY AND PERFORMANCE OF THE EVENT METADATA SYSTEM FOR THE BM@N EXPERIMENT

K. Gertsenberger^a, P. Klimai^{b,c}, I. Dunaev^c,

A. Degtyarev^c, O. Nemova^{c,*}

*^a Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region,
141980, Russia*

*^b Institute for Nuclear Research, Russian Academy of Sciences, 60th October
Anniversary Prospect 7a, Moscow, 117312, Russia*

*^c Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny,
Moscow region, 141701, Russia*

*e-mail: olyanemova36@gmail.com

Received November 22, 2024

Abstract – Event indexing, or event metadata systems are common for particle physics experiments. Their main goal is to keep a searchable catalogue of physics events, a subset of which can be retrieved based on given filtering criteria. The Event Metadata System (EMS) of the BM@N experiment has been previously designed, developed and deployed and is being improved now to increase its performance, convenience for users, as well as fault tolerance. In this paper, the architecture of the current version of the BM@N EMS is reviewed and some recent improvements that have been applied are presented.

INTRODUCTION

The Event Metadata System of the BM@N [1] experiment is one of the information systems that are being developed to support data collection, storage, processing and subsequent physics analysis [2, 3]. The main goal of the EMS is to provide a searchable event catalogue, where the physics events can be selected and obtained based on stored summary properties (metadata) and then passed to a particular analysis, thus speeding up the whole process. Some additional benefits of using the system include event quality control, data integrity, self-consistency checks and provision of useful statistics. The first iteration of the BM@N EMS had been developed and presented earlier [4]. Since then, several aspects and components of the system were added and improved. The main goal of this paper is to describe recent progress made, including high availability solution, automatic deployment system, and improved performance due to using a set of database indexes.

HIGH AVAILABILITY SOLUTION FOR THE EMS

High availability is a pressing issue for such an information service used in modern HEP experiments as the EMS, which is designed to speed up the process of analyzing physics event data. The task of ensuring high availability can be divided into two subtasks, such as setting up data replication and ensuring access to replicas from the client side in case the main server fails.

The PostgreSQL DBMS [5] that is used as event catalogue database in the EMS provides an implementation of data replication using write ahead logging. This approach involves allocating a main database server to serve clients, and backup servers to receive and apply changes made to the database from the main server. The general principle of the replication in PostgreSQL is implemented in three different options with different degree of synchronicity.

- Segmented replication, also known as file-based log shipping: changes recorded on the main server are accumulated until they form a 16 MB log file (the default value). Then the file is sent to the backup servers in order to replicate the changes.

- Streaming replication: a more granular approach, in which changes are sent as they arrive, but without waiting for acknowledgement from the backups. There is still some risk of data loss in this case, but smaller than for segmented replication.

- Synchronous replication: data is stored in the write ahead log (WAL, also known as a redo log) until the main server receives confirmation from the backups that the data has been received by them. The method is slower but guarantees that data will not be lost.

The streaming replication has been chosen as a reasonable compromise between performance and protecting against data loss. Additional control of the presence of a complete set of data in all database instances can be further implemented using the EMS monitoring and statistics collection system.

To ensure continuous access, or fast restoration of access to the database service for users of the EMS, several solutions can be proposed. One solution would be to provide main and backup server addresses to the client applications and make them responsible for finding a working server. This approach has significant drawbacks, including the need to modify the existing code of all client applications by adding some monitoring and switchover logic. So instead, a new solution for server-side redundancy has been implemented based on the Virtual Router Redundancy Protocol (VRRP) [6]. The protocol involves combining several physical hosts into one virtual one and assigning a virtual IP address (VIP) to it. Clients always connect to VIP and need to be aware only of this address. Thus, using the VRRP protocol

makes it possible to provide the redundant service with automatic switchover. The Keepalived [7] package is used for implementing a fault-tolerant cluster based on the VRRP. The overall EMS cluster scheme is shown in Fig. 1.

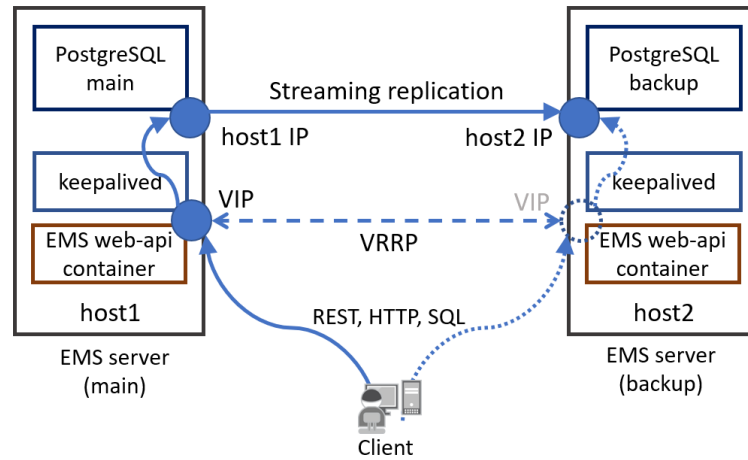


Fig. 1. High availability solution for the Event Metadata System

The described solution is quite simple and at the same time provides the necessary level of reliability and control. In particular, the backup server, when switchover to it happens after main server failure, is initially available only for reading, but can be promoted to the read-write mode by the administrator's command or by a command from the monitoring system; the latter functionality is being developed as a part of the monitoring service [8] for the BM@N information systems.

AUTOMATIC DEPLOYMENT SOLUTION

In some circumstances, a new instance of the EMS needs to be deployed. The scenarios include using the system for new experiments, development purposes, performing migrations or recovery.

To enable the rapid deployment, automatic deployment scripts based on Ansible [9] playbooks have been developed for the EMS. The system components, such as the EMS Web and API Interfaces, Keepalived and PostgreSQL database are deployed and configured on assigned hosts according to the provided settings file. This includes the ability to deploy a highly available cluster (see Fig. 1) with streaming replication and VRRP enabled.

ADDING INDEXES FOR THE EMS DATABASE

The search time for a required set of physics events in the catalogue in the worst case is determined by the time of a full scan of the database table with event metadata. With a fixed hardware configuration of the database server, the search can be optimized by adding indexes

built on certain fields. A set of tests to measure response times of the event catalogue was performed for typical catalogue queries with different indexing options (no index, BTree-based index, block-range index [BRIN]) and different table columns being indexed. Some of the obtained results are shown in Fig. 2. It is seen that in most cases using proper index allows for a significant reduction in the response time (exceptions include very heavy queries that actually require running through almost all the catalogue records, such as option 3 presented). In particular, using the BRIN index for the *period_number* column improved query times by at least factor of 2 for the queries requesting events for a given run period of the BM@N experiment. An additional advantage of BRIN over BTree is using significantly less drive space.

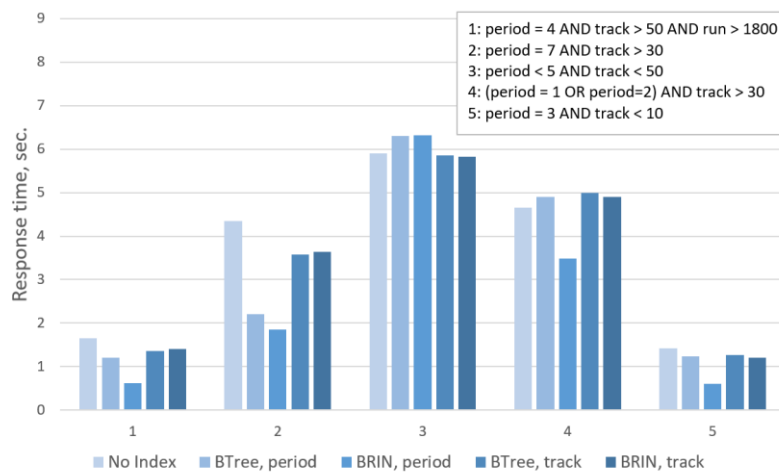


Fig. 2. An example of the impact of PostgreSQL indexes on the execution time of queries to the event catalogue with 48 million records

The efficiency of using the BRIN index depends on the distribution of the indexed value: the higher the correlation between the value and the row number, the higher the probability of skipping of an unnecessary part of the table when scanning. So, indexing for values that change almost arbitrary with every event, such as track number (Fig. 2) is not effective.

Taking into account the test results, the following overall conclusion has been made: it is advisable to use three independent BRIN indexes: on the *period_number*, *run_number*, and *software_id* fields of the event catalogue database.

CONCLUSIONS

A set of the improvements, described in the paper, has been made to the BM@N EMS that enables better availability and performance of the system. The improvements include the

ability to run EMS in highly available mode and ensure automated system deployment. Furthermore, a set of indexes has been added to make typical database queries significantly faster. Plans for future development of the EMS include its integration with the event display system [10] and development of automated indexing of new experimental events and query caching service.

REFERENCES

1. *Afanasiev S. et al.* [BM@N Collaboration] The BM@N spectrometer at the NICA accelerator complex // Nucl. Instrum. Meth. A. 2024 V. 1065. P. 169532.
2. *Gertsenberger K., Alexandrov I., Filozova I., Alexandrov E., Moshkin A., Chebotov A., Mineev M., Pryahina D., Shestakova G., Yakovlev A., Nozik A., Klimai P.* Development of Information Systems for Online and Offline Data Processing in the NICA Experiments // Phys. Part. Nucl. 2021. V. 52. P. 801.
3. *Alexandrov E., Alexandrov I., Degtyarev A., Gertsenberger K., Filozova I., Klimai P., Nozik A., Yakovlev A.* Design of the Event Metadata System for the Experiments at NICA // Phys. Part. Nucl. Lett. 2021. V. 18. P. 603.
4. *Alexandrov E., Alexandrov I., Chebotov A., Degtyarev A., Filozova I., Gertsenberger K., Klimai P., Yakovlev A.* Implementation of the Event Metadata System for physics analysis in the NICA experiments // J. Phys. Conf. Ser. 2023. V. 2438. P. 012046.
5. PostgreSQL project, “PostgreSQL” [software]. Available from <https://www.postgresql.org> [accessed 2024-11-06].
6. *Internet Engineering Task Force (IETF).* Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6. Available from <https://datatracker.ietf.org/doc/html/rfc5798> [accessed 2024-11-06].
7. Keepalived project, “Keepalived” [software]. Available from <https://github.com/acassen/keepalived> [accessed 2024-11-06].
8. *Gertsenberger K., Klimai P., Nemova O.* Development of Monitoring Service for BM@N Information Systems // Phys. Part. Nucl. Lett. 2024. V. 21. P. 793.
9. Ansible project, “Ansible” [software]. Available from <https://github.com/ansible/ansible> [accessed 2024-11-06].
10. *Blinova E., Dunaev I., Gertsenberger K., Klimai P., Nozik A.* Development of Next-Generation Event Visualization Platform for the BM@N Experiment. Phys. Part. Nucl. Lett. 2024. V. 21. P. 785.