

BM@N Run 8 data processing on a distributed
infrastructure with DIRAC
Обработка данных 8-го сеанса BM@N на
распределённой инфраструктуре под управлением
DIRAC

K. Gertsenberger^{a,1}, I. Pelevanyuk^{a,2}
К.В. Герценбергер^{a,1}, И.С. Пелеванюк^{a,2}

^a Joint Institute for Nuclear Research

^a Объединённый институт ядерных исследований

The BM@N 8th run using xenon ion beams was successfully completed in February 2023, resulting in approximately 600 million events. They were recorded in the form of around 31000 files, with a combined size exceeding 400TB. To process all these data JINR DIRAC platform was chosen. The data reconstruction of BM@N 8th physics run was the first time DIRAC had been used for data reconstruction in JINR. A set of approaches, systems, and methods were developed during this campaign, which aid in reducing the efforts required for future data reconstructions at JINR.

PACS: 07.05.-t;

Introduction

BM@N (Baryonic Matter at Nuclotron) is the first experiment undertaken at the accelerator complex of NICA-Nuclotron [1]. The aim of the BM@N experiment is to study hot and dense baryonic matter produced in interactions of relativistic heavy ion beams with fixed targets. The scientific program of the BM@N experiment comprises studies of the nuclear matter in the intermediate energy range between experiments at the SIS, other NICA experiments, and FAIR facilities.

The 8th BM@N run in the xenon ion beam was conducted between December 2022 and February 2023. BM@N collected about 500M Xe+CsI interactions at the beam kinetic energy of 3.8 AGeV and 50M interactions at the kinetic energy of 3 AGeV. All these events were recorded in the form of around 31000 files with total size of 400 TB on the EOS storage system of NICA cluster.

85% of all files have size between 14 GB and 16 GB. 15% of files have size less than 14 GB because of end of writing to a particular file when moving from run to run.

¹ORCID: 0000-0002-5753-1852

²ORCID: 0000-0002-4353-493X, E-mail: pelevanyuk@jinr.ru

The files that were originally obtained from BM@N facility is called raw files. They are recorded in a binary format by the Data Acquisition System. The raw files contain information on registered signals and metadata from different detectors within the BM@N facility. In order to allow physics analysis each file is converted from RAW format to DST (data summary tape) format. According to BM@N data processing model this transition is done in two steps. On the first step raw files converted and decoded to event trees with detector digits in the CERN ROOT [2] format (generally called DIGI format). On the second step, special macro restores information on the particles registered by the detectors, their tracks and other parameters that are relevant for further physics analysis and saves them in the DST format. The general schema is shown in Fig. 1.

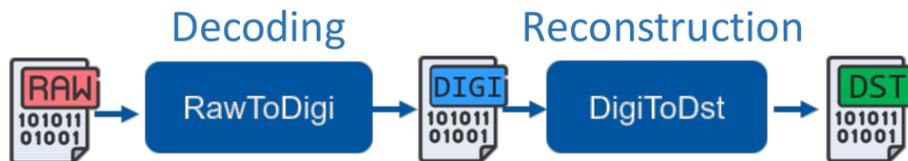


Fig. 1. Data preparation for physics analysis

The main task of the work was to develop and apply a fast and repeatable method to perform BM@N mass data processing for Run8. The DIRAC infrastructure in JINR had been chosen for that task. It provides both access to all major computing and storage resources available for the BM@N collaboration and a set of basic tools for workload and data management [3]. The EOS storage in Meshcheryakov Laboratory of Information Technologies (MLIT) is the main storage within JINR DIRAC infrastructure. So, an additional step is required before performing data processing in DIRAC, that is initial data transfer of the RAW files from the NICA cluster to EOS in MLIT and registering all the files in DIRAC FileCatalog.

Data transfer from the NICA cluster to MLIT EOS

The standard way to move files between two EOS storages in our infrastructure is by using *xrootd* protocol. DIRAC provides a command to upload file to storage and register it in FileCatalog in one transaction. There are two problems. Single *xrootd* transfer stream in our infrastructure is limited to 100 MB/s. It is possible to overcome this issue by increasing number of streams for a transfer. But, this would require some changes in the code of uploading scripts, and also will lead to a server's network overload. Second problem is a limited throughput of a server from which the transfer is performed. So, in order to transfer data fast it is necessary to transfer files from many servers simultaneously and it is possible to achieve using DIRAC.

Special DIRAC jobs were formed to perform parallel transfers. The jobs were submitted in a pack of 20. Each job starts knowing its pack size, number within the pack and a list of files that should be transferred. Each

running job follows the same principle to determine which files need to be transferred. These jobs were submitted to the NICA cluster where they were evenly distributed by the batch system. The average transfer speed was around 1.92 GB/s and the full transfer took 63 hours.

RAW to DIGI file conversion campaign

Before starting RAW data processing in a distributed infrastructure it is crucial to understand the requirements of the jobs in terms of RAM, disk and CPU requirements. With the specified information it is possible to choose computing resources that can handle the anticipated workload. Several jobs

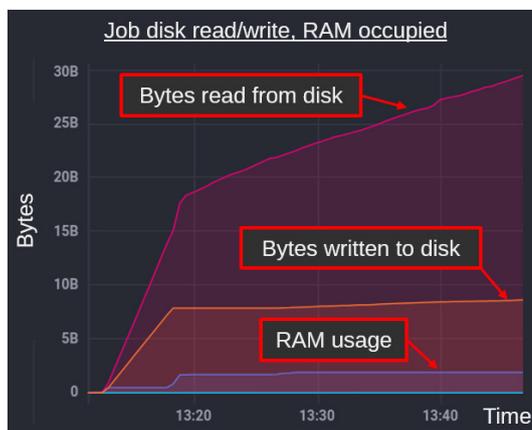


Fig. 2. Profile of the RAW to DIGI job disk and RAM usage

on conversion from RAW to DIGI were submitted to Govorun supercomputer together with a special script for job monitoring. The monitoring demonstrated (see Fig. 2) that 15 GB RAW file is processed in two steps. At the first step the raw binary data are converted to DAQ digits in the ROOT tree file with the size of about 8 GB. On the second step the DIGI file is formed after decoding the input data to detector digits (and performing detector mappings, calculating pedestals, clearing noisy channels, etc.). The average size of the DIGI files is around 1 GB. Observed RAM load was within 2.5 GB. And CPU usage was always around 100% which indicated the efficient CPU usage.

The most limiting requirement for this process is a disk size. Not all computing resources can provide 25 GB of free disk space for a job and therefore allocate the amount of the jobs corresponding to the amount of available CPU cores. This problem is even more severe for files that are larger than 16 GB. As a result, only Govorun supercomputer can handle them efficiently.

With these requirements two computing resources were chosen for this campaign: Tier1 centre and NICA cluster. Tier1 cluster provides 1500 CPU cores for the NICA experiments. The NICA cluster provides 250 CPU cores, but it was limited by 100 simultaneously running jobs to prevent disk over-load.

Before the job submission it was possible to simulate the jobs execution. For that purpose a specially developed simulation system was used [4]. It allows simulating job execution on a distributed infrastructure taking into consideration network throughput of different components, and CPU cores performance on different clusters. The following simulation parameters were used. Maximum network speed of the EOS storage, Tier1 and NICA cluster is 10 GB/s. Tier1 provides two sub-clusters with 1125 old cores (13 DB12 benchmark) and 375 new cores (24 DB12 benchmark). The NICA cluster provides of 100 cores with 24 DB12 benchmark. The amount of simulated jobs was 30000. Each job downloads 15 GB, processes 57000 DB12*seconds (calculated from test job executions), and uploads 1 GB. The results of simulation are present in Fig. 3.

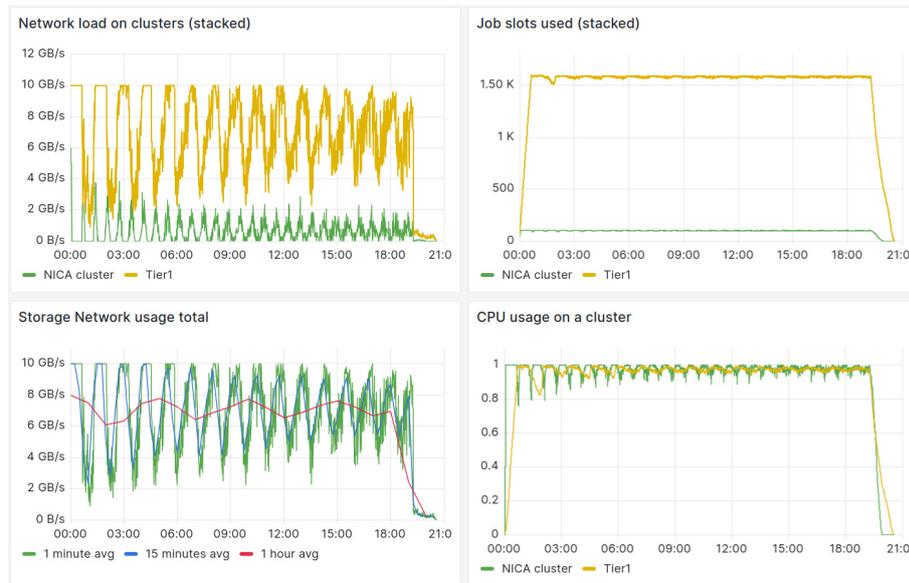


Fig. 3. Simulation of BM@N raw data processing

The graph shows occasional full network usage, but CPU usage is still acceptable. No severe "cold start" issues were observed with the default DIRAC pilot submission rate. So, the full raw data processing for BM@N Run 8 was launched. The execution took 35 hours. At the beginning, Tier1 was occupied by another workloads, but at the peak there were 1500 jobs on Tier1 and 100 jobs on the NICA cluster. The maximum transfer speed between worknodes and MLIT EOS was 7.5 GB/s. Average job execution time was around 90 minutes. The total size of the DIGI files was 23 TB. The campaign on the raw data processing for BM@N became the most data intensive workload that was executed by DIRAC in JINR.

BM@N reconstruction campaign

The profiling of the BM@N reconstruction jobs was performed in a similar way that it was done during raw data processing. The reconstruction jobs

generate DST files from the DIGI files received on the previous step. The resulting DST files usually have the doubled size of the DIGI files. So, the disk requirements were around 4 GB per running job. The profiling demonstrated similar RAM and CPU requirements.

With obtained requirements it was possible to submit the reconstruction jobs to all major computing resources available for the BM@N collaboration: to Tier1, Tier2, NICA cluster and Govorun supercomputer. The peak of amount of executing jobs was around 3100 that were distributed as follows: Tier1 - 1500 jobs, Tier2 - 1000 jobs, NICA cluster - 200 jobs, Govorun - 400 jobs. The average execution time job was around 3 hours. The full reconstruction campaign took around 30 hours. The total size of the DST files is 53 TB.

Once the DST files were produced it was necessary to upload them back to the NICA cluster for physics analysis. It was done within a day in a manner similar to what was described in "Data transfer from NICA cluster to MLIT EOS" section.

Conclusion

The task of development and application of a fast and repeatable method to perform BM@N mass data processing for Run8 was successfully completed. It was the first usage of DIRAC infrastructure in JINR for working with experimental data. Reconstruction process for obtained raw data is crucial for successful operation of experimental facilities. During this mass data processing we recorded different metrics related to the distributed nature of workload execution and studied the behavior of the infrastructure with data intensive tasks.

The developed DIRAC computing infrastructure proved to be flexible, efficient and powerful for massive workloads. The developed and applied methods of jobs execution and monitoring will be also used for future campaigns.

REFERENCES

1. *Kapishin M.* Studies of baryonic matter at the BM@N experiment (JINR) // Nuclear Physics A. — 2019. — V. 982. — P. 967–970.
2. *Brun R., Rademakers F.* ROOT: An object oriented data analysis framework // Nucl. Instrum. Meth. A. — 1997. — V. 389. — P. 81–86.
3. *Korenkov V., Pelevanyuk I., Tsaregorodtsev A.* DIRAC at JINR as a general purpose system for massive computations // Journal of Physics: Conference Series. — 2023. — feb. — V. 2438, no. 1. — P. 012029.
4. *Pelevanyuk I.S., Campis D.* Simulation of Job Execution in Distributed Heterogeneous Computing Infrastructures // Physics of Particles and Nuclei Letters. — 2023. — Oct. — V. 20, no. 5. — P. 1276–1278.