ISSN 1063-7796, Physics of Particles and Nuclei, 2021, Vol. 52, No. 4, pp. 835–841. © Pleiades Publishing, Ltd., 2021.

Integration of Distributed Heterogeneous Computing Resources for the MPD Experiment with DIRAC Interware

N. Kutovskiy[a], V. Mitsyn[a], A. Moshkin[a], I. Pelevanyuk[a, *], D. Podgayny[a], O. Rogachevsky[a], B. Shchinov[a], V. Trofimov[a], and A. Tsaregorodtsev[b]

[a] *Joint Institute for Nuclear Research, Dubna, Moscow oblast, 141980 Russia*

[b] *CPPM, Aix-Marseille University, CNRS/IN2P3, Marseille, France*

**e-mail: pelevanyuk@jinr.ru*

Received October 20, 2020; revised November 23, 2020; accepted November 23, 2020

Abstract—Computing and storage resources are essential for efficient Monte Carlo generation. In JINR there are several types of computing resources: Tier1 and Tier2 grid clusters, Govorun supercomputer, JINR Cloud, and NICA cluster. There are also EOS disk and dCache tape storage systems. In order to use them, users have to be aware of many details and differences between resources and keep track of load on all of them. The DIRAC Interware was adopted, configured, and expanded to fulfill requirements of massive centralized Monte Carlo generation for the multipurpose detector. For a year, the all the infrastructure was used via DIRAC to run successfully around 500000 jobs with an average duration of 5 h each. The use of DIRAC allowed for unified data access, performance estimation, and accounting across all resources.

DOI: 10.1134/S1063779621040419

1. MULTIPURPOSE DETECTOR EXPERIMENT

The multipurpose detector (MPD) is designed to study heavy-ion collisions at the Nuclotron-based heavy Ion Collider fAcility (NICA) at JINR, Dubna [1]. A conceptual design of the MultiPurpose Detector (MPD) is proposed for the study of hot and dense baryonic matter in collisions of heavy ions over the atomic mass range A = 1–197 at center-of-mass energy up to $\sqrt{s_{NN}} = 11$ GeV (for Au79+).

The detector is in the development stage right now. The computing system for MPD is also being developed. When the detector will be put into operation the average data flow is expected to be around 10 GB/s. Until that time, physicists need data from Monte Carlo simulation for preliminary analysis and software testing. Right now, data from Monte Carlo simulation are the best choice to test both the software and the infrastructure. It is required to generate and reconstruct billions of events. Each event has the size of around 1 MB, which means that petabytes of data have to be generated, transferred, stored, and analyzed.

2. COMPUTING RESOURCES AT JINR

Several types of computing resources are available at the Joint Institute for Nuclear Research. They all have different purposes, features, and areas of use. A big group of resources is integrated in the multifunctional information and computing complex (MICC). This infrastructure was built to give access to computing and storage resources for a wide range of users, scientific groups, and virtual organizations including MPD. It consists of the following computing resources: Tier1/Tier2 clusters, Govorun supercomputer, JINR computing cloud. Storage resources in MICC are represented by two systems: EOS for disk-only storage and dCache for disk and tape storage. Apart from MICC, there is a dedicated cluster for all NICA experiments: NICA offline cluster. It groups together both computing and storage resources.

2.1 Tier1/Tier2 Clusters

The Tier2 cluster has been in operation at JINR since 2003 [2]. It provides access to computing with grid protocols for many virtual organizations: ATLAS, CMS, ALICE, LHCb, BES-III, NOVA, and others. The CREAM software was used for the computing element (CE) service. Now transition to the ARC CE is in process. The operating system on the worker nodes is Scientific Linux.

A separate queue for the MPD users was created. The peak amount of slots available for MPD in Tier1 is around 450 and on Tier2 it is around 500. Users need to have a grid certificate and should be registered in the MPD virtual organization (VO) in order to use these resources. It is possible to use it directly with the gLite middleware client via a command-line interface.

2.2. Govorun Supercomputer

The supercomputer Govorun was created in 2018 [3]. Its main purpose is to run parallel jobs or jobs with special requirements of Input/Output. It is equipped with powerful processors, a big amount of RAM and fast disks. This makes the Govorun supercomputer almost a universal tool for a wide range of computational tasks. The SLURM software is used as a workload management system on the supercomputer. The Scientific Linux operating system is installed on the worker nodes.

A separate queue was created for the Monte Carlo generation for the MPD experiment. Generally, around 380 job slots are available, but in case of supercomputer underload, the quota may be increased temporarily up to 670. Users have to be registered in the SLURM system to be able to submit jobs. The authorization to run jobs in the MPD queue is determined by the groups to which the users belong.

2.3. JINR Cloud

The JINR cloud service running on the OpenNebula platform is built upon the Infrastructure as a Service (IaaS) model [4]. It is a flexible resource that is used for development, service hosting, and also computational tasks. There are several approaches to perform computational tasks on the cloud: run it directly on virtual machines which are created manually, or setup a workload management system on virtual resources, and submit jobs via its help. Some JINR cloud resources were dedicated to the MPD workflows testing (40 single CPU core virtual machines in total). But during a production period, MPD jobs can potentially occupy up to 2000 CPU cores in the JINR cloud while it is underloaded.

2.4. NICA Offline Cluster (Computing and Storage)

The NICA cluster is a resource that is fully dedicated to experiments on NICA including the MPD. It was put into operation in 2019. Since that time more than 1.5 million tasks have been successfully executed and more than 800 TB of data recorded by members of rgw MPD collaboration. It allows running more than 5000 jobs simultaneously. Sun Grid Engine is used as a workload management system on the cluster. EOS is used as a local storage system, total size is 12 PB. It is accessible on all worker nodes of the NICA cluster.

In terms of access, the NICA cluster and the Govorun supercomputer are similar. The MPD users have to be registered on the cluster to submit jobs. Jobs are submitted directly to the workload management system. By default, every user may have up to 250 jobs running. The operating system on the worker nodes is CentOS.

2.5. MICC EOS Storage

EOS storage on MICC is originally connected to all MICC computing resources. It is used as disk storage. The MPD users have a big quota shared across all of them. There are two main modes of accessing the storage: local access from MICC components and worldwide access through xRootD protocol. Authentication and authorization may be done either through Kerberos or with an x509 grid certificate with the VOMS extensions.

2.6. MICC dCache Storage

The dCache storage is similar to EOS. The main difference is that apart from access to disks it allows using tape storage as well. The main access mode is through grid protocols. It supports a wide range of protocols: DCAP, SRM, GridFTP, and xRootD. The tape works through dedicated disk buffer servers. The time required to select the right tape and transfer data from tape to disk depends on the tape library task queue. Generally, the time varies from 20 s up to several minutes. Tapes are ideal for long term storage (archiving) of raw data which will be received from the MPD detector.

The described resources are mostly heterogeneous. They have different hardware, different access protocols, different authentication and authorization procedures, a different set of associated storage resources available on them. Centralized use of all of them by the MPD experiment is not possible without some additional efforts. That is why we had to develop an approach for their combined usage. Low-level integration of all of them is possible but would need a substantial amount of work to allow interconnection between all the components. Another approach is to use some additional tools for the integration of all resources into a bigger system.

3. DIRAC INTERWARE AT JINR

The DIRAC Interware (later on just DIRAC) is a software developed originally by the LHCb experiment to support its computing operations. Since 2008 it is developed as a general-purpose tool for various scientific groups. The DIRAC project provides a development framework and a large number of ready-to-use components to build distributed computing systems of arbitrary complexity. DIRAC services ensure the integration of computing and storage resources of different types and provide all the necessary tools for managing user tasks and data in distributed environments [5].

The DIRAC instance at JINR has been installed, studied, and gradually improved since 2017. There were four main reasons for us to choose DIRAC as a candidate for the integration of resources:

(1) It is a single tool for all aspects of computing for an experiment: workload management, data management, access and shares management, accounting, workflow management, monitoring.

(2) Its community is wide and active. It is used by many different scientific groups which represent several experiments in physics, astronomy, biological research, and medicine.

(3) It demonstrates good performance [6].

(4) It is possible to extend it with custom components if needed.

Integration by DIRAC does not mean just the integration of different computing and storage resources into the DIRAC infrastructure. It also means the integration of user workflows. This may be achieved only by the collaborative work of all the parties: users, resource administrators, and DIRAC administrators.

3.1. Integration of Computing Resources

Integration of a computing resource in DIRAC means the ability of the resource to accept and run DIRAC pilot jobs. The DIRAC pilot job is a special job that works as a wrapper for a user workload. After starting on the worker node, the pilot job performs a set of operations. It sets up the DIRAC environment, performs basic checks and tests, requests a user job from the DIRAC job queue. A user job is matched to the pilot by the DIRAC service only if the resource is suitable for the job requirements. Tests and checks provide all relevant information for taking this decision. The DIRAC environment on the worker node provides all the necessary tools for the job to download and upload data, change status, and communicate with various DIRAC services.

Tier1 and Tier2 being originally designed to run grid jobs have been integrated in the first place. MPD VO and the VOMS server were configured to allow execution on grid Computing Elements. At that time, the CREAM CE was used. Now the transfer to the ARC CE is being done and Tier1 jobs are submitted to the ARC CE already.

The JINR Cloud integration was started in parallel. Everything related to running a DIRAC pilot also applies to clouds, but to run a pilot a dedicated virtual machine should be instantiated by DIRAC in advance. The standard approach at that time was to use the rOCCI command-line tool to communicate with different clouds over the OCCI protocol for the creation and deletion of virtual machines for pilots. This approach suffered from a substantial overhead caused by the need to use the rOCCI tool itself and supporting the OCCI server on all clouds. Even though this approach worked it was decided to extend DIRAC to create and delete virtual machines in the JINR cloud using the OpenNebula native XML-RPC protocol. A special module was developed and included in the DIRAC source code to allow direct use of OpenNebula clouds [7]. DIRAC at JINR was the first installation to use it. Now it is used also by the BES-III and JUNO installations.

Development of the OpenNebula module for DIRAC allowed integration of all JINR Member States clouds in the JINR DIRAC installation. Now, for the Member States, it is a good opportunity to participate in computing for the MPD experiment. Right now the main issue with using cloud resources is software distribution. A single effective way to distribute it on clouds is by using CVMFS. Unfortunately, this was adopted for the MPD Monte Carlo generation only recently and not all generation campaigns use it.

Next, the Govorun supercomputer was integrated into DIRAC. That required to perform several big steps. A dedicated user account was created and included in the MPD group. Initial tests demonstrated that by default pilots work extremely slow. The cause of that was the use of a shared file system by pilots to run on. High load on the IO system during pilot job initialization prevented running a substantial amount of jobs simultaneously. Pilot initialization requires to extract around 10000 files from a 200MB tar archive. With a large number of jobs, the shared file system could be even blocked. To overcome this issue the access to /tmp directory which is placed directly on disk was granted to the DIRAC pilots. That let us run on the Govorun up to 700 DIRAC jobs simultaneously without any issues.

The NICA cluster became operational in DIRAC just recently. The integration of the NICA cluster went similarly to the Govorun supercomputer integration. A dedicated user account was created and access to the /tmp directory was granted. The use of the /tmp directory requires additional attention from the resource administrators. By default, DIRAC pilots clean all directories after execution. But sometimes, when pilots are stopped due to external circumstances, some parts may be left over. In case of the use of a shared file system, they may be deleted manually by the DIRAC administrator. In case of running in the /tmp directory the rights to access and clean this directory on worker nodes belong to resource administrators.

The most distant resource added and tested in the JINR DIRAC infrastructure up to now is a cluster from the National Autonomous University of Mexico (UNAM). UNAM is participating in the MPD collaboration. The batch system in UNAM is Torque. A thousand of MPD jobs were completed there.

3.2. Integration of Storage Resources

The integration of storage resources is crucial for effective operations. They provide the possibility to read and write files on a worldwide scale preserving authentication and authorization methods used in DIRAC. Their integration in DIRAC is less flexible compared with the computing resources. Storage

Table 1. (Color online) Amount of slots and average CPU core performance of the different resources

Resource	Current slots	Average DB12
Govorun	375	24.2
LHEP cluster	250	21.2
Tier1	440	14.05
Tier2	500	14.75
JINR Cloud	40	18.5

resources to be integrated have to support at least one of the grid transfer protocols. If the storage does not support any of those protocols it is possible to setup a service that will work as a proxy. DIRAC provides this kind of service called DIRAC Storage Element. This is a viable temporary solution, but it would not work effectively in case of a high load and would require additional support.

In the beginning, the dCache system in JINR fully supported grid access to storage. Two protocols were tested and applied: GridFTP and xRootD. It gave access to both disk storage and tape storage.

Later on, the EOS storage was installed in MICC. Once the access with grid certificates and VOMSes was configured, it became possible to download and upload data from all resources integrated into DIRAC. Now MICC EOS is used by DIRAC to access to disks and dCache/Enstore [8] used to access tapes.

3.3. Integration of the Centralized MPD Monte Carlo Generation Workflow

Initially, centralized Monte Carlo generation for MPD was performed directly on the Govorun supercomputer. The software for Monte Carlo was built for the Govorun supercomputer. The data produced during the Monte Carlo generation was stored in the local Govorun storage. But the number of jobs that had to be completed required to find and use additional computing resources. By that moment, Tier1 and Tier2 clusters were integrated into DIRAC which made them a good resource to try.

The first step was to receive an RDIG certificate and join a virtual organization. The second step was to learn how jobs may be modified to be executed inside the DIRAC pilot. To run jobs on Tier1/Tier2 the software had to be built for them and should be available on local storage on Tiers. AFS worked as storage with software accessible across all Tiers' worker nodes. The results of jobs should be uploaded to MICC EOS storage. DIRAC provided all the necessary tools for the data upload. The advantage of MICC EOS is that all data uploaded using grid protocols during a DIRAC job execution are also accessible locally on the resources where MICC EOS is mounted. That helped to overcome some initial issues with grid access to the EOS storage. Now, the EOS storage in MICC works reliably. Direct access to the data on EOS is used mostly for analysis purposes.

The third step of the MPD workflow integration was the use of the DIRAC Data Management System which is closely related to the use of the DIRAC File Catalog. This service provides a single logical namespace for all files uploaded with DIRAC. It allows abstracting away from physical names and different paths on different storage resources. Another feature of the DIRAC File Catalog is the ability to assign arbitrary meta-information about data. For the MPD data it is information about beams collided, energy, generators, and others. That allows flexible filtering which is not bound to directories structure.

4. RESULTS, 15 MONTHS OF CENTRALIZED MPD MONTE CARLO GENERATION WITH DIRAC

The MPD experiment has been using DIRAC at JINR for centralized MPD Monte-Carlo generation since August 2019. By the moment of writing this article, October 2020, the total amount of wall time consumed by these jobs is around 325 yr, or, in other words, approximately 2.85 millions h. Normalized CPU time is shown in Fig. 1. Almost 500 millions MPD Monte Carlo jobs have been successfully completed during the centralized generation campaign (Fig. 2). On average, an MPD job lasted for 5 h and a half. These jobs generated 1.23 millions files. The total size of data accounted for by DIRAC is 130 TB.

The centralized use of resources allowed unified accounting of consumed resources. The overall use and performance of a resource depends mostly on three factors: the number of job slots provided, the performance of CPU cores given to jobs, the amount of jobs submitted by users to a particular resource. The total amount of slots and the average CPU core performance are listed pin Table 1. All data are received from DIRAC accounting and pilot jobs finished on resources. The fastest cores right now are on the Govorun supercomputer. The LHEP cluster is not too far from the supercomputer in terms of single-core performance.

Before running a user job, the DIRAC pilot performs a fast benchmark of the core it is running on. The benchmark that DIRAC uses is DIRAC Benchmark 2012 or DB12. It well correlates with the HEP-SPEC2006 benchmark but takes much less time to complete (less than 1 min). This information is crucial for a resource performance estimation. It also gives a hint about resource structure and user workload features.

The bigger estimation of the DB12 benchmark test, the less time it will take to complete the user job on the

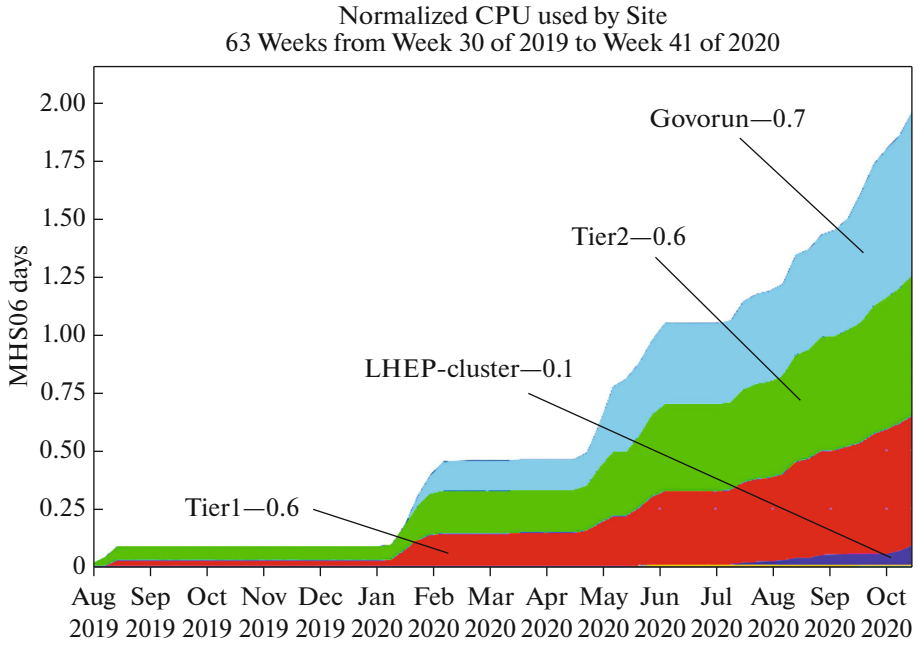

Fig. 1. Cumulative amount of normalized CPU time provided by site in millions of HS06 days. Time span from August 2019 to October 2020.

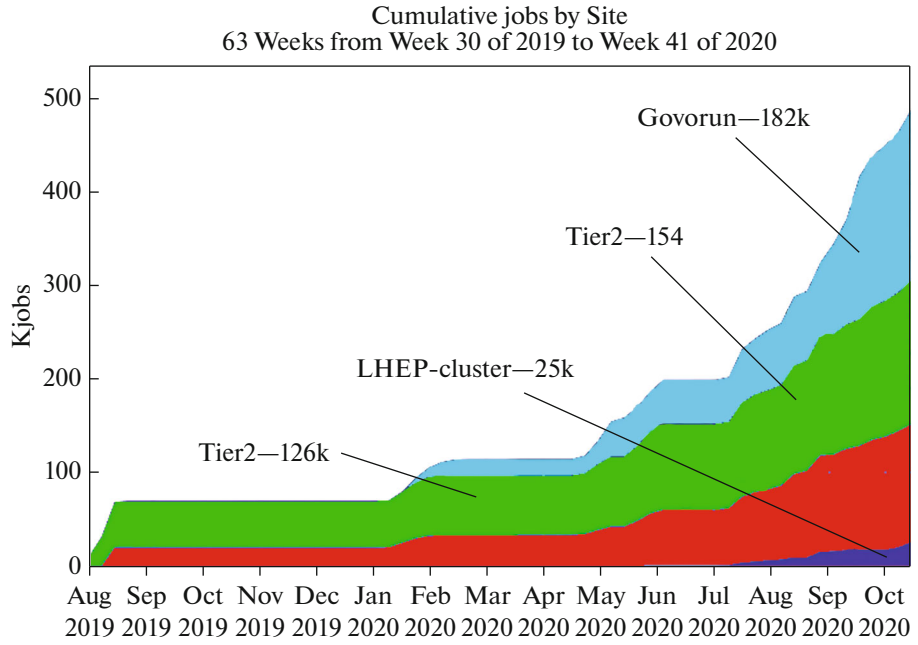

Fig. 2. Cumulative number of jobs done by site in thousands jobs. Time span from August 2019 to October 2020.

resource. An example of the correlation plot between DB12 benchmark and wall time is presented in Fig. 3. It is known that at the time of the plot generation jobs of the same complexity were submitted to all resources. It is clearly visible that the Tier2 cluster consists of several "subclusters". It is also noticeable that the LHEP cluster has a bigger dispersion of wall time as a function of DB12values. The reason is that three other resources on the plot are part of the MICC infrastructure. They are in the same place with a similar connection to MICC EOS storage. Data generated on the LHEP cluster have to go with a longer network path and sometimes it results in additional wall time for the jobs.

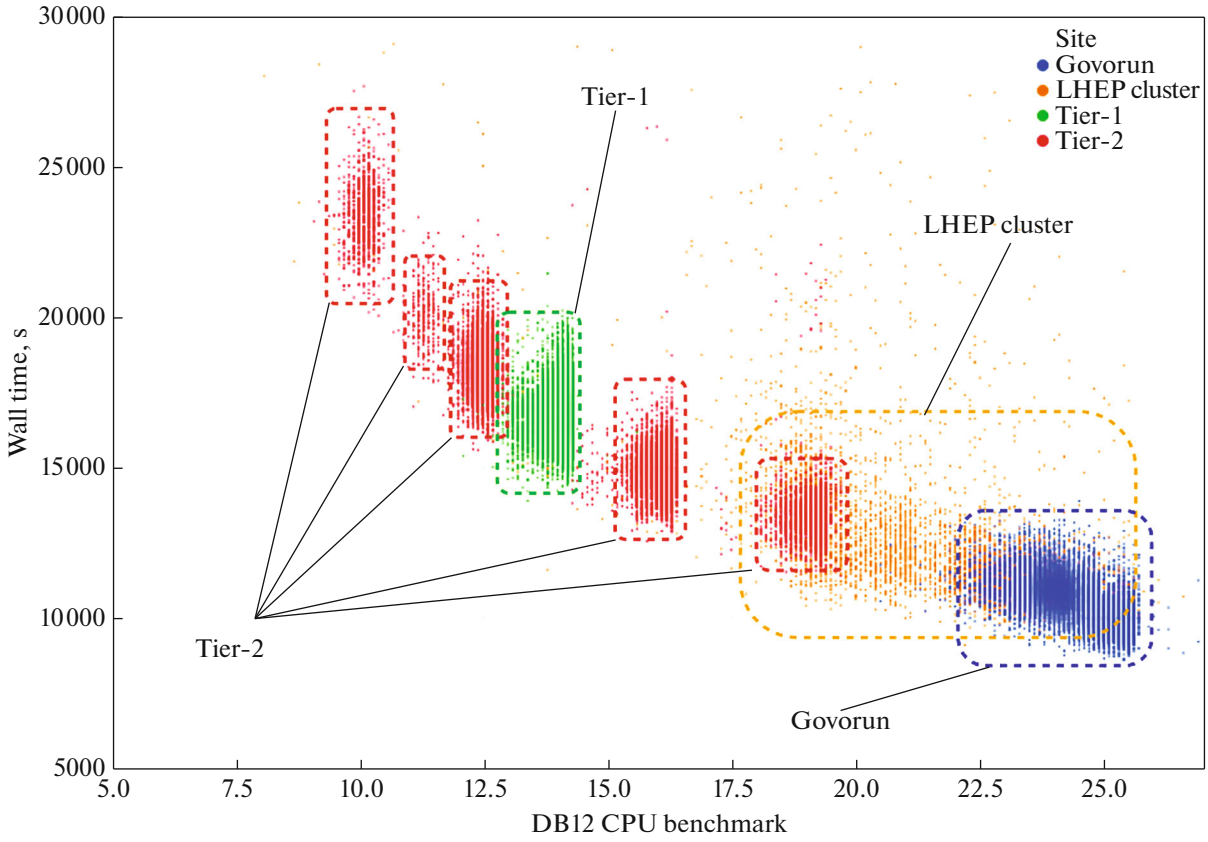

Fig. 3. User jobs duration depending on the results of DB12 benchmarks.

The plot provides a good overview of the structure, performance, user workload, network connection between resource and storage [9].

5. CONCLUSIONS

Unification of access to distributed resources in JINR, its Member States, and collaborators allowed a substantial increase of computing power available through "one window." For users, this simplifies the job submission and management process. For resources, it increases their utilization.

The DIRAC Interware was used to integrate heterogeneous resources into one large system. It provided all the necessary tools for integration into a single system to support all the computing tasks including workload management and data management. The features of DIRAC allow efficient use by different users.

The MPD experiment was the first to adopt DIRAC at JINR as a tool for centralized Monte Carlo generation. Processes and methods of using different resources were established. Right now MPD is the biggest computing resource consumer on MICC comparing to other NICA experiments. In total: 500000 jobs were executed successfully, it took around 5–6 h of wall time to complete each job. More than 1 million files with a total volume of 130 TB were generated.

ACKNOWLEDGMENTS

We wish to acknowledge the contribution of UNAM, especially our colleague Luciano Diaz. We wish to acknowledge the support of all our colleagues from teams responsible for different resources. We wish to acknowledge the directorate support on all levels during these years.

FUNDING

This work was supported by grant no. 18-02-40101 from the Russian Foundation for Basic Research.

REFERENCES

1. MPD Collab., Phys. At. Nucl. **76**, 1 (2013).
2. N. S. Astakhov et al., "JINR Grid Tier-1@Tier-2," CEUR Workshop Proc. **2023**, 68 (2017).
3. G. Adam et al., "IT-ecosystem of the HybriLIT heterogeneous platform for high-performance computing and training of IT-specialists," CEUR Workshop Proc. **2267**, 638 (2018).

4. A. V. Baranov, N. A. Balashov, N. A. Kutovskiy, and R. N. Semenov, "JINR cloud infrastructure evolution," Phys. Part. Nucl. Lett. **13**, 672 (2016).
5. V. Gergel, V. Korenkov, I. Pelevanyuk, M. Sapunov, A. Tsaregorodtsev, and P. Zrelov, "Hybrid distributed computing service based on the Dirac interware," in *Data Analytics and Management in Data Intensive Domains* (Cham, Springer, 2017).
6. P. Charpentier, "Benchmarking worker nodes using LHCB productions and comparing with HEPSPEC06," J. Phys.: Conf. Ser. **898**, 082011 (2017).
7. N. Balashov, R. Kuchumov, N. Kutovskiy, I. Pelevanyuk, V. Petrunin, and A. Tsaregorodtsev, "Cloud integration within the Dirac interware," CEUR Workshop Proc. **2507**, 256 (2019).
8. A. Baginyan, A. Balandin, S. Belov, A. Dolbilov, A. Golunov, N. Gromova, I. Kadochnikov, I. Kashunin, V. Korenkov, V. Mitsyn, I. Pelevanyuk, S. Shmatov, T. Strizh, V. Trofimov, N. Voytishin, and V. Zhiltsov, "The CMS Tier1 at JINR: Five years of operations," CEUR Workshop Proc. **2267**, 1 (2018).
9. V. Korenkov, I. Pelevanyuk, and A. Tsaregorodtsev, "Integration of the JINR hybrid computing resources with the Dirac interware for data intensive applications," Commun. Comput. Inf. Sci. **1223**, 31 (2020).