JINR WLCG TIER1 & TIER2/CICC ACCOUNTING SYSTEM

I.A. Kashunin[1,a], V.V. Mitsyn[1], T.A. Strizh[1]

[1] *Meshcheryakov Laboratory of Information Technologies, Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russia*

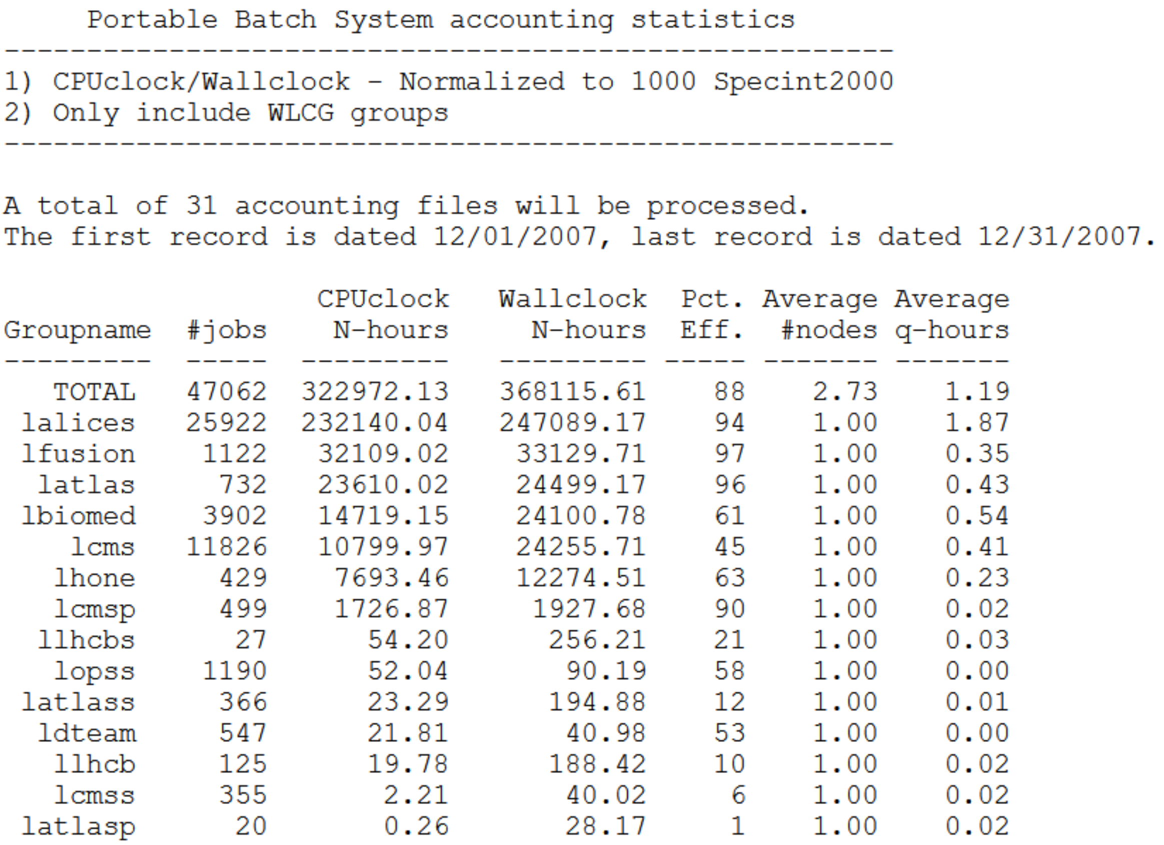
E-mail: [a] miramir@jinr.ru

The problem of evaluating the efficiency of the JINR MLIT grid sites has always been topical. At the beginning of 2021, a new accounting system was created, it managed to fully cover the functionality of the previous system and further expand it. The article provides detailed information on the implemented accounting system.

Keywords: Accounting, WLCG, Grid, Monitoring

1. Original accounting system

Since the beginning of 2001, MLIT JINR started creating a distributed system for processing, storing and analyzing experimental data from the experiments at the Large Hadron Collider (LHC) using grid technologies [1]. In 2003, the Russian segment of the LCG Global Infrastructure was organized under the Russian Data Intensive Grid Consortium (RDIG) [2], and the Tier2 grid site for data processing within the distributed computing infrastructure began functioning at JINR. The Linux operating system, the Torque batch processing system [3], the Maui task scheduler [4] and the software stack that ensures consistent work within the distributed grid infrastructure are the main software of the computing cluster. The Torque and Maui systems were finalized at MLIT and, by default, had a built-in script for collecting statistics on the batch system. The script was run with various parameters at a certain time and generated text files (Fig. 1) containing information about the operation of the system. The collected data was used to evaluate the effectiveness of jobs performed on the grid site, to account for them by users and compile various reports. Jobs were grouped according to their belonging to experiments and virtual organizations included in the RDIG, data processing from which was carried out on the site (lalice – Alice experiment at the LHC, lcms – CMS experiment at the LHC, etc.).



```
     Portable Batch System accounting statistics
-------------------------------------------------------
1) CPUclock/Wallclock - Normalized to 1000 Specint2000
2) Only include WLCG groups
-------------------------------------------------------

A total of 31 accounting files will be processed.
The first record is dated 12/01/2007, last record is dated 12/31/2007.

                      CPUclock   Wallclock  Pct. Average Average
Groupname  #jobs       N-hours     N-hours  Eff.  #nodes q-hours
---------  -----     ---------   --------- ----- ------- -------
   TOTAL   47062     322972.13   368115.61    88    2.73    1.19
  lalices  25922     232140.04   247089.17    94    1.00    1.87
  lfusion   1122      32109.02    33129.71    97    1.00    0.35
   latlas    732      23610.02    24499.17    96    1.00    0.43
  lbiomed   3902      14719.15    24100.78    61    1.00    0.54
     lcms  11826      10799.97    24255.71    45    1.00    0.41
    lhone    429       7693.46    12274.51    63    1.00    0.23
    lcmsp    499       1726.87     1927.68    90    1.00    0.02
   llhcbs     27         54.20      256.21    21    1.00    0.03
    lopss   1190         52.04       90.19    58    1.00    0.00
  latlass    366         23.29      194.88    12    1.00    0.01
   ldteam    547         21.81       40.98    53    1.00    0.00
    llhcb    125         19.78      188.42    10    1.00    0.02
    lcmss    355          2.21       40.02     6    1.00    0.02
  latlasp     20          0.26       28.17     1    1.00    0.02
```

Figure 1. Original accounting system

2. Prerequisites for the development of a new system

In mid-2020, due to the obsolescence and lack of support for Torque and Maui, it was decided to switch to a new cluster management and task dispatching system for large and small Linux clusters, SLURM [5]. Unlike Torque and Maui, SLURM does not allow a number of different default parameters to be displayed:

- CPUclock – CPU time spent on job execution for a certain period of time;
- Wallclock – total astronomical time spent by the job for a certain period of time;
- efficiency of using the cluster computing resources by the job.

These parameters can be calculated from the database (DB) of the SLURM system, which is formed during the operation of the system. To do it, there is a need to write a special script and create your own accounting system, which will allow obtaining the necessary parameters upon request.

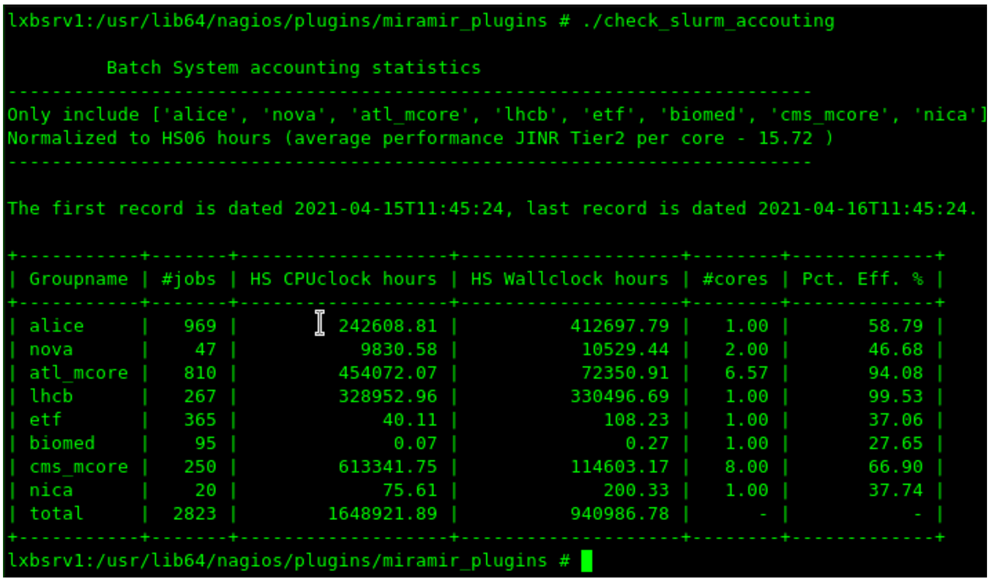
3. Development of an accounting system

At the end of 2020, the SLURM system was put into operation. At the same time, the development of an accounting system started. The initial task was to develop a console version that would display data for a certain period of time in a tabular format on the screen. Similar to the original accounting system, this data can be stored in the form of reports. The main parameters that the accounting system should display for a certain period of time are:

- number of jobs grouped by the affiliation to an experiment/group;
- CPUclock of grouped jobs;
- Wallclock of grouped jobs;
- average number of cores used by the job in a group;
- efficiency of using the cluster computing resources by a job group.

Terms such as CPUclock and Wallclock have analogs, i.e. total_cpu_time and elapsed_time in SLURM, respectively. SLURM has a special command *sacct* to display various system parameters on the screen. The output is in a tabular format, where the required parameters are displayed for each job. Due to this, it is possible to calculate the amount of CPUclock for a job group as the sum of total_cpu_time for each job for a certain period of time. Wallclock is calculated in the same way, only in this case the amount is calculated from elapsed_time. The efficiency of using the cluster computing resources by a job group can be calculated as the percentage of these parameters to the number of cores allocated for each job. This parameter is especially useful for detecting cases when a job reserves cores without using all the processor power. Thus, the poorly optimized code can be identified. Batch queues can also comprise jobs with a different number of cores required for computing. The script of the accounting system was developed taking into account this feature.

As a result, an accounting system script was created, it allows displaying the main parameters on the screen (Fig. 2).



```
lxbsrv1:/usr/lib64/nagios/plugins/miramir_plugins # ./check_slurm_accouting

         Batch System accounting statistics
------------------------------------------------------------------------
Only include ['alice', 'nova', 'atl_mcore', 'lhcb', 'etf', 'biomed', 'cms_mcore', 'nica']
Normalized to HS06 hours (average performance JINR Tier2 per core - 15.72 )
------------------------------------------------------------------------

The first record is dated 2021-04-15T11:45:24, last record is dated 2021-04-16T11:45:24.

+-----------+-------+-------------------+--------------------+--------+-------------+
| Groupname | #jobs | HS CPUclock hours | HS Wallclock hours | #cores | Pct. Eff. % |
+-----------+-------+-------------------+--------------------+--------+-------------+
| alice     |   969 |         242608.81 |          412697.79 |   1.00 |       58.79 |
| nova      |    47 |           9830.58 |           10529.44 |   2.00 |       46.68 |
| atl_mcore |   810 |         454072.07 |           72350.91 |   6.57 |       94.08 |
| lhcb      |   267 |         328952.96 |          330496.69 |   1.00 |       99.53 |
| etf       |   365 |             40.11 |             108.23 |   1.00 |       37.06 |
| biomed    |    95 |              0.07 |               0.27 |   1.00 |       27.65 |
| cms_mcore |   250 |         613341.75 |          114603.17 |   8.00 |       66.90 |
| nica      |    20 |             75.61 |             200.33 |   1.00 |       37.74 |
| total     |  2823 |        1648921.89 |          940986.78 |      - |           - |
+-----------+-------+-------------------+--------------------+--------+-------------+
lxbsrv1:/usr/lib64/nagios/plugins/miramir_plugins #
```

Figure 2. Console view of the accounting system

The next task is to develop a visualization system. Currently, there are various options for choosing software for data visualization systems. Since 2014, a monitoring system [6] has been operating at the Multifunctional Information and Computing Complex (MICC) [7], and one of its components is the Grafana visualization system [8]. It was decided to build a visualization system on its basis. That enabled easier integration into the existing monitoring system.

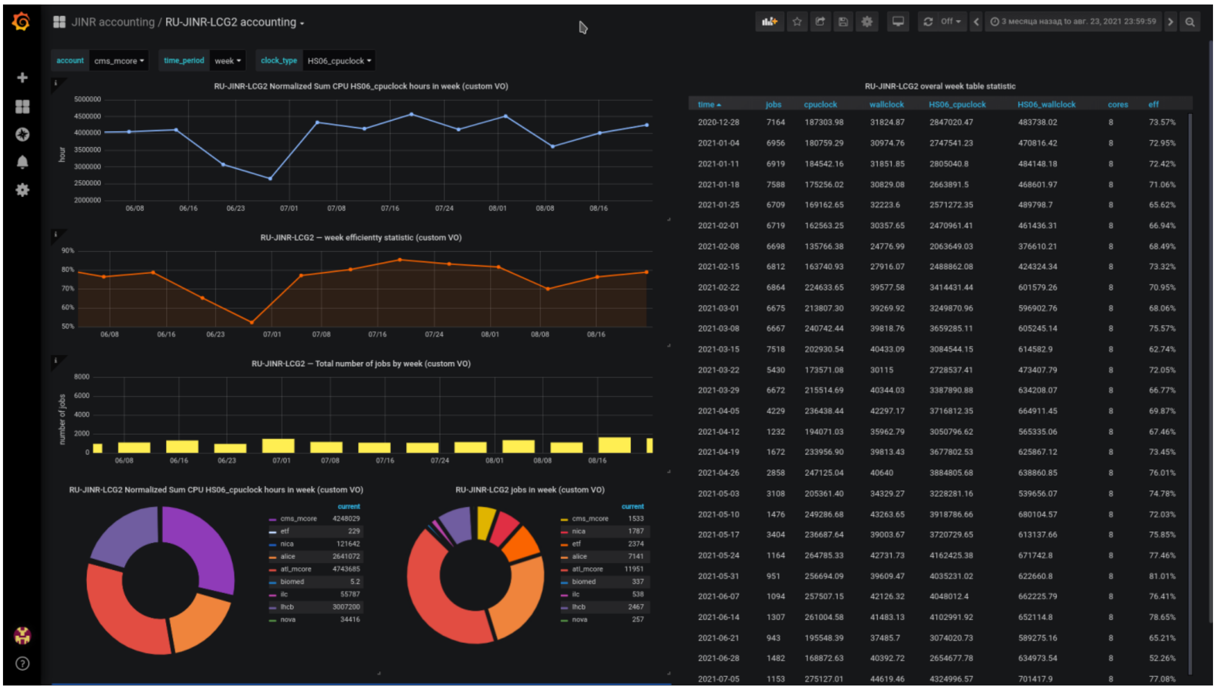
To display statistics, the Grafana system receives data from a specific resource, i.e. a backend (database, software, socket, etc.). One of them is the MariaDB database [9]. To connect the accounting system with Grafana, special support for writing data to the accounting system script was added. The script is launched by the operating system through the standard service *cron* [10] and writes the parameters for the day, week, month and year to the database.

4. Vizualization system

An interactive visualization system was created on the basis of Grafana, it allows providing the user with the most up-to-date analytical data:

- graphs and pie charts of CPUclock and Wallclock for a certain period of time;
- graphs for the use of the cluster computing resources by jobs;
- graph of the number of completed jobs;
- tabular version of the displayed data.

As a result, an information display (Fig. 3) was created.



RU-JINR-LCG2 overal week table statistic

time ▲	jobs	cpuclock	wallclock	HS06_cpuclock	HS06_wallclock	cores	eff
2020-12-28	7164	187303.98	31824.87	2847020.47	483738.02	8	73.57%
2021-01-04	6956	180759.29	30974.76	2747541.23	470816.42	8	72.95%
2021-01-11	6919	184542.16	31851.85	2805040.8	484148.18	8	72.42%
2021-01-18	7588	175256.02	30829.08	2663891.5	468601.97	8	71.06%
2021-01-25	6709	169162.65	32223.6	2571272.35	489798.7	8	65.62%
2021-02-01	6719	162563.25	30357.65	2470961.41	461436.31	8	66.94%
2021-02-08	6698	135766.38	24776.99	2063649.03	376610.21	8	68.49%
2021-02-15	6812	163740.93	27916.07	2488862.08	424324.34	8	73.32%
2021-02-22	6864	224633.65	39577.58	3414431.44	601579.26	8	70.95%
2021-03-01	6675	213807.30	39269.92	3249870.96	596902.76	8	68.06%
2021-03-08	6667	240742.44	39818.76	3659285.11	605245.14	8	75.57%
2021-03-15	7518	202930.54	40433.09	3084544.15	614582.9	8	62.74%
2021-03-22	5430	173571.08	30115	2728537.41	473407.79	8	72.05%
2021-03-29	6672	215514.69	40344.03	3387890.88	634208.07	8	66.77%
2021-04-05	4229	236438.44	42297.17	3716812.35	664911.45	8	69.87%
2021-04-12	1232	194071.03	35962.79	3050796.62	565335.06	8	67.46%
2021-04-19	1672	233956.90	39813.43	3677802.53	625867.12	8	73.45%
2021-04-26	2858	247125.04	40640	3884805.68	638860.85	8	76.01%
2021-05-03	3108	205361.40	34329.27	3228281.16	539656.07	8	74.78%
2021-05-10	1476	249286.68	43263.65	3918786.66	680104.57	8	72.03%
2021-05-17	3404	236687.64	39003.67	3720729.65	613137.66	8	75.85%
2021-05-24	1164	264785.33	42731.73	4162425.38	671742.8	8	77.46%
2021-05-31	951	256694.09	39609.47	4035231.02	622660.8	8	81.01%
2021-06-07	1094	257507.15	42126.32	4048012.4	662225.79	8	76.41%
2021-06-14	1307	261004.58	41483.13	4102991.92	652114.8	8	78.65%
2021-06-21	943	195548.39	37485.7	3074020.73	589275.16	8	65.21%
2021-06-28	1482	168872.63	40392.72	2654677.78	634973.54	8	52.26%
2021-07-05	1153	275127.01	44619.46	4324996.57	701417.9	8	77.08%

Figure 3. Graphical representation of the accounting system

To display all the information in the visualization system, panels that generate special menus were added.

- Account

  Activating the panel displays a list of names of user accounts/user groups, on behalf of which the job is launched, in the form of a drop-down menu. When a specific account is selected, all information on it is displayed in accordance with the specified parameter. The account value can be set as ‘total’ to display general information for the entire computing cluster.
- Time_period

All calculations of the accounting system are carried out taking into account only completed jobs. The longer the time interval for completed jobs accounting is, the more accurate the report data is.

- Clock_type
  It displays the change in the type of parameters, such as Wallclock, CPUclock and their versions translated to HepSpec06 [11]. After changing this parameter, the top-most graph and pie charts change according to the selection.
- Data – time interval
  The date of both the beginning and the end of the data display can be set. Changing the date affects the display of all graphs and tables.

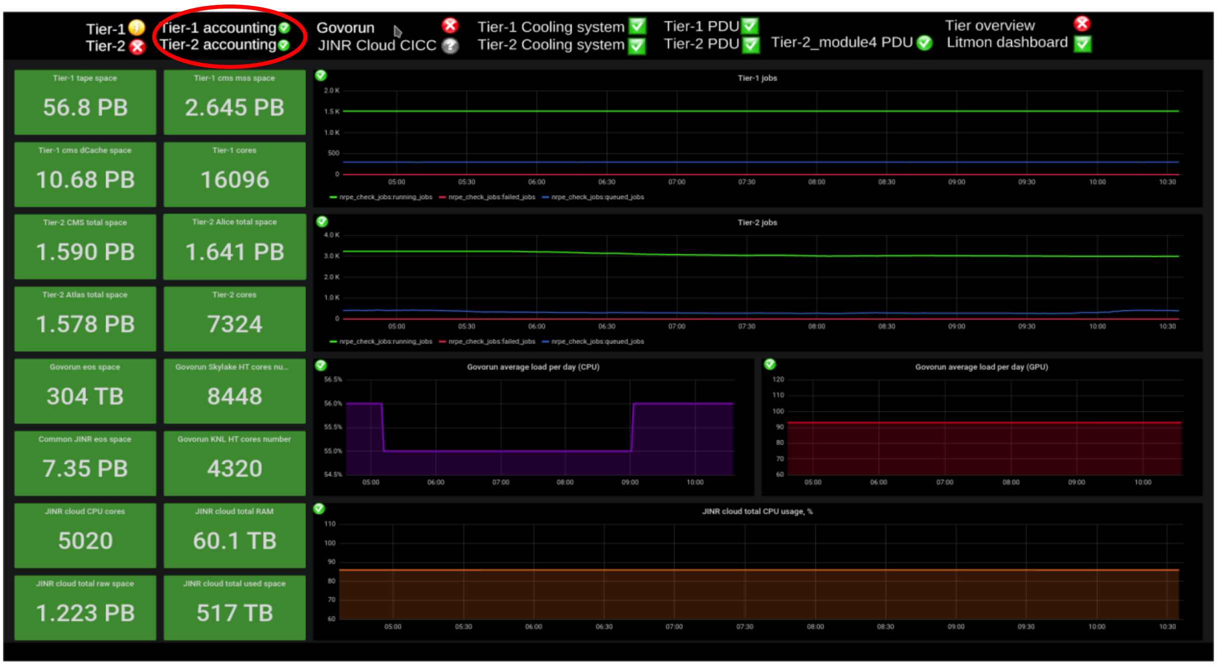
Thus, by changing various parameters, the set of data required for displaying can be configured.

5. Conclusion

In February 2021, various tests and the verification of the functionality of the accounting system, including data verification with the EGI Federation account [12], were performed. The test results showed that the difference between the calculations in the systems was 0.4%, which can be caused by different time frames and calculation algorithms.

In March 2021, the accounting system was integrated into the general monitoring system of the MICC.

Access to the pages of the accounting system for the WLCG (Worldwide LHC Computing Grid) sites [13], Tier1, Tier2 of MLIT JINR, is carried out via the main page of the Litmon monitoring system (Fig. 4). As a result, it became possible to back up the readings of the other accounting systems, as well as to display JINR specific tasks that are performed on the sites and related to tasks within the NICA project. The Tier1, Tier2 accounting system was put into operation. It completely covered the functionality of the original system and significantly expanded it due to the flexible configuration of the visualization system. The data collected by the original accounting system since 2018 was imported into the visualization system.



Tier-1 tape space 56.8 PB

Tier-1 cms mss space 2.645 PB

Tier-1 cms dCache space 10.68 PB

Tier-1 cores 16096

Tier-2 CMS total space 1.590 PB

Tier-2 Alice total space 1.641 PB

Tier-2 Atlas total space 1.578 PB

Tier-2 cores 7324

Govorun eos space 304 TB

Govorun Skylake HT cores nu... 8448

Common JINR eos space 7.35 PB

Govorun KNL HT cores number 4320

JINR cloud CPU cores 5020

JINR cloud total RAM 60.1 TB

JINR cloud total raw space 1.223 PB

JINR cloud total used space 517 TB

Figure 4. Main page of the Litmon monitoring system

References

[1] V. Korenkov, E. Tikhonenko, GRID Concept and Computer Technologies in the LHC Era, Physics of Elementary Particles and Atomic Nuclei (PEPAN), vol.32, p.6, 2001, in Russian.

[2] A. Soldatov, V. Korenkov, V. Ilyin, Russian Segment of the LCG Global Infrastructure, Open Systems, N1, 2003, in Russian.

[3] Torque. Available at: http://dipc.ehu.es/cc/computing_resources/jobs/batch_systems/torque/ (accessed 15.07.2021).

[4] Maui. Available at: http://dipc.ehu.es/cc/computing_resources/jobs/batch_systems/torque/ (accessed 15.07.2021).

[5] SLURM. Available at: https://sLURM.schedmd.com/documentation.html (accessed 15.07.2021).

[6] I. Kashunin, V. Mitsyn, V. Trofimov, A. Dolbilov. Integration of the Cluster Monitoring System Based on Icinga2 at JINR LIT MICC // PEPAN Letters: http://www1.jinr.ru/Pepan_letters/panl_2020_3/14_kashunin.pdf, 2020.

[7] Dolbilov A., et al. Multifunctional Information and Computing Complex of JINR: Status and Perspectives // CEUR Workshop Proc. 2019. V. 2507. P. 16 – 22.

[8] Grafana. Available at: https://grafana.com/ (accessed 15.07.2021).

[9] MariaDb. Available at: https://mariadb.org/ (accessed 15.07.2021).

[10] Crontab. Available at: https://ru.wikipedia.org/wiki/Cron (accessed 15.07.2021).

[11] HepSpec06. Available at: https://www.gridpp.ac.uk/wiki/HEPSPEC06 (accessed 15.07.2021).

[12] EGI. Available at: https://accounting.egi.eu/ (accessed 15.07.2021).

[13] Worldwide LHC Computing Grid. Available at: https://wlcg.web.cern.ch/ (accessed 15.07.2021).