JINR GRID INFRASTRUCTURE STATUS AND PLANS

Andrey Baginyan, Anton Balandin, Andrey Dolbilov, Aleksey Golunov, Natalia Gromova, Ivan Kashunin, Vladimir Korenkov, Valery Mitsyn, Igor Pelevanyuk, Sergey Shmatov, Tatiana Strizh, Vladimir Trofimov, Alexey Vorontsov, Nikolay Voytishin

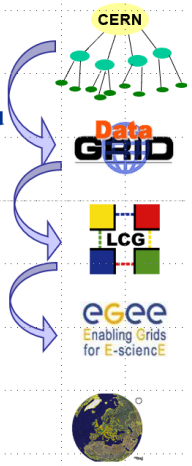
JOINT INSTITUTE FOR NUCLEAR RESEARCH
Meshcheryakov Laboratory of Information Technologies

- How we started
- Main functions
- Infrastructure
- Network and telecommunication channels
- Servers to support the grid-WLCG environment
- Resources
- Monitoring
- How it works
- Where are we going

How we started

Some history

- 1999 – Monarc Project
  - Early discussions on how to organise distributed computing for LHC
- 2001-2003 - EU DataGrid project
  - middleware & testbed for an operational grid
- 2002-2005 – LHC Computing Grid – LCG
  - deploying the results of DataGrid to provide a production facility for LHC experiments
- 2004-2006 – EU EGEE project phase 1
  - starts from the LCG grid
  - shared production infrastructure
  - expanding to other communities and sciences
- 2006-2008 – EU EGEE-II
  - Building on phase 1
  - Expanding applications and communities …



LHC Computing Grid Project (LCG)

The protocol between CERN, Russia and JINR on a participation in LCG Project has been approved in 2003.

The tasks of the Russian institutes in the LCG:

- ✓ **LCG software testing;**
- ✓ **evaluation of new Grid technologies (e.g. Globus toolkit 3) in a context of using in the LCG;**
- ✓ **event generators repository, data base of physical events: support and development;**

Структура комплекса

130 CPU

18TB RAID-5

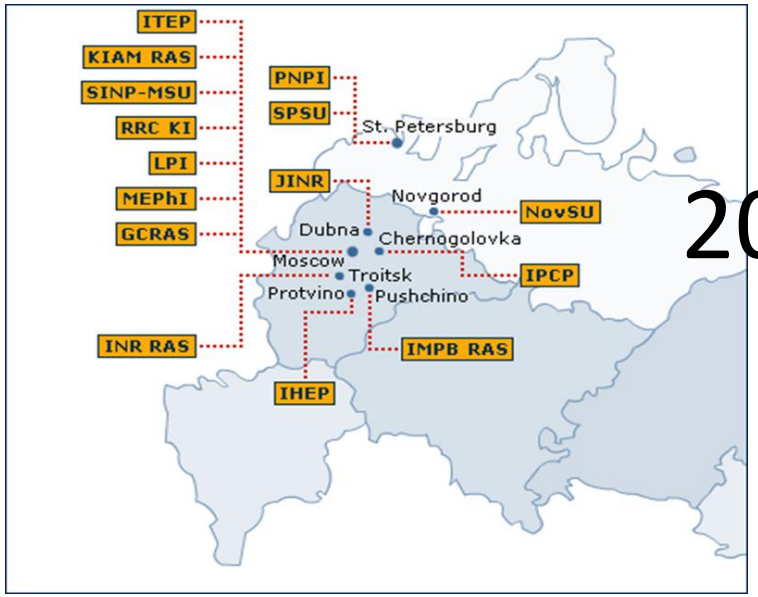
ATL~ 5 (15) TB

6 – Interactive
18 – Common PC-farm
30 – LHC
14 – MYRINET (Parallel)
20 – LCG
20 – File servers
8 – LCG-user interface

The Russian consortium RDIG

The Russian consortium RDIG (Russian Data Intensive Grid, was set up in September 2003 as a national federation in the EGEE project.

IHEP - Institute of High Energy Physics (Protvino),
IMPB RAS - Institute of Mathematical Problems in Biology (Pushchino),
ITEP - Institute of Theoretical and Experimental Physics
JINR - Joint Institute for Nuclear Research (Dubna),
KIAM RAS - Keldysh Institute of Applied Mathematics
PNPI - Petersburg Nuclear Physics Institute (Gatchina),
RRC KI - Russian Research Center "Kurchatov Institute"
SINP-MSU - Skobeltsyn Institute of Nuclear Physics,MSU,

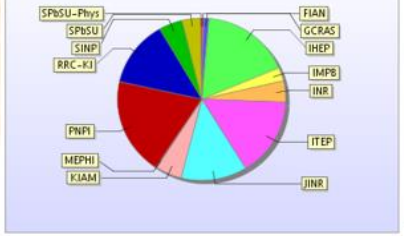


2006

RDIG accounting

Total number of records in Database of RDIG accounting System – 1 384 800

Number of Job Records per Site



Number of Job Records per VO

How we started

Joint NRC "Kurchatov Institute" – JINR Tier1 Computing Centre

- Proposal to create the WLCG Tier1 center in Russia: March 2011, accepted in October 2012
- *The Federal Target Programme Project:* «Creation of the automated system of data processing for experiments at the LHC of Tier1 level and maintenance of Grid services for a distributed analysis of these data»

Duration: 2011 – 2013

Russia Tier1 full scope start-up in WLCG in 2014

NRC "Kurchatov Institute" supports ATLAS, ALICE and LHCb, JINR supports CMS (Compact Muon Solenoid)

Systematic increase of computing capacity and data storage is needed in accordance with the experiment requirements


ORGANISATION EUROPEENNE POUR LA RECHERCHE NUCLEAIRE
EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

Subject: Acceptance of the proposal to build Tier 1 centres in Russia

Dear Directors,

As you know, the proposals from the National Research Centre – "Kurchatov Institute" and the Joint Institute for Nuclear Research, Dubna, to build Tier 1 centres for LHC data analysis were discussed in the recent WLCG Overview Board held on September 28. I am very happy to report that the proposals were well received by the members of the board, and that the decision was made to accept the Russian sites as a new "Associate Tier 1". This decision will be noted in the formal minutes of the meeting.

The next step is now to proceed to signing the WLCG Memorandum of Understanding. The WLCG project office will assist in drafting this MoU, which should be signed by the relevant funding agencies for the two Russian institutes, or their designated agents.

I am at your disposal for any assistance or to provide further details of the process.

Yours Sincerely,

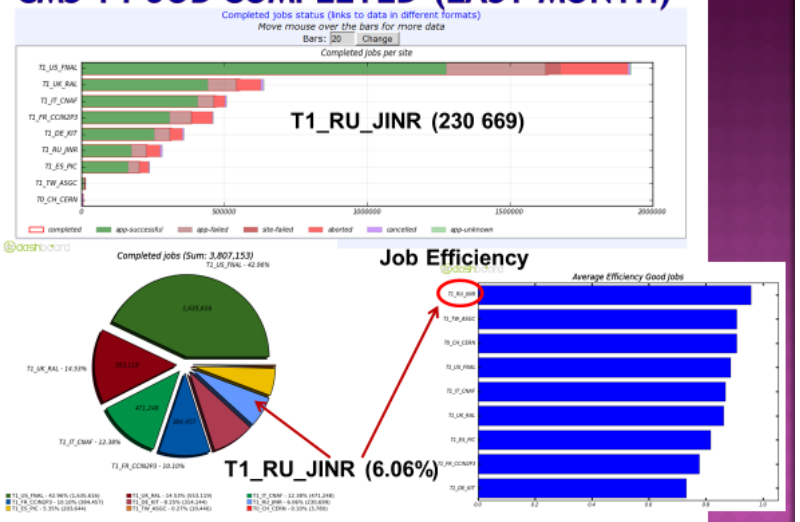
Dr. Ian Bird
LHC Computing Grid Project Leader
IT Department
CERN

Cc: Prof. Sergio Bertolucci, Dr. Viacheslav Ilyin

JINR: current state

- Our Tier-1 currently supports CMS as the tape-less Tier-1 since October 2013;
- Our resources were fully validated;
- Our Tier-1 participates in CMS Multicore job scheduling project:
  - Able to run multicore glideins (12-cores) through two queues as of April 29th
  - Large scale tests at Tier1s
- Our Tier-1 was tested for high memory (6GB) jobs

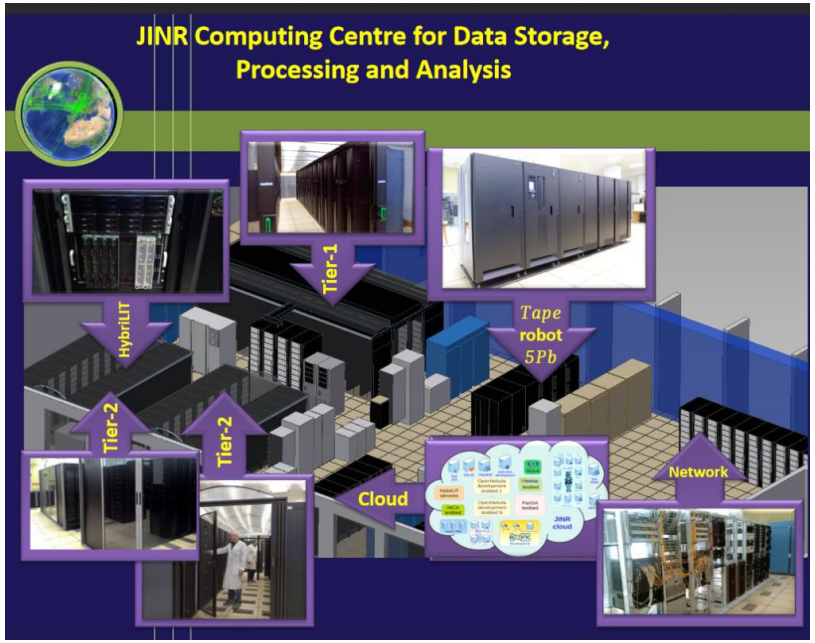
CMS T1 JOB COMPLETED (LAST MONTH)



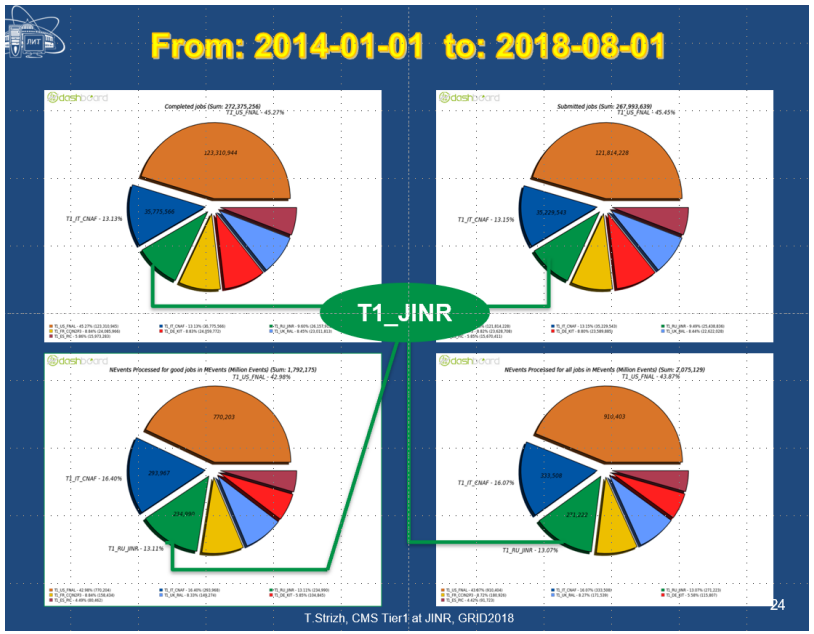
CMS Tier-1

March 2015 – CMS Tier1 Inauguration
LHCOPN – 10Gbps, 2400 cores (~ 30 kHS06),
5 PB tapes (IBM TS3500), 2.4 PB disk
Close-coupled, chilled water cooling InRow
Hot and cold air containment system
MGE Galaxy 7000 – 2x300 kW energy efficient solutions
3Ph power protection with high adaptability

JINR Computing Centre for Data Storage, Processing and Analysis



From: 2014-01-01 to: 2018-08-01



Main functions

Tier1

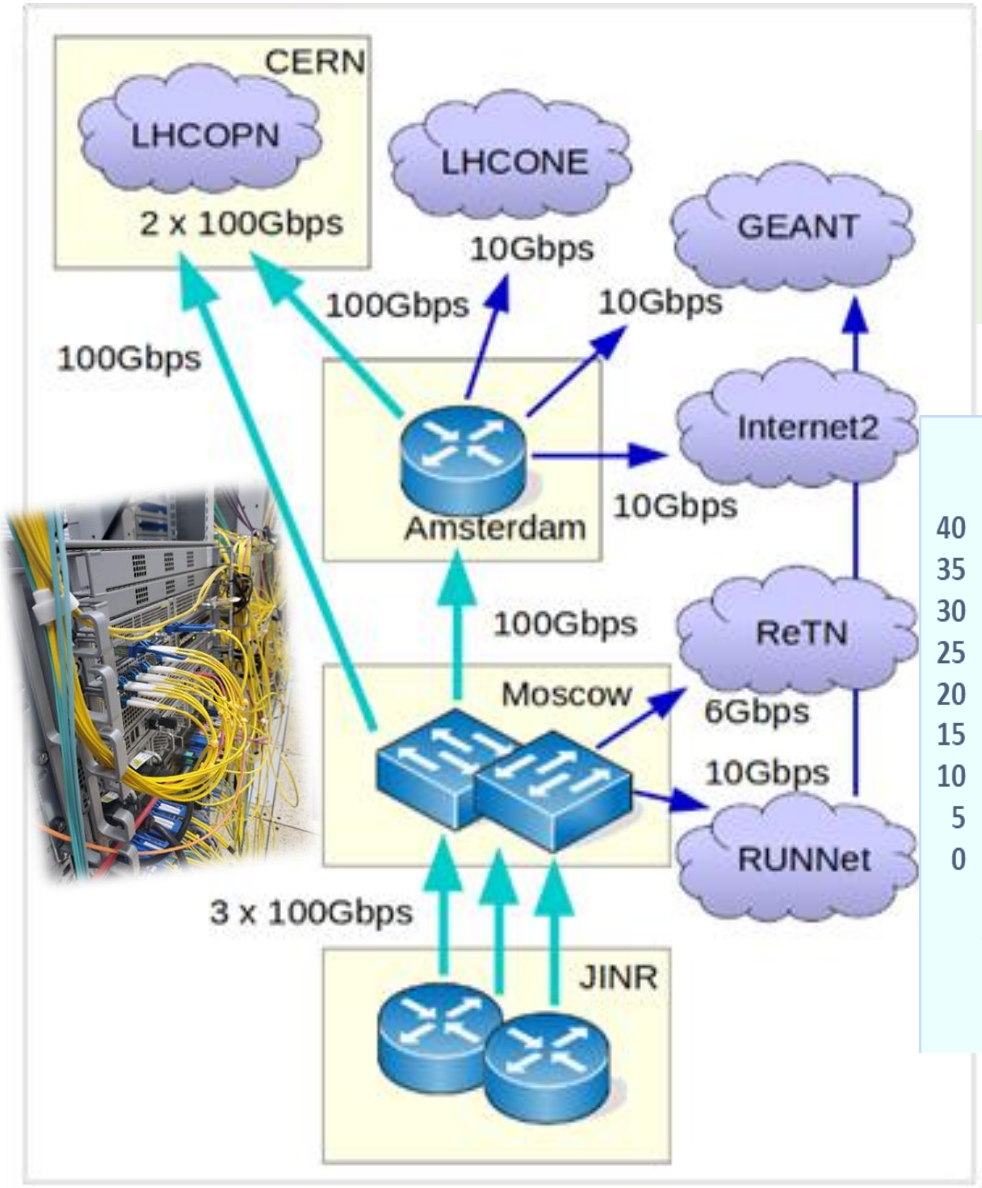
Receiving of experimental data from a Tier0 site in the volume determined by the WLCG agreement
Archiving and custodial storage of part of experimental RAW data
Consecutive and continuous data processing
Additional processing (skimming) of RAW, RECO (RECOnstructed) and AOD (Analysis Object Data) data
Data reprocessing with the use of new software or new calibration and alignment constants of parts of the CMS setup
Making available AOD data-sets
Serving RECO and AOD datasets to other Tier1/Tier2/Tier3 sites for their duplicated storage (replication) and physical analysis
Running production reprocessing with the use of new software and new calibration and alignment constants of parts of the CMS setup, protected storage of the simulated events
Production of simulated data and data analysis recorded by the CMS experiment
Production of simulated data and data analysis for the NICA experiments (MPD, BM@N, SPD)

Tier2

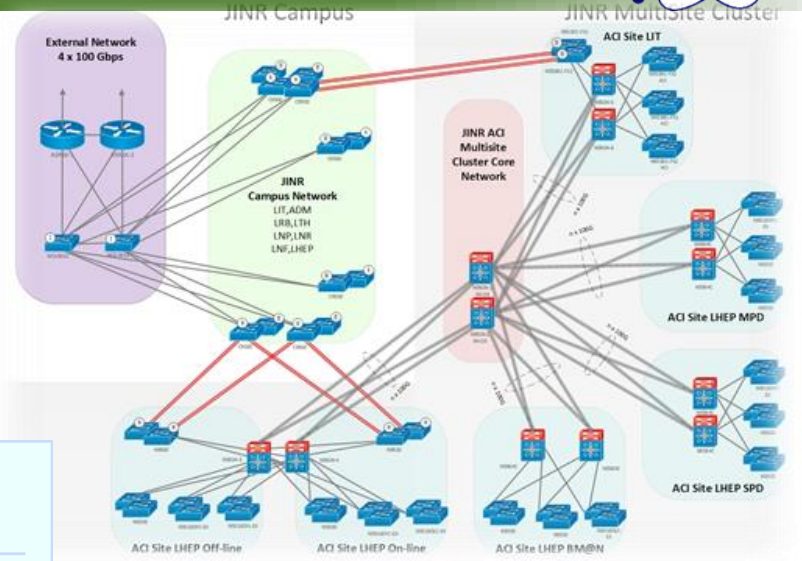
Provides services for local communities
Production of simulated data and data analysis for all virtual organisations registered on RDIG and of the grid using experiments with JINR participation
Production of simulated data and data analysis for the NICA experiments (MPD, BM@N, SPD)

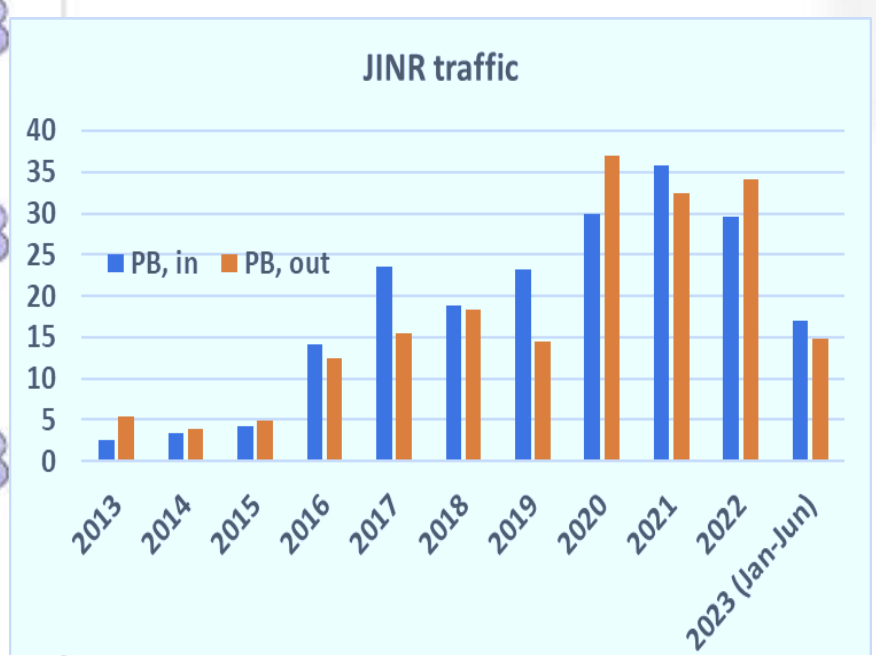
Infrastructure

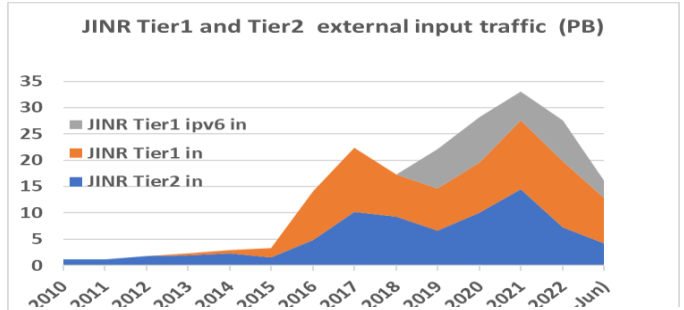
Network and telecommunication channels

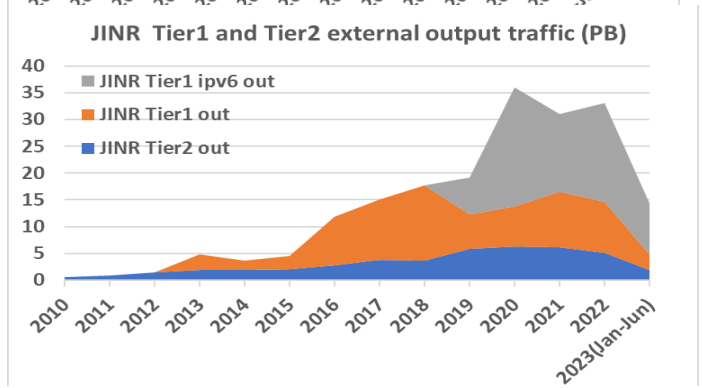


Wide Area Network 3x100 Gbps
Cluster Backbone 4x100 Gbps
Campus Backbone 2x100 Gbps









Infrastructure and services

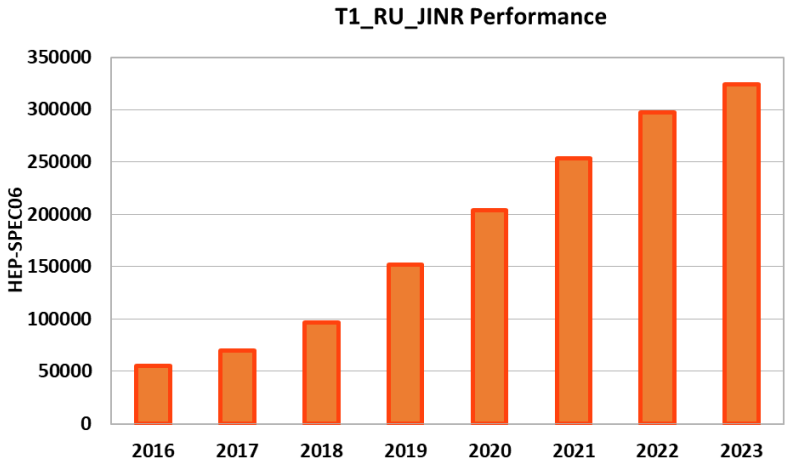
The infrastructure and services of the Tier1 (JINR-T1) and Tier2 (JINR-LCG2) sites ensure the operation of:

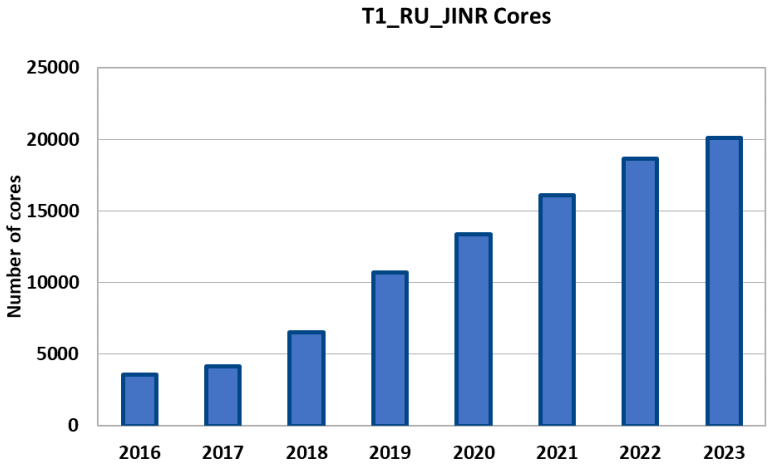
- computing service,
- data storage service,
- service of access to user home directories,
- service of access to user software versions,
- GRID support service,
- data transfer service,
- distributed computing control system,
- information service (monitoring, information sites).

Common services for most components:

- kerberos, VOMS - authentication and authorization of access;
- AFS - user home directories, installation and access to user and group software, available world-wide like local FS with POSIX access;
- CVMFS (CernVM-File System) servers (stratum0/1) - installation and storage of collaboration and groups software with many software versions, available world-wide like local FS with POSIX access
- CVMFS clients and caching - access to collaborations and groupssoftware (read-only), used to access local CVMFS and global repositories from all over the world;
- EOS - storage and access to experimental data over large volume, available on interactive and calculating machines like local FS with POSIX access, world-wide access via xroot and http protocols;
- GIT - service for building and testing collaboration software and groups for subsequent installation in CVMFS.

Infrastructure and services (Tier1 2023)





Computing farm (CE)

323820.54 HEP-SPEC06, 20096 cores
Average HEP-SPEC06 per Core = 16.11
468 hosts
CMS (16-cores pilot):
Max: 20096 cores
NICA (from DIRAC)
Max: 4000 cores

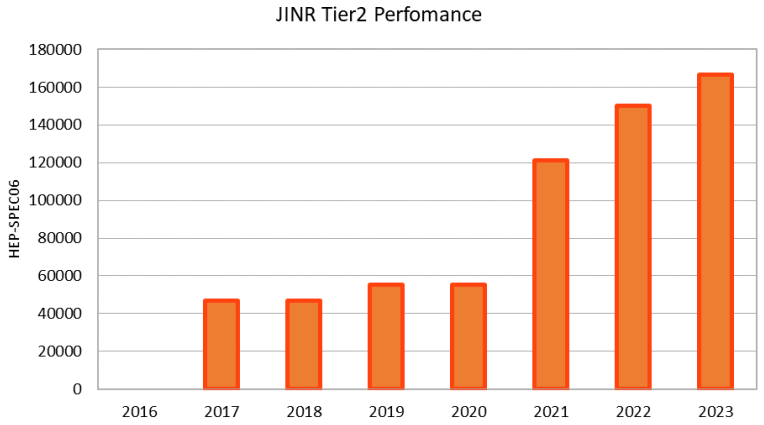
Storage Systems (SE)

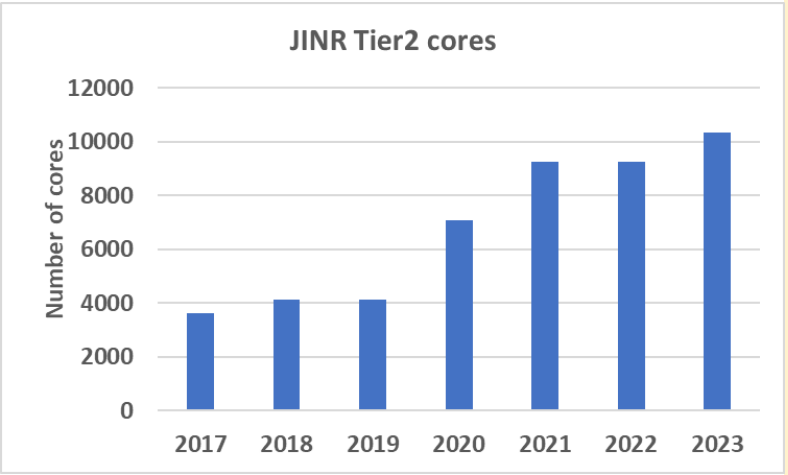
dCache: SE disks: 11763.44 PB
CMS @ dcache mss Total: 2642.24 TB
Tapes@Enstore: 35562,00 TB
Tape robots: 51.5PB, IBM TS3500(11.5PB) + IBM T4500(40PB)
EOS: 21829.01 TB
CVMFS
2 squid servers cache CVMFS

Software :

OC: Scientific Linux release 7.9.
EOS 5.1.23
dCache 8.2,
Enstore 6.3.
Slurm 20.11.
grid UMD4 + EPEL (current versions)
ARC-CE
FairSoft
FairRoot
MPDroot

Infrastructure and services (Tier2 2023)





Computing Resources (CE):

Interactive cluster: lxpub [01-05] .jinr.ru
User interface lxui [01-04] .jinr.ru (gateway for external connections)
Computing farm.

485 hosts
10356 cores
166788.4 HEP-SPEC06
16.11 HEP-SPEC06 average per core

Storage Systems (SE)

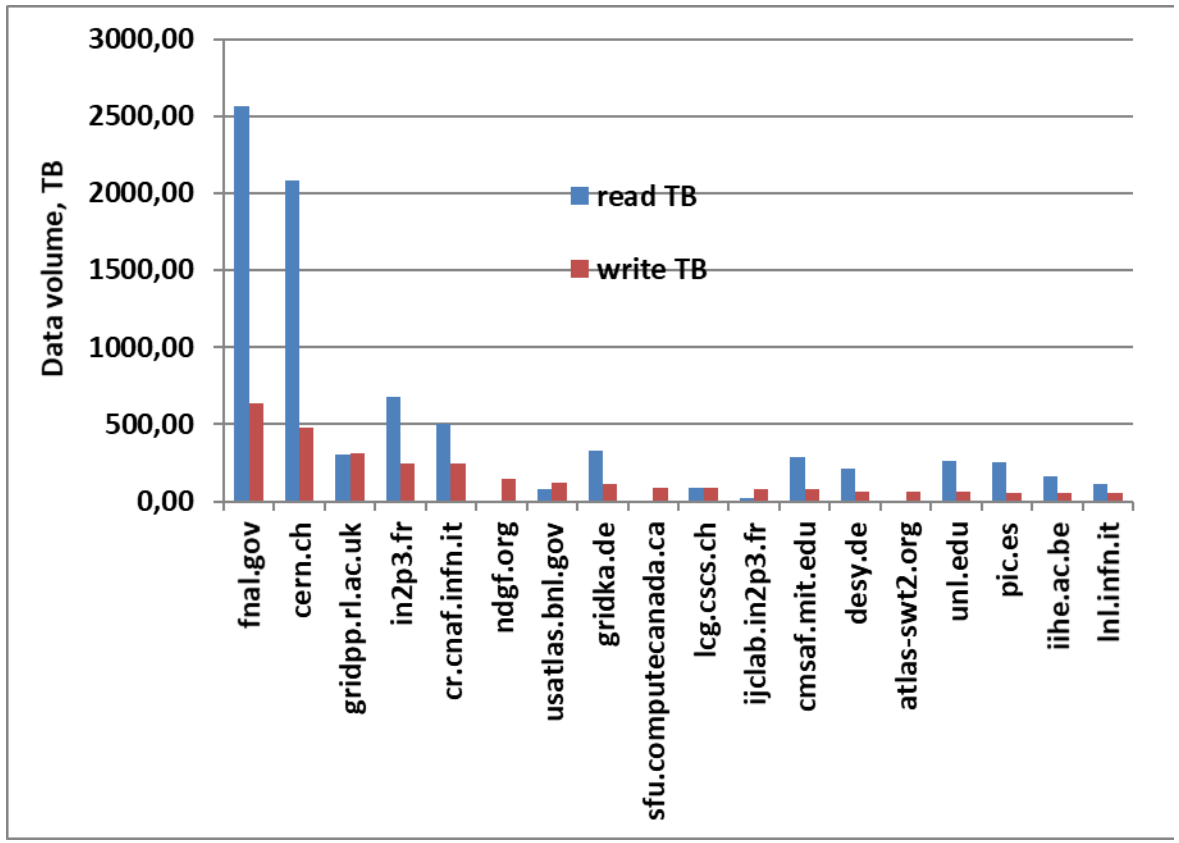
EOS=21829.04 TB
ALICE @ EOS Total: 1653.24 TB
AFS: ~12.5TB (user home directories, workspaces)
CVMFS: 3 machines : 1 stratum0, 2 stratum1, 2 squid servers cache CVMFS (VOs: NICA (MPD, B@MN, SPD), dstau, jjnano, juno, baikalgvd).
dCache : SE disks = 3753,69 TB
for CMS: 1903.2695 TB
for ATLAS: 1850.4248 TB
Local & EGI @ dcache2 Total: 256.91 TB

Software :

OC: Scientific Linux release 7.9.
EOS 5.1.23
dCache 8.2
BATCH: Slurm 20.11 with adaptation to kerberos and AFS
grid UMD4 + EPEL (current versions)
ARC-CE
FairSoft
FairRoot
MPDroot

Long-term storage system

Storage system exchange rate, statistics since the beginning of 2023



Total exchanges - 230 millions files

Presented are external customers who have accessed more than 1 million times.

Storage and data.

TS3200 is only used for tests.

TS3500 on standby, currently connected to CTA

TS4500 runs on CMS, half capacity reserved for NICA

Tape storage volumes.

T1mss tape. 20 PB allocated, 11 PB occupied, 7 PB available (remaini space on tapes only).

TS3500 12 PB free

TS4500 total 40 PB, of which 20 CMS, 20 reserve

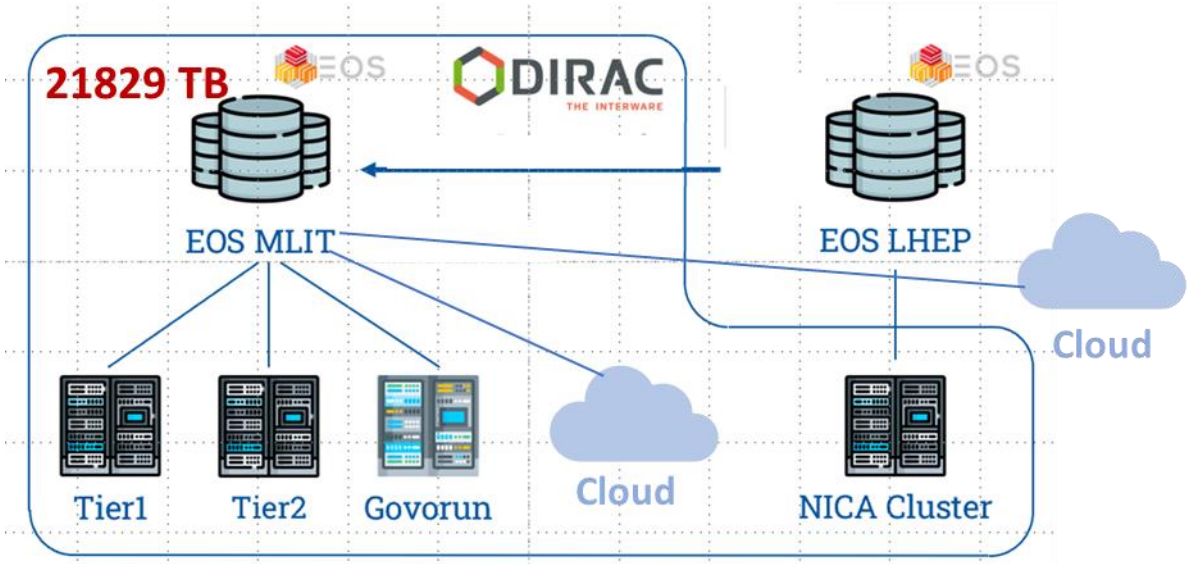
Detailed volumes in PB since the beginning of 2023

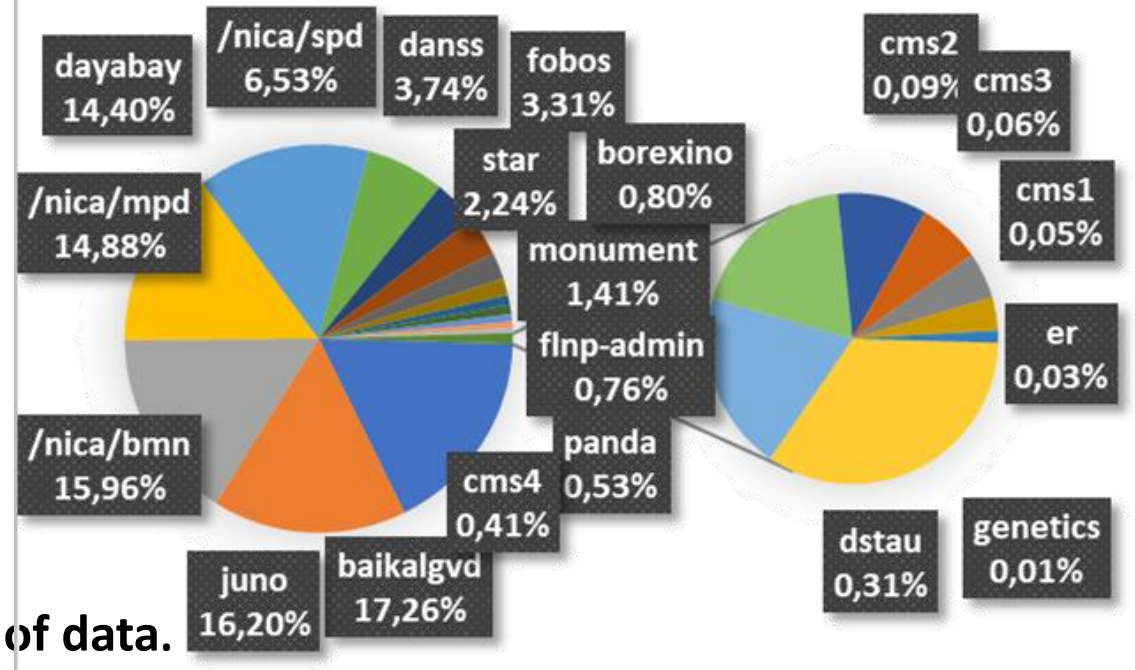
	T2	T1	T1mss
write	3.0	4.3	2.0
read	8.2	19.8	1.1

Development for next one-two years

- TS3500 11,5 PB, 12 LTO6 drives is being used as a testing ground for the EOSCTA installation. Will be a repository for non-WLCG experiments
- TS4500 40 PB 12 drive 3592-60F Jaguar will be divided into 2 logical libraries
  - 20 PB 6 drives managed by Enstore for CMS
  - 20 PB 6 drives under EOSCTA for NICA

Middle-term storage system

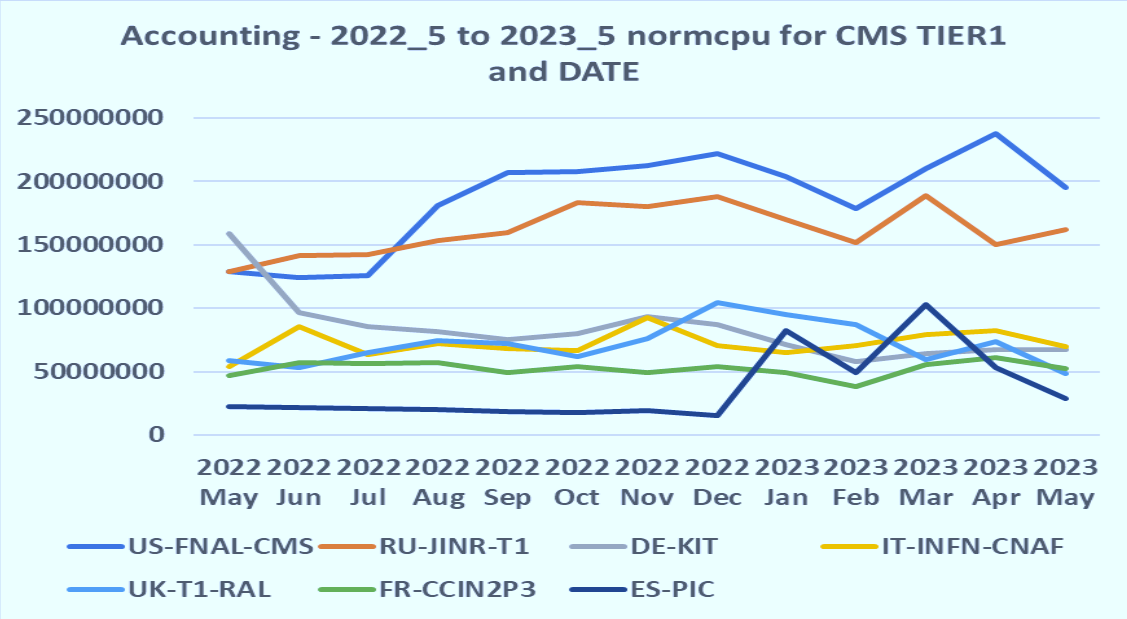




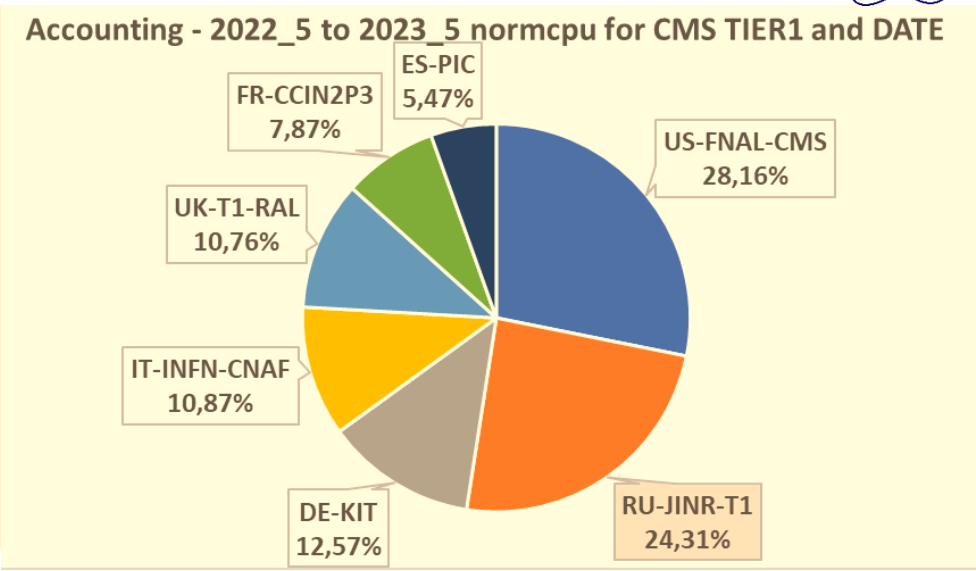
EOS is definitely a storage system for extra large amounts of data.

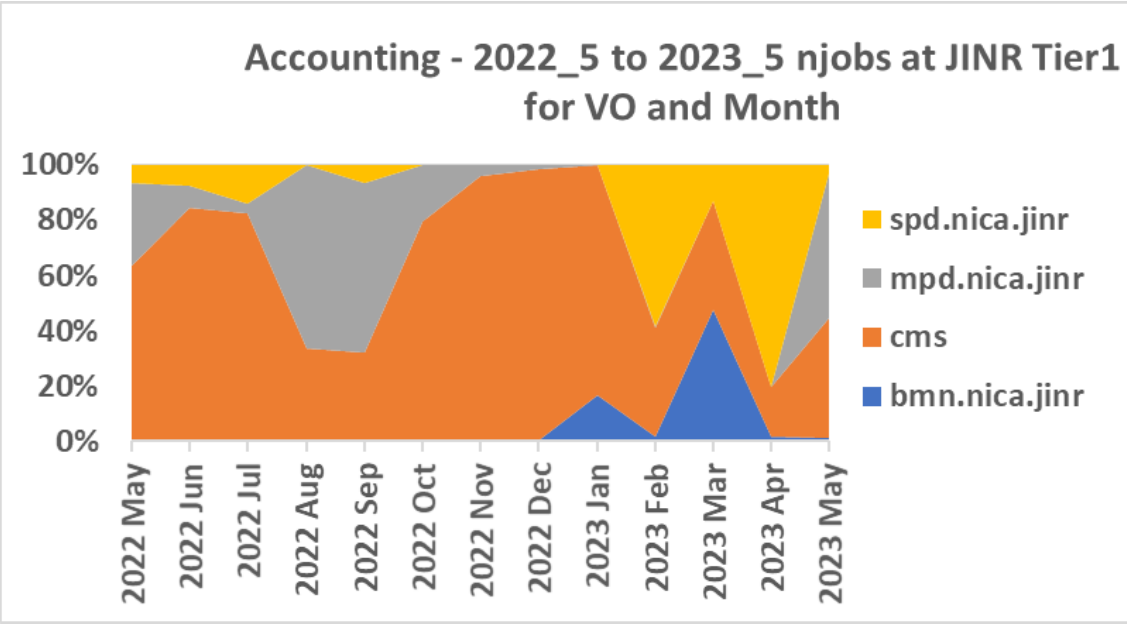
- Optimal in terms of cost / volume of storage,
- Convenient for users almost like a local file system.
- Supports many access protocols: **POSIX** when mounted on user machine; **xroot** and **http** for fast remote access.
- High reliability of data storage due to duplication on different servers, storage on different servers in the format vertical RAID with checksums.
- High data access speed due to parallel copying from many servers.
- Protecting Data with an Extended Access Mod List Set groups and individual users.

JINR Tier1 (last year)

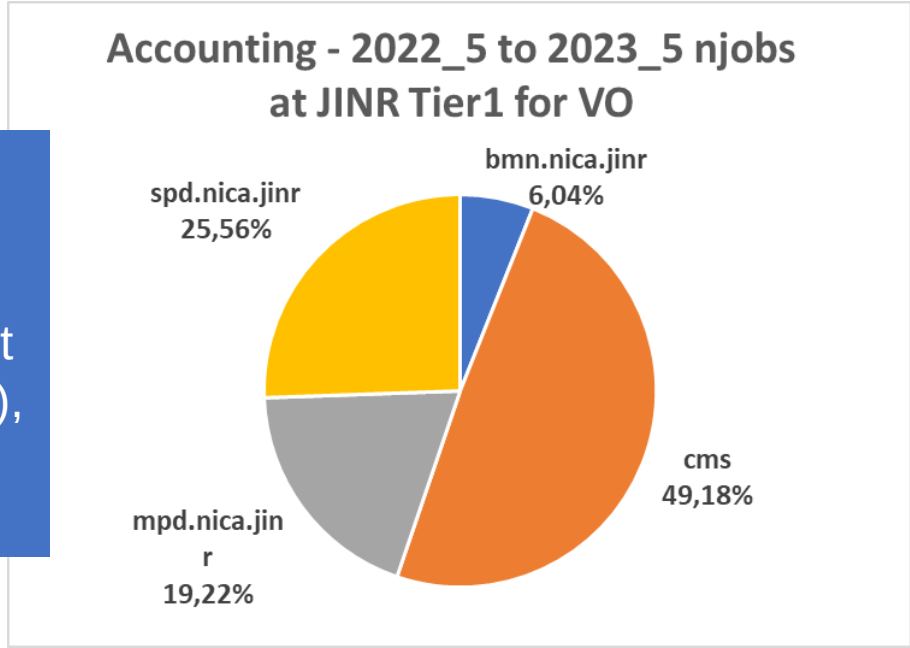


JINR Tier1 is regularly ranked on top among world Tier1 sites that process data from the CMS experiment at the LHC.

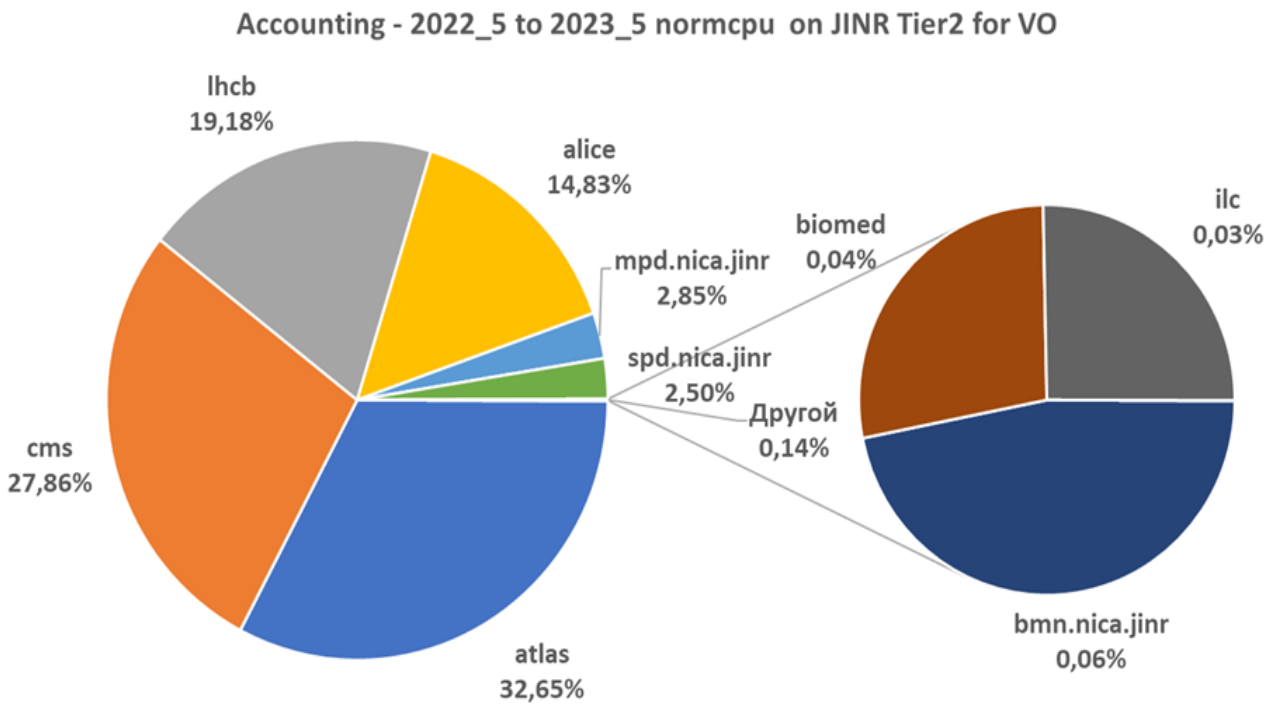


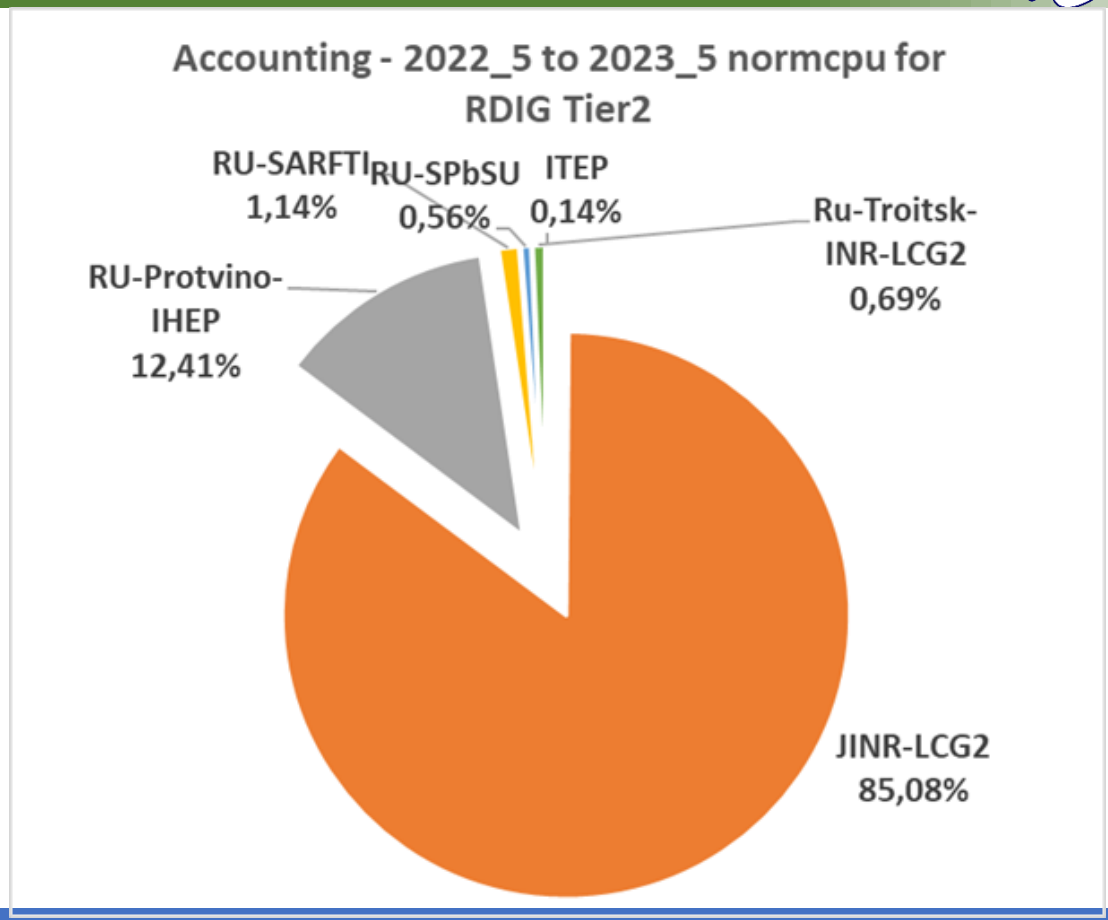


Since 2019, the JINR Tier1 center has demonstrated stable operation not only for CMS (LHC), but also for the NICA experiments.



JINR Tier2 (last year)

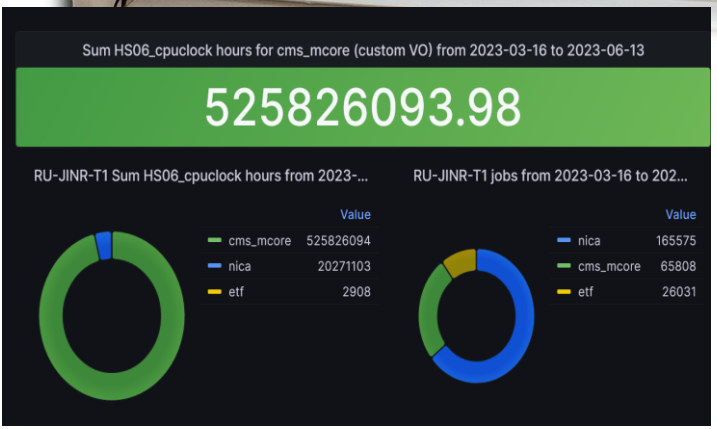


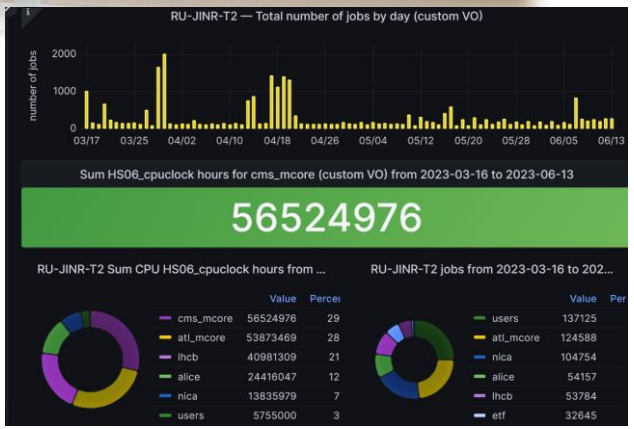


Tier2 at JINR provides computing power and data storage and access systems for the majority of JINR users and user groups, as well as for users of virtual organizations (VOs) of the grid environment (LHC, NICA, etc.).

JINR Tier2 is the most productive in the Russian Data Intensive Grid (RDIG) Federation.
More than 80% of the total CPU time in the RDIG is used for computing on our site.

MICC Monitoring @Accounting





- **3 monitoring servers**
- **About 1800 nodes**
- **About 16000 service checks**

The successful functioning of the computing complex is ensured by the system that monitors all MICC components. We must

- expand the monitoring system by integrating local monitoring systems for power supply systems into it (diesel generators, power distribution units, transformers and uninterruptible power supplies);
- organize the monitoring of the cooling system (cooling towers, pumps, hot and cold water circuits, heat exchangers, chillers);
- create an engineering infrastructure control center (special information panels for visualizing all statuses of the MICC engineering infrastructure in a single access point);
- account each user job on each MICC component.

It is required to develop intelligent systems that will enable to detect anomalies in time series on the basis of training samples, which will result in the need to create a special analytical system within the monitoring system to automate the process.

How it works

Detailed Monthly Site Reliability

Site	Jul-2022	Aug-2022	Sep-2022	Oct-2022	Nov-2022	Dec-2022
T0_CH_CERN	97%	98%	97%	97%	99%	99%
T1_DE_KIT	99%	100%	95%	100%	100%	96%
T1_ES_PIC	100%	99%	98%	99%	99%	99%
T1_FR_CCIN2P3	99%	99%	96%	98%	97%	94%
T1_IT_CNAF	100%	99%	90%	100%	100%	99%
T1_RU_JINR	98%	98%	98%	99%	98%	99%
T1_UK_RAL	98%	94%	95%	86%	99%	99%
T1_US_FNAL	99%	96%	96%	96%	96%	96%
Target	97%	97%	97%	97%	97%	97%

Availability of WLCG Tier-0 + Tier-1 Sites

CMS

Target Availability for each site is 97.0%. Target for 8 best sites is 98.0%

Availability Algorithm: (CREAM-CE + ARC-CE + HTCONDOR-CE) * all SRM

Jul-2022 - Dec-2022

T0_CH_CERN Avail: 98% Unkn: 0% T1_DE_KIT Avail: 98% Unkn: 1% T1_ES_PIC Avail: 98% Unkn: 0% T1_FR_CCIN2P3 Avail: 97% Unkn: 2%

T1_IT_CNAF Avail: 98% Unkn: 0% T1_RU_JINR Avail: 98% Unkn: 0% T1_UK_RAL Avail: 95% Unkn: 0% T1_US_FNAL Avail: 96% Unkn: 2%

Availability of WLCG Tier-0 + Tier-1 Sites

CMS

Target Availability for each site is 97.0%. Target for 8 best sites is 98.0%

Availability Algorithm: (CREAM-CE + ARC-CE + HTCONDOR-CE) * all SRM

Dec-2022 - May-2023

T0_CH_CERN Avail: 99% Unkn: 0% T1_DE_KIT Avail: 98% Unkn: 1% T1_ES_PIC Avail: 97% Unkn: 0% T1_FR_CCIN2P3 Avail: 95% Unkn: 3%

T1_IT_CNAF Avail: 99% Unkn: 0% T1_RU_JINR Avail: 98% Unkn: 0% T1_UK_RAL Avail: 95% Unkn: 0% T1_US_FNAL Avail: 97% Unkn: 1%

Availability of WLCG Tier-0 + Tier-1 Sites

CMS

May 2023

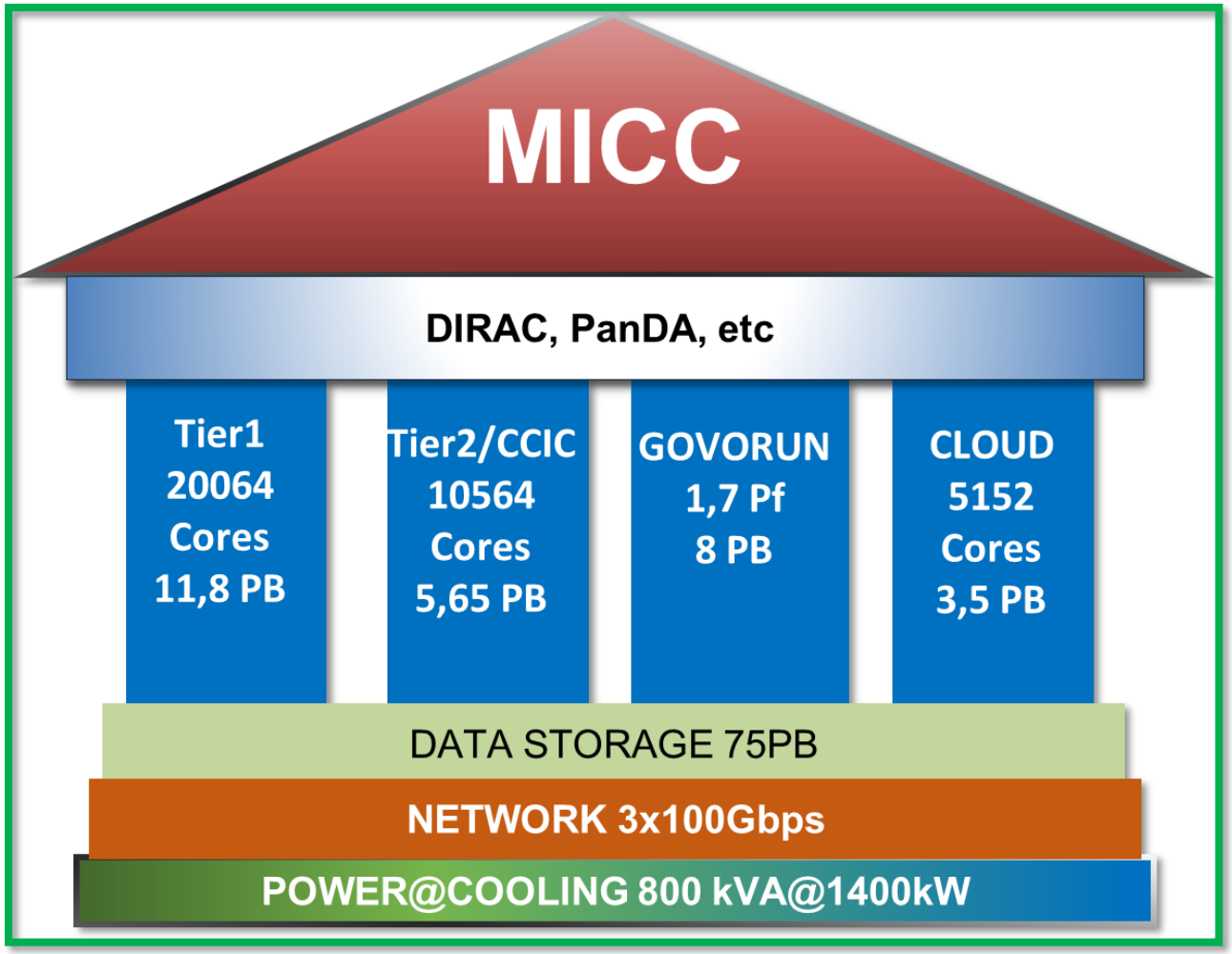
Target Availability for each site is 97.0%. Target for 8 best sites is 98.0%

Availability Algorithm: (CREAM-CE + ARC-CE + HTCONDOR-CE) * all SRM

T0_CH_CERN Avail: 99% Unkn: 0% T1_DE_KIT Avail: 98% Unkn: 0% T1_ES_PIC Avail: 99% Unkn: 0% T1_FR_CCIN2P3 Avail: 99% Unkn: 0%

T1_IT_CNAF Avail: 99% Unkn: 0% T1_RU_JINR Avail: 98% Unkn: 0% T1_UK_RAL Avail: 91% Unkn: 1% T1_US_FNAL Avail: 98% Unkn: 0%

Multifunctional Information and Computing Complex (MICC)



4 advanced software and hardware components

- Tier1 grid site
- Tier2 grid site
- hyperconverged “Govorun” supercomputer
- cloud infrastructure

Distributed multi-layer data storage system

- Disks
- Robotized tape library

Engineering infrastructure

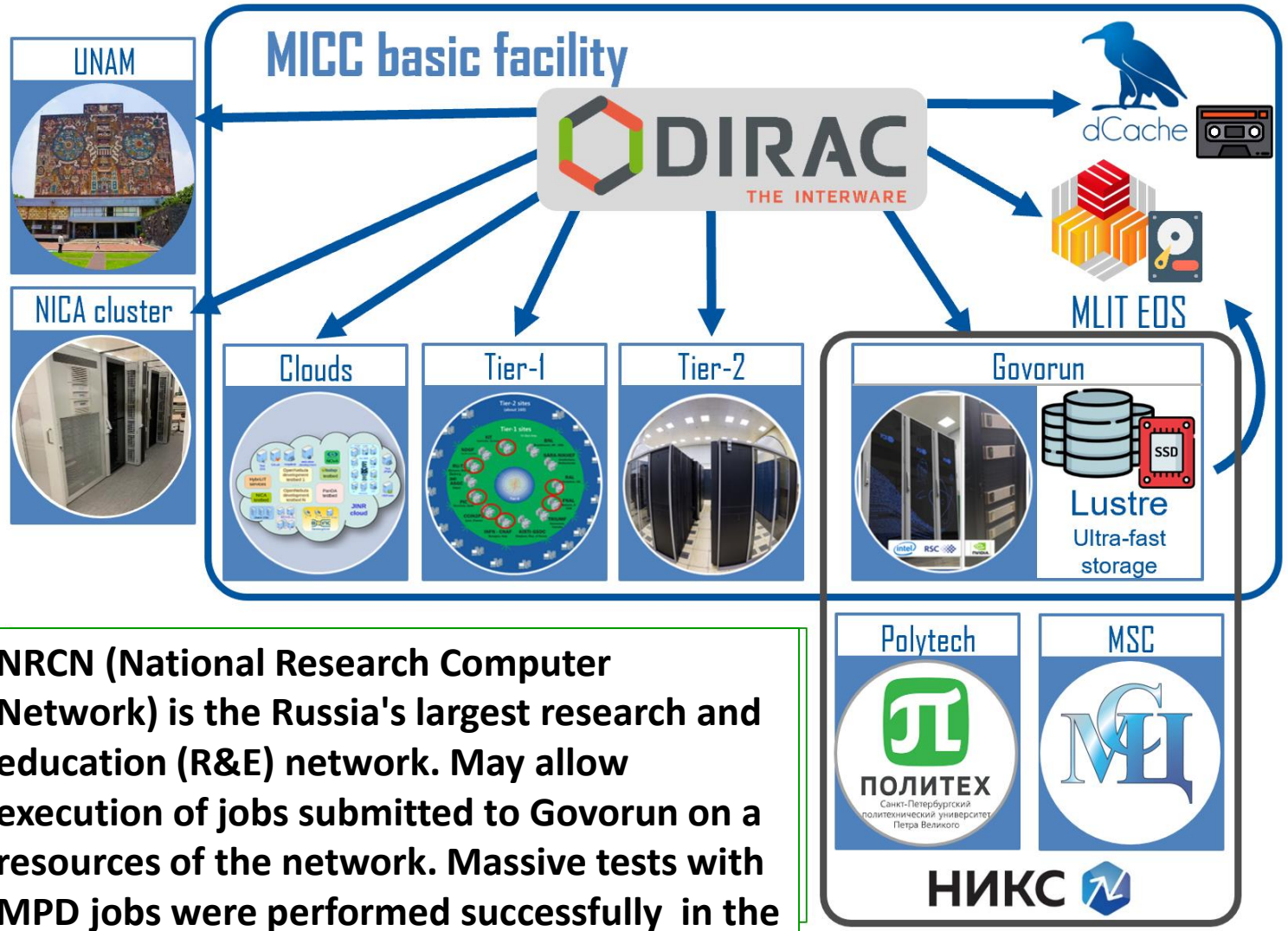
- Power
- Cooling

Network

- Wide Area Network
- Local Area Network

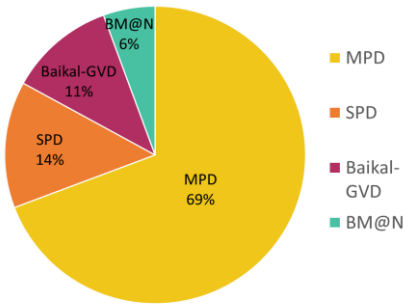
The main objective of the project is to ensure multifunctionality, scalability, high performance, reliability and availability in 24x7x365 mode for different user groups that carry out scientific studies within the JINR Topical Plan.

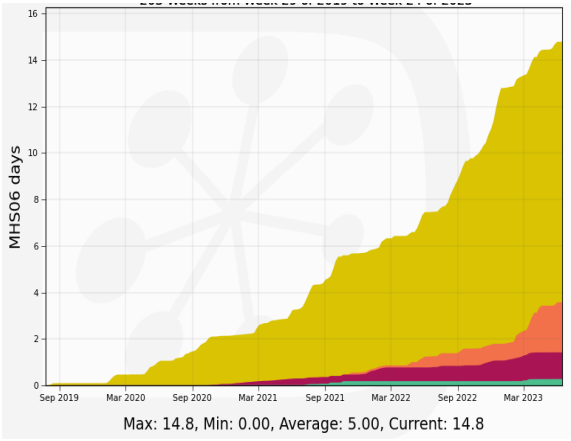
DIRAC @ JINR



NRCN (National Research Computer Network) is the Russia's largest research and education (R&E) network. May allow execution of jobs submitted to Govorun on a resources of the network. Massive tests with MPD jobs were performed successfully in the beginning of 2022

Usage of the DIRAC platform by experiments in 2019-2022

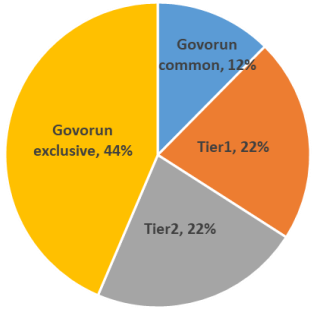


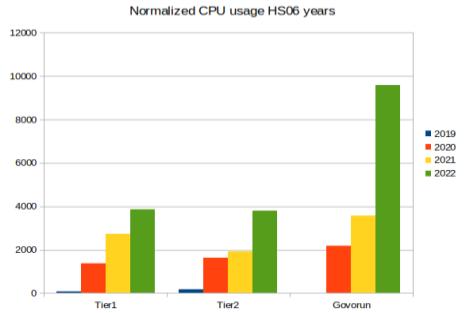


The major user of the distributed platform is the MPD experiment.

MPD Monte-Carlo production

Computing work of MPD using DIRAC 2022





Development of the JINR Grid sites @ 7-year plan

The Seven-Year Plan provides for the creation of a long-term data storage center on the MICC resources at MLIT.

1. 1. The process of modeling, processing and analyzing experimental data obtained from the BM@N, MPD and SPD detectors will be implemented in a distributed computing environment based on the MICC and the computing centers of VBLHEP and collaboration member countries.
2. Data center dedicated to Monte Carlo production, data storage and processing for the JUNO experiment. This data center is expected to be one of three European data centers managing JUNO data. The requested numbers are needed for the processing and storage of the JUNO data and were approved by the parties within "Memorandum of Understanding for Collaboration in the Deployment and
3. Exploitation of the JUNO Computing Grid" signed between IHEP and JINR on September 1, 2022.
4. To continue as Tier1 and Tier2 for LHC (HL-LHC)

The information and computer unit of the NICA complex embraces:

1. **online NICA cluster**,
2. **offline NICA cluster at VBLHEP**,
3. **all MICC components** (Tier0, Tier1, Tier2, "Govorun" supercomputer, cloud computing),
4. multi-layer **data storage system**,
5. **distributed computing network**.

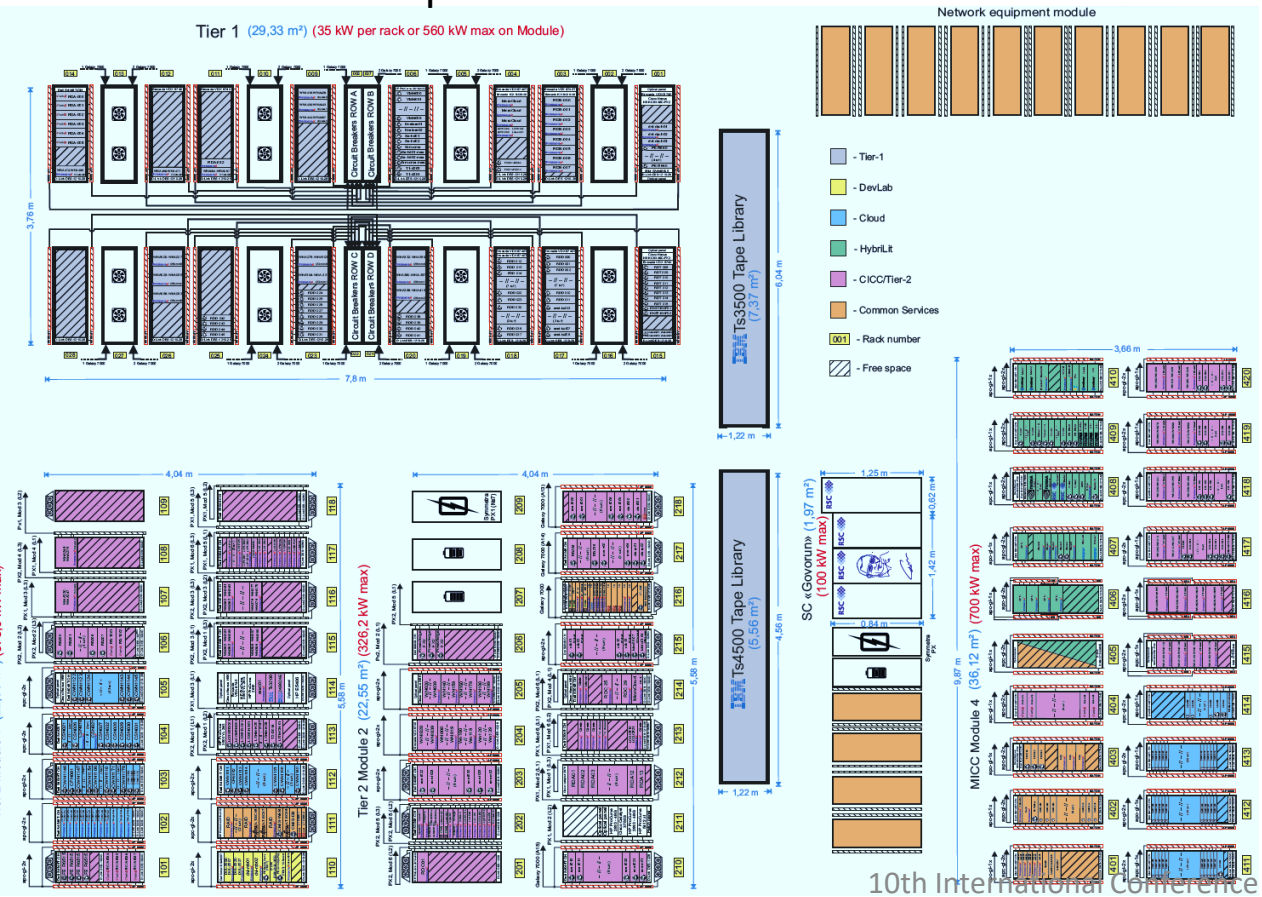
NICA Tier 0,1,2	2024	2025	2026	2027	2028	2029	2030
CPU (PFlops)	2.2	2.6	8.6	8.6	15.6	15.6	15.6
DISK (PB)	17	24	47	75	96	119	142
TAPE (PB)	45	88	170	226	352	444	536
NETWORK (Gbps)	400	400	800	800	800	1000	1000

It should be underlined that the resources given in the table can be approximately satisfied by 20-25% of the budget allocated for the MICC.

MICC Server Halls

Present (1000 kW)

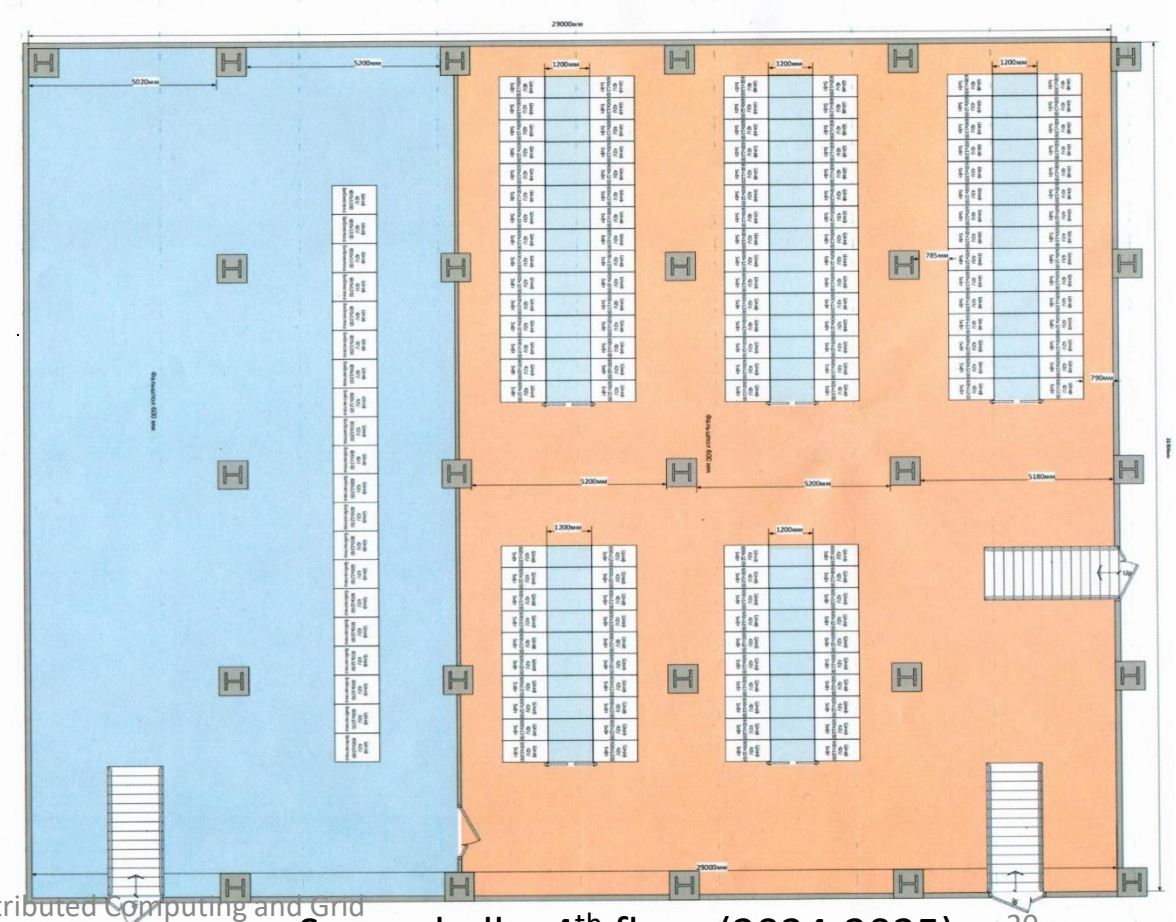
69 racks for servers
4 racks for the “Govorun” SC
10 racks for network equipment
4 racks for administrative services
2 robotic tape libraries



Server hall – 2nd floor (2023)

Planning for the future – new server hall for the MICC (600 kW)

containment area for robotic tape libraries
130 racks for servers



Server hall – 4th floor (2024-2025)

Joint Institute for Nuclear Research
Meshcheryakov Laboratory of Information Technologies

GRID2023
3-7 July 2023

10th International Conference
"Distributed Computing and Grid Technologies in Science and Education"

MANY THANKS TO YOU ALL !!!

Estimation of the Grid Resources of the MICC Components

	2024	2025	2026	2027	2028	2029	2030
Tier1 grid site							
Tier1 performance HEPS06	350000	400000	500000	550000	650000	750000	850000
Total number of CPU cores	22000	23000	30000	32000	38000	45000	50000
Total data storage capacity, TB	14500	16000	18000	20000	22000	23000	25000
Tier2 grid site							
Tier2 performance HEPS06l	187000	204000	221000	238000	306000	408000	510000
Total number of CPU cores	11000	12000	13000	14000	18000	24000	30000
Data storage system							
Total volume of the Data Lake on EOS, PB	27	35	38	53	58	71	83
Total robotic tape storage capacity, PB	70	90	130	130	170	170	190

Resources for JUNO

JINR CONTRIBUTION: COMPUTING

Data centers

- Dubna is expected to be one of the data storage and data processing centers
- Data rate: 3 PB/year
- Memorandum of Understanding for computing is signed by JINR
- IHEP is able to facilitate construction of high speed channel on Chinese side

Resources requirements, from MoU

JINR	Planned to be pledged*				
	2023	2024	2025	2026	2027
Tape (PB)	5	5	5	5	5
Disk (PB)	5	5	5	5	5
CPU	36	36	30	20	10

*numbers are *not* cumulative

The JUNO Project
Dmitry Naumov
Dzhelepov Laboratory for Nuclear Problems
PAC for Particle Physics, June 21, 2023