НИЯУ МИФИ

JINR MLIT

Распределенные вычисления в ФВЭ

Кореньков Владимир Васильевич

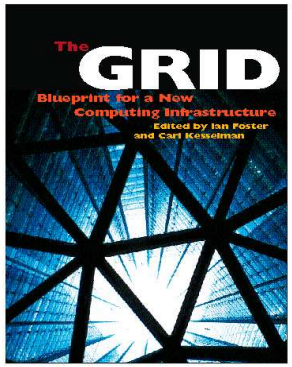
Научный руководитель
Лаборатории информационных технологий
имени М.Г. Мещерякова ОИЯИ

Рабочее совещание МИФИ-ОИЯИ:
Компьютинг для мегапроекта NICA
12 декабря 2023 года

Grids, clouds, supercomputers, Big data

Grids

- Collaborative environment
- Distributed resources

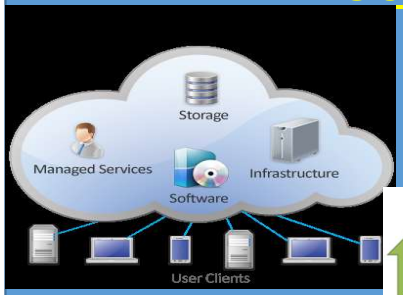


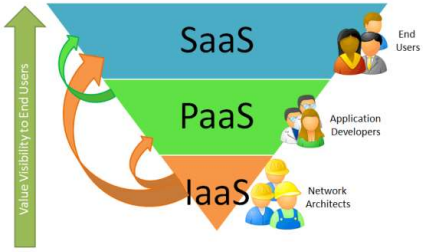
Supercomputers

Titan System (Cray XK7)			
Peak Performance	27.1 PF 18,688 compute nodes	24.5 PF GPU	2.6 PF CPU
System memory	710 TB total memory		
Interconnect	Gemini High Speed Interconnect	3D Torus	
Storage	Lustre Filesystem	32 PB	
Archive	High-Performance Storage System (HPSS)	29 PB	
I/O Nodes	512 Service and I/O nodes		

12 OLCF | 20

Clouds

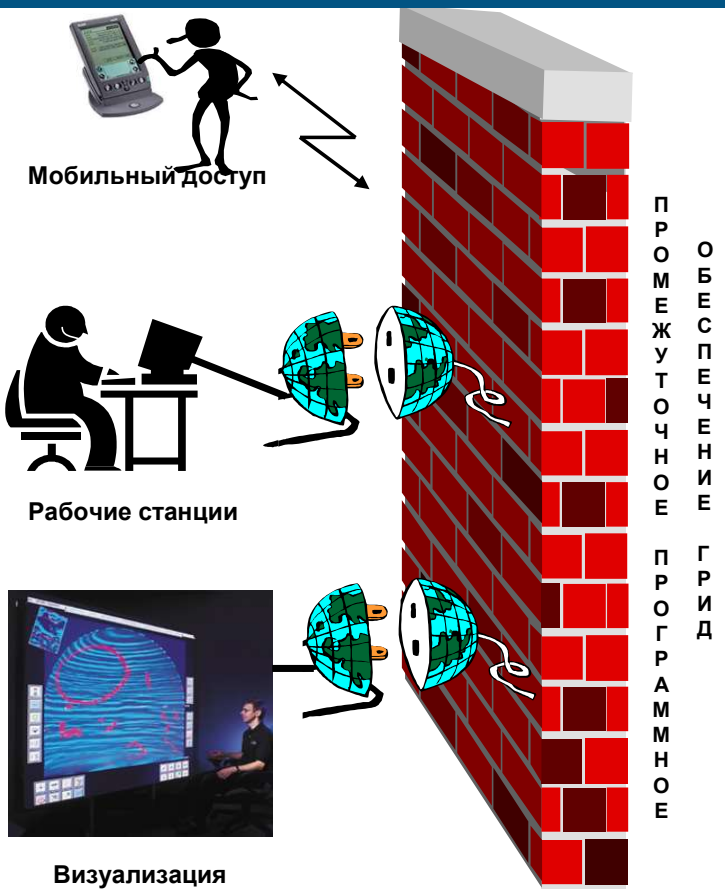


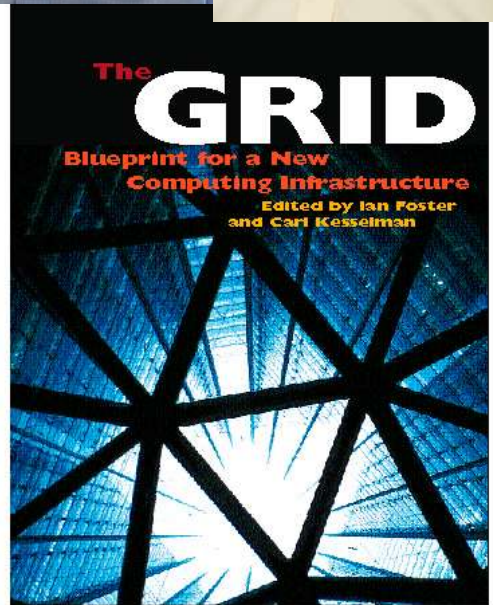


Big Data

- *Volume*
- *Velocity*
- *Variety*

Грид - это средство для совместного использования вычислительных мощностей и хранилищ данных посредством интернета





Грид технологии – путь к успеху

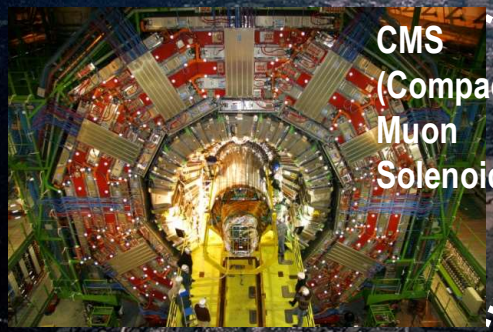
На торжестве по поводу получения Нобелевской премии за открытие бозона Хиггса директор ЦЕРНа Рольф Хойер прямо назвал **грид-технологии одним из трех столпов успеха** (наряду с ускорителем LHC и физическими установками).

Без организации грид-инфраструктуры на LHC было бы невозможно обрабатывать и хранить колоссальный объем данных, поступающих с коллайдера, а значит, совершать научные открытия.

Сегодня уже ни один крупный проект не осуществим без использования распределенной инфраструктуры для обработки данных.

Large Hadron Collider

The Large Hadron Collider (LHC), one of the largest and truly global scientific projects ever, is the most exciting turning point in particle physics.



CMS (Compact Muon Solenoid)

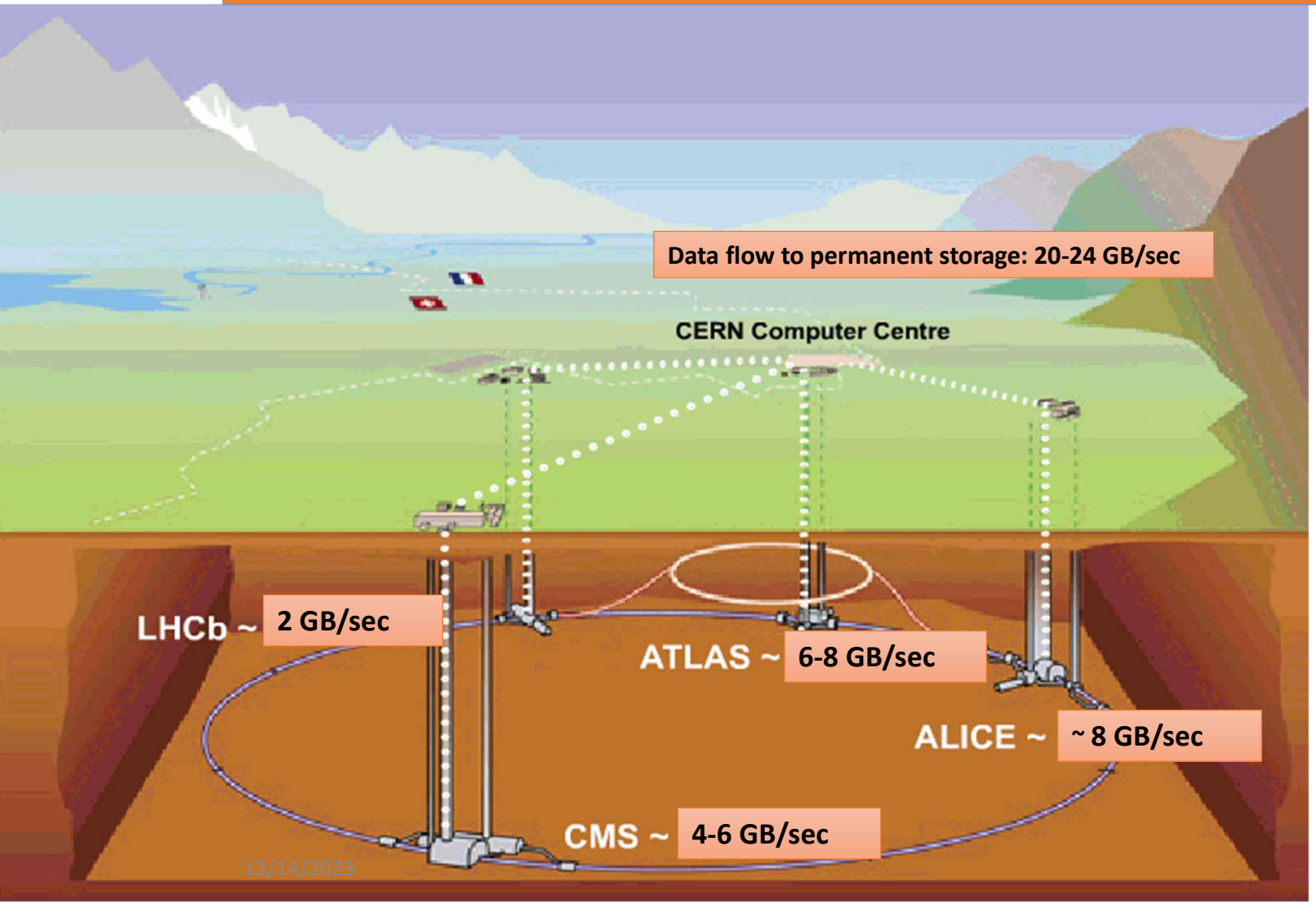
LHCb (Large Hadron Collider beauty experiment)

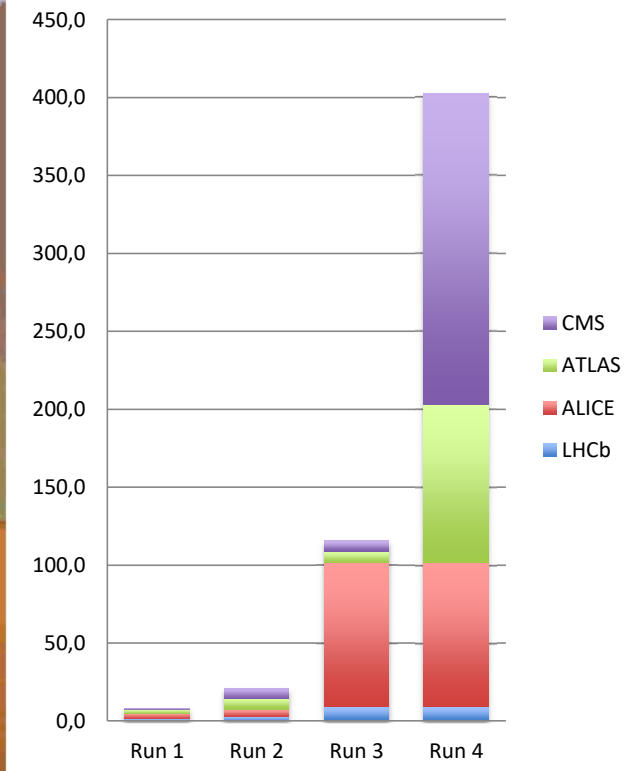
Data flow to permanent storage: 4-6 GB/sec

ALICE (A Large Ion Collider Experiment)

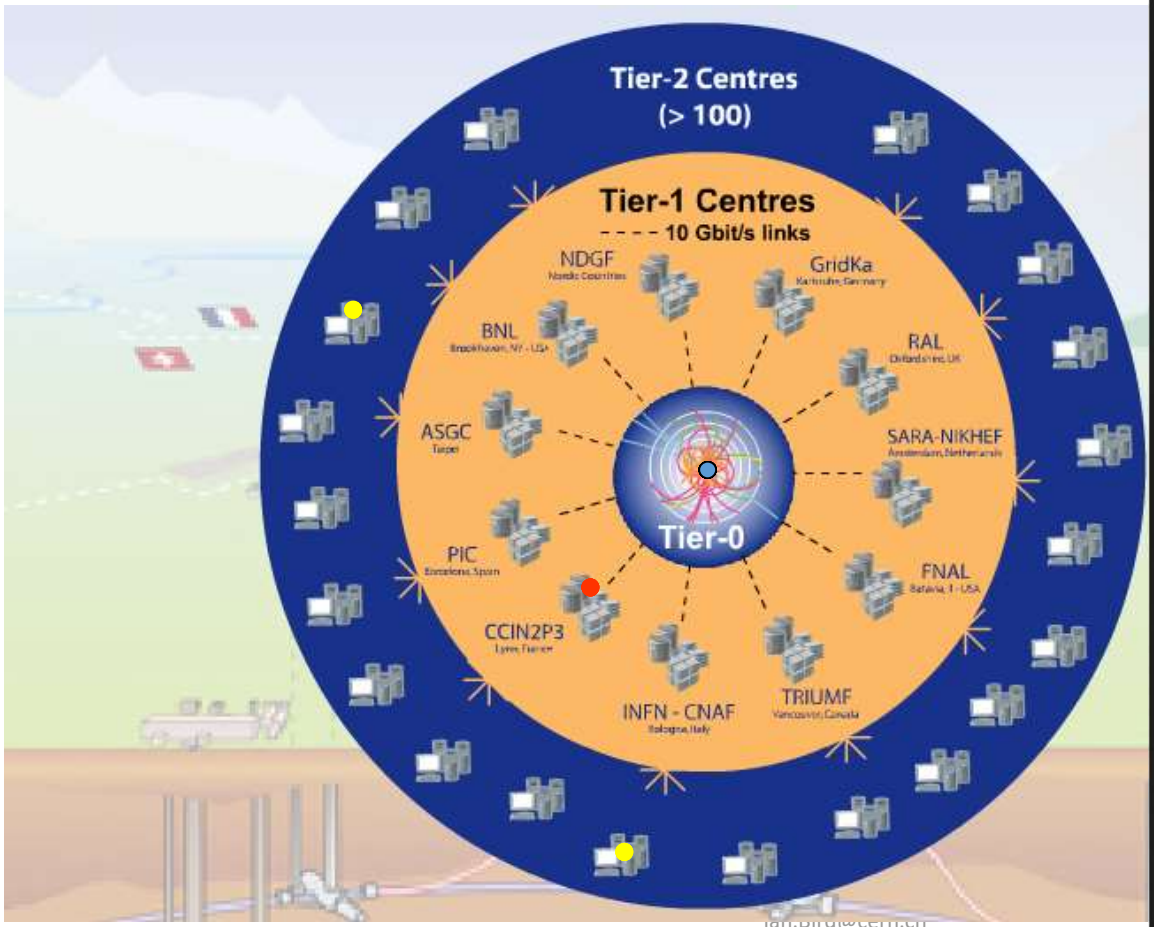
ATLAS (A Toroidal LHC ApparatuS)

Data Collection and Archiving at CERN





Tier Structure of GRID Distributed Computing: Tier-0/Tier-1/Tier-2



Tier-0 (CERN):

- accepts data from the CMS Online Data Acquisition and Trigger System
- archives RAW data
- the first pass of reconstruction and performs Prompt Calibration
- data distribution to Tier-1

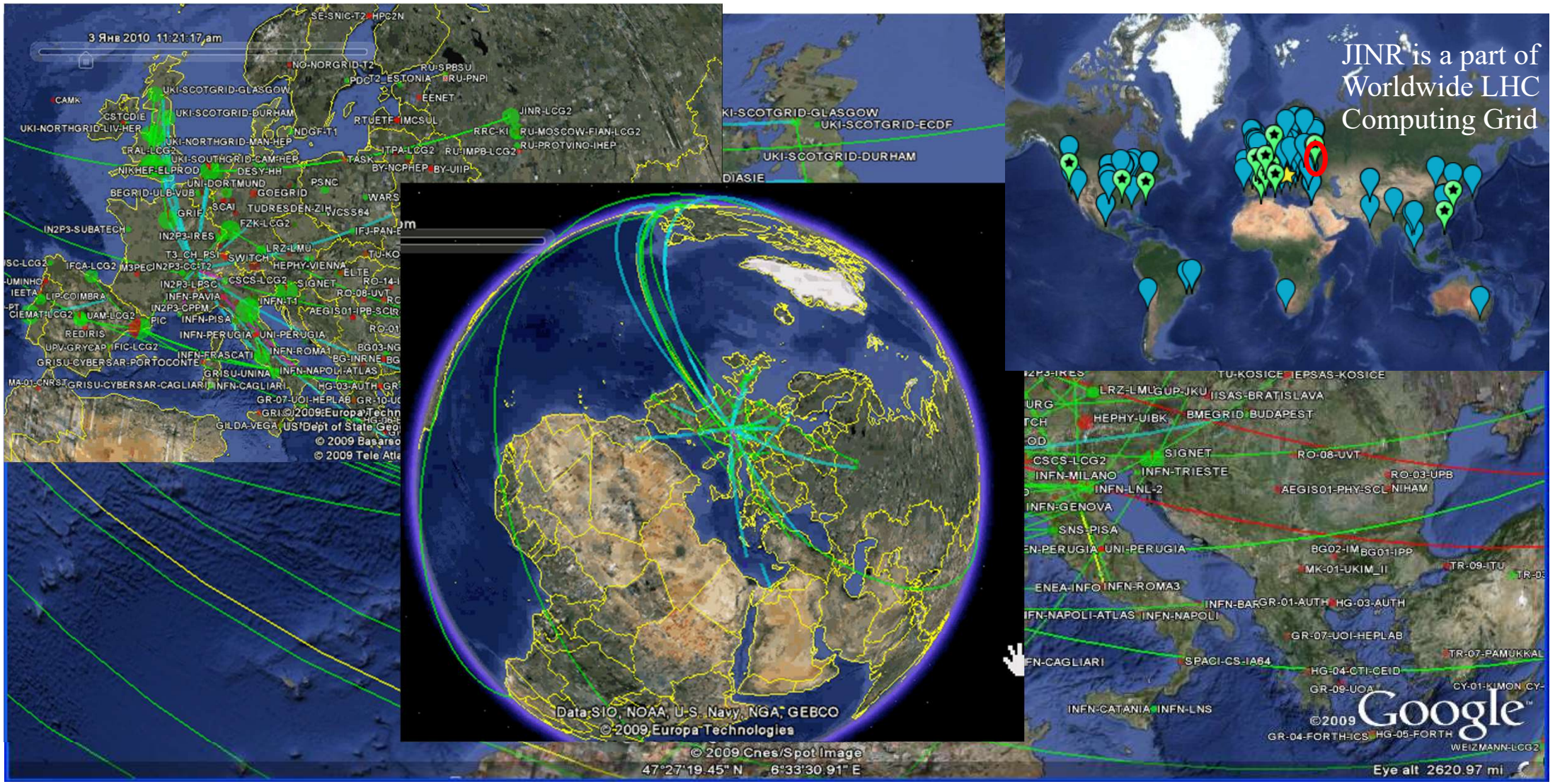
Tier-1 (11 centers):

- receives a data from the Tier-0
- data processing (re-reconstruction, skimming , calibration etc)
- distributes data and MC to the other Tier-1 and Tier-2
- secure storage and redistribution for data and MC

Tier-2 (>200 centers):

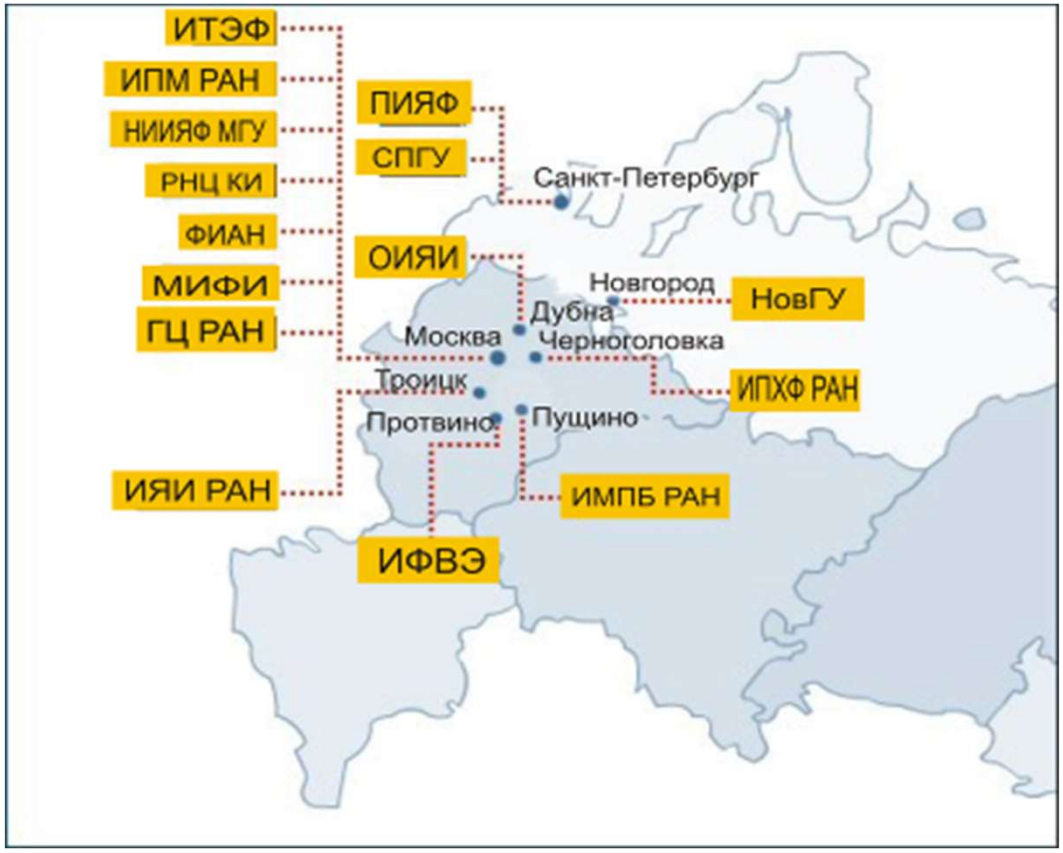
- simulation
- user physics analysis

The Worldwide LHC Computing Grid (WLCG)



Russian Data Intensive Grid infrastructure (RDIG)

The Russian consortium RDIG (Russian Data Intensive Grid), was set up in September 2003 as a national federation in the EGEE project.
A protocol between CERN, Russia and JINR on participation in the LCG project was signed in 2003.
MoU on participation in the WLCG project was signed in 2007.



RDIG Resource Centres:

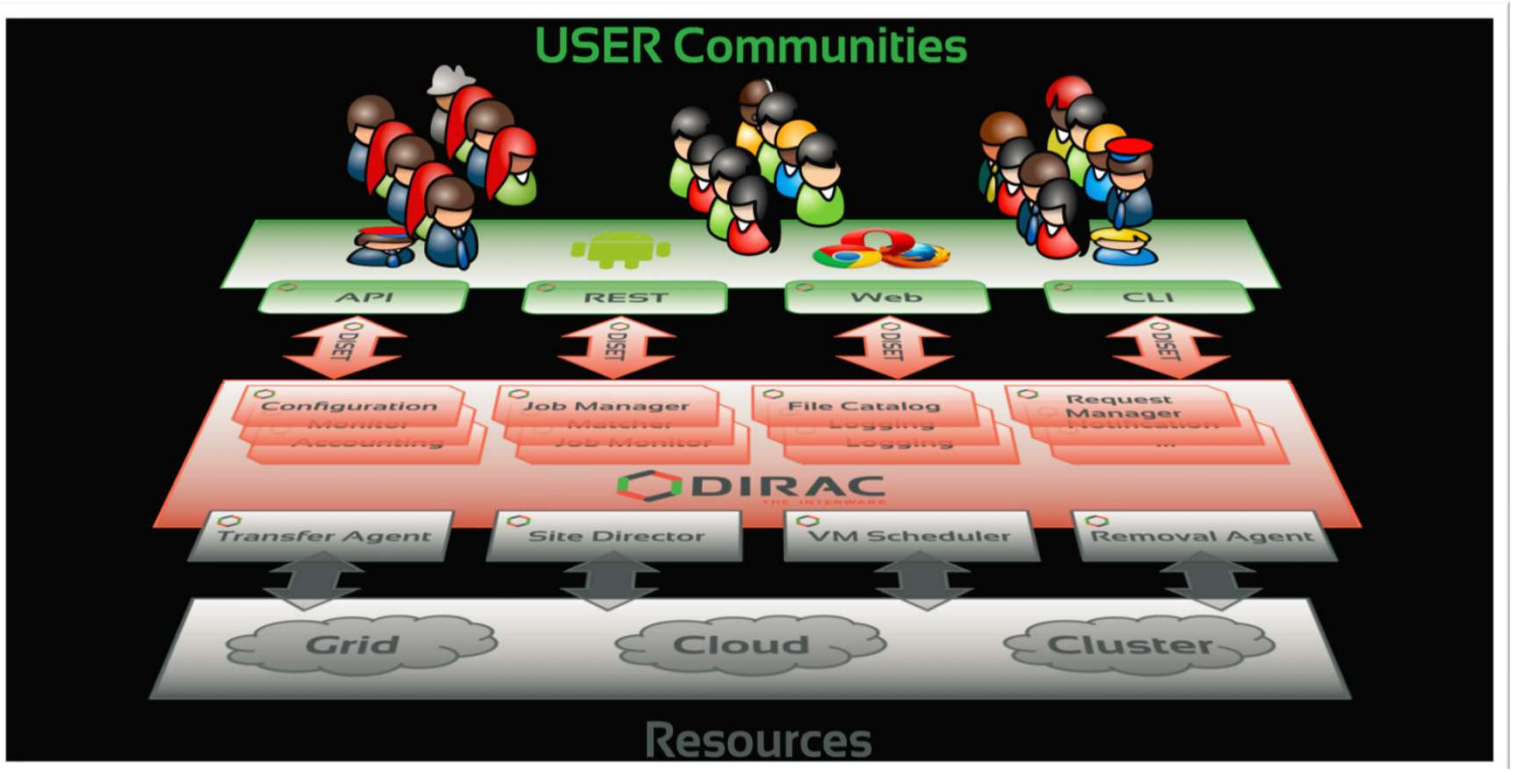
- ITEP
- JINR-LCG2 (Dubna)
- RRC-KI
- RU-Moscow-KIAM
- RU-Phys-SPbSU
- RU-Protvino-IHEP
- RU-SPbSU
- Ru-Troitsk-INR
- ru-IMPB-LCG2
- ru-Moscow-FIAN
- ru-Moscow-MEPHI
- ru-PNPI-LCG2 (Gatchina)
- ru-Moscow-SINP
- Kharkov-KIPT (UA)
- BY-NCPHEP (Minsk)
- UA-KNU
- UA-BITP

Эволюция модели компьютинга

- Расширение компьютерных ресурсов за счет использования внешних невыделенных ресурсов (HLT, Clouds, HPC…)
- Изменения модели компьютинга в каждом эксперименте, с целью оптимизации использования ресурсов
- Значительные усилия вкладываются в развитие программного обеспечения, чтобы улучшить общую производительность при использовании современных архитектур (многоядерность, GPU…)
- Оптимизации процессов обработки, количество хранящихся реплик данных и др.

Платформа DIRAC

- DIRAC has all the necessary components to build ad-hoc grid infrastructures **interconnecting** computing resources of different types, allowing **interoperability** and simplifying **interfaces**.
- This allows to speak about the DIRAC ***interware***.



* PanDa в эксперименте ATLAS

В эксперименте ATLAS на Большом адронном коллайдере разработана платформа для управления вычислительными ресурсами PanDA Workload Management System (WMS), которая обладает следующими возможностями:

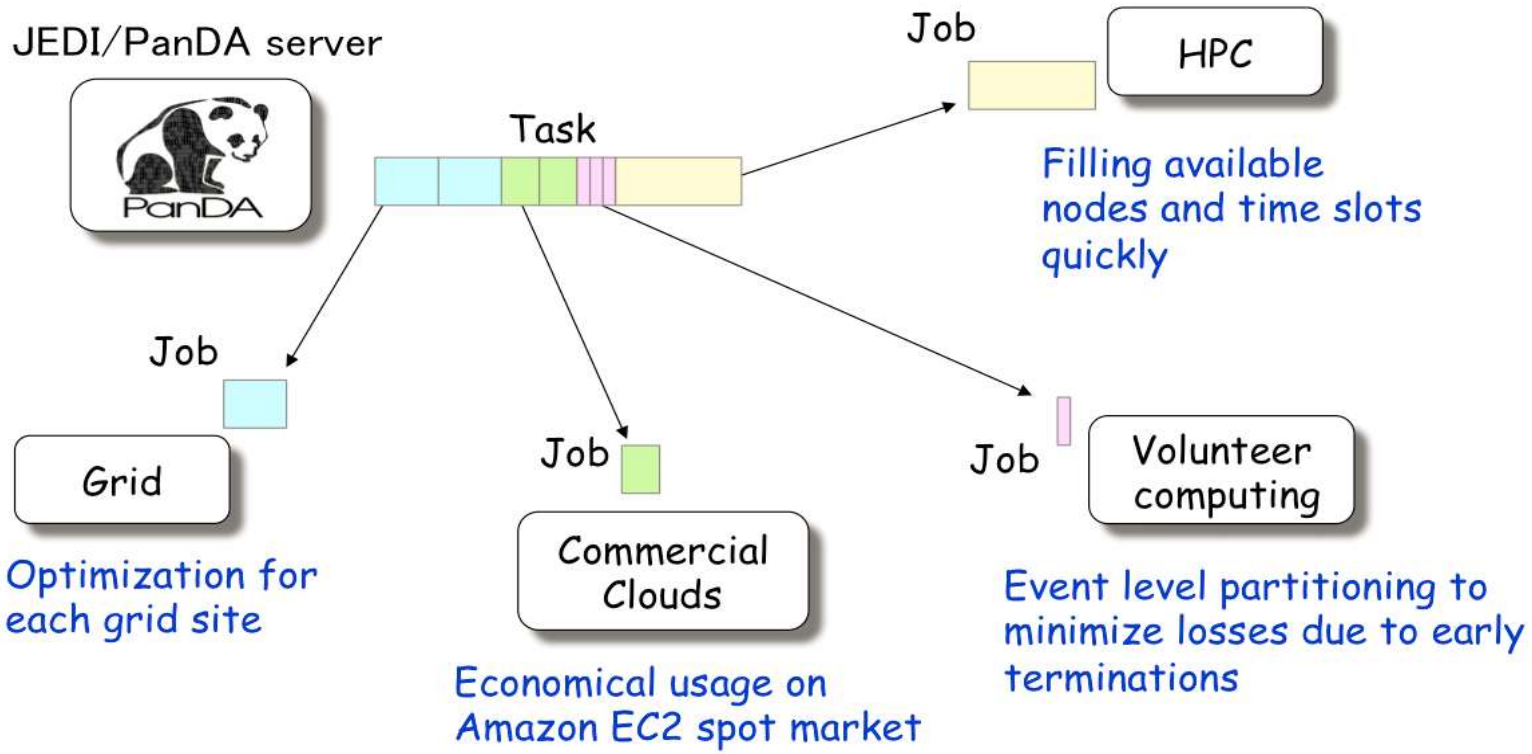
- Проект PanDA начался в 2005 году группами BNL и UTA - **P**roduction **an**d **D**ata **A**nalysis system.
- Автоматизированная и гибкая система управления заданиями, которая может оптимально сделать распределенные ресурсы доступными пользователю.
- С помощью PanDA, физики видят единый вычислительный ресурс, который предназначен для обработки данных эксперимента, даже если дата-центры разбросаны по всему миру
- PanDa изолирует физиков от аппаратного обеспечения, системного и промежуточного программного обеспечения и других технологических сложностей, связанных с конфигурированием сети и оборудования.
- Вычислительные задачи автоматически отслеживаются и выполняются. Могут выполняться групповые задачи физиков

В настоящее время PanDa контролирует:

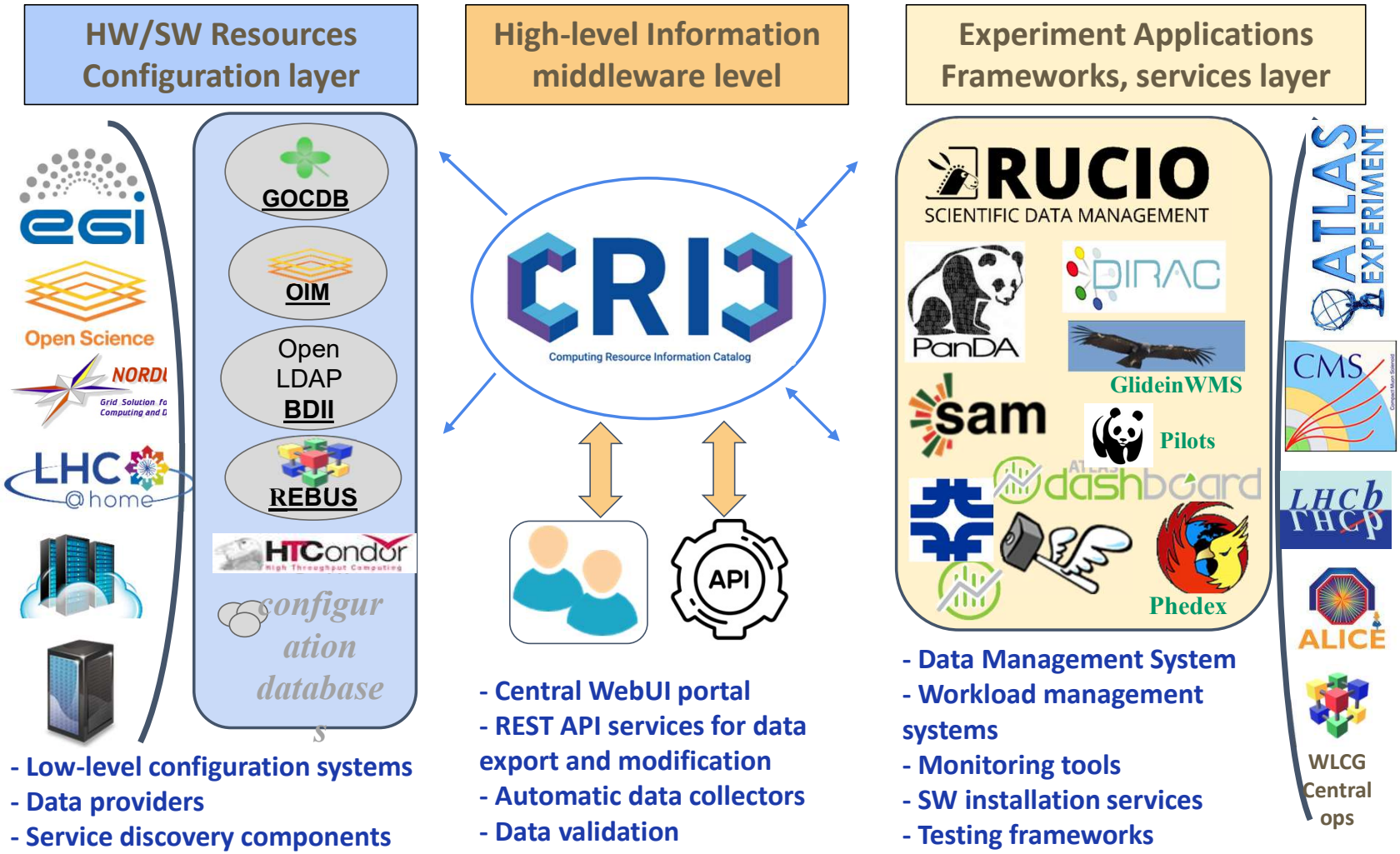
- **сотни дата - центров в 50 странах мира**
- **сотни тысяч вычислительных узлов**
- **сотни миллионов заданий в год**
- **тысячи пользователей**

Dynamic Job Definition in PanDA

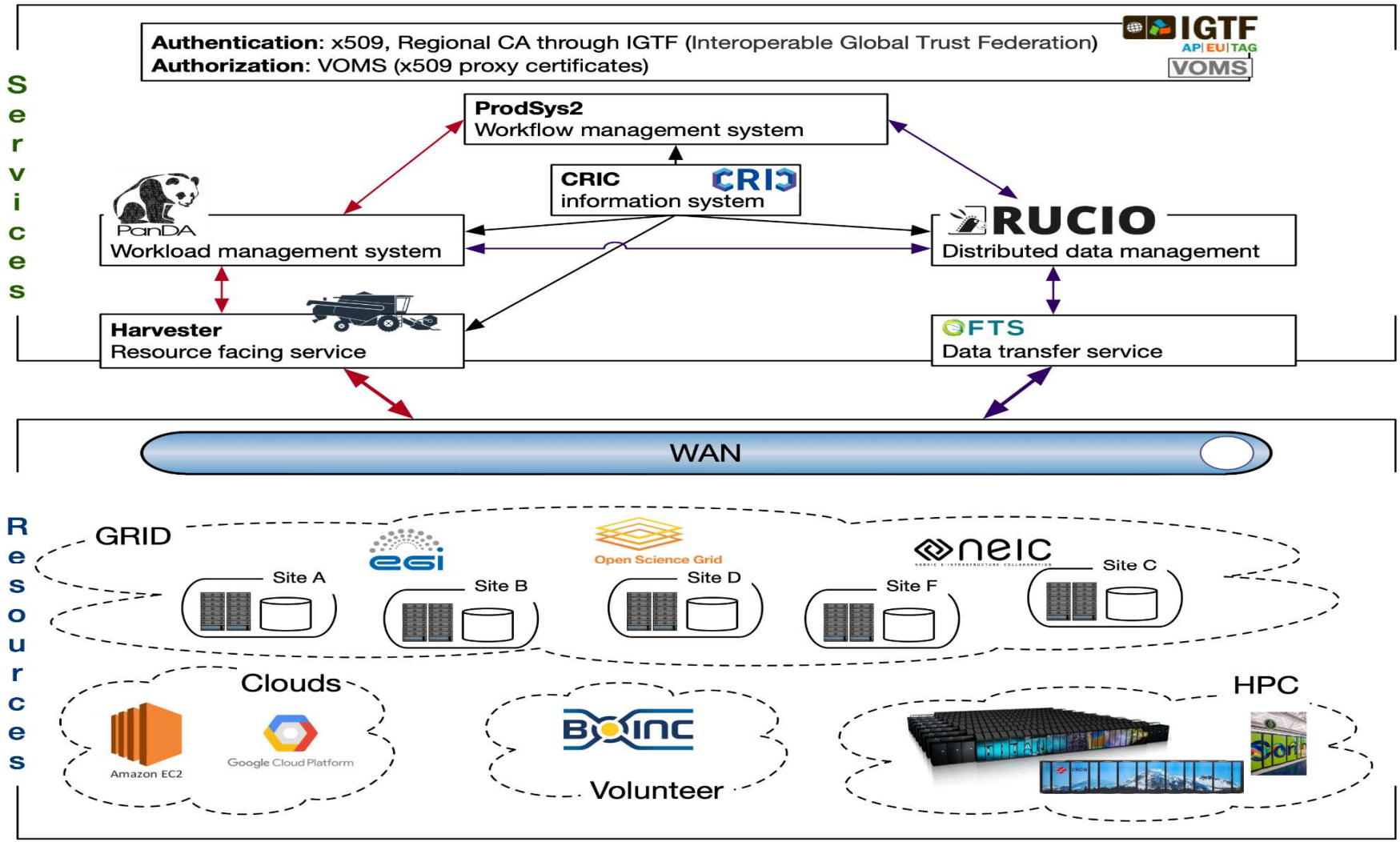
- Workload partitioning for traditional and opportunistic resources



CRIC: a unified topology system for a large scale, heterogeneous and dynamic computing infrastructure

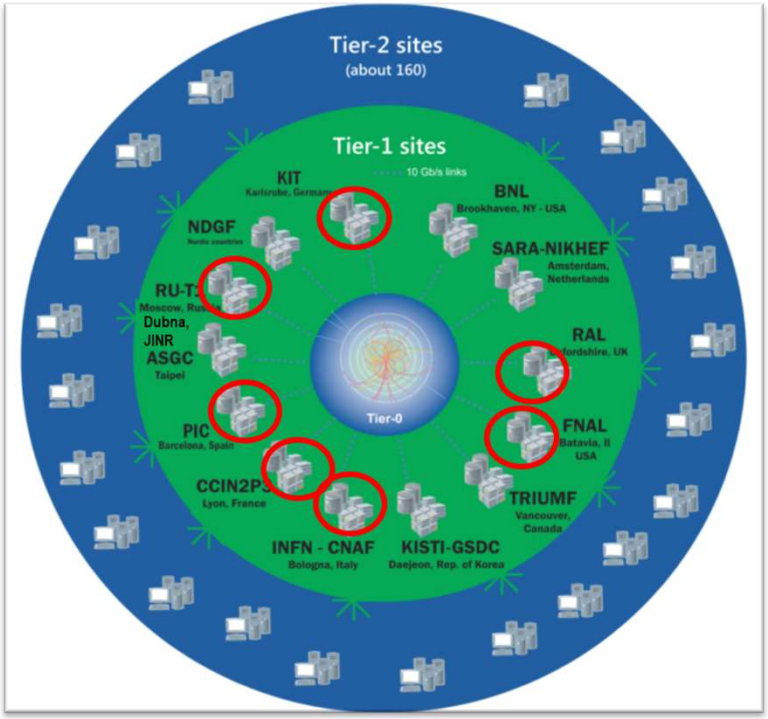


ATLAS computing



The Worldwide LHC Computing Grid

WLCG: an International collaboration to distribute and analyse LHC data. Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists



The mission of the WLCG project is to provide global computing resources to store, distribute and analyze the ~250-300 Petabytes of data expected every year of operations from the Large Hadron Collider.

WLCG computing enabled physicists to announce the discovery of the Higgs Boson.

180 sites

42 countries

> 12k physicists

~1.6 M CPU cores

~2 EB of storage (1 EB - CERN)

> 3 million jobs/day

100-400 Gb/s links

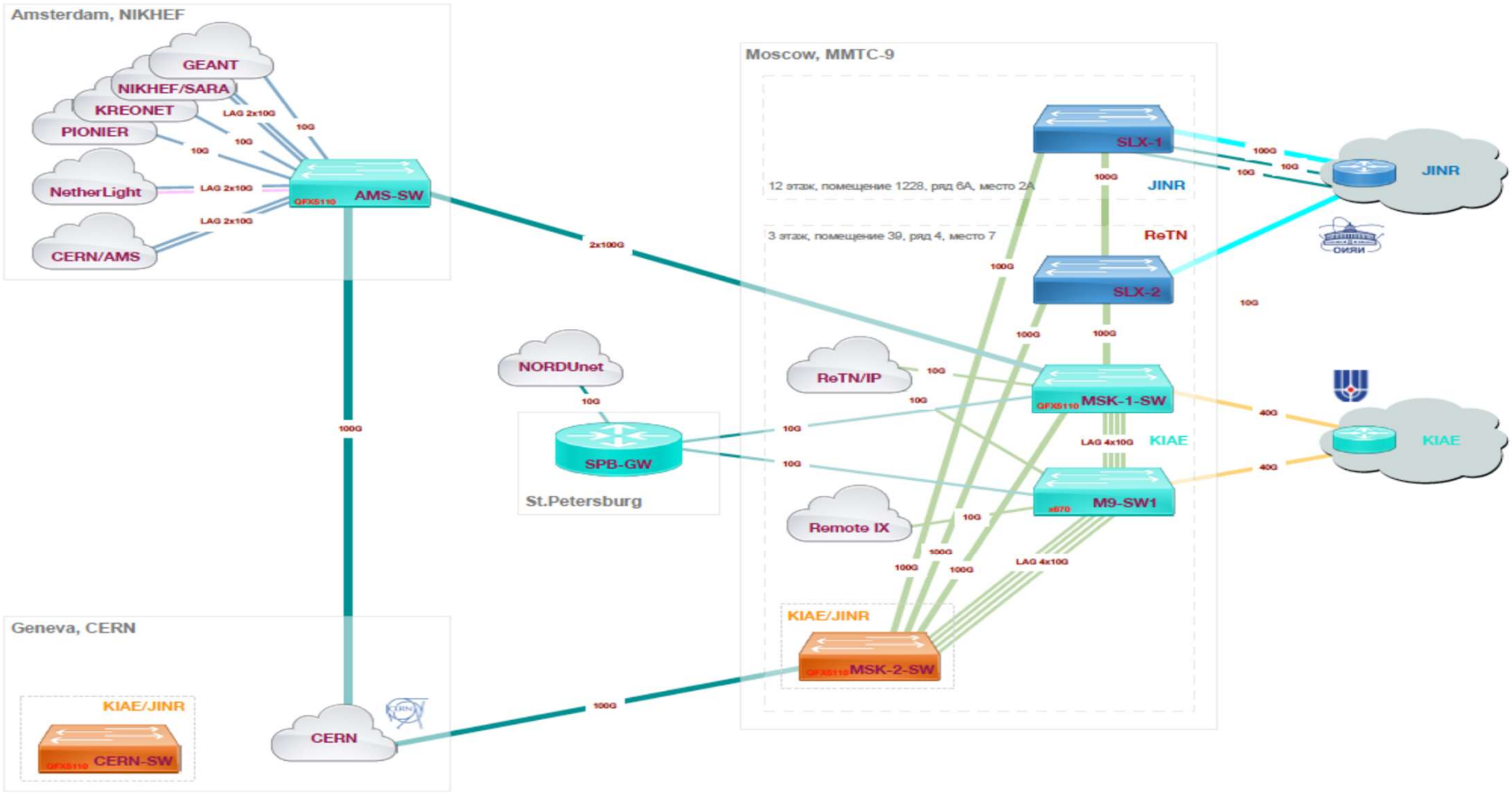
Worldwide LHC Computing Grid - 2023

Tier0 (CERN): data recording, reconstruction and distribution

Tier1: permanent storage, re-processing, analysis

Tier2: Simulation, end-user analysis

Сеть RDIG-M для мегасайенс проектов



Tier1 – Tier2 in Russia 2023 (Sum CPU in HS23 hours)

• **JINR-T1**	**1,906,119,856**	**59.33%**
• **RRC-KI-T1**	**686,785,972**	**21.38%**
• **JINR-LCG2**	**560,810,281**	**17.46%**
• **RU-Protvino-IHEP**	**40,125,942**	**1.25%**
• **ru-PNPI**	**13,372,211**	**0.42%**
• **Ru-Troitsk-INR-LCG2**	**4,891,806**	**0.15%**
• **RU-SARFTI**	**311,101**	**0.01%**
• **RU-SPbSU**	**205,437**	**0.01%**

International Large-scale projects

Russian research institutes and universities actively participate in international large-scale projects:

- LHC, CERN (experiments: ATLAS, ALICE, LHCb, CMS)
- XFEL, DESY (European free electron laser)
- ESRF, France (European synchrotron center)
- FAIR, GSI, Germany (CBM, PANDA experiments)
- ITER, France …

International large-scale projects are being prepared in Russia:

- **NICA**, JINR, Dubna (proton and heavy ion collider)
- **PIK**, PNPI, Gatchina (high-flow reactor complex)
- **SKIF**, INP SB RAS Novosibirsk (Siberian ring photon source)
- **Super S-Tau Fabric**, Sarov (electron-positron collider)
- **Нейтринная программа (Байкал, JUNO, NOVA, DUNE …)**
- **синхротронно-нейтронная программа, науки о жизни**

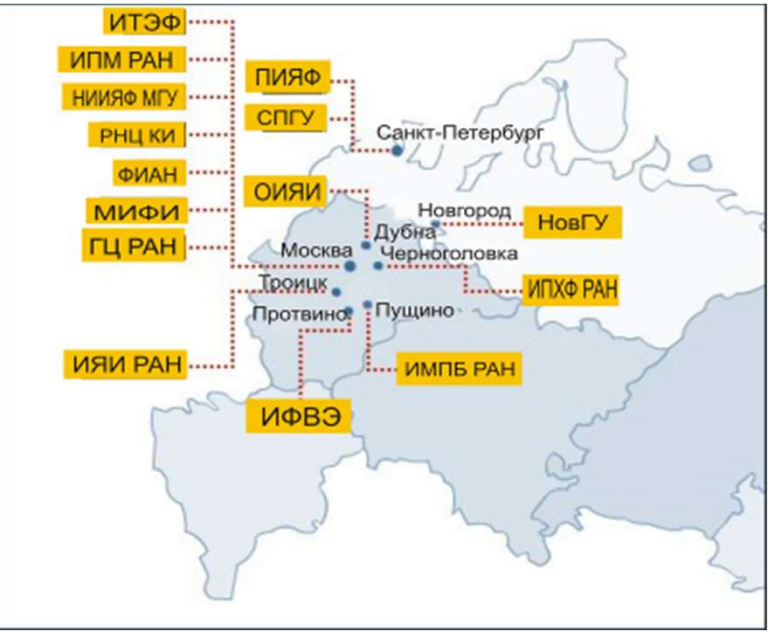
Joint Institute for Nuclear Research
SCIENCE BRINGING NATIONS TOGETHER

NICA

СКИФ
СИБИРСКИЙ КОЛЬЦЕВОЙ ИСТОЧНИК ФОТОНОВ

Институт ядерной физики имени Г. И. Будкера СО РАН

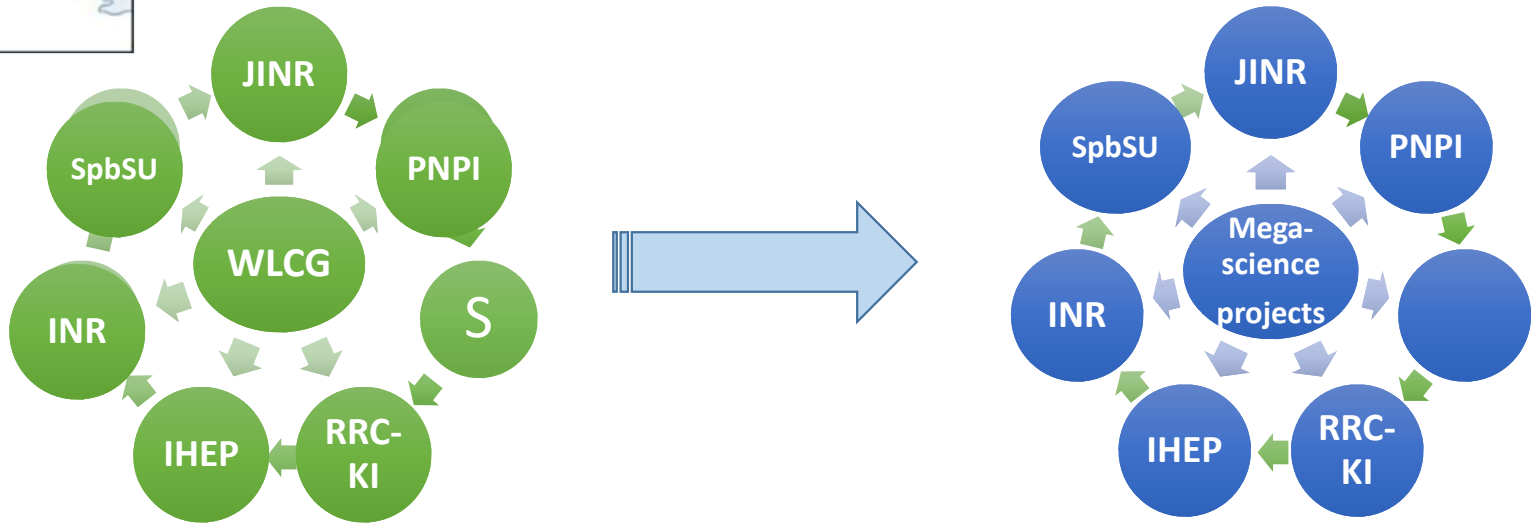
From RDIG to RDIG-M



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

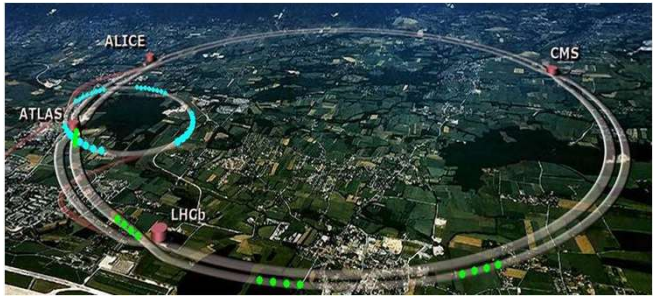
A protocol between CERN, Russia and JINR on participation in the LCG project was signed in 2003. MoU on participation in the WLCG project was signed in 2007.

Consortium RDIG-M – Russian Data Intensive GRID for Megascience projects

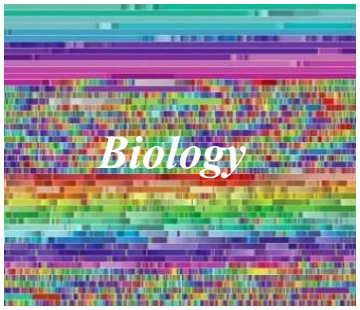


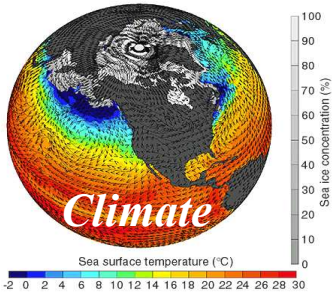
HPC+Big Data+Artificial intelligence

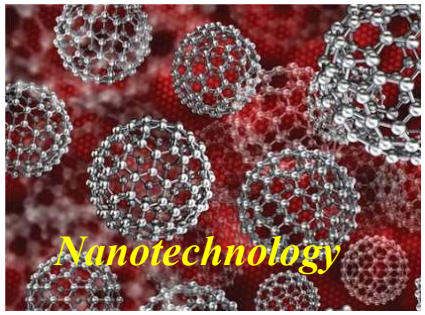
High Energy Physics



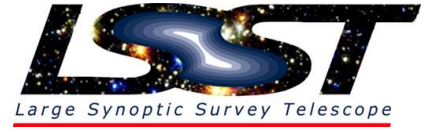
CERN Large Hadron Collider > 600 Pb/Year

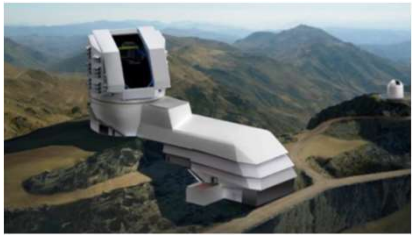






Square Kilometer Array radio telescope > 1 Eb/Year ra data (estimatic

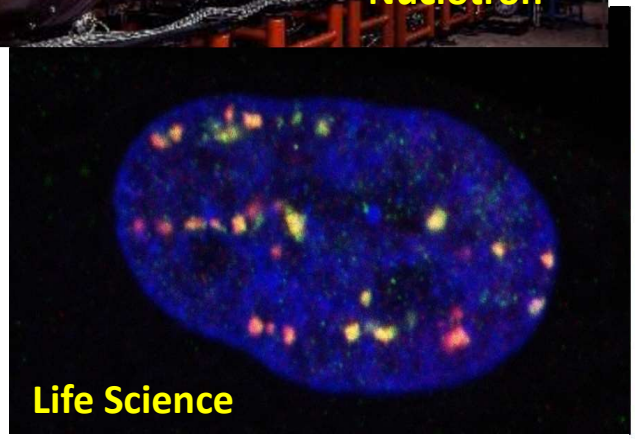




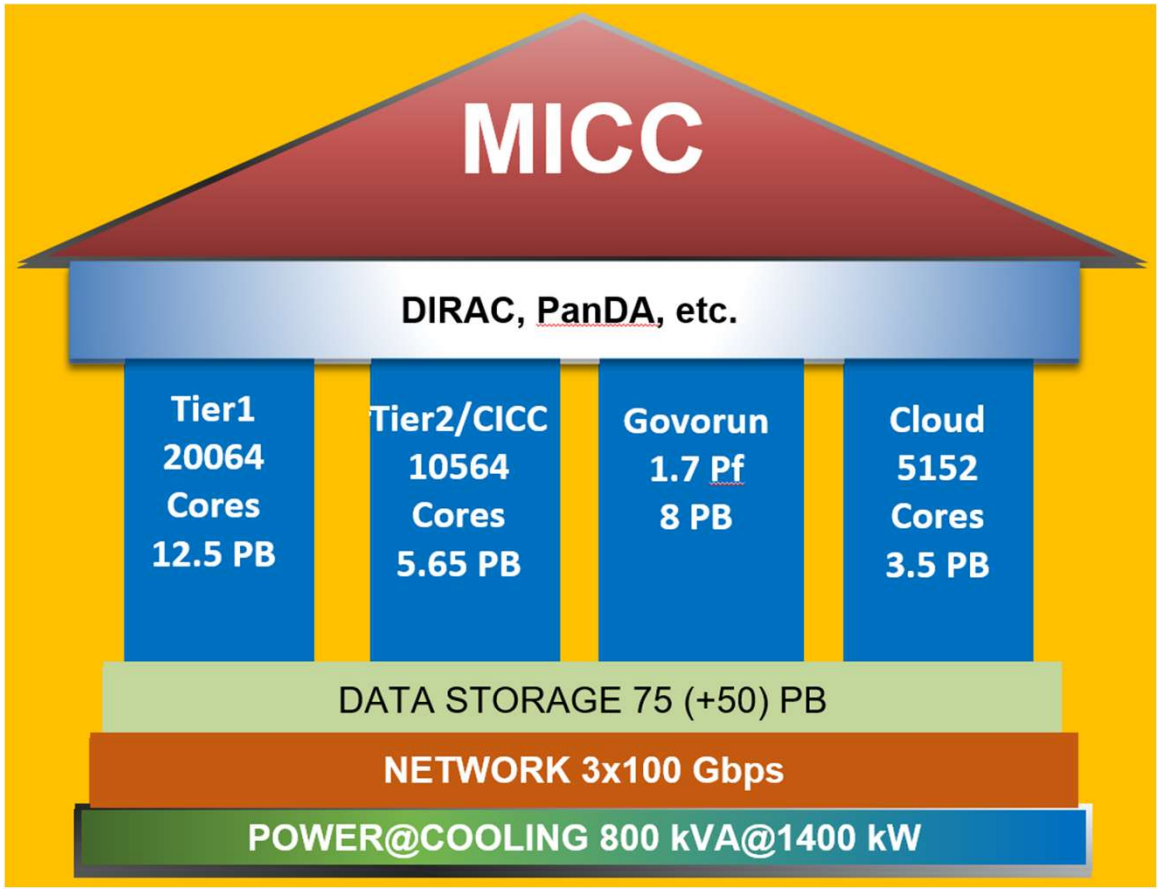
Large Synoptic Survey Telescope > 10 Pb/Year (estimation)

... et cetera

Implementation of the JINR Development Program



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

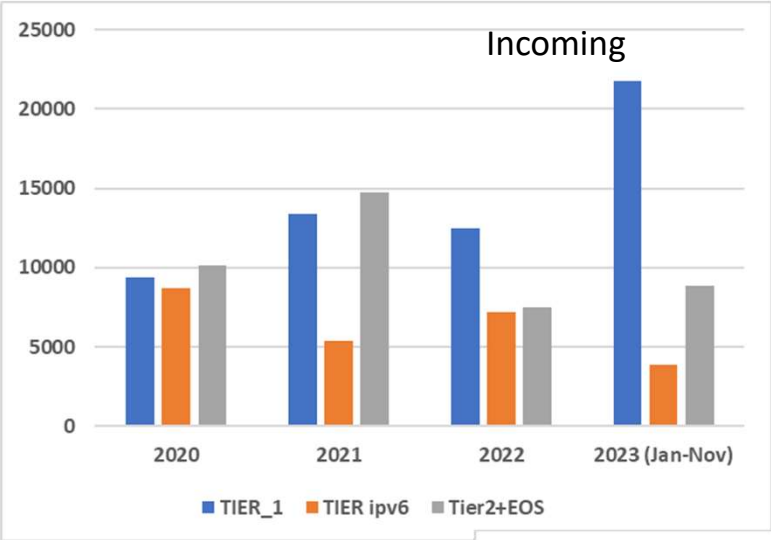
Network

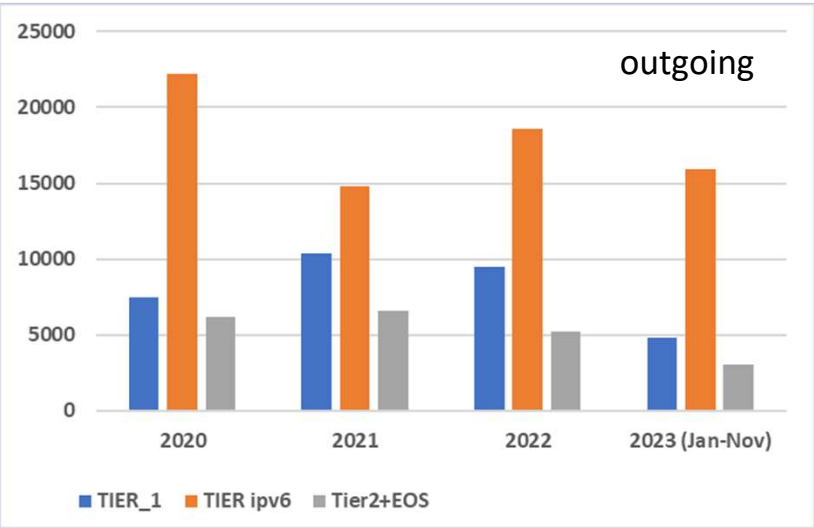
- Wide Area Networkr
- Local Area Network

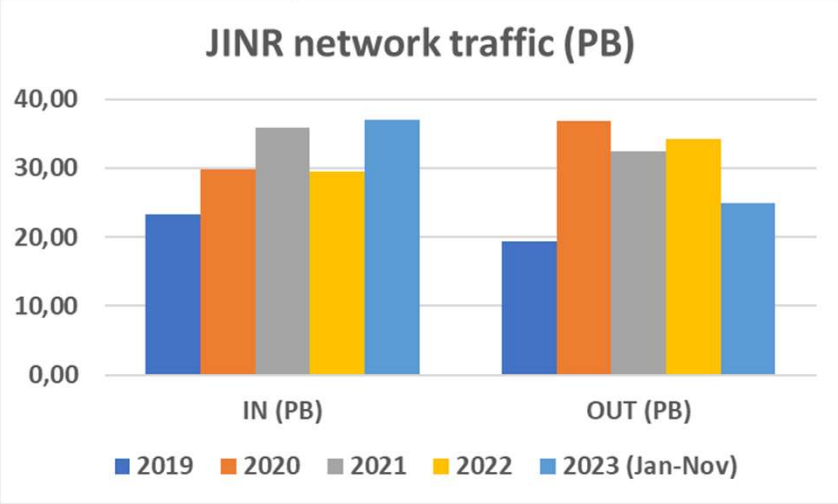
The main objective of the project is to ensure multifunctionality, scalability, high performance, reliability and availability in 24x7x365 mode for different user groups that carry out scientific studies within the JINR Topical Plan

Networking @ Traffic

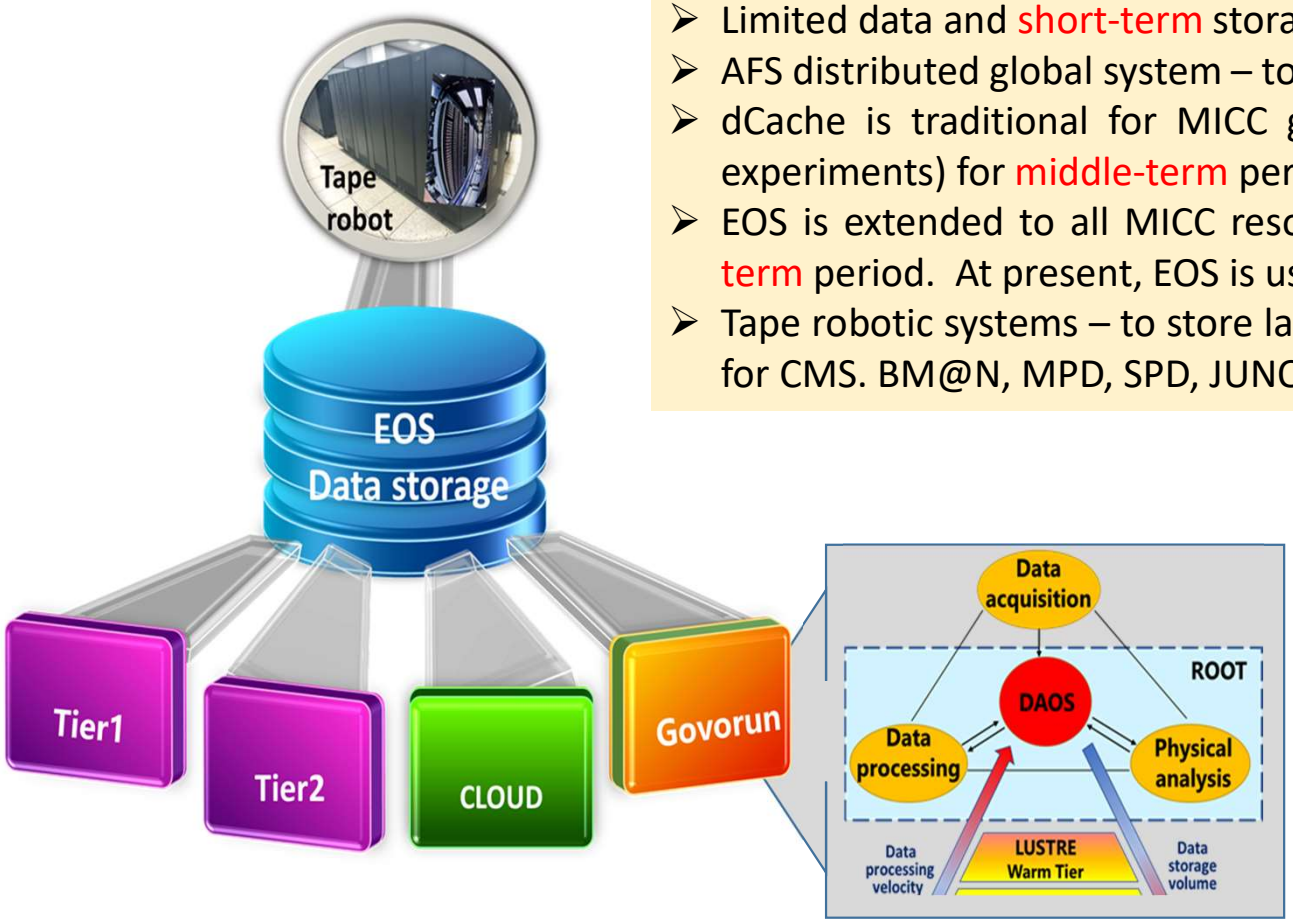
Distribution of the incoming and outgoing traffics by the JINR MICC in 2020-2023 (TB)







Distributed Multilayered Data Storage System



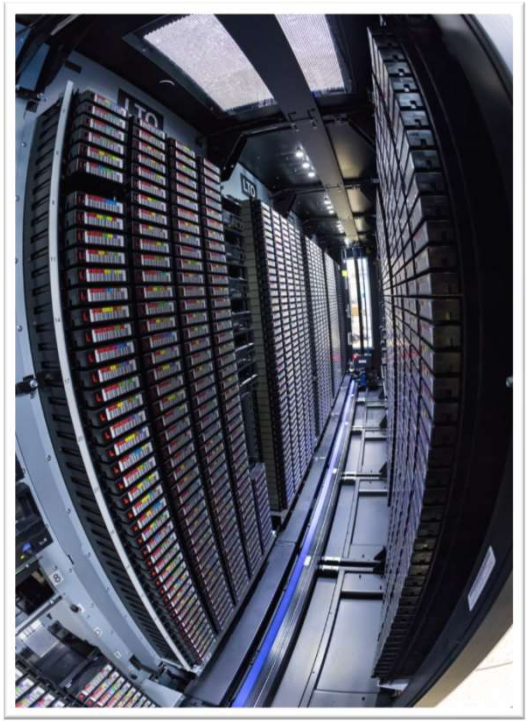
- Limited data and short-term storage – to store OS itself, temporary user files
- AFS distributed global system – to store user home directories and software
- dCache is traditional for MICC grid sites – to large amounts of data (mainly LHC experiments) for middle-term period
- EOS is extended to all MICC resources – to store large amounts of data for middle-term period. At present, EOS is used for storage by BM@N, MPD, SPD, BaikalGVD, etc.
- Tape robotic systems – to store large amounts of data for long-term period. At present for CMS. BM@N, MPD, SPD, JUNO – in progress.

Special hierarchical data processing and storage system with a software-defined architecture was developed and implemented on the “Govorun” supercomputer.

According to the speed of accessing data there are next layers:

- ✓ very hot data (DAOS (Distributed Asynchronous Object Storage)) ,
- ✓ the most demanded data (fastest access),
- ✓ hot data
- ✓ warm data (LUSTRE).

JINR Tier1 for CMS (LHC) and NICA



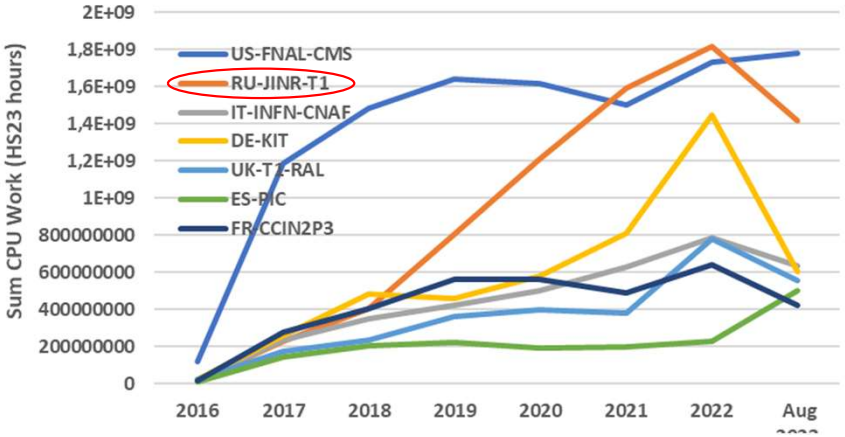
- 20064 cores
- 360 kHS06
- 12.5 PB disks
- 50.6 (+ 50) PB tapes
- 100% reliability and availability

Since the beginning of 2015, a full-scale WLCG Tier1 site for the CMS experiment has been operating at MLIT JINR.

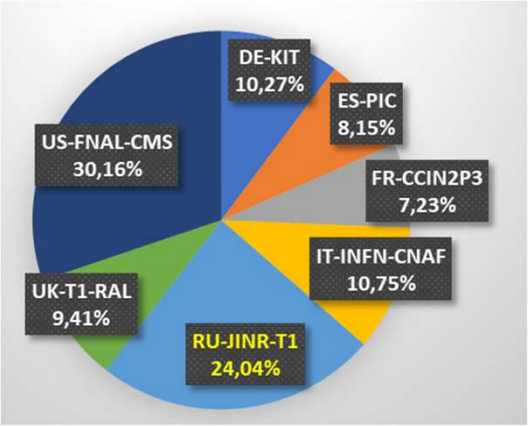
The importance of developing, modernizing and expanding the computing performance and data storage systems of this center is dictated by the research program of the CMS experiment, in which JINR physicists take an active part within the RDMS CMS collaboration.

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

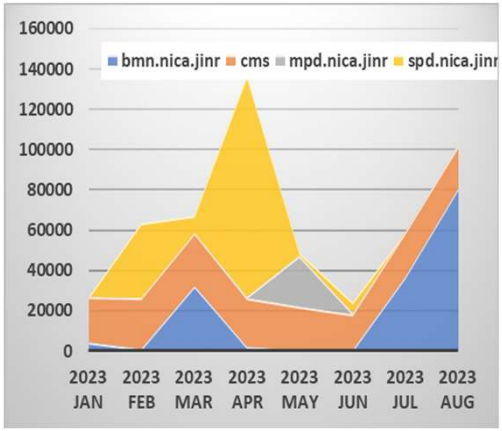
Since 2021 the JINR Tier1 center has demonstrated stable work not only for CMS (LHC), but also for NICA experiments.



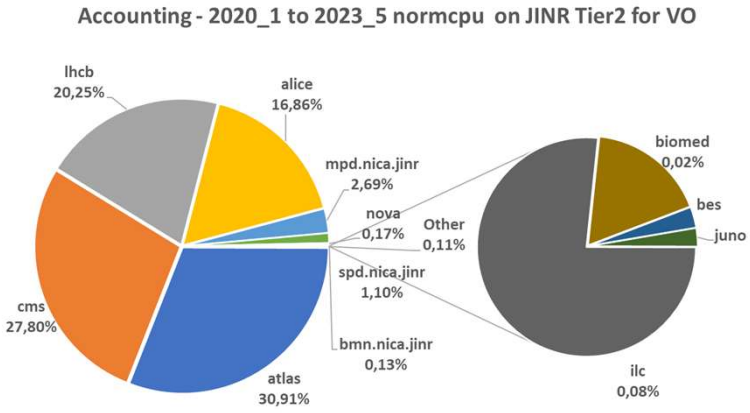
Distribution by CPU Work (HS23 hours) among CMS Tier1 worldwide



Distribution by the number of jobs completed on Tier1 by CMS, BM@N, MPD and SPD

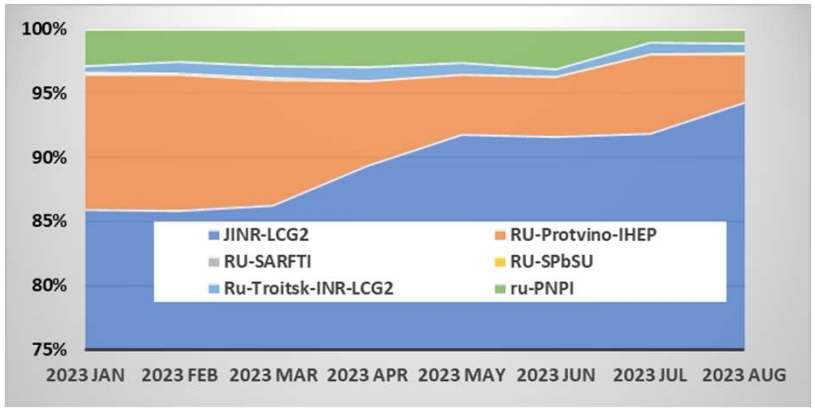


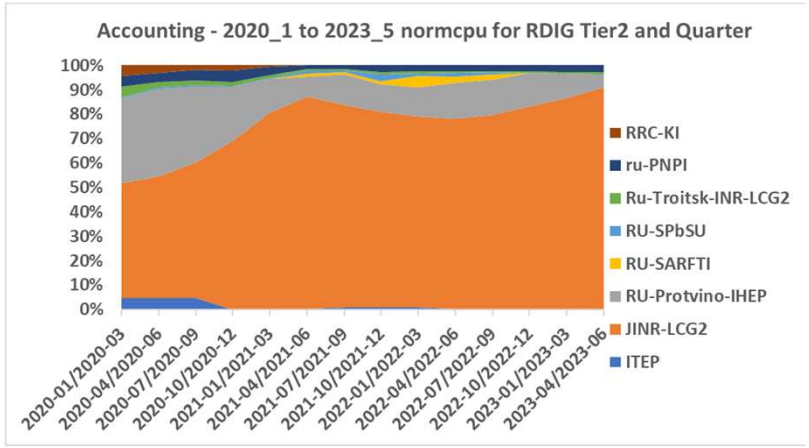
Tier2 at JINR



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, FAIR, etc.).

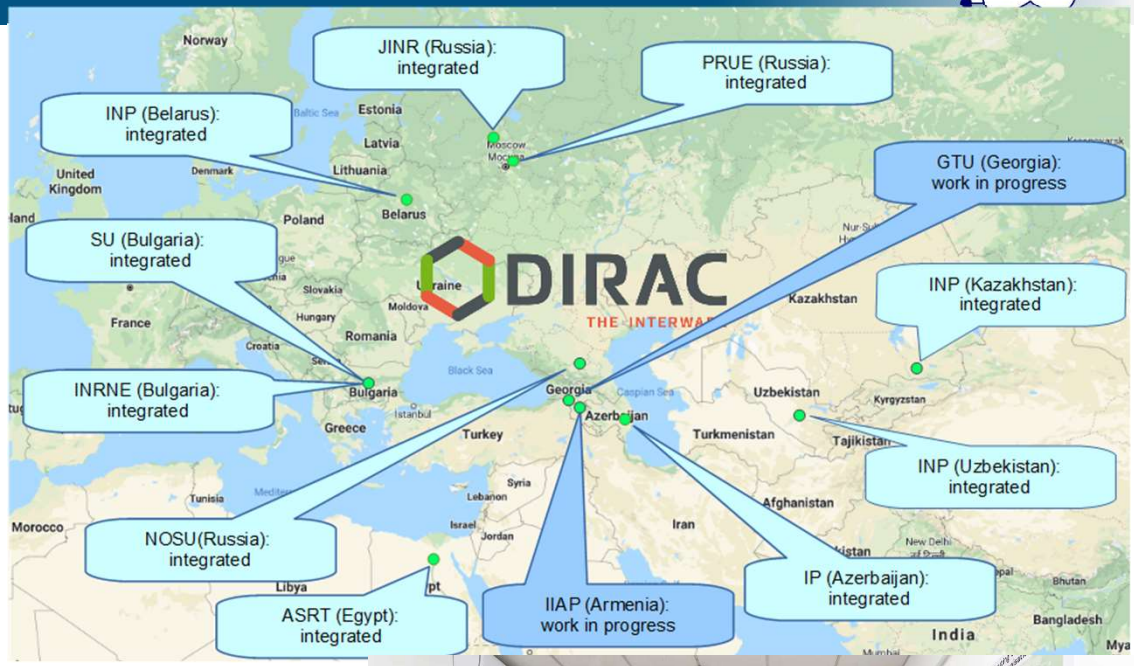
The JINR Tier2 output is the highest (89.3%) in the Russian Data Intensive Grid (RDIG) Federation.





Cloud Infrastructure

DIRAC-based distributed information and computing environment (DICE) that integrates the JINR Member State organizations' clouds

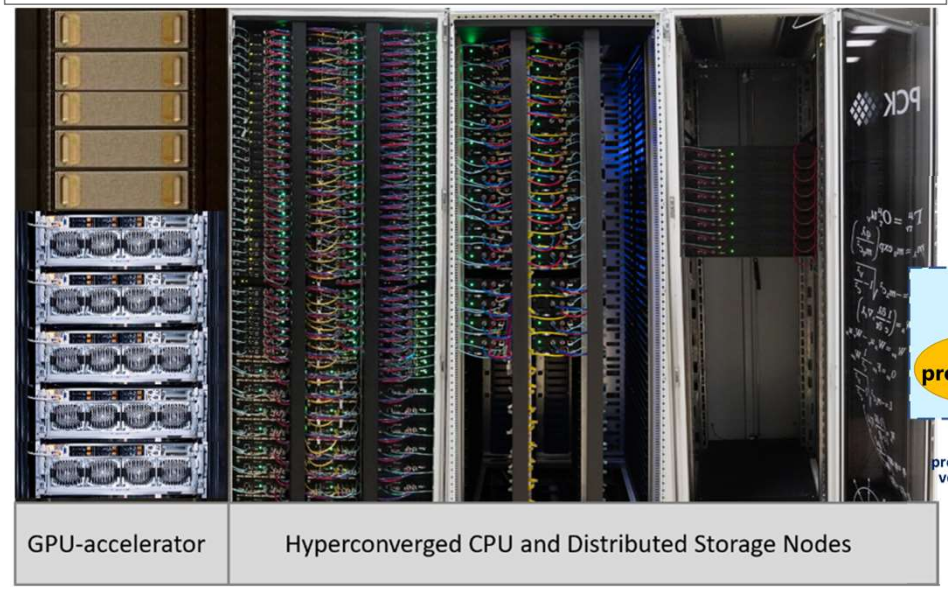


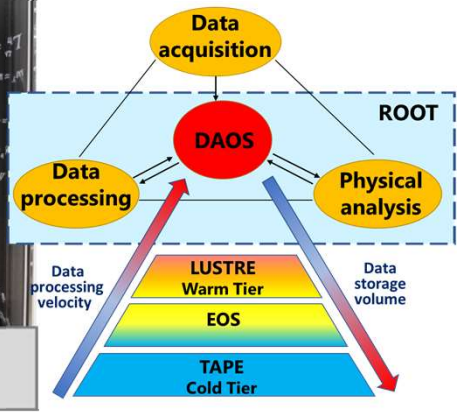
- Cloud Platform - OpenNebula
- Virtualization - KVM
- Storage (Local disks, Ceph)
- Total Resources
  ~ 5,152 CPU cores; 80 TB RAM;
  3.5 PB of raw ceph-based storage

- Computational resources for neutrino experiments
- Testbeds for research and development in IT
- COMPASS production system services
- Data management system of the UNECE ICP Vegetation
- Scientific and engineering computing
- Service for data visualization
- VMs for JINR users

“Govorun” Supercomputer

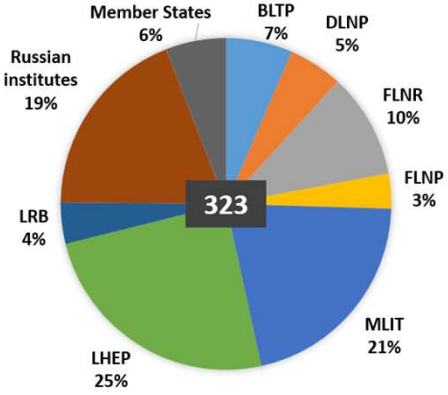
- Hyper-converged software-defined system
- Hierarchical data processing and storage system
- Scalable solution Storage-on-demand
- Total peak performance: 1.7 PFlops DP
- GPU component based on NVIDIA Tesla V100&A100
- CPU component based on RSC “Tornado” liquid cooling solutions
- The most energy-efficient center in Russia (PUE = 1.06)
- Storage performance >300 GB/s





Key projects that use the resources of the SC “Govorun”:

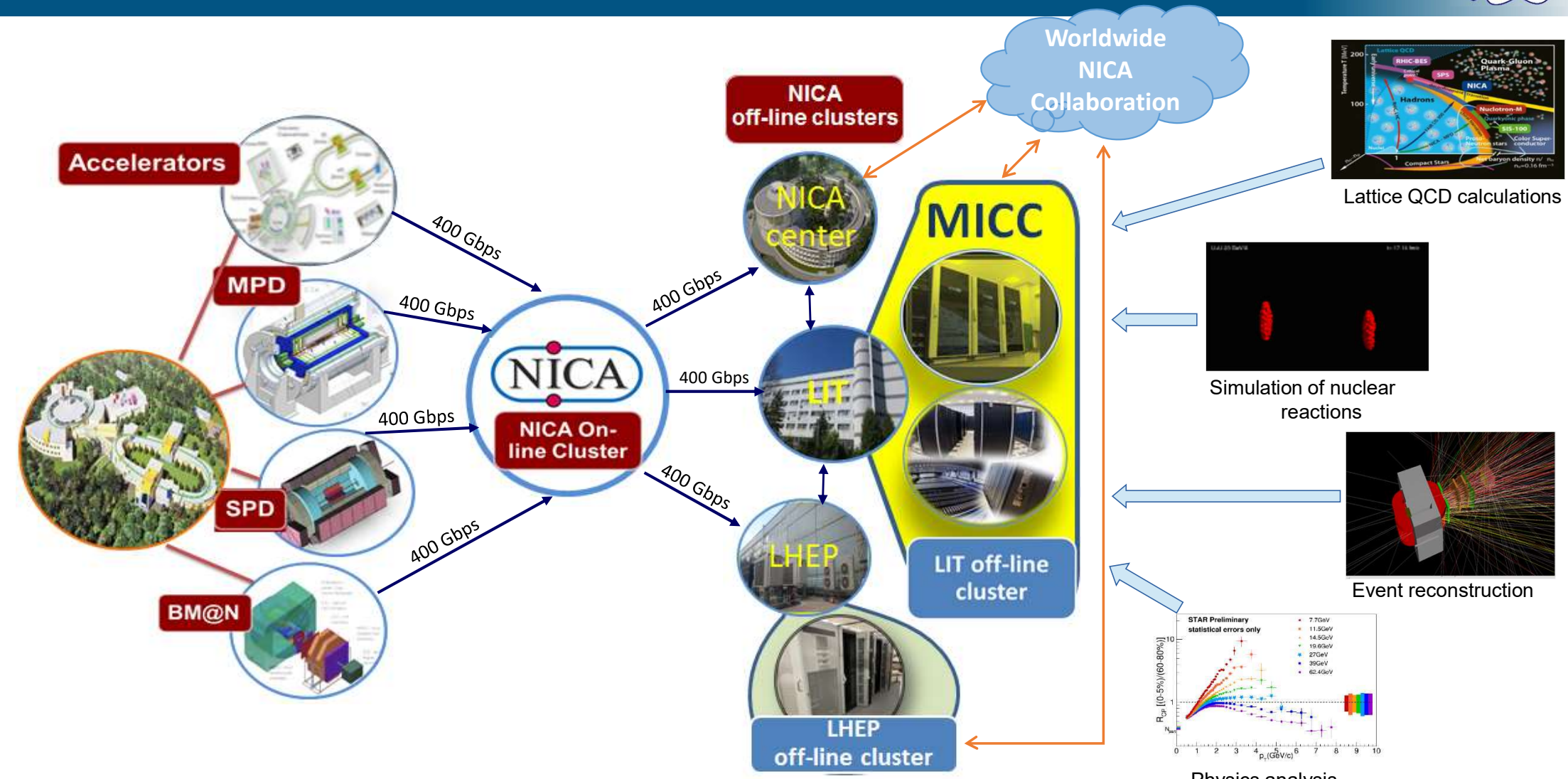
- NICA megaproject,
- calculations of lattice quantum chromodynamics,
- computations of the properties of atoms of superheavy elements,
- studies in the field of radiation biology,
- calculations of the radiation safety of JINR’s facilities.



Total number of users : 323

> 250 user papers (two in Nature Physics)

NICA Computing Concept & Challenges

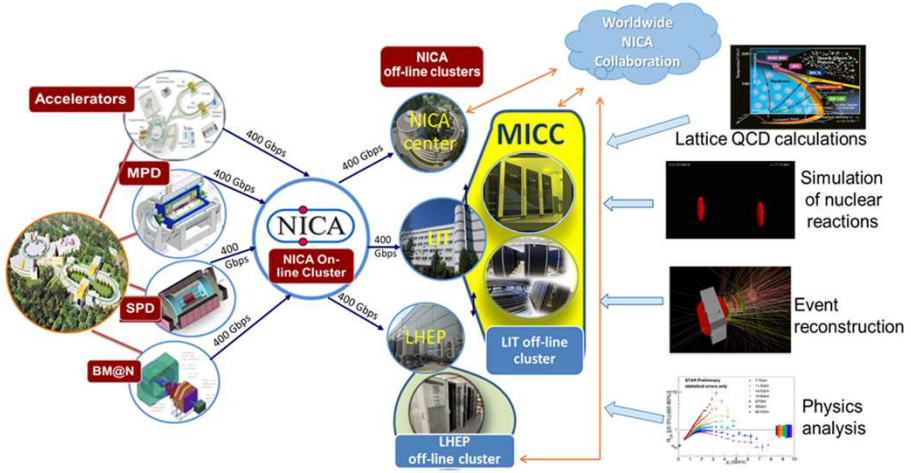


Development of the NICA Information and Computer Complex

The Seven-Year Plan provides for the creation of a long-term data storage center on the MICC resources at MLIT (Tier0). 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.

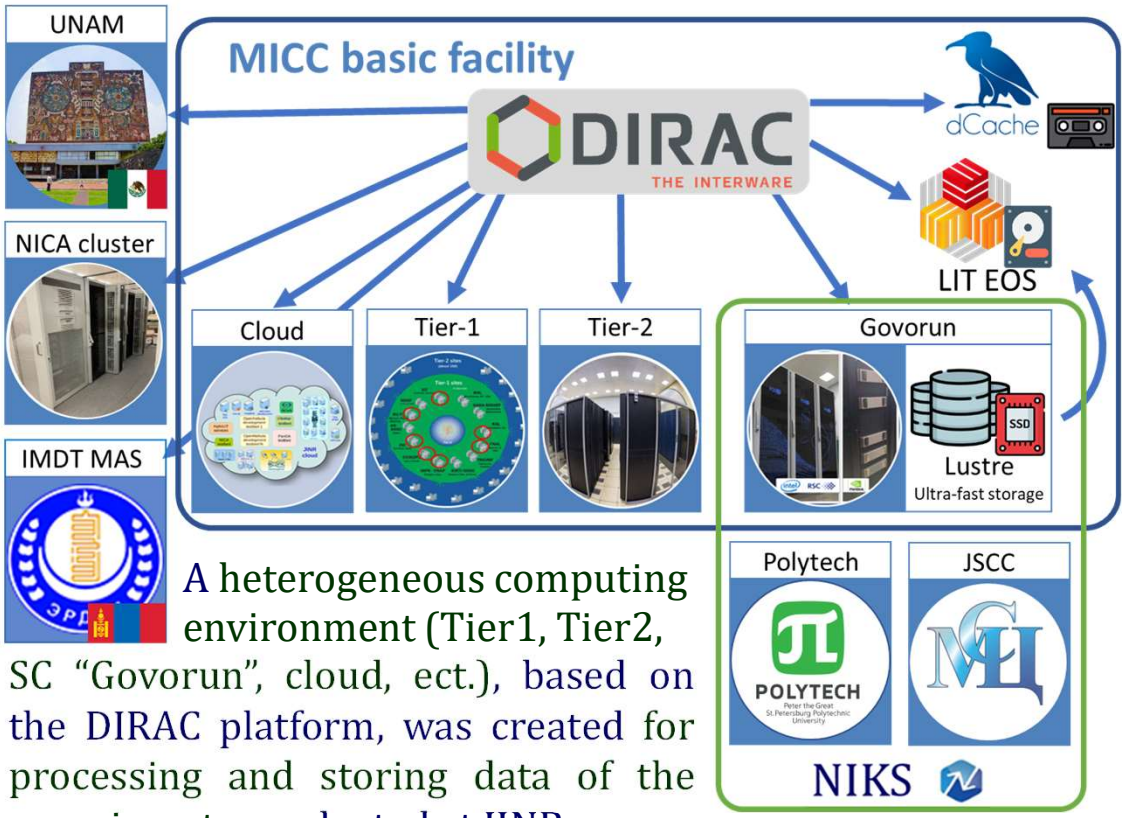
The information and computer unit of the NICA complex embraces:

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

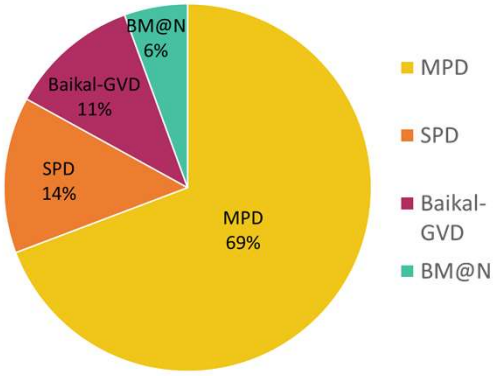
DIRAC-based distributed heterogeneous environment



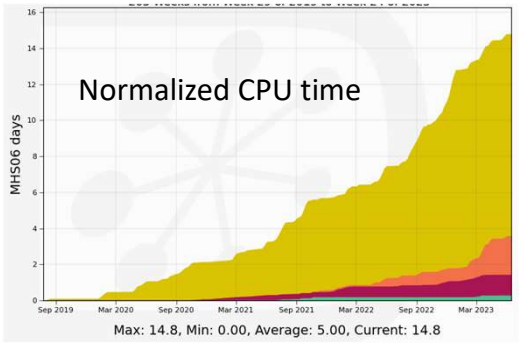
A heterogeneous computing environment (Tier1, Tier2, SC "Govorun", cloud, ect.), based on the DIRAC platform, was created for processing and storing data of the experiments conducted at JINR.

The distributed infrastructure is used by the MPD, Baikal-GVD, BM@N, SPD.

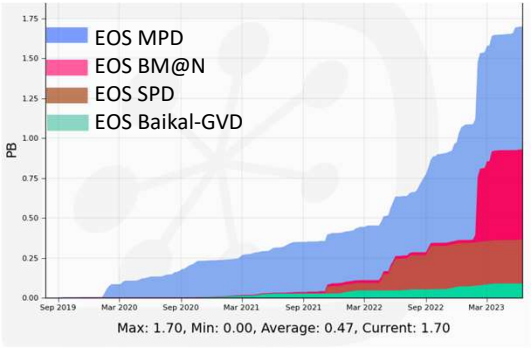
Use of DIRAC platform by experiments in 2019-2022



Total Number of executed jobs



Data processed by experiments



The major user of the distributed platform is the MPD experiment

Summary statistics of using the DIRAC platform for MPD tasks in 2019-2022

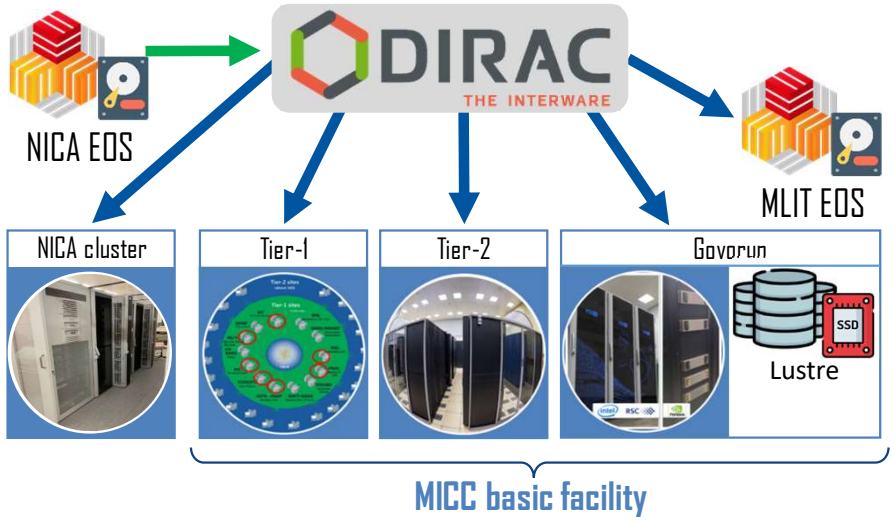
31 data generation campaigns	1.283 B events generated	440M events reconstructed	1.5M jobs completed	1557 years total computation time	1.3 PB MPD data produced

BM@N Run 8 Data

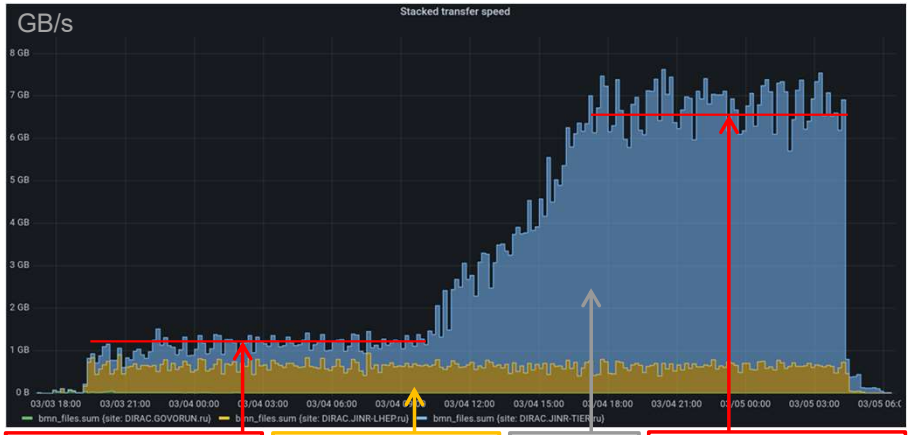
The 8th BM@N physics run was the first time at JINR when the entire computing infrastructure, integrated by DIRAC, was used for the complete reconstruction of raw experimental data. During the session, there were received about 550 million events, which were written in 31,306 files with a total size of more than 430 TB.

The reconstruction process was carried out in two stages:

1. Raw → DIGI (99% processed on Tier1 and the NICA cluster, large files (16 - 250 GB) could only be processed on the “Govorun” SC)
2. DIGI → DST (Tier1, Tier2, NICA cluster and “Govorun” SC)



Raw → DIGI: High disk and network system load



300 jobs
4 MB/s per job

NICA cluster

Tier1

1,580 jobs
4.1 MB/s per job

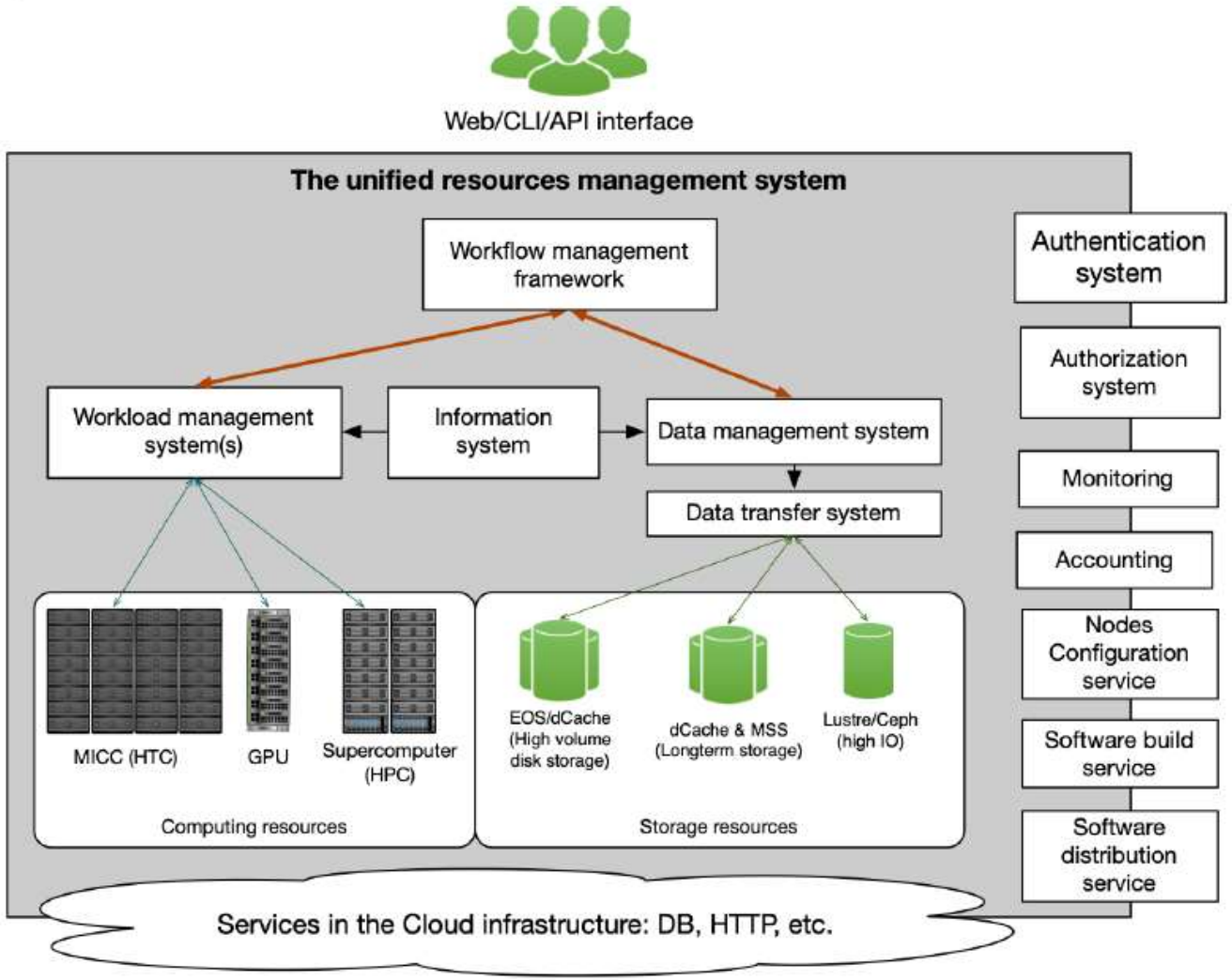
Maximal data transfer speed (Read+Write) with EOS (MLIT) – 7.5 GB/s

Jobs completed
62612

Real time
~48 h

RAW
436 TB
→
DIGI
23 TB
→
DST
53 TB

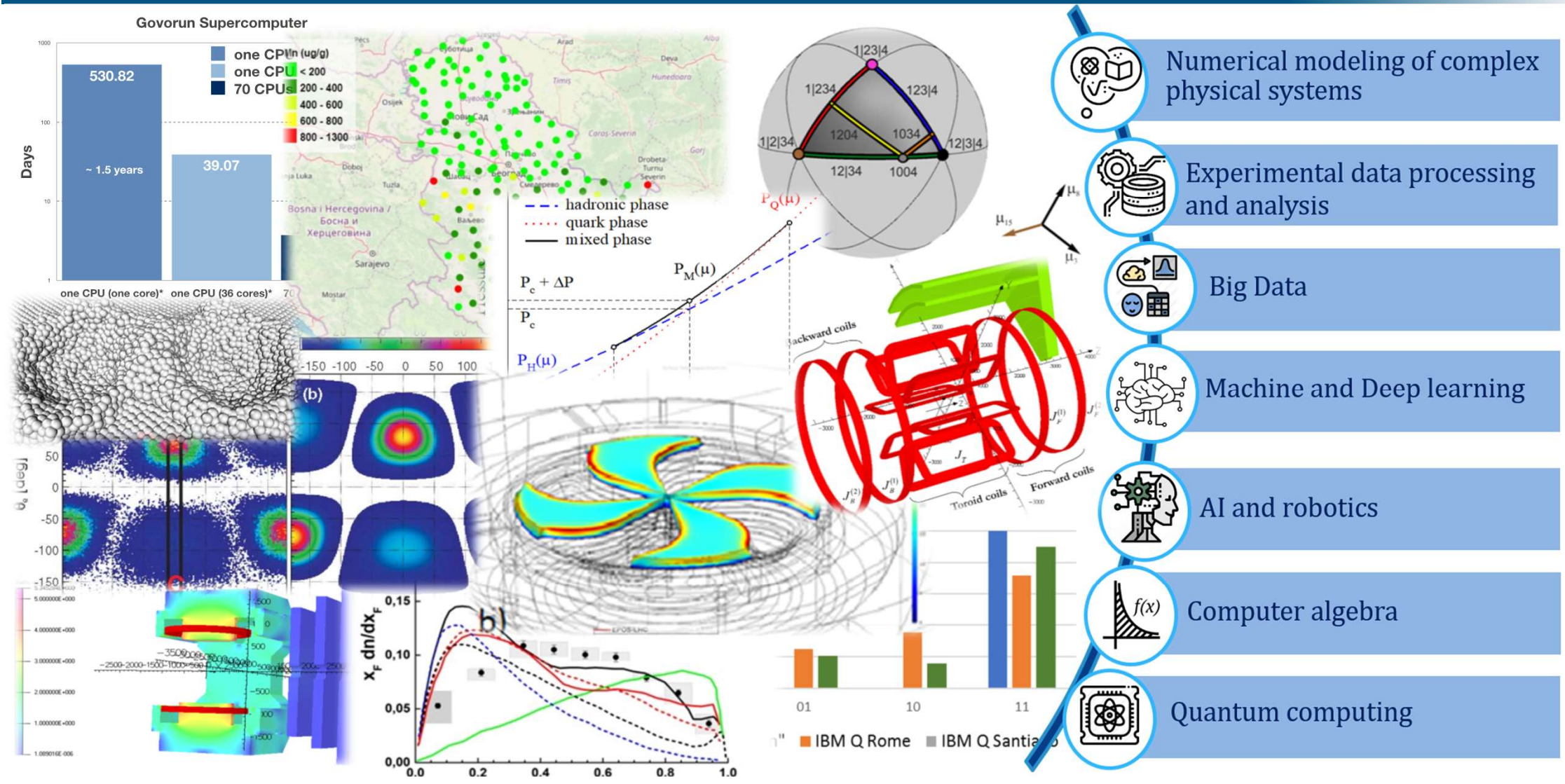
MICC Unified Resource Management System



The main objectives of the unified resource management system are:

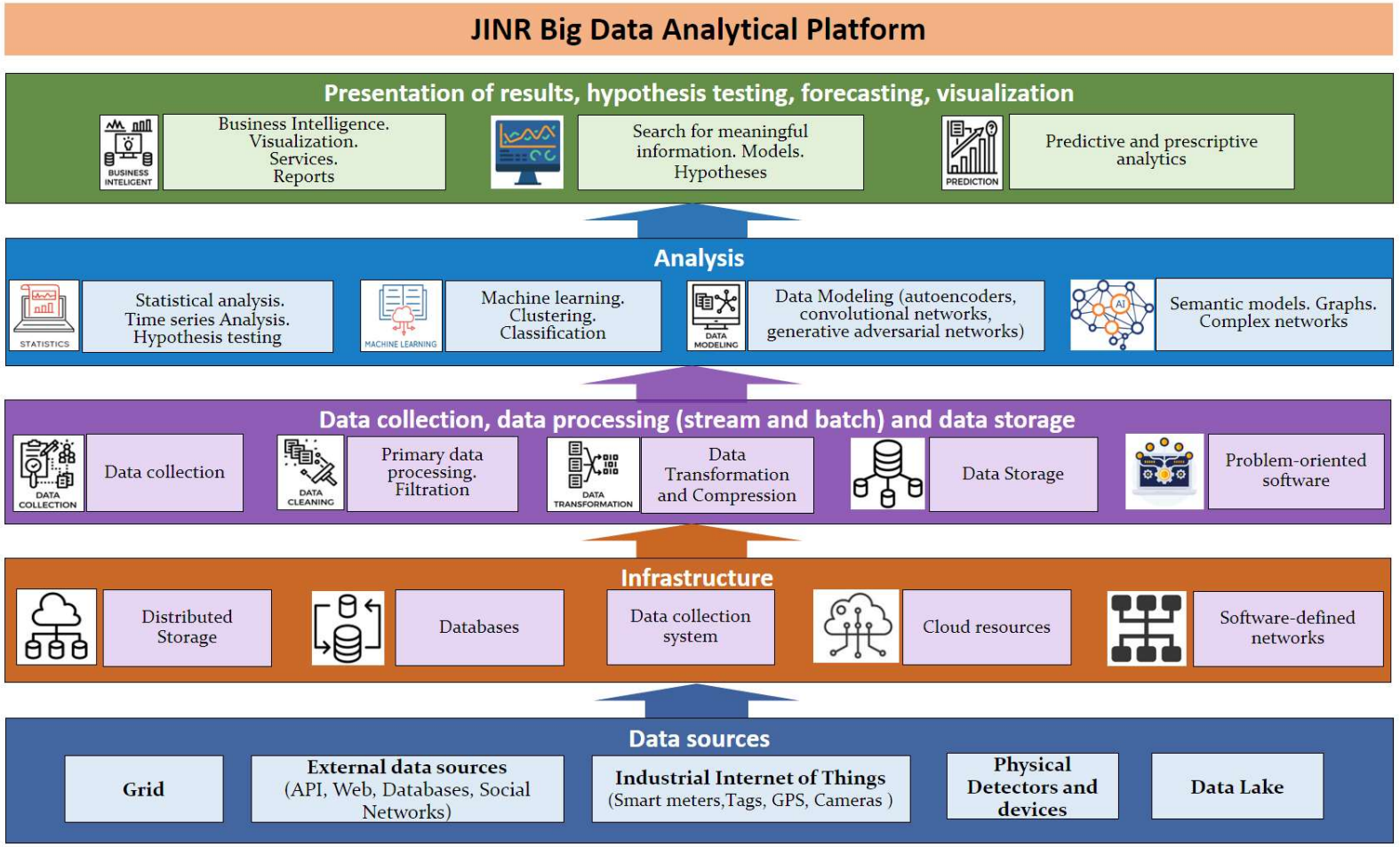
- to provide the ability to process large amounts of data
- to enable the organization of massive computing tasks
- to optimize the efficiency of using computing and storage resources
- to effectively monitor resource loading
- to consolidate resource accounting
- to provide a unified interface for accessing resources

Methods, Algorithms and Software



JINR Big Data Analytical Platform

- Bringing best of Big Data approaches to JINR practices
- Providing the Big Data infrastructure for users



Development of the system for training and retraining IT specialists

Training courses, master classes and lectures

MLIT staff and leading scientists from JINR and its Member States

Leading manufacturers of modern computing architectures and software

Parallel programming technologies

OpenMP

MPI

OpenCL

NVIDIA CUDA

Tools for debugging and profiling parallel applications

Intel Cluster Studio

Work with applied software packages

COMSOL MULTIPHYSICS

Wolfram Mathematica

ROOT Data Analysis Framework

MATLAB

GEANT4

Maple

Frameworks and tools for ML/DL tasks

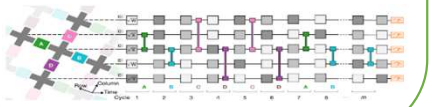
jupyterhub

TensorFlow

NumPy

scikit learn

Quantum algorithms, quantum programming and quantum control



The International Conference "Distributed Computing and Grid Technologies in Science and Education"

- Distributed computing systems
- Computing for MegaScience Projects
- Distributed computing applications
- Data Management, Organisation and Access
- HPC
- Virtualization
- Big data Analytics and Machine learning
- Research infrastructure

The International Symposium Nuclear Electronics and Computing

- Detector & Nuclear Electronics
- Triggering, Data Acquisition, Control Systems
- Distributed Computing, GRID and Cloud Computing
- Machine Learning Algorithms and Big Data Analytics new!
- Research Data Infrastructures
- Computations with Hybrid Systems (CPU, GPU, coprocessors)
- Computing for Large Scale Facilities (LHC, FAIR, NICA, SKA, PIC, XFEL, ELI, etc.)
- Innovative IT Education

MATHEMATICAL MODELING AND COMPUTATIONAL PHYSICS

- ❑methods, software and program packages for data processing and analysis;
- ❑mathematical methods and tools for modeling complex physical and technical systems, computational biochemistry and bioinformatics;
- ❑methods of computer algebra, quantum computing and quantum information processing;
- ❑ machine learning and big data analytics;
- ❑ algorithms for parallel and hybrid calculations.

MLIT Schools

JINR School of Information Technology 2022

60 students from 13 universities

- Dubna State University
- Far Eastern Federal University
- National Research Nuclear University MEPhI
- North Ossetian State University after K.L. Khetagurov
- Plekhanov Russian University of Economics
- St. Petersburg University
- The Bauman Moscow State Technical University
- The National University of Science and Technology (MISIS)
- The Peoples' Friendship University of Russia
- Tomsk Polytechnic University
- Tula State University
- Tver State University
- Vitus Bering Kamchatka State University

JINR School of Information Technology 2022

60 students from 13 Russian universities

SCHOOL JINR

JINR School of Information Technology 2023

50 students from 11 Russian universities

Joint Institute for Nuclear Research
Meshcheryakov Laboratory of Information Technologies

10 GRID2023

3-7 July 2023

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

More than **275** participants
In person - 216
Remotely - 60

30 Plenary reports
135 Sessional reports

17 Countries: Azerbaijan, Armenia, Belarus, Bulgaria, the Czech Republic, Egypt, Germany, Georgia, Iran, Kazakhstan, Mexico, Moldova, Mongolia, Serbia, CERN and Uzbekistan. **Russia** was represented by participants **from 41 universities and research centers**.

Conference Topics:

1. Distributed Computing Systems
2. HPC
3. Distributed Computing and HPC Application
4. Cloud Technologies
5. Computing for MegaScience Projects
6. Quantum Informatics and Computing
7. Big Data, M/D Learning, Artificial Intelligence
8. Student session

Workshop "Computing for radiobiology and medicine"

Workshop "Modern approaches to the modeling of research reactors, creation of the "digital twins" of complex systems"

Round table "RDIG-M - Russian distributed infrastructure for large-scale scientific projects in Russia"

Round table on IT technologies in education