INTERNATIONAL INTERGOVERNMENTAL ORGANIZATION
МЕЖДУНАРОДНАЯ МЕЖПРАВИТЕЛЬСТВЕННАЯ ОРГАНИЗАЦИЯ

JOINT INSTITUTE FOR NUCLEAR RESEARCH
ОБЪЕДИНЕННЫЙ ИНСТИТУТ ЯДЕРНЫХ ИССЛЕДОВАНИЙ


Laboratory of Information Technologies

# JINR Multifunctional Information and Computing Complex: Status and Perspectives

KORENKOV Vladimir, DOLBILOV Andrey, MITSYN Valery, STRIZH Tatiana, KUTOVSKIY Nikolay, PODGAINY Dmitry, STRELTSOVA Oksana, TROFIMOV Vladimir

Joint Institute for Nuclear Research, Dubna

GRID-2018, Dubna, 10 September 2018

# JINR Laboratory of Information Technologies

Laboratory of Information Technologies of the Joint Institute for Nuclear Research in Dubna was founded in August 1966.

The main directions of the activities at the Laboratory are connected with the provision of networks, computer and information resources, as well as mathematical support of a wide range of research at JINR in high energy physics, nuclear physics, condensed matter physics, etc.

M.G. Mescheryakov

N.N. Govorun

## Computing is an integral part of theory, experiment, technology development

Many recent successes only possible because of significant community effort to develop and advance the necessary computing tools!

# LIT Fundamentals

- Provide IT services necessary for the fulfillment of the JINR Topical Plan on Research and International Cooperation in an efficient and effective manner
- Building world-class competence in IT and computational physics
- 24/7 support of computing infrastructure and services such availability is called nonstop service

- User policies
- Computing infrastructure
- Mathematical and software support
- Corporative information system
- Training, education and user support

**IT-infrastructure is the one of JINR basic facilities**

# IT-services

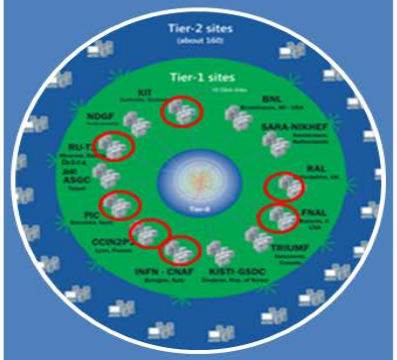
| | |
|---|---|
| **Network** | Telecommunication channels, JINR LAN, JINR IXP, JINR LAN Remote Access, Datacenter Network, Device registration, DHCP, DNS, IPDB, Network Registration & Connection, Network Monitoring, Technical Network, WIFI, LHCONE, LHCOPN |
| **Primary** | Account Management, JINR Certificate Authority, Computer Security Controls, Security Firewall, Single Sign On, SSH (Secure SHell), E-mail, Resources, Portal |
| **Collaboration** | Audio Conferencing, Eduroam, Indico, Video Conferencing, Webcast and Recording, Project Management, GIT, JINR Disk, JINR Document Server (Invenio) |
| **Database Services** | Administration Database Service, ADB2, ISS, 1.C EPR, EDH “Dubna”, General Purpose Database Service |

**Computer Science &Physics Computing**

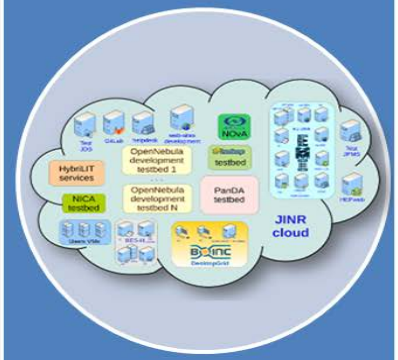
| | |
|---|---|
| **Development** | **Research** |
| **Computing** | **Applied software** |

| | | | | | | |
|---|---|---|---|---|---|---|
| **GRID Tier1 Tier2** | **CLOUD** | **HPC** | **JINRLIB** | **Quantum computing** | **Software support** | **BigData analytics** |

# Multifunctional Information and Computing Complex
## Main components




**Tier1:**
4720 cores
8.3 PB disk
9 PB tape




**Cloud:**
1100 CPU
5.7 TB RAM
896 TB disk


**HPC Govorun**
Peak ~0.5 Pflops
**HybriLIT:**
~142 Tflops

**CICC/Tier2:**
4128 cores
2.7 PB disk
**NICA** off-line cluster and storage system



Network infrastructure: LAN: 10 Gbps WAN: 100 Gbps +2x10 Gbps

Engineering infrastructure

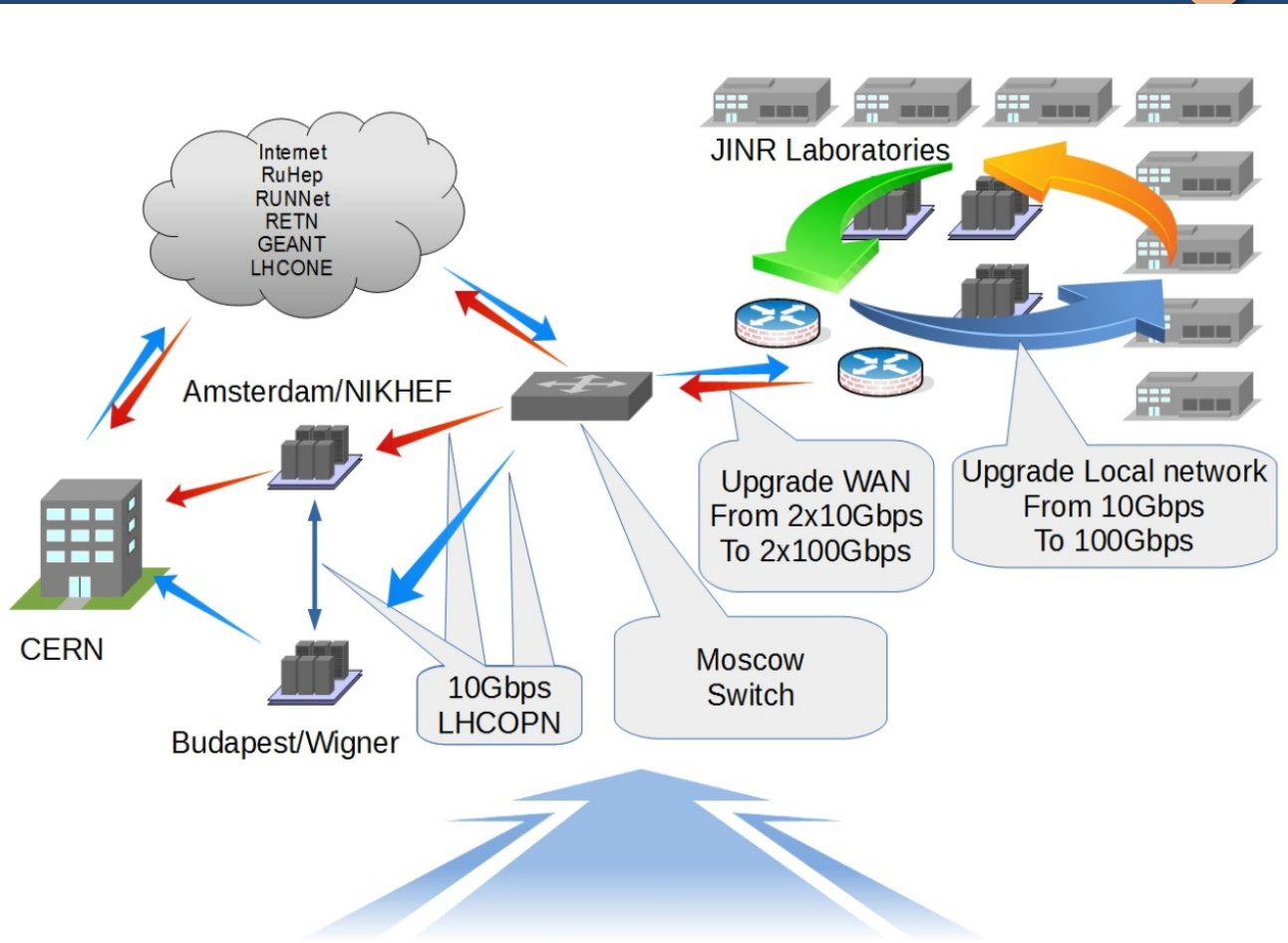
**LIT IT-infrastructure is the one of JINR basic facilities**

# Engineering infrastructure

The main results of developing the engineering infrastructure:

- Installation of two new transformers (2.5 MW)
- New cooling system for Tier-1 complex
- New water cooling system for supercomputer
- Installation of two power supply using a diesel generator
- Installation of four power supply systems (1.2 MW)

# Networking

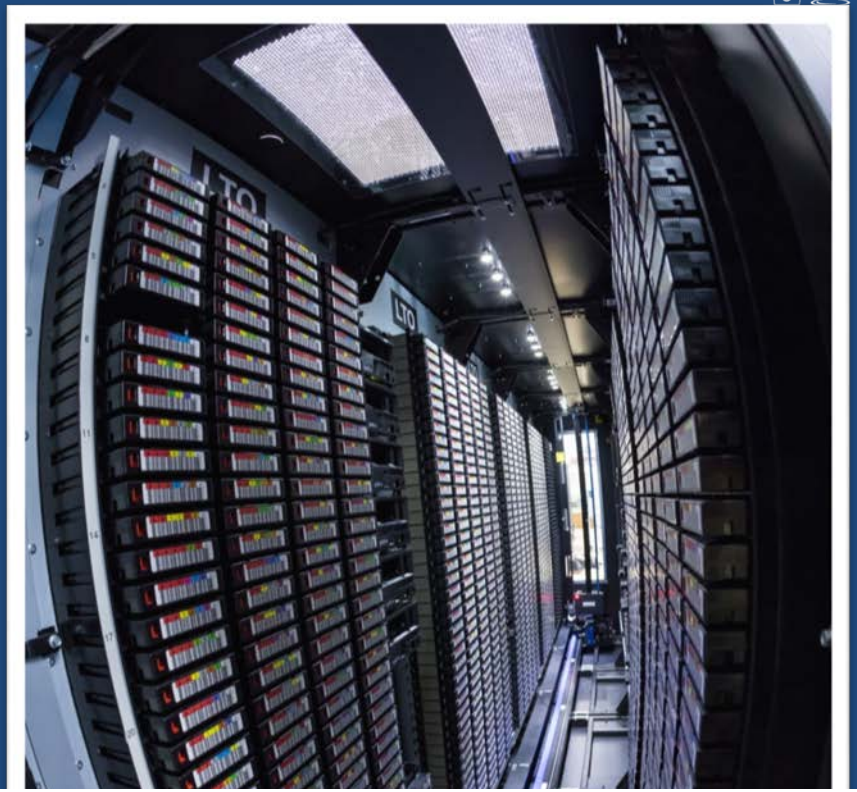



One of the most important components of JINR and MICC providing access to resources and the possibility to work with the big data is a network infrastructure.

**Users - 6524**
**IP addresses - 19805**
**EDUROAM - 389**
**E-library - 1573**
**AFS - 867**
**VOIP - 128**
**E-mail - 2550**
**2017 год: 40 PB**
**IN ~ 24 PB**
**OUT ~ 16 PB**

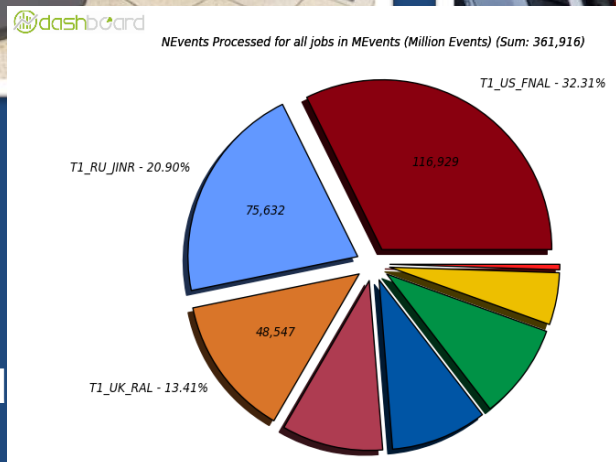
**Local Area Network – 2x10 Gbps, planned upgrade to 100 Gbps**
**Wide Area Network – 2x10 Gbps, 100 Gbps**
**Upgrade WAN to n x 100 Gbps planned**

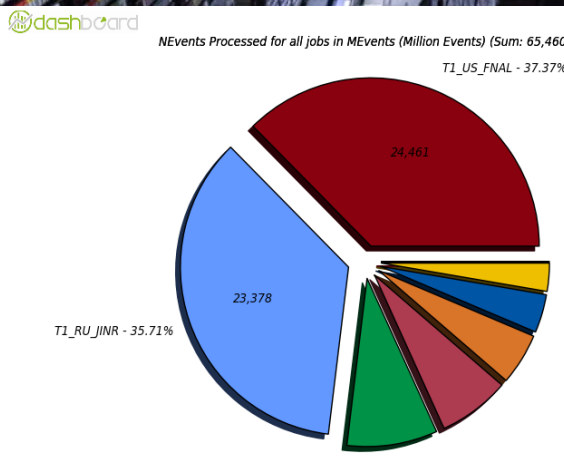
# JINR Tier1 for CMS



**Now**

- 4720 cores
- 9 PB tapes
- 8.3 PB disk
- 100% reliability and availability

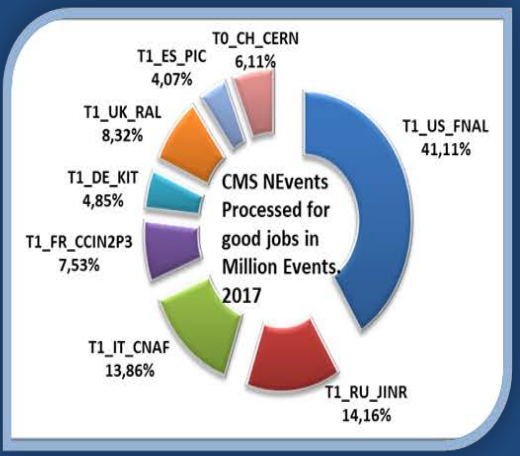


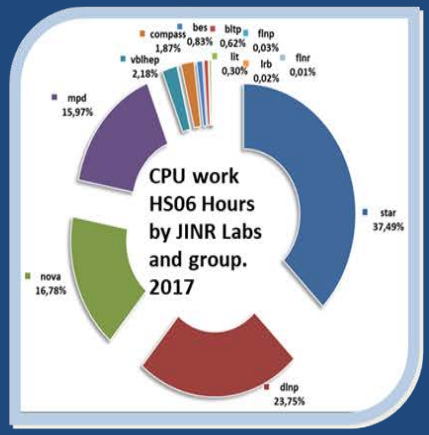
# JINR Tier2 site for LHC and other experiments

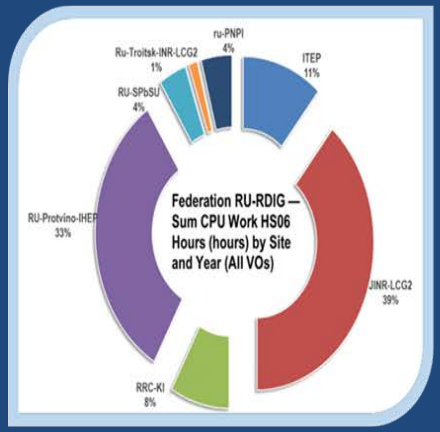
**Experiments@VO**

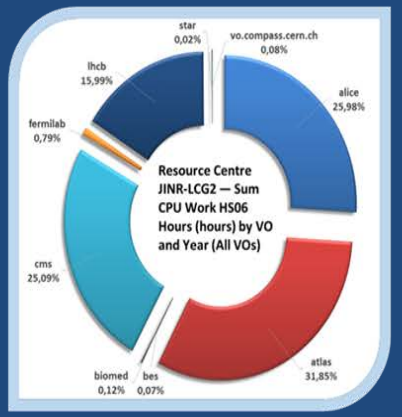
BM@N, MPD, CMS, ATLAS, ALICE, LHCb, COMPASS, PANDA, CBM,STAR, NOvA, BESIII, DIRAC,OPERA, NEMO, Mu2e, NUCLON, HONE, FUSION, BIOMED







# JINR Tier1 for CMS Resources

**Computing Elements (CE)**

***Worker Node (WN)***

Typically SuperMicro Blade

275 64-bit machines: 2 x CPU, 6-10 core/CPU

**Total: 4720 core/slots for batch**

**Storage Elements(SE)**

**Storage System: dCache**

Typically Supermicro and DELL

1st - Disk Only: **7.2 PB**

2nd - support Mass Storage System:**1.1PB**.

1 tape robot: IBM TS3500, 3440xLTO-6 data cartridges; 12xLTO-6 tape drives FC8, **9 PB**

***Software***

OS: Scientific Linux release 6 x86_64.

BATCH : Torque 4.2.10 (home made)

Maui 3.3.2 (home made)

CMS Phedex

dCache-3.2

Enstore 4.2.2 for tape robot.

# JINR Tier2 and CCIC Resources

**Computing Elements (CE)**

***Worker Node (WN)***

Typically SuperMicro Blade, SuperMicroTwin2, Dell FX

240 machines: 2 x CPU Xeon, 4-14/ CPU

**Total: 4128 core/slots for batch**

Storage Elements(SE)

Storage System: dCache

Typically Supermicro and DELL

1st - Disk for ATLAS & CMS 2070TB

2nd – Disk for ALICE, EOS 712TB

3-rd Disk for some EGI VOs & local users 147TB

***Software***

OS: Scientific Linux release 6 x86_64

BATCH: Torque 4.2.10 (home made)

Maui 3.3.2 (home made)

CMS Phedex

ALICE Vobox

dCache-3.2

EOS aquamarine

# From HybriLIT cluster to HybriLIT platform

**The supercomputer is a natural continuation of heterogeneous platform and leads to a significant increase in the performance of both CPU and GPU components.**

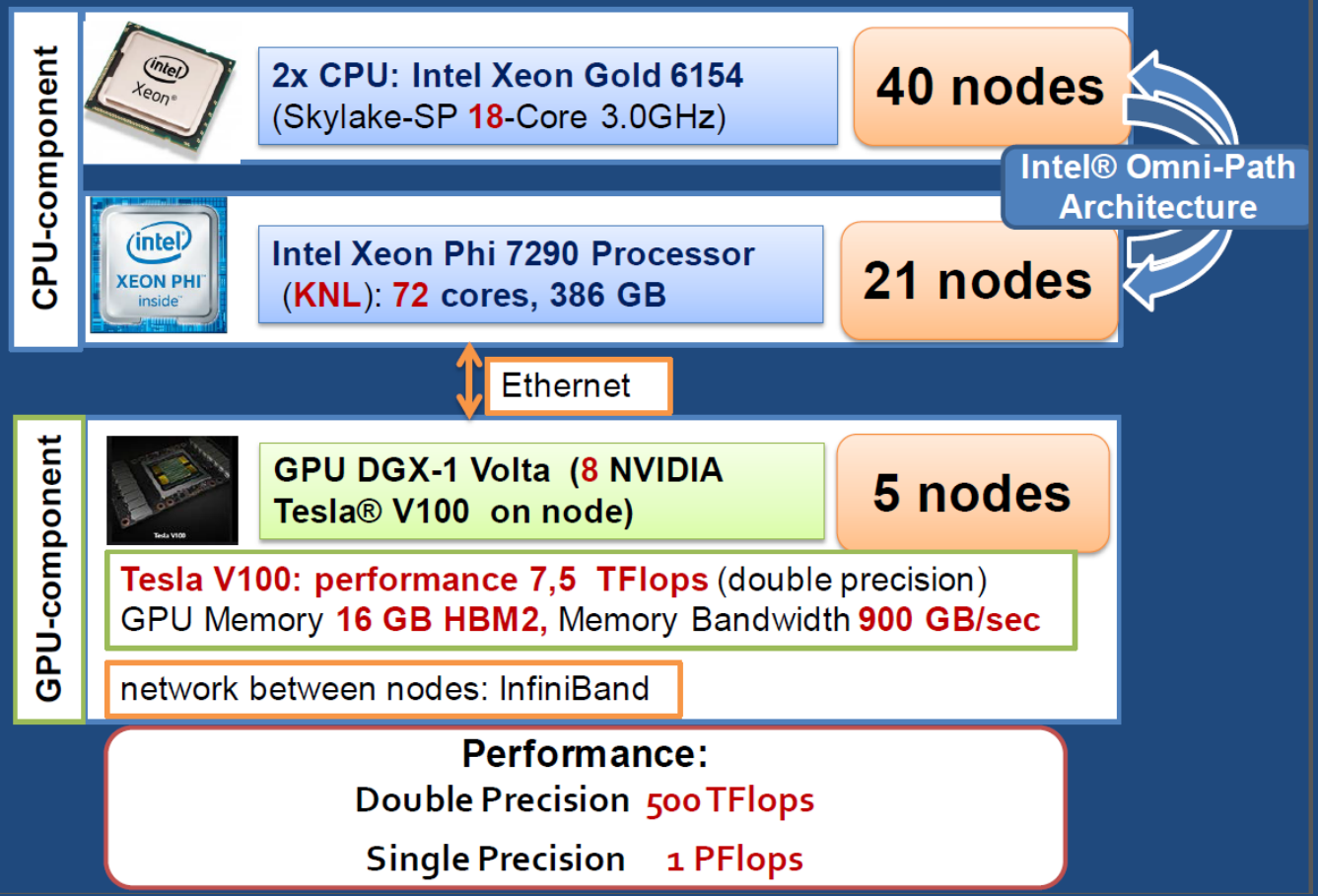
**2014**

**Total peak performance:**
**140 TFlops for single precision;**
**50 TFlops for double precision**

**2018**

**Total peak performance:**
**1000 TFlops for single precision;**
**500 TFlops for double precision**

# Supercomputer “GOVORUN”

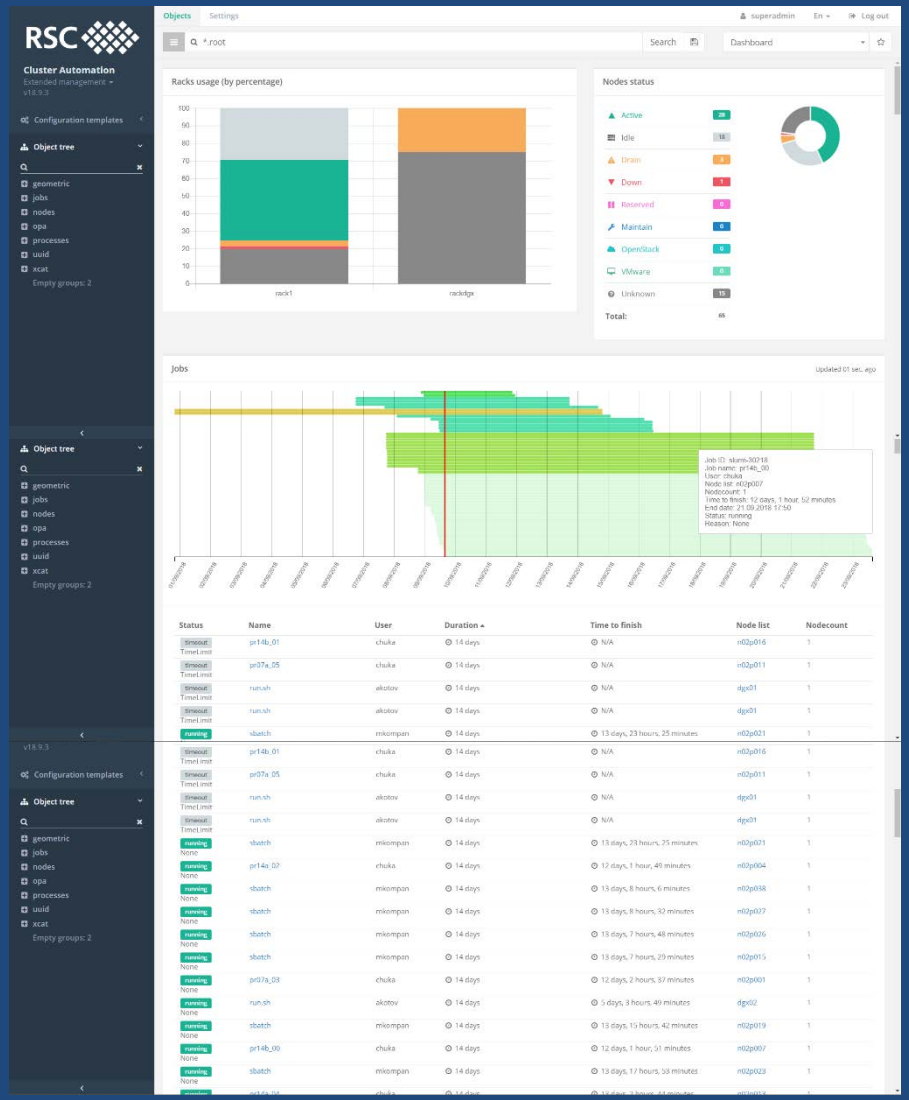


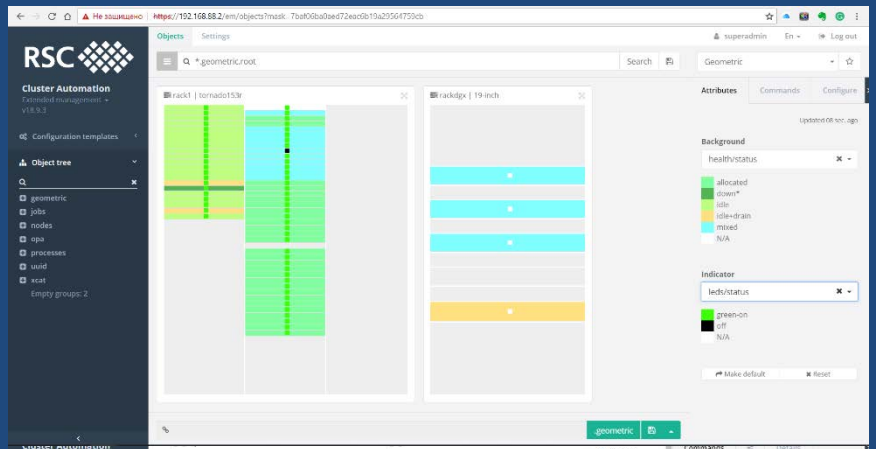

**CPU-component:**

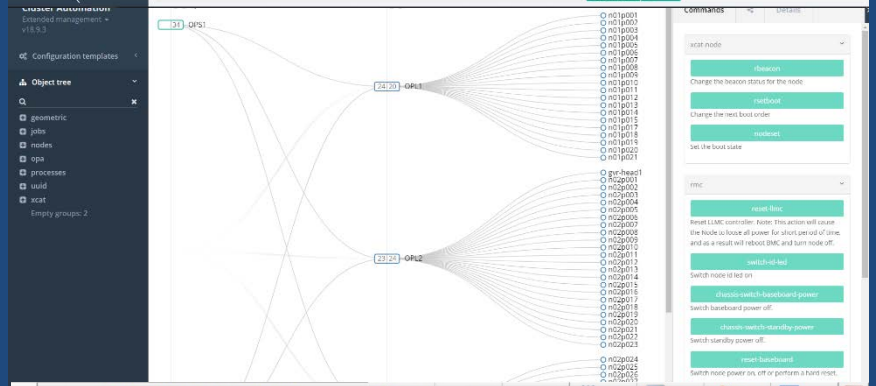
**«RSC Tornado» based on Intel® Xeon® Scalable and Intel® Xeon Phi™**

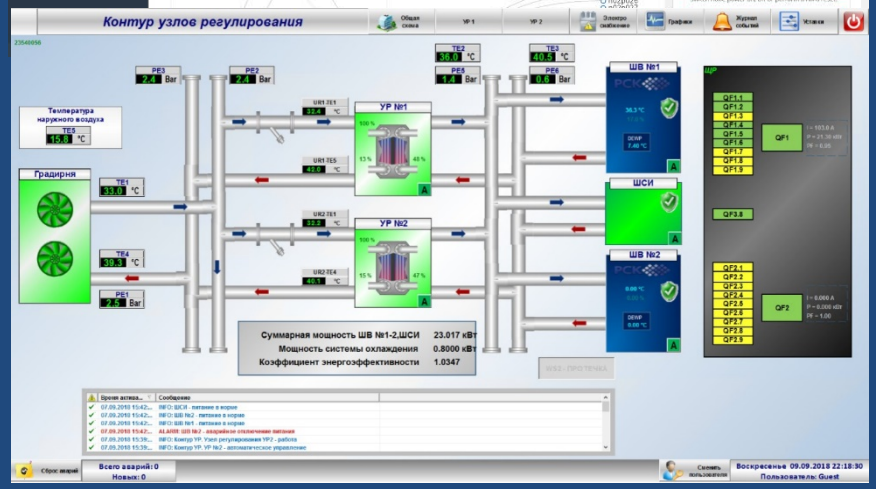
- Unique heterogeneous and hyper-converged system
- Multipurpose high performance system with direct hot liquid cooling of all system components
- The most energy-efficient system in Russian (PUE = 1.02)
- First 100% hot liquid cooling of Intel® Omni-Path interconnect

# Supercomputer Govorun. Management



## IO-500

This is the official list from ISC-HPC 2018. The list shows the best result for a given combination of system/institution/filesystem.

| # | information | | | | | | io500 | | |
|---|---|---|---|---|---|---|---|---|---|
| | system | institution | filesystem | storage vendor | client nodes | data | score | bw | md |
| | | | | | | | | GiB/s | kIOP/s |
| 1 | Oakforest-PACS | JCAHPC | IME | DDN | 2048 | zip | 137.78 | 560.10 | 33.89 |
| 2 | ShaheenII | KAUST | DataWarp | Cray | 1024 | zip | 77.37 | 496.81 | 12.05 |
| 3 | ShaheenII | KAUST | Lustre | Cray | 1000 | | 41.00* | 54.17 | 31.03* |
| 4 | JURON | JSC | BeeGFS | ThinkparQ | 8 | | 35.77* | 14.24 | 89.81* |
| 5 | Mistral | DKRZ | Lustre2 | Seagate | 100 | | 32.15 | 22.77 | 45.39 |
| 6 | Sonasad | IBM | Spectrum Scale | IBM | 10 | zip | 24.24 | 4.57 | 128.61 |
| 7 | Seislab | Fraunhofer | BeeGFS | ThinkparQ | 24 | | 16.96 | 5.13 | 56.14 |
| 8 | Mistral | DKRZ | Lustre1 | Seagate | 100 | zip | 15.47 | 12.68 | 18.88 |
| 9 | Govorun | Joint Institute for Nuclear Research | Lustre | RSC | 24 | zip | 12.08 | 3.34 | 43.65 |
| 10 | EMSL Cascade | PNNL | Lustre | | 126 | | 11.12 | 4.88 | 25.33 |
| 11 | Serrano | SNL | Spectrum Scale | IBM | 16 | | 4.25* | 0.65 | 27.98* |
| 12 | Jasmin/Lotus | STFC | PanFS | Panasas | 64 | zip | 2.33 | 0.26 | 20.93 |

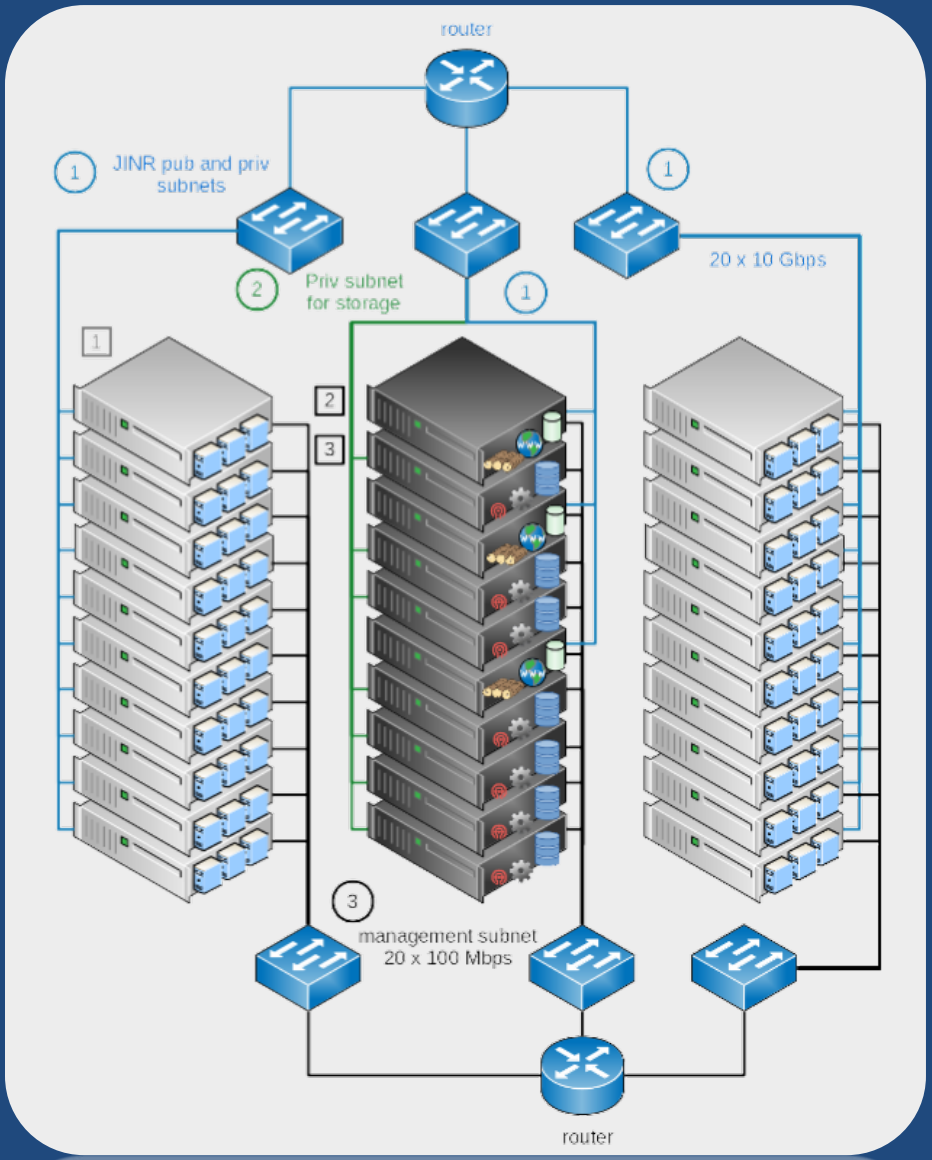
*Values with* indicate that a value for the computation was missing.*

Further lists with more detail can be found on the navigation menu. For example, a list with radar chart and configurable scoring.


РСК


РСК

# JINR cloud




**Purpose:**

- increase the efficiency of hardware and proprietary software utilization,
- improve IT-services management.

**Implementation:**

- Cloud platform: OpenNebula
- Two types of virtualization: OpenVZ containers (linux only) and KVM VM (any OS)
- Storage back-end for KVM VM images: ceph block-device
- User interfaces: web GUI and command line interface
- Authentication in the cloud web-GUI : JINR central user database (LDAP+Kerberos)
- VM/CT access: rsa/dsa-key or Kerberos credentials

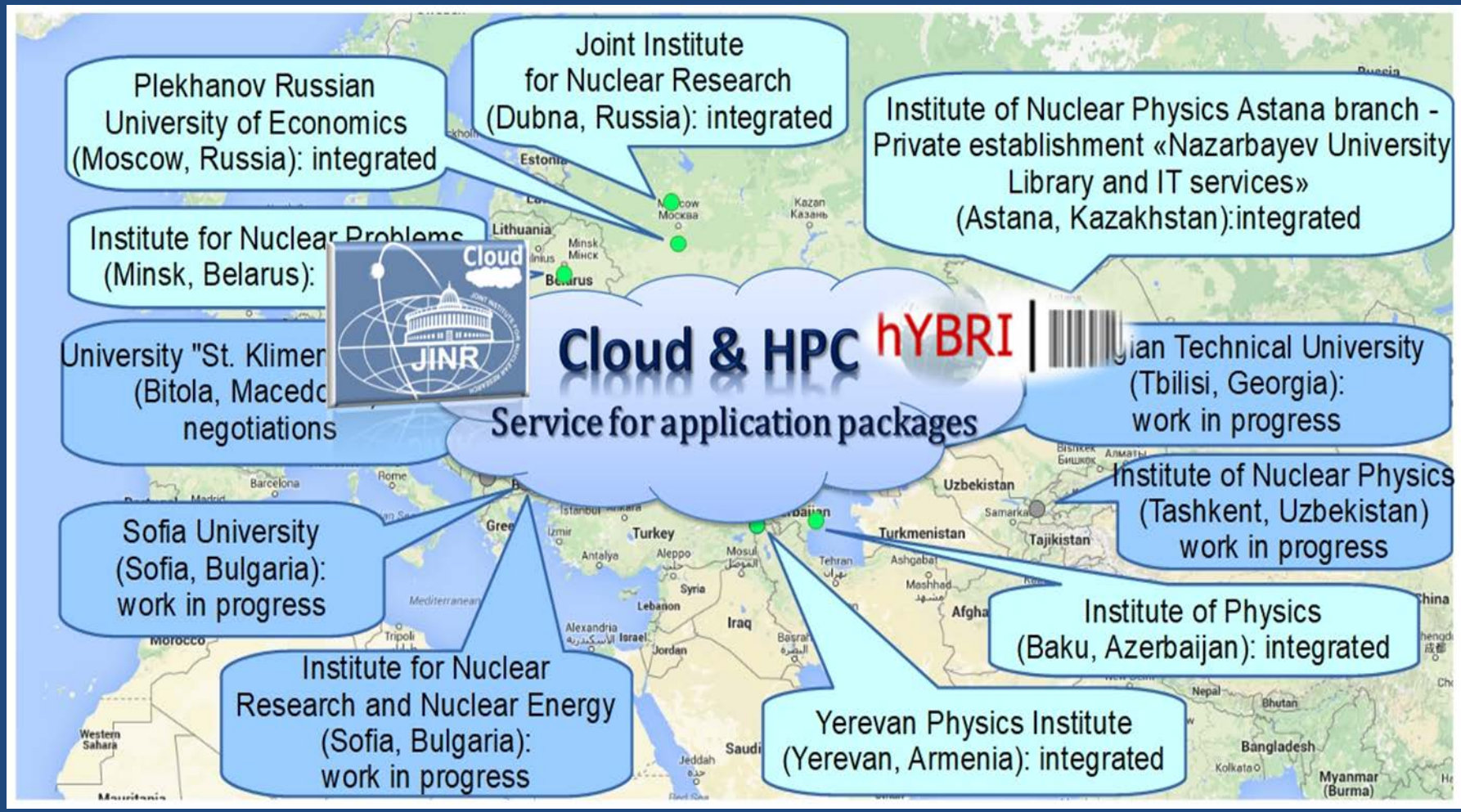
**Resources:** ~1100 cores, ~5.7 TB RAM, 896 TB disk storage

**Utilization:**

- VMs&CTs for JINR users
- Computational resources for Baikal-GVD, BESIII, Daya Bay, JUNO, NOvA experiments
- Testbeds for development and R&D in IT

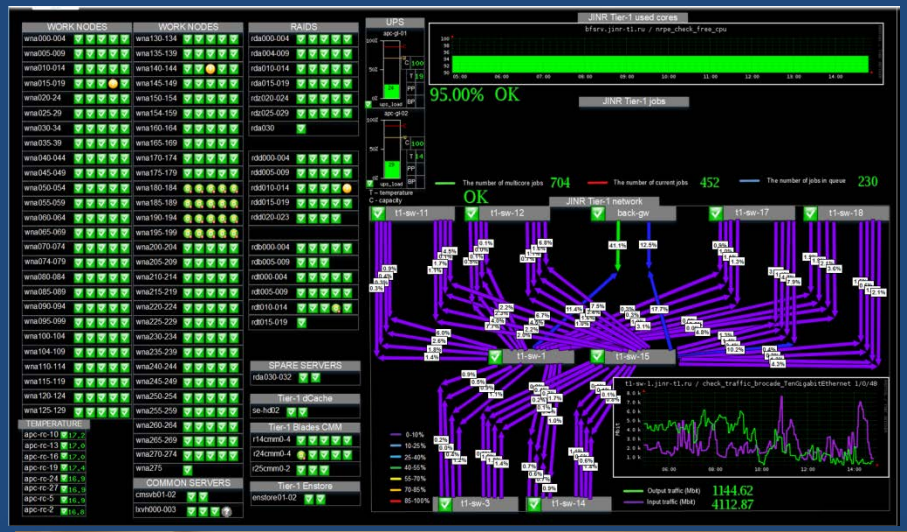
# JINR Cloud + HPC

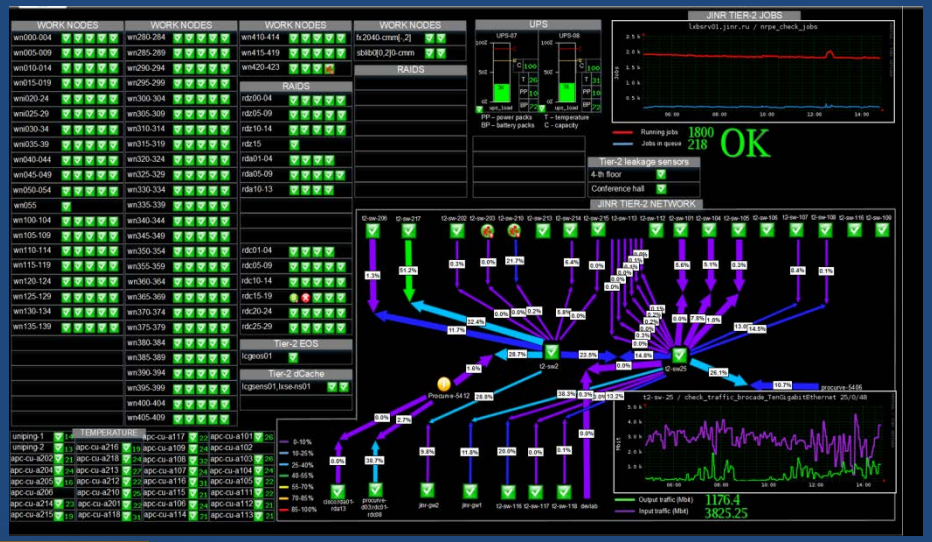
**New challenge – integration JINR Member States clouds and the Supercomputer into unified distributed computational infrastructure**



# MICC Monitoring

For a robust performance of the complex it is necessary to monitor the state of all nodes and services - from the supply system to the robotized tape library.




The system allows one, in a real time mode, to observe the whole computing complex state and send the system alerts to users via e-mail, sms, etc. ~ 850 elements are under observation
~ 8000 checks in real time
~ 100 scripts

# Examples of participation of LIT JINR in some important experiments and megaprojects

# JINR activity at WLCG project

We have two WLCG Data Centers at the JINR Laboratory of Information Technologies:

- Tier1 (T1_RU_JINR) provides one of 7 dedicated real-time computing facilities for experimental data from the CMS detector at the CERN
- Tier2 (JINR-LCG2) provides one of more then 170 dedicated real-time computing facilities for experimental data from the ATLAS, ALICE, CMS, LHCb detectors, and other VO from RDIG community

We participate in some other WLCG activities:

- Software development for ATLAS, ALICE, CMS
- WLCG Dashboard (WLCG Google Earth monitoring, Global data transfer monitoring, Local and global Monitoring of Tier3 centers, xROOTd Monitoring,…)
- NoSQL data storage integration (Hadoop, ElasticSearch,…)
- Integration of Grid, Clouds and HPC
- PanDA WMS Development (Pilot2, Harvester), Compass Production System
- GENSER & MCDB
- DataLakes

# CHALLENGES: distributed data storage evolution: DATALAKES

**GOAL:**

•to provide a computing infrastructure to the experiments and the community to store and analyze data,

•to achieve storage consolidation where geographically distributed storage centers (potentially deploying different storage technologies) are operated and accessed as a single entity.

EOS - a CERN open-source storage software solution to manage multi PB storage.

XRootD - core of the implementation framework providing a feature-rich remote access protocol.

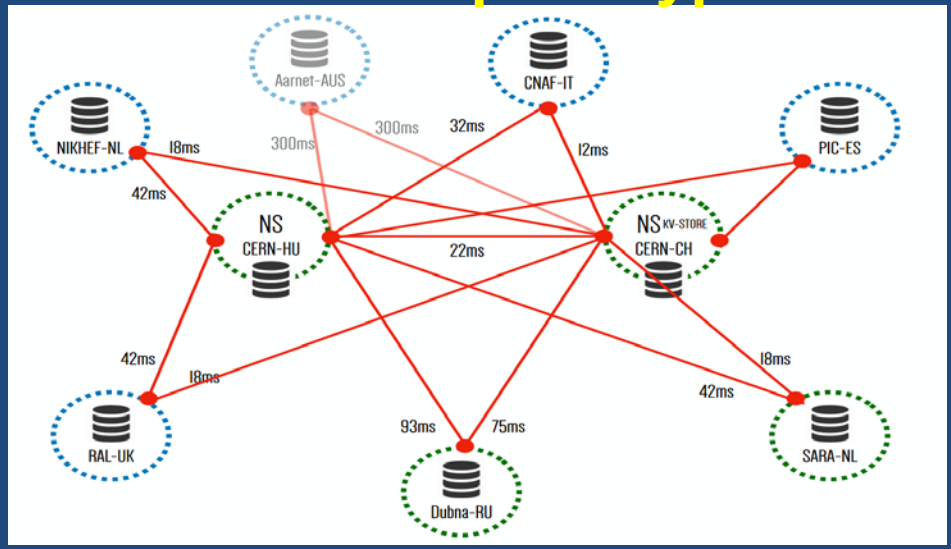
Improvement of already existing production quality Data Management services.

Scalable technologies for federating storage resources and managing data in highly distributed computing environments.

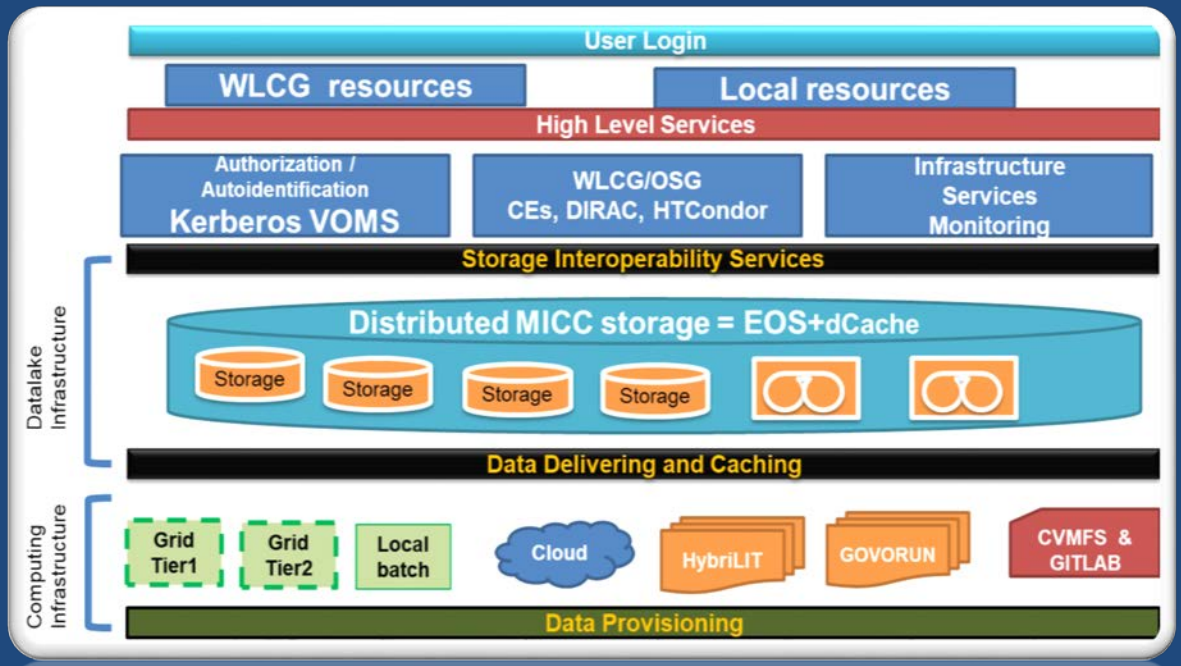
# JINR in Data(Lakes)

## eulake prototype

1. We start the work in collaboration in WLCG eulake R&D . Prototype is based on EOS Distributed storage endpoints at SARA, NIKHEF, RAL, JINR DUBNA, Kurchatov, PIC, CNAF and Aarnet.




2. As a next step we start to develop the common EOS based data storage for MICC components.
Total space: 3740TB User space: 1870TB (2 replicas)



# New COMPASS Production System

**COMPASS Grid Production System** was developed and provides automated task processing from definition till archiving

**Key features:**

Production management via Web UI, allows one to define a task, send, follow and manage it during lifecycle.

Via **PanDA** job execution layer jobs are being sent to any available type of computing resource: **Condor, LSF, PBS, etc.**

**Computing sites:** CERN Tier1, JINR Tier2, BlueWaters HPC

**Storage:** EOS and CASTOR at CERN

**All management services deployed at JINR Cloud Service**

In production since August, statistics:

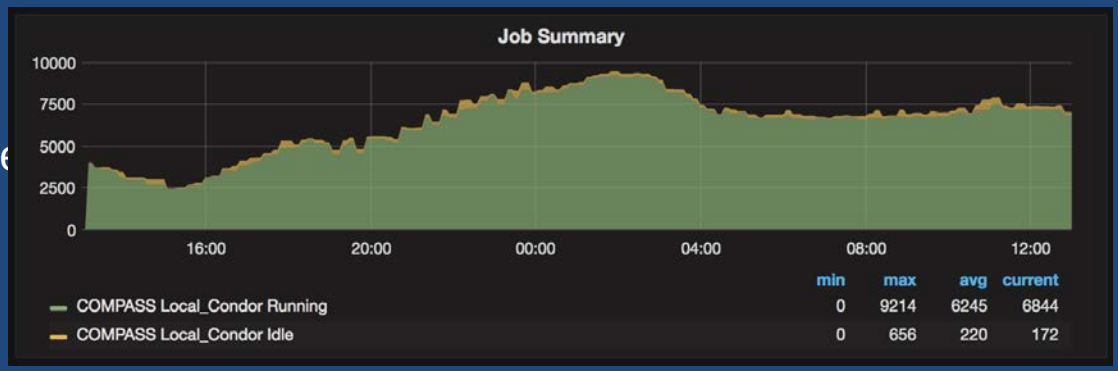
- ~2 000 000 chunks of raw data processe
- ~60 billions of events,
- ~400 TB of data produced,
- ~4 000 000 jobs processed: reconstruction, ddd filtering, merging of mDST, histograms and event dumps.

**Tasks requests layer: Web UI**

**Task definition layer: ProdSys**

**Job definition layer: ProdSys**

**Job execution layer: PanDA**



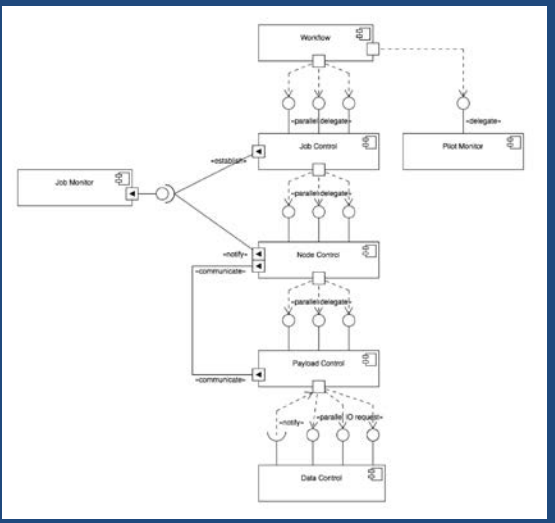

Number of COMPASS jobs processed simultaneously on 2018-06-02

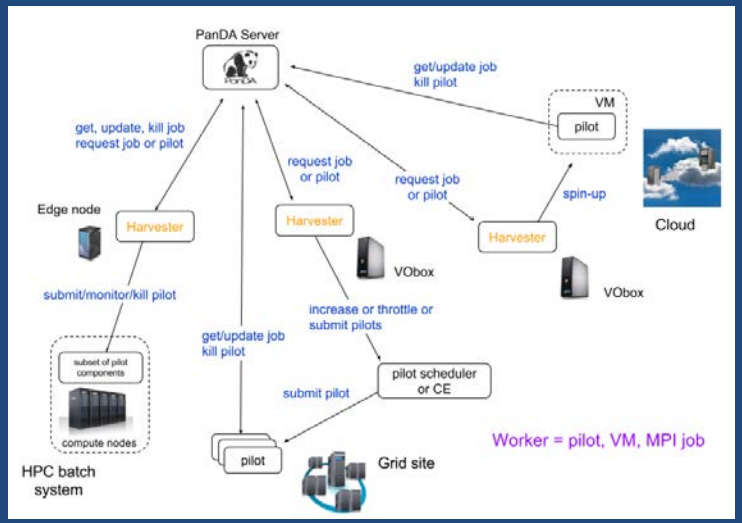
# New PanDA applications

The collaborative work on next-generation PanDA applications is in progress.

## Pilot 2

The PanDA Pilot has been used by ATLAS and other experiments for well over a decade. To meet the demands of extending PanDA beyond grids and ATLAS, the original Pilot is being reimplemented.
More details can de found on poster: ***«The next generation PanDA Pilot for and beyond the ATLAS experiment»*** *(10 Jul 2018, 15:00)*




## Harvester

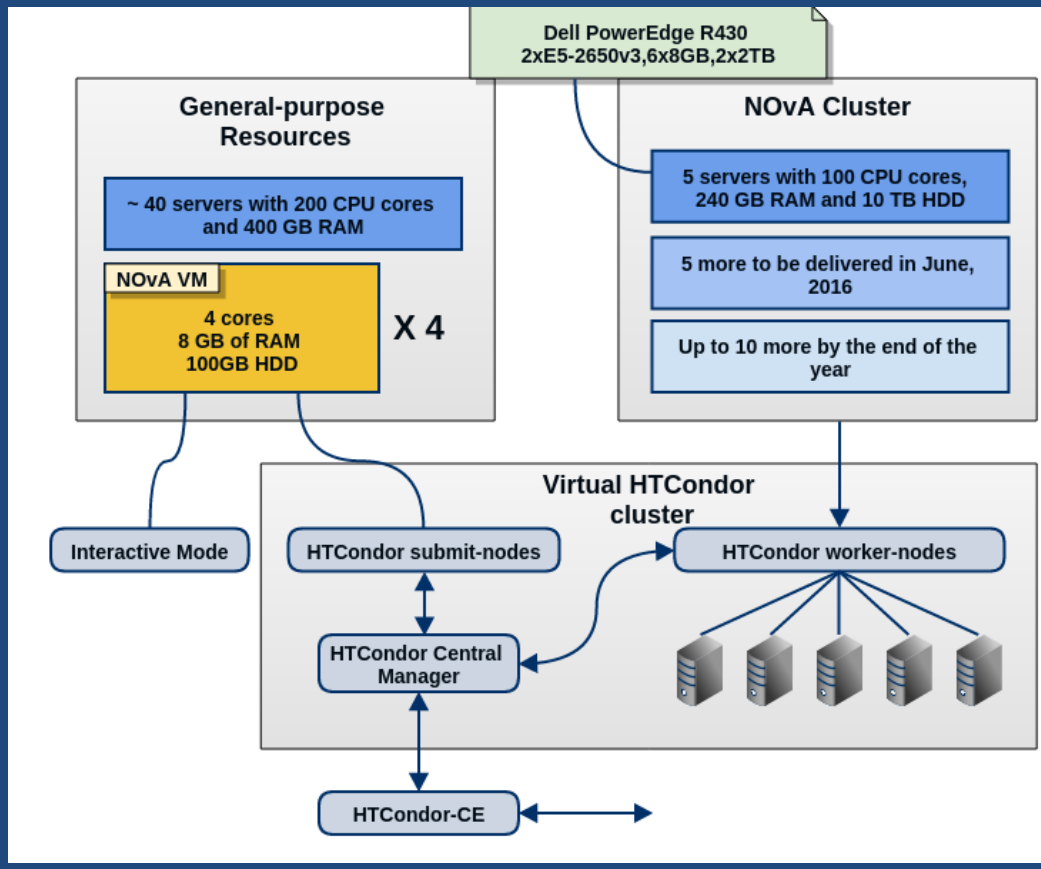
Harvester is a resource-facing service between the PanDA server and collection of pilots for resource provisioning and workload shaping.
More details in talk: ***«Harvester : an edge service harvesting heterogeneous resources for ATLAS»*** *(12 Jul 2018, 10:15, Hall 7 (National Palace of Culture))*

# Computing Support for Neutrino Projects

NOvA (Fermilab, USA) is the first neutrino experiment actively using JINR Cloud:

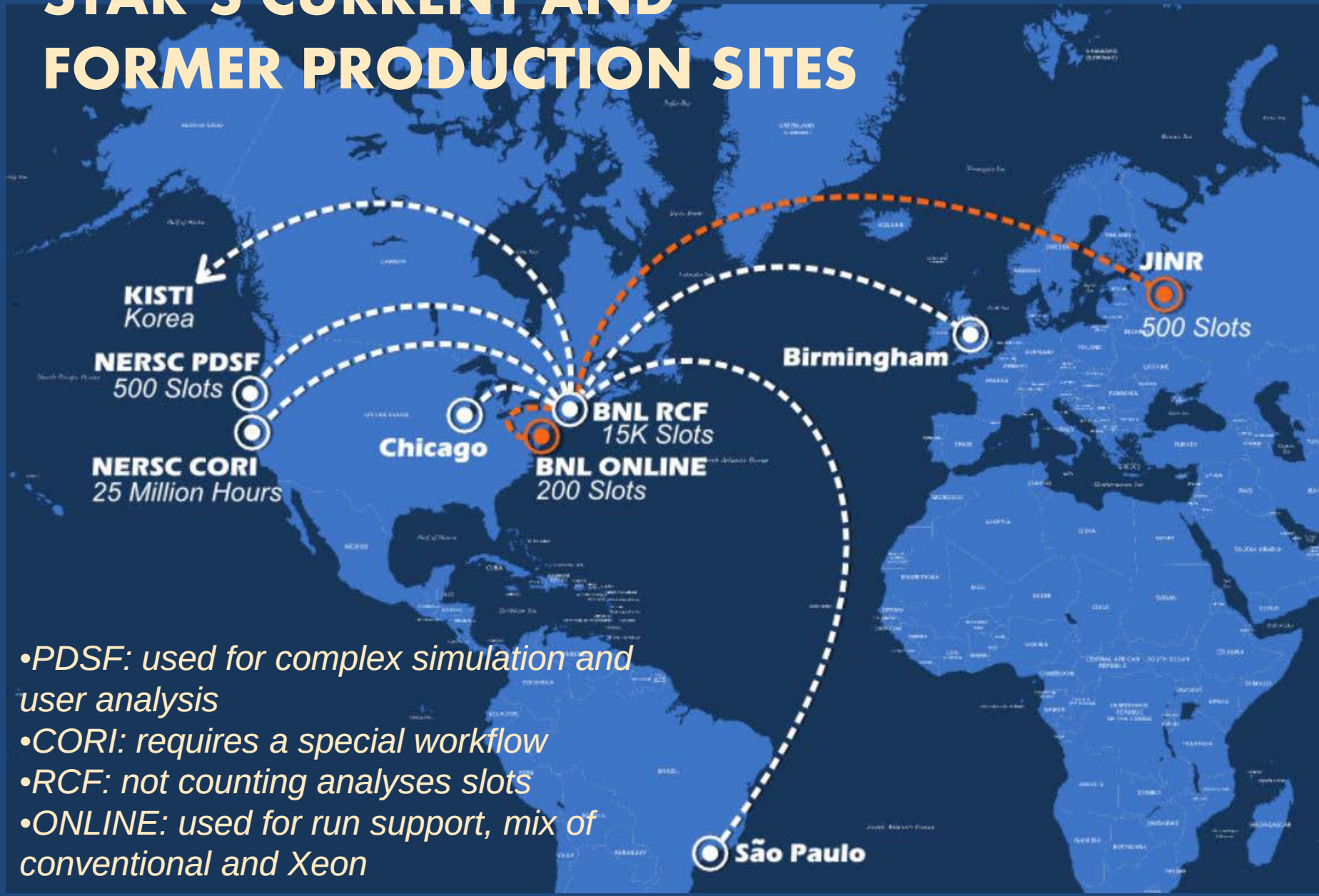
- ✓ Computing support team was formed including physicists and IT specialists

These resources may also be used by other future experiments at Fermilab, such as DUNE and mu2e.




Reactor neutrino experiments Daya Bay and JUNO also showed its interest in using JINR cloud resources. At the moment the experiments' tasks and required computing capacities are being discussed.

# STAR'S CURRENT AND FORMER PRODUCTION SITES

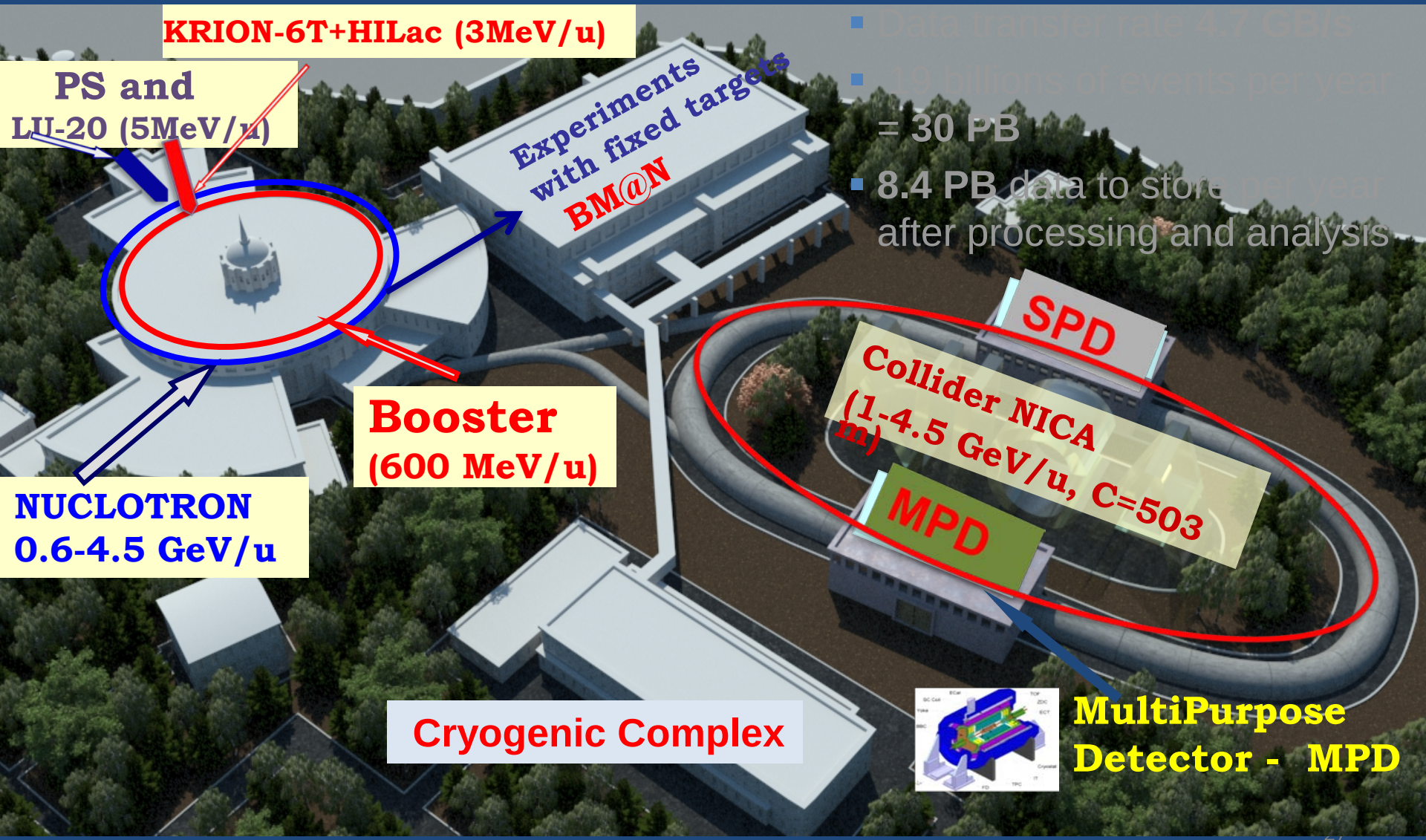



- *PDSF: used for complex simulation and user analysis*
- *CORI: requires a special workflow*
- *RCF: not counting analyses slots*
- *ONLINE: used for run support, mix of conventional and Xeon*

# NICA Complex: *New era in the hot dense matter science*

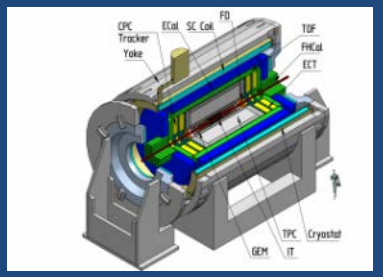
**Collider basic parameters:**

$\sqrt{s_{NN}}$ = 4-11 GeV; *beams: from* p to Au; L~$10^{27}$ cm$^{-2}$ c$^{-1}$ (Au), ~$10^{32}$ cm$^{-2}$ c$^{-1}$ (p)

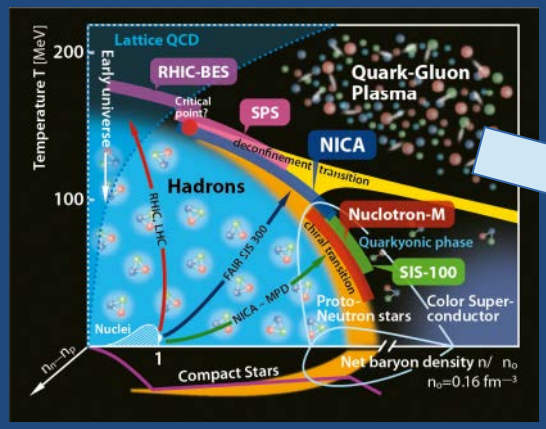



- Data transfer rate 4.7 GB/s
- 19 billions of events per year = 30 PB
- 8.4 PB data to store per year after processing and analysis

# NICA computing challenge

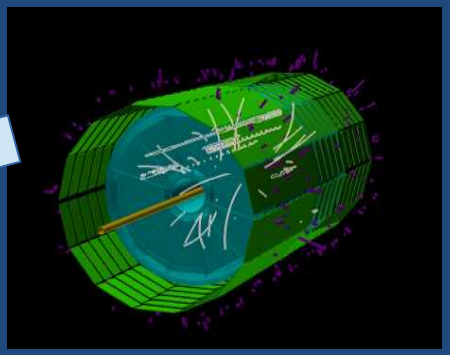



MPD experiment

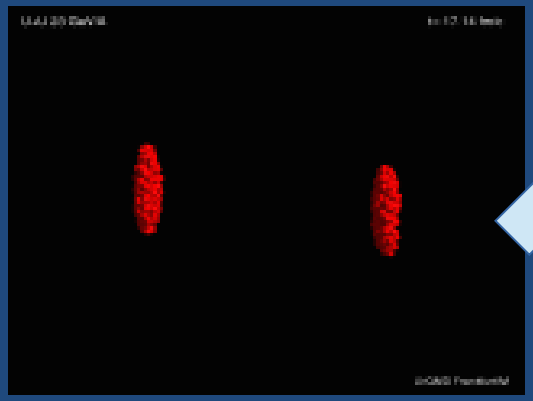



QCD phase diagram

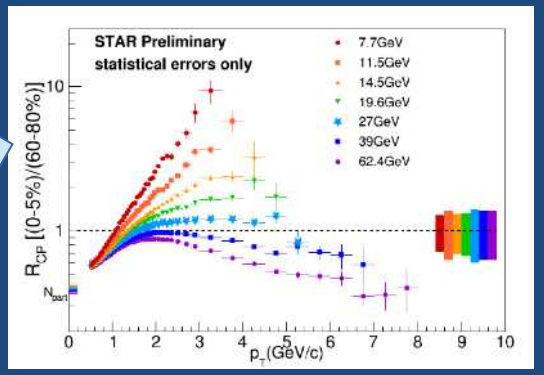
GOVORUN



Events reconstruction

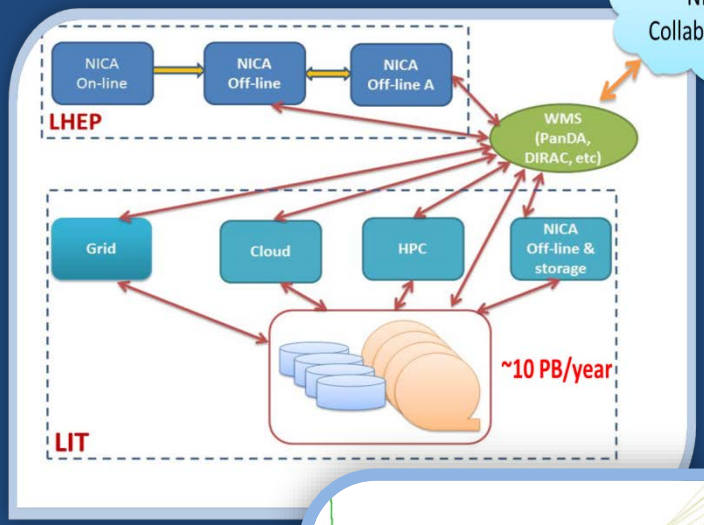


Simulations




Physics analysis

# Computing for MPD-BM&N




## GOAL:

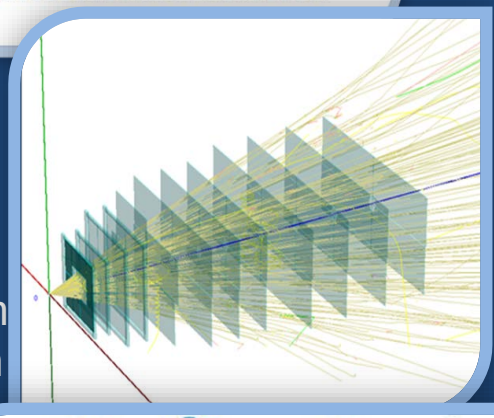
integration of various computing resources from NICA VBLHEP and MICC LIT for the development of computer infrastructure of the NICA.

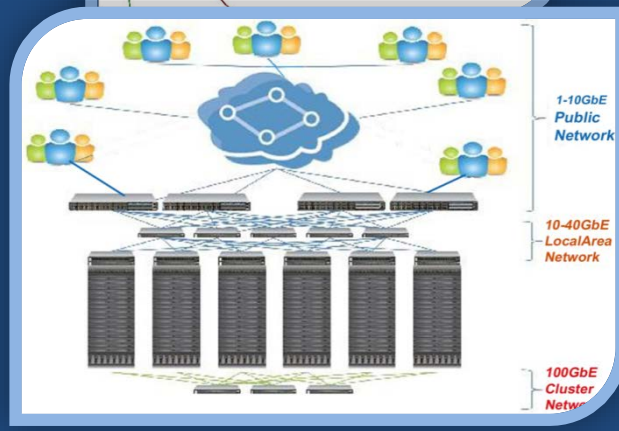
## REQUIREMENTS:

reliability, scalability, interoperability, adaptability to new technical solutions, year-round operation in 24x7 mode

## CURRENT DEVELOPMENT:

- ✓ maintenance and development of a management system for the NICA project
- ✓ local network of the NICA project
- ✓ optimization of Distributed Data Processing System using simulation
- ✓ software for BM@N GEM (Gas Electron Multiplier) tracker based on deep learning methods
- ✓ Nuclotron beam momentum estimation in BM@N experiment





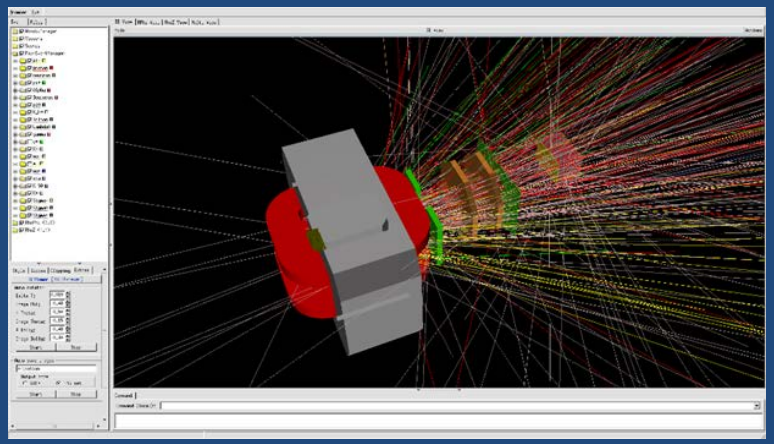

## HOT TOPICS:

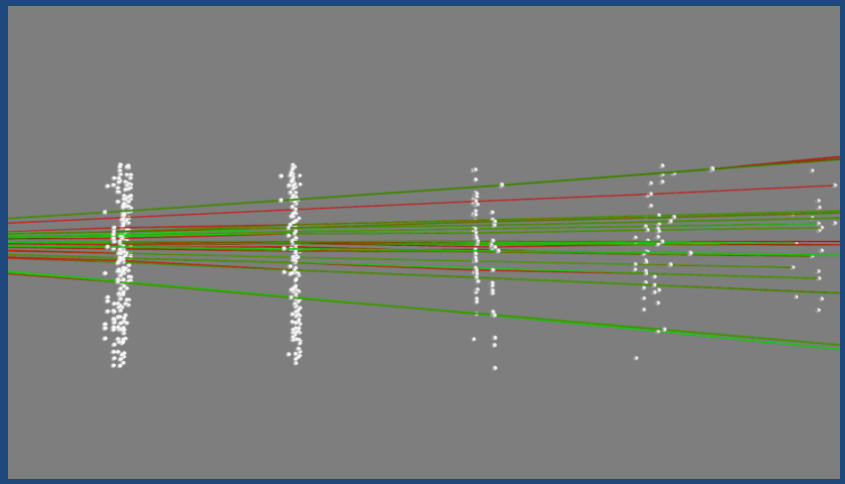
- ➢ Software for simulation and data processing
- ➢ HPC for theoretical investigation of dense hadronic matter
- ➢ Integrated computing resources for data processing
- ➢ Data storage system – datalake concept based on EOS

# ML Tracks Reconstruction BM@N

Machine learning algorithms bring a lot of potential to the tracks reconstruction problem due to their capability to learn effective representations of high-dimensional data through training, and to parallelize on supercomputer GOVORUN.

Simulation data





Input data for the first step algorithm were simulated by GEANT in MPDRoot framework for the real BM@N configuration.

Real track

True found track

Ghost track

White dots are both hits and fakes

Efficiency 97,5%

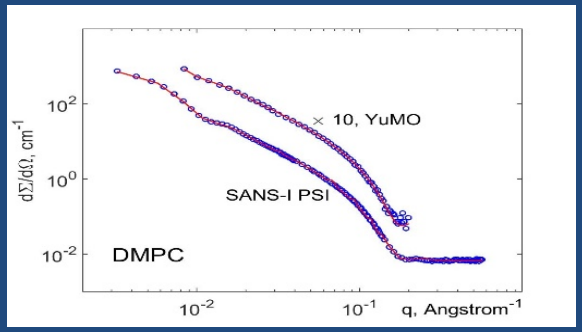
*G.Ososkov et al, Deep learning for tracking NEC-2017*

# Efficient parallelization of complex numerical problems in computational physics

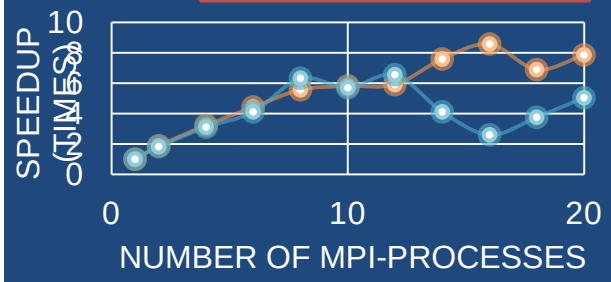
## Determination of the Vesicular Systems Parameters: Parallel Implementation and Analysis of the PTNS Vesicle Structure.

Basic parameters of phospholipid vesicles are estimated within the separated form factors method (SFF) by means of the least square minimization procedure.

M. Bashashin, E. Zemlyanaya, E. Zhabitskaya, M. Kiselev, T. Sapozhnikova. European Physics Journal – Web of Conferences, Vol. 173 (2018) 05003



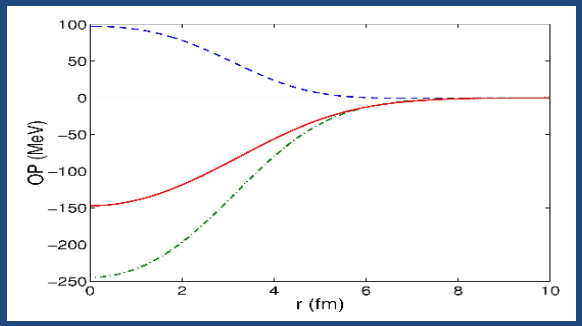

Reduction of computation time by ~**10 times**



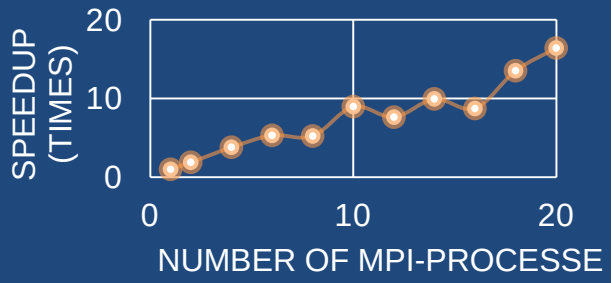

## MPI calculation of microscopic nucleus-nucleus optical potential within the frame of the double-folding model.

The double folding microscopic nucleus-nucleus optical potential is constructed by means of numerical solution of nonlinear integral equation with help of respective iteration procedure.

E.Zemlyanaya, K.Lukyanov, M. Bashashin. System analysis in science and education, Num 3 (2016)



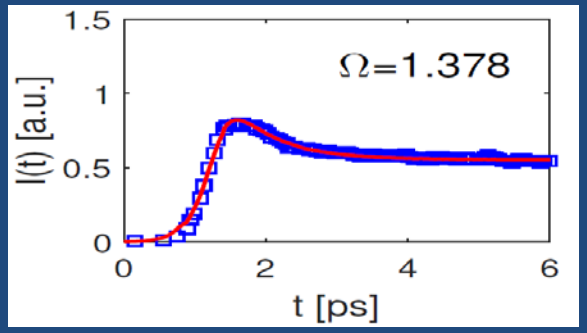

Reduction of computation time by ~**18 times**



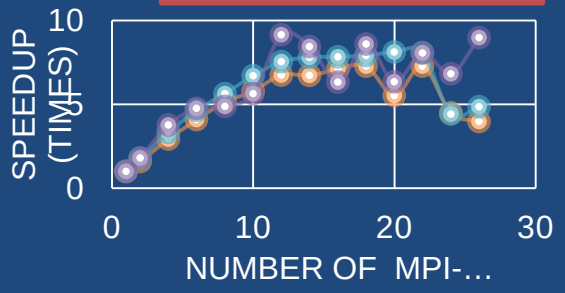

## Numerical Simulation of the Formation of Hydrated Electron States.

The model is described by a system of nonlinear partial differential equations which are solved by means of implicit method with the use of partition algorithm.

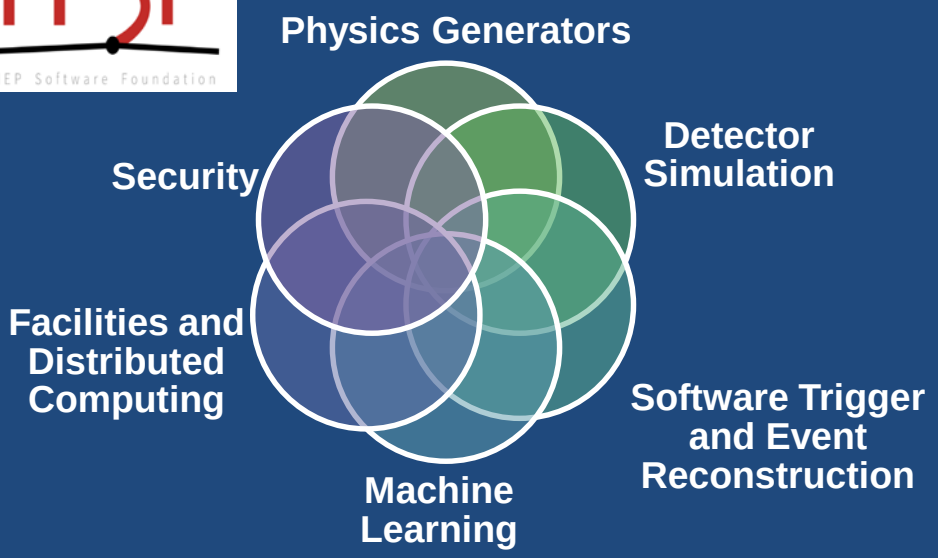
A. Volokhova, E. Zemlyanaya, V. Lakhno, I. Amirkhanov, M. Bashashin, I. Puzynin, T. Puzynina. European Physics Journal – Web of Conferences, Vol. 173 (2018) 06013

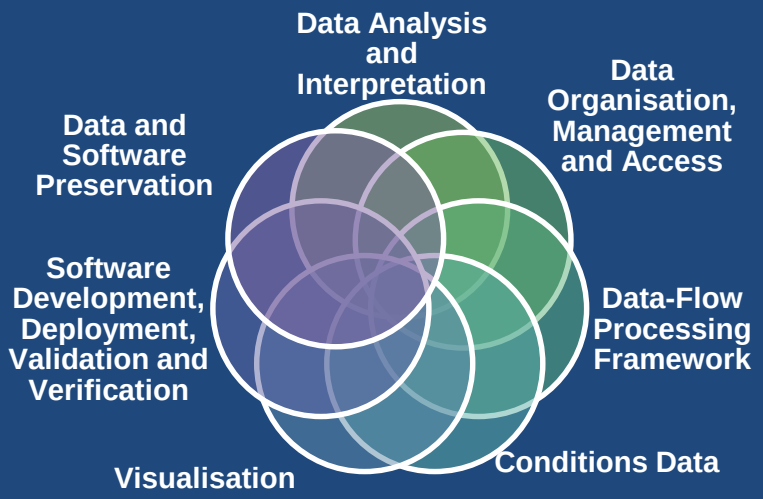


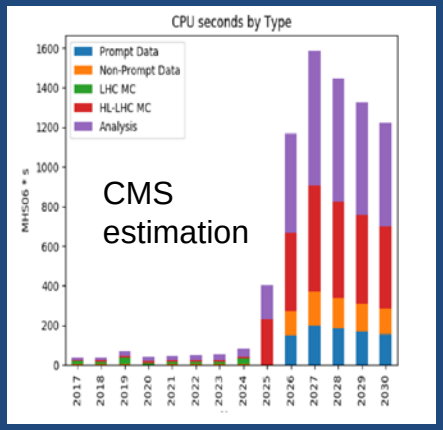

Reduction of computation time by ~**10 times**

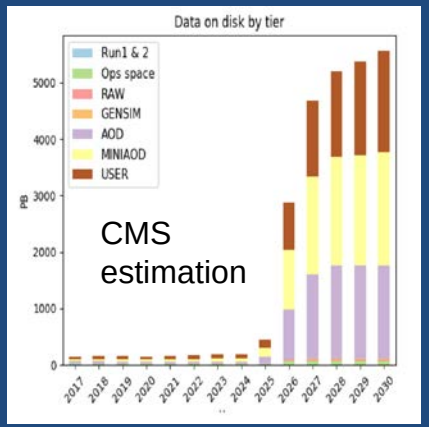


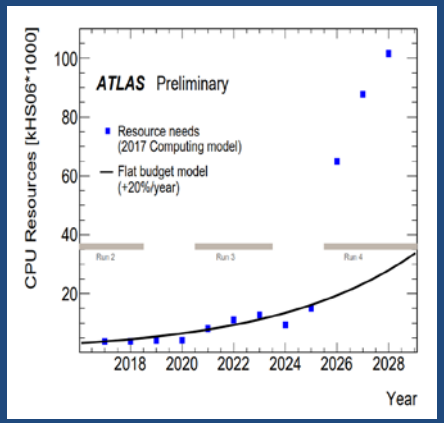
# CHALLENGE: R&D of software to acquire, manage, process, and analyse the big amounts of data to be recorded

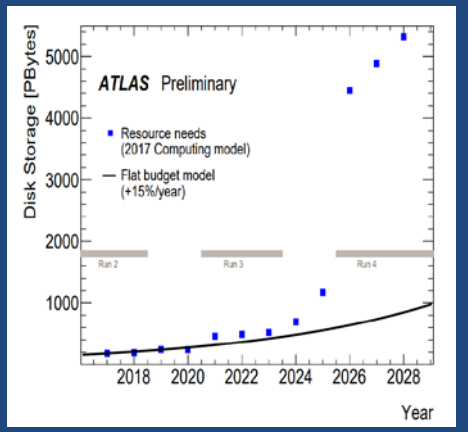



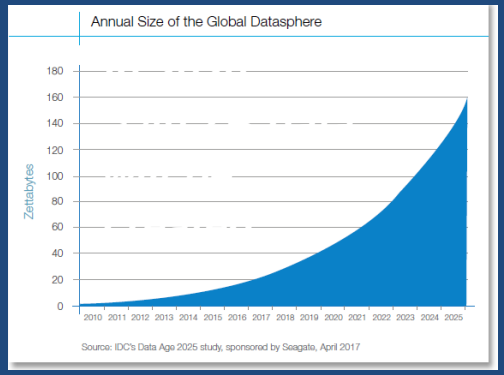



# Big Data + HPC (HPDA - High Performance Data Analysis)

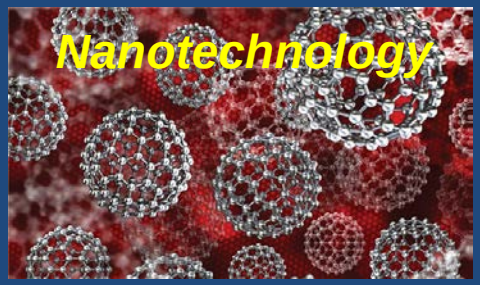



***Annual data production follows to exponential law.***

***High Energy Physics***


***Biology***

***Nanotechnology***

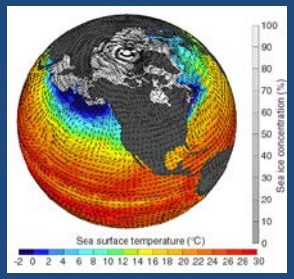


**Science**

**CERN Large Hadron Collider**
**> 20 Pb/Year, > 200 Pb stored**

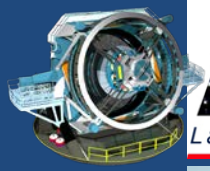
***Astrophysics***

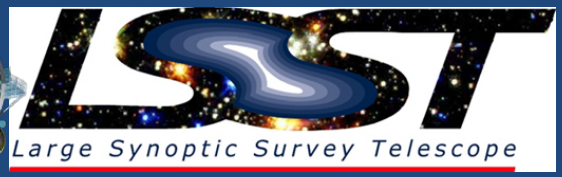
***Climate***

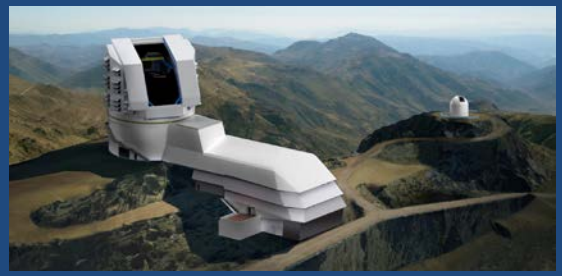




**Square Kilometre Array radio telescope (SKA)**
**> 20 Pb/Day (estimation)**



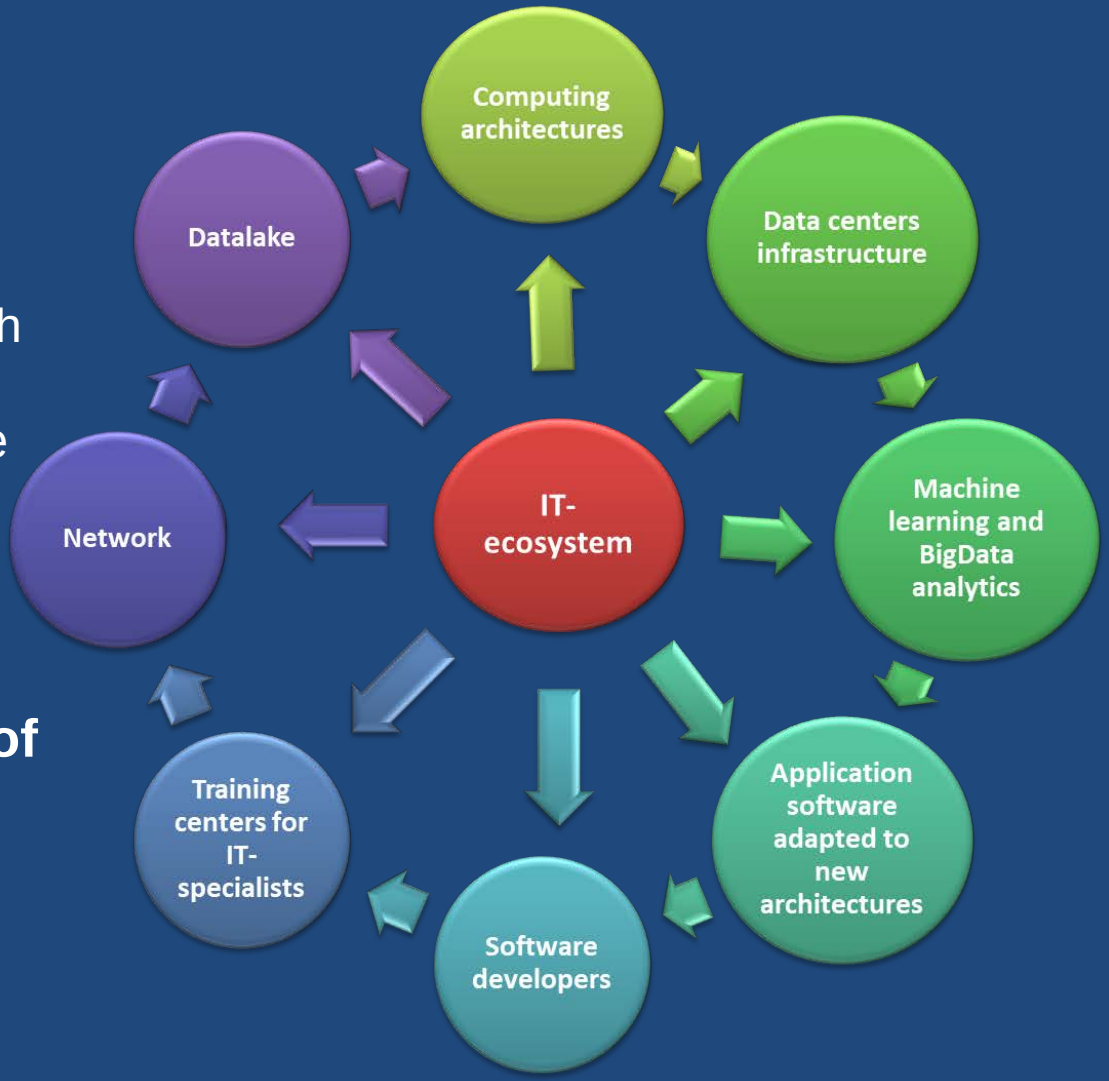
***...et cetera***

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

# Strategy of Information Technologies and Scientific Computing in JINR

## CONCEPT

of the development of IT-technologies and scientific computing aimed at solving strategic tasks of JINR through the introduction and development of a whole range of advanced IT solutions, integrated into a unified computing environment – **scientific IT ecosystem that combines a variety of technological solutions, concepts and methodologies**.

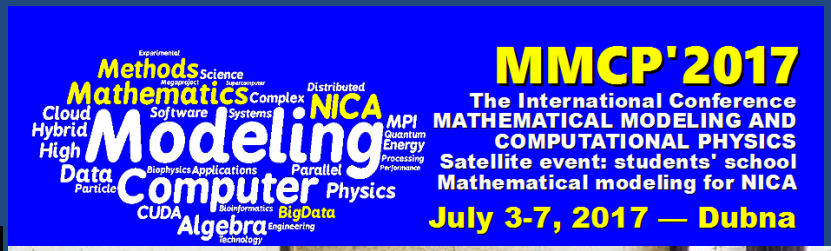


# Projects in framework Distributed computing

- **JINR-CERN cooperation (Worldwide LHC Computing Grid (WLCG), CMS, ATLAS, ALICE, COMPASS, NA62…)**
- **RDIG Development**
- **BigPANDA cooperation "Next Generation Workload Management System for BigData"**
- **Tier1 Center in Russia (NRC KI, LIT JINR)**
- **BMBF grant "Development computing and software environment for large-scale projects (NICA, FAIR, LHC) based on Grid, cloud and HPC infrastructures to provide joint investigations performed with participation of JINR and German research centers"**
- **"Development of grid segment for the LHC experiments" with South Africa;**
- **JINR - FZU AS Czech Republic Project "The grid for the physics experiments"**
- **JINR-BITP(Kiev) project "Development and implementation of cloud computing technologies on grid-sites at LIT JINR and BITP for ALICE experiment"**
- **JINR-Romania cooperation Hulubei - Meshcheryakov programme**
- **JINR- Mongolia cooperation (MongolGRID)**
- **JINR-Bulgaria cooperation in cloud computing (INRNE, SU)**
- **JINR- Kazakhstan cooperation in cloud computing (NU, ENU)**
- **JINR-Belarus cooperation in grid and cloud computing (BU, BTU)**
- **JINR- Azerbaijan cooperation in cloud computing (IP)**
- **JINR-China cooperation (BES-III)**
- **JINR-DESY cooperation (dCache)**
- **JINR- FNAL cooperation in computing (NOVA)**
- **JINR-BNL cooperation in computing (STAR)**
- **Cooperation with Armenia, Georgia, Moldova, Poland, Slovakia, Egipt…**

# LIT traditional conferences


Distributed Computing and Grid-technologies in Science and Education




IFIN-HH

ND ADVANCED INFORMATION SYSTEMS

UEF SAV

# LIT schools


# Summary

- We develop the Multifunctional Information and Computing Complex (MICC) based on grid, cloud and HPC technologies.
- We start the integration clouds of JINR Member States organizations to join resources for solving common tasks as well as to distribute a peak load across resources of partner organizations
- Development of  MICC as a center for scientific computing within the multidisciplinary research environment of JINR and JINR Member States, and particularly for megascience projects, such as NICA, is our goal.

# THANK YOU FOR YOUR ATTENTION !