MICC – new targets for information technology and computing in JINR

Gh. Adam[1, 2], V.V. Korenkov[1], T.A. Strizh[1]

[1]LIT-JINR Dubna, Russia

[2]IFIN-HH, Bucharest-Magurele, Romania

RO-LCG 2016 Conference , Bucharest-Magurele, Romania, October 26-28, 2016

Preamble:

The acronym **MICC** stands for the name of the project called **Multifunctional Information and Computing Complex**, approved for development in the Laboratory of Information Technologies (LIT) of the Joint Institute for Nuclear Research (JINR) in Dubna.

It is planned to be implemented at the state-of-the-art information technology with parameters enabling integral fulfillment of the computing needs of the JINR scientific projects in accordance with the new Seven-Year Plan of Development for 2017-2023.

The presentation in the sequel is heavily based on that made by the LIT Director, Dr. V.V. Korenkov, at the 45th Meeting of the PAC for Particle Physics, June 20, 2016

GOAL OF THE PROJECT:

- **development of the network, information and computing infrastructure of JINR for scientific and production activity of the Institute and its Member States on the basis of state-of-art information technologies according to the seven-year plan of JINR development for 2017-2023**

Countries and Organizations:	
Armenia (IIAP NAS RA, YSU) Azerbaijan (IP ANAS) Belarus (NC PHEP BSU, BNTU, JIPNR-Sosny NASB) Bulgaria (INRNE BAS, SU) CERN China (IHEP) Czech Republic (IP ASCR) Egypt (CU) France (CPPM) Georgia (GRENA, TSU, GTU) Germany (GSI, DESY, KIT) Moldova (ASM, IMCS ASM, IAP ASM, RENAM) Mongolia (NUM)	Poland (CYFRONET) Romania (IFA, IFIN-HH, INCDTIM) Russia (FRC"Computer Science and Control" RAS, IITP RAS, ISP RAS, ITEP, KIAM RAS, MPEI, MSU, RCC MSU, RIPN, NRC KI, RSCC, SINP MSU, INR RAS, SCC IPCP RAS, LITP RAS, Dubna Univ., SEZ "Dubna", SCC "Dubna", PNPI, UNN, BINP SB RAS, PSI RAS, IHEP, IMPB RAS, SSAU, ITMO, SPbSU, SPbSPU) Slovakia (IEP SAS) South Africa (UCT) Sweden (LU) USA (UTA, Fermilab, BNL) Ukraine (BITP NASU, NTUU KPI, KFTI)
Collaborations:	WLCG, RDMS CMS, RDIG

REQUIREMENTS

- Multi-functionality
- High performance
- Task adapted data storage system
- High reliability and availability
- Information security
- Scalability
- User customized hardware-software environment
- High-speed telecommunications and modern local network infrastructure

KEY ISSUES

- Creation of dedicated new computing infrastructure, modernization and development of existing one
- Development and improvement of the JINR telecommunication and network infrastructure
- Modernization of the MICC engineering infrastructure
- Exhaustive monitoring and control of the functioning of all the MICC elements

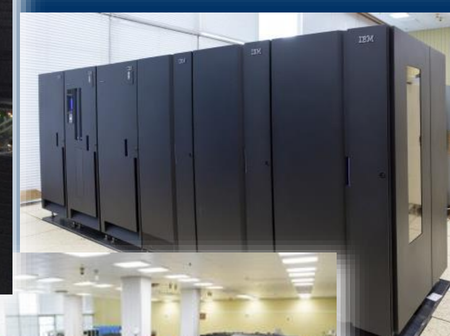
Part One: Computing@MICC

ACHIEVEMENTS DURING 2014-2016

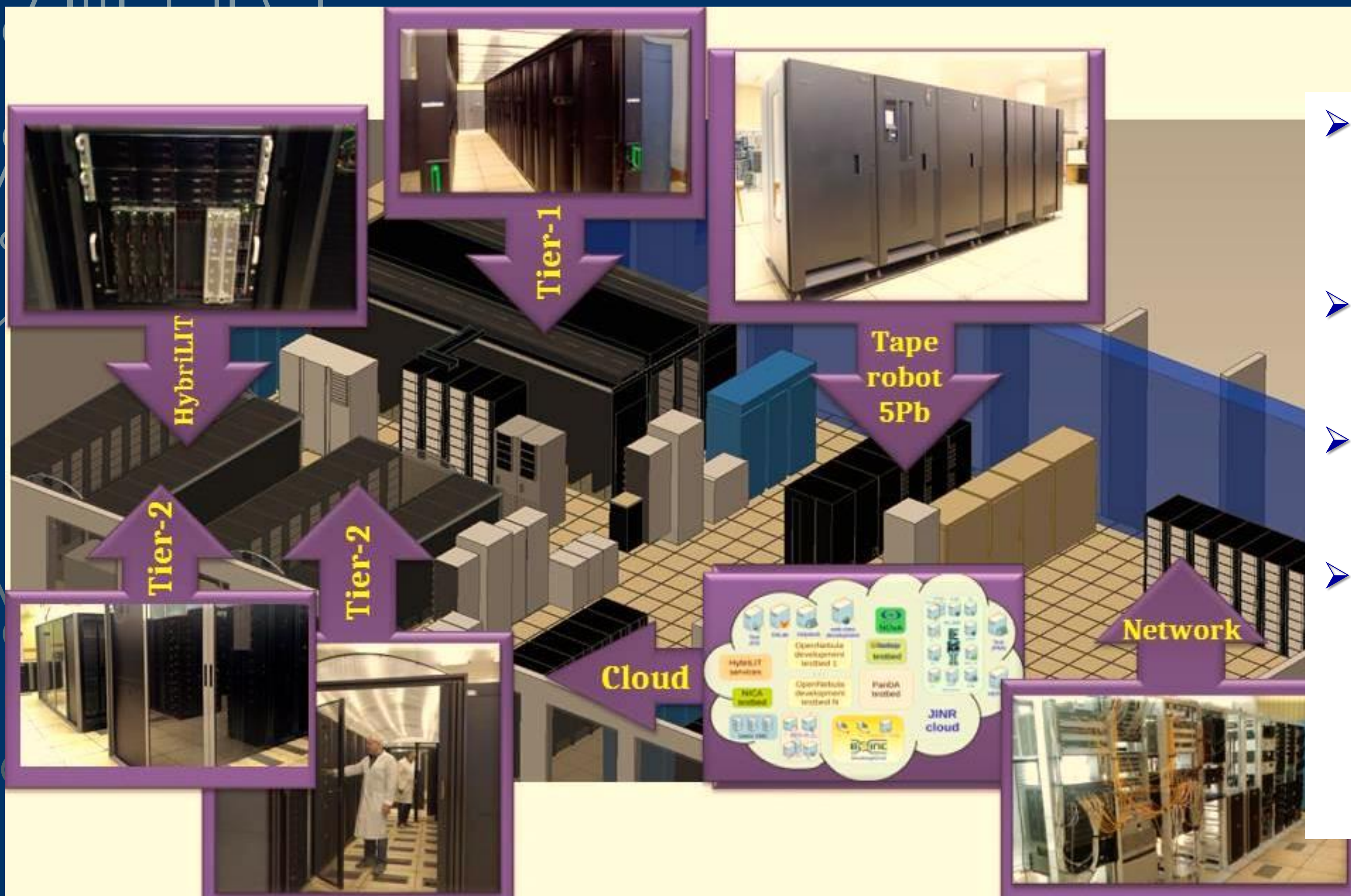
Put into operation of three new CICC components:

- CMS Tier-1 level center, the 7-th world one for the CMS experiment
- JINR cloud infrastructure
- Heterogeneous computing cluster HybriLIT

IT-infrastructure is one of the JINR basic facilities



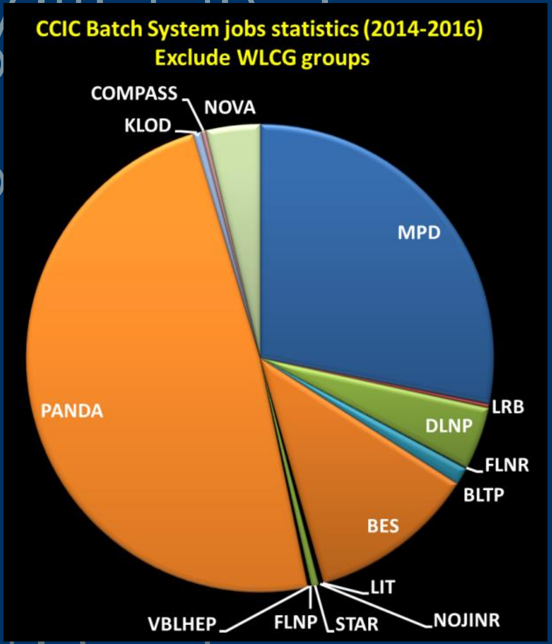
JINR Computing Centre at a Glance

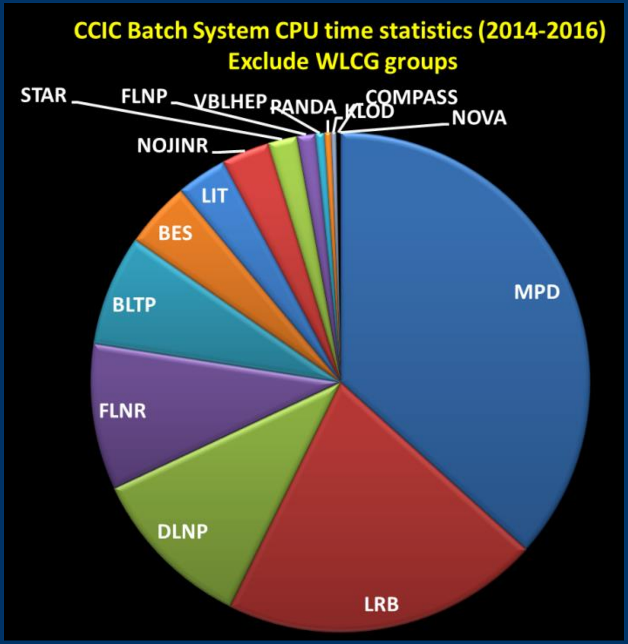


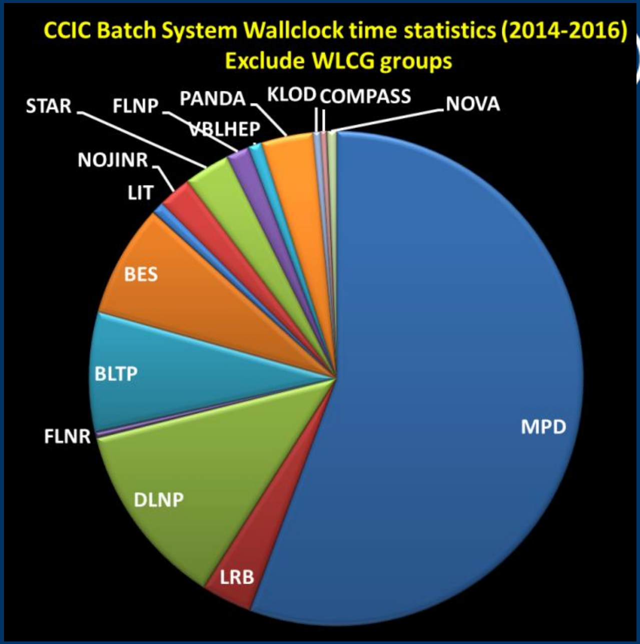
- Increase of computational nodes by ~3 times: **to 6160**
- Increase of disk file system by ~3 times: **to 5000 TByte**
- Tape robot for **5000 TByte** was put into operation
- **100 Gps** telecommunication link to Moscow will be put into operation by the end of 2016

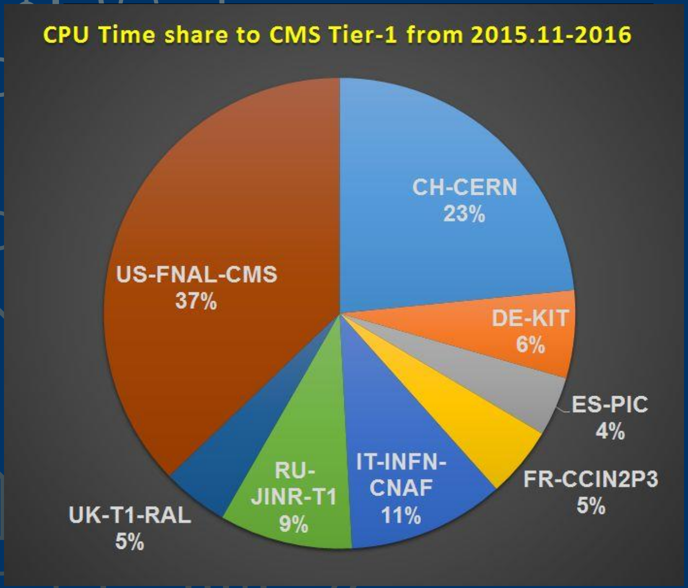
During the last three years more than **15 million** tasks have been carried out at the JINR CICC

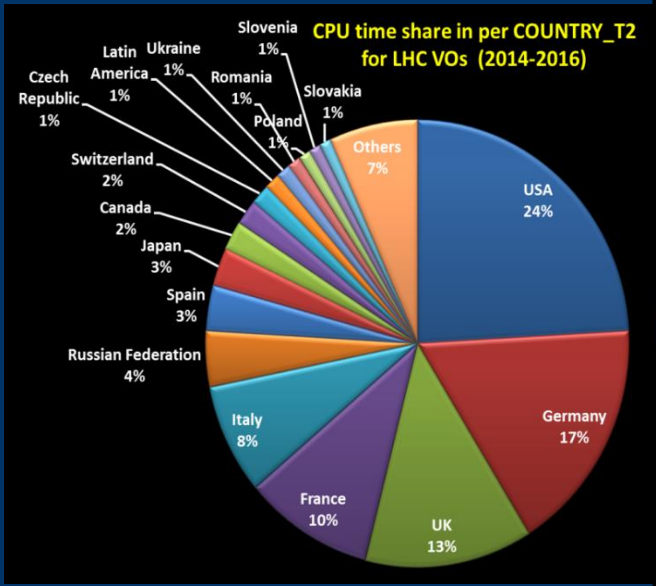
Resource usage 2014-2016

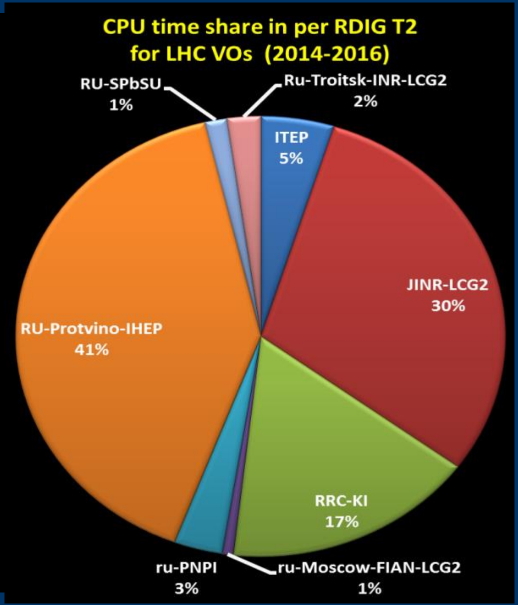












MICC MAIN COMPONENTS

JINR grid sites of WLCG/EGI
Tier-1 for CMS
Tier-2 for ALICE, ATLAS, CMS, LHCb, STAR, PANDA, BES, biomed, fermilab

Cloud infrastructure

Heterogeneous modular (CPU + GPU) computing cluster HybriLIT

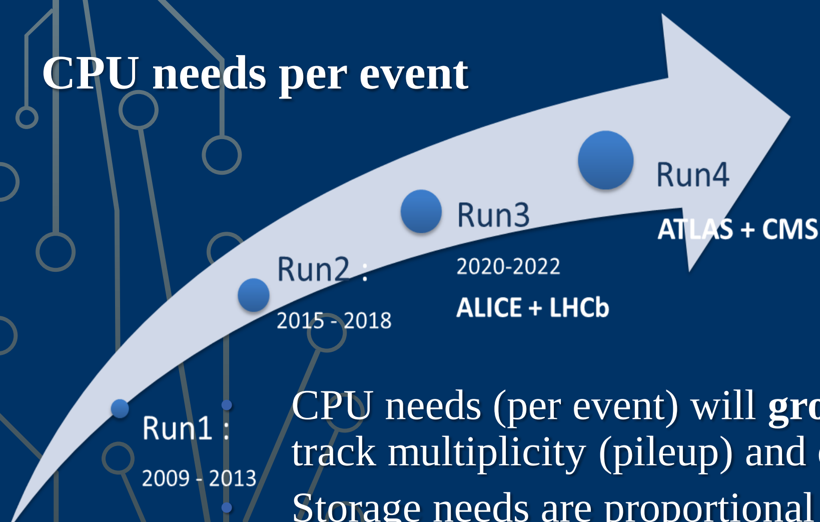
New: Cluster for off-line comprehensive data handling for BM@N, MPD, SPD. Related storage and computing facilities.

Future developments of the Tier-1 centre

Planned yearly growth of Tier-1**resources**.
Absolute values and percentage growth over previous year

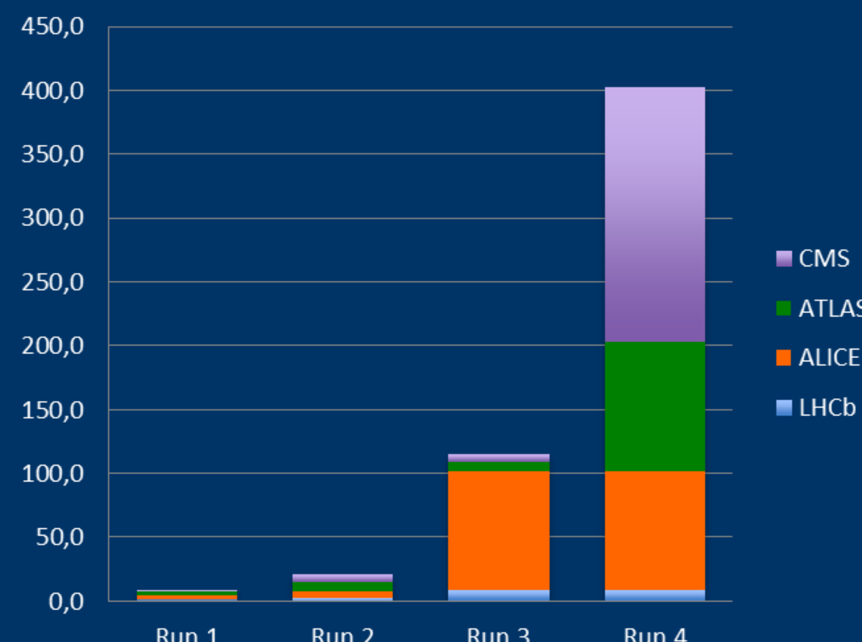
	2016	2017	2018	2019
Processor capacity of the core/kHS06	3400/54,4	4200/67,2 (24%)	5200/83,2 (23%)	10000/160 (52%)
Disc storage (TB)	3390	5070 (49%)	6100 (20%)	8000 (80%)
Tape storage (TB)	10000	20000 (100%)	20000 (0%)	20000 (0%)

CPU needs per event



- CPU needs (per event) will **grow** with track multiplicity (pileup) and energy
- Storage needs are proportional to accumulated luminosity
- Grid resources are limited by funding and fully installed capacity

LHC Upgrade 2019-2021. Computing Needs



MICC/TIER-2 FORESEEN DEVELOPMENTS

Increase of the MICC/Tier-2 resources and destination of their usage:

- Organization and support of the Tier-2 level sites for CMS, ATLAS, ALICE, LHCb
- Provision of countable resources as well as data storage and data access for third-party collaborations, local user groups and individual JINR users

A principal direction is the creation of a computing infrastructure devoted to the support of the BM@N, MPD, SPD experiments and the whole NICA project at all its stages. Within this work, it is planned to design a hardware-software installation DevLab for testing new hardware solutions and software systems for designing and creating a data processing complex for NICA.

Planned yearly growth of the MICC/Tier-2 resources

	2016	2017	2018	2019
Comp.cores / kHS06	2700/43,2	3700/59,2	4700/75,2	6000/96,0
Disk (TB)	2690	2970	3400	5000

STATUS AND PROSPECTS OF DEVELOPMENT OF THE MICC DATA STORAGE SYSTEMS

The data storage systems MICC/Tier-1/Tier-2 at LIT are intended to provide storage of the results of physical experiments and their processing. Basic parameters of the storage provision:

- **enough resources provided to the users,**
- **reliable storage,**
- **access to the formats used in experiments.**

Characterization of storage power and amount of stored data

Storage segment	engaged	allocated
T1-Disk	0.9 PB	2.6 PB
T1-Buffer	0.3 PB	0.5 PB (size of tape robot –5PB)
T2	0.7 PB	1.1 PB
dCacheII	56 TB	153 TB
XROOTD	323 TB	438 TB

Organization of data storage for the NICA experiment

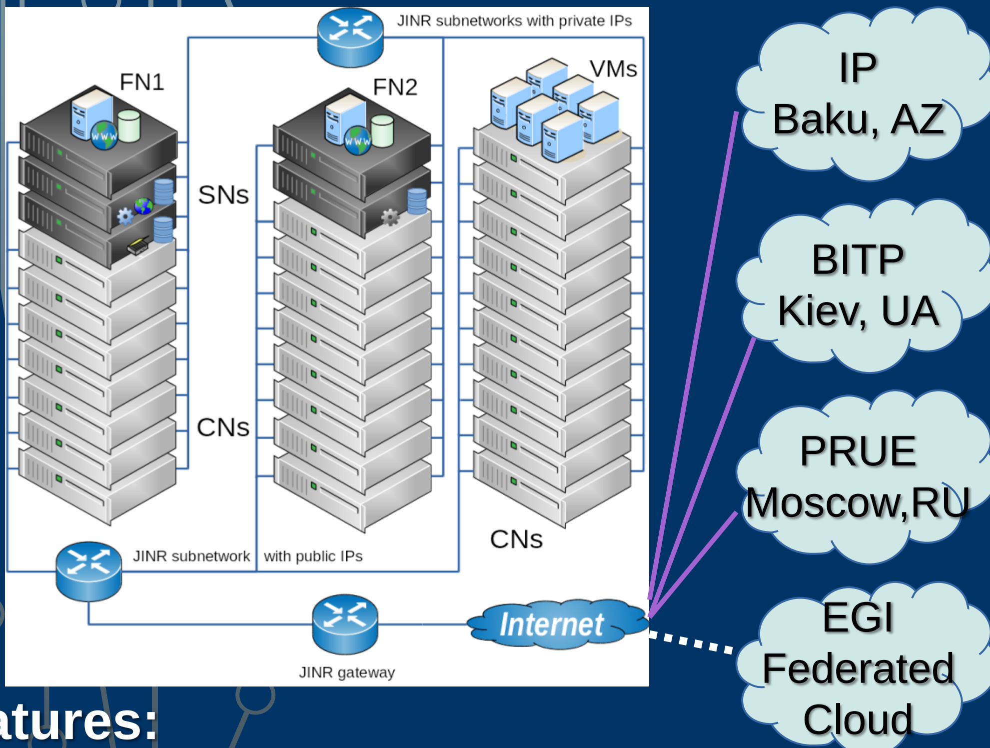
A first priority task is to install a two-tier (disks-tapes) storage system for the NICA experiments able to cope with the significant amounts of data storage (up to 2.5 PB per year) after the launch of the first phase of NICA project.

This task is subdivided into two subtasks :

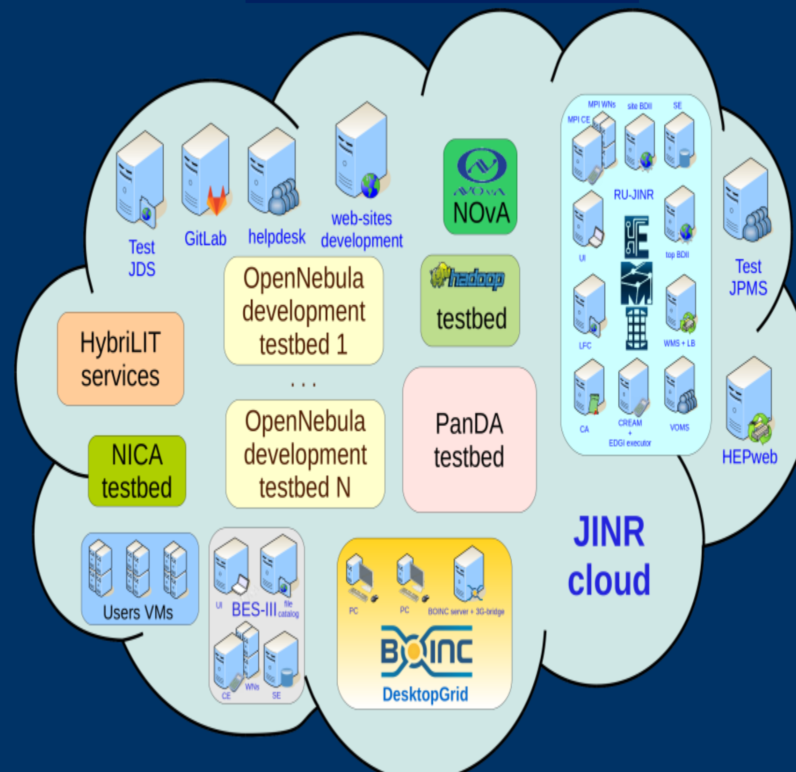
1) Provision of a system of information handling using the existing infrastructure,
2) Development of a full-scale storage project obeying the NICA data processing model which is being worked out.

OPEN NEBULA – BASED JINR CLOUD INFRASTRUCTURE

Architecture and connections



Utilization



Features:

- Servers: 40
- CPU cores: 200
- Total RAM: 400 GB
- Total DNFS disk capacity: 16 TB
- Total local disk capacity for VM/CT deployment: 20TB

Registered users: 80

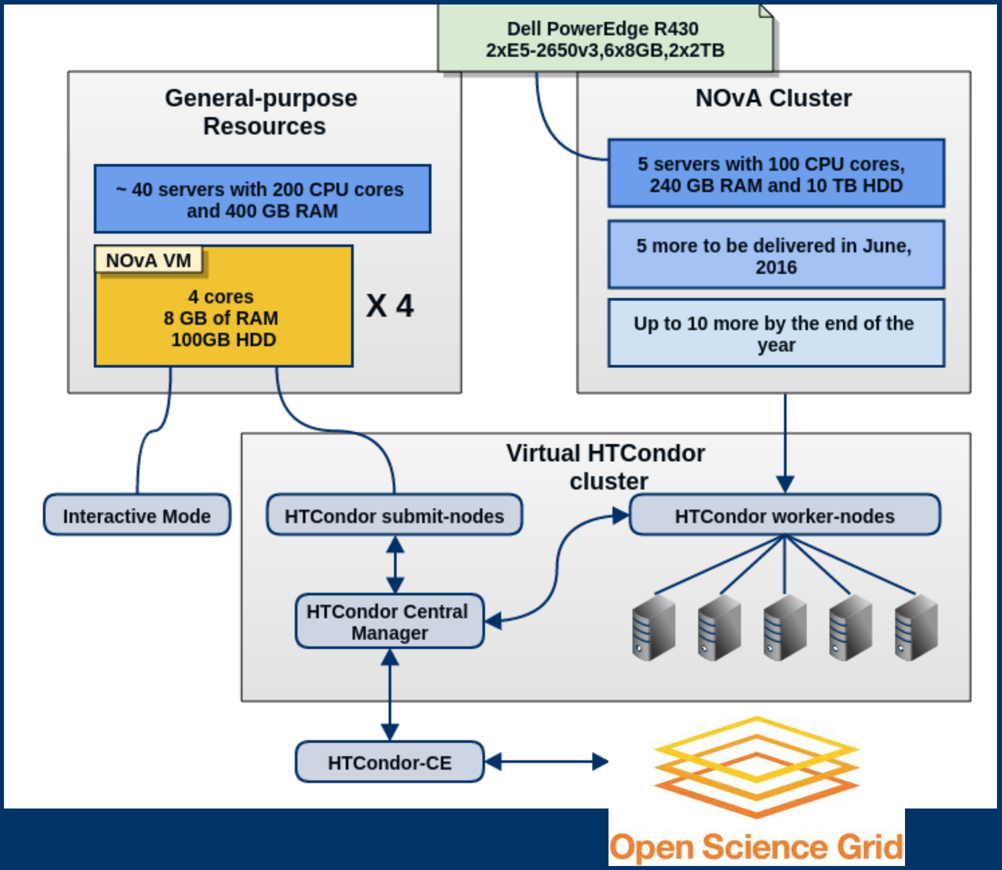
# of running VMs: 118

Computing Support for Neutrino Projects within JINR Cloud Infrastructure

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

- ✓ 4 VMs for interactive/batch processing used by local JINR NOvA team
- ✓ Virtual batch-cluster based on HTCondor and connected to OSG
- ✓ 100 CPU, 240 GB RAM and 10 TB HDD already available
- ✓ Up to 400 CPU, 1 TB RAM and 80 TB HDD by the end of the year
- ✓ Computing support team was formed including physicists and IT specialists

Cloud 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 interest in using JINR cloud resources. At the moment the experiments' tasks and required computing capacities are being discussed.

Heterogeneous computing cluster HybriLIT

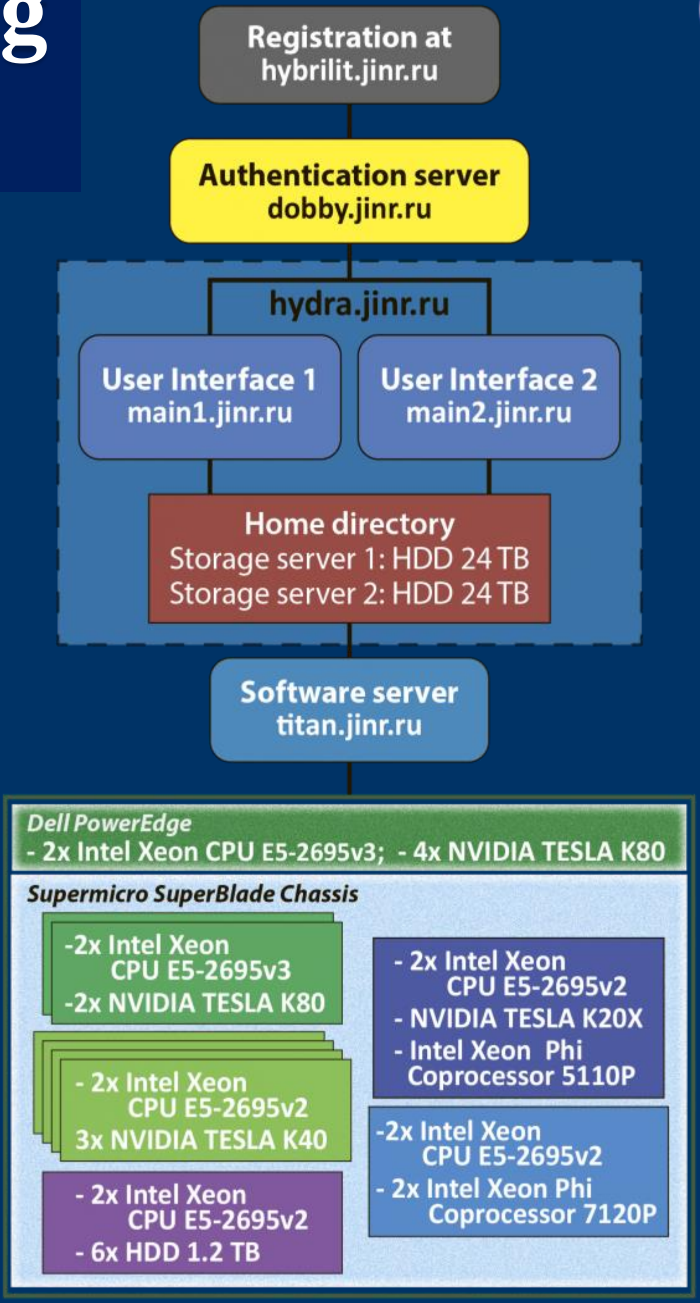
• **<u>Purpose</u>**: The HybriLIT can be characterized as a **modular** heterogeneous High Performance Computing (HPC) complex answering to three basic tasks:

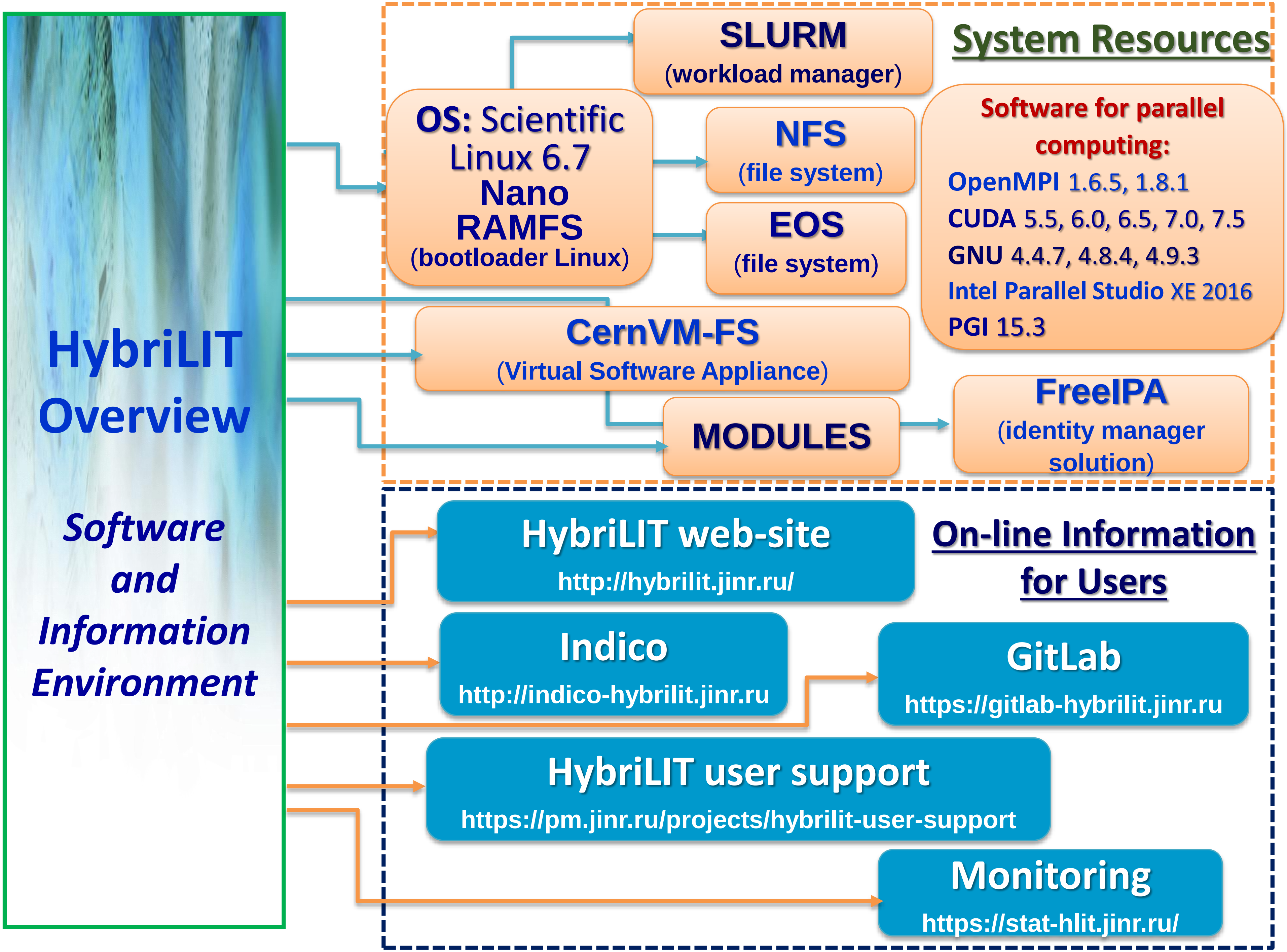
- • *Task* 1: Design and implementation of parallel software for computing intensive research;
- • *Task* 2: Porting to the cluster open software packages, numerical libraries, and programs which are already tuned for hybrid architectures;
- • *Task* 3: Development of new mathematical methods and parallel algorithms adapted to heterogeneous architectures.

Heterogeneous computing cluster HybriLIT

Cluster State by end of June 2016

# of CPU cores	252	# of nodes involving CPU	10
# of GPU cores	77184	# of nodes involving GPU accelerators	8
# of MIC cores	182	# of nodes involving CPU co-processors	2
Total RAM	2.4 TB	InfiniBand	16x FDR 56 Gbps
Total HDD	55.2 TB	Ethernet	10 Gbps
Peak performance in single precision floating point arithmetic	142 TFLOPS	Peak performance in double precision floating point arithmetic	50 TFLOPS
|| Usual power consumption: 5 – 10 kW || Energy efficiency: 4.56 GFlops/W || Peak power demand: 20 kW			




HybriLIT Overview
Software and Information Environment
System Resources
SLURM
(workload manager)
OS: Scientific Linux 6.7
Nano RAMFS
(bootloader Linux)
NFS
(file system)
EOS
(file system)
Software for parallel computing:
OpenMPI 1.6.5, 1.8.1
CUDA 5.5, 6.0, 6.5, 7.0, 7.5
GNU 4.4.7, 4.8.4, 4.9.3
Intel Parallel Studio XE 2016
PGI 15.3
CernVM-FS
(Virtual Software Appliance)
MODULES
FreeIPA
(identity manager solution)
On-line Information for Users
HybriLIT web-site
http://hybrilit.jinr.ru/
Indico
http://indico-hybrilit.jinr.ru
GitLab
https://gitlab-hybrilit.jinr.ru
HybriLIT user support
https://pm.jinr.ru/projects/hybrilit-user-support
Monitoring
https://stat-hlit.jinr.ru/

HybriLIT: User training

At present the cluster is used by: 120 registered users,
including **26** from JINR Member States and **19** – from Russian Universities

Tutorials on the **HybriLIT** use:

- **Frequent tutorials** on parallel programming techniques for the institute staff and, under JINR-UC organization, for students and young scientists from JINR Member-States;
- **Specialized courses** from leading software developers.

Specialized courses and **seminars** within JINR-organized conferences and schools: HybriLIT-Indico site mentions, mostly for 2015 and 2016:
21 *tutorials* (1 in 2014, 12 in 2015, 8 in 2016); **16** *specialized* **meetings**;
18 *series of lectures on parallel programming techniques, etc.*

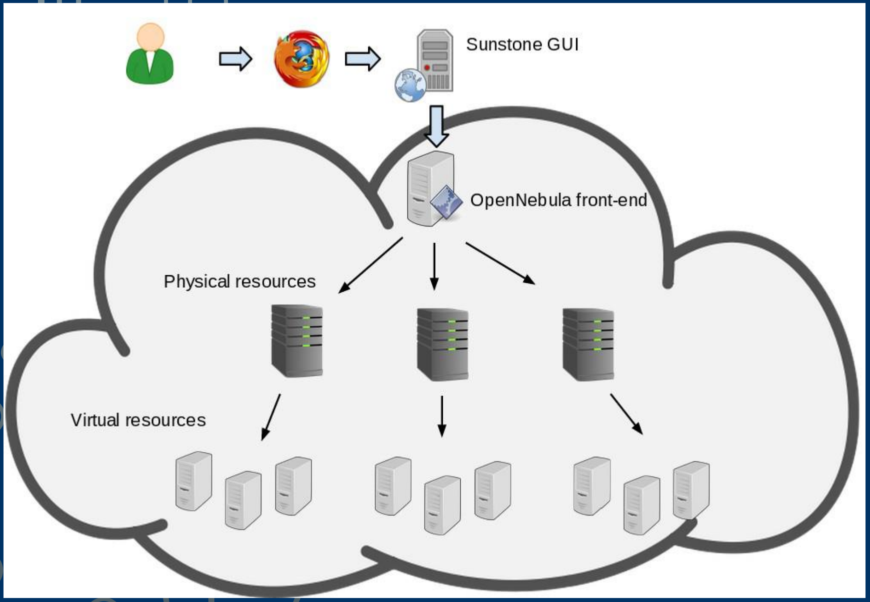
Participants (over 300, mostly young) from: JINR, Austria, Germany, India, Ireland, Japan, Romania, Russia, Slovakia, Ukraine, etc.

All information (lectures, materials) of past and upcoming events can be found at
http://indico-hybrilit.jinr.ru/

Prospects on cloud and heterogeneous computing

Advanced cloud infrastructures

– Dynamically reconfigurable computing services
– Large-scale open data repository and access services



	2016	2017	2018	2019
Cores	1000	1400	1800	2200
RAM/GB	4240	6160	8080	10000
Disk serves/TB	384	576	768	960

Advanced heterogeneous computing

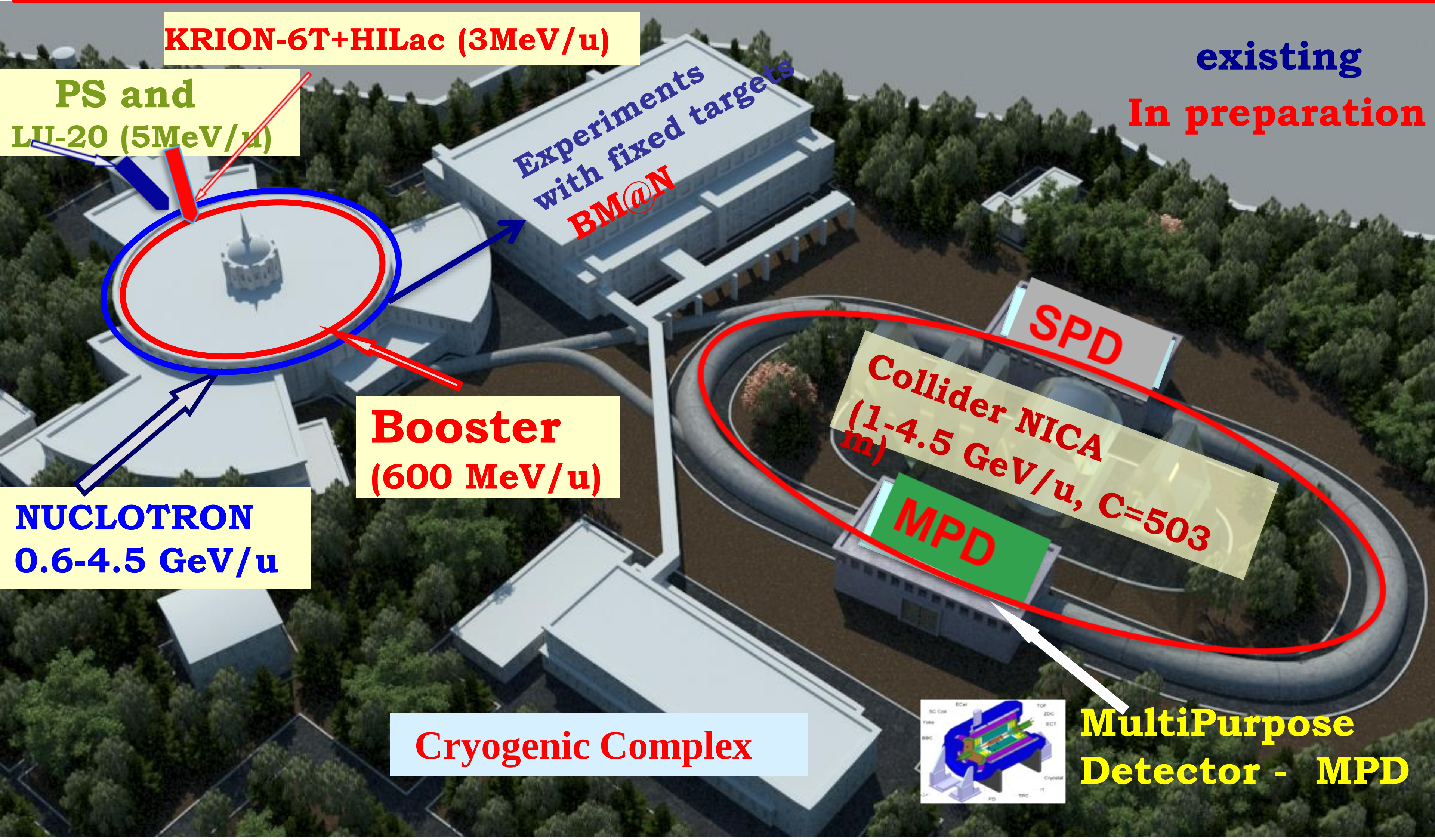
– User friendly information-computing environment
– New methods and algorithms for parallel hybrid computations
– Infrastructure for tutorials on parallel programming techniques

- **Annual increases in:**
  - computing resources - 90 Tflops
  - disk storage - 20TB
- **Systematic follow up of the advances in new hardware modules for high-performance computing and cluster upgrade with last hour modules**

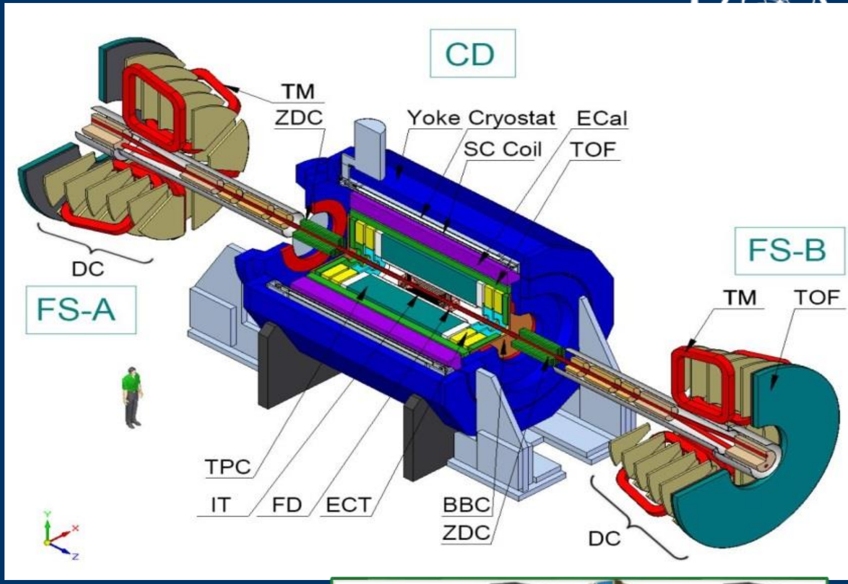
NICA Complex: *New era in the hot dense matter science*

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)



MICC modeling of computing at JINR-NICA Collider



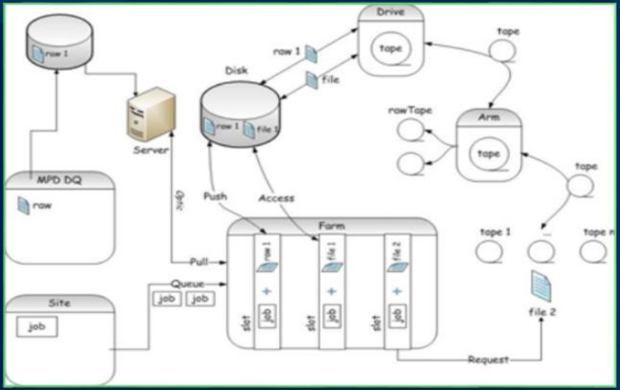
Main characteristics of the expected NICA data flow:

- Data acquisition speed up to 6 kHz
- Creation of ~ 1000 charged particles in central Au-Au collisions at NICA energies
- Expected amount of simultaneous events ~ 19 billion
- Expected raw data annual volume ~ 30PB, or, after data processing, ~ 8.4 PB

Modeling NICA distributed computer infrastructure

A model for process studies was created:

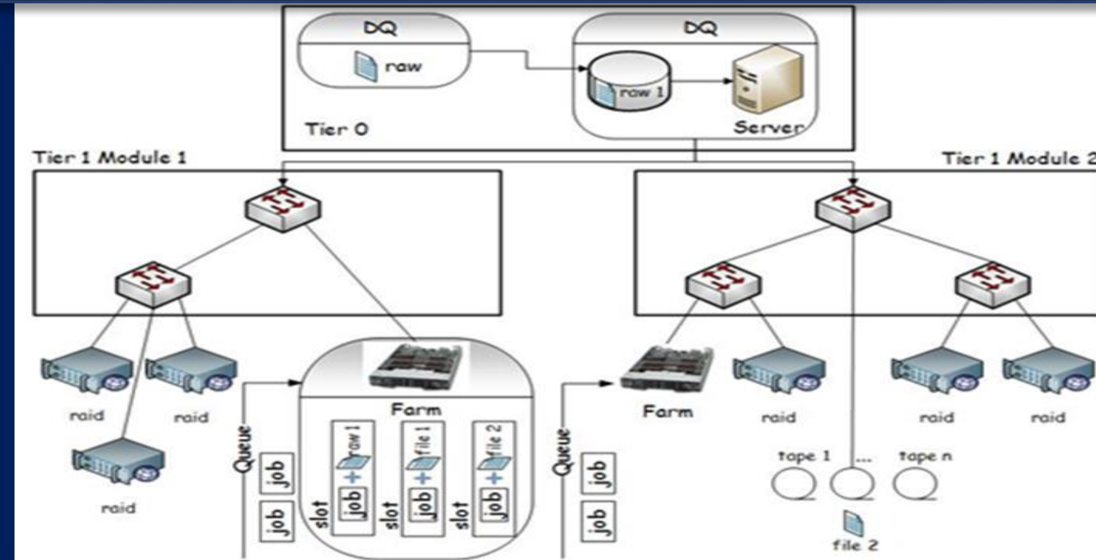
✓Tape robot,
✓Disk array,
✓CPU Cluster



Simulation of NICA-MPD-SPD Tier0-Tier1 computing facilities

Challenges to be faced in the NICA MPD-SPD experiment simulations:

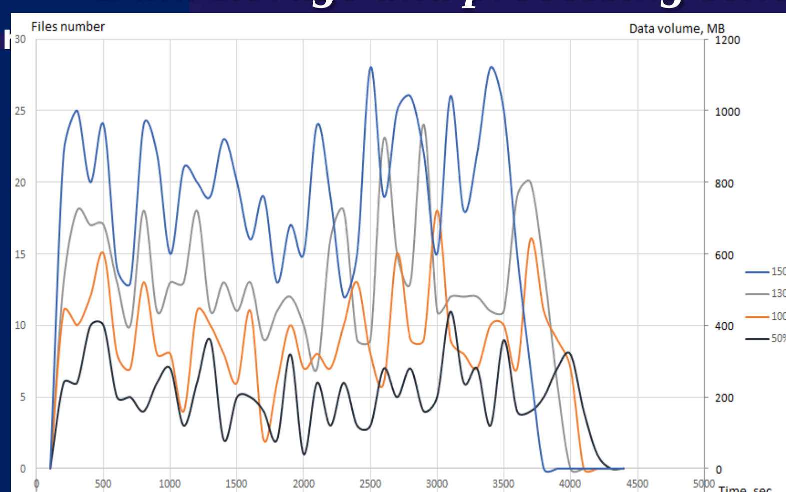
- large CPU and network resources
- combined grid and cloud access
- Intelligent dynamic data placement
- distributed parallel computing
- need of simulation a



Data storage and processing scheme of Tier0-Tier1 level

The program SyMSim (**S**ynthesis of **M**onitoring and **Sim**ulation) for the simulation of grid-cloud structures was developed.

Its **originality** consists in **combining a simulation program with a real monitoring**

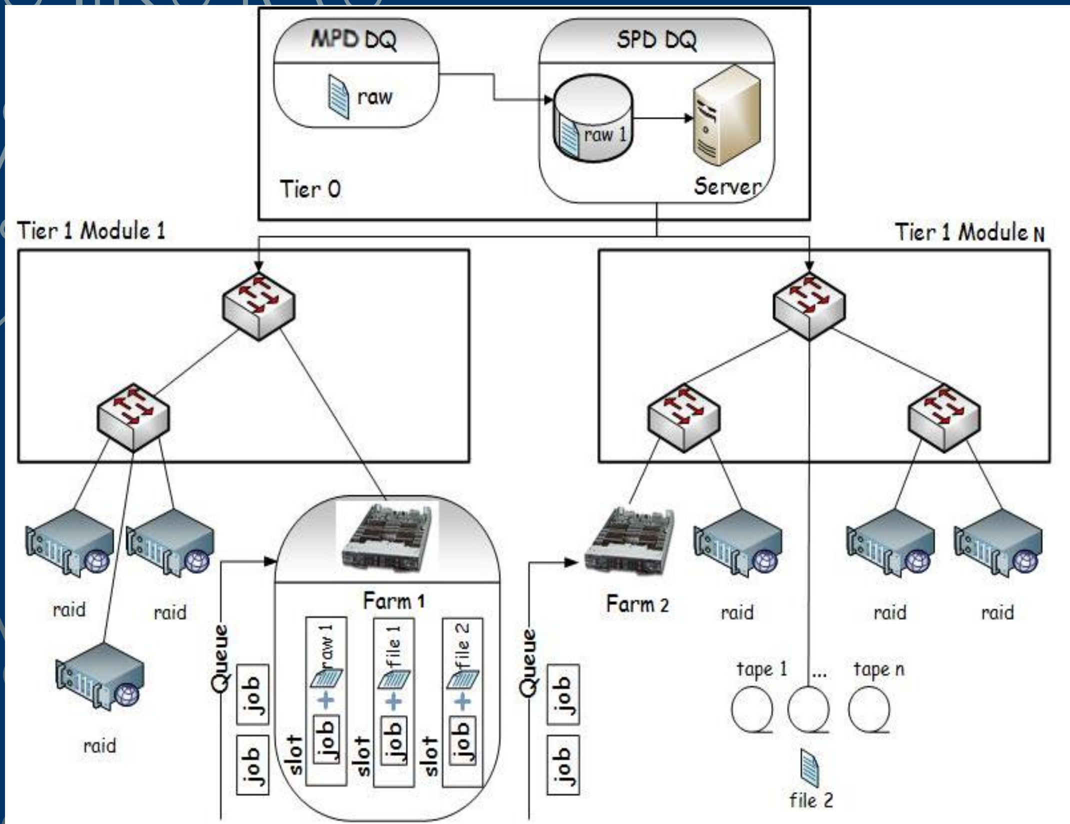


Number of DAQ data files stored on output disk buffer for growing data volumes

Estimate of the needed system capacities under variation of the intensity of the input stream.

SyMSim is sufficiently general and flexible to allow more realistic future assumptions

Toward a Computing Complex for Off-Line Data Handling in NICA Experiments



Basic goal:

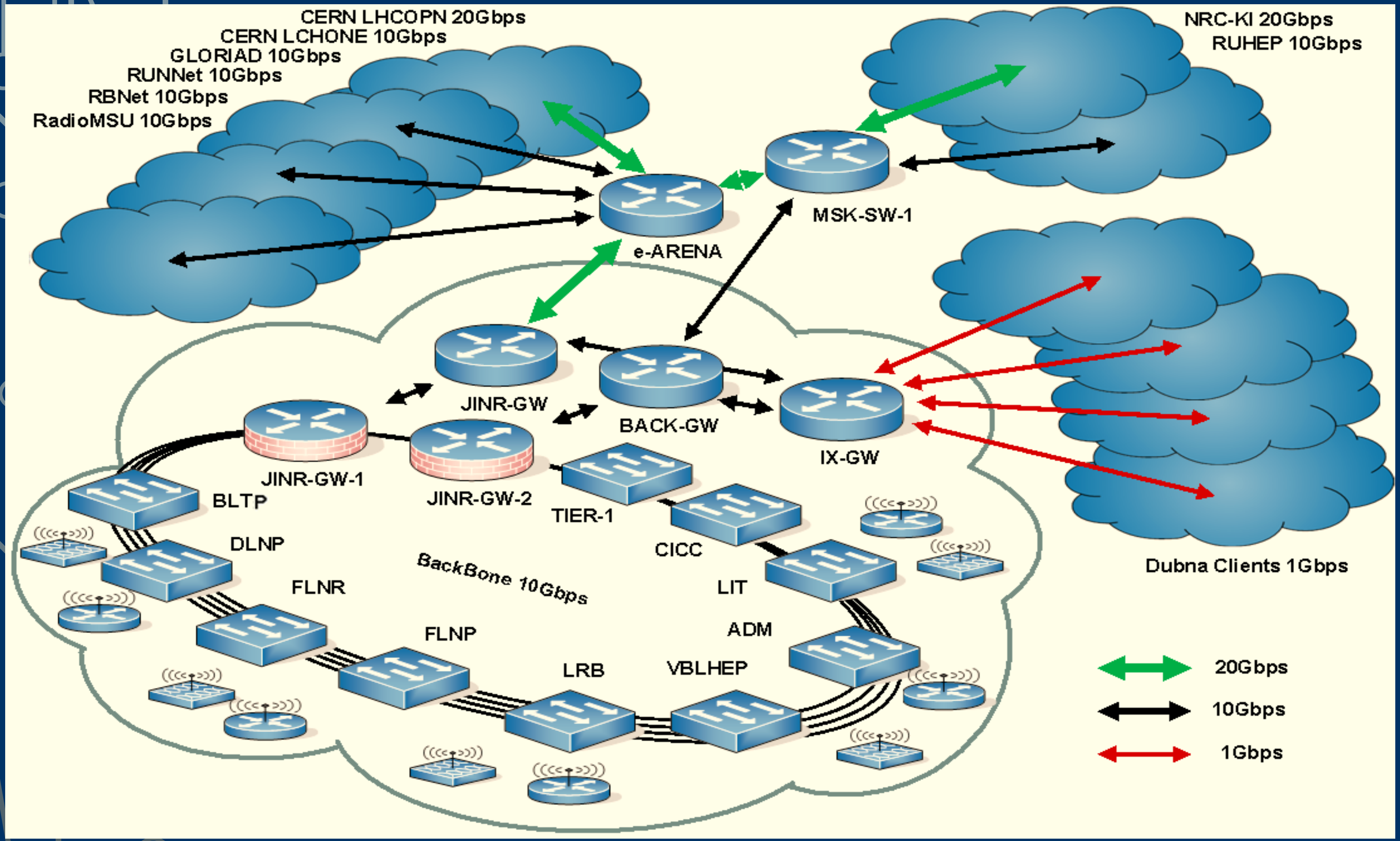
To create a MICC component for physical data storage and processing which takes into account specific parameters of the experiments. To this aim, it is necessary to properly describe and predict the performance and limitations of the developed storage and processing system for NICA data.
An answer is to be got to the question what kind of system architecture is preferable from the point of view of the realization of a reasonable balance of time, financial and technological costs.

SyMSim Simulation System > NICA > 3 stages

1. **M**odeling description of data generation processes, their volumes and storage conditions.
2. Modeling data processing - use of resources such as CPU, memory and I/O between concurrently running tasks.
3. Modeling communication processes of data traffic for different protocols in local and global networks.

Part Two: Connectivity

JINR Network and telecommunication



JINR Local Area Network
High-speed transport (10 Gb/s)
Comprises 8073 computers & nodes
Users – 4216, IP – 13267
Remote VPN users – 646
E-library- 1475, mail.jinr.ru-2000

Controlled-access at network entrance
Basic authorization services (Kerberos, AFS, batch systems, JINR LAN remote access, etc.)
IPDB database - registration of the network elements and users, visualization of statistics of the network traffic flow, etc.

INFORMATION SECURITY

From the point of view of the network structure, the safety system foresees a 3-level implementation:

- **On the edge of the network** by means of border routers;
- **At the core of the JINR local area network** – by means of the protection of all network servers;
- **On user (destination) level** – by means of operating systems, internetwork filters, workstations and specialized anti-virus tools

Development plans for the network infrastructure:

Increasing the channel capacity of the external JINR data link: 2 x 100 Gbps

Modernization of optical backbone of the local area network of JINR: 100 Gbps

Development of network services:

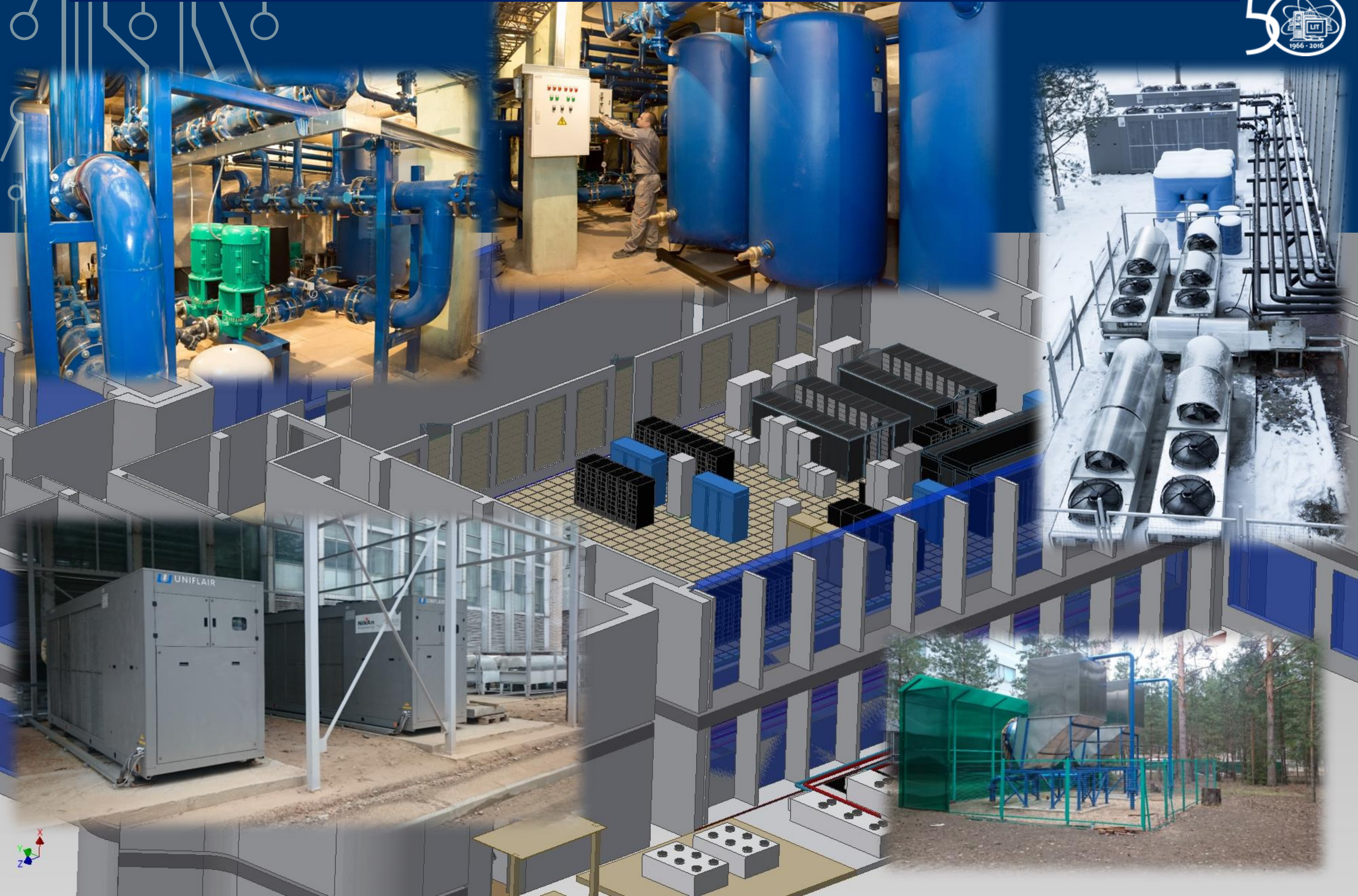
- Implement IPv6
- The use of new data transfer protocols
- Improved email service
- Wi-Fi authorization service
- Project “Personal office”

Local network of the NICA project:

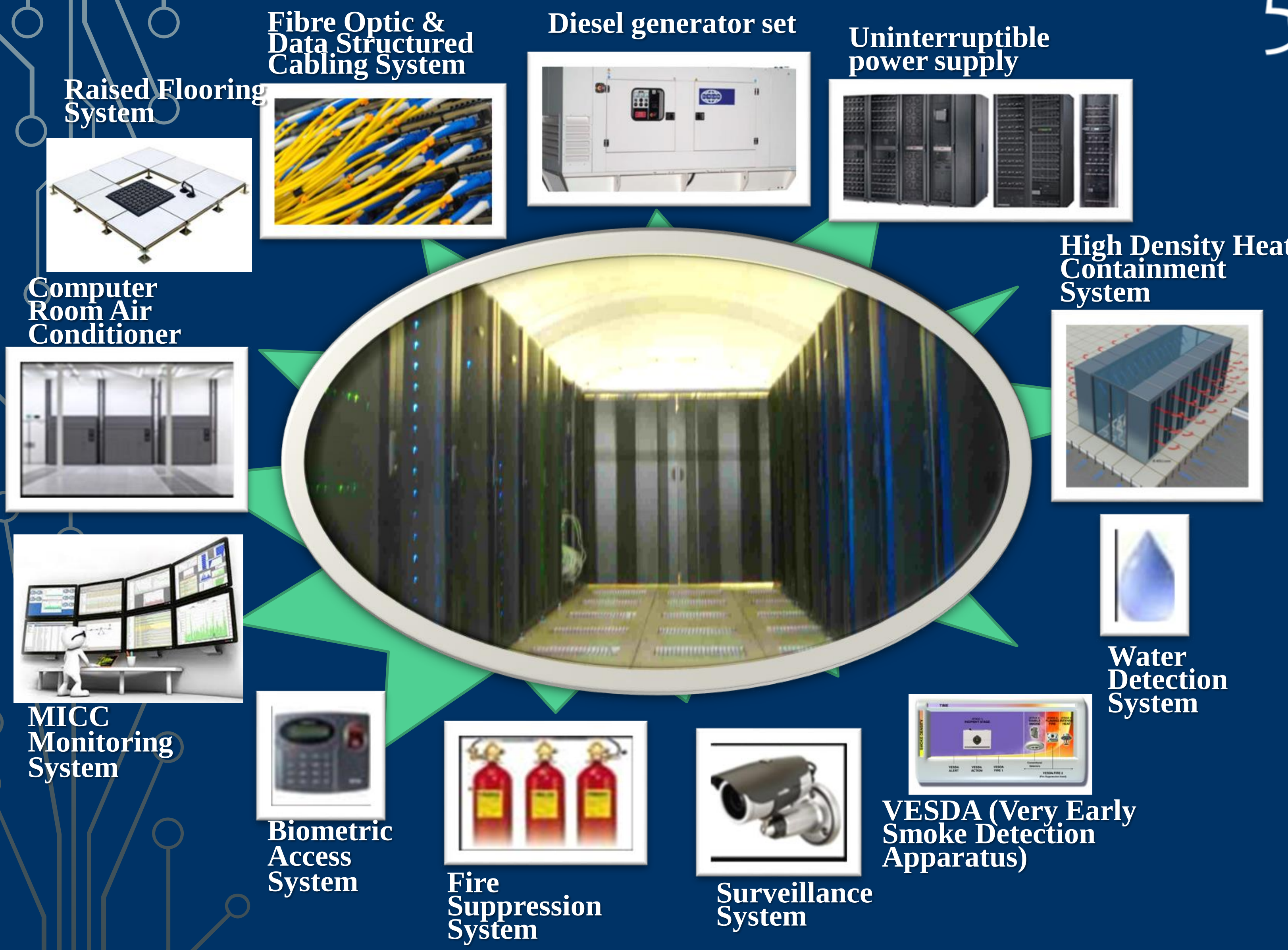
The projected capacity is planned as a data transmission channel with a throughput of 100 GbE

Part Three: Engineering Infrastructure

ENGINEERING INFRASTRUCTURE

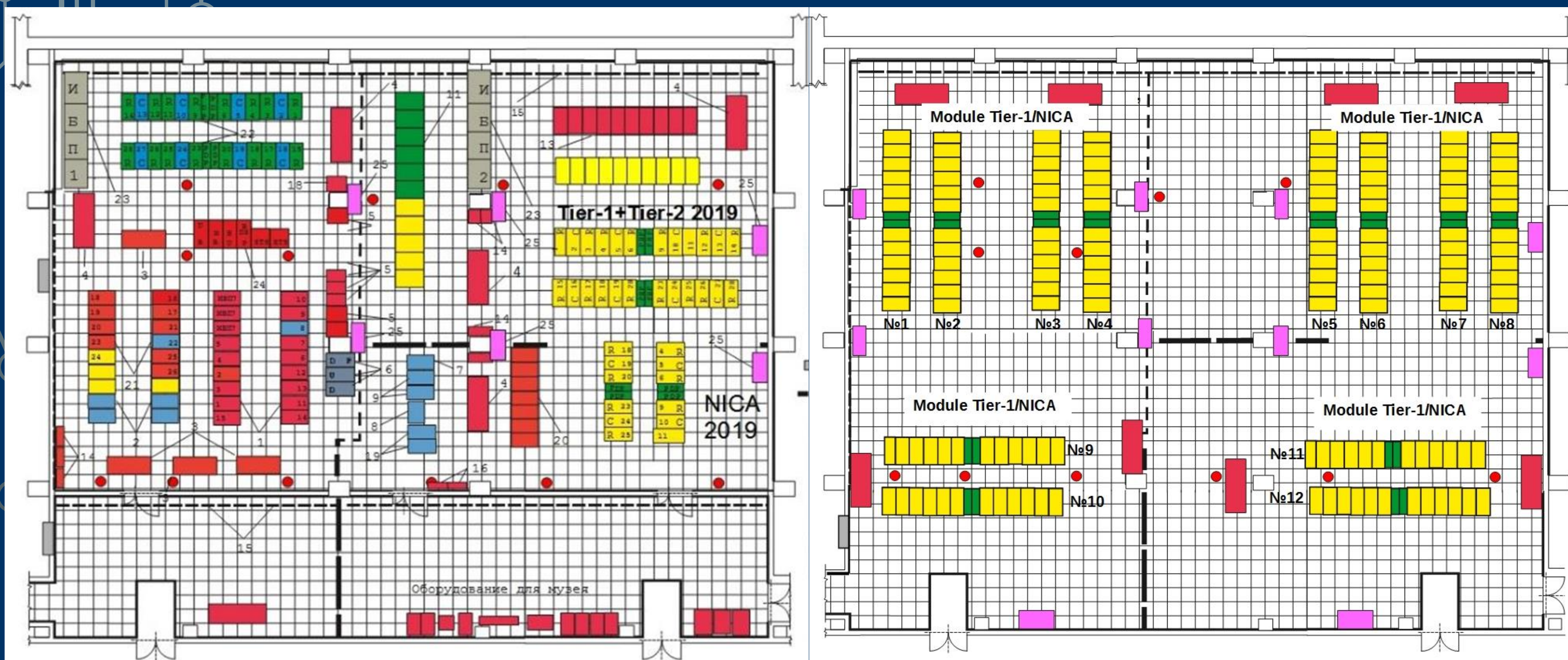


MICC engineering infrastructure



Goals and tasks of the MICC engineering infrastructure

The engineering infrastructure is to provide reliable functioning of the Complex 24 hours a day, 7 days a week round-the-year



Schematic arrangement of the equipment in the computer hall at the 2-nd and 4-th floors

Engineering infrastructure development targets

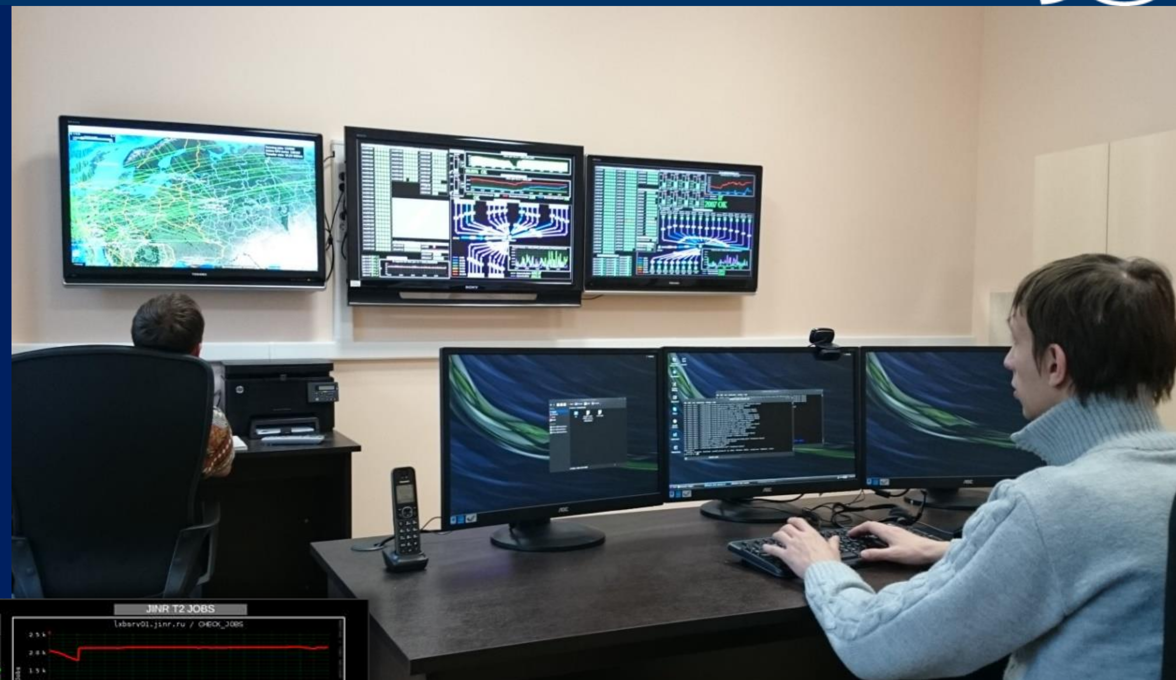
First and foremost, it is planned the modernization of the LIT power supply system including:

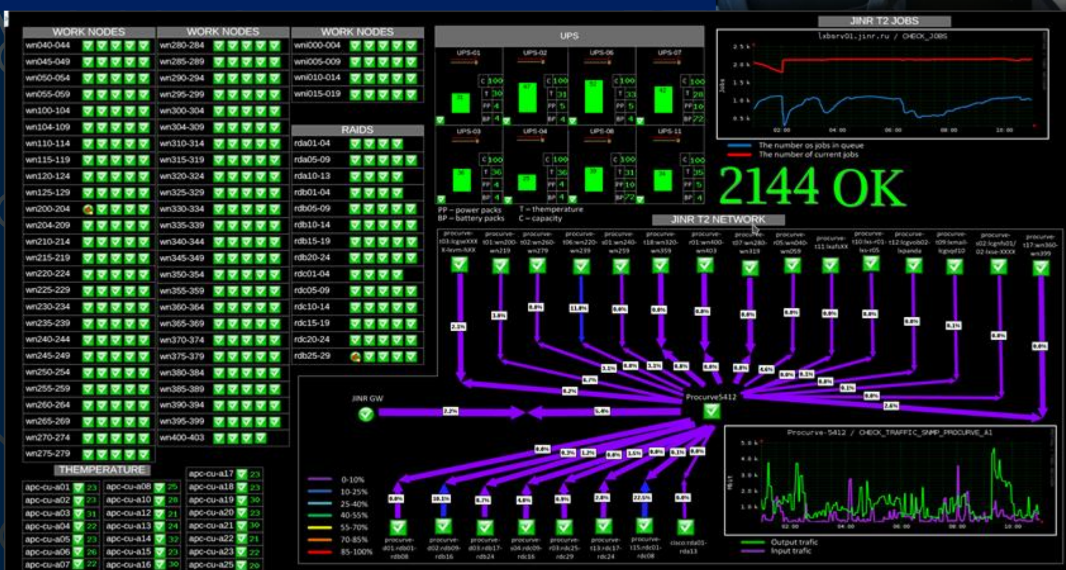
- the replacement of the existing 1 MW transformers by 2.5 MW ones and
- the purchase of a DGS unit

Part Four: Monitoring@MICC

Monitoring System Development

Robust performance needs monitoring the states of all nodes and services.
For the time being, **real time check is done for 690 elements, 3497 checks are simultaneously made.**
In case of emergency, alerts are sent to habilitated staff via e-mail, SMS, etc.





Prospects: The development of a **monitoring system** that integrates the monitoring of **all MICC components**: Tier-1, CICC/Tier-2, cloud environment, heterogeneous cluster, and engineering infrastructure.

The monitoring system: Principle of work

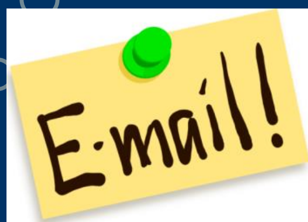
Visualization

Nagios

Plugins

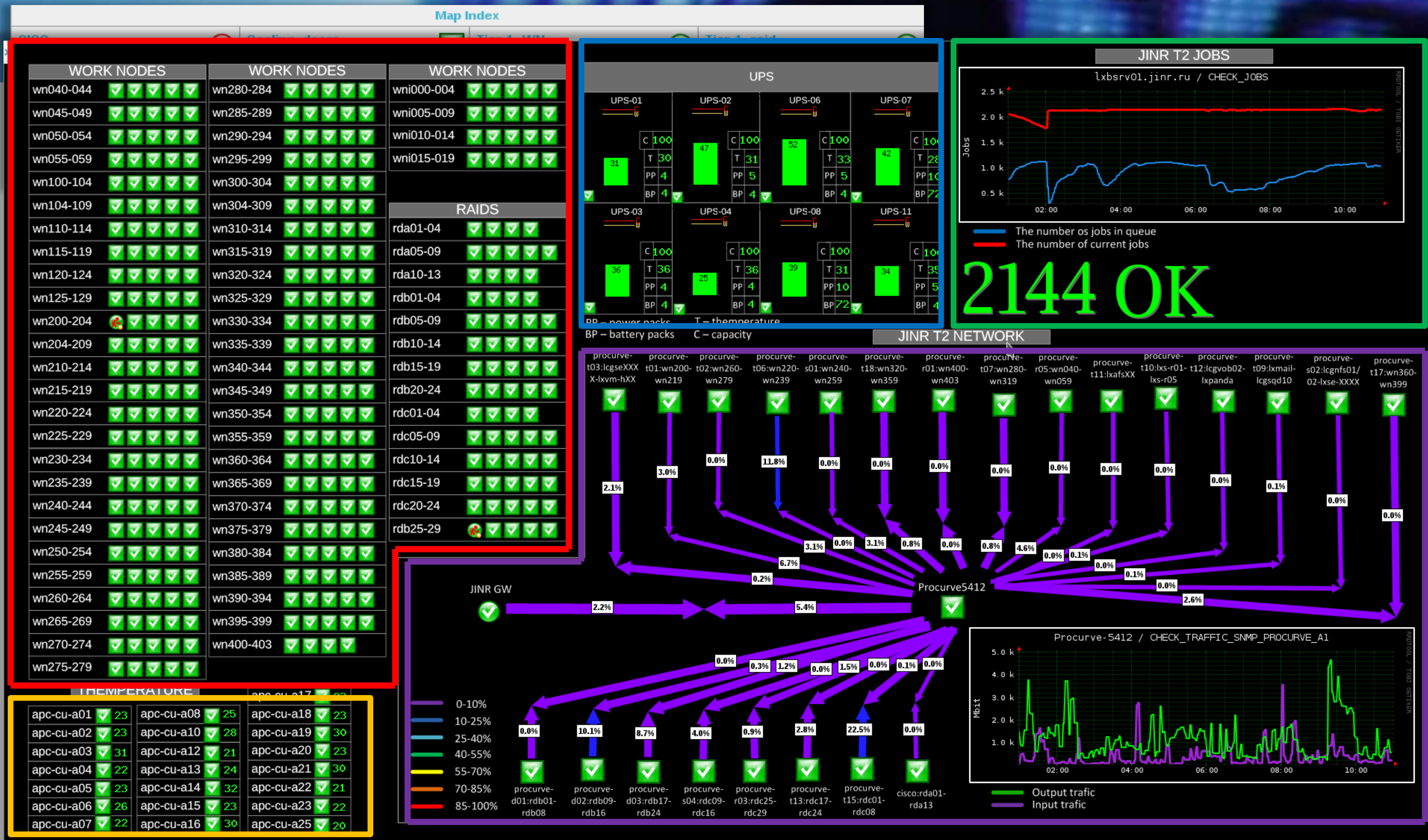
Hardware

Notification



Data storage system

Informational displays



Computing and storage servers

UPS

Network

Cumputing cluster load

Cooling system

Thank you for your attention !