MICC – new targets for information technology and computing in JINR

Gh. Adam\textsuperscript{1, 2}, V.V. Korenkov\textsuperscript{1}, T.A. Strizh\textsuperscript{1}

\textsuperscript{1}LIT-JINR Dubna, Russia
\textsuperscript{2}IFIN-HH, Bucharest-Magurele, Romania
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:

<table>
<thead>
<tr>
<th>Armenia (IIAP NAS RA, YSU)</th>
<th>Poland (CYFRONET)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azerbaijan (IP ANAS)</td>
<td>Romania (IFA, IFIN-HH, INCDTIM)</td>
</tr>
<tr>
<td>Belarus (NC PHEP BSU, BNTU, JIPNR-Sosny NASB)</td>
<td>Russia (FRC&quot;Computer Science and Control&quot; 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 &quot;Dubna&quot;, SCC &quot;Dubna&quot;, PNPI, UNN, BINP SB RAS, PSI RAS, IHEP, IMPB RAS, SSAU, ITMO, SPbSU, SPbSPU)</td>
</tr>
<tr>
<td>Bulgaria (INRNE BAS, SU)</td>
<td>Slovakia (IEP SAS)</td>
</tr>
<tr>
<td>CERN</td>
<td>South Africa (UCT)</td>
</tr>
<tr>
<td>China (IHEP)</td>
<td>Sweden (LU)</td>
</tr>
<tr>
<td>Czech Republic (IP ASCR)</td>
<td>USA (UTA, Fermilab, BNL)</td>
</tr>
<tr>
<td>Egypt (CU)</td>
<td>Ukraine (BITP NASU, NTUU KPI, KFTI)</td>
</tr>
<tr>
<td>France (CPPM)</td>
<td>Mongolian (NUM)</td>
</tr>
<tr>
<td>Georgia (GRENA, TSU, GTU)</td>
<td>Mongolia (NUM)</td>
</tr>
<tr>
<td>Germany (GSI, DESY, KIT )</td>
<td>Mongolia (NUM)</td>
</tr>
<tr>
<td>Moldova (ASM, IMCS ASM, IAP ASM, RENAM)</td>
<td>Mongolia (NUM)</td>
</tr>
<tr>
<td>RAL, Montan Heartlands</td>
<td>Mongolia (NUM)</td>
</tr>
<tr>
<td>Collaborations:</td>
<td>WLCG, RDMS CMS, RDIG</td>
</tr>
</tbody>
</table>
**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
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

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

**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.

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

<table>
<thead>
<tr>
<th></th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
<th>2019</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comp.cores / kHS06</strong></td>
<td>2700/43,2</td>
<td>3700/59,2</td>
<td>4700/75,2</td>
<td>6000/96,0</td>
</tr>
<tr>
<td><strong>Disk (TB)</strong></td>
<td>2690</td>
<td>2970</td>
<td>3400</td>
<td>5000</td>
</tr>
</tbody>
</table>

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
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

<table>
<thead>
<tr>
<th>Storage segment</th>
<th>engaged</th>
<th>allocated</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1-Disk</td>
<td>0.9 PB</td>
<td>2.6 PB</td>
</tr>
<tr>
<td>T1-Buffer</td>
<td>0.3 PB</td>
<td>0.5 PB (size of tape robot –5PB)</td>
</tr>
<tr>
<td>T2</td>
<td>0.7 PB</td>
<td>1.1 PB</td>
</tr>
<tr>
<td>dCacheII</td>
<td>56 TB</td>
<td>153 TB</td>
</tr>
<tr>
<td>XROOTD</td>
<td>323 TB</td>
<td>438 TB</td>
</tr>
</tbody>
</table>

**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

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

Utilization

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.
**Purpose**: 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.
### Cluster State by end of June 2016

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

**Hybrid computing cluster HybriLIT**

- Registration at hybrilit.jinr.ru
- Authentication server: dobbie.jinr.ru
- User Interface 1: main1.jinr.ru
- User Interface 2: main2.jinr.ru
- Software server: titan.jinr.ru

**Supermicro SuperBlade Chassis**

- 2x Intel Xeon CPU E5-2695v3
- 4x NVIDIA TESLA K80
- 2x NVIDIA TESLA K20X
- Intel Xeon Phi Coprocessor 5110P
- 2x Intel Xeon CPU E5-2695v2
- 3x NVIDIA TESLA K40
- 2x Intel Xeon CPU E5-2695v2
- 2x Intel Xeon Phi Coprocessor 7120P
- 2x Intel Xeon CPU E5-2695v2
- 6x HDD 1.2 TB
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)

CernVM-FS
(Virtual Software Appliance)

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

FreeIPA
(Identity manager solution)

HybriLIT web-site
http://hybrilit.jinr.ru/

Indico
http://indico-hybrilit.jinr.ru

HybriLIT user support
https://pm.jinr.ru/projects/hybrilit-user-support

Monitoring
https://stat-hlit.jinr.ru/

On-line Information for Users

GitLab
https://gitlab-hybrilit.jinr.ru

HybriLIT Overview

Software and Information Environment
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/](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

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Cores</th>
<th>RAM/GB</th>
<th>Disk serves/TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>1000</td>
<td>4240</td>
<td>384</td>
</tr>
<tr>
<td>2017</td>
<td>1400</td>
<td>6160</td>
<td>576</td>
</tr>
<tr>
<td>2018</td>
<td>1800</td>
<td>8080</td>
<td>768</td>
</tr>
<tr>
<td>2019</td>
<td>2200</td>
<td>10000</td>
<td>960</td>
</tr>
</tbody>
</table>

- 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:
\[ \sqrt{s_{NN}} = 4-11 \text{ GeV}; \text{ beams: from p to Au}; \ L \sim 10^{27} \text{ cm}^{-2} \text{ c}^{-1} (\text{Au}), \sim 10^{32} \text{ cm}^{-2} \text{ c}^{-1} (\text{p}) \]
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 ~ 30 PB, 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
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 and analysis software renewal

The program SyMSim (Synthesis of Monitoring and Simulation) for the simulation of grid–cloud structures was developed. Its originality consists in combining a simulation program with a real monitoring system. 

Data storage and processing scheme of Tier0-Tier1 level

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.
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. Modeling 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 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.
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
MICC engineering infrastructure

- Computer Room Air Conditioner
- Fibre Optic & Data Structured Cabling System
- Diesel generator set
- Uninterruptible power supply
- Raised Flooring System
- High Density Heat Containment System
- MICC Monitoring System
- Biometric Access System
- Fire Suppression System
- Surveillance System
- Water Detection System
- VESDA (Very Early Smoke Detection Apparatus)
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
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
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
- Data storage system
- Notification
- Hardware

I. Kashunin, Communication at ROLCG2015
### Informational displays

#### Computing and storage servers
- Wm04-044
- Wm28-284
- Wm10-004
- Wm01-104
- Wm10-109
- Wm11-114
- Wm15-119
- Wm20-004
- Wm20-200
- Wm21-214
- Wm21-219
- Wm22-224
- Wm22-229
- Wm23-234
- Wm23-239
- Wm24-244
- Wm24-249
- Wm25-254
- Wm25-259
- Wm26-264
- Wm26-269
- Wm27-274
- Wm27-279

#### RAID systems
- TDB01-04
- TDB05-09
- TDB01-10
- TDB00-09
- TDB03-09
- TDB00-10
- TDB01-10
- TDB02-08
- TDB03-09
- TDB05-10
- TDB01-11
- TDB02-09
- TDB03-10
- TDB05-11

#### UPS
- UPS-01
- UPS-02
- UPS-03
- UPS-04
- UPS-06
- UPS-12

#### Network
- JINR T2 NETWORK

#### Cooling system
- Process
- Output traffic

#### Computing cluster load
- CPU load

---

I. Kashunin, Communication at ROLCG2015
Thank you for your attention!