TECHNICAL SPECIFICATION
(PASSPORT)

of the object “Computer Unit and Computer Networks of the NICA Complex” (Project of the Computer Unit of the NICA Complex)

The present document is an independent section of the Technical Project of the NICA Complex. The passport of the project of the NICA complex is available at http://nica.jinr.ru/megaproject.

25 december 2019
Under the general editorship of Director of the Laboratory of High Energy Physics (VBLHEP) JINR V. Kekelidze, Deputy Director of VBLHEP JINR Yu. Potrebenikov, Head of the Sector of Information and Network Technologies A. Dolbilov. Responsible for the sections of the document are shown in the table.

<table>
<thead>
<tr>
<th>Section</th>
<th>Editors of the section</th>
<th>Signature</th>
</tr>
</thead>
</table>
| Purpose of the computer unit and its general design | Yu. Potrebenikov, coordinator of work on the creation of the computer unit of the NICA complex  
A. Dolbilov, head of work on the creation of the computer unit of the NICA complex  
T. Strizh, coordinator of work related to the MICC | potreb@jinr.ru  
tel. 21-63441  
dolbilov@jinr.ru  
tel. 21-63429 |
| Computer networks of the NICA complex              | Yu. Minaev, chief technical specialist of SINT                                        | minaev@jinr.ru  
tel. 21-62889 |
| NICA VBLHEP on-line cluster                        | Yu. Minaev                                                                             | minaev@jinr.ru  
tel. 21-62889 |
| NICA VBLHEP off-line cluster                       | B. Shchinov, advisor to the Directorate of VBLHEP                                     | sbg@jinr.ru  
tel. 21-65080 |
| NICA LIT off-line cluster                          | A. Dolbilov                                                                           | dolbilov@jinr.ru  
tel. 21-63429 |
| Off-line cluster of the NICA Center                | A. Dolbilov                                                                           | dolbilov@jinr.ru  
tel. 21-63429 |
| Operating and file systems                         | V. Mitsyn, system administrator of the JINR network  
B. Shchinov,  
Yu. Minaev                                          | vvm@jinr.ru  
tel. 21-64317  
sbg@jinr.ru  
tel. 21-65080  
minaev@jinr.ru  
tel. 21-62889 |
| Application software                               | O. Rogachevsky, software coordinator                                                  | rogachevsky@jinr.ru  
tel. 21-65015 |
| Infrastructure for the creation and operation of the computer unit | A. Dolbilov  
B. Shchinov  
Yu. Minaev                                          | dolbilov@jinr.ru  
tel. 21-63429  
sbg@jinr.ru  
tel. 21-65080  
minaev@jinr.ru  
tel. 21-62889 |

The present document contains a description of the capabilities of the NICA complex computer unit, the technical specifications of its main systems and subsystems, as well as a description of their elements.
GENERAL INFORMATION

Customer – Joint Institute for Nuclear Research (JINR).

The computer unit and computer networks of the NICA complex (hereinafter referred to as the Computer Unit) are implemented by JINR; the Laboratory of High Energy Physics (VBLHEP) with the support of JINR services has the main obligations for its implementation; contractor organizations are involved in the implementation.

The computer unit is distributed geographically and functionally. Its main technological elements are located in four specialized rooms, three of which are situated at the site of the Laboratory of High Energy Physics (VBLHEP), one is in the Laboratory of Information Technologies (LIT) at the DLNP site (see Passport of the “NICA complex” project).

Total power consumption:
- at the stage of assembly and setup - 0.6 MW;
- after commissioning - 1.9 MW.

The construction of the Computer Unit is performed in three stages:

• first stage - “Starting configuration of the Computer Unit”, ends with the construction of prototypes of On-line and Off-line Clusters;

• second stage - “Basic configuration of the Computer Unit”, is implemented after the construction of the basic elements of distributed On-line and Off-line clusters;

• third stage - "Design configuration of the Computer Unit", is determined by the complete configuration of computing, disk, tape and network systems necessary for the full functioning of information and computing resources of the NICA complex.
CONTENT
1 Purpose and general structure of the information and computing (computer) unit of the “NICA Complex” project
2 Basic computing network of the computer unit of the NICA complex
   2.1 Structure and purpose of the computer network of the NICA complex
   2.2 Infrastructure of the network of the distributed computer unit of the NICA complex
3 Main computing elements of the computer unit of the NICA complex
   3.1 NICA on-line cluster
   3.2 Special solutions for building the off-line cluster of the NICA complex
   3.3 NICA VBLHEP off-line cluster
   3.4 NICA LIT off-line cluster
   3.5 Off-line cluster of the NICA center
4 File system of the NICA complex
5 Software of the computer unit of the NICA complex
   5.1 Software of the BM@N, MPD, SPD experiments
   5.2 Monitoring the operation of the main elements of the computer unit of the NICA complex
6 Infrastructure for the creation and operation of the computer unit of the NICA complex
7 Plan for the creation of objects of the computer unit of the NICA complex
1. **Purpose and general structure of the information and computing (computer) unit of the “NICA Complex” project**

The information and computing (hereinafter referred to as computer) unit of the “NICA Complex” project is aimed to carry out modeling of experiments, to process, analyze and store data obtained from facilities and the accelerator unit of the NICA complex, as well as to monitor the operation of the complex itself. The main physical problems, for the solution of which it is designed and created, are determined and described in the Passport of the “NICA Complex” project (see http://nica.jinr.ru/megaproject.php).

The solution of scientific problems, at which the “NICA Complex” megaproject is aimed, is impossible without using novel achievements and developments in the field of computer and telecommunication technologies, high-performance computing systems and programming, and requires the creation and development of a distributed heterogeneous grid-cloud information and computing infrastructure designed to process, analyze and store petabyte flows of experimental data. Considering that the work on the NICA project is performed and will be performed within wide international cooperation, such an infrastructure should provide informational access to accumulated data not only at the Joint Institute for Nuclear Research (hereinafter referred to as JINR or the Institute), where the NICA complex is created, but also for all organizations participating in the megaproject.

The computer unit of the NICA complex is created as part of the existing information and computing infrastructure of JINR (Fig. 1.1). The computing part of this unit at JINR is distributed geographically and functionally. Its main technological elements (Fig. 1.2) are located in four specialized rooms, three of which are situated at the site of the Laboratory of High Energy Physics (hereinafter referred to as VBLHEP), one is in the Laboratory of Information Technologies (hereinafter LIT) at the site of the Laboratory of Nuclear Problems (hereinafter DLNP). The VBLHEP on-line cluster is located in Building 14, its prototype is situated in Building 201; the VBLHEP off-line cluster is located in Building 216, room 115, its prototype – in Building 215, the off-line cluster of the NICA center is situated in the new building of the NICA center. At the DLNP site, the LIT NICA off-line cluster is located in Building 134 as part of the current Multifunctional Information and Computing Complex (MICC) of JINR (see the general view of the MICC in Fig. 1.3).

The basic component of this unit is the data transmission computing network, being created within the JINR local area network, including the elements of its interaction with technological networks of the facilities and individual computers of the complex, with networks of other organizations, involved in the implementation of the “NICA Complex” project, at the required speeds. High-speed channels of this network are also shown schematically in Fig. 1.2.

The JINR local area network integrates the campus network and the data transmission network. The campus network is based on Cisco Catalyst c9500 with 100 Gb/s ports, the data transmission network is built on Cisco Nexus 9504 ACI (Application-Centric Infrastructure) Multisite Fabric with 100 Gb/s ports with the possibility of cloud-based network management. The network covers all laboratories, services and subdivisions of the Institute; it is supported by LIT specialists and system administrators of the network of individual independent subdivisions of the Institute. The network can be upgraded to a speed of 8 x 100 Gb/s.
Fig. 1.1. Scheme of the JINR local area network.

Fig. 1.2. Scheme of the distributed computer unit of the NICA complex.

The maximum amount of data for processing, storage and analysis, obtained from simulation and experimental facilities of the NICA complex, in its **basic configuration** in 2020-2021 amounts to 50 TB/24 hours and 3-10 PB per year, and in the **design configuration** of the complex (starting from 2023) will be up to 300 TB/24 hours and 30-70 PB per year. Correspondingly, resources of the computer unit of the NICA complex should be sufficient to effectively work with such volumes of data.
Fig. 1.3. General view of the JINR MICC.

Data on the main subsystems of the complex, used to calculate the major characteristics of the computing unit of the NICA complex, are shown in Table 1.1.

<table>
<thead>
<tr>
<th>Subsystems of the NICA complex</th>
<th>Technological data acquisition rate (GB/s)</th>
<th>Event generation rate (kHz)</th>
<th>Event size (MB)</th>
<th>Total event size per second (GB/s)</th>
<th>Average data transmission rate (Gb/s)</th>
<th>Data volume per day (TB/24 hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerators</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2019 - 2020</td>
<td>0,5</td>
<td>0,1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;2020</td>
<td>1,5</td>
<td></td>
<td></td>
<td>0,3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>BM@N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2019-2021</td>
<td>30</td>
<td>0,5</td>
<td>15</td>
<td>20</td>
<td>100</td>
<td>300</td>
</tr>
<tr>
<td>&gt;2021</td>
<td>50</td>
<td>0,7</td>
<td>35</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021-2022</td>
<td>0,1</td>
<td>1</td>
<td>0,1</td>
<td>10</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>&gt;2022</td>
<td>6</td>
<td>2</td>
<td>12</td>
<td>100</td>
<td></td>
<td>600</td>
</tr>
<tr>
<td>SPD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;2023</td>
<td>50</td>
<td>0,5</td>
<td>25</td>
<td>100</td>
<td></td>
<td>1000</td>
</tr>
</tbody>
</table>
2. **Basic computing network of the computer unit of the NICA complex**

2.1. **Structure and purpose of the computer network of the NICA complex**

The computer network of the computer unit of the NICA complex is one of its uppermost components. Integrated into the JINR local area network, it combines technological and computing clusters of the complex, servers inside the clusters and the computer network of VBLHEP physical buildings (1, 1A, 1B, 2, 4, 14, 17, 32, 42, 201, 202, 203A, 203B, 205, 215, 216, 217). The central connecting element of the JINR local area network, i.e. the module Cisco Catalyst c9500 capable of providing data transmission within the network and with external computing networks at a speed of up to 400 Gb/s with a possible expansion of up to 800 Gb/s, is located in Building 134 LIT; the module Cisco Nexus 9504 ACI (Application-Centric Infrastructure) Multisite Fabric is situated in the on-line cluster at the VBLHEP site.

When organizing a computer network of the NICA complex, interfaces with transmission rates of 100 Gbit/s are used in its on-line and off-line components, providing a bandwidth of 400 Gbit/s and higher with the use of several parallel N x 100 Gbit/s connected channels, which ensures additional reliability and fault-tolerance in general.

Main goals and objectives of the computer network of the NICA complex:

- creation of a unified information space for access to computing and information resources, as well as to data storage systems of the complex;
- creation of a unified information space for the participants of the NICA project, providing the possibility of data exchange between JINR subdivisions and other organizations involved in the project;
- organization of efficient and secure access to both centralized resources of the Institute’s IT structure and resources of individual subdivisions;
- provision of high-speed access to Internet resources;
- provision of maximum reliability and performance of the network;
- support of general and special network services, such as authorization and authentication, e-mail, DNS, the proxy-service, the IPDB network elements database, NMIS systems for monitoring the status of active network elements of the infrastructure.

The equipment of the central telecommunication node, i.e. the core of the switching and routing system, is implemented on two multifunctional switches Cisco Nexus 9504 ACI Multisite Fabric with a full-mesh topology for maximum reliability and performance (Fig. 2.1). The ACI Multisite Fabric provides network security (right levels of control and protection against malicious agents) regardless of how or where users connect. Other nodes of the computing network of the NICA complex, of on-line and off-line clusters, of experimental facilities use switches of the Cisco Nexus family of the same or lower models.
Fig. 2.1. Scheme of interaction between the networks of VBLHEP and LIT.

Seven nodes for all data sources of the NICA complex are organized on the basis of the ACI technology:

• ACI Site on-line VBLHEP
• ACI Site off-line VBLHEP
• ACI Site off-line of the NICA Center
• ACI Site MPD
• ACI Site BM@N
• ACI Site SPD
• ACI Site off-line LIT

Such a scheme is aimed at combining physical and virtual computing infrastructures and creating an environment that is automated, scalable, programmable, cost-effective and meets both the application requirements and network security.

To create a reliable computer network of the NICA complex, which meets modern standards and requirements, the following work must be done:

− its core is built;
− a telecommunication channel between the sites, providing a data transmission rate of 400 Gb/s with the ability to easily expand to 800 Gb/s, is launched and put into operation;
− switches with access to the ACI Multisite Fabric are installed;
− preparatory work to connect all elements of the ACI MultiSite network cluster is done;
− the VBLHEP off-line and LIT off-line computing clusters are connected at a speed of 200 Gb/s.

Computer complexes of the systems for data acquisition and management of individual basic elements of the NICA complex, such as the accelerator unit, the experimental facilities BM@N, MPD, SPD, as well as prototypes of the detectors and subsystems used in them, are combined into technological networks of these units, being technological networks of the computer network of the
NICA complex, and have limited access to the JINR local area network via the Ethernet technology through their direct connection to the ACI Multisite Fabric as shown in Fig. 2.1. Monitoring and control over their functioning are performed at the point of connection of the technological networks.

The computer network of the NICA information and computing complex refers to the basic configuration of the NICA complex.

2.2. **Infrastructure of the network of the distributed computer unit of the NICA complex**

The most important stage in the development of the network infrastructure of the NICA complex is the organization of optical backbones between the sites of DLNP and VBLHEP. To ensure maximum reliability and fault-tolerance of the given network, it is necessary to geographically separate two independent backbone routes for organizing communication between on-line and off-line objects of the “NICA Complex” project and the JINR MICC at speeds of N x 100 Gb/s.

Fig. 2.2 illustrates a plan for the construction of two physically independent optical backbones between the DLNP and VBLHEP sites.

An optical cable infrastructure with the ability to connect the network equipment via two geographically independent routes is being built at the VBLHEP site. Both the existing scheme of underground telephone routes (dash-dotted lines) and newly constructed sections (blue lines) are presented in Fig. 2.3.

The infrastructure of the computer network of the NICA complex should provide its maximum reliability, performance and fault-tolerance by creating special optical telecommunication channels and a system of underground telephone routes. For this purpose, an optical cable infrastructure should be built at the VBLHEP site, and an optical cable infrastructure should be created between the DLNP and VBLHEP sites.
Fig. 2.2. Scheme of two optical telecommunication channels for the computing network of the NICA complex.

Fig. 2.3. System of underground telephone routes.
3. Main computing elements of the computer unit of the NICA complex

As all large-scale centers for processing and storing data from large physical facilities, the computer unit of the NICA complex comprises geographically distributed on-line and off-line clusters connected by a unified high-speed local area network. In the basic configuration, this unit will include one on-line cluster and three off-line clusters: the NICA VBLHEP off-line cluster, the NICA LIT off-line cluster and the off-line cluster of the NICA center, geographically located at the VBLHEP and DLNP sites.

3.1. NICA on-line cluster

The on-line cluster has several primary and secondary functions. The primary function is to get data from the NICA data acquisition system (DAQ). The function of not less importance is to acquire data for monitoring and diagnostics of the main units and physical facilities of the NICA complex.

Goals and objectives of the NICA on-line cluster:

— obtaining information from systems of acquiring data of physical facilities;
— sorting and packaging “raw” data;
— express data processing (not more than 5~10% of the total volume);
— temporary storage of “packed” data (no more than 24 hours);
— data transfer for further storage and processing on off-line clusters;
— acquiring data for monitoring and diagnostics of the physical facilities functioning.

The data flow from physical facilities is heterogeneous in data transmission rates from 1 to 100 Gb/s with an average flow of 10 Gb/s. To acquire and store data on the NICA on-line cluster, a data storage system (DSS) built on fast SSD NVMe disks is required, since “spindle” disks will not be able to acquire such a data flow without delay.

Based on the maximum storage time, which is 24 hours, and an average flow rate of 10 Gb/s, the volume of temporarily stored data on the NICA on-line cluster amounts to 0.5 PB.

When organizing the network computer infrastructure of the NICA on-line cluster, it is necessary to ensure maximum bandwidth, reliability and fault-tolerance of the given network. All network computer elements of the NICA on-line cluster will be connected by interfaces at a speed of 100 Gb/s and interconnected by several links as shown in Fig. 3.1.
3.2. Special solutions for building the off-line cluster of the NICA complex

Increasing the efficiency and performance of the NICA computer unit is possible using special equipment and optimization of exchange protocols. The main exchange technology, instead of traditional TCP/IP exchange, becomes RDMA (remote direct memory access), which allows one to transfer data between servers directly from the memory of one application to the memory of another without the participation of central processors. Until recently, RDMA was available only in InfiniBand structures. However, a new technology RoCE (RDMA over Converged Ethernet) has been developed, and the benefits of RDMA are now available to data centers that rely on Ethernet equipment. RoCE is a technology of efficient data transmission with very low latency in Ethernet networks without losses. One of the leading manufacturers of equipment compliant with the RoCE requirements is the Mellanox company. It produces a complete set of such equipment:

- switches SN2700/SN2100/SN2410 with 32/16/8 100 GbE ports;
- ConnectX-4 adapters 100GbE;
- optical and cooper 100GbE cables;
- software components (drivers) for operating systems.

The Mellanox 100GbE equipment and operating systems with integrated RoCE are used in the NICA VBLHEP off-line cluster and the NICA center off-line cluster. Fig. 3.2 shows the comparison of delays in data transmission via the RoCE and TCP/IP protocols.
To reduce the cost of expensive 100 Gb/s equipment, the Mellanox company developed a new technology, i.e. MultiHost Technology. On the basis of the ConnectX-4 interface, a device was manufactured; it multiplexes the PCIe bus into four completely independent PCIe buses, which are inserted into servers instead of expensive Ethernet adapters using inexpensive repeaters. The scheme of their connection is shown in Fig. 3.3.

Each node can be active or inactive at any time, regardless of other nodes, and receive its own bandwidth. Full bandwidth is shared between nodes either uniformly (by default) or based on the configurable quality of service (QOS), depending on the needs of the data center. Thus,
expenses are significantly reduced, as instead of four NICs cables and switch ports, only one each is used, and overall power consumption is also reduced. The Mellanox MultiHost device is illustrated in Fig. 3.4.

![Mellanox MultiHost](image)

**Fig. 3.4. Device Mellanox Multi-Host.**

### 3.3. NICA VBLHEP off-line cluster

The NICA VBLHEP off-line cluster is aimed at processing, storing and analyzing data from experimental setups of the NICA complex. At the VBLHEP site, it is located in Building 216, room 115, and connected with the N x 100 Gb/s local area network as all components of the NICA computer unit. To balance the load on the entire Ethernet segment of the computer network, the NICA VBLHEP off-line cluster is structured by access speeds. The scheme of its connection, as well as the connection of other off-line clusters of the NICA complex, is presented in Fig. 3.5.

![Diagram of Ethernet structured network](image)

**Fig. 3.5. Scheme of the Ethernet structured network of the NICA VBLHEP off-line cluster.**

The NICA VBLHEP off-line cluster (Fig. 3.6) is designed for 5000 CPU cores and 10 PB of disk arrays. The cluster consists of four identical modules, each of which consists of two racks (cabinets) with servers and an air-water cooling device (LCP) between the racks. To create a “cold corridor”, the modules stand in two lines. Above, between the racks, there is a glass roof 1200 mm
wide and side glass doors. The equipment in the racks receive power from two uninterruptible power supplies (UPS), which, in turn, receive power through cables from the switchboard. There are also temperature, humidity, leakage sensors and a GSM unit. The arrangement of the equipment in Building 216, room 115, is shown in Fig. 3.7.

![Fig. 3.6. NICA VBLHEP off-line cluster.](image1)

To reduce the load on chillers or in case of their failure, the general air-conditioning system with the installation of two-channel inverter split air conditioners Toshiba Digital Inverter RAV-SM2804AT8-E with a total cooling capacity of 50 kW is provided. In addition, the internal free cooler of the chiller must operate with a turned-off or failed compressor.

The parameters of the NICA VBLHEP off-line cluster are shown in Table 3.1.

![Fig. 3.7. Arrangement of the equipment in Building 216, room 115.](image2)
Table 3.1.
Parameters of the NICA VBLHEP off-line cluster

<table>
<thead>
<tr>
<th>Term</th>
<th>Required parameters</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>End of 2017</td>
<td>0.5PB disks (with replication), 1K CPU cores</td>
<td>NICA VBLHEP off-line cluster</td>
</tr>
<tr>
<td>End of 2019</td>
<td>4-5PB disks (with replication), 4K CPU cores</td>
<td>NICA VBLHEP off-line cluster</td>
</tr>
</tbody>
</table>

3.4. **NICA LIT off-line cluster**

The NICA computing and information off-line cluster in LIT is organized on the basis of the JINR Multifunctional Information and Computing Complex as a distributed scalable hybrid cluster. The given approach allows one to organize computing for the NICA complex efficiently and without additional labor costs at the request of a different class of tasks and users. The main objective of the NICA LIT off-line cluster is to create a two-layer (disk and tape) storage system for the NICA experiments, since after the first stage of these experiments significant storage volumes (from 2.5 PB to 70 PB per year) will be required.

The JINR Multifunctional Information and Computing Complex currently has the following main components:

- JINR Central Information and Computing Complex (CICC) with built-in computing and memorizing elements;
- Tier2 cluster for all experiments at the Large Hadron Collider (LHC) and other virtual organizations (VO) in the grid environment;
- Tier1 cluster for the CMS experiment;
- HybridLIT heterogeneous platform for high-performance computing (HPC) with the “Govorun” supercomputer;
- cloud infrastructure;
- data storage system based on the EOS file system;
- tape robotized library based on dCache.

The NICA LIT computing and information off-line cluster is organized on the basis of Tier1, Tier2/CICC, the “Govorun” supercomputer, the EOS-based storage system and dCache. Due to the fact that the developed computing models, taking into account current trends in the development of network solutions, computing architectures and IT solutions, allow combining supercomputer (heterogeneous), grid and cloud technologies, and creating distributed software-configurable HPC platforms on their basis.

The implementation of different computing models for the “NICA Complex” megaproject requires the confirmation of their performance, i.e. meeting the requirements for the temporal characteristics of acquiring data from detectors with their subsequent transfer to processing, analysis and storage, as well as the requirements for the efficiency of event simulation and processing in the experiment. For these purposes, it is necessary to carry out tests in a real software and hardware environment, which should contain all the required components. The “Govorun” supercomputer commissioned as part of the MICC, containing the most up-to-date computing and data storage...
resources, including the ultrafast data storage system that provides a high speed of data acquisition up to hundreds of gigabytes per second, with the possibility of linear expansion of system performance and capacity, can become such an environment.

The “Govorun” supercomputer is used as part of the NICA LIT distributed and scalable off-line cluster to solve problems requiring massive parallel calculations in lattice quantum chromodynamics for the study of properties of hadron matter at high energy density and baryon charge and in the presence of supramaximal electromagnetic fields, for mathematical modeling of interactions of antiprotons with protons and nuclei using DPM, FTF and UrQMD+SMM generators developed at JINR and of interest for the NICA-MPD experiment, for the simulation of dynamics of collisions of relativistic heavy ions. In addition, the work related to the development of computing for the “NICA Complex” megaproject is carried out on the basis of the supercomputer.

Another component of the NICA LIT off-line cluster is the ultrafast data storage system (UDSS) implemented in the “Govorun” supercomputer under the Lustre file system. UDSS has 12 data storage servers with 12 SSD drives with the NVMe connection technology, which reduces the time of access to data. The total capacity of UDSS is currently 256 TB and the data acquisition/transfer rate is up to 200 GB per second. It is noteworthy that UDSS has the ability to linearly increase its productivity (speed of working with data) and the storage volume without changing the principles of the architectural design of the system. At the same time, hyperconvergence and the software-defined architecture of UDSS allow providing maximum flexibility of data storage system configurations, including creating data storage domains, which meet the requirements of specific tasks, as well as connecting external disk fields to expand and develop the UDSS capacities.

The existing data storage system in LIT, on the basis of which the system of the NICA LIT off-line cluster will be organized, is illustrated in Fig. 3.8.

Fig. 3.8. Data storage system of the LIT cluster on tapes (left) and ultrafast disk memory system (right).

At the first stage of modeling computing for the “NICA Complex” megaproject, it is supposed to use computing resources of the “Govorun” supercomputer, the Skylake and KNL sections to
generate data, save them to UDSS and transfer to “semi-cold” storages managed by the EOS and ZFS file systems (Fig. 3.9). It will allow checking the basic stack of data storage and transfer technologies, as well as modeling data flows, choosing optimal distributed file systems and increasing the efficiency of event modeling and processing. This work has been carried out to simulate events of the MPD experiment, in addition, it is planned to use the DIRAC software to manage jobs and the process of reading out/writing data from different types of storages and file systems (Fig. 3.9).

Fig. 3.9. Data transfer scheme on the “Govorun” supercomputer to model calculations for the “NICA Complex” megaproject and simulate events for the MPD experiment, which was implemented using the DIRAC software.

More than 85% of the total number of jobs are carried out on the “Govorun” supercomputer for the NICA experiments. Fig. 3.10 shows the distribution of jobs performed on the supercomputer within the MPD, BM@N and SPD projects.

Fig. 3.10. Distribution of jobs performed on the supercomputer within the MPD, BM@N and SPD projects.
The required sets of events have been generated and reconstructed for the MPD experiment using different types of generators. To solve this task, the CPU and UDSS components have been used.

In addition to the computing resources of the “Govorun” supercomputer, computing resources allocated for NICA on the Tier1 and Tier2 components of the MICC were integrated on the basis of DIRAC, and there are also procedures for writing and reading data stored on the tape robot.

The DIRAC Interware is a product for integrating heterogeneous computing and data storage resources into a unified platform. Resource integration is based on the use of standard data access protocols (xRootD, GridFTP, etc.) and pilot jobs. Thanks to this, a unified environment, in which is possible to run jobs, manage data, build processes and control their implementation, is provided to users. In the framework of DIRAC, batch processing systems, grid computing elements, clouds, supercomputers and even separate computing nodes can act as computing resources. Storage resources are limited only to those that support file transfer protocols used in grid systems.

Pilot jobs are an important concept in DIRAC. One can integrate almost any computing resource with their help. The idea of a pilot job is that a “pilot” job, which is always the same, is sent to the computing resource queue instead of user jobs. Only the way pilots can be run on different resources can differ. Once on the working node, the pilot loads basic data utilities, checks the environment and asks the DIRAC central service for the job that the resources occupied by the pilot would do. Having received the job, the pilot transfers control to it. After completing the job, the pilot can either ask for the next job or complete its work. Having a pilot guarantees the job some basic set of functions with which the job can work. If the resource on which the pilot works cannot provide these functions, the job will not even start and the error of the pilot launch will not hinder the user. Due to it, users see only errors related to their jobs, and only administrators should worry about the pilots not starting on the resource.

When working with data, DIRAC provides all the required set of commands. For the correct operation of all commands, the storage system should support grid data transfer protocols. It allows the pilot to acquire data on any of the resources and download them back. However, the user can always send his job to the specific computing resource and then work with the local file system on the resource. In this case, additional efforts may be required to make data accessible from anywhere. The integration of computing resources and different data storage systems of the JINR MICC is presented schematically in Fig. 3.11.
With the help of DIRAC, the computing resources of the JINR MICC, i.e. Tier1/Tier2, the “Govorun” supercomputer, the JINR cloud, and storage resources, such as UDSS Lustre, dCache and EOS, are combined. As part of Monte-Carlo data generation for the NICA experiments, jobs are performed on the MICC Tier1/Tier2 components using the DIRAC platform and on the computing resources of the “Govorun” supercomputer using UDSS. Results of data generation are sent to dCache.

In the basic configuration, the NICA LIT hybrid off-line cluster is organized on the basis of the MICC resources and includes 500 CPUs (slots in batch) on the Tier2/CICC component of the MICC and 688 cores on the Tier1 component. 500 TB is reserved on the MICC EOS storage for the MPD experiment, 300 TB for BM@N and 250 TB for SPD. The capacity of the tape robot in the basic configuration is 40 PB, in which 16 PB is reserved for the NICA complex.

The main objective for the complete configuration of the NICA computer unit is to combine computing resources and storage/data access resources from all off-line clusters to conveniently use for the needs of the NICA experiments. It is supposed to use resources located in two JINR laboratories, i.e. LIT and VBLHEP, for modeling, reconstruction and analysis of the NICA experiments. For convenient job launch and data access, the existing infrastructure requires refinement, modification and unification of different components of user authentication and authorization when accessing different services of data processing systems. It is planned to expand the use of resources with the possibility of attracting the resources of organizations that are somehow interested in accessing data from the NICA experiments.

In the basic configuration, the following stages are used to integrate the LIT and VBLHEP computing resources into a unified processing system of the NICA experiments:

1st stage. Integration based on DIRAC. At this stage, access to data is possible in two ways:
1. using the xRootD or GridFTP protocols and the x509 user certificate, the user can access data from dCache and EOS.

2. Local access to data with pseudo-user rights.

2nd stage. Installation and configuration of CVMFS, i.e. unified storage location for all versions of software of the NICA complex experiments.

3rd stage. Integration with JINR Kerberos/LDAP. This procedure significantly simplifies the work with data in EOS, data are available for reading and writing without copying, the job sees EOS storage as a locally mounted space.

4th stage. Consolidation of the MICC LIT and VBLHEP EOS storages. Areas for storing data of three NICA experiments, i.e. BMN, MPD and SPD, are distributed in MICC EOS. Data are stored with two replicas on different EOS disk servers. In the future, one of the replicas will go to EOS servers in VBLHEP, which will make it possible to access the nearest replica of jobs launched at different JINR computing complexes.

The order of numbering the stages does not apply their strictly sequential implementation. Stages 1, 2 and 3 can be performed in parallel. However, the implementation of the 4th stage requires the completion of the 3rd one.

3.5. Off-line cluster of the NICA center

The off-line cluster of the NICA center is designed for modeling, processing and storing data obtained from the experiments of the NICA complex. The NICA center is located in the new building of the NICA complex (Fig. 3.12).

Fig. 3.12. General view of the NICA center building.

To achieve the required cluster parameters (20 PB disks, 20K CPU cores), two methods for constructing computing complexes, i.e. with air and liquid cooling, will be implemented together.
The first one allows placing equipment with a total capacity of up to 25 kW in one rack. The second one, using the development of companies such as RSK (Russian Supercomputers), allows placing equipment with a capacity of up to 100 kW in one rack. The general view of the computer of the second type is presented in Fig. 3.13. Up to 112 (7 blocks of 16 modules) two processor computing modules (224 processors, up to 4928 CPU cores) with direct liquid cooling using a cooling plate, a water-air subsystem and an external free cooler, can be installed in the cabinet 42U. This architecture does not require the installation of air conditioning systems, has high energy efficiency and low noise. Two SSD disks can be installed in each module. It is planned to install only one disk for the system in the module, and the necessary large disk space will be provided by the installation of a modern data storage system (DSS).

It should also be noted that the second option for the NICA center cluster requires connection to the new building and a detailed study of the cooling and power supply system.

Fig. 3.13. View of the element of the NICA center off-line cluster.

One of the successfully used DSS is the Seagate ClusterStor L300N/L300/L6000/L9000 series. ClusterStor L300N (Fig. 3.14) is a parallel scalable hybrid system that uses both hard and solid state drives; it is optimized for working with variable-length units. The given DSS is a full-featured solution that includes all the necessary hardware and software components. The system is located in a standard rack 42U and can include up to seven (first) and eight (second) units with 82 disks of maximum (at present) capacity, i.e. for 8 TB disks, one unit can contain approximately 0.5 PB of data, and more than 1 PB of data and almost 10 PB in one rack for 16TB disks.
The planned parameters of the NICA center off-line cluster from 2021 to 2023 with the necessary server equipment are shown in Table 3.2. The equipment needed to achieve these parameters is designed under the assumption that 2.5” SSD disks will have a capacity of 16 TB. The number of CPU cores is 24. The data are given for two options of the cluster equipment: option 1 – based on the quadro-twin architecture (High Density – four dual processor modules in a server 2U high), option 2 – based on the architecture of a mini data center from RSK.

### Table 3.2.

<table>
<thead>
<tr>
<th>Term</th>
<th>Required parameters</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>End of 2021</td>
<td>8-10PB disks (with replication),</td>
<td>NICA Center</td>
</tr>
<tr>
<td></td>
<td>5K CPU cores</td>
<td></td>
</tr>
<tr>
<td>End of 2022</td>
<td>12-15PB disks (with replication),</td>
<td>NICA Center</td>
</tr>
<tr>
<td></td>
<td>10K CPU cores</td>
<td></td>
</tr>
<tr>
<td>End of 2023</td>
<td>20PB disks (with replication),</td>
<td>NICA Center</td>
</tr>
<tr>
<td></td>
<td>20K CPU cores</td>
<td></td>
</tr>
</tbody>
</table>

The hardware of the computer unit of the NICA center will be located in the new NICA center building and will occupy an area of about 200 m². The computing unit will have its own 800 kW power supply system, climate control and automation systems. The computing network of the computer unit of the NICA center is implemented on telecommunication equipment at N x 100 Gb/s. The design and construction of the “NICA Center” building is carried out in accordance with the tasks set forth in the Passport of the object “NICA Complex” ([see](http://nica.jinr.ru/docs/TDR_spec_Fin0_for_site_short.pdf)).

4. **File system of the NICA complex**

One of the main components of the computer unit of the “NICA Complex” project is a cluster file system. Currently, several systems are used in this capacity: GPFS, Lustre, dCache, Ceph, EOS, GlusterFS, etc. The EOS file system specially developed for high-energy physics is the best option.
for on-line and off-line clusters. EOS is a distributed, parallel, linearly scalable file system with the possibility of protecting it from failures by replicating data.

The EOS structure integrated into the MICC structure is presented in Fig. 4.1. EOS is displayed as a local file system on the MICC working nodes and allows authorized users to read and write data.

Special attention is paid to novel perspective directions in creating distributed data storages (DataLake), integrating Big Data and supercomputer technologies, methods of machine learning. The fundamental problem is to create a geographically distributed infrastructure, i.e. scientific data lakes, based on supercomputer centers, resource centers of high-performance computing (grids) and cloud centers, which allows one to deploy an environment in which the computing infrastructure looks logically unified for the end user. To level the differences in architectural solutions of computing centers, one needs a system of managing flows of tasks for processing and analyzing data. In addition, the lake infrastructure requires a monitoring and control system, as well as tracking the functioning of the distributed computing infrastructure using classical methods of visual analytics.

Basic principles of the architecture of a scientific data lake:

- logical separation of the computing infrastructure and data storage;
- presence of high-level services (control of job flows, load and data) that interact with all elements of the infrastructure and manage the use of resources;
- presence of the hierarchical geographically distributed structure from regional data lakes of different sizes with a specific network topology, internal balancing mechanisms, support for integrity and necessary data redundancy;
- presence of “smart” services for data transfer between all components of the infrastructure, as well as services for defining and predicting the amount of computing resources required to perform job flows.

Fig. 4.1. Scheme of integrating EOS in the LIT MICC.
In the design configuration, a significant increase in the amount of information, which is needed to be stored and processed, is expected. At the same time, it is very difficult to evaluate the requirements to storage systems due to the evolution of data and processing models. However, the following requirements should be met:

- to provide a sufficient resource for storage and fast access to information during processing;
- to provide a constantly expandable resource for long-term data storage, the volume and speed of which should be balanced with the corresponding flows of information;
- to provide the ability to use a data management system that automates the processes of interaction with storage systems;
- to automate support for the storage system to optimize and minimize costs.

In the coming years, it is necessary to significantly increase the information storage capacity on the EOS system. This data storage and access system should become the major system for all components of the NICA computer complex and, in the future, for all JINR computing resources. The system is already being used to store data from BM@N. Storing uppermost data or data distributed geographically requires the creation of replicas, which reduces the available space multiple to the number of replicas.

One of the options for developing storages is to combine and import local site installations into a “Data Lake”. Such projects are already at the development and testing stage.

Due to the significant increase in the needs of the NICA experiments and the extended use of EOS/MICC, it is necessary to plan a substantial growth in the storage capacity. For long-term storage, it is expected to use tape libraries (tape robots) with the possibility of expanding, if necessary, up to 40-50 PB/year.

Software should be placed in a cached file system so that it is available in all data processing centers.

Storage systems (Fig. 4.2) are installed in the dCache software. One of the dCache installations is used only with disk servers and for online data storage with fast access to them. The second cache unit comprises disk servers and a tape robot. The disks serve as a buffer zone for exchange with tapes, while the tape robot is designed for long-term, almost eternal storage of data from the NICA complex facilities. In total, the facilities have 3.4 PB of effective disk space, and the tape robot has 40 PB of storage capacity. To support storage and access to data 8 physical and 14 virtual machines are required.
5. Software of the computer unit of the NICA complex

At present, the process of authentication and authorization of users of the NICA complex is unified on the following LIT MICC resources: the CICC/Tier2 computing cluster, the Tier1 cluster, cloud infrastructure resources, EOS data storage and access resources, CVMFS software update and access resources. Access to all of the given resources is authorized through Kerberos5 protocols for all users registered in the Kerberos and LDAP databases of JINR. This access option is most convenient for data processing and modeling, as it does not require any additional data manipulations from the user (copying from EOS storage to local disks, rewriting from local disks to EOS).

An important component of the computing and information system for the NICA complex is the availability of a unified common software repository. For these purposes, in 2019, the CVMFS subsystem was installed. The NICA MPD and BM@N software packages are placed in CVMFS. The repository stores several versions of the same software package, and only an authorized person can change it on the CVMFS server. Read access from the repository is configured on all MICC working nodes, and it can be used around the world using the CVMFS client. The MPD and BM@N software repository occupies 9.5 GB on the CVMFS MICC servers.

Software for modeling, reconstruction and analysis of particle physics data is an important part of the computer unit of the NICA megaproject. For experiments on the BM@N, MPD and SPD facilities, software frameworks based on the Root package, such as BmnRoot, MpdRoot and Spdroot respectively, have been developed or are being developed and will be used.

5.1. Software of the BM@N, MPD, SPD experiments

The purpose of such computing systems is to provide modeling of primary particles, ion interactions and the resulting response of the detector, reconstruction and analysis of data from simulated and real interactions. When constructing detectors, the optimization of the hardware
design, preparation of the code and the computing infrastructure require a reliable chain of modeling and reconstruction implemented by the distributed computing environment. The basics of BmnRoot, MpdRoot and SpdRoot are used to carry out research to simulate the technical design of all subdetectors of BM@N, MPD and SPD in order to optimize their constructions. They are used to model physical processes occurring in the detectors, as well as to study the feasibility of physical problems on the facilities.

5.2. Monitoring the operation of the main elements of the computer unit of the NICA complex

Modern high-energy physics is directly related to the modeling and storage of large amounts of raw data. It requires the availability of extensive computing power that can provide operational processing and data storage. To increase the fault-tolerance of computing equipment, it is necessary to respond to different failures in a timely manner and quickly eliminate them. The presence of monitoring systems allows reducing the response time to the failure and thus increase the fault-tolerance of equipment, which will improve the reliability of the computing complex as a whole.

The distributed computer unit of the NICA complex should function as a unified structure. This approach should provide the unity of management and access to computing power and other resources. Depending on the capabilities and level of access to servers, there are different approaches to organizing a unified system of its monitoring.

Fig. 5.1 illustrates an approach that includes the organization of a unified management center and its many satellites.

![Diagram of the unified management center](image)

Fig. 5.1. General scheme for building a cluster of the unified management center.

6. Infrastructure for the creation and operation of the computer unit of the NICA complex

Modern computing centers, such as the computing clusters of the NICA complex (NICA on-line cluster, NICA VBLHEP off-line cluster, off-line cluster of the NICA center, NICA LIT off-line cluster) must have not only state-of-the-art software, a sufficient amount of computing resources, a management and data storage system, but also a reliable engineering infrastructure, which ensures the uninterruptible functioning of the center in a 24/7/365 mode. The equipment used to create an
infrastructure must have the necessary level of reliability and maintainability. The engineering infrastructure must be modular, scalable, adaptive and expandable.

The engineering infrastructure comprises:

1. Power supply system, including:
   1.1. General electricity supply and distribution system
       - NICA on-line cluster — 300 kW,
       - NICA VBLHEP off-line cluster — 400 kW,
       - Off-line cluster of the NICA center — 800 kW,
       - NICA LIT off-line cluster — 1600 kW.
   1.2. Uninterruptible power supply system;
   1.3. Guaranteed power supply system;
   1.4. Grounding system.

2. Climate control system, including:
   2.1. Cold supply system;
   2.2. Industrial air conditioning system;
   2.3. Supply and exhaust ventilation system;
   2.4. Smoke and gas exhaust system.

3. Fire safety complexes, including:
   3.1. Fire alarm;
   3.2. Early fire detection system of a VESDA type;
   3.3. Modular gas fire suppression system.

4. Automated dispatching and engineering infrastructure management system, including:
   4.1. Engineering automation;
   4.2. Dispatching system;
   4.3. Basic parameters monitoring system;
   4.4. Technological television surveillance system.

5. Complexes of machine halls of computer equipment in the buildings of the NICA computing complex, including:
   5.1. Structured cabling system (SCS) with the cable manhole system;
   5.2. System of hardware racks and isolation of corridors;
   5.3. System of raised floors and technical trays.

7. Plan for the creation of objects of the computer unit of the NICA complex


The plan for the creation of the computer unit of the NICA complex by years is presented in Table 7.1.
Table 7.1.

Plan for the creation of the computer unit of the NICA complex by years

<table>
<thead>
<tr>
<th>Passport section</th>
<th>Name of the object</th>
<th>Date of completion, year</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.</td>
<td>Computer unit and computer networks</td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>On-line cluster</td>
<td>S</td>
</tr>
<tr>
<td>7.2</td>
<td>VBLHEP off-line cluster</td>
<td>S</td>
</tr>
<tr>
<td>7.3</td>
<td>Off-line cluster of the NICA Center</td>
<td>S</td>
</tr>
<tr>
<td>7.4</td>
<td>LIT off-line cluster</td>
<td>S</td>
</tr>
</tbody>
</table>

(Notation: S – Starting configuration, B – Basic configuration, C – Complete configuration of the object)

Table 7.2 shows the costs for creating objects of the computer unit of the NICA complex by years.

Table 7.2.

Expenses for the information and computer complex in millions of US dollars by years

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Computer unit and computer networks</td>
<td>0.129</td>
<td>0.217</td>
<td>0.365</td>
<td>1.040</td>
<td>7.127</td>
<td>2.123</td>
<td>1.709</td>
<td>6.594</td>
<td>2.807</td>
<td>4.149</td>
<td>26.262</td>
</tr>
<tr>
<td>1.1 On-line cluster</td>
<td>0.037</td>
<td>0.0</td>
<td>0.015</td>
<td>0.053</td>
<td>2.619</td>
<td>0.075</td>
<td>0.850</td>
<td>0.046</td>
<td>0.0</td>
<td>0.0</td>
<td>3.700</td>
</tr>
<tr>
<td>1.2 VBLHEP off-line cluster</td>
<td>0.0</td>
<td>0.047</td>
<td>0.167</td>
<td>0.140</td>
<td>1.229</td>
<td>0.613</td>
<td>0.059</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
<td>3.684</td>
</tr>
<tr>
<td>1.3 Off-line cluster of the NICA Center</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>5.927</td>
<td>2.397</td>
<td>3.786</td>
<td>12.110</td>
</tr>
<tr>
<td>1.4 LIT off-line cluster</td>
<td>0.0</td>
<td>0.001</td>
<td>0.0</td>
<td>0.265</td>
<td>0.281</td>
<td>0.880</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.428</td>
</tr>
<tr>
<td>1.5 Other services</td>
<td>0.087</td>
<td>0.138</td>
<td>0.179</td>
<td>0.344</td>
<td>2.109</td>
<td>0.217</td>
<td>0.800</td>
<td>0.607</td>
<td>0.397</td>
<td>0.350</td>
<td>5.229</td>
</tr>
</tbody>
</table>