Network and computing infrastructure for the NICA complex at JINR.

Technical Design

Version 1.03

A. Dolbilov, Yu. Minaev, V. Mitsyn , A. Nechaevskiy,

Yu.Potrebenikov, D.Pryahina, O.Rogachevsky, B.Shchinov, T.Strizh, V. Trofimov

Laboratory of High Energy Physics
Laboratory of Information Technologies

Joint Institute for Nuclear Research

Dubna

May 15, 2018

ABSTRACT

This document is a part of serial of TDR documents describing sub-systems of the accelerator-experimental complex NICA (Nuclotron-based Ion Collider fAcility), which is realised at the Joint Institute for Nuclear Research (JINR) as a mega-science project with a title "NICA Complex". It describes an organization of the common computing infrastructure for this complex at JINR aimed to physical data accumulation and storage from the basic nodes of the Complex: accelerators, detectors BM@N, MPD and in a future - SPD, experimental setups for innovation and applied researches, to their processing, analysis, monitoring and to simulation. It describes as well a common local network of the NICA Complex, its position in the joint JINR network and its interaction with technological networks and computers of these nodes.

The computing of the NICA Complex at JINR is territorially and functionally distributed. Its main technological elements are located in four specialized rooms, three of which are located at the site of the Laboratory of High Energy Physics (LHEP), one - in the Laboratory of Information Technologies (LIT). Description of the structure, purpose and equipment of each of them is also presented in this document.

The computing for the NICA Complex is based on the modern and affordable technical solutions, is scalable, open for connection of different computer resources located outside of JINR, and is aimed at working with data volumes of 2 - 5 petabytes per year in 2019 - 2020 and 10 - 20 petabytes per year in 2021-2023 years.

Table of content

1. Introduction

One of the most important tasks related to the realization of a mega-project "NICA Complex" [1] is a construction of its computing infrastructure. Solving the scientific problem at which the NICA megaproject is targeted, is impossible without the use of the latest achievements and the development of new techniques in the field of computer and telecommunication technologies, high-performance computing systems and programming, it requires the development of a distributed heterogeneous grid-cloud information and computing system for modeling, processing, analysis and storage of petabyte-scale data streams in the experiments at the NICA complex. Since the activities on the NICA project are underway and will be conducted in frames of the broad international cooperation, there is a need not only to store and process the experimental data at JINR, but also to provide access to them for all the organizations participating in this mega-project.

An example of the current most successful computer systems for processing enormous data in high energy physics is models of data processing of the LHC experiments. Computer models, structures and data flows of each of the 4 LHC experiments have their own peculiarities related to physical tasks solved in the course of the research. However, their common part should be noted which is a technology of distributed computing. This approach was laid when designing the accelerator and facilities, because it was getting clear that the provision of data processing and access to them for the members of collaborations of thousands of scientists, even within such a powerful organization like CERN, looks impossible for many years to come.

The use of the computing models based on the grid-technologies for LHC in frames of WLCG has been incredibly successful for creating an environment that yielded physical results of the experiments at the LHC. Nevertheless, for nearly 10 years that passed since its creation, the distributed computing has greatly evolved and the collaborations have gained a great experience of work in such an operating environment. It has been shown that it is possible to build an environment that allows one to provide access to data for all the physicists involved with the experiments and that the collaborations can use the resources regardless of their location. It was clear that the data management service should be improved in order to optimize the use of the resources of both hardware and staffing levels.

Currently, much attention for solving such problems is paid to the development of systems of task management (preparation and analysis of experimental data, simulation, etc.), possessing a scalable and flexible architecture, which provides opportunities to adapt the system to changing computing, storage and network resources. This allows one to combine within a single computing environment a variety of heterogeneous computer systems of different hard- and software architecture. For these purposes one have to conduct a scientific research in the field of intensive operations with large data volumes in distributed systems (Big Data).

The main direction is a desire of the experiments of the most effectively use of all the opportunities provided by the grid sites, and departing a strict adherence to the roles of grid sites in the original strictly hierarchical model of distributed computing. There is a great potential in the emerging cloud paradigm which simplifies the intermediate layer required for grid as a common view on the heterogeneous and distributed resources. The attitude also has changed to the network infrastructure, which is an invaluable resource, and the appropriate use of the network for data access allows one to optimize the cost of the whole computing.

The computer simulation is of particular importance as significant computing resources are used for this purpose. The main objective is to improve the efficiency of the overall simulation, its speed up where it is possible and to use a variety of computing resources.

One has to take into account as well the evolution in technology, architecture of processors and storage systems that are rapidly evolving and will continue their changing over the coming years and which are very difficult to assess and to predict on this time scale.

Following present-day trends in the field of data processing in high energy physics and taking into consideration the fact that the NICA data processing is expected to involve several computing complexes with a variety of resources and architectures (grid, supercomputers, clouds, clusters), the processing should be considered as a distributed one and the environment - as heterogeneous. At present, for the data processing in such environments, there are workload (task) management systems like PanDA (Production and Distributed Analysis System), Dirac, Alien, etc. Today, the PanDA possesses a confirmed ability to work with all types of the listed above resources. However, this does not restrict research and use of other systems, work on which is also underway.

In the heterogeneous environment, the distribution of processing by the types of tasks looks logical: for simulation to maximize the use of supercomputer resources of the collaboration organizations (one of the candidates is a HybriLIT cluster at LIT), mass processing should be performed in the environment like Tier 1 at JINR, user analysis on the cloud and Tier-2.

As to storage systems, a perspective way is to focus on a federal solution with built-in replication, this would remove load from the distributed data management component and would simplify the users work. The currently existing federated repositories cannot work with tapes. If the use of tape storages becomes economically viable at some stage, it is necessary to conduct research and experiments. However, the development of the data storage technologies can afford to abandon using the tape storages for archival data storage.

The concept includes topics related to network infrastructure of the project, the engineering infrastructure of computing centers (online-offline), data acquisition from the detectors, their initial on-line and off-line processing, tasks of simulating physical processes on the NICA setups, with the development of a model of data storage and processing, creation of the system of long-term storage of experimental and model data, with using of the distributed computing technologies, information system and a unified monitoring system. The solution of these problems, which are a necessary step to expand and consolidate the participation of numerous organizations of various fields in creating and using for the purposes which are the JINR priority, the unique features of this distributed infrastructure at JINR, is one of the defining moments of the successful implementation of the project.

2. NICA computing network at JINR

The NICA Complex is constructing on the LHEP site. So, the network infrastructure of this laboratory should provide a number of functions thus generating are designed to ensure the implementation of the scientific and thematic plan of LHEP. Hence the requirements for the nomenclature of network services and the technical characteristics of the hardware and software complex of the network infrastructure. The usage of LIT's computer capabilities in the computer support of the NICA project requires also the upgrading of the computer network of the JINR as a whole.

Below is a brief statement of the goals and objectives of the network infrastructure of LHEP and LIT, a description of the main network subsystems according to their current state. The main activities for the modernization of common network systems excluding technological ones connectet to the experimentsl setups of the NICA Complex are listed as well.

Modernization of the considered network subsystems should be carried out in the next few years in order to improve reliability, security level, speed and other characteristics of the network infrastructure of the Institute.

2.1. Goals and tasks of the NICA network infrastructure of LHEP

Main goals and tasks of the NICA network infrastructure of LHEP are:

- a construction of a common information space development of the existing LHEP resources: computing, information, and data storage;

- a construction of a common information space for the staff of the LHEP and collaborating institutions working in the Lab, providing the possibility of data exchange between units of JINR and other Institutions involved in the NICA project;

- organization of access to both the centralized resources of the IT structure of the Institute, and to the resources of individual set-up’s networks, and divisions;

- providing high-speed access to Internet resources;

- support of network services, such as authorization and authentication, NMIS-systems for monitoring the status of active NICA network infrastructure elements. Ofcourse the common network services like e-mail, DNS, proxy-service, IPDB network element databases and so on should be supported as well.

2.2. Current state and main modernizations of the network subsystems.

Currently, the main optical transport data transmission line LIT - LHEP is organized in the form of an optical ring and operates at a speed of 10 Gbit/s (Fig. 1.).

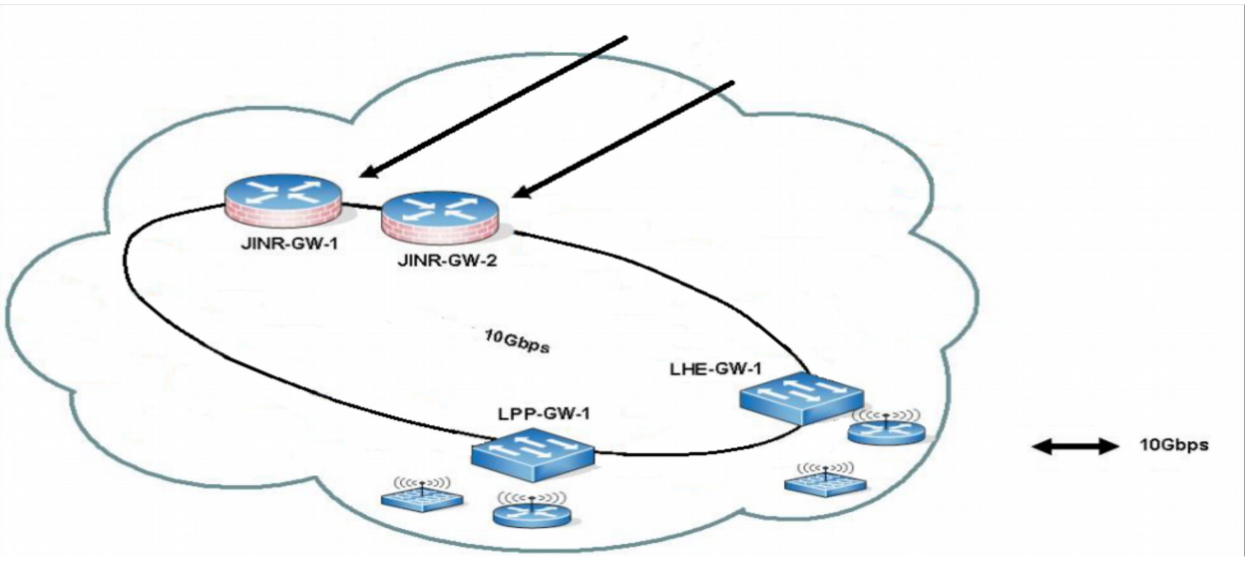


Fig. 1. Optical transport data transmission line LIT-LHEP.

It is based on two network switches of the third level (L3) according to the reference network model ISO OSI (International Standard Organization Open System Interface). Two Cisco Catalyst 6509 devices (configuration logical names - JINR-GW-1 and JINR-GW-2) and two Cisco 3560-S devices of the LHEP (configuration logical names - LHE-GW-1) are used as switches in the core of the LIT network structure and LPP-GW1.

To meet of the future needs of the NICA project (see description below), this transition plan to increase to 100 Gbit/s. It is also necessary that the new network equipment supports both time-tested technologies and protocols and new ones such as MPLS, VxVlan. So, MPLS (Multiprotol Label Switching) is a mechanism in a high-performance telecommunications network, it is scalable and independent of any protocols by the data transfer mechanism. The MPLS technology makes it possible to implement the services of virtual private networks (VPN) of the new generation. VxVlan (Virtual Extensible LAN) is a network virtualization technology designed to solve scalability problems in large systems, increases the scalability of up to 16 millions of logical networks and allows networks of the 2 layers to simultaneously coexist over IP.

The new equipment for the NICA central telecommunication nodes, the core transport between the LNP/LIT and LHEP sites, switching and routing, is proposed for four multifunctional switches of the Cisco Nexus 9504 family with full-mesh connection topologies for maximum reliability and performance. (see Fig. 2.)

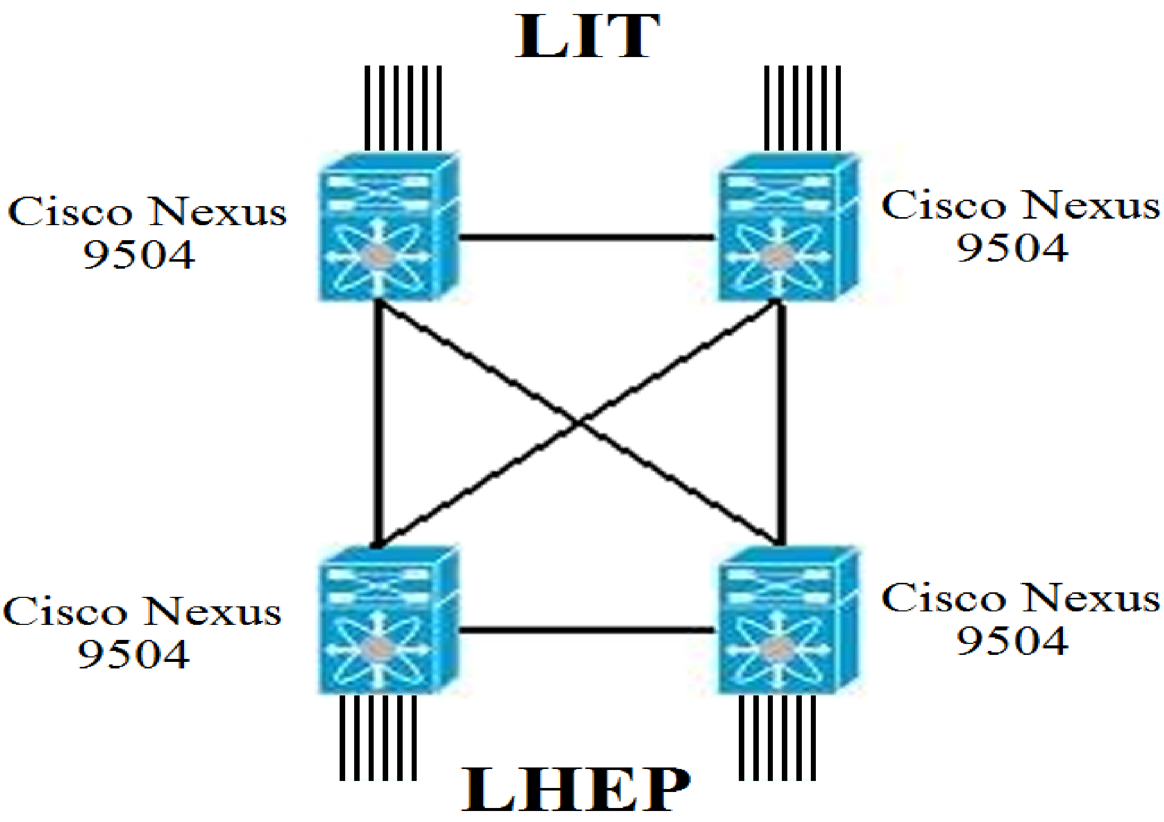


Fig. 2. The core of the system: switching is performing by the Cisco Nexus 9504.

The most important stage in the development of the NICA network infrastructure is the creation and laying of new optical highways between the two sites: LIT and LHEP. It is necessary to distribute independent branches geographically for the organization of communication between on-line and off-line objects of the NICA Complex and MIVC project at a speed of 100 Gbit/s and to ensure maximum reliability and fault tolerance of this network. Fig. 3 shows a plan for constructing of two independent branches of optical highways between the sites of LNP/LIT and LHEP.

At the LHEP site, it is also necessary to build an optical-cable infrastructure with the ability to connect network equipment to two territorially independent routes. Fig. 4 represents the existing system of telephone ducts, dot-dashed lines, and the newly constructed parts of telephone ducts, blue lines.

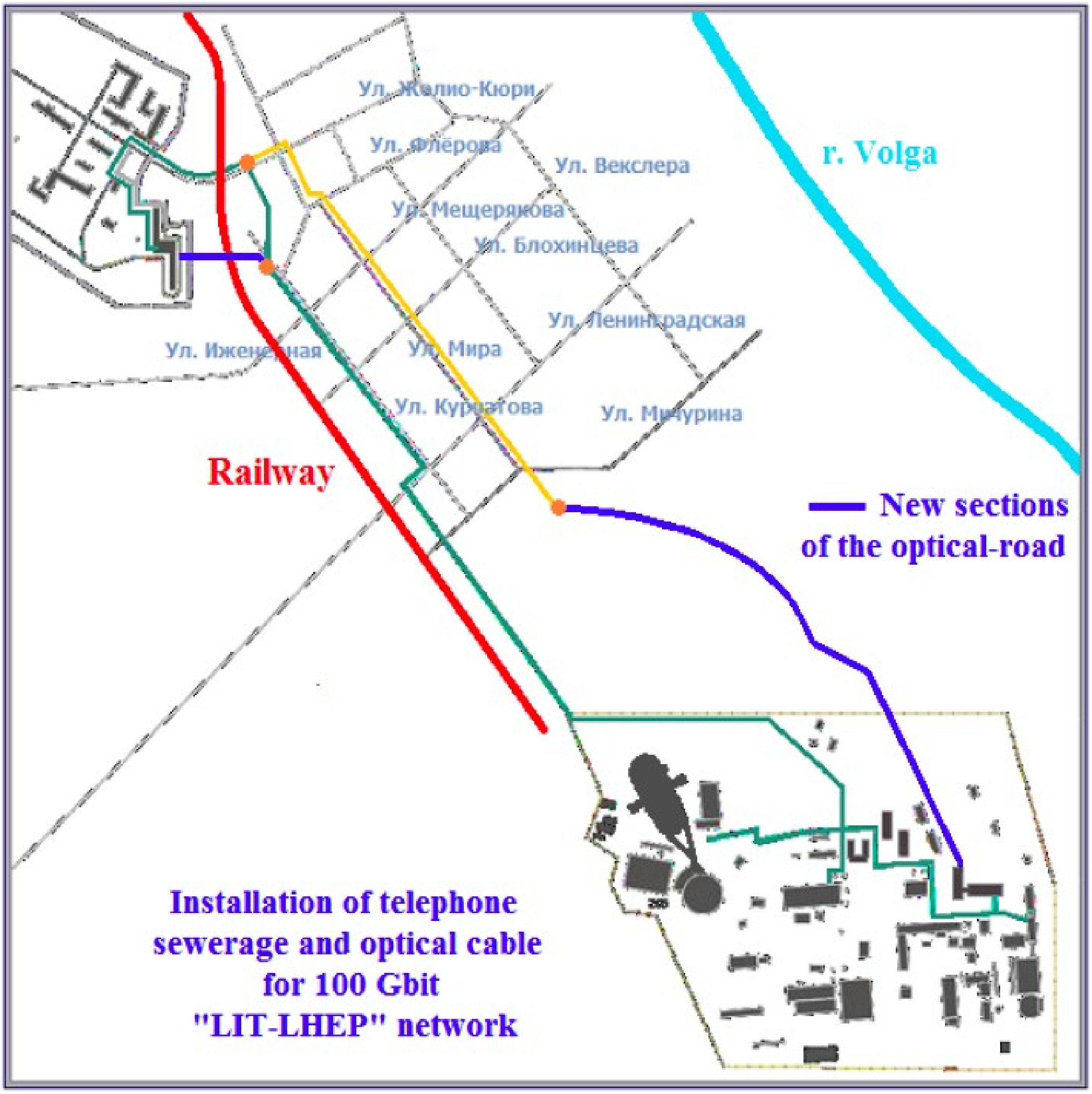


Fig. 3. The optical Scheme of the two tracks between the platforms LIT and LHEP

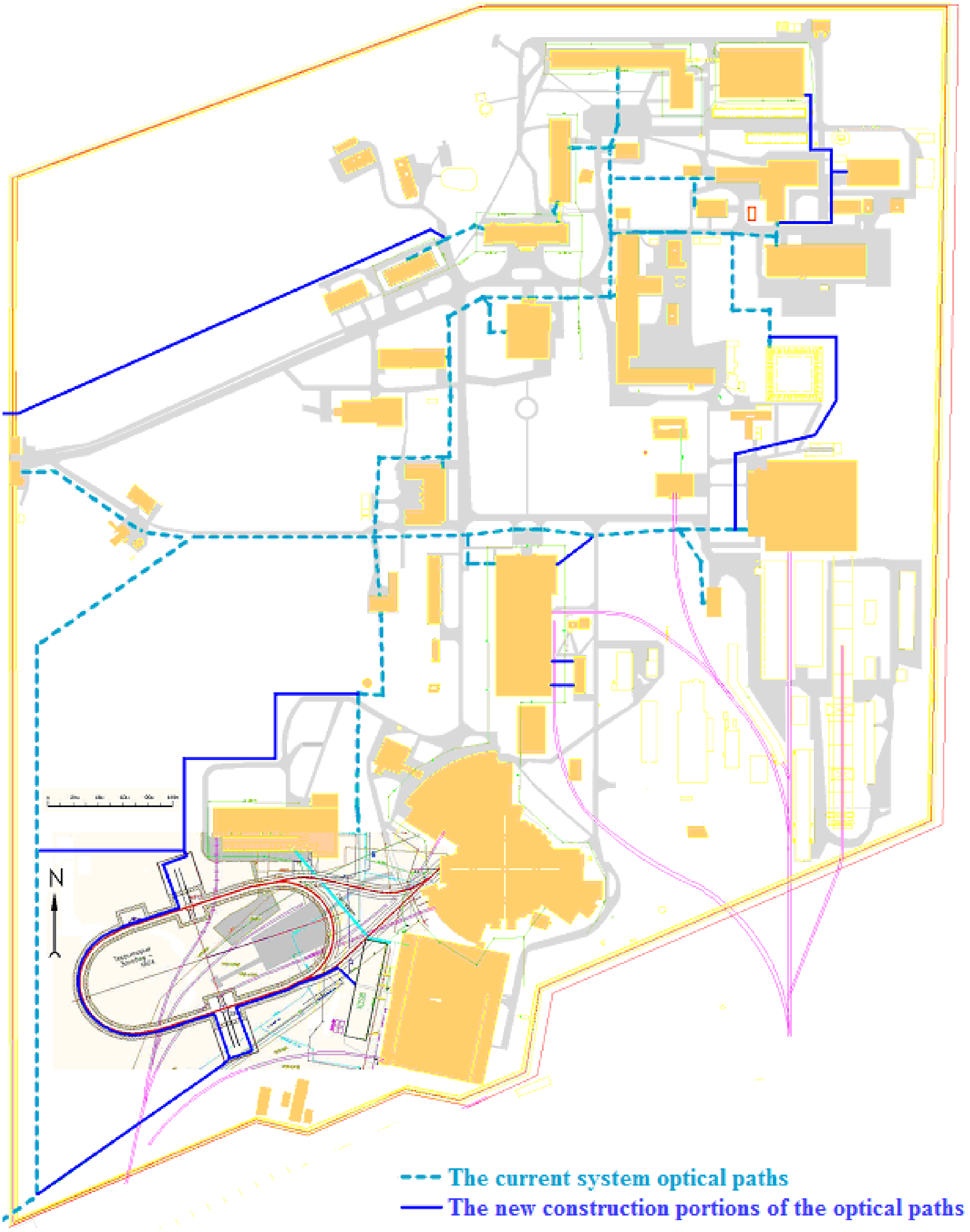

N
--- The current system optical paths
— The new construction portions of the optical paths

Fig. 4. Schematic view of the optical paths at the site of LHEP.

3. A structure of a distributed NICA computer cluster at JINR

A NICA computer cluster is intended for simulation, processing, storage and data analysis obtained from the experimental set-ups of the NICA Complex. As well as all large centers of processing and data storage from big physical installations, such cluster is planned to create as geographically distributed (multisite), integrating all components located both on the LHEP site, and on the LNP/LIT site, by a local computer network of hundreds Gbit/s.

On the LHEP site the on-line cluster will be located in the building 14 (prototype is located in the building 201). A first stage of the off-line cluster will be located in the building 216, room 115 (further – cluster-216), and the second one - in a new building of NICA Center of the NICA Complex (further – cluster-A). On LNP/LIT site the off-line cluster is located in the building 134 as part of JINR Multifunctional Information and Computing Complex (MICC) .

The diagram of a possible useage of the distributed computers system at the NICA clusters is presented in Fig. 5.

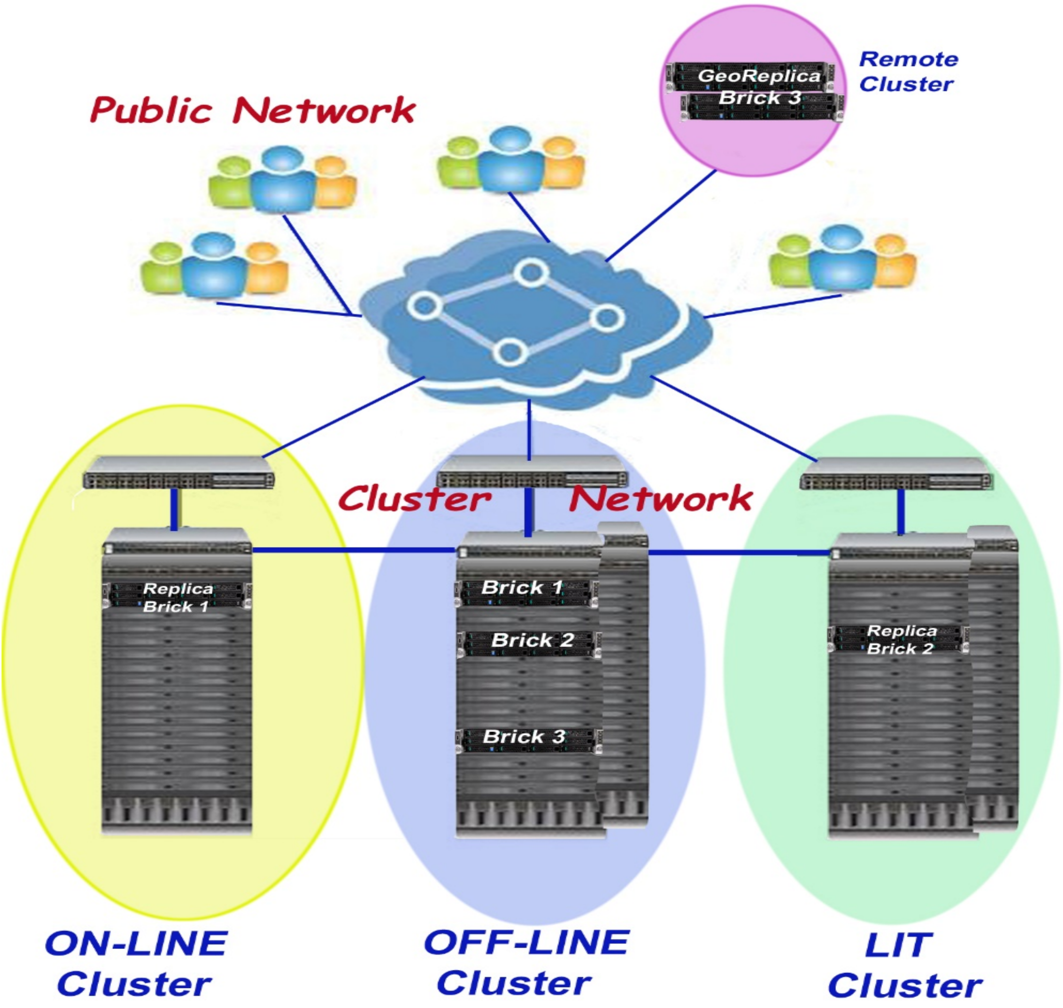


Fig. 5. A diagram of the distributed resourses usage at the NICA cluster.

3.1. On-line cluster for a fast physical data analysis, intermediate data storage and physical data monitoring

The on-line cluster is a specialized PC-farm. It has several basic and secondary functions. The main function is to receive data from different subsystem of the NICA Complex including data from the Data Acquisition (DAQ) system of the NICA experimental set-ups. In the process of receiving and transferring data for further processing into the off-line complex, using its resources, partial processing of the recorded data and their physical on-line monitoring will be carried out.

Modern systems for collecting data from large physical setups use very actively Ethernet networking technology. The DAQ systems of the NICA experimental set-ups (BM@N, MPD, SPD) have to have own technological Ethrnet networks, described in detail in the corresponding DAQ TDRs. For example, a network connection diagram of the electronics of the MPD DAQ using Ethernet network technology to the on-line farm is shown in Fig. 6 [2] and Fig. 7. Such connection will be made using switches with optical interfaces and optical cables, with interface speeds 100 Gbit/s.

Goals and tasks of the NICA on-line farm are the following:

- receiving raw events or raw data from the main subsystems of the NICA Complex;
- packing and sorting these data;
- storing data at the temporary disk space (up to 24 hours);
- carrying out express processing and analysis of this data and 5~10% received events;
- transfering data to the off-line NICA clusters;
- receiving data for monitoring the computer network.

The estimated size of events and raw data from the main subsystem of the NICA Complex is presented in Table №1.

with an average event rate of 6 kHz, and processed by the DAQ system arising from heavy-ion collisions.

Table № 1

NICA subsystem	Technical data rate (GB/s)	Event rate (kHz)	Event size (MB)	Full event size (GB)	Mean data transfer rate (Gb/s)	Data volume (TB/24 hours)
Accelerators						
2019-2020	0.5				0.1	4
>2020	1.5				0.3	10
BM@N						
2019-2020		30	0.5	15	20	100
>2020		50	0.7	35	100	300
MPD						
2021-2022		0.1	1	0.1	10	200
>2022		6	2	12	100	600
SPD						
>2023		50	0.5	25	100	1000

When calculating the required computational resources, we select the CPU a 16-Core Intel® Xeon® Processor E5-2697A v4 2.60 GHz 40MB Cache (145W) (average price/performance) processing flow at 50 Mb/s per core.

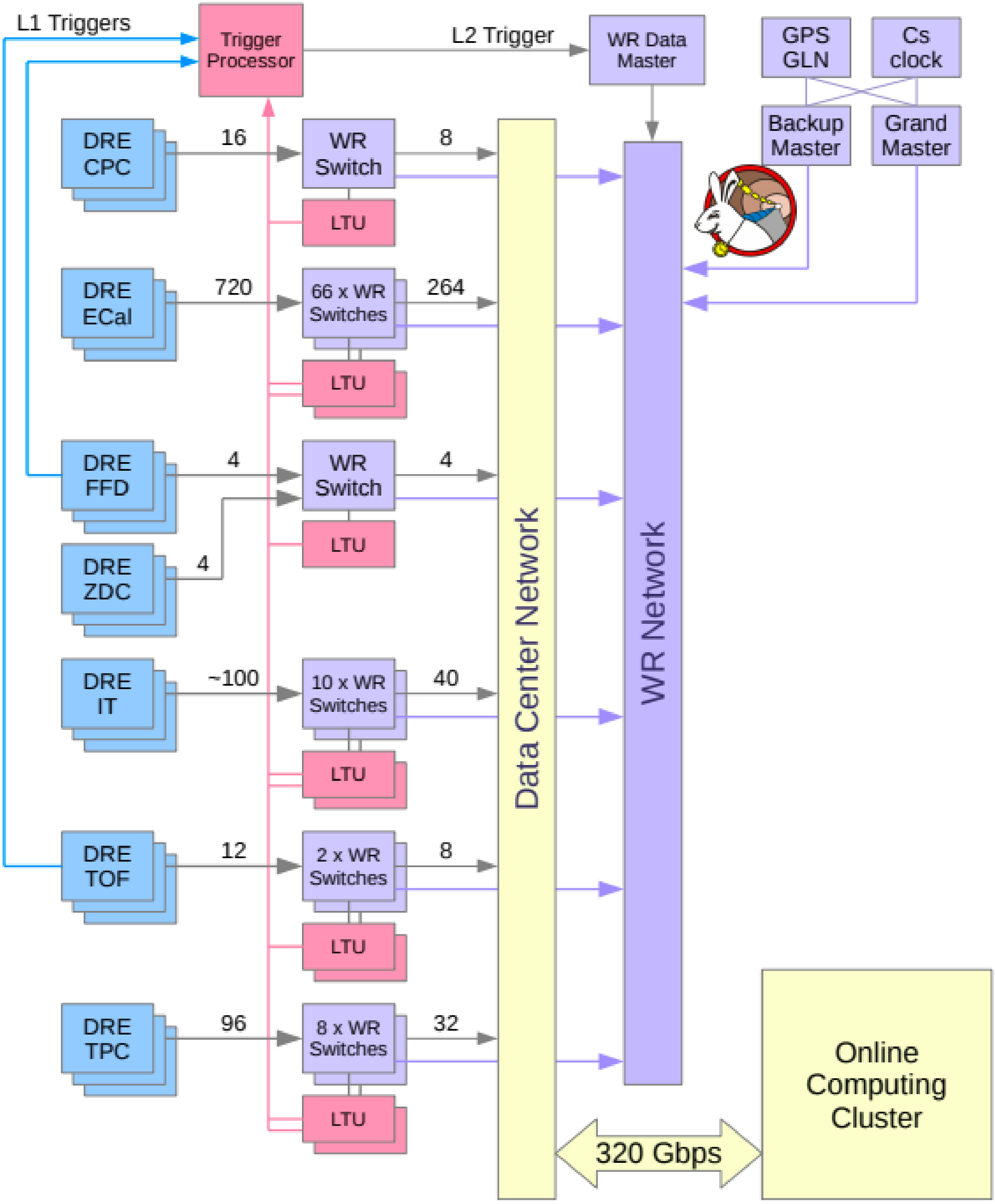


Fig. 6. MPD DAQ Trigger, Timing and Data Links.

As mentioned earlier, the on-Line cluster will be located in building No. 14 and will meet the need for the placement of computing resources of experimental physicists associated with the experiment MPD and SPD, servers of equipment for control and monitoring of accelerators, as well as all other equipment that requires guaranteed power and climate control system.

Building No. 14 is being built according to the project, which has passed all the approvals and checks.

The main tasks of the On-Line cluster are guaranteed acceptance of data from the DAQ physical setups, provide computer resources for express processing and analysis of received data to physicists and provide the data to the Off-Line cluster.

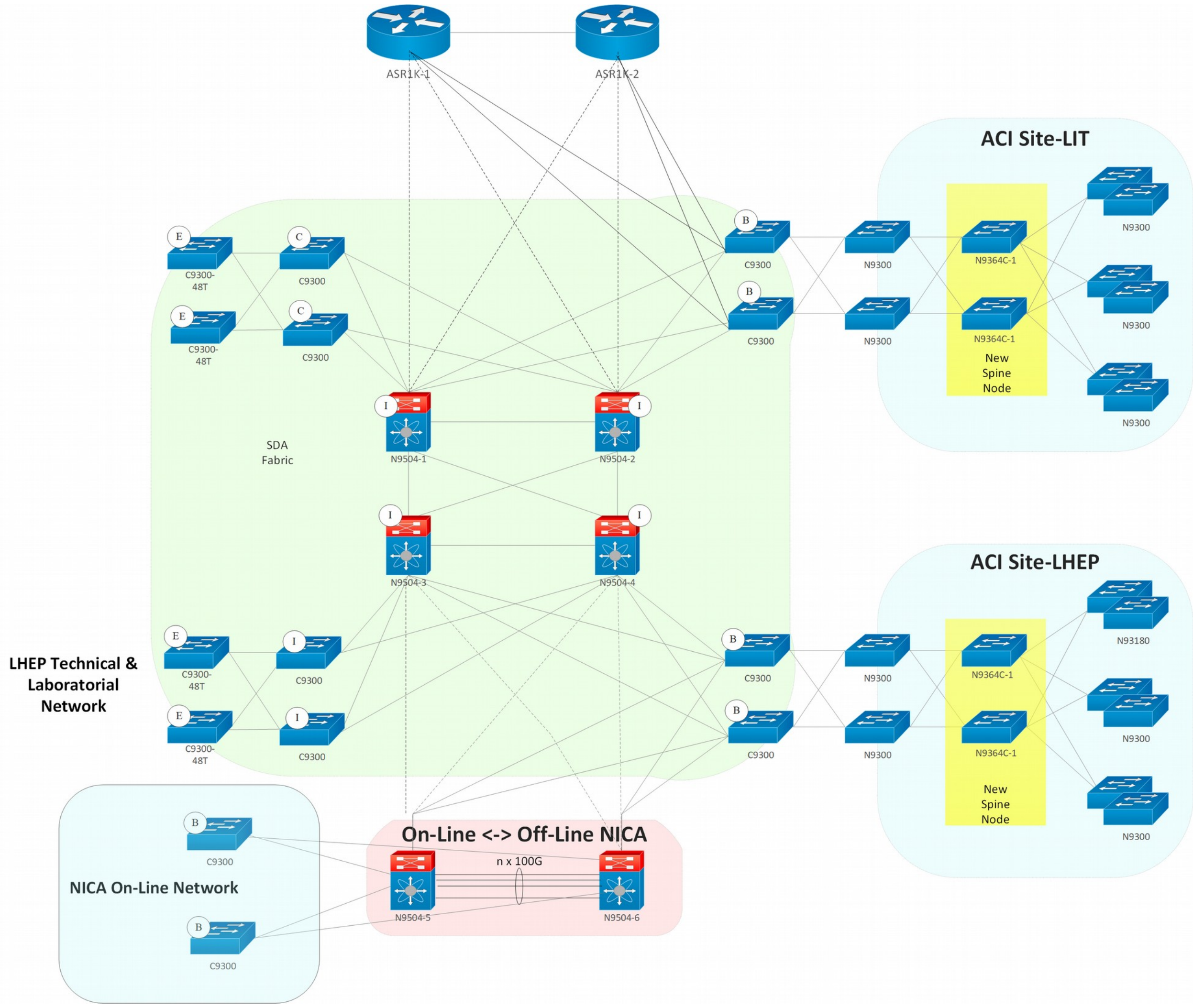


Fig. 7. Wiring diagram of the switches of the NICA network

As it mentioned above, the NICA On-line cluster will be located in the building 14. It should satisfy user requests for computing resources associated with the accelerators of the NICA Complex, experiments BM@N, MPD, SPD and applied experiments. In particular, it should guaranteed recieving of data from the DAQs of physical setups, provide computer resources for their express processing for physical on-line monitoring and transfering data to the NICA Off-line clusters. It should be equipped with servers for network monitoring and support of common network services.

To date, a project of the building 14 has been developed and it has passed all the approvals and checks.

3.1.1. Network computer infrastructure of On-line farm

At organization a networked computer infrastructure of the On-line farm it is necessary to provides maximum throughput, reliability and fault tolerance of the network. All network computer elements of the On-line farms will be associated with interfaces at a speed of 100 Gbit/s and multiple links connected to each other as shown in Fig. 8.

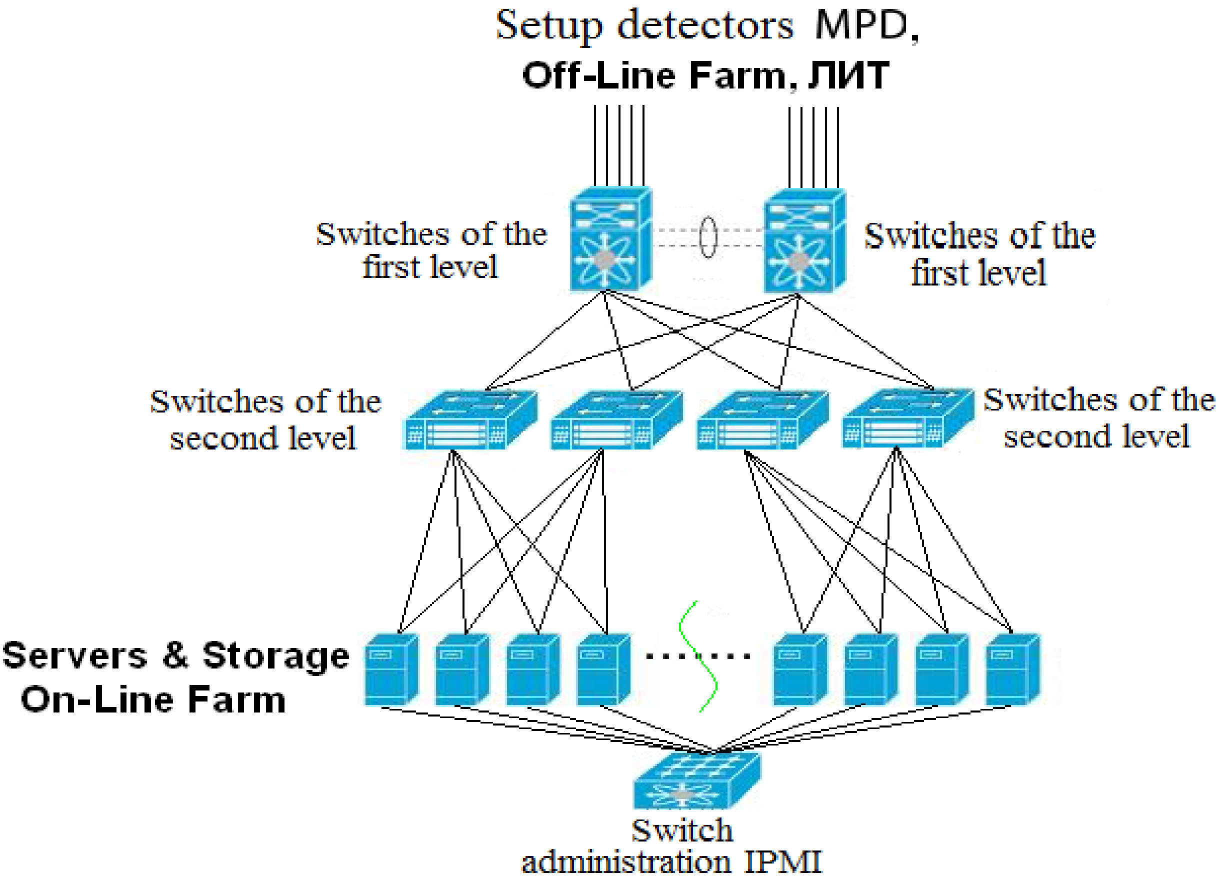

Setup detectors MPD,
Off-Line Farm, ЛИТ
Switches of the first level
Switches of the first level
Switches of the second level
Switches of the second level
Servers & Storage On-Line Farm
Switch administration IPMI

Fig. 8. Schematic view of the inner networking structure of the NICA On-line farm.

Table № 2 presents data on the consumption of electrical power by various computer elements of the On-Line farm.

Table № 2

Name	Amount	Consumption electric power 1 unit, kW	The total consumed electric power, kW
Server such as Super Micro 2U TwinPro	10	0.8	8
Server such as Super Micro 2028R-DN2R24L	22	1.0	22
The switch the first level of the Cisco Nexus 9504	2	3.0	6
The switch of the second level of the Cisco Nexus 9336	6	1.5	9
The IPMI management switch of the HPE Aruba 3810M 48G	2	0.5	1
		Total:	46

Assuming 5 kW in one telecommunication rack all power system of the NICA On-line cluster will be allocated in 20 racks.

3.1.2. Engineering infrastructure of the NICA On-Line farm.

To implement the NICA On-line farm project the building 14 was allocated (see Fig.9).

Fig. 9. The building 14.

Modern computer centers, including the NICA On-Line farm, should have modern software, a sufficient amount of computing resources, a system for managing and storing data and also a reliable and developed engineering infrastructure that allows for uninterrupted operation of the center in the 24/7/365 mode. The equipment used to developed such of infrastructure should have the necessary level of reliability and maintainability. The engineering infrastructure should be modular, scalable, adaptive and expandable.

The main objective of the NICA On-line project is to create an engineering infrastructure that consists of:

1. Power supply systems and uninterruptible power supplies, including:
   1.1. General power supply system and power distribution;
   1.2. Uninterrupted power supply system;
   1.3. A system of guaranteed power supply;
   1.4. A grounding system;
   1.5. Structured cabling system (SCS) with the cable wells.;

2. Systems of conditioning and ventilation, including:
   2.1. Refrigeration system;
   2.2. Industrial air conditioning systems;
   2.3. Ventilation system of;
   2.4. Smoke and gas removal system.

3. The fire safety complex, including:
   3.1 Fire alarm;
   3.2 System for very early fire detection type VESDA;

3.3 System of modular gas fire.

4. The automated dispatching system and control of engineering infrastructure, including:
   4.1. Engineering automation;
   4.2. Dispatching system;
   4.3. Monitoring system for baseline parameters;
   4.5. Technological video surveillance.
5. Hardware racks and aisle containment system.
6. Raised floors.
7. Building 14 with computer room and computer equipment.

3.1.3. Status of the engineering infrastructure of the NICA On-line cluster.

Building 14 needs a reconstruction with a constructing of an engineering infrastructure on the basis of the project. When developing the project it is necessary guided by the following regulations:

- TIA–942 "Telecommunications Infrastructure Standard for Data Centers" (Telecommunications infrastructure of data centers);
- SN 512-78 "manual for the design of buildings and rooms for electronic computers" (with amendments dated March 1, 1989 February 24, 2000);
- SNiP 2.08.02-89 "Public buildings and constructions";
- SNiP 41-01-2003. "Heating, ventilation and air conditioning".
- SNiP 2.04.01-85. "Domestic water supply and Sewerage of buildings".
- EIC "Rules for electrical installation, edition 7";
- SNiP 3.05.06-85 "Electrotechnical devices";
- GOST P50571.21-2000. "Electrical installations of buildings. Earthing devices and systems of equalization of potentials in electrical installations containing equipment data processing";
- NTPD-90 "norms of technological design of diesel power plants";
- GOST R 53246-2008. "Information technology. Structured cable system. Design of major components of the system";
- Federal law of the Russian Federation of 22 July 2008 N 123-FZ "Technical regulations on fire safety requirements" (edition of 13.07.2015);
- SP5.13130.2009 "fire protection Systems. Installation of fire alarm and fire fighting automatic. Norms and rules of design";
- SP6 13130.2009 "fire protection Systems. Electrical equipment. Fire safety requirements";
- RD 78.36.003.2002. Guidance document. "Engineering and technical strengthening. Technical means of protection. Requirements and design standards to protect objects from criminal encroachments";
- GOST 24.104-85 "Automated control systems. General requirements";
- SPDS "System of design documents for construction";
- Other regulatory documents.

3.1.3.1. Energy management for the NICA On-line cluster.

1.1 General power supply system (ES)

Two independent feeders should be provides all equipment of the NICA On-line cluster. They should pass in independent and non-crossing ways before entering the building from the main distribution panel (MDB). Maximal permitted/allocated power for each input is not less than 300 kW.

- power supply redundancy of it equipment OF the 2N;
- complete the water-distribution device (IDD) with automatic transfer switch (ATS) and the system of technical accounting of electricity consumption;
- internal Cabinet lighting IDD from the power supply cables;
- the ability later to connect the diesel generator set (DGS) without replacing the IDD;
- at least 5% of the provision for the automatic switches, as well as a backup place for further expansion in all low-voltage distribution switchboard (LVDS) (switchboards);
- laying of cable lines, supply and distribution networks indoors
- the ability to increase capacity without replacing and the dismantling of the designed devices;
- the location of the main LVDS up in a special room outside the main room;
- the lighting device to the led lamps;
- the earthing system;
- additional contacts at the key automatic breakers and switches for remote monitoring of their situation;
- quality control of electricity transfer capability;
- using wire and cable with flame retardant insulation having a low level of smoke and gas, like FRLS.

1.2. Uninterrupted power supply system (UPS)

- smooth operation of it equipment and infrastructure for at least 15 minutes with a capacity of 300 kW;
- build on the basis of the power modules with double-conversion power;
- N+1 redundancy;
- power supply integrated security systems from dedicated shield power UPS KSB;
- the Location of UPS and BATTERY racks in a special room outside the main room OF;
- system power conditioning technology areas (SCTA) to be carried out through a system of uninterrupted power supplies (UPSs);
- Intelligent power distribution.

PDU include:

- power distribution to end consumers (it equipment of the Customer);
- the possibility of remote monitoring;
- installation directly to each server rack;
- in the amount of 2 pieces.

3.1.3.2. Conditioning system technology areas (SCTA) of the NICA On-Line cluster.

SCTA is constructed on the basis of air conditioners must provide:

- round-the-clock and year-round operability of it equipment;
- stable and reliable operation in outdoor temperatures from ?40°C to +45°C;
- total cooling capacity not less than 200 kW;

- redundancy is not below N+1;
- optimum distribution of air flows;
- to provide for the uniform operation of air conditioners by hours by controlling via the central unit;
- consider implementing liquid cooling of server racks;
- to provide for the retirement of air conditioning systems in case of fire.

3.1.3.3. System of automatic gas fire suppression (SAGFS) NICA On-Line cluster.

SAGFS must provide:

- the elimination of possible fire in the main space and in other isolated from the main volume spaces (raised floors, suspended ceiling, etc.);
- the presence of lead fire protection fluid;
- ndividual protection of personnel;
- imagesonline as well as removal of combustion products by fixed installation;
- ventilation (purge) after implementation of fire suppression;
- the prohibition of entry in the case of fire protection fluid start to the end of the ventilation system (purge space);
- the ban on entrance to carry through MSMA;
- fire protection fluid Freon 125

The gas fire fighting modules are placed outside the main room of the On-line farm.

3.1.3.4. Aspiration system for very early fire detection (ASPS).

ASPS provides that:

- early detection of fire;
- a high degree of protection against false signal recognition system;
- three alarm thresholds;
- a set of components certified to EN 54-20;
- modularity;
- the function of dual detection to control the premises and equipment;
- of ASPS is not intended to start SAGFS.

3.1.3.5. The system of cable support structures and eavesdropping devices NICA On-Line cluster.

The following constructions for cabling:

- trays of ladder type for the laying of power cables inside;
- mesh type trays for laying data cables inside;
- trays of ladder type for pipes of the coolant inside and outside the refrigeration machinery and/or condensing units and air-conditioning system;
- track laying electrical and data cables must be independent, the distance between the trays is not less than 300 mm.

The cable support structures are grounded. Reserve cable support structures and embedded devices is at least 30%.

3.1.3.6. SKS NICA On-Line cluster

SKS should include:

- further switching equipment in a single computing environment
- the strip topology star
- the use of fiber-optic and copper UTP cables cat 5e
- a common point of switching with the use of optic boxes and patch panels
- optic cable OM4 standard

3.1.3.7. Monitoring system and managements of access (MSMA)

MSMA must ensure

- control authorized access to the premises, in accordance with the object mode.
- integration into complex security system
- the possibility to register new access cards

3.1.3.8. Video surveillance system (VSS)

VSS provides the ability to:

- visual control of the situation in the server room and in adjacent rooms and outside the building
- view and analyze data from the archive of the events
- video storage
- view video over the network simultaneously from all cameras in real time
- schedule recording, on demand
- motion detection, setting of zones and sensitivity
- connection of external alarm sensors
- integration in the system is security-the disturbing alarm system..

Cameras resolution not less than Full HD.

3.1.3.9. System monitoring and alarm (SM)

SM provides::

- tracking of environmental parameters in the server room:
  - temperature, humidity;
  - leaks as a system of conditioning/cooling, and outside spaces;
  - smoke, fire, triggering APPT;
  - unauthorized entry;
- timely, remotely over the network or using GSM module alert for action to prevent or eliminate abnormal/emergency situations;
- integration into the existing monitoring system of the building.

The layout of the equipment of the engineering infrastructure and telecommunication boxes of the On-Line farm is shown in Fig. 10.

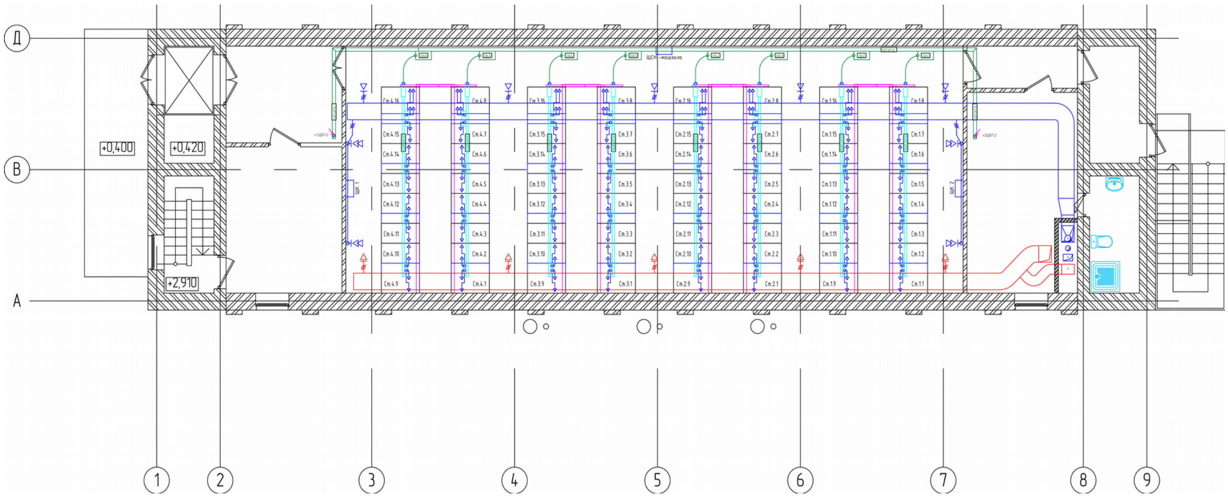


Fig. 10. Plan of the 1st floor with location of telecommunications closets in the building 14 (top view).

3.2.3.10. Software of the NICA On-Line cluster.

Software to be used at the NICA On-line farm is described in sufficient detail in the MPD DAQ TDR [2]. As a selected storage, a distributed POSIX-compliant file system Ceph FS has been selected. It's at open source elastic design easily scalable petabyte-scale storage. It is based on the Association of the storage spaces of several tens of servers in the object storage, which allows one to realize flexible multiple pseudorandom data redundancy.

3.2. Off-line NICA computer clusters

3.2.1. LHEP Off-line cluster

During design the off-line cluster an experience of the large JINR MICC center development results of simulation and an experience accumulated during an off-line LHEP cluster prototype construction is considered. The LHEP cluster prototype consists of 15 servers, about 100 TB of disk space, 768 cores of the CPU (more precisely - logical kernels of the CPU or flows). About 150 registered users can logged in on the cluster via 4 interactive machines by ssh and use 11 machines for batch processing.

One of the principal components of a cluster is the clustered file system. Several clustered file systems are applied now: GPFS, Lustre, EOS, dCache, Ceph, GlusterFS etc. The GlusterFS file system is suitable best of all in our opinion, for off-line cluster. The GlusterFS is distributed, parallel, linearly scalable file system with a possibility of protection against failures. The GlusterFS can integrate by means of RDMA (Remote Direct Memory Access) or TCP/IP the data storages, which are located on different servers in the different sites in one parallel network file system. The GlusterFS works in the user space by means of FUSE technology therefore doesn't require support from operating system kernel, working over the existing file systems (ext3, ext4, XFS, reiserfs, etc.). Unlike other distributed file systems, such as Lustre and Ceph, operation of the GlusterFS doesn't require the separate server for storage of metadata that improves scalability and reliability of system. The GlusterFS has the convenient mechanism of a synchronous data replication between the server of a local area network and a mechanism of asynchronous data replication on geographically remote servers (geo-replication).

The second important component of the cluster is the computer network, which integrates servers in one cluster. 10 - 40 GB/s Ethernet network are used now at the LHEP cluster prototype, at the cluster-216 and the cluster-A it is planned to use 100 GB/s local area network to increase throughput by 2.5 times. The off-line cluster is structured by access rates to provide a load balance of all Ethernet segment. The diagram of an Ethernet segment usage is presented in Fig. 11.

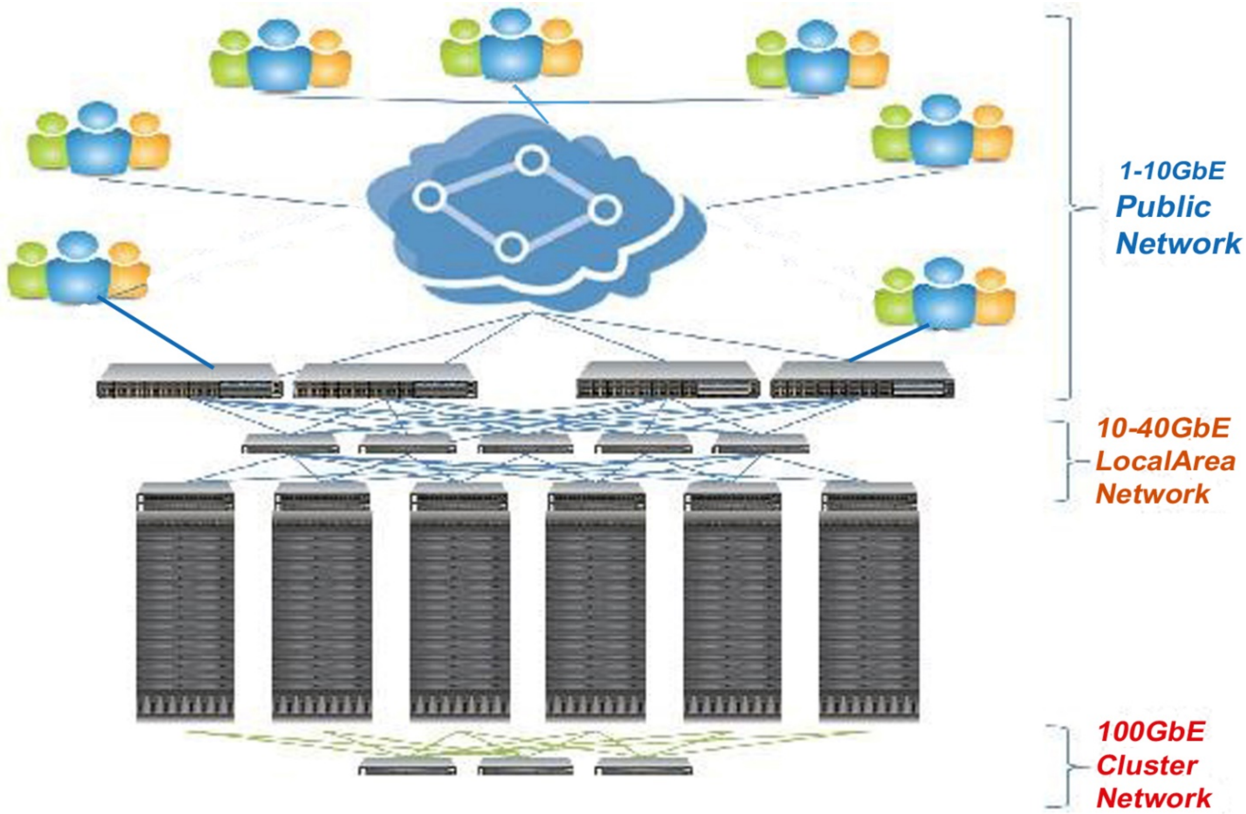


Fig. 11. Diagram of the Structured Ethernet Network

Further increasing of a data transfer efficiency is possible only due to use of the special equipment and optimization of exchange protocols. RDMA (remote direct memory access) which allows to transfer data between servers directly from memory of one application to memory of another without involvement of central processors becomes the main technology of exchange, instead of traditional exchange on TCP/IP. Until recently, RDMA was available only in structures of InfiniBand. However, the new RoCE technology (RDMA over Converged Ethernet) has been produced recent years and RDMA advantages are available now for data processing centers which are based on the Ethernet equipment. RoCE is a technology of effective data transfer with very low time delay into Ethernet networks without loss. One of the leading firm in production of the equipment compatible to requirements of RoCE is Mellanox. It releases a complete set of such equipment: SN2700/SN2100/SN2410 switches with 32/16/8 100 GbE ports, ConnectX-4 100GbE adapters, optical and copper 100GbE cables and also necessary program components (drivers) for operating systems. Mellanox 10 - 40GbE equipment, SLC 6.X OS and drivers from Mellanox is used now at the cluster prototype. At the cluster-216 and cluster-A the latest version of CentOS supported of RoCE is built already and will be used. Time delays of data transfers according to the RoCE and TCP/IP protocols are shown in Fig. 12.

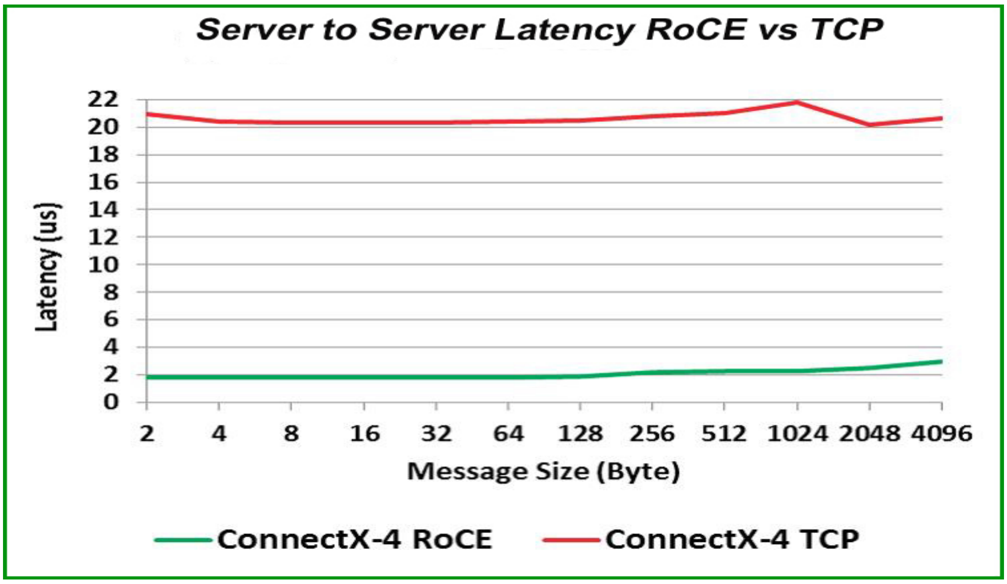


Fig. 12. Comparing time delay for data transfer between servers with RoCE and TCP/IP technologies.

The Mellanox company developed a new interesting technology for cost reduction for expensive 100 GbE equipment – a Multi-Host Technology. On the basis of the ConnectX-4 interface a new device has been produced, to multiplexing the bus PCIe at 4 completely independent buses PCIe which by means of inexpensive repeaters are inserted into servers instead of expensive Ethernet adapters. The diagram of such connection is shown in Fig. 13.

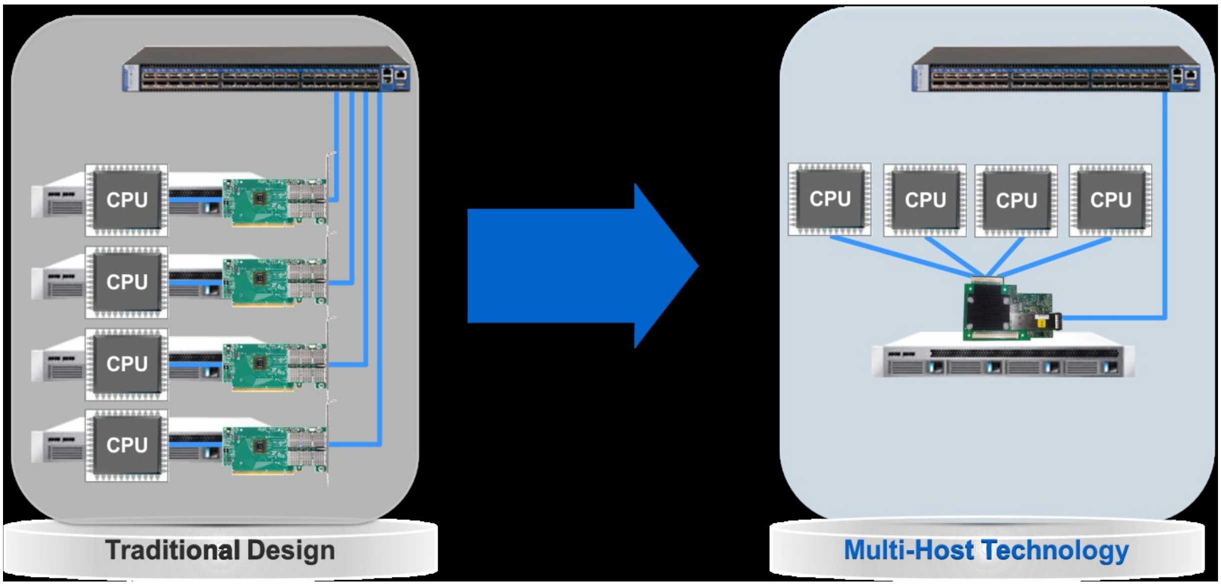


Fig. 13. The diagram of a usage of the Mellanox Multi-Host Device.

Each node can be active or inactive at any time irrespective of other nodes to receive own throughput. Complete throughput is partitioned between nodes ether uniformly (value by default), or on a basis of the configured quality of service (QOS) depending on needs of a data processing center. Thus capital expenditures are significantly reduced because instead of four

NICs, cables and switch ports it is used only on one, and also complete energy consumption is decreased. The Mellanox Multi-Host device is presented in Fig. 14.

Fig. 14. The Mellanox Multi-Host device.

The cooling system of the prototype is based on traditional couple Chiller-LCP – air- and water-cooling. The equipment of the Rittal company (Germany) is used. A common view of the LHEP off-line cluster prototype is shown in Fig. 15.

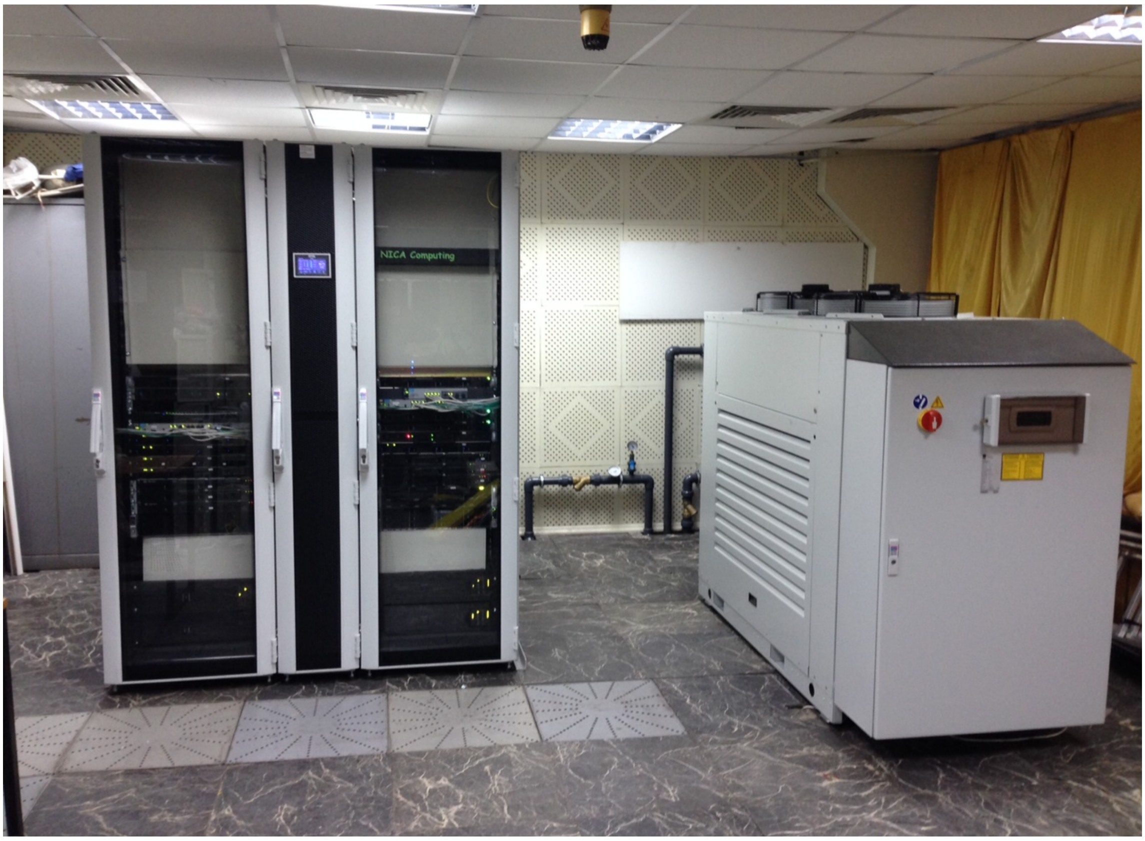


Fig. 15. A common view of the LHEP off-line cluster prototype .

The cluster-216 with the parameters of 0.5 PB of disks space and 1K of CPU cores (1-st stage, 2017), and 4 - 5 PB of disks space and 4K of CPU cores (the 2-nd stage, 2019) will be located in a new location – in the room 115, building 216, LHEP. It will consist of 8 IT racks 42U, 4 LCP with a power of 40 - 55 kW each, 2 80 kW chillers with an internal free-cooling to 120 kW located outside of the building. The racks settle down in two ranks with a cold aisle between. Besides, two 96 kW uninterruptible power supply unit (UPS) will be set. Temperature, humidity and leaks sensors as well as a GSM unit will be installed also in a set. All equipment should be produced by the Rittal company (Germany). An approximate appearance of the cluster-216 is shown in Fig. 16.

Fig.16. An approximate appearance of the LHEP off-line cluster-216.

A general conditioning system for server room with installation of dual-channel inverter split air conditioners of Toshiba Digital Inverter RAV-SM2804AT8-E with total power of cooling of 50 kW is provided to reduce of the chiller load (or for case of their failure). Besides, the internal free-cooler of a chiller can work in case of switched-off or faulted compressor.

Concerning a cooling system for the cluster-A which will be located in the new building of Centre of innovative project developments of the NICA Complex: two options are possible for achievement of the required parameters (20 PB of disks space, 20K of CPU cores). The first one – scaling of all equipment used in cluster-216 by 1.5 times - 12 enclosures, 6 LCP, 2 120 kW chillers, two 128 kW UPS, etc. The second one – to use products of the RSC-company (Russian SuperComputers), for example, a mini-DPC (Data-processing center) architecture. A general view of such mini-DPC is presented in Fig. 17. In an racks 42U it can be set up to 112 (7 units on 16 modules) 2-CPU computing modules (224 CPUs, up to 4928 cores of the CPU) with direct liquid cooling by means of the cooling plate, an air-and-water subsystem and an external free-cooler. Such architecture doesn't require installation of a room air conditioning systems, has high energy efficiency and low level of a noise. In each Tornado module it is possible to set 2 SSD disks. Because now a price for 8 TB SSD disk is 10 times more then a price for HDD, it is reasonable to set into the modules only in one SSD disk for system and to provide a necessary big disk space with installation of the modern data storage system (DSS). If, as it is expected, the capacity of solid-state disks will considerably increased and the price will considerably fall, it is possible to provide the required disk space using only internal disks of modules (however, for example, in mini-DPC are 112 modules, i.e. 224 disks, so to obtain 20 PB it is necessary to have SSD disks with a capacity about 100 TB each!).

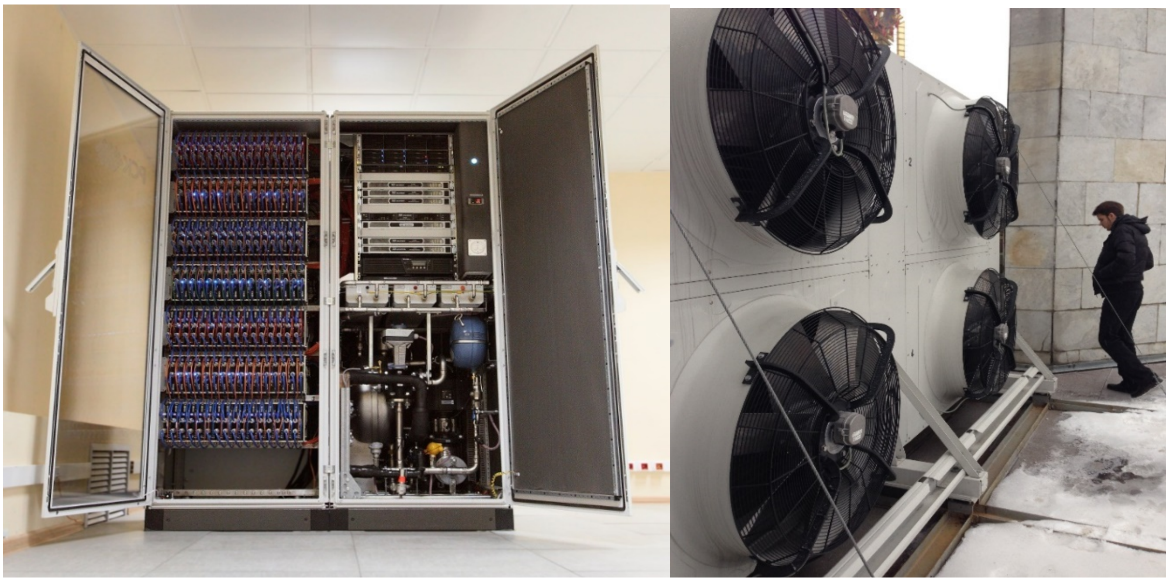


Fig. 17. Mini-DPC of RSC and free-cooler.

One of successfully used DSS is the ClusterStor L300N/L300/L6000/L9000 series of Seagate. The ClusterStor L300N (see Fig. 18) is a parallel scalable hybrid system in which both rigid and solid-state drives is used. This system is optimized for operation with units of variable length. This DSS is a full-function decision including all necessary hardware and software components. The system is placed in a standard racks 42U and can include up to 7 (first) and 8 (subsequent) units with 82 HDD of a maximum (at the moment) capacity, i.e. for disks 8 TB in one unit it can contains about 0.5 PB of data, and for disks 16 TB in one unit – more than 1 PB of data and nearly 10 PB in one rack.

The schedule of a purchase of the necessary equipment with corresponding parameters for the NICA LHEP off-line cluster for 2017 - 2023 is specified in Table №4. The equipment necessary for achievement of these parameters was choosen in the assumption that in 2018 a hard drives 3.5" with a capacity of 16 TB will be available, and in 2020 - with a capacity of 32 TB, and a 2.5" SSD will have capacity 16 TB. Quantity of cores of the CPU: in 2018 – 18, 2019 - 20, 2020 – 22, 2021 - 24, 2023 – 26. Data are provided for two versions of the equipment cluster-A: option 1 – on a basis of 2 CPUs Intel R2000WT servers and 8 CPUs R2000KP (High Density – four dual-processor modules in the 2U); option 2 – on a basis of architecture of RSC mini-DPC. Option 1 is the main for the cluster-216 because the first half of the engineering equipment is already received and will be put into operation in the 2nd quarter of this year; the second half should be received by the end of 2017.

It is also necessary to mark that option 2 for cluster-A requires a binding to the new building and detail study of the cooling system and power supply system.

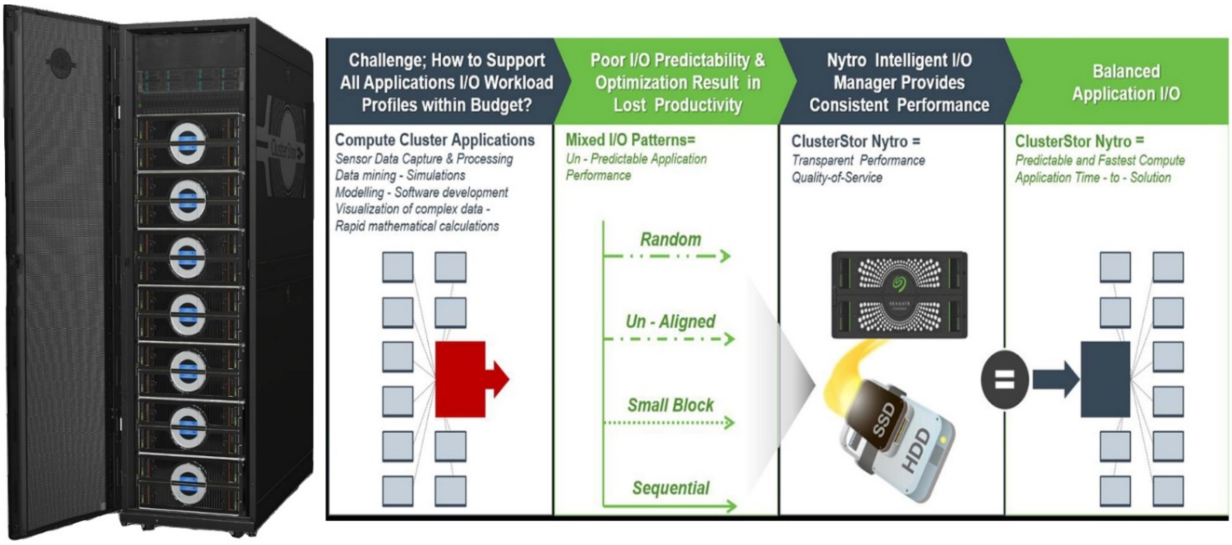

Challenge; How to Support All Applications I/O Workload Profiles within Budget?
Compute Cluster Applications
Sensor Data Capture & Processing
Data mining - Simulations
Modelling - Software development
Visualization of complex data -
Rapid mathematical calculations
Poor I/O Predictability & Optimization Result in Lost Productivity
Mixed I/O Patterns=
Un - Predictable Application Performance
Random
Un - Aligned
Small Block
Sequential
Nytro Intelligent I/O Manager Provides Consistent Performance
ClusterStor Nytro =
Transparent Performance Quality-of-Service
SSD
HDD
Balanced Application I/O
ClusterStor Nytro =
Predictable and Fastest Compute Application Time - to - Solution

Fig. 18. The Data Storage System (DSS) Seagate ClusterStor L300N.

The schedule of a purchase of the necessary equipment with corresponding parameters for the NICA LHEP off-line cluster

Table 3.

Period	Necessary parameters	Locations	Necessary server hardware (option 1 – INTEL servers)	Necessary server hardware (option 2 – architecture of RSC)
End of 2017г.	0.5PB HDD (with replication), 1К CPU cores	Cluster-216	14 servers INTEL R2000WT 18x2x2x14=1008 CPUs cores 12x8x14=1344TB disk space	See Note 1.
End of 2019г.	4-5PB HDD (with replication), 4К CPU cores	Cluster-216	80 servers INTEL R2000WT 20x2x2x80=6400 CPUs cores 12x12x60=8640TB + + 12x16x20=3840TB == 12480TB disk space	See Note 2.
End of 2020г.	8-10PB HDD (with replication), 5К CPU cores	Cluster-A	60 servers INTEL R2000WT 22x2x2x60=5280 CPUs cores 12x32x60=23040TB disk space	64 modules (4 units) 22x2x2x64=5632 CPUs cores SDS ClusterStor L300N: 4 SSU: 82x32x4=10496TB disk space
End of 2021г.	12-15PB HDD (with replication), 10К CPU cores	Cluster-A	30 servers INTEL **R2000KP** (HD: 4 x 2CPU at 2U) 24x4x30x4=11520 CPUs cores 12x32x30=11520TB disk space	48 modules (3 units) 24x2x2x48=4608 CPUs cores SDS ClusterStor L300N: 2 SSU: 82x32x2=5248TB disk space
End of 2023г	20PB HDD (with replication), 20К CPU cores	Cluster-A	30 servers INTEL **R2000KP** (HD: 4 x 2CPU at 2U) 26x4x30x4=11520 CPUs cores	48 modules (3 units) 26x2x2x48=4992 CPUs cores SDS ClusterStor L300N:

			12x32x30=11520TB disk space	2 SSU: 82x32x2=5248TB disk space

For commissioning of an object "Offline a computer cluster of LHEP in 216-115" it is necessary to prepare a set of the project documentation of engineering infrastructure. Detailed requirements of subsystems of power supply, cooling, conditioning, fire extinguishing and a raised floor will be included in the technical specification (TS), and also the diagram of arrangement of the equipment is provided. Here we will tell about only briefly.

Power supply – the server and communication hardware, and also electronics of control of chillers are connected to two 96 kW Concept power DPA 150 UPSs, to which the racks of power distribution of PDR with the set modules of secondary distribution of power supply of PDM having the principal switch and the protected 3-phase outputs connected to racks with the equipment. In turn, UPSs receive power supply on cables from electro baffle (location - 216-113).

Cooling – air-to-water on a basis of two Rittal chillers with a power of cooling of 80 kW and an internal free-cooling to 120 kW, and 4 LCP InLine Rittal heat exchangers of 40 - 55 kW. Reserve cooling and the general conditioning of servers room consists of two inverter a split system of Toshiba Digital Inverter RAV-SM2804AT8-E with 50 kW total power of cooling.

Fire detection and fire extinguishing – the modern installations on a basis of automatic (smoke, thermal, combined, etc.) fire annunciators; the integrated systems of the security and fire warning, fire detection and fire extinguishing; two-component gas fire extinguishing system: intra rack-mount – Novec gas, the general indoors – R-125 freon.

The diagram of the equipment arrangement is presented in Fig. 19.

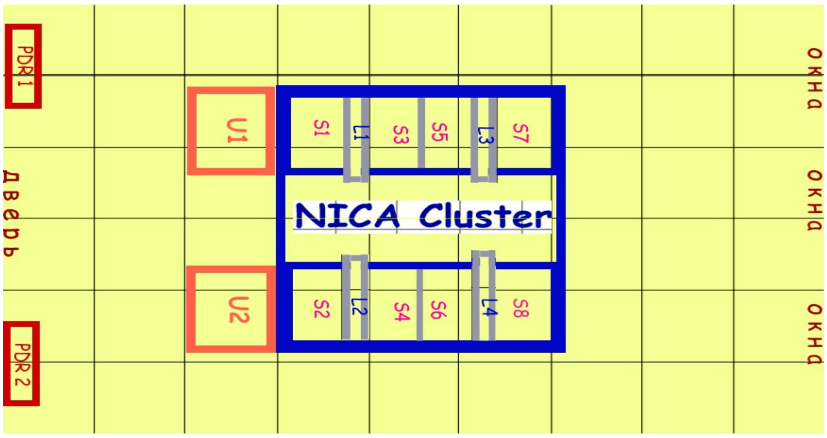


Fig. 19. A diagram of the of the Cluster 216-115 equipment arrangement.

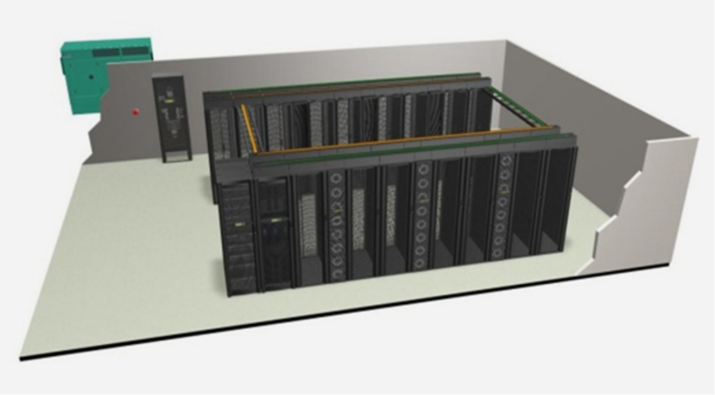


Fig. 20. Prototype model of the Off-Line cluster

The enclosures (S1-S8) for servers have the sizes 1200x600x2000mm,
The LCP (L1-L4) – 1200x300x2000mm,
The UPS (U1-U2) – 1000x800x2000mm.
Weight of racks: 300 (racks) + 20 servers (20-25kg) > = 800 kg.
LCP weight: 300 kg.
UPS weight: about 1100 kg. (407+2.65x240=1043kg)

Raised floor – Lindner Nortec U 36 ST/PVC of a plate from calcium sulfate, from below the sheet of steel. Raised floor height: 430 mm. C-profiles (C3-2mm) are set under racks frames to provide a possibility of removing a part (or all) of plates for access from racks to underground space. All pedestals are fastened by special stringers.

Note 1. The equivalent configuration: 16 modules of the Tornado (1 unit) of 18x2x2x16=1152 CPU core, ClusterStor L300N SDS, 1 SSU: 82x8=656TB of disks. However, comparing isn't included in the table since architecture for a cluster-216 is option 1.
Note 2. The equivalent configuration: 64 modules of the Tornado (4 units) of 20x2x2x64=5120 CPU core, ClusterStor L300N SDS, 4 SSU: 82x16x4=5248TB of disks. However, comparing isn't included in the table since architecture for a cluster-216 is option.

3.2.2 Off-Line cluster of “NICA Center”

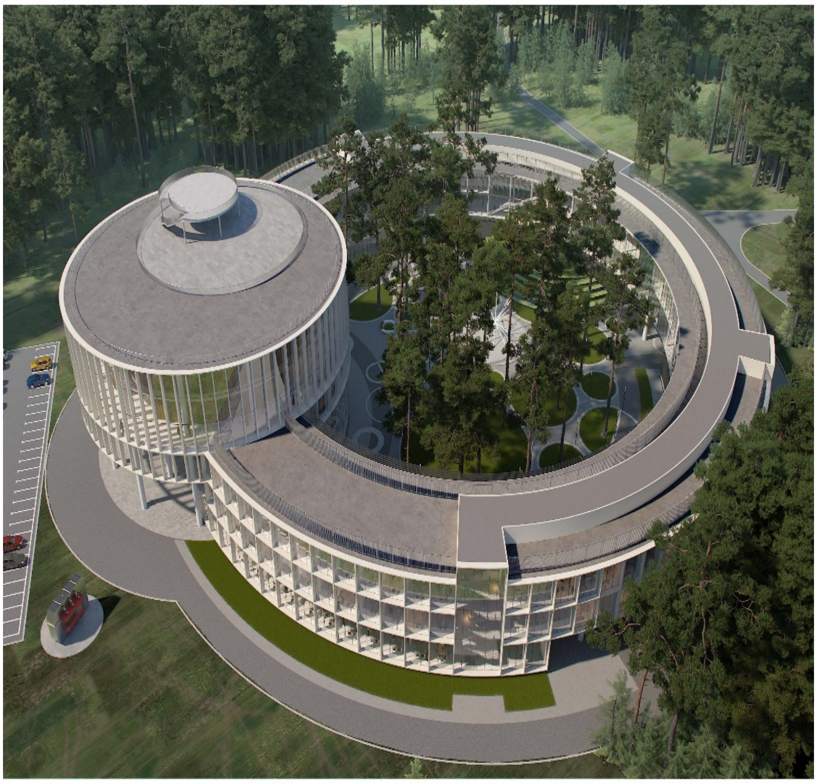
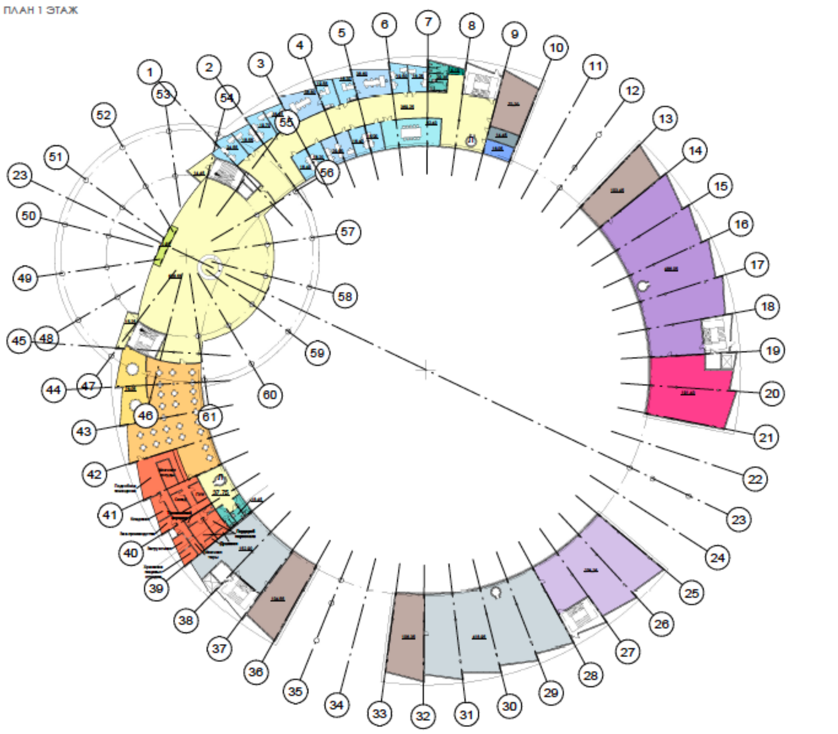


Fig. 21. NICA Center building



ЭКСПЛИКАЦИЯ:

	Имя	Площадь	Кол-во
🟦	Архив	14,45	1
🟫	Венткамеры	386,55	4
🟥	Диспетчерская	191,40	1
🟦	Документация и печать	16,05	1
🟦	Зал для совещаний (лаборатория)	62,40	1
⬜	Инженерные помещения	571,85	2
🟧	Кухня при столовой	197,10	17
🟦	Офис	222,10	12
🟩	Офис 2	107,25	3
🟩	Помещение уборочного инвентаря	4,10	1
🟩	СУ	38,75	2
🟪	Склад	276,25	1
🟧	Столовая	264,40	1
🟨	Столовая для VIP	76,05	1
🟩	Тамбур	12,55	1
🟨	Фойе	1075,30	5
🟪	ЦОД	486,05	1

Fig. 22. Computing hall of NICA center

3.2.3. LIT-Off-line cluster

3.2.3.1. Development of MICC network structure at JINR

According the plans of development of the resource component of MICC center it will be needed to provide network support for the second module, which includes 80 disk servers (160 ports of 10 Gbps in the aggregation mode), 15 blades (30 ports of 10 Gbps in the aggregation mode), 60 servers of infrastructure (40 ports of 10Gbs and 40 ports of 1 Gbps in the aggregation mode). Thus, it is necessary to have a network segment of 230 ports of 10 Gbps and 40 ports of 1 Gbps.

Interaction of the two modules (aggregation 2 Virtual Cluster Switching (VCS) fabric) will be performed at the 3-d level of the OSI model, using the OSPF routing protocol. Each distribution layer switch will be connected to the multi-service IP/MPLS SDN Cisco Nexus 9504 router with the help of high-speed communication channel of 200 Gbps. The data rate of 200 Gbps can be achieved by using the protocol Link Aggregation Control Protocol (LACP) or Port Aggregation Protocol (PAgP). The proposed aggregation of two modules diagram is shown in Figure 23.

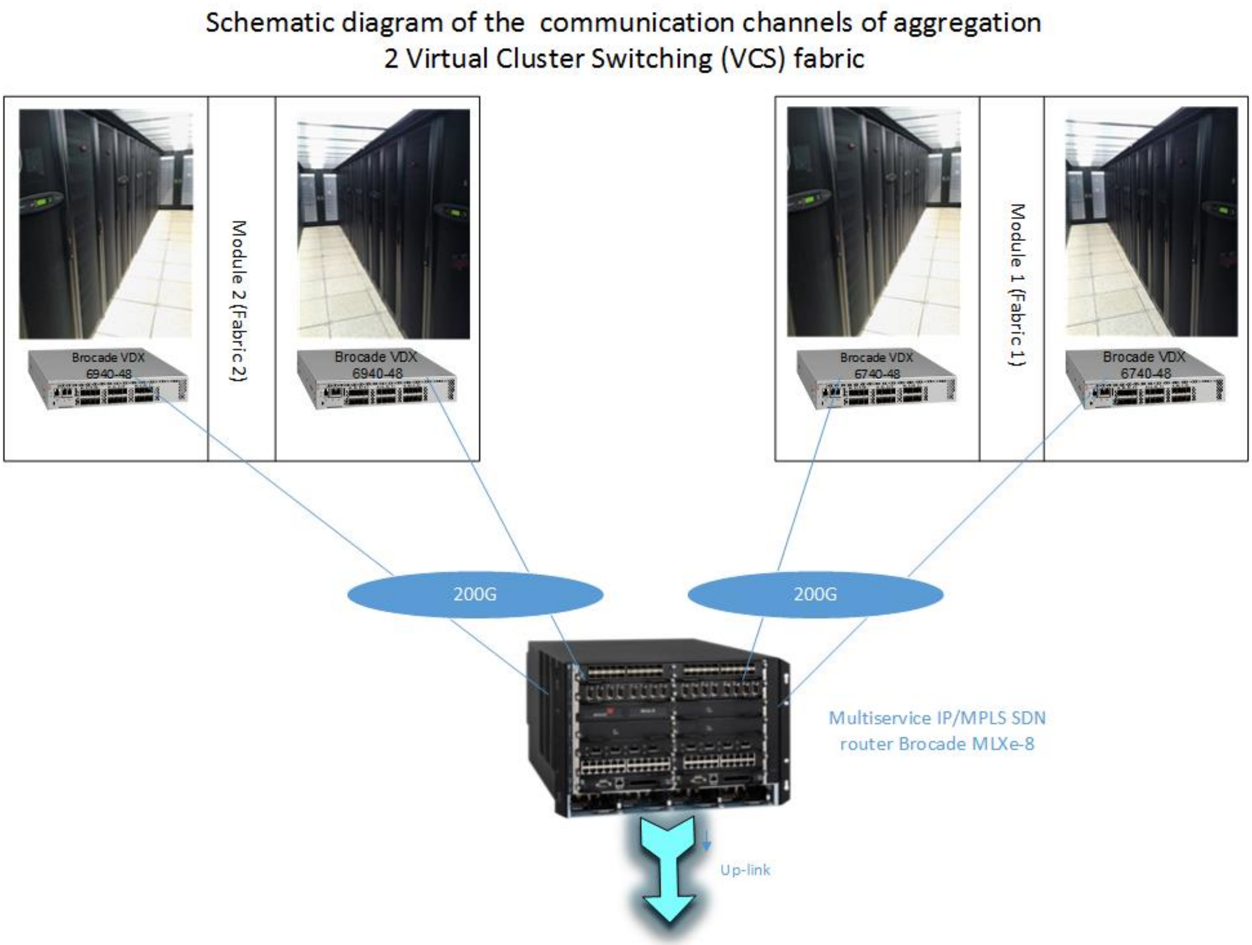

Schematic diagram of the communication channels of aggregation 2 Virtual Cluster Switching (VCS) fabric
Brocade VDX 6940-48
Module 2 (Fabric 2)
Brocade VDX 6940-48
Brocade VDX 6740-48
Module 1 (Fabric 1)
Brocade VDX 6740-48
200G
200G
Multiservice IP/MPLS SDN router Brocade MLXe-8
Up-link

Fig. 23. Diagram of proposed aggregation of two modules

3.2.3.2. Development of the network structure MICC

Improving the performance of network systems segment MICC will increase the speed of access to data stored in the databases of the experiments. At the same time MICC is not required high reliability of the communication channels, therefore, the new data network will be simpler than that of Tier–1, but will be built on the same principle. The scheme of the MICC network is given in Figure 24.

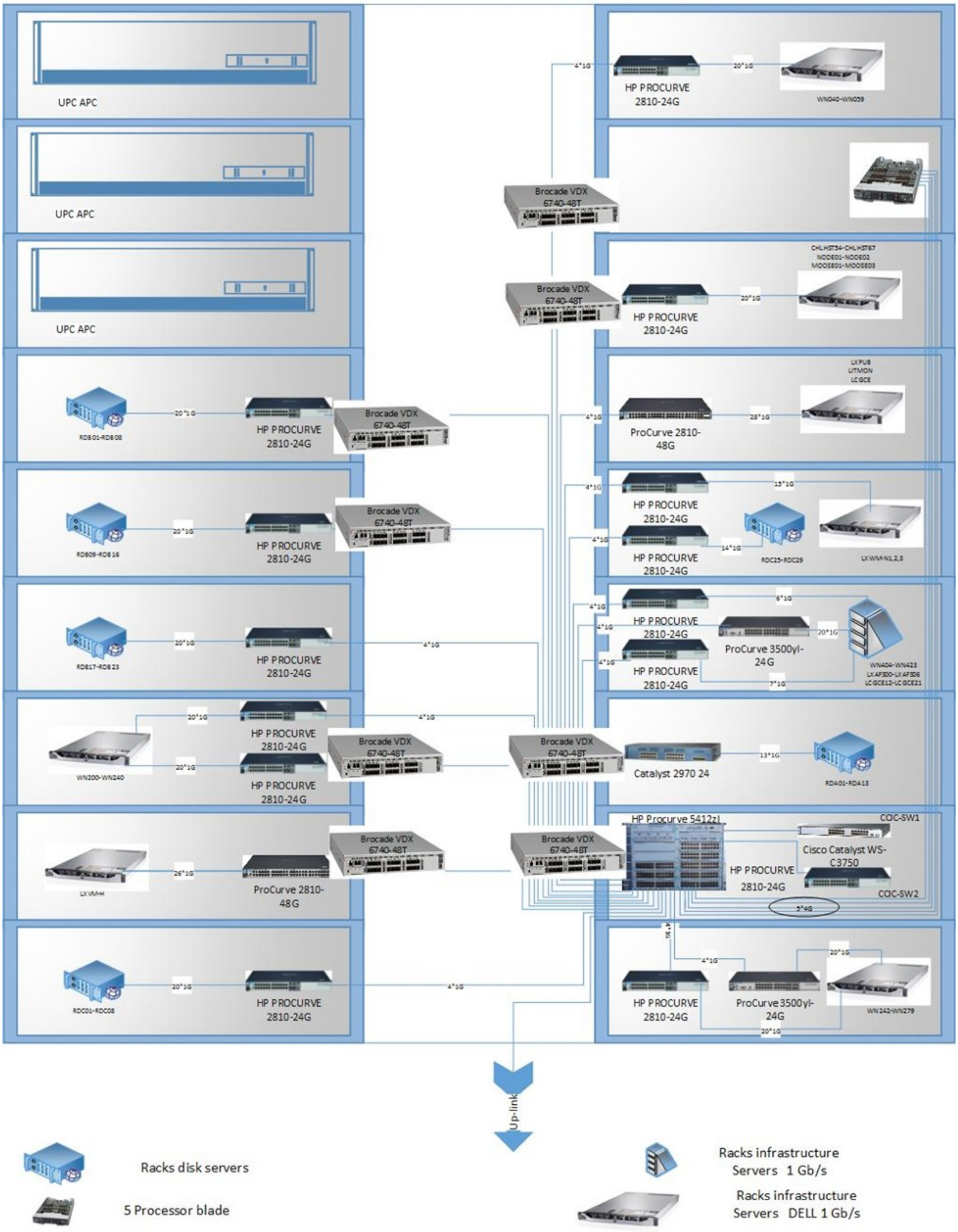


Fig. 24. Proposed scheme of the MICC network

3.2.3.3. Organization of the data storage for the NICA experiment

The primary objective is the creation of a two-level (disks-tapes) of the storage system for the NICA experiment because after a first phase of running this experiment significant amounts of storage (up to 2.5 PB per year) will require. This task is subdivided into 2 subtasks:

1 provision in the nearest future of taking information using the existing infrastructure,

2 design of a comprehensive project of a repository in accordance with the model of data processing which is currently in the process of the development.

To implement a first subtask, if to take into account 2-2.5 PB per year during a first run, it is necessary to allocate or purchase RAID of the volume of no more than 120 TB (preferable 2x60TB), connecting these servers to the already installed drives in the library with the following allocation in the library of a group of slots (partitioning into logical libraries). Purchasing of additional frames with tapes is shown on the Table №4.

Table №4

Product	**Description**	**Quantity**
3584-S54	TS3500 HD Frames for LTO Drives	2
1646	HD COD for S54/S55	2
3589-550	2.5 TB Ultrium Tape Cartridge Labeled	1
5500	2.5 TB Labeled 20-pack	66

3.2.3.4. LIT Off-Line cluster File system

The software should be placed on a cached file system so that it is available on all data centers.

Storage systems (Fig. 25.) have been installed in dCache software. One of the dCache installations is only used with disk servers and used for operational data storage with fast access to them. The second dCache unit includes disk servers and a tape robot. The disks serve as a buffer zone for exchange with tapes, while the tape robot is intended for a long-time, practically eternal, storage of data from the LHC. Totally, 2 installations have now 3.4 PB of effective disk space, and the tape robot has a 5.4 PB of data storage capacity. To support the storage and access to data, 8 physical and 14 virtual machines have been installed.

Fig. 25 Type robot of the storage system

3.3. Off-line software

3.3.1. Computing frameworks for the NICA experiments

The objectives of the experiment computing frameworks are the simulation of the primary interactions with the realistic detector response and the reconstruction and analysis of the data coming from simulated and real interactions. The frameworks for each experiment of the NICA project - MPD, BM@N and SPD, are named as MPDRoot, BMNRoot and SPDRoot correspondingly, and differ with only description of different sets of detectors. These frameworks are inherited from the former CBMRoot framework, have the same structure and use the same external packages like ROOT, FAIRRoot, FAIRsoft. FAIRsoft package includes external packages for the software development like BOOST, GSL, GEANT4(3), Millepede and ZeroMQ. All of these packages are free, available under the LGPL license and works at Linux flavors operational systems.

3.3.2. Software for Simulation

Many Monte-Carlo generators are available for the physics simulation of heavy ion collisions: UrQMD, QGSM, pHSD, Hybrid UrQMD and THESEUS. Mainly these generators are used to study detectors responses but some of them can be also used for investigation the feasibility of some physics signatures.

3.3.3. Databases

During the experimental runs there is a lot of information, which must be stored for the further offline experimental data analysis. These databases include:

- Detector Construction Database (DCDB): used by individual sub-detector groups during the production and integration phase and containing static information on detector elements performance, localization and identification;

- Experiment Control System (ECS) database: contains information on the active sub-detector partition during data taking and the function of this partition;
- Data Acquisition (DAQ) database: a repository for data acquisition related parameters and for resource assignments to data acquisition tasks: current and stored configurations, current and stored run parameters;
- Trigger database: contains the experiment trigger classes and the definition of the trigger masks;
- Detector Control System (DCS) databases: configuration DB, containing the configuration parameters for systems and devices (modules and channels), and the front-end configuration (busses and thresholds);
- Archive DB containing the monitored detectors and device parameters.
- High-Level Trigger (HLT) database: a TAG/ESD database containing HLT information relevant for physics studies and offline event selection;
- NICA Machine database: machine status and beam parameters.

To work with these relative data for the NICA experiments it plans to use the open source relational database PostgreSQL.

At the same time many commercial software systems are used for the different tasks of the NICA project. For designing and creating accelerator systems and detectors of the NICA complex are used CAD systems and Autodesk Inventor. Special systems as Vidyo, Wowza

Media Systems are used for organizing effective computer video-conferences in the process of project implementation. The Microsoft software is used on computers with specialized systems for analytical calculations and modeling, systems for NICA project management (ADB2-EVM), in guest rooms and video-conference rooms of the complex.

3.3.4. Applied equipment and software

As part of the NICA Complex, the construction of the NICA Centre building is being designed, where in addition to the off-line computing cluster described above, 6 video conference rooms and a large transformable conference hall have been designed. Similar halls are already used in works on the project "NICA Complex" in buildings 215, 3, 20, 201. All these rooms and halls are and should be equipped with computer video communication systems, supported by appropriate software, projectors, video monitors and screens. A significant number of guest rooms in the building of the NICA Centre and the existing buildings of the LHEP site are also equipped with computer and network equipment in an amount sufficient for the work of the specialists who are participants of international teams implementing the NICA project.

4. Schedule and cost estimation

The quarterly schedule for commissioning the information and computer complex NICA is shown in Table 5.

Table 5

	2017-IV	2018-I	2018-II	2018-III	2018-IV	2019-I	2019-II	2019-III	2019-IV	2020-I	2020-II	2020-III	2020-IV
On-line Cluster NICA	10%	-	-	50%	-	70%	-	-	100%	-	-	-	-
Off-line Cluster NICA	30%	10%	-	-	60%	-	-	-	100%	-	-	-	-
Off-line Cluster LIT NICA	5%	-	-	30%	-	-	70%	-	100%	-	-	-	-
Cluster Center NICA	0%	-	-	-	-	-	-	-	-	-	40%		60%

References

1. Agapov N N, Kekelidze V D, Kovalenko A D, Lednitsky R, Matveev V A, Meshkov I N, Nikitin V A, Potrebennikov Yu K, Sorin A S, Trubnikov G V “Relativistic nuclear physics at JINR: from the synchrophasotron to the NICA collider” *Phys. Usp.* **59** 383–402 (2016).

2. A. Baskakov, S. Bazylev, A. Fediunin, I. Filippov, S. Kuklin, Yu. Minaev, A. Shchipunov, A. Shutov, I. Slepnev, V. Slepnev, N. Tarasov, A. Terletskiy. MPD Data Acquisition System Technical Design Report Version 0.6. http://mpd.jinr.ru/wp-content/uploads/2017/06/mpd_daq_tdr_v0.6.pdf. 18 May, 2017.