Information Technologies at JINR

Executive Summary

The mission of the Laboratory of Information Technology, LIT, is two-fold:

a) To serve the scientists of JINR and its Member States in the pursuit of their research projects by developing *'Methods, Algorithms and Software for Modelling Physical Systems, Mathematical Processing and Analysis of Experimental Data'*

b) To assure that the IT-infrastructure and IT know-how of JINR experts is always of latest state of the art

Since JINR and other big science centres depend on the commodity of IT-hardware, a long-term view must look for great flexibility in the system. Presently, the JINR Infrastructure has been developed in close connection with CERN and other Institutes of Nuclear and High Energy Physics. During the last 15 years, a distributed computing infrastructure has been created for processing and storing data from the LHC experiments. Each of the four outstanding scientific experimental facilities at the LHC, namely ATLAS, CMS, Alice and LHCb, are being run by collaborations, each of which has several thousand scientists from several hundred institutes distributed worldwide.

Therefore, a geographically distributed computing environment like grid for processing and storing experimental data of all LHC experiments, i.e. the Worldwide LHC Computing Grid (WLCG), has been created. About 1 000 000 processors are geographically distributed in more than 170 data processing centres from 42 countries. They are combined into a unified computing environment within the WLCG infrastructure, capable of managing hundreds of petabytes of data, and providing access to computing resources and data storage systems for the entire community, and integrating national and international structures. JINR is a strong part in this system.

Projecting the development of Information Technologies of the Institute into the future must take into account the latest development of some large experiments, where huge computing power is or will be installed at the front end of the detectors, thus substantially reducing the amount of data transmitted to the distributed computing centres in the World. At the same time, data taking rates in the LHC experiments are steadily increasing, leading to an increase in data volumes.

The original concept of grid as an implementation of the HTC (High Throughput Computing) concept has changed to a complex, heterogeneous computing system combining computing resources of various concepts: HTC, HPC (High Performance Computing), computing resources provided on a voluntary basis (Volunteer computing), commercial and non-commercial cloud computing resources.

Currently, research on the development of novel architectural and functional principles of the WLCG distributed computing infrastructure is carried out by the community of LHC computer specialists [1] and other scientific world centres in the field of high-energy physics; in 2020, it is planned to complete the preparation of relevant technical projects (Computing TDR). JINR specialists are part of today's community of experts.

It should be noted that the JINR research program for the next decades is aimed at conducting ambiguous and large-scale experiments at the Institute basic facilities and in frames of worldwide collaborations. This program is connected with the implementation of the NICA mega-project, the construction of new experimental facilities, the JINR neutrino program, the modernization of the LHC experimental facilities (CMS, ATLAS, ALICE), the programs on condensed matter physics and nuclear physics. The implementation of the projects mentioned above requires adequate and commensurable investments in the IT systems providing the processing and storage of increasing data volumes. The experience of recent years shows that the progress in obtaining research results directly depends on the performance and efficiency of computing resources. In this regard, the further development and performance extension of the JINR LIT Multifunctional Information and Computing Complex (MICC) [2,3] as well as the provision of novel IT-solutions to the Complex users and the increase in its operation efficiency are the uppermost tasks of the Laboratory of Information Technologies.

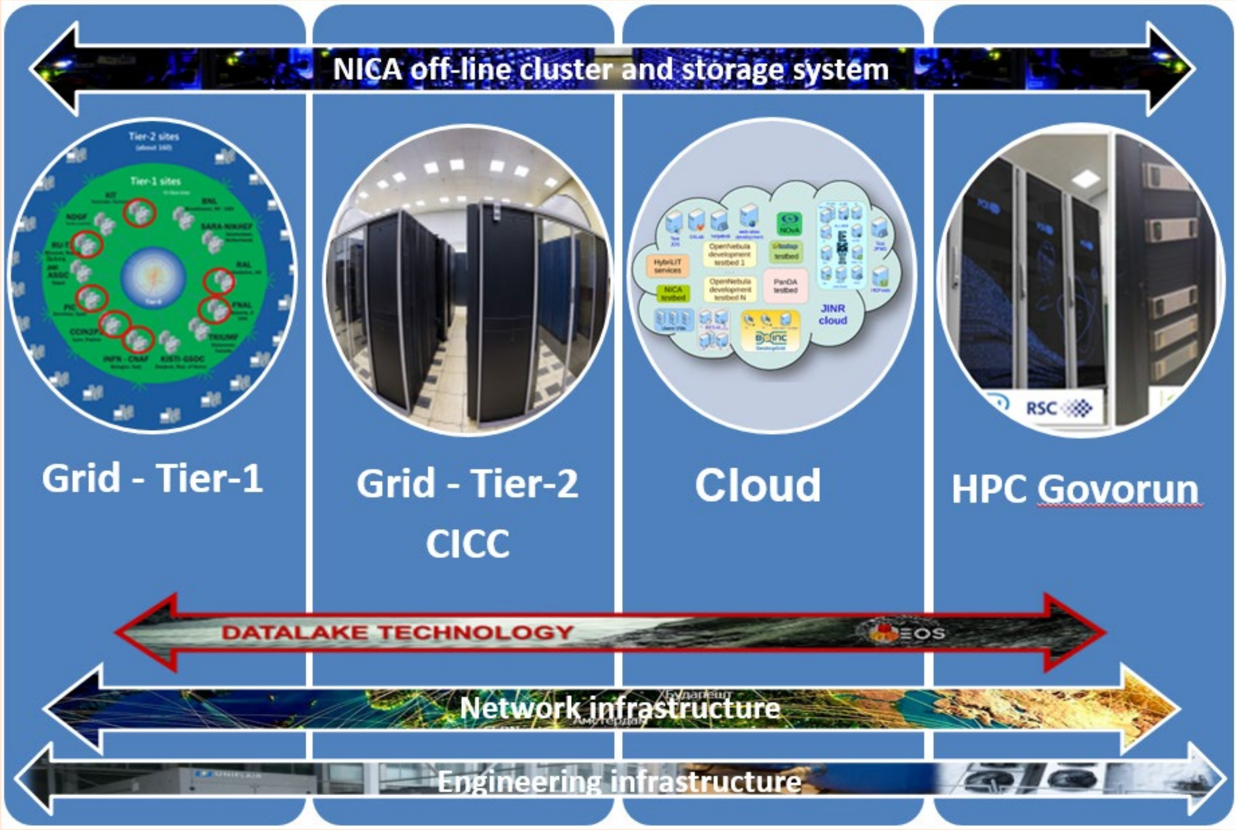
The JINR computing infrastructure consists of numerous computing components and IT-technologies to solve JINR tasks, from theoretical studies to experimental data processing, storage and analysis. The JINR MICC is the key element of this infrastructure and plays a defining role in research, which requires modern computing power and data storage systems. They are the IT-ecosystem for the NICA project (BM@N, MPD, SPD), Tier1 of the CMS experiment at JINR, Tier2/CICC providing support to the experiments at the LHC (ATLAS, ALICE, CMS), FAIR (CBM, PANDA) and other large-scale experiments as well as support to users of JINR Laboratories and the JINR Member States (MPD/NICA, BESIII, LRB, FLNR, DLNP, BLTP, LNP); the integrated cloud environment of the JINR Member States [4] for support of JINR users and experiments (NICA, ALICE, BESIII, NOvA, Daya Bay, JUNO, etc.); the HybriLIT platform [5] with the "Govorun" supercomputer as a major resource for high-performance hybrid computing.

In summary, LIT plans to establish an IT-ecosystem, i.e. a dynamically growing IT platform, which responds to the rapidly developing IT-World. The connection to CERN and FAIR will influence the JINR strategy. Different to most other IT-groups in the World, JINR LIT is providing and will provide forefront service to scientists working in JINR collaborations, on- and off-site of Dubna, by continuing developing '*Methods, Algorithms and Software for Modelling Physical Systems, Mathematical Processing and Analysis of Experimental Data'.* Although many of these tasks are being done by individual experts in experiment collaborations, LIT as a base of these experts can guarantee a high level of professionalism strongly needed in Data- and Code-handling.

The LIT manpower development must reflect the fast-changing world of IT by providing attractive positions to both promising young and highly qualified IT-specialists.

Today's Computing at JINR

The JINR computing complex has undergone numerous transformations over the years. Today the Multifunctional Information and Computing Complex (MICC) of the Laboratory of Information Technologies (LIT) at JINR is the key element of the computing infrastructure and plays a defining role in research, which requires modern computing power and data storage systems. The MICC is regarded as the JINR basic facility "Networks and ECM", which is a combination of complexes, subsystems and other organizational units, including WLCG grid centres of Tier1 and Tier2 levels for data processing and storage; the cloud infrastructure; the HybriLIT platform with the "Govorun" supercomputer, data storage; the data transmission network; the engineering infrastructure; the monitoring system. Today the main goal is the effective operation of all the components and the centre as a whole, which is able to store a diverse set of equipment.



Structure of the MICC

A modern JINR computing complex should meet the following requirements: multi-functionality, high performance, high reliability, fault-tolerance and availability, information security, scalability as well as a customized software environment for different user groups. To fulfil these requirements in a 24x7x365 mode, the MICC needs a constant upgrade and expansion of the complex's possibilities. The rapid development of information technologies does not allow one to fully define specific solutions that will determine the MICC development for the coming years; however, the trends for this upgrade are clear enough.

The experience gained in LIT shows the necessity of taking into account the expected technological evolution in meeting future user needs. Moreover, there is a tendency to more extensive use of specialized equipment to perform complex tasks computationally. The latter means that the deployment of heterogeneous equipment will be required in data processing centres, and experience and skills will be needed to support these resources.

Network infrastructure at JINR

A high-performance network infrastructure is crucial for the development of IT in the modern world. In the last year, the following projects were carried out at JINR:

- Work related to increasing the bandwidth of the Moscow-JINR telecommunication channel to 3x100 Gb/s,
- Installation and configuration of the equipment of the Institute backbone computing network to 2x100 Gb/s and the distributed computing cluster network between JINR facilities to 400 Gb/s.

The development of the JINR network infrastructure is dependent on the modernization and increase in the bandwidth of external telecommunication channels, the JINR local area network, the MICC network infrastructure and the NICA megaproject. In the near future, it is planned to develop the communication component of network subsystems of the MICC and the NICA megaproject using the Multisite cluster network technologies at a speed of more than 400 Gb/s to create a network infrastructure for acquiring and transmitting data between the BM@N, MPD, SPD facilities and on/off-line clusters. The given direction presupposes laying new optical network links between LIT and VBLHEP as well as ensuring maximum reliability and fault tolerance of this network.

Increasing the performance of the network segment of the MICC/Tier2 system will allow speeding up access to data stored in the databases of the LHC experiments.

It is noteworthy that the increase in the bandwidth of external and internal telecommunication channels beyond the present plan will be defined by the needs of network-connected facilities, including the needs of the NICA megaproject.

One more important direction is the development of network services supporting the network functioning, namely Network services (NOC Cluster) of e-mail, name management (DNS), data caching (Proxy), resource management (IPDB), authorization (Radius, Tacacs, Kerberos), monitoring (NMIS); Service of Single Sign-On (SSO); Information security system; System of testing of users for the knowledge of the operation rules in the JINR computing network, etc.

Grid at JINR

The grid centre resources of the JINR MICC are part of the global grid infrastructure WLCG (Worldwide LHC Computing Grid), developed for the LHC experiments. JINR LIT actively participates in the WLCG global project. The work on the use of the grid infrastructure within the WLCG project is carried out in cooperation with the collaborations such as CMS, ATLAS, Alice and major international centres, which operate as Tier1 centres of the CMS experiment (CH-CERN, DE-KIT, ES-PIC, FR-CCIN2P3, IT-INFN-CNAF, US-FNAL-CMS) and as Tier2 grid centres located in more than 170 computing centres of 42 countries worldwide. Since the beginning of 2015, a full-scale WLCG Tier1 site for the CMS experiment at the LHC has been operating in JINR LIT.

The importance of developing, upgrading and expanding the computing performance and data storage systems of the centre is dictated by the research program of the CMS experiment, in which JINR physicists take an active part within the RDMS CMS collaboration. The fact of the creation and support of the work of the JINR Tier1 site demonstrates recognition of the high qualification level of the JINR LIT staff ensuring the functioning of this MICC component. There are only seven similar CMS centres in the world, and the JINR site is regularly ranked second in the number of processed events showing almost 100% level of availability and reliability.

Tier2 at JINR operates within the MICC computing component, which currently provides the majority of JINR users and user groups, as well as users of virtual organizations (VO) of the grid environment, with computing power, storage systems and access to data. For supporting a batch processing system, a special server with a cluster resource allocation system and a job scheduler is installed; it allows processing data for all four experiments at the LHC (Alice, ATLAS, CMS, LHCb) and supports some virtual organizations working within grid (BESIII, BIOMED, COMPASS, FUSION, MPD, NOvA, STAR, etc.).

In terms of hardware, a linear increase in the characteristics of Tier1, Tier2/CICC is planned.

It is planned to launch Run3 at the LHC by the summer of 2021. The CMS and ATLAS plans do not require a cardinal increase in computing and storage resources at this stage. Alice plans to increase its resource needs for Run3 almost by a factor of two. The fulfilment of the seven-year plan figures will ensure the required level of resources for all LHC collaborations at Tier1 and Tier2 in JINR.

In the coming years, it is necessary to increase the information storage capacity on the EOS system, which was put into operation at the beginning of 2019. This data storage and access system should become the major system for all MICC components, and later for all JINR computing resources. There are prospects of using EOS for WLCG collaborations.

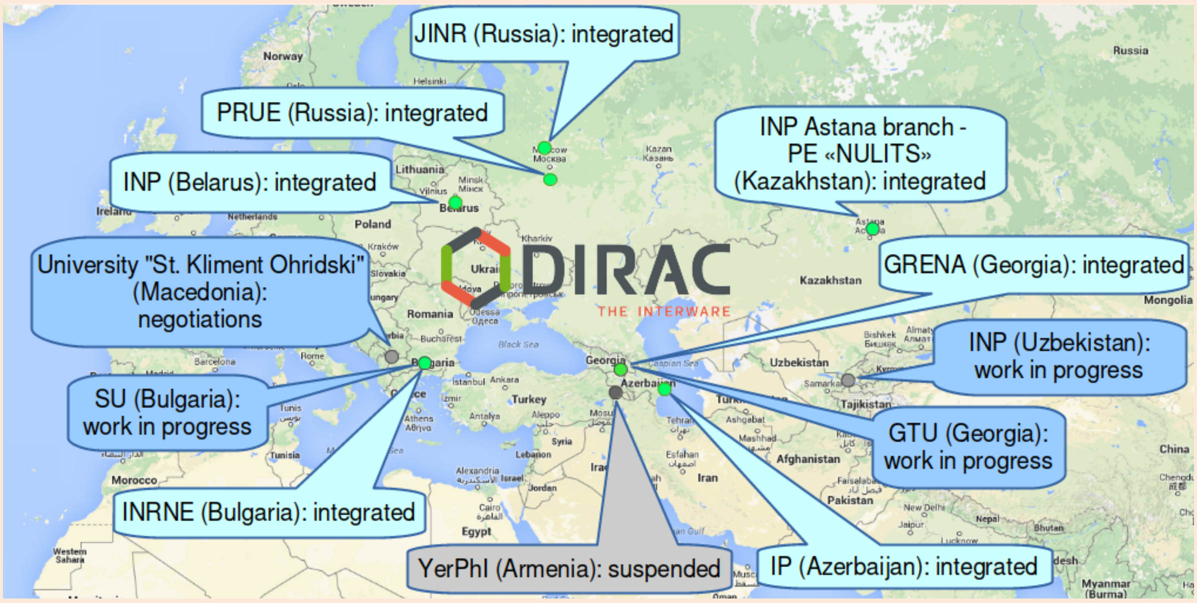
Experiments that are processed by the MICC within and outside grid

The development of the data storage system on robot tape libraries is required, however, the development of other storage media is to be closely watched. In addition to this facility for the CMS experiment at Tier1, it is required to create a facility for the NICA experiments in 2020–2021, as well as other WLCG collaborations and user groups. The update of the existing tape library to increase its capacity is needed.

Cloud infrastructure at JINR

As the resource base of the JINR cloud develops, its computing resources and storage disk space are planned to be provided to individual JINR users, as well as to various scientific projects, for which JINR has certain obligations. A significant increase in the demand for cloud resources is expected from neutrino experiments (NOvA, JUNO, Daya Bay, Baikal-GVD, etc.). The transition of users of the mentioned above experiments to the use of centralized cloud resources will eliminate the burden for local scientific groups to independently support computing resources and data storages, and they will be able to fully concentrate on their research.

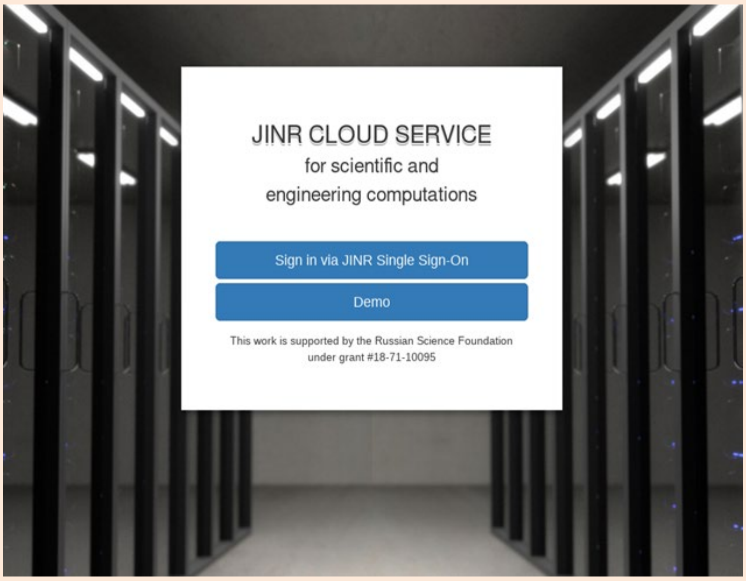
One of the directions of the cloud technology development worldwide is the integration of cloud resources of geographically distant organizations.



Geographical location of the organizations of the JINR Member States involved in the JINR distributed information and computing environment

The work in a similar direction is also being carried out at JINR LIT together with the organizations from its Member States, such as the Astana branch of the Institute of Nuclear Physics, i.e. the private establishment "Nazarbayev University Library and IT Services" (INP Astana, Kazakhstan), the Scientific Research Institute for Nuclear Problems of the Belarusian State University (INP, Minsk, Belarus), Yerevan Physical Institute (YerPhi, Yerevan, Armenia), the Institute of Physics of the National Academy of Sciences of Azerbaijan (IP, Baku, Azerbaijan), Plekhanov Russian University of Economics (PRUE, Moscow, Russia), the Institute for Nuclear Research and Nuclear Energy of BAS (INRNE, Sofia, Bulgaria), Sofia University "St. Kliment Ohridski" (SU, Sofia, Bulgaria), Georgian Research and Educational Networking Association (GRENA, Tbilisi, Georgia), Georgian Technical University (GTU, Tbilisi, Georgia), the Institute of Nuclear Physics (INP, Tashkent, Uzbekistan) and the University "St. Kliment Ohridski" (Bitola, Macedonia). The figure shows a map indicating the geographical location of the participating organizations of the **C**loud **C**omputing and **I**nformation **P**latform (CCIP). At present, the integration of the JINR Member State organizations' clouds into the DIRAC-based distributed platform is at different stages. Their technical integration has been accomplished, while the work on testing and debugging is in progress.

The work on enabling the use of resources of the given CCIP by users of experiments at the NICA accelerator complex and neutrino programs, in which JINR takes part and which do not have their own distributed infrastructure (for example, Baikal-GVD), is in progress.



Screenshot of the login page of the cloud service for scientific and engineering calculations

With regard to building CCIP, it seems promising to study the possibility of integrating the distributed information and computing environment of JINR and the organizations of its Member States into the EOSC infrastructure with support of scientific services provided by the platform.

HPC and "Govorun" supercomputer

JINR LIT is closely observing the development of Computing projects in the EU. There are two major trends of the HPC development, i.e. a significant increase in performance up to a post-exaflop scale and the integration of existing supercomputers.

In particular, within the EuroHPC project created in 2018 with the participation of 25 European countries by the end of 2020 it is planned to equip the European Union with a world-class supercomputer infrastructure, which will be available to users from science, industry, small and medium enterprises and the public sector. It is planned to put into operation two exaflop systems by 2023. The EU has also launched a project in the field of combining supercomputers including grid technologies for creating a supercomputer infrastructure PRACE (Partnership for Advanced Computing in Europe). By 2028, according to the document Update on European HPC from March 2018, which outlines the landscape of the European HPC ecosystem, post-exaflop supercomputers, as well as hybrid HPC+Quantum Computing systems, are expected to emerge. JINR HPC specialists are going to evaluate the experience and best practices of these European projects. LIT is working on a possible cooperation to provide interoperability between the "Govorun" supercomputer and other supercomputer centres.

In the USA, there are a number of programs aimed at the development of supercomputer systems including:

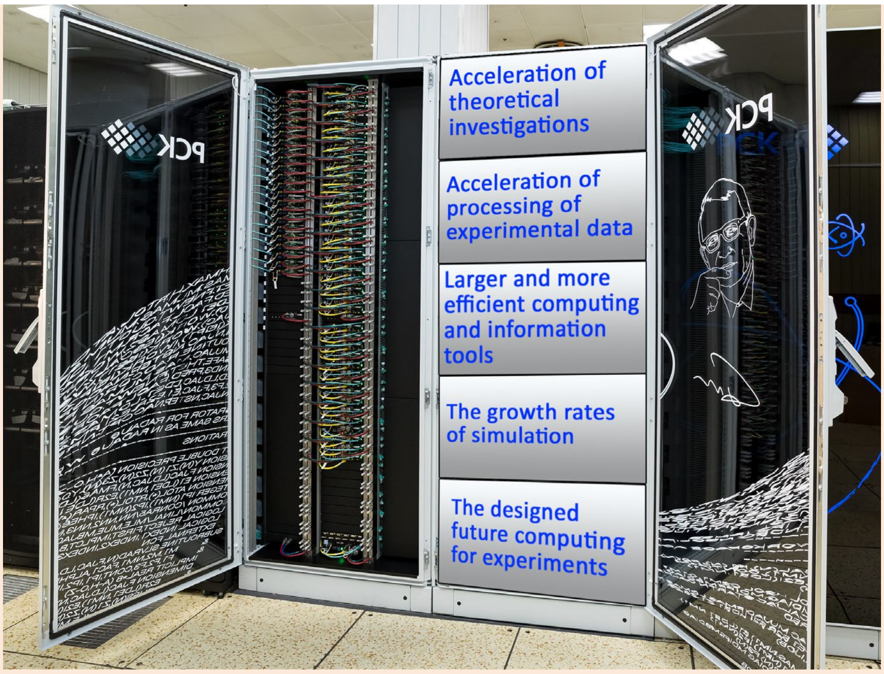
- CORAL (Collaboration of Oak Ridge, Argonne, and Livermore), the goal of the program is to build in the USA three supercomputers with a capacity of more than 1 exaFLOPS each. Machines for ALCF and OLCF have already been announced, but they are expected to be commissioned in 2021–2022;
- ASCR (Advanced Scientific Computing Research), a program aimed at the development of computing for research needs of the Department of Energy;
- ECP (Exascale Computing Project), a project for preparing the next generation of supercomputers for the use. It is focused on the development of compatible software.

The major goal of modern experiments on the collision of ultra-relativistic heavy ions, in which JINR participates, is to study the properties of quark-gluon matter at extremely high energy density and/or baryon charge density in the presence of strong electromagnetic fields. The RHIC and LHC experiments study quark-gluon matter at extremely high energy density and almost zero baryon density. In contrast to these experiments, the main purpose of the experimental programs of the accelerator and detector complexes NICA (JINR, Dubna) and FAIR (GSI, Darmstadt) will be to study the properties of quark-gluon matter at high baryon density and relatively moderate energy density.

Theoretical studies in this field of physics are aimed at substantiating physical programs of the given experimental projects. They are complex and related to resource-intensive extra-massively parallel calculations. These theoretical studies are based on computing of different observables via several methods, e.g.:

- quantum chromodynamics on a lattice (LQCD),
- quantum field continuum approaches to the study of critical phenomena based on the functional renormalization group (FRG) and equations of Dyson-Schwinger (DSE), Bethe-Saltpeter (BS) and Faddeev,
- modelling of the dynamics of heavy ion collisions (statistical, hydrodynamic and hybrid models),
- calculations in quantum field models of quantum chromodynamics vacuum.
- And others

An important feature of modern lattice computing in LQCD is the necessity to model dynamic quarks, the contribution of which is taken into account using the fermion determinant. The lattice calculation with the fermion determinant is the most resource-intensive operation in modelling. In modern computing this operation is performed on NVIDIA GPU graphics cards using the CUDA technology.



“Govorun” supercomputer

To support the development of the JINR BLTP project “Theory of hadron matter under extreme conditions”, especially for lattice calculations, numerical modelling of the collision dynamics and computations within modern functional approaches, the “Govorun” supercomputer has been created at JINR LIT; it has a module structure including nodes with graphics accelerators NVIDIA V100, multi-core processors Intel Skylake and Intel Xeon Phi, which allowed starting such computing at JINR.

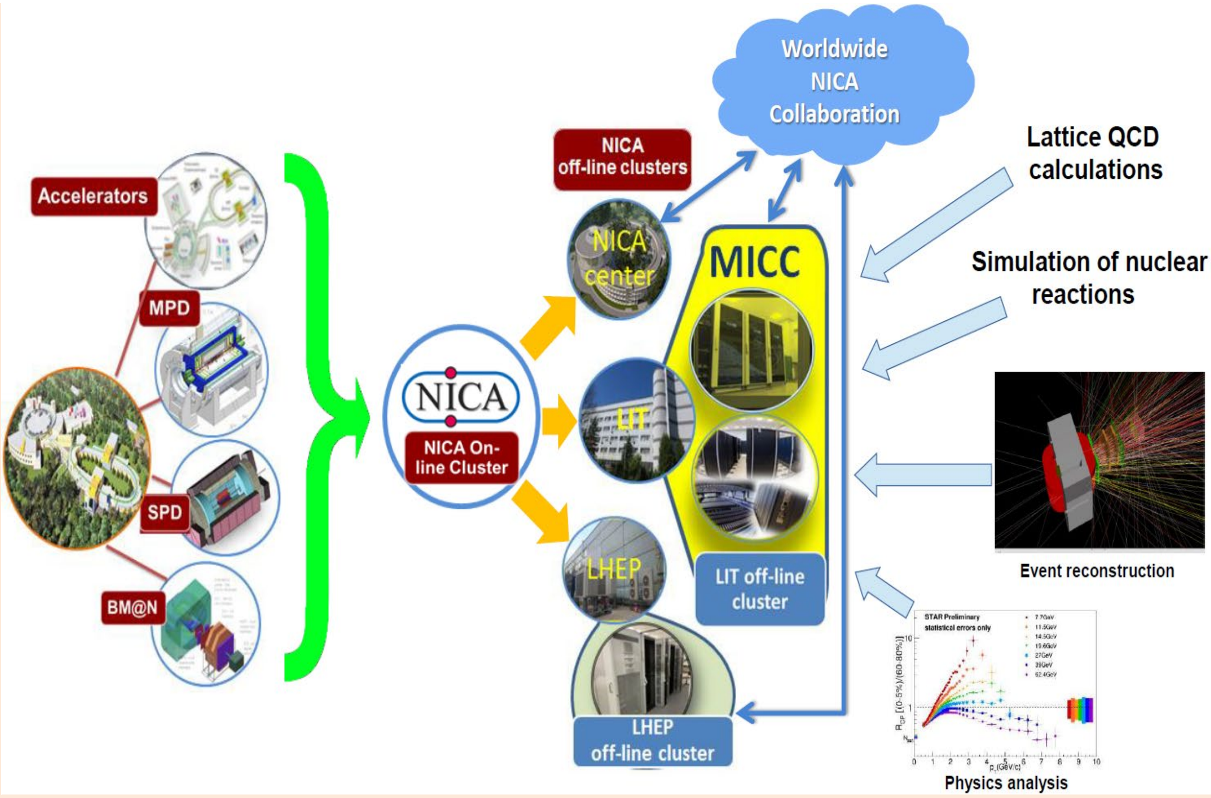
It is very important to ensure a high data exchange rate between computing units and a large amount of disk space for data storage, which is crucial for, e.g., lattice calculations. Thus, the scalable “Govorun” supercomputer included in the MICC meets all requirements for modern high-performance (HPC) systems by its technical characteristics, which allowed significantly speeding up theoretical studies and brought JINR computing capabilities in this field of research to a competitive level.

Planned future studies require more numerous or massive computing resources, i.e. multi-GPU architectures and resources with many thousands of graphics processors or novel multi-core CPU architectures.

At the same time, it is necessary to ensure proper **support for the software development**: new software will be needed to fully use novel hardware architectures; moreover, most of them will have more complicated memory hierarchy. In addition, significant progress can be achieved with the development of novel and optimized algorithms.

“Govorun” and NICA

Besides carrying out massively parallel calculations, it is planned to use the “Govorun” supercomputer for modelling an entire system of computing for all experiments of the NICA complex.



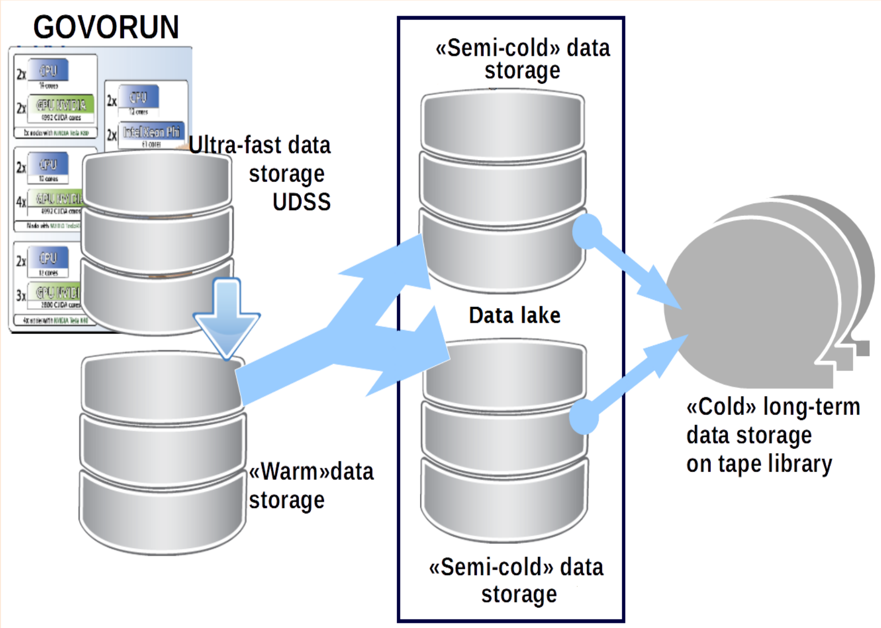
NICA computing concept and challenges

The implementation of various computing models of the NICA megaproject requires confirmation of the models’ performance, i.e. meeting the requirements for temporal characteristics of data acquisition from detectors with their subsequent transfer to processing, analysis and storage, as well as the requirements for the effectiveness of modelling and processing of events in the experiment. System tests in a real software and hardware environment must contain all the required components. At the moment, the “Govorun” supercomputer is such an environment. It contains the latest computing resources.

However, to perform a full cycle of testing, e.g., the MPD-computing framework, including the acquisition and processing of all data of the MPD detector, as well as of the simulated data from theories into the computing environment, the supercomputer should be supplemented with an ultra-fast data storage system (UDSS) that ensures a high data acquisition speed. Such a solution is proposed to be implemented on the basis of novel products Intel 3DNAND and Intel Optane. The proposed UDSS will allow saving and processing experimental data with a speed of hundreds of gigabytes per second for volumes of 0,5–5 PB, with the possibility of the linear extension of performance and capacity of the system up to 1000 times.

It is noteworthy that UDSS has the potential for linearly upgraded productivity (speed of working with data) and for increasing the storage volume without changing the principles of the architectural design of the system. At the same time, the hyper-convergence and software-defined architecture of UDSS allows a maximum of flexibility in data storage system configurations.

At present, the “Govorun” supercomputer is used for experiment simulations, i.e. the Skylake and KNL components for generating data of events from the MPD detector and saving them on UDSS with a subsequent transfer to semi-cold storages managed by the EOS and ZFS file systems. The DIRAC software is used to manage jobs and the process of Reading out and Writing data from different types of storages and cold long-term data storage on the tape library.



Scheme of using UDSS implemented in the "Govorun" supercomputer

This approach will allow one to check a basic stack of data storage and transmission technologies, as well as to simulate data flows, choose optimal distributed file systems and increase the efficiency of event modelling and processing.

Data storage systems

Novel computing and network technologies such as distributed cloud systems and the availability of broadband software-defined networks open up new possibilities for international cooperation; however, they require the further development of computing models. Recently, the trend of using and working with data storage systems has shifted from a simple approach separating organization, management, access and its hierarchy to the DOMA (Data Organization, Management and Access) model, a more flexible system combining all aspects.

The aspects of storage and access to data can be divided into the following main categories:

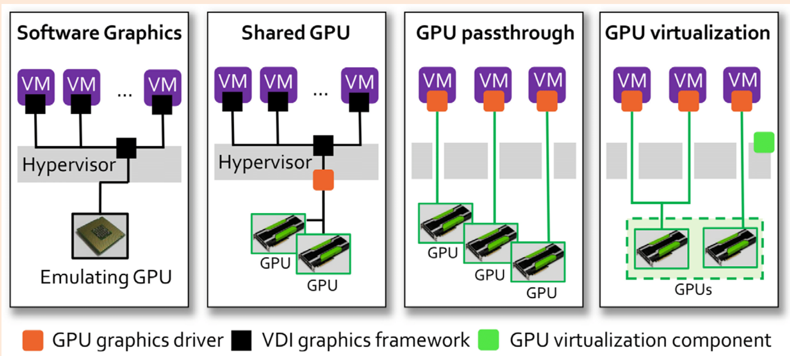
- Data storage volumes;
- Access speed of read-out and writing;
- Reliability of data storage;
- Security issues while accessing and defining the authority of access to data by different users and user groups;
- Organization of storage, update and access to unified software for users and user groups.

At present, for storing data of local experiments, users and collaborations, in which JINR has obligations, LIT uses about 10 petabytes on disks and 11 petabytes on a tape robot. In the next three years it is necessary to increase volumes of storing such data up to 80 petabytes on disks and 80 petabytes on tape robots. So far, requests for such volumes are foreseen from 3 major sources of data acquisition: the LHC experiments (Alice, ATLAS, CMS), the NICA experiments (B@MN, MPD, SPD) and neutrino experiments. In 5 years, due to the upgrade of the LHC to the HL-LHC, the launch of NICA at full luminosity and the development of cooperation with other collaborations worldwide, necessary data storage volumes will increase by an order of magnitude from the today's level.

Software at JINR

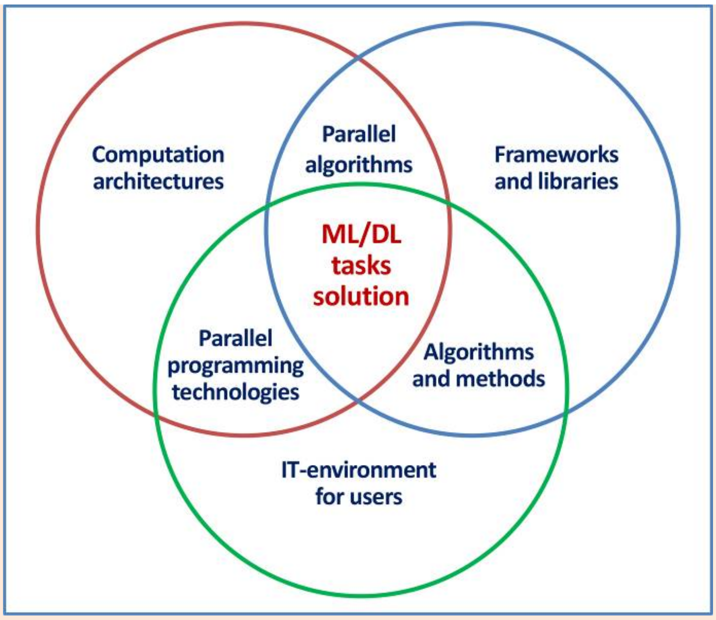
To effectively use modern computing equipment, parallel programming is necessary. Required software skills exceed those that can currently be assumed from an average physicist. This situation requires increased efforts in the field of personnel training on modern programming technologies, as well as the development of appropriate data processing systems by experts. In recent years, the successful experience in working with software packages within the framework of FairROOT developed by GSI has been gained. This new structure with an open source code will be the basis for developing software for specific experiments in the next decade. The given structure allows experiments to fully use the capabilities of modern computing systems and reduces the cost of development and maintenance. The transition from single-core to multi-core architectures with wide vector processing units, accelerator cards such as graphics processors, the availability of commercial high-speed networks, as well as multi-level memory and storage solutions, have changed the way of the implementation of algorithms. Skilled software experts need to be trained in this area, be it at the commercial hardware provider or be it in LIT.

At JINR, the software and information environment of the HybriLIT heterogeneous platform contains the HLIT-VDI service, which is designed to work with applied software using advanced graphics interfaces. This service allows one to work with such application packages as Wolfram Mathematica, Maple, Matlab, GEANT4, etc. through remote access to virtual machines, which ensures performing small computing inside VMs, as well as resource-intensive calculations on the computing nodes of the "Govorun" supercomputer.



Schematic representation of implementations of VMs with a remote access function, with support of the central processor and the graphics accelerator

The active implementation of the neural network approach, methods and algorithms of Machine Learning and Deep Learning (ML/DL) for solving a wide range of problems is defined by many factors. The development of computing architectures, especially while using DL methods for training convolutional neural networks, the development of libraries, in which various algorithms are implemented, and frameworks, which allow building different models of neural networks (see the Venn diagram shown below), can be referred to as key points. To provide all the possibilities both for developing mathematical models and algorithms and carrying out resource-intensive computations including graphics accelerators, which significantly reduce the calculation time, an ecosystem for tasks of ML/DL and data analysis has been created and is actively being developed for HybriLIT platform users.



Venn diagram on the implementation of the neural network approach, methods and algorithms of ML/DL to solve applied tasks

The created ecosystem has two components:

- the computing component is aimed at carrying out resource-intensive, massively parallel tasks of neural network training using NVIDIA graphics accelerators;
- The component for the development of models and algorithms on the JupyterHub basis, i.e. a multi-user platform for working with Jupyter Notebook (known as IPython with the possibility to work in a web browser), including several libraries and frameworks.

Infrastructure

The engineering infrastructure of computing centres must ensure the reliable, uninterrupted and fault-tolerant operation of the information and computing system and the network infrastructure.

Another integral feature of the engineering infrastructure is its scalability, the possibility of reacting to computing equipment growth prospects for the coming 3–5 years.

Modern and future computing complexes are characterized by high performance combined with low power consumption. An example is the “Govorun” supercomputer, which is a hyper-convergent calculator built on 100% liquid cooling in a “hot water” mode and has an energy efficiency of less than 1,06.

Big Data at JINR

Big Data are often defined as datasets that exceed traditional storage and processing systems in volume, speed and diversity.

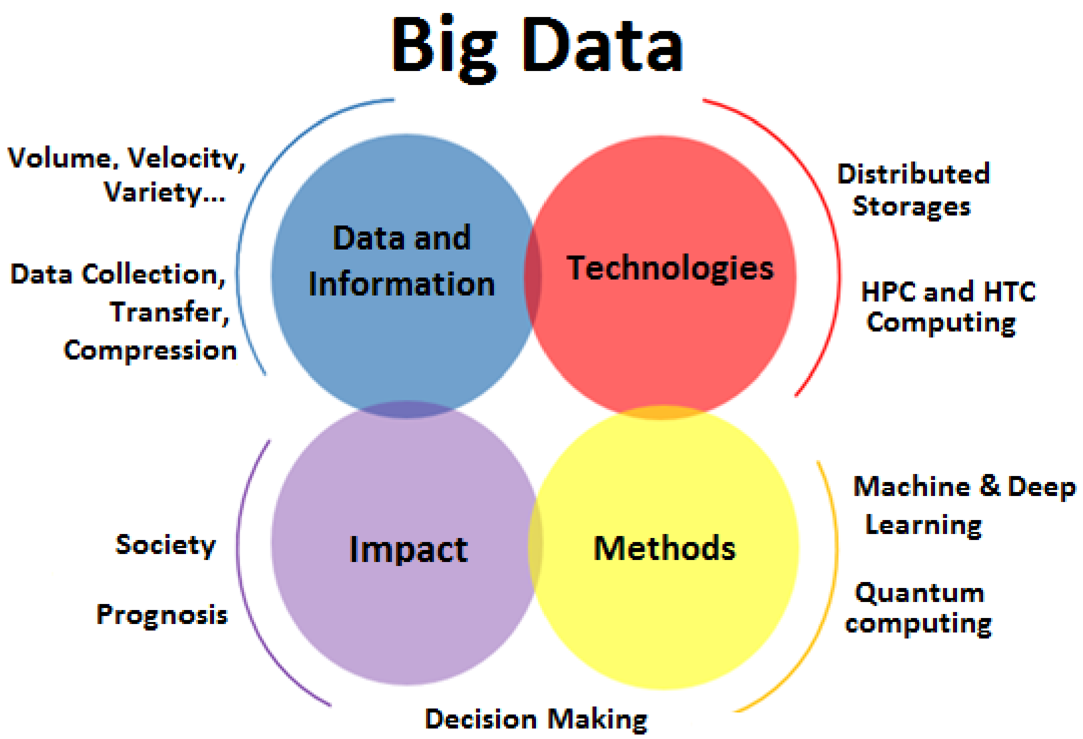
In the modern world, such amount of data is produced everywhere. The contribution of Science to the data growth increases all the time. Thus, in the course of the Large Hadron Collider (LHC) operation at CERN, particle collisions are detected by millions of sensors generating huge amounts of data. The total amount of information produced by four experimental facilities at the LHC is about 20 PB per year, the amount of accumulated data is currently about 200 PB, which is comparable and sometimes even exceeds the data accumulated in storages of traditional generators and Big Data providers such as, for example, Facebook, whose photo and video storage amount is at least 100 PB. It is noteworthy that the data flow after the LHC upgrade to *High Luminosity LHC* will grow by more than an order of magnitude.

Big Data are not only data themselves, but also processing technologies and usage scenarios, which can be divided into several main categories:

- new algorithms suitable for providing search in huge data sets and their processing;
- new data management technologies that allow working with complex, heterogeneous and distributed data sources;
- high-performance computing systems;
- architectures and algorithms that allow processing stream data coming from high-speed networks, instruments and sensors;
- high-performance and highly reliable distributed file systems capable of working with petabyte range data;
- a set of techniques to integrate heterogeneous data from different sources, i.e. data fusion and integration.

It should be underlined that methods, algorithms and technologies forming a group of notions "Artificial Intelligence" (AI) are an integral part of Big Data, an important component of the ecosystem. At present, "Artificial Intelligence" presented by methods, technologies and Machine Learning (ML) systems based on so-called Deep Learning (DL) neural networks is very rapidly and successfully developing, being at the peak of interest not only in IT itself, but also in economics, industry, other applied fields, and in the general public.

In more recent decades, science experiments such as at CERN have produced data on similar scales to current commercial "big data". However, due to differences in data structures and required precision, scientists have tended to analyse their data using specialized custom-built high-performance computing (supercomputing) clusters and grids, rather than clouds of cheap commodity computers as in the current commercial wave, implying a difference in both culture and technology stack. (https://en.wikipedia.org/wiki/Big_data)



An important trend in the development of Big Data is the transition from descriptive and diagnostic analytics (that is, mainly from establishing cause-and-effect relationships) through predictive analytics (what will happen in the future if certain actions are taken) to prescriptive analytics (choosing alternative actions based on an assessment of likely outcomes) and further to cognitive analytics. Cognitive analytics presupposes modelling and forecasting results using machine-learning methods that allow one to extract new knowledge from mixed data. The application of Big Data technologies together with "data lakes" and supercomputer technologies provides new opportunities for the development of large-scale mega science research projects (such as the MPD, SPD and BM@N experiments at the NICA accelerator complex and a group of experiments of the JINR Neutrino Program). Within the JINR information and computing infrastructure the creation of a software and analytical platform for solving a wide range of actual problems of physical analysis and modelling, as well as a number of operational issues of distributed computing systems, will allow solving certain types of tasks at various stages of conducting a physical experiment on a new level - from modelling of a facility and its various systems to the reconstruction of events, physical analysis, monitoring of the functioning of detectors, the prediction of possible nonstandard situations, etc.

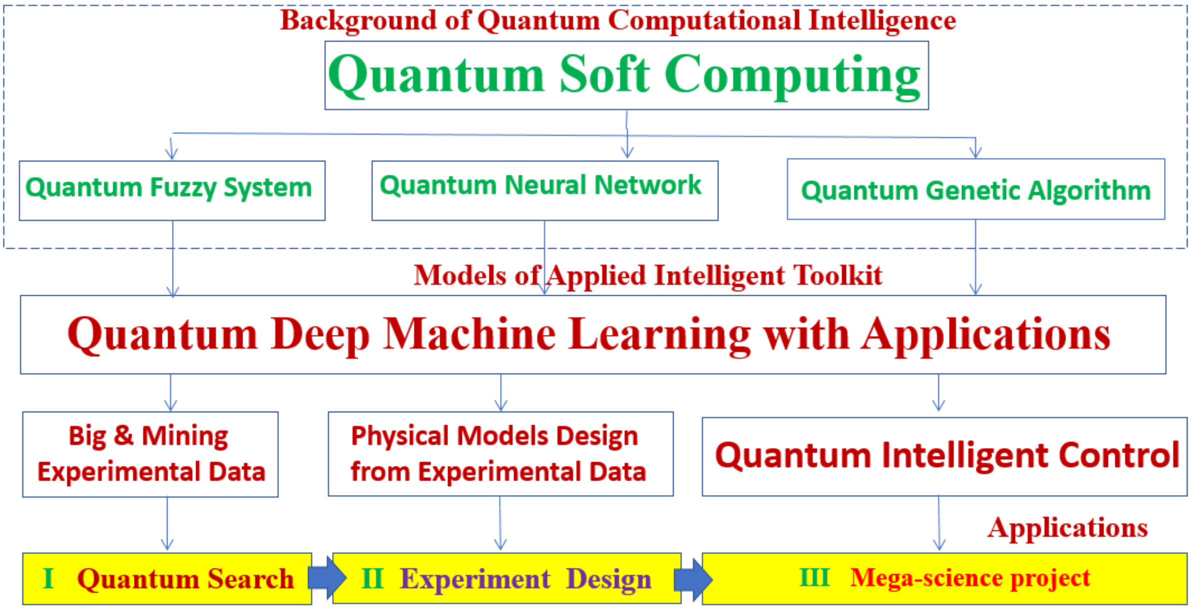
Quantum computing

Other promising directions of modern information technologies are quantum computing and quantum computer science [6].

Quantum applications are the following

- Quantum decision making and quantum search algorithms in intelligent control and unstructured large DB (Big Data information processing),
- Quantum intelligent control in mega-science projects,
- Intelligent cognitive robotics,
- Quantum soft computational intelligence toolkit.

Quantum computing, which has been rapidly developing in recent years, could offer new possibilities for processing big data from the LHC. Improving the reliability of the system, reducing the time of preparation and adjustment of the experimental mode is possible due to the involvement of end-to-end quantum information technologies of intelligent control and increasing the level of intelligence of the subsystems of the accelerator complex.



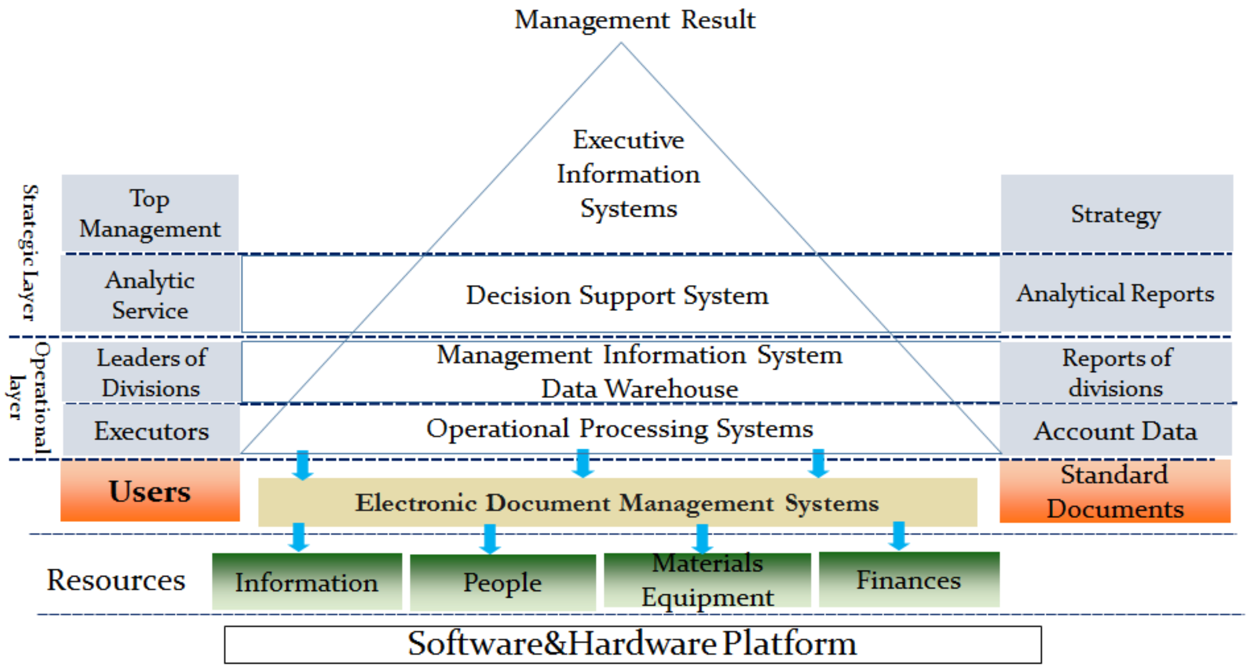
The growth rate of the number of qubits, which characterizes the maximum possible amount of input data for a quantum computer, is clearly higher than the linear one, which shows the accelerated progress in creating the equipment for quantum computers.

The JINR IT will actively observe this development in close collaboration with the CERN IT Division.

Information system at JINR

Information technologies can help to solve the task of complex automation of all activities of an organization. In the modern world, e.g., a **C**orporate **I**nformation **S**ystem (CIS) is an essential tool for organizing business in any sphere of human activities, as it allows automating business processes of financial, personnel and material management, providing a holistic picture of organization activities and helping to make management decisions. The objective of CIS is the effective management of all resources of the organization (material, technical, financial, technological, informational, human, intellectual) for the successful accomplishment of the mission (for scientific organizations it is achieving scientific results, meeting the information needs of the scientific community). Currently there are no standards nor a clear definition of the notion of a corporate information system (CIS).

Structurally, CIS is presented in the form of a two-level management pyramid: the lower layer is operational, the upper layer is strategic. At the entrance, CIS receives information from bottom-up about main resources (financial, material, personnel, informational), which must be managed, and the output is the result of main activities of the organization. As it moves up along the pyramid, primary information is structured, convolved and filtered. Thus, for senior management, reports reflecting the indicators needed for elaborating strategic decisions on management and development are generated and submitted top-down.



Structure of the corporate information system

Different information systems, covering such areas as financial management, accounting and taxation, personnel and salary accounting, scientific information systems and publication archives, project management, service centre, business analytics, materials and equipment management, operate at JINR:

- 1C information platform,
- JINR video portal,
- JINR Document Server, i.e. an Open Access (OA) repository of articles, preprints and other documents reflecting and contributing to research activities underway at JINR,
- platform for organizing and supporting scientific events,
- personal information system (PIN),
- monitoring and accounting systems,
- tool for planning activities for the NICA experiment,

- project management system,
- Information system of technical support Help Desk.

The CIS component, referred to as CRIS & OAR, which includes the Current Research Information System (CRIS) and the Open Access Repository (OAR), is highly important for a scientific organization.

The aggregation of institutional information systems, namely personnel, project, fund and other management systems, is necessary for the exchange and effective data reuse. This task cannot be solved only at the system level. A person is required to continuously manage all resources: personalities (personnel manager), projects (research project manager), finances (financial director), materials and equipment (facility manager), bibliographic information (bibliography specialist, librarian, content manager, identification manager). The key mechanism of the aggregation is the Authority Control subsystem, which provides communication and control of objects of authoritative records with resources of the corporate system.

Personnel training at JINR

To implement the strategy for the development of IT-technologies at JINR, it is necessary to systematically train and retrain specialists in modern information technologies, based on all the existing and future components of the MICC. At present, the education and testing polygon of the HybriLIT platform is used for training students, post-graduate students and young scientists:

- in parallel programming technologies to develop programs allowing one to perform calculations on hybrid computing architectures,
- in modern tools for development, debugging and profiling of parallel applications,
- tools and frameworks for solving problems of machine and deep learning,
- data analysis, as well as in the usage of specialized software packages for solving problems that require massively parallel calculations in physics, biophysics, etc.
- Special attention will be given to the directions of IT-specialists' training for solving JINR tasks related to data processing and analysis for experiments of mega science projects including the NICA project and the neutrino program.

The following training courses are planned:

- regular courses on modern IT-technologies both for the Institute staff and students and young scientists from the JINR Member States within the practices organized by the UC (JINR University Centre);
- special courses from leading software developers;
- special courses and seminars within conferences and schools organized by JINR;
- special courses organized in the JINR Member States within international cooperation programs.

One of the key points for keeping excellent staff at LIT and JINR is to offer conditions competitive with the industrial world. Since so many JINR scientists and projects are dependent on the support of LIT, not only a special salary structure, but also an atmosphere where creative individuals can develop their ideas with full responsibility, must be provided.

Objectives on medium-term prospects (3–4 years):

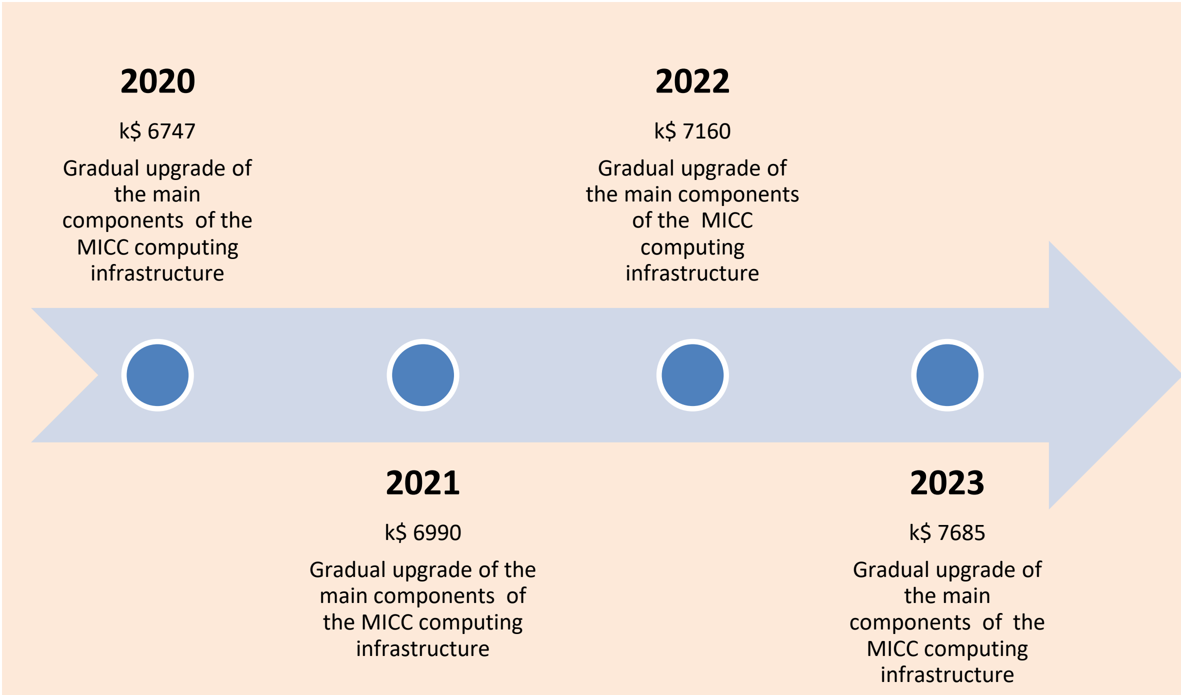
- Development and improvement of the JINR telecommunication and network infrastructure.
- Stage-by-stage modernization of the JINR MICC engineering infrastructure.
- Modernization and development of the IT-infrastructure of the NICA project.
- Extension of the performance and capacity of storage systems of the Tier1 data processing centre for the CMS experiment. Modernization and development of the resources being

part of the Tier-2/CICC integral component which provides support for the experiments using the grid environment and cooperating with physical groups in JINR, as well as for non-grid JINR users and its Member States (MPD/NICA, BESIII, LRB, FLNR, DLNP, BLTP, LNP). Extension of the cloud component in order to enlarge a range of services provided to users, as well as to create an integrated cloud environment for experiments of JINR (NICA, ALICE, BESIII, NOvA, Daya Bay, JUNO, etc.) and its Member States using the containerization technology.

- Enlargement of the HybriLIT heterogeneous platform with the "Govorun" supercomputer.
- Significant extension of the resources of the MICC components to meet the requirements of neutrino experiments.
- Development of a unified system for computing resource management aimed at big data processing and of a unified data management system for all MICC components (JINR data lake).
- Development of the concept and stage-by-stage implementation of a scalable software and analytical platform within the Big Data approach for acquiring, storing, processing and analysing, finding relevant information and visualizing results for the experiments at the NICA accelerator complex and a group of experiments within the JINR neutrino program.
- Development of methods and software for the effective application of Big Data analytics for resource-intensive calculations on coprocessors and graphics processors for modelling, reconstructing and processing experimental data (at the NICA accelerator complex, within the JINR neutrino program, the experiments at the LHC).
- Development of techniques for the reconstruction of events and intelligent monitoring of detectors using methods and technologies of Big Data and machine learning, novel mathematical methods such as DQC (Dynamic Quantum Clustering), Complex Networks, etc.
- Use of technologies of machine learning and artificial intelligence to optimize the functioning of distributed computing of physical experiments and intelligent monitoring of distributed computing systems.
- The educational program "Data Science" within the "International IT-School on Big Data" organized by JINR is highly important for the training of specialists for JINR, including for solving computing problems. It includes lectures and tutorials on methods and technologies for processing and analysing Big Data, distributed computing and machine learning in the framework of both university programs and international schools for students, young scientists and specialists.

Budget

Given the previous experience of the development of the MICC components, the financial support requested on the project implementation for medium-term prospects (4 years) comprises about 30 000 kUSD. The financial support for an additional expansion of data storage systems for the neutrino program and the NICA project will be provided within the budget of specific experiments.



Financial support requested on the project implementation for medium-term prospects

Long-Term Concept of a Scientific IT-ecosystem at JINR

Based on the above, a concept for the development of IT-technologies and scientific computing is being formulated as a scientific IT-ecosystem, combining a great number of various technological solutions, concepts and principles.

The creation of this scientific IT-ecosystem will enable to significantly reduce the time spent on the implementation of JINR experimental projects that require theoretical, applied calculations during the project stage, the work on constructing computing models for experiments, the solution of problems related to data processing and analysis, as well as the creation of an information environment for managing all project components. Thus, the overall work efficiency could be improved at JINR.

For developing this IT-ecosystem, information and communication support based on the latest technological network solutions is needed for providing a higher bandwidth and reliability of the network operation. The continually developed network infrastructure will respond to the increasing needs for effective and fast processing and storage of data, obtained both from the LHC and future experimental projects where JINR scientists are involved.

Another key basis of the scientific IT-ecosystem is a distributed software-configured HPC platform combining supercomputer (heterogeneous), grid and cloud technologies. This allows one to optimally solve various scientific and applied problems requiring both massively parallel computing and the use of Big Data methods and technologies. The proposed platform will enable to use or switch to novel computing architectures in the most effective way.

In addition, within the scientific IT-ecosystem, a task will be solving the problem of managing team robotic devices based on artificial intelligence methods. The development of this direction requires the involvement of the latest computing architectures, fundamental aspects of artificial intelligence for robotics, including learning how to learn, combining advanced pattern recognition and model-based reasoning. Some promising direction in robotic technologies is the creation of so-called “robot swarms”. Robot swarms allow simpler, less expensive, modular robotic units to be reconfigured into a team depending on the task at hand, while being as effective as a larger, task-specific, monolithic robot, which may be more expensive and have to be rebuilt depending on the task. The robot swarm can be used at macro-, micro-, and nanoscales with a plethora of application areas. The creation of centralized, distributed and decentralized manage systems of the robotic team or robot swarms will allow servicing experimental installations in places that are not accessible to humans, including installations of the mega science class such as NICA.

The development of the scientific IT-ecosystem depends on novel technologies for data acquisition and analysis, as well as for data sharing. This system must be open to new computing methods such as quantum, cognitive calculations, machine learning methods and data mining, as well as to any developments of new algorithmic bases.

The IT-ecosystem will be a basic platform for training IT-specialists, capable of developing algorithmic and software solutions in all fields needed at JINR.

In summary, the following topics are of strategic importance for the long-term JINR IT-ecosystem:

Telecommunication technologies:

- Networks with terabit data transmission rates;
- Networks based on new principles of the organization including cognitive, hybrid, adaptive, reconfigurable and heterogeneous (on-demand networks, networks of networks, software-configured networks, etc.);
- Network systems with a guaranteed dynamic resource allocation;
- Research systems of a new generation allowing to transfer a large amount of data;
- Networks allowing performing distributed processing of exabyte data volumes.

Computing systems:

- Distributed software-configured HPC platform combining supercomputer (heterogeneous), grid and cloud technologies in order to effectively use novel computing architectures.

Algorithms and software for:

- Distributed solution of individual classes of complex computing problems;
- Machine learning;
- Formalization and extraction of knowledge from poorly structured and unstructured data;
- Computing based on quantum formalism;

  software – algorithmic robot control platform based on artificial intelligence methods.

Technologies of data processing and analysis:

- Services for distributed and parallel computing (meta-computing) for increasing the research efficiency through the use of supercomputers;
- Development of scalable algorithms and software for processing multi-parameter, multi-dimensional, hierarchical and multi-series data sets of an exabyte volume;
- Analytics systems of a new generation based on effective methods and algorithms for formalizing and extracting knowledge and big data processing;
- Predictive modelling of computer models of mega-experiments and functioning of perspective systems.

Information security:

- Protection of computer infrastructure systems based on fundamentally new paradigms including quantum cryptography and computing, neurocognitive principles;
- Use of perspective tools and software systems for data protection taking into account new principles of data organization and interaction of data objects including the global integration of information systems, universal access to

applications, new Internet protocols, virtualization, social networks, mobile device data and geo-location.

Concise SWOT[1] analysis

Strengths

- The staff members of LIT cope with the implementation of current tasks, the modernization and upgrade of computing components and data storage systems.
- Multi-year successful experience in the WLCG project for processing data from the LHC experiments.
- The MICC components operate at the level of the best international standards in a 24x7 mode.
- The modern hyper-convergent supercomputer has been built on liquid cooling and modern processors.
- The network infrastructure is updated.
- The mechanism for monitoring the functioning of all MICC components has been established.
- Cooperation with MICC users.

Weaknesses

- Weak control over users' actions.
- Absence of a centralized user support system.
- The procurement system is slow and burocratic.
- Slow modernization of the elements of the engineering and network infrastructure.
- Volatility of prices on the equipment from leading manufacturers of computing and related equipment in the region.

Opportunities

- Understanding the need for investment and support of the developed IT-infrastructure by the JINR Directorate.
- Resources of students from Dubna University and other institutes as a potential source of personnel for servicing the MICC components.

Threats

- Rapid pace of obsolescence of the computer and network equipment.
- Virus and hacker attacks outside and inside, often due to users' carelessness.
- Depreciation and obsolescence of the engineering equipment, the modernization of which is delayed due to the excessive bureaucracy of the decision-making procedure.

References

1. P. Charpentier LHC Computing: past, present and future EPJ Web of Conferences 214, 09009 (2019) DOI: https://doi.org/10.1051/epjconf/201921409009

[1]SWOT – Strengths, Weaknesses, Opportunities, Threats

2. V. Korenkov, A. Dolbilov, V. Mitsyn, I. Kashunin, N. Kutovskiy, D. Podgainy, O. Streltsova, T. Strizh, V. Trofimov and P. Zrelov. The JINR distributed computing environment. EPJ Web of Conferences 214, 03009 (2019), https://doi.org/10.1051/epjconf/201921403009
3. A. Dolbilov, I. Kashunin, V. Korenkov, N. Kutovskiy, V. Mitsyn, D. Podgainy, O. Stretsova, T. Strizh, V. Trofimov, A.Vorontsov. Multifunctional Information and Computing Complex of JINR: Status and Perspectives. CEUR Workshop Proceedings (CEUR-WS.org), ISSN:1613-0073, v.2507, pp.16-22 (2019)
4. A.V. Baranov, N.A. Balashov, N.A. Kutovskiy, R.N. Semenov JINR cloud infrastructure evolution // Physics of Particles and Nuclei Letters, ISSN 1547-4771, eISSN: 1531-8567, 2016, vol. 13, No. 5, pp. 672–675. DOI: 10.1134/S1547477116050071.
5. Gh. Adam, M. Bashashin, D. Belyakov, M. Kirakosyan, M. Matveev, D. Podgainy, T. Sapozhnikova, O. Streltsova, Sh. Torosyan, M. Vala, L. Valova, A. Vorontsov, T. Zaikina, E. Zemlyanaya, M. Zuev. IT-ecosystem of the HybriLIT heterogeneous platform for high-performance computing and training of IT-specialists. Selected Papers of the 8th International Conference "Distributed Computing and Grid-technologies in Science and Education» (GRID 2018), Dubna, Russia, September 10-14, 2018, CEUR-WS.org/Vol-2267"
6. J. Amundson, E. Sexton-Kennedy Quantum Computing EPJ Web of Conferences 214, 09010 (2019) DOI: https://doi.org/10.1051/epjconf/201921409010