JINR Network Infrastructure for Megascience Projects

Andrey Baginyan
Laboratory of information technologies
Joint Institute for Nuclear Research
Dubna, Russia
bag@jinr.ru

Anton Balandin
Laboratory of information technologies
Joint Institute for Nuclear Research
Dubna, Russia
golter@jinr.ru

Sergey Belov*
Laboratory of cloud technologies and Big Data analysis
Plekhanov Russian University of Economics
Moscow, Russia
belov@jinr.ru

Andrey Dolbilov
Laboratory of information technologies
Joint Institute for Nuclear Research
Dubna, Russia
dolbilov@jinr.ru

Ivan Kadochnikov*
Laboratory of cloud technologies and Big Data Analysis
Plekhanov Russian University of Economics
Moscow, Russia
kadivas@jinr.ru

Vladimir Korenkov§‡
Laboratory of information technologies
Joint Institute for Nuclear Research
Dubna, Russia
korenkov@jinr.ru

Petr Zrelov§
Laboratory of information technologies
Joint Institute for Nuclear Research
Dubna, Russia
zrelov@jinr.ru

***Abstract*— Megascience projects like the Worldwide LHC Computing Grid (WLCG) usually require global, naturally distributed systems to operate the vast experimental data. For data processing, storage, and analysis, WLCG unites the assets of about 180 data centers in 50 countries, and the total storage volume exceeds 1 Exabyte. Joint Institute for Nuclear Research is profoundly engaged in the integration and development of distributed heterogeneous resources and the development of Big Data technologies to provide the infrastructure and the methods for contemporary scientific megaprojects in such data- and computing-intensive science domains as high energy physics, astrophysics, bioinformatics, and others. One of the network's critical tasks here is to provide high-speed and reliable access to the distributed resources worldwide and to allow using Big Data analytical platforms to the full. The paper provides a review of the JINR network infrastructure and its main parts - JINR Local Area Network (Backbone), Network of the Multifunctional Information and Computing Complex, and JINR Telecommunication Channels, as a crucial part of JINR Informational and Computer Infrastructure which serves needs of numerous large scale projects and experiments of megascience level.**

Keywords—megascience, network, traffic modelling, Big Data, grid technologies

I. INTRODUCTION

The Joint Institute for Nuclear Research (JINR) [1] is an international intergovernmental organization. It is developing as a big multidisciplinary international scientific center that combines fundamental research in modern high energy and nuclear physics, development and application of high technologies, and university education in the respective areas of expertise. Currently, there are eighteen Member States and six countries participating in JINR activities, basing on bilateral governmental agreements.

The research program of the JINR is focused on ambitious and large-scale experiments on the Institute's core facilities and in frames of the broad international cooperation. The program is also concerned with the implementation of the NICA (Nuclotron-based Ion Collider fAcility) megaproject [2], the building of new experimental facilities, the JINR's neutrino program, the upgrade of the Large Hadron Collider's (LHC) [3] detectors and other systems (CMS, ATLAS, Alice), research activities in nuclear physics and condensed matter physics. The recent years' experience shows that the progress in obtaining research results depends directly on computing resources' performance and efficiency. JINR has an information and computing complex that has evolved into stand-alone structures with a joint core engineering and networking facilities.

The JINR computing infrastructure combines a broad spectrum of computer systems, data networks, and IT technologies, providing the opportunity to solve various scientific and engineering tasks of a large-scale level [4].

JINR Network infrastructure consists of three main parts: 1) JINR Local Area Network (Backbone), 2) MICC (Multifunctional Information and Computing Complex) Network and external channels - 3) JINR Telecommunication Channels.

The bandwidth of the Moscow-JINR telecommunication channel is 3x100 Gbps; the Institute's local area network's backbone is 2x100 Gbps distributed computing cluster network between the JINR facilities has a capacity of 400

The study in part of the creation of a traffic analysis platform and methods was carried out at the expense of the Russian Science Foundation grant (project No.19-71-30008).

* Joint Institute for Nuclear Research, Dubna, Russia

§ Plekhanov Russian University of Economics, Moscow, Russia

‡ Dubna State University, Dubna, Russia

978-1-7281-6460-1/20/$31.00 ©2020 IEEE

Gbps with double redundancy to increase the reliability of the backbone network. At first, the JINR-CERN direct one and a backup one, which passes through MMTS-9 in Moscow and Amsterdam, and ensures the operation of LHCOPN (JINR-CERN) for the connection between Tier-0 (CERN) and Tier-1 (JINR) and the LHCONE external overlay network designed for the JINR Tier-2 center. The second group is direct channels with the collaboration of RUHEP research centers and RUNNet networks, RETN using the RUVRF technology.

II. JINR Local Area Network Structure

In 2019, it was continued the work on developing and improving the JINR IT infrastructure's network components. The Cisco ACI factory is based on the equipment Cisco Nexus 9504 and Cisco Nexus C9336'FX2, allowing one to connect the MICC components at speeds of up to 100 Gbps and more, was put into operation. To ensure the necessary bandwidth and the possibility of redundancy, the factory is connected to the JINR backbone network by 4 channels of 100 Gbps. The Tier-1 network was transferred to the ACI factory and had an overall connection of 160 Gbps (4 channels of 40 Gbps). The EOS distributed file system, the HybriLIT heterogeneous platform, the Govorun supercomputer, WEB services, Tier-2, and Cloud computing networks are connected to the ACI factory. The modernization of the network cluster of virtual services of the JINR network operations center (NOC) was in progress. The NOC virtualization cluster was created using the Proxmox open source software under the license with the open-source code GNU AGPL v3, enabling the free use of the code in creating cluster solutions. The cluster serves virtual machines of network services of the JINR network, such as DNS, DHCP, RELAYs, different network databases, as well as numerous services of LIT (Laboratory of Information Technology) and UC (University Center). The functionality of the system for the network traffic analysis was expanded with the help of new scripts, which can identify infected and hacked user computers. The support of the Wi-Fi eduroam network is provided at LIT, the Dubna hotel, the House of International Meetings, the House of Scientists, and the JINR University Center's hostel. The status of 560 hosts, more than 150 services, and conditions are being monitored in the network monitoring system. Several types of notifications, namely, e-mail messages and SMS alerts, are used. JINR LAN comprises 8169 network elements and 15505 IP-addresses, 7512 network users, 2465 users of the mail.jinr.ru service, 1531 users of electronic libraries, and 358 users of the remote access service.

III. MICC Network Structure

Multifunctional Information and Computing Complex (MICC) of JINR currently has the following basic components:

- the Central Information and Computing Complex (CICC) of JINR with computing and mass storage elements and Tier-2 for all the LHC experiments and other virtual organizations (VOs) in the grid environment,
- Tier-1 for CMS experiment [5],
- High Performance Computing (HPC) heterogeneous platform HybriLIT, including the “Govorun” supercomputer [6],
- the Cloud infrastructure [7].

Joint Institute for Nuclear Research also takes part CMS (Compact Muon Solenoid), in the multipurpose experiment at LHC in the European Organization for Nuclear Research (CERN).

One of the new data processing and storage centers was commissioned in JINR before the second launch (RUN 2) of the Large Hadron Collider. Following the technical collaboration task, the CMS network segment provides continuous interaction between 160 disk servers, 25 blade servers, 100 infrastructure servers, and a tape robot. The first module contains 80 disk storage servers (with 160 10G-ports in bonding mode), 15 blade servers (30 of 10G-ports in bonding mode), 60 servers infrastructure (with 40 10G-ports and 40 1G-ports in bonding mode). Finally, the data center network segment provides 230 10G-ports and 40 1G-ports. The same approach is applied for the second stage of the project to be commissioned in 2021.

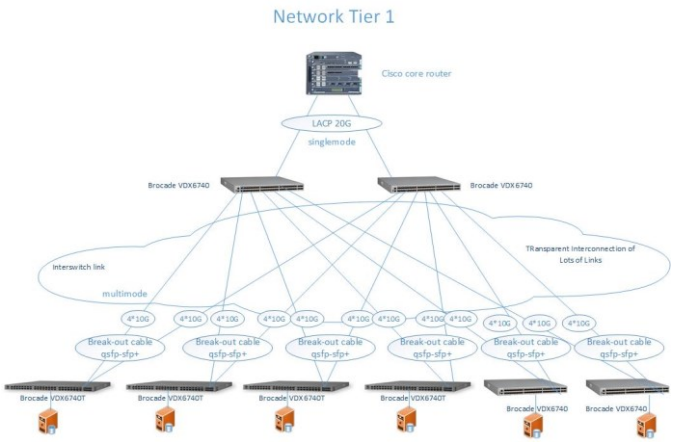

Fig. 1. A three-tier model of network segment of Tier-1 at JINR.

Figure 1 shows the scheme of the network architecture of the first module of Tier-1 center at JINR. Full redundancy of links is provided at all levels.

JINR Tier-2 center serves all experiments at the Large Hadron Collider (LHC) – CMS, ATLAS, ALICE, LHCb, and other virtual organizations (VOs) in the grid environment. The Network architecture of the JINR Tier-2 center is shown in figure 2.

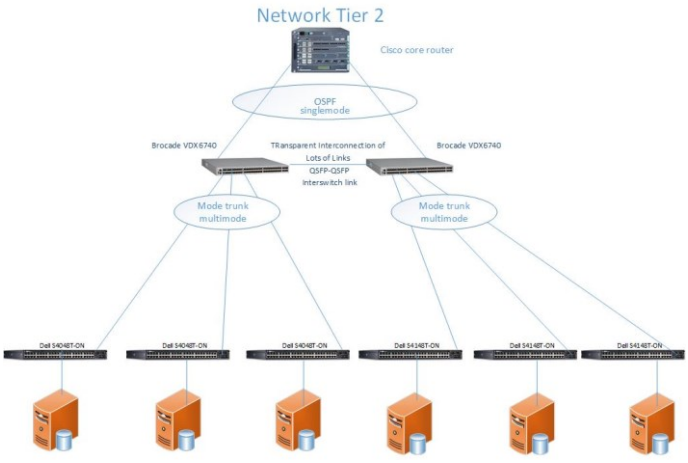

Fig. 2. Three levels of Tier-2 center’s network segment of at JINR.

It is planned to reconstruct the network architecture of Tier-1 data center with a two-way path between the access level and the server level on the Brocade equipment. Each server will have access to the network segment via two links 10G, each with total traffic capacity 20G.

To provide 160G data transfer capacity, the links between the access and the distribution levels will have four 40G paths.

IV. JINR Telecommunication Channels

In 2019, JINR carried out a significant modernization of the network infrastructure. It was implemented the project of increasing the bandwidth of the Moscow-JINR telecommunication channel from 100 to 3×100 Gbps. The bandwidth of the backbone of the Institute's local area network was increased to 2×100 Gbps. To improve the backbone network's reliability, it was built a distributed dual-redundant computing cluster network between the VBLHEP and the LIT sites with a capacity of 400 Gbps. At present, the external distributed network of JINR (Fig.5) includes several main channels. At first, the JINR-CERN direct one and a backup one, which passes through MMTS-9 in Moscow and Amsterdam, and ensures the operation of LHCOPN (JINR-CERN) for the connection between Tier-0 (CERN) and Tier-1 (JINR) and the LHCONE external overlay network designed for the JINR Tier-2 center. The second group is direct channels with the collaboration of RUHEP research centers and RUNNet networks, RETN using the RUVRF technology. An average rate of incoming and outgoing traffic over the JINR laboratories in 2019 (exceeding 25 TB by the incoming traffic) is shown in figures 3 and 4, respectively. In 2019, the overall incoming traffic of JINR, counting the general-purpose services, Tier-1 and Tier-2 clusters, and the computing complex, was about 56 PB. The traffic from scientific and educational networks is the principal one and is 96.4% of all traffic.

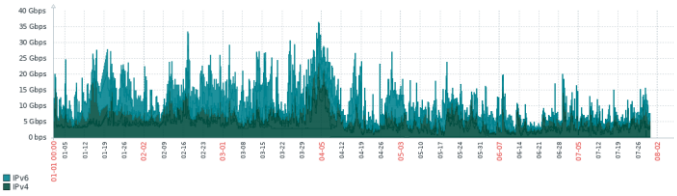

Fig. 3. Total incoming IPv4 / IPv6 traffic of JINR for seven months from the beginning of 2020

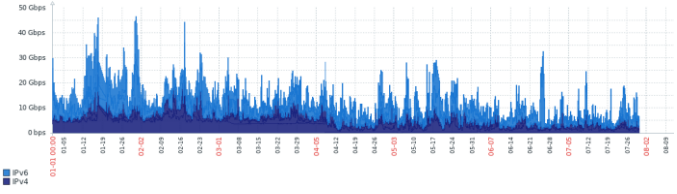

Fig. 4. Total outgoing IPv4 / IPv6 traffic of JINR for seven months from the beginning of 2020

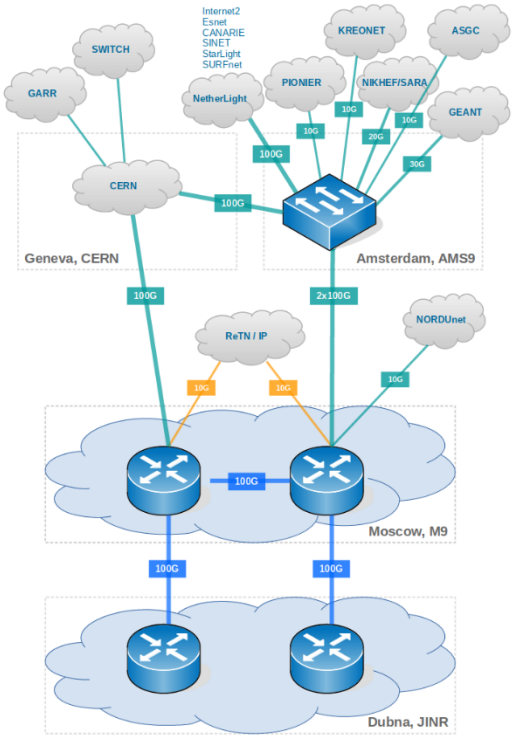

Fig. 5. The scheme of JINR external communication channels

V. Monitoring System

To guarantee the reliable performance of the MICC, a multi-level monitoring system was created and is expanded. It operates in a 24x365 mode and allows monitoring of the data center's life support systems as power supply, climate control, local networking hardware, telecommunication channels, computing nodes, disk and tape storage systems, and also computing jobs. The monitoring is based on several technologies, including Nagios, Icinga2, Grafana, NMIS, ZABBIX and products developed in LIT. For real-time monitoring of the whole JINR grid infrastructure, a dedicated operation center has been launched. It is a reliable place where operators can effectively control the MICC in a wide range of critical situations like a global power cut or a network failure.

At present, the monitoring system controls all types of MICC equipment and send system alerts to users via e-mail, SMS, etc. in a real-time mode. The number of nodes included in monitoring amounts to more than 2000. The monitoring provides a set of dashboards on the web site for the key system's state parameters.

To monitor the state of network links, the software package 10-Strike is used. The SNMP (Simple Network Management Protocol) software is employed for creating graphs [8]. The 10-Strike gathers statistical data for particular time periods and allows to present them graphically. Pre-defined templates are in use to deliver statistics on CPU usage, memory allocation, the number of running processes, and the volume of incoming and outgoing network traffic.

Figure 6 presents data on loading I/O interfaces between the links of the distribution and access levels.

Fig. 6. I/O interfaces load between the links of the distribution level and the access level.

The alert system is also configured to send e-mail and SMS notifications in case of network devices' failures or the peak download links.

VI. Big Data Stream Processing Platform

One of the essential subjects of analysis in different fields is streaming data, i.e. continuously incoming new information. It is necessary to carry out a historical analysis and make operational decisions. At present, using Big Data technologies seems to be the most effective for studying streaming data. A prototype of the analytical platform was built at the Laboratory of Information Technologies for these purposes – and first of all, for network traffic analysis and anomaly detection. The following services were deployed: 1) Apache Spark for analyzing incoming data in memory; 2) streaming data processing and storage system. Apache Kafka for uninterrupted data transfer and intermediate storage; 3) Mesos for managing the resources of the computing cluster; 4) Elasticsearch for primary profiling and analysis, and incoming data storage; 5) Docker server for deploying supporting services.

The project aimed to allow online transformation, filtering and flexible analysis of Ethernet packets with long-term storage and visualization of the results. The structure of the system in application to packet analysis is shown in Fig. 7.

Data stream analysis using Big Data tools is a promising approach to implementing flexible network traffic analysis and anomaly detection. Conventional network monitoring and intrusion detection tools, such as Snort, Squil and Suricata are functional and performant, but they focus on immediate practical application for security and failure detection. Long-term storage and in-depth analysis require large-scale processing and orchestration, which can be provided by so-called Big Data tools. At the same time, modern Big Data frameworks, such as Spark, natively support stream processing, which allows quick response comparable to network monitoring tools. Apache Metron is a Big Data principles framework; it is particularly designed for network security. Our stream processing system prototype is a general-purpose one and not intended to compete with such industrial solutions; packet analysis merely provides convenient sample problems to test and demonstrate the prototype's capabilities.

Apache Kafka queues and Apache Flume managed data transfer in streams between the system's components. For the initial data quality checks or processed data visualization, the Elasticsearch document database and Kibana can be used. For arbitrary online data processing and analysis, the system contains a Spark cluster with a Zeppelin notebook-style interface. After processing or filtering, data is to be stored in Ceph filesystem and then undergo bulk processing in Spark. To manage computing resources and run all the prototype components on five physical servers, we used Docker containerization and Apache Mesos

Raw traffic on a local network was ingested and parsed with tshark, which provides JSON metadata output suitable for processing with many standard Big Data frameworks. A sample is to be saved in Elasticsearch for initial profiling. Compressed packet metadata was stored in Ceph and bulk processed in Spark. As a promising approach used to cluster malicious traffic, packet data was transformed into AGgregate and Mode (AGM) format and visualized.

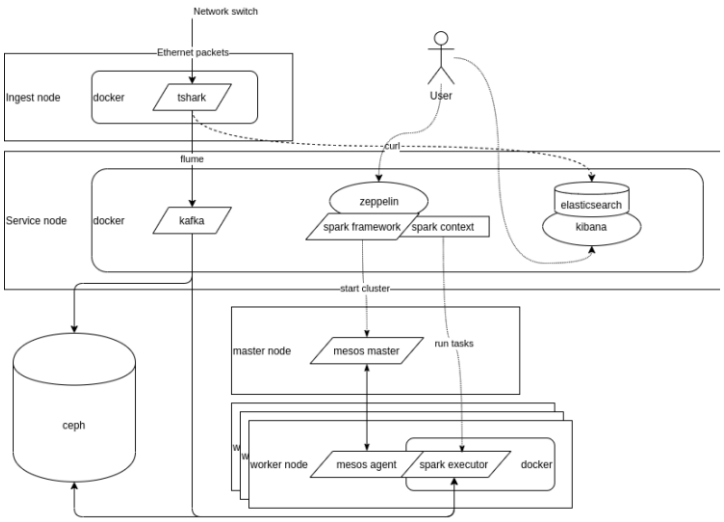

Fig. 7. The structure of the analytical platform in application to traffic analysis

In the future, there are plans to develop the project cluster of a multi-site network of network segments using EVPN MP-BGP technologies (Multi-Protocol Edge Gateway Protocol for Virtual Private Ethernet Networks) together with an external controller ACI (Application-Oriented Infrastructure) based on VXLAN technologies, using a multicast broadcast domain. This scheme will allow you to combine distributed sites. Analysis of technologies shows that this is the most acceptable choice.

A new Tier-1 module is being built based on Multi-site Network Cluster, and part of the work on integrating the Tier-2 factory has already been implemented.

VII. Modeling Network Data Traffic

Modeling network data traffic is one of the most essential tasks in designing and constructing new network centers and campus networks. The results of the analysis of models can be applied in the reorganization of existing centers and in the configuration of data routing protocols based on the use of links.

In frames of the project “Modeling network data traffic using the infrastructure of data centers at JINR” it will be shown how constant monitoring of the main directions of data transfer allows optimizing the payload of links by methods of increasing the priority of a different type of traffic. There are some basic elements for solving this problem, which are various ways of coloring data. Today it can be implemented with the help of variable length subnet masks, additional fields in the transmitted frame "quality of service" (QoS), and Deep Packet Inspection (DPI). One of the newest ways is mirroring the network at the application level (OSI level 7 model).

The project will present a plan for deploying a similar system in Tier-1 and Tier-2 center at JINR using the Ixia TrafficREWIND [9] technology as an example. An initial analysis of the traffic distribution in the data center will be made, graphs are shown, and conclusions are drawn on the implementation of the necessary measures to reduce the use of links.

The network infrastructure of the first stage of Tier-1 center in JINR, as notiecd, is based on Brocade’s company hardware. For multichannel data transfers, the modern protocol TRILL [10, 11] is in use.

At present, the authors are developing a testbed for a similar experiment using a traffic generator. It is likely that, to prove the gathered data, it will be necessary to build a similar network fabric that uses TRILL protocol on the vendor's hardware. This work is of great importance, as the collected data will be used while building the Data Processing Center for the NICA megaproject [2].

VIII. Conclusion

Network infrastructure is a basis for modern large scale projects as it provides high-speed access to experimental data to users in many countries. This article presents the main directions of development of the JINR network infrastructure for megascience projects. Special attention is paid to the development of traffic monitoring, forecasting and modelling tools for optimizing data flows and improving network security.

References

[1] Joint Institute for Nuclear Research. Web: http://www.jinr.ru/

[2] NICA (Nuclotron-based Ion Collider fAcility). Web: http://nica.jinr.ru/

[3] LHC (Large Hadron Collider). Web: https://home.cern/science/accelerators/large-hadron-collider

[4] V. Korenkov, The JINR Distributed Computing Environment, 2018 International Scientific and Technical Conference Modern Computer Network Technologies, MoNeTeC 2018 – Proceedings 10 December 2018, Article # 8572157.

[5] A.S. Baginyan, A.I. Balandin, A.G. Dolbilov, et al., Proc. of the 27th International Symposium on Nuclear Electronics & Computing (NEC’2019), http://ceur-ws.org/Vol-2507/321-325-paper-58.pdf, (2019).

[6] D.V. Belyakov, A.G. Dolbilov, A.N. Moshkin, I.S. Pelevanyuk, D.V.Podgainy, O.V. Rogachevsky, O.I. Streltsova, M.I. Zuev, Proc. of the 27th International Symposium on Nuclear Electronics & Computing (NEC’2019), http://ceur-ws.org/Vol-2507/316-320-paper-57.pdf, (2019).

[7] N.A. Balashov, A.V. Baranov, N.A.Kutovskiy, a, A.N. Makhalkin, Ye.M.Mazhitova, I.S. Pelevanyuk, R.N.Semenov, Proc. of the 27th International Symposium on Nuclear Electronics & Computing (NEC’2019), http://ceur-ws.org/Vol-2507/185-189-paper-32.pdf, (2019).

[8] J. Case, K. McCloghrie, M. Rose, S. Waldbusser RFC 1448 – Protocol Operations for version 2 of the Simple Network Management Protocol / April 1993.

[9] TrafficREWIND Regenerate Production Network Dynamics in the Lab. Web: https://www.ixiacom.com/products/trafficrewind

[10] Touch, J. and R. Perlman, "Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement," RFC 5556, May 2009.

[11] A.Baginyan, A.Dolbilov, I.Kashunin, V.Korenkov Equal-cost multi-pathing in high power systems with TRILL, EPJ Web Conf., Volume 214, 2019, 08013.