А. Г. Долбилов, П. В. Зрелов, В. В. Кореньков, Н. А. Кутовский, В. В. Мицын, Д. В. Подгайный, О. И. Стрельцова, Т. А. Стриж, В. В. Трофимов

Многофункциональный информационно-вычислительный комплекс в Лаборатории информационных технологий

Многофункциональный информационно-вычислительный комплекс (МИВК) удовлетворяет следующим требованиям, предъявляемым к современному высокопроизводительному научному вычислительному комплексу:

— многофункциональность,
— высокая производительность,
— развитая система хранения данных,
— высокая надежность и доступность,
— информационная безопасность,
— масштабируемость,
— развитая программная среда для различных групп пользователей,
— высокоскоростные телекоммуникации и современная локальная сетевая инфраструктура.

Внешний канал ОИЯИ построен по технологии DWDM (Dense Wave Division Multiplexing — спектрального мультиплексирования по длине волны) и использует две лямбды (две частоты) по 10 Гбит/с каждая. Для построения канала используется оптическое оборудование Nortel: терминал Nortel Optical Multiservice Edge (OME) 6500, усилитель и мультиплексор Nortel Common Photonic Layer (CPL).

К внешней распределенной сети ОИЯИ относятся: внешняя наложенная сеть LHCOPN (ОИЯИ–ЦЕРН), проходящая через МГТС-9 в Москве, Будапеште и

A. G. Dolbilov, P. V. Zrelov, V. V. Korenkov, N. A. Kutovsky, V. V. Mitsyn, D. V. Podgainy, O. I. Streltsova, T. A. Strizh, V. V. Trofimov

Multifunctional Information and Computing Complex of the Laboratory of Information Technologies

The JINR Multifunctional Information and Computing Complex (MICC) satisfies the following requirements for the modern high-performance scientific computing facility:

— versatility,
— high performance,
— developed data storage system,
— high-level reliability and availability,
— information security,
— scalability,
— developed software environment for different user groups,
— high-speed telecommunications and a modern network infrastructure.

External JINR channel is built on DWDM technology (Dense Wave Division Multiplexing — WDM wavelength) and uses two lambdas (two frequencies) of 10 Gbps each. For the construction of the channel, optical equipment of Nortel is used: Nortel Optical Multiservice Edge terminal (OME) 6500, an amplifier and a multiplexer Nortel Common Photonic Layer (CPL).

Амстердаме, для связи центров Tier-0 (ЦЕРН) и Tier-1 (ОИЯИ) и внешняя наложенная сеть LHCONE, проходящая таким же маршрутом, предназначенная для центра Tier-2 ОИЯИ; прямые каналы для связи по технологии RU-VRF с коллаборацией научных центров RUHEP (Гатчина, НИЦ «Курчатовский институт», Протвино), а также с сетями Runnet, RASnet.

В настоящее время основная оптическая транспортная магистраль локальной вычислительной сети ОИЯИ работает на скорости 10 Гбит/с.

Установлены системные сетевые сервисы: DNS, DHCP, SMTP, SNMP, регистрация пользователей, авторизация устройств, аутентификация пользователей, коммутация, маршрутизация, безопасность, связь в режиме видеоконференции, VoIP, IPDB (Internet Protocol Data Base), WebMail и др.

Центр Tier-1 CMS в ОИЯИ состоит из следующих главных систем.

1. Система обработки данных поддерживает 248 64-разрядных 12- и 20-ядерных рабочих узлов (WNs), что в общей сложности составляет 4160 ядер. Задания обслуживаются в пакетном режиме. Для поддержки системы пакетной обработки установлен специальный сервер с системой распределения ресурсов кластера и планировщиком.

2. Система хранения данных обслуживается с помощью программного обеспечениея dCache. Одна из установок dCache работает только с дисковыми серверами и используется для оперативного хранения данных с быстрым доступом к ним. Вторая установка dCache обслуживает специальные дисковые серверы и ленточного робота. Дисковые серверы являются буферной зоной для обмена с лентами, тогда как ленточный робот предназначен для длительного, практически вечного, хранения данных CMS. В общей сложности 2 установки имеют сейчас 6,4 Пбайт эффективного дискового пространства, а ленточный робот IBM TS3500 имеет 9 Пбайт для хранения данных. Для поддержки хранения и доступа к данным было установлено 8 физических и 14 виртуальных машин.

3. Система поддержки сервисов обеспечивает функционирование вычислительного сервиса, сервиса хранения данных, грид-сервисов, сервиса пересылки данных (FTS File Transfer System), системы управления распределенными вычислениями (Portable Batch System (PBS)), информационного сервиса (мониторинг сервисов, серверов хранения, передачи данных, информационные сайты). Грид-сервис VOBOX предназначен для переноса данных между грид-сайтами CMS посредством FTS; также сконфигурирован и ис-

---

External overlay network LHCOPN (JINR–CERN) goes through MGTS-9 in Moscow, Budapest, and Amsterdam network nodes to link Tier-0 (CERN) centre with Tier-1 centre (JINR), meanwhile, external overlay network LHCONE, passing the same route, is intended for system Tier-2 of JINR; direct channels by technology RU-VRF (virtual routing forwarding) with collaboration of RUHEP research centres: Gatchina, KI, Protvino; the networks Runnet, RASnet.

The optical backbone of the local JINR network (backbone) currently operates at a speed of 10 Gbps.

There is a system of network services: DNS, DHCP, SMTP, SNMP, user registration, authorization of devices, user authentication, switching, routing, security, videoconferencing communications, VoIP, IPDB (Internet Protocol Data Base), WebMail, etc.

Tier-1 CMS at JINR consists of the following main elements:

1. The data processing system supports 248 64-bit 12- and 20-core work nodes (WNs), which in total gives 4160 cores. Jobs are serviced in a batch mode. To support the batch system, a special server was installed with a system to allocate the resources of the cluster and a scheduler.

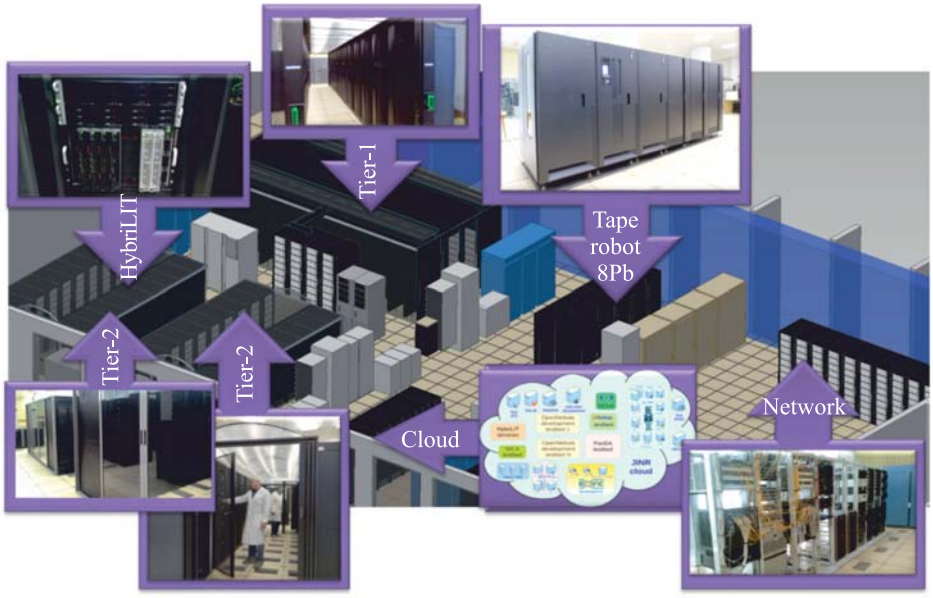
2. Storage systems have been installed in dCache software. One of the dCache installations is only used with disk servers for operational data storage with fast access to them. The second dCache unit includes disk servers and a tape robot. The disks serve as a buffer zone for exchange with tapes, while the tape robot is intended for a long-time, practically eternal, storage of data from the LHC. Totally, two installations have now 6.4 PB of effective disk space, and the tape robot has a data storage capacity of 9 PB. To support the storage and access to data, 8 physical and 14 virtual machines have been installed.

3. The system to support services ensures the operation of the computing service, storage service, grid services, data transfer service (File Transfer System, FTS), management system for distributed computing (Portable Batch System, PBS), information service (monitoring of services, storage servers, data transfer, information sites). Grid service VOBOX is designed for transferring data between the CMS grid sites by means of FTS; the proxy server SQUID required for work with specialized CMS databases (conditions DB) is also configured and used. The FTS service is used to reliably transfer files between large data stores, primarily, between the centres of Tier-0

пользуется прокси-сервер SQUID, который необходим при работе со специализированными базами данных CMS (conditions DB). Сервис FTS используется для надежной пересылки файлов между крупными хранилищами данных, в первую очередь между центрами уровня Tier-0 и Tier-1. Кроме того, сервис FTS обеспечивает контроль и мониторинг передач, распределение ресурсов сайта между различными организациями, управление запросами пользователей.

Интегральный компонент Tier-2/ЦИВК в настоящее время обеспечивает вычислительными мощностями и системами хранения и доступа к данным большинство пользователей и групп пользователей ОИЯИ и виртуальные организации (ВО). На кластере открыты очереди выполнения задач для четырех коллабораций LHC и нескольких ВО других физических экспериментов в России и за рубежом. Компонент Tier-2 включает следующие элементы.

1. Вычислительный кластер, общий для всех пользователей и коллабораций LHC. Часть кластера используется для традиционных параллельных вычислений, эта же часть вместе со всеми остальными вычислительными ресурсами кластера обычно используется для последовательных задач. В настоящее время кластер состоит из 279 физических машин, 20 из которых соединены высокоскоростной сетью Infiniband для обмена сообщениями внутри параллельных задач. Количество процессорных ядер равно 3640. Кластер обслуживает задачи в режиме пакетной обработки. Для поддержки системы пакетной обработки установлен специальный сервер с системой распределения ресурсов кластера и планировщиком задач.



Многофункциональный информационно-вычислительный комплекс ОИЯИ

The JINR Multifunctional Information and Computing Complex

and Tier-1 levels. Additionally, the service FTS provides control and monitoring of transmission, distribution of the site resources among different organizations, as well as user requests management.

The Central Information and Computing Complex (CICC) and the Tier-2 centre of JINR currently provide computing as well as storage and data access systems for most users and user groups of JINR and virtual organizations (VO). On the JINR CICC job queues were opened for four LHC collaborations and several VO collaborations of other physical experiments in Russia and abroad. The CICC/Tier-2 comprises:

1. The computing cluster shared by all users and collaborations. Part of the cluster is used for traditional parallel computing. This part, together with all other computing resources of the cluster, is typically used for sequential tasks. At present, the cluster consists of 279 physical machines, 20 of which are interconnected by a high-speed Infiniband network for message exchange inside the parallel tasks. The number of processor cores (slots to perform tasks) is equal to 3640. The cluster serves the tasks in a batch mode. To support the batch processing system, a special server with a system of the cluster resource allocation and a scheduler has been installed.

2. The storage systems installed in two versions of software: two dCache installations, two installations of XRootD. One of the dCache installations is used by VO CMS and ATLAS. The second dCache installation is used by users and user groups of JINR, including the NICA experiment (MPD). This facility also stores data of some third-party experiments (BIOMED, BES, FUSION). The first installation XRootD is used for ALICE. The second XRootD is used within the project FAIR by collaboration PANDA. In total, the systems of storage and data access utilize 29 disk servers with a total capacity of usable (accessible to users) disk space of 1909.8 TB. The storage

2. Системы хранения данных, установленные в двух вариантах программного исполнения: две установки dCache, две установки XRootD. Одна из установок dCache используется CMS и ATLAS. Вторая установка dCache используется пользователями и группами пользователей ОИЯИ, в том числе и для эксперимента NICA (MPD), также на этой установке хранятся данные нескольких сторонних экспериментов (BIOMED, BES, FUSION). Одна установка XRootD используется ALICE, вторая — в проекте FAIR коллаборацией PANDA. Объем систем хранения и доступа к данным составляет 1909,8 Тбайт. Системы хранения обслуживают 19 серверов, организующих распределение данных, авторизацию доступа к данным и протоколы работы с данными.

3. Серверы поддержки грид-окружения WLCG для различных ВО. Часть сервисов WLCG установлены на физических машинах, часть — на виртуальных. Сервисы WLCG оснащены программным обеспечением EMI-3 для совместимости с программной средой грид в WLCG. В настоящее время установлено 23 сервиса WLCG. Сервисы обеспечивают всю инфраструктуру удаленной работы с грид: авторизацию пользователей и ВО; запуск задач из удаленных сервисов ВО; информационную систему WLCG; различные алгоритмы удаленного тестирования и проверки среды обслуживания на локальных ресурсах. Имеются пять установок пользовательского интерфейса (UI) для запуска задач в распределенную грид-среду. Все перечисленные сервисы вместе с вычислительным кластером и системами хранения и доступа к данным образуют грид-сайт 2-го уровня (Tier-2) глобальной инфраструктуры WLCG, который доступен для коллабораций ALICE, ATLAS, CMS и LHCb.

Основными целями облачной инфраструктуры ОИЯИ являются: использование облачных вычислений для выполнения обязательств ОИЯИ в различных научно-исследовательских проектах, связанных с применением современных подходов в области построения и эксплуатации распределенных информационных и вычислительных систем; обучение сотрудников из организаций стран-участниц ОИЯИ в области облачных технологий; создание в этих организациях облачных инфраструктур с их последующей интеграцией в облако ОИЯИ и/или в глобальную распределенную информационно-вычислительную инфраструктуру.

Облачная инфраструктура ОИЯИ функционирует на основе программного обеспечения OpenNebula, а облачные ресурсы ОИЯИ используются в трех основных направлениях: образование, научные исследова-

---

systems serve 19 servers organizing data distribution, authorization of access to data and data operation protocols.

3. Servers supporting the WLCG grid environment for various VO. Part of the WLCG services were installed on physical machines, some on virtual ones. The WLCG services are installed with EMI-3 software for compatibility with the grid software environment in the WLCG. Twenty-three WLCG services have currently been installed. The services provide the entire infrastructure of remote work with grid: authorization of users and VO; job runs from remote services; WLCG information system; various algorithms for remote testing and verification of the service environment on local resources. There are five installations of the user interface (UI) to run jobs into a distributed grid environment. All of these services, together with the computing cluster and the systems of storage and data access, form a grid site of the 2nd level (Tier-2) of the global WLCG infrastructure which is available for the collaborations ALICE, ATLAS, CMS and LHCb.

The main aims of the cloud infrastructure of JINR are: the use of cloud computing for the implementation of JINR's obligations in various research projects related to the application of the present-day approaches in the field of construction and exploitation of the distributed information and computing systems; training of the employees from the JINR Member-State organizations in the field of cloud technologies; the creation in these organizations of the cloud infrastructures with their subsequent integration into the JINR cloud and/or global distributed information-computational infrastructure.

The cloud infrastructure of JINR operates on the basis of the OpenNebula software. At present the JINR cloud resources are used in three main areas: for education, research, and test tasks in various projects; to accommodate services with high availability and reliability; as computing resources and as an extension of the computational capabilities of grid infrastructures.

Below is a list of some services and polygons that are currently deployed in the JINR cloud:

— polygon PanDA (for PanDA product development and its use for solving tasks of the experiments ATLAS and COMPASS);

— polygon on the basis of the DIRAC middleware (is used for development of means for monitoring the distributed computing infrastructure of the experiment BES-III and one of its computing resources);

ния и тестовые задания в различных проектах; размещение сервисов с высоким уровнем доступности и надежности; в качестве вычислительных ресурсов и как расширение вычислительных мощностей в грид-инфраструктурах.

Ниже приводится перечень некоторых сервисов и полигонов, которые в настоящее время развернуты в облаке ОИЯИ:

— полигон PanDA для развития программного обеспечения PanDA и его использования для решения задач экспериментов ATLAS и COMPASS;

— полигон на основе промежуточного программного обеспечения для разработки средств мониторинга распределенной вычислительной инфраструктуры эксперимента BES-III, который может использоваться как один из ее вычислительных ресурсов;

— набор контейнеров для участников эксперимента NOνA для выполнения задач моделирования и анализа;

— полигон для разработки промежуточного программного обеспечения, создаваемого для построения вычислительной инфраструктуры проекта NICA;

— GitLab — локальный сервис для пользователей ОИЯИ;

— полигон на базе программного обеспечения Hadoop;

— набор виртуальных машин VM/CT для собственных нужд пользователей;

— контейнеры для оценки существующих систем мониторинга и разработки на их основе системы мониторинга сервисов для Tier-1;

— испытательный стенд EOS для исследования гетерогенной киберинфраструктуры и разработки прототипа федеративной вычислительной инфраструктуры, создаваемой на основе объединения высокопроизводительных вычислений, облачных вычислений и суперкомпьютерных вычислений для хранения, обработки и анализа больших данных (Big Data);

— автономная реализация программной платформы Spark для машинного обучения и анализа больших данных (Big Data).

В настоящее время с облаком ОИЯИ интегрированы облака Института физики НАН Азербайджана (Баку); Института теоретической физики им. Н. Н. Боголюбова НАН Украины (Киев); РЭУ им. Г. В. Плеханова (Москва).

Вычислительный компонент гетерогенного вычислительного кластера HybriLIT состоит из 10 узлов с графическими ускорителями NVIDIA Tesla K20X,

---

— a set of containers for users participating in the experiment NO*ν*A (simulation and analysis);

— polygon for the study and evaluation of the middleware to build a computing infrastructure of NICA;

— GitLab — local GitLab service installation for all JINR users;

— polygon on the basis of Hadoop software;

— set VM/CT of users for their own needs;

— containers for the assessment of existing monitoring systems and development on their basis of a monitoring system of the JINR Tier-1;

— EOS testbed for research on heterogeneous cyber-infrastructures, computing federation prototype creation and development based on high-performance computing, cloud computing and supercomputing for Big Data storage, processing and analysis;

— a standalone Spark instance for Machine Learning and Big Data analysis.

At present, the clouds of the following partner organizations from JINR Member States are integrated with the JINR cloud: the Institute of Physics of Azerbaijan NAS (Baku, Azerbaijan); the Bogolyubov Institute for Theoretical Physics of the NAS of Ukraine (Kiev, Ukraine); the Plekhanov Russian University of Economics (Moscow, Russia).

At the moment, the HybriLIT heterogeneous cluster, a computing component of JINR MICC, contains 10 nodes with NVIDIA Tesla K80 graphical processors, NVIDIA Tesla K40 accelerators, Intel Xeon Phi7120P coprocessors and two types of computing accelerators NVIDIA Tesla K20x and Intel Xeon Phi 5110P. All the nodes have two multi-core processors Intel Xeon. Overall, the cluster contains 252 CPU cores, 77184 GPU cores, and 182 PHI-cores; it has 2.5 TB RAM and 57.6 TB HDD, and its total capacity is 140 Tflops for operations with single precision and 50 TFlops for double precision.

A virtual desktop system has been installed to support a user work with applied packages — 48 virtual machines with Linux Centos 7.4 operation system with a personal access to the desktop to provide work with such graphical applications as Maple, MatLAB, Mathematica, etc.

The cluster resources are used for calculations in the field of quantum chromodynamics, quantum mechanics and molecular dynamics, and software PandaRoot, MpdRoot installed on the cluster allows one to perform calculations in high energy physics.

K40, K80, сопроцессорами Intel Xeon Phi 5110P, 7120 и процессорами Intel Xeon E5-2695 V2 и V3. Общее количество ядер CUDA — 77184, процессорных ядер — 252, ядер сопроцессора — 182, общий объем памяти — 2,5 Тбайт, общая производительность при вычислениях с одинарной точностью — 140 Тфлопс, с двойной — 50 Тфлопс.

Для пользователей организованы виртуальные рабочие столы — 48 виртуальных машин с ОС Linux Centos 7.4 с персональным доступом к рабочему столу для работы с графическими приложениями, такими как Maple, MatLAB, Mathematica и т. п.

Ресурсы кластера используются для расчетов в области квантовой хромодинамики, квантовой механики и молекулярной динамики, физики высоких энергий.

В 2018 г. с целью многократного увеличения вычислительной мощности предусмотрено развитие гетерогенного кластера HybriLIT, необходимого для кардинального ускорения комплексных теоретических исследований, проводимых в ОИЯИ.

Существенное наращивание вычислительных ресурсов гетерогенного кластера HybriLIT заключается в увеличении производительности как CPU-, так и GPU-компонента кластера. Модернизированный вычислительный кластер позволит проводить ресурсоемкие, массивно-параллельные расчеты в решеточной КХД для исследования свойств адронной материи при высокой барионной плотности, высокой температуре и в присутствии сильных электромагнитных полей, качественно повысит оперативность моделирования динамики столкновений релятивистских тяжелых ионов, а также позволит разрабатывать и адаптировать программное обеспечение для проекта NICA к новым вычислительным архитектурам от основных лидеров рынка HPC — Intel и NVIDIA, создавать программно-аппаратную среду на базе HPC и готовить IT-специалистов по всем необходимым направлениям. Такая модернизация кластера HybriLIT приблизит вычислительные возможности ОИЯИ в этой области исследований к мировому уровню.

Расширение CPU-компонента будет осуществлено на базе уже создаваемой в ЛИТ специализированной для HPC инженерной инфраструктуры, которая базируется на технологии контактного жидкостного охлаждения, реализуемой российской компанией ЗАО «РСК Технологии». Расширение GPU-компонента кластера планируется за счет приобретения вычислительных серверов последнего поколения с графическими ускорителями NVIDIA Volta.

In 2018, the development of the heterogeneous cluster HybriLIT will be directed at a multiple increase in the computing power required for dramatic acceleration of the complex theoretical research at JINR.

The significant increase in the computing resources of heterogeneous cluster HybriLIT is due to growing performance of both CPU and GPU components of the cluster. The upgraded computing cluster will allow massively parallel calculations in lattice QCD for research on the properties of hadron matter at a high baryon density, high temperature and in the presence of strong electromagnetic fields. It will positively increase the efficiency of simulations of the dynamics of relativistic heavy ion collisions as well as allow one to develop and adapt software for the mega-project NICA on new computing architectures of the major HPC market leaders Intel and NVIDIA, to create a hard- and software environment on the HPC basis and to prepare IT-specialists in all areas needed. Such a modernization of the HybriLIT cluster will bring the computational capabilities of JINR in this research field to the global level.

Extension of the CPU component will be carried out on the basis of an engineering infrastructure specialized for HPC being created at LIT which is based on the technology of contact liquid cooling, implemented by the Russian JSC “RSC Technologies”. The extension of the GPU component of the cluster is planned to be done by purchasing computing servers of the latest generation with graphics accelerators NVIDIA Volta.