Д. В. Подгайный, О. И. Стрельцова, О. А. Горбачев

Производительность суперкомпьютера «Говорун» в ОИЯИ достигла 1,1 Пфлопс

14 ноября в ОИЯИ состоялась презентация модернизированного суперкомпьютера «Говорун», производительность которого выросла на 23,5% и достигла уровня 1,1 Пфлопс. Презентация проходила в рамках ИТ-школы, для участия в которой в Дубну приехали более 60 студентов из 13 университетов России.

Основой для создания суперкомпьютера «Говорун» послужил опыт, накопленный в результате эксплуатации гетерогенного кластера HybriLIT, входящего в состав Многофункционального информационно-вычислительного комплекса ЛИТ ОИЯИ. HybriLIT показал свою востребованность при решении задач КХД на решетках, радиационной биологии, в прикладных исследованиях и др. Постоянный рост числа пользователей и расширение круга решаемых задач потребовали не просто существенно нарастить вычислительные возможности кластера, а разработать и внедрить новые технологии, что привело к созданию в 2018 г. новой вычислительной системы — суперкомпьютера «Говорун». СК «Говорун» создавался как высокопроизводительная масштабируемая система с жидкостным охлаждением, обладающая гиперконвергентной и программно-определяемой архитектурой. В текущую конфигурацию СК «Говорун» входят вычислительные модули, содержащие компоненты GPU и CPU, а также иерархическая система обработки и хранения данных.

Для CPU-компонента суперкомпьютера была выбрана технология прямого жидкостного охлаждения компании ЗАО «РСК Технологии», которая является ведущим в России разработчиком и интегратором «полного цикла» суперкомпьютерных решений и обладает целым рядом собственных инновационных разработок. Благодаря внедрению этих технологий для СК «Говорун» удалось достичь рекордной плотности размещения вычислительных узлов на шкаф (153 узла

D. V. Podgainy, O. I. Streltsova, O. A. Gorbachev

The Performance of the “Govorun” Supercomputer at JINR Reached 1.1 PFlops

Within the IT School, which was held on 14–19 November at the Joint Institute for Nuclear Research in Dubna, bringing together over 60 students of 13 universities from Russia, a presentation of the modernized “Govorun” supercomputer took place. The performance of this high-performance system enhanced by 23.5% and reached 1.1 PFlops.

The “Govorun” supercomputer was created on top of the experience gained during the operation of the HybriLIT heterogeneous cluster, which is part of the JINR MLIT Multifunctional Information and Computing Complex. HybriLIT has shown its relevance in solving tasks of QCD on lattices, radiation biology, applied research, etc. The continuous growth in the number of users and the expansion of the range of tasks to be solved entailed not only a significant increase in the computing capabilities of the cluster, but the development and implementation of novel technologies, which resulted in the creation in 2018 of a new computing system, the “Govorun” supercomputer. The “Govorun” supercomputer was created as a high-performance, scalable liquid-cooled system with a hyperconverged and software-defined architecture. The current configuration of the “Govorun” supercomputer involves computing modules containing GPU and CPU components, as well as a hierarchical data processing and storage system.

The technology of direct liquid cooling of CJSC “RSC Technologies”, which is the leading Russian developer and integrator of the “full cycle” of supercomputer solutions and has a number of its own innovative developments, was chosen for the CPU component of the su-

Лаборатория информационных технологий им. М. Г. Мещерякова. Сотрудник ЗАО «РСК Технологии» Ю. Н. Мигаль рассказывает о текущем этапе модернизации суперкомпьютера «Говорун»

The Meshcheryakov Laboratory of Information Technologies. Yu. N. Migal (CJSC "RSC Technologies") speaks about the current stage of the modernization of the "Govorun" supercomputer

percomputer. Thanks to the introduction of these technologies, the "Govorun" supercomputer managed to achieve a record density of placement of compute nodes per rack (153 nodes vs 25 nodes for air cooling), and the operation in the "hot water" cooling mode made it possible to use the year-round free cooling mode ($24 \times 7 \times 365$). In addition to high-energy efficiency, this approach enabled scientists to significantly simplify the infrastructure of the supercomputer centre, i.e., the cooling system of the "Govorun" supercomputer was created using only dry cooling towers that cool the liquid using ambient air. Due to liquid cooling, the average annual PUE indicator of the system, reflecting the level of energy efficiency, is less than 1.06. That is, less than 6% of the total electricity consumed is spent on cooling, which is an outstanding result for the high-performance computing (HPC) industry. The given system is the first system in the world with 100% liquid cooling; all components, namely, compute nodes, network switches and the data storage system, are cooled.

The current stage of the modernization, related to the expansion of the CPU component, was implemented within a hyperconverged approach to building a computing complex, which underlies the "Govorun" supercomputer. Hyperconvergence allows orchestrating computing resources and data storage elements, as well as creating computing systems on demand, taking into account the needs of user applications, with the help of the RSC BasIS software. The notion "orchestration" means the logical disintegration of a compute node into separate components, such as compute cores, data storage elements (SSDs), with their subsequent integration into the configuration. Thus, computing elements (CPU cores and graphics accelerators) and data storage elements (SSDs) form independent sets of resources (pools). Due to orchestration, the user can allocate for his task the required number and type of compute nodes (including the required number of graphics accelerators), the required volume and type of data storage systems, as well as automatically configure the required software, including parallel file systems. After the task is completed, the compute nodes and storage elements are returned to their corresponding pools and are ready for the next use. This feature enables to effectively solve user tasks of different types, to enhance the level of confidentiality of working with data and avoid system errors that occur when crossing the resources for different user tasks. The storage-on-demand system implemented on the hyperconverged nodes of the first modification under the management of the Lustre file system allowed the "Govorun" supercomputer to take the 9th place in the IO500 world rating (June 2018) for HPC storage systems.

против 25 узлов для воздушного охлаждения), а работа в режиме охлаждения «горячей водой» позволила использовать круглогодичный режим free cooling (24 × 7 × 365). Помимо высокой энергоэффективности, такой подход позволил существенно упростить инфраструктуру суперкомпьютерного центра — система охлаждения СК «Говорун» создана с применением только сухих градирен, охлаждающих жидкость при помощи окружающего воздуха. За счет применения жидкостного охлаждения среднегодовой показатель PUE-системы, отражающий уровень эффективности использования электроэнергии, составляет менее чем 1,06. То есть на охлаждение расходуется менее 6% всего потребляемого СК «Говорун» электричества, что является выдающимся результатом для HPC-индустрии. Построенная система является первой в мире системой со 100%-м жидкостным охлаждением, т. е. жидкостным образом охлаждаются все компоненты: вычислительные узлы, сетевые коммутаторы и система хранения данных.

Текущий этап модернизации, связанный с расширением CPU-компонента, осуществлен в рамках гиперконвергентного подхода к построению вычислительного комплекса, положенного в основу СК «Говорун». Гиперконвергентность позволяет «оркестрировать» вычислительными ресурсами и элементами хранения данных и создавать, используя программное обеспечение РСК БазИС, вычислительные системы, конфигурации которых зависят от потребностей пользовательских приложений. Под термином оркестрация подразумевается логическая дезинтеграция вычислительного узла на отдельные компоненты, такие как вычислительные ядра, элементы хранения данных (SSD-накопители), с последующим их объединением в конфигурацию. Таким образом, вычислительные элементы (CPU-ядра и графические ускорители) и элементы хранения данных (SSD-диски) образуют независимые наборы ресурсов (пулы). Благодаря «оркестрации» пользователь может под свою задачу аллоцировать необходимое число и тип вычислительных узлов (и число графических ускорителей), необходимый объем и тип систем хранения данных, а также автоматически настроить необходимое ПО, в том числе параллельные файловые системы. После завершения задачи вычислительные ядра и элементы хранения возвращаются в соответствующие пулы и готовы к следующему использованию. Это свойство позволяет эффективно решать пользовательские задачи разных типов, повысить уровень конфиденциальности работы с данными, избежать системных ошибок, возникающих при пере-

---

As a result of the modernization, the CPU component was extended to 32 hyperconverged compute nodes based on two Intel Xeon Platinum 8368Q processors (frequency 2.6 GHz, 38 cores, cache 57 MB, TDP 270 W) each, DDR4 RAM modules (256 GB per node), 8 Intel Optane DC Persistent Memory modules (2 TB per node), 4 EDSFF E1.S NVMe SSDs (16 TB per node), and an M.2 NVMe SSD with a capacity of 128 GB. In addition, each node is equipped with two 100-Gb/s Intel Omni-Path adapters.

The uniqueness of the upgrade makes it possible to create a layer of “very hot” data with a capacity of 8 PB in the hierarchical data processing and storage system of the “Govorun” supercomputer. The creation of such a hierarchical system was defined by its relevance for working with Big Data, primarily for the NICA megaproject. According to the speed of accessing data, the system is divided into layers, namely, “very hot data”, the most demanded data to which it is currently required to provide the fastest access, “hot data”, and “warm data”. Each layer of the developed data storage system can be used both independently and as part of data processing workflows. At the moment, as a layer of “very hot data”, the latest DAOS (Distributed Asynchronous Object Storage) technology for Big Data processing, which has shown its promise for deep learning tasks and for the operation of quantum simulators to emulate a larger number of qubits, is being implemented on the “Govorun” supercomputer. For the high-speed data processing and storage system, the “Govorun” supercomputer received the prestigious Russian DC Awards 2020 in nomination “The Best IT Solution for Data Centers” .

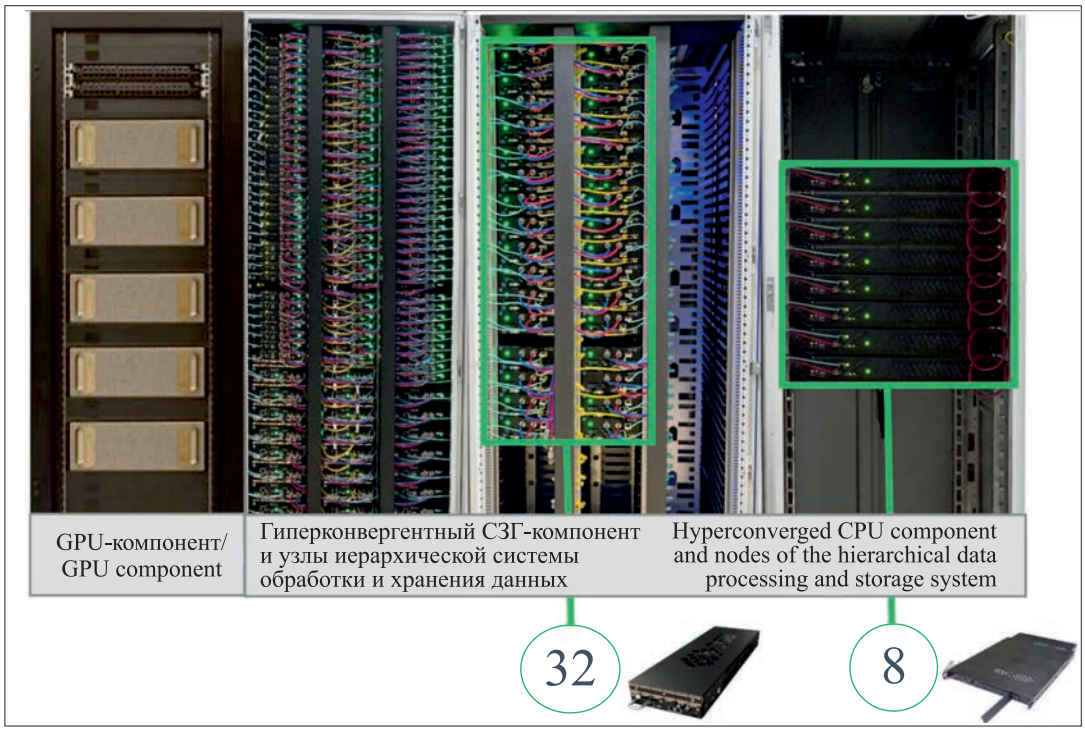
The hyperconvergence of new compute nodes has already enabled their use for the tasks of mass generation and data reconstruction within the NICA MPD experiment. At the same time, at different stages of generation and reconstruction, there is a need for different access rates to data; for example, for long-term storage tasks, access speed is not an important factor, however, for reconstruction tasks, speed plays a relevant role. In addition, for a number of MPD tasks, there was a need for a large amount of random-access memory (RAM), which is satisfied by new nodes. Thus, methodologically, to ensure all workflows for the tasks of the NICA megaproject, a system that combines both computing architectures of different types and the developed hierarchical data processing and storage system was created on the “Govorun” supercomputer. The computing resources and the hierarchical data processing

сечении ресурсов для различных пользовательских задач. Реализованная на гиперконвергентных узлах первой очереди система хранения данных по требованию (storage-on-demand) под управлением файловой системы Lustre позволила суперкомпьютеру «Говорун» занять 9-е место в мировом рейтинге IO500 (июнь 2018 г.) для систем хранения данных HPC-класса.

В результате модернизации CPU-компонент был расширен на 32 гиперконвергентных вычислительных узла на базе двух процессоров Intel Xeon Platinum 8368Q (частота 2,6 ГГц, 38 ядер, кэш 57 Мбайт, тепловыделение 270 Вт) в каждом, модулей оперативной памяти DDR4 — 256 ГБ на узел, 8 модулей энергонезависимой памяти Intel Optane DC Persistent Memory — 2 ТБ на узел, четырех твердотельных дисков SSD в формфакторе EDSFF E1.S (рулер) с интерфейсом NVMe — 16 ТБ на узел, а также с твердотельным диском SSD NVMe в формате M.2 емкостью 128 ГБ. Кроме того, каждый узел снабжен двумя адаптерами Intel Omni-Path с пропускной способностью 100 Гбит/с.

Уникальность такого расширения позволяет создать слой «очень горячих» данных емкостью 8 ПБ в иерархической системе обработки и хранения данных СК «Говорун». Создание такой иерархической системы было обусловлено ее востребованностью для работы с большими данными, прежде всего для мегапроекта NICA. По скорости доступа к данным система разделена на уровни: «очень горячие» данные — наиболее востребованные данные, к которым в настоящий момент требуется обеспечить самый быстрый доступ, «горячие» данные и «теплые» данные. Каждый уровень разработанной системы хранения данных может использоваться как самостоятельно, так и в составе рабочих процессов обработки данных. В настоящее время на СК «Говорун» в качестве слоя «очень горячих» данных осуществляется внедрение новейшей технологии DAOS (Distributed Asynchronous Object Storage) для обработки больших данных, показавшей свою перспективность для задач глубокого обучения и для работы квантовых симуляторов при эмуляции большего числа кубитов. За высокоскоростную систему обработки и хранения данных СК «Говорун» получил престижную премию Russian DC Awards 2020 в номинации «Лучшее ИТ-решение для центров обработки данных».

Гиперконвергентность новых вычислительных узлов уже позволила задействовать их для задач массовой генерации и реконструкции данных эксперимента MPD NICA. При этом на разных этапах генерации и реконструкции возникает потребность в разной скорости доступа к данным, например, для задач долговременного хранения скорость доступа не является важным фактором, а для задач реконструкции играет существенную роль. Также для ряда задач MPD возникла потребность в большом объеме оперативной памяти, которому удовлетворяют новые узлы. Таким образом,



Расположение новых узлов СК «Говорун» Placement of new nodes of the “Govorun” supercomputer

методологически, для обеспечения всех рабочих процессов для задач мегапроекта NICA на СК «Говорун» создана система, сочетающая в себе как вычислительные архитектуры различных типов, так и развитую иерархическую систему обработки и хранения данных. Вычислительные ресурсы и иерархическая система обработки и хранения данных СК «Говорун» были интегрированы на базе платформы DIRAC в распределенную гетерогенную среду, включающую в себя ресурсы ОИЯИ и стран-участниц. Практика использования различных вычислительных ресурсов ОИЯИ и других институтов коллаборации MPD показала, что на сегодня наиболее эффективным является использование ресурсов именно СК «Говорун».

Проведенная модернизация СК «Говорун» позволит ускорить исследования в области решеточной квантовой хромодинамики, качественно повысить эффективность моделирования динамики столкновений релятивистских тяжелых ионов, провести расчеты радиационной безопасности экспериментальных установок ОИЯИ и повысить эффективность решения прикладных задач.

Обновленный СК «Говорун» дает возможность не только проводить расчеты, но и использовать его как научно-исследовательский полигон для выработки программно-аппаратных и ИТ-решений. Это свойство позволило развернуть полигоны для квантовых вычислений и для обработки экспериментальных данных ЛРБ, включить ресурсы СК «Говорун» в единую гетерогенную среду на основе платформы DIRAC для проекта NICA и задействовать его ресурсы для реализации программы сеансов массового моделирования данных эксперимента MPD. Следует отметить, что некоторые задачи для моделирования данных эксперимента MPD возможно выполнить только на ресурсах СК «Говорун».

---

and storage system of the “Govorun” supercomputer were integrated into a DIRAC-based distributed heterogeneous environment that includes the resources of JINR and its Member States. The experience of using different computing resources of JINR and other MPD Collaboration institutes has shown that at present, the use of the “Govorun” supercomputer resources is the most efficient.

The above modernization of the “Govorun” supercomputer will make it possible to speed up studies in the field of lattice quantum chromodynamics, to qualitatively increase the efficiency of modeling the dynamics of relativistic heavy ion collisions, to carry out calculations of the radiation safety of JINR experimental facilities and to enhance the efficiency of solving applied tasks.

The modernized “Govorun” supercomputer enables not only to perform computing, but also to use the supercomputer as a research polygon for developing software, hardware and IT solutions for JINR tasks. This feature made it possible to deploy polygons for quantum computing and LRB experimental data processing, to integrate the resources of the “Govorun” supercomputer into a unified heterogeneous environment based on the DIRAC platform for the NICA project and to use its resources to implement the programme of runs of the mass simulation of MPD experiment data. It is noteworthy that some tasks for the MPD experiment data simulation can only be performed on the resourses of the “Govorun” supercomputer.