ЯДЕРНАЯ ФИЗИКА, 2020, том 83, № 6, с. 534–538

ЭЛЕМЕНТАРНЫЕ ЧАСТИЦЫ И ПОЛЯ

ТЕНДЕНЦИИ И ПЕРСПЕКТИВЫ РАЗВИТИЯ РАСПРЕДЕЛЕННЫХ ВЫЧИСЛЕНИЙ И АНАЛИТИКИ БОЛЬШИХ ДАННЫХ ДЛЯ ПОДДЕРЖКИ ПРОЕКТОВ КЛАССА МЕГАСАЙЕНС

© 2020 г. В. В. Кореньков*

Объединенный институт ядерных исследований, Лаборатория информационных технологий, Дубна, Россия

Поступила в редакцию 26.04.2020 г.; после доработки 26.04.2020 г.; принята к публикации 26.04.2020 г.

Важнейшей частью проектов класса мегасайенс является создание и развитие компьютерных систем для обработки, хранения и анализа экспериментальных данных, алгоритмов поиска и доступа к данным. Информационно-вычислительные инфраструктуры, необходимые для выполнения исследовательских задач проектов класса мегасайенс, являются сложными распределенными, гетерогенными системами, включая системы экстрамассивного параллелизма, и системами распределенного хранения огромных массивов данных.

DOI: 10.31857/S0044002720050153

Российские исследовательские институты и университеты активно участвуют в международных мегапроектах: эксперименты ATLAS, ALICE, LHCb, CMS на Большом адронном коллайдере (LHC) в Европейской организации ядерных исследований (ЦЕРН), Европейский рентгеновский лазер на свободных электронах (XFEL) в Немецком исследовательском центре по физике частиц (DESY), Европейский источник синхротронного излучения (ESRF) в Гренобле, эксперименты CBM, PANDA на комплексе по исследованию ионов и антипротонов (FAIR) в немецком центре по изучению тяжелых ионов имени Гельмгольца (GSI), Международный экспериментальный термоядерный реактор (ITER) в исследовательском центре Кадараш (Франция) и др. В России идет подготовка проектов класса мегасайенс: НИКА (Коллайдер протонов и тяжелых ионов) в ОИЯИ, Дубна; ПИК (Высокопоточный реакторный комплекс) в НИЦ ПИЯФ, Гатчина; СКИФ (Сибирский кольцевой источник фотонов) в ИЯФ СО РАН, Новосибирск; и другие. Реализуется нейтринная программа: проекты в России (Байкал), Китае (JUNO), США (NOvA, DUNE) и другие масштабные проекты.

Для обработки, хранения и анализа данных экспериментов на Большом адронном коллайдере в ЦЕРНе создана распределенная инфраструктура на основе грид-технологий, которая называется Всемирный вычислительный грид для LHC (Worldwide LHC Computing Grid — WLCG) [1]. WLCG сегодня объединяет ~1 000 000 процессорных ядер, ~0.6 эксабайт хранилища на дисках и ~0.8 эксабайт на ленточных роботах, обеспечивающих долговременное хранилище данных, которые географически распределены по 170 центрам обработки данных в 42 странах. Ежедневно эта система обрабатывает более 2 миллионов заданий и управляет сотнями петабайт данных. Инфраструктура WLCG была одним из факторов успеха первой фазы LHC, обеспечившим открытие бозона Хиггса.

Эксперименты на LHC играют ведущую роль в научных исследованиях не только в физике элементарных частиц и ядерной физике, но и в области аналитики Больших данных. За эти годы модель компьютинга на базе грид-технологий для LHC претерпела ряд изменений, которые позволили ей в большей мере удовлетворять запросам научного сообщества. От строго иерархической модели обработки [2], в которой весь процесс сбора и обработки данных распределялся по вычислительным центрам определенного уровня: Tier0 — основной центр в ЦЕРНе для сбора всех необработанных данных и их первичная реконструкция со всех экспериментальных установок и распределения данных по 14 центрам первого уровня Tier1 для их долговременного хранения, переобработки и анализа и обеспечения доступа к этим данным центров второго уровня Tier2, которые предназначены для проведения этапа моделирования и анализа данных конечными пользователями. При этом каждый Tier1 был связан с определенными Tier2 центрами. В новой модели центры уровня Tier1 и Tier2 вза-

*E-mail: korenkov@jinr.ru

имодействуют друг с другом. Кроме того, сегодня обработка и анализ данных ведется с использованием высокопроизводительных комплексов, академических, национальных и коммерческих ресурсов облачных вычислений, суперкомпьютеров и других ресурсов [3].

Российские центры, в первую очередь НИЦ КИ и ОИЯИ, активно участвуют в интеграции распределенных неоднородных ресурсов и развитии технологий Больших данных для обеспечения современных мегапроектов в таких высокоинтенсивных областях науки, как физика высоких энергий, астрофизика, биоинформатика и другие. В ОИЯИ активно ведутся работы по сооружению уникального ускорительного комплекса НИКА [4], который требует новых подходов к реализации распределенной инфраструктуры для обработки и анализа экспериментальных данных.

Следует отметить, что первоначально грид-технологии реализовали концепцию HTC (High-throughput computing), а в результате эволюции модели компьютинга произошло объединение различных технологий: HTC, HPC (High Performance Computing), добровольные вычисления (Volunteer computing), коммерческие и некоммерческие облачные вычислительные ресурсы. Такой подход необходим для удовлетворения требований экспериментов класса мегасайенс как по производительности систем обработки, так и по объему хранения данных. Кроме того, требуются дальнейшие изменения модели компьютинга в каждом эксперименте с целью оптимизации использования ресурсов. Сегодня значительные усилия вкладываются в развитие программного обеспечения, чтобы улучшить общую производительность при использовании современных архитектур (многоядерность, графические процессоры и другое). Необходима оптимизация процессов обработки, систем хранения и количества хранящихся реплик данных.

Ключевым моментом в организации таких инфраструктур, в частности WLCG, является связующее промежуточное программное обеспечение (платформа), позволяющее осуществлять совместную работу в информационно-вычислительных системах. Например, в эксперименте ATLAS на LHC разработана платформа для управления вычислительными ресурсами PanDA (Production and Distributed Analysis) Workload Management System (WMS) [5], которая является автоматизированной и настраиваемой системой управления заданиями и оптимизирует доступ пользователей к распределенным ресурсам. С помощью PanDA пользователи видят единый вычислительный ресурс, который предназначен для обработки данных эксперимента, хотя ресурсные центры разбросаны по всему миру. PanDA изолирует физиков от аппаратного обеспечения, системного и промежуточного программного обеспечения и других технологических сложностей, связанных с конфигурированием сети и оборудования. Вычислительные задачи автоматически отслеживаются и выполняются. В настоящее время PanDA контролирует сотни вычислительных центров в 50 странах мира, сотни тысяч вычислительных узлов, сотни миллионов заданий в год, тысячи пользователей.

Другим вариантом связующего промежуточного программного обеспечения является DIRAC Interware [6] — продукт для интеграции гетерогенных вычислительных ресурсов и ресурсов хранения данных в единую платформу. Интеграция ресурсов основана на использовании стандартных протоколов доступа к данным (xRootD, GridFTP и других) и пилотных задач. Благодаря этому пользователю предоставляется единая среда, в которой можно запускать задачи, управлять данными, выстраивать процессы и контролировать их выполнение. В рамках DIRAC в качестве вычислительных ресурсов могут выступать системы пакетной обработки, грид-сайты, облака, суперкомпьютеры и даже отдельно стоящие вычислительные узлы. Важной концепцией в DIRAC являются пилотные задачи. Именно благодаря им можно интегрировать практически любой вычислительный ресурс. При работе с данными DIRAC предоставляет весь необходимый набор команд. Для корректного функционирования всех команд система хранения должна поддерживать грид-протоколы передачи данных.

В настоящее время большое внимание уделяется новым перспективным направлениям в создании распределенных хранилищ данных (DataLake) [7], что позволяет существенно повысить эффективность хранения больших данных в сочетании с высокой скоростью доступа к данным.

Большую роль в развитии компьютинга для проектов класса мегасайенс играет Лаборатория информационных технологий ОИЯИ, основной задачей которой является развитие сетевой, информационно-вычислительной инфраструктуры ОИЯИ для научно-производственной деятельности института [8]. Активно развивается многофункциональный информационно-вычислительный комплекс (МИВК) ОИЯИ [9], который отвечает требованиям, предъявляемым к современному высокоэффективному научно-вычислительному комплексу: многофункциональность, высокая производительность, многоуровневая система хранения данных, высокая надежность и доступность, информационная безопасность, масштабируемость, индивидуальная программная среда для различных групп пользователей, высокопроизводительные телекоммуникации и современная

локальная вычислительная сеть. Многофункциональный информационно-вычислительный комплекс ОИЯИ в настоящее время имеет следующие основные компоненты:

1. центральный информационно-вычислительный комплекс (ЦИВК) ОИЯИ со встроенными вычислительными и запоминающими элементами,

2. кластер Tier2 для всех экспериментов на Большом адронном коллайдере (LHC) и других виртуальных организаций (VOs) в грид-среде [10] (4128 ядер, общая полезная емкость дисковых серверов составляет 2.929 петабайта),

3. кластер Tier1 для эксперимента CMS [11] (10 688 ядер, полезная емкость дисковых серверов — 10.4 петабайт, ленточных роботов — 51 петабайт),

4. гетерогенная платформа HybriLIT для высокопроизводительных вычислений (HPC) с суперкомпьютером "Говорун" [12] (совокупная пиковая производительность суперкомпьютера 860 терафлопс для операций с двойной точностью),

5. облачная инфраструктура [13] (1564 ядра),

6. система хранения данных на базе файловой системы EOS (3.740 петабайта дискового пространства).

В последнее время основное внимание для мегасайенс-проектов уделяется развитию телекоммуникационной и сетевой инфраструктуры, включая модернизацию локальной вычислительной сети с целью обеспечения ресурсов хранения и обработки данных.

В настоящий момент эксплуатационные характеристики и системы хранения данных базового грид-компонента МИВК — сайта CMS Tier1 ОИЯИ обеспечивают ему устойчивое второе место в мире среди других сайтов CMS Tier1 по количеству обработанных событий.

Активно развивался сайт ОИЯИ уровня Tier2. Он обеспечивает обработку данных четырех экспериментов на LHC (ALICE, ATLAS, CMS, LHCb), а также целого ряда виртуальных организаций, не связанных с LHC (BESIII, BIOMED, COMPASS, MPD, NOvA, STAR, ILC). МИВК также обеспечивает вычислительную мощность для вычислений, выполняемых вне грид-среды. Это очень важно для таких экспериментов, как NOvA, PANDA, BESIII, NICA/MPD/BM@N и пользователей из всех лабораторий ОИЯИ.

Еще одна компонента МИВК — облачная инфраструктура ОИЯИ. В рамках этой инфраструктуры была проведена интеграция облачных структур государств-членов ОИЯИ.

Важной частью МИВК является гетерогенная вычислительная платформа HybriLIT, состоящая из учебно-тестового полигона и суперкомпьютера "Говорун", совместно использующих единое программное обеспечение и информационную среду. Суперкомпьютер "Говорун" предназначен для проведения ресурсоемких и массивно параллельных вычислений при решении широкого круга задач, стоящих перед ОИЯИ, что становится возможным благодаря неоднородности (наличию различных типов вычислительных ускорителей) аппаратной архитектуры суперкомпьютера.

Для расширения возможностей разработки математических моделей и алгоритмов и проведения ресурсоемких вычислений, в том числе на графических ускорителях, значительно сокращающих вычислительное время, была создана и активно развивается экосистема для задач машинного/глубокого обучения и анализа данных для пользователей платформы HybriLIT.

Проект МИВК оказался успешным объединением всех вычислительных и инфраструктурных ресурсов. Он обеспечивает надежную и хорошо построенную вычислительную среду для проведения научных исследований учеными ОИЯИ и его государств-членов. Наличие таких вычислительных средств высшего уровня, как суперкомпьютер "Говорун" и центр CMS Tier1, способствует значительному повышению узнаваемости ОИЯИ во всем мире.

Разработанная комплексная система мониторинга [14] МИВК позволяет получать информацию от различных компонентов вычислительного комплекса: инженерной инфраструктуры, сети, вычислительных узлов, систем запуска задач, элементов хранения данных, грид-сервисов, что гарантирует высокий уровень надежности МИВК.

Для мегасайенс-проекта НИКА создается гетерогенный распределенный информационно-вычислительный кластер, что позволяет наиболее полно удовлетворить требования участников проекта как в области теоретических исследований, так и в области обработки, хранения и анализа экспериментальных данных детекторов BM@N, MPD и SPD. Распределенный информационно-вычислительный кластер комплекса НИКА в его базовой конфигурации должен обеспечить обработку и хранение до 10 петабайт данных в год. Комплекс состоит из территориально распределенных on-line и off-line кластеров, связанных между собой высокоскоростной компьютерной сетью с пропускной способностью 4×100 Гбит/с.

Разрабатываемые модели компьютинга должны учитывать тенденции развития сетевых решений, вычислительных архитектур и ИТ-решений, позволяющих объединять суперкомпьютерные (гетерогенные), грид- и облачные технологии и создавать на этой основе распределенные,

программно-конфигурируемые HTC и HPC платформы. Для экспериментов на ускорительном комплексе НИКА создан распределенный масштабируемый гибридный кластер, который можно легко реконфигурировать по требованию различного класса задач и пользователей. Важным компонентом этого кластера является распределенная двухуровневая (диско-ленточная) система хранения.

Суперкомпьютер "Говорун" используется в составе распределенного кластера НИКА для решения задач, требующих массивных параллельных вычислений в решеточной квантовой хромодинамике для изучения свойств адронной материи при высокой плотности энергии, для математического моделирования взаимодействий антипротонов с протонами и ядрами с использованием генераторов DPM, FTF и UrQMD + SMM, разработанных в ОИЯИ и представляющих интерес для эксперимента НИКА-MPD, для моделирования динамики столкновений релятивистских тяжелых ионов.

Еще одним компонентом кластера НИКА является сверхбыстрая система хранения данных (ССХД), реализованная в суперкомпьютере "Говорун" под управлением файловой системы Lustre. В настоящее время ССХД организована на твердотельных накопителях с технологией подключения NVMe, что сокращает время доступа к данным и обеспечивает скорость ввода/вывода более 300 гигабайт в секунду.

С помощью программной платформы DIRAC были объединены вычислительные ресурсы МИВК ОИЯИ: Tier1/Tier2, суперкомпьютер "Говорун", облако ОИЯИ и ресурсы хранения: ССХД Lustre, dCache и EOS. Эти результаты вносят существенный вклад в развитие цифровой платформы для проектов класса мегасайенс.

СПИСОК ЛИТЕРАТУРЫ

1. The Worldwide LHC Computing Grid (WLCG): http://wlcg.web.cern.ch/LCG
2. LHC Computing Grid: Technical Design Report, document LCG-TDR-001, CERN-LHCC-2005-024 (CERN, 2005).
3. Ph. Charpentier, in *Proceedings of the 23rd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2018), Sofia, 2018,* Ed. by A. Forti, L. Betev, M. Litmaath, O. Smirnova, and P. Hristov, EPJ Web Conf. **214**, 09009 (2019); https://doi.org/10.1051/epjconf/201921409009
4. Мегапроект НИКА: https://nica.jinr.ru/ru/
5. T. Maeno, J. Phys.: Conf. Ser. **119**, 062036 (2008).
6. F. Stagni, A. Tsaregorodtsev, Ch. Haen, Ph. Charpentier, Z. Mathe, W. J. Krzemien, and V. Romanovskiy, in *Proceedings of the 23rd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2018), Sofia, 2018,* Ed. by A. Forti, L. Betev, M. Litmaath, O. Smirnova, and P. Hristov, EPJ Web Conf. **214**, 03012 (2019); https://doi.org/10.1051/epjconf/201921403012
7. I. Bird, S. Campana, M. Girone, X. Espinal, G. McCance, and J. Schovancová, in *Proceedings of the 23rd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2018), Sofia, 2018,* Ed. by A. Forti, L. Betev, M. Litmaath, O. Smirnova, and P. Hristov, EPJ Web Conf. **214**, 04024 (2019); https://doi.org/10.1051/epjconf/201921404024
8. V. Korenkov, A. Dolbilov, V. Mitsyn, I. Kashunin, N. Kutovskiy, D. Podgainy, O. Streltsova, T. Strizh, V. Trofimov, and P. Zrelov, in *Proceedings of the 23rd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2018), Sofia, 2018,* Ed. by A. Forti, L. Betev, M. Litmaath, O. Smirnova, and P. Hristov, EPJ Web Conf. **214**, 03009 (2019); https://doi.org/10.1051/epjconf/201921403009
9. A. Dolbilov, I. Kashunin, V. Korenkov, N. Kutovskiy, V. Mitsyn, D. Podgainy, O. Stretsova, T. Strizh, V. Trofimov, and A. Vorontsov, in *Proceedings of the 27th Symposium on Nuclear Electronics and Computing, Montenegro, Budva, 2019*, Ed. by V. Korenkov, T. Strizh, A. Nechaevskiy, and T. Zaikina, CEUR Workshop Proceedings **2507**, 16 (2019).
10. A. Baginyan, A. Balandin, A. Dolbilov, A. Golunov, N. Gromova, I. Kadochnikov, I. Kashunin, V. Korenkov, V. Mitsyn, D. Oleynik, I. Pelevanyuk, A. Petrosyan, S. Shmatov, T. Strizh, A. Vorontsov, V. Trofimov, *et al.*, in *Proceedings of the 27th Symposium on Nuclear Electronics and Computing, Montenegro, Budva, 2019*, Ed. by V. Korenkov, T. Strizh, A. Nechaevskiy, and T. Zaikina, CEUR Workshop Proceedings **2507**, 321 (2019).
11. N. Astakhov, A. Baginyan, S. Belov, A. Dolbilov, A. Golunov, I. Gorbunov, N. Gromova, I. Kadochnikov, I. Kashunin, V. Korenkov, V. Mitsyn, I. Pelevanyuk, S. Shmatov, T. Strizh, E. Tikhonenko, V. Trofimov, *et al.*, Phys. Part. Nucl. Lett. **13**, 714 (2016); A. Baginyan, A. Balandin, S. Belov, A. Dolbilov, A. Golunov, N. Gromova, I. Kadochnikov, I. Kashunin, V. Korenkov, V. Mitsyn, I. Pelevanyuk, S. Shmatov, T. Strizh, V. Trofimov, N. Voytishin, and V. Zhiltsov, in *Proceedings of the 8th International Conference "Distributed Computing and Grid-technologies in Science and Education", Dubna, 2018*, Ed. by V. Korenkov, A. Nechaevskiy, T. Zaikina, and E. Mazhitova, CEUR Workshop Proceedings **2267**, 1 (2018).

12. 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, and M. Zuev, in *Proceedings of the 8th International Conference "Distributed Computing and Grid-technologies in Science and Education", Dubna, 2018*, Ed. by V. Korenkov, A. Nechaevskiy, T. Zaikina, and E. Mazhitova, CEUR Workshop Proceedings **2267**, 638 (2018).
13. N. Balashov, A. Baranov, N. Kutovskiy, A. Makhalkin, Y. Mazhitova, I. Pelevanyuk, and R. Semenov, in *Proceedings of the 27th Symposium on Nuclear Electronics and Computing, Montenegro, Budva, 2019*, Ed. by V. Korenkov, T. Strizh, A. Nechaevskiy, and T. Zaikina, CEUR Workshop Proceedings **2507**, 185 (2019).
14. A. Baginyan, N. Balashov, A. Baranov, S. Belov, D. Belyakov, Y. Butenko, A. Dolbilov, A. Golunov, I. Kadochnikov, I. Kashunin, V. Korenkov, N. Kutovskiy, A. Mayorov, V. Mitsyn, I. Pelevanyuk, R. Semenov, *et al.*, in *Proceedings of the 26th International Symposium on Nuclear Electronics and Computing (NEC 2017), Budva, 2017*, Ed. by V. Korenkov and A. Nechaevskiy, CEUR Workshop Proceedings **2023**, 226 (2017).

TRENDS AND PROSPECTS FOR THE DEVELOPMENT OF DISTRIBUTED COMPUTING AND BIG DATA ANALYTICS TO SUPPORT MEGASCIENCE PROJECTS

V. V. Korenkov

Joint Institute for Nuclear Research, the Laboratory of Information Technologies, Dubna, Russia

The creation and development of computer systems for experimental data processing, storage and analysis, of search algorithms and data access are crucial to megascience projects. Information and computing infrastructures, necessary for carrying out the research tasks of megascience projects, represent complex distributed, heterogeneous systems, including systems of extra-massive parallelism, and systems of distributed storage of big data arrays.