УДК 004.4

Применение технологий платформы HybriLIT для построения вычислительных комплексов в различных организациях

М. В. Башашин[1, 2], **М. А. Матвеев**[1], **М. И. Зуев**[1]

[1] *Лаборатория информационных технологий, Объединённый институт ядерных исследований, ул. Жолио-Кюри 6, Дубна, Московская область, Россия, 141980*
[2] *Государственное бюджетное образовательное учреждение высшего образования Московской области Университет «Дубна», ул. Университетская 19, Дубна, Московская область, Россия, 141980*

Email: bashashinmv@jinr.ru, matveevma@jinr.ru, zuev@jinr.ru

В настоящее время в мире наблюдается рост потребности в вычислительных комплексах. Множество вычислительных серверов, объединенных высокоскоростной сетью, еще называют суперкомпьютером.

В 2018 году, в Дубне запущен в эксплуатацию суперкомпьютер «Говорун», который входит в состав платформы HybriLIT.

Технологии платформы HybriLIT получили своё распространение в создании других вычислительных комплексов. Первым, кто решил воспользоваться наработками группы HybriLIT был Институт Математики и Цифровой Технологии, Улан-Батор, Монголия в 2015 году. В 2018 добавился Российский Экономический Университет им. Г. В. Плеханова, Москва. В 2021 году стек технологий гетерогенной платформы HybriLIT был реализован на мобильной версии платформы для работы в Государственном Университете «Дубна».

В докладе представлено описание основных технологий для построения вычислительного комплекса малой и средней производительности.

Ключевые слова: суперкомпьютер, вычислительный комплекс, гипервизор

1. Введение

В 2018 году, в Дубне запущен в эксплуатацию суперкомпьютер «Говорун» [1]. Суперкомпьютер входит в состав платформы HybriLIT, одной из важнейших функций которой, является внедрение передовых технологий для взаимодействия пользователей с вычислительными устройствами нового поколения, поэтому особое внимание уделяется мониторингу и безопасности комплекса.

Чтобы приступить к работе опытным пользователям достаточно получить доступ к VPN, заполнить форму регистрации на платформе с указанием типов ресурсов и их количества. Новичкам следует начать своё ознакомление с инструкции по доступу и работе на учебно-тестовом полигоне, который входит в состав платформы. Инструкция размещена на сайте платформы – hlit.jinr.ru.

Обучение пользователей начинается с выбора режима доступа к ресурсам вычислительных комплексов (ВК): графического или терминального.

При работе в терминале пользователям доступен расширенный список компиляторов и библиотек, но нет возможности работать с проприетарным ПО. В терминальном режиме доступны все ресурсы очередей, в то время как в графическом – ресурсы ограничены интерактивными.

В графическом режиме доступны такие интерфейсы, как VDI (Virtual Desktop Infrastructure) и JHUB (Jupyter Hub). Интерфейс VDI представляет из себя удаленный рабочий стол с графическим доступом к проприетарному ПО: Comsol, Matlab, Maple, Mathematica. Интерфейс JHUB является программно-аппаратной средой для задач машинного обучения. В отличии от терминального доступа, эти интерфейсы взаимодействуют только с интерактивными ресурсами, лимит которых, сильно ограничен по сравнению с ресурсами, доступными в очередях.

За несколько лет эксплуатации суперкомпьютер «Говорун» модернизировался дважды. Весь накопленный опыт позволил выделить главные компоненты конструкции в самостоятельный сегмент с перспективой дальнейшего масштабирования по аналогии с облачными технологиями, используемыми в ОИЯИ [2].

На данный момент, несколько учреждений участвовали в создании ВК, аналогичных HybriLIT. Первыми, кто решил воспользоваться наработками группы HybriLIT были Институт Математики и Цифровой Технологии, Улан-Батор, Монголия в 2015 году [3]. В 2018 к ним присоединился Российский Экономический Университет им. Г. В. Плеханова, Москва [4]. В 2021 году стек технологий гетерогенной платформы HybriLIT был реализован на мобильной версии платформы для работы в Государственном Университете «Дубна» [5].

В данной статье описаны ключевые технологии, необходимые при построении безопасного и отказоустойчивого вычислительного комплекса.

2. Технологии

На начальной стадии построения вычислительного комплекса необходимо оценить масштабы с помощью примерного количества пользователей и их интересов. Например, один сервер с 4 графическими ускорителями, на задачах машинного обучения мог бы обслуживать до 4 пользователей, одновременно.

Как правило, в учебных заведениях с ресурсами ВК работают студенты и используют не много машинного времени, поэтому линейная зависимость количества GPU от количества пользователей не совсем корректна. На таком сервере свободно смогли бы работать несколько десятков пользователей с небольшими периодом ожидания доступа к вычислительному ресурсу.

С другой стороны, без программных ограничений, один пользователь способен загрузить вычислениями десятки серверов. Поэтому, еще один важный критерий при построении ВК – это уровень компетенций пользователей.

Количество и компетенции пользователей в первую очередь повлияли на направление развития вычислительных технологий в ЛИТ ОИЯИ, так как институт занимается передовыми разработками в различных областях науки, является международным, а также имеет постоянный поток студентов. Поэтому, для нужд начинающих и продвинутых пользователей целесообразно иметь различные категории ресурсов. На сегодняшний день, в ЛИТ ОИЯИ функционируют учебно-тестовый полигон HybriLIT для обучения сотрудников и студентов, а также суперкомпьютер «Говорун» для решения ресурсоёмких задач.

Суперкомпьютер «Говорун» содержит наиболее передовые устройства для вычислений, такие как графические ускорители (GPU) Nvidia Tesla V100 и A100. Еще несколько моделей GPU прошлых поколений входят в состав учебного полигона – Nvidia Tesla K20X, K40, K80.

Кроме графической компоненты пользователям доступны для расчетов несколько моделей многоядерных CPU Intel Xeon Platinum 8268, 8280, Phi 7190. В состав учебного полигона входят Intel Xeon E5-2695 v2, E5-2695 v3, E5-2680 v3, Phi 5110P, 7290P.

Программно-аппаратная среда для машинного обучения JHUB содержит графические ускорители Nvidia V100 и процессоры Intel E5-2698v4.

Виртуальные рабочие столы, с возможностью запуска проприетарного программного обеспечения (ПО), содержат графические ускорители Nvidia M60 и процессоры Intel E5-2697A.

Для сетевого взаимодействия используется InfiniBand, Omni-Path и Ethernet от 10 до 40 Гбит/сек.

Данные всех проектов хранятся на серверах с общим дисковым пространством 9 Пб. Для доступа к сетевым файлам используется ZFS. В качестве распределенной файловой системы – Lustre.

Локальный репозиторий платформы HybriLIT постоянно расширяется и оперативно обновляется. В терминальном режиме доступ к пакетам организован с помощью модулей. В интерактивном режиме на рабочем столе подготовлены ярлыки для запуска проприетарного ПО.

Далее будут перечислены еще несколько решений, необходимых для лёгкого и безопасного масштабирования всех компонентов суперкомпьютера: в качестве системы для управления пользователями и их доступом используется LDAP; для распределения нагрузки на виртуальные машины, с которыми взаимодействуют десятки пользователей одновременно, используется Haproxy (roundrobin); с 2022 года доступ к вычислительным ресурсам ЛИТ ОИЯИ возможен только через VPN (L2TP).

3. Безопасность

Информационной безопасности (ИБ) сегодня уделяется повышенное внимание. Следует выделить две составляющие ИБ: продуманные системными администраторами IT решения, а также регламент доступа и использования суперкомпьютера. Если в разработке регламента системные администраторы участвуют наравне с другими вовлеченными в проект сотрудниками, то при внедрении IT решений вся ответственность ложится на системных администраторов.

Можно выделить несколько первостепенных решений, которые позволят уменьшить количество инцидентов:

— В целях информационной безопасности, в первую очередь следует ограничить прямой доступ к вычислительным ресурсам через интернет. Поэтому, необходим VPN для доступа во внутреннюю сеть. Доступ пользователей к данным и вычислительным ресурсам в локальной сети следует ограничить на всех уровнях: VPN (маршрутизация), LDAP (политики), firewall.
— Такой подход позволит обезопасить комплекс от несанкционированных действий заинтересованных пользователей, которые получили доступ в сеть, пройдя регламентный фильтр. Однако, большим источником инцидентов является начинающая группа пользователей, в связи с недостаточным уровнем знаний в области взаимодействия с терминалом. Например, случаи с логическим искажением данных чаще всего контролируется регламентом и не более. Случаи с физическим искажением данных контролируется с помощью распределенной файловой системы, программными и аппаратными рейдами или резервными копиями данных.
— Сопровождение суперкомпьютера неизбежно связано с постоянным мониторингом всех его компонентов: датчики, нагрузка на сеть, CPU, GPU, RAM, диски и визуализация собранной статистики.

Комбинация описанных решений позволит уменьшить количество инцидентов и уделить больше времени на внедрение новых сервисов или масштабирования имеющихся компонентов.

4. Применение

Многие из перечисленных решений нашли своё применение при организации вычислительной инфраструктуры в других научных учреждениях, в рамках сотрудничества с ОИЯИ. Первыми, кто решил воспользоваться наработками группы HybriLIT был Институт Математики и Цифровой Технологии, Улан-Батор, Монголия в 2015 году.

Вычислительный компонент комплекса в Улан-Баторе содержит семь серверов с процессорами Intel Xeon Gold 6130, 6240R, 6252, 6312U, и графическими ускорителями Nvidia Tesla V100, A40. Объем дискового пространства хранилища данных 90Тб. Для доступа к домашним каталогам и репозиторию программного обеспечения используется сетевая файловая система NFS.

Один сервер используется в качестве гипервизора для хостинга виртуальных машин, которые необходимы для работы основных служб:

— Для безопасного доступа к вычислительному комплексу через интернет настроен OpenVPN с дополнительными настройками, ограничивающими доступ пользователей к IP адресам во внутренней сети.
— Для управление пользователями и их доступом к домашним каталогам настроен LDAP с управлением через CLI и web интерфейс – FreeIPA с политиками ограничения доступа к узлам.
— Для равномерного распределения нагрузки, которую создают пользователи на виртуальных машинах, настроен Haproxy.
— Для мониторинга серверов и виртуальных машин настроен Telegraf.
— Для визуализации агрегированных данных настроен Chronograf
— В качестве планировщика задач используется SLURM.

В 2018 тот же подход был применен в Российском Экономическом Университете им. Г. В. Плеханова, Москва. Вычислительный сервер в РЭУ содержит процессоры Intel Xeon Gold 6130 и графические ускорители Nvidia Tesla V100. В качестве сетевой файловой системы используется CephFS. Для управления доступом к вычислительным ресурсам используется планировщик задач SLURM. Для интерактивного доступа к графическим ускорителям в задачах машинного обучения настроен JupyterHub.

В 2021 году стек технологий гетерогенной платформы HybriLIT был реализован на мобильной версии платформы для работы в Государственном Университете «Дубна». Три мини ПК содержат процессоры Intel Core i9-9900, i7-8700T и видеокарты Nvidia Quadro P1000. Для управление пользователями настроен web интерфейс – FreeIPA. Для управления доступом к вычислительным ресурсам используется планировщик задач SLURM.

5. Заключение

В докладе описан минимальный набор технологий для построения вычислительного комплекса до тысячи пользователей. Полученный опыт переноса основных технологий на вычислительные комплексы малых масштабов позволил некоторым организациям внедрить высокопроизводительные технологии в своих учреждениях.

При должной настройке серверное оборудование имеет многолетний запас срока эксплуатации в режиме 24/7. Контроль параметров износа оборудования помогает своевременно находить проблемные комплектующие и менять их. Если замена комплектующих вычислительных серверов или целого сервера не влияет на работу всего комплекса, то вывод из строя сервера с системными сервисами может парализовать работу всех пользователей, на время замены.

Избежать подобной ситуации можно с помощью технологий отказоустойчивости и возможности кластеризации основных серверов. Комбинация из трёх и более серверов с распределенной файловой системой, например, CephFS позволит добиться отказоустойчивости основных сервисов за счет достижения консенсуса между серверами в случае выхода из строя одного из них.

Литература

1. Гетерогенная платформа HybriLIT, URL: http://hlit.jinr.ru (Дата обращения: 05.03.2024).
2. Облачная инфраструктура, URL: https://cloud-info.jinr.ru/ (Дата обращения: 05.03.2024).
3. Institute Of Mathematics and Digital Technology, URL: https://imdt.ac.mn (Дата обращения: 05.03.2024).
4. Российский Экономический Университет им. Г. В. Плеханова, URL: https://www.rea.ru/ru/org/managements/unitscires/Laboratorija-Oblachnykh-tekhnologijj-i-analitiki-Bolshikh-dannykh/Pages/IT-infrastructure.aspx (Дата обращения: 05.03.2024).
5. Государственный Университет "Дубна", URL: https://uni-dubna.ru/ (Дата обращения: 05.03.2024).

UDC 004.4

HybriLIT platform technologies to setup computing systems in external organizations

M. V. Bashashin[1, 2], **M. A. Matveev**[1], **M. I. Zuev**[1]

[1] *Laboratory of Information Technologies*
Joint Institute for Nuclear Research
Joliot-Curie 6, Dubna, Moscow region, Russia, 141980
[2] *Dubna State University*
Universitetskaya 19, Dubna, Moscow region, Russia, 141980

Email: bashashinmv@jinr.ru, matveevma@jinr.ru, zuev@jinr.ru

Currently, the world is experiencing an increase in the urge various computing systems. Computing servers connected by main network are called supercomputers.

In 2018, the GOVORUN supercomputer, which is a part of the HybriLIT platform, was launched in Dubna.

HybriLIT platform technologies have become widespread in the creation of other computing systems. The first who decided to use the developments of the HybriLIT group was the Institute of Mathematics and Digital Technology, Ulaanbaatar, Mongolia in 2015. In 2018, the Russian Economic University named after. G. V. Plekhanov, Moscow. In 2021, the technology stack of the heterogeneous HybriLIT platform was implemented on the mobile version of the platform for work at the Dubna State University.

The report provides a description of the main technologies for building a low- and medium-performance computing complex.

Key words and phrases: supercomputer, computing complex, hypervisor