За пределами суперкомпьютеров

На прошедшей в ноябре 2019 г. в Денвере (США) суперкомпьютерной конференции – SC'19 – компания HPE объявила о начале эры эксамасштаба. Сделано это было на фоне анонса о реализации новой аппаратной платформы Shasta в трёх эксакомпьютерах.

Вячеслав Елагин – специалист по продажам высокопроизводительных систем и систем искусственного интеллекта, HPE Россия.

Введение

«Beyound super» – с таким лозунгом компания HPE приняла участие в суперкомпьютерной конференции в Денвере, состоявшейся в середине ноября 2019 г. (SuperComputing'19, Денвер, Колорадо, США), объявив начало эксамасштабной эры. Время для этого заявления было выбрано неслучайно – эксакомпьютеры станут реальностью в недалёком будущем. Компания выполняет три проекта в интересах Министерства энергетики США, результатом которых станет появление трёх кластеров с производительностью эксафлопс и более. Был сделан также ряд важных заявлений, касающихся подходов к построению эксамасштабных систем, которые, по словам представителей HPE, должны быть «производительны, как суперкомпьютеры и работать как облако».

Необходимость радикального пересмотра подходов к реализации высокопроизводительных систем, успешно существовавших с момента выхода Cray-1 – первого коммерческого суперкомпьютера, продиктована требованиями рынка, в частности, развитием технологий машинного обучения и аналитики больших данных. Это означает, что суперкомпьютерные системы следующего поколения должны одинаково хорошо «переваривать» три типа нагрузок: традиционное моделирование и симуляции, требующие традиционных HPC аппаратных платформ, традиционных HPC систем хранения данных и традиционного HPC-ПО, а также нагрузки машинного обучения и аналитики больших данных. Для нагрузок этого типа традиционные HPC-аппаратные платформы, HPC-системы хранения данных, а также методы доставки HPC-ПО радикально иные. Поэтому и необходимо пересмотреть подходы к тому, как ПО доставляется и исполняется на вычислительных кластерных системах, а также пересмотреть существующее представление о самих вычислителях, о том, как они взаимодействуют с системами хранения данных и между собой.

Взрыв интереса ИТ-рынка к получению значимых результатов из анализа больших данных произошёл благодаря распространению и доступности самих данных, порождаемых миллиардами устройств, и технологий их обработки, поэтому на HPC-рынке наблюдается устойчивый рост. Ожидается, что в течение следующих трех лет сегмент рынка высокопроизводительных вычислений и связанные с ним рынки систем хранения данных и услуг вырастет, приблизительно, с $28 млрд в 2018 году, примерно, до $35 млрд в 2021 году, а совокупный годовой темп роста (ист.: Market data as of October 5, 2018) составит около 9%. При этом ожидается, что на проекты эксамасштаба в ближайшие пять лет будет потрачено более $4 млрд.

Решение сложных задач и продвижение критических научных исследований, в том числе таких, как обеспечение прорывных медицинских открытий, поиск новых материалов, развитие неуглеводородной энергетики требуют значительных вычислительных возможностей, вплоть до архитектуры эксамасштабного уровня. Экасмасштабные системы позволяют решать эти проблемы с гораздо большей точностью и глубиной понимания, поэтому только суперкомпьютеров для решения этих сверхмасштабных задач уже недостаточно. Нам нужно покинуть эру супер-вычислений и вступить в эру эксамасштаба.

HPC-системы для будущих применений

«Beyound super» – не просто лозунг, с которым мы выступили на Supercomputing 2019, мы объявили начало эксамасштабной эры. Сделано это было потому, что компания уже ведёт работы над тремя такими системами. Эти эксакомпьютеры скоро станут реальностью, и о них уже можно рассказывать. Эксакомпьютер – не просто очень большой вычислительный кластер. Для организации его работы необходимо пересмотреть подходы к:

- построению вычислителей и интерконнекта (потому что для свербольших систем интерконнект – очень сложная часть системы);
- организации взаимодействия с системами хранения;
- способу доставки ПО;
- инженерной инфраструктуре, электроснабжению и охлаждению.

Все эти вызовы встали перед командой разработчиков.

Объединение разработок HPE и Cray обеспечивает заказчиков системами, которые производительны, как суперкомпьютеры и работают, как облако. Особенно это важно на пороге эксамасштабной эры. Эксамасштаб должен радикально изменить все подходы к HPC, которые нас устраивали всё это время. Мы уже стали свидетелями, как за 20 лет HPC совершили переход с RICS и векторных архитектур на x86, как Unix заменил Linux, пусть даже и Unix-подобный... Все эти технологии изменяли ИТ-отрасль. Теперь мы стоим на пороге нового изменения – конвергенции вычислительных нагрузок. Помимо традиционных высокопроизводительных вычислительных нагрузок, эти системы должны исполнять также нагрузки машинного обучения и аналитики больших данных. Все три типа нагрузок должны работать на одной системе – и работать одинаково хорошо. Это – требование рынка и новый вызов к разработчикам систем и их эсплуатантам.

Эпоха, определяемая стремительным ростом объёмов данных, конвергентными рабочими нагрузками и цифровым преобразованием, знаменует начало новой главы исследований и анализа для корпоративных и государственных заказчиков. Эра эксамасштаба требует революционных инноваций в программном обеспечении и инфраструктуре, чтобы клиенты могли раскрыть весь потенциал своих данных и ускорить свои инновации, чтобы получить конкурентное преимущество. То, что в HPC-сообществе мы называем «эксамасштабом» в корпоративном сегменте зовётся «цифровой трансформацией». И мы планируем принести эту цифровую трансформацию в как можно большее количество ЦОДов по всему миру.

HPE CRAY Shasta – эксакомпьютер новой эры

В конце 2018 года HPE впервые представила концепцию своей новой аппаратной платформы Shasta, которая отвечает современным вызовам по поддержке конвергентных HPC-рабочих нагрузок и объединяет в себе новые вычислители, интерконнект, инженерную инфраструктуру и т.д. В ноябре 2019 объявлено об её аппаратной реализации для первых трёх эксафлопсных систем в США. Инженерная реализация Shasta позволяет выбрать, но не смешивать, на уровне шкафа тип охлаждения – жидкостное или воздушное. К ключевым характеристикам, которые инженеры-разработчики заложили в основу новой системы, относится возможность модернизации. Не секрет, что инженерная инфраструктура служит в несколько раз долее эффективнее, нежели ИТ-оборудование, установленное в них, поэтому возможность многократной модернизации учитывалась при разработке самих шкафов. И это – история об экономике: совокупной стоимости владения, экономической эффективности, возврате инвестиций и т.д. Кроме того, учитывался тренд в элементной базе на увеличение

термопакета для вычислительных устройств: инженеры ориентировались на появление в будущем электронных компонентов с энергопотреблением в 500+ Вт, которые смогли бы быть установлены в новые вычислительные платформы Shasta.

Вычислители Shasta могут быть построены на процессорах различных архитектур – x86 или ARM, содержать ускорители на ГПУ, программируемых матрицах (FPGA) или даже на специальных интегральных схемах (ASIC), обеспечивая гибкость предоставления аппаратной платформы исполняемым приложениям. Эти разные по своей архитектуре вычислители могут быть охлаждаемы жидкостью или воздухом и устанавливаются в привычный 19-дюймовый шкаф с жидкостным или воздушным охлаждением, соответственно. В случае воздушного хладагента продув воздуха – от фронтальной панели к тылу, как принято в отрасли.

Slingshot – интерконнект для микро- и эксамасштабных систем

Для поддержки работоспособности программ, интенсивно работающих с данными, необходимо было пересмотреть идеологию интерконнекта. При его разработке необходимо было предусмотреть бесшовную масштабируемость – от одной (коммерческий ЦОД) до сотен (эксакомпьютер) стоек, подключение к корпоративным Ethernet-ресурсам и возможность обеспечения адекватной производительности для конвергентных (HPC+AI+HPA) рабочих нагрузок. Поэтому из мира HPC мы взяли высокую пропускную способность и низкие задержки и добавили такие «корпоративные» функции, как например, обеспечение качества обслуживания (QoS) или управление перегрузками (congestion management). Так появился Slingshot – 8-й по счёту интерконнект, разработанный Cray. Пропускная способность начинается с 200 Гб/с, а его топология позволяет эксакомпьютеру добираться до нужных данных в другом узле не более чем за три «прыжка» (hops).

Slingshot позволяет существенно снизить эффект «шумных соседей» – когда, соседние узлы начинают интенсивный обмен данными. Почему для HPC-систем важно управление перегрузками? Группа исследователей решила уменьшить разрешение климатической модели на 7 км. Ожидаемое время счёта должно было составить 55 минут, но фактически счёт шёл на 20 минут дольше. Аналитики выявили, что большинство шагов рассчитывается менее чем за секунду, а некоторые исполняются более 20-ти. Выяснилось, что асинхронный ввод-вывод, исполняющийся одновременно с моделью, вызывал появление перегрузок (congestion, пробок) в сети, и это вело к появлению ненужных задержек. Эта проблема была решена путём изменения кода в LNET. И это обновление в LNET стало доступно всем пользователям систем Cray XC. Введение управления перегрузками (congestion management) помогло команде сократить время счёта модели с более высоким разрешением до 47 минут, то есть на 8 минут улучшить ожидаемый результат. И эта технология есть в Slingshot.

Также важно сказать, что Slingshot не является закрытым (proprietary) интерконнектом, и он не просто совместим с Ethernet. Это и есть Ethernet! Мы просто сделали свою реализацию и оптимизировали его под конвергентные HPC-нагрузки. Любой производитель мог это сделать.

Но Shasta – не только Slingshot. Платформа поддерживает HDR Infiniband. Поддержка Omni-Path Architecture следующего поколения (OPA-2) будет также возможна, если партнёры представят её рынку.

Cray ClusterStor E1000 – СХД для данных эксамасштаба

Мы строим компьютер производительностью эксафлопс, чтобы обработать эксабайт данных. Для эры эксамасштабных вычислений, помимо вычислителей и интерконнекта, необходима эксасистема хранения данных. И такая система хранения была объявлена незадолго до Supercomputing 2019 – ClusterStor E1000. Требования конвергентных HPC-нагрузок к доставке данных превышают возможности корпоративных систем хранения данных, поэтому для HPC-мира существуют параллельные файловые системы и программно-аппаратные комплексы хранения данных для HPC-систем. Среди них – Lustre – самая распространенная на рынке параллельная файловая система. Параллельные файловые системы прекрасно справляются с такими нагрузками, как «одно задание, исполняемое на десятках тысяч узлов, читает и пишет в одну файловую систему» или «десятки тысяч исполняемых заданий независимо читают и пишут в одну файловую систему». Это типичные нагрузки для параллельной файловой системы, характерные для высокопроизводительных вычислений и совершенно не характерные для систем хранения данных корпоративного сегмента. Ключевой характеристикой для параллельной файловой системы является пропускная способность, которую необходимо устойчиво предоставлять вплоть до эксабайтного масштаба. Традиционным HPC необходима скорость, задачам машинного обучения и аналитики больших данных – объём, и эти требования равнозначны для конвергентных HPC-систем эксамасштабной эры.

Ответом на вызов построения эксабайтной системы хранения для эксафлопсного вычислителя стала новая аппаратная платформа в ClusterStor E1000 – первые в индустрии NVMe Gen4 контроллеры систем хранения, разработанные Cray, новые высокоплотные полки для установки твердотельных или шпиндельных дисков и, конечно, интерконнект Slingshot – с управлением перегрузками, классификацией трафика, пропускной способностью 200 Гб/с и полностью Ethernet. Всё это уже обеспечивается новой функциональностью Lustre 2.12. Ядром новой ClusterStore E1000, связывающей всё в единый комплекс, является ПО, реализующее интеллектуальную оркестрацию потока данных в зависимости от рабочей нагрузки.

Ожидается, что пропускная способность системы ClusterStor для эксакомпьютера Frontier, строящегося сейчас в Окриджской национальной лаборатории, составит 10 ТБ/с при ёмкости системы в 1 эксабайт.

ПО эксамасштабной эры

Последний столп начала эксафлопсной эры – программное обеспечение. Оно должно поддерживать контейнеризированную и мультиарендную (multitenant) архитектуру, позволяя конвергентным HPC- и ИИ-нагрузкам работать одновременно на одной системе и давая администраторам и разработчикам возможность работать с эксакомпьютером, как с простой облачной системой.

Это программное обеспечение должно предоставлять доступ к вычислительным ресурсам по принципу облака для пользователей, обеспечивать доступный инструментарий для разработчиков, а администраторам – необходимые данные для сопровождения сложного вычислительного комплекса. И за это в ответе – программный пакет HPE для высокопроизводительных вычислений. Этот пакет охватывает наиболее распространённые корпоративные дистрибутивы Linux – Red Hat и SUSE, включает Cray Linux Environment и некоторые дистрибутивы от сообщества, например, CentOS. Программное обеспечение фабрик – Infiniband, OPA, Slingshot. Управление аппаратными платформами осуществляется Cray System Management и HPE Performance Cluster Mananger, для платформ Cray и HPE, соответственно, или Cluster Manager от компании Bright Computing. В ближайшей перспективе ПО управления пока будет разным для платформ Shasta и Apollo.

Платформа управления данными – Data Management Framework – позволяет управлять десятками петабайт данных в течение десятков лет. Она не зависит от технологий хранения и поддерживает параллельные файловые системы Lustre или CXFS. Мы предлагаем Weka IO в качестве файловой системы для задач машинного обучения и ClusterStore для конвергентных HPC-нагрузок. Мы поддерживаем целый ряд планировщиков для пакетного запуска задач и обширный инструментарий разработчиков, включая Cray Programming Environment и свой MPI.

Cray Programming Environment (CPE) позволяет упростить разработку приложений для высокопроизводительных вычислений и ИИ, которые запускаются на различных процессорах и ускорителях. CPE предоставляет полную и полностью поддерживаемую среду разработки – интегрированный программный пакет, предлагающий компиляторы и языки программирования, инструментарий и библиотеки, которые повышают продуктивность работы программистов, масштабируемость приложений и их производительность. Он предназначен для простого переноса (porting) существующих приложений с минимальной необходимостью перекодировки, позволяет сократить изменения в существующих моделях программирования и упрощает переход разработчика к новой аппаратной парадигме.

Программное обеспечение HPE Performance Cluster Management (HPCM) гарантирует полную поддержку, управление и мониторинг кластеров масштабом до 100 000 узлов. Оно позволяет осуществлять быструю настройку системы с нуля, комплексный мониторинг и управление оборудованием, управление образами,

Рис. 1. Три эксафлопных суперкомпьютера, разрабатываемых HPE. Планируемое завершение работ – 2021 г.

обновление программного обеспечения и управление электропитанием. Интеграция HPCM с планировщиками позволяет заказчикам воспользоваться синергетическим эффектом от сквозной программной интеграции от задачи до управления электричеством для снижения совокупной стоимости владения вычислительной системой.

Эксамасштабная тиара

Эксамасштабная эра уже у нас на пороге – компания ведёт три проекта построения эксафлопсных систем (рис. 1). Все они выполняются в интересах Министерства энергетики США.

Первый – проект Aurora. Исполняется в Аргонской национальной лаборатории. Его достигнутая производительность (R_{max}) будет около 1 эксафлопса. В основе его вычислительной составляющей – процессоры Intel Xeon следующего поколения и графические ускорители следующего поколения Intel X^e.

Второй – проект Frontier. Он исполняется для Окриджской национальной лаборатории. Его достигнутая производительность будет на уровне 1,5 эксафлопс. Он будет построен на элементной базе компании AMD – процессорах EPYC следующего поколения и ускорителях Radeon следующего поколения.

Третий – проект El Capitan. Он будет установлен в Лоуренс-Ливерморской национальной лаборатории. Его производительность составит также около 1,5 эксафлопс.

Все эти проекты реализуются на базе аппаратной концепции Shasta и содержат Slingshot в качестве интерконнекта – сотня вычислительных шкафов и три «прыжка» до самого дальнего вычислительного узла! Все эти системы будут одинаково хорошо справляться с HPC-, ИИ- и рабочими нагрузками, связанными с поддержкой задач аналитики больших данных.

Вячеслав Елагин,
HPE Россия.

HPE Container Platform

Декабрь 2019 г. – Hewlett Packard Enterprise (HPE) объявила о выпуске HPE Container Platform (*https://www.hpe.com/us/en/solutions/container-platform.html*), первой в отрасли контейнерной платформы корпоративного класса на базе Kubernetes, предназначенной как для приложений, разворачиваемых в облачных средах, так и для монолитных приложений, работающих с постоянным хранилищем данных. С помощью платформы HPE Container Platform корпоративные заказчики смогут ускорить разработку новых и существующих приложений в практически любых средах – на серверах без гипервизора, с использованием виртуальных машин, в публичном облаке и вне центров обработке данных – на границе сети.

Контейнерная платформа HPE основана на проверенных инновационных разработках BlueData и MapR, приобретенных HPE, а также на полностью открытой системе Kubernetes. Это решение нового поколения позволяет значительно снизить затраты и упростить процессы за счет работы контейнерных приложений на «голом» железе, а также обеспечивает гибкость при развертывании на виртуальных машинах и в облаке. Благодаря «свертыванию стека» и устранению необходимости в виртуализации, заказчики получают большую эффективность, рост утилизации и повышение производительности.

Новая платформа отвечает всем требованиям, предъявляемым корпоративными пользователями к крупномасштабным развертываниям на базе Kubernetes для различных целей: от машинного обучения и аналитики на границе сети до CI/CD и разработки приложений. ИТ-специалисты могут управлять несколькими кластерами Kubernetes с возможностью мультиарендной изоляции контейнеров и преднастроенного постоянного хранения данных.

Сегодня компании понимают, что для осуществления трансформации и поддержания конкурентоспособности в своей отрасли им необходимо быстрее внедрять инновации и обновлять свои приложения. Чтобы ускорить разработку новых приложений и стимулировать развитие цифровых инноваций, они используют контейнеры и Kubernetes для создания собственных облачных приложений в формате микросервисов. Исследования отраслевых аналитиков подтверждают, что компании все больше внедряют контейнеры:

- по оценкам Gartner, к 2022 году более 75% организаций по всему миру будут использовать контейнерные приложения в производстве, по сравнению с менее чем 30% сегодня (*Gartner: «Эволюция виртуализации: виртуальные машины, контейнеры, бессерверная архитектура – что и когда использовать?», 26 сентября 2019 г.*);
- исследования IDC показывают, что 55% крупных предприятий США используют Kubernetes для оркестрации контейнеров (*IDC: Опрос по управлению контейрнерами и облаком, 2019 г.*);
- недавнее исследование 451 Research показывает, что 95% новых приложений будут использовать контейнеры (*451 Research: Голос компании: DevOps, 1 кв. 2019 г.*).

По мере того как корпорации расширяют применение контейнеров и Kubernetes за рамки привычных задач разработки и тестирования, все более актуальными становятся безопасность, возможность управления несколькими кластерами одновременно и балансировка нагрузок. В качестве новых вариантов применимости контейнеров можно назвать граничные вычисления и базы данных. Кроме того, значительная часть корпоративных приложений и систем не является облачной – эти традиционные монолитные приложения затратны в обслуживании, и многие из них только выиграют от контейнеризации. Но реорганизация или преобразование существующих приложений в облачные – процесс трудоемкий и дорогостоящий. Более того, у этих приложений есть требования, которые делают переход на Kubernetes сложным, такие как сохранение корневой файловой системы и миграция.

Контейнерная платформа HPE решает эти проблемы благодаря программному обеспечению BlueData, выступающему в качестве платформы для управления контейнерами, распределенной файловой системе MapR с постоянным хранилищем данных самих контейнеров и Kubernetes для оркестрации. Такой подход позволяет использовать контейнеры не только для облачных приложений в формате микросервисов, но и для контейниризации монолитных корпоративных приложений с постоянным хранилищем данных.

Ключевые преимущества HPE Container Platform включают в себя:

- модернизацию монолитных приложений за счет применения современных облачных подходов без изменения их архитектуры;
- возможность создавать приложения единожды и запускать их где угодно – в локальной инфраструктуре, публичных облаках или на граничных вычислителях;
- повышение производительности разработчиков и ускоренный выпуск новых версий кода с возможностью упрощенного развертывания через Kubernetes и управления несколькими кластерами одновременно;
- безопасность, производительность и надежность корпоративного класса с меньшими затратами за счет контейнеров, работающих на «голом» железе и с постоянным хранилищем данных.

Это новое решение дополняет существующие услуги HPE по сопровождению клиентов при реализации стратегий контейниризации, модернизации приложений и перехода на гибридные облака. HPE Pointnext предоставляет консультационные услуги, основанные на опыте реализации более чем одной тысячи проектов с гибридным облаками, а также на передовых практиках, полученных вместе с приобретением компаний Cloud Technology Partners и RedPixie.

Программное обеспечение HPE Container Platform будет доступно для заказа в начале 2020 года.