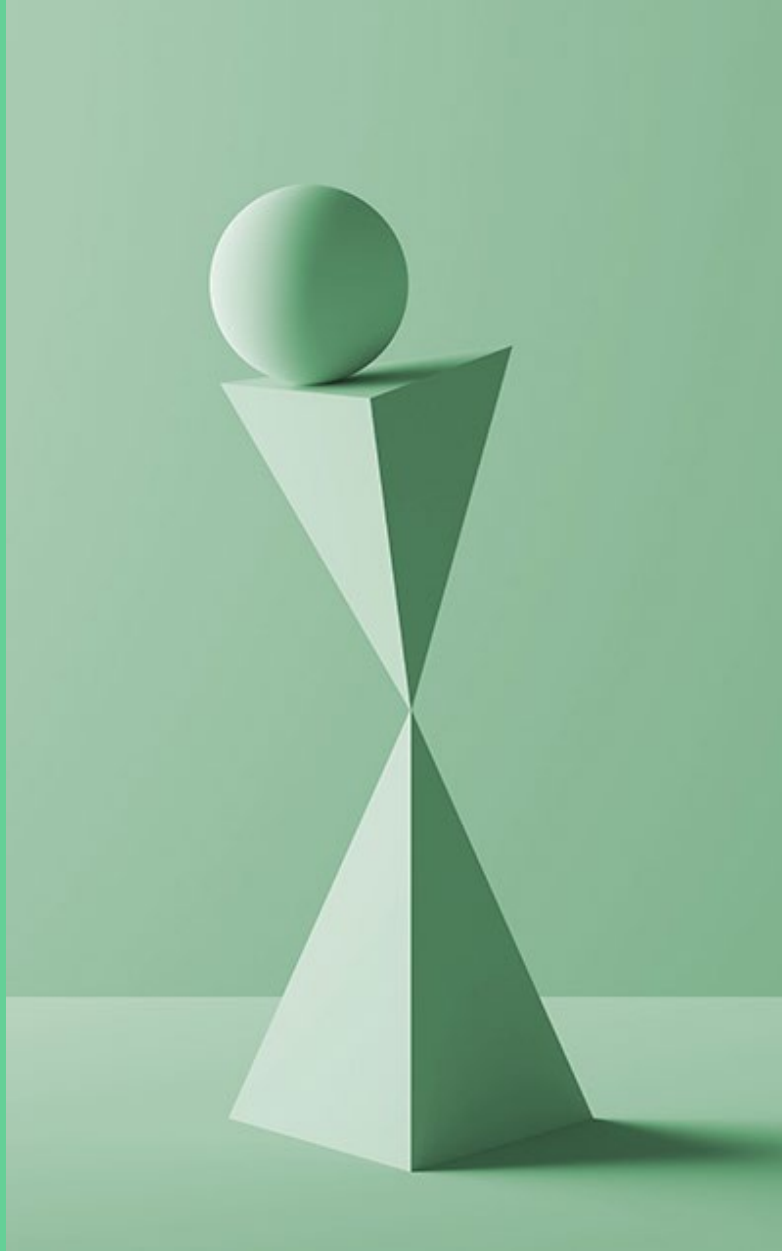
ОПИСАНИЕ РЕШЕНИЯ

NetApp ONTAP AI

Упрощение и ускорение работы, а также интеграция конвейера данных для процессов глубокого обучения с помощью решений NetApp и NVIDIA



NetApp | NVIDIA

Сложности ИИ-инфраструктуры
Искусственный интеллект (ИИ) и глубокое обучение позволяют предприятиям распознавать попытки мошенничества, повышать качество взаимоотношений с заказчиками, оптимизировать цепочки поставок, а также предлагать инновационные продукты и сервисы на рынке в условиях непрерывно растущей конкуренции. Возможно, вы, как и многие другие, уже используете новые подходы, основанные на процессах глубокого обучения, чтобы ускорить переход на цифровые технологии и повысить свои конкурентные преимущества. Однако, чтобы получить максимальные преимущества от глубокого обучения, сначала необходимо решить несколько ключевых задач.

Варианты самостоятельной интеграции довольно сложны. Система для работы с глубоким обучением, состоящая из типовых компонентов, которые обеспечивают вычисления, хранение данных, сетевое и программное взаимодействие будет обладать сложной архитектурой и потребует дополнительного времени на развертывание. В результате при интеграции всех этих компонентов теряется польза ценных ресурсов анализа данных.

Обеспечение предсказуемой и масштабируемой производительности также представляет серьезную проблему. Передовые практики в сфере глубокого обучения свидетельствуют, что организации начинают с малого и постепенно масштабируются. Традиционно для ИИ использовались вычислительные системы с непосредственно подключаемой СХД. Однако, масштабирование может вызывать перебои и простои в работе традиционной СХД.

Сбои в работе влияют на эффективность работы специалистов по обработке данных. Инфраструктура ГО включает многочисленные аппаратные и программные взаимосвязи, поэтому для поддержания такой инфраструктуры в рабочем состоянии требуются глубокие и комплексные знания в области ИИ. Простои и низкая производительность ИИ могут вызвать цепную реакцию, которая снижает производительность разработчиков и непредсказуемо повышающую операционные расходы.

Решение
Сейчас вы можете полностью реализовать потенциал ИИ и глубокого обучения, упрощая и ускоряя работу и интегрируя конвейер данных с апробированной архитектурой NetApp® ONTAP® AI на базе систем NVIDIA DGX™ и флеш-СХД NetApp с возможностью подключения к облаку. Фабрика данных, охватывающая пространство от периферии до ЦОД и облака, поможет надежно оптимизировать обработку данных, ускорить анализ, обучение и логический вывод.

Основные преимущества

Снижение рисков с помощью гибких проверенных решений
- Ускоренное развертывание за счет простой архитектуры, не требующей предварительной отладки.
- Оптимизация настройки и развертывания с помощью доступных предварительно настроенных решений.

Обеспечение необходимой производительности и масштабируемости
- Быстрое начало и беспроблемное расширение.
- Ускорение работы посредством высокопроизводительных технологий.

Интеграция конвейера данных
- Автоматическое управление данными от периферии до ЦОД и облака с помощью интегрированного конвейера данных..
- Развертывание решения, опираясь на опыт специалистов в области искусственного интеллекта и удобные варианты поддержки.

Унификация рабочих нагрузок ИИ
- Устранение разрозненности инфраструктуры данных.
- Гибкий отклик на потребности бизнеса.

NetApp ONTAP AI — один из первых стеков конвергентной инфраструктуры, в котором используется первая в мире ИИ-система с производительностью 5 петафлопс NVIDIA DGX A100, и высокопроизводительные Ethernet-коммутаторы NVIDIA Mellanox. Воспользуйтесь унифицированными рабочими нагрузками ИИ, упрощенным развертыванием и быстрой окупаемостью инвестиций.

«Глубокое обучение в корне меняет практически каждый сегмент, в котором мы работаем. Мы применяем эти технологии на различных рынках, расширяя границы возможного. NetApp ONTAP AI на системах NVIDIA DGX и СХД NetApp all-flash упрощает и ускоряет работу конвейера данных в процессах глубокого обучения».

Тим Энсор, директор по технологиям искусственного интеллекта Cambridge Consultants

Рисунок 1. Архитектуры ONTAP AI на основе систем DGX A100, конфигурации с 2, 4 и 8 узлами.

Снижение рисков с помощью гибких проверенных решений

Технология искусственного интеллекта эволюционирует довольно быстро, поэтому разработка эффективной ИИ-инфраструктуры также является сложной задачей. ONTAP AI позволяет отказаться от догадок и быстрее начинать работу благодаря проверенной на практике эталонной архитектуре. Или упростить проектирование и управление, используя предварительно сконфигурированному и быстро разворачиваемому решению.

Интегрированное решение ONTAP AI доступно в трех предварительно сконфигурированных вариантах с увеличенной емкостью и дополнительным расширенным программным обеспечением. Управление этим интегрированным решением еще более упрощено благоларя «единому окну» сервиса, начиная с инсталляции и заканчивая выяаление и устранение неполадок.

Обеспечение необходимой производительности и масштабируемости

Обучающие процессы для глубокого обучения требуют высокой вычислительной мощности. Ускорение обучения распознаванию изображений поможет сократить общие расходы на вычисления, одновременно ускоряя внедрение технологий искусственного интеллекта и повышая производительность.

Система DGX A100, созданная с использованием новой архитектуры NVIDIA Ampere, повышает производительность обучения почти в 6 раз по сравнению с предыдущим поколением. Вы получаете эквивалент вычислительной инфраструктуры ЦОД для аналитики, обучения и вывода в эксплуатацию, объединенной в единую систему. По сравнению с системами на основе центральных процессоров, DGX A100 занимает в 25 раз меньше места, потребляет в 20 раз меньше энергии и при этом дешевле в 10 раз.

Инвестирование в передовые вычислительные технологии требует ультрасовременной СХД, которая может обрабатывать тысячи обучающих образов в секунду. Этого можно добиться с помощью высокопроизводительных сервисов обработки данных, способных обеспечивать самые высокие рабочие нагрузки, связанные с глубоким обучением.

С помощью СХД NetApp all-flash вы можете добиться производительности вычислений выше 2 ГБ/с (пиковое значение 5 ГБ/с) при задержке менее 1 мс, и использовании графических ускорителей (GPU) более чем на 95 %. Одна система NetApp All Flash A800 поддерживает пропускную способность 25 ГБ/с при последовательном чтении и 1 млн операций «ввод-вывод» в секунду для небольших задач произвольного чтения при латентности менее 500 микросекунд для рабочих нагрузок NAS.

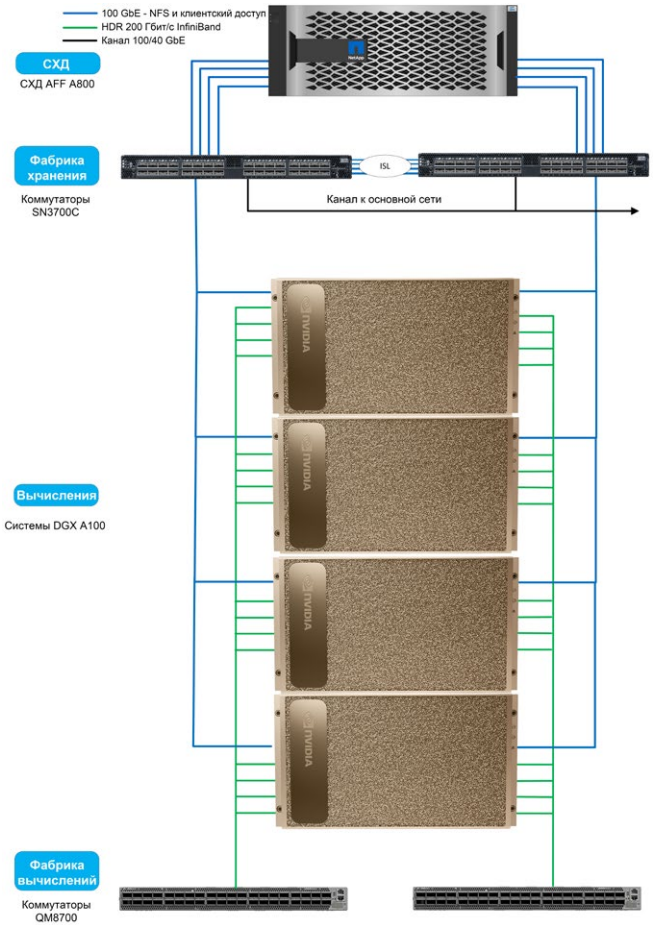


Рисунок 2. Конфигурация ONTAP AI с 4 узлами с коммутаторами 100GbE Mellanox Spectrum.

Стоечная архитектура NetApp позволяет масштабировать системы от десятков терабайт до десятков петабайт с помощью хранилища all-flash. Благодаря NetApp ONTAP FlexGroup одна область имен объемом до 20 ПБ способна обрабатывать более 400 млрд файлов.

Создайте и интегрируйте конвейер данных от периферии до ЦОД и облака

ONTAP AI с помощью фабрики данных NetApp унифицирует управление по всему конвейеру данных на единой платформе. Используйте одни и те же средства для надежного контроля и защиты данных во время переноса, использования или хранения, уверенно соблюдая все нормативные требования. В случае возникновения проблем в среде глубокого обучения вы всегда можете положиться на нашу апробированную модель поддержки, чтобы провести диагностики неисправностей и получить рекомендации по их устранению.

Унификация рабочих нагрузок ИИ

Теперь вы можете устранить разрозненность в инфраструктуре данных, которая вызывает недоиспользование или, наоборот, нехватку ресурсов для рабочих нагрузок ИИ. ONTAP AI — универсальное решение для ИИ-инфраструктуры на основе систем DGX A100, предназначенное для консолидации аналитики, обучения и логических выводов на единой платформе, которая гибко реагирует на потребности бизнеса. Вы также получаете преимущество в виде меньшей совокупной стоимости владения, чем у предыдущих архитектур.

NetApp и NVIDIA: внедряем инновации вместе

Ядром решения ONTAP AI является DGX A100 — универсальный компонент для создания ИИ ЦОД, обеспечивающий поддержку обучения DL, получение информации, анализ и обработку данных и выполнение других высокопроизводительных рабочих нагрузок в рамках одной платформы. Каждая система DGX A100 оснащена восемью графическими процессорами NVIDIA A100 с тензорными ядрами и двумя процессорами AMD EPYC™ 2-го поколения, которые включают новейшие высокоскоростные межкомпонентные адаптеры NVIDIA Mellanox ConnectX-6 100/200 Гбит с поддержкой Ethernet и InfiniBand.

Чтобы ускорить исполнение нескольких небольших рабочих нагрузок, в системе DGX A100 можно выделить до 56 экземпляров на систему с помощью новой технологии многоэкземплярных GPU. Такое ускорение позволяет вам очень эффективно распределять мощности графического процессора в ONTAP AI. Ваши специалисты по обработке и анализу данных смогут быстрее проводить итерации, автоматизировать воспроизводимость и внедрять ИИ-проекты быстрее, сокращая реализацию на срок до 3 месяцев, при более высоком качестве.

Системы NetApp AFF передают данные в обучающие процессы посредством самой быстрой и гибкой в отрасли СХД all-flash, которая обеспечивает работу первых в мире сквозных технологий NVMe. AFF A800 может передавать данные в системы DGX до 4 раз быстрее, чем любые конкурирующие решения.[1]

В ONTAP AI используются Ethernet-коммутаторы Mellanox Spectrum, которые обеспечивают низкую латентность, высокую плотность и производительность, а также эффективное использование мощностей, которое необходимо в средах ИИ.

1. Производительность чтения составляет до 300 ГБ/с на кластер all-flash по сравнению с 75 ГБ/с, обеспечиваемыми ведущим конкурентом.

Фабрика данных на основе технологий NetApp обеспечивает лучшие в своем сегменте возможности управления данными и облачной интеграции, которые ускоряют внедрение глубокого обучения, одновременно управляя важными данными и защищая их. ONTAP предлагает беспрецедентное соотношение сокращения объема данных — 22:1, а также снижение TCO на величину до 54% по сравнению с непосредственно подключаемой СХД.

Система DGX A100 основана на программном стеке NVIDIA DGX, который включает оптимизированное программное обеспечение для рабочих нагрузок ИИ и обработки данных. Вы получаете максимальную производительность и более быструю окупаемость инвестиций в ИИ-инфраструктуру.

NetApp AI Control Plane помогает упростить управление ИИ-данными за счет интеграции Kubernetes и Kubeflowc предоставляемой NetApp фабрикой данных, обеспечивая оптимальную доступность и переносимость данных от периферии в ЦОД и облако. Для расширения функций AI Control Plane используется набор инструментов для анализа данных Data Science Toolkit. Это библиотека Python, которая упрощает выполнение многочисленных задач по управлению данными для инженеров и для специалистов по обработке данных. Например, они могут выделить новый том данных, мгновенно клонировать его и создать снимок тома данных NetApp™, чтобы отслеживать и определять базовые значения.

Использование правильных инструментов — это залог успеха. Именно поэтому ONTAP AI пользуется признанием ведущих разработчиков программного обеспечения для машинного обучения (MLOps), включая Domino Data Lab, Iguazio и другие. Ваши сотрудники смогут использовать знакомые инструменты, чтобы получить максимальную отдачу от ИИ-среды и ускорить получение аналитических сведений.

Компоненты решения

- Системы NVIDIA DGX A100
- СХД NetApp AFF A-Series с ONTAP 9
- NVIDIA Mellanox Spectrum SN3700C, NVIDIA Mellanox Quantum QM8700 и/или NVIDIA Mellanox Spectrum SN3700-V
- Программный стек NVIDIA DGX
- NetApp AI Control Plane
- Набор инструментов для анализа данных Data Science Toolkit

Эталонные архитектуры

Компания NetApp выпустила следующие эталонные архитектуры на основе ONTAP AI, предназначенные для использования в различных сценариях в определенных отраслях:

- Эталонная архитектура ONTAP AI для здравоохранения: диагностическая визуализация
- Эталонная архитектура ONTAP AI для рабочих нагрузок беспилотных автомобилей: проектное решение
- Эталонная архитектура ONTAP AI для рабочих нагрузок в сфере финансовых услуг: проектное решение

О компании NVIDIA

Изобретение компанией NVIDIA графического процессора в 1999 г. повлекло рост рынка игровых ПК, заставило производителей переосмыслить современную компьютерную графику и произвело революцию в области параллельных вычислений. Недавно глубокое обучение на основе GPU поспособствовало развитию современного искусственного интеллекта — следующей эпохи вычислений — где графический процессор выполняет роль мозга компьютеров, роботов и автомобилей с автономным управлением, наделяя их возможностью воспринимать и понимать окружающий мир.
Более подробную информацию можно найти на сайте www.nvidia.com.

О компании NetApp

В отличие от многих NetApp не является экспертом широкого профиля, но это истинный специалист в своей области. Наша специализация — помочь вам извлечь максимум преимуществ из ваших данных. Мы предлагаем облачные сервисы по хранению данных корпоративного класса, которые не подведут вас в нужный момент, а также инструменты для удобной и гибкой работы с вашими ЦОД в «облачном» формате. Нами передовые решения подойдут для любой пользовательской среды и совместимы с крупнейшими публичными облаками.

Компания NetApp всегда специализировалась на разработке программного обеспечения для работы с облаками и обработки данных. Поэтому именно наши продукты помогут вам создать собственную фабрику данных или упростить управление вашей облачной инфраструктурой и наладить взаимодействие между ее элементами. А если вы поставщик цифровых решений, будь то данные, сервисы или приложения, то мы обеспечим их точную и безопасную поставку клиентам в любых условиях. www.netapp.com/ru

NetApp