White paper

Программный RAID раскрывает возможности NVMe для enterprise-задач

Содержание

Основные положения

Появление протокола NVMe расширило горизонты возможностей для систем хранения данных. Новые накопители не только сохраняют все полезные свойства традиционных SSD, но и демонстрируют совершенно новый уровень производительности и времени отклика.

Благодаря своим характеристикам NVMe-накопители являются отличным решением для многих передовых технологий. Многомерная аналитика, комплексные системы планирования ресурсов предприятия, машинное обучение и автономные производственные цепочки по-прежнему требуют серьезных долгосрочных инвестиций, но теперь их внедрение становится значительно проще и доступнее.

При этом повсеместное распространение NVMe-накопителей имеет ряд объективных ограничений, с которыми сталкиваются все производители СХД и системные интеграторы. Для enterprise-сегмента критически важным является отказоустойчивость массива и использование технологии NVMe-oF для его сетевого подключения к клиентам.

Существующие программные и аппаратные технологии создания отказоустойчивых массивов либо используют чрезмерную избыточность в виде зеркалирования, либо теряют до 65% производительности при работе в RAID 5 или RAID 6. При этом далеко не все из них имеют возможность реализовать функционал сетевого доступа к накопителям через NVMe-oF.

RAIDIX ERA является программным массивом, который разрабатывался с учетом всех особенностей NVMe. Благодаря ряду технологических инноваций и внимательному отношению к возможностям нового протокола, наш программный RAID способен продемонстрировать в RAID 6 до 97% суммарной производительности накопителей, обеспечивая минимальное время задержки даже в условиях смешанных нагрузок.

В этом документе мы рассмотрим основные особенности протокола NVMe, оценим его сильные стороны и недостатки, расскажем о возможностях RAIDIX ERA и сценариях его использования с различными типами прикладных приложений.

Возможности и ограничения протокола NVMe

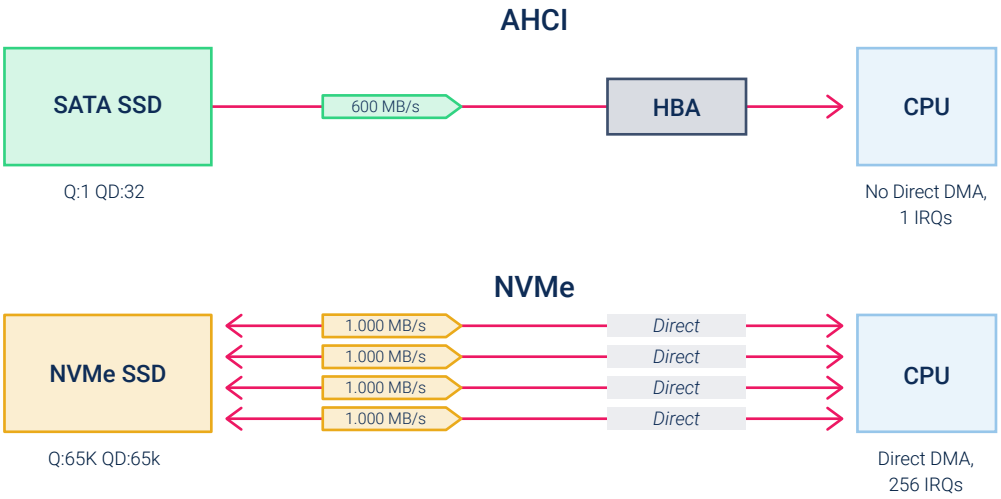
Протокол NVMe (Non-Volatile Memory Express) — это первый протокол, созданный специально для работы с твердотельными накопителями. Традиционные SAS и SATA разрабатывались с учетом сложной механики жесткого диска. Эти протоколы используют большое количество служебных команд, необходимых для корректного управления подвижными элементами и их бесперебойную работу.

Для flash-памяти такие команды обычно сложны и избыточны. SAS и SATA протоколы обращаются к данным на SSD по процедуре работы с жесткими дисками, которая не учитывает физические особенности flash-памяти. Протокол NVMe, наоборот, опирается на эти особенности и использует их для достижения максимальной производительности.

Высокая скорость и низкая задержка

NVMe работает на основе интерфейса PCIe. Протокол использует набор PCIe команд, с помощью которых приложения взаимодействуют с центральным процессором. Благодаря этому между накопителем и CPU нет уровня HBA, который увеличивает data path за счет преобразования SAS/SATA команды в PCIe.

Рисунок 1. Сравнение программных интерфейсов AHCI и NVMe



Помимо этого контроллер NVMe-накопителя поддерживает до 65 тысяч очередей передачи данных в процессор, каждая из которых может содержать более 65 тысяч команд. В SATA/SAS интерфейсах возможна лишь одна очередь с 32 или 254 командами соответственно.

Такие характеристики позволяют NVMe SSD снизить уровень задержки в два раза и в несколько раз увеличить скорость чтения и записи на случайных и смешанных нагрузках по сравнению с традиционными SSD. Благодаря этому NVMe-накопители являются основой для эффективной работы с требовательными бизнес-приложениями и такими технологиями как AI & ML, OLTP, Edge-computing и др.

NVMe-oF для сетевого доступа к данным

NVMe SSD имеет неоспоримые преимущества над SATA/SAS SSD при прямом размещении в сервере на шине PCIe. Но локальное использование накопителей подходит далеко не для всех бизнес-задач: при нем не выполняются минимальные требования по отказоустойчивости, и существуют ограничения по масштабированию и гибкости распределения ресурсов.

Поэтому технологию стоит оценивать также в контексте сетевого использования, где из-за распространения протоколов предыдущего поколения NVMe не может демонстрировать свою эффективность. Для решения этой задачи разработано расширение NVMe-oF (NVMe Over Fabric), которое позволяет использовать NVMe с транспортом Ethernet, Fiber Channel или InfiniBand.

NVMe-oF позволяет приложению на сервере взаимодействовать с внешним NVMe-массивом практически с локальным уровнем задержки. Технология предполагает также сохранение высокой производительности в сети, но на данным этапе это ограничивается возможностями сетевых адаптеров.

Поэтому для полноценного использования протокола NVMe в enterprise-сегменте важнейшим условием является поддержка NVMe-oF другими компонентами существующей инфраструктуры.

Программный RAID для NVMe

Если говорить про использование NVMe за пределами пользовательских устройств, то практически во всех сценариях требуется RAID-массив, обеспечивающий защиту данных при сбое диска. Эту цель можно достигнуть с помощью программных инструментов на уровне ОС или аппаратных RAID-контроллеров, к которым подключаются накопители. При этом, с учетом стоимости накопителей, разумным выбором будет RAID 5 или RAID 6.

Но в таком случае потребуются ресурсы на вычисление контрольных сумм и общая производительность массива будет существенно ниже суммарной производительности накопителей.

1 NVMe	6 NVMe	6 NVMe в RAID 6
800 000 IOps	4 800 000 IOps	до 2 400 000 IOps

В компании «Рэйдикс» мы разработали RAIDIX ERA — собственный программный RAID для NVMe и SSD, в котором расчет четности учитывает основные особенности нового протокола. Это позволяет обеспечить массиву до 97% суммарной производительности NVMe-накопителей.

1 NVMe	6 NVMe	6 NVMe в RAID 6 (RAIDIX ERA)
800 000 IOps	4 800 000 IOps	4 600 000 IOps

Быстрый RAID для Linux

RAIDIX ERA представляет собой модуль ядра Linux, который создает из накопителей массив и предоставляет его клиенту в виде блочного устройства. Поддержка POSIX API обеспечивает ему универсальность использования, устраняя необходимость модифицировать приложения или файловые системы.

RAIDIX ERA использует технологию параллелизации вычислений и lockless-архитектуру для снижения барьеров внутри массива, благодаря чему собранный RAID может демонстрировать скорость работы до 10M IOps и 55 GBps. Массив сохраняет высокую производительность и низкую задержку (< 0.5ms) даже в режиме смешанной нагрузки.

Для обеспечения отказоустойчивости RAIDIX ERA предлагает разные уровни массива: RAID 1/0/5/6/7.3/50/60/70. При этом при отказе накопителя массив показывает минимальную просадку производительности, которая не угрожает качеству работы приложений. Это достигается за счет инновационного подхода к векторным вычислениям в erasure coding.

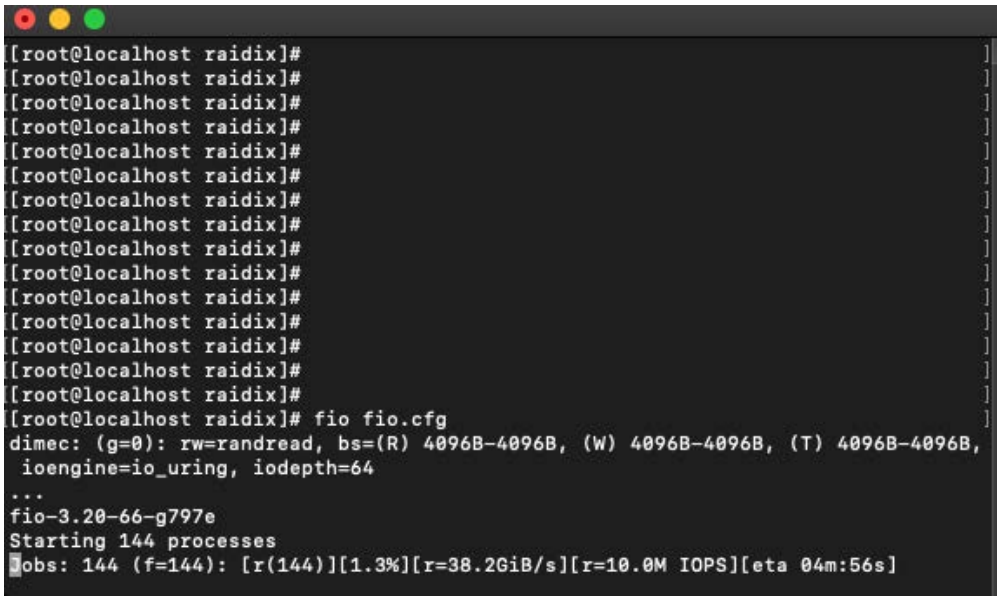


Рисунок 2. Скриншот терминала с результатами тестирования RAIDIX ERA

Сравнение RAIDIX ERA с другими решениями

Мы разрабатывали RAIDIX ERA с четким пониманием того, что продукт должен быть не только производительным, но и удобным для использования. Именно поэтому он свободно работает практически на любом серверном оборудовании, не привязан к какому-то определенному бренду накопителей и легко встраивается в состав уже существующего программно-аппаратного решения.

Также в RAIDIX ERA решена главная проблема программных массивов — чрезмерное потребление вычислительных ресурсов системы. При максимальной производительности массива нагрузка на CPU не превышает 20%, а для эффективной работы требуется менее 4 GB оперативной памяти. Также для более эффективного использования системы в RAIDIX ERA предусмотрена возможность выставления процента потребления RAM.

	Аппаратный RAID-контроллер	Программный RAID	**RAIDIX ERA**
Производительность	Средняя	Средняя	**Высокая**
Приобретение оборудования	Да	Нет	**Нет**
Физический износ и моральное устаревание	Да	Нет	**Нет**
Неограниченное кол-во подключаемых накопителей	Нет	Да	**Да**
Поддержка NVMe-oF	Нет	Да	**Да**
Использование ресурсов RAM	—	Высокое	**Низкое**
Использование ресурсов CPU	—	Высокое	**Низкое**
Настройка приоритета использования вычислительных ресурсов	—	Нет	**Да**
Совместимость с различными серверными платформами	Нет	Да	**Да**
Совместимость с накопителями любых производителей	Нет	Да	**Да**
Легкая интеграция в решение другого вендора	Нет	Опционально	**Да**

Использование RAIDIX ERA для enterprise-приложений

RAIDIX ERA позволяет создавать из NVMe-накопителей быстрый массив с высокими показателями отказоустойчивости. Он хорошо подходит для работы с требовательными enterprise-приложениями, эффективность которых напрямую зависит от времени задержки и пропускной способности back-end инфраструктуры.

Базы данных для аналитики и исследований

Анализ больших массивов данных является ключевым аспектом эффективного управления в крупных современных предприятиях и при проведении масштабных фундаментальных исследований.

Большинство аналитических приложений для таких задач имеют характерный паттерн нагрузки на систему хранения: небольшие и регулярные запросы на запись и постоянные запросы на чтение больших блоков данных. За счет высокой пропускной способности (до 55 ГБ/с) RAIDIX ERA позволяет базам данных быстрее передавать аналитическим приложениям запрашиваемые массивы данных. Это повышает итоговую эффективность всего процесса и сокращает время получения аналитических результатов.

Обработка транзакций в режиме реального времени (OLTP)

OLTP-приложения являются одним из ключевых технологических решений в финансовом секторе, автоматизированных производственных цепочках крупных предприятий, системах учета и планирования ресурсов. С точки зрения инфраструктуры, для эффективной работы таких приложений требуется низкое время отклика и высокая скорость обработки множества мелких запросов.

NVMe-массив на базе RAIDIX ERA в качестве back-end устройства обеспечивает минимальное время отклика для транзакционных запросов и позволяет приложению выполнять большее количество операций в секунду. Помимо этого, он сохраняет высокую производительность даже в условиях интенсивных смешанных нагрузок.

Производство видео в разрешении 8K

Распространение разрешения 8K, технология HDR и повышение частоты кадров до 60 fps серьезно увеличили требования к технологическому оснащению пост-продакшн студий. Хранилище видеоматериала является в нем узким местом, от пропускной способности которого зависит не только комфорт и скорость работы команды, но и защита от возможной потери кадров при монтаже.

RAIDIX ERA используется в решениях для видеопроизводства в качестве подсистемы хранения, обеспечивая высокую скорость чтения и записи данных для одновременной работы с нескольких монтажных станций. Использование RAIDIX ERA позволяет устранить какие-либо задержки при совместном редактировании материала и предотвратить возможную потерю кадров при его обработке.

Edge-компьютинг в IoT

Периферийные вычисления или Edge-компьютинг представляет собой один из вариантов организации интернета вещей (IoT, Internet of Things). В нем вычислительные ресурсы распределенной системы размещаются в непосредственной близости от датчиков и сенсоров действующего объекта. Такая технология используется в беспилотных автомобилях и позволяет искусственному интеллекту получать сведения о дорожной ситуации с минимальной задержкой.

RAIDIX ERA применяется в беспилотных автомобилях в качестве компонента вычислительной платформы. ПО управляет NVMe-массивом и позволяет демонстрировать минимальный уровень задержки и высокую производительность даже с ограниченного количеств накопителей.

Заключение

Протокол NVMe работает на основе шины PCIe, используя более подходящий для flash-накопителей набор команд. Это дает возможность передавать данные через 65 тысяч параллельных потоков и избегать дополнительных барьеров при работе с запросами приложения.

Чтобы сохранить максимум производительности NVMe-накопителей в отказоустойчивом массиве, мы разработали RAIDIX ERA — программный RAID, который демонстрирует высокую скорость работы и низкую задержку даже в условиях интенсивной смешанной нагрузки. Продукт легко встраивается в программную среду существующих систем и способен работать с широким перечнем серверным платформ и накопителей.

Применение RAIDIX ERA в enterprise-сегменте открывает бизнесу доступ к преимуществам передовых технологий без лишних затрат на избыточную инфраструктуру. Для производителей СХД и системных интеграторов RAIDIX ERA является простым и доступным способом получить технологическое преимущество при создании производительных NVMe-решений.

www.raidix.ru
request@raidix.ru
+7 (812) 622 16 80