Флеш-массивы EMC

В течение последних месяцев (начиная с сентября 2013 г.) стали доступны новые решения на базе флеш-технологий от EMC. Среди них: второе поколение решений для серверного кэширования на базе PCIe флеш-карт, а также новые флеш-массивы (анонсированные еще в 2012 г.). Первое поколение массивов в продаже с конца ноября 2013 г., второе – емкостью до 80 Тбайт – будет доступно в первой половине 2014 г.

Тимофей Григорьев – консультант по технологиям, Уральский регион, EMC Россия.

Введение

Первая волна использования флеш-памяти началась в 2007 г. с появлением SSD-накопителей (Solid State Disks). Корпорация EMC первая анонсировала доступность с 1 кв. 2008 г. для системы хранения уровня enterprise – EMC Symmetrix® DMX-4 – модулей расширения, построенных на твердотельных дисках, параллельно с традиционными. Введение SSD-накопителей в архитектуру СХД в 2008 г. позволяло поднять производительность в 300 раз в сравнении с 15K HDD. В настоящее время SSD-накопители используются EMC во всех семействах СХД: EMC VMAX, VMAXe, VNX, VNXe и Isilon.

Второй прилив флеш-технологий был ознаменован выходом на рынок PCIe-флеш-карт в 2011 г. В феврале 2012 г. EMC анонсировала решение VFCache, которое осуществляло управление кэшированием данных в составе сервера.

VFCache представляло собой PCIe-карту с установленной флеш-памятью на основе NAND-технологии (34 нм SLC) – продукт от третьих фирм. Решение VFCache дополняло аппаратные и программные решения EMC в области уровневого хранения данных и вводило самый высокий уровень кэширования данных в семействе решений EMC (уровни строились на SSD-дисках, FC/SAS-дисках и SATA/NL-SAS-дисках). Введение уровня хранения на основе VFCache (шина PCIe Gen2, x8) позволило еще на порядок увеличить производительность ввода/вывода – до 4000 раз в сравнении с 15K HDD. В этом решении кэширование данных осуществлялось только на операциях чтения. Кэширование записи осуществлялось на уровне СХД.

Одновременно с решением VFCache EMC также анонсировала "Project Thunder" для оптимизации высокоинтенсивной read/write нагрузки, требующей низкой задержки. В рамках проекта Thunder предполагалась разработка специализированного сервера, масштабируемого PCIe-картами EMC VFCache. Управление процедурами перемещения данных планировалось на базе ПО EMC FAST (Fully Automated Storage Tiering), которое должно было обеспечивать автоматическое перемещение "горячих" данных на PCIe флеш-память/SSD-накопители, а менее используемых данных – на SAS/FC и NL-SAS/SATA-диски, одновременно улучшая производительность приложений и снижая стоимость хранения.

По прошествии одного года проект Thunder успешно трансформировался в Project X – комбинацию решений из XtremIO, XtremSF и XtremCache, где XtremIO – кластерная СХД, построенная на высокофункциональных двухконтроллерных модулях со встроенной DRAM-памятью 512 Гбайт и 25 SSD-дисками (с общей неформатированной емкостью 10 Тбайт). Для справки: PCIe флеш-карты позволяют достичь задержек на операциях ввода-вывода менее 100 мкс, соответственно, флеш-массивы – около 500 мкс. Однако, отказавшись от построения внешнего массива на PCIe флеш-картах, EMC существенно расширила функциональность своего SSD-массива (в сравнении с существующими на рынке), сохранив при этом возможность построения решений с ускорителями на PCIe флеш-картах.

Все три вышеназванных решения были анонсированы в марте 2013 г.:

- *EMC XtremSF: серверное флеш-оборудование PCIe;*
- *EMC XtremSW Suite: пакет программного обеспечения для флеш-технологий следующего поколения;*
- *EMC XtremSW Cache: программное обеспечение серверного кэширования на основе флеш-памяти;*
- *EMC XtremIO: массив полностью на основе флеш-дисков.*

В рамках данного объявления был осуществлен ребрендинг старых продуктов:

- *EMC VFStore — новое название: XtremSF;*
- *EMC VFSoftware — новое название: EMC XtremSW Suite;*
- *EMC VFCache — новое название: EMC XtremSW Cache (в составе XtremSW Suite);*
- *Project X — новое название: EMC XtremIO.*

Контроллер XtremSF

Решение XtremSF (прежнее название – VFStore) не претерпело значительных изменений с момента выхода на рынок в 2012 г. Карта XtremSF с половинной высотой и половинной длиной представляет собой отдельную карту низкого профиля, которую можно установить в любом стоечном сервере, используя мощность одного слота PCIe. Карта XtremSF обеспечивает значительное повышение производительности приложений за счет сокращения задержек и повышения пропускной способности без ухудшения характеристик в течение всего срока службы этого устройства даже при его заполнении.

XtremSF выпускается в шести конфигурациях: с флеш-накопителями eMLC (многоуровневые ячейки корпоративного класса) и SLC (одноуровневые ячейки) емкостью от 350 Гбайт до 2,2 Тбайт (рис. 1). XtremSF можно использовать в двух ипостасях:

- как устройство для локального хранения данных для повышения производительности чтения и записи;
- для кэширования XtremSW Cache (прежнее название — EMC VFCache) в сочетании с серверным программным обеспечением для повышения производительности чтения с защитой данных.

Для максимального увеличения срока службы флеш-накопителя XtremSF производит динамическое глобальное и локальное выравнивание износа. При необходимости XtremSF перемещает данные в менее всего использовавшиеся зоны флеш-накопителя. Более того, сложные алгоритмы планирования позволяют выполнять задачи управления флеш-накопителем в то время, когда они не будут влиять на производительность приложений. Флеш-карта XtremSF специально разработана для минимизации дополнительной нагрузки на ЦП за счет выгрузки этих операций управления флеш-накопителем с серверного ЦП на карту PCIe. Кроме того, карта XtremSF обеспечивает бесперебойную работу в случае ошибок в самом флеш-накопителе. Благодаря использованию массива с резервированием и независимой технологией NAND

Рис. 1. Конфигурации XtremSF.

(RAIN), флеш-компоненты на карте можно выделить как отдельные секторы и распределить по группам RAIN. Это обеспечивает возможность бесперебойного доступа приложения к данным даже при нескольких ошибках. Когда устройство автоматически обнаруживает сбой одного элемента хранения, выполняется полное прозрачное восстановление данных. Восстанавливается избыточность данных без негативных последствий для любых существующих пользовательских данных или производительности дисков.

Используя схему контроля по четности RAID, карта XtremSF обеспечивает защиту целостности данных на уровне флеш-памяти. Однако флеш-карта XtremSF не предоставляет сервисы защиты критически важных данных. Для защиты данных следует использовать ПО XtremSW Cache или другие программные средства.

EMC провела тестирование возможностей XtremSW Cache. При этом архитектура решения была следующая: Microsoft SQL Server 2008 R2, стоечный сервер Cisco UCS C-460/M1, система EMC VNX5300. При тестировании использовалась стандартная рабочая нагрузка OLTP (аналогичная условиям TPC-E) с базой данных 750 Гбайт и соотношением операций чтения/записи 90:10%. ПО XtremSW Cache было включено на всех логических устройствах хранения данных, кроме логических устройств для ведения журналов. Настройка базы данных SQL не выполнялась. Сведения о конфигурации: кэш-память чтения на VNX5300 = 700 Мбайт, кэш-память записи на VNX5300 = 2000 Мбайт; 20 логических устройств хранения данных емкостью 260 Гбайт на 100 дисках SAS емкостью 300 Гбайт, 10 000 об/мин (RAID 5), 1 логическое устройство для ведения журналов емкостью 500 Гбайт на 4 дисках SAS емкостью 300 Гбайт, 10 000 об/мин (RAID 10); буферная кэш-память базы данных 10 Гбайт; сервер, подключенный к системе хранения VNX с помощью 4 портов Fibre Channel 8 Гбит/с. Тестирование показало: производительность приложений возросла в 3, 6 раза; время оклика приложений улучшилось на 87%.

ПО XtremSW Cache

XtremSW Cache входит в состав более широкого пакета программного обеспечения EMC для флеш-технологий XtremSW Suite, который использует оборудование XtremSF. XtremSW Suite представляет собой программное обеспечение EMC для флеш-технологий нового поколения, которое создается на основе продукта XtremSW Cache, обеспечивающего функциональность кэширования для флеш-устройств PCIe. В этот пакет входят расширенные службы данных, работающие с флеш-накопителями как с памятью и как с системой хранения данных, подсоединенной непосредственно к серверу. Первый выпуск пакета XtremSW Suite предоставляет заказчикам возможность использования пулов, обеспечивает согласованность кэш-памяти, интеграцию с массивом хранения EMC и специальные улучшения для сред VMware.

XtremSW Cache — это единственное решение на рынке серверов с флеш-технологией, которое обеспечивает дедупликацию данных на картах PCIe. Дедупликация кэшируемых данных обеспечивает следующие преимущества:

– *снижение стоимости в расчете на гигабайт*. На карте с функцией дедупликации можно хранить больше данных, поскольку на ней не хранятся дублирующие копии одинаковых данных;

– *увеличение ожидаемого срока службы карт*. Поскольку на флеш-карту не записываются дублирующие данные, снижается количество операций записи на карту и, соответственно, снижается износ карты.

При внедрении баз данных Oracle дедупликация данных в кэш-памяти не приводит к значимому улучшению. Поэтому использовать дедупликацию в таких средах не рекомендуется.

В начале сентября 2013 г. стала доступна новая версия XtremSW Cache 2.0, которая обеспечивает:

– совместимость с функциями VMware vCenter High Availability (HA) и Distributed Resource Scheduler (DRS). XtremSW Cache теперь поддерживает автоматические (встроенные) функции vMotion в среде VMware. Локальная флеш-память используется кластером в виде распределенного ресурса, поэтому функции VMware vCenter DRS и HA будут работать естественным образом;

– совместимость с IBM AIX OS. Теперь XtremSW Cache поддерживается ОС IBM AIX, в частности IBM AIX 6.1 и 7.1. на серверах Power 7. Реализована поддержка стандартной версии Power VM и встроенной кластеризации через PowerHA. Сертифицированные SSD AIX поддерживаются в качестве базового аппаратного ресурса для ПО XtremSW Cache, которое теперь не привязано к определенному типу СХД;

– совместимость с любым серверным флеш-оборудованием, включая Fusion-io. В дополнение к XtremSF, XtremSW Cache теперь можно использовать с любым серверным флеш-оборудованием – SSD, картами PCIe и даже Fusion-io. Это означает, что заказчики могут получить все преимущества использования XtremSW Cache в своих блейд-серверах. Пользователи, которым требуется более экономичное решение, могут использовать SSD (актуальный список поддерживаемого оборудования можно найти на сайте E-Lab Interoperability Navigator);

– совместимость с Oracle RAC, которая позволяет использовать СХД общего пользования типа “активный-активный” в среде Oracle, используя алгоритм согласования распределенной кэш-памяти. Это обеспечивает одноранговую связь между узлами. XtremSW Cache поддерживает Oracle Database 11g для Microsoft Windows, Red Hat Enterprise Linux (RHEL), или Oracle Enterprise Linux (OEL) с использованием Oracle Clusterware 11g с подключением через Ethernet, до восьми узлов в кластере;

– более тесную интеграцию с массивами EMC. Используя VMAX, можно напрямую отслеживать состояние XtremSW Cache через Unisphere, которая выдает рекомендации по кэшированию логических устройств на основе анализа тенденций и предоставляет отчет о производительности. При использовании VMAX включаются дополнительные функции, такие как предварительная выборка, чтение всех дорожек, координация кэша или оптимизированная обработка промахов. Чтение всех дорожек позволяет более эффективно готовить кэш-память и отслеживать узкие места, что дает возможность увеличить скорость операций ввода-вывода до 25%. Оптимизированная обработка промахов перемещает уровень кэша на хост, освобождая ресурсы для дальнейшего использования, и увеличивает количество операций ввода-вывода в секунду до 2,5 раз;

– контроль и эффективность с помощью XtremSW Management Center. Данными в XtremSW Cache можно управлять через новую систему XtremSW Management Center, которая обеспечивает контроль и эффективность при развертывании нескольких экземпляров XtremSW Cache с единой консоли, содержащей данные о состоянии и производительности. С появлением XtremSW Management Center, XtremSW Cache может развертываться через Unisphere Remote, подключаемый модуль VSI, традиционный интерфейс или интерфейс командной строки. XtremSW Management Center поддерживает REST API, что позволяет администратору или сторонним разработчикам взаимодействовать с ним.

XtremIO: горизонтально-масштабируемый массив на твердотельных дисках

Выход массивов XtremIO первого поколения (до 40 Тбайт физической емкости и более 250 Тбайт логической емкости) состоялся в конце ноября 2013 г. Второе поколение массивов XtremIO (до 80 Тбайт физической емкости и с дополнительными функциями “тонкого” выделения, дедупликацией “на лету” и технологией защиты данных XDP) будет доступно в первой половине января 2014 г.

Современные внешние массивы на базе флеш-технологий, в основном, делятся на два класса: полностью построенные на PCIe флеш-картах и построенные на SSD-дисках. Особенностью первых является более высокая производительность (превышение до 10 раз), но и более высокая цена за гигабайт. Вторые имеют более высокую общую емкость (что позволяет их использовать для постоянного хранения активных данных) и, соответственно, меньшую стоимость за гигабайт. Массивы XtremIO относятся ко второму классу и отличаются высокой функциональностью (предоставляя тонкое выделение ресурсов, онлайн дедупликацию, создание мгновенных снимков и др.), что заметно выделяет их на рынке. Помимо этого, двухконтроллерная пара в базовом блоке имеет общую DRAM-память емкостью 512 Гбайт и 32 процессорных ядра (до 128 ядер в кластере), что позволяет “на лету” выполнять не только очень ресурсоемкие операции, например, дедупликацию, но и

поддерживать высокую производительность ввода/вывода – до 600 000 IOPs при смешанной нагрузке и до 1 000 000 IOPs при чтении.

EMC® XtremIO™ — массив на флеш-дисках, который обеспечивает стабильно высокую прогнозируемую производительность для любых рабочих нагрузок за любой период времени независимо от статуса и заполнения массива. Этого удалось достичь благодаря использованию в EMC XtremIO нескольких уникальных инноваций в области флеш-технологий:

– горизонтально масштабируемой многоконтроллерной архитектуре с линейной масштабируемостью;

– встроенной и постоянно действующей дедупликации и защите данных, превосходящей в 6 раз по эффективности и в 4 раза по производительности традиционные RAID.

EMC XtremIO отличается по архитектуре от любых других флеш-массивов. Чтобы добиться максимальной производительности без ущерба для эффективности и долговечности используется согласованная работа четырех технологий:

– *размещение данных на основе содержания,* что позволяет естественным образом обеспечить максимальную производительность (при минимальном износе флеш-дисков) и отсутствие горячих точек с высоким потоком данных или метаданных;

– *двухэтапный механизм обработки метаданных*;

– *алгоритм защиты данных XtremIO Data Protection (XDP)*;

– *совместно используемые метаданные (shared in-memory metadata)*.

Архитектура XtremIO

Системы XtremIO — это горизонтально масштабируемые кластеры, состоящие из 1–4 модулей, называемых X-Brick (рис. 2). Каждый модуль X-Brick (высота 5U) — это массив хранения с высоким уровнем доступности для высокопроизводительной сети хранения данных, который состоит из следующих компонентов (рис. 3):

– два контроллера системы хранения данных с резервированием в режиме “активный–активный”;

– одна 2,5-дюймовая оперативно заменяемая полка для дисков с 25 твердотельными дисками eMLC;

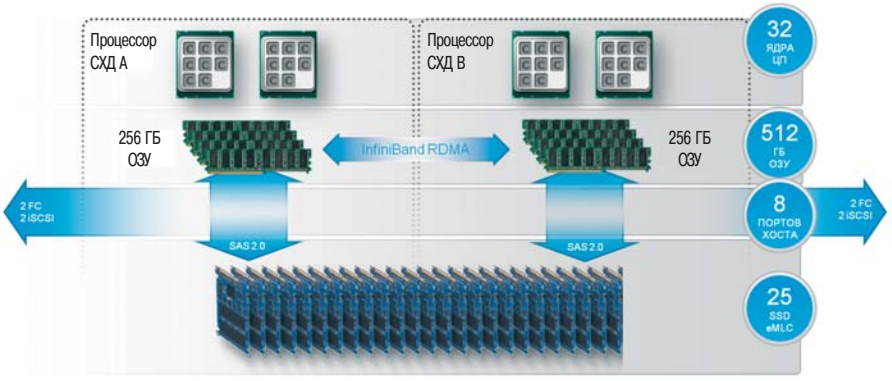

Рис. 3. Каждый модуль X-Brick состоит из двух контроллеров, содержащих 32 ядра и 512 Гбайт общей DRAM-памяти, с подсоединенной по SAS 2.0 дисковой полкой, которая включает 25 твердотельных eMLC-дисков.

– четыре порта Fibre Channel 8 Гбит/с и четыре порта iSCSI 10 Гбит/с;

– резервные компоненты (например, блоки питания, модули ввода-вывода SAS, кабели и т. д.);

– два резервных аккумулятора для защиты кэшированных данных в случае сбоя электропитания;

– две резервные фабрики InfiniBand (IB) 40 Гбит/с (QDR) с сетевыми коммутаторами фабрики RDMA (для систем XtremIO с двумя или более модулями X-Brick).

Для каждого модуля X-Brick предусмотрена емкость 10 Тбайт и 20 Тбайт. Один кластер XtremIO масштабируется с двух до восьми контроллеров и до 128 ядер и может работать с любой базой данных OLTP, виртуальным сервером и инфраструктурой VDI со всеми активными сервисами данных.

В случае кластеров, состоящих из двух и более модулей X-Brick, в системе хранения XtremIO используется высокодоступная внутренняя сеть InfiniBand (40 Гбайт/с, QDR) с резервированием и очень малым временем отклика. Сеть InfiniBand — это полностью управляемый компонент массива XtremIO. Для ее использования администраторам систем XtremIO не нужно обладать специальными знаниями по технологии InfiniBand.

Массив XtremIO работает по тому же принципу, что и любой другой блочный массив хранения. Он интегрируется с существующими сетями хранения данных и поддерживает подключение к хостам по протоколам Fibre Channel 8 Гбит/с или Ethernet iSCSI 10Гбит/с (SFP+). Тем не менее, в отличие от других блочных массивов, система хранения XtremIO специально разработана для флеш-дисков, обеспечивая высокую производительность, предоставляя удобные в использовании расширенные услуги по управлению данными. В качестве основной платформы каждого контроллера системы хранения данных в массиве XtremIO используется специально настроенный упрощенный дистрибутив Linux. ОС XtremIO (XIOS) запускается в ОС Linux в качестве оболочки и обрабатывает все операции в контроллере системы хранения. ОС XIOS оптимизирована для работы в условиях высокой интенсивности операций ввода-вывода и управляет функциональными модулями системы, соединением RDMA по операциям InfiniBand, мониторингом и пулами памяти.

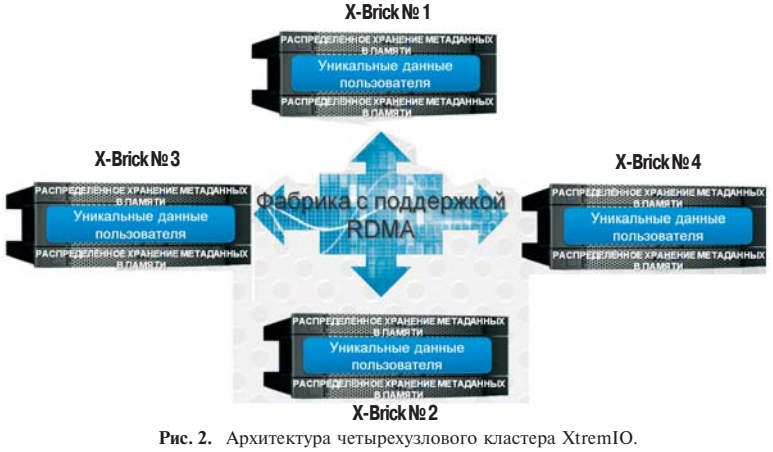

Рис. 2. Архитектура четырехузлового кластера XtremIO.

В ОС XIOS применяется проприетарный алгоритм планирования и обработки процессов, который позволяет обеспечить соответствие техническим требованиям подсистем хранения с малым временем отклика, высокой производительностью и поддержкой анализа содержания.

Службы XIOS предоставляют следующее:

– *планирование обработки процессов с низким временем отклика* — эффективное переключение контекста подпроцессов, оптимизацию планирования и минимальное времени ожидания;

– *линейную масштабируемость ЦП* — возможность полноценного использования всех ресурсов ЦП, включая многоядерные ЦП;

– *ограниченную межъядерную синхронизацию ЦП* — оптимизация взаимодействия внутренних подпроцессов и передачи данных;

– *отсутствие межразъемной синхронизации ЦП* — минимальное число задач синхронизации и зависимостей между подпроцессами, которые используют разные разъемы;

– *поддержку строк кэш-памяти* — оптимизация времени отклика и доступа к данным.

Контроллеры системы хранения данных в каждом модуле X-Brick сопоставлены с определенной дисковой полкой, которая подключается к ним через резервные модули SAS. Контроллеры систем хранения также подключены к резервным высокодоступным фабрикам InfiniBand. Независимо от того, какой контроллер системы хранения принимает от хоста запрос ввода-вывода, для его обработки используется сразу несколько контролле-

ров системы хранения данных на нескольких модулях X-Brick. Структура данных в системе хранения XtremIO обеспечивает естественное распределение нагрузки между всеми компонентами и равномерно задействует их в обработке операций ввода-вывода.

Принципы работы XtremIO

Массив хранения данных XtremIO автоматически сокращает количество данных (выполняет дедупликацию) как только эти данные попадают в систему, обрабатывая их блоками по 4 Кбайт. Процесс дедупликации выполняется глобально (по всей системе), перманентно и в режиме реального времени (и никогда не запускается как операция постобработки).

В системе хранения XtremIO используется глобальная кэш-память, в которой выполняется дедупликация данных и естественное равномерное распределение содержания по всему массиву. Все тома данных доступны во всех модулях X-Brick и со всех серверных портов массивов хранения данных.

Система использует высокодоступную внутреннюю сеть InfiniBand (поставляемую корпорацией EMC), в которой обеспечивается высокая скорость передачи данных со сверхнизкими задержками и удаленным прямым доступом к памяти (RDMA) между всеми контроллерами системы хранения данных в кластере. Используя RDMA, система XtremIO, по сути, образует одно общее пространство памяти, которое распространяется на все контроллеры системы хранения данных.

Эффективная логическая емкость одного модуля X-Brick меняется в зависимости от набора хранимых данных. При обработке информации с большим числом дублирующихся данных (что типично для виртуализированных сред) эффективно используемая емкость значительно выше, чем доступная физическая емкость флеш-дисков. В таких средах легко достигается коэффициент дедупликации от 3:1 до 10:1.

Размещение данных на основе содержания (“отпечатков”)

На каждом контроллере системы хранения данных есть таблица для записи местоположения каждого блока данных на твердотельный диск (табл. 3). Таблица состоит из двух частей:

– в первой части таблицы адрес LBA хоста сопоставляется с соответствующим “отпечатком” содержания;

– во второй части таблицы “отпечаток” содержания сопоставляется с его местоположением на твердотельном диске (SSD).

Вторая часть таблицы дает системе XtremIO уникальную возможность равномерно распределять данные по всему массиву и размещать каждый блок данных в наиболее подходящем месте на твердотельном диске. Также система может пропускать диск, который не отвечает на обращения, и выбрать место для записи новых блоков, когда массив почти заполнен и в нем нет пустых страйпов для записи.

При выполнении типичной операции записи поток входящих данных поступает на

Таблица 1. Пример таблицы сопоставления.

	Смещение LBA		Отпечаток		Смещение на твердотельном диске / физическое расположение	
4 K	0	→	20147A8	→	40	4 K
4 K	4	→	AB45CB7	→	8	4 K
4 K	8	→	F3AFBA3	→	88	4 K
4 K	12	→	963FE7B	→	24	4 K
4 K	16	→	0325F7A	→	64	4 K
4 K	20	→	134F871	→	128	4 K
4 K	24	→	CA38C90	→	516	4 K
4 K	28	→	963FE7B	–	Дедупликация	– X

Примечание.
Цвета блоков данных соответствуют их содержанию. Уникальное содержание представлено различными цветами, дубликат содержания представлен одним и тем же цветом (красный).

любой из контроллеров системы хранения данных, работающий в режиме “активные–активные”, блоками по 4 Кбайт. Для каждого блока данных размером 4 Кбайт в массиве создается “отпечаток” в виде уникального идентификатора.

Массив хранит таблицу этих “отпечатков” (см. табл. 1), чтобы определить, нет ли уже таких входящих записей в массиве. Эти “отпечатки” также используются для определения места хранения данных. Сопоставление адреса LBA с “отпечатком” содержания сохраняется в метаданных в памяти контроллера системы хранения данных.

Система проверяет, был ли ранее сохранен “отпечаток” и соответствующий блок размером 4 Кбайт. Если “отпечаток” новый, то система выполняет следующие действия:

– выбирает место в массиве для записи блока (по “отпечатку”, а не по адресу LBA);

– создает сопоставление между “отпечатком” и физическим местоположением;

– увеличивает значение счетчика ссылок на “отпечаток” на единицу;

– выполняет операцию записи.

В случае “дублирующейся” записи система записывает новое сопоставление адреса LBA и “отпечатка” и увеличивает значение счетчика ссылок для этого идентификатора. Так как данные уже находятся в массиве, нет необходимости ни изменять сопоставление “отпечатка” с физическим местоположением, ни записывать что-либо на твердотельный диск (SSD). Все изменения метаданных выполняются в памяти. Таким образом, запись по дедупликации выполняется быстрее, чем первая запись уникального блока данных. Это одно из уникальных преимуществ сокращения объема данных “на лету” XtremIO, когда дедупликация, по сути, повышает производительность записи.

Фактическая запись блока размером 4 Кбайт на твердотельный диск (SSD) выполняется асинхронно. В момент выполнения приложением операции записи система помещает блок размером 4 Кбайт в буфер записи в памяти (который защищен с помощью репликации на разные контроллеры системы хранения данных посредством RDMA) и сразу же возвращает подтверждение на хост. Когда в буфере накапливается достаточное количество блоков, система записывает их в страйпы XDP (технология защиты данных XtremIO) твердотельного диска (этот процесс, выполняемый наиболее эффективным способом, подробно описан в Белой книге “Защита данных XtremIO”, *прим. ред.*).

Из-за особенностей алгоритма вычисления “отпечатков” идентификаторы кажутся совершенно случайными числами и равномерно распределяются в пределах возможного диапазона значений «отпечатков». В результате, блоки данных равномерно распределяются по всему кластеру и всем твердотельным дискам массива (без применения внутренних процессов чистки, называемых также “сбор мусора”). Иными словами, в системе XtremIO отсутствует необходимость в проверке уровней использования пространства на различных твердотельных дисках и активном управлении для одинакового распределения операций записи по всем твердотельным дискам. Это позволяет XtremIO избежать свойственных другим флеш-массивам снижения показателей производительности в IOPS на 50%, повышения задержки на 1000% и 10-кратного сокращения срока службы флеш-дисков.

Все метаданные сохраняются системой в памяти контроллеров системы хранения данных и защищаются с помощью зеркального копирования журналов изменений между различными контроллерами системы хранения посредством RDMA. Метаданные периодически сохраняются на твердотельные диски.

Алгоритм защиты данных XtremIO (XDP)

Система хранения данных XtremIO обеспечивает высокоэффективную защиту данных с двойным контролем четности и самовосстановлением.

Система расходует очень мало емкости на метаданные и защиту данных. Отсутствует также необходимость и в выделенных резервных дисках для восстановления избыточности данных. Вместо этого, в системе используется технология “горячего” резерва, которая подразумевает, что для восстановления данных с неисправных дисков можно использовать любое свободное пространство массива. Система всегда резервирует достаточную распределенную емкость для выполнения одной операции восстановления избыточности данных.

Система XtremIO сохраняет свою производительность с минимальными издержками по емкости даже при высоком значении коэффициента использования ресурса хранения. В системе отсутствует необходимость в использовании схем зеркального копирования (и в соответствующих 100%-ных издержках по емкости).

Системе XtremIO требуется намного меньшая емкость для защиты данных, хранения метаданных, снимков, резервных дисков и обеспечения резерва производительности, благодаря чему остается намного больше места для пользователь-

ских данных. Это снижает стоимость полезного гигабайта данных. XtremIO обеспечивает следующие преимущества:

– защита данных по схеме N+2, аналогичная алгоритму RAID 6;

– низкие издержки по емкости, используемой для защиты данных, на уровне 8%;

– высокая производительность по сравнению с любым алгоритмом RAID (наиболее эффективному для записи алгоритму RAID-массивов, RAID 1, требуется на 60% больше операций записи, чем технологии XDP);

– высокий по сравнению с любым алгоритмом RAID срок службы флеш-дисков, благодаря меньшему количеству операций записи и равномерному распределению данных;

– автоматическое восстановление в случае сбоя диска и короткий период восстановления избыточности данных по сравнению с традиционными алгоритмами RAID;

– высокая отказоустойчивость и масштабируемые алгоритмы, которые полностью защищают входящие данные, даже если в системе присутствуют диски, в которых произошел сбой;

– простота администрирования, благодаря поддержке устранения проблем на местах;

– поддержка механизма сохранения работоспособности при случайном удалении дисков.

Эффективная поддержка функционала гипервизора VMware

Совместно используемые метаданные (shared in-memory metadata) позволяют массиву обеспечивать широкий спектр показателей производительности и быстро клонировать хранимую в массиве информацию, что обеспечивает резкое ускорение выполнения таких стандартных задач, как развертывание виртуальных машин и т. п. Клонирование виртуальных машин производится со скоростью до 20 раз превышающей ширину полосы пропускания между хостом и массивом, выполняется в несколько раз быстрее и с меньшим влиянием на производственные виртуальные машины, чем в других массивах на флеш-дисках.

Массив XtremIO полностью совместим с VAAI, что позволяет ему обмениваться данными непосредственно с vSphere и использовать такие функциональные возможности для ускорения хранения, как vMotion, выделение ресурсов для виртуальной машины и “тонкое” выделение ресурсов.

Кроме того, интеграция XtremIO с интерфейсом VAAI дополнительно повышает эффективность X-copy благодаря возможности свести все операции к обработке метаданных. Благодаря функции сокращения объема данных “на лету” и обработке метаданных в оперативной памяти в массиве XtremIO копирование фактических блоков данных во время выполнения команды X-copy не выполняется. В системе только создаются новые указатели для существующих данных, а весь процесс осуществляется в памяти контроллера системы хранения данных. Таким образом, ресурсы массива системы хранения не используются, а выполнение этого процесса не влияет на производительность системы.

Например, при помощи массива XtremIO можно мгновенно создать клон виртуальной машины (даже несколько раз).

Это стало возможным только благодаря таким функциям XtremIO, как сокращение данных “на лету” и обработка метаданных в оперативной памяти. На других флеш-дисках, где реализована поддержка VAAI, но отсутствует функция дедупликации “на лету”, сначала выполняется запись X-COPY и лишь затем — дедупликация. В массивах, не поддерживающих обработку метаданных в оперативной памяти, необходимо выполнить поиск на твердотельном диске для выполнения команды X-COPY, что отрицательно сказывается на выполнении операций ввода-вывода на существующих активных виртуальных машинах. Только в массиве XtremIO этот процесс выполняется быстро и без записи на твердотельный диск, не оказывая влияния на операции ввода-вывода на существующих виртуальных машинах.

В массиве XtremIO предусмотрен ряд функций, обеспечивающих поддержку интерфейса VAAI:

– *Zero Blocks/Write Same* – используется для очищения областей диска (термин VMware: HardwareAcceleratedInit). Эта функция обеспечивает ускоренное форматирование тома;

– *Clone Blocks/Full Copy/XCOPY* – используется для копирования или миграции данных в пределах одного физического массива (термин VMware: HardwareAcceleratedMove). Эта функция обеспечивает практически мгновенное клонирование виртуальной машины в массиве XtremIO без ущерба для пользовательских операций ввода-вывода на активных виртуальных машинах;

– *Record based locking/Atomic Test & Set (ATS)* – используется при создании и блокировке файлов в томе VMFS, например, во время отключения/включения виртуальных машин (термин VMware: HardwareAcceleratedLocking). Это позволяет использовать большие тома и кластеры ESX без возникновения конфликтов;

– *Block Delete/UNMAP/TRIM* – позволяет повторно выделять неиспользуемое пространство с помощью функции SCSI UNMAP (термин VMware: BlockDelete; только для vSphere 5.x).

Внедрения

Компания Boston Scientific – разработчик медицинского оборудования с ежегодным оборотом $7,25 млрд – испытывала большие сложности в управлении крупномасштабной средой VDI (на базе VMware Horizon View с гипервизором VMware vSphere с общим числом приложений более 2500 и обязательным полным копированием постоянных рабочих мест). Помимо этого, были частые жалобы пользователей VDI на медленную реакцию приложений и высокие затраты на хранение данных.

Для решения вышеназванных проблем было предложено использование массива на твердотельных дисках XtremIO с одним модулем X-Brick для каждого ”рабочего стола.

Результаты:

– бОльшие возможности рабочих мест для конечных пользователей;

– более низкая стоимость на рабочее место, чем в предыдущих массивах;

– быстрое развертывание рабочих мест в рабочее время (раньше требовалось 8 мин для развертывания одного виртуального образа рабочего стола размером 40 Гбайт, теперь – 17 сек);

– более простое и успешное развертывание сложной инфраструктуры VDI.

Заключение

Состоявшиеся последние объявления EMC в области массивов на базе флеш-технологий существенно расширяют возможности существующих ИТ-инфраструктур в области повышения производительности приложений, включая гетерогенные среды с использованием аппаратных и программных решений других вендоров.

Архитектура, заложенная в основу массива XtermIO, оптимизирована для всех подсистем корпоративной системы хранения данных на твердотельных дисках с учетом функционала, развертываемого на XtermIO.

XtremIO это высокомасштабируемое решение, позволяющее приобретать дополнительную емкость и повышать производительность по мере необходимости; обладает высокой производительностью с сотнями тысяч операций ввода-вывода в секунду, низкими задержками (доли миллисекунд), дедупликацией (с обработкой данных “на лету”), высокой доступностью, а также функциями “тонкого” выделения ресурсов, моментальных снимков и поддержки интерфейса VAAI.

Система XtremIO также предлагает уникальную запатентованную схему, в которой для обеспечения эффективного и мощного механизма защиты, способного защитить данные в случае двух одновременных и нескольких последовательных сбоев, используются уникальные особенности твердотельного диска.

В системе XtremIO предусмотрен комплексный, интуитивно понятный и удобный интерфейс, который включает в себя как графический интерфейс пользователя, так и возможности командной строки, и ориентирован на простоту использования для эффективного управления системой.

Массив XtremIO хорошо интегрируется с другими продуктами EMC – PowerPath, VPLEX, RecoverPoint и др.

Тимофей Григорьев,
EMC Россия