Белая книга

ОБЩИЕ СВЕДЕНИЯ О МАССИВЕ ХРАНЕНИЯ ДАННЫХ EMC XTREMIO (версия 3.0)

Подробный обзор

Аннотация

Эта белая книга знакомит с массивом хранения данных EMC XtremIO. Она содержит подробные описания архитектуры системы, принципов работы и функций. В ней также объясняется, как уникальные функции массива XtremIO, например сокращение объема данных (включая дедупликацию на лету и сжатие данных), масштабируемая производительность, защита данных и прочие возможности) обеспечивают решение проблем хранения данных, которые невозможно устранить с помощью любой другой системы.

Июль 2014

EMC²

Арт. H11752.5 (ред. 06)

Оглавлоение

Краткий обзор

Система хранения на флэш-дисках — привлекательный способ повышения производительности при выполнении операций ввода-вывода в центрах обработки данных. Но он всегда связан со значительными затратами за счет высокой стоимости и потери такой функциональности, как масштабируемость, высокая доступность и корпоративные функциональные возможности.

XtremIO — это масштабируемый массив хранения на твердотельных дисках корпоративного класса, который обеспечивает не только высокий уровень производительности и масштабируемости, но и существенно повышает удобство использования систем хранения на базе сети хранения данных, открывая новый мир невиданных ранее возможностей.

Архитектура массива XtremIO на твердотельных дисках изначально создавалась для обеспечения максимальной производительности и согласованно малого времени отклика. Особое внимание обращалось на функциональность для обеспечения высокой доступности корпоративного класса, сокращение объема данных на лету в реальном времени с целью существенного сокращения затрат, а также такие дополнительные функции, как «тонкое» выделение ресурсов, тесная интеграция с VMware, снимки файловой системы, клоны томов и непревзойденная защита данных.

При этом стоимость владения массивом остается вполне конкурентной. Архитектура продукта соответствует всем требованиям к системе хранения на основе флэш-дисков. Она, в частности, удлиняет срок службы флэш-дисков, снижает фактическую стоимость их емкости, а также обеспечивает производительность и масштабируемость, эффективность эксплуатации и расширенную функциональность массива хранения данных.

Эта белая книга в общих чертах знакомит с различными аспектами массива хранения данных XtremIO. В ней подробно описана архитектура системы, принципы ее работы и всевозможные функции.

Введение

XtremIO — это массив хранения данных на твердотельных дисках, который изначально разрабатывался в качестве системы, способной раскрыть весь потенциал производительности флэш-дисков и предоставить функциональные возможности массива, разработанные с учетом уникальных характеристик твердотельных дисков.

Для обеспечения беспрецедентных уровней производительности в массиве XtremIO используются компоненты и проприетарное интеллектуальное ПО, соответствующие отраслевым стандартам. Массив в состоянии обеспечить производительность в диапазоне от сотен тысяч до миллионов операций ввода-вывода в секунду при стабильно низком времени отклика менее одной миллисекунды.*

Конструкция системы позволяет также свести к минимуму планирование. Удобный интерфейс значительно облегчает выделение ресурсов и управление массивом.

В массиве XtremIO используются флэш-диски, которые дают ценные преимущества с точки зрения следующих основных характеристик.

- **Производительность.** Время отклика и пропускная способность массива всегда остаются прогнозируемыми и постоянными, независимо от степени занятости и коэффициента использования ресурса хранения. Время отклика для запросов ввода-вывода в пределах массива, как правило, гораздо меньше одной миллисекунды*.
- **Масштабируемость.** Система хранения XtremIO создана на основе масштабируемой архитектуры. Сначала система состоит из одного строительного блока — модуля X-Brick. Если требуется дополнительная производительность и емкость, система масштабируется путем добавления модулей X-Brick. Производительность масштабируется линейно, то есть конфигурация с двумя блоками X-Brick обеспечивает в два раза больше операций ввода-вывода в секунду, с четырьмя блоками X-Brick — в четыре раза больше операций, а с шестью блоками X-Brick — в шесть раз больше операций ввода-вывода в секунду, чем конфигурация с одним блоком X-Brick. При горизонтальном масштабировании системы время отклика остается согласованно неизменным.
- **Эффективность.** Основная подсистема обеспечивает сокращение объема данных на лету на основе содержания. Массив хранения XtremIO автоматически сокращает количество данных (выполняет дедупликацию) на лету по мере их поступления в систему. Это уменьшает количество данных, записываемых на флэш-диск, увеличивая срок службы носителей и снижая затраты. Массивы XtremIO выделяют емкость для томов небольшими фрагментами по требованию. Для томов всегда производится «тонкое» выделение ресурсов без потери производительности, избыточное выделение

* По данным измерений для блоков малых размеров. Для блоков ввода-вывода большого размера в любой системе хранения характерно более длительное время отклика.

емкости или фрагментация. После внедрения дедупликации на лету на основе содержания оставшиеся данные сжимаются еще сильнее, что сокращает объем данных, записываемых на флэш-диски. Дедуплицированные (уникальные) блоки данных сжимаются «на лету».

Преимущества сокращения объема операций записи:

- улучшение производительности за счет сокращения объема данных;
- повышение общей долговечности твердотельных дисков в массиве;
- уменьшение требуемой физической емкости хранения данных, что повышает эффективность массивов хранения и значительно снижает себестоимость ресурсов хранения.

- **Защита данных.** В массиве XtremIO используется проприетарный алгоритм защиты данных, оптимизированный для флэш-дисков (функция защиты данных XtremIO, или XDP). Этот алгоритм обеспечивает производительность, превосходящую все существующие алгоритмы RAID. Оптимизация XDP также приводит к уменьшению количества операций записи на флэш-диски в целях защиты данных.
- **Функциональность.** XtremIO поддерживает высокую производительность и компактные моментальные снимки, сокращение объема данных (включая дедупликацию на лету и сжатие данных), «тонкое» выделение ресурсов и полную интеграцию с интерфейсом VMware VAAI, а также поддержку протоколов Fibre Channel и iSCSI.

Обзор системы

Массив хранения XtremIO — это система на твердотельных дисках, в основе которой лежит масштабируемая архитектура. В системе используются строительные блоки X-Brick, которые по мере необходимости можно объединять в кластеры для увеличения производительности и емкости, как показано на Рис. 2.

Управление системными операциями осуществляется с помощью выделенного автономного сервера на платформе ОС Linux, который называется управляющим сервером XtremIO (XMS). Для каждого кластера XtremIO необходим отдельный хост XMS, который может быть как физическим, так и виртуальным сервером. Массив продолжает работу, даже если он отключен от XMS, но в этом случае невозможно настроить его конфигурацию и осуществлять мониторинг его работы.

Специально разработанная архитектура массива XtremIO позволяет полностью раскрыть весь потенциал производительности флэш-дисков, обеспечивая при этом сбалансированное линейное масштабирование всех ресурсов, таких как ЦП, ОЗУ, твердотельные диски и серверные порты. Это позволяет достичь любого желаемого уровня производительности массива, сохранив при этом согласованность уровней производительности,

что играет критически важную роль с точки зрения прогнозируемого поведения приложений.

Система хранения данных XtremIO обеспечивает очень высокий уровень производительности, который остается согласованным с течением времени, при изменении состояния системы и схем доступа. Эта система создана для обслуживания случайных операций ввода-вывода.

На уровень производительности системы не влияют ни коэффициент использования ресурса хранения, ни количество томов, ни эффект износа. Кроме того, для обеспечения производительности не используется архитектура «общей кэш-памяти». Поэтому производительность не зависит от размера набора данных или схемы доступа к информации.

Благодаря учитывающей содержание архитектуре системы хранения массив XtremIO обеспечивает следующие преимущества:

- равномерное распределение блоков данных, которое естественным образом обеспечивает максимальную производительность при минимальном износе флэш-дисков;
- равномерное распределение метаданных;
- отсутствие горячих точек с высоким потоком данных или метаданных;
- простота установки без настроек;
- расширенная функциональность системы хранения, в том числе сокращение объема данных (дедупликация и сжатие данных), «тонкое» выделение ресурсов, расширенная защита данных (XDP), моментальные снимки и многое другое.

X-Brick

На Рис. 1 изображен модуль X-Brick.

Рис. 1 - X-Brick

X-Brick — это основной строительный блок массива XtremIO.

Каждый модуль X-Brick включает следующие компоненты:

- одна дисковая полка форм-фактора 2U, содержащая следующие компоненты:
  - 25 твердотельных дисков eMLC (стандартный модуль X-Brick) или 13 твердотельных дисков eMLC (модуль Starter X-Brick [5 Тбайт] на 10 Тбайт);
  - два резервных блока питания;
  - два резервных соединительных модуля SAS;
- одна батарея аварийного питания;
- два контроллера системы хранения данных (резервные процессоры СХД) в форм-факторе 1U.

  Каждый контроллер системы хранения данных состоит из следующих компонентов:
  - два резервных блока питания;
  - два порта Fibre Channel 8 Гбит/с;
  - два порта 10GbE iSCSI;
  - два порта InfiniBand 40 Гбит/с;
  - один порт управления/IPMI со скоростью 1 Гбит/с

В Таблица 1 указаны технические характеристики каждого модуля X-Brick.

Таблица 1. Системные характеристики (на X-Brick)

Функция	Технические характеристики (на модуль X-Brick)
Физические среды	• 5U • 13 твердотельных дисков eMLC (модуль Starter X-Brick [5 Тбайт] на 10 Тбайт) • 25 твердотельных дисков eMLC (обычный модуль X-Brick)
Высокая доступность	• Резервированные • Компоненты с возможностью «горячей» замены • Нет критических точек отказа (отказ в которых приведет к отказу всей системы)
Доступ хостов	Симметричный в режиме «активный-активный». Доступ ко всем томам можно получать параллельно с любого целевого порта на любом контроллере с аналогичной производительностью. Нет необходимости в ALUA.
Порты хостов	• 4 порта FC 8 Гбит/с • 4 порта Ethernet iSCSI 10 Гбит/с
Полезная емкость*	• Модуль Starter X-Brick (5 Тбайт) на 10 Тбайт: - 3,16 Тбайт (13 твердотельных дисков, без сокращения объемов данных); - 6,99 Тбайт (25 твердотельных дисков, без сокращения объемов данных). • Модуль X-Brick на 10 Тбайт: 7,47 Тбайт (без сокращения объема данных). • Модуль X-Brick на 20 Тбайт: 14,94 Тбайт (без сокращения объема данных).
Время отклика	Менее одной миллисекунды†

* Полезная емкость — это объем уникальных несжимаемых данных, которые можно записать в массив. Эффективная емкость обычно значительно больше благодаря сокращению объемов данных на лету в системе XtremIO. Окончательные показатели могут несколько отличаться от указанных.
† Задержка менее 1 мс характерна для блоков типовых размеров. В случае малых или крупных блоков задержка может быть выше.

Архитектура Scale-Out

Система хранения XtremIO может состоять из одного модуля X-Brick или представлять собой кластер из нескольких модулей X-Brick, как показано на Рис. 2 и в Табл. 2.*

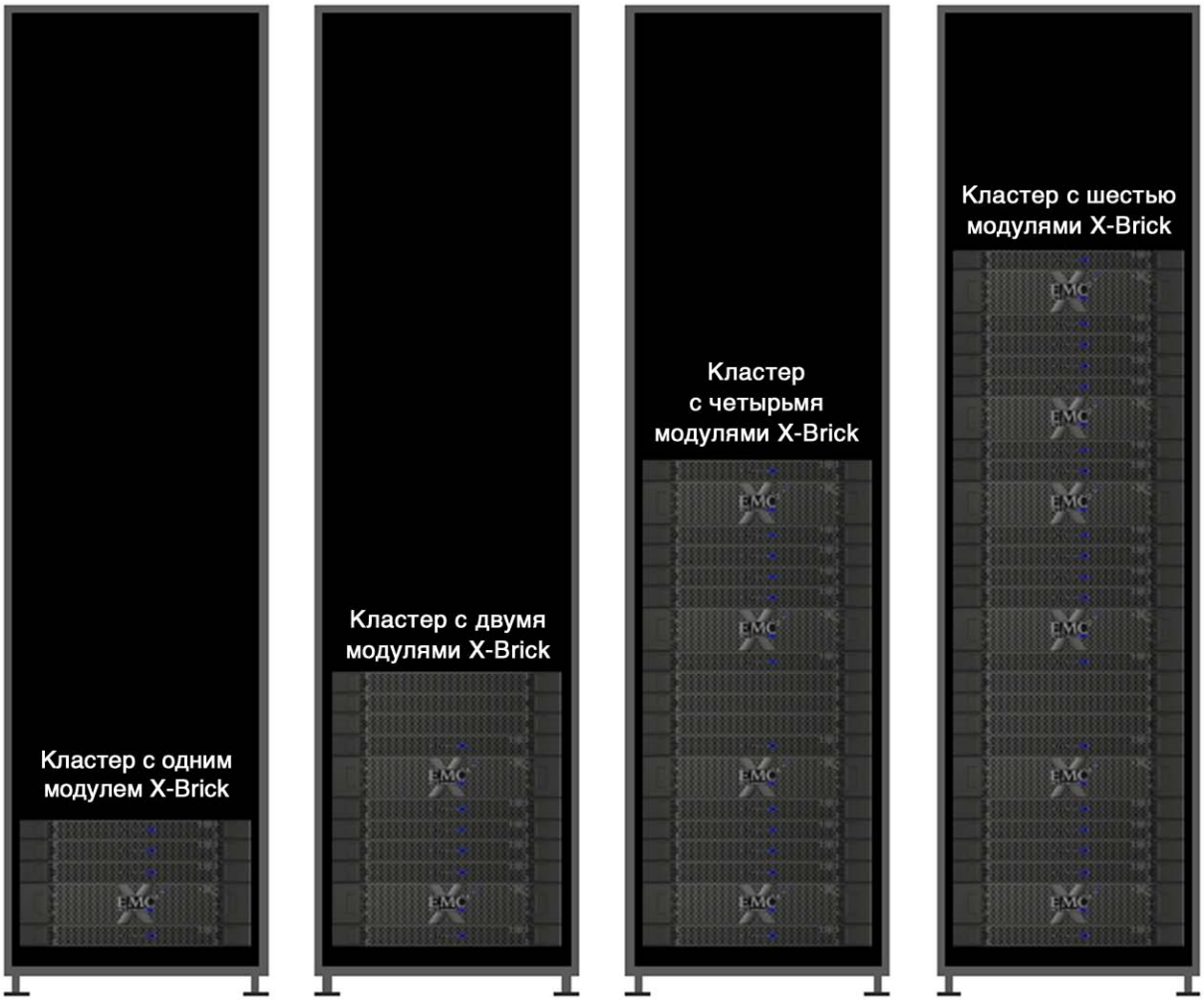


Рис. 2. Конфигурации систем из одного блока и в виде кластеров из нескольких блоков X-Brick.

В случае кластеров из двух и более модулей X-Brick в системе хранения XtremIO используется высокодоступная внутренняя сеть InfiniBand (40 Гбит/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 о технологии InfiniBand.

* Система версии 3.0 поддерживает до шести модулей X-Brick в кластере (доступна с 4-го квартала 2014 г.). В последующих версиях ОС XtremIO их количество продолжит расти.

Кластер с одним модулем X-Brick состоит из следующих компонентов:

- один модуль X-Brick;
- один дополнительный резервный аккумулятор.

Кластер с несколькими модулями X-Brick состоит из следующих компонентов:

- два или четыре модуля X-Brick;
- два коммутатора InfiniBand.

Табл. 2. Конфигурации систем из одного блока и в виде кластеров из нескольких блоков X-Brick.

	10 Тбайт Starter X-Brick (5 Тбайт)	Кластер с одним модулем X-Brick	Кластер с двумя модулями X-Brick	Кластер с четырьмя модулями X-Brick	Кластер с шестью модулям и X-Brick
Количество модулей X-Brick	1	1	2	4	6
Количество коммутаторов InfiniBand	0	0	2	2	2
Количество дополнительных батарей аварийного питания	1	1	0	0	0

10 Тбайт Starter X-Brick (5 Тбайт)

Кластер XtremIO Starter X-Brick (5 Тбайт) на 10 Тбайт идентичен стандартному кластеру X-Brick, но оснащен всего 13 твердотельными дисками eMLC вместо 25. XtremIO Starter X-Brick (5 Тбайт) на 10 Тбайт можно расширить до обычного кластера X-Brick путем добавления 12 твердотельных дисков.

Архитектура системы

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 10 Гбит/с (SFP+).

Тем не менее, в отличие от других блочных массивов, система хранения XtremIO специально разработана для флэш-дисков, обеспечивая непревзойденную производительность и 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. Операционная система XtremIO Operating System (XIOS) работает на основе Linux и обслуживает все операции контроллера системы хранения данных, как показано на Рис. 3. XIOS оптимизирована для обработки большого количества операций вывода-вывода. Операционная система управляет функциональными модулями системы, операциями RDMA поверх InfiniBand, средствами мониторинга и пулами памяти.

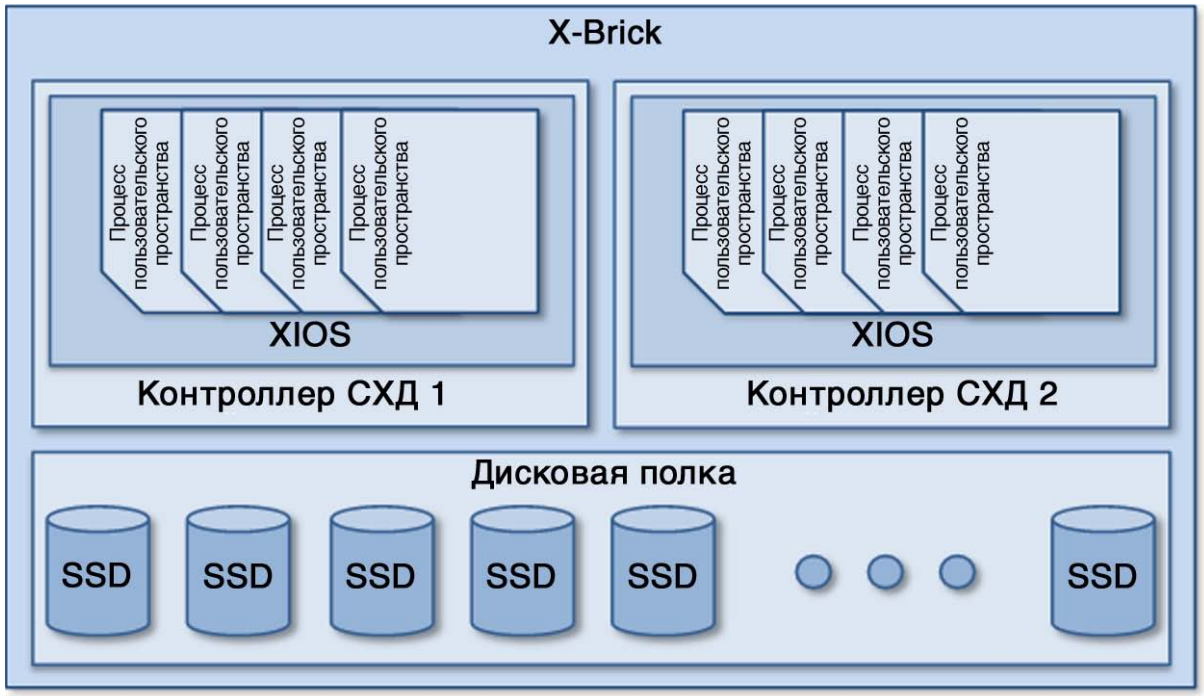


Рис. 3. Блок-схема X-Brick.

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.
- Линейная масштабируемость ЦП — возможность полноценного использования всех ресурсов ЦП, включая многоядерные процессоры.
- Ограниченная межъядерная синхронизация ЦП — оптимизация взаимодействия внутренних подпроцессов и передачи данных.
- Отсутствие межразъемной синхронизации ЦП — минимальное число задач синхронизации и зависимостей между подпроцессами, которые используют разные разъемы.
- Поддержка 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 контроллеров 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

Массив хранения данных XtremIO автоматически сокращает количество данных (выполняет дедупликацию), как только эти данные попадают в систему, и обрабатывает их поблочно. Процесс дедупликации выполняется глобально (по всей системе), перманентно и в режиме реального времени (и никогда не запускается как операция постобработки). После дедупликации данные сжимаются «на лету», а затем записываются на твердотельные диски.

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 клонированных сред, например VDI) эффективная используемая емкость значительно выше, чем доступная физическая емкость флэш-дисков. В таких средах легко достигается коэффициент дедупликации от 5:1 до 10:1.
- В случае сжимаемых данных (как правило, это базы данных и данные приложений) коэффициент сжатия находится в диапазоне от 2:1 до 3:1.
- В случае с системами, где хорошо работает как сжатие, так и дедупликация данных (например, VSI), типичный коэффициент сжатия составляет 6:1.

Таблица сопоставления

В каждом контроллере системы хранения данных хранится таблица для записи местоположения каждого блока данных на твердотельный диск, как показано в Табл. 3 (стр. 15).

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).

Эта 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.

Порядок операций ввода-вывода при записи

При выполнении типичной операции записи поток входящих данных поблочно поступает на любой из контроллеров системы хранения данных, работающий в режиме «активный-активный». Для каждого блока данных в массиве создается «отпечаток» в виде уникального идентификатора.

Массив хранит таблицу этих «отпечатков» (как показано в Табл. 3), 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.

Табл. 3. Пример таблицы сопоставления.

		Смещение LBA		Отпечаток		Смещение на твердотельном диске / физическое расположение		
Данные	→	Address 0	→	20147A8	→	40	→	Данные
Данные	→	Address 1	→	AB45CB7	→	8	→	Данные
Данные	→	Address 2	→	F3AFBA3	→	88	→	Данные
Данные	→	Address 3	→	963FE7B	→	24	→	Данные
Данные	→	Address 4	→	0325F7A	→	64	→	Данные
Данные	→	Address 5	→	134F871	→	128	→	Данные
Данные	→	Address 6	→	CA38C90	→	516	→	Данные

		Смещение LBA		Отпечаток		Смещение на твердотельном диске / физическое расположение		
Данные	→	Address 7	→	963FE7B	–	Дедупликация	–	X

Примечание.
В Табл. 3 цвета блоков данных соответствуют их содержанию. Уникальное содержание представлено различными цветами, дубликат содержания представлен одним и тем же цветом (красным).

Система проверяет, был ли ранее сохранен «отпечаток» и соответствующий блок.

Если «отпечаток» новый, то система выполняет следующие действия:

- сжимает данные;
- 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 на твердотельный диск (SSD) выполняется асинхронно. В момент выполнения приложением операции записи система помещает блок 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».

После вызова команды I/O write (запись в порт) в массиве происходит следующее.

1. Система анализирует входящие данные и разделяет их на блоки, как показано на Рис. 4.

Рис. 4. Данные, разделенные на блоки фиксированного размера.

2. Для каждого блока данных в массиве выделяется уникальный «отпечаток» данных, как показано на Рис. 5.

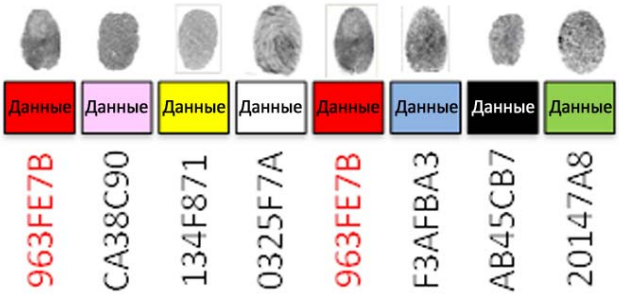


Рис. 5. «Отпечаток», выделенный для каждого блока.

В массиве хранится таблица с этим «отпечатком», позволяющая в дальнейшем определять, нет ли в массиве записываемых данных, как показано в Табл. 3 (стр. 15).

- Если такой блок данных в системе не существует, выполняющий обработку контроллер системы хранения данных заносит в журнал данные о том, что он намерен записать блок на другие контроллеры системы хранения данных, используя «отпечаток» для определения местоположения данных.
- Если блок данных уже существует в системе, запись не выполняется, как показано на Рис. 6.

Рис. 6. Дедупликация существующего или повторяющегося блока.

3. Для каждого блока массив увеличивает значение счетчика.
4. Согласованное распределение и сопоставление позволяет направлять каждый блок на контроллер системы хранения данных, который отвечает за адресное пространство соответствующего «отпечатка».

   Согласованное распределенное сопоставление выполняется по «отпечатку» содержания. Математический процесс вычисления «отпечатков» обеспечивает равномерное распределение значений «отпечатков», а их сопоставление равномерно распределяется между всеми контроллерами системы хранения данных в кластере, как показано на Рис. 7.

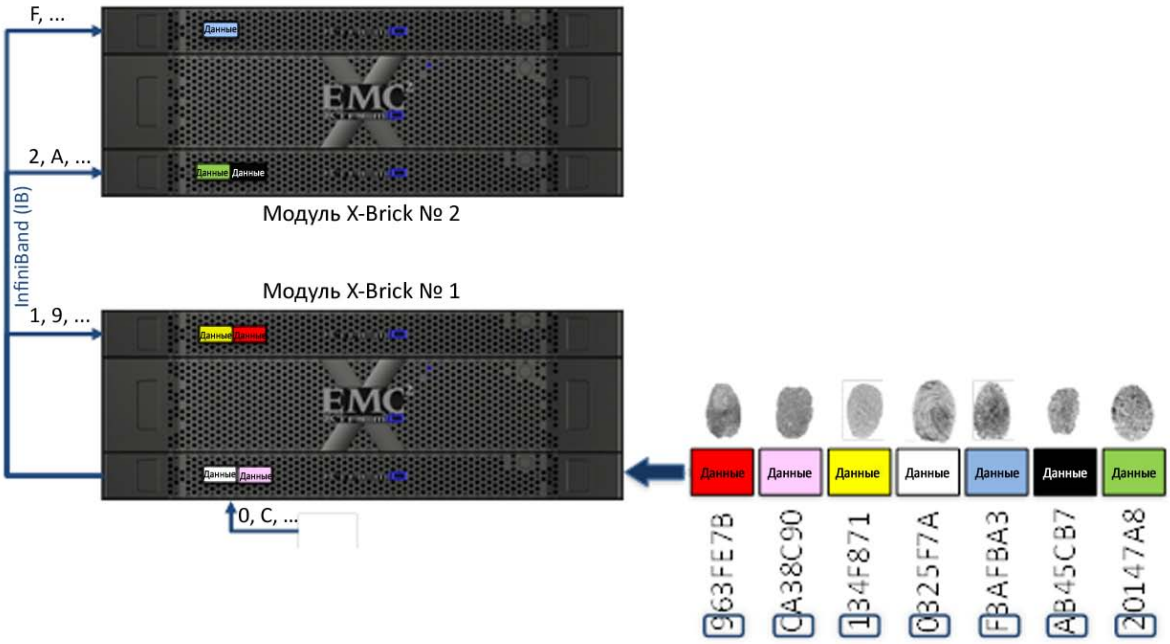


Рис. 7. Распределение данных по всему кластеру.

Примечание.
Передача данных по всему кластеру выполняется посредством RDMA с помощью высокоскоростной сети InfiniBand с низким временем отклика, как показано на Рис. 7.

---

5. Система передает подтверждение обратно на хост.

6. Благодаря равномерному распределению процесса создания «отпечатков» все контроллеры хранения данных в кластере получают одинаковые доли блоков данных. После поступления дополнительных блоков эти блоки распределяются по страйпам, как показано на Рис. 8.

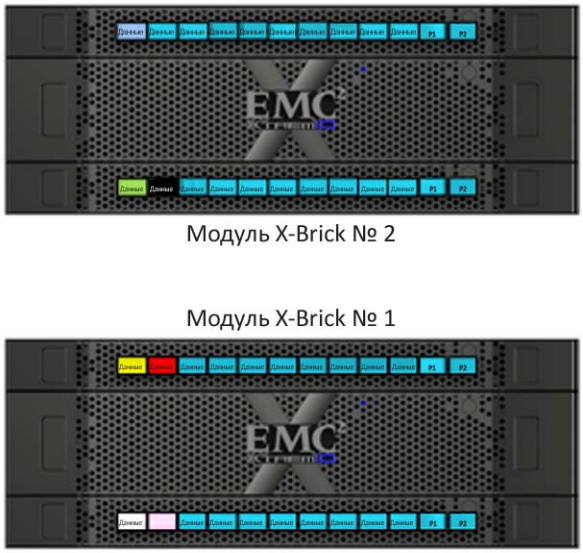


Рис. 8. Дополнительные блоки, распределенные по всем страйпам.

7. Система сжимает блоки данных, чтобы еще больше уменьшить их размер.
8. После накопления в контроллере системы хранения данных достаточного количества блоков данных, чтобы заполнить наименее заполненные страйпы в массиве (или весь пустующий страйп, при его наличии), контроллер передает блоки из кэш-памяти на твердотельный диск, как показано на Рис. 9.

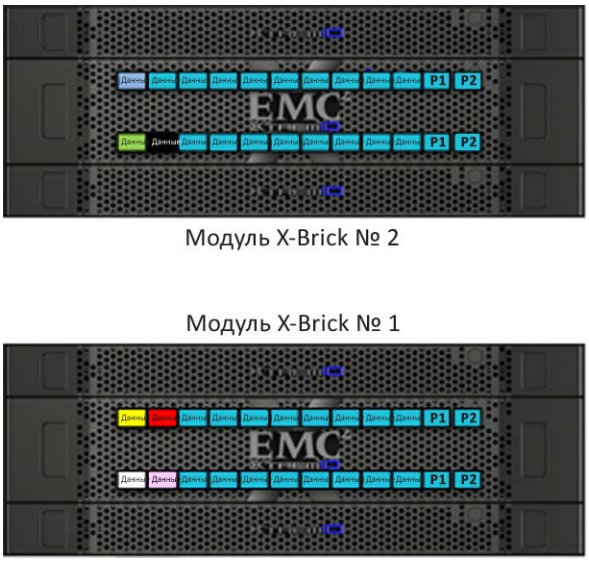


Рис. 9. Страйпы на твердотельных дисках.

Порядок операций ввода-вывода при чтении

При чтении блоков система производит поиск логического адреса в таблице LBA, чтобы сопоставить его с «отпечатком». Как только «отпечаток» найден, система выполняет поиск соответствия «отпечатка» физическому местоположению блока и извлекает этот блок из найденного физического местоположения. Поскольку данные равномерно записываются по всему кластеру и твердотельным дискам, нагрузка, связанная с операциями чтения, также распределяется равномерно.

В каждом контроллере системы хранения данных XtremIO есть кэш-память чтения, расположенная в оперативной памяти.

- В традиционных массивах кэш-память чтения заполняется по логическим адресам. Блоки по адресам с наибольшей вероятностью считывания помещаются в кэш-память чтения.
- В массиве XtremIO кэш-память чтения заполняется по «отпечаткам» содержания. В кэш-память помещаются блоки, у содержания которых (представленного идентификаторами «отпечатков») наиболее высокая вероятность считывания.

Это обеспечивает поддержку дедупликации кэш-памяти чтения в массиве XtremIO, благодаря чему кэш-память чтения относительно небольшого размера оказывается намного вместительнее традиционной кэш-памяти той же емкости.

Если запрашиваемый размер блока больше размера записанных блоков данных, система XtremIO выполняет одновременное чтение блоков данных по всему кластеру и собирает их в более крупные блоки перед возвратом приложению.

Сжатые блоки данных перед отправкой восстанавливаются.

После вызова команды I/O read (чтение с порта) в массиве происходит следующее.

1. Система анализирует входящий запрос, чтобы определить адрес LBA для каждого блока данных, и создает буфер для хранения этих данных.
2. Следующие процессы выполняются параллельно.
   - Массив находит хранимый «отпечаток» для каждого блока данных. Местоположение блока данных в массиве X-Brick определяется с помощью «отпечатка». Для более интенсивных операций ввода-вывода (например, со скоростью 256 Кбит/с) каждый блок данных извлекается несколькими массивами X-Brick.
   - Система передает запрошенные данные операции чтения по сети InfiniBand на выполняющий обработку контроллер системы хранения данных посредством RDMA.
3. Система отправляет полностью заполненный буфер данных обратно на хост.

Функциональность системы

Массив хранения XtremIO Storage Array предлагает широкий спектр функций, которые доступны без специальных лицензий.

Функции системы

- Функции обработки данных, применяемые последовательно (в указанном ниже порядке) ко всем операциям записи.
  - «Тонкое» выделение ресурсов
  - Сокращение объема данных «на лету»:
    - Дедупликация данных «на лету»
    - Сжатие данных «на лету»
  - Защита данных XtremIO (XDP)
  - Шифрование данных в состоянии покоя
  - Снимки файловой системы
- Общесистемные функции
  - Масштабируемая производительность
  - Равномерное распределение данных
  - Высокая доступность
- Другие функции
  - Обновление без прерывания работы
  - Интеграция с VMware VAAI

«Тонкое» выделение ресурсов

В системе хранения XtremIO изначально предусмотрено «тонкое» выделение ресурсов с использованием внутренних блоков небольшого размера. Такое выделение обеспечивает мелкоструктурное разбиение пространства с «тонким» выделением ресурсов.

Во всех томах системы предусмотрено «тонкое» выделение ресурсов. Это означает, что система использует емкость только по мере необходимости. Система XtremIO определяет, куда поместить уникальные блоки данных в пределах кластера после вычисления идентификаторов «отпечатков». Таким образом, до наступления момента записи в массиве пространство системы хранения никогда заранее не распределяется и полностью не выделяется.

Благодаря архитектуре обработки содержания XtremIO блоки данных могут храниться в любом месте системы (их местоположение определяется только метаданными), и данные записываются только при получении уникальных блоков.

Таким образом, в отличие от множества архитектур, ориентированных на жесткие диски, «тонкое» выделение ресурсов в архитектуре XtremIO исключает пустую трату пространства и сборку мусора. Кроме того, в системе XtremIO отсутствует длительная фрагментация тома (блоки разбросаны по всему массиву с произвольным доступом). Также нет необходимости в программах дефрагментации.

Естественное «тонкое» выделение ресурсов в системе XtremIO обеспечивает также согласованную производительность и управление данными на протяжении всего жизненного цикла томов, независимо от коэффициента использования ресурсов хранения системы или схем записи.

Сокращение объема данных «на лету»

Уникальное сокращение объемов данных «на лету» в системе XtremIO достигается благодаря использованию указанных ниже методов.

- Дедупликация данных «на лету».
- Сжатие данных «на лету».

Дедупликация данных «на лету»

Дедупликация данных «на лету» — это устранение избыточных данных **перед** их записью на флэш-диски.

Система XtremIO автоматически выполняет глобальную дедупликацию данных при их попадании в систему. Дедупликация выполняется в режиме реального времени, а не во время постобработки. В системе XtremIO отсутствуют ресурсоемкие фоновые процессы и дополнительные операции чтения или записи (связанные с постобработкой). Таким образом, это не оказывает отрицательного влияния на производительность массива хранения данных, не приводит к трате доступных ресурсов, выделенных для серверных операций ввода-вывода, а также не вызывает излишнего износа твердотельных дисков.

Блоки данных в системе XtremIO хранятся в соответствии с их содержанием, а не согласно адресу в томе на уровне пользователя. Это обеспечивает идеальную балансировку нагрузки между всеми устройствами в системе с точки зрения емкости и производительности. При каждом изменении блок данных может быть помещен в любой набор твердотельных дисков в системе или же не будет записываться вообще, если его содержание уже есть в системе.

Система естественным образом равномерно распределяет данные по массиву, используя все твердотельные диски и обеспечивая равномерный износ. Даже в случае повторной записи по одному и тому же адресу LBA на хосте все операции записи направляются в разные места в пределах массива XtremIO. Данные, для которых хост выполняет многократное повторение записи, будут дедуплицированы без выполнения дополнительных операций записи на флэш-накопители.

Для высокоэффективной дедупликации данных система XtremIO использует глобальную дедуплицированную кэш-память с учетом содержания. Уникальная архитектура системы хранения с учетом содержания позволяет добиться существенно большего размера кэш-памяти при меньшем объеме памяти DRAM. Таким образом, система XtremIO является идеальным решением для реализации сложных моделей доступа к данным, таких как модель «шквала пользовательских загрузок», которые распространены в средах виртуальных рабочих мест.

«Отпечатки» содержимого используются системой не только для дедупликации данных «на лету», но и для равномерного распределения по массиву блоков данных. Это обеспечивает естественную равномерную балансировку нагрузки и повышает эффективность с точки зрения равномерности износа твердотельных дисков, поскольку данные не нуждаются в перезаписи или перебалансировке.

Выполнение этого процесса «на лету» и по всему массиву приводит к уменьшению количества операций записи на твердотельный диск. Уменьшение количества операций записи продлевает срок службы твердотельных дисков и предотвращает снижение производительности, характерное для дедупликации при постобработке.

Дедупликация «на лету» и интеллектуальный процесс хранения данных в системе XtremIO обеспечивают следующие преимущества:

- сбалансированное использование системных ресурсов и максимальная производительность системы;
- уменьшение количества операций и продление срока службы флэш-дисков;
- равномерное распределение данных и обусловленный этим равномерный износ твердотельных дисков системы;
- отсутствие сборки мусора на уровне системы (в отличие от сокращения объема данных путем постобработки);
- интеллектуальное использование емкости твердотельных дисков и сокращение затрат на хранение.

Сжатие данных «на лету»

Сжатие данных «на лету» выполняется после дедупликации и **до** записи этих данных на флэш-диски.

После удаления всех дубликатов XtremIO автоматически сжимает данные. Благодаря этому сжимаются только уникальные блоки данных. Сжатие данных выполняется в режиме реального времени, а не во время постобработки.

Общий коэффициент сжатия зависит от природы набора данных. Затем сжатый блок данных сохраняется в массиве.

Сжатие сокращает общий объем физических данных, записываемый на твердотельный диск. Это снижает эффект увеличения объема записи (Write Amplification) на твердотельные диски, увеличивая таким образом надежность массива флэш-дисков.

Сжатие данных «на лету» в системах XtremIO обеспечивает целый ряд преимуществ.

- Сжатие данных всегда выполняется «на лету», а не после их сохранения. Таким образом, данные всегда записываются только один раз.
- Поддерживается сжатие различных наборов данных (например, баз данных, сред VDI, VSI и т. п.).
- Во многих случаях сжатие данных дополняет дедупликацию. Например, в среде VDI дедупликация значительно сокращает емкость, требуемую для хранения клонированных рабочих мест. В свою очередь, сжатие позволяет сократить объем данных, хранимых отдельными пользователями. В результате один модуль X-Brick может управлять еще большим количеством рабочих мест VDI.
- Сжатие экономит емкость благодаря максимальной эффективности хранения данных.
- В сочетании с возможностями создания моментальных снимков система XtremIO может легко хранить петабайты полезных данных приложений.

Общее сокращение объемов данных

Функции дедупликации и сжатия данных в системе XtremIO дополняют друг друга. Дедупликация позволяет снизить объем физических данных за счет устранения избыточных блоков. Сжатие еще больше снижает объем данных за счет устранения избыточности на уровне двоичного кода каждого блока.

На Рис. 10 показаны преимущества совместного использования дедупликации и сжатия, позволяющего уменьшить общий объем данных.

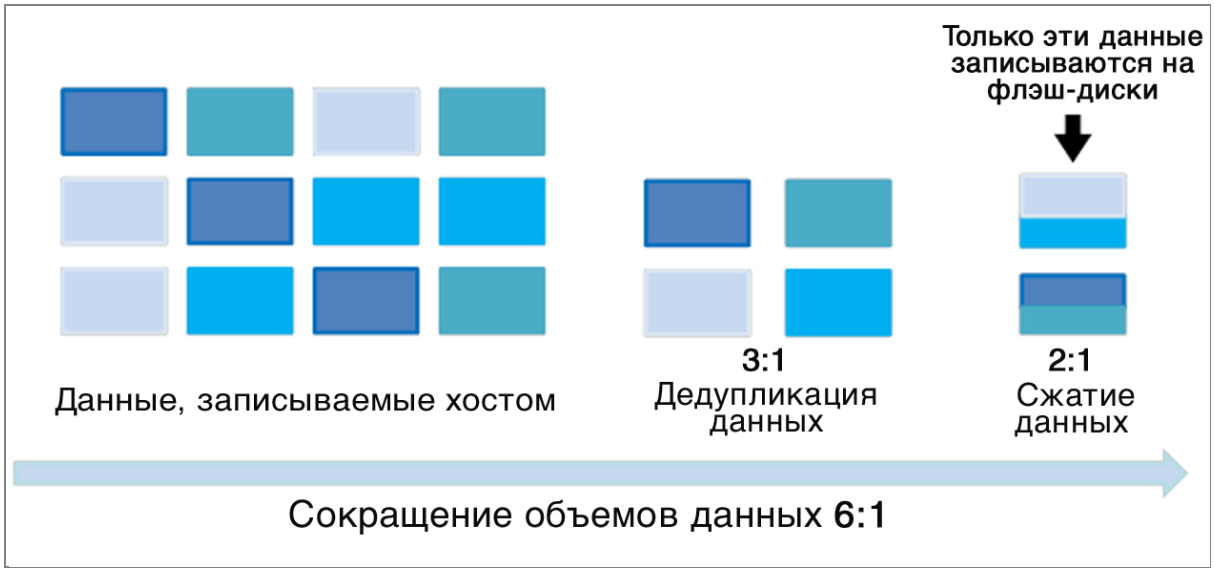


Рис. 10. Совместное использование дедупликации и сжатия данных.

В примере выше 12 блоков данных, записываемых хостом, сначала благодаря дедупликации превращаются в 4. Т. е., коэффициент дедупликации составляет 3:1. Затем каждый из 4 блоков данных подвергается сжатию с коэффициентом 2:1. Таким образом, общий коэффициент уменьшения объема данных составляет 6:1.

Защита 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 стопроцентных 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 пользовательских данных. Это снижает стоимость полезного гигабайта данных.

Система хранения данных XtremIO обеспечивает следующие преимущества:

- защита данных по схеме N+1;
- невероятно низкие издержки по емкости, используемой для защиты данных, на уровне 8 %;
- превосходная производительность по сравнению с любым алгоритмом RAID (RAID 1, наиболее эффективному для записи алгоритму RAID-массивов, требуется на 60 % больше операций записи, чем технологии XDP);
- превосходный по сравнению с любым алгоритмом RAID срок службы флэш-дисков, возможный 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;
- превосходная отказоустойчивость и масштабируемые алгоритмы, которые полностью защищают входящие данные, даже если в системе есть диски, в которых произошел сбой;
- простота администрирования благодаря поддержке устранения проблем на местах.

* Текущая версия ПО поддерживает восстановление избыточности данных на одном диске в определенный момент времени. Двойное параллельное восстановление избыточности данных станет доступно в следующем доработанном выпуске.

Табл. 4. Сравнение защиты данных в массиве XtremIO и схемах RAID.

Алгоритм	Производ-ительность	Защита данных	Издержки по емкости	Число операций чтения на обно в-ление страйпа	Улучш-ение по сравне-нию с традиц-ионными алгори-тмами при чтении	Число опера-ций записи на обнов-ление страйпа	Улучш-ение по сравне-нию с традиц-ионными алгори-тмами при записи
RAID 1	Высокая	1 отказ	50 %	0	–	2 (64 %)	В 1,6 раза
RAID 5	Средняя	1 отказ	25 % (3+1)	2 (64 %)	В 1,6 раза	2 (64 %)	В 1,6 раза
RAID 6	Низкая	2 отказа	20 % (8+2)	3 (146 %)	В 2,4 раза	3 (146 %)	В 2,4 раза
XtremIO XDP	**На 60 % лучше, чем RAID 1**	**1 отказ на модуль X-Brick**	**Сверх-низкая 8 % (23+2)**	**1,22**	**–**	**1,22**	**–**

Принцип работы XDP

В функцию защиты данных XtremIO (XDP) изначально заложена возможность использования преимуществ определенных свойств флэш-дисков и архитектуры системы хранения данных с адресацией по содержанию (Content Addressed Storage, CAS).

Используя преимущество возможности контролировать места хранения данных без издержек, технология XDP обеспечивает высокий уровень защиты и небольшое количество служебных протокольных данных, но с более высокой производительностью, чем у массива RAID 1. У технологии XDP есть также дополнительное преимущество, которое состоит в том, что она значительно повышает срок службы флэш-диска по сравнению с любым предыдущим алгоритмом RAID. Это важный фактор для массива на твердотельных дисках корпоративного класса.

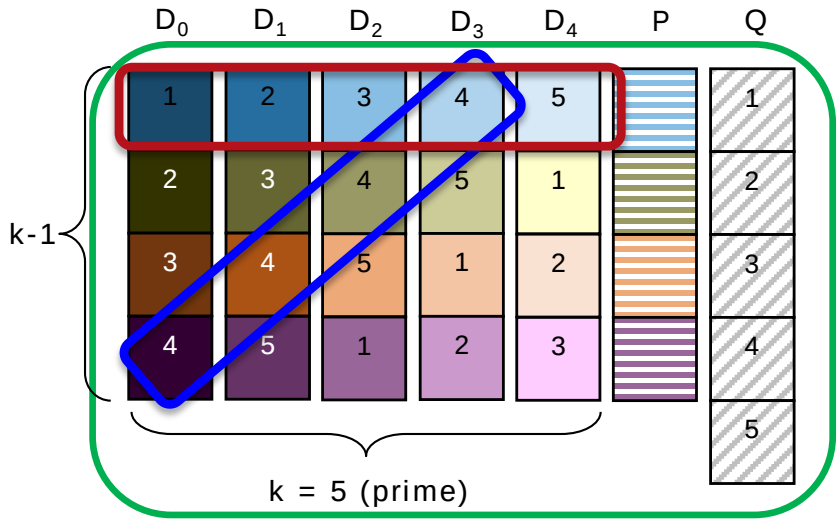


Рис. 11. Четность по рядам и диагоналям.

Защита данных XDP использует вариацию строк N+2 и диагональный контроль четности, как показано на Рис. 11, что обеспечивает защиту от возникновения двух одновременных ошибок на твердотельном диске. При использовании массивов, состоящих из 25 твердотельных дисков, на служебные протокольные данные расходуется всего 8 % емкости.

Традиционные массивы обновляют логические адреса блоков (LBA) в том же физическом расположении на диске (что приводит к возникновению большого количества служебных протокольных данных при выполнении операций ввода-вывода для обновления страйпов). В массиве хранения XtremIO данные всегда размещаются в наименее заполненном страйпе. Запись данных в наименее заполненный страйп эффективно уменьшает количество служебных протокольных данных для операций ввода-вывода при чтении и записи для каждого обновления страйпа и доступна только в архитектуре массива XtremIO на твердотельных дисках с учетом содержимого. Этот процесс обеспечивает согласованность выполнения операций системы XtremIO по мере заполнения массива и выполняется на протяжении длительного периода времени до тех пор, пока перезапись и частичное обновление страйпов не станут регулярной операцией.

В массиве XtremIO также предусмотрен процесс восстановления избыточности данных. Если в обычном массиве RAID 6 происходит один сбой диска, для восстановления избыточности данных используются методы RAID 5, то есть выполняется считывание каждого страйпа, а недостающие ячейки вычисляются на основании данных, хранящихся в других ячейках страйпа. Вместо этого в массиве XtremIO для восстановления избыточности недостающей информации используется контроль честности P и Q. При этом применяется тщательно проработанный алгоритм, который обеспечивает считывание только той информации, которая необходима для следующей операции восстановления избыточности данных в ячейке.

Табл. 5. Сравнение количества операций чтения XDP для восстановления избыточности данных после замены неисправного диска с аналогичными показателями различных схем RAID.

Алгоритм	Операции чтения для восстановления страйпа неисправного диска с шириной K	Улучшение по сравнению с традиционными алгоритмами
XtremIO XDP	3K/4	–
RAID 1	1	Отсутствует
RAID 5	K	33 %
RAID 6	K	33 %

Примечание.
Более подробные сведения о системе защиты данных XDP см. в белой книге «Защита данных XtremIO».

Шифрование данных в состоянии покоя

Шифрование данных в состоянии покоя (DARE) защищает критически важные данные даже в случае извлечения диска из массива. В массивах XtremIO используется высокопроизводительный метод шифрования «на лету», который гарантирует невозможность чтения данных при извлечении твердотельного диска. Это предотвращает несанкционированный доступ в случае кражи или утери диска при транспортировке, а также позволяет выполнять операции возврата и обмена дисков, содержащих конфиденциальные данные.

Технология DARE обязательна к применению в ряде отраслей, включая здравоохранение (защита медицинских карт пациентов), банковском деле (обеспечение безопасности финансовых данных) и во многих государственных учреждениях.

Основа решения XtremIO DARE — технология твердотельных дисков с самошифрованием (SED). SED использует специальное оборудование, которое выполняет шифрование и дешифрование данных при записи на твердотельный диск и чтении с него. Перенос операций шифрования в твердотельный диск дает возможность использовать в XtremIO одну и ту же программную архитектуру вне зависимости от того, включено шифрование в массиве или отключено. Все функции и сервисы XtremIO, включая сокращение объемов данных «на лету», защиту данных XtremIO (XDP), «тонкое» выделение ресурсов и моментальные снимки, доступны как в зашифрованном, так и в незашифрованном кластере.

При изготовлении диска создается уникальный ключ шифрования данных (DEK). Ключ никогда не передается за пределы диска. Ключ DEK можно удалить или изменить, но в таком случае данные будет невозможно прочитать, и прежний ключ DEK восстановлению не подлежит. Чтобы гарантировать, что данные на диски SED записывают только авторизованные хосты, ключ DEK защищен ключом аутентификации (AK). Без этого ключа ключ DEK остается зашифрованным, и его нельзя использовать для шифрования и дешифрования данных.

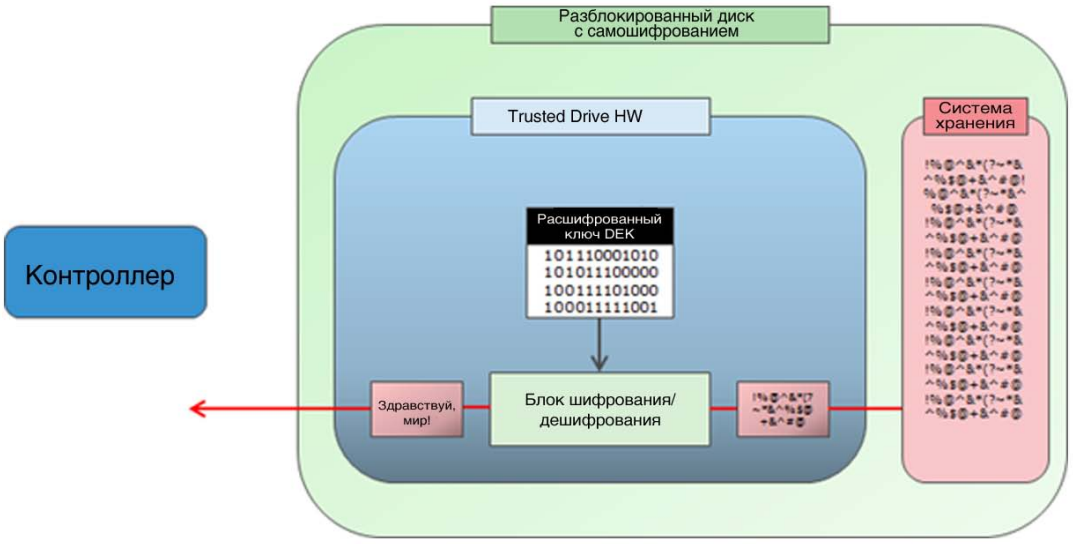


Рис. 12. Незаблокированный диск SED.

Диски SED поставляются с завода в незаблокированном состоянии, и доступ к данным на диске может осуществлять любой хост. На незаблокированных дисках данные всегда зашифрованы, но ключ DEK хранится в незашифрованном состоянии, и аутентификация не требуется.

Блокировка диска производится путем измерения его ключа AK, заданного по умолчанию, на новый конфиденциальный ключ AK, а также изменения настроек диска SED таким образом, чтобы он оставался заблокированным после перезагрузки или отключения питания (например, при извлечении твердотельного диска из массива). При извлечении твердотельного диска из массива этот диск отключается, и после загрузки ему потребуется ключ AK. Без правильного ключа AK данные на твердотельном диске невозможно прочитать, и они надежно защищены.

Чтобы получить доступ к данным, хосты должны предоставить правильный ключ AK, который разблокирует ключ DEK и разрешает доступ к данным. Иногда этот процесс называется «получение владения диском».

Диск SED разблокируется только во время загрузки и остается в разблокированном состоянии, пока работает массив. Поскольку шифрование или расшифровка данных выполняется во всех случаях аппаратными средствами, блокирование диска SED не влияет на производительность.

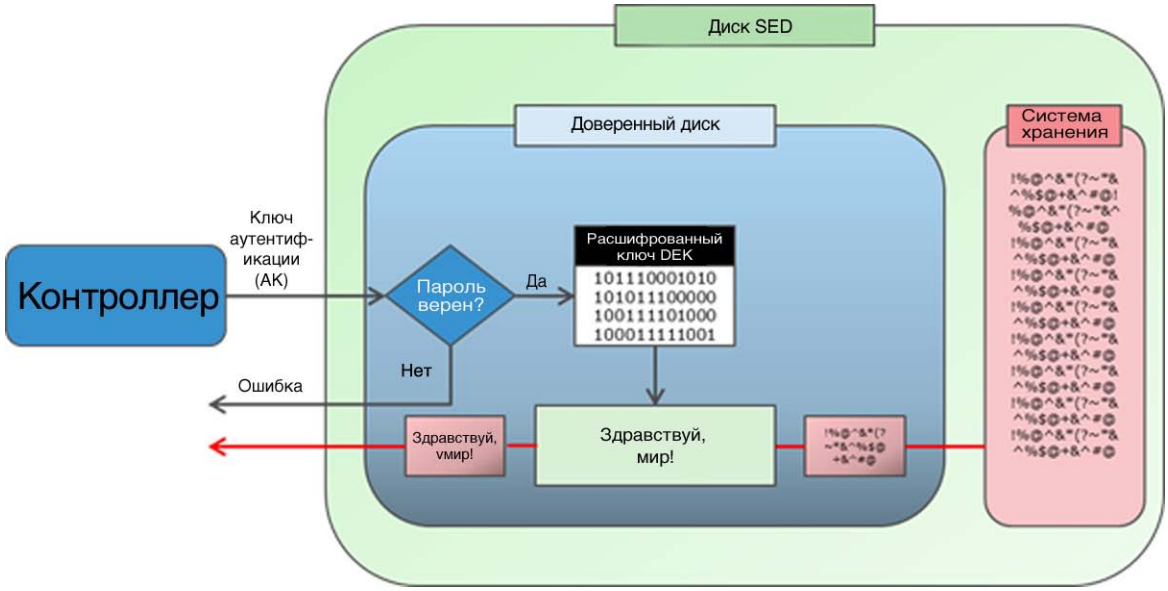


Рис. 13. Режим работы диска SED.

Массив на твердотельных дисках XtremIO выполняет шифрование данных на следующих твердотельных дисках:

- все твердотельные диски с пользовательскими данными;
- твердотельные диски контроллера системы хранения, которые могут содержать двоичные образы журналов пользовательских данных.

Снимки файловой системы

Моментальные снимки создаются путем фиксации состояния данных в томах в определенный момент времени и предоставления пользователям доступа к данным по мере необходимости, даже если исходный том был изменен. Моментальные снимки XtremIO — изначально доступны для записи, но их можно перевести в режим доступа «только для чтения», чтобы обеспечить неизменность данных. Моментальные снимки можно создавать из исходного тома или из любого снимка исходного тома.

Моментальные снимки можно использовать в ряде сценариев использования, в том числе следующих.

- Защита от логического повреждения данных

  XtremIO позволяет создавать снимки достаточно часто (в зависимости от требуемых интервалов создания целевых точек восстановления (RPO)) и использовать их для восстановления после любого повреждения логических данных. Моментальные снимки могут храниться в системе столько времени, сколько нужно. В случае повреждения логических данных восстановить состояние приложения на определенный момент времени можно, воспользовавшись последними моментальными снимками состояния приложения (до повреждения логических данных).

- Резервное копирование

  Созданные моментальные снимки можно предоставить серверу или агенту резервного копирования. Их можно использовать для выгрузки процесса резервного копирования с производственного сервера.

- Разработка и тестирование

  Система позволяет пользователю создавать снимки производственных данных и несколько (компактных и высокопроизводительных) копий производственной системы, а затем предоставлять их для разработки и тестирования.

- Клоны

  В системе XtremIO можно, используя постоянные перезаписываемые моментальные снимки, обеспечить функциональность, похожую на клонирование. Эти снимки можно использовать для предоставления клона производственного тома нескольким серверам. Производительность клона и производственного тома будет идентичной.

- Автономная обработка

  Моментальные снимки можно использовать в качестве средств переноса процесса обработки данных с производственного сервера. Например, если нужно запустить ресурсоемкий процесс обработки данных, который может повлиять на производительность производственного сервера, можно использовать моментальные снимки, чтобы создать актуальную копию производственных данных и подключить ее на другом сервере. После этого процесс может быть запущен на другом сервере без использования ресурсов производственного сервера.

Технология создания моментальных снимков XtremIO реализована с использованием возможностей учета содержания (сокращение объема данных «на лету»), оптимизированных для твердотельных дисков с уникальной древовидной структурой метаданных, которая направляет операции ввода-вывода к правильной метке времени данных. Благодаря этой технологии можно эффективно создавать моментальные снимки с сохранением высокой производительности при максимальном сроке службы носителей. Эффективность обеспечена как возможностью создания нескольких моментальных снимков, так количеством операций ввода-вывода, которые может поддерживать моментальный снимок.

При создании снимка система генерирует указатель на первичные метаданные (фактических данных в системе). Таким образом, создание моментального снимка является очень быстрым процессом, который не влияет на систему и не занимает место в памяти. В процессе создания моментальных снимков память расходуется, только если для внесения изменений необходимо записать новый уникальный блок.

Метаданные созданного снимка идентичны метаданным первичного тома. При записи нового блока в первичный том в системе метаданные первичного тома обновляются с учетом новой записи, а блок сохраняется в системе с помощью стандартного процесса записи. Этот блок не удаляется из системы после записи, пока он совместно используется снимками и первичным томом. Это касается записи в новом местоположении на томе (запись по неиспользованному адресу LBA) и перезаписи в уже записанном местоположении.

Для управления метаданными моментальных снимков и первичными метаданными в системе используется древовидная структура. В этой структуре моментальные снимки и первичные тома представлены в виде листьев, как показано на Рис. 14.

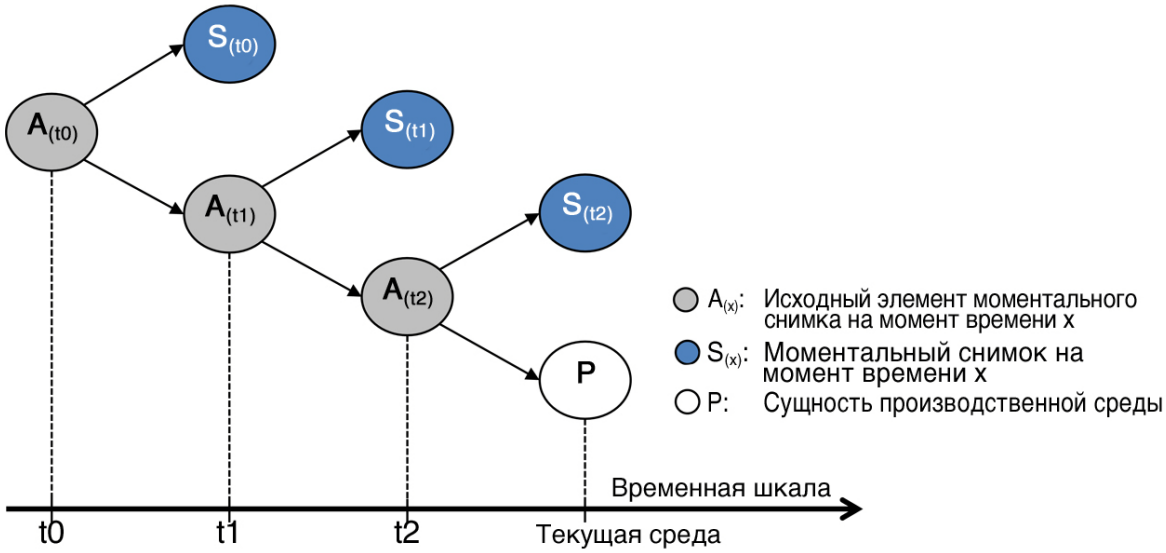


Рис. 14. Структура дерева метаданных.

Метаданные являются общими для всех блоков моментальных снимков, которые не были изменены (по сравнению с исходным моментальным снимком). Моментальный снимок содержит уникальные метаданные только для адреса LBA, блоки данных которых отличаются от первичных. Это обеспечивает экономное управление метаданными.

При создании нового моментального снимка в системе всегда создается два листа (два дочерних элемента) сущности, на основе которой создан моментальный снимок. Один из листьев соответствует моментальному снимку, а другой становится исходной сущностью. Сущность, на основе которой создан моментальный снимок, больше не используется напрямую и хранится в системе только для целей, связанных с управлением метаданными.

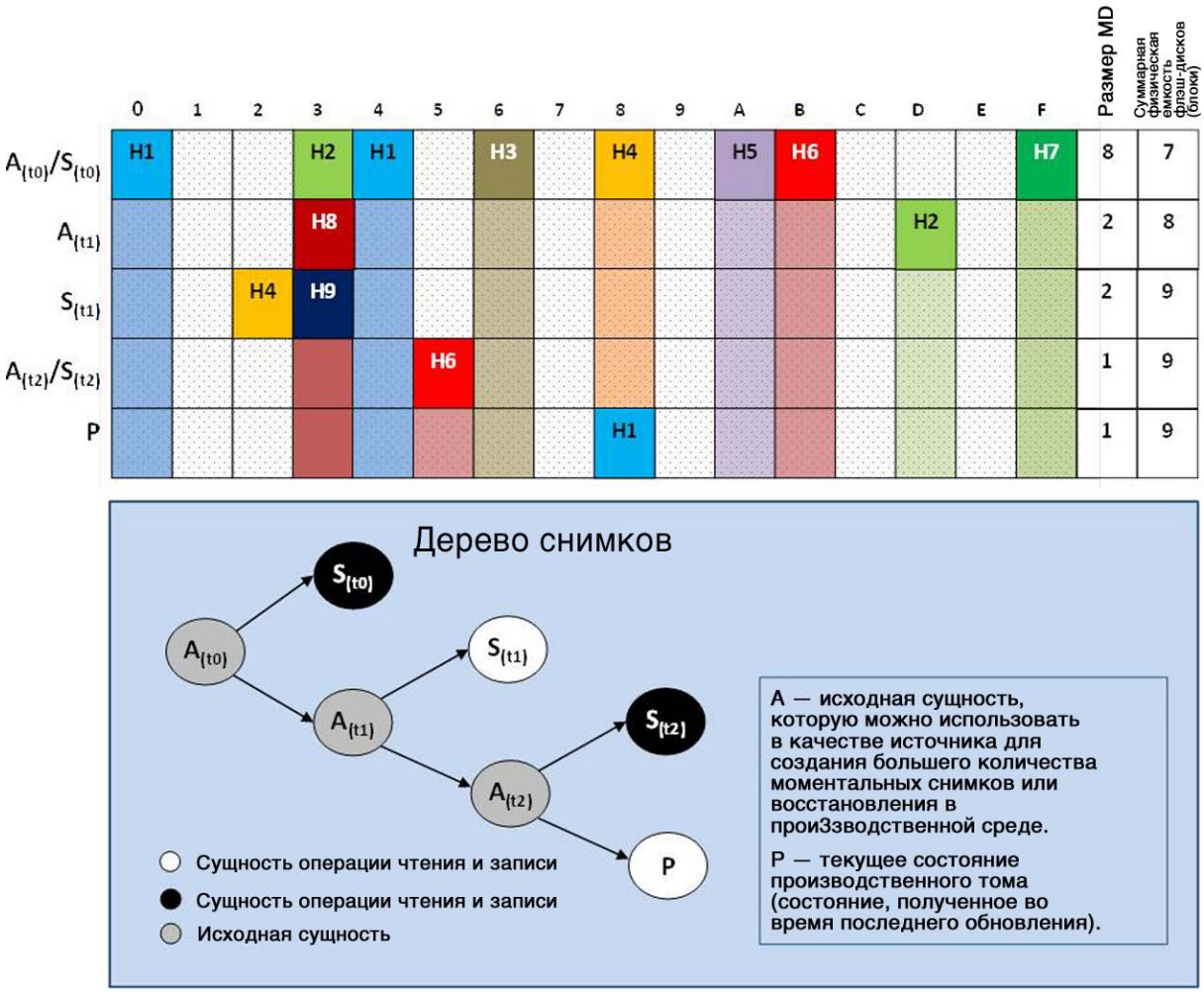


Рис. 15. Создание моментальных снимков.

На Рис. 15 изображен 16-блочный том в системе XtremIO. В первом ряду ($A_{(t0)/S}$(t0)) показан том во время создания первого моментального снимка (t0). В момент t0 у исходного элемента ($A_{(t0)}$) и моментального снимка ($S_{(t0)}$) данные и метаданные совпадают, потому что $S_{(t0)}$ — это моментальный снимок элемента $A_{(t0)}$ с доступом только для чтения (содержит те же данные, что и его исходный элемент).

Примечание.
Из 16 блоков используются только 8. В результате дедупликации блоки 0 и 4 занимают в физической памяти только один блок. Пустые точечные блоки представляют блоки с «тонким» выделением ресурсом, которые не занимают места в физической памяти.

На Рис. 15 перед созданием моментального снимка в момент $S_{(t1)}$ два новых блока записываются в P:

- H8 перезаписывает H2;
- H2 записывается в блок D, но этот фрагмент не занимает дополнительную физическую емкость, поскольку он является копией фрагмента H2, хранящегося в блоке 3 в $A_{(t0)}$.

$S_{(t1)}$ — это моментальный снимок с доступом для чтения и записи. Он содержит два дополнительных блока (2 и 3), которые отличаются от первичного элемента.

В отличие от традиционных моментальных снимков (в которых необходимо резервировать пространство для измененных блоков и хранить полную копию метаданных для каждого моментального снимка), массив XtremIO не требует зарезервированного места для моментальных снимков и не накапливает огромное количество метаданных.

Для моментального снимка XtremIO всегда требуются только уникальные метаданные, которые используются только для блоков, не используемых несколькими первичными элементами моментального снимка. Это позволяет системе эффективно поддерживать большое количество моментальных снимков, используя небольшое количество служебных протокольных данных системы, которые являются динамическими и пропорциональными количеству изменений в сущностях.

Например, в момент времени t2 блоки 0, 3, 4, 6, 8, A, B, D и F являются общими для первичных сущностей. Для этого моментального снимка уникален только блок 5. Поэтому в системе XtremIO используется только один блок метаданных. Остальные блоки являются общими для первичных элементов, и для получения правильных данных и структуры тома к ним применяется структура данных первичного элемента.

Система поддерживает создание моментальных снимков для наборов томов. Все моментальные снимки из томов в наборе согласованы между собой и содержат идентичное время для всех томов. Эти моментальные снимки можно создать вручную, выбрав набор томов для создания моментальных снимков или поместив том в контейнер группы консистентности и создав моментальный снимок этой группы.

Создание моментального снимка не влияет на производительность системы или ее общее время отклика (производительность остается неизменной). Производительность не зависит от количества моментальных снимков в системе или размера дерева моментальных снимков.

Удаление моментальных снимков выполняется легко, а объем удаляемых данных пропорционален количеству измененных блоков, которыми отличаются сущности. Для обработки операций удаления моментальных снимков в системе используется функциональность учета содержимого. У каждого блока данных есть счетчик, который указывает количество экземпляров этого блока в системе. После удаление блока значение счетчика уменьшается на единицу. Любой блок, значение счетчика которого равно нулю (логический адрес блока (LBA) отсутствует во всех относящихся к этому блоку томах или моментальных снимках в системе), перезаписывается при помощи технологии XDP при попадании новых уникальных данных в систему.

Для удаления дочернего элемента, у которого нет своих дочерних элементов, дополнительная обработка системой не требуется.

Удаление моментального снимка в середине дерева вызывает асинхронный процесс. Этот процесс объединяет метаданные удаленных дочерних элементов сущности с метаданными прародителя, что обеспечивает отсутствие фрагментации в структуре дерева.

В массиве XtremIO каждый блок, который необходимо удалить, тотчас же будет обозначен как освобожденный. Таким образом, в системе не выполняется сборка мусора, а для поиска и удаления изолированных блоков не требуется выполнять сканирование. Кроме того, при использовании массива XtremIO удаление моментальных снимков не влияет на производительность системы и износостойкость твердотельных дисков.

Реализация функции моментального снимка полностью основана на метаданных и использует функцию сокращения объема данных массива «на лету» для предотвращения копирования данных в пределах массива. Это обеспечивает поддержку большого количества моментальных снимков.

Снимки системы XtremIO обладают следующими преимуществами:

- отсутствие потребности в резервировании пространства для моментальных снимков;
- возможность создания неизменных копий и/или доступных для записи клонов исходного тома;
- мгновенное создание;
- незначительное влияние на производительность исходного тома и самого моментального снимка.

Примечание.
Более подробные сведения о моментальных снимках см. в белой книге «Моментальные снимки XtremIO».

Масштабируемая производительность

Структура массива XtremIO позволяет горизонтально масштабировать его в соответствии с будущими требованиями к производительности и емкости не только для новых, но и для уже развернутых приложений. Архитектура системы XtremIO позволяет увеличить производительность и емкость путем добавления строительных блоков (модулей X-Brick), сохраняя при этом единый центр управления и балансировку ресурсов по всей системе.

Горизонтальное масштабирование является неотъемлемым свойством архитектуры XtremIO и может быть выполнено без полной модернизации имеющегося оборудования и замедления передачи данных.

Если требуется дополнительная производительность или емкость, система хранения данных XtremIO может масштабироваться путем добавления модулей X-Brick. Несколько модулей X-Brick объединяются вместе с помощью высокодоступной сети InfiniBand с резервированием и очень низким временем отклика.

При расширении системы балансировка ресурсов сохраняется, а данные в массиве равномерно распределяются по всем модулям X-Brick для сохранения согласованной производительности и равномерного износа твердотельных дисков.

Расширение системы осуществляется без какой-либо настройки или перемещения томов вручную.* Система XtremIO использует алгоритм согласованного создания «отпечатков», который сводит к минимуму необходимость повторного сопоставления. Новый модуль X-Brick добавляется во внутреннюю схему балансировки нагрузки, и в новую дисковую полку передаются только имеющиеся актуальные данные.

Емкость и производительность масштабируются линейно, то есть конфигурация с двумя блоками X-Brick обеспечивает в два раза больше операций ввода-вывода в секунду, с четырьмя блоками X-Brick — в четыре раза больше операций, а с шестью блоками X-Brick — в шесть раз больше операций ввода-вывода в секунду, чем конфигурация с одним блоком X-Brick. Тем не менее, при масштабировании системы значение времени отклика остается стабильно низким (менее 1 мс), как показано на Рис. 16.

* В текущей версии программного обеспечения динамическое горизонтальное масштабирование не поддерживается.

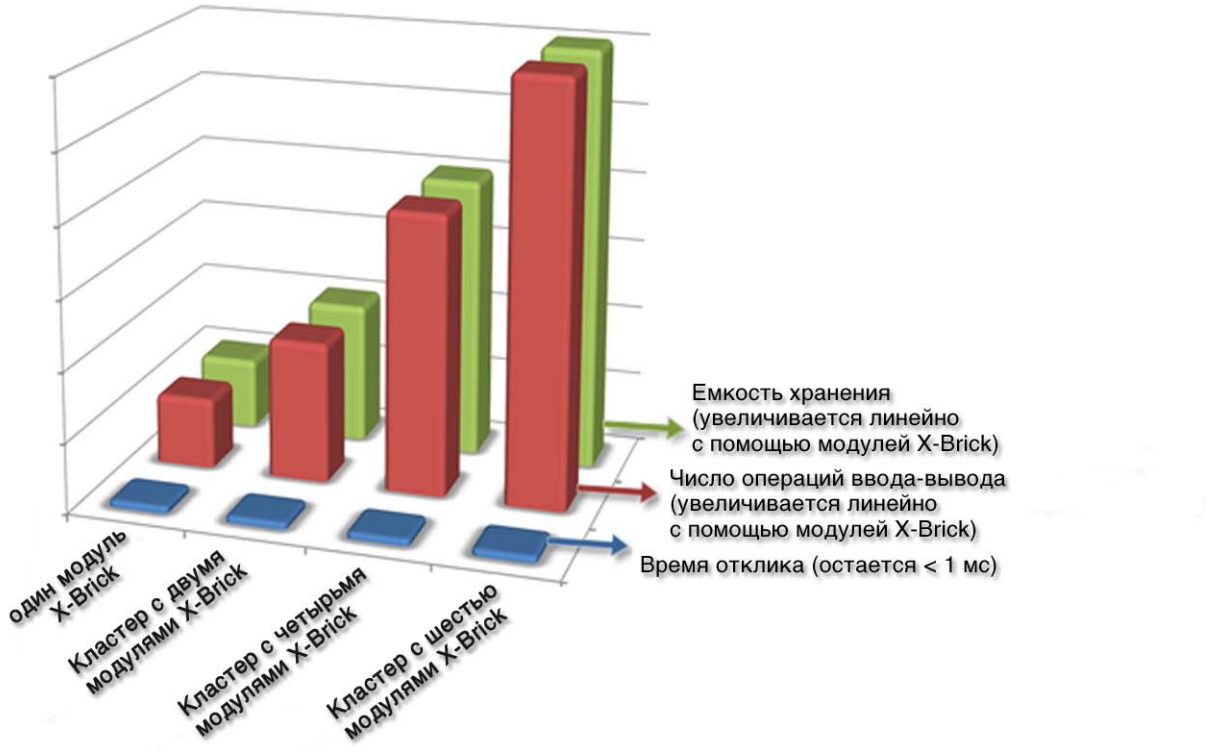


Рис. 16. Линейная масштабируемость производительности со стабильно низким значением времени отклика.

Поскольку система XtremIO специально разработана с учетом возможности масштабирования, в ее программном обеспечении нет внутренних ограничений на размер кластера.* В архитектуре системы также используется наиболее эффективный подход к обеспечению надлежащего времени отклика. Программное обеспечение обладает модульной структурой. На каждом контроллере системы хранения данных запущены разные наборы модулей. Общая нагрузка распределяется между всеми контроллерами. Эти распределенные программные модули (на разных контроллерах системы хранения данных) обрабатывают отдельные транзитные операции ввода-вывода по кластеру. Система XtremIO обрабатывает каждый запрос ввода-вывода по двум программным модулям (в 2 шага), независимо от того, используется ли система из одного блока X-Brick или кластер из нескольких блоков X-Brick. Таким образом, значение времени отклика всегда остается неизменным независимо от размера кластера.

Примечание.
Время отклика менее одной миллисекунды подтверждено фактическими результатами тестирования и определяется для наихудшего сценария работы.†

* Максимальный размер кластера зависит от тестируемых и поддерживаемых в данный момент конфигураций.
† Задержка менее 1 мс характерна для блоков типовых размеров. В случае малых или крупных блоков задержка может быть выше.

Сеть InfiniBand играет важную роль в архитектуре системы XtremIO. В массиве XtremIO используется два типа связи через коммутационную панель InfiniBand: вызов удаленных процедур (RPC) для управляющих сообщений и удаленный прямой доступ к памяти (RDMA) для перемещения блоков данных.

Сеть InfiniBand отличается не только очень высокой пропускной способностью по сравнению с любыми технологиями соединений (40 Гбит/с для одного QDR-подключения), но и самым низким значением времени отклика. Задержка «запрос-отклик» при передаче блока данных между двумя контроллерами системы хранения данных XtremIO с помощью RDMA составляет около 7 микросекунд. По сравнению с допустимым значением времени отклика 500 микросекунд на каждую операцию ввода-вывода указанным значением времени практически можно пренебречь. Благодаря этому программное обеспечение может выбирать любые необходимые ресурсы контроллеров системы хранения данных и твердотельных дисков, независимо от того, являются ли они локальными или удаленными (через сеть InfiniBand) по отношению к контроллеру системы хранения данных, на который поступает операция ввода-вывода.

Все корпоративные функциональные возможности XtremIO (включая сокращение объема данных на лету, снимки, XDP, высокую доступность и т. д.) были разработаны с учетом масштабируемости архитектуры. Все данные и метаданные равномерно распределены по всему кластеру. Операции ввода-вывода поступают в массив через все серверные порты с использованием решения по управлению путями ввода-вывода и зон сети хранения данных. Таким образом, возникновение в системе узких мест по производительности практически невозможно, так как все рабочие нагрузки равномерно распределяются между контроллерами и твердотельными дисками.

Ниже приведен список преимуществ системы XtremIO.

- Процессоры, ОЗУ, твердотельные диски и порты подключения масштабируются все одновременно, обеспечивая масштабируемость производительности с идеальной балансировкой.
- Внутренняя связь осуществляется с помощью высокодоступной внутренней фабрики InfiniBand (40 Гбит/с, QDR).
- Для кластера применяется модель резервирования N-узлов в режиме «активный», обеспечивающая возможность доступа к тому с любого серверного порта любого контроллера системы хранения данных на любом модуле X-Brick с эквивалентной производительностью.
- Доступ к данным без копирования посредством RDMA обеспечивает эквивалентность операций ввода-вывода на локальных или удаленных твердотельных дисках, независимо от размера кластера.
- Данные балансируются по всем модулям X-Brick по мере расширения системы.

- Обеспечивается более высокий уровень резервирования, а сам кластер более устойчив к отказам при возникновении аппаратных и программных сбоев. Если в кластере с моделью резервирования N-узлов в режиме «активный» один контроллер СХД выходит из строя, система теряет только 1/N-ю часть общей производительности.
- Систему просто модернизировать и, в отличие от традиционных двухконтроллерных систем, горизонтально масштабируемая модель XtremIO дает заказчикам возможность наращивать емкость и производительность системы хранения по мере увеличения рабочей нагрузки.

Равномерное распределение данных

Для внешних приложений система XtremIO по своему поведению и характеристикам выглядит, как стандартный блочный массив хранения данных. Благодаря уникальной архитектуре в массиве применяется принципиально иной подход к внутренней организации данных. При определении места, в которое необходимо поместить блоки, система XtremIO вместо логических адресов использует содержание этих блоков.

Система XtremIO использует блоки данных для внутренних операций. Во время операции записи в массив блоки данных, размер которых превышает стандартный, разбиваются на фрагменты стандартного размера. Используя специальный математический алгоритм, система вычисляет уникальный «отпечаток» для каждого входящего блока данных.

Этот уникальный идентификатор используется для двух основных задач:

- определение местоположения блока данных в массиве;
- сокращение объема данных «на лету» (см. стр. 23).

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. 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. Массив XtremIO естественным образом обеспечивает равномерное распределение данных, размещая блоки по их уникальным идентификаторам (см. Рис. 7 на стр. 18).

В системе XtremIO используются следующие метаданные:

- сопоставление логического адреса (LBA) с идентификатором отпечатка;
- сопоставление идентификаторов «отпечатков» и физических местоположений;
- данные счетчика ссылок для каждого идентификатора «отпечатков».

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.

Благодаря хранению всех метаданных в оперативной памяти массив XtremIO обладает рядом уникальных преимуществ.

- Отсутствие поиска на твердотельном диске

  Благодаря отсутствию операций поиска на твердотельных дисках увеличивается число серверных операций ввода-вывода.

- Моментальный снимок файловой системы

  Операции создания снимка выполняются мгновенно, поскольку процесс полностью выполняется в памяти (см. стр. 32).

- Мгновенное клонирование виртуальных машин

  Сокращение объема данных на лету и интерфейс VAAI в сочетании с обработкой метаданных в оперативной памяти позволяют системе XtremIO клонировать виртуальные машины исключительно путем выполнения операций в памяти.

- Стабильная производительность

  Физическое местоположение данных, большие тома и широкие диапазоны адресов LBA никак не влияют на производительность системы.

Высокая доступность

Одними из основных особенностей архитектуры массива хранения XtremIO на твердотельных дисках являются предотвращение потери данных и поддержание обслуживания при множественных сбоях.

С точки зрения оборудования ни один из компонентов не является критической точкой отказа. Каждый контроллер системы хранения данных, дисковая полка и коммутатор InfiniBand системы оснащены двумя блоками питания. Система также оснащена двумя батареями аварийного питания, двумя сетевыми портами и портами данных (в каждом контроллере системы хранения данных). Два коммутатора InfiniBand соединяются перекрестно и образуют две фабрики данных. В системе осуществляется постоянный мониторинг мощности на входе и различных путей передачи данных, и при каждом сбое предпринимается попытка восстановления или аварийного переключения на резервный ресурс.

Архитектура программного обеспечения построена аналогичным образом. Каждая не помещенная на твердотельный диск часть информации сохраняется в нескольких местоположениях, называемых журналами. У каждого программного модуля есть свой журнал, который хранится на другом контроллере системы хранения данных. В случае неожиданного сбоя такой журнал можно использовать для восстановления данных.

Журналы считаются очень важными, поэтому они всегда хранятся в контроллерах системы хранения данных, оснащенных резервными источниками питания. В случае возникновения проблем с батареей аварийного питания журнал переключается на другой контроллер системы хранения данных. В случае глобального сбоя электропитания батареи аварийного питания гарантируют дальнейшую запись всех журналов на диски хранилища в контроллерах системы хранения данных, а также возможность дальнейшего функционирования системы.

Кроме того, благодаря масштабируемой архитектуре и алгоритму защиты данных XDP все модули X-Brick предварительно настроены как единая группа резервирования. Благодаря этому нет необходимости выбирать, настраивать и регулировать группы резервирования.

В архитектуру XtremIO с режимом работы «активный-активный» изначально заложена возможность обеспечения максимальной производительности и согласованного времени отклика. Система оснащена механизмом самовосстановления, который выполняет попытки восстановления после сбоев и полностью возобновляет функциональность. Попытка перезапуска вышедшего из строя компонента предпринимается однократно перед аварийным переключением на резервный ресурс. Аварийное переключение контроллера системы хранения данных на резервный ресурс выполняется в самом крайнем случае. В зависимости от типа сбоя в системе выполняются попытки аварийного переключения соответствующего программного компонента на резервный ресурс с сохранением работоспособности остальных компонентов, что позволяет свести к минимуму влияние на производительность. Весь контроллер системы хранения данных переключается на резервный ресурс, только если не удалось успешно выполнить восстановление, или если в системе должны предприниматься все возможные меры для защиты от потери данных.

Если компонент, который был временно недоступен, восстанавливается, в системе выполняется восстановление после сбоя. Этот процесс осуществляется на уровне компонента программного обеспечения или контроллера системы хранения данных. Предусмотренный в массиве механизм защиты от поспешного возврата предохраняет систему от восстановления после сбоя на нестабильном или обслуживаемом компоненте.

Система XtremIO построена на основе стандартного оборудования, поэтому ее возможности не ограничиваются одним лишь аппаратным обнаружением ошибок. В ней используется проприетарный алгоритм, обеспечивающий обнаружение, исправление и маркировку поврежденных областей. Любой сценарий повреждения данных, который не обрабатывается автоматически на твердотельном диске, реализуется с помощью механизма XDP в массиве или нескольких копий, хранящихся в журналах. В качестве безопасного и надежного механизма обеспечения целостности данных во время выполнения операций чтения используется «отпечаток» содержимого, позволяющий избежать ошибок из-за незаметного повреждения данных. В случае несоответствия в ожидаемом отпечатке данные в массиве восстанавливаются путем повторного считывания или восстановления из группы резервирования XDP.

Обновление без прерывания работы

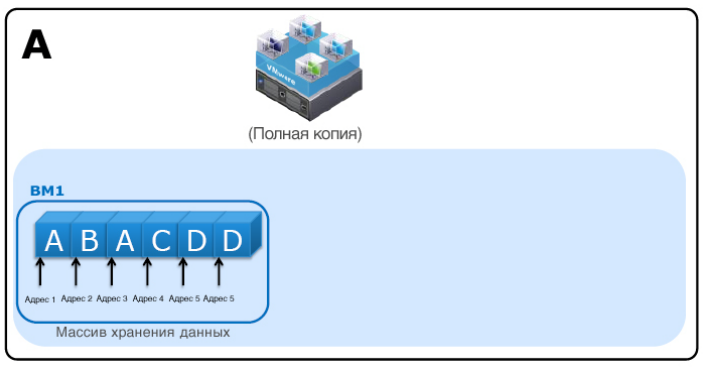
Во время модернизации без прерывания работы ОС XtremIO соответствующие процедуры в системе выполняются на активном кластере. При этом модернизируются все контроллеры системы хранения данных в кластере, а также выполняется перезапуск работающих приложений, который занимает менее 10 с. Поскольку основное ядро Linux остается активным на протяжении всего процесса модернизации, при перезапуске приложения хосты не обнаруживают пути отключения.

В редких случаях модернизации ядра Linux или встроенного ПО весь массив XtremIO на твердотельных дисках можно модернизировать без прерывания обслуживания и риска потери данных. Процедура модернизации без прерывания работы запускается с управляющего сервера XtremIO и может предусматривать модернизацию ПО XtremIO, основной ОС и встроенного ПО.

Во время модернизации Linux или встроенного ПО без прерывания работы система автоматически переключается на другой компонент и выполняет модернизацию ПО. После завершения модернизации и верификации состояния компонента система выполняет процедуру восстановления на этом компоненте, а затем этот процесс повторяется на других компонентах. Во время модернизации система полностью доступна, не происходит потери данных и влияние на производительность остается минимальным.

Интеграция с VMware VAAI

Для улучшения серверного клонирования виртуальных машин был внедрен программный интерфейс vSphere Storage API for Array Integration (VAAI). Чтобы без VAAI клонировать полную виртуальную машину, хосту нужно поочередно считать все блоки данных и записать их на новые адреса, где находится клонированная виртуальная машина, как показано на Рис. 17. Это затратная операция, которая нагружает хост, массив и сеть хранения данных (SAN).



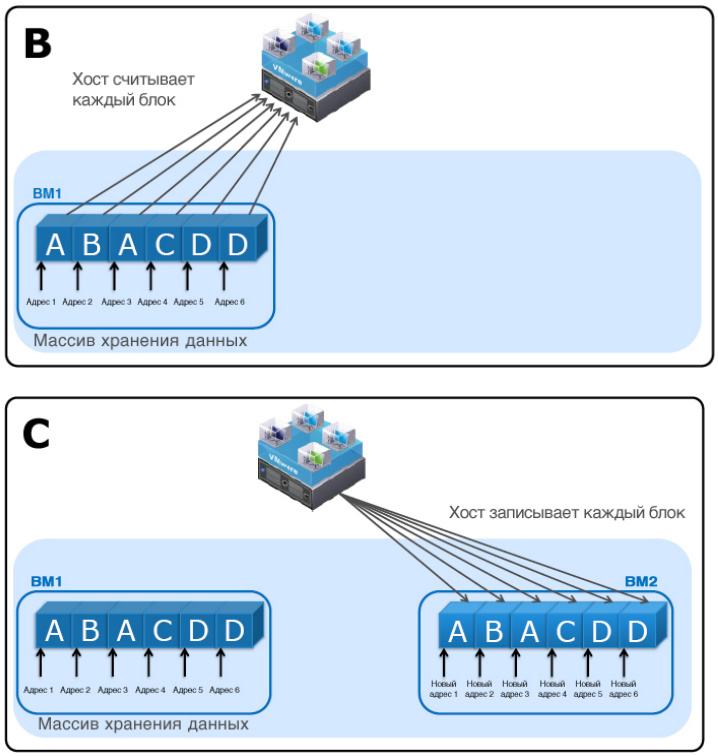


Рис. 17. Полное копирование без использования VAAI.

При использовании VAAI рабочая нагрузка во время клонирования виртуальной машины переносится на массив хранения. Хосту нужно только отдать команду «X-copy», и массив скопирует блоки данных на адрес новой виртуальной машины, как показано на Рис. 18. Этот процесс сокращает нагрузку на ресурсы хоста и сети. Тем не менее, процесс по-прежнему потребляет ресурсы массива хранения.

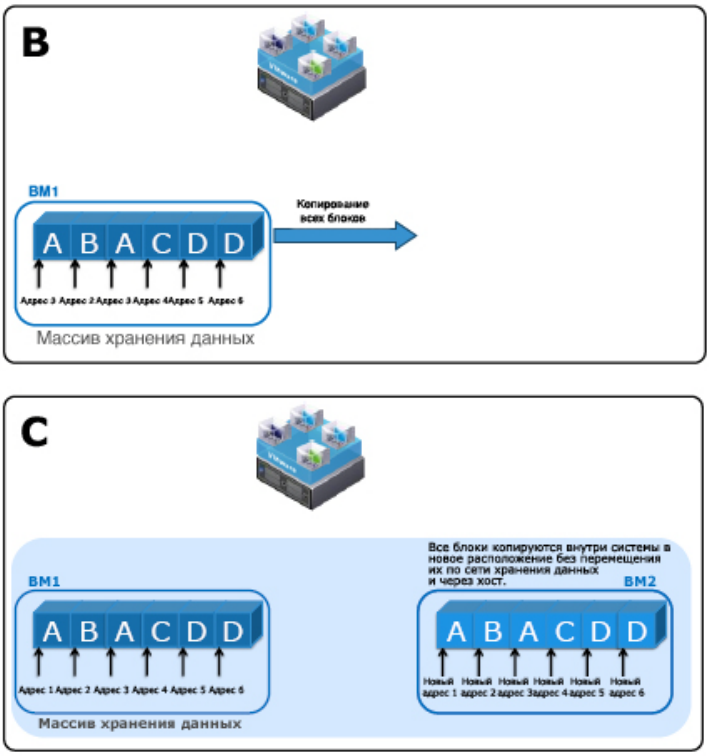


Рис. 18. Полное копирование при использовании VAAI.

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. Система только создает указатели на существующие данные, и весь процесс осуществляется в памяти контроллера системы хранения данных, как показано на Рис. 19. Поэтому ресурсы массива хранения данных не используются, а значит, влияние на производительность системы отсутствует.

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

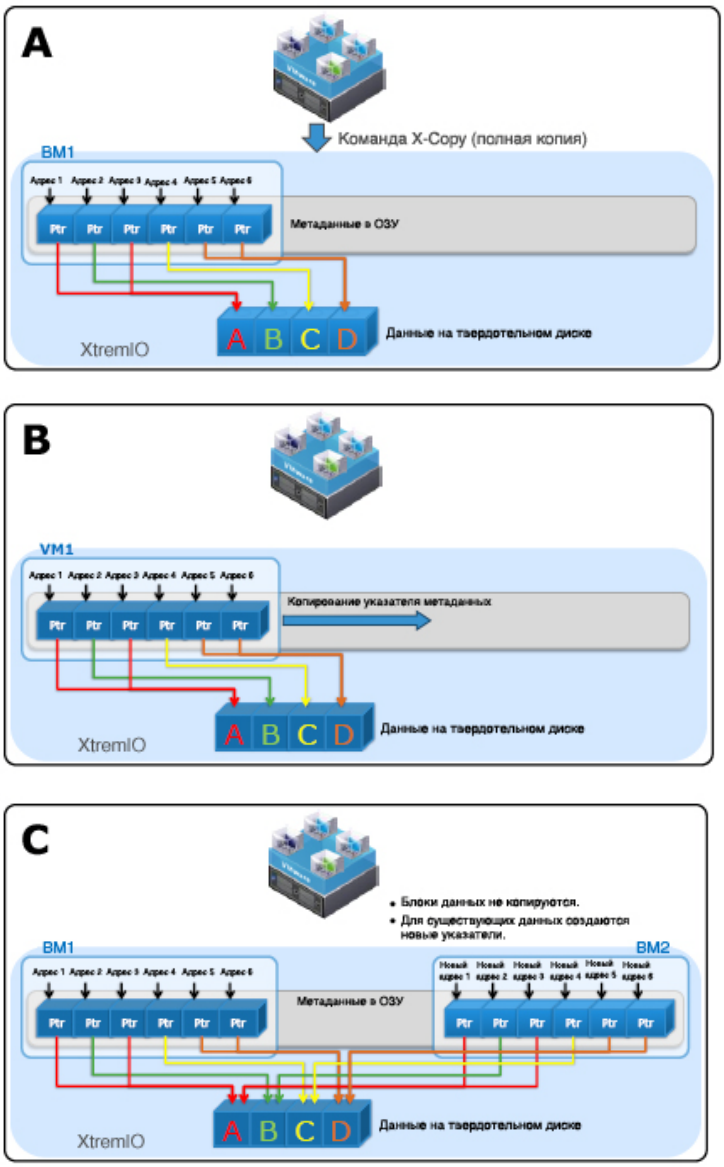


Рис. 19. Полное копирование при использовании XtremIO.

Это стало возможным только благодаря таким функциям XtremIO, как сокращение данных на лету и обработка метаданных в оперативной памяти. На других флэш-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).

Управляющий сервер XtremIO (XMS)

Сервер XMS позволяет контролировать систему и управлять ею, в том числе:

- создавать, инициализировать и форматировать новые системы;
- осуществлять мониторинг событий и состояния системы;
- осуществлять мониторинг производительности системы;
- обслуживать базу данных журналов статистики производительности;
- предоставлять клиентам службы графического интерфейса пользователя и интерфейса командной строки;
- реализовать логику управления томами и операций с группами защиты данных;
- обслуживать систему (останавливать, запускать и перезагружать).

Сервер XMS предустановлен с графическим интерфейсом пользователя и интерфейсом командной строки. XMS можно установить на выделенном физическом сервере в центре обработки данных или в качестве виртуальной машины на платформе VMware.

Для сервера XMS необходим доступ ко всем портам управления на контроллерах системы хранения данных X-Brick. Кроме того, он сам должен быть доступен для любого клиентского хоста с графическим интерфейсом пользователя и интерфейсом командной строки. Поскольку для обмена данными всегда используются стандартные подключения TCP/IP, сервер XMS может быть расположен в любом месте, которое соответствует вышеуказанным требованиям к подключениям.

Поскольку сервер XMS не находится на пути передачи данных, его отключение от кластера XtremIO никак не повлияет на операции ввода-вывода. Сбой на сервер XMS влияет только на операции мониторинга и настройки конфигурации, такие как создание и удаление томов. Тем не менее, при использовании виртуальной топологии XMS с такими сбоями можно легко справиться, воспользовавшись преимуществами функций высокой доступности VMware vSphere.

Графический интерфейс пользователя системы

На Рис. 20 показана взаимосвязь между графическим интерфейсом пользователя и другими компонентами сети.

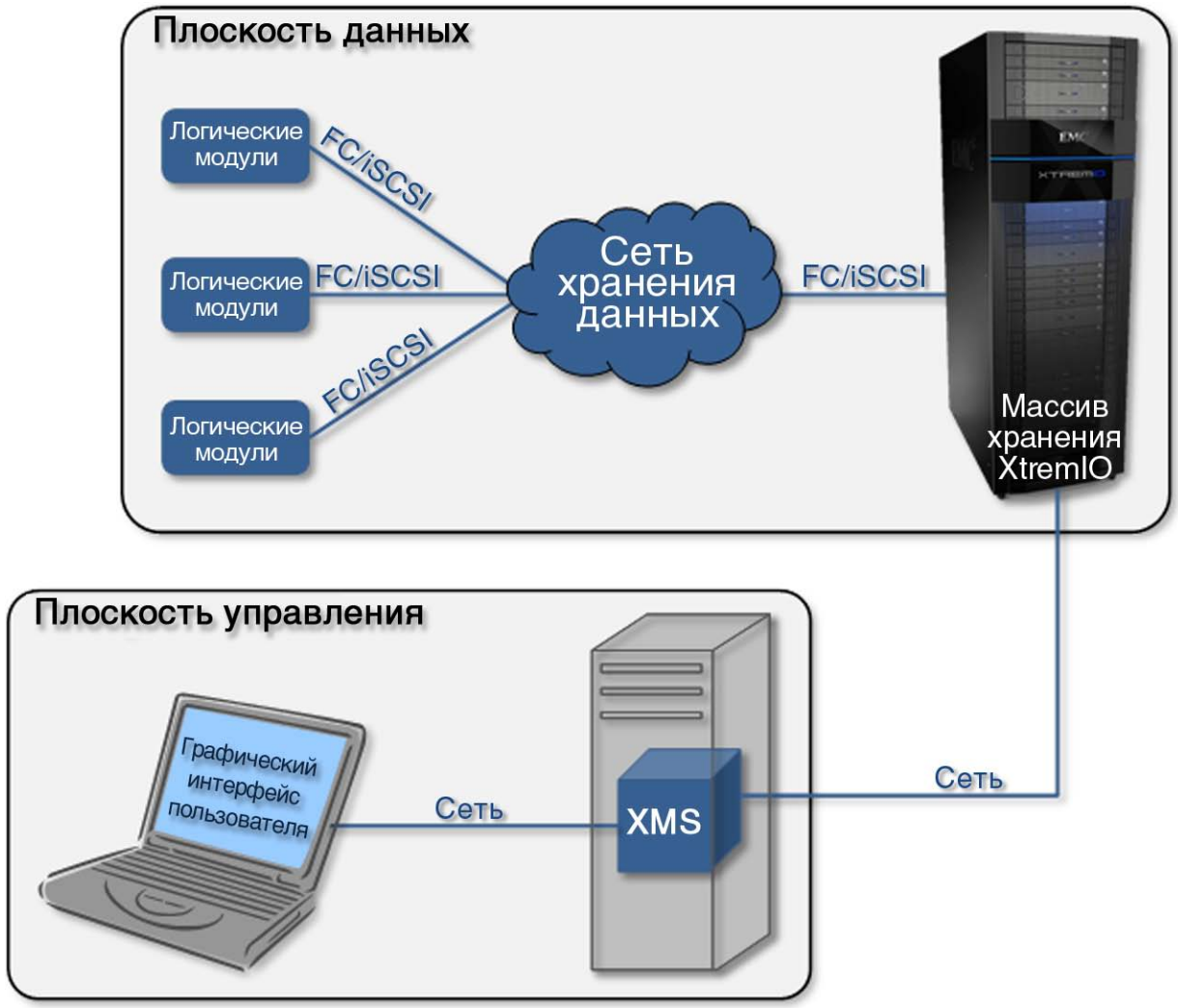


Рис. 20. Взаимосвязь между графическим интерфейсом пользователя и другими сетевыми компонентами.

Графический интерфейс пользователя системы реализован с помощью клиента Java. ПО клиента графического интерфейса пользователя взаимодействует с сервером XMS, используя стандартные протоколы TCP/IP, и может использоваться в любом месте, из которого клиент может получить доступ к серверу XMS.

Графический интерфейс пользователя предоставляет удобные в использовании инструменты для выполнения большинства системных операций (определенные операции управления все же должны выполняться с помощью интерфейса командной строки). Кроме того, операции над несколькими компонентами, такие как создание нескольких томов, могут быть выполнены только с помощью графического интерфейса пользователя.

На Рис. 21 показана привязка тома к группе инициаторов всего за несколько шагов при помощи графического интерфейса пользователя.

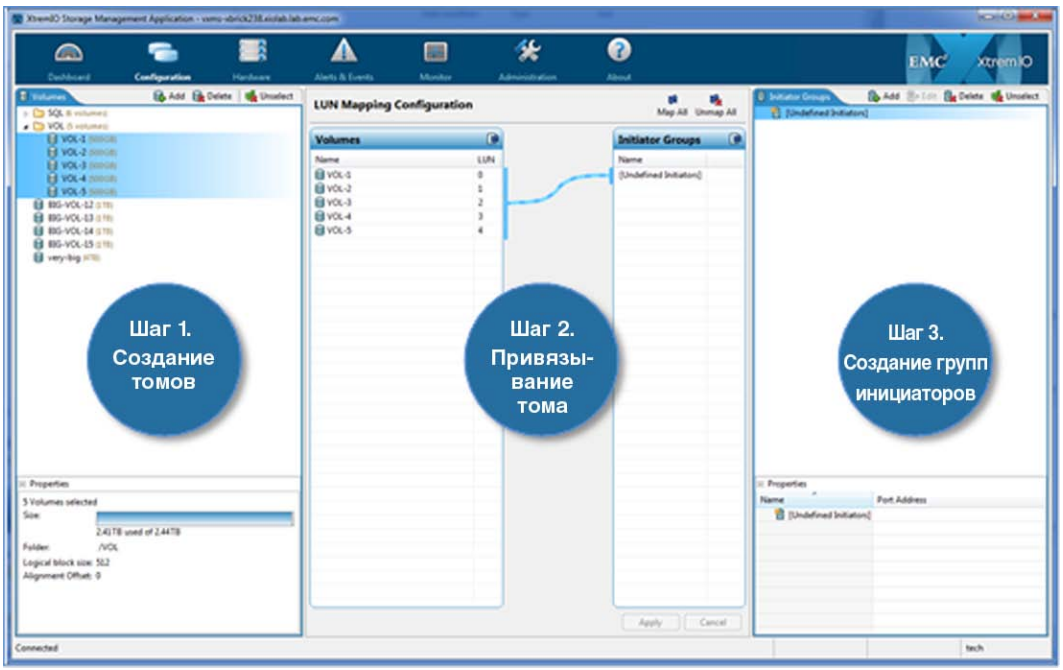


Рис. 21. Привязка томов к группам инициаторов с помощью графического интерфейса пользователя.

На Рис. 22 показана панель управления в графическом интерфейсе пользователя, которая позволяет пользователю контролировать хранение, производительность, оповещения и состояние оборудования системы.

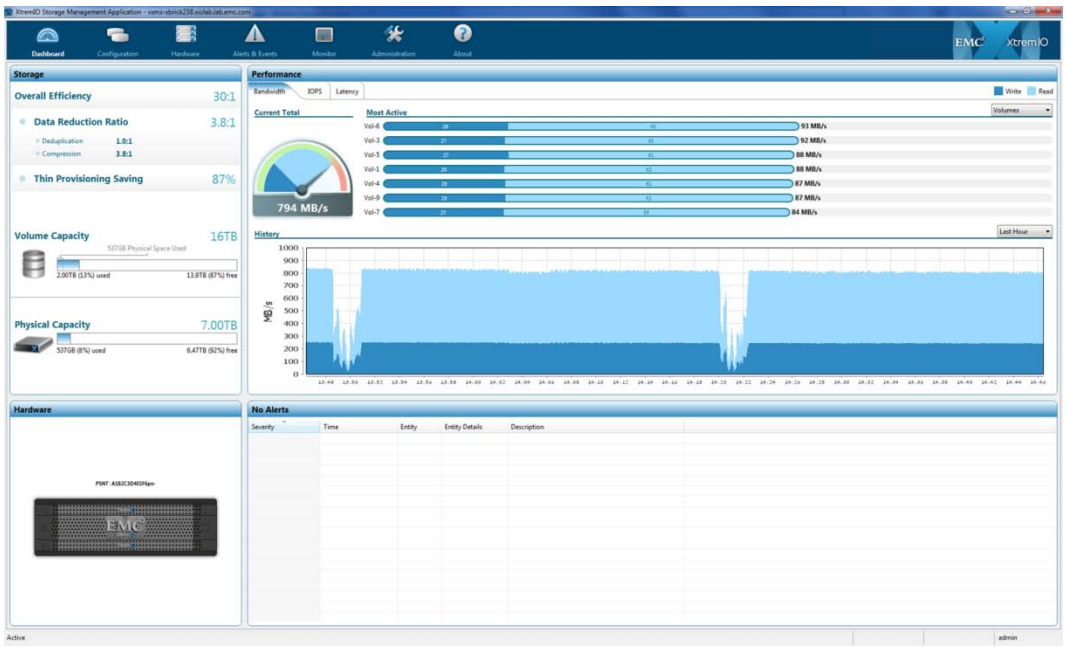


Рис. 22. Мониторинг системы с помощью графического интерфейса пользователя.

Интерфейс командной строки

Интерфейс командной строки системы (CLI) позволяет администраторам и другим пользователям системы выполнять поддерживаемые операции управления. Этот интерфейс предварительно установлен на сервере XMS. К нему можно получить доступ с помощью стандартного протокола SSH.

Доступен также клиентский пакет CLI, который обращается к управляющему серверу XMS через стандартное подключение по протоколу TCP/IP и может быть установлен на хосте Linux CentOS с доступом к XMS.

Программный интерфейс RESTful API

Программный интерфейс RESTful API массива XtremIO позволяет использовать интерфейс на основе подключения по протоколу HTTP для автоматизации, координации, выполнения запросов и выделения ресурсов системы. Программный интерфейс (API) позволяет использовать приложения сторонних производителей, чтобы полностью управлять массивом и осуществлять его администрирование. Таким образом, с его помощью можно разрабатывать гибкие решения для управления массивом XtremIO.

LDAP/LDAPS

Массив хранения данных XtremIO поддерживает аутентификацию пользователей посредством протокола LDAP как из графического интерфейса, так и из командной строки. После настройки аутентификации по протоколу LDAP сервер XMS перенаправляет процесс аутентификации пользователей на сервер с настроенным протоколом LDAP или Active Directory (AD) и разрешает доступ только аутентифицированным пользователям. Пользовательские разрешения для сервера XMS определяются на основе соответствия между группами пользователей в LDAP/AD и ролями XMS.

Функциональность конфигурации LDAP позволяет выполнять аутентификацию внешних пользователей на сервере XMS с использованием одного или нескольких серверов.

Операция LDAP выполняется только один раз при входе внешнего пользователя на сервер XMS со своими учетными данными. Сервер XMS выступает в качестве клиента LDAP и подключается к соответствующему сервису на внешнем сервере. После этого выполняется операция «LDAP Search» с использованием предварительно заданного профиля конфигурации LDAP и учетных данных внешнего пользователя.

В случае успешной аутентификации внешний пользователь входит на сервер XMS и получает доступ к его полной или ограниченной функциональности (согласно роли XMS, назначенной группе пользователей LDAP).

Массив XtremIO также поддерживает безопасную аутентификацию посредством протокола LDAPS.

Простота управления

Массив XtremIO очень прост в настройке и удобен в управлении. Кроме того, он не требует настройки или тщательного планирования.

Пользователю массива XtremIO не нужно выбирать между различными вариантами RAID для оптимизации системы. При инициализации системы функция XDP (см. стр. 27) уже настроена как единая группа резервирования. Все пользовательские данные распределены по всем модулям X-Brick. Кроме того, не нужно настраивать многоуровневое хранение и производительность. Все операции ввода-вывода обрабатываются одинаково При создании все тома привязываются ко всем портам (FC и iSCSI), и в массиве не выполняется многоуровневое хранение данных. Это избавляет от необходимости настраивать производительность и параметры оптимизации вручную, а также упрощает управление системой, ее настройку и использование.

Службы XtremIO предоставляют следующее:

- минимум планирования;
  - не нужна настройка конфигурации RAID;
  - минимум усилий для определения конфигурации при клонировании и создании моментальных снимков;
- без многоуровневого хранения;
  - одноуровневый массив на твердотельных дисках;
- не нужна настройка производительности;
  - отсутствие зависимости от схемы доступа для выполнения операций ввода-вывода, показателя попаданий в кэш-память, решений многоуровневого хранения данных и т. д.

Интеграция с другими продуктами EMC

Массив XtremIO прекрасно интегрируется с другими продуктами EMC. В последующих выпусках XtremIO будет добавлено больше точек интеграции, благодаря чему этот массив станет еще более ценным для заказчиков EMC.

PowerPath

EMC PowerPath — это программное обеспечение на базе хостов, которое обеспечивает автоматизированное управление путями данных и возможности балансировки нагрузки для разнородных серверных, сетевых ресурсов и ресурсов хранения данных, развернутых в физических и виртуальных средах. Это ПО позволяет пользователям обеспечить соблюдение уровней обслуживания благодаря высокой доступности и производительности приложений. PowerPath автоматизирует переключение на резервный путь и операции восстановления для обеспечения высокой доступности в случае ошибки или сбоя, а также оптимизирует производительность благодаря балансировке нагрузки, связанной с операциями ввода-вывода по нескольким путям. Система XtremIO поддерживается программным обеспечением PowerPath как непосредственно, так при виртуализации системы XtremIO, использующей решения VPLEX.

VPLEX

Семейство EMC VPLEX — это следующее поколение устройств, которые обеспечивают мобильность данных и доступ к ним как внутри центров обработки данных, так и между ними. Платформа предоставляет возможность локального и распределенного объединения.

- Локальное объединение обеспечивает прозрачное взаимодействие физических элементов в пределах площадки.
- Распределенное объединение расширяет доступ между двумя площадками, которые находятся на определенном расстоянии.

Решение VPLEX устраняет физические барьеры и позволяет пользователям получить доступ к согласованной копии данных, требующей согласованности кэш-памяти, в разных географических точках. Кроме того, оно позволяет использовать географически распределенные физические или виртуальные кластеры хостов. Благодаря этому выполняется прозрачное распределение нагрузки между несколькими площадками, при котором задействуются гибкие возможности перемещения рабочих нагрузок между площадками при подготовке к запланированным событиям. Более того, в случае внепланового события, которое может вызвать прерывание работы в одном из центров обработки данных, можно возобновить предоставление услуг на уцелевшей площадке.

VPLEX поддерживает две конфигурации: Local и Metro. Если используется решение VPLEX Metro с дополнительным компонентом VPLEX Witness и конфигурацией с перекрестными подключениями, приложения продолжат работу на неповрежденной площадке без прерываний или простоев. Ресурсы хранения, виртуализированные системой VPLEX, взаимодействуют через стек и могут динамически переносить приложения и данные между разными географическими точками и поставщиками услуг.

Систему XtremIO можно использовать в качестве высокопроизводительного пула хранения в кластере VPLEX Local или VPLEX Metro. При использовании в сочетании с решением VPLEX система XtremIO получает преимущества всех сервисов управления данными VPLEX, в том числе поддержку операционной системы сервера, мобильность данных, защиту данных, репликацию и перемещение рабочих нагрузок.

RecoverPoint

Семейство EMC RecoverPoint предлагает экономичные решения по локальной непрерывной защите данных (CDP), непрерывной удаленной репликации (CRR) и непрерывной параллельной локальной и удаленной репликации (CLR), которые позволяют восстановить состояние данных на любой момент времени. RecoverPoint/EX поддерживает локальную и удаленную репликацию для EMC Symnmetrix® VMAX™ 10K, Symmetrix VMAX 20K, Summetrix VMAX 40K, VPLEX™, XtremIO (при виртуализации с помощью VPLEX; встроенную поддержку RecoverPoint планируется внедрить в одном из следующих выпусков), серии VNX и массивов Clariion CX3 и CX4.

Этот продукт позволяет заказчикам централизовать и упростить управление защитой данных и предоставляет возможности локальной непрерывной защиты данных и/или удаленной репликации.

- Удаленная и локальная репликация на уровне блоков
- Синхронная, асинхронная или динамическая синхронная удаленная репликация
- Репликация на основе политик позволяет оптимизировать сетевые ресурсы и ресурсы хранения данных, получив желаемые показатели целевой точки (RPO) и целевого времени восстановления (RTO)
- Интеграция с поддержкой приложений
- Поддержка кластера Geo в среде Windows (при использовании RecoverPoint/CE)

RecoverPoint для виртуальных машин — это полностью виртуализированное решение для репликации, работающее на уровне гипервизора и построенное на базе полностью виртуализированного механизма EMC RecoverPoint.

RecoverPoint для виртуальных машин позволяет:

- оптимизировать целевую точку и целевое время восстановления (RPO/RTO) в средах VMware при более низкой совокупной стоимости владения;
- упростить операционное и аварийное восстановление, а также повысить оперативность бизнеса;
- предоставить ИТ-отделам и поставщикам услуг решение для защиты данных с поддержкой облачных сред, позволяющее реализовать защиту от аварий как услуг в частных, публичных и гибридных облаках.

Краткое описание решения

PowerPath, VPLEX, RecoverPoint и XtremIO можно объединить,* чтобы создать надежное мощное и высокопроизводительное решение для блочного хранения данных.

- PowerPath — устанавливается на хосты и обеспечивает переключение на резервный путь, балансировку нагрузки и оптимизацию производительности модулей VPLEX (или непосредственно в массив XtremIO, если решение VPLEX не используется).
- Кластер VPLEX Metro — позволяет совместно использовать услуги по хранению между распределенными виртуальными томами и обеспечивает одновременный доступ для чтения и записи на площадках Metro и за пределами массива.
- Кластер VPLEX Local — используется на целевой площадке, виртуализирует устройства хранения данных от корпорации EMC и других производителей, одновременно повышая коэффициент использования ресурсов.
- ПО RecoverPoint/EX — в любом устройстве, инкапсулированном в VPLEX (включая массив XtremIO), можно использовать услуги RecoverPoint для асинхронной, синхронной или динамической синхронной репликации данных.

* Требуется одобрение RPQ. Обратитесь к представителю EMC.

Например:

У организации есть три центра обработки данных — в Нью-Джерси, Нью-Йорке и Айове, — как показано на Рис. 23.

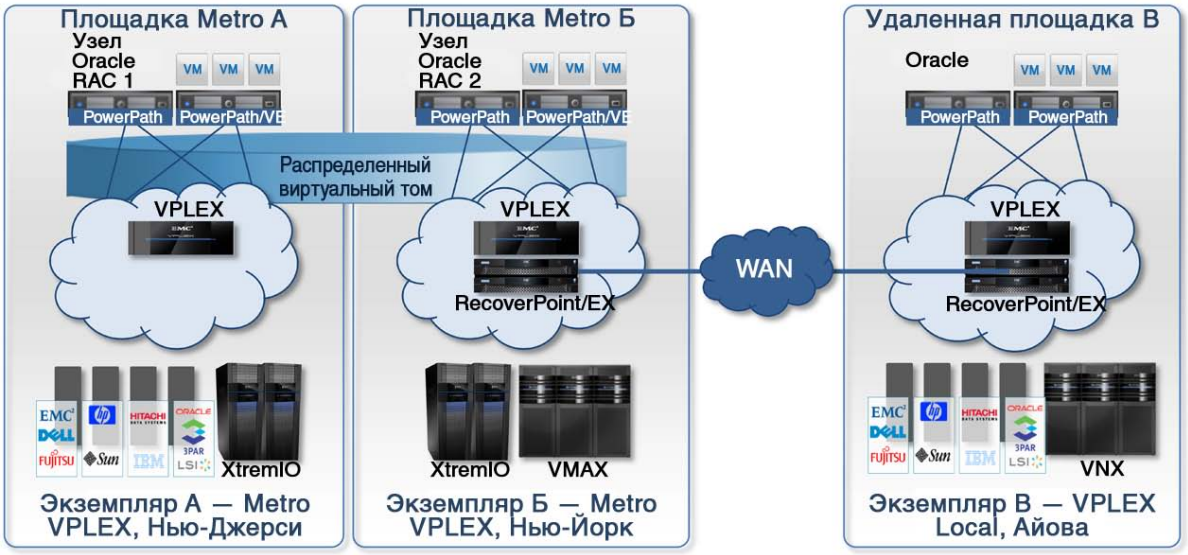


Рис. 23. Интегрированное решение с использованием XtremIO, PowerPath, VPLEX и RecoverPoint

Узлы Oracle RAC и VMware HA рассредоточены между площадками в Нью-Джерси и в Нью-Йорке, и данные часто перемещаются между всеми площадками.

Организация решила применить к своей инфраструктуре хранения данных стратегию сотрудничества с различными производителями.

- Система хранения XtremIO используется для инфраструктуры VDI организации и других высокопроизводительных приложений.
- Кластер VPLEX Metro применяется для обеспечения мобильности данных и доступа к площадкам в Нью-Джерси и Нью-Йорке. Кластер VPLEX Metro обеспечивает организацию функциональностью доступа везде и отовсюду, когда к виртуально распределенным томам можно получить доступ для чтения и записи на обеих площадках.
- Решение для аварийного восстановления реализуется с помощью ПО RecoverPoint для непрерывной асинхронной удаленной репликации между площадкой кластера Metro и площадкой в Айове.
- Кластер VPLEX Metro используется на площадке в Айове для улучшения использования ресурсов, предоставляя возможность выполнить репликацию из системы хранения EMC в систему хранения другого производителя.

Решение EMC (как в примере выше) обладает рядом уникальных и ценных качеств, в том числе:

- высокой доступностью и оптимизацией производительности при управлении путями ввода-вывода в среде высокопроизводительной системы хранения данных;
- высокопроизводительной ориентированной на содержание системой хранения на твердотельных дисках, которая поддерживает сотни тысяч операций ввода-вывода в секунду с низким временем отклика и высокой пропускной способностью;
- географически распределенными кластерами без целевой точки восстановления (RPO);
- автоматическим восстановлением с почти нулевым целевым временем восстановления (RTO);
- высокой доступностью в центрах обработки данных на базе VPLEX Metro и между ними;
- увеличением производительности благодаря распределению рабочей нагрузки между площадками;
- непрерывной удаленной репликацией (или непрерывной защитой данных, или параллельной локальной и удаленной репликацией) систем XtremIO.

Интеграция с OpenStack

OpenStack — это открытая платформа управления частными и публичными облаками. Она позволяет размещать ресурсы хранения данных в любом месте облака и предоставлять их по требованию. Cinder — это сервис блочного хранения данных для OpenStack.

Драйвер XtremIO Cinder позволяет облакам на основе OpenStack осуществлять доступ к массивам хранения XtremIO. Драйвер управления XtremIO Cinder управляет созданием и удалением томов в массиве XtremIO, а также подключает тома к экземплярам и виртуальным машинам и отключает эти тома от них. Драйвер автоматизирует привязку инициаторов к томам. Эта привязка позволяет экземплярам OpenStack осуществлять доступ к массивам хранения XtremIO. Все эти операции выполняются по требованию в зависимости от потребностей облака OpenStack.

Драйвер OpenStack XtremIO Cinder использует программный интерфейс (API) RESTful, чтобы передавать команды управления OpenStack в массив XtremIO.

Облако OpenStack может работать с массивом хранения XtremIO по протоколу iSCSI или Fibre Channel.

Заключение

В основу массива XtermIO положена революционная архитектура, оптимизированная для всех подсистем корпоративной системы хранения данных на твердотельных дисках. Система XtremIO обладает обширным набором функций, в которых используется и оптимизируется функциональность твердотельных дисков. Эти функции были специально разработаны для создания беспрецедентных решений, учитывающих потребности и требования корпоративных заказчиков.

Среди особенностей системы XtremIO можно выделить по-настоящему масштабируемые решения (возможность приобретения дополнительной емкости и повышения производительности по мере необходимости), высокую производительность с сотнями тысяч операций ввода-вывода в секунду, задержку, измеряемую долями миллисекунд, сокращение объема данных «на лету» с учетом содержимого, высокую доступность, «тонкое» выделение ресурсов, моментальные снимки и поддержку интерфейса VAAI.

Массив 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 особенности твердотельных дисков.

Кроме того, в XtremIO предусмотрен универсальный, интуитивно понятный и удобный интерфейс, который включает в себя как графический интерфейс пользователя, так и режимы командной строки и ориентирован на простоту использования и эффективное управление системой.

Массив XtremIO — это идеальное решение для корпоративных сетей хранения данных на твердотельных дисках, которое отличается превосходной совокупной стоимостью владения.