Учебная программа для сертификации Huawei в области систем хранения данных

Учебное руководство для получения

сертификата HCIA в области СХД

Версия 4.5



HUAWEI TECHNOLOGIES CO., LTD.

**© Huawei Technologies Co., Ltd., 2020 г. Все права защищены.**

Запрещается воспроизводить или передавать любые фрагменты данного документа в любой форме и любым способом без предварительного письменного согласия компании Huawei Technologies Co., Ltd.

**Товарные знаки и разрешения**

 и другие товарные знаки Huawei являются товарными знаками компании Huawei Technologies Co., Ltd.

Все остальные товарные знаки и торговые наименования, упоминаемые в настоящем документе, являются собственностью соответствующих владельцев.

**Примечание**

Перечень приобретаемых изделий, сервисов и функций приводится в договоре, заключаемом между компанией Huawei и клиентом. Изделия, сервисы и функции, описываемые в настоящем документе, могут не входить в объем закупок или использования. Если иное не установлено условиями договора, все утверждения, информация и рекомендации в настоящем документе приводятся по принципу «КАК ЕСТЬ» без каких-либо явных или подразумеваемых гарантий или заявлений.

Информация, приведенная в данном документе, может быть изменена без предварительного уведомления. Составители настоящего документа приняли все возможные меры, чтобы обеспечить достоверность и точность его содержания, однако приведенные в нем утверждения, информация и рекомендации не содержат каких бы то ни было явных или подразумеваемых гарантий.

**HUAWEI TECHNOLOGIES CO., LTD.**

Адрес: Промышленная база Huawei Баньтянь, Лунган, Шэньчжэнь, 518129

Китайская Народная Республика

Веб-сайт: http://[e](http://e.huawei.com/).huawei.com

**Система сертификации Huawei**

Программа профессиональной сертификации Huawei является частью стратегии развития «Платформа + экосистема», которая представляет собой новый подход к инфраструктуре ИКТ, основанный на принципах совместной работы и взаимодействия облака, каналов передачи и устройств. Компания Huawei разработала полноценную систему сертификации, охватывающую три категории: сертификация в области инфраструктуры ИКТ, сертификация в области платформы и услуг, а также отраслевая сертификация ИКТ. Это единственная система технической сертификации в отрасли, охватывающая все эти области.

Компания Huawei предоставляет сертификаты трех уровней: Начинающий сертифицированный специалист Huawei (HCIA), Сертифицированный специалист Huawei продвинутого уровня (HCIP) и Сертифицированный эксперт Huawei (HCIE).

Программа сертификации начинающих специалистов в области СХД (HCIA-СХД) предназначена для инженеров Huawei, учащихся и специалистов отрасли ИКТ. Сертификация HCIA-СХД охватывает тенденции в области технологий хранения данных, базовые и распространенные передовые технологии хранения, решения для поддержки непрерывности бизнеса, а также системы управления эксплуатацией и техническим обслуживанием СХД.

Сертификат HCIA-СХД станет визитной карточкой в отрасли и на рынках СХД, выделяющей вас среди коллег, а соответствующая учебная программа научит разбираться в отраслевых тенденциях.



**Содержание**

[1 Тенденции в области технологий хранения данных 6](#_Toc68029717)

[1.1 Тенденции в области технологий хранения данных 6](#_Toc68029718)

[1.1.1 Данные и информация 6](#_Toc68029719)

[1.1.2 Хранение данных 8](#_Toc68029720)

[1.1.3 Развитие технологий хранения 10](#_Toc68029721)

[1.1.4 Тенденции развития продуктов для хранения данных 13](#_Toc68029722)

[2 Базовые технологии хранения данных 18](#_Toc68029723)

[2.1 Компоненты интеллектуального хранилища 18](#_Toc68029724)

[2.1.1 Контроллерная полка 18](#_Toc68029725)

[2.1.2 Дисковая полка 19](#_Toc68029726)

[2.1.3 Модуль расширения 19](#_Toc68029727)

[2.1.4 Жесткий диск 20](#_Toc68029728)

[2.1.5 Твердотельный накопитель (SSD) 25](#_Toc68029729)

[2.1.6 Интерфейсный модуль 28](#_Toc68029730)

[2.2 Технологии RAID 30](#_Toc68029731)

[2.2.1 Традиционные технологии RAID 30](#_Toc68029732)

[2.2.2 RAID 2.0+ 38](#_Toc68029733)

[2.2.3 Прочие технологии RAID 41](#_Toc68029734)

[2.3 Стандартные протоколы хранения 42](#_Toc68029735)

[2.3.1 SCSI 42](#_Toc68029736)

[2.3.2 iSCSI, FC и FCoE 44](#_Toc68029737)

[2.3.3 SAS и SATA 48](#_Toc68029738)

[2.3.4 PCIe и NVMe 50](#_Toc68029739)

[2.3.5 RDMA и IB 53](#_Toc68029740)

[2.3.6 CIFS, NFS и NDMP 55](#_Toc68029741)

[2.4 Архитектура системы хранения данных 56](#_Toc68029742)

[2.4.1 Развитие архитектуры систем хранения данных 56](#_Toc68029743)

[2.4.2 Способы расширения систем хранения данных 59](#_Toc68029744)

[2.4.3 Архитектура продуктов для хранения данных Huawei 61](#_Toc68029745)

[2.5 Архитектура сетей хранения 65](#_Toc68029746)

[2.5.1 DAS 65](#_Toc68029747)

[2.5.2 NAS 65](#_Toc68029748)

[2.5.3 SAN 67](#_Toc68029749)

[2.5.4 Распределенная архитектура 71](#_Toc68029750)

[2.6 Знакомство с интеллектуальными продуктами Huawei для хранения данных 72](#_Toc68029751)

[2.6.1 Система хранения данных на базе флеш-накопителей 72](#_Toc68029752)

[2.6.2 Гибридные флеш-системы хранения данных 75](#_Toc68029753)

[2.6.3 Распределенное хранилище 76](#_Toc68029754)

[2.6.4 Решение для периферийного хранения данных (FusionCube) 78](#_Toc68029755)

[3 Передовые технологии хранения 80](#_Toc68029756)

[3.1 Технологии и приложения для настройки ресурсов хранения 80](#_Toc68029757)

[3.1.1 SmartThin (интеллектуальное динамическое выделение емкости) 80](#_Toc68029758)

[3.1.2 SmartTier (интеллектуальное многоуровневое хранение данных) 81](#_Toc68029759)

[3.1.3 SmartQoS (интеллектуальный контроль качества услуг) 84](#_Toc68029760)

[3.1.4 SmartDedupe (интеллектуальная дедупликация «на лету») 86](#_Toc68029761)

[3.1.5 SmartCompression (интеллектуальное сжатие «на лету») 87](#_Toc68029762)

[3.1.6 SmartMigration (интеллектуальная миграция) 88](#_Toc68029763)

[3.2. Приложения и технологии для защиты данных СХД 89](#_Toc68029764)

[3.2.1 HyperSnap (моментальные снимки) 89](#_Toc68029765)

[3.2.2 HyperClone (клонирование LUN) 92](#_Toc68029766)

[3.2.3 HyperReplication (удаленная репликация) 94](#_Toc68029767)

[3.2.4 HyperMetro 97](#_Toc68029768)

[4 Решения для поддержки непрерывности бизнеса 100](#_Toc68029769)

[4.1 Общая информация о решениях для резервного копирования 100](#_Toc68029770)

[4.1.1 Обзор 100](#_Toc68029771)

[4.1.2 Архитектура 101](#_Toc68029772)

[4.1.3 Режимы сетевой передачи данных 103](#_Toc68029773)

[4.1.4 Основные технологии резервного копирования 104](#_Toc68029774)

[4.1.5 Приложения 110](#_Toc68029775)

[4.2 Общая информация о решениях для аварийного восстановления 112](#_Toc68029776)

[4.2.1 Обзор решений для аварийного восстановления 112](#_Toc68029777)

[4.2.2 Архитектура решений для аварийного восстановления 116](#_Toc68029778)

[4.2.3 Основные технологии аварийного восстановления 118](#_Toc68029779)

[4.2.4 Сценарии применения решений для аварийного восстановления 120](#_Toc68029780)

[5 Управление эксплуатацией и техническим обслуживанием систем хранения данных 122](#_Toc68029781)

[5.1 Управление эксплуатацией и техническим обслуживанием систем хранения данных 122](#_Toc68029782)

[5.1.1 Обзор управления системами хранения данных 122](#_Toc68029783)

[5.1.2 Общая информация об инструментах управления системами хранения данных 123](#_Toc68029784)

[5.1.3 Основные операции управления 124](#_Toc68029785)

[5.2 Управление эксплуатацией и техническим обслуживанием систем хранения данных 124](#_Toc68029786)

[5.2.1 Обзор процессов эксплуатации и технического обслуживания (ЭиТО) 124](#_Toc68029787)

[5.2.2 Инструмент управления эксплуатацией и техническим обслуживанием 125](#_Toc68029788)

[5.2.3 Примеры процессов эксплуатации и технического обслуживания 125](#_Toc68029789)

|  |
| --- |
| 1 Тенденции в области технологий хранения данных |
|  |

## 1.1 Тенденции в области технологий хранения данных

### 1.1.1 Данные и информация

#### 1.1.1.1 Определение данных

Под данными понимаются распознаваемые символы, описывающие те или иные события. Это физические символы или их комбинации, несущие информацию о свойствах и состояниях событий, а также отношениях между ними. В узком смысле под данными понимаются числа. В широком смысле это комбинации символов, букв и цифр, графиков, изображений, видео- и аудиоматериалов, несущие определенное значение, или абстрактное представление атрибутов, количества, местоположения объектов и взаимосвязей между ними. Примеры данных: 0, 1, 2, ветреная погода, дождь, понижение температуры, успеваемость учащихся и транспортировка товаров.

В информатике данные — это общий термин для всех пригодных для обработки компьютерными программами форм представления информации, таких как числа, буквы, символы и аналоговые параметры. Компьютеры хранят и обрабатывают данные, генерируемые широким спектром объектов.

Ассоциация управления данными DAMA определяет данные как представление фактов в форме текста, чисел, графики, изображений, аудио- и видеоматериалов.

#### 1.1.1.2 Типы данных

В зависимости от режимов хранения и управления данные делятся на структурированные, слабоструктурированные и неструктурированные.

Структурированные данные можно хранить в реляционных базах данных, где они обычно представлены в виде двухмерных таблиц. Примеры таких баз данных: SQL Server, MySQL и Oracle.

Формат слабоструктурированных данных не соответствует структуре реляционных баз данных или других таблиц данных. При этом для разделения семантических элементов или поддержки иерархии записей и полей в таких данных используются теги. К слабоструктурированным данным относятся форматы XML, HTML и JSON.

Неструктурированные данные не характеризуются постоянной/полноценной структурой или не имеют предопределенной модели данных. К таким данным относятся тексты, рисунки, отчеты, изображения, аудио- и видеоматериалы.

#### 1.1.1.3 Цикл обработки данных

Обработка данных — это выполняемый людьми или машинами процесс реорганизации или переупорядочивания данных с целью извлечения полезной информации. Цикл обработки данных включает три основных этапа: ввод, обработку и вывод.

* + - * + Ввод: ввод данных в определенном формате, в зависимости от механизма обработки. При использовании компьютера, например, входные данные могут быть записаны на различных типах носителей, таких как диски и ленты.
        + Обработка: выполнение различных действий с входными данными для извлечения полезной информации. Пример обработки: расчет заработной платы на основе карточек учета рабочего времени или расчет заказов на продажу для создания отчетов о продажах.
        + Вывод: генерация и вывод результатов обработки данных. Формат выходных данных зависит от сценария использования. Например, выходными данными может считаться сумма заработной платы сотрудника.

#### 1.1.1.4 Определение информации

Под информацией понимаются сведения об объектах, передаваемые и обрабатываемые с использованием голоса, сообщений и систем связи. Этому определению соответствует весь контент, распространяемый в человеческом обществе. Собирая и анализируя информацию о природе и обществе, человек получает возможность различать понятия, изучать и преобразовывать мир. Информация является универсальным элементом во всех системах связи и управления. В 1948 году математик Клод Элвуд Шеннон (Claude Elwood Shannon) в статье «*Математическая теория связи*» сказал, что сущность информации — это сокращение случайной неопределенности.

Информация — это данные в контексте. К контексту относится следующее:

* + - * + Прикладные значения элементов данных и связанных терминов
        + Формат выражения данных
        + Временной диапазон данных
        + Актуальность данных для конкретного сценария использования

В общем случае понятие «данные» является более объективным и представляет нечто, не зависящее от воли людей. Информация — это обработанные данные, которые имеют определенную ценность и определенное значение.

Например, с точки зрения футбольного фаната, история футбола, футбольные матчи, тренеры, игроки и даже правила ФИФА представляют собой данные о футболе. А данные любимой команды человека, его любимого игрока и интересующих футбольных событий — это информация.

Невозможно получить «все данные», но можно располагать «достаточной информацией», которая позволит принимать решения.

#### 1.1.1.5 Данные и информация

Данные — это необработанные и неорганизованные сведения, представляющие какой-либо факт. Чтобы такой факт получил значение, соответствующие данные необходимо обработать. Информация же — это набор данных, обработанных адекватным образом в соответствии с заданными требованиями.

Данные не имеют какой-либо конкретной цели, тогда как информация несет значение, определенное путем интерпретации данных.

Сами по себе данные не имеют значения, в то время как информация важна сама по себе. Данные не зависят от информации, но информация зависит от данных.

Данные измеряются в битах и байтах, а информация — в значимых единицах, таких как время и количество.

Данные могут быть структурированными, представленными в виде таблиц, диаграмм или деревьев данных. Информация — это язык, идеи и мысли, основанные на данных.

Данные — это сведения, отражающие атрибуты объекта. Это определенная форма, несущая информацию. Данные преобразуются в информацию в процессе обработки. Информацию необходимо преобразовать в цифровые данные для хранения и передачи.

#### 1.1.1.4 Управление жизненным циклом информации

Управление жизненным циклом информации — это не просто продукт или решение, а стратегия и концепция в сфере информационных технологий, которых рекомендуется придерживаться корпоративным пользователям. Данные — ключ к распространению информационных технологий и важнейшее конкурентное преимущество предприятия. Цикл начинается с момента создания информации. Полный жизненный цикл включает процессы создания, защиты, миграции, архивирования и уничтожения данных, а также доступа к ним. Все эти процессы требуют качественного управления и сотрудничества. Плохо организованное управление может привести к напрасной трате больших объемов ресурсов или неэффективной работе вследствие их недостатка.

Специалисты EMC рекомендуют клиентам компании внедрять системы управления жизненным циклом информации в три этапа. Этап 1: внедрение автоматического сетевого хранилища и оптимизация инфраструктуры хранения. Этап 2: повышение уровня обслуживания и оптимизация управления информацией. Этап 3: внедрение интегрированной среды управления жизненным циклом.

Управление данными в рамках управления жизненным циклом информации обычно делится на следующие этапы:

Этап создания данных: генерация данных терминалами и их сохранение на устройствах хранения.

* + - * + Этап защиты данных: применение различных технологий в зависимости от уровня данных и прикладной системы с целью обеспечения эффективной и своевременной защиты различных типов данных и информации. Система хранения поддерживает такие функции защиты данных, как RAID, обеспечение высокой доступности, аварийное восстановление и управление правами доступа.
        + Этап доступа к данным: обеспечение удобного доступа к информации и обмена ею между отделами и приложениями внутри предприятия с целью максимизации бизнес-потенциала.
        + Этап миграции данных: поддержка переноса данных на новые ИТ-устройства при их обновлении или замене.
        + Этап архивирования данных: системы архивирования данных поддерживают бизнес-деятельность предприятий, предоставляя возможности доступа к историческим данным о транзакциях и фактах принятия решений. На этом этапе часто используются технологии дедупликации и сжатия.
        + Этап уничтожения данных: по прошествии определенного времени некоторые данные больше не требуется хранить. На этом этапе такие данные необходимо уничтожить или иным способом удалить из систем хранения и хранилищ.

### 1.1.2 Хранение данных

#### 1.1.2.1 Определение хранилища данных

В узком смысле под хранилищем данных понимаются физические носители с функциями резервирования, защиты и миграции данных, такие как гибкие диски, компакт-диски, DVD-диски, накопители и даже ленты.

В широком смысле это портфель решений для обеспечения доступа к информации, ее защиты, оптимизации и использования в корпоративной среде. Это основа ориентированной на данные информационной архитектуры.

В рамках этого курса хранилища данных рассматриваются в широком смысле.

#### 1.1.2.2 Система хранения данных (СХД)

Технологии хранения данных работают в комплексе. Полноценная система хранения включает несколько компонентов.

Система хранения данных состоит из аппаратных и программных компонентов, а также решений для хранения. К аппаратным компонентам относятся устройства хранения данных и подключения хранилищ, такие как массивы накопителей, ленточные библиотеки и коммутаторы Fibre Channel. Программное обеспечение для хранения данных значительно повышает доступность устройств хранения. Программное обеспечение позволяет реализовать такие функции, как зеркалирование, репликация и автоматическое резервное копирование данных.

#### 1.1.2.3 Физическая структура систем хранения данных

Типичная система хранения включает подсистемы накопителей, контроля, соединений и программного обеспечения для управления хранилищем.

На нижнем уровне физической структуры СХД находятся накопители, которые подключаются к внутренним платам и контроллерам системы с помощью таких соединителей, как оптоволоконные и последовательные кабели.

Для обеспечения доступа к данным СХД подключается к хостам через интерфейсные платы и коммутаторы сети хранения.

Программное обеспечение для управления хранилищем используется для настройки, мониторинга состояния и оптимизации подсистем и соединителей СХД.

#### 1.1.2.4 Типы систем хранения данных

В зависимости от взаимного расположения устройств хранения и хостов системы хранения можно разделить на внутренние и внешние.

Внутренняя система хранения напрямую подключена к главной шине. Она оснащена высокоскоростным кэшем и памятью, необходимыми для работы ЦП, а также накопителями и приводами CD-ROM, которые, в свою очередь, напрямую подключены к материнским платам компьютеров. Емкость таких систем, как правило, невелика, и ее трудно увеличить.

В зависимости от режима подключения внешние системы хранения делятся на СХД с прямым подключением (DAS) и СХД, подключаемые в сетевую инфраструктуру (FAS).

СХД FAS, в свою очередь, делятся на сетевые хранилища (NAS) и сети хранения данных (SAN) в зависимости от используемого протокола передачи.

#### 1.1.2.5 Развитие технологий управления данными

Управление данными — это процесс эффективного сбора, хранения, обработки и использования данных с применением компьютерного оборудования и программных технологий. Цель управления данными — максимально эффективное использование данных. Ключевым элементом этого процесса является организация данных.

Технологии управления данными обеспечивают классификацию, организацию, кодирование, ввод, хранение, извлечение, обслуживание и вывод данных. Развитие устройств хранения данных и компьютерных прикладных систем стимулирует развитие баз данных и технологий управления данными. Управление данными в компьютерной системе включает четыре категории: ручное управление, управление файловой системой, управление традиционной системой базы данных и управление большими данными.

#### 1.1.2.6 Сценарии использования систем хранения данных

Данные, генерируемые отдельными лицами и организациями, обрабатываются вычислительными системами и сохраняются в системах хранения данных.

В эпоху ИКТ системы хранения в основном используются для обеспечения доступа к данным, их защиты в целях поддержки безопасности, а также управления данными.

В хранилищах данных постоянной доступности устройства хранения и сами хранимые данные доступны пользователям в любое время, а скорость доступа к данным соответствует требованиям вычислительной платформы. При этом рабочий режим аналогичен использованию дисковых систем хранения на ПК.

В таких системах используются дорогие, но высокопроизводительные накопители и массивы накопителей.

Автономные хранилища используются для резервного копирования данных хранилища постоянной доступности с целью предотвращения возможных потерь. Доступ к данным в автономном хранилище осуществляется редко, а их чтение и запись выполняются последовательно. При использовании автономных хранилищ на базе ленточных библиотек для поиска необходимых данных требуется перемотка лент. А для редактирования уже записанных данных необходимо перезаписать весь объем данных. Таким образом, скорость доступа к автономному хранилищу и его эффективность оказываются достаточно низкими. Типичные продукты для автономного хранения — относительно дешевые ленточные библиотеки.

Хранилища типа Nearline (вторичное хранилище данных) предоставляют пользователям более широкие возможности выбора. По стоимости и производительности они находятся где-то между хранилищами постоянной доступности и автономными хранилищами. Если данные используются редко или объем данных, к которым осуществляется доступ, невелик, их можно сохранить на устройствах Nearline. Такие устройства по-прежнему предоставляют возможности быстрой адресации и обеспечивают высокую скорость передачи. Хранилища Nearline, например, подходят для размещения архивных файлов, которые могут не использоваться в течение длительного времени. Таким образом, хранилища Nearline подходят для сценариев, где важна относительно высокая скорость доступа, но не общая высокая производительность.

### 1.1.3 Развитие технологий хранения

#### 1.1.3.1 Архитектура систем хранения данных

С точки зрения архитектуры, современные СХД прошли путь развития от традиционных хранилищ, внешних хранилищ и сетей хранения данных до распределенных и облачных хранилищ.

Традиционные хранилища состоят из дисков. В 1956 году компания IBM создала первый в мире жесткий диск емкостью всего 5 МБ с 50 пластинами диаметром 24 дюйма. Он был размером с два холодильника и весил более тонны. Этот жесткий диск не зависел от основной вычислительной системы (мэйнфрейма) и использовался в промышленной сфере.

Внешние хранилища также называются хранилищами с прямым подключением. Их самый ранний образец — системы JBOD (Just a Bundle Of Disks — просто набор дисков), объединяющие в себе несколько жестких дисков, каждый из которых виден для хостов как независимый диск. Такой подход лишь увеличивает емкость, но не обеспечивает безопасность данных.

Диски, устанавливаемые в серверы, имеют следующие недостатки: ограниченное число слотов и недостаточная емкость; низкая надежность, вследствие хранения данных на независимых дисках; диски становятся узким местом производительности системы; низкий коэффициент использования пространства хранения; разрозненность данных, хранящихся на разных серверах.

Системы JBOD позволяют в определенной степени решить проблему ограниченного числа слотов, а технология RAID повышает надежность и производительность всей системы. Внешние хранилища постепенно превращаются в массивы хранения с контроллерами. Контроллеры оснащены кэш-памятью и поддерживают технологию RAID.

Кроме того, в таких системах можно использовать специальное программное обеспечение для управления. Хост-системы видят массивы хранения как высокопроизводительный диск большого объема с поддержкой резервирования.

СХД с прямым подключением характеризуются разрозненностью данных и низким коэффициентом использования пространства хранения.

По мере стремительного роста объемов данных, генерируемых и обрабатываемых в нашем обществе, на первый план выходят такие требования к СХД, как гибкие возможности совместного использования данных, высокая эффективность использовании ресурсов и увеличенное расстояние передачи. Появление сетей придало новый импульс развитию технологий хранения.

SAN подразумевает развертывание сетей между устройствами хранения и серверами для реализации функций блочного хранилища.

В рамках систем NAS создаются сети между серверами и устройствами хранения с файловыми системами для предоставления возможности хранения файлов.

С 2011 года популярность набирают унифицированные хранилища, поддерживающие оба подхода — SAN и NAS. Проявляется новая тенденция: конвергенция NAS и SAN. Объединенные системы поддерживают функции как базы данных, так и обмена файлами, что упрощает управление хранилищем и повышает эффективность использования ресурсов хранения.

SAN — это типичная сеть хранения. Ее первые версии — FC SAN — использовали для передачи данных сеть Fibre Channel. Впоследствии появилась поддержка IP-сетей — IP SAN.

В рамках распределенного хранилища используется универсальное серверное оборудование для создания пулов ресурсов хранения. Такие хранилища подходят для облачных вычислений. С помощью программного обеспечения физические ресурсы объединяются в надежный и высокопроизводительный логический пул хранения, предоставляющий возможность использования различных служб хранения.

Как правило, при использовании таких хранилищ данные распределяются по нескольким независимым серверам в рамках масштабируемой системы. СХД использует серверы хранения для распределения нагрузок при хранении и серверы определения местоположения для поиска информации о системе хранения. Характеристики распределенной архитектуры хранилища: универсальное оборудование, унифицированная архитектура и разделение уровней сети и хранения; возможность линейного расширения производительности и емкости до тысяч узлов; эластичное масштабирование ресурсов и высокая эффективность их использования.

Технология виртуализации системы хранения позволяет объединять устройства хранения в логические ресурсы, обеспечивая предоставление комплексных и унифицированных возможностей хранения. Унифицированные функции предоставляются независимо от форм хранения и типов используемых устройств.

Облачное хранилище объединяет в себе несколько устройств хранения, приложений и служб. В нем используется многопользовательская инфраструктура с высокой степенью виртуализации для предоставления предприятиям масштабируемых ресурсов хранения. Такие ресурсы можно динамически настраивать в зависимости от фактических требований организации.

Концепция облачного хранилища сформировалась на основе облачных вычислений. Это новая технология сетевого хранилища. Облачные хранилища созданы на базе таких технологий, как кластерные приложения, сетевые технологии и распределенные файловые системы, и используют прикладное программное обеспечение для поддержки эффективного взаимодействия различных типов устройств хранения в сетях с целью предоставления функций внешнего хранилища данных и доступа к сервисам. Для хранения огромных объемов данных и управления ими облачной системе требуется соответствующее количество устройств хранения. При этом облачная система превращается в облачное хранилище. Поэтому облачное хранилище можно рассматривать как систему облачных вычислений, в основе которой лежат функции хранения данных и управления ими. Одним словом, облачное хранилище — это новое решение, которое объединяет ресурсы хранения в облаке для упрощения доступа к ним. Благодаря этому пользователи могут получать доступ к данным в облаке в любое время и в любом месте с любого сетевого устройства.

#### 1.1.3.2 Накопители

История развития жестких дисков:

* + - * + С 1970 по 1991 годы плотность хранения дисковых пластин ежегодно увеличивалась на 25–30 %.
        + Начиная с 1991 года ежегодное увеличение плотности хранения выросло до 60–80 %.
        + С 1997 года ежегодные темпы роста выросли до 100 % и даже 200 % благодаря технологии IBM Giant Magneto Resistive (GMR, гигантское магнитосопротивление), которая позволила еще сильнее повысить чувствительность головки диска и плотность хранения.
        + IBM 1301: использование технологии Air Bearing (воздушная подушка) для устранения трения между головкой и пластинами диска и повышение емкости до 28 МБ.
        + IBM 3340: жесткий диск с пластинами и считывающими головками в герметичном корпусе емкостью 30 МБ. Он также получил название «Винчестер» (Winchester) в честь винтовки Winchester 30-30, поскольку в нем планировалось использовать два модуля по 30 МБ.
        + В 1992 году были представлены 1,8-дюймовые жесткие диски.

История SSD-накопителей:

* + - * + Транзистор с плавающим затвором, изобретенный Давоном Кангом (Dawon Kahng) и Саймоном Ши Минем (Simon Min Sze) в 1967 году, стал основой технологии флеш-памяти NAND. Такие транзисторы похожи на МОП-транзисторы, за исключением наличия плавающего затвора в центре конструкции. Это и обуславливает их название. В транзисторе используются материалы с высоким сопротивлением, а плавающий затвор полностью гальванически изолирован от всех областей транзистора, чтобы сохранить заряды, которые попадают на него вследствие эффекта квантового туннелирования.
        + В 1976 году компания Dataram выпустила твердотельные накопители под названием Bulk Core. Их емкость составляла 2 МБ (и была очень большой по меркам того времени). В накопителях использовалось восемь планок RAM-памяти объемом 256 КБ каждая.
        + В конце 1990-х некоторые поставщики начали использовать флеш-память при производстве SSD-накопителей. В 1997 году компания Altec Computer Systeme выпустила SSD-накопитель на базе флеш-памяти с параллельным интерфейсом SCSI. В 1999 году компания BiTMICRO выпустила SSD-накопитель на базе флеш-памяти емкостью 18 ГБ. С тех пор SSD-накопители на флеш-памяти постепенно выместили продукты на основе RAM с рынка твердотельных накопителей. Как и жесткие диски, накопители на флеш-памяти способны сохранять данные даже в случае сбоя питания.
        + В мае 2005 года компания Samsung Electronics объявила о выпуске SSD-накопителей и стала первым IT-гигантом, вышедшим на этот рынок. Она также стала первым широко признанным поставщиком SSD.
        + В 2006 году компания NextCom начала оснащать SSD-накопителями свои ноутбуки. Компания Samsung выпустила SSD-накопитель емкостью 32 ГБ. По данным Samsung, объем рынка твердотельных накопителей в 2007 году составил 1,3 млрд долл. США, а в 2010 году — 4,5 млрд долл. США. В сентябре Samsung представила SSD-накопители на памяти PRAM (память на основе фазового перехода) — технологию, призванную заменить флеш-память NOR. В ноябре вышла ОС Windows Vista от Microsoft, ставшая первой операционной системой для ПК, поддерживающая особенности SSD-накопителей.
        + В 2009 году твердотельные накопители сравнялись по емкости с жесткими дисками. 2,5-дюймовый твердотельный накопитель PureSilicon емкостью 1 ТБ включал 128 планок памяти MLC NAND емкостью по 64 Гбит. С этого момента SSD-накопители обеспечивают ту же емкость, что и жесткие диски аналогичного размера. Это очень важно. Производители когда-то считали, что емкость жестких дисков можно с легкость и при низких затратах увеличить за счет повышения плотности дисков. Однако удвоить емкость SSD-накопителей можно было только при удвоении числа внутренних чипов, что было непросто сделать. Однако SSD на базе памяти MLC доказали, что емкость накопителей можно удвоить за счет увеличения числа битов, сохраняемых в одной ячейке. Кроме того, производительность SSD-накопителей значительно превышает возможности жестких дисков. Пропускная способность SSD-накопителей при чтении достигает 240 МБ/с, а при записи — 215 МБ/с, их задержка при чтении — всего 100 микросекунд, и они обеспечивают 50 000 IOPS (операций ввода/вывода в секунду) при чтении и 10 000 IOPS при записи. Таким образом, поставщики жестких дисков столкнулись с серьезной угрозой.

Флеш-чипы SSD-накопителей прошли развитие от SLC, где каждая ячейка памяти хранит один бит информации, MLC с двумя битами и TLC с тремя битами. Сегодня активно развиваются накопители на памяти QLC, где каждая ячейка хранит четыре бита информации.

#### 1.1.3.3 Протоколы обмена данными

Под протоколами обмена данными понимаются режимы связи и требования, которым должны соответствовать интерфейсы.

Интерфейсы используются для передачи данных между кэшем диска и памятью хоста. Различные дисковые интерфейсы определяют скорость соединения между дисками и контроллерами.

По мере развития протоколов хранения скорость передачи данных увеличивается. По мере развития накопителей от жестких дисков к твердотельным накопителям на смену протоколу SCSI пришла спецификация NVMe, в том числе протокол NVMe на основе PCIe и протокол NVMe over Fabrics (NVMe-oF), обеспечивающий подключение к хост-сети.

В рамках спецификации NVMe-oF используются протоколы передачи со сверхнизким значением задержки, такие как RDMA (удаленный прямой доступ к памяти), обеспечивающий удаленный доступ к SSD-накопителям и позволяющий избежать компромисса между производительностью, функциональностью и емкостью при горизонтальном масштабировании центров обработки данных нового поколения.

Выпущенная в 2016 году спецификация NVMe-oF поддерживает протоколы Fibre Channel и RDMA. В структурах на основе RDMA сеть InfiniBand поддерживает конвергентный Ethernet и iWARP (Internet Wide Area RDMA Protocol).

В спецификации NVMe-oF 1.1, выпущенной в ноябре 2018 года, была добавлена возможность создания архитектуры на базе протокола TCP — RoCE (RDMA over Converged Ethernet — RDMA через конвергентный Ethernet). При использовании протокола RoCE не требуется использовать кэш, и ЦП могут напрямую обращаться к дискам.

NVMe — это стандарт интерфейса контроллера SSD. Он разработан для твердотельных накопителей с интерфейсом PCIe с целью максимального повышения производительности флеш-памяти. Он поддерживает интенсивные вычисления для рабочих нагрузок корпоративного уровня в отраслях, требующих обработки больших объемов данных, таких как медико-биологические науки, финансы, мультимедиа и развлечения.

SSD-накопители NVMe обычно используются для баз данных. Высокая скорость и низкая задержка позволяют использовать NVMe в файловых системах и флеш-массивах хранения (СХД типа All-Flash) для обеспечения превосходной производительности чтения/записи. СХД типа All-Flash на базе SSD-накопителей NVMe обеспечивает эффективное хранение данных, коммутацию сети и обмен метаданными.

### 1.1.4 Тенденции развития продуктов для хранения данных

#### 1.1.4.1 История развития продуктов для хранения данных

Общая история развития:

* + - * + СХД типа All-Flash: что касается носителей информации, то цены на флеш-чипы снижаются год от года, а основным сценарием использования жестких дисков постепенно становится хранение холодных и архивных данных (аналогично лентам).
        + Переход на облачные технологии: с точки зрения тенденций архитектуры хранения распределенная архитектура обеспечивает эластичное масштабирование, а перемещение рабочих нагрузок в облако помогает снизить совокупную стоимость владения (TCO).
        + Интеллектуальные возможности: в отношении эксплуатации и технического обслуживания интеллектуальное оборудование, такое как умные дисковые полки, предоставляет интеллектуальные программные возможности.

#### 1.1.4.2 Новые требования к хранению данных в эпоху интеллектуальных технологий

История человечества насчитывает три технологические революции, связанные с изобретением парового двигателя, развитием электроэнергетики и информационных технологий. Каждый период оказал существенное влияние на работу и повседневную жизнь людей. Сейчас мы проходим через четвертую революцию и вступаем в эпоху интеллектуальных технологий. Новые технологии, такие как искусственный интеллект (ИИ), облачные вычисления, Интернет вещей и большие данные, используются для крупномасштабной цифровой трансформации различных отраслей. Новые сервисы основаны на данных и интеллекте и способствуют внедрению таких новых подходов, как ориентация на обслуживание, сотрудничество посредством сетей, интеллектуальное производство и индивидуальная настройка. Вместе мы можем полностью раскрыть потенциал данных, извлечь полезную информацию и повысить их ценность.

Человеческое общество стремительно преобразуется в интеллектуальное. В ходе этого процесса объемы данных стремительно растут. В среднем каждый пользователь в день использует более 1 ГБ мобильного трафика. В процессе обучения каждое автономное транспортное средство ежедневно генерирует 64 ТБ данных. Согласно докладу Huawei Global Industry Vision 2025, с 2018 по 2025 годы глобальный объем данных увеличится с 33 до 180 зеттабайт. Данные становятся основным бизнес-активом предприятий и даже целых государств. Такие ориентирующиеся на эффективное использование данных инициативы, как умное правительство, умные финансы и умное производство, значительно повышают эффективность всего общества. Все больше предприятий осознают, что инфраструктура данных — это ключ к успешному внедрению интеллектуальных технологий, а системы хранения данных — фундамент этой инфраструктуры. Раньше мы классифицировали системы хранения по инновационным технологиям, технической архитектуре и накопителям. По мере того как экономика и общество переходят от цифровой трансформации к интеллектуальной, развивается и новый тип хранилища.

Существует несколько тенденций этого развития:

Первая — интеллектуальные возможности. В этом отношении компания Huawei выделяет концепции «СХД для ИИ» и «ИИ в СХД». Концепция «СХД для ИИ» предполагает, что в будущем хранилище будет предоставлять предприятиям больше возможностей для обучения и применения ИИ. Подход «ИИ в СХД» означает использование технологий ИИ в системах хранения, а также их интеграцию в управление жизненным циклом хранилища с целью обеспечения максимальной эффективности этого процесса, а также повышения производительности, эффективности и стабильности СХД.

Вторая — трансформация СХД. Например, преобразование массивов хранения в СХД типа All-Flash. В будущем все больше приложений будут требовать низкой задержки, высокой надежности и низкой совокупной стоимости владения, а СХД типа All-Flash имеют все эти характеристики. Несмотря на появление новых конкурирующих технологий, в будущем флеш-накопители станут основными носителями информации. На сегодняшний день они еще не достигли такого положения на рынке СХД.

Третья тенденция — развитие распределенного хранилища. В эпоху интеллектуальных технологий 5G для поддержки сценариев высокопроизводительных приложений, таких как искусственный интеллект, высокопроизводительные вычисления и автономное вождение, а также для хранения огромных объемов генерируемых данных необходимы распределенные системы хранения. Благодаря выделенному оборудованию они смогут обеспечить эффективное и экономичное хранение данных в объемах, исчисляемых в эксабайтах. Способы повышения эффективности и масштабного расширения, а также возможные изменения в работе чипов и алгоритмов в будущем — это те проблемы, с которыми столкнулись распределенные хранилища. Ученые стремятся разработать такие чипы, алгоритмы и шины, которые позволят преодолеть барьеры архитектуры фон Неймана и повысить вычислительные мощности базовой инфраструктуры данных, чтобы создать эффективные и недорогие накопители и сократить разрыв между СХД и вычислительными системами. Эти проблемы может решить выделенное оборудование для хранения данных. Такие концепции, как Memory Fabric (коммутационная матрица памяти), также способствуют изменению архитектуры хранилища.

Последняя тенденция — конвергенция. В будущем хранилище будет интегрировано с инфраструктурой данных для поддержки неоднородных вычислений на чипах, оптимизации разнообразных протоколов и обеспечения совместной работы с обработкой данных и аналитикой больших данных для снижения затрат на обработку данных и повышения эффективности. Например, по сравнению с хранилищем, обеспечиваемым серверами общего назначения, интеграция данных и хранилища снизит совокупную стоимость владения, поскольку задачи обработки данных переносятся с серверов на само хранилище. Объекты, большие данные и другие протоколы объединены и взаимодействуют для поддержки больших данных без необходимости в миграции. Такая конвергенция сильно влияет на разработку систем хранения и является ключом к повышению их эффективности.

#### 1.1.4.3 Тенденции в области систем хранения данных

В эпоху интеллектуальных технологий необходимо сосредоточиться на инновациях в оборудовании, протоколах и технологиях. По мере развития от мэйнфреймов IBM до систем x86 и виртуализации использование флеш-накопителей и сетевых протоколов на базе IP стало одной из основных тенденций.

В эпоху интеллектуальных технологий на основе новых сверхбыстрых протоколов взаимодействия были разработаны система согласования кэша Huawei (Huawei Cache Coherence System, HCCS) и стандарт Compute Express Link (CXL), которые поддерживают высокоскоростное взаимодействие между функционально различными процессорами ЦП и ИИ-ускорителями (neural processing unit, NPU). Протоколы RoCE и NVMe поддерживают технологии высокоскоростной передачи данных и контейнеризации. Кроме того, новое оборудование и технологии предоставляют широкий выбор служб хранения данных. Архитектура Memory Fabric позволяет объединять ресурсы флеш-памяти и памяти накопительного класса (SCM) в пулы, обеспечивая обработку данных с задержкой на уровне микросекунд. К памяти SCM относятся технологии Optane, MRAM, ReRAM, FRAM и Fast NAND. Повышение надежности предполагает реконструкцию систем и использование миграции данных. По мере развития архитектуры хранилищ типа All-Flash на уровне микросхем приложения верхнего уровня перестанут получать информацию об аппаратном обеспечении СХД.

В настоящее время скорость доступа SSD-накопителей в 100 раз превышает аналогичный показатель жестких дисков. А скорость доступа SSD-накопителей NVMe уже в 10 000 раз выше, чем у жестких дисков. Хотя задержку накопителей удалось значительно снизить, отношение задержки сети к общей задержке выросло с менее чем 5 % до примерно 65 %. Другими словами, более чем половину рабочего времени накопители бездействуют, ожидая сетевой связи. Уменьшение задержки сети — ключ к эффективному повышению показателя IOPS.

Развитие накопителей

Перейдем к СХД на основе Blu-ray. Технологии оптических накопителей начали развиваться в конце 1960-х годов и пережили три поколения продуктов: CD, DVD и BD. Диски Blu-ray (или BD) — относительно новый член семейства оптических накопителей. Они способны хранить данные в течение от 50 до 100 лет, но даже это не соответствует современным требованиям к хранению. Нам требуется хранить данные еще дольше. Стеклянный композитный материал на основе наночастиц золота может стабильно хранить данные более 600 лет.

Кроме того, появляются такие технологии, как хранилища данных на основе ДНК и квантовое хранилище.

По мере развития науки и технологий емкость дисков увеличивается, а размер уменьшается. Жесткий диск значительно больше ДНК по размеру, но при этом способен хранить гораздо меньше информации. Поэтому ученые начинают использовать ДНК для хранения данных. Сначала несколько групп специалистов попытались записать данные в геном живых клеток. Однако у этого подхода есть недостатки. Клетки воспроизводятся, и со временем проявляются новые мутации, которые могут изменить данные. Более того, клетки умирают, что приводит к потере данных. Затем была предпринята попытка хранить данные на искусственно синтезированной ДНК вне клеток. Хотя плотность хранения на ДНК сейчас достаточно высока и небольшое количество искусственной ДНК может хранить большие объемы данных, чтение и запись в таких системах не отличаются эффективностью. К тому же синтез молекул ДНК весьма дорогостоящ. Однако можно ожидать, что с развитием технологий секвенирования генов эта стоимость будет снижаться.

Ссылки:

Дж. Боханнон (Bohannon, J.), 2012. Статья «DNA: The Ultimate Hard Drive» (ДНК: идеальный жесткий диск). Журнал «Science». Ссылка на статью: https[://w](http://www.sciencemag.org/news/2012/08/dna-ultimate-hard-drive)ww[.sc](http://www.sciencemag.org/news/2012/08/dna-ultimate-hard-drive)i[encemag.org/news/2012/08/dna-ultimate-hard-drive](http://www.sciencemag.org/news/2012/08/dna-ultimate-hard-drive)

Акрам Ф. (Akram F), Хак И. У. (Haq IU), Али Х. (Ali H), Лагари А. Т. (Laghari AT), октябрь 2018. Публикация «Trends to store digital data in DNA: an overview» (Обзор тенденций хранения цифровых данных с помощью ДНК). Журнал «Molecular Biology Reports». 45 (5): 1479–1490. doi:10.1007/s11033-018- 4280-y

Атомное хранилище данных — относительно новая технология, но не новая концепция.

В начале декабря 1959 года физик Ричард Фейнман на ежегодном собрании Американского физического общества в Калифорнийском технологическом институте прочитал лекцию под названием «There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics» (Внизу много места: приглашение войти в новую область физики). В этой лекции Фейнман рассмотрел возможность использования отдельных атомов в качестве базовых единиц для хранения информации.

В июле 2016 года исследователи из Делфтского технологического университета (Нидерланды) опубликовали статью в журнале Nature Nanotechnology. Они использовали атомы хлора на медных пластинах для хранения 1 килобайта перезаписываемых данных. Однако такая память пока что может работать только в очень чистом вакууме или в жидкоазотной среде при температуре -196 °C (77 K).

Ссылки:

Эрвин С. (Erwin, S.), публикация «A picture worth a thousand bytes» (Картина ценностью в тысячу байт). Журнад «Nature Nanotechnology», 11, 919–920 (2016). https://doi.org/10.1038/nnano.2016.141

Каллф Ф. (Kalff, F.), Реберген М. (Rebergen, M.), Фаренфорт Э. (Fahrenfort, E.) и др. Публикация «A kilobyte rewritable atomic memory» (Килобайт перезаписываемой атомной памяти). Журнад «Nature Nanotechnology», 11, 926–929 (2016). https://doi.org/10.1038/nnano.2016.131

Из-за чрезвычайно малого размера атома емкость атомного хранилища будет значительно больше, чем у любого существующего накопителя того же размера. Благодаря развитию науки и технологий в последние годы идея Фейнмана стала реальностью. Чтобы отдать должное этой великой идее, некоторые исследовательские группы записали его лекцию в атомную память. Хотя идея атомного хранилища великолепна, а его создание становится возможным, атомная память предъявляет строгие требования к операционной среде. Атомы движутся, и даже атомы внутри твердых тел колеблются в окружающей среде, поэтому в обычных условиях трудно поддерживать их в упорядоченном состоянии. Атомное хранилище можно использовать только в условиях низких температур, жидкоазотной среде или вакууме.

Если хранилище на ДНК и атомное хранилище связаны с уменьшением размера накопителя и увеличения его емкости, квантовое хранилище предназначено для повышения производительности и скорости работы.

После многих лет исследований эффективность хранения и жизненный цикл квантовой памяти увеличились, но применить такую память на практике все еще довольно сложно. Элементы квантовой памяти имеют такие недостатки, как неэффективность, высокий уровень шума, короткий срок службы и сложность работы при комнатной температуре. Чтобы выпустить продукты квантовой памяти на рынок, необходимо решить эти проблемы.

Квантовое состояние элементов быстро теряется из-за влияния внешней среды. Кроме того, сложно обеспечить 100 % точность при создании квантового состояния и выполнении квантовых операций.

Ссылки:

Ван Ю. (Wang, Y.), Ли Цз. (Li, J.), Чжан Ш. (Zhang, S.) и др. Публикация «Efficient quantum memory for single-photon polarization qubits» (Эффективная квантовая память для кубитов с однофотонной поляризацией). Журнал «Nature Photonics», 13, 346–351 (2019). https://doi.org/10.1038/s41566-019-0368-8

Доу Цзянь-Пэн (Dou Jian-Peng), Ли Хан (Li Hang), Пан Сяо-Лин (Pang Xiao-Ling), Чжан Чао-Ни (Zhang Chao-Ni), Ян Тянь-Хуай (Yang Tian-Huai), Цзинь Сянь-Минь (Jin Xian-Min). Публикация «Research progress of quantum memory» (Прогресс исследований квантовой памяти). Журнал «Acta Physica Sinica», 2019, 68(3): 030307. doi: 10.7498/aps.68.20190039

Развитие сетей хранения данных

В традиционных центрах обработки данных сети IP SAN используют технологию Ethernet для формирования многоузловой симметричной сетевой архитектуры и стек сетевых протоколов TCP/IP для передачи данных. Для обеспечения передачи данных в сети FC SAN требуется независимая сеть FC.

Хотя традиционные сети TCP/IP или Fibre Channel после многих лет разработки близки к зрелости, их техническая архитектура ограничивает возможности применения искусственного интеллекта и распределенного хранилища.

Для уменьшения сетевой задержки и загрузки ЦП разрабатывается технология удаленного прямого доступа к памяти (RDMA), используемая на серверах для предоставления соответствующей функции. RDMA напрямую передает данные из памяти одного компьютера в память другого. Данные быстро перемещаются из одной системы в удаленное системное хранилище без вмешательства операционных систем обоих компьютеров и длительной обработки процессорами. Благодаря этому система имеет высокую пропускную способность, низкую задержку и характеризуется высокой эффективностью использования ресурсов.

#### История продуктов для хранения данных компании Huawei

Huawei занимается разработкой технологий хранения данных с 2002 года и объединяет высококлассных инженеров со всего мира. В эпоху интеллектуальных технологий компания опирается на инновации в области хранения данных, а также исследования и разработки. Благодаря этому продукты Huawei широко признаны заказчиками и организациями по стандартизации по всему миру за их высочайшее качество.

Ориентированная на интеллектуальные возможности система хранения Huawei OceanStor предоставляет инновационную архитектуру, основанную на ИИ, инновационных аппаратных компонентах и алгоритмах. Она включает уровень максимальной производительности на основе памяти/SCM и технологии Memory Fabric, а также высокопроизводительный уровень хранения на основе протоколов IP и флеш-памяти. Это позволяет создать интеллектуальную многоуровневую систему хранения данных. Поддержка технологий объединения вычислительных ресурсов в пулы и интеллектуального планирования на уровне десятков тысяч ядер для центральных, нейронных и графических процессоров реализована на основе инновационных алгоритмов и протоколов высокоскоростного взаимодействия. Кроме того, неоднородные микросервисы на основе контейнеров адаптированы для бизнеса и позволяют выйти за рамки производительности памяти, вычислительной мощности и протоколов. Наконец, на протяжении всего жизненного цикла данных предоставляется интеллектуальная система управления, позволяющая создавать инновационные продукты для хранения данных, соответствующие требованиям мира умных устройств, объединенных в единую сеть.

|  |
| --- |
| 2 Базовые технологии хранения данных |
|  |

## 2.1 Компоненты интеллектуального хранилища

### 2.1.1 Контроллерная полка

#### 2.1.1.1 Конструкция контроллерной полки

Контроллерная полка, в которую устанавливаются контроллеры, является основным компонентом системы хранения.

Контроллерная полка имеет модульную конструкцию и включает корпус, контроллеры (со встроенными модулями вентиляторов), резервные батарейные блоки, модули питания, модули управления и интерфейсные модули.

* + - * + Корпус оснащен объединительной платой, обеспечивающей распределение сигналов и питания между модулями.
        + Контроллер — основной модуль обработки служб системы хранения.
        + Резервные батарейные блоки обеспечивают защиту данных СХД, предоставляя резервное питание при сбоях внешнего источника питания.
        + Модуль питания переменного тока подает питание на контроллерную полку, поддерживая его нормальную работу при максимальной мощности.
        + Модуль управления предоставляет функции управления и обслуживания, а также последовательные порты.
        + Интерфейсные модули предоставляют сервисные порты или порты управления и поддерживают замену на месте. В информатике данные — это общий термин для всех пригодных для обработки компьютерными программами форм представления информации, таких как числа, буквы, символы и аналоговые параметры. Компьютеры хранят и обрабатывают данные, генерируемые широким спектром объектов.

#### 2.1.1.2 Компоненты контроллерной полки

* + - * + Контроллер — это основной компонент системы хранения данных. Он обрабатывает службы хранения, принимает команды управления конфигурацией, сохраняет данные конфигурации, устанавливает соединение с дисками и сохраняет критически важные данные на коффер-дисках.

ЦП контроллера с помощью кэша обрабатывает запросы ввода/вывода от хоста и управляет RAID-массивом СХД.

Каждый контроллер оснащен встроенными дисками для хранения системных данных. На эти диски также сохраняются данные кэш-памяти во время сбоев питания. Диски на разных контроллерах дублируют друг друга.

* + - * + Внешние порты (FE) обеспечивают передачу служебных данных между серверами приложений и системой хранения данных для обработки операций ввода/вывода хоста.
        + Внутренние порты (BE) обеспечивают соединение контроллерной полки с дисковыми полками и предоставляют дискам каналы для чтения и записи данных.
        + Кэш — это микросхема памяти на контроллере диска. Он обеспечивает быстрый доступ к данным и является буфером между внутренним хранилищем и внешними интерфейсами.
        + «Движок» (engine) — ключевой компонент программы или системы разработки на электронной платформе. Обычно он служит для поддержки программ или набора систем.
        + На коффер-дисках хранятся пользовательские данные, конфигурации системы, журналы и «грязные» (измененные) данные в кэш-памяти для защиты от неожиданных отключений электроэнергии.

Встроенный коффер-диск: каждый контроллер Huawei OceanStor Dorado V6 оснащен одним или двумя встроенными SSD-накопителями, используемыми в качестве коффер-дисков. Подробную информацию см. в документации по продукту.

Внешний коффер-диск: система хранения автоматически выбирает четыре диска для использования в качестве коффер-дисков. Каждый коффер-диск предоставляет 2 ГБ памяти для формирования группы RAID 1. В оставшемся пространстве можно хранить служебные данные. В случае неисправности коффер-диска система автоматически выбирает другой диск для использования в качестве коффер-диска с целью обеспечения резервирования.

* + - * + Модуль питания: контроллерная полка оснащена модулем питания переменного тока, поддерживающим ее нормальную работу.

Контроллерная полка высотой 4U оснащена четырьмя модулями питания (БП 0, БП 1, БП 2 и БП 3). БП 0 и БП 1 образуют слой питания контроллеров A и C и поддерживают взаимное резервирование. БП 2 и БП 3 образуют слой питания контроллеров B и D и поддерживают взаимное резервирование. Для обеспечения максимальной надежности рекомендуется подключать БП 0 и БП 2 к одному блоку распределения питания, а БП 1 и БП 3 — к другому.

Контроллерная полка высотой 2U оснащена двумя модулями питания (БП 0 и БП 1) для питания контроллеров A и B. Два модуля питания образуют слой питания и поддерживают взаимное резервирование. Для обеспечения максимальной надежности рекомендуется подключать БП 0 и БП 1 к разным блокам распределения питания.

### 2.1.2 Дисковая полка

#### 2.1.2.1 Конструкция дисковой полки

Дисковая полка имеет модульную конструкцию и включает корпус, модули расширения, модули питания и собственно диски.

* + - * + Корпус оснащен объединительной платой, обеспечивающей распределение сигналов и питания между модулями.
        + Модуль расширения предоставляет дополнительные порты для подключения к контроллерной полке или другой дисковой полке с целью передачи данных.
        + Модуль питания подает питание на дисковую полку, поддерживая ее нормальную работу при максимальной мощности.
        + Диски предоставляют системе хранения пространство для хранения служебных данных и системных данных, а также данных кэш-памяти. Определенные диски используются в качестве коффер-дисков.

### 2.1.3 Модуль расширения

#### 2.1.3.1 Модуль расширения

Каждый модуль расширения предоставляет порты расширения P0 и P1 для подключения к контроллерной или дисковой полке с целью передачи данных.

#### 2.1.3.2 Коммутаторы серии CE

Фиксированные коммутаторы Huawei CloudEngine — это коммутаторы Ethernet нового поколения для центров обработки данных, обеспечивающие высокую производительность, высокую плотность портов и низкие значения задержки. В коммутаторах используется гибкая система охлаждения с движением воздуха от фронтальной панели к тыльной или в обратном порядке. Они поддерживают сети IP SAN и распределенные сети хранения.

#### 2.1.3.3 Коммутатор Fibre Channel

Коммутаторы Fibre Channel — это высокоскоростные сетевые релейные устройства, которые передают данные по оптоволоконным кабелям. Они ускоряют передачу данных и обеспечивают защиту от помех. Коммутаторы Fibre Channel используются в сетях FC SAN.

#### 2.1.3.4 Кабели устройств

Последовательный кабель соединяет систему хранения данных с терминалом техобслуживания через последовательный порт.

Кабели Mini SAS HD соединяют контроллерные и дисковые полки через порты расширения. Существуют электрические и оптические кабели mini SAS HD.

Активный оптический кабель (AOC) соединяет контроллерную полку с коммутатором данных через порт PCIe.

Кабели 100G QSFP28 предназначены для прямого соединения контроллеров или подключения к умным дисковым полкам.

Кабели 25G SFP28 предназначены для внешних сетевых подключений.

Кабели Fourteen Data Rate (FDR) предназначены для подключения интерфейсных модулей IB 56 Гбит/с.

Оптоволоконные кабели соединяют систему хранения с коммутаторами Fibre Channel. Один конец такого кабеля подключается к контроллеру главной шины Fibre Channel (HBA), а другой — к коммутатору Fibre Channel или СХД. На обоих концах оптоволоконного кабеля используются разъемы LC. Оптоволоконные кабели MPO-4\*DLC предназначены для интерфейсных модулей Fibre Channel 8 Гбит/с с 8 портами и интерфейсных модулей Fibre Channel 16 Гбит/с с 8 портами. Они используются для подключения СХД к коммутаторам Fibre Channel.

### 2.1.4 Жесткий диск

#### 2.1.4.1 Структура жестких дисков

* + - * + Обе поверхности пластин жесткого диска покрыты магнитными материалами, на которых путем поляризации кодируются единицы двоичной информации — биты.
        + Магнитные головки осуществляют чтение и запись данных. Для сохранения данных они изменяют полярность магнитного материала на поверхности диска.
        + Рычаг устройства позиционирования перемещает головку чтения/записи в указанное положение.
        + Шпиндель диска подключен к двигателю и оснащен подшипником. Шпиндель вращает пластины жесткого диска, обеспечивая расположение необходимого сектора под головкой.
        + Цепь управления контролирует скорость вращения пластин и движение рычага перемещения головок, а также подает команды на магнитную головку.

#### 2.1.4.2 Конструкция жестких дисков

Каждой магнитной пластине диска соответствует по две головки чтения/записи — по одной для каждой из сторон пластины.

Воздушный поток предотвращает соприкосновение головки с пластинами, обеспечивая высокую скорость перемещения головки между дорожками. Слишком большое расстояние между головкой и пластиной диска приводит к ослаблению сигнала, а слишком малое может привести к соприкосновению головки чтения/записи с пластиной и их повреждению. Поэтому поверхность пластины диска должна быть гладкой и ровной. Любые посторонние предметы или пыль сокращают расстояние, вследствие чего головка может соприкасаться с магнитной поверхностью. А это приводит к необратимому повреждению данных.

Принцип работы:

* + - * + Изначально магнитная головка удерживается в «зоне парковки» рядом с пластинами.
        + Шпиндель соединен со всеми пластинами и подключен к двигателю. Двигатель шпинделя обеспечивает его вращение с постоянной скоростью и приводит в движение пластины.
        + При вращении шпинделя между магнитной головкой и пластинами диска остается небольшой зазор. Он называется высотой полета головки.
        + Головки размещаются на конце рычага, который перемещает их в необходимое для чтения/записи данных положение над пластиной.
        + Головка считывает данные с поверхности пластины или записывает их на эту поверхность в двоичном формате. Считанные данные сохраняются во флеш-чипе диска, а затем передаются в соответствующую программу.

#### 2.1.4.3 Организация данных на диске

* + - * + Поверхность пластин: хранить данные можно на обеих сторонах каждой пластины диска. Все рабочие поверхности нумеруются последовательно, начиная с 0 (самая верхняя). Номер поверхности в дисковой системе также называется номером головки, поскольку каждой рабочей поверхности соответствует отдельная магнитная головка.
        + Дорожка: дорожки представляют собой концентрические круги на пластине вокруг шпинделя, предназначенные для записи данных. Нумерация дорожек начинается с внешнего круга и 0. На каждой поверхности пластины доступно от 300 до 1024 дорожек. Поверхности пластин новых типов дисков большой емкости имеют еще больше дорожек. Для оценки плотности дорожек на пластине обычно используется показатель числа дорожек на дюйм (tracks per inch, TPI). Дорожки — это намагниченные области на поверхности пластин, невидимые для человеческого глаза.
        + Цилиндр: совокупность дорожек с одинаковыми номерами на всех поверхностях пластин диска. Нумерация магнитных головок в отношении цилиндров идет сверху вниз, начиная с 0. Данные считываются и записываются по цилиндрам. Головка 0 первой осуществляет чтение данных соответствующего цилиндра или их запись, после чего другие головки последовательно читают и записывают данные в том же цилиндре. После завершения чтения и записи всеми головками в одном цилиндре они переходят к следующему цилиндру. Выбор цилиндров — это механический процесс переключения, называемый поиском. Положение головок на диске обычно указывается с помощью номера цилиндра, а не номера дорожки.
        + Сектор: для упорядоченного размещения данных каждая дорожка разделена на меньшие блоки, называемые секторами. Сектор — это минимальная адресуемая единица хранения информации на диске.

Количество секторов в дорожках может различаться. Обычно сектор может содержать 512 байт пользовательских данных, но в некоторых новых дисках используются более крупные секторы объемом 4 КБ.

#### 2.1.4.4 Емкость диска

Жесткие диски могут иметь одну или несколько пластин. Однако в каждый момент времени читать и записывать данные может только одна магнитная головка. Поэтому увеличение количества пластин и головок позволяет повысить лишь емкость диска. Пропускная способность или производительность ввода/вывода диска при этом не изменится.

Емкость диска = Число цилиндров x Число головок x Число секторов x Размер сектора. Емкость измеряется в МБ или ГБ. Емкость диска определяется емкостью одной пластины и количеством пластин.

Скорость обработки данных ЦП значительно выше, чем у жестких дисков, вследствие чего ЦП приходится ждать, пока диск завершит операцию чтения/записи, прежде чем выдать новую команду. Решить эту проблему позволяет добавление кэша на диск для повышения скорости чтения/записи.

#### 2.1.4.5 Показатели производительности диска

* + - * + Скорость вращения: число оборотов пластин диска в минуту (об/мин). В процессе чтения или записи данных пластины диска вращаются, а магнитные головки остаются неподвижными. Быстрое вращение пластин сокращает время передачи данных. При обработке операций последовательного доступа рычаг устройства позиционирования переключается нечасто, поэтому скорость вращения является основным фактором, определяющим пропускную способность и производительности IOPS диска.
        + Скорость поиска: при обработке операций случайного доступа рычаг перемещения головок должен часто переключаться между дорожками. Переключение между дорожками занимает гораздо больше времени, чем передача данных. Таким образом, рычаг с более высокой скоростью поиска позволяет увеличить показатель IOPS при обработке операций случайного доступа.
        + Емкость одной пластины: большая емкость одной пластины позволяет хранить больше данных на единицу пространства, то есть повышает плотность данных. Повышение плотности данных при неизменных скорости вращения и поиска обеспечивает увеличение производительности диска.
        + Скорость порта: теоретически текущей скорости порта достаточно для поддержки максимальной пропускной способности внешней передачи дисков. Узким местом в отношении операций ввода/вывода случайного доступа является скорость поиска, а скорость порта практически не влияет на производительность.

#### 2.1.4.6 Среднее время доступа

* + - * + Среднее время поиска — это среднее время, затрачиваемое на перемещение магнитной головки из исходного положения к запрашиваемой дорожке пластины диска. Это важный показатель внутренней скорости передачи данных жесткого диска, а его значение должно быть как можно ниже.
        + Среднее время задержки — это время, в течение которого магнитная головка после перемещения к запрашиваемой дорожке должна ожидать поворота пластин, чтобы необходимый сектор оказался под ней. Среднее время задержки обычно равно времени, за которое совершается половина полного оборота пластины диска. Таким образом, более быстрое вращение пластин приводит к снижению задержки.

#### 2.1.4.7 Скорость передачи данных

Скорость передачи данных жесткого диска описывает то, насколько быстро диск может читать и записывать данные. Этот показатель охватывает внутреннюю и внешнюю скорость передачи данных и измеряется в МБ/с.

* + - * + Внутренняя скорость передачи также называется скоростью непрерывной передачи (sustained transfer rate). Это максимальная скорость, с которой магнитная головка считывает и записывает данные. При ее расчете не учитывается время поиска и задержка вращения пластин диска. Этот показатель описывает идеальные условия, при которых магнитной головке не требуется менять дорожку или считывать определенный сектор, то есть когда чтение/запись охватывает все секторы одной дорожки последовательно и циклически.
        + Внешняя скорость передачи данных также называется скоростью импульсной передачи или скоростью передачи интерфейса. Она описывает скорость передачи данных между системной шиной и буфером диска и зависит от типа дискового интерфейса и размера буфера.

#### 2.1.4.8 Показатели IOPS и пропускной способности диска

Показатель IOPS рассчитывается на основе времени поиска, задержки вращения и времени передачи данных.

* + - * + Время поиска: чем короче время поиска, тем выше показатель IOPS. На текущий момент среднее время поиска составляет от 3 до 15 мс.
        + Задержка вращения: это время поворота пластины диска таким образом, чтобы запрашиваемый сектор оказался под магнитной головкой. Задержка вращения зависит от скорости вращения. Среднее время задержки обычно равно времени, за которое совершается половина полного оборота пластины диска. Например, средняя задержка вращения для дисков 7200 об/мин составляет около 4,17 мс (60 x 1000/7200/2), а средняя задержка вращения для дисков 15000 об/мин — около 2 мс.
        + Время передачи данных: это время, необходимое для передачи запрошенных данных. Оно рассчитывается путем деления объема передаваемых данных на скорость их передачи. Скорость передачи данных в дисках с интерфейсом IDE/ATA, например, может достигать 133 МБ/с, а в дисках с интерфейсом SATA II — 300 МБ/с.
        + Для обработки операций ввода/вывода случайного доступа магнитные головки должны часто переключаться между дорожками. Время передачи данных значительно короче, чем время переключения между дорожками. Поэтому в этом случае временем передачи данных можно пренебречь.

Теоретически максимальное значение IOPS диска можно рассчитать по следующей формуле: IOPS = 1000 мс/(время поиска + задержка вращения). Время передачи данных при этом не рассматривается.

Например, если среднее время поиска составляет 3 мс, теоретическое максимальное значение IOPS дисков со скоростью вращения 7200, 10 000 и 15 000 об/мин составляет 140, 167 и 200 соответственно.

#### 2.1.4.9 Режим передачи

Параллельная передача:

* + - * + Параллельная передача отличается высокой эффективностью, малым расстоянием передачи и низкой частотой.
        + При передаче на большие расстояния использование нескольких линий оказывается дороже использования одной.
        + Для передачи на большие расстояния требуются более толстые электрические провода для уменьшения затухания сигнала, но их сложно объединить в одном кабеле.
        + При передаче на большие расстояния время передачи по каждой линии различается в зависимости от сопротивления проводов или других факторов. Последующую передачу можно инициировать только после того, как данные по всем линиям достигли партнера по связи.
        + Высокая частота передачи вызывает серьезные колебания в цепи и создает помехи между линиями. Поэтому частоту параллельной передачи необходимо тщательно настраивать.

Последовательная передача:

* + - * + Последовательная передача менее эффективна, чем параллельная, но, как правило, быстрее за счет потенциального увеличения скорости передачи за счет увеличения ее частоты.
        + Последовательная передача используется для передачи на большие расстояния. В настоящее время интерфейсы PCI используют последовательную передачу. Интерфейс PCIe является типичным примером последовательного интерфейса. Скорость передачи одной линии достигает 2,5 Гбит/с.

#### 2.1.4.10 Дисковые интерфейсы

По используемым портам жесткие диски делятся на диски IDE, SCSI, SATA, SAS и Fibre Channel. Они также различаются по своей механической основе.

Диски IDE и SATA основаны на технологии ATA и подходят для однозадачной обработки.

Диски SCSI, SAS и Fibre Channel созданы на базе стандарта SCSI и подходят для многозадачной обработки.

Сравнение:

* + - * + Диски SCSI обеспечивают более быструю обработку, чем диски ATA, при высокой пропускной способности.
        + Диски ATA перегреваются во время многозадачной обработки из-за частого движения магнитной головки.
        + Диски SCSI более надежны, чем диски ATA.

Дисковый интерфейс IDE:

* + - * + Выпущено несколько версий ATA, включая ATA-1 (IDE), ATA-2 (Enhanced IDE/Fast ATA), ATA-3 (Fast ATA-2), ATA-4 (ATA33), ATA-5 (ATA66), ATA-6 (ATA100) и ATA-7 (ATA133).
        + У портов ATA есть несколько преимуществ и недостатков:

Их сильные стороны — невысокая цена и хорошая совместимость.

К их недостаткам относятся низкая скорость, ограниченные варианты применения и строгие ограничения по длине кабеля.

Скорость передачи порта PATA также не может удовлетворить текущие потребности пользователя.

Интерфейс SATA:

* + - * + Во время передачи линии передачи данных и сигналов разделены и используют независимую тактовую частоту передачи. Скорость передачи через интерфейс SATA в 30 раз выше, чем при использовании PATA.
        + Преимущества:

В устройствах SATA обычно используются два разъема (7-контактный и 15-контактный) и один канал передачи данных. Интерфейс SATA поддерживает более высокую скорость передачи, чем ATA.

Для обеспечения надежности передачи данных устройства SATA применяет циклический избыточный код (CRC) к инструкциям и пакетам данных.

В отношении защиты от помех интерфейсы SATA превосходят ATA.

Интерфейс SCSI:

* + - * + Жесткие диски SCSI были разработаны для замещения устройств IDE. Они обеспечивают более высокие скорости вращения и передачи данных. SCSI изначально был шинным интерфейсом, работавшим независимо от системной шины.
        + Преимущества:

Поддерживает широкий спектр устройств. К одной плате контроллера SCSI можно одновременно подключить 15 устройств.

Стандарт обеспечивает высокую производительность при многозадачной обработке, низкую загрузку ЦП, а также поддерживает высокие скорость вращения и передачи.

Жесткие диски SCSI можно использовать в качестве внешних или встроенных компонентов. Они поддерживают различные приложения и возможность горячей замены.

* + - * + Недостатки:

Высокая стоимость и сложность установки и настройки.

Интерфейс SAS:

* + - * + В интерфейсе SAS так же, как и в SATA, используется последовательная архитектура, обеспечивающая высокую скорость передачи и оптимизацию внутреннего пространства за счет более коротких внутренних соединений.
        + Интерфейсы SAS повышают эффективность, доступность и расширяют возможности масштабирования системы хранения. Они поддерживают обратную совместимость с интерфейсами SATA на физическом уровне и уровне протоколов.
        + Преимущества:

SAS превосходит SCSI по скорости передачи, помехоустойчивости и поддерживаемому расстоянию подключения.

* + - * + Недостатки:

Более высокая стоимость дисков SAS.

Интерфейс Fibre Channel:

* + - * + Fibre Channel изначально был разработан для передачи по сети, а не для подключения накопителей. Постепенно он стал применяться в дисковых системах для повышения скорости передачи.
        + Преимущества:

Легкость модернизации. Поддержка оптоволоконных кабелей длиной более 10 км.

Большая пропускная способность.

Высокая степень универсальности.

* + - * + Недостатки:

Высокая стоимость.

Сложность сборки.

### 2.1.5 Твердотельный накопитель (SSD)

#### 2.1.5.1 Обзор SSD-накопителей

В отличие от традиционных жестких дисков, в которых для хранения данных используются магнитные материалы, в твердотельных накопителях данные хранятся в ячейках флеш-памяти NAND. Флеш-память NAND — это энергонезависимая память с произвольным доступом, обеспечивающая сохранение данных после отключения питания. Она позволяет быстро сохранять цифровую информацию на компактных устройствах.

В SSD-накопителях не используются вращающиеся на высокой скорости компоненты, что позволяет повысить производительность, снизить энергопотребление и обеспечить нулевой уровень шума.

В SSD-накопителях нет механических элементов, но это не означает, что их жизненный цикл бесконечен. Поскольку флеш-память NAND является энергонезависимой, для записи новых данных в ее ячейки необходимо сначала удалить исходные данные. Однако число циклов перезаписи каждой ячейки ограничено. После достижения этого ограничения чтение и запись данных для соответствующей ячейки становятся невозможны.

#### 2.1.5.2 Архитектура SSD-накопителей

Интерфейс хоста — это протокол и физический интерфейс, посредством которых хост осуществляет доступ к SSD-накопителю. Обычно используются интерфейсы SATA, SAS и PCIe.

Контроллер SSD — это основной компонент SSD-накопителя, обеспечивающий взаимодействие для чтения и записи между хостом и внутренним накопителем, а также отвечающий за преобразование протоколов, управления записями в таблицах, кэширование и проверку данных.

Память DRAM используется в качестве кэша для записей и данных уровня флеш-преобразования (FTL).

В энергонезависимой флеш-памяти с произвольным доступом NAND хранятся данные.

При этом одновременно используется несколько каналов с временным мультиплексированием адресов в канале. Также поддерживаются технологии TCQ и NCQ для обеспечения одновременного ответа на множество запросов ввода/вывода.

#### 2.1.5.3 Флеш-память NAND

К единицам флеш-памяти NAND относятся логические модули (LUN), плоскости, блоки, страницы и ячейки.

Флеш-память NAND хранит данные с помощью транзисторов с плавающим затвором. Пороговое напряжение изменяется в зависимости от величины электрического заряда, сохраненного в плавающем затворе. Данные представляются величиной заряда и считываются путем измерения порогового напряжения транзистора.

* + - * + LUN — это наименьший физический блок, который может быть изолирован и обычно содержит несколько плоскостей.
        + Плоскость имеет независимый регистр страниц. Обычно он содержит 1000 или 2000 нечетных или четных блоков.
        + Блок — это минимальная единица стирания, обычно состоящая из нескольких страниц.
        + Страница — это минимальная единица программирования (записи) и чтения, объем которой обычно равен 16 КБ.
        + Ячейка — это наименьшая стираемая, программируемая и читаемая единица на странице. Ячейка соответствует транзистору с плавающим затвором, который хранит один или несколько битов.

Страница — это основная единица программирования и чтения, а блок — основная единица стирания.

Каждый цикл программирования/стирания (P/E) вызывает некоторое повреждение изоляционного слоя транзистора с плавающим затвором. Если не удается выполнить стирание или программирование блока, блок помечается как дефектный. Когда количество дефектных блоков достигает порогового значения (4 %), срок службы флеш-памяти NAND подходит к концу.

#### 2.1.5.4 Технологии SLC, MLC, TLC и QLC

В зависимости от количества битов, хранящихся в ячейке, флеш-чипы NAND делятся на следующие категории:

* + - * + Одноуровневая ячейка (SLC) может хранить один бит данных: 0 или 1.
        + Многоуровневая ячейка (MLC) может хранить два бита данных: 00, 01, 10 и 11.
        + Трехуровневая ячейка (TLC) может хранить три бита данных: 000, 001, 010, 011, 100, 101, 110 и 111.
        + Четырехуровневая ячейка (QLC) может хранить четыре бита данных: 0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110 и 1111.

Размер этих ячеек не сильно различается, но они способны хранить разный объем данных. Первоначально емкость SSD-накопителей составляла до 64 ГБ. Теперь твердотельный накопитель TLC может хранить до 2 ТБ данных. Однако у каждого типа ячеек своя продолжительность жизненного цикла, что обуславливает разный уровень надежности соответствующих SSD-накопителей. Жизненный цикл также является важным фактором при выборе твердотельных накопителей.

#### 2.1.5.5 Структура данных флеш-чипа

На этом слайде представлена логическая схема флеш-чипа (Toshiba 3D-TLC).

* + - * + Страница состоит из 146 688 ячеек. Каждая страница может хранить 16 КБ содержимого и 1952 байта данных кода коррекции ошибок (ECC). Страница — минимальная единица ввода/вывода для флеш-чипа.
        + Каждые 768 страниц образуют блок. Каждые 1478 блоков образуют плоскость.
        + Флеш-чип состоит из двух плоскостей, одна из которых содержит блоки с нечетными порядковыми номерами, другая — с четными. Обе плоскости могут работать одновременно.

Код ECC применяется к данным во флеш-памяти NAND, вследствие чего размер страницы флеш-памяти NAND не равен в точности 16 КБ, а включает также дополнительную группу байтов. Например, если фактический размер страницы 16 КБ составляет 16 384 + 1 952 байта, то 16 384 байта используются для хранения данных, а 1 952 байта — для хранения кодов проверки данных ECC.

#### 2.1.5.6 Управление преобразованием адресов

Под логическим адресом блока (LBA) может пониматься как адрес блока данных, так и сам блок данных, на который указывает адрес.

PBA: физический адрес блока.

Хост обращается к SSD-накопителю через логические адреса LBA. Каждый LBA обычно представляет сектор размером 512 байт. ОС хоста обращается к массивам ячеек SSD-накопителя, вмещающим 4 КБ данных. Такой массив называется страницей хоста и является базовой единицей доступа хоста к SSD.

Базовой единицей доступа контроллера SSD к флеш-чипу является страница флеш-памяти. Она также называется физической страницей. Каждый раз, когда хост записывает страницу хоста, контроллер SSD записывает ее на физическую страницу и сопоставляет логический и физический адреса записываемой страницы, занося их в специальную таблицу.

Когда хост считывает страницу хоста, SSD находит запрошенные данные по таблице преобразования адресов.

#### 2.1.5.7 Процессы чтения и записи данных

Процесс записи на SSD-накопитель:

* + - * + Контроллер SSD-накопителя по восьми каналам подключается к восьми кристаллам флеш-памяти. Для упрощения на рисунке представлено по одному блоку в каждом кристалле. Каждый элемент блока размером 4 КБ представляет страницу.

Хост записывает 4 килобайта данных в блок канала 0 и тем самым занимает одну страницу.

Хост продолжает записывать еще 16 килобайт данных. В данном примере порции данных по 4 килобайта записываются в каждый блок каналов с 1 по 4.

Хост продолжает записывать данные в блоки, пока все блоки не будут заполнены.

* + - * + После заполнения блоков на всех каналах контроллер SSD выбирает новый блок для аналогичного процесса записи данных.
        + Зеленым цветом обозначаются актуальные данные, а красным — устаревшие. Блоки с ненужными данными помечаются как устаревшие или недействительные, а соответствующая запись о соответствии адресов заменяется.
        + Например, страница хоста A изначально была записана в страницу флеш-памяти X — существовала запись о соответствии адресов A–X. Позже хост перезаписал эту страницу. Данные во флеш-памяти не перезаписываются в обычном смысле: новые данные записываются на новую страницу Y, и создается запись о соответствии адресов A–Y. Исходная запись о соответствии при этом отклоняется. Страница X помечается как содержащая устаревшие данные, также называемые «мусором».
        + Хост продолжает записывать данные на SSD, пока накопитель не будет заполнен. В этом случае хост не сможет продолжать записывать данные, пока не будет удален «мусор».

Процесс чтения с SSD-накопителя:

* + - * + Если считываемые данные распределены в блоках каждого канала равномерно, можно добиться восьмикратного увеличения скорости чтения. Если, например, 32 КБ данных хранятся в блоках каналов с 1 по 4, будет возможно только четырехкратное увеличение скорости чтения. Поэтому файлы меньшего размера считываются медленнее.

#### 2.1.5.8 Преимущества SSD-накопителей в области производительности

Короткое время отклика: традиционные жесткие диски отличаются более низкой эффективностью при передаче данных, поскольку тратят дополнительное время на поиск и имеют механическую задержку. В твердотельных накопителях используется флеш-память NAND, исключающая поиск и механическую задержку и обеспечивающая более быстрый отклик на запросы чтения и записи.

Высокая эффективность чтения/записи: жесткие диски выполняют операции произвольного чтения/записи за счет перемещения магнитной головки, что обуславливает низкую эффективность этих процессов. В твердотельных накопителях места хранения данных определяет внутренний контроллер, сокращая потребность в механических операциях и упрощая обработку чтения/записи.

Кроме того, крупные массивы SSD-накопителей отличаются очень высокой энергоэффективностью.

### 2.1.6 Интерфейсный модуль

#### 2.1.6.1 Интерфейсные модули GE

Модуль электрического интерфейса GE оснащен четырьмя электрическими портами 1 Гбит/с и используется для развертывания сети кворума в рамках решения HyperMetro.

Интерфейсный модуль 40GE предоставляет два оптических порта 40 Гбит/с для подключения устройств хранения к серверам приложений.

Интерфейсный модуль 100GE предоставляет два оптических порта 100 Гбит/с для подключения устройств хранения к серверам приложений.

#### 2.1.6.2 Модуль расширения SAS и интерфейсный модуль RDMA

Интерфейсный модуль RDMA 25 Гбит/с предоставляет четыре оптических порта 25 Гбит/с для установки прямого соединения между двумя контроллерными полками.

Интерфейсный модуль RDMA 100 Гбит/с предоставляет два оптических порта 100 Гбит/с для подключения контроллерных полок к коммутаторам для горизонтального масштабирования или умным дисковым полкам. На этикетках SO означает горизонтальное масштабирование, а BE — внутренние подключения.

Модуль расширения SAS 12 Гбит/с предоставляет четыре порта расширения mini SAS HD со скоростью передачи 4 x 12 Гбит/с для подключения контроллерных полок к дисковым полкам SAS высотой 2U.

#### 2.1.6.3 Интерфейсные модули SmartIO

Интерфейсный модуль SmartIO поддерживает оптические модули 8, 10, 16, 25 и 32 Гбит/с, которые предоставляют порты Fibre Channel 8 Гбит/с, 16 Гбит/с и 32 Гбит/с, а также порты 10GE и 25GE. Через интерфейсные модули SmartIO устройства хранения подключаются к серверам приложений.

Скорость порта оптического модуля должна соответствовать скорости, указанной на этикетке интерфейсного модуля. В противном случае система хранения выдаст аварийный сигнал, а соответствующий порт станет недоступным.

#### 2.1.6.4 Интерфейсные модули PCIe и IB 56 Гбит/с

Интерфейсный модуль PCIe предоставляет два порта PCIe для подключения контроллерных полок к коммутаторам PCIe и обмена потоками управления и данных между контроллерными полками.

Индикаторы:

* + - * + Индикатор питания интерфейсного модуля
        + Индикатор линии связи/скорости соединения порта PCIe
        + Порт PCIe
        + Ручка модуля

Интерфейсный модуль IB 56 Гбит/с предоставляет два порта IB со скоростью передачи 4 x 14 Гбит/с.

Индикаторы:

* + - * + Индикатор питания/кнопка горячей замены
        + Индикатор линии связи порта IB 56 Гбит/с
        + Индикатор активного состояния порта IB 56 Гбит/с
        + Порт IB 56 Гбит/с
        + Ручка модуля

#### 2.1.6.5 Интерфейсные модули Fibre Channel и FCoE

Интерфейсный модуль Fibre Channel 16 Гбит/с предоставляет два физических порта, которые с помощью специальных кабелей преобразуются в восемь портов Fibre Channel 16 Гбит/с. Каждый порт поддерживает скорость передачи 16 Гбит/с. Они служат в качестве служебных портов и поддерживают передачу команд обмена данными от сервера приложений в систему хранения данных.

Интерфейсный модуль FCoE 10 Гбит/с предоставляет два порта FCoE 10 Гбит/с, которые обеспечивают подключение системы хранения к серверу приложений для передачи данных.

Интерфейсный модуль FCoE 10 Гбит/с поддерживает только прямые соединения.

## 2.2 Технологии RAID

### 2.2.1 Традиционные технологии RAID

#### 2.2.1.1 Базовая концепция RAID

Технология RAID (избыточный массив независимых дисков) позволяет различными способами объединить нескольких физических дисковых устройств в логический модуль с целью повышения производительности чтения/записи и отказоустойчивости.

Функциональные возможности RAID:

* + - * + Объединяет несколько физических дисков в один логический массив для увеличения емкости.
        + Разделяет данные на блоки и поддерживает параллельные операции записи/чтения с использованием нескольких дисков для повышения эффективности доступа к ним.
        + Обеспечивает повышенную отказоустойчивость за счет зеркалирования или проверок четности.

Реализовать технологию RAID в устройствах хранения можно аппаратным и программным способами.

* + - * + Аппаратный способ подразумевает использование специального адаптера RAID, контроллера диска или процессора хранилища. RAID-контроллер оснащен встроенными процессором, процессором ввода/вывода и памятью для повышения эффективности использования ресурсов и скорости передачи данных. RAID-контроллер управляет маршрутами и буферами, а также контролирует потоки данных между хост-системой и RAID-массивом. Аппаратные RAID-массивы обычно используются на серверах.
        + При программном способе реализации RAID используется не встроенный процессор или процессор ввода/вывода, а процессор хост-системы. При этом низкоскоростные ЦП не способны удовлетворить требования в отношении реализации RAID. Программный RAID обычно используется в устройствах хранения корпоративного класса.

Чередование дисков: пространство каждого диска делится на несколько полос определенного размера. При записи данные также делятся на блоки в зависимости от размера полосы.

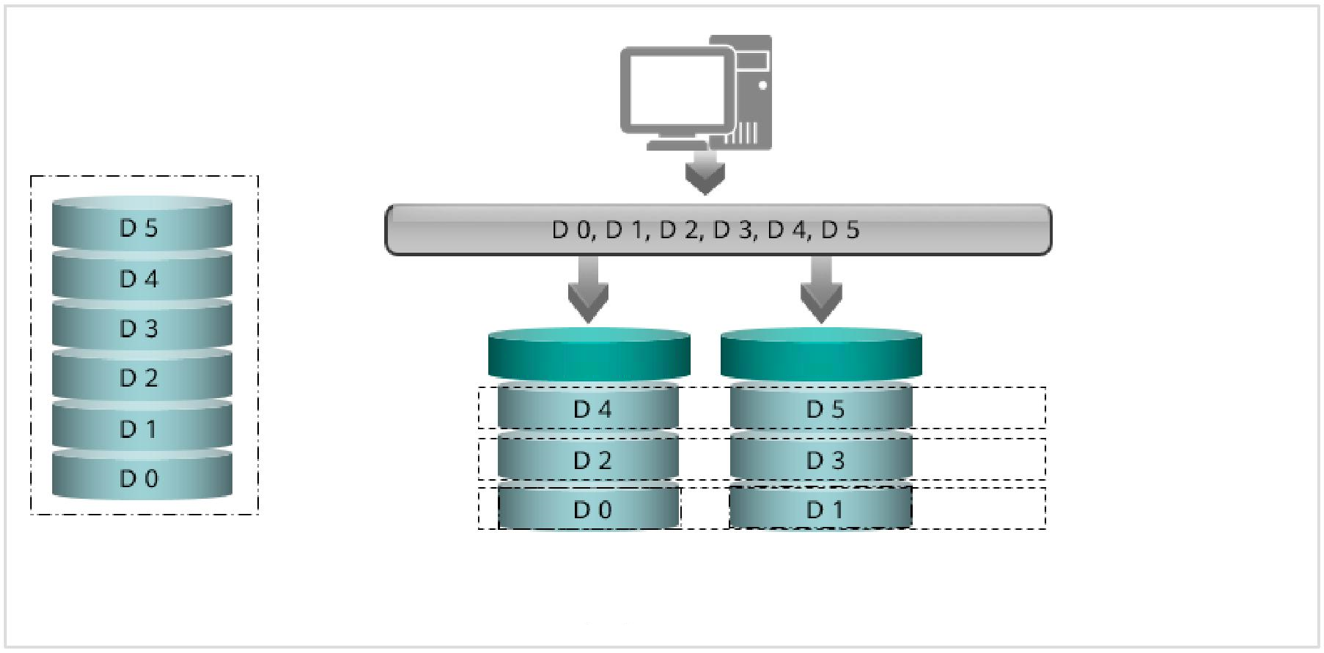
* + - * + Полоса: полоса включает один или несколько последовательных секторов на диске, а несколько полос образуют страйп.
        + Страйп: страйп состоит из полос с одинаковым расположением или ID (идентификатором) на нескольких дисках в одном массиве.

RAID-массивы обычно предоставляют два метода защиты данных.

* + - * + Первый предполагает хранение копии данных на резервном диске для повышения отказоустойчивости и производительности чтения.
        + Второй — проверка четности. Данные о четности — это дополнительная информация, рассчитываемая на основе пользовательских данных. В RAID-массивах, поддерживающих проверки четности, используется выделенный диск четности. Для проверки четности используется алгоритм XOR (символ: ⊕).

#### 2.2.1.2 RAID 0

RAID 0, также называемый чередованием (striping), обеспечивает лучшую производительность хранилища среди всех уровней RAID. В RAID 0 используется технология чередования для распределения данных по всем дискам в RAID-массиве.



Чтение данных из полос D 2, D 3...

Чтение данных из полосы D 1.

Чтение данных из полосы D 0.

Запись данных в полосы D 2, D 3...

Запись данных в полосу D 1.

Запись данных в полосу D 0.

Страйп 2

Страйп 1

Страйп 0

Чередование без проверки ошибок

Полосы данных в диске

Полосы данных в диске

Логический диск

Физический диск 2

Физический диск 1

##### Рисунок 2-1. Принцип работы RAID 0

Массив RAID 0 должен включать не менее двух дисков. Массив RAID 0 разделяет данные на блоки разного размера от 512 байт до мегабайт (обычно кратных 512 байтам) и одновременно записывает такие блоки данных на разные диски. На рисунке выше представлен массив RAID 0 из двух дисков (накопителей). Первые два блока данных записываются в страйп 0: первый блок данных записывается в D 0 на диске 1, а второй блок данных — в D 1 на диске 2. Затем следующий блок данных записывается в D 2 на диске 1 и т. д. В этом режиме нагрузки ввода/вывода распределяются между всеми дисками в RAID-массиве. Поскольку скорость передачи данных по шине намного выше, чем скорость чтения и записи данных на диски, можно считать, что операции чтения и записи данных на диски обрабатываются одновременно.

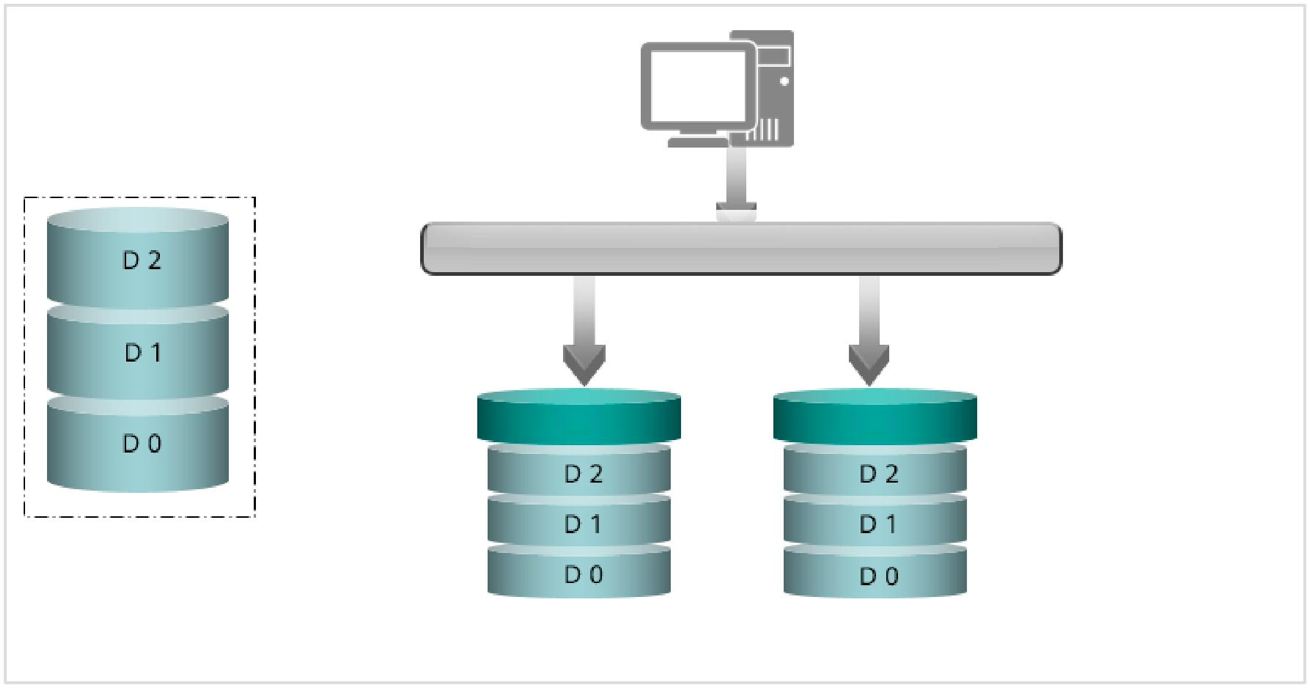
Массив RAID 0 предоставляет диск большой емкости с высокой производительностью обработки ввода/вывода. До появления RAID 0 использовалась аналогичная технология Just a Bundle Of Disks (JBOD). JBOD — это виртуальный диск большой емкости, состоящий из нескольких физических дисков. В отличие от RAID 0, JBOD не выполняет запись блоков данных одновременно на разные диски. В рамках JBOD переход к следующему диску только после того, как закончится место на первом диске. Таким образом, общая емкость массива JBOD равна сумме емкостей всех составляющих этот массив дисков, но его производительность остается на уровне отдельных накопителей.

В массивах RAID 0, в свою очередь, после получения запроса на чтение поиск соответствующих блоков и чтение данных выполняются сразу со всех дисков массива. На рисунке выше представлен процесс чтения данных. Производительность чтения/записи массива RAID 0 прямо пропорциональна количеству дисков, составляющих этот массив.

#### 2.2.1.3 RAID 1

RAID 1, также называемый зеркалированием (mirroring), обеспечивает максимальную степень безопасности данных. В массивах RAID 1 используются диски, являющиеся полными копиями друг друга. При записи данных на диск их копия также сохраняется на втором диске. В случае сбоя исходного (физического) диска зеркальный диск принимает на себя его задачи, поддерживая непрерывность обслуживания. Зеркальный диск используется в качестве резервного для обеспечения высокой надежности.

Объем массива RAID 1 из двух дисков фактически равен объему одного из них, а на втором диске сохраняются копии данных. Таким образом, для хранения каждого гигабайта данных в массиве RAID 1 требуется 2 гигабайта дискового пространства. Коэффициент использования пространства в массиве RAID 1 из двух дисков составляет 50 %.



Зеркалирование полос D 0, D 1 и D 2

Чтение/запись данных в полосе D 2.

Чтение/запись данных в полосе D 1.

Чтение/запись данных в полосе D 0.

Физический диск 2

Физический диск 1

Логический диск

Дисковый массив с зеркалированием

##### Рисунок 2-2. Принцип работы RAID 1

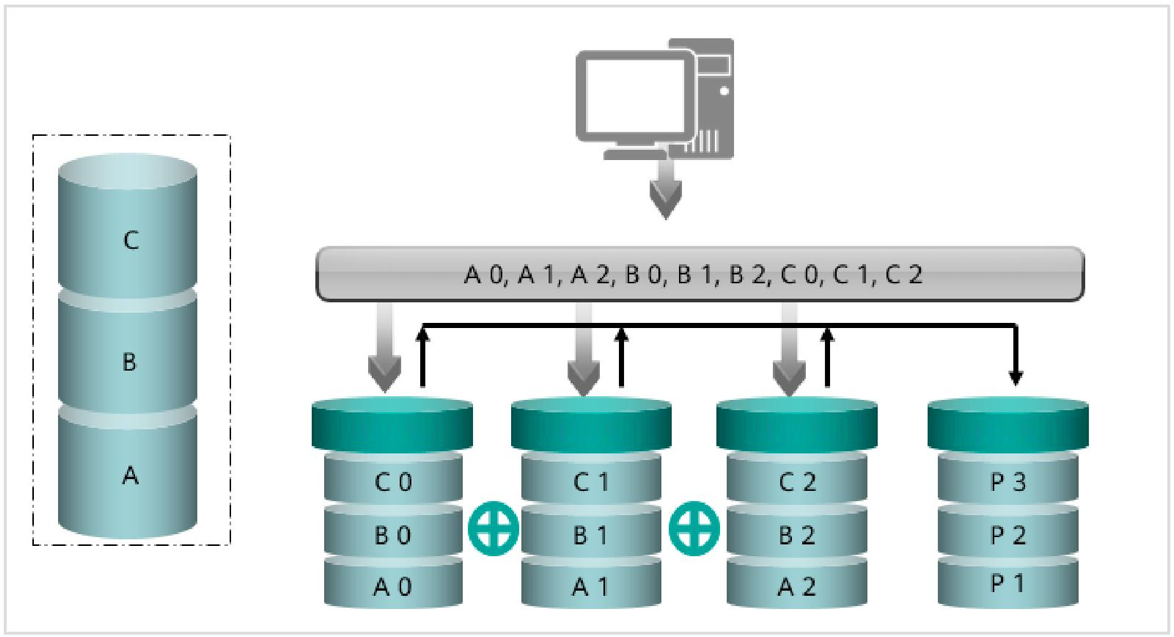
В отличие от RAID 0, где используется чередование для параллельной записи разных данных на разные диски, в массиве RAID 1 одни и те же данные записываются на каждый диск, что позволяет обеспечить их согласованность. На рисунке выше блоки данных записываются на диски в D 0, D 1 и D 2. Блок данных записывается в D 0 каждого диска (диск 1 и диск 2) одновременно. Прочие блоки данных записываются в массив RAID 1 аналогичным образом с поддержкой зеркалирования. Как правило, производительность записи массива RAID 1 остается на уровне одного диска.

Производительность чтения же увеличивается, поскольку в массиве RAID 1 данные одновременно считываются с обоих дисков. В случае сбоя одного диска данные можно считать с другого.

Производительность чтения для массива RAID 1 равна сумме производительности чтения двух дисков. При сбое одного из дисков производительность чтения для RAID-массива снижается вдвое.

#### 2.2.1.4 RAID 3

Принцип работы RAID 3 похож на RAID 0, но при этом также используются выделенные области четности. В массиве RAID 3 используется выделенный диск (диск четности) для хранения данных четности других дисков в одном страйпе. Используя данные четности, можно исправить ошибки в данных при их обнаружении или восстановить данные неисправного диска. RAID 3 применяется в средах, требующих интенсивного использования данных, или в однопользовательских средах, где к блокам данных требуется осуществлять непрерывный доступ в течение длительного времени. В рамках массива RAID 3 данные записываются на все диски данных. Однако при записи новых данных на любой из дисков контроллер RAID 3 пересчитывает и перезаписывает данные четности. Следовательно, при записи большого объема данных приложением диск четности в массиве RAID 3 сталкивается с серьезной нагрузкой. Операции с четностью оказывают определенное влияние на производительность чтения и записи массива RAID 3. Кроме того, из-за больших рабочих нагрузок диск четности характеризуется наибольшей интенсивностью отказов среди всех дисков в массиве RAID 3. Этот уровень RAID не очень эффективен при записи на несколько дисков небольшого объема данных — эффективность чтения в этом случае не отличается от чтения с отдельного диска.



Запись данных в полосу C.

Запись данных в полосу B.

Запись данных в полосу A.

Чтение данных.

Физический диск 1

Диск четности

Физический диск 3

Физический диск 2

Логический диск

##### Рисунок 2-3. Принцип работы RAID 3

В массиве RAID 3 один из дисков используется для поддержки отказоустойчивости, а передача данных выполняется параллельно. В RAID 3 используется чередование для разделения данных на блоки и алгоритм XOR для записи данных четности на соответствующий диск (диск четности).

Производительность записи массива RAID 3 зависит от объема изменяемых данных, числа дисков и времени, затрачиваемого на расчет и сохранение данных четности. Если массив RAID 3 состоит из *N* дисков с одинаковой скоростью вращения, а штраф за запись не учитывается, производительность последовательного ввода/вывода для такого массива при записи данных сразу на все диски теоретически будет превышать производительность записи на один диск чуть менее, чем в N − 1 раз (дополнительное время уходит на выполнение проверки на избыточность).

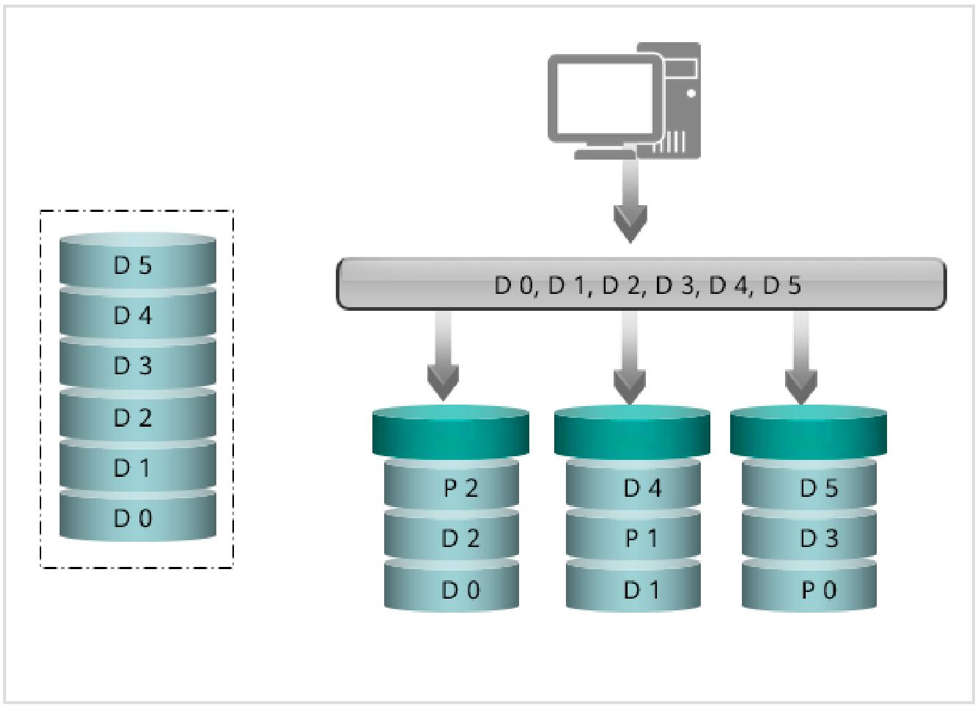
Данные считываются из массива RAID 3 страйпами. Блоки данных в страйпе могут считываться параллельно, так как контроллер управляет одновременно всеми дисками.

RAID 3 выполняет чтение и запись данных параллельно. Производительность чтения массива RAID 3 зависит от объема считываемых данных и количества составляющих массив дисков.

#### 2.2.1.5 RAID 5

RAID 5 — улучшенная технология на основе RAID 3, предполагающая чередование и проверку четности. В массиве RAID 5 данные записываются на диски с чередованием. В массиве RAID 5 данные четности распределяются между дисками данных без использования диска четности.

Как и в RAID 3, эффективность записи падает при записи небольшого объема данных.



Физический диск 2

Физический диск 3

Физический диск 1

Независимая дисковая структура с распределением кода проверки четности

Логический диск

Запись данных.

Чтение данных.

##### Рисунок 2-4. Принцип работы RAID 5

Производительность записи массива RAID 5 зависит от объема записываемых данных и числа составляющих массив дисков. Если массив RAID 5 состоит из *N* дисков с одинаковой скоростью вращения, а штраф за запись не учитывается, производительность последовательного ввода/вывода для такого массива при записи данных сразу на все диски теоретически будет превышать производительность записи на один диск чуть менее, чем в N − 1 раз (дополнительное время уходит на выполнение проверки на избыточность).

В случае сбоя одного из дисков в массивах RAID 3 или RAID 5 массив переходит из оперативного (нормального) состояния в состояние пониженной производительности до полного восстановления неисправного диска. В случае сбоя второго диска данные в массиве будут потеряны.

#### 2.2.1.6 RAID 6

Механизмы защиты данных во всех рассмотренных выше RAID-массивах (за исключением RAID 0) учитывали только отказы отдельных дисков. Время восстановления дисков увеличивается по мере роста их емкости. Восстановление массива RAID 5 из дисков большой емкости может занять несколько дней, а не часов. Во время восстановления массив будет находиться в состоянии пониженной производительности, а отказ любого дополнительного диска приведет к отказу всего массива и потере данных. По этой причине некоторым организациям или подразделениям требуется система двойного резервирования. То есть им нужен RAID-массив, способный выдерживать отказы до двух дисков, поддерживая при этом нормальный доступ к данным. Подобную систему защиты данных с двойным резервированием можно реализовать следующими способами:

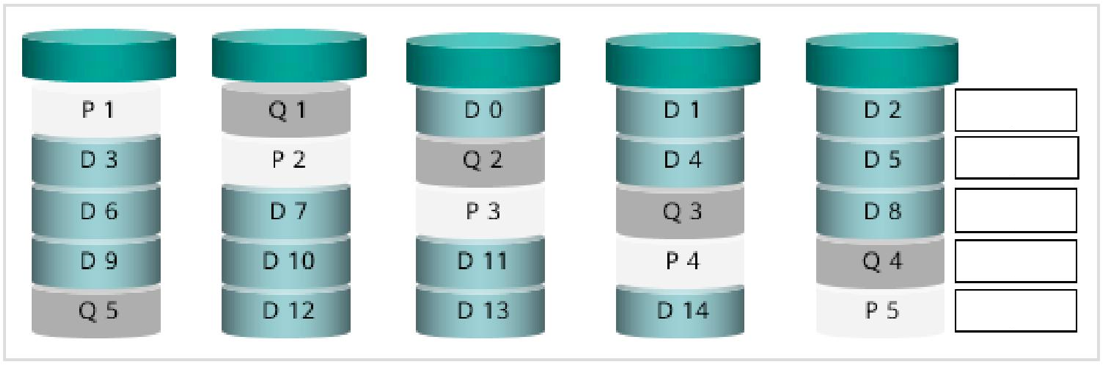
* + - * + Первый — множественное зеркалирование. Множественное зеркалирование подразумевает сохранение нескольких копий блока данных, записываемого на основной диск, на резервных дисках. Однако это приводит к существенным накладным расходам.
        + Второй способ — массив RAID 6. Массив RAID 6 обеспечивает защиту данных за счет выдерживания отказов до двух дисков даже одновременно.

Официальное название RAID 6 — распределенный RAID с двойной четностью (DP). По сути, это улучшенная технология RAID 5, которая также поддерживает чередование и распределенную четность. RAID 6 также поддерживает двойную четность, что означает следующее:

* + - * + При записи пользовательских данных расчет четности выполняется дважды. Поэтому RAID 6 характеризуется наименьшей скоростью записи среди всех уровней RAID.
        + Дополнительные данные четности сохраняются на двух дисках. Поэтому RAID 6 также считается RAID-массивом N + 2.

В настоящее время массив RAID 6 можно организовать разными способами. Для получения данных о четности используются разные методы.

RAID 6 P+Q



Физический диск 3

Физический диск 4

Страйп 0

Страйп 1

Страйп 2

Страйп 3

Страйп 4

Физический диск 2

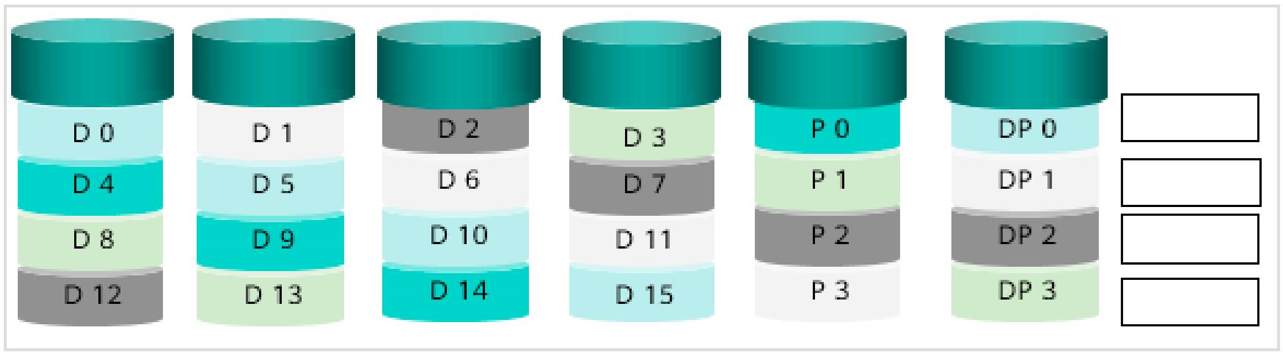
Физический диск 1

Физический диск 5

##### Рисунок 2-5. Принцип работы RAID 6 P+Q

* + - * + Если в массиве RAID 6 используется контроль четности P+Q, P и Q представляют два независимых набора данных четности. Для расчета данных четности P и Q используются разные алгоритмы. Пользовательские данные и данные четности сохраняются в одном страйпе на всех дисках.
        + Как показано на рисунке, значение P 1 рассчитывается путем выполнения операции XOR для данных в D 0, D 1 и D 2 в страйпе 0, P 2 — путем выполнения операции XOR для данных в D 3, D 4 и D 5 в страйпе 1, а P 3 — путем выполнения операции XOR для данных в D 6, D 7 и D 8 в страйпе 2.
        + Значение Q 1 рассчитывается путем выполнения преобразования GF, а затем операции XOR для данных в D 0, D 1 и D 2 в страйпе 0, Q 2 — путем выполнения преобразования GF, а затем операции XOR для данных в D 3, D 4, и D 5 в страйпе 1, а Q 3 — путем выполнения преобразования GF, а затем операции XOR для данных в D 6, D 7 и D 8 в страйпе 2.
        + В случае повреждения области памяти данные на соответствующем диске можно восстановить с помощью значения четности P. При этом нужно выполнить операцию XOR для значения четности P и других дисков данных. Если одновременно выходят из строя области памяти двух дисков в одном страйпе, то в зависимости от сценария применяются разные решения. Если ни на одном из двух неисправных дисков не хранится значение четности Q, данные можно восставить, а затем пересчитать данные четности. Если на одном из двух неисправных дисков хранится значение четности Q, данные на двух неисправных дисках можно восстановить с использованием обеих формул.

RAID 6 DP



Физический диск 2

Физический диск 1

Диск четности по горизонтали

Диск четности по диагонали

Физический диск 4

Физический диск 3

Страйп 0

Страйп 1

Страйп 2

Страйп 3

##### Рисунок 2-6. Принцип работы RAID 6 DP

* + - * + В массивах RAID 6 DP также может использовать два независимых типа данных четности. Первый тип соответствует данным четности P в рамках RAID 6 P+Q. Второй тип — это данные четности по диагонали, рассчитываемые путем выполнения операции XOR по диагонали. Данные четности по горизонтали рассчитываются путем выполнения операции XOR над пользовательскими данными в одном страйпе. Как показано на рисунке выше, значение P 0 рассчитывается путем выполнения операции XOR для данных в D 0, D 1, D 2 и D 3 в страйпе 0, а P 1 — путем выполнения операции XOR для данных в D4, D5, D6, и D 7 в страйпе 1. Таким образом, P 0 = D 0 ⊕ D 1 ⊕ D 2 ⊕ D 3, P 1 = D 4 ⊕ D 5 ⊕ D 6 ⊕ D 7 и так далее.
        + Второй тип данных четности рассчитывается путем выполнения операции XOR над диагональными блоками данных в массиве. Процесс выбора блоков данных относительно сложен. Значение DP 0 рассчитывается путем выполнения операции XOR для D 0 на диске 1 в страйпе 0, D 5 на диске 2 в страйпе 1, D 10 на диске 3 в страйпе 2 и D 15 на диске 4 в страйпе 3. Значение DP 1 рассчитывается путем выполнения операции XOR для D 1 на диске 2 в страйпе 0, D 6 на диске 3 в страйпе 1, D 11 на диске 4 в страйпе 2 и значения P 3 на первом диске четности в страйпе 3. Значение DP 2 рассчитывается путем выполнения операции XOR для D 2 на диске 3 в страйпе 0, D 7 на диске 4 в страйпе 1, значения P 2 на первом диске четности в страйпе 2 и D 12 на диске 1 в страйпе 3. Таким образом, DP 0 = D 0 ⊕ D 5 ⊕ D 10 ⊕ D 15, DP 1 = D 1 ⊕ D 6 ⊕ D 11 ⊕ P 3 и так далее.
        + Массив RAID 6 выдерживает отказ до двух дисков.
        + Независимо от способа реализации (DP или P+Q), производительность массива RAID 6 относительно низка. Поэтому RAID 6 в основном применяется в следующих двух сценариях:

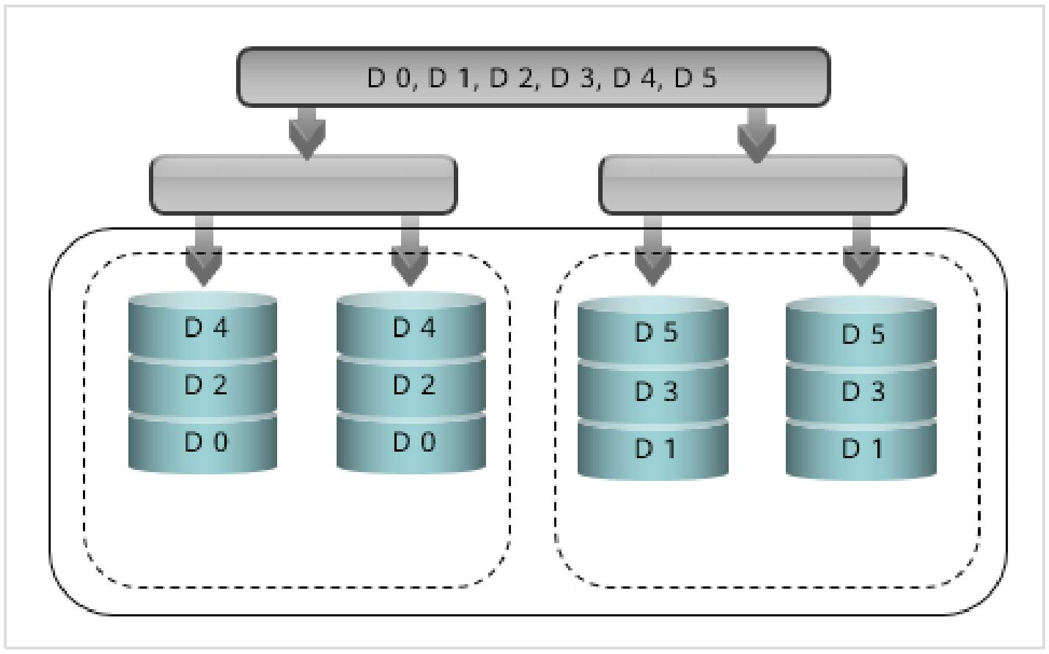
Данные имеют критически важное значение и должны быть доступны постоянно.

Используются диски большой емкости (обычно больше 2 ТБ). Реконструкция диска большой емкости занимает много времени. Поэтому данные будут недоступны в течение длительного времени, если одновременно выйдут из строя два диска. Массив RAID 6 допускает отказ второго диска во время восстановления первого. Некоторые предприятия используют RAID-массив с двойной избыточностью для обеспечения защиты данных на дисках большой емкости.

#### 2.2.1.7 RAID 10

Для большинства предприятий массивы RAID 0 — не самый практичный выбор, а RAID 1 не удовлетворяют требованиям в отношении эффективности использования дискового пространства. Технология RAID 10 — оптимальное решение, объединяющее RAID 1 и RAID 0. В частности, RAID 10 обеспечивает повышенную производительность за счет устранения штрафов при произвольной записи.

Массив RAID 10 всегда включает четное число дисков. На одну половину дисков записываются пользовательские данные, а на другую — их полные копии. Зеркалирование выполняется по страйпам.



Зеркалирование диска

Зеркалирование диска

Пользовательские данные

RAID 1

RAID 0

Физический диск 4

Физический диск 1

Физический диск 2

Физический диск 3

RAID 1

##### Рисунок 2-7. Принцип работы RAID 10

Как показано на рисунке, каждая пара физических дисков (диски 1 и 2, диски 3 и 4) образует отдельный массив RAID 1. Два массива RAID 1 образуют массив RAID 0.

При записи данных в массив RAID 10 блоки данных параллельно записываются в составляющие его массивы с зеркалированием. Как показано на рисунке, данные записываются в D 0 на физический диск 1, а их точная копия — на физический диск 2.

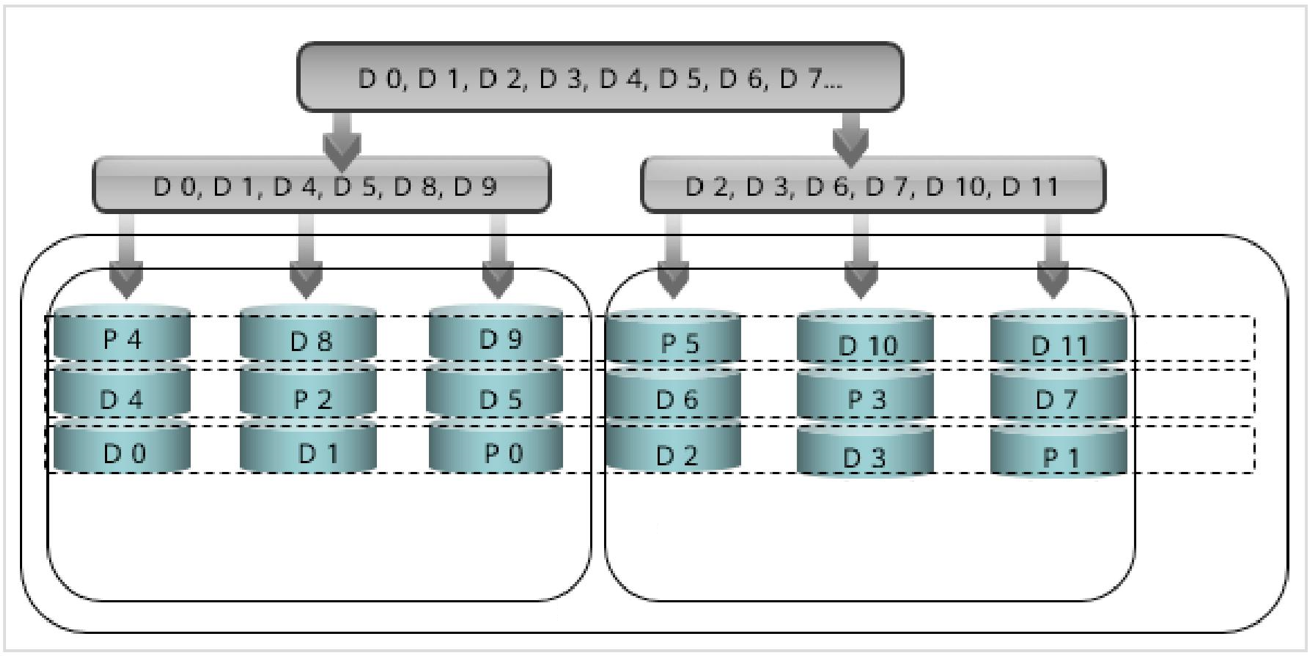
Если диски (например, диск 2 и диск 4) в обоих массивах RAID 1 выйдут из строя, к данным в массиве RAID 10 все равно можно будет получить доступ обычным способом. Это связано с тем, что полноценные копии данных неисправных дисков 2 и 4 сохраняются на двух других дисках (например, на диске 3 и диске 1).

Однако в случае отказа обоих дисков (например, дисков 1 и 2) в одном массиве RAID 1 данные будут недоступны.

Теоретически массив RAID 10 выдерживает отказ половины физических дисков. Однако в худшем случае отказ двух дисков в одном массиве может привести к потере данных. Как правило, массив RAID 10 обеспечивает защиту данных от сбоя одного диска.

#### 2.2.1.8 RAID 50

RAID 50 сочетает технологии RAID 0 и RAID 5. Два массива RAID 5 образуют массив RAID 0. Два массива RAID 5 независимы друг от друга. Для создания массива RAID 50 требуется как минимум шесть дисков, поскольку для организации каждого массива RAID 5 нужно не менее трех дисков.



RAID 5

RAID 5

Страйп 2

Страйп 1

Страйп 0

RAID 0

Физический диск 6

Физический диск 5

Физический диск 4

Физический диск 3

Физический диск 2

Физический диск 1

##### Рисунок 2-8. Принцип работы RAID 50

Как показано на рисунке, диски 1, 2 и 3 образуют массив RAID 5, а диски 4, 5 и 6 — второй массив RAID 5. Два массива RAID 5 образуют массив RAID 0.

Массив RAID 50 выдерживает одновременный отказ нескольких дисков. Однако отказ двух дисков в одном массиве RAID 5 приведет к потере данных.

### 2.2.2 RAID 2.0+

#### 2.2.2.1 Развитие технологий RAID

Технология RAID — высокоразвитый и надежный стандарт защиты данных на дисковых накопителях, вследствие чего она всегда использовалась в качестве основы систем хранения данных. Однако традиционные RAID-массивы теряют свою эффективность, в частности в отношении восстановления дисков большой емкости, по мере постоянного роста требований к хранилищам и емкости дисков.

Недостатки традиционных RAID-массивов связаны с высоким риском потери данных и влияния на предоставление услуг.

* + - * + Высокий риск потери данных: постоянно увеличивающаяся емкость дисков приводит к увеличению времени восстановления и увеличению риска потери данных. Двойное резервирование не действует во время восстановления, и в случае выхода из строя любого дополнительного диска или блока данных данные будут потеряны. Таким образом, увеличение времени восстановления приводит к увеличению риска потери данных.
        + Существенное влияние на предоставление услуг: во время восстановления диски в RAID-массиве заняты этим процессом и работают в режиме низкой производительности, что затрагивает работу сервисов верхнего уровня.

Для решения вышеописанных проблем традиционных RAID-массивов и эффективного использования технологий виртуализации появились следующие альтернативные решения:

* + - * + Виртуализация LUN: традиционный RAID-массив делится на блоки небольшого размера. Эти блоки распределяются по местам хранения, доступным хостам.
        + Виртуализация блоков: диски в пуле хранения делятся на небольшие блоки данных. RAID-массив создается на основе этих блоков данных, благодаря чему данные можно равномерно распределить по всем дискам в пуле. Управление ресурсами в этом случае осуществляется на основе блоков данных.

#### 2.2.2.2 Основные принципы RAID 2.0+

В рамках RAID 2.0+ физический диск делится на блоки (Chunks, CK). Блоки на разных дисках образуют группы блоков (Chunk group, CKG). Группы блоков связаны отношениями RAID между собой. Несколько групп блоков образуют крупный пул ресурсов хранения. Ресурсы такого пула выделяются хостам.

Механизм реализации RAID 2.0+:

* + - * + Несколько SSD-накопителей образуют пул хранения.
        + Каждый SSD-накопитель делится на блоки фиксированного размера (обычно 4 МБ) для управления логическим пространством.
        + Блоки на разных накопителях образуют группы блоков на основе политики RAID, выбранной в DeviceManager.
        + Группы блоков делятся на более мелкие составляющие (grains — зерна), объем которых обычно составляет 8 КБ. Такие зерна сопоставляются с логическими модулями LUN для поддержки управления ресурсами хранения с высокой степенью детализации.

Массивы RAID 2.0+ превосходят традиционные массивы RAID по следующим параметрам:

* + - * + Балансировка сервисной нагрузки, снижающая интенсивность отказов: данные равномерно распределяются по всем дискам в пуле ресурсов хранения, что позволяет уменьшить количество операций записи на каждый диск и, таким образом, продлить срок их службы.
        + Быстрое восстановление, снижающее риск отказа дополнительных дисков: в процессе восстановления после отказа диска восстанавливаемые данные распределяются по всем остальным работающим дискам в пуле (быстрое восстановление по типу «многие ко многим»), что, по сути, восстанавливает конфигурацию резервирования.
        + Нагрузка в процессе восстановления равномерно распределяется между всеми дисками в пуле ресурсов, что позволяет минимизировать влияние отказа на работу приложений верхнего уровня.

#### 2.2.2.3 Элементы RAID 2.0+

1. Дисковые домены

Дисковый домен (Disk Domain) объединяет несколько дисков (или даже все диски в массиве). После объединения дисков и выделения места для поддержки емкости горячего резерва каждый дисковый домен выделяет ресурсы в пул хранения.

При использовании традиционной технологии RAID сначала необходимо создать RAID-массив для выделения дискового пространства узлам служб. Однако в отношении создания RAID-массива существуют определенные ограничения и требования: RAID-массив должен состоять из дисков одного типа и размера, поддерживающих одинаковую скорость вращения, и не может включать более 12 дисков.

RAID 2.0+ от Huawei создается иным способом. Сначала создается дисковый домен. Дисковый домен представляет массив дисков. Диск может принадлежать только к одному дисковому домену. В системе хранения данных OceanStor можно создать один или несколько дисковых доменов. Дисковый домен напоминает RAID-массив. Оба состоят из дисков, но на самом деле существенно различаются. RAID-массив состоит из дисков одного типа и размера, поддерживающих одинаковую скорость вращения. При этом такие диски связаны с определенным уровнем RAID. Дисковый домен, в свою очередь, может включать более 100 дисков трех разных типов. Каждый тип диска связан с уровнем хранения. Твердотельные накопители относятся к уровню высокой производительности, диски SAS — к уровню производительности, а диски NL-SAS — к уровню емкости. Если в дисковом домене нет дисков какого-либо типа, то нет и соответствующего уровня хранения. Дисковые домены разделяют массивы дисков, обеспечивая полную изоляцию в отношении отказов и поддерживая независимость их производительности и ресурсов хранения. При создании дискового домена не указываются уровни RAID. То есть не указываются методы резервирования и защиты данных. Фактически RAID 2.0+ предоставляет более гибкие и подходящие для конкретных приложений методы защиты данных. Пространство хранения, предоставляемое дисками в дисковом домене, делится на меньшие пулы хранения и емкость горячего резерва, совместно используемое всеми уровнями хранения. Система автоматически настраивает емкость горячего резерва на основе выбранной администратором для дискового домена политики горячего резервирования (высокий, низкий или нулевой объем), а также количества дисков на каждом уровне хранения в дисковом домене. В традиционном RAID-массиве администратор указывает конкретный диск в качестве диска горячего резерва.

1. Пулы и уровни хранения

Пул хранения — это контейнер ресурсов хранения. Все ресурсы хранения, которые используются серверами приложений, поступают из пулов хранения данных.

Уровень хранения представляет набор накопителей одного уровня производительности в пуле хранения. В рамках различных уровней хранения осуществляется управление накопителями соответствующего типа и выделения пространства хранения для приложений с различными требованиями к производительности.

Блоки хранения динамически выделяются из пула хранения соответствующего дискового домена и формируют группы блоков в соответствии с заданной для каждого уровня хранения политикой RAID с целью предоставления приложениям защищенных технологией RAID ресурсов хранения.

Пул хранения можно разделить на несколько уровней в зависимости от типов используемых дисков.

При создании пула хранения можно задать его емкость, уровень хранения и соответствующую политику RAID.

Системы хранения OceanStor поддерживают технологии RAID 1, RAID 10, RAID 3, RAID 5, RAID 50 и RAID 6 и соответствующие политики RAID.

К уровню емкости относятся диски SATA и NL-SAS большой емкости. Рекомендуется использовать DP RAID 6.

1. Группа дисков

Система хранения OceanStor автоматически распределяет диски каждого типа в каждом дисковом домене в одну или несколько групп дисков (DG) в зависимости от количества дисков.

Группа дисков может включать диски только одного типа.

Блоки хранения выделяются в группы блоков с разных дисков в группе дисков.

Группы дисков — это внутренние объекты системы хранения OceanStor, которые настраиваются автоматически и обычно используются для изоляции отказов. Группы дисков не имеют внешнего представления.

1. Логический диск

Логический диск (LD) — это управляемый СХД диск, соответствующий физическому диску.

1. Блок

Блок (Chunk, CK) — это дисковое пространство указанного размера, выделенное из пула хранения. Это базовая единица хранения в RAID-массиве.

1. Группа блоков

Группа блоков данных (Chunk Groups, CKG) — это логическая единица хранения, состоящая из блоков, выделяемых разными дисками в одной группе дисков в соответствии с алгоритмом RAID. Это минимальная единица выделения ресурсов из дискового домена в пул хранения.

Все блоки в группы блоков выделяются дисками в одной группе дисков. Группа блоков имеет атрибуты RAID, которые фактически настраиваются для соответствующих уровней хранения. Блоки и группы блоков — это внутренние объекты систем хранения, которые настраиваются автоматически. Они не имеют внешнего представления.

1. Экстенты

Каждая группа блоков делится на логические элементы пространства хранения определенного настраиваемого размера. Такие элементы называются экстентами. Экстент — это минимальная единица (степень детализации) для миграции и статистики горячих данных. Это также минимальная единица использования и освобождения пространства в пуле хранения.

Экстент принадлежит тóму или LUN. При создании пула хранения можно задать размер экстента. Впоследствии размер экстента нельзя будет изменить. Различные пулы хранения могут быть разбиты на экстенты разного размера, но один пул хранения включает экстенты только одного размера.

1. Зерна

При создании тонкого LUN экстенты делятся на блоки размером 64 КБ, которые называются зернами. Тонкий LUN выделяет пространство хранения зернами. Логические адреса блоков (LBA) в зерне последовательны.

Зерна сопоставляются с тонкими LUN. Толстые LUN не используют зерна.

1. Том и LUN

Том — это внутренний объект управления в системе хранения данных.

LUN — это единица хранения, которую можно напрямую сопоставить с хостом для чтения и записи данных. LUN — это внешнее воплощение тома.

Том организует все экстенты и зерна LUN, запрашивает и освобождает экстенты для увеличения или уменьшения фактического пространства, занимаемого томом.

### 2.2.3 Прочие технологии RAID

#### 2.2.3.1 Алгоритм динамического RAID Huawei

При отказе компонента флеш-памяти алгоритм динамического RAID Huawei может в упреждающем порядке восстановить данные этого компонента, продолжая поддерживать защиту данных по технологии RAID.

Этот алгоритм RAID динамически регулирует количество блоков данных в RAID-массиве в соответствии с требованиями к надежности и емкости системы. Если блок отказал, а на дисках за пределами дискового домена блоки недоступны, система динамически восстанавливает исходные блоки N + M в блоках (N − 1) + M. При добавлении нового SSD-накопителя система переносит данные из блоков (N − 1) + M во вновь созданные блоки N + M для обеспечения эффективного использования дискового пространства.

В динамическом RAID используется алгоритм стирающего кода (erasure coding, EC), который позволяет динамически регулировать количество блоков в группе блоков при использовании SSD-накопителей для удовлетворения требований к надежности и емкости системы.

#### 2.2.3.2 RAID-TP

Технология RAID имеет важнейшее значение для поддержания неизменно высоких уровней надежности и производительности СХД. Однако неконтролируемый рост времени создания RAID-массивов вследствие резкого увеличения емкости накопителей ставит ее надежность под сомнение.

Технология RAID-TP позволяет добиться оптимальных производительности, надежности и эффективности использования емкости.

Для замены существующих накопителей при обновлении системы заказчикам приходится приобретать диски большей емкости. В результате одна система может включать диски разной емкости. Но как поддержать высокую эффективность использования ресурсов в системе с дисками разной емкости?

В массивах RAID-TP используется оптимизированный алгоритм FlexEC от Huawei, который позволяет системе выдерживать отказы до трех дисков, повышая ее надежность при увеличении необходимого времени восстановления.

Технология RAID-TP в сочетании с алгоритмом FlexEC позволяет сократить объем данных, считываемых с одного диска, на 70 % по сравнению с традиционным RAID-массивом, что сводит к минимуму влияние на производительность системы.

В типичном массиве RAID 6 (4 блока данных: 2 блока четности) показатель эффективности использования ресурсов хранения составляет около 67 %. С учетом этого коэффициент использования емкости флеш-системы хранения Huawei OceanStor с 25 дисками увеличен 20 %.

## 2.3 Стандартные протоколы хранения

### 2.3.1 SCSI

#### 2.3.1.1 Протокол SCSI

SCSI (интерфейс малых компьютерных систем) — это широкий набор стандартов. Протокол SCSI определяет модели и необходимые наборы команд для подключения и передачи данных между различными устройствами.

Стандарты SCSI охватывают устройства, модели и интерфейсы подключения.

* + - * + В документации по архитектуре SCSI рассматриваются базовые модели архитектуры SAM и SPC и подробно описывается архитектура SCSI, включая такие темы, как модель очереди задач и базовая модель общих команд.
        + Документация по устройствам SCSI охватывает реализацию конкретных устройств, в том числе наборы команд для устройств прямого (дисков) и последовательного (ленты) доступа — SBC и SSC.
        + В документации о видах транспорта SCSI рассматриваются протоколы FCP, SAS, iSCSI и FCoE, а также подробно описывается реализация протокола SCSI на носителях.

#### 2.3.1.2 Логическая топология SCSI

Логическая топология SCSI включает инициаторы, целевые устройства и LUN.

* + - * + Инициатор: по сути, SCSI представляет собой архитектуру «клиент-сервер», в рамках которой клиент выступает в качестве инициатора взаимодействия и отправляет команды запросов на целевое устройство SCSI. Обычно инициатором является хост-система.
        + Целевое устройство: устройство, обрабатывающее команды SCSI. Целевое устройство получает команды от хоста и анализирует их. Примером целевого устройства может служить массив хранения.
        + LUN: номер логического раздела целевого устройства на шине SCSI. Целевое устройство может быть представлено несколькими LUN с разными атрибутами. Например, LUN#0 может относиться к диску, а LUN#1 — к другому устройству.

Инициатор и целевое устройство взаимодействуют по принципу стандартной модели «клиент-сервер». При этом каждая команда обрабатывается в запросно-ответном режиме. Инициатор посылает запрос SCSI на целевое устройство. Целевое устройство отвечает на такой запрос, предоставляет службы хранения при помощи LUN и осуществляет управление задачами.

#### 2.3.1.3 Модель инициатора SCSI

Логические уровни инициатора SCSI в разных операционных системах:

* + - * + В Windows инициатор SCSI имеет три логических уровня: драйвер накопителя/ЗУ на магнитной ленте, порт SCSI и минипорт. Порт SCSI поддерживает базовые операции SCSI, такие как обнаружение устройств и сканирование пространства имен.
        + В Linux инициатор SCSI имеет три логических уровня: драйвер устройства SCSI, драйвер среднего уровня scsi\_mod и драйвер адаптера SCSI (HBA). Драйвер среднего уровня scsi\_mod обрабатывает не связанные с устройствами или адаптером SCSI процессы, такие как исключения и обслуживание пространства имен. Драйвер HBA предоставляет подробные сведения о реализации транспорта SCSI, в том числе об упаковке и распаковке команд SCSI. Драйвера устройств, такие как драйвер накопителя SCSI, драйвер ЗУ на магнитной ленте SCSI и драйвер CD-ROM SCSI, поддерживают работу отдельных устройств SCSI.
        + Уровни инициатора SCSI в системе Solaris аналогичны структуре в Linux/Windows: драйвер устройства SCSI, драйвер среднего уровня SSA и драйвер адаптера SCSI.
        + В ОС AIX инициатор SCSI также имеет три уровня: драйвер устройства SCSI, драйвер среднего уровня SCSI и драйвер адаптера SCSI.

#### 2.3.1.4 Модель целевого устройства SCSI

На основе архитектуры SCSI целевые устройства имеют три уровня: уровень порта, средний уровень и уровень устройства.

* + - * + На уровне порта в целевом устройстве осуществляется упаковка или распаковка команд SCSI, передаваемых по соответствующим каналам. Например, на уровне порта команды могут упаковываться в пакеты FPC, iSCSI или SAS, а также распаковываться из таких пакетов.
        + На уровне устройства осуществляется анализ команд SCSI. Устройство сообщает инициатору о том, к какому устройству относится текущий LUN, с использованием команды INQUIRY, а также обрабатывает команды ввода/вывода READ/WRITE.
        + Средний уровень обеспечивает обслуживание пространства LUN, наборов задач и задач (команд). Существует два способа обслуживания пространства LUN: через глобальный LUN для всех портов или через отдельное пространство LUN для каждого порта.

#### 2.3.1.5 Протокол SCSI и система хранения данных

Протокол SCSI — это основной протокол взаимодействия хостов и устройств хранения.

Контроллер отправляет запрос процессору шины на использование шины. По принятии такого запроса высокопроизводительный кэш контроллера начинает отправку данных. При этом шину будет использовать контроллер, а другие подключенные к ней устройства не смогут ее использовать. Однако контроллер шины может прервать передачу данных в любой момент, чтобы предоставить шину другим устройствам для выполнения операций с более высоким приоритетом.

Контроллер SCSI — это своего рода процессор с собственным набором команд и кэш-памятью. Архитектура шины SCSI обеспечивает динамическое выделение ресурсов на обработку задач несколькими устройствами на компьютере. Это позволяет одновременно обрабатывать несколько задач.

#### 2.3.1.6 Адресация в протоколе SCSI

Традиционный контроллер SCSI подключается к одной шине, вследствие чего определяется только один ID шины. Сервер корпоративного уровня может использовать несколько контроллеров и, соответственно, шин SCSI. В сети хранения каждый сетевой адаптер FC HBA или iSCSI подключен к шине. Чтобы эти шины можно было различить, каждой из них должен быть присвоен отдельный ID.

Для адресации устройств, подключенных к шине SCSI, используются ID устройств SCSI и номера LUN. Каждому устройству, подключенному к шине SCSI, должен быть присвоен уникальный ID. Адаптер HBA на сервере также имеет собственный ID устройства: 7. Каждая шина поддерживает до 8 или 16 ID устройств, включая ID адаптера шины. ID устройства используется для адресации устройств и определения их приоритета на шине.

Каждое устройство хранения может включать вспомогательные устройства, такие как виртуальные диски и ленточные накопители. Для адресации таких устройств используются номера LUN.

Для идентификации целевого устройства SCSI используется комбинация из трех идентификаторов: ID шины, ID устройства и номер LUN.

### 2.3.2 iSCSI, FC и FCoE

#### 2.3.2.1 Протокол iSCSI

Первоначально протокол iSCSI был представлен компаниями IBM, Cisco и HP. С 2004 года протокол iSCSI является официальным стандартом IETF. Текущая версия протокола iSCSI основана на модели SAM2 (SCSI Architecture Model-2 — модель архитектуры SCSI 2).

iSCSI — это сокращение от Internet Small Computer System Interface (сетевой интерфейс малых компьютерных систем). Это стандарт сети хранения, использующий технологии IP для поддержки взаимодействия устройств хранения данных. Он обеспечивает доступ к устройствам хранения на уровне блоков за счет передачи команды SCSI по сети TCP/IP.

Протокол iSCSI упаковывает команды SCSI и данные блоков в пакеты TCP для передачи через IP-сеть. iSCSI — один из транспортных протоколов SCSI, в основе которого лежат развитые IP-технологии, поддерживающие и расширяющие возможности сетей SAN. Уровень протокола SCSI генерирует блок дескриптора команды (CDB), а затем отправляет его на уровень протокола iSCSI. На уровне iSCSI блок CDB упаковывается в протокольный блок данных (PDU), который передается по IP-сети.

#### 2.3.2.2 Инициатор и целевое устройство iSCSI

Сети iSCSI перенимают некоторые характеристики сетей SCSI. Во взаимодействии по протоколу iSCSI участвует инициатор, который отправляет запросы ввода/вывода, и целевое устройство, отвечающее на эти запросы и выполняющее соответствующие операции. После установки соединения между инициатором и целевым устройством последнее принимает на себя управление процессом в качестве основного устройства.

* + - * + Существует три типа инициаторов iSCSI: программно-управляемый инициатор, аппаратная сетевая плата TOE (TCP offload engine) и адаптер шины iSCSI. В указанном порядке растет и их производительность.
        + Целевыми устройствами iSCSI обычно выступают дисковые массивы или ленточные библиотеки iSCSI.

Протокол iSCSI определяет набор методов именования и адресации инициаторов и целевых устройств iSCSI. Все узлы iSCSI определяются уникальными именами iSCSI. Данный принцип позволяет четко разграничить имена iSCSI и имена хостов.

#### 2.3.2.3 Архитектура iSCSI

В системе на базе iSCSI пользователь отправляет команду чтения или записи данных на устройство хранения SCSI. Операционная система преобразует эту команду в одну или несколько команд SCSI и отправляет их на контроллер целевого устройства SCSI. Узел iSCSI упаковывает команды и данные в пакеты iSCSI и отправляет их на уровень TCP/IP, где такие пакеты упаковываются в пакеты IP для передачи по сети. Также существует возможность зашифровать команды SCSI для передачи по незащищенной сети.

Пакеты данных можно передавать по локальной сети или через Интернет. Контроллер получающего данные устройства хранения анализирует пакеты данных и отправляет управляющие команды SCSI и данные из пакетов iSCSI на соответствующие диски. Затем диски выполняют запрошенные хостом или приложением операции. В случае запросов на получение данных данные считываются с дисков, а затем пересылаются в хост-систему. Данный процесс полностью прозрачен для пользователей. Обработку команд SCSI и подготовку данных можно реализовать с использованием программного обеспечения сетевого контроллера по модели TCP/IP, однако это может оказать значительное влияние на производительность ЦП хост-системы. Использование выделенных устройств для этих целей позволит свести к минимуму влияние на производительность системы. Адаптер iSCSI сочетает в себе функции сетевой платы и адаптера HBA. Адаптер iSCSI получает данные блоками, классифицирует и обрабатывает их с помощью механизмов TCP/IP, после чего отправляет полученные пакеты IP-данных по IP-сети. Сочетание функций позволяет пользователям создавать сети IP SAN без ущерба для производительности сервера.

#### 2.3.2.4 Протокол FC

Термин «FC» используется для обозначения протокола FC, сети FC или соединения FC. FC характеризуется высокой производительностью и поэтому все чаще используется для обеспечения внешнего доступа хостов в сетях «точка-точка» и сетях на основе коммутаторов. В наборе протоколов FC используются многие концепции из наборов протоколов TCP/IP и Ethernet, в том числе коммутация FC, коммутатор FC, маршрутизация FC, маршрутизатор FC и алгоритм маршрутизации SPF.

Структура протокола FC:

* + - * + FC-0: определяет физические соединения и осуществляет выбор различных физических носителей и скорости передачи данных для работы протокола. Это позволяет обеспечить максимальную гибкость системы и использовать существующие кабели и различные технологии для удовлетворения требований разных систем. Обычно используются медные и оптические кабели.
        + FC-1: записывает 8-битный или 10-битный код передачи для балансировки потока битов передачи. Такой код также может служить для передачи данных и выявления ошибок. Возможность 8-битного/10-битного кодирования позволяет снизить затраты на проектирование компонентов и обеспечивает оптимальную плотность передачи для улучшения показателей восстановления синхронизации. Примечание: 8-битное/10-битное кодирование также поддерживается при использовании интерфейса IBM ESCON.
        + FC-2: определяет правила передачи данных по сети:

Принцип разбиения данных на блоки — кадры

Объем отправляемых в определенный момент данных (управление потоком)

Место отправления данных (включая определение уровней обслуживания на основе приложений)

* + - * + FC-3: определяет расширенные функции, такие как чередование (передача данных по нескольким каналам), многоадресная передача (отправка одного сообщения нескольким адресатам) и групповой запрос (сопоставление нескольких портов с одним узлом). Если FC-2 определяет функции для одного порта, то FC-3 может описать функции для всех портов.
        + FC-4: поддерживает взаимодействие с протоколами верхнего уровня. FC поддерживает инкапсуляцию протоколов IP, SCSI, ATM и др. SCSI — это разновидность протокола FC.

Как и протокол Ethernet, FC поддерживает следующие топологии:

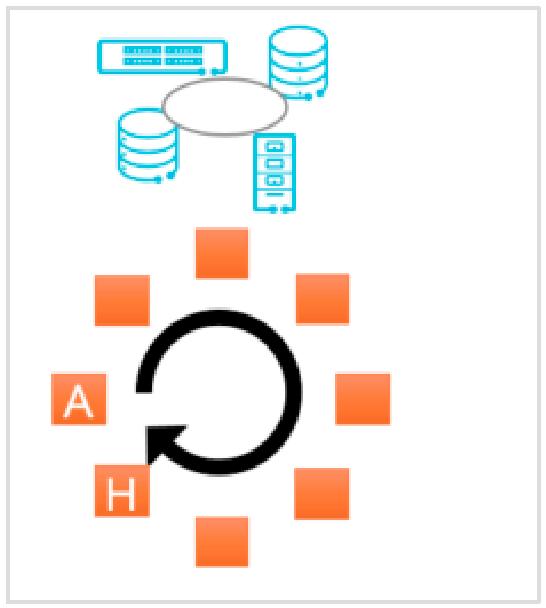
* + - * + Точка-точка:

Простейшая топология, обеспечивающая прямую связь между двумя узлами (обычно устройством хранения и сервером).

* + - * + Управляемая петля:

Аналогична топологии Ethernet «общая шина», но работает в режиме управляемой петли, а не подключения к шине. Каждое устройство подключается к следующему, образуя петлю.

Кадры данных передаются от узла к узлу в рамках управляемой петли, при этом передача возможна только в одном направлении. На рисунке ниже узлу A необходимо установить связь с узлом H. После того как узел A выиграет арбитраж, он посылает кадры данных на узел H. Однако передача осуществляется по часовой стрелке в последовательности B-C-D-E-F-G-H, что неэффективно.



##### Рис. 2-9:

* + - * + Связная архитектура:

Подобно коммутируемой сети Ethernet, связная архитектура представляет собой коммутационную матрицу ячеистой сети.

Эффективность пересылки в ней намного выше, чем при использовании управляемой петли.

Устройства FC подключаются к коммутаторам с использованием оптоволоконных или медных кабелей для установки связи между узлами по принципу «точка-точка».

FC позволяет снять с рабочей станции нагрузку, связанную с управлением каждым портом. Каждый порт самостоятельно управляет своим двухсторонним подключением к сети, а прочие сетевые функции поддерживают коммутаторы FC. В сетях FC используется семь типов портов.

* + - * + Порты устройства (узла):

N\_Port: порт узла. Порт сетевого устройства с поддержкой прямого подключения.

NL\_Port: порт петли узлов. Порт устройства с поддержкой топологии «управляемая петля».

* + - * + Порты коммутатора:

E\_Port: порт расширения (соединение коммутаторов).

F\_Port: сетевой порт для подключения портов N\_Port к коммутатору.

FL\_Port: сетевой порт с поддержкой топологии «управляемая петля».

G\_Port: общий порт, который может определяться как порт типа E\_Port или F\_Port.

U\_Port: универсальный порт для описания автоматически определяемых портов.

#### 2.3.2.5 Протокол FCoE

Протокол FCoE определяет взаимодействие протоколов FC и IEEE 802.3 Ethernet. Он предполагает использование физического уровня и уровня передачи данных Ethernet, а также сетевого уровня, уровня услуг и уровня протокола FC.

Особенности протокола FCoE:

* + - * + Организация стандартов: протокол был представлен для утверждения комитету T11 ANSI (American National Standards Institute — Американский национальный институт стандартизации) в 2008 году. Требуется тесное сотрудничество с IEEE (Институт инженеров электротехники и электроники).
        + Основная задача: использование возможностей масштабирования сети Ethernet, сохраняя при этом высокую степень надежности и эффективности, характерную для FC.
        + Прочие задачи: при использовании FC в сочетании с Ethernet могут возникать проблемы, связанные с потерей пакетов, дублированием маршрутов, отказоустойчивостью, разделением и восстановлением кадров, а также неблокирующей передачей.
        + Протокол FC отличается низкой степенью совместимости и не поддерживает передачу на большие расстояния. FCoE тоже не удалось решить эти проблемы.

FCoE сохраняет стек протоколов выше уровня FC-2, а вместо уровней FC-0 и FC-1 использует уровень передачи данных Ethernet. Исходный уровень FC2 делится на следующие элементы:

* + - * + FC-2V: виртуальный подуровень FC2
        + FC-2M: подуровень мультиплексора FC2
        + FC-2P: виртуальный физический уровень FC2

Протокол сопоставления FC\_BB\_E требует, чтобы FCoE использовал на нижнем уровне стандарты lossless Ethernet для передачи данных и передавал кадры FC в полнодуплексном режиме без потерь. В физическом канале передачи используется протокол Ethernet.

Сравнение протоколов FC и FCoE:

* + - * + Уровень FC-0 определяет тип носителя, а уровень FC-1 — режим кодирования/декодирования кадров. При передаче по сети FC SAN эти два уровня необходимо определить. FCoE, в свою очередь, работает в сети Ethernet. Поэтому уровень передачи данных Ethernet заменяет вышеописанные уровни.
        + Различная рабочая среда: протокол FC использует традиционную сеть хранения FC SAN, а протокол FCoE — сеть Ethernet.
        + Различные рабочие каналы: протокол FC работает в сети FC, и все пакеты передаются по каналам FC. В сети Ethernet одновременно используются пакеты различных протоколов, например пакеты IP и ARP. Для передачи пакетов FCoE в сети Ethernet необходимо создать виртуальный канал FC.
        + В отличие от протокола FC, FCoE использует протокол инициализации FIP для построения VLAN, а также создания и обслуживания виртуальных каналов с механизмами пересылки FCF.

FCoE требует поддержки других протоколов. Ethernet, в отличие от FC, допускает потерю пакетов. Несмотря на работу через сеть Ethernet, в основе FCoE лежит протокол FC, поэтому FCoE также должен передавать пакеты без потерь. Таким образом, для нормальной работы FCoE через Ethernet требуется улучшение протокола Ethernet для предотвращения потери пакетов. Такое улучшение существует и называется Converged Enhanced Ethernet (CEE — конвергентный улучшенный Ethernet).

### 2.3.3 SAS и SATA

#### 2.3.3.1 Протокол SAS

SAS — это последовательный интерфейс и протокол, пришедший на замену протоколу SCSI. Последовательный порт отличается простой структурой, поддержкой горячей замены, а также высокой скоростью передачи и производительностью.

Как правило, большие кабели параллельных портов вызывают электронные помехи. Кабельная структура SAS позволяет решить эту проблему. Кабели SAS занимают меньше места, благодаря чему улучшаются показатели теплоотвода и вентиляции серверов, в которых используются диски SAS.

Преимущества SAS:

* + - * + Низкая стоимость:

Объединительная плата SAS поддерживает диски SAS и SATA, что позволяет снизить затраты, связанные с использованием различных типов дисков.

Не требуется проектировать разные продукты на основе стандартов SCSI и SATA. Кроме того, снижаются сложность прокладки кабелей и количество слоев печатной платы, что приводит к дополнительному снижению затрат.

Системным интеграторам не требуется приобретать разные объединительные платы и кабели для разных дисков.

Можно подключить больше устройств.

Специальные расширители SAS (SAS Expanders) позволяют системе SAS поддерживать больше устройств. Каждый расширитель можно подключить к нескольким портам, а каждый из таких портов, в свою очередь, — к устройству SAS, хосту или другому расширителю SAS.

* + - * + Высокая надежность:

Степень надежности протокола SAS аналогична дискам SCSI и FC и превышает возможности дисков SATA.

SAS поддерживает проверенный набор команд SCSI.

* + - * + Высокая производительность:
* Высокая скорость однонаправленной передачи.
  + - * + Совместимость с SATA:

Среды SAS поддерживают прямую установку дисков SATA.

Диски SATA и SAS можно использовать в одной системе, что соответствует современным тенденциям к использованию многоуровневых СХД.

Архитектура SAS включает шесть уровней: физический уровень, физический уровень передачи данных (phy), уровень каналов, уровень портов, транспортный уровень и уровень приложений. Каждый уровень поддерживает определенные функции.

* + - * + Физический уровень: определяет используемое оборудование, такое как кабели, разъемы и трансиверы.
        + Уровень phy: включает протоколы нижнего уровня, такие как схемы кодирования и последовательности питания/сброса.
        + Уровень каналов: осуществляет управление соединениями уровня phy, базисными элементами, CRC, скремблированием и дескремблированием, а также согласованием скорости передачи.
        + Уровень портов: описывает интерфейсы уровня каналов и транспортного уровня, включая способы запроса, прерывания и установки соединений.
        + Транспортный уровень: определяет процессы упаковки передаваемых команд, состояний и данных в кадры SAS, а также распаковки таких кадров.
        + Уровень приложений: описывает использование протокола SAS в различных типах приложений.

Характеристики протокола SAS:

* + - * + SAS использует полнодуплексный (двунаправленный) режим связи. При подключении через традиционный параллельный интерфейс SCSI передача возможна только в одном направлении. При параллельном подключении, когда устройство получает пакет данных, для ответа потребуется разорвать существующий канал связи SCSI и установить новый. Однако каждый кабель SAS включает по паре входных и выходных кабелей. Благодаря этому при использовании SAS можно одновременно считывать и записывать данные, что значительно повышает эффективность их передачи.
        + Преимущества SAS по сравнению со SCSI:

Более высокая пропускная способность за счет последовательной связи и возможность повышения производительности в будущем.

Четыре «узких» порта можно объединить в «широкий» для обеспечения более высокой пропускной способности.

Возможности масштабирования SAS

* + - * + В архитектуре SAS для расширения интерфейсов используются специальные расширители. Один домен SAS поддерживает до 16 384 накопителей.
        + Расширитель SAS — это устройство взаимодействия в домене SAS, аналогичное коммутаторам Ethernet. Расширители SAS позволяют подключать большее количество устройств и уменьшить затраты на HBA. К каждому расширителю можно подключить до 128 терминалов или расширителей. Домен SAS включает следующие основные компоненты: расширители SAS, терминалы и устройства подключения (соединительные кабели SAS).

Расширитель SAS использует таблицу маршрутизации для отслеживания адресов всех дисков SAS.

Терминалы могут выступать в качестве инициаторов (обычно это HBA SAS) или целевых устройств (диск SAS/SATA или HBA в режиме целевого устройства).

* + - * + Для целей корректного обнаружения устройств формирование петель в домене SAS не поддерживается.
        + В действительности к расширителю подключается значительно меньше 128 терминалов вследствие ограничений полосы пропускания.

Кабельные подключения SAS

* + - * + Большинство производителей устройств хранения используют кабели SAS для подключения дисковых полок между собой или к контроллерным полкам. В кабеле SAS четыре независимых канала (узких порта) объединены в широкий порт с более высокой пропускной способностью. Каждый из четырех независимых каналов поддерживает скорость передачи 12 Гбит/с. Таким образом, пропускная способность широкого порта составляет 48 Гбит/с. Чтобы объемы передаваемых по кабелям SAS данных не превышали значение максимальной пропускной способности такого кабеля, необходимо ограничить общее число дисков, подключенных к петле SAS.
        + Для устройств хранения Huawei максимальное поддерживаемое число дисков в петле составляет 168. Таким образом, петлю могут образовать до семи дисковых полок, поддерживающих до 24 накопителей каждая.

Однако такая конфигурация возможна, только если все диски относятся к традиционным жестким дискам SAS. Сейчас все большую популярность набирают SSD-накопители, которые обеспечивают гораздо более высокую скорость передачи, чем диски SAS. Поддерживается объединение в петлю до 96 SSD-накопителей: до четырех дисковых полок с 24 накопителями каждая.

* + - * + В кабелях SAS используются разъемы мини-SAS, поддерживающие скорость одного канала 6 Гбит/с, и разъемы мини-SAS HD, поддерживающие скорость одного канала 12 Гбит/с.

#### 2.3.3.2 Протокол SATA

SATA (Serial ATA) — это тип компьютерной шины, используемый для передачи данных между материнской платой и устройствами хранения (дисками и приводами CD-ROM). SATA предполагает использование совершенно новой конструкции шины, а не просто улучшение используемой ранее шины PATA.

На уровне разъемов SAS и SATA совместимы. Диски SATA можно использовать в среде SAS. С точки зрения стандартов портов, SATA является второстепенным стандартом SAS. Поэтому контроллер SAS может напрямую управлять дисками SATA. Однако диски SAS использовать напрямую в среде SATA нельзя, поскольку контроллер SATA не поддерживает управление дисками SAS.

На уровне протокола SAS включает три типа протоколов, используемых для обмена данными с различными устройствами.

* + - * + Последовательный протокол SCSI (SSP) используется для передачи команд SCSI.
        + Протокол управления SCSI (SMP) используется для обслуживания подключенных устройств и управления ими.
        + Протокол канала SATA (STP) используется для обмена данными между SAS и SATA.

Сочетание этих трех протоколов позволяет использовать устройства SAS в сочетании с дисками SATA и некоторыми устройствами SCSI.

### 2.3.4 PCIe и NVMe

#### 2.3.4.1 Протокол PCIe

Компания Intel представила концепцию PCI в 1991 году. Характеристики шины PCI:

* + - * + Простая структура шины, низкая стоимость, простота конструкции.
        + Параллельная шина поддерживала ограниченное количество устройств, а ее возможности масштабирования оставляли желать лучшего.
        + При подключении нескольких устройств значительно снижались пропускная способность шины и скорость передачи данных.

С учетом скорости развития современных процессорных технологий поиск способов замены параллельных шин высокоскоростными дифференциальными шинами был неизбежен. В отличие от асимметричных параллельных сигналов высокоскоростные дифференциальные сигналы поддерживают более высокие тактовые частоты. Это и обусловило появление шины PCIe.

PCIe (PCI Express) — это стандарт высокопроизводительного последовательного соединения с высокой пропускной способностью. Впервые он был представлен Intel, а впоследствии разрабатывался организацией PCI-SIG (Peripheral Component Interconnect Special Interest Group) с целью замены шинных коммуникационных архитектур.

По сравнению с традиционной шиной PCI, PCIe имеет следующие преимущества:

* + - * + Двунаправленные соединения (линии), высокая пропускная способность и высокая скорость передачи: PCIe поддерживает полнодуплексный режим передачи (RX и TX разделены). Кроме того, шина поддерживает более высокую скорость передачи. Скорость передачи одной линии PCIe 1.0 составляет 2,5 Гбит/с, PCIe 2.0 — 5 Гбит/с, PCIe 3.0 — 8 Гбит/с, PCIe 4.0 — 16 Гбит/с, а PCIe 5.0 — до 32 Гбит/с.
        + Совместимость: PCIe поддерживает совместимость с PCI на программном уровне, хотя программное обеспечение PCIe более совершенно.
        + Простота использования: поддержка горячей замены. Слоты интерфейса PCIe поддерживают сигнал обнаружения горячей замены.
        + Обработка ошибок и составление отчетов: программный уровень многоуровневой структуры шины PCIe поддерживает обработку ошибок и составление отчетов об ошибках.
        + Виртуальные каналы физических соединений: каждый физический канал поддерживает несколько виртуальных каналов (теоретически для независимого управления передачей доступно до восьми виртуальных каналов) с высоким качеством передачи и управления трафиком.
        + Сокращение необходимого количества каналов ввода/вывода, уменьшение площади платы и снижение уровня перекрестных помех: для линии данных стандартной шины PCI требуется не менее 50 каналов ввода/вывода, в то время как PCIe XL требует только четырех. Сокращение количества каналов ввода/вывода позволяет уменьшить платы и увеличить расстояние между каналами, что, в свою очередь, приводит к снижению перекрестных помех.

Почему PCIe? Стандарт PCIe ориентирован на будущее и продолжает развиваться для поддержки все более высокой пропускной способности. PCIe позволяет добиться увеличения пропускной способности за счет применения новейших технологий, а благодаря многоуровневым протоколам и дискам, поддерживающим совместимость программных моделей, переход от PCI к PCIe не будет вызывать трудностей. В рамках протокола PCIe используются соединения типа «Точка-точка», он обеспечивает высокую надежность, поддержку древовидной структуры, полнодуплексную передачу и передачу на основе кадров.

Протокол PCIe включает физический уровень, уровень каналов передачи данных, уровень транзакций и уровень приложений.

* + - * + Физический уровень в архитектуре шины PCIe определяет физические характеристики шины. В будущем производительность шины PCIe можно будет увеличить за счет увеличения скорости передачи или изменения режимов кодирования или декодирования. Такие изменения затронут только физический уровень, что облегчает процесс обновления.
        + Уровень каналов передачи данных поддерживает корректность и надежность передачи пакетов данных по шине PCIe. На этом уровне проверяется полнота и корректность упаковки пакетов данных, к данным добавляется порядковый номер и код CRC, а также используется протокол взаимной идентификации ack/nack для обнаружения и коррекции ошибок.
        + Уровень транзакций принимает запросы на чтение и запись с уровня приложений или создает пакеты инкапсуляции запроса и передает их на уровень каналов передачи данных. Пакеты этого типа называются пакетами уровня транзакций (transaction layer packet, TLP). TLP принимает пакеты уровня каналов передачи данных (data link layer packets, DLLP), сопоставляет их с соответствующим запросом и передает на уровень приложений для обработки.
        + Уровень приложений представляет реальные приложения. Другие уровни должны соответствовать требованиям протокола.

#### 2.3.4.2 Протокол NVMe

NVMe (Non-Volatile Memory Express) — это спецификация на протоколы доступа к SSD-накопителям, подключенным по шине PCI Express. Прямое соединение между специализированным каналом PCIe и ЦП позволяет избежать задержки, вызванной необходимостью обмена данными между внешним контроллером (PCH) интерфейсов SATA/SAS и ЦП.

В процессах хранения NVMe рассматривается не только как логический интерфейс, но также как новый набор команд и определенный протокол. Низкая задержка и высокий параллелизм каналов PCIe в сочетании с возможностями современных многоядерных процессоров, платформ и приложений позволяют серьезно повысить производительность чтения и записи твердотельных накопителей при адекватном уровне затрат. NVMe приходит на замену протоколу AHCI (Advanced Host Controller Interface), позволяя забыть о задержках, связанных с использованием последнего, и раскрыть весь потенциал SSD-накопителей в эпоху SATA.

Стек протоколов NVMe:

* + - * + На уровне передачи во флеш-массивах SAS данные ввода/вывода передаются от внешнего сервера на ЦП с использованием протокола внешнего интерфейса FC/IP, используемого устройством хранения. Затем с помощью каналов и коммутаторов PCIe они передаются на чип SAS, расширитель SAS и, наконец, на твердотельный накопитель SAS.
        + СХД Huawei типа «All-flash» на базе NVMe отличаются сквозной поддержкой NVMe. Запросы на чтение/запись данных передаются от внешнего сервера на ЦП с использованием протокола внешнего интерфейса FC-NVMe/NVMe Over RDMA, используемого устройством хранения. Внутренние данные передаются напрямую на твердотельные накопители NVMe через интерфейс RDMA 100 Гбит/с. Центральный процессор СХД типа «All-Flash» на базе NVMe напрямую взаимодействует с SSD-накопителями NVMe по наиболее короткому пути, что позволяет повысить эффективность и сократить задержку при передаче.
        + На аналитическом уровне программного протокола системы хранения типа «All-Flash» на базе SAS и NVMe сильно отличаются друг от друга в области взаимодействия протоколов для записи данных. В системах SAS используется протокол SCSI, и для обработки запроса на запись данных требуется четыре взаимодействия. СХД Huawei типа «All-flash» на базе NVMe в два раза более эффективны при обработке запросов на запись, чем аналогичные системы на основе SAS, поскольку требуют всего двух взаимодействий.

Преимущества NVMe:

* + - * + Низкая задержка: при выполнении команд данные не считываются из регистров, что обуславливает низкую задержку ввода/вывода.
        + Высокая пропускная способность: PCIe на 4 линии обеспечивает пропускную способность на уровне до 4 Гбит/с для одного накопителя.
        + Высокий показатель IOPS: глубина очереди в рамках NVMe увеличена с 32 до 64 000. Также NVMe позволяет значительно увеличить показатели IOPS SSD-накопителей.
        + Снижение энергопотребления: NVMe поддерживает автоматическое переключение режимов энергопотребления и динамическое управление питанием, что позволяет значительно снизить энергопотребление.
        + Высокая степень совместимости драйверов: решена проблема совместимости драйверов различных SSD-накопителей PCIe.

В СХД типа «All-Flash» OceanStor Dorado от Huawei используется протокол NVMe Over Fabric для распределения ресурсов SSD-накопителей, а также протоколы FC-NVMe 32 Гбит/с и NVMe over RDMA 100 Гбит/с для поддержки внешних подключений. То есть один сетевой протокол используется для внешних сетевых подключений, внутренних подключений дисковых полок и подключения контроллеров для горизонтального масштабирования.

Оборудование и сетевые технологии RDMA позволяют сетевым платам серверов получать прямой доступ к памяти, что приводит к повышению пропускной способности, а также снижению задержки и сокращению потребления ресурсов. Однако сетевая архитектура IB, ориентированная на RDMA, не подходит для использования в рабочих сетях, что обуславливает высокие затраты. Эту проблему решает протокол RoCE. RoCE — это сетевой протокол, поддерживающий возможность удаленного доступа к памяти (RDMA) через сеть Ethernet. Существует две версии протокола RoCE. RoCEv1 — это протокол канала передачи данных, ограниченный одним широковещательным доменом.

RoCEv2 — это протокол сетевого уровня, поддерживающий маршрутизацию.

### 2.3.5 RDMA и IB

#### 2.3.5.1 Протокол RDMA

RDMA (Remote Direct Memory Access — удаленный прямой доступ к памяти) — это метод передачи данных по сети между приложениями на двух серверах с использованием буферов.

Сравнение традиционных методов и RDMA:

* + - * + В отличие от передачи данных ввода/вывода по внутренней шине в рамках традиционного режима DMA (direct memory access — прямой доступ к памяти), RDMA предполагает прямую передачу данных между приложениями двух конечных точек по сети с использованием буферов.
        + В отличие от традиционной передачи по сети, RDMA не требует вмешательства операционных систем или стеков протоколов партнеров по связи.
        + Метод RDMA позволяет добиться сверхнизкой задержки и сверхвысокой пропускной способности при передаче между конечными точками, не требуя при этом больших объемов ресурсов ЦП и ОС. На обработку и миграцию данных уходит лишь небольшое количество ресурсов.

В настоящее время существует три типа сетей, поддерживающих RDMA: IB, RoCE и iWARP. Сети IB ориентированы на использование RDMA и обеспечивают надежность передачи на аппаратном уровне. RoCE и iWARP представляют собой основанные на Ethernet технологии EDMA (Enhanced direct memory access — расширенный прямой доступ к памяти), поддерживающие соответствующие интерфейсы команд.

* + - * + IB — это сетевой протокол нового поколения, изначально поддерживающий RDMA. Для развертывания сети необходимы сетевые платы и коммутаторы, поддерживающие эту технологию.
        + RoCE — это сетевой протокол, поддерживающий передачу данных RDMA по сетям Ethernet. На более низком уровне используется сетевой заголовок Ethernet, а на более высоком (включая данные) — заголовок IB. RoCE позволяет использовать технологию RDMA в стандартной инфраструктуре Ethernet на основе коммутаторов. Сетевые платы при этом должны поддерживать протокол RoCE. RoCEv1 — это протокол RDMA, реализованный на уровне канала передачи Ethernet. Для поддержания надежности передачи на физическом уровне коммутаторы должны поддерживать технологии управления потоком, такие как PFС (Priority-based flow control — управление потоком на основе приоритетов). Версия RoCEv2 реализована на уровне протокола UDP в стеке Ethernet TCP/IP.
        + Протокол iWARP обеспечивает передачу данных RDMA по сетям TCP. iWARP не поддерживает функции, характерные для IB и RoCE. Однако iWARP также позволяет использовать RDMA в стандартной инфраструктуре Ethernet на основе коммутаторов. Соответствующие сетевые платы должны поддерживать протокол iWARP (если используется функция разгрузки ЦП). В противном случае все функции iWARP реализуются с помощью программного обеспечения, но при этом теряется бóльшая часть преимуществ производительности, предоставляемых RDMA.

#### 2.3.5.2 Протокол Infiniband (IB)

Технология IB специально разработана для серверных соединений и широко используется для установки связи между серверами (например, для выполнения репликации и распределенной работы), между серверами и СХД (например, сетями SAN и системами DAS), а также между серверами и сетями (такими как LAN, WAN и Интернет).

Стандарт IB определяет набор устройств для системной связи, включая адаптеры каналов, коммутаторы и маршрутизаторы, используемые для подключения к другим устройствам, такие как канальные адаптеры хоста (HCA) и канальные адаптеры периферийных устройств (TCA). Особенности протокола IB:

* + - * + Протокол на основе стандартов: развитием и стандартизацией технологий Infiniband занимается основанная в 1999 году ассоциация InfiniBand Trade Association, в которую входит 225 компаний. В состав ассоциации входят компании Agilent, Dell, HP, IBM, InfiniSwitch, Intel, Mellanox, Network Appliance и Sun Microsystems. Более 100 других компаний оказали содействие в разработке и продвижении стандарта IB.
        + Скорость: IB обеспечивает высокую скорость передачи данных.
        + Память: серверы с поддержкой InfiniBand используют HCA-адаптеры для преобразования протоколов IB и поддержки передачи данных по внутренней шине PCI-X или PCI-Xpress сервера. HCA-адаптер поддерживает RDMA — это иногда называют обходом ядра. Механизм RDMA особенно полезен при использовании в кластерах. Он использует решение для виртуальной адресации, позволяющее серверу определять и использовать ресурсы памяти других серверов без необходимости вмешательства со стороны ядра операционной системы.
        + Разгрузка процессора: RDMA позволяет перенести нагрузку по маршрутизации пакетов данных с ОС на уровень чипа, снижая тем самым рабочую нагрузку на процессор. Для обработки данных на скорости 10 Гбит/с в ОС требуется процессор 80 ГГц.

Система IB включает канальные адаптеры, коммутаторы, маршрутизаторы, ретрансляторы и каналы связи. К канальным адаптерам относятся HCA и TCA.

* + - * + HCA-адаптер используется для подключения процессора хост-системы к сети IB.
        + TCA-адаптер служит для подключения адаптеров ввода/вывода к сети IB.

IB в сфере СХД: внешняя сеть IB используется для обмена данными с клиентами. Передача данных осуществляется на основе протокола IPoIB. Внутренняя сеть IB используется для обмена данными между узлами устройства хранения. Модуль RPC (Remote procedure call — удаленный вызов процедур) обеспечивает синхронизацию данных между узлами.

Архитектура IB включает несколько уровней: уровень приложений, транспортный уровень, сетевой уровень, уровень каналов и физический уровень. Ниже приведено описание функций каждого уровня.

* + - * + Транспортный уровень: отвечает за упорядоченное распределение и разделение пакетов, мультиплексирование каналов и передачу данных. На этом уровне также осуществляются отправка, получение и восстановление сегментов пакетов данных.
        + Сетевой уровень: поддерживает механизм маршрутизации пакетов между узлами сети. У каждого пакета маршрутизации конечных узлов, участвующих в обмене, есть заголовок глобальной маршрутизации (Global routing header, GRH) и 128-битный адрес IPv6. Сетевой уровень также включает стандартный глобальный 64-битный идентификатор, уникальный для всех подсетей. Технология адресации с использованием этих идентификаторов позволяет передавать данные через множество подсетей.
        + Уровень каналов: на этом уровне поддерживаются функции проектирования пакетов, установки соединения «точка-точка» и коммутации пакетов в локальной подсистеме. На уровне пакетной передачи пакеты делятся на два специализированных типа — пакеты передачи данных и пакеты управления сетью. Пакет управления сетью поддерживает функции управления операциями, индикации подсетей и обеспечения отказоустойчивости для целей перечисления устройств. Пакеты передачи данных используются для собственно передачи данных. Максимальный размер каждого такого пакета составляет 4 КБ. В каждой подсети определенного устройства за адресацию и обмен пакетами отвечает менеджер локальной подсети с 16-битным идентификатором.
        + Физический уровень: на этом уровне используются кабельные соединения со скоростями 1X, 4X и 12X. Скорость передачи данных для них составляет 2,5 Гбит/с, 10 Гбит/с и 30 Гбит/с соответственно. Таким образом, архитектура IB поддерживает объединение соединений для увеличения их скорости вплоть до 30 Гбит/с. Благодаря использованию последовательной передачи в полнодуплексном режиме для создания одной базовой двунаправленной шины требуется всего 4 провода. Соответственно, для шины 12X требуется всего 48 проводов.

### 2.3.6 CIFS, NFS и NDMP

#### 2.3.6.1 Протокол CIFS

В 1996 году компания Microsoft переименовала протокол SMB в CIFS, добавив в него множество новых функций. В настоящее время существуют версии протокола SMB1, SMB2 и SMB3.0.

CIFS предъявляет высокие требования к надежности передачи по сети, вследствие чего обычно использует протоколы TCP/IP. CIFS в основном используется для подключения к Интернету, а также хост-системами Windows для доступа к файлам или другим ресурсам через Интернет. CIFS позволяет клиентам под управлением Windows определять общие ресурсы и осуществлять к ним доступ. С помощью CIFS клиенты могут быстро считывать и записывать данные, а также создавать файлы в СХД, как на локальном компьютере. CIFS поддерживает высокую скорость доступа и быстрый отклик системы даже при одновременном обращении множества пользователей к одному общему файлу.

#### 2.3.6.2 Протокол NFS

NFS (Network File System — сетевая файловая система) — это разработанный IETF протокол совместного доступа к файлам в сети, который нашел широкое применение в средах Linux/Unix.

NFS работает по принципу «клиент-сервер» и использует удаленный вызов процедур (RPC) для обеспечения связи между компьютерами. NFS позволяет пользователям сохранять и редактировать файлы в сетевом хранилище NAS так же, как на локальных компьютерах. При этом клиент NFS должен подключиться к серверу NFS. NFS используется для независимой передачи, поэтому для транспорта использует протоколы TCP или UDP. С помощью NFS пользователи или системные администраторы могут подключать (монтировать) файловые системы или их части (части любой структуры каталогов). Управление доступом к подключенной файловой системе осуществляется на основе прав доступа, например права доступа только для чтения или чтения и записи.

Различия версий протокола NFSv3 и NFSv4:

* + - * + NFSv4 поддерживает отслеживание состояния. В нем реализована функция блокировки файлов и доступна возможность получить корневой узел файловой системы без использования протоколов NLM и MOUNT. NFSv3 — это протокол без сохранения информации о состоянии. В рамках NFSv3 для блокировки файлов потребуется использовать протокол NLM.
        + NFSv4 отличается повышенной степенью безопасности и поддерживает аутентификацию личности RPCSEC-GSS.
        + NFSv4 включает только два типа запросов — NULL и COMPOUND. Все операции интегрированы в тип COMPOUND. Для повышения гибкости клиент может включать несколько операций в один запрос COMPOUND в соответствии с фактическими потребностями.
        + Пространство команд в файловой системе NFSv4 претерпело изменения. На сервере должна быть установлена корневая файловая система (fsid = 0), а прочие файловые системы монтируются в корневую файловую систему для экспорта.
        + NFSv4 также отличается расширенными кроссплатформенными функциями по сравнению с NFSv3.

#### 2.3.6.3 Протокол NDMP

В традиционном хранилище NAS резервное копирование организовано следующим образом:

* + - * + Устройство NAS — это закрытая система хранения. Агент клиента программного обеспечения для резервного копирования можно установить только в производственной системе, а не на устройстве NAS. В рамках традиционного процесса сетевого резервного копирования данные считываются с устройства NAS с использованием протокола совместного использования данных CIFS или NFS, а затем передаются по сети на сервер резервного копирования.
        + Этот механизм занимает ресурсы сети, производственной системы и сервера резервного копирования, что приводит к снижению производительности и не позволяет удовлетворить требования в отношении резервного копирования больших объемов данных.

Протокол NDMP разработан для поддержки системы резервного копирования данных устройств NAS. Он позволяет устройствам NAS передавать данные напрямую на подключенные диски или серверы резервного копирования в сети без необходимости в использовании какого-либо агента клиента резервного копирования.

Протокол NDMP поддерживает два режима работы:

* + - * + В двухстороннем режиме целевое устройство резервного копирования подключается непосредственно к сетевому хранилищу NAS, а не к серверу резервирования. В процессе резервного копирования сервер резервного копирования отправляет соответствующую команду в систему NAS по сети Ethernet. После получения такой команды система копирует соответствующие данные на подключенную ленточную библиотеку.

В рамках двухстороннего режима резервного копирования NDMP потоки данных передаются напрямую на целевое устройство, что позволяет существенно повысить эффективность передачи и снизить нагрузку на сервер. Однако для этого ленточная библиотека (целевое устройство) должна быть напрямую подключена к устройству хранения NAS, а значит, резервное копирование данных будет поддерживаться только для этого устройства NAS.

Ленточные библиотеки дорого стоят. Чтобы обеспечить совместное использование ленточных хранилищ различными устройствами хранения NAS, NDMP также поддерживает трехсторонний режим резервного копирования.

* + - * + В рамках трехстороннего режима резервного копирования система хранения NAS сначала передает подлежащие копированию данные на подключенное к ленточной библиотеке устройство хранения NAS через выделенную сеть резервного копирования. После получения данных устройство хранения выполняет их резервное копирование на ленточную библиотеку.

## 2.4 Архитектура системы хранения данных

### 2.4.1 Развитие архитектуры систем хранения данных

На ранних этапах своего развития системы хранения данных перешли от использования одного контроллера к СХД с двумя контроллерами и взаимным резервированием, которые обрабатывали отдельные задачи до перехода на параллельную обработку данных. Затем поддержка параллельной симметричной обработки была реализована и для систем с несколькими контроллерами. Благодаря развитию облачных вычислений и технологий обработки больших данных начали широко распространяться распределенные хранилища.

В настоящее время СХД с одним контроллером встречаются редко. В большинстве СХД начального и среднего уровня используется архитектура с двумя контроллерами, а в большинстве СХД для критически важных задач — несколько контроллеров.

СХД с одним контроллером:

* + - * Внешний дисковый массив с RAID-контроллерами: с помощью технологий виртуализации и контроллеров RAID физические накопители в дисковом массиве объединяются в логические диски, после чего массив подключается к интерфейсу SCS в хосте через внешний интерфейс SCSI.
      * В СХД с одним контроллером контроллер становится единой точкой отказа.

СХД с двумя контроллерами:

* + - * В настоящее время архитектура с двумя контроллерами в основном используется в популярных СХД начального и среднего уровня.
      * Они работают в одном из двух режимов: «активный-резервный» или «активный-активный».
        + Активный-резервный

Этот режим также называется режимом высокой доступности (HA). В каждый момент времени основные функции выполняет только один контроллер, пока второй находится в режиме ожидания, синхронизируя данные и отслеживая работу служб. В случае отказа активного контроллера резервный контроллер принимает на себя его функции. Кроме того, при переключении на резерв активный контроллер отключается или перезагружается, чтобы избежать эффекта разделения вычислительных мощностей (split brain). Активный контроллер освобождает используемые шины, после чего внешние и внутренние шины переходят под управление резервного контроллера.

* + - * + Активный-активный

В этом режиме два контроллера работают одновременно. При этом каждый подключен ко всем внутренним шинам, однако управление каждой шиной может осуществлять только один контроллер. Каждый контроллер управляет половиной всех внутренних шин. В случае отказа одного из контроллеров второй принимает на себя управление оставшимися шинами. Этот режим отличается более высокой эффективностью, чем режим «активный-резервный».

Развитие архитектуры СХД среднего уровня:

* + - * В системах хранения среднего уровня всегда используется независимая архитектура с двумя контроллерами. При этом контроллеры обычно представляют собой модульное оборудование.
      * Развитие СХД среднего уровня в основном связано с увеличением скорости интерфейсов хоста и дисковых интерфейсов, а также увеличением числа портов.
      * Одна из главных тенденций в этом сегменте — конвергенция SAN и NAS.

СХД с несколькими контроллерами:

* + - * В большинстве СХД, предназначенных для критически важных задач, используется архитектура на основе нескольких контроллеров.
      * Используются следующие основные модели архитектуры:
        + Шинная архитектура
        + Иерархическая звездообразная архитектура
        + Архитектура с прямым подключением
        + Виртуальная матричная архитектура

Развитие архитектуры СХД для критически важных задач:

* + - * В 1990 году компания EMC представила систему хранения Symmetrix на основе шинной архитектуры. В этой системе параллельная шина, соединявшая модули внешних интерфейсов, кэш-памяти и внутренних дисковых интерфейсов, обеспечивала обмен данными и сигналами в режиме временного мультиплексирования.
      * В 2000 году компания HDS анонсировала продукты Lightning 9900 на основе коммутируемой архитектуры. Модули внешних интерфейсов, кэш-памяти и внутренних дисковых интерфейсов подключались к двум коммутируемым сетям с резервированием, что позволило в десятки раз увеличить количество каналов связи по сравнению с шинной архитектурой. Внутренняя шина перестала быть узким местом производительности.
      * В 2003 году компания EMC выпустила системы серии DMX на основе полносвязной архитектуры. Все модули подключались в режиме «точка-точка», что в теории увеличивало пропускную способность внутри системы, но усложняло ее и ограничивало возможности масштабирования.
      * В 2009 году с целью снижения затрат на разработку оборудования компания EMC представила распределенную коммутируемую архитектуру, предполагающую подключение отдельного модуля коммутации к системам хранения среднего уровня с двумя контроллерами и высокой степенью взаимосвязанности. Это позволило сбалансировать затраты и возможности масштабирования.
      * В 2012 году компания Huawei выпустила серию СХД OceanStor 18000 для критически важных задач, также основанную на распределенной коммутируемой архитектуре.

Развитие программных технологий хранения:

Системы хранения на основе ненадежных и низкопроизводительных дисков способны обеспечить высокую степень надежности и производительности хранения за счет эффективного управления. Системы хранения поддерживают совместное использование данных, простые инструменты управления и удобные функции защиты данных. Программное обеспечение СХД прошло путь развития от поддержки базовых технологий RAID и кэширования до реализации продвинутых функций защиты данных, таких как моментальные снимки и репликация, динамического управления ресурсами с повышенной эффективностью управления данными, а также дедупликации и высокоэффективного многоуровневого хранения.

Распределенная архитектура хранилища:

* + - * В распределенной системе хранения данных локальные жесткие диски и твердотельные накопители серверов общего назначения объединяются в крупномасштабный пул ресурсов хранения, после чего данные распределяются по нескольким серверам хранения.
      * В своих распределенных СХД компания Huawei перенимает опыт Google, создавая распределенную файловую систему для нескольких серверов, а затем развертывая службы хранения в этой файловой системе.
      * В качестве большинства узлов хранения выступают серверы общего назначения. СХД OceanStor 100D от Huawei совместима со множеством серверов общего назначения x86 и серверами Arm.
        + Протокол: уровень протокола хранения. Система обеспечивает локальный доступ через подключение к службам блочного, объектного и файлового хранения, а также хранилищу на базе файловой системы Hadoop (HDFS) с использованием протоколов iSCSI или VSC, S3/Swift, NFS и HDFS соответственно.
        + VBS (служба виртуальных блоков): уровень доступа к блокам FusionStorage. Пользовательские данные передаются на уровень VBS с использованием протоколов iSCSI или SCSI.
        + EDS-B: предоставляет службы блочного хранения с поддержкой функций корпоративного уровня, а также получает и обрабатывает данные с уровня VBS.
        + EDS-F: предоставляет службы хранения на основе файловой системы HDFS.
        + Контроллер метаданных (MDC): устройство управления метаданными отслеживает состояние узлов распределенного кластера, правила распределения и восстановления данных.
        + Устройство объектного хранения (OSD): устройство для хранения пользовательских данных в распределенных кластерах устройств объектного хранения.
        + Диспетчер кластера (CM): управляет информацией о кластере.

### 2.4.2 Способы расширения систем хранения данных

Объемы служебных данных продолжают расти по мере развития корпоративных информационных систем и постоянного расширения масштабов предоставления услуг. Зачастую исходной конфигурации СХД оказывается недостаточно для удовлетворения растущих потребностей. Расширение емкости системы хранения стало серьезной проблемой для системных администраторов. Существует два метода ее решения: вертикальное и горизонтальное масштабирование.

Вертикальное масштабирование:

* + - * Традиционная схема вертикального масштабирования предполагает постоянное добавление устройств хранения в существующие СХД для удовлетворения новых потребностей.
      * Преимущество: простота на начальном этапе.
      * Недостаток: по мере увеличения масштаба системы хранения дополнительное расширение ее емкости затрудняется.

Горизонтальное масштабирование:

* + - * Схема горизонтального масштабирования предполагает добавление контроллеров для удовлетворения новых требований.
      * Преимущество: по мере увеличения масштаба сокращается цена за единицу и повышается эффективность.
      * Недостаток: возрастает сложность программного обеспечения и процессов управления.

Рассмотрим в качестве примера дисковую полку SAS от Huawei.

* + - * Согласованность портов: в цепи для соединения двух дисковых полок используется порт EXP полки верхнего уровня и порт PRI полки нижнего уровня.
      * Двухуровневая сеть: модуль расширения A подключается к контроллеру A, а модуль расширения B — к контроллеру B.
      * Симметричная сеть: одна дисковая полка подключается к симметричным портам и слотам контроллеров A и B.
      * Прямое и обратное подключение: модуль расширения A использует прямое подключение, а модуль расширения B — обратное.
      * Глубина каскадирования: число дисковых полок в цепи с каскадным подключением не может превышать заданного ограничения.

Рассмотрим в качестве примера умную дисковую полку Huawei.

* + - * Согласованность портов: в цепи для соединения двух дисковых полок используется порт EXP (P1) полки верхнего уровня и порт PRI (P0) полки нижнего уровня.
      * Двухуровневая сеть: плата расширения A подключается к контроллеру A, а плата расширения B — к контроллеру B.
      * Симметричная сеть: одна дисковая полка подключается к симметричным портам и слотам контроллеров A и B.
      * Прямое подключение: оба модуля расширения A и B используют прямое подключение.
      * Глубина каскадирования: число дисковых полок в цепи с каскадным подключением не может превышать заданного ограничения.

В СХД Huawei OceanStor V3 и V5 начального и среднего уровня, Huawei OceanStor V5 на базе Kunpeng и Huawei OceanStor Dorado V6 используется архитектура горизонтального масштабирования на основе протокола IP. Эта технология объединяет стек протоколов TCP/IP, механизм RDMA и протокол iWARP для поддержки переключения служб между контроллерами, что соответствует тенденции перевода сетей ЦОД на протоколы IP.

В СХД Huawei OceanStor 18000 V3 и V5, а также в Huawei OceanStor Dorado V3 используется архитектура горизонтального масштабирования на основе шины PCIe. Эта технология использует каналы PCIe и механизм RDMA для поддержки переключения служб между контроллерами.

Горизонтальное масштабирование на основе PCIe характеризуется высокой пропускной способностью и низкими значениями задержки.

В рамках горизонтального масштабирования на основе протокола IP используются стандартные технологии (например, протоколы ETH, TCP/IP и iWARP) и инфраструктура центров обработки данных. Использование этой архитектуры позволяет Huawei быстрее разрабатывать собственные чипы для продуктов начального и среднего уровня.

Рассмотрим операции чтения и записи, выполняемые хостом. Существуют следующие сценарии:

* + - * Локальная запись
        + Хост передает запрос на запись в подсистему 0.
        + Подсистема 0 записывает данные в локальный кэш, выполняет зеркалирование и возвращает хосту сообщение об успешной записи данных.
        + Подсистема 0 сбрасывает грязные данные на диск. Если целевой диск находится на локальном компьютере, подсистема 0 выполняет запись напрямую.
        + Если целевой диск находится на удаленном устройстве, подсистема 0 передает запрос в подсистему (например, подсистему 1), которой соответствует целевой диск.
        + Подсистема 1 записывает грязные данные на диски.
      * Нелокальная запись
        + Хост передает запрос на запись в подсистему 2.
        + После того, как станет ясно, что запрашиваемый LUN относится к подсистеме 0, подсистема 2 пересылает запрос на запись в подсистему 0.
        + Подсистема 0 записывает данные в локальный кэш, выполняет зеркалирование и возвращает в подсистему 2 сообщение об успешной записи данных.
        + Подсистема 2 возвращает хосту сообщение об успешной записи данных.
        + Подсистема 0 сбрасывает грязные данные на диск. Если целевой диск находится на локальном компьютере, подсистема 0 выполняет запись напрямую.
        + Если целевой диск находится на удаленном устройстве, подсистема 0 передает запрос в подсистему (например, подсистему 1), которой соответствует целевой диск.
        + Подсистема 1 записывает грязные данные на диски.
      * Локальное чтение
        + Хост передает запрос на чтение в подсистему 0.
        + Если запрашиваемые данные доступны в кэше подсистемы 0, подсистема возвращает их хосту.
        + Если запрашиваемых данных нет в кэше подсистемы 0, подсистема считывает данные с диска. Если целевой диск находится на локальном компьютере, подсистема 0 считывает данные с диска напрямую.
        + После записи считанных данных в локальный кэш подсистема 0 возвращает данные хосту.
        + Если целевой диск находится на удаленном устройстве, подсистема 0 передает запрос в подсистему (например, подсистему 1), которой соответствует целевой диск.
        + Подсистема 1 считывает данные с диска.
        + Подсистема 1 выполняет чтение данных.
        + Подсистема 1 возвращает данные в подсистему 0, после чего подсистема 0 возвращает их хосту.
      * Нелокальное чтение
        + Хост направляет запрос на чтение в подсистему 2, но запрашиваемый LUN не относится к этой подсистеме.
        + После того, как станет ясно, что запрашиваемый LUN относится к подсистеме 0, подсистема 2 пересылает запрос на чтение в подсистему 0.
        + Если запрашиваемые данные доступны в кэше подсистемы 0, подсистема возвращает их в подсистему 2.
        + Подсистема 2 возвращает данные хосту.
        + Если запрашиваемых данных нет в кэше подсистемы 0, подсистема считывает данные с диска. Если целевой диск находится на локальном компьютере, подсистема 0 считывает данные с диска напрямую.
        + После записи считанных данных в локальный кэш подсистема 0 возвращает их в подсистему 2, которая затем передает их хосту.
        + Если целевой диск находится на удаленном устройстве, подсистема 0 передает запрос в подсистему (например, подсистему 1), которой соответствует целевой диск.
        + Подсистема 1 считывает данные с диска.
        + Подсистема 1 выполняет чтение данных.
        + Подсистема 1 возвращает данные в подсистему 0. Подсистема 0 пересылает данные в подсистему 2, которая затем возвращает их хосту.

### 2.4.3 Архитектура продуктов для хранения данных Huawei

По умолчанию в продуктах для хранения данных Huawei начального и среднего уровня используется архитектура с двумя контроллерами. В СХД Huawei для критически важных задач используется архитектура с несколькими контроллерами. Архитектура SmartMatrix, используемая в СХД OceanStor Dorado V6, сочетает в себе преимущества архитектур горизонтального и вертикального масштабирования. Одна система может поддерживать до 32 контроллеров, что значительно повышает ее надежность. Архитектура предотвращает любые прерывания обслуживания даже при выходе из строя семи из восьми контроллеров, обеспечивая доступность на уровне 99,9999 %. Это идеальный выбор для поддержки ключевых сервисных приложений предприятий финансовой, производственной и транспортной отраслей.

SmartMatrix — это новая революция в области архитектуры СХД для критически важных задач; обеспечиваются разделение вычислительных ресурсов и ресурсов хранения. Контроллерные полки полностью изолированы от дисковых полок и подключаются к ним напрямую. Главное преимущество — возможность масштабирования и обновления контроллеров и устройств хранения независимо друг от друга, что значительно повышает гибкость СХД, обеспечивает защиту инвестиций клиентов в долгосрочной перспективе, снижает связанные с хранением риски и гарантирует непрерывность обслуживания.

* + - * Полное взаимодействие внешних интерфейсов
        + СХД Dorado 8000 и 18000 V6 поддерживают соединительные модули ввода/вывода на клиентской стороне (FIM), к которым могут одновременно обращаться четыре контроллера в одном корпусе.
        + После получения запросов ввода/вывода от хоста модуль FIM напрямую распределяет операции между соответствующими контроллерами.
      * Полное взаимодействие контроллеров
        + Контроллеры в контроллерной полке подключаются через каналы RDMA 100 Гбит/с (40 Гбит/с в случае СХД Dorado 3000 V6) объединительной платы.
        + Для целей горизонтального масштабирования до нескольких контроллерных полок любые два контроллера можно подключить напрямую, чтобы избежать пересылки данных.
      * Полное взаимодействие внутренних интерфейсов
        + Dorado 8000 и 18000 V6 поддерживают соединительные модули ввода/вывода на серверной стороне (BIM), которые позволяют подключать умные дисковые полки к двум контроллерным полкам и поддерживают одновременное обращение к дисковой полке с восьми контроллеров. Эта технология в сочетании с функцией непрерывного зеркалирования позволяет системе выдерживать одновременный отказ 7 из 8 контроллеров.
        + СХД Dorado 3000, 5000 и 6000 V6 не поддерживают модули BIM. К дисковым полкам, подключенным к СХД Dorado 3000, 5000 и 6000 V6, получить доступ может только одна контроллерная полка. Непрерывное зеркалирование также не поддерживается.

Система хранения поддерживает три типа дисковых полок: SAS, smart SAS и smart NVMe. В настоящее время их нельзя использовать одновременно в одной системе хранения. Дисковые полки Smart SAS и Smart NVMe поддерживают единый сетевой режим. В этом режиме для подключения к дисковой полке контроллерная полка использует общий интерфейсный модуль RDMA 100 Гбит/с с двумя портами. Каждый интерфейсный модуль подключается к четырем контроллерам в корпусе по шине PCIe 3.0 x16. Таким образом, к каждой дисковой полке могут одновременно обращаться все четыре контроллера, что обеспечивает полное взаимодействие между дисковой полкой и этими контроллерами. Умные дисковые полки оснащены двумя группами восходящих портов и поддерживают одновременное подключение к двум контроллерным полкам. Это позволяет двум контроллерным полкам (то есть восьми контроллерам) одновременно обращаться к дисковой полке, обеспечивая полное взаимодействие между дисковой полкой и контроллерами. Полное взаимодействие между дисковыми полками и восемью контроллерами позволяет системе использовать функцию непрерывного зеркалирования и выдерживать отказ 7 из 8 контроллеров без прерывания обслуживания.

СХД Huawei предоставляют возможность совместного использования глобальных ресурсов:

* + - * Симметричная архитектура
        + Все продукты поддерживают доступ к хосту в режиме «активный-активный». Запросы равномерно распределяются между внешними каналами.
        + LUN больше не принадлежат отдельным контроллерам, что упрощает их использование и обеспечивает балансировку нагрузок. Для этой цели LUN делятся на несколько сегментов (slices), которые затем равномерно распределяются по контроллерам с использованием алгоритма DHT.
        + В продуктах для критически важных задач интеллектуальные модули FIM разделяют LUN на сегменты для запросов ввода/вывода хостов и направляют запросы на целевой контроллер, что позволяет сократить задержки.
      * Общий порт
        + Четыре контроллера в контроллерной полке используют один порт совместно.
        + Балансировка нагрузок обеспечивается без необходимости в программном обеспечении хоста для передачи данных по нескольким каналам.
      * Глобальный кэш
        + При получении запроса ввода/вывода система напрямую записывает соответствующие данные (в одном или двух сегментах) в кэш соответствующего контроллера и отправляет хосту подтверждение.
        + В процессах предварительной выборки и поиска в кэше любых данных и метаданных LUN участвует интеллектуальный кэш чтения всех контроллеров.

В модулях FIM СХД Huawei OceanStor Dorado 8000 и 18000 V6 используется разработанный Huawei чип Hi1822 для подключения ко всем контроллерам в контроллерной полке через четыре внутренних канала, при этом каждый внешний порт предоставляет канал связи для хоста. В случае перезагрузки любого контроллера во время обновления переключение служб на другой контроллер выполняется без влияния на хост-систему и без прерывания связи. Хост-система не получает данных о неисправности контроллера. Переключение занимает всего 1 секунду.

Особенности модулей FIM:

* + - * Отказ контроллера не приводит к отключению внешнего канала связи, и хост не получает данных о сбое.
      * Канал PCIe между модулем FIM и контроллером разрывается, и FIM определяет сбой контроллера.
      * Затем выполняется переключение служб между контроллерами, а модуль FIM перераспределяет запросы хоста по другим контроллерам.
      * Переключение занимает около 1 секунды, что намного лучше, чем при использовании программного обеспечения для передачи данных по нескольким каналам (10–30 секунд).

В режиме глобального кэширования данные хоста записываются непосредственно в журналы с линейным адресным пространством, а эти журналы напрямую копируют данные хоста в память нескольких контроллеров с использованием механизма RDMA на основе предварительно заданной политики копирования. Глобальный кэш включает два типа памяти:

* + - * Глобальная память: память всех контроллеров (четыре контроллера на рисунке). Управление этим типом памяти осуществляется посредством единого адресного пространства памяти. Этот тип предоставляет линейное адресное пространство для верхнего уровня на основе политики резервирования.
      * WAL (Write-Ahead Logging — упреждающая запись в журнал): новый кэш записи типа «журнал».

Глобальный пул поддерживает RAID 2.0+, запись новых данных во все полосы и общие группы RAID для нескольких полос.

Еще одна функция — совместное использование ресурсов на серверной стороне. Она включает совместное использование модулей внутренних интерфейсов внутри корпуса и совместное использование дисковых полок контроллерами.

Архитектура «активный-активный» с полной балансировкой нагрузки:

* + - * Равномерное распределение LUN без привязки
        + Данные в рамках LUN делятся на сегменты размером 64 МБ. Такие сегменты распределяются по разным виртуальным узлам в соответствии с результатом работы алгоритма хеширования (ID LUN + LBA).
      * Балансировка внешней нагрузки
        + Технология UltraPath выбирает подходящие физические каналы для отправки каждого сегмента данных на соответствующий виртуальный узел.
        + Соединительные модули ввода/вывода на клиентской стороне перенаправляют сегменты на соответствующие виртуальные узлы.
        + Внешний интерфейс: если не используются технология UltraPath или модуль FIM, команды ввода/вывода на соответствующие виртуальные узлы перенаправляют контроллеры.
      * Балансировка нагрузки на глобальный кэш записи
        + Выполняется балансировка объема данных.
        + Выполняется балансировка горячих областей данных.
      * Балансировка нагрузки глобального пула хранения
        + Выполняется балансировка использования дисков.
        + Выполняется балансировка степени износа и срока службы дисков.
        + Выполняется равномерное распределение данных.
        + Выполняется балансировка данных в горячих областях.
      * Три копии кэша
        + Система поддерживает две или три копии кэша записи.
        + Для использования трех копий потребуется дополнительная лицензия.
        + Три копии поддерживают только СХД для критически важных задач.
      * При использовании трех копий допускается одновременный отказ двух контроллеров.
        + Выход из строя двух контроллеров не приводит к потере данных или прерыванию обслуживания.
      * При использовании трех копий допускается отказ одной контроллерной полки.
        + При использовании трех копий выполняется зеркалирование данных как внутри контроллерной полки, так и между такими полками.
        + Выход из строя контроллерной полки не приводит к потере данных или прерыванию обслуживания.

Ключевые технологии обеспечения надежности в продуктах Huawei для хранения данных:

* + - * Непрерывное зеркалирование
        + СХД для критически важных задач Dorado V6 поддерживают непрерывное зеркалирование. В случае отказа контроллера система автоматически выбирает новые контроллеры для зеркалирования.
        + Непрерывное зеркалирование охватывает все взаимосвязанные внутренние устройства.
      * Полное взаимодействие внутренних интерфейсов
        + Контроллеры подключаются к дисковым полкам напрямую.
        + СХД для критически важных задач Dorado V6 поддерживают полное взаимодействие внутренних интерфейсов.
        + Модули BIM + две группы восходящих портов на дисковых полках обеспечивают полное взаимодействие полок с восемью контроллерами.
      * Непрерывное зеркалирование и полное взаимодействие внутренних интерфейсов позволяют системе выдерживать отказ семи из восьми контроллеров.

Переключение обслуживания хоста при отказе одного контроллера: при использовании модулей FIM отказ контроллера не приводит к разрыву внешних соединений с хостами, и хосты не получат данных об отказе контроллера. Это позволяет обеспечить высокую доступность системы. При сбое контроллера чип порта модуля FIM определяет разрыв канала PCIe между FIM и контроллером. Затем выполняется переключение служб между контроллерами, а модуль FIM перераспределяет запросы хоста по другим контроллерам. Этот процесс занимает всего около секунды и не влияет на обслуживание хоста. Для сравнения, при использовании интерфейсных модулей, которые не поддерживают общий доступ, в случае сбоя контроллера за переключение канала отвечает программное обеспечение хоста для передачи данных по нескольким каналам. При этом переключение занимает гораздо больше времени (от 10 до 30 секунд), что снижает надежность системы.

## 2.5 Архитектура сетей хранения

### 2.5.1 DAS

СХД с прямым подключением (DAS) обеспечивает подключение одного или нескольких устройств хранения к серверам. Такие устройства хранения поддерживают доступ серверов к данным на уровне блоков. В зависимости от расположения устройств хранения и серверов СХД DAS делятся на внутренние и внешние. В СХД DAS для подключения хостов и устройств хранения используются кабели SCSI.

Технология JBOD (Just a Bunch Of Disks — просто набор дисков) объединяет несколько физических дисков в логический для увеличения емкости, но не предоставляет функций защиты данных. JBOD позволяет решить проблему расширения емкости СХД, связанную с ограниченным количеством дисковых слотов во внутреннем хранилище. Однако эта технология не поддерживает резервирование, а значит, не обеспечивает достаточную надежность.

Контроллер интеллектуального дискового массива предоставляет функции RAID и кэш-память большой емкости, а также позволяет массиву выполнять различные функций. Кроме того, контроллеры оснащены специализированным программным обеспечением для управления.

### 2.5.2 NAS

Предприятиям требуется хранить большие объемы данных и обмениваться ими по сети. Поэтому отличным решением для них будет использование сетевых хранилищ NAS. СХД NAS обеспечивает подключение устройств хранения к рабочей сети и предоставляет службы обработки и передачи данных, а также файловые службы.

Сервер NAS — это внешнее устройство, которое можно гибко подключить к сети. Кроме того, NAS предоставляет возможность совместного использования файлов, а не блоков, что упрощает для клиентов сетевой доступ к NAS. Пользователи UNIX и Microsoft Windows могут беспрепятственно обмениваться данными посредством NAS или протокола передачи файлов FTP. Для работы с NAS в UNIX применяется протокол NFS, а в Windows — CIFS.

Характеристики NAS:

* + - * NAS предоставляет ресурсы хранения с доступом к данным на уровне файлов, что позволяет пользователям с легкостью обмениваться файлами при минимальных затратах на управление хранилищем.
      * NAS — это популярное решение для хранения файлов и обмена ими, не требующее использования нескольких файловых серверов.
      * NAS также помогает устранить узкие места в отношении доступа пользователей к серверам общего назначения.
      * Для архивирования и хранения данных в NAS используются сетевые протоколы и протоколы обмена файлами. К ним относятся протоколы TCP/IP для передачи данных, а также CIFS и NFS для предоставления удаленных файловых служб.

Серверы общего назначения можно использовать для поддержки любого приложения и запуска операционной системы общего назначения. В отличие от таких серверов, системы NAS предназначены для файловых служб и предоставляют функции обмена файлами другим операционным системам с использованием открытых протоколов. Устройства NAS оптимизированы на основе серверов общего назначения в таких аспектах, как файловые службы, хранение и поиск данных. Для повышения доступности устройств NAS некоторые поставщики также поддерживают функцию кластеризации NAS.

Компоненты устройства NAS:

* + - * Подсистема NAS (ЦП и память)
      * Одна или несколько сетевых плат, поддерживающих сетевые подключения, например платы GE и 10GE.
      * Операционная система, оптимизированная для управления функциями NAS.
      * Протоколы NFS и CIFS.
      * Дисковые ресурсы на базе стандартных протоколов хранения, таких как ATA, SCSI и Fibre Channel.

В рамках NAS используются такие протоколы, как NFS, CIFS, FTP, HTTP и NDMP.

* + - * NFS — это стандартный протокол обмена файлами в среде UNIX. Это протокол, работающий без сохранения информации о состоянии. При сбое соединения NFS можно восстановить автоматически.
      * CIFS — это стандартный протокол обмена файлами в среде Microsoft. Это протокол с отслеживанием состояния, основанный на протоколе Server Message Block (SMB). При сбое соединения CIFS нельзя восстановить автоматически. Протокол CIFS по умолчанию поддерживается операционной системой и не требует дополнительного программного обеспечения. Более того, при передаче CIFS отправляет небольшой объем избыточных данных, поэтому эффективность передачи для него выше, чем у NFS.
      * FTP — один из протоколов в стеке TCP/IP. Протокол построен на архитектуре «клиент-сервер». FTP-сервер используется для хранения файлов. С помощью FTP-клиента пользователи получают доступ к ресурсам FTP-сервера через FTP.
      * HTTP (Протокол передачи гипертекста) — прикладной протокол, используемый для передачи гипермедийных документов (например, HTML-документов). Он предназначен для обеспечения связи между веб-браузером и веб-сервером, но также может использоваться для других целей.
      * NDMP (Протокол управления сетевыми данными) — это открытый стандарт резервного копирования данных устройств NAS. NDMP позволяет напрямую записывать данные на ленточные накопители без участия серверов резервного копирования, что повышает скорость и эффективность резервирования данных в системах NAS.

Принципы работы протокола NFS: как и прочие протоколы обмена файлами, NFS основан на архитектуре «клиент-сервер». Однако NFS предоставляет только базовую функцию обработки файлов и не поддерживает функций передачи данных TCP/IP. Функцию передачи данных TCP/IP можно реализовать только с использованием протокола удаленного вызова процедур (RPC). Файловые системы NFS полностью прозрачны для клиентов. Доступ к файлам или каталогам в файловой системе NFS осуществляется так же, как доступ к локальным файлам или каталогам.

Одна программа может с помощью RPC запросить обслуживание у программы на другом компьютере через сеть без необходимости в понимании базовых сетевых протоколов. RPC предполагает наличие протокола передачи, такого как TCP или UDP, для передачи данных сообщения между взаимодействующими программами. В сетевой модели OSI RPC охватывает транспортный и прикладной уровни. RPC упрощает разработку приложений.

В основе RPC лежит архитектура «клиент/сервер». Инициатором запроса является клиент, а поставщиком услуг — сервер. Клиент отправляет запрос на установку соединения с параметрами на сервер RPC и ожидает ответа. На стороне сервера процесс ожидает поступления запроса в спящем режиме. С запросом сервер получает параметры процесса, проводит вычисления и отправляет ответ клиенту. Затем сервер входит в режим ожидания следующего запроса. Клиент получает ответ и результаты вызова.

Зачастую серверы NFS используются облачными платформами в качестве внутреннего общего хранилища. Клиент NFS оптимизируется для облачных вычислений с целью повышения производительности и надежности. Программное обеспечение облачной виртуализации (например, VMware) оптимизирует клиент NFS таким образом, чтобы пространство хранения виртуальной машины можно было создать в общем пространстве сервера NFS.

Принципы работы CIFS: CIFS осуществляет передачу поверх TCP/IP и позволяет компьютерам под управлением Windows получать доступ к файлам на компьютерах UNIX по сети.

Протокол CIFS поддерживает обмен файлами. Два типичных сценария его применения:

* + - * Служба обмена файлами
        + CIFS часто используется для реализации служб обмена файлами, например в корпоративных файлообменниках.
      * Виртуальные машины Hyper-V
        + SMB можно использовать для совместного использования копий виртуальных машин Hyper-V от Microsoft. В этом сценарии для обеспечения непрерывности обслуживания при отказе узла и поддержания надежности виртуальных машин требуется функция аварийного переключения SMB 3.0.

### 2.5.3 SAN

#### 2.5.3.1 Технологии сетей SAN на основе протокола IP

Сетевые платы + программное обеспечение инициаторов: хост-системы, такие как серверы и рабочие станции, используют стандартные сетевые платы для подключения к коммутаторам Ethernet. Устройства хранения iSCSI также подключаются к коммутаторам Ethernet или к сетевым платам хост-систем. Программное обеспечение инициатора, установленное в хост-системе, с помощью технологий виртуализации представляет сетевые платы в виде плат iSCSI. Платы iSCSI принимают и отправляют пакеты данных iSCSI, поддерживая передачу данных через iSCSI и TCP/IP между хостами и устройствами iSCSI. В этом режиме используются стандартные сетевые платы и коммутаторы Ethernet, а значит, нет необходимости в приобретении каких-либо специализированных адаптеров. По этой причине данный режим является наиболее экономичным. Однако в этом режиме для преобразования пакетов iSCSI в пакеты TCP/IP требуется использование ресурсов хоста, что увеличивает расходы на работу хоста и снижает производительность системы. Режим «Сетевые платы + программное обеспечение инициаторов» подходит для сценариев, в которых не требуются высокие показатели IOPS и высокая пропускная способность для доступа к данным.

Сетевая плата с технологией TOE + программное обеспечение инициатора: сетевая плата с технологией TOE (TCP Offload Engine) обрабатывает функции уровня протокола TCP/IP, а хост — функции уровня протокола iSCSI. За счет этого использование сетевых плат TOE позволяет значительно повысить скорость передачи данных. По сравнению с полностью программным режимом, данный режим снижает расходы на работу хоста и требует минимальных затрат на построение сети. Это компромиссное решение.

HBA-адаптер iSCSI:

* + - * + Хосты оснащаются HBA-адаптерами iSCSI для обеспечения эффективного обмена данными с коммутаторами и устройствами хранения. HBA-адаптер хоста обрабатывает функции как уровня протокола iSCSI, так и стека протоколов TCP/IP, что позволяет минимизировать нагрузку на ЦП. Этот режим характеризуется наивысшей производительностью передачи данных, но требует больших затрат.
        + Сети iSCSI перенимают некоторые характеристики сетей SCSI. Во взаимодействии по протоколу iSCSI участвует инициатор, который отправляет запросы ввода/вывода, и целевое устройство, отвечающее на эти запросы и выполняющее соответствующие операции. После установки соединения между инициатором и целевым устройством последнее принимает на себя управление процессом в качестве основного устройства. Целевыми устройствами iSCSI обычно выступают дисковые массивы или ленточные библиотеки iSCSI.
        + Протокол iSCSI определяет набор методов именования и адресации инициаторов и целевых устройств iSCSI. Все узлы iSCSI определяются уникальными именами iSCSI. Данный принцип позволяет четко разграничить имена iSCSI и имена хостов.
        + Для идентификации инициаторов и целевых устройств в системе iSCSI используются полные имена iSCSI (iSCSI Qualified Name, IQN). При перемещении инициаторов или целевых устройств их адреса меняются, но имена остаются неизменными. При установке соединения инициатор отправляет запрос. После того как целевое устройство получает запрос, оно проверяет, согласуется ли имя iSCSI, содержащееся в запросе, с именем, связанным с целевым устройством. Если имена iSCSI совпадают, соединение устанавливается. Каждый узел iSCSI имеет уникальное имя iSCSI. Одно имя iSCSI может использоваться в соединениях одного инициатора с несколькими целевыми устройствами. Несколько имен iSCSI могут использоваться в соединениях одного целевого устройства с несколькими инициаторами.

На основе объединенных портов, портов VLAN или портов Ethernet создаются логические порты. Логические порты — это виртуальные порты, поддерживающие обслуживание хоста. Для обеспечения предоставления услуг каждому логическому порту назначается уникальный IP-адрес.

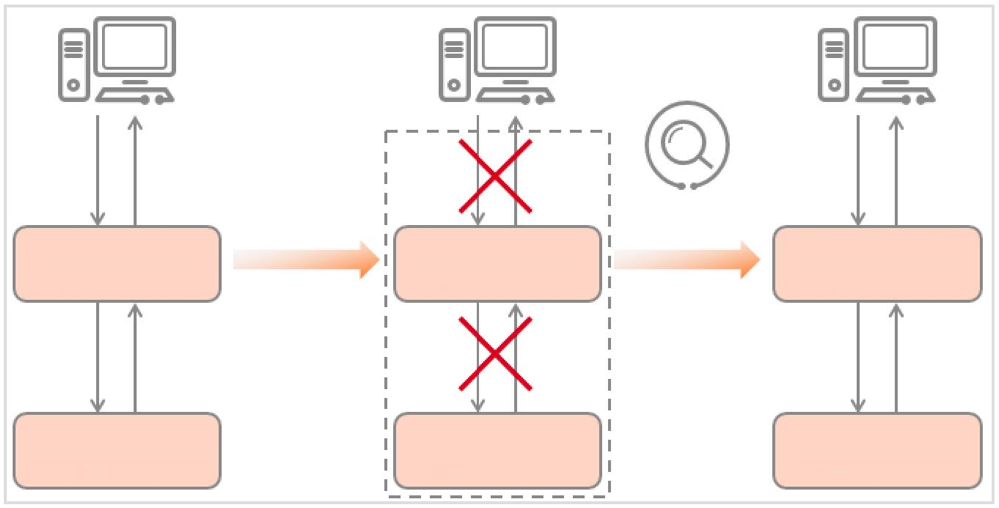
* + - * + Объединенный порт: для повышения надежности доступа к файловым системам и увеличения пропускной способности несколько портов Ethernet одного интерфейсного модуля можно объединить в один.
        + VLAN: виртуальные локальные сети (VLAN) обеспечивают логическое разделение физических портов Ethernet или объединенных портов системы хранения по нескольким широковещательным доменам. При отправке или приеме служебные данные в VLAN получают соответствующий ID VLAN, что позволяет изолировать сети и службы VLAN для повышения безопасности и надежности передачи служебных данных.
        + Порт Ethernet: физические порты Ethernet интерфейсного модуля СХД. Объединенные порты, VLAN и логические порты создаются на основе портов Ethernet.

Аварийное переключение IP-адреса: логический IP-адрес переназначается с неисправного порта на доступный. Таким образом службы переключаются с неисправного порта на доступный без прерывания. После восстановления исходный порт вновь принимает на себя обслуживание. Переключение можно выполнять в автоматическом или ручном режиме. Аварийное переключение IP-адреса поддерживается в сетях IP SAN и сетевых хранилищах NAS.

При аварийном переключении IP-адреса службы переключаются с неисправного порта на доступный, что позволяет обеспечить непрерывность обслуживания и повышает надежность доступа к файловым системам. Этот процесс проходит незаметно для пользователей.

Суть аварийного переключения IP-адресов — переключение служб между портами Ethernet, объединенным портами или портами VLAN.

* + - * + Аварийное переключение IP-адреса для портов Ethernet: для повышения надежности доступа к файловым системам на основе портов Ethernet можно создать логические порты.



Логический порт A

Логический порт A (IP-адрес a)

Ethernet-порт B

Ethernet-порт A

Ethernet-порт A

Логический порт A (IP-адрес a)

Переключение порта

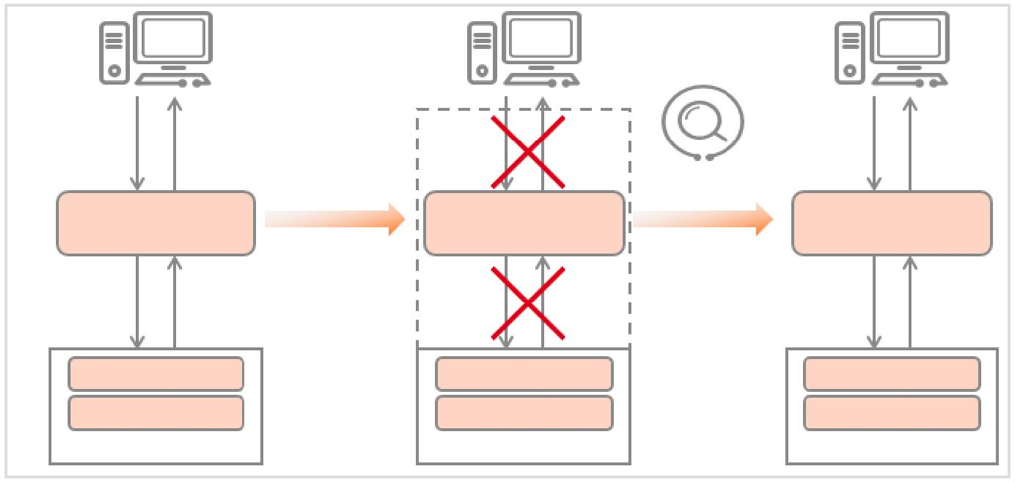
Поиск

Сбой порта

##### Рис. 2-10

Хост получает доступ к службам через логический порт A Ethernet-порта A. Соответствующий IP-адрес — «a». Ethernet-порт A выходит из строя и поэтому не может поддержать обслуживание. После активации аварийного переключения IP-адреса система хранения автоматически определит доступный Ethernet-порт B, очистит конфигурацию логического порта A, соответствующую Ethernet-порту A, а затем создаст и настроит логический порт A на Ethernet-порте B. Это позволит быстро переключить службы на логический порт A Ethernet-порта B. Этот процесс проходит незаметно для пользователей.

* + - * + Аварийное переключение IP-адреса для объединенных портов: для повышения надежности доступа к файловым системам несколько портов Ethernet можно объединить в один. В случае сбоя одного из портов Ethernet в рамках объединенного порта службы продолжат работать через объединенный порт. Аварийное переключение IP-адреса выполняется только при отказе всех портов Ethernet в рамках объединенного порта.



Высоко-скоростная передача данных

Ethernet-порт A

Ethernet-порт B

Объединенный порт A

Ethernet-порт A

Ethernet-порт B

Объединенный порт A

Ethernet-порт C

Ethernet-порт D

Объединенный порт A

Переключение порта

Логический порт A (IP-адрес a)

Поиск

Сбой порта

Логический порт A

Логический порт A (IP-адрес a)

##### Рис. 2-11

Несколько портов Ethernet формируют объединенный порт A. Логический порт A создан на основе объединенного порта A и поддерживает высокоскоростную передачу данных. Если оба порта Ethernet (A и B) выйдут из строя, система хранения автоматически определит доступный объединенный порт B, удалит логический порт A и создаст логический порт A на объединенном порте B. Это позволит переключить службы с объединенного порта A на объединенный порт B. После восстановления Ethernet-портов A и B службы будут переключены обратно на объединенный порт A, если используется функция восстановления конфигурации. Переключение служб выполняется быстро и проходит незаметно для пользователей.

* + - * + Аварийное переключение IP-адреса на основе VLAN: для изоляции различных служб можно создать сети VLAN.

Для поддержки аварийного переключения IP-адресов на основе VLAN необходимо создать сети VLAN, назначить уникальные ID внутри каждой такой сети, а затем изолировать различные службы с помощью VLAN. При сбое порта Ethernet в сети VLAN система хранения автоматически определит доступный порт Ethernet с тем же ID VLAN и переключит службы на такой порт. После восстановления исходный порт вновь принимает на себя обслуживание.

Простые имена VLAN, такие как VLAN A и VLAN B, генерируются автоматически при создании сети VLAN. Фактические имена VLAN зависят от версии системы хранения.

Порты Ethernet и соответствующие порты коммутатора делятся по нескольким VLAN и получают уникальные ID. Сети VLAN используются для изоляции различных служб. Сеть VLAN A создается на основе Ethernet-порта A, который получает ID VLAN «1». Логический порт A, созданный на основе сети VLAN A, используется для изоляции служб. При сбое Ethernet-порта A система хранения автоматически определит сеть VLAN B и порт с ID VLAN «1», удалит логический порт A и создаст логический порт A в сети VLAN B. Это позволит переключить службы на порт в сети VLAN B. После восстановления Ethernet-порта A службы будут вновь переключены на порт в сети VLAN A, если используется функция восстановления конфигурации.

Порт Ethernet может относиться к нескольким VLAN. При сбое такого порта Ethernet откажут все соответствующие сети VLAN. Службы потребуется переключить на порты других доступных VLAN. Переключение служб выполняется быстро и проходит незаметно для пользователей.

#### 2.5.3.2 Сети SAN на основе протокола FC

HBA-адаптер FC: HBA-адаптер FC преобразует пакеты SCSI в пакеты Fibre Channel без создания дополнительной нагрузки на хост-систему.

Основные принципы работы сетей Fibre Channel:

* + - * + Маршрутизация Fibre Channel (Fibre Channel Routing, FCR) обеспечивает подключение устройств, принадлежащих разным фабрикам, без необходимости в их объединении. В отличие от каскадирования через порты расширения E\_Port в обычных коммутаторах, после подключения коммутаторов через коммутатор FCR две фабрики не объединяются, сохраняя независимость. Коммутатор канала между двумя фабриками работает в качестве маршрутизатора.
        + Маршрутизатор FC: коммутатор, поддерживающий службу маршрутизации FC-FC.
        + EX\_Port: тип порта, функционирующий аналогично портам E\_Port, но не передающий служебные данные фабрики или информацию о топологии маршрутизации из одной фабрики в другую.
        + Backbone fabric (базовая фабрика): фабрика коммутатора, поддерживающего службу маршрутизатора Fibre Channel.
        + Edge Fabric (периферийная фабрика): фабрика, к которой подключен маршрутизатор Fibre Channel.
        + Inter fabric link (IFL): канал связи между портами E\_Port и EX-Port или VE\_Port и VEX-Port.

Еще одно важное понятие — зонирование. Зона — это набор портов или устройств, которые взаимодействуют друг с другом. Участник зоны имеет доступ только к другим участникам той же зоны. Устройство может принадлежать нескольким зонам. Посредством настройки основных зон можно осуществлять управление правами доступа каждого устройства или порта. Также можно настроить зоны изоляции трафика. При наличии нескольких каналов ISL (портов E\_Port) каждый канал ISL передает только трафик, предназначенный для портов в одной зоне изоляции трафика.

#### 2.5.3.3 Сравнение сетей IP SAN и FC SAN

Сначала вернемся к концепции SAN.

* + - * + Используемые протоколы: Fibre Channel/iSCSI. Эти протоколы используются в двух видах архитектуры SAN соответственно — FC SAN и IP SAN.
        + Прямой доступ к устройству: подходит для традиционных сценариев доступа к базе данных.
        + Файловый доступ зависит от хоста приложения. Для обеспечения общего доступа требуется поддержка кластерного программного обеспечения, однако при обработке конфликтов доступа оно требует больших затрат ресурсов, что приводит к снижению производительности системы. Кроме того, в разнородных средах трудно поддерживать совместное использование ресурсов.
        + Высокая производительность, высокая пропускная способность и низкая задержка, но высокая стоимость и недостаточные возможности масштабирования.

Теперь сравним FC SAN и IP SAN.

* + - * + Для решения проблемы масштабирования СХД DAS устройства хранения можно объединить в сеть FC SAN, которая поддерживает подключения к более чем 100 серверам.
        + Сеть IP SAN разработана для решения проблем FC SAN, связанных со сложностью управления и высокими затратами. Для развертывания IP SAN требуется лишь небольшое количество широко используемого оборудования. Поэтому затраты на сети IP SAN намного ниже, чем при использовании FC SAN. Большинство хостов используют сетевые платы и коммутаторы, которые также подходят (хотя и не идеально) для передачи по протоколу iSCSI. Для создания высокопроизводительной сети SAN на базе протокола IP требуются выделенные HBA-адаптеры iSCSI и высокотехнологичные коммутаторы.

### 2.5.4 Распределенная архитектура

В распределенной системе хранения данных локальные жесткие диски и твердотельные накопители серверов общего назначения объединяются в крупномасштабный пул ресурсов хранения, после чего данные распределяются по нескольким серверам хранения.

Обычно в качестве внутренних сетей распределенного хранилища используются сети 10GE, 25GE и IB. На стороне клиента, как правило, используются соединения GE, 10GE или 25GE.

Ниже описываются плоскости сети и соответствующие функции:

* + - * Сеть управления: взаимодействует с сетью управления клиента для поддержки управления системой и ее обслуживания.
      * Сеть BMC (контроллер управления материнской платой): подключается к портам Mgmt узлов управления или хранения с целью удаленного управления устройствами.
      * Сеть хранения: внутренний уровень, на котором осуществляется передача служебных данных между всеми узлами в системе хранения.
      * Сервисная сеть: взаимодействует с клиентскими приложениями и осуществляет доступ к устройствам хранения посредством стандартных протоколов, таких как iSCSI и HDFS.
      * Сеть репликации: поддерживает синхронизацию и репликацию данных между узлами репликации.
      * Сеть арбитража: взаимодействует с кворум-сервером HyperMetro. Этот уровень рассматривается, только если служба блочного хранения поддерживает функцию HyperMetro.

Ниже описываются ключевые программные компоненты и соответствующие функции:

* + - * FSM: процесс управления распределенным хранилищем Huawei, предоставляющий такие функции эксплуатации и технического обслуживания, как управление аварийными сигналами, мониторинг, управление журналами и настройка системы. Рекомендуется развернуть этот модуль на двух узлах в режиме «активный-резервный».
      * Служба виртуальных блоков (VBS): процесс, поддерживающий службу точки доступа к распределенному хранилищу через интерфейсы SCSI или iSCSI и позволяющий серверам приложений получать доступ к ресурсам распределенного хранилища.
      * Устройство объектного хранения (OSD): компонент распределенного хранилища Huawei, предназначенный для хранения пользовательских данных в распределенных кластерах.
      * REP: сеть репликации данных.
      * Enterprise Data Service (EDS): компонент для обработки запросов ввода/вывода, отправляемых службой VBS.

## 2.6 Знакомство с интеллектуальными продуктами Huawei для хранения данных

### 2.6.1 Система хранения данных на базе флеш-накопителей

#### 2.6.1.1 Серии продуктов для хранения данных на базе флеш-накопителей

Флеш-системы хранения данных OceanStor Dorado от Huawei — это высокопроизводительные системы хранения нового поколения, предназначенные для использования в средних и крупных центрах обработки данных или для поддержки критически важных бизнес-служб. Эти системы ориентированы на основные службы, предоставляемые средними и крупными предприятиями (например, центры обработки данных корпоративного уровня, виртуальные и облачные центры обработки данных). При этом они также отвечают требованиям к высокой производительности, надежности и эффективности, предъявляемым средними и крупными центрами обработки данных. Эффективное передовое оборудование, высокая производительность во всех аспектах за счет использования флеш-технологий и интеллектуальные функции управления системой хранения позволяют OceanStor Dorado предоставлять предприятиям службы хранения высочайшего класса, отвечающие требованиям различных корпоративных приложений, таких как крупные базы данных OLTP/OLAP, облачные вычисления, виртуализация серверов и рабочих столов.

#### 2.6.1.2 Особенности продуктов

В сбалансированной полносвязной архитектуре SmartMatrix 3.0 используется высокопроизводительная полносвязная пассивная объединительная плата на основе матрицы. Она поддерживает подключение множества контроллеров. Интерфейсные модули подключаются к объединительной плате в полносвязном режиме, благодаря чему хосты могут осуществлять доступ к любому контроллеру через любой порт. Архитектура SmartMatrix поддерживает тесное взаимодействие контроллеров и позволяет упросить программную модель системы, обеспечивая точную балансировку нагрузки в режиме «активный-активный», высокую эффективность, низкую задержку и широкие возможности совместной работы.

Интеллектуальный алгоритм FlashLink® обеспечивает высокую скорость параллельной обработки операций ввода/вывода, а также стабильно низкую задержку. Он предоставляет несколько механизмов оптимизации работы флеш-памяти. FlashLink® поддерживает связь ЦП контроллера и ЦП SSD-накопителя для координации работы алгоритмов SSD, что позволяет обеспечить высокую производительность и надежность системы.

В SSD-накопителях в качестве постоянной памяти используется флеш-память NAND. SSD-накопители работают быстрее традиционных жестких дисков, потребляют меньше энергии и характеризуются меньшей задержкой. Кроме того, они меньше, легче и устойчивы к ударам.

Высокая производительность

* + - * + СХД, полностью состоящие из SSD-накопителей, отличаются высоким показателем IOPS и низкой задержкой.
        + Поддержка алгоритма FlashLink®, в том числе технологий интеллектуальной многоядерной оптимизации (intelligent multi-core), повышения эффективности RAID, разделения горячих и холодных данных и обеспечения низкой задержки.

Высокая надежность

* + - * + Защита от отказа компонентов, поддержка двойного резервирования и режима «активный-активный»; полносвязная архитектура SmartMatrix 3.0 для обеспечения высокой эффективности, низкой задержки и широких возможностей совместной работы.
        + Поддержка двойного резервирования, защита от отключения питания и применение коффер-дисков.
        + Передовые технологии защиты данных: HyperSnap (моментальные снимки), HyperReplication (удаленная репликация), HyperClone (клонирование LUN) и HyperMetro.
        + Поддержка технологии виртуализация RAID 2.0+.

Высокая доступность

* + - * + Поддержка горячей замены таких компонентов, как контроллеры, блоки питания, интерфейсные модули и диски.
        + Поддержка технологии disk roaming (переключение соединений дисков), позволяющая системе хранения автоматически определять перемещенные диски и возобновлять их работу.
        + Централизованное управление ресурсами в сторонних системах хранения.

#### 2.6.1.3 Архитектура продукта

Используется следующая архитектура:

* + - * + Аппаратная платформа Pangea V6 Arm.
        + Процессор: Kunpeng 920.
        + Контроллерная полка высотой 2U со встроенными дисками.
        + Поддержка контроллерных полок на 25 дисков 2,5 дюйма и на 36 дисков NVMe размером с ладонь.
        + Два контроллера, поддерживающие работу в режиме «активный-активный».

#### 2.6.1.4 Компоненты

Контроллерная полка имеет модульную конструкцию и включает корпус, контроллеры (со встроенными модулями вентиляторов), модули питания, резервные батарейные блоки и дисковые модули.

#### 2.6.1.5 Программная архитектура

Программное обеспечение системы хранения управляет устройствами хранения и хранимыми данными, а также поддерживает операции с данными, выполняемые серверами приложений.

Пакет ПО систем хранения данных Huawei OceanStor Dorado 3000 V6, Dorado 5000 V6 и Dorado 6000 V6 включает программное обеспечение СХД, программное обеспечение терминала обслуживания и программное обеспечение, выполняемое сервером приложений. ПО этих трех типов взаимодействует между собой для интеллектуального, эффективного и экономичного предоставления различных служб хранения, резервного копирования и аварийного восстановления.

#### 2.6.1.6 ИИ-ускорители

Ключевые технологии FlashLink®:

Интеллектуальная многоядерная оптимизация

В системе хранения используются процессоры собственной разработки Huawei. СХД оснащены большим количеством ЦП и большим количеством ядер ЦП на контроллер, чем любая другая система в отрасли. Технология интеллектуальной многоядерной оптимизации позволяет линейно увеличивать производительность хранилища по мере роста числа процессоров и ядер.

Повышение эффективности RAID

В системе хранения используется технология перенаправления при записи (redirect-on-write, ROW) для обеспечения эффективной записи во все страйпы. Эти технологии позволяют записывать все новые данные в новые блоки без перезаписи существующих. Это позволяет значительно сократить потребление ресурсов ЦП контроллера и нагрузки чтения/записи на SSD-накопители в процессе записи, благодаря чему повышается общая производительность системы на различных уровнях RAID.

Разделение горячих и холодных данных

Система хранения способна выявлять и разделять горячие и холодные данные для повышения производительности сборки мусора, сокращения числа циклов программирования/стирания (P/E) SSD-накопителя и продления срока его службы.

Обеспечение низкой задержки

В системе хранения используются новейшие SSD-накопители, разработанные Huawei, и более эффективный протокол для оптимизации обработки ввода/вывода и поддержания низкого значения задержки ввода/вывода.

Умная дисковая полка

Система хранения поддерживает умные дисковые полки нового поколения, разработанные специалистами Huawei. Умные дисковые полки оснащаются ЦП и памятью, что позволяет им брать на себя такие задачи, как восстановление данных при отказе диска. Это помогает снизить нагрузку на контроллеры и устранить влияние подобных задач на производительность системы.

Эффективная технология временных точек

В системе хранения используется технология защиты данных на основе распределенных временных точек. Операции чтения и записи, запрашиваемые хостами, включают информацию о временной точке, которая помогает быстро найти необходимые метаданные, обеспечивая тем самым повышение производительности доступа.

Технологии выравнивания износа и противодействия износу

Глобальное выравнивание износа: при неравномерном распределении данных по твердотельным накопителям некоторые накопители могут использоваться чаще и, соответственно, будут быстрее изнашиваться. В результате они могут выйти из строя намного раньше ожидаемого срока, что приведет к дополнительным затратам на техническое обслуживание. В системе хранения используется технология глобального выравнивания износа, которая позволяет выровнять степень износа всех твердотельных накопителей и повысить надежность их работы.

Глобальное противодействие износу: если для нескольких твердотельных накопителей степень износа приближается к пороговому значению, система хранения начинает отдавать приоритет при записи только определенным накопителям. Такие SSD-накопители изнашиваются быстрее прочих. Подобный подход позволяет предотвратить одновременный отказ нескольких твердотельных накопителей.

#### 2.6.1.7 Стандартный сценарий применения — ускорение работы критически важных служб

Часто задаваемые вопросы:

В условиях быстрого развития мобильного Интернета эффективный анализ больших объемов пользовательских данных и обработка быстро растущих объемов данных о транзакциях зависят от эффективного сбора, анализа, консолидации и извлечения данных. Эти процессы имеют ключевое значение для реализации стратегий, ориентированных на данные. Существующие ИТ-системы необходимо модернизировать. Например, на обработку данных и интеграцию хранилищ данных в системы выставления счетов и учета банков и крупных предприятий уходит несколько часов. Это не позволяет своевременно получить доступ к таким службам, как анализ операций или запрос обслуживания.

Решение:

Высокопроизводительные флеш-системы хранения Huawei позволяют решить вышеописанные проблемы. Передовые флеш-системы хранения поддерживают нескольких основных приложений (таких как база данных системы обработки транзакций). СХД Huawei обеспечивают в два раза более быструю обработку данных, сокращение задержки ответа и повышение эффективности обслуживания в несколько раз.

### 2.6.2 Гибридные флеш-системы хранения данных

#### 2.6.2.1 Серия гибридных флеш-систем хранения данных

Гибридное флеш-хранилище Huawei OceanStor сочетает передовую аппаратную структуру и интегрированную архитектуру, объединяющую блочные и файловые службы хранения с продвинутыми технологиями использования и защиты данных. Это позволяет удовлетворить требования средних и крупных предприятий к высокой производительности, масштабируемости, надежности и доступности хранилищ.

Конвергентное хранилище:

* + - * + Конвергенция технологий хранения SAN и NAS.
        + Поддержка сетевых протоколов хранения, таких как iSCSI, Fibre Channel, NFS, CIFS и FTP.

Высокая производительность:

* + - * + Высокопроизводительный процессор, высокоскоростная кэш-память большой емкости и различные высокоскоростные интерфейсные модули позволяют добиться беспрецедентной производительности СХД.
        + Поддержка ускорения SSD-накопителей, значительно повышающая производительность хранилища.

Гибкие возможности масштабирования:

* + - * + Поддержка различных типов дисков.
        + Поддержка различных интерфейсных модулей.
        + Поддержка различных технологий масштабирования.

Высокая надежность:

* + - * + Полносвязная архитектура SmartMatrix, поддержка резервирования всех компонентов, режима работы «активный-активный», а также технологии RAID 2.0+.
        + Несколько технологий защиты данных, в том числе защита от сбоев питания, предварительное копирование данных, коффер-диски и восстановление дефектных секторов.

Высокая доступность:

* + - * + Несколько передовых технологий защиты данных, в том числе создание моментальных снимков, копирование LUN, удаленная репликация, клонирование, зеркалирование томов и работа в режиме «активный-активный», а также поддержка протокола NDMP.

Интеллектуальные возможности и повышение эффективности:

* + - * + Различные функции для обеспечения точного контроля и управления, в том числе SmartTier, SmartQoS и SmartThin.
        + Приложение DeviceManager поддерживает работу в графическом интерфейсе.
        + Платформа eService предоставляет интеллектуальные возможности управления эксплуатацией и техническим обслуживанием с поддержкой самообслуживания.

#### 2.6.2.2 Архитектура продукта

Система хранения включает контроллерные и дисковые полки. Она предоставляет клиентам интеллектуальную платформу хранения, которая отличается высокой надежностью, производительностью и емкостью.

В зависимости от модели можно использовать разные типы контроллерных и дисковых полок.

#### 2.6.2.3 Конвергенция SAN и NAS

Конвергенция технологий хранения SAN и NAS: единая система хранения одновременно поддерживает службы SAN и NAS, предоставляя ресурсы хранения в их общее пользование. Хосты могут получить доступ к любому LUN или любой файловой системе через любой внешний порт любого контроллера. На протяжении жизненного цикла данных горячие данные постепенно становятся холодными. Длительное хранение холодных данных в кэше или на твердотельном накопителе приводит к неэффективному использованию ресурсов и может повлиять на производительность системы хранения в долгосрочной перспективе. В системе хранения Huawei используется интеллектуальная технология многоуровневого хранения для гибкого выделения памяти в фоновом режиме.

Интеллектуальную технологию многоуровневого хранения следует развертывать на устройствах с различными типами носителей. Она отслеживает данные в режиме реального времени. Данные, к которым в течение длительного времени не осуществляется доступ, помечаются как «холодные» и постепенно передаются с высокопроизводительных накопителей на более медленные. Это позволяет избежать увеличения времени ответа устройств на запрос обслуживания. Если холодные данные вновь потребуется часто использовать, их можно быстро переместить на высокопроизводительный накопитель для обеспечения стабильного уровня производительности системы.

Политики миграции можно активировать вручную или автоматически.

#### 2.6.2.4 Поддержка разнообразных сценариев обслуживания

Гибридная флеш-система хранения Huawei OceanStor объединяет в себе службы хранения SAN и NAS и поддерживает несколько протоколов хранения. Эти системы подходят для использования в широком диапазоне сценариев общего назначения, включая, помимо прочего, предприятия государственного и финансового секторов, телекоммуникационной отрасли, производственной сферы, а также сценарии, требующие функций резервного копирования и аварийного восстановления.

#### 2.6.2.5 Сценарий применения — центры обработки данных, работающие по схеме «активный-активный»

Распределение нагрузки между контроллерами

RPO (recovery point objective — допустимая потеря данных) = 0 и RTO (recovery time objective — допустимое время восстановления данных) ≈ 0 для критически важных сервисов.

Конвергенция SAN и NAS: службы SAN и NAS в режиме «активный-активный» можно развернуть на одном устройстве. Поддерживается локальное переключение в случае отказа одного контроллера.

Решение обеспечивает бесперебойное предоставление услуг для клиентов.

Решение, работающее в режиме «активный-активный», подходит для таких отраслей, как здравоохранение, финансы и социальное обеспечение.

### 2.6.3 Распределенное хранилище

#### 2.6.3.1 Распределенное хранилище

Развитие облачных вычислений и искусственного интеллекта привело к экспоненциальному росту данных. Новые приложения, такие как высокоскоростная связь 5G, видео высокой четкости (HD) 4K/8K, автономное вождение и анализ больших данных, повышают требования к хранению данных.

Huawei OceanStor 100D — это распределенное хранилище с возможностью горизонтального масштабирования для удовлетворения как текущих, так и будущих потребностей бизнеса. Оно предоставляет гибкие сервисы по требованию на базе облачной инфраструктуры и способно поддерживать как критически важные, так и развивающиеся рабочие нагрузки.

#### 2.6.3.2 Особенности продуктов

Для использования блочного хранилища предоставляются стандартные интерфейсы SCSI и iSCSI. Это идеальная платформа хранения для частных облаков, контейнеров, платформ виртуализации и приложений баз данных.

Хранилище HDFS — это решение для обработки больших данных с разделенными модулями вычисления и хранения и поддержкой протокола HDFS. Используемая в нем интеллектуальная система многоуровневого хранения позволяет снизить совокупную стоимость владения и обеспечить стабильное качество взаимодействия с пользователями.

Объектное хранилище поддерживает взаимодействие с основными облачными экосистемами посредством стандартных API-интерфейсов для хранения контента, резервного копирования и архивирования с поддержкой облака, а также использования общедоступных облачных хранилищ.

Файловое хранилище предоставляет пользователям общие службы файлового хранения с использованием протокола NFS.

#### 2.6.3.3 Примеры аппаратных узлов

Масштабируемое файловое хранилище Huawei OceanStor 9000 — это распределенная платформа файлового хранения, предназначенная для больших данных. Huawei OceanStor 9000 поддерживает следующие функции:

Разработанная Huawei распределенная файловая система OceanStor DFS обеспечивает хранение больших объемов неструктурированных данных и использование объединенного пространства имен.

Поддерживает взаимодействие с платформами FusionInsight Hadoop и Cloudera Hadoop на основе открытых компонентов Hadoop, позволяя пользователям с легкостью создавать платформы анализа больших данных корпоративного уровня.

Оптимизирует протокол NFS. На NFS-клиенте можно настроить несколько сетевых портов и установить плагин оптимизации протокола NFS — DFSClient. Он позволяет активировать поддержку одновременного подключения и оптимизации кэширования для нескольких сетевых портов, что значительно повышает производительность клиента.

#### 2.6.3.4 Программная архитектура системы

Уровень интерфейсов хранилища: предоставляет приложениям стандартные интерфейсы для доступа к системе хранения и поддерживает протоколы SCSI, iSCSI, Hadoop и объектные протоколы.

Уровень служб хранения: предоставляет службы блочного и объектного хранения, хранения на базе HDFS, а также расширенные функции корпоративного уровня.

Уровень подсистемы хранения: использует интерфейс Plog и технологию перенаправления при записи (redirect-on-write, ROW) с конкатенацией для поддержки функций многократного копирования, EC, восстановления и балансировки данных, управления дисками и чтения/записи данных.

Управление хранилищем: управляет использованием системы и обслуживает ее, а также предоставляет такие функции, как установка системы, развертывание, настройка служб, управление устройствами, создание отчетов об аварийных сигналах, мониторинг, обновление и расширение.

#### 2.6.3.5 Сценарий применения — пул облачных ресурсов

В качестве примера рассмотрим СХД Huawei OceanStor 100D. Она применяется в сценариях частного облака и виртуализации.

Huawei OceanStor 100D позволяет использовать сверхвысокое количество пулов ресурсов хранения данных, поддерживает выделение ресурсов по запросу и гибкое расширение емкости хранилища в средах виртуализации и частных облаках. Эти технологии повышают эффективность развертывания, масштабирования, эксплуатации и технического обслуживания системы хранения при использовании серверов общего назначения. К стандартным сценариям применения этой СХД относятся облака доступа к каналам сферы Интернет-финансов, облака разработки и тестирования, облачные сервисы, пулы облачных ресурсов B2B в доменах спецификаций операторов связи и облачные сервисы электронного правительства.

Критически важные базы данных

СХД Huawei OceanStor 100D предоставляет службы корпоративного уровня, такие как распределенное хранение данных в режиме «активный-активный» с низкими задержками, для поддержки эффективной и стабильной работы хранилищ данных и критически важных баз данных, включая OLAP (Online Analytical Processing, оперативная аналитическая обработка) и OLTP (Online Transaction Processing — оперативная обработка транзакций).

Аналитика больших данных

СХД OceanStor 100D — это ведущее в отрасли решение для обработки больших данных с разделенными модулями вычисления и хранения, которое объединяет традиционные хранилища данных и создает на их основе единый пул ресурсов больших данных, доступный для предприятий. В ней также реализованы функции корпоративного уровня, такие как технология Elastic Erasure Coding, позволяющая максимизировать используемую долю дискового пространства, возможности развертывания по запросу и расширения изолированных вычислительных ресурсов и ресурсов хранения. Все это призвано повысить эффективность обработки больших данных и снизить совокупную стоимость владения системой. К стандартным сценариям применения этой СХД относится аналитика больших данных в финансовой сфере, операторском бизнесе (хранение журналов) и государственном секторе.

Хранение контента и архивирование резервных копий

СХД OceanStor 100D предоставляет высокопроизводительные и высоконадежные пулы ресурсов объектного хранилища для удовлетворения требований к пропускной способности, частому доступу к данным горячей области, а также к долгосрочному хранению и доступу через Интернет, предъявляемых онлайн-сервисами реального времени, связанными с передачей данных, аудио- и видеоматериалов через Интернет и доступом к корпоративным веб-дискам. К стандартным сценариям применения этой СХД относятся хранение, резервное копирование и архивирование электронных финансовых чеков, аудио- и видеозаписей, медицинских снимков, правительственных и корпоративных электронных документов и данных Интернета транспортных средств (IoV).

### 2.6.4 Решение для периферийного хранения данных (FusionCube)

#### 2.6.4.1 Решение для периферийного хранения данных (FusionCube)

Решение FusionCube от Huawei — это инфраструктурная ИТ-платформа на базе гиперконвергентной архитектуры. FusionCube — предварительно собранное решение, объединяющее вычислительные, сетевые устройства и устройства хранения в готовом к использованию комплексе. Пользователи FusionCube избавлены от необходимости приобретать дополнительные СХД или сетевые устройства. Программно-аппаратный комплекс FusionCube объединяет серверы и устройства хранения. В него по умолчанию интегрирована подсистема распределенного хранилища, платформа виртуализации и программное обеспечение для управления облаком. Комплекс поддерживает планирование ресурсов по требованию и линейное расширение. В основном FusionCube используется в центрах обработки данных, поддерживающих различные типы гибридных рабочих нагрузок, таких как базы данных, виртуальные рабочие столы, контейнеры и виртуализация.

Huawei FusionCube 1000 — это интегрированное решение для развертывания периферийной ИТ-инфраструктуры. Решение поставляется в формате интегрированного шкафа. FusionCube 1000 в основном применяется в периферийных центрах обработки данных и таких сценариях использования периферийных приложений в вертикальных отраслях, как заправочные станции, кампусы, угольные шахты и электростанции. Для обеспечения удаленного централизованного управления комплексами FusionCube 1000, развернутыми в офисах, можно использовать систему FusionCube Center. Комплекс FusionCube 1000 в сочетании с системой FusionCube Center Vision — это интегрированное в шкаф решение, поддерживающее развертывание услуг, управление эксплуатацией и техническим обслуживанием, а также централизованное устранение неполадок. Применение FusionCube 1000 позволяет значительно ускорить развертывание системы и снизить затраты на ее эксплуатацию и техническое обслуживание.

Гиперконвергентная инфраструктура (HCI) — это набор устройств, объединяющий не только вычислительные и сетевые ресурсы, ресурсы хранения и виртуализации серверов, но также такие элементы, как программное обеспечение для резервного копирования, технологии моментальных снимков, дедупликации данных и сжатия «на лету». Несколько наборов устройств можно объединить в сеть для поддержки удобного горизонтального масштабирования по модульному принципу и формирования единого пула ресурсов.

#### 2.6.4.2 Примеры аппаратных моделей

Портфель продуктов:

Блейд-сервер: например, Huawei E9000, объединяющий в себе вычислительные и сетевые ресурсы с хранилищем данных. В корпус E9000 высотой 12U можно установить блейд-серверы Huawei серии E9000, узлы хранения данных и узлы расширения емкости.

Сервер высокой плотности: например, Huawei X6800 и X6000, отличающиеся высокой плотностью узлов. Huawei X6800 — это сервер высотой 4U с четырьмя узлами, а Huawei X6000 — сервер высотой 2U с четырьмя узлами.

Стоечный сервер: например, Huawei RH или TaiShan. TaiShan 2280 — это двухсокетный стоечный сервер высотой 2U. Он отличается высокой производительностью вычислений, хранилищем большой емкости, низким энергопотреблением, простотой управления и развертывания. Сервер ориентирован на работу с Интернетом, предоставление служб распределенного хранилища, облачные вычисления, обработку больших данных и предоставление корпоративных услуг.

#### 2.6.4.3 СХД

Распределенная система хранения — это программный пакет для развертывания распределенного блочного хранилища. Он устанавливается на сервер для виртуализации всех локальных дисков этого сервера и их объединения в пул ресурсов хранения. Это позволяет предоставлять службы блочного хранилища.

#### 2.6.4.4 Программная архитектура

К аппаратному обеспечению решения FusionCube относятся серверы, коммутаторы, твердотельные накопители и источники бесперебойного питания (ИБП).

Программное обеспечение решения FusionCube включает системы FusionCube Builder и FusionCube Center, ПО для аварийного восстановления и резервного копирования, а также программное обеспечение распределенного хранилища.

#### 2.6.4.5 Сценарий применения — периферийный центр обработки данных

Автоматическая настройка: полноценный периферийный центр обработки данных, поставляемый в виде интегрированного шкафа. Не требует настройки на месте и поддерживает функцию автоматической настройки.

Автоматическая эксплуатация: система унифицированного и централизованного управления периферийными центрами обработки данных, не требующая вмешательства специализированного персонала и позволяющая значительно сократить затраты на эксплуатацию и техническое обслуживание.

Взаимодействие облачных и периферийных систем: экосистема корпоративных приложений, развернутая в облаке для быстрой доставки приложений из центрального центра обработки данных в периферийные, что позволяет ускорить разработку новых услуг для клиентов.

|  |
| --- |
| 3 Передовые технологии хранения |
|  |

## 3.1 Технологии и приложения для настройки ресурсов хранения

### 3.1.1 SmartThin (интеллектуальное динамическое выделение емкости)

#### 3.1.1.1 Обзор

Разработанная Huawei система OceanStor SmartThin для серии СХД OceanStor предоставляет функцию автоматического тонкого выделения ресурсов. Она позволяет решить проблемы, связанные с развертыванием традиционных систем хранения.

В отличие от систем, где вся доступная емкость выделяется на раннем этапе, SmartThin выделяет емкость по запросу. Это более эффективно с экономической точки зрения, поскольку позволяет клиентам на ранней стадии использовать лишь несколько дисков и добавлять емкость в соответствии с фактическими требованиями. При этом снижаются как первоначальные расходы на приобретение, так и совокупная стоимость владения.

#### 3.1.1.2 Принципы работы

Система SmartThin виртуализирует ресурсы хранения.

Она осуществляет управление ресурсами по запросу. SmartThin не выделяет всю емкость системы хранения сразу. Система предоставляет виртуальное пространство, объем которого может превышать объем физически доступного пространства хранения. Физическую емкость SmartThin выделяет по требованию пользователей. При необходимости емкость внутренних пулов хранения может быть увеличена. При этом не требуется отключать систему хранения, а процесс расширения емкости полностью прозрачен для пользователей.

SmartThin создает тонкие LUN (логические тома) на основе технологии RAID 2.0+, то есть в одном пуле ресурсов хранения одновременно существуют тонкие и толстые LUN. Тонкий LUN — это логическая единица в пуле хранения. Тонкий LUN можно сопоставить с хостом для предоставления хосту доступа к этому LUN. Емкость тонкого LUN не зависит от выделенной для него физической емкости. Его емкость виртуализирована. Физически доступное пространство пула ресурсов хранения выделяется тонкому LUN только после того, как он начнет обрабатывать запрос ввода/вывода.

Толстый LUN — это логический диск, к которому хост может получить доступ. Емкость толстого LUN указывается при его создании. SmartThin позволяет СХД выделять ресурсы хранения толстым LUN посредством технологии автоматической настройки ресурсов.

#### 3.1.1.3 Чтение/запись

SmartThin использует технологии выделения емкости при записи (capacity-on-write) и перенаправления по времени (direct-on-time) для упрощения обработки отправляемых хостами запросов к тонким LUN на чтение и запись. Технология Capacity-on-write используется для выделения пространства при записи, а direct-on-time —для перенаправления запросов на чтение и запись данных.

Capacity-on-write: после получения запроса на запись от хоста тонкий LUN использует технологию direct-on-time, чтобы проверить, выделено ли для запрашиваемого логического хранилища достаточное физическое пространство хранения. Если это не так, инициируется задача выделения емкости, в рамках которой емкость выделяется минимальными единицами — зернами (64 КБ). После этого данные записываются во вновь выделенное физическое пространство хранения.

Direct-on-time: при использовании функции capacity-on-write соотношение между фактической и логической областями хранения данных не рассчитывается по какой-либо фиксированной формуле, а определяется случайным сопоставлением на основе принципа выделения емкости при записи. Поэтому при обработке операций чтения/записи данных для тонкого LUN соответствующий запрос необходимо перенаправить в фактическую область хранения, которая определяется на основе вышеописанного сопоставления.

Таблица соответствия: таблица с данными о сопоставлении фактической и логической областей хранения. Таблица соответствия динамически обновляется во время записи данных и запрашивается при их чтении.

#### 3.1.1.4 Сценарии применения

SmartThin выделяет пространство хранения по запросу. Система хранения выделяет емкость серверам приложений по мере необходимости в пределах определенного порогового значения, что позволяет максимизировать эффективность использования ресурсов хранения.

SmartThin подходит для следующих сценариев:

SmartThin позволит увеличивать емкость систем банковских транзакций в режиме реального времени и без прерывания обслуживания.

SmartThin позволит по запросу динамически выделять физическое дисковое пространство для служб электронной почты и хранилищ данных постоянной доступности.

SmartThin поможет оптимизировать конфигурацию хранения оператора связи, позволив различным службам, предоставляемым этим оператором, конкурировать за физическое пространство хранения.

#### 3.1.1.5 Настройка

Для использования тонких LUN необходимо импортировать на соответствующее устройство хранения файл лицензии SmartThin и активировать лицензию.

Если после создания тонкого LUN выводится аварийное сообщение о нехватке свободного пространства в пуле хранения, рекомендуется как можно скорее расширить пул хранения.

В противном случае тонкий LUN может войти в режим сквозной записи, что приведет к снижению производительности системы.

### 3.1.2 SmartTier (интеллектуальное многоуровневое хранение данных)

#### 3.1.2.1 Обзор

Функция SmartTier также называется интеллектуальным многоуровневым хранилищем. Она поддерживает интеллектуальное управление ресурсами хранения данных — выполняет анализ действий с данными для автоматического сопоставления данных с наиболее подходящими для их хранения носителями.

SmartTier переносит горячие данные на высокопроизводительные устройства (например, твердотельные накопители), а неиспользуемые данные — на более экономичные накопители (например, диски NL-SAS) с большей емкостью. Это обеспечивает быстрый отклик и высокий показатель IOPS при работе с горячими данными, повышая производительность всей системы хранения.

#### 3.1.2.2 Уровни хранения

Уровень хранения в пуле хранения — это совокупность накопителей с одинаковым уровнем производительности. В зависимости от этого уровня SmartTier выделяет уровни высокой производительности, производительности и емкости. На каждом уровне хранения используются диски одного типа и один уровень RAID.

* + - * 1. Уровень высокой производительности

Тип диска: твердотельные накопители

Характеристики диска: твердотельные накопители отличаются высоким показателем IOPS и быстрым откликом на запросы ввода/вывода. Однако стоимость пространства хранения на один диск в их случае высока.

Характеристики приложения: на этом уровне обычно развертываются приложения, предполагающие интенсивные запросы произвольного доступа.

Характеристики данных: хранение наиболее активных данных (горячих данных).

* + - * 1. Уровень производительности

Тип диска: диски SAS

Характеристики диска: диски этого уровня обеспечивают высокую пропускную способность для обработки рабочих нагрузок. На запросы ввода/вывода они отвечают относительно быстро. Если кэш не содержит данных, запись данных выполняется медленнее, чем чтение.

Характеристики приложения: на этом уровне обычно развертываются приложения, предполагающие запросы доступа средней частоты.

Характеристики данных: хранение горячих данных (активные данные).

* + - * 1. Уровень емкости

Тип диска: диски NL-SAS

Характеристики диска: диски NL-SAS имеют низкий показатель IOPS и медленно реагируют на запросы ввода/вывода. Однако их цена за единицу в отношении обработки запросов на хранение высока.

Характеристики приложения: подходит для приложений, предполагающих небольшой объем запросов доступа.

Характеристики данных: хранение холодных данных (неиспользуемые данные).

Количество уровней хранения определяется количеством типов дисков в пуле хранения.

#### 3.1.2.3 Три этапа внедрения SmartTier

Если пул хранения включает более одного типа дисков, функция SmartTier позволит максимизировать эффективность использования ресурсов хранения. Во время миграции данных пул хранения оценивает уровень активности блоков данных и переносит блоки на наиболее подходящий уровень хранения.

Внедрение SmartTier включает три этапа: мониторинг ввода/вывода, анализ данных и миграцию данных.

За мониторинг ввода/вывода отвечает модуль мониторинга ввода/вывода. Система хранения выявляет горячие и холодные блоки, сравнивая уровни их активности. Уровень активности блока данных рассчитывается на основе частоты доступа к нему и объема операций ввода/вывода.

Модуль анализа данных выполняет анализ данных. Система определяет пороговое число операций ввода/вывода для каждого уровня хранения в пуле хранения на основе емкости каждого такого уровня, показателей каждого блока данных, определенных модулем мониторинга ввода/вывода, и частоты доступа к блокам данных. Модуль анализа данных присваивает приоритет всем блокам (в одном пуле хранения) в соответствии с уровнем их активности, и в первую очередь переносятся самые горячие блоки данных.

Модуль миграции данных выполняет миграцию данных. SmartTier переносит блоки данных в соответствии с уровнем их активности и политикой миграции. Блоки данных с более высоким приоритетом переносятся на более высокие уровни (обычно на уровень высокой производительности или уровень производительности), а блоки данных с меньшим приоритетом — на более низкие уровни (обычно уровень производительности или уровень емкости).

#### 3.1.2.4 Ключевые технологии

Первоначальное выделение емкости: система хранения распределяет новые данные по соответствующим уровням хранения в соответствии с политикой первоначального выделения емкости.

Политика SmartTier определяет направление миграции данных.

Пул хранения может включать несколько LUN с разными политиками SmartTier. Однако мониторинг ввода/вывода выполняется в отношении сразу всего пула хранения. Поэтому рекомендуется активировать функцию мониторинга ввода/вывода, чтобы убедиться, что LUN, созданные в соответствии с политикой автоматической миграции, способны выполнить задачи миграции.

Режим миграции данных определяется в плане миграции. Пользователь может настроить режим миграции данных в соответствии с требованиями к обслуживанию.

Степень детализации миграции данных: SmartTier разделяет данные в пуле хранения на основе степени детализации миграции данных или по блокам данных. Во время миграции данных пул хранения оценивает уровень активности блоков данных и переносит блоки на наиболее подходящий уровень хранения.

Скорость миграции данных позволяет оценить ход выполнения переноса данных на разные уровни хранения.

#### 3.1.2.5 Сценарии применения

SmartTier можно применять в широком спектре сред обслуживания. В этом разделе в качестве примера используется служба базы данных Oracle.

Поскольку холодные данные хранятся на дисках NL-SAS, пространство высокопроизводительных твердотельных накопителей освобождается для размещения горячих данных. Твердотельные накопители обеспечивают быстрый отклик и высокий показатель IOPS при работе с горячими данными. Это позволяет значительно повысить общую производительность системы хранения.

#### 3.1.2.6 Настройка

Процесс развертывания SmartTier в системе хранения включает этапы проверки лицензии, настройки SmartTier на уровне системы хранения, настройки SmartTier на уровне пула хранения и настройки SmartTier на уровне LUN.

Лицензия предоставляет право на использование определенных дополнительных функций. Перед развертыванием SmartTier убедитесь, что файл лицензии содержит соответствующую информацию.

Настройка на уровне системы хранения включает настройку скорости миграции данных, которая распространяется на все пулы хранения в системе хранения.

Настройка на уровне пула хранения включает настройку степени детализации миграции данных, уровня RAID, плана миграции данных, мониторинга ввода/вывода и анализа прогнозов. Эти настройки применяются к соответствующему пулу хранения.

Настройка на уровне LUN включает выбор политики первоначального выделения емкости и политики SmartTier, которые применяются к соответствующему LUN.

Чтобы функция SmartTier могла эффективно работать, требуется дополнительное пространство для обмена данными в процессе динамической миграции данных. Поэтому в пуле хранения, для которого развернута система SmartTier, необходимо зарезервировать дополнительное свободное пространство.

### 3.1.3 SmartQoS (интеллектуальный контроль качества услуг)

#### 3.1.3.1 Обзор

Функция SmartQoS динамически выделяет ресурсы хранения для поддержки определенных показателей производительности отдельных приложений.

По мере развития технологий хранения емкость СХД растет. При этом все больше пользователей стремятся развертывать несколько приложений на одном устройстве хранения. В таких условиях различные приложения конкурируют за пропускную способность системы и ресурсы IOPS, определяющие производительность критически важных приложений.

SmartQoS помогает пользователям надлежащим образом использовать ресурсы системы хранения и поддерживает высокую производительность критически важных служб.

SmartQoS позволяет пользователям задавать для определенных приложений целевые показатели производительности, такие как IOPS или пропускная способность. Система хранения динамически выделяет ресурсы для удовлетворения требований приложений к качеству обслуживания в соответствии с заданными целевыми показателями. Приоритет отдается приложениям с высокими требованиями к качеству обслуживания.

#### 3.1.3.2 Расстановка приоритетов операций ввода/вывода

Технология расстановки приоритетов операций ввода/вывода SmartQoS использует приоритеты LUN, файловых систем или моментальных снимков. Эти приоритеты определяются пользователями в зависимости от важности служб, связанных с этими LUN, файловыми системами или моментальными снимками.

Каждый LUN или файловая система имеет заданный пользователем приоритет, который сохраняется в системе хранения. При поступлении в систему запроса ввода/вывода система хранения назначает ему приоритет в соответствии с приоритетом LUN, файловой системы или моментального снимка, для которых будет обрабатываться этот запрос. Приоритет операции ввода/вывода сохраняется на протяжении всей процедуры ее обработки. В случае нехватки ресурсов система будет в первую очередь обрабатывать высокоприоритетные операции ввода/вывода для повышения производительности высокоприоритетных LUN, файловых систем или моментальных снимков.

Расстановка приоритетов операций ввода/вывода предназначена для планирования выделения ресурсов системы хранения, включая время вычислений ЦП и ресурсы кэш-памяти.

#### 3.1.3.3 Контроль трафика ввода/вывода

Контроль трафика SmartQoS предполагает управление очередями операций ввода/вывода, распределение токенов и управление удалением из очереди контролируемых объектов.

Контроль трафика ввода/вывода реализован через систему на основе токенов. Когда пользователь устанавливает для группы контроля трафика ограничение на количество операций ввода/вывода в секунду, это ограничение преобразуется в количество соответствующих токенов. В системе хранения с ограниченными ресурсами IOPS каждая операция ввода/вывода соответствует токену. Если пропускная способность ограничена, токены распределяются по секторам.

Система хранения регулирует количество токенов в очереди операций ввода/вывода в соответствии с приоритетом LUN, файловой системы или моментального снимка. Чем больше токенов имеет очередь операций ввода/вывода, связанная с LUN, файловой системой или моментальным снимком, тем больше ресурсов система выделяет этой очереди. Система хранения в первую очередь обрабатывает запросы ввода/вывода в очереди, связанной с LUN, файловой системой или моментальным снимком.

Например, если пользователь активирует политику SmartQoS для двух LUN в системе хранения и задаст в ней целевые показатели производительности, система хранения может ограничить выделение ресурсов одному LUN, чтобы поддержать целевой уровень производительности LUN с более высоким приоритетом. SmartQoS рассматривает такие целевые показатели производительности, как пропускная способность, количество операций ввода/вывода в секунду и время задержки.

#### 3.1.3.4 Сценарии применения

SmartQoS динамически выделяет ресурсы хранения для поддержки надлежащего уровня производительности критически важных служб и пользователей с высоким приоритетом.

1. Поддержка производительности критически важных служб

Если службы OLTP и резервного копирования и архивирования одновременно используют одно устройство хранения, для поддержания работы обеих служб требуются достаточные системные ресурсы.

* 1. Служба OLTP имеет критически важное значение и предъявляет высокие требования к производительности в реальном времени.
  2. Служба резервного копирования работает с большим объемом данных, но менее требовательна к задержке.

SmartQoS определяет целевые показатели производительности для разных служб, чтобы гарантировать производительность тех из них, что имеют критически важное значение. Для поддержки производительности критически важных служб можно использовать любой из следующих методов: задайте приоритеты операций ввода/вывода для удовлетворения требований службы OLTP; настройте политики контроля трафика SmartQoS для удовлетворения требований к обслуживанию.

1. Обеспечение качества обслуживания для пользователей с высоким приоритетом

Для снижения затрат некоторые пользователи предпочитают не развертывать собственные выделенные системы хранения. Они используют платформы хранения, предлагаемые соответствующими поставщиками. Это позволяет снизить совокупную стоимость владения (TCO) и поддержать непрерывность обслуживания. На подобных общих платформах хранения службы различных типов и функций конкурируют за ресурсы хранения, вследствие чего важные пользователи могут не получить доступ к требуемым ресурсам.

SmartQoS позволяет создать политики SmartQoS и задать приоритеты ввода/вывода для разных подписчиков. При этом в случае нехватки ресурсов в первую очередь обслуживание будет предоставлено подписчикам с высоким приоритетом для удовлетворения их требований к качеству.

#### 3.1.3.5 Настройка

Перед развертыванием системы SmartQoS ознакомьтесь с руководством по настройке и проверьте файл лицензии. Функция мониторинга служб, предоставляемая системой хранения, позволяет отслеживать характеристики ввода/вывода LUN или файловых систем и создавать на их основе политики SmartQoS. Политики SmartQoS регулируют и контролируют работу приложений с целью обеспечения непрерывности работы критически важных служб.

### 3.1.4 SmartDedupe (интеллектуальная дедупликация «на лету»)

#### 3.1.4.1 Обзор

Функция SmartDedupe позволяет удалить избыточные данные из системы хранения. Эта технология дедупликации сокращает занимаемое данными физическое пространство хранения, высвобождая ресурсы для обслуживания.

СХД OceanStor от Huawei имеют функцию дедупликации «на лету», обеспечивающую дедупликацию новых записываемых в систему данных.

Эта функция удаляет дубликаты данных перед записью на накопители.

Функция дедупликации на основе поиска соответствия анализирует записанные данные и выявляет повторяющиеся и похожие блоки данных на основе отличительных меток (отпечатков), после чего выполняет собственно дедупликацию.

#### 3.1.4.2 Принцип работы дедупликации «на лету»

Размер блока данных дедупликации определяет степень детализации при дедупликации.

Отпечаток — это двоичное значение фиксированной длины, представляющее блок данных. Для расчета отпечатков любых блоков данных в СХД OceanStor Dorado V6 используется слабый алгоритм хеширования. СХД сохраняет данные о соответствии между отпечатками и местами хранения блоков данных в библиотеке отпечатков.

Принцип работы дедупликации «на лету»:

1. В СХД OceanStor Dorado V6 используется слабый алгоритм хеширования для расчета отпечатков любых новых блоков данных, записываемых в систему.
2. СХД сравнивает отпечаток нового блока данных с отпечатками в библиотеке отпечатков.
   * Если выявлено соответствие, СХД выполняет побайтовое сравнение.
   * В противном случае система принимает новые записанные блоки данных как не являющиеся дубликатами, записывает их на диски, после чего заносит информацию о соответствии между отпечатком и местом хранения блока данных в библиотеку отпечатков.
3. СХД выполняет побайтовое сравнение, чтобы проверить, полностью ли новый блок данных соответствует уже существующему.
   * Если это так, система рассматривает новый блок данных как дубликат существующего, удаляет новый блок и связывает его отпечаток с местом хранения исходного блока, занося соответствующую информацию в библиотеку отпечатков.
   * В противном случае СХД записывает новый блок на диск. Обработка в этом случае выполняется так же, как если бы функция дедупликации была отключена.

#### 3.1.4.3 Принцип работы постдедупликации на основе поиска соответствия

Отпечатки блоков данных и информация о месте их хранения заносится в оперативную таблицу (opportunity table) для выявления наиболее активных отпечатков.

Процесс удаления похожих данных включает следующие этапы:

1. Выполняется запись данных и расчет отпечатков, после чего соответствующая информация заносится в оперативную таблицу.

СХД делит новые записываемые данные на блоки, использует соответствующий алгоритм для расчета похожих отпечатков (similar fingerprint, SFP) этих блоков, записывает блоки данных на диск, а затем заносит информацию об отпечатках и местах хранения блоков данных в оперативную таблицу.

1. После записи данных система периодически выполняет дедупликацию на основе поиска соответствия.
   1. СХД периодически проверяет оперативную таблицу на предмет наличия похожих отпечатков.
      * Если они существуют, выполняется побайтовое сравнение блоков данных.

* В противном случае продолжаются периодические проверки.
  1. СХД выполняет побайтовое сравнение, чтобы проверить, являются ли похожие блоки данных полностью идентичными.
     + Если это так, СХД рассматривает новый блок данных как дубликат существующего, удаляет новый блок и связывает его отпечаток с местом хранения исходного блока.
     + Если блоки просто похожи, но не идентичны, система выполняет их разностное сжатие, добавляет отпечатки блоков в библиотеку отпечатков, обновляет отпечатки в метаданных блоков, после чего освобождает занимаемое ими физическое пространство для повторного использования.

#### 3.1.4.4 Сценарии применения

В приложениях инфраструктуры виртуальных рабочих столов (VDI) пользователи создают множество виртуальных образов на физическом устройстве хранения. Эти образы содержат большие объемы повторяющихся данных. По мере увеличения объемов таких данных системе хранения становится все труднее удовлетворять требования к обслуживанию. Функция SmartDedupe помогает удалять повторяющиеся данные образов для высвобождения ресурсов и поддержки хранения больших объемов служебных данных.

#### 3.1.4.5 Настройка

При создании LUN необходимо выбрать тип приложения. Функция дедупликации, относящаяся к этому типу настраивается по умолчанию. С помощью команд пользователи могут просматривать статус дедупликации и сжатия для каждого типа приложения.

### 3.1.5 SmartCompression (интеллектуальное сжатие «на лету»)

#### 3.1.5.1 Обзор

Функция SmartCompression реорганизует данные для экономии места и повышает эффективность передачи, обработки и хранения данных без потери данных. СХД OceanStor Dorado V6 поддерживают режимы сжатия «на лету» и сжатия после обработки.

Сжатие «на лету»: система выполняет дедупликацию и сжатие данных до их записи на диски. Пользовательские данные обрабатываются в реальном времени.

Сжатие после обработки: данные сначала записываются на диски, а затем, во время простоя системы, выполняется их чтение и сжатие.

#### 3.1.5.2 Принципы работы

В СХД используется алгоритм сжатия без потерь LZ77. Этот алгоритм заменяет подлежащие кодированию/декодированию данные ссылкой на данные, которые только что прошли кодирование/декодирование. Сжатие достигается за счет использования в текущих данных ссылок на уже доступную информацию.

В алгоритме LZ77 используется принцип скользящего окна. При сканировании строки данных кодировщик учитывает уже просмотренную часть сообщения длиной N — это и есть скользящее окно. Если для сканируемого фрагмента есть соответствие в скользящем окне, то вместо него будет использоваться ссылка на повторяющийся фрагмент (смещение в окне, длина совпадения).

#### 3.1.5.3 Сценарии применения

По мере увеличения объемов данных, сжимаемых системой хранения, возрастает нагрузка на ЦП.

1. Базы данных: сжатие данных — идеальный выбор для сценариев баз данных. Большинство пользователей с радостью пойдут на небольшое снижение производительности ради освобождения более 65 % емкости хранилища.
2. Файловые службы: сжатие данных также применяется и в файловых хранилищах. Пиковые часы занимают половину общего периода обслуживания, а степень сжатия набора данных в системе достигает 50 %. В таких сценариях функция SmartCompression приводит к незначительному снижению показателя IOPS системы.
3. Хранение технических и сейсмогеологических данных: в этом случае требования аналогичны сценариям резервного копирования баз данных. Данные этого типа хранятся в том же формате, но характеризуются не таким большим объемом повторений. Эти данные можно сжать для экономии дискового пространства.

SmartDedupe + SmartCompression

Для еще большей экономии пространства хранения функции дедупликации и сжатия можно использовать одновременно. Применение SmartDedupe в сочетании со SmartCompression подходит для СХД в системах тестирования или разработки, систем файлового хранения и систем хранения технических данных.

#### 3.1.5.4 Настройка

Процесс развертывания SmartCompression включает проверку лицензии и собственно активацию функции SmartCompression. Проверка лицензии позволяет убедиться в доступности функции SmartCompression.

### 3.1.6 SmartMigration (интеллектуальная миграция)

#### 3.1.6.1 Обзор

SmartMigration — это ключевая технология миграции служб. Служебные данные можно переносить как внутри системы хранения, так и между различными системами.

Слово «согласованный» в данном контексте означает, что после завершения миграции служб целевой LUN содержит полные копии всех соответствующих данных исходного LUN.

#### 3.1.6.2 Принципы работы

SmartMigration синхронизирует и разделяет служебные данные для переноса всех данных с исходного LUN на целевой LUN.

#### 3.1.6.3 Синхронизация служебных данных SmartMigration

Пара: в SmartMigration парой называются исходный и целевой LUN, на который выполняется перенос данных. Пара может включать только по одному исходному и целевому LUN.

В СХД управление функцией SmartMigration осуществляют модули LM.

В процессе миграции служебных данных технология двойной записи (Dual-write) осуществляет редактирование и запись данных одновременно на исходный и целевой LUN.

Изменения в данных на исходном LUN заносятся в журнал (LOG), чтобы определить, выполняется ли одновременная запись в целевой LUN с использованием технологии двойной записи.

В журнал изменений в данных (data change log, DCL) записываются дифференциальные данные, которые не удалось записать в целевой LUN в процессе синхронизации изменений в данных.

Процесс синхронизации служебных данных на исходном и целевом LUN включает первичную синхронизацию и синхронизацию изменений в данных. Эти два режима синхронизации не зависят друг от друга. Их можно выполнять одновременно, чтобы обеспечить синхронизацию изменений в служебных данных хоста с данными на исходном и целевом LUN.

#### 3.1.6.4 Разделение пар SmartMigration

Разделение применяется к отдельным парам. Процесс разделения пар предполагает остановку синхронизации служебных данных на исходном и целевом LUN в одной паре, обмен данными между LUN и последующее удаление отношений миграции данных между рассматриваемыми LUN.

В течение процесса разделения обслуживание хоста приостанавливается. После завершения обмена данными обслуживание возобновляется через целевой LUN. Миграция служб выполняется незаметно для пользователей.

1. Обмен данными между LUN

Прежде чем целевой LUN сможет взять на себя службы исходного LUN, эти LUN должны синхронизироваться, а затем обменяться данными.

1. Разделение пар

Разделение пар: после завершения обмена данными между исходным и целевым LUN отношения миграции данных между этими LUN удаляются.

Функция согласованного разделения пар SmartMigration позволяет одновременно выполнять обмен данными между LUN и удаление отношений миграции данных для нескольких пар. Это позволяет поддержать согласованность данных в любой момент времени — как до, так и после разделения пар.

#### 3.1.6.5 Настройка

Процесс развертывания SmartMigration в системе хранения данных включает проверку файла лицензии, создание задачи SmartMigration и разделение пар SmartMigration.

На целевом LUN копия данных исходного LUN на определенный момент времени будет доступна только после разделения пар. Эту копию можно использовать для восстановления данных исходного LUN в случае его повреждения. Кроме того, копию данных можно использовать, например, при тестировании приложений и анализе данных.

## 3.2. Приложения и технологии для защиты данных СХД

### 3.2.1 HyperSnap (моментальные снимки)

#### 3.2.1.1 Обзор

По мере развития информационных технологий и увеличения объемов глобальных данных люди придают все большее значение безопасности и защите этих данных. В отношении защиты больших объемов данных традиционные решения для резервного копирования данных сталкиваются со следующими проблемами:

Необходимость резервного копирования огромных и все возрастающих объемов данных. При этом окно резервного копирования не изменяется.

Необходимость минимизации влияния процессов резервного копирования на производительность рабочих систем. Требования пользователей в отношении значений RPO (recovery point objective — допустимая потеря данных) и RTO (recovery time objective — допустимое время восстановления данных) продолжают расти.

* + - * + Окно резервного копирования — это интервал, в течение которого резервное копирование данных выполняется без серьезного влияния на приложения.
        + RTO — это политика переключения служб, которая обеспечивает кратчайшее время аварийного восстановления. С помощью этой политики определяется допустимое время восстановления данных и обеспечивается моментальное переключение на систему аварийного восстановления.
        + RPO — это политика переключения служб, которая обеспечивает наименьшую потерю данных. С помощью этой политики определяется допустимая потеря данных и обеспечивается использование наиболее актуальных данных для аварийного восстановления.

Повышение требований стимулирует развитие и модернизацию технологий резервного копирования. Все больше образцов ПО для резервного копирования предоставляют технологию создания моментальных снимков.

Создание моментальных снимков — это технология резервного копирования данных, напоминающая фотосъемку. Эта технология записывает состояния объектов в определенный момент времени. Технология создания моментальных снимков позволяет системным администраторам ликвидировать окно резервного копирования и удовлетворить требования предприятий к непрерывности обслуживания и надежности данных.

Функция HyperSnap позволяет создать согласованную копию исходных данных на определенный момент времени. Моментальный снимок — это доступная для использования копия определенного набора данных. Копия содержит статический образ исходных данных на момент их копирования.

#### 3.2.1.2 Принципы работы

Основной технологией, используемой для создания моментальных снимков, является перенаправления при записи (redirect-on-write, ROW). В случае перезаписи данных для новых данных выделяется новое пространство хранения. После успешного завершения записи исходное пространство очищается.

Организация данных: LUN, созданные в пуле хранения OceanStor Dorado V6, состоят из томов данных и метаданных.

* + - * + В том метаданных записывается информация об организации данных (адрес логического блока (LBA), версия и ID клона) и атрибуты данных. Тома метаданных организованы в виде древовидной структуры. LBA — это адрес логического блока. Версия указывает на время создания моментального снимка, а ID клона — на число копий данных.
        + Том данных содержит данные, записанные в LUN.

В исходном томе хранятся исходные данные, для которых создается моментальный снимок. Такой том виден пользователям как LUN.

Том моментального снимка — это логическая копия данных, создаваемая после создания моментального снимка исходного тома. Такая копия видна пользователям как LUN моментального снимка.

Неактивно: моментальный снимок в неактивном состоянии. В этом состоянии моментальный снимок недоступен. Его использование станет возможным только после активации.

#### 3.2.1.3 Поддержка работы без потерь

Процесс активации моментального снимка предполагает сохранение состояния данных исходного объекта на момент активации. На данном этапе выполняются такие операции, как создание зеркала для исходного объекта снимка и сопоставление такого зеркала с активированным снимком. Принципы создания зеркала для исходного объекта: общие LUN и моментальные снимки во флеш-системе хранения OceanStor Dorado V6 используют режим чтения и записи на основе технологии ROW. В этом режиме каждый раз, когда хост записывает новые данные, система выделяет пространство для их хранения и обновляет в таблице соответствия LUN ссылку на новое пространство. Как показано на слайде, L0–L4 — логические адреса, P0–P8 — физические адреса, а A–I — данные.

#### 3.2.1.4 Откат к моментальному снимку

При необходимости можно использовать функцию отката для немедленного восстановления состояния данных системы хранения на момент создания моментального снимка. Это позволяет компенсировать повреждение или потерю данных, вызванную некорректными операциями или вирусами после создания моментального снимка.

Функция отката копирует данные моментального снимка в целевой объект (исходный LUN или моментальный снимок). После запуска процедуры отката целевой объект мгновенно активирует эти данные (то есть данные моментального снимка). Для обеспечения немедленной доступности целевого объекта сразу после начала процедуры отката необходимо выполнить следующие операции:

* + - * + Выполните операцию перенаправления чтения для целевого объекта, которому требуется откат чтения.
        + Выполните откат перед записью для целевого объекта, которому требуется откат записи.

Остановка процедуры отката не приведет к восстановлению состояния данных целевого объекта на момент до начала этой процедуры. Поэтому для обеспечения дополнительной защиты данных рекомендуется создать моментальный снимок перед запуском процедуры отката. Остановка процедуры отката останавливает только процесс копирования, но не приводит к восстановлению состояния данных исходного объекта на момент до запуска отката.

#### 3.2.1.5 Каскадирование моментальных снимков и межуровневый откат

СХД Huawei OceanStor Dorado V6 поддерживает функцию межуровневого отката. Эта функция предполагает откат обычных и каскадных моментальных снимков одного корневого LUN. Для выполнения отката можно использовать любые два моментальных снимка или моментальный снимок и корневой LUN. Процесс межуровневого отката представлен на этом слайде. Можно выбрать любые два снимка среди snapshot0, snapshot1, Snapshot1.snapshot0 и Snapshot1.snapshot1, чтобы выполнить взаимный откат, или выполнить откат любого из снимков через корневой LUN0.

#### 3.2.1.6 Ключевые технологии: дублирование

Дубликат моментального снимка обеспечивает сохранение копии данных моментального снимка на момент его активации. Копирование пользовательских данных, добавленных в моментальный снимок после активации, не выполняется. Дубликат моментального снимка и исходный моментальный снимок совместно используют пространство тома COW исходного LUN, но личная область памяти у каждого снимка своя. Дубликат моментального снимка представляет собой доступный для записи моментальный снимок, не зависящий от исходного моментального снимка.

Чтение и запись в отношении дубликата моментального снимка выполняются так же, как и в случае с обычным моментальным снимком.

Технология дублирования моментальных снимков позволяет получить несколько копий одного моментального снимка.

#### 3.2.1.7 Ключевые технологии: откат перед записью

Функция HyperSnap конвергентных систем хранения Huawei OceanStor предоставляет возможность отката второго уровня посредством технологии отката перед записью. После запуска процедуры отката к моментальному снимку исходный LUN моментально становится доступным для чтения и записи. Суть технологии отката перед записью состоит в том, что если хост выполняет чтение данных из исходного LUN или запись данных в него во время отката, сначала в исходный LUN копируются блоки данных моментального снимка и только затем — данные хоста. Если хосты не выполняют чтение/запись данных, данные моментального снимка копируются в исходный том последовательно.

#### 3.2.1.8 Сценарии применения

Моментальные снимки нашли широкое применение. Они используются для резервного копирования, анализа данных, проверки состояния приложений или даже репликации данных. Примеры применения:

Функция моментальных снимков периодически создает моментальные снимки файловых систем с доступом только для чтения и копирует их данные на локальные серверы резервного копирования для упрощения последующего запроса и восстановления данных. Таким образом, моментальные снимки служат для резервного копирования, что позволяет ликвидировать окно резервного копирования и обеспечить непрерывную работу файловой системы. Данные хранятся удаленно, что повышает их надежность.

Если данные файловой системы будут повреждены или станут недоступны из-за вирусов или некорректных операций, функция моментального снимка позволит быстро восстановить состояние этих данных на момент создания соответствующего моментального снимка.

Моментальные снимки поддерживают прямой доступ на чтение и запись, что позволяет использовать их для анализа данных и тестирования без влияния на работу служб.

#### 3.2.1.9 Настройка

Процесс развертывания функции моментальных снимков включает проверку соответствующего файла лицензии, создание LUN и собственно моментальных снимков.

Если в пуле хранения нет LUN, который можно использовать в качестве исходного для моментальных снимков, потребуется сначала создать такой LUN.

Создание моментального снимка подразумевает создание согласованной копии данных исходного LUN на определенный момент времени. Данные такой копии можно считывать, не оказывая никакого влияния на исходные данные. Если в пуле хранения, к которому принадлежит исходный LUN, нет доступного пространства, создать моментальный снимок будет невозможно.

LUN моментального снимка можно сопоставить с хостом в обычном порядке.

### 3.2.2 HyperClone (клонирование LUN)

#### 3.2.2.1 Обзор

Функция HyperClone позволяет создавать полные копии LUN без прерывания обслуживания хоста. Такие копии можно использовать для резервного копирования и восстановления, воспроизведения и анализа данных.

#### 3.2.2.2 Принципы работы

HyperClone позволяет создать полную копию данных исходного LUN прямо в момент начала синхронизации данных. Целевой LUN сразу становится доступным для чтения и записи — нет необходимости дожидаться завершения копирования. После синхронизации исходный и целевой LUN физически изолируются друг от друга. При этом операции с одним LUN не будут влиять на другой LUN. В случае повреждения данных исходного LUN их можно восстановить путем обратной синхронизации с целевым LUN. Для поддержки последующей инкрементной синхронизации изменения данных в обоих LUN записывают в битовый массив.

#### 3.2.2.3 Синхронизация

В момент начала синхронизации данных в паре HyperClone система немедленно создает моментальный снимок исходного LUN, синхронизирует его данные с целевым LUN и записывает последующие операции записи в таблицу.

При последующей синхронизации система сравнивает данные исходного и целевого LUN и синхронизирует только различающиеся данные. Данные, записанные в целевой LUN в промежутке между двумя синхронизациями, перезаписываются. Чтобы сохранить изменения в целевом LUN, пользователи могут создать его моментальный снимок перед запуском синхронизации.

Связанные понятия:

1. Пара: пара HyperClone состоит из исходного и целевого LUN. Пара предполагает зеркальную связь между исходным и целевым LUN. Каждый исходный LUN может входить в несколько пар HyperClone с различными целевыми LUN. Каждый целевой LUN может входить только в одну пару HyperClone.
2. Синхронизация: копирование данных из исходного LUN в целевой.
3. Обратная синхронизация: при необходимости восстановления данных исходного LUN это можно сделать путем обратной синхронизации с целевым LUN.
4. Инкрементное копирование: механизм, который в процессе синхронизации позволяет копировать с одного LUN на другой только различающиеся данные в соответствии с битовым массивом.

#### 3.2.2.4 Обратная синхронизация

В случае повреждения данных исходного LUN их можно восстановить с помощью данных целевого LUN. При этом на исходный LUN можно скопировать сразу все или только различающиеся данные.

При запуске обратной синхронизации система создает моментальный снимок целевого LUN и синхронизирует все данные такого снимка с исходным LUN. В случае инкрементной обратной синхронизации система сравнивает данные исходного и целевого LUN и копирует только различающиеся данные.

Исходный и целевой LUN становятся доступными для чтения и записи сразу после запуска обратной синхронизации.

#### 3.2.2.5 Ограничение в отношении используемых функций

Использование HyperClone накладывает определенные ограничения в отношении применения других функций СХД Huawei OceanStor Dorado V3.

#### 3.2.2.6 Сценарии применения

Функция HyperClone широко применяется. Она используется для резервного копирования, анализа данных и проверки состояния приложений. После завершения процесса синхронизации исходный и целевой LUN пары HyperClone физически изолируются друг от друга. При этом операции с одним LUN не будут влиять на другой LUN.

Сценарий использования 1: резервное копирование и восстановление данных

HyperClone позволяет создать одну или несколько резервных копий исходных данных на определенный момент времени, которые можно использовать для восстановления данных исходного LUN в случае их повреждения. Использование целевого LUN позволяет оперативно создавать резервные копии и восстанавливать данные.

Сценарий использования 2: анализ и воспроизведение данных

Анализ данных предполагает исследование больших объемов данных с целью извлечения из них полезной информации, формирования выводов и поддержки принятия решений. Службы анализа могут использовать данные целевых LUN, чтобы не допустить конкуренции с рабочими службами за ресурсы исходного LUN и поддержать стабильную производительность системы.

Функция воспроизведения данных HyperClone позволяет копировать данные одного исходного LUN в несколько целевых LUN.

#### 3.2.2.7 Настройка

Настройка HyperClone выполняется в системе DeviceManager.

Создайте клон для копирования данных в целевой LUN.

Создайте группу защиты для защиты данных. Перед выполнением этого действия необходимо создать соответствующие объекты, в том числе группы защиты и LUN. Группа защиты могут включать один или несколько LUN.

Создайте группу согласованности клонов для унификации операции с клонами, повышения эффективности и обеспечения согласованности данных всех клонов в группе в определенный момент времени.

### 3.2.3 HyperReplication (удаленная репликация)

#### 3.2.3.1 Обзор

По мере развития цифровых технологий в различных отраслях данные стали критически важными для эффективной работы предприятий, а пользователи предъявляют все более строгие требования к стабильности систем хранения. Несмотря на то, что во многих предприятиях развернуты высокостабильные системы хранения, восстановление данных после повреждений, вызванных стихийными бедствиями, все равно представляет для них проблему. Для обеспечения непрерывности предоставления услуг, возможности восстановления и высокой доступности служебных данных были разработаны решения для удаленного аварийного восстановления. HyperReplication — одна из ключевых технологий в этой области.

HyperReplication — это функция удаленной репликации, разработанная Huawei. HyperReplication предоставляет гибкие и эффективные возможности репликации данных, облегчающие удаленное резервное копирование и восстановление данных, обеспечивающие постоянную доступность служебных данных и поддерживающие функции аварийного восстановления.

Первичный объект — это рабочий центр, в котором развернуты первичные системы хранения, серверы приложений и каналы связи.

Вторичный объект — это резервный центр со вторичными системами хранения, серверами приложений и каналами связи.

Функция HyperReplication поддерживает два режима работы:

* + - * + Синхронная удаленная репликация LUN: данные первичных и вторичных LUN синхронизируются в реальном времени. Этот режим позволяет избежать потери данных в случае возникновения аварийных ситуаций. Однако при этом задержка передачи данных между первичным и вторичным LUN влияет на производительность рабочих служб.
        + Асинхронная удаленная репликация LUN: данные первичных и вторичных LUN синхронизируются периодически. При этом задержка передачи данных между первичным и вторичным LUN не влияет на производительность рабочих служб. Однако в случае возникновения аварийных ситуаций возможна потеря данных.

#### 3.2.3.2 Общая информация об аварийном восстановлении и резервном копировании

При использовании функции HyperReplication в двух центрах обработки данных они будут работать в режиме «активный-резервный». При этом один центр (первичный) поддерживает работу служб, а второй находится в резервном режиме.

Режим аварийного восстановления: если в первичном ЦОД откажет устройство или выйдет из строя весь первичный ЦОД, службы автоматически переключаются на резервный ЦОД.

Режим резервного копирования: резервный ЦОД используется только для резервного копирования данных первичного ЦОД и не принимает на себя обслуживание в случае сбоя в первичном ЦОД.

#### 3.2.3.3 Связанные понятия

* + - * + Пара: характеризует отношения репликации данных между первичным и вторичным LUN. Первичный и вторичный LUN в одной паре должны относиться к разным системам хранения.
        + Состояние данных: HyperReplication определяет состояние данных в рассматриваемой паре на основании различий в данных между первичным и вторичным LUN. При возникновении аварийной ситуации HyperReplication разрешает переключение между первичным и вторичным LUN в паре в зависимости от состояния данных пары. Существует два состояния данных пары: полные и неполные.
        + Вторичный LUN, доступный для записи: отправляемые хостом данные можно записать во вторичный LUN. Разрешить запись во вторичный LUN можно в одном из двух случаев:

Первичный LUN выходит из строя, и разрываются каналы связи HyperReplication. В этом случае во вторичной системе хранения можно разрешить запись во вторичный LUN.

Первичный LUN выходит из строя, но каналы связи HyperReplication продолжают работать в штатном режиме. Прежде чем разрешить запись во вторичный LUN во вторичной или первичной СХД, необходимо выполнить разделение пары.

* + - * + Группа согласованности: набор пар, поддерживающих связанные службы. Предположим, в первичной системе хранения есть три первичных LUN, на которых хранятся служебные данные, журналы и информация об отслеживании изменений базы данных соответственно. Если данные на любом из этих LUN недействительны, недействительными становятся все данные на трех LUN. Пары, в которые входят эти три LUN, составляют группу согласованности.
        + Синхронизация: репликация данных первичного LUN во вторичный. HyperReplication поддерживает первичную синхронизацию и инкрементную синхронизацию.
        + Разделение: приостановка синхронизации данных первичного и вторичного LUN. После разделения первичный и вторичный LUN сохраняют связь между собой. Права доступа хостов к обоим LUN также не меняются. В определенной ситуации пользователям может быть полезно перестать копировать данные первичного LUN во вторичный. Например, в случае недостаточной пропускной способности для поддержки критически важных служб может потребоваться приостановить синхронизацию данных HyperReplication. В этом случае можно выполнить разделение, чтобы приостановить синхронизацию.
        + Переключение между первичным и вторичным LUN: процесс, в ходе которого первичный и вторичный LUN в паре удаленной репликации меняются ролями. Этот процесс меняет связи между LUN в паре HyperReplication.

#### 3.2.3.4 Принцип работы

Удаленная репликация HyperReplication предполагает четыре типа операций: создание отношений HyperReplication, синхронизация данных, переключение служб и восстановление данных.

1. Создайте пару HyperReplication.
2. Синхронизируйте все данные первичного и вторичного LUN в паре HyperReplication вручную или автоматически. Кроме того, нужно настроить периодическую синхронизацию новых или измененных данных первичного LUN со вторичным LUN.
3. Проверьте состояние данных в паре HyperReplication и разрешения на доступ ко вторичному LUN для чтения и записи, чтобы определить, можно ли выполнять переключение между первичным и вторичным LUN. Выполните переключение между первичным и вторичным LUN, чтобы создать новую пару HyperReplication.
4. Синхронизируйте данные первичной и вторичной систем хранения. Затем выполните переключение между первичным и вторичным LUN, чтобы восстановить исходные отношения.
5. Состояние пары

В зависимости от состояния пары HyperReplication пользователям доступны функции синхронизации, разделения и переключения между первичным и вторичным LUN. После выполнения соответствующей операции можно просмотреть состояние пары, чтобы убедиться, что операция была выполнена успешно.

1. Принцип работы асинхронной удаленной репликации

В рамках асинхронной удаленной репликации данные из первичной системы хранения периодически копируются во вторичную систему хранения.

Функция асинхронной удаленной репликации в системах хранения Huawei OceanStor использует инновационную технологию кэширования с несколькими временными точками.

1. Переключение служб в рамках HyperReplication

В случае возникновения аварийной ситуации на первичном объекте услуги можно переключить на вторичный объект для обеспечения непрерывности обслуживания. При переключении необходимо учитывать показатели RPO и RPO.

Требования в отношении обработки служб во вторичной системе хранения:

* + До возникновения аварийной ситуации необходимо убедиться, что данные первичных и вторичных LUN согласованы. Если данные вторичного LUN неполны, переключение служб может оказаться невозможным.
  + Службы первичного объекта настроены и на вторичном объекте.
  + Вторичная система хранения предоставляет хосту право на доступ к LUN в группе LUN, сопоставленной с этим хостом.

Если первичный объект выходит из строя в случае аварийной ситуации, каналы связи HyperReplication между первичными и вторичными LUN разрываются. В этом случае для успешного переключения служб администратору потребуется вручную разрешить доступ ко вторичному LUN на запись.

1. Восстановление данных

После сбоя на первичном объекте вторичный объект временно принимает на себя функции первичного. После восстановления первичного объекта службы переключаются обратно.

После восстановления первичного объекта после аварии необходимо восстановить связи HyperReplication между первичной и вторичной системами хранения, после чего восстановить данные первичной системы на основе данных вторичной.

1. Функции групп согласованности

Пользователи могут выполнять синхронизацию, разделение и переключение между первичным и вторичным LUN применительно как к одной, так и сразу к нескольким парам HyperReplication. Для этого используются группы согласованности. Добавление соответствующих LUN в группу согласованности HyperReplication может предотвратить потери данных.

#### 3.2.3.5 Сценарии применения

Функция HyperReplication используется для аварийного восстановления и резервного копирования. К стандартным сценариям ее применения относятся системы централизованного аварийного восстановления и резервного копирования, географического резервирования, а также системы аварийного восстановления на основе ПО BCManager eReplication.

Различные режимы HyperReplication подходят для разных ситуаций.

Асинхронная удаленная репликация применяется в сценариях резервного копирования и аварийного восстановления, когда пропускная способность сети ограничена или первичный объект находится на значительном удалении от вторичного (например, в другой стране или другом регионе).

#### 3.2.3.6 Настройка

Для синхронизации данных между двумя объектами (первичным и вторичным) можно объединить их в пару.

Если не указано иное, представленные на слайде операции поддерживаются только для одного из устройств в паре — первичного или вторичного. Если необходимо выполнить операцию для второго устройства, следует произвести переключение между первичным и вторичным LUN.

### 3.2.4 HyperMetro

#### 3.2.4.1 Обзор

HyperMetro — это решение Huawei для хранения данных с архитектурой «активный-активный». Два центра обработки данных с развернутой технологией HyperMetro не только обеспечивают взаимное резервирование, но и поддерживают предоставление услуг. Если в одном ЦОД откажет устройство или выйдет из строя весь центр обработки данных, службы автоматически переключаются на второй ЦОД, что позволяет решить связанные с переключением проблемы традиционных центров аварийного восстановления. Это позволяет обеспечить высокую степень доступности данных и бесперебойность обслуживания, а также повышает эффективность использования ресурсов СХД.

#### 3.2.4.2 Принципы работы

1. Развертывание в локальном ЦОД: в большинстве случаев хосты развертываются в разных аппаратных одного индустриального парка. Развертывание выполняется в кластерном режиме. Хосты и устройства хранения взаимодействуют друг с другом посредством коммутаторов. Поддерживаются коммутаторы Fibre Channel и IP. Кроме того, для обеспечения корректной работы служб HyperMetro между системами хранения развернуты каналы зеркалирования с поддержкой двойной записи.
2. Развертывание в нескольких ЦОД: обычно хосты развертываются в двух центрах обработки данных в одном городе (или соседних городах). Физическое расстояние между двумя такими центрами не превышает 300 км. Оба ЦОД работают в активном режиме и могут одновременно предоставлять одни и те же услуги, что расширяет общие возможности обслуживания и повышает эффективность использования системных ресурсов. В случае сбоя одного из центров обработки данных службы автоматически переключаются на второй.

Если при развертывании в нескольких центрах обработки данных предполагается передача данных на большие расстояния (не менее 80 км для IP-сетей и 25 км для сетей Fibre Channel), то для поддержания низкой задержки передачи необходимо использовать устройства мультиплексирования с разделением по длине волны (DWDM). Кроме того, для обеспечения корректной работы служб HyperMetro между системами хранения развертываются каналы мониторинга HyperMetro.

Особенности решения HyperMetro:

* + - * + Технология двойной записи данных для обеспечения резервирования хранилища. Позволяет избежать потери данных в случае выхода из строя СХД или всего ЦОД. Поддерживается быстрое переключение служб для обеспечения бесперебойного обслуживания клиентов. Решение удовлетворяет требованиям к нулевому времени восстановления (RTO = 0) и нулевой потере данных (RPO = 0).
        + HyperMetro можно развернуть в сочетании с решением SmartVirtualization для поддержки разнородных хранилищ и объединения ресурсов на сетевом уровне с целью защиты существующих инвестиций клиента.
        + Решение легко модернизировать для реализации поддержки трех ЦОД с помощью компонента HyperReplication.

Все эти функции делают решение HyperMetro отличным вариантом для предприятий в таких отраслях, как здравоохранение, финансы и социальное обеспечение.

#### 3.2.4.3 Режимы кворума

В случае разрыва канала связи между двумя центрами обработки данных или сбоя одного из них данные ЦОД будет невозможно синхронизировать в реальном времени. В этом случае продолжить предоставлять услуги сможет только один участник пары HyperMetro или группы согласованности HyperMetro. Для обеспечения согласованности данных в решении HyperMetro используется механизм арбитража, позволяющий определить приоритеты служб в ЦОД.

HyperMetro поддерживает два режима кворума.

Режим статического приоритета подходит для сценариев без кворум-сервера.

Если кворум-сервер не настроен или недоступен, HyperMetro работает в режиме статического приоритета. При возникновении конфликта объект с более высоким приоритетом выигрывает арбитраж и предоставляет обслуживание.

Режим кворум-сервера (рекомендуется) можно использовать, если настроен соответствующий сервер.

В качестве кворум-сервера используется независимый физический сервер или виртуальная машина. Кворум-сервер рекомендуется развертывать на отдельной площадке, не относящейся ни к одному из двух рассматриваемых ЦОД. В этом случае кворум-сервер продолжит работу даже в случае сбоя одного из ЦОД.

При использовании режима кворум-сервера в случае сбоя в центре обработки данных или разрыва канала связи между системами хранения каждая система хранения отправляет запрос арбитража на сервер кворума. Только выигравшая арбитраж система продолжит предоставлять услуги. Объект с более высоким приоритетом имеет преимущество при арбитраже.

#### 3.2.4.4 Принцип работы двойной записи

Для поддержки согласованности данных в системах хранения двух ЦОД в HyperMetro используется технология двойной записи и журнал изменений в данных (data change log, DCL). Оба центра обработки данных одновременно предоставляют хостам возможности чтения и записи данных.

Механизмы блокировки и двойной записи имеют ключевое значение для поддержания согласованности данных двух систем хранения.

Две СХД с поддержкой HyperMetro параллельно обрабатывают запросы ввода/вывода хостов. Для предотвращения конфликтов доступа, возникающих при получении обеими СХД запросов на редактирование одного блока данных, записывать данные может только та система хранения, которая получила разрешение от механизма блокировки. При этом вторая система должна будет дождаться снятия блокировки, после чего получить разрешение на запись.

#### 3.2.4.5 Высокая степень согласованности данных

При использовании HyperMetro для целей аварийного восстановления две СХД могут одновременно обрабатывать операции чтения и записи. При одновременном обращении к одному адресу памяти для чтения/записи данных за согласованность данных в обеих системах отвечает уровень хранения.

Согласованность данных на уровне приложений: базы данных, развернутые на нескольких объектах; развернутые в кластере приложения; архитектура общего хранилища.

Согласованность данных на уровне хранения

При нормальных условиях запросы приложений на запись обрабатываются обеими СХД одновременно, обеспечивая согласованность данных между ними.

Если одно из устройств хранения недоступно, также записываются изменения данных. Если одно из устройств хранения недоступно, запись обрабатывается только стабильно работающей системой, а изменения в данных, внесенные за это время, записываются в DCL. После восстановления неисправной системы ее данные синхронизируются на основе записей DCL.

Распределенное управление блокировками (DLM): при одновременном обращении нескольких хостов к одному адресу памяти только один хост получает право записывать данные. Это позволяет обеспечить согласованность данных.

#### 3.2.4.6 Широкие возможности масштабирования

Два центра обработки данных с развернутой технологией HyperMetro не только обеспечивают взаимное резервирование, но и поддерживают предоставление услуг. Если в одном ЦОД откажет устройство или выйдет из строя весь центр обработки данных, службы автоматически переключаются на второй ЦОД, что позволяет решить связанные с переключением проблемы традиционных центров аварийного восстановления. Это позволяет обеспечить высокую степень доступности данных и бесперебойность обслуживания, а также повышает эффективность использования ресурсов СХД. Решение HyperMetro легко модернизировать для реализации поддержки трех ЦОД с помощью компонента HyperReplication.

В таком модернизированном решении используются три центра обработки данных для обеспечения непрерывности обслуживания в случае отказа любых двух из них, что значительно повышает степень доступности служб для клиентов. Под тремя центрами обработки данных понимаются рабочий центр, местный центр аварийного восстановления и удаленный центр аварийного восстановления. Они развертываются в двух разных регионах для поддержки географического резервирования.

#### 3.2.4.7 Стандартный сценарий применения

Решение HyperMetro подходит для таких отраслей, как здравоохранение, финансы и социальное обеспечение.

#### 3.2.4.8 Настройка

Решение HyperMetro необходимо настроить в обеих рассматриваемых СХД.

Для СХД B в центре обработки данных B необходимо проверить файл лицензии, настроить IP-адрес порта кворума, создать кворум-сервер и настроить сопоставление.

Если при создании пары HyperMetro не была создана группа согласованности, такую группу потребуется создать отдельно.

|  |
| --- |
| 4 Решения для поддержки непрерывности бизнеса |
|  |

## 4.1 Общая информация о решениях для резервного копирования

### 4.1.1 Обзор

#### 4.1.1.1 Общая информация и определения

Зачастую угрозы безопасности данных невозможно предотвратить. Рано или поздно возникают аварийные ситуации, которые приводят к повреждению целостности данных и систем, обеспечивающих доступ к этим данным. Можно выделить следующие основные причины потери данных и повреждения систем:

* + - * 1. Неисправности программного обеспечения и платформ, используемых для обработки и доступа к данным
        2. Намеренно или непреднамеренно созданные уязвимости операционных систем
        3. Неисправности оборудования
        4. Ошибочные действия
        5. Незаконный доступ злоумышленников
        6. Перебои питания

Важные данные, архивы и исторические записи, хранящиеся в компьютерных системах, имеют критическое значение как для предприятий, так и для частных лиц. Любая потеря данных может привести к значительным экономическим убыткам, обнулению нескольких лет исследований и разработок и даже нарушению нормальной работы и деятельности предприятий.

Для поддержки нормальной производственной, маркетинговой деятельности, а также деятельности в области НИОКР предприятия должны принимать передовые и эффективные меры для резервного копирования своих данных в целях профилактики аварийных ситуаций.

В областях ИТ и управления данными под резервной копией понимается копия данных в файловой системе или базе данных, с помощью которой можно быстро восстановить актуальные данные и возобновить нормальную работу системы в случае аварии или ошибок.

#### 4.1.1.2 Резервное копирование, аварийное восстановление и архивирование

Аварийное восстановление и резервное копирование — это разные концепции. Системы аварийного восстановления поддерживают непрерывность бизнеса, обеспечивая нормальную работу ИТ-систем в случае возникновения аварийной ситуации. Системы резервного копирования ориентированы на предотвращение потерь данных при авариях. До появления интегрированных систем аварийного восстановления и резервного копирования эти функции выполняли отдельные решения. Интегрированная система аварийного восстановления и резервного копирования помогает предприятиям защищать свои данные как в случае небольших сбоев, связанных, например, с человеческими ошибками, неисправностями программного обеспечения и вирусами, так и от серьезных аварий вследствие отказа оборудования и стихийных бедствий. Архивирование файлов — еще один популярный метод защиты данных. В системах архивирования применяются недорогие носители информации (например, ленты), а соответствующие операции можно выполнять в автономном режиме. Таким образом, архивирование помогает сократить затраты и упростить хранение. Система архивирования файлов также может сохранять файлы в соответствии с их атрибутами. Таким атрибутом может быть создатель файла, дата его редактирования или иные пользовательские метки. Система архивирования сохраняет файлы вместе с их метаданными и атрибутами. Кроме того, она предоставляет функции сжатия данных. Наконец, архивирование подразумевает долгосрочное автономное хранение резервных копий данных, которые больше не предполагается часто использовать или обновлять. Такие данные можно снабдить специальной меткой в соответствии с конкретными атрибутами для упрощения поиска в будущем.

Как правило, под резервным копированием понимается резервное копирование данных или систем, а аварийное восстановление связано с резервным копированием данных или приложений во всех используемых аппаратных. Для резервного копирования используется соответствующее программное обеспечение, а для аварийного восстановления — программное обеспечение для репликации или зеркалирования. Две описанные системы также имеют следующие различия:

* + - * 1. Системы аварийного восстановления предназначены для защиты данных от стихийных бедствий, таких как пожары и землетрясения. Поэтому соответствующий резервный центр должен быть развернут на определенном удалении от рабочего центра. В свою очередь, резервное копирование данных осуществляется прямо в центре обработки данных.
        2. Система аварийного восстановления не только защищает данные, но и поддерживает непрерывность бизнеса. Системы резервного копирования ориентированы только на защиту данных.
        3. Система аварийного восстановления поддерживает целостность данных. Система резервного копирования позволяет восстановить состояние данных только на момент создания резервной копии.
        4. Аварийное восстановление выполняется в реальном времени, а резервное копирование — в автономном режиме.
        5. Данные двух объектов, которые поддерживает система аварийного восстановления, синхронизируются в реальном времени, в то время как данные резервных копий зависят от времени.
        6. При возникновении неисправности процессы аварийного восстановления выполняются за секунды или минуты, а у системы резервного копирования на восстановление данных могут уйти десятки часов.

Системы резервного копирования и архивирования предназначены для защиты данных различными способами, и сочетание этих двух систем повысит эффективность защиты данных. Системы резервного копирования обеспечивают защиту данных за счет сохранения их копий. Системы архивирования организуют данные и поддерживают их долгосрочное хранение в соответствии с принципами управления данными. Другими словами, резервное копирование можно рассматривать как краткосрочное хранение копий данных, а архивирование — как долгосрочное хранение файлов. На практике исходные данные не удаляются после резервного копирования. Однако после архивирования исходные данные можно удалить, так как не предполагается, что они могут понадобиться в ближайшее время. Сочетание систем резервного копирования и архивирования данных позволяет повысить степень защиты данных.

### 4.1.2 Архитектура

#### 4.1.2.1 Компоненты

Система резервного копирования обычно включает три компонента: программное обеспечение для резервного копирования, устройства хранения резервных копий и сервер резервного копирования.

Программное обеспечение для резервного копирования позволяет настраивать политики резервного копирования, управлять носителями и добавлять функции. Программное обеспечение для резервного копирования — это основа системы резервного копирования. С его помощью выполняется копирование рабочих данных на резервные носители и управления созданными копиями. Некоторые образцы программного обеспечения для резервного копирования могут иметь ряд дополнительных функций, связанных с защитой, резервным копированием, архивированием и восстановлением данных.

К устройствам хранения резервных копий относятся физические и виртуальные ленточные библиотеки, а также дисковые массивы. Виртуальная ленточная библиотека — это дисковый массив, в котором устройства хранения данных представлены как ленточные библиотеки с помощью технологий виртуализации. В отличие от механических лент виртуальные ленточные библиотеки совместимы с программным обеспечением для управления резервным копированием на ленточные накопители, а также с традиционными процессами резервного копирования, что значительно повышает их доступность и надежность.

Сервер резервного копирования поддерживает реализацию политик резервного копирования. Сервер резервного копирования размещает и выполняет программное обеспечение для резервного копирования. Обычно для поддержки резервного копирования на хосте требуется установить агент клиента ПО для резервного копирования.

В системах резервного копирования используется три целевых показателя: окно резервного копирования, допустимая потеря данных (RPO) и допустимое время восстановления данных (RTO).

Окно резервного копирования — это интервал времени, в течение которого резервное копирование данных выполняется без серьезного влияния на приложения.

RPO применяется с целью использования наиболее актуальных данных для аварийного восстановления. Чем меньше значение RPO, тем меньше потери данных.

RTO характеризует допустимое время восстановления, а также уровень обслуживания, в пределах которых необходимо восстановить работу бизнес-процесса для минимизации влияния прерывания на обслуживание.

#### 4.1.2.2 Обзор решений для резервного копирования

Компания Huawei предоставляет решения для резервного копирования в формате «все в одном», решения для централизованного резервного копирования, а также решения для резервного копирования в облако.

Решение Huawei Data Protection Appliance (Программно-аппаратный комплекс для защиты данных) — это продукт для защиты данных и управления ими, в котором сочетаются программное обеспечение для резервного копирования, сервер резервного копирования и хранилище резервных копий. Благодаря распределенной архитектуре комплекс Huawei Data Protection Appliance поддерживает линейное увеличение емкости и производительности. Для поддержки процессов защиты и создания пользовательских данных и приложений, а также управления ими достаточно развернуть всего одну систему. Это повышает качество защиты данных для пользователей, позволяет сэкономить средства на защиту данных и упрощает процесс управления данными. Комплекс Huawei Data Protection Appliance применим в широком спектре сценариев, но особенно хорошо подходит для государственного сектора, финансовой отрасли, операторского бизнеса, а также сфер здравоохранения и производства.

Этот программно-аппаратный комплекс предоставляет систему управления с графическим интерфейсом для централизованного управления программным и аппаратным обеспечением системы резервного копирования, а также его обслуживания.

Решение для централизованного резервного копирования предполагает использование специализированных узлов для управления локальными и удаленными центрами обработки данных и централизованного планирования задач резервного копирования, что значительно упрощает эксплуатацию системы и позволяет пользователям эффективно использовать ее с учетом общего состояния.

Преимущества решения для централизованного резервного копирования:

* + - * 1. Централизованное управление задачами резервного копирования и восстановления данных из различных приложений, позволяющее сформировать единую политику управления.
        2. Объединение ресурсов резервного копирования для оптимизации их использования.
        3. Гибкие возможности масштабирования за счет поддержки добавления любых ленточных накопителей, клиентов и ленточных библиотек.
        4. Упрощенное правление — для управления большим количеством устройств и систем требуются усилия меньшего числа специалистов.

Решения для резервного копирования в облако поддерживают копирование данных из локального рабочего центра в центр обработки данных (главный ЦОД предприятия или ЦОД поставщика услуг) с использованием стандартного сетевого протокола и глобальной сети WAN. Резервное копирование в облако предоставляется как доступная из любого места услуга. Такие системы характеризуются гибкостью и безопасностью и поддерживают совместный доступ, а также доступ по запросу. Резервное копирование в облако — это услуга нового поколения, разработанная на базе технологий широкополосного Интернета и с учетом увеличения емкости систем хранения. Соответствующие решения предоставляют услуги хранения и резервного копирования данных посредством различных функций, таких как кластерные приложения, сетевые технологии и распределенные файловые системы, а также интеграции различных устройств хранения в сети с помощью прикладного программного обеспечения.

### 4.1.3 Режимы сетевой передачи данных

Стандартные режимы сетевой передачи данных: по локальной сети; с ограниченным использованием локальной сети; с ограниченным использованием сервера; без участия сервера. Режим передачи по локальной сети поддерживают все типы СХД, а режимы с ограниченным использованием локальной сети или сервера — только сети хранения данных SAN.

Передача по локальной сети

Резервное копирование по локальной сети предполагает передачу как потоков данных, так и потоков управления по локальной сети, поэтому характеризуется высокой нагрузкой на сеть. Вследствие этого резервное копирование больших объемов данных за короткий промежуток времени может вызвать перегрузку сети.

Процесс резервного копирования: сервер резервного копирования по локальной сети отправляет поток управления на сервер приложений, на котором установлен агент. Сервер приложений отвечает на запрос и отправляет данные на сервер резервного копирования. Сервер резервного копирования получает данные и сохраняет их на устройстве хранения. Таким образом, процесс резервного копирования завершается.

Преимущества:

* Система резервного копирования и система приложений не зависят друг от друга, что позволяет сократить потребление аппаратных ресурсов серверов приложений в процессе резервного копирования.

Недостатки:

* Развертывание дополнительных серверов резервного копирования приводит к увеличению затрат на оборудование.
* Работа агентов резервного копирования негативно влияет на производительность серверов приложений.
* Данные резервного копирования передаются по локальной сети, что отрицательно сказывается на ее производительности.
* Службы резервного копирования требуется поддерживать отдельно, что усложняет процессы управления и обслуживания.
* Обрабатывать службы резервного копирования должны высококвалифицированные пользователи. Передача с ограниченным использованием локальной сети

В этом режиме по локальной сети передаются только потоки управления, но не потоки данных. При резервном копировании с ограниченным использованием локальной сети потоки данных передаются через сеть хранения данных SAN, а не по локальной сети. Сервер, для которого необходимо создать резервную копию, подключается к устройству хранения резервных копий по сети SAN. Клиент резервного копирования активирует медиа-сервер, который считывает данные, подлежащие резервному копированию, и копирует их на общее устройство хранения резервных копий.

Процесс резервного копирования: сервер резервного копирования отправляет поток управления по локальной сети на сервер приложений, на котором установлен агент. Сервер приложений отвечает на запрос и считывает рабочие данные. Затем медиа-сервер считывает данные с сервера приложений и передает их на устройство хранения резервных копий. Таким образом процесс резервного копирования завершается.

Преимущества:

* Передача данных резервного копирования не создает нагрузку на локальную сеть, что значительно повышает производительность резервного копирования, поддерживая при этом высокую производительность сети.

Недостатки:

* Работа агентов резервного копирования негативно влияет на производительность серверов приложений.
* Резервное копирование с ограниченным использованием локальной сети требует высоких затрат.
* Устройства должны соответствовать определенным требованиям.

Передача с ограниченным использованием сервера

Резервное копирование с ограниченным использованием сервера во многом предоставляет те же преимущества, что и резервное копирование с ограниченным использованием локальной сети. Основные компоненты канала резервного копирования — это исходное устройство, целевое устройство и устройство SAN. Сервер по-прежнему участвует в процессе резервного копирования, но обрабатывает гораздо меньше операций, поскольку не относится к основному каналу резервного копирования, не обрабатывает рабочие нагрузки, а отвечает только за подачу команд. В этом режиме по локальной сети передаются только потоки управления, но не потоки данных.

Процесс резервного копирования: данные резервного копирования передаются по независимой сети в обход рабочего сервера.

Преимущества:

* Передача данных резервного копирования не создает нагрузку на локальную сеть и не влияет на ее производительность.
* Резервное копирование почти не влияет на работу служб хоста.
* Обеспечивает высочайшую производительность резервного копирования.

Недостатки:

* Резервное копирование с ограниченным использованием сервера требует высоких затрат.
* Устройства должны соответствовать строгим требованиям.

Передача без участия сервера

В процессах резервного копирования без участия сервера используется протокол NDMP (протокол управления сетевыми данными). NDMP — это стандартный протокол для выполнения операций резервного копирования. Он обеспечивает связь между интеллектуальными устройствами хранения данных, ленточными библиотеками и приложениями резервного копирования. После получения команды NDMP от сервера устройство хранения, поддерживающее протокол NDMP, может напрямую передавать данные на другие устройства в обход хоста.

### 4.1.4 Основные технологии резервного копирования

#### 4.1.4.1 Типы и политики резервного копирования

Основные типы резервного копирования: полное резервное копирование, кумулятивное инкрементное резервное копирование и дифференциальное инкрементное резервное копирование.

Полное резервное копирование предполагает создание копии всех данных на определенный момент времени.

Преимущества:

* Возможность быстрого восстановления данных на основе последней резервной копии.
* Высокая скорость восстановления.

Недостатки:

* Высокая потребность в пространстве для хранения.
* Длительный процесс резервного копирования.

В процессе кумулятивного инкрементного резервного копирования (или просто инкрементного резервного копирования) данные добавляются к последней полной резервной копии. Если такая копия отсутствует, выполняется резервное копирование всех данных.

Преимущества:

* Для создания каждой последующей копии требуется меньше дискового пространства.
* Высокая скорость восстановления и резервного копирования.

Недостатки:

* Для полноценного восстановления всех данных требуются как последняя полная, так и текущая инкрементная резервные копии.
* Восстановление выполняется быстрее, чем в случае дифференциального резервного копирования.

В процессе дифференциального инкрементного резервного копирования (или дифференциального резервного копирования) данные добавляются к последней резервной копии, независимо от ее типа. Если такая копия отсутствует, выполняется резервное копирование всех данных.

Преимущества:

* Наибольшая экономия дискового пространства и уменьшение окна резервного копирования.

Недостатки:

* Для полноценного восстановления всех данных требуются как последняя полная, так и все дифференциальные резервные копии, что обуславливает высокую длительность процесса восстановления.

Политика резервного копирования определяет время и метод резервного копирования, а также данные, подлежащие копированию.

Выбор политики резервного копирования:

В отношении операционных систем и приложений: следует выполнять полное резервное копирование при обновлении операционных систем или установке нового программного обеспечения.

В отношении критически важных данных приложений, которые характеризуются небольшим общим объемом, но высокой частотой обновлений в день: следует ежедневно выполнять полное резервное копирование с учетом окна резервного копирования.

В отношении критически важных данных приложений с не очень высокой частотой обновлений в день: следует еженедельно или ежемесячно выполнять полное резервное копирование. Также следует часто выполнять инкрементное резервное копирование.

#### 4.1.4.2 Дедупликация

Цифровые преобразования предприятий привели к резкому росту объемов служебных данных. Общий объем резервных копий, требующих защиты, также быстро увеличивается. При этом в процессах резервного копирования и архивирования генерируется все больше повторяющихся данных. Такие данные занимают ресурсы хранения и пропускной способности, приводя к увеличению окон резервного копирования, что, в свою очередь, снижает степень доступности систем обслуживания.

Комплекс Huawei Data Protection Appliance поддерживает дедупликацию на стороне источника, а также параллельную дедупликацию. Дедупликация выполняется перед сохранением данных резервной копии на носителе, что позволяет значительно повысить производительность резервного копирования.

Дедупликация на стороне источника

Данные или файлы делятся на блоки с помощью интеллектуального алгоритма дедупликации на основе анализа содержимого. Затем путем хеширования рассчитываются отпечатки блоков данных, которые сравниваются с содержимым библиотеки отпечатков на предмет наличия дубликатов. Если они существуют, значит, на медиа-сервере хранятся одинаковые блоки данных. В целях обеспечения экономии дискового пространства и ресурсов пропускной способности, а также упрощения передачи и хранения данных будут использоваться только уже существующие блоки данных.

Принцип работы:

* + - * 1. С помощью алгоритма хеширования создается отпечаток блока данных.
        2. Система проверяет библиотеку отпечатков в программно-аппаратном комплексе на предмет наличия такого же отпечатка. Если он существует, новый записываемый блок данных рассматривается как дубликат и не отправляет в СХД программно-аппаратного комплекса. В противном случае блок отправляется в программно-аппаратный комплекс и записывается в пул хранения резервных копий. Затем отпечаток такого блока заносится в библиотеку отпечатков.

Параллельная дедупликация

Большинство традиционных режимов дедупликации предполагают использование одного узла и могут столкнуться с проблемами низкой эффективности доступа к данным, низкой производительности обработки и недостаточной емкости хранилища в эпоху больших данных.

В комплексе Huawei Data Protection Appliance используется технология параллельной дедупликации, в рамках которой библиотеки отпечатков развертываются на нескольких узлах. Отпечатки также обрабатываются всеми участвующими узлами в параллельном режиме. Это позволяет решить проблемы производительности и емкости хранилища, с которыми сталкиваются решения на базе одного узла.

Принцип работы:

После расчета отпечатков для блоков данных система применяет алгоритм группирования для поиска определенных серверных узлов. Разные отпечатки равномерно распределяются между разными узлами. Таким образом система позволяет одновременно проверять разные узлы на наличие дубликатов для поддержки параллельной дедупликации.

Библиотеки отпечатков позволяют сэкономить пространство за счет использования ссылок на уже существующие данные вместо записи новых данных. Такой подход сокращает время опроса всех отпечатков при выполнении глобальной дедупликации и позволяет максимально эффективно использовать кэш чтения СХД, чтобы снизить частоту поиска по диску при случайном чтении и повысить эффективность восстановления.

#### 4.1.4.3 Режимы резервного копирования

Резервное копирование на основе моментальных снимков

В этом режиме резервное копирование выполняется с использованием моментальных снимков системы хранения и агентов.

Высокая скорость и широкие возможности восстановления:

* Позволяет просматривать информацию резервных копий и быстро восстанавливать выбранные объекты.
* В фоновом режиме объединяет инкрементные копии в полную копию для поддержки быстрого восстановления данных.
* Позволяет восстанавливать данные на основе моментальных снимков оборудования, а также восстанавливать отдельные объекты.

Восстановление копий:

Защита массива хранения:

* Встроенные функции
* Автоматическое обнаружение СХД
* Полная интеграция (без использования скриптов)
* Поддержка моментальных снимков

Восстановление массива хранения:

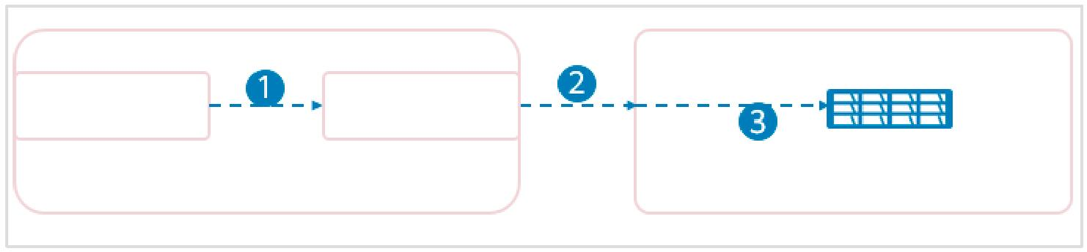
* Позволяет восстанавливать, клонировать и монтировать (подключать) тома.
* Позволяет восстанавливать данные из копий.

Стандартное резервное копирование

Стандартное резервное копирование — это механизм защиты данных, выполняющий соответствующие операции в соответствии с определенным расписанием. Задача резервного копирования автоматически инициализируется в указанное время в соответствии с политикой и планом резервного копирования, выполняет чтение данных и их запись на устройство хранения резервных копий.

Принцип работы:

Стандартный процесс резервного копирования включает три этапа:



Huawei CDM Data Protection Appliance

Рабочая среда

Клиент резервного копирования

##### Рис. 4-1

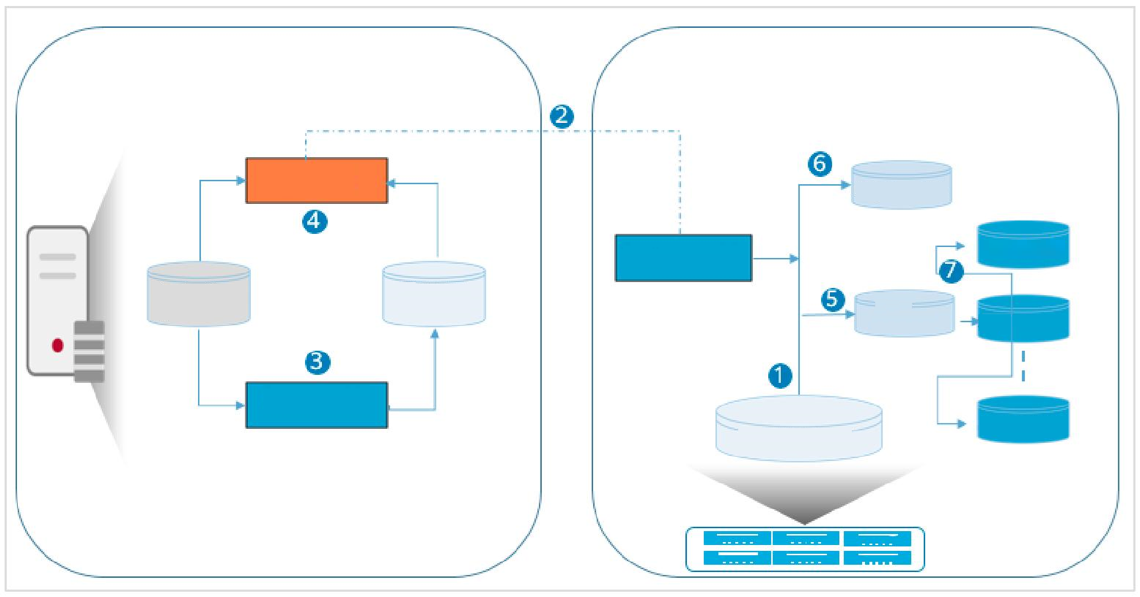
* + - * 1. Чтение данных, подлежащих копированию, с помощью клиента резервного копирования. В зависимости от приложения можно использовать клиент, развернутый на рабочем сервере (резервное копирование с использованием агента) или встроенный в комплекс Huawei Data Protection Appliance (резервное копирование без агента).
        2. Передача данных из рабочей системы в комплекс Data Protection Appliance через сеть TCP.
        3. Получение данных комплексом Data Protection Appliance и их сохранения в хранилище резервных копий.

В разных режимах резервного копирования (полное, инкрементное, постоянное инкрементное или дифференциальное резервное копирование) данные передаются по-разному. Могут передаваться сразу все данные или только уникальные данные, оставшиеся после дедупликации.

Если требуется поддержка удаленного аварийного восстановления, функция удаленной репликации позволит выполнить репликацию данных резервных копий в удаленные центры обработки данных.

Непрерывное резервное копирование

Процесс непрерывного резервного копирования данных рабочих узлов на устройства хранения резервных копий. В основе этой функции лежит технология непрерывной защиты данных на уровне блоков. При этом на рабочие узлы устанавливаются клиенты агента резервного копирования. Данные рабочих узлов в непрерывном режиме копируются в пул хранения моментальных снимков внутренней СХД комплекса Data Protection Appliance, где хранятся в исходном формате. При определенных условиях в пуле хранения моментальных снимков создаются моментальные снимки, которые позволяют управлять состояниями данных в различные моменты времени.



Снимок T0

Снимок T1

Снимок TN

Пул хранения

Основной том

Том журналов

Сервер CDP

Клиент

Рабочий том

Пул памяти

Драйвер CDP

Data Protection Appliance

Рабочий узел

##### Рис. 4-2

1. Пул хранения моментальных снимков выделяет ресурсы под основной том.
2. Клиент агента непрерывного резервного копирования подключается к серверу Data Protection Appliance.
3. Отдельный драйвер непрерывной защиты данных в разделе рабочего узла постоянно регистрирует изменения в данных и записывает их в кэш пула памяти.
4. Клиент агента для непрерывного резервного копирования постоянно передает данные на устройство хранения в пуле хранения моментальных снимков комплекса Data Protection Appliance.
5. Исходные данные рабочего узла записываются в основной том.
6. Изменения в данных рабочего узла сначала записываются в том журналов, а затем — в основной том, где хранятся исходные данные.
7. Управление моментальными снимками основного тома осуществляется в соответствии с политикой хранения данных в рамках непрерывного резервного копирования.

Расширенное резервное копирование

Функция расширенного резервного копирования, доступная в комплексе Huawei Data Protection Appliance, сочетает многолетний опыт специалистов в областях резервного копирования и аварийного восстановления со специализированной системой хранения копий данных для обеспечения согласованности данных приложений. Она включает возможности автоматизации и аварийного восстановления на основе политик, а также инструменты автоматизации для разработчиков, поддерживает работу с разнородными хранилищами рабочих данных и позволяет развернуть полноценную систему управления данными копий.

Принцип работы:

Сбор рабочих данных: данные собираются в исходном формате. Преобразование формата не требуется. Доступ к данным возможен после их монтирования. Политику SLA можно настроить в соответствии с используемыми приложениями. Система в удобном формате выводит информацию о сроке и местах хранения данных, а также показатели RPO и RTO.

Управление копированием

Постоянное инкрементное резервное копирование предполагает выполнение первоначального полного резервного копирования и *N* операций инкрементного резервного копирования. В момент выполнения инкрементного копирования создается полная копия данных на соответствующий момент времени. Повреждение копии данных на момент инкрементного резервного копирования не препятствует восстановлению данных на любой другой момент времени.

Без необходимости в откате: копии данных на определенный момент времени, созданные с помощью виртуального клона, предоставляют доступ как к исходным данным, так и к текущим добавленным данным. Такие копии можно напрямую использовать для восстановления.

Доступ к копиям и их использование

Не требует перемещения данных: подключение данных выполняется за минуты, а их объем не влияет на эффективность восстановления.

Виртуальную копию можно смонтировать на несколько хостов.

Можно восстановить состояние данных на любой момент времени.

После монтирования виртуальной копии хост автоматически принимает управление исходными рабочими приложениями.

### 4.1.5 Приложения

Базы данных

Базы данных имеют критически важное значение для рабочих систем. Встроенная функция резервного копирования баз данных требует сложных ручных операций. Кроме того, необходимо обеспечивать защиту различных баз данных на разных платформах, что требует использования продуктов для резервного копирования с высокой степенью совместимости.

Комплекс Data Protection Appliance включает мастер резервного копирования с графическим интерфейсом. Это позволяет пользователям отказаться от ручного выполнения операций резервного копирования и восстановления, что значительно упрощает соответствующие процессы. Процесс резервного копирования базы данных включает следующие этапы:

Установка агента клиента резервного копирования на рабочий сервер, данные которого необходимо защитить, и подключение этого агента к консоли управления. Агент клиента резервного копирования определяет данные базы данных на рабочем сервере, считывает файлы и данные с этого сервера посредством API резервного копирования и передает эти файлы и данные в СХД комплекса Data Protection Appliance для резервного копирования. Консоль управления Data Protection Appliance отправляет управляющую информацию клиенту и серверу Data Protection Appliance и осуществляет управление задачей резервного копирования.

Процесс резервного копирования: агент клиента резервного копирования вызывает API резервного копирования базы данных через API для чтения данных в базе данных, выполняет дедупликацию или шифрование, а затем отправляет данные в СХД Data Protection Appliance для резервного копирования.

Процесс восстановления: консоль управления отправляет соответствующую команду агенту клиента резервного копирования на рабочем сервере. Этот агент вызывает API восстановления базы данных через API для чтения данных с сервера резервного копирования, а затем отправляет данные в API восстановления.

Комплекс Data Protection Appliance подключается к базе данных через специализированный API для резервного копирования. API зависит от используемой базы данных. Например, в БД Oracle используется интерфейс RMAN, а в SQL Server — интерфейс VDI.

Платформы виртуализации

Рост популярности технологий виртуализации приводит к тому, что все больше предприятий стремятся хранить свои основные данные в виртуальной среде. Поэтому предприятиям необходимы системы защиты данных в виртуальной среде. В частности, особое внимание уделяется эффективности резервного копирования и восстановления данных.

Data Protection Appliance — это комплексное и эффективное решение для защиты платформ виртуализации, которое предоставляет следующие преимущества:

* Защита виртуальных данных для повышения эффективности резервного копирования.
* Унифицированная система защиты физических и виртуальных сред для упрощения эксплуатации и технического обслуживания.
* Гибкие возможности восстановления данных, позволяющие избежать прерывания обслуживания.
* Возможность резервного копирования без использования агентов для минимизации нагрузки на хосты и оптимизации производительности рабочей среды.

Процесс резервного копирования включает следующие этапы: создание моментального снимка виртуальной машины (ВМ). Резервное копирование данных ВМ, включая информацию о конфигурации и данные на виртуальных дисках. В процессе резервного копирования можно использовать технологию CBT для копирования действительных блоков данных или измененных блоков данных с дисков виртуальных машин. В режиме полного резервного копирования с виртуальных дисков копируются действительные блоки данных. В режиме инкрементного копирования с виртуальных дисков копируются измененные блоки данных. Удаление моментального снимка ВМ, созданного на первом этапе.

Процесс восстановления: для восстановления данных на новой виртуальной машине вручную создайте виртуальную машину на основе конфигурации исходной виртуальной машины. Для восстановления путем перезаписи виртуальной машины вручную настройте перезапись на соответствующей ВМ. Система в определенный момент времени считывает данные дисков с медиа-сервера и записывает их на диски виртуальной машины, указанной на первом этапе. Если ВМ FusionCompute использует решение FusionStorage, а не виртуализированное хранилище, процессы резервного копирования и восстановления будут отличаться. При использовании FusionStorage диски ВМ сопоставляются с томами LUN FusionStorage. Для получения измененных блоков данных с виртуальных дисков используется соответствующий битовый массив FusionStorage.

Файловые системы

Модуль резервного копирования файловых систем в комплексе Data Protection Appliance позволяет выполнять резервное копирование неструктурированных файловых систем. Особенности модуля:

* Поддерживаемые типы резервного копирования: полное, инкрементное и постоянное инкрементное резервное копирование.
* Степень детализации резервного копирования и восстановления: по одному файлу, по одному каталогу или весь диск.
* Дедупликация на уровне блоков: сокращает объем передаваемых при резервном копировании данных, позволяя уменьшить окно резервного копирования и сэкономить сетевые ресурсы и ресурсы хранения.
* Место восстановления: позволяет восстановить файловую систему в том же месте на исходном хосте, в указанном месте на исходном хосте или на другом хосте.
* Инкрементное резервное копирование: поддержка инкрементного резервного копирования на уровне файлов. Система сравнивает состояние файла во время резервного копирования с его состоянием во время последнего изменения, чтобы определить, нужно ли выполнять резервное копирование такого файла. Если да, система выполняет инкрементное резервное копирование в соответствии с результатами сравнения.

В режиме резервного копирования файловой системы доступно четыре режима фильтрации источников данных, которые упрощают для пользователей выбор файлов для резервного копирования.

Процесс резервного копирования:

1. Клиент, развернутый в рабочей системе, считывает данные, подлежащие резервному копированию.
2. Клиент передает считанные данные по сети.
3. Комплекс Data Protection Appliance получает данные и сохраняет их на физические носители. Таким образом процесс резервного копирования завершается.
4. Выполняется проверка состояния хоста (Windows и Linux) Windows:

После установки клиента нажмите сочетание клавиш **Windows+R**. В открывшемся диалоговом окне «**Выполнить**» введите **services.msc**, чтобы открыть окно управления службами. Проверьте, запущена ли служба клиента должным образом. Если это так, проверьте, подключен ли клиент к серверу. Если подключение установлено, можно создать стандартный план резервного копирования файловой системы.

Linux:

Требования к резервному копированию файловой системы Linux аналогичны требованиям в отношении файловой системы Windows. Проверьте состояние службы или процесса. Если клиент подключен к серверу, можно создать стандартный план резервного копирования файловой системы. В дистрибутиве CentOS 7 для проверки используются следующие команды:

**systemctl status HWClientService.service**

**ps -ef|grep esf**

Операционные системы

Процесс резервного копирования операционной системы:

1. Установите клиент, чтобы собрать данные об операционной системе.
2. В Windows вызовите интерфейс VSS, чтобы создать моментальный снимок тома, на котором размещена операционная система. В Linux выберите источник данных для резервного копирования.
3. Считайте данные тома, на котором размещена операционная система, и создайте резервную копию этих данных в СХД Data Protection Appliance.
4. На этом процесс резервного копирования завершается (в Windows также потребуется удалить созданный моментальный снимок).

Процесс восстановления:

1. Загрузите среду предустановки Windows WinPE или носитель LiveCD, чтобы загрузить среду восстановления. В Linux установите клиент.
2. В Windows вручную разбейте диск на разделы.
3. Выполните восстановление данных операционной системы с носителя в указанный системный том.
4. В Windows используйте системный API, чтобы загрузить драйвер, изменить значения в реестре и исправить BSOD. В Linux отредактируйте файл конфигурации.
5. Перезагрузите операционную систему по завершении операции восстановления.

## 4.2 Общая информация о решениях для аварийного восстановления

### 4.2.1 Обзор решений для аварийного восстановления

#### 4.2.1.1 Определение аварийного восстановления

Развертывание систем аварийного восстановления необходимо для минимизации последствий аварийных ситуаций.

Согласно статистике международной организации, в 2004 году стихийные бедствия и антропогенные катастрофы по всему миру привели к прямым финансовым убыткам в размере 123 миллиардов долларов США.

В 2005 году в мире произошло 400 катастроф, ущерб от которых составил более 230 миллиардов долларов США.

В 2006 году финансовые убытки от стихийных бедствий и антропогенных катастроф оказались ниже ожидаемых и составили 48 миллиардов долларов США.

Частота стихийных бедствий, которые поддаются измерению, в 1990-х годах была в три раза выше, чем в 1960-х, а финансовые убытки при этом были в девять раз больше.

Также нельзя игнорировать огромные убытки, связанные с маловероятными стихийными бедствиями.

По данным IDC, 55 % компаний, столкнувшихся с бедствиями за последнее десятилетие XX века, не смогли их пережить, 29 % закрылись в течение 2 лет после происшествий из-за потери данных, и только 16 % смогли выжить.

Высокая доступность системы означает, что приложения все еще будут доступны при выходе из строя одного из ее компонентов, независимо от того, связан ли сбой с нарушением работы самой службы, неисправностью на физическом объекте или ошибкой программного/  
аппаратного обеспечения ИТ.

В идеальной ситуации при выходе из строя какого-либо компонента в центре обработки данных пользователь услуг этого центра даже не узнает о сбое. Однако в случае отказа сервера в центре обработки данных переключение соответствующих служб на резервные мощности занимает некоторое время. А значит, пользователям станет известно о сбое.

Ключевой показатель высокой доступности — собственно доступность. Она рассчитывается по следующей формуле: [1 − (время простоя)/(время простоя + время работы)]. Обычно доступность измеряется числом девяток:

4 девятки: 99,99 % = 0,01 % x 365 x 24 x 60 = 52,56 минут/год

5 девяток: 99,999 % = 0,001 % x 365 = 5,265 минут/год

6 девяток: 99,9999 % = 0,0001 % x 365 = 31 секунд/год

Для поддержания высокой доступности обычно используется общее хранилище. Оно позволяет снизить показатель RPO до нуля. Кроме того, часто применяется режим «активный-активный», который обеспечивает показатель RTO почти на уровне нуля. При использовании режима «активный-резервный» следует приложить усилия, чтобы минимизировать показатель RTO.

Для обеспечения высокой доступности серверы необходимо объединить в кластеры, в которых будут развертываться приложения и службы. Системы обеспечения высокой доступности можно разделить на следующие типы:

Системы обеспечения высокой доступности с архитектурой «активный-резервный»

Кластер включает всего два узла (активный и резервный). В этой конфигурации для предоставления служб система использует активные и резервные серверы. Фактически службы предоставляются только активным устройством.

В случае сбоя активного устройства службы запускаются на резервном устройстве для продолжения обслуживания.

Обычно для управления переключением служб между активными и резервными устройствами, а также для предоставления службам виртуального IP-адреса можно использовать программное обеспечение CRM, такое как Pacemaker.

Системы обеспечения высокой доступности с архитектурой «активный-активный»

Кластер из двух активных узлов называется кластером «активный-активный». Кластер из множества активных узлов называется мультиактивным.

В этой конфигурации на всех серверах в кластере выполняются одинаковые нагрузки.

В качестве примера рассмотрим базы данных. При обновлении экземпляра БД данные всех экземпляров будут синхронизироваться.

В этой конфигурации для предоставления службам виртуальных IP-адресов используется программное обеспечение для балансировки нагрузки, такое как HAProxy.

Pacemaker — это диспетчер кластера. В нем используются функции сообщений и участников, предоставляемые используемой инфраструктурой кластера (OpenAIS или heartbeat), для диагностики неисправностей вторичных узлов и систем, что позволяет обеспечить высокую доступность служб кластера (ресурсов кластера).

HAProxy — бесплатное открытое программное обеспечение, написанное на языке C. Оно предоставляет функции обеспечения высокой доступности и балансировки нагрузки, а также прокси-сервер приложений на основе TCP и HTTP. HAProxy особенно подходит для веб-сайтов с высокими нагрузками, где обычно требуются сохранение сеансов.

Аварийная ситуация (происшествие, бедствие) — это непредвиденное событие, вызванное человеческими или естественными факторами, которое приводит к возникновению серьезных сбоев или выходу систем из строя в центре обработки данных. Такие ситуации могут привести к значительному снижению качества обслуживания или его прерыванию. Если в определенный момент степень недоступности системы достигнет определенного уровня, потребуется переключение на резервный объект.

Под аварийным восстановлением понимается возможность восстановления данных, приложений и служб в удаленных центрах обработки данных в случае повреждения систем в рабочем центре вследствие аварийной ситуации.

В рамках мер по аварийному восстановлению в дополнение к рабочему объекту развертывается резервный объект. При возникновении аварийной ситуации и повреждении систем рабочего объекта службы переключаются на резервный объект для поддержания непрерывности обслуживания. Для повышения доступности систем некоторые пользователи даже развертывают несколько резервных объектов.

Основные целевые показатели эффективности систем аварийного восстановления:

RPO (recovery point objective) — определяет максимальную допустимую потерю данных в случае аварийной ситуации.

RTO (recovery time objective) — определяет время, необходимое для восстановления системы.

Чем ниже значения RPO и RTO, тем выше доступность системы. Однако это предполагает высокие затраты для пользователей.

#### 4.2.1.2 Уровень системы аварийного восстановления

Системы аварийного восстановления — важный технический аспект для предприятий, который играет ключевую роль в защите корпоративных данных. В отношении аварийного восстановления многие руководители ИТ-отделов на первый план ставят удаленное аварийное восстановление на уровне приложений. Они также стремятся развернуть системы удаленного аварийного восстановления с нулевой потерей данных и автоматическим переключением приложений на самом высоком уровне. Однако здесь имеет место заблуждение. Важность аварийного восстановления трудно переоценить, однако это не означает, что нужно развертывать системы для восстановления на уровне приложений. Самое главное — выбрать подходящую систему аварийного восстановления в соответствии с актуальными потребностями.

В общем случае системы аварийного восстановления делятся на три уровня: уровень данных, уровень приложений и уровень служб. Уровни данных и приложений относятся к ИТ-системе. Что касается уровня служб, на нем также учитываются некоторые факторы за пределами ИТ-систем, например географическое положение резервного объекта и его персонал.

Аварийное восстановление на уровне данных направлено на защиту от потери или повреждения данных после возникновения аварийной ситуации. Система аварийного восстановления на уровне данных может предполагать ручное сохранение резервных копий на удаленной площадке. Например, можно просто перевозить ленточные накопители с резервными копиями на такую площадку. В передовых решениях для аварийного восстановления используются сетевые инструменты репликации данных для поддержки их асинхронной или синхронной передачи между рабочим центром и центром аварийного восстановления. Например, можно использовать функцию репликации данных дисковых массивов.

Система аварийного восстановления на уровне приложений подразумевает развертывание на резервном объекте хостов и приложений в соответствии с вышеописанным уровнем данных. Такая система включает систему резервного копирования данных, резервную систему обработки данных и резервную сетевую систему. Система аварийного восстановления на уровне приложений позволяет в случае сбоя переключать приложения на резервные мощности. То есть при отказе систем в рабочем центре обработки данных приложения можно будет переключить на ЦОД аварийного восстановления для минимизации простоев системы и поддержки бесперебойности обслуживания.

Компания SHARE, информационная ИТ-организация, созданная IBM в 1955 году, на 78-й конференции в 1992 году представила стандарт аварийного восстановления SHARE 78. Этот стандарт широко признан во всем мире.

В рамках стандарта SHARE 78 план аварийного восстановления можно разделить на восемь уровней в соответствии со следующими аспектами:

Охват стратегии резервного копирования и восстановления

Состояние плана аварийного восстановления

Расстояние между местом развертывания приложений и резервным объектом

Соединение между местом развертывания приложений и резервным объектом

Передача данных между двумя объектами

Допустимая потеря данных

Возможность обновления резервных копий

Возможность запуска задачи резервного копирования с резервного объекта

Системы аварийного восстановления делятся на семь уровней:

Локальное хранение резервных копий

Хранение резервных копий за пределами объекта

Хранение резервных копий за пределами объекта + поддержка горячего резервирования

Копирование и восстановление с использованием сети

Плановое резервное копирование

Действующий центр аварийного восстановления с поддержкой синхронизации

Нулевая потеря данных

Кроме того, стандарт ISO 27001 Международной организации по стандартизации (ISO) требует поддержки хранения соответствующих данных в течение 1–5 лет.

#### 4.2.1.3 Обзор решений Huawei для обеспечения непрерывности бизнеса и аварийного восстановления

Решения Huawei для обеспечения непрерывности бизнеса и аварийного восстановления разработаны для поддержки непрерывной работы предприятий и защиты данных их клиентов. Компания Huawei предлагает четыре основных решения для развертывания системы аварийного восстановления, охватывающей локальный рабочий центр, а также местный и удаленный центры аварийного восстановления. Кроме того, Huawei предоставляет профессиональные консалтинговые услуги в области аварийного восстановления с целью поддержки непрерывности обслуживания и защиты данных.

Локальное решение высокой доступности: обеспечивает высокую доступность ключевых служб в центре обработки данных и предотвращает прерывание обслуживания и потерю данных вследствие отказов отдельных компонентов.

Решение для аварийного восстановления с архитектурой «активный-резервный»: поддержка местного и удаленного центров аварийного восстановления. В случае возникновения аварийной ситуации службы можно быстро переключить на центр аварийного восстановления, чтобы обеспечить бесперебойность обслуживания.

Решение для аварийного восстановления с архитектурой «активный-активный»: нагрузки ключевых служб обрабатываются как в рабочем, так и в местном резервном центре, что позволяет полностью избежать прерывания обслуживания и потери данных в случае сбоя в рабочем центре обработки данных.

Решение для аварийного восстановления с географическим резервированием: обеспечивает защиту в случае отказа всего центра обработки данных или аварии регионального уровня, повышает степень непрерывности обслуживания в отношении критически важных служб. Обычно используется следующая конфигурация: местный центр в режиме «активный-резервный» + удаленный центр в режиме «активный-резервный» или местный центр в режиме «активный-активный» + удаленный центр в режиме «активный-резервный».

### 4.2.2 Архитектура решений для аварийного восстановления

#### 4.2.2.1 Режимы аварийного восстановления

Решение Huawei для аварийного восстановления и резервного копирования предоставляет традиционные функции аварийного восстановления и резервного копирования уровней обычных и облачных центров обработки данных. Оно поддерживает аварийное восстановление и резервное копирование на уровне приложений и данных, а также механизмы защиты данных на уровне приложений. Кроме того, доступны системы аварийного восстановления для приложений разного уровня, позволяющие обеспечить баланс между требованиями к аварийному восстановлению и общими затратами. Решение Huawei для аварийного восстановления и резервного копирования соответствует стратегиям обслуживания и развития наших клиентов и включает профессиональные услуги, в том числе стратегический консалтинг, планирование исследований и разработок, внедрение и непрерывное управление эксплуатацией.

Функции удаленной репликации в основном используются для резервного копирования данных и аварийного восстановления. Различные режимы удаленной репликации подходят для разных ситуаций.

Синхронная удаленная репликация применяется в сценариях резервного копирования и аварийного восстановления, когда первичный и вторичный объекты расположены недалеко друг от друга, например в одном городе.

Асинхронная удаленная репликация применяется в сценариях резервного копирования и аварийного восстановления, когда пропускная способность сети ограничена или первичный объект находится на значительном удалении от вторичного (например, в другой стране или другом регионе).

В зависимости от приложения может потребоваться учитывать расстояние и режим удаленной репликации. Эта функция применяется в сценариях централизованной резервной копии и площадок для аварийного восстановления, а также в сценариях площадок для аварийного восстановления с архитектурой «активный-активный».

Система аварийного восстановления с архитектурой «активный-резервный» предполагает развертывание центра аварийного восстановления на отдельном объекте в дополнение к рабочему центру. Такой центр аварийного восстановления обеспечивает защиту данных на уровне данных или приложений за счет создания их полных копий. По сравнению с более сложными конфигурациями, такими как географическое резервирование и централизованное аварийное восстановление, аварийное восстановление с архитектурой «активный-резервный» является наиболее широко используемым режимом.

Решение для аварийного восстановления с архитектурой «активный-резервный» полезно в двух типах сценариев: когда на обоих объектах (первичном и вторичном) применяются СХД Huawei или когда на рабочем объекте развернута система от поставщика аналогичного уровня.

При использовании решения Huawei с архитектурой «активный-активный» оба центра обработки данных одновременно поддерживают предоставление услуг, что повышает производительность ЦОД и эффективность использования ресурсов систем.

В настоящее время центры обработки данных поддерживают работу в режиме «активный-резервный» или в режиме «активный-активный».

В режиме «активный-резервный» одна часть служб в основном обрабатывается в ЦОД A, в то время как ЦОД B в их отношении находится в режиме горячего резерва, и наоборот. Это позволяет в какой-то степени добиться эффекта режима «активный-активный».

В режиме «активный-активный» все запросы ввода/вывода могут обрабатываться соответствующими LUN обоих ЦОД. Это позволяет сбалансировать нагрузку и обеспечить незаметное переключение на резерв при сбое.

В решении Huawei для аварийного восстановления используется архитектура «активный-активный», ведущие в отрасли технологии HyperMetro, а также устройства для поддержки сетевых функций, кластеров баз данных, балансировки нагрузки и передачи данных. Это сочетание позволяет предоставить клиентам комплексное решение для центров обработки данных, расположенных в пределах 100 км друг от друга, обеспечивающее переключение услуг и непрерывность обслуживания даже в случае отказа устройства или всего центра обработки данных.

#### 4.2.2.2 Развитие новых режимов аварийного восстановления в среде облачных вычислений

Чтобы помочь корпоративным клиентам развернуть высококачественные ИТ-системы и удовлетворить требования в отношении развития услуг, компания Huawei предоставляет профессиональные услуги в областях хранения данных, облачных вычислений и серверов на базе ИТ-продуктов.

В отношении хранения данных Huawei предоставляет профессиональные услуги по миграции данных, аварийному восстановлению, резервному копированию и виртуализации, способные удовлетворить требования корпоративных клиентов к модернизации хранилищ, защите данных и унифицированному управлению хранилищами.

В отношении облачных вычислений Huawei предоставляет профессиональные услуги планирования и проектирования облачных систем, внедрения решений FusionSphere и решений FusionCloud для рабочих столов, миграции служб FusionSphere, планирования и внедрения систем обработки больших данных. Эти услуги позволят удовлетворить требования корпоративных клиентов в отношении планирования, проектирования и внедрения систем виртуализации, миграции данных, а также планирования и внедрения систем обработки больших данных.

На уровне центра обработки данных Huawei предоставляет профессиональные услуги по консолидации центров обработки данных для удовлетворения требований клиентов к уровням ЦОД L1 и L2, защите данных, планированию ЦОД и миграции данных. L1 — это уровень физической инфраструктуры, включающий планировку этажей, системы питания, охлаждения, пожаротушения, физической безопасности, а также кабельную систему. L2 — это уровень ИТ-инфраструктуры на основе системы облачных вычислений, который включает подсистемы вычислений, сети, СХД, безопасности, поддержания непрерывности обслуживания, аварийного восстановления и резервного копирования.

Корпоративная служба аварийного восстановления и резервного копирования Huawei включает несколько сервисных продуктов, в том числе миграцию данных хранилища, аварийное восстановление, резервное копирование, миграцию виртуальных машин и услуги по внедрению облачных решений. Эти услуги применимы в различных отраслях, таких как государственный сектор, энергетика, финансы и образование. Huawei предлагает профессиональные сервисные решения для конкретных отраслей, позволяющие удовлетворить потребности корпоративных клиентов в обновлении ИТ-инфраструктуры, защите данных и трансформации технологий.

Основываясь на требованиях клиентов и жизненном цикле услуг, мы добавили в корпоративную службу аварийного восстановления и резервного копирования Huawei универсальные профессиональные услуги, включая управление проектами, планирование, проектирование, интеграционное тестирование, собственно интеграцию, проверку интеграции и оптимизацию.

Кроме того, мы используем разнообразные профессиональные инструменты для быстрого сбора и анализа проектной информации, поддержки разработки и внедрения решений, а также настройки и предоставления наиболее подходящих профессиональных услуг для наших клиентов.

### 4.2.3 Основные технологии аварийного восстановления

#### 4.2.3.1 Технология аварийного восстановления на уровне хоста

**Технология аварийного восстановления на уровне хоста**

Для поддержки удаленной репликации на серверах в рабочем центре и центре аварийного восстановления устанавливается специальное программное обеспечение для репликации данных, такое как ПО для репликации томов. При этом между двумя центрами должно быть установлено сетевое соединение. На сервере также можно развернуть программное обеспечение для удаленного переключения приложений, чтобы создать комплексное решение для аварийного восстановления на уровне приложений.

Этот режим репликации данных требует меньших затрат, которые в основном будут связаны с приобретением программного обеспечения. Поддерживается совместимость с серверами и устройствами хранения разных марок, что будет удобно для пользователей, использующих разнородное оборудование. Однако в этом режиме синхронизация на сервере выполняется с помощью программного обеспечения, что создает высокую нагрузку на хост и сеть.

#### 4.2.3.2 Технология аварийного восстановления на уровне сети

**Технология аварийного восстановления на уровне сети**

Технология репликации данных на уровне сети SAN предполагает добавление шлюзов хранения между внешними серверами приложений и внутренними системами хранения. Такой шлюз подключается к серверам и устройствам хранения.

Шлюзы позволяют настроить зеркальную связь между двумя томами на разных устройствах хранения. При этом данные, записываемые на основной том, будут одновременно записываться на резервный том.

Для поддержания бесперебойности обслуживания в случае отказа основного устройства службы переключаются на резервное, а резервный том активируется.

Принцип работы:

Хост в рабочем центре передает данные на локальный шлюз виртуализации.

Этот шлюз записывает данные в локальный том журналов.

После успешного завершения записи в том журналов шлюз виртуализации в рабочем центре отправляет на локальный хост соответствующее подтверждение.

Затем шлюз виртуализации в рабочем центре записывает данные в локальный рабочий том и отправляет запрос на запись данных в шлюз виртуализации на стороне резервного устройства.

После получения такого запроса шлюз виртуализации на стороне резервного устройства отправляет соответствующее подтверждение на шлюз в рабочем центре.

Затем шлюз виртуализации на стороне резервного устройства записывает данные в резервный том.

После успешной записи данных в резервный том в центре аварийного восстановления шлюз виртуализации на стороне этого центра отправляет соответствующее подтверждение на шлюз в рабочем центре.

#### 4.2.3.3 Технология аварийного восстановления на уровне хранения

**Технология аварийного восстановления на уровне хранения**

В рамках аварийного восстановления на уровне хранения используется технология репликации данных локального массива хранения для создания доступной копии данных в резервном массиве хранения. Для поддержания бесперебойности обслуживания в случае отказа основного массива хранения службы можно быстро переключить на резервный массив.

**Принципы синхронной репликации SAN**

Процесс синхронизации:

Рабочая СХД получает от узла запрос на запись. Система HyperReplication регистрирует этот запрос в журналах. При этом HyperReplication записывает только информацию об адресах.

HyperReplication передает запрос в первичный и вторичный LUN. Обычно LUN используют политику отложенной записи, и данные сначала записываются в кэш.

Система HyperReplication ожидает сообщения от первичного и вторичного LUN о результатах записи. В случае успешной записи как в первичный, так и во вторичный LUN HyperReplication удаляет запрос из журналов. В противном случае запрос сохраняется со статусом сбоя соединения. При последующей синхронизации HyperReplication копирует блоки данных в соответствии с информацией об адресах, записанной в журналах.

HyperReplication возвращает на узел результат записи в первичный LUN.

Разделение:

В этом режиме записываемые рабочим хостом данные сохраняются только на первичном LUN, а разница в данных между первичным и вторичным LUN регистрируется в соответствующем журнале. Для согласования данных между первичным и вторичным LUN можно вручную запустить процесс синхронизации. В рамках этого процесса будет выполнено инкрементное копирование блоков данных, помеченных в соответствующем журнале как различающиеся, с первичного на вторичный LUN. Обработка операций ввода-вывода выполняется аналогично первоначальной синхронизации.

**Принципы асинхронной репликации SAN**

Временной сегмент — это логическое пространство в кэше для записи данных за период времени (без ограничения на объем данных).

При низких значениях показателя RPO удаленная асинхронная репликация не занимает много времени. Все данные могут храниться в кэше системы хранения OceanStor в нескольких временных сегментах. Однако если возникает проблема с пропускной способностью хоста или резервной системы, которая приводит к увеличению времени репликации или ее прерыванию, то в целях обеспечения согласованности данные из кэша автоматически записываются на диски в соответствии с политикой сбрасывания данных на диск. При репликации данные считываются с дисков.

В зависимости от заданного пользователем периода репликации (от 3 секунд до 1440 минут), система автоматически запускает процедуру синхронизации для инкрементной синхронизации данных основного и резервного объектов (если настроена ручная синхронизация, пользователям потребуется запустить ее вручную). В момент начала периода репликации система создает в кэше основного LUN (LUN A) и резервного LUN (LUN B) временные сегменты TPN+1 и TPX+1 соответственно.

Основная система получает от рабочего хоста запрос на запись.

Основная система записывает соответствующие данные во временной сегмент TPN+1 в кэше LUN A и возвращает хосту соответствующее подтверждение.

В процессе синхронизации система считывает данные, сгенерированные за предыдущий период репликации из временного сегмента TPN в кэше LUN A, передает их в резервную систему и записывает данные во временной сегмент TPX+1 в кэше LUN B. После заполнения кэша LUN A до определенного порога система автоматически начинает записывать данные на диски. В этом случае на дисках создается моментальный снимок данных во временном сегменте TPN. В процессе синхронизации система считывает данные моментального снимка на дисках и копирует их в LUN B.

После завершения синхронизации система записывает данные из временных сегментов TPN и TPX+1 в кэше LUN A и LUN B на диски в соответствии с политикой сбрасывания данных на диски (моментальные снимки удаляются автоматически), после чего ожидает начала следующего периода репликации.

Переключение:

Если пара синхронной удаленной репликации находится в нормальном состоянии, в ней можно выполнить переключение между первичным и вторичным LUN.

В состоянии разделения переключение между первичным и вторичным LUN можно выполнить только в том случае, если для вторичного LUN настроена возможность записи.

Асинхронная удаленная репликация выполняется в состоянии разделения.

В этом состоянии для вторичного LUN необходимо настроить возможность записи.

**Принципы асинхронной репликации NAS**

В начале каждого периода асинхронной удаленной репликации файловой системы создается моментальный снимок основной файловой системы. С учетом добавочной информации, полученной с момента завершения предыдущего периода репликации, система считывает данные моментального снимка и копирует их в резервную файловую систему. После завершения инкрементной репликации данные в резервной файловой системе будут полностью соответствовать данным основной файловой системы. При этом в резервной файловой системе будут созданы точки согласованности данных.

Поддерживается удаленная репликация данных файловых систем. Репликация отдельных каталогов или файлов не поддерживается.

Одна файловая система может относиться только к одной задаче репликации, но каждая задача репликации может обрабатывать несколько файловых систем.

В отношении файловых систем поддерживается только полная репликация. Не допускается указывать файловую систему одновременно в качестве источника и цели репликации. Каскадная репликация и возможность использования трех ЦОД не поддерживаются.

Минимальная единица инкрементной репликации равна размеру блока файловой системы (от 4 до 64 КБ). Минимальный период синхронизации при асинхронной репликации составляет 5 минут.

Поддерживается возобновляемая загрузка.

### 4.2.4 Сценарии применения решений для аварийного восстановления

#### 4.2.4.1 Виртуальное аварийное восстановление

Проблемы для клиентов

Клиент использует виртуальный центр обработки данных vSphere, а теперь хочет развернуть новый центр обработки данных для аварийного восстановления.

Требуется обеспечить низкую совокупную стоимость владения и высокую окупаемость инвестиций (ROI).

Решение Huawei

В центре аварийного восстановления развернута ИТ-система, в том числе устройства хранения, службы, сети и платформы виртуализации.

В рабочем центре, а также в центре аварийного восстановления установлена система поддержки аварийного восстановления UltraVR от Huawei.

Для обеспечения защиты виртуальных машин на уровне приложений на соответствующих хост-компьютерах установлены агенты ConsistentAgent.

Преимущества для клиента

Не требуется реконструкция архитектуры действующей сети

Гибкая настройка политик аварийного восстановления и восстановление одним нажатием

Возможность испытания системы аварийного восстановления и обратного переключения

#### 4.2.4.2 Решение для аварийного восстановления на уровне приложений

Проблемы для клиентов

Текущая ИТ-система клиента не способна удовлетворить требования в отношении развития услуг и обеспечить непрерывность предоставления онлайн-услуг.

Процессы эксплуатации и технического обслуживания ИТ-систем затруднены, требуют больших затрат энергии и обуславливают низкую эффективность использования ресурсов систем.

Решение Huawei

Миграция системы обслуживания на облачную платформу Huawei.

Развертывание устройств и программного обеспечения для непрерывной защиты данных в двух центрах обработки данных. Развертывание в двух ЦОД в одном городе системы аварийного восстановления на уровне приложений на базе технологий непрерывной защиты данных.

Преимущества для клиента

Гибкое выделение ресурсов и возможность их повторного использования, повышение эффективности использования ресурсов и сокращение затрат на эксплуатацию и техническое обслуживание.

Снижение до нуля показателей RTO и RPO для ключевых служб. Для обеспечения бесперебойности обслуживания в случае отказа рабочего центра службы и данные можно переключить на центр аварийного восстановления.

|  |
| --- |
| 5 Управление эксплуатацией и техническим обслуживанием систем хранения данных |
|  |

## 5.1 Управление эксплуатацией и техническим обслуживанием систем хранения данных

### 5.1.1 Обзор управления системами хранения данных

DeviceManager — это часть интегрированного программного обеспечения для управления СХД, разработанного Huawei. Это ПО предварительно устанавливается на системы хранения. DeviceManager поддерживает доступ через веб-браузер или с планшета.

Интерфейс командной строки СХД позволяет запрашивать состояние СХД, выполнять ее настройку и обслуживание, а также управлять ею. Для входа в интерфейс командной строки можно использовать клиент PuTTY, установленный на любом терминале обслуживания, подключенном к системе хранения. Он обеспечит доступ к IP-адресу сетевого порта управления на контроллере системы хранения по протоколу SSH. Протокол SSH поддерживает два режима аутентификации: по имени и паролю пользователя, а также по открытому ключу.

Поддерживаются следующие способы входа в систему:

**Вход в систему с использованием последовательного порта**

При подключении контроллерной полки к терминалу обслуживания через последовательные кабели для входа в интерфейс командной строки СХД можно использовать установленную на терминал программу (например, PuTTY).

**Вход в систему с использованием сетевого порта управления**

Выполнить вход в интерфейс командной строки СХД можно с использованием адресов IPv4 или IPv6.

При подключении контроллерной полки к терминалу обслуживания через сетевые кабели для входа в интерфейс СХД можно использовать любое программное обеспечение для удаленного входа в систему, поддерживающее протокол SSH.

В контроллерных полках высотой 2U по умолчанию используются следующие IP-адреса сетевых портов управления: 192.168.128.101 для контроллера A и 192.168.128.102 для контроллера B. Маска подсети по умолчанию: 255.255.0.0. В контроллерных полках высотой 3U/6U по умолчанию используются следующие IP-адреса сетевых портов управления: 192.168.128.101 для модуля управления 0 и 192.168.128.102 для модуля управления 1. Маска подсети по умолчанию: 255.255.0.0.

IP-адрес сетевого порта управления контроллерной полки должен принадлежать к тому же сегменту сети, что и терминал обслуживания. В противном случае потребуется изменить IP-адрес сетевого порта управления через последовательный порт, выполнив команду **change system management\_ip**.

### 5.1.2 Общая информация об инструментах управления системами хранения данных

DeviceManager — это интегрированное программное обеспечение для управления СХД, разработанное Huawei и предназначенное для сценариев с одной системой хранения. DeviceManager облегчит настройку и обслуживание устройств хранения, а также управление ими.

Пользователи могут запрашивать состояние СХД, выполнять ее настройку и обслуживание, а также управлять ею с помощью DeviceManager и интерфейса командной строки. Такие инструменты, как SmartKit и eService, повышают эффективность эксплуатации и технического обслуживания системы.

Перед использованием DeviceManager убедитесь, что терминал обслуживания соответствует следующим требованиям:

DeviceManager поддерживает используемые версии операционной системы и браузера терминала.

DeviceManager поддерживает различные операционные системы и браузеры. Для получения подробной информации о совместимости используйте инструмент Huawei Storage Interoperability Navigator.

Терминал обслуживания надлежащим образом подключен к СХД.

Суперпользователь может выполнить вход в интерфейс СХД только с использованием этого режима аутентификации.

Перед входом в DeviceManager в качестве пользователя домена LDAP необходимо настроить сервер домена LDAP, а затем добавить его в домен LDAP в параметрах системы хранения и создать пользователя домена LDAP.

По умолчанию DeviceManager поддерживает одновременный вход в систему 32 пользователей.

В СХД доступны роли пользователей по умолчанию, а также возможность настраивать пользовательские роли.

Для ролей по умолчанию в системе предварительно определены права доступа, указанные в таблице. К ролям по умолчанию относятся следующие: суперпользователь, администратор и пользователь с доступом только на чтение.

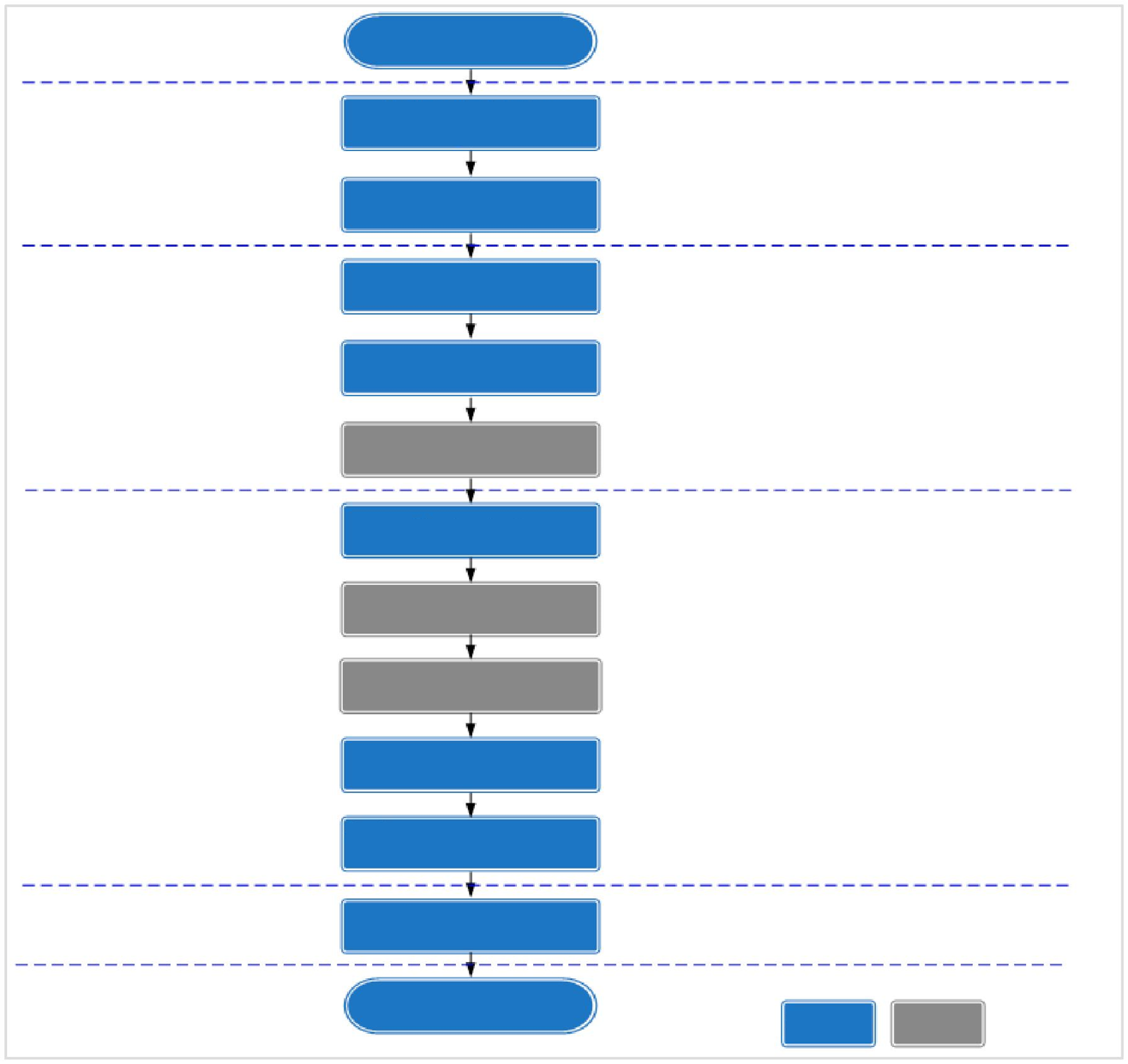
Права доступа для пользовательских ролей можно настроить в соответствии с фактическими требованиями.

Для обеспечения контроля прав доступа при использовании нескольких клиентов система хранения делит роли по умолчанию на две группы: системная группа и группа клиентов. Эти группы имеют следующие различия:

Группа клиентов: роли из этой группы используются только в представлении клиента (управление этим представлением становится доступным после входа в DeviceManager с использованием учетной записи клиента).

Системная группа: роли из этой группы используются только в системном представлении (управление этим представлением становится доступным после входа в DeviceManager с использованием системной учетной записи).

### 5.1.3 Основные операции управления



Настройка подключения между хостом и СХД

Создание хоста

Проверка перед настройкой

Использование пространства хранения сервером приложений

Обязательно

Необязательно

Конец

Сопоставление

Создание групп портов

Создание групп хостов

Создание групп LUN

Создание LUN

Создание пула хранения

Вход в DeviceManager

Начало

**4. Использование пространства хранения.**

**3. Настройка сопоставления и подключения хоста.**

**2. Создание пространства хранения.**

**1. Подготовка к настройке.**

Если приложение развернуто в кластере из нескольких хостов, эти хосты будут обращаться к данным приложения одновременно. В этой ситуации хосты можно объединить в группу.

Если для связи между системой хранения и сервером приложений требуется использовать определенные порты, можно создать группу портов.

Если данные приложения хранятся в нескольких LUN, эти LUN можно объединить в группу для упрощения управления ими.

##### Рис. 5-1. Процесс настройки

## 5.2 Управление эксплуатацией и техническим обслуживанием систем хранения данных

### 5.2.1 Обзор процессов эксплуатации и технического обслуживания (ЭиТО)

ITIL

Библиотека инфраструктуры информационных технологий (Information Technology Infrastructure Library, ITIL) — это признанный во всем мире набор практических рекомендаций по эффективному управлению ИТ-услугами. С 1980 года Министерство государственной торговли Соединенного Королевства постепенно разрабатывало и улучшало набор методов оценки качества ИТ-услуг, который получил название ITIL. Этот набор был предназначен для решения проблемы низкого качества ИТ-услуг. В 2001 году на форуме IT Service Management Forum (itSMF) Британский институт стандартов официально выпустил национальный стандарт BS15000, в основе которого лежала библиотека ITIL. Это стало историческим событием для сферы управления ИТ-услугами.

Если раньше ИТ-услуги играли лишь вспомогательную роль, то теперь они окончательно стали отдельным типом обслуживания. ITIL вызвала жаркие обсуждения по всему миру, направленные на достижение целей сокращения затрат, повышения производительности и улучшения качества обслуживания. Множество известных международных компаний, таких как IBM, HP, Microsoft, P&G и HSBC, принимают активное участие в разработке ITIL. По мере того как отрасль постепенно переводит основное внимание с технологий на услуги, требования предприятий к управлению ИТ-услугами также растут, что в значительной степени помогает стандартизировать ИТ-процессы, поддерживать их соответствие требованиям бизнеса и повышать эффективность обработки.

ITIL широко признана в Великобритании, других странах Европы, Северной Америки, Новой Зеландии и Австралии. Использование предприятием методов ITIL рассматривается как ключевой показатель при определении того, допускается ли поставщик услуг или подрядчик по аутсорсингу к участию в торгах.

### 5.2.2 Инструмент управления эксплуатацией и техническим обслуживанием

Для поддержки эксплуатации и технического обслуживания систем хранения используются следующие инструменты:

DeviceManager: программное обеспечение для эксплуатации и технического обслуживания отдельных устройств.

SmartKit: профессиональный инструмент для инженеров технической поддержки Huawei, поддерживающий оценку совместимости, планирование и проектирование, сбор информации об ошибках одним нажатием, проверку, обновление и замену легкосъемных модулей (FRU).

eSight: ориентированный на клиентов пакет ПО, предназначенный для обслуживания нескольких устройств и поддерживающий мониторинг неисправностей, а также визуализацию процессов ЭиТО.

DME: ориентированное на клиентов программное обеспечение, поддерживающее унифицированное управление ресурсами хранения, организацию каталогов услуг и предоставление услуг хранения и обработки данных по запросу.

Клиент eService: ПО, устанавливаемое в среде клиента. Позволяет в реальном времени выявлять ошибки устройств хранения и уведомлять о них центр обслуживания Huawei.

Облачная платформа eService: развертывается в центре обслуживания Huawei с целью мониторинга устройств во всей сети в режиме реального времени. Позволяет перейти от реактивного обслуживания к проактивному и создать систему обслуживания на основе агентов.

### 5.2.3 Примеры процессов эксплуатации и технического обслуживания

Обзор операций эксплуатации и технического обслуживания

Системный администратор может выполнять проверку среды и состояния устройств (конкретные операции и периоды их выполнения указаны в таблицах ниже). В случае возникновения ошибки системный администратор может своевременно обработать ее и провести обслуживание устройства для обеспечения непрерывной работы системы хранения в штатном режиме.

Операции первоначального обслуживания

| **Операция** | **Необходимые действия** |
| --- | --- |
| Проверка установки SmartKit | Проверьте корректность установки SmartKit и соответствующих дополнительных инструментов с помощью терминала обслуживания. Дополнительные инструменты предоставляют следующие функции:  Сбор архивных данных устройств  Сбор информации  Анализ работоспособности диска  Проверка  Установка обновлений |
| Проверка установки и конфигурации eService | Проверьте корректность установки инструмента eService и настройки политики аварийных сигналов с помощью терминала обслуживания. |
| Проверка конфигурации политики аварийных сигналов | Проверьте, настроена ли политика аварийных сигналов должным образом с помощью DeviceManager. После настройки политики аварийных сигналов сообщения о таких сигналах будут отправляться на сервер или мобильный телефон клиента для обеспечения их своевременной обработки. Политика аварийных сигналов включает следующие элементы:  Уведомления по электронной почте  Уведомления в SMS-сообщениях  Системные уведомления  Дамп аварийных сигналов  Управление IP-адресом для системного прерывания  Управление пользователями USM (унифицированный диспетчер сеансов связи)  Маскирование аварийных сигналов  Уведомления в системном журнале |

Ежедневные операции обслуживания

Необходимо проверять наличие аварийных сигналов и обрабатывать их. Войдите в DeviceManager или используйте выбранный режим оповещения об аварийных сигналах для их просмотра и своевременной обработки в соответствии с рекомендациями.

Еженедельные операции обслуживания

| **Операция** | **Необходимые действия** |
| --- | --- |
| Проверка устройств хранения | Используйте инструмент проверки SmartKit на терминале обслуживания. Проверка включает следующие элементы:   * Проверка состояния оборудования * Проверка состояния программного обеспечения * Проверка состояния дополнительных служб * Проверка аварийных сигналов   **Примечание:**  Если рекомендации SmartKit не помогают решить проблему, используйте SmartKit для сбора соответствующей информации и связи с технической поддержкой Huawei. |
| Проверка среды в аппаратной | Проверьте состояние среды в аппаратном помещении в соответствии с надлежащими методами проверки.  **Примечание:**  Если среда не соответствует требованиям, скорректируйте ее соответствующим образом. |
| Проверка внутренней среды стойки | Проверьте состояние внутренней среды стойки.  **Примечание:**  Если среда не соответствует требованиям, скорректируйте ее соответствующим образом. |

Сбор информации

В рамках ЭиТО собирается основная информация о системе, а также информация о неисправностях, об устройствах хранения, о сети и сервере приложений.

| **Тип информации** | **Элемент** | **Описание** |
| --- | --- | --- |
| Основная информация | Серийный номер и версия устройства | Серийный номер и версия устройства хранения.  **Примечание:**  Получить информацию о серийном номере и версии устройства хранения можно в разделе «**Basic Information**» (Основная информация) в DeviceManager. |
| Информация о клиенте | Контактная информация клиента. |
| Информация о неисправностях | Время возникновения неисправности | Время возникновения неисправности. |
| Признак | Признак возникновения неисправности, например диалоговое окно с сообщением об ошибке или полученное уведомление. |
| Операции, выполнявшиеся непосредственно перед возникновением неисправности | Данные об операциях, выполнявшихся непосредственно перед возникновением неисправности. |
| Операции, выполнявшиеся после возникновения неисправности | Данные об операциях, выполнявшихся с момента возникновения неисправности вплоть до сообщения о ней обслуживающему персоналу. |
| Информация об устройстве хранения | Конфигурация аппаратных модулей | Информация о конфигурации оборудования устройства хранения. |
| Состояние индикаторов | Информация о состоянии индикаторов устройства хранения (особенно об оранжевых и красных).  Подробную информацию о состоянии индикаторов каждого компонента устройства хранения см. в «*Описании*» соответствующей модели. |
| Данные системы хранения | Вручную выполните экспорт текущих параметров СХД и системных журналов. |
| Аварийные сигналы и журналы | Вручную выполните экспорт данных об аварийных сигналах, а также журналов устройства хранения. |
| Информация о сети | Режим подключения | Тип подключения сервера приложений к устройству хранения, например Fibre Channel или iSCSI. |
| Модель коммутатора | Модель коммутатора, если таковой используется в сети. |
| Данные диагностики коммутатора | Вручную выполните экспорт данных диагностики рабочего коммутатора, включая его исходную и текущую конфигурации, информацию об интерфейсах, время и версию системы. |
| Топология сети | Данные о топологии сети или топологии соединения между сервером приложений и устройством хранения. |
| IP-адрес | Правила использования IP-адресов или список их предоставления, если сервер приложений подключен к устройству хранения через сеть iSCSI. |
| Информация о сервере приложений | Версия операционной системы | Тип и версия операционной системы, под управлением которой работает сервер приложений. |
| Скорость порта | Скорость портов сервера приложений, подключенного к устройству хранения. Подробную информацию об оценке скорости портов см. в «*Онлайн-справке*». |
| Журнал ОС | Просмотр и экспорт журнала ОС. |