Реализация кластерных технологий средствами Huawei

Булыгин О., Модуль-Проекты

一(yat). Пример реализации облачных вычислений в МФТИ (ГУ):

32 узла 2x2630v3/256GB/10GbE/8GFC в 12U E9000v3

и отказоустойчивая кеширующая СХД OceanStor 5500, выполняющая роль шлюза 10GbE/FC

二 (yee). Создание линейно масштабируемых программно- управляемых систем хранения данных (FusionStorage/Ceph)

三 (sam). Аппаратные средства Huawei для организации отказоустойчивых решений

Почему Huawei?

- Авторитет и рекомендации в области телекоммуникации более 20 лет
- Практически полный портфель собственных решений для ИТ (инфраструктура, вычисления, хранение, сети, управление…)
- Перенос технологий 9999 во все охваченные отрасли
- Высочайший уровень разработок и сопровождения
- Предсказуемость в партнерстве с РФ
- Впечатляющие результаты 2014…

Почему Huawei?

Выучить китайский просто: 一 二 三 (yat, yee, sam)…

Постановка задачи

- Минимизация затрат по хранению 500+ТБ данных
- Линейное масштабирование по стоимости, объему хранения, производительности
- Самовосстановление, высокая надежность автобалансировка отсутствие единой точки отказа
- Легкая миграция при замене платформы
- Относительная независимость от производителя и его условий сопровождения

- Предлагается решение на базе серверов Huawei 2288v3 и ПО Ceph/ FusionSorage

Описание решения

- **7* Huawei RH2288v3,** каждый:
- **2U** шасси
- 2* iE5-2630v3 (8*2.4GHz)
- 8* 16GB (до 24* 32GB)
- 2* 750W platinum БП
- 2* 2* 10GE
- 2* 2* 40GE (или 100GE/100IB/56IB)
- 2* 16* 12G SAS HBA
- 1* 2* 128GB (32GB) SATA DOM
- 1* 8GB UDisk
- 2* SSD (600/960GB PCIe или 2.5")
- 12* 3.5" HDD (6TB или 8ТБ)

Коммутация:

- **Huawei CE8860-4C-EI**

  (32* 100GE QSFP28/ 64*QSFP+,

  6.4Tb/s, 2976Mpps)
- **Huawei CE7850-32Q-EI**

  (32* 40GE QSFP+, 2.56Tb/s, 1440Mpps)
- Mellanox SB7800

  (36* EDR IB, 7Tb/s, 7020Mpps, 90ns)
- Mellanox SX6036/6025

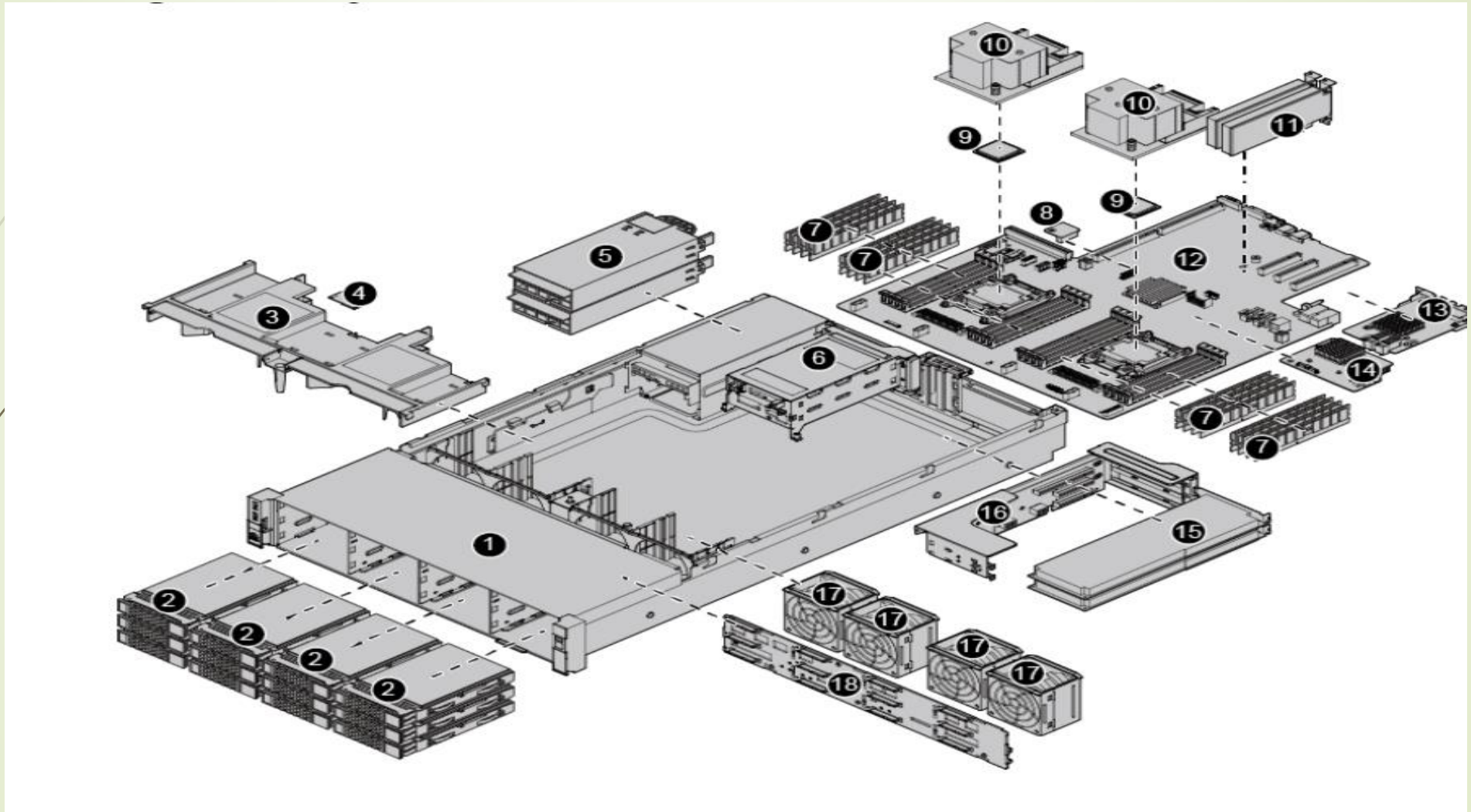
  (36* FDR IB, 4Tb/s, 200ns)

ПО:

- Red Hat Ceph Storage, Premium (до 256TB, 12* физ.узлов, RS00036 до 512TB, 25* физ.узлов, RS00037)
- Huawei FusionSorage

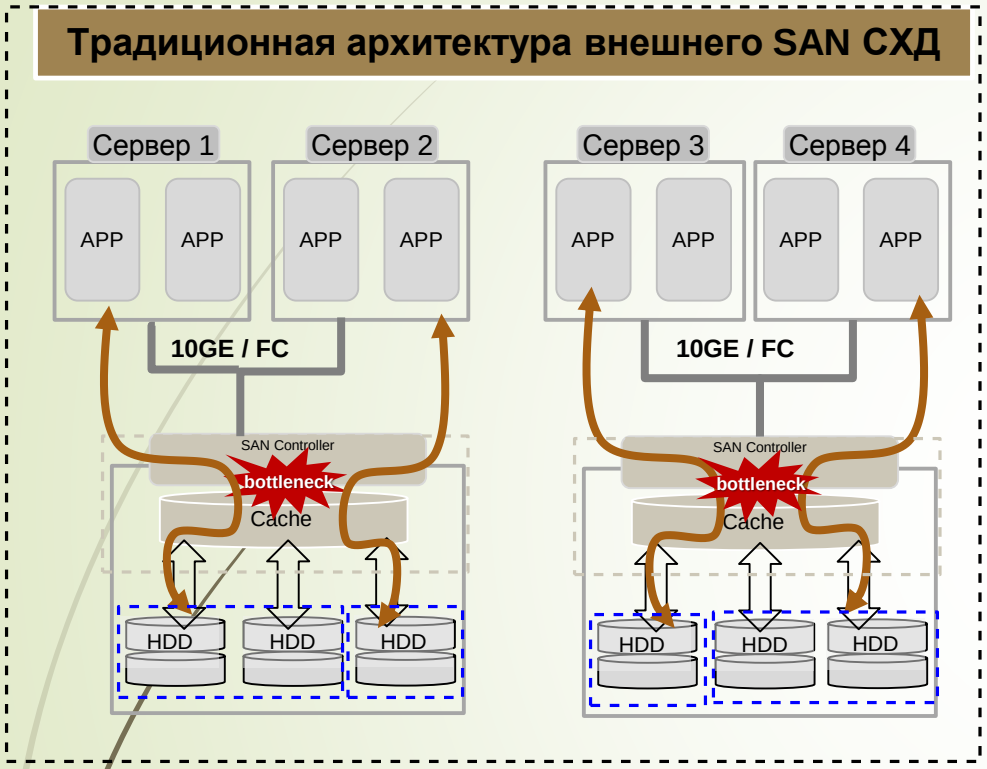
Huawei RH2288v3

структура

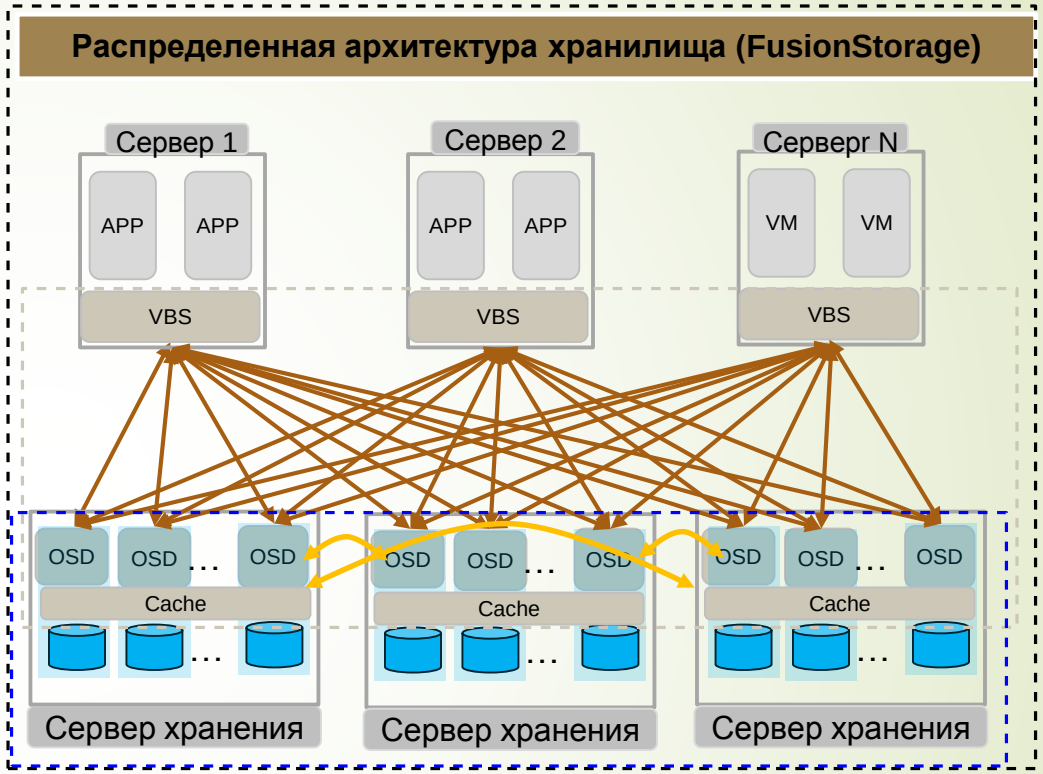


Основные принципы

Эластичная архитектура



VS.

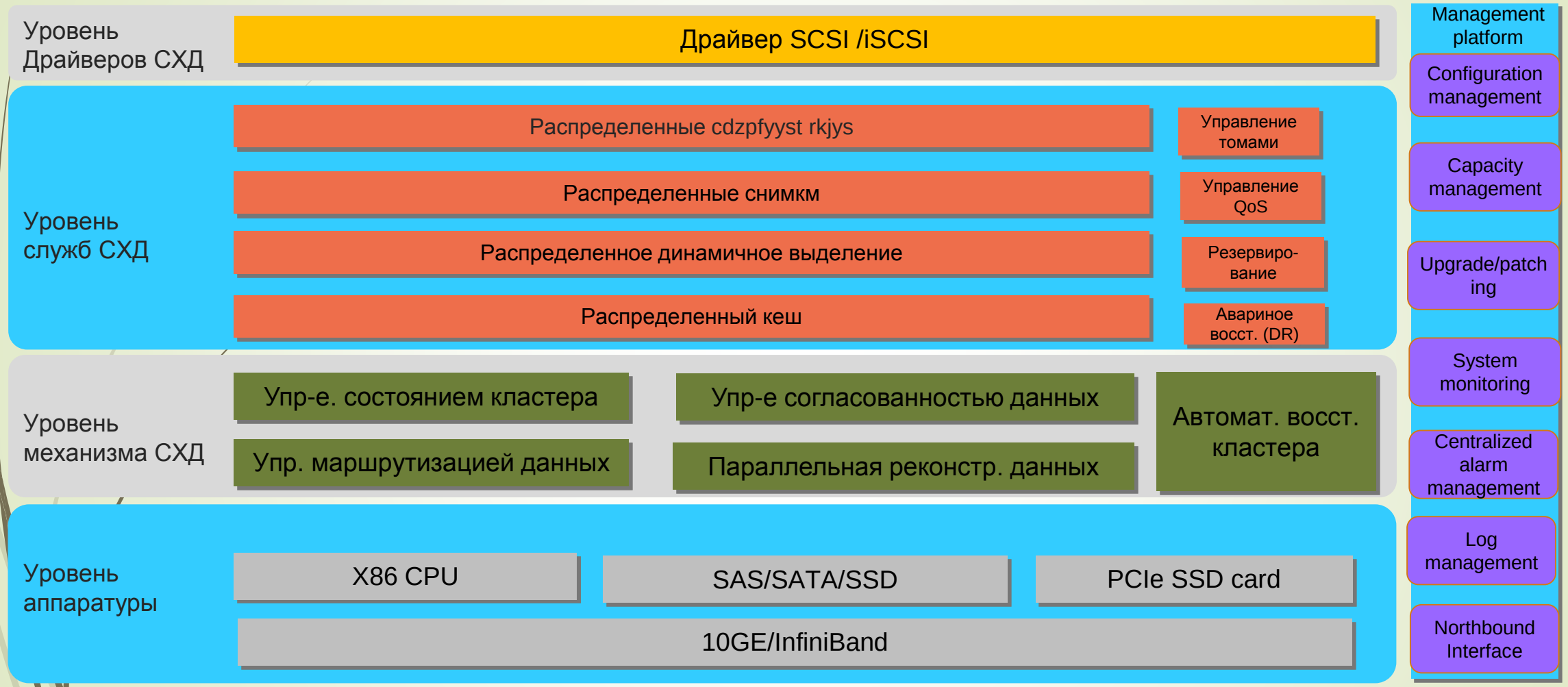


- Существуют узкие места I/O в контроллерах хранилища и в кеше, отсутствует масштабирование
- Приложения связаны только с частью дисков, как результат – низкое распараллеливание I/O и плохая эластичность.
- Нагрузка служб балансируется между группами RAID, которые имеют низкий уровень использования и сложен в планировании.

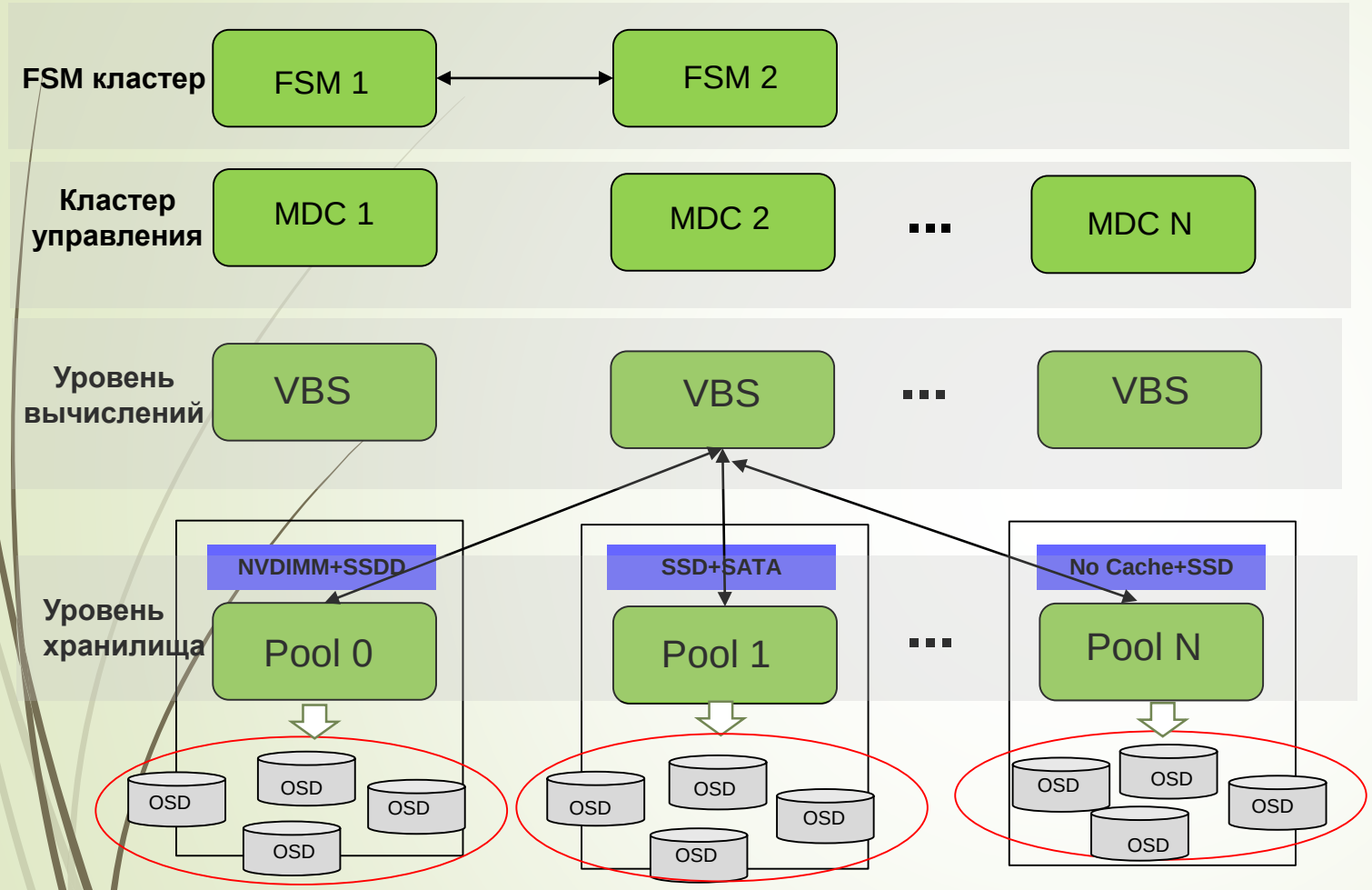
- Распределенные, контроллеры кеш и выровненная сетевая среда исключают узкие места I/O.
- Приложения связаны со всеми дисками в пуле, гарантируя высокий параллелизм I/O и лучшую эластичность. Взрывной рост MBPS в 3-5 раз.
- Сверхкрупные ресурсы пулов гарантируют балансировку нагрузки, повышая использование ресурсов и облегчая планирование.

HUAWEI

Архитектура ПО распределенного СХД FusionStorage

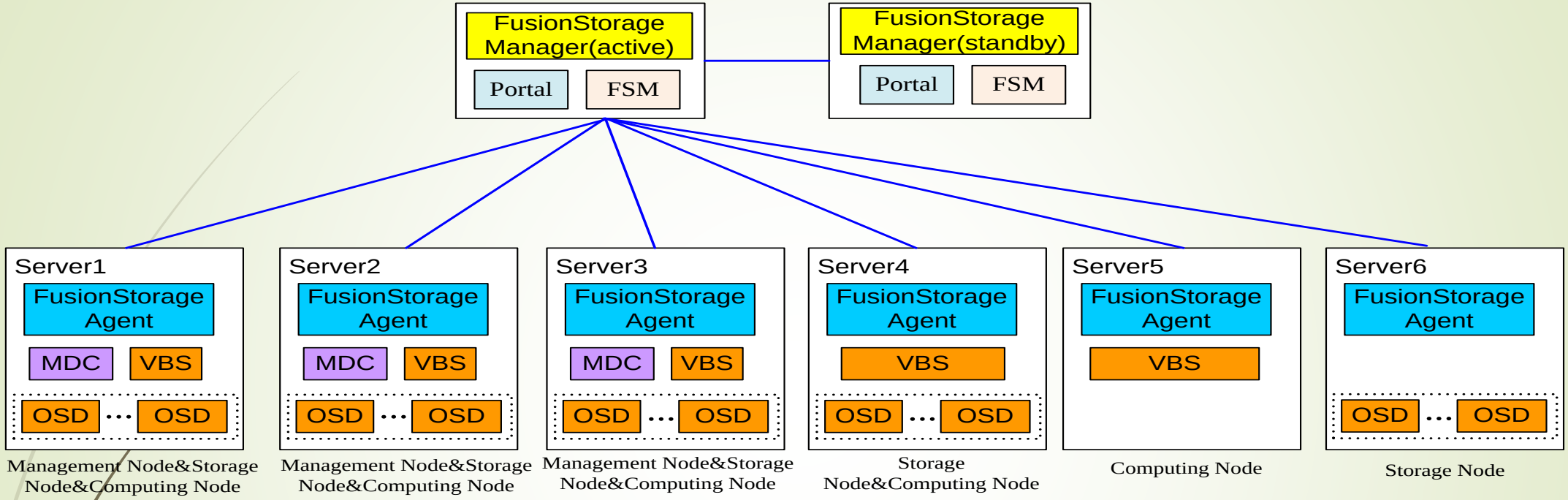


Архитектура многоресурсного пула FusionStorage



- **Спецификация FusionStorage Manager (FSM) :** А Система FusionStorage может иметь два узла FSM для работы в активном/резервном режиме. Они могут быть развернуты на VM или физических серверах.
- **Спецификация Metadata Controller (MDC)**: FusionStorage поддерживает до 96 MDC, причемкаждый MDC может управлять двумя пулами или до 2000 дисков.
- **Спецификация Virtual Block System (VBS):** Система FusionStorage может поддерживать работу до 10,240 процессов VBS, причем каждый из них может иметь доступ к любому пулу хранения в системе.
- **Спецификация пулов:** Система FusionStorage поддерживает до 128 пулов которые не зависят друг от друга. Эти пулы используют различные копии установок и устройств кеша.
- **Требования поддержки пулов хранения с различной производительностью**
- **Изоляция отказов для повышения надежности**
  Пулы ресурсов не зависят друг от друга с применением механизмов изоляции отказов и множества MDC для гарантирования того, что отказы в одном пуле не повлияют неблагоприятно на другие пулы.
- **Ограничения:**
  1) Пул ресурсов не может содержать основные носители или носители кеша разного типа.
  2) Пул ресурсов не может содержать два различных сетевых протокола.

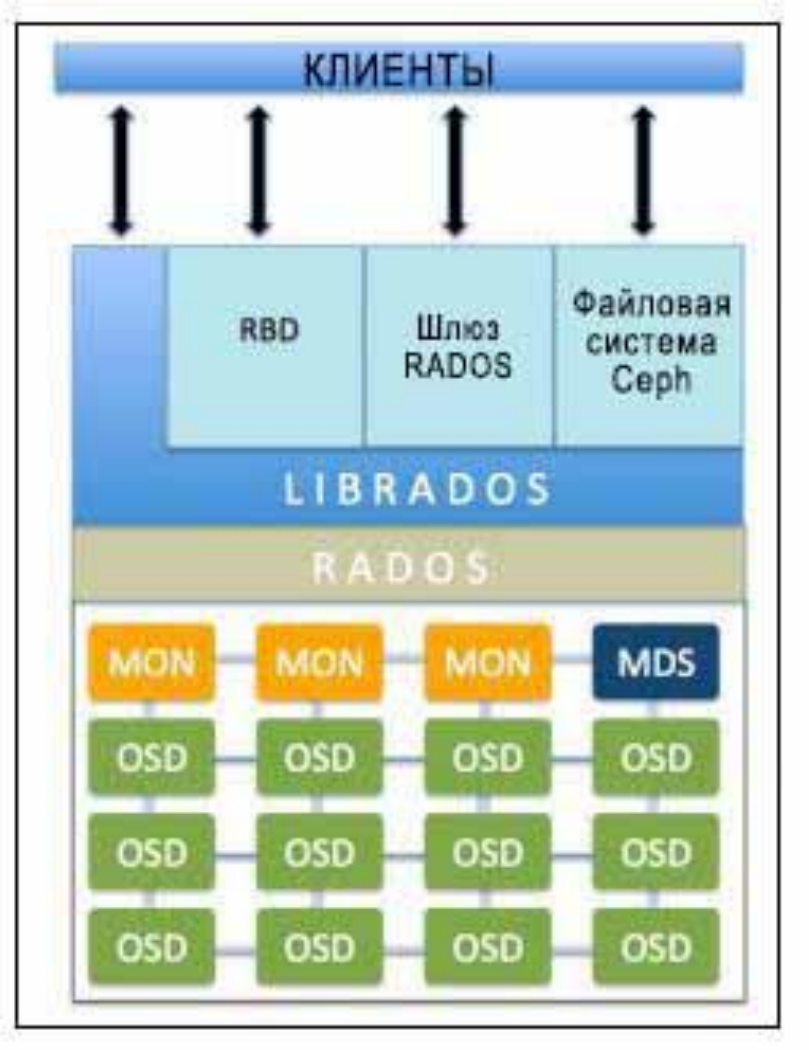
Развертывание программных модулей FusionStorage



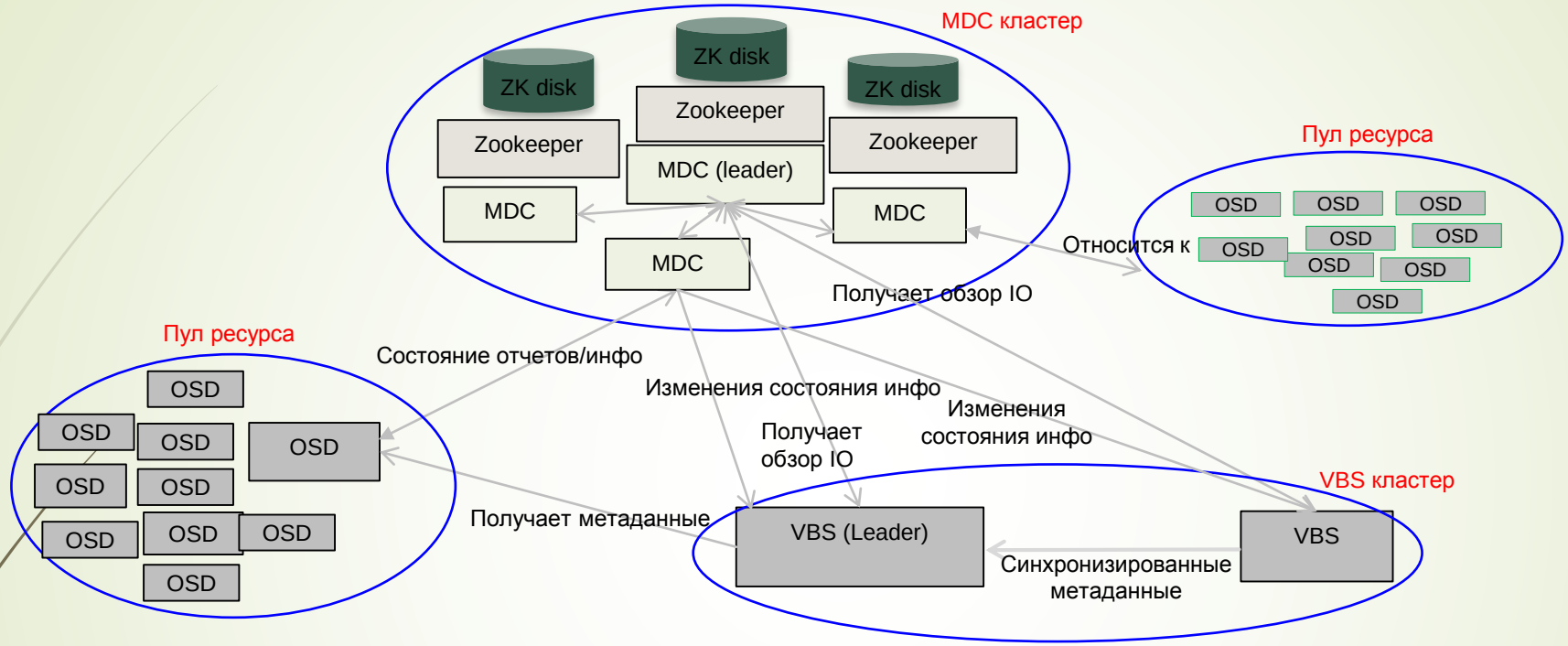
Модуль	Функция
FusionStorage Manager	Модуль управления FusionStorage, который разворачивается на активных и резервных узлах и обеспечивает работу и управляющие функции, включая управление сигнализацией, мониторинг, управление протоколированием и настройкой.
FusionStorage Agent	Модуль агента FusionStorage, который разворачивается на каждом узле (сервере) для обеспечения соединения между узлами сервера и FusionStorage Manager.
MDC	Компонента управления метаданными, контролирует состояние распределенного кластера, правила распределенных данных и правила реконструкции данных. По крайней мере три узла MDC чтобы сформировать кластер MDC.
VBS	Компонента управления Virtual block storage, которая управляет метаданными тома и предоставляет службу точки доступа распределенного кластера делай возможным доступ вычислительных ресурсов к ресурсам распределенного хранилища посредством VBS. A VBS process is deployed on each server to form a VBS cluster.
OSD	Устройство хранения на основе объектов, которое реализует определенные операции I/O. На каждом сервере развертывается множество процессов OSD, причем каждый процесс OSD разворачивается на один диск.

Сравниваем: Ceph

- **RBD-** блочные устройства RADOS
- **RGW-** шлюз RADOS (шлюз объектов), RESTful интерфейс, совместимый с ***AmazonS3***, ***Swift*** +***API***
- **CephFS-** файловая система. Через библиотеку libcephfs предоставляет доступ через ***API***, ядро (Ceph/***Fuse***), а также ***NFS***, ***CIFS***, ***SMB***
- **RADOS-** Reliable Autonomic Distributed Object Store, Безотказное автономное распределенное хранилище объектов
- **MON-** мониторы Ceph, поддерживают ***карты*** **OSD**, **MON**, **PG** (групп размещения) и **CRUSH** (Controlled Replication Under Scalable Hashing , Управляемые масштабируемые хешированием репликации)
- **OSD-** Object Storage Device, устройстве хранения объектов Ceph
- **MDS**-сервер метаданных

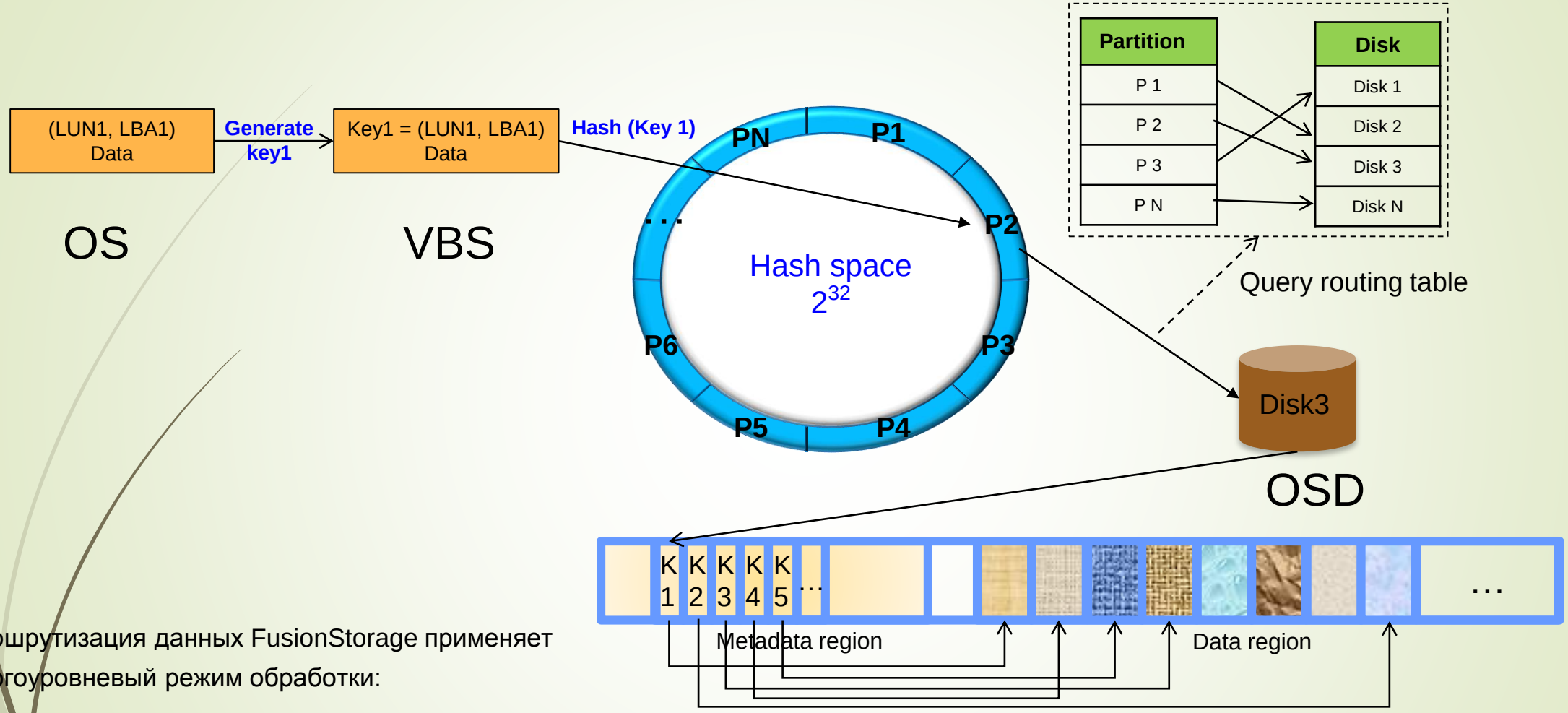


Основные принципыFusionStorage — Управление кластером



- При запуске системы MDCs взаимодействует с ZooKeeper (ZK) для определения ведущего MDC. Ведущий MDC и прочие MDCs отслеживают состояние друг друга путем взаимодействия heartbeat. Ведущий MDC определяет MDC, который сможет принять на себя обслуживание если MDC откажет. Если другой MDC определяет отказ ведущего MDS, оставшиеся MDC взаимодействуют с ZK для выбора нового ведущего MDC.
- При запуске OSD, он ищет свой домашний MDC и сообщает свое состояние этому MDC. Домашний MDC отсылает изменения состояния OSD на подключенные VBS. Если домашний MDC OSD отказывает, ведущий MDC определяет MDC для замены отказавшего. К одному MDC максимальномогут быть подключены два ресурсных пула.
- При запуске VBS он ищет ведущий MDC и регистрируется на нем, а также запрашивает у ведущего MDC является ли он ведущим. (Ведущий MDC устанавливает динамичный список VBS и синхронизует список VBS с другими MDC, так что MDC могут сообщать изменения состояния OSD отображенным VBS.)

Основные принципы FusionStorage — Маршрутизация данных

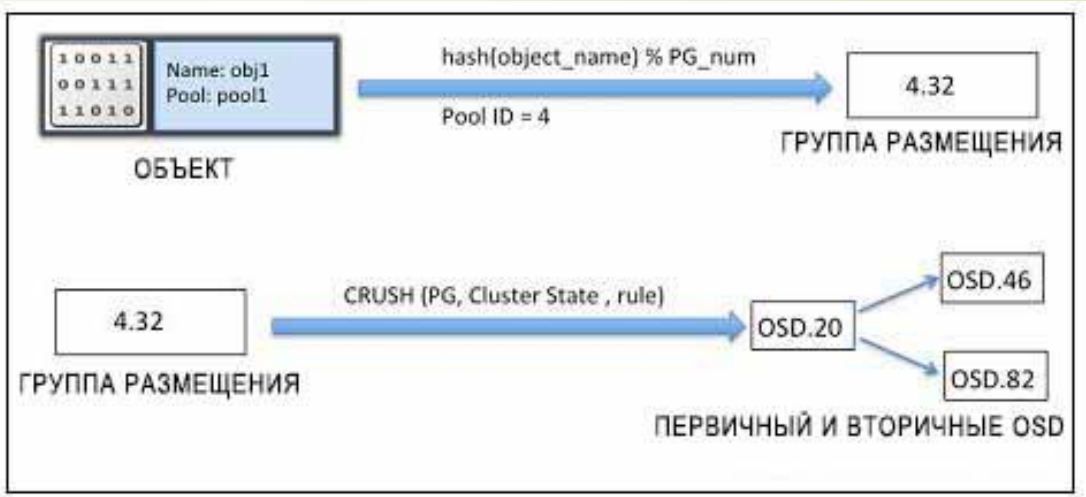


Маршрутизация данных FusionStorage применяет Многоуровневый режим обработки:

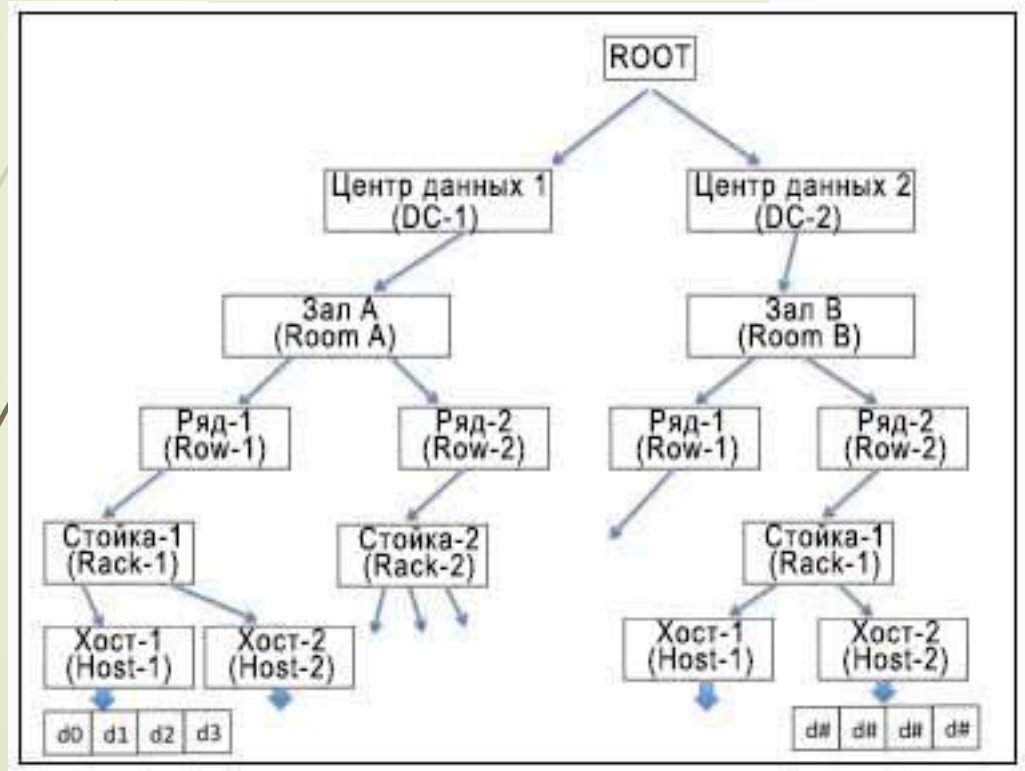
1. VBS определяет диск сервера где должны быть сохранены данные.
2. OSD устанавливает определенное положение на диске в котором должны быть сохранены данные.

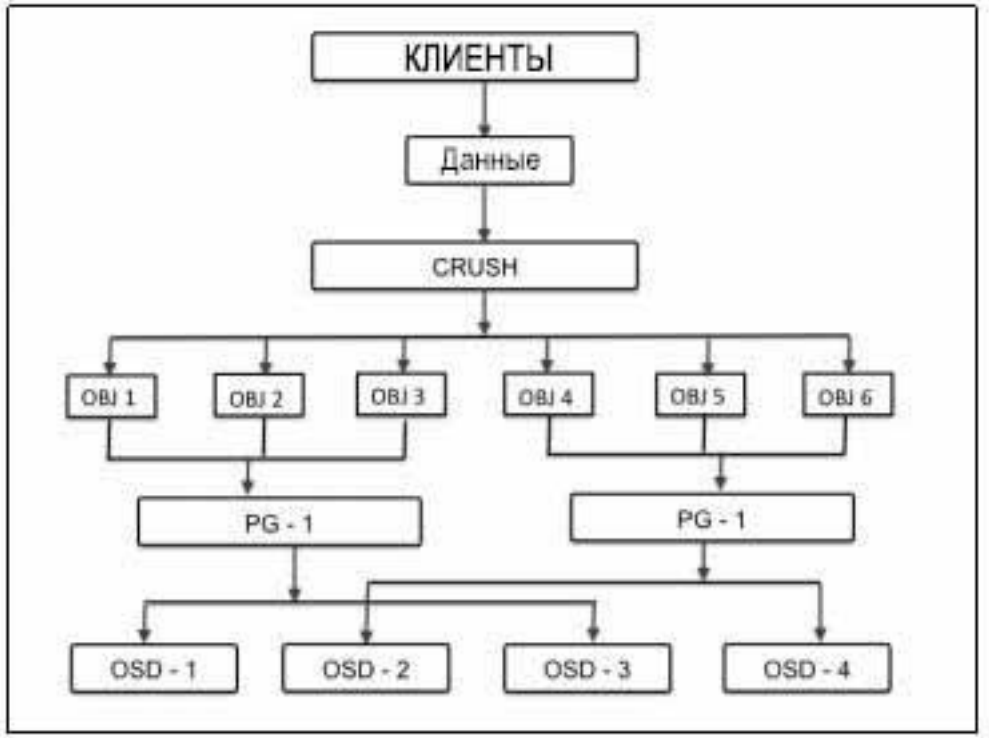
Ceph, Группа размещения:
способ группировки адресации OSD

-:

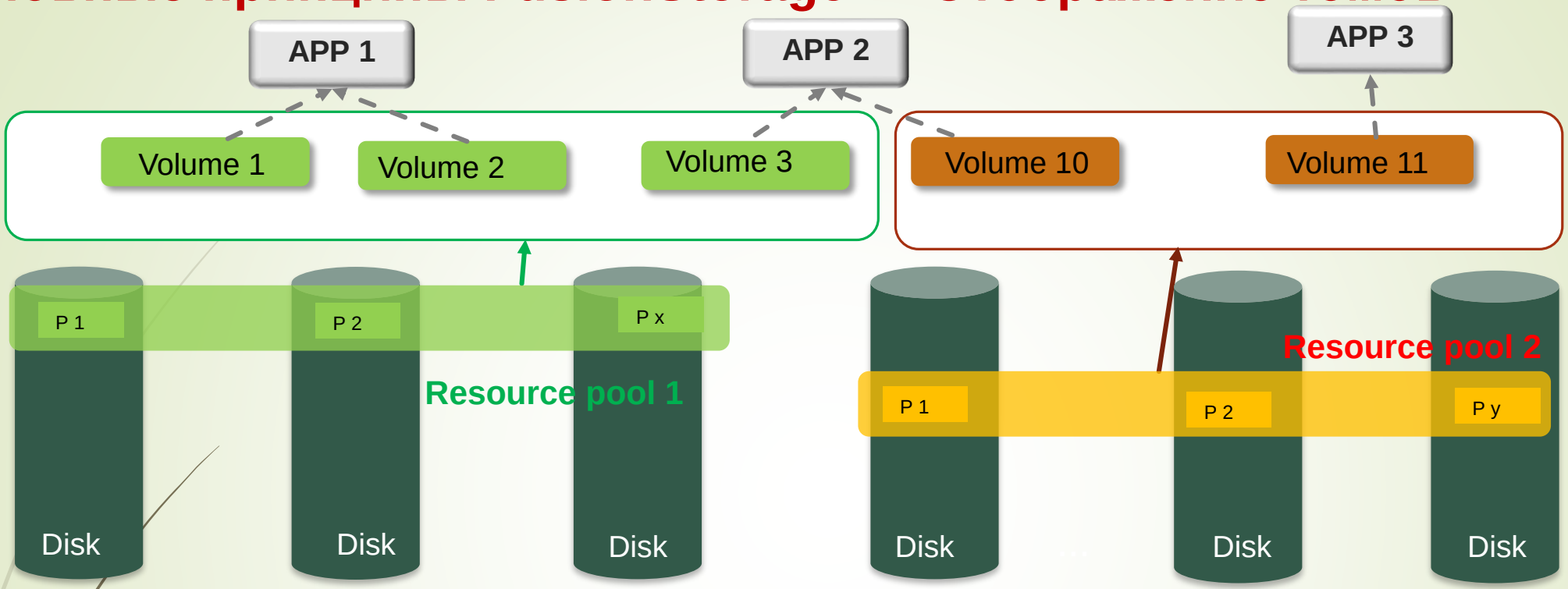


карта CRUSH:





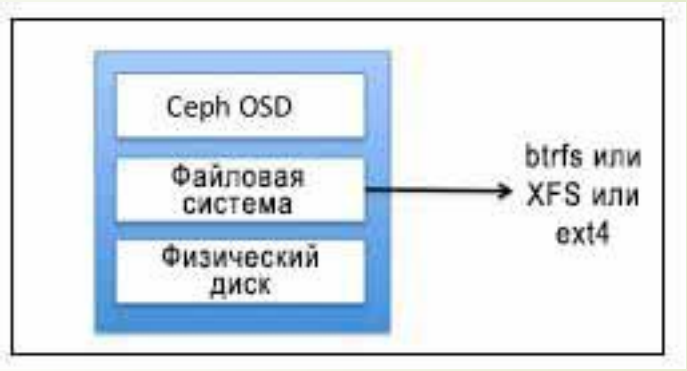
Основные принципы FusionStorage — Отображение томов



Пул ресурсов: аналогичен группе RAID group в SAN-устройстве. По сравнению с группой RAID, пул ресурсов имеет следующие преимущества:

- **Ширина полосы**: Поддерживает максимально 96 дисков (две копии), тем самым предоставляя гтгантское пространство хранения и избегая недостатка пространства на некоторых часто используемых дисках.
- **Динамичное горячее резервирование**: Все диски в пуле ресурса могут быть использованы в качестве дисков горячего резерва пула.
- **Простая структура**: Пул ресурса не применяет структуру logical unit number (LUN). Вместо этого, он только подразделяет ресурсы хранения на тома. Серверы могут получать прямой доступ к томам в пуле.

Использование дисковой файловой системы: XATTR и журналирование.



二 Функциональность

FusionStorage Функциональность и Спецификация

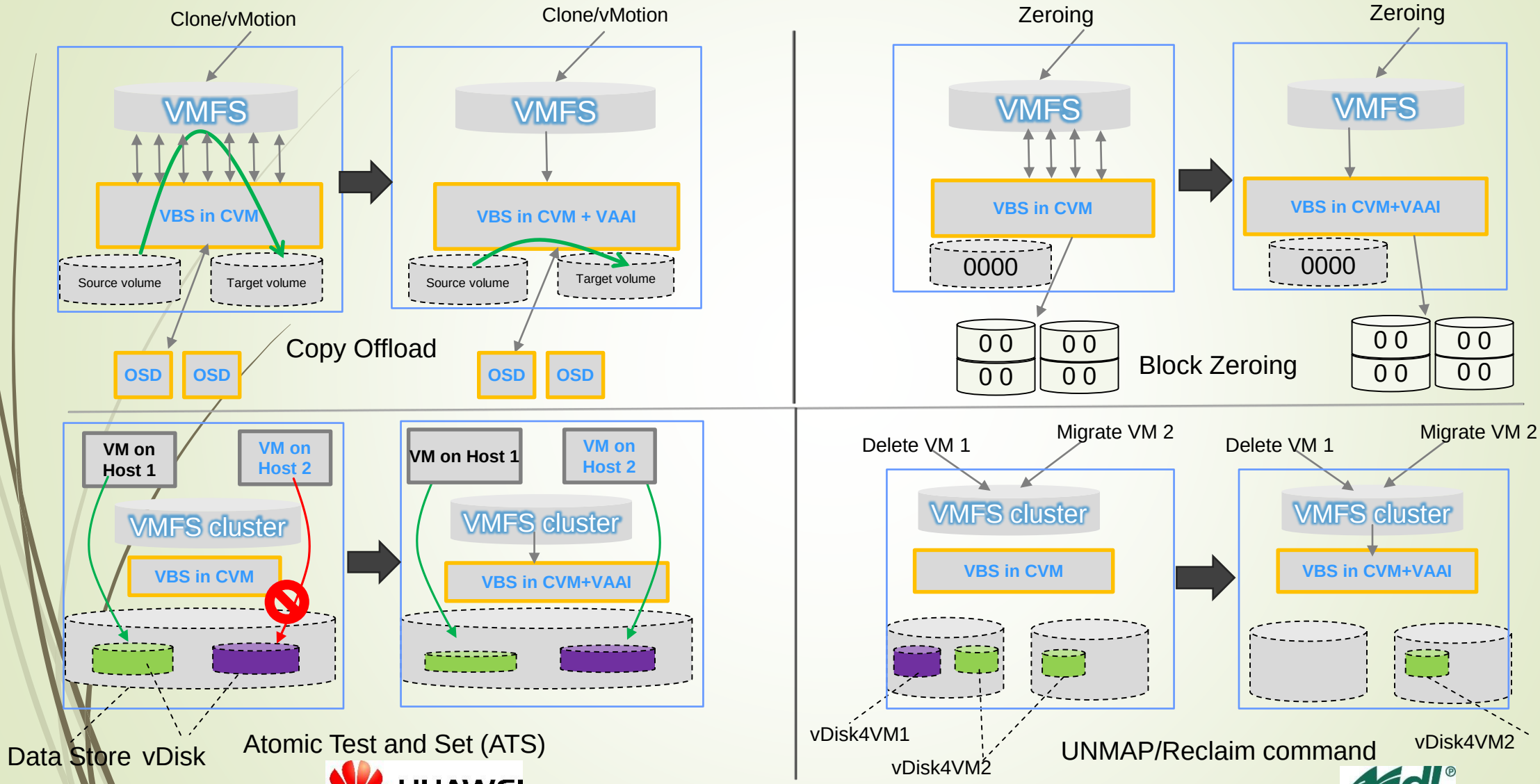
Функции ПО

- Блочные хранилища, поддерживающие интерфейсы SCSI и iSCSI.
- Множественные пулы ресурсов
- Репликации VM на основе хостов
- Снимки хранилищ
- Связанные клоны хранилищ
- SSD кеш и автоматическое построение иерархий хранимых данных
- Резервное копирование данных
- Поддерживает отображение iSCSI LUN, маскирование LUN, иVAAI
- Холодная миграция томов между пулами ресурсов
- Расширение томов в подключенном и отключенном состояниях

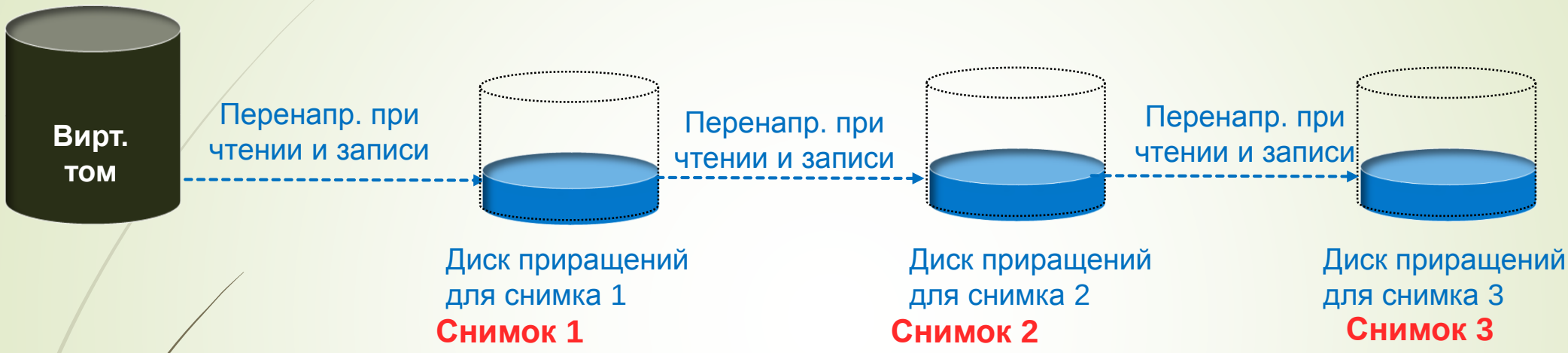
Основные характеристики

Показатель	Значение
Максимальное число узлов, поддержи-ваемое множеством пулов ресурсов	4096
Максимальноечисло узлов хранения подд. пулами ресурсов	256
Максимальное число дисков поддерживаемое пулами ресурсов	•12 до 96 дисков (2 коии) •12 до 2000 дисков (3 копии)
Ммаксимальная емкость тома	256 TB
Restoration duration for 1 TB data	30 минут
Режимы избыточности данных	2 копии и 3 копии
Максимальноечисло логических томов подд. пулами ресурсов	65,000
Степень связанных клонов тома	256
Максимальное число пулов ресурсов, поддерживаемое системой	128

VAAI (API хранилища VMware vSphere— Интеграция массивов)

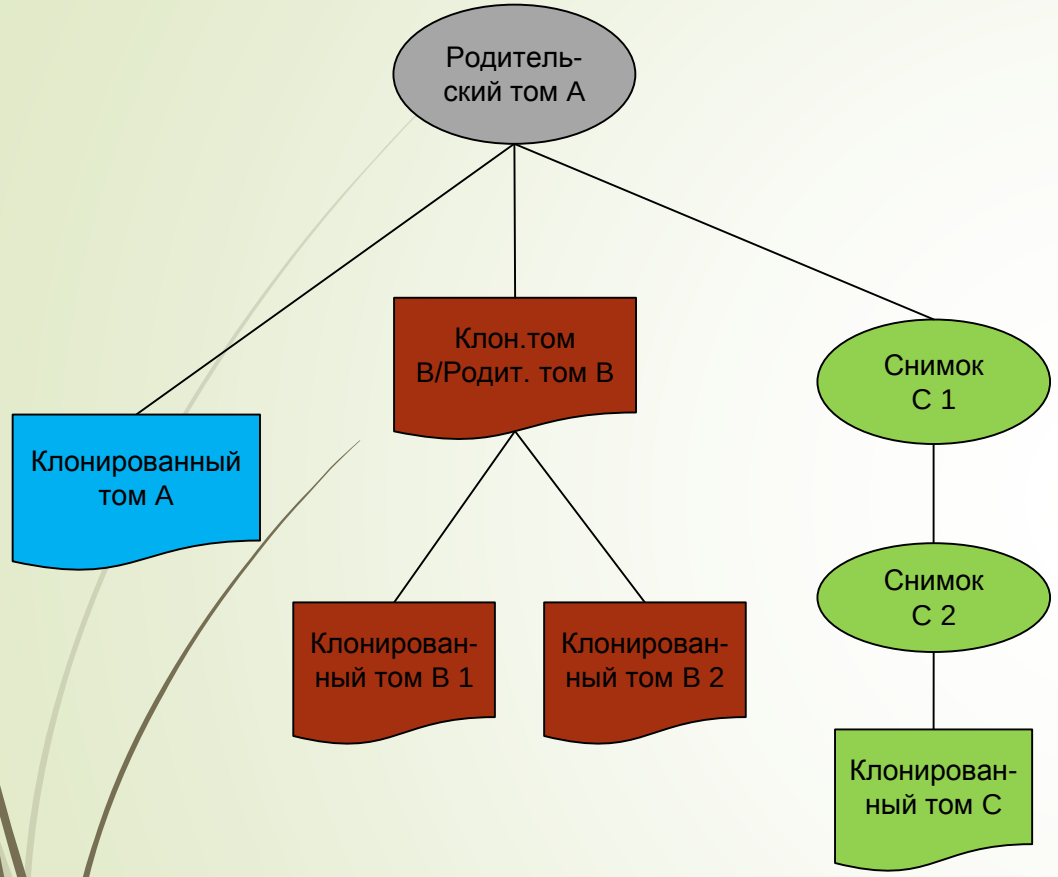


Высокопроизводительная технология инкрементальных снимков



- **Мин. задержки производит-ти**: Данные снимка могут быть быстро найдены с применением механизма DHT.
- **Неограниченное число снимков**: Метаданные снимка сохраняются распред. образом и могут масштабировать-ся без ограничений. Теоретически количество снимков не ограничено.
- **Быстрое восстановление тома**: С применением снимков том может быть восстановлен в пределах1сек без миграции данных. В противовес на устройствах SAN восстановление занимает несколько часов.

Высокопроизводительное связанное клонирование на базе снимков



Клон тома.

Клонированные тома выступают в роле родительских для клонирования.

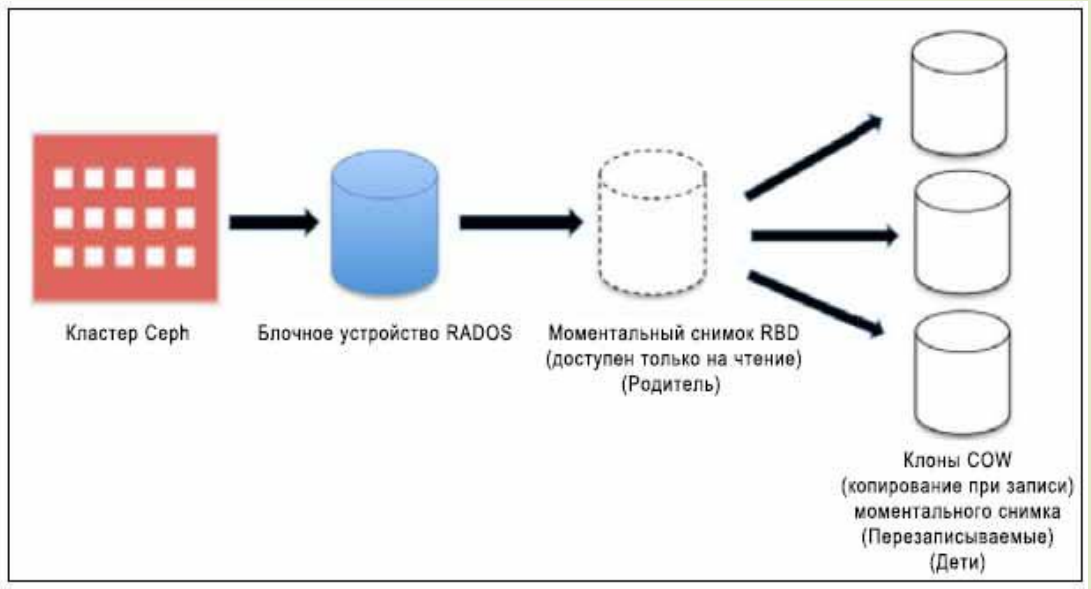
Клонирование тома при помощи снимков.

- **Клонирование тома без деградации произв.:** Том может быть клонирован с применением снимка и может быть быстро найден с механизмом DHT.
- **Высокопроизв. доступ к родит. тому:** Данные родит. тома кешируются в памяти, повышая произв. доступа в 3-5 раз. Более того, данные сохраняются распред. образом, что приводит к отсутствию узких мест, которые возникают при централизованном режиме хранения.
- **Имеет те же функции что и обычные тома:** Клонированный том также поддерживает снимки, восстановление с ними, а также клонирование в качестве родительского тома.

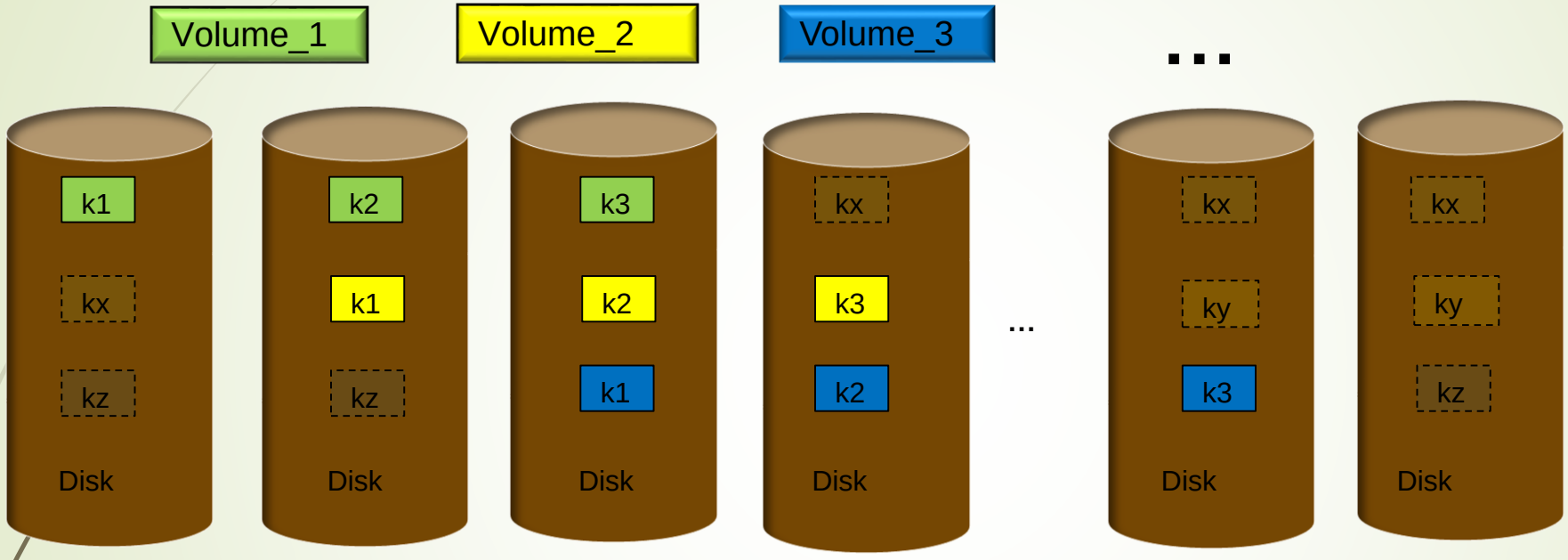
HUAWEI

Сравниваем, Ceph: снимки и клоны

Аналогичная функциональность

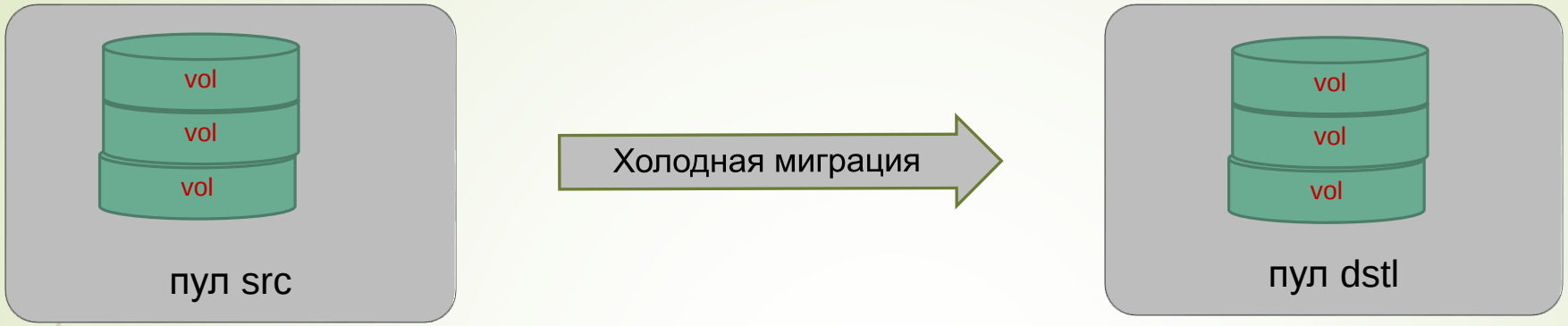


Интеллектуальное динамичное выделение без ухудшения производительности



- Для поддержки динамичного выделения применяется алгоритм распределенного хеша (DHT ring). Не требуется предварителное выделение пространства.
- Динамичное выделение не ухудшает производительность. В случае применения устройств SAN, расширение ресурса может ухудшать производительность СХД.

Холодная миграция томов между пулами ресурсов



Функция холодной миграции: Мгирация томов из пула src в пул dst.

Процедура:

Создание томов получателей> Копир. данных томов> Удаление исх. томов> Переим.томов получателей> Выполнено.

1. При холодной миграции, запись данных на исходный том не допускается.
2. Интерфейсы копирования могут запускаться инструментами миграции.

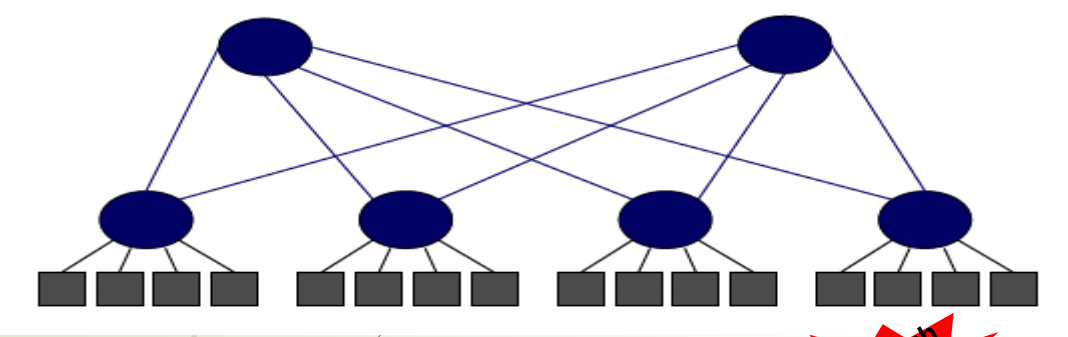
Сценарии пользователей:

Сценарий 1: Балансировка емкости между заполненным пулом и пустым пулом.

Сценарий 2: Миграция тома bмеждупулами ресурсов с различной производительностью, например, миграция томов с пула с низкой производительностью на высокопроизводительный пул.

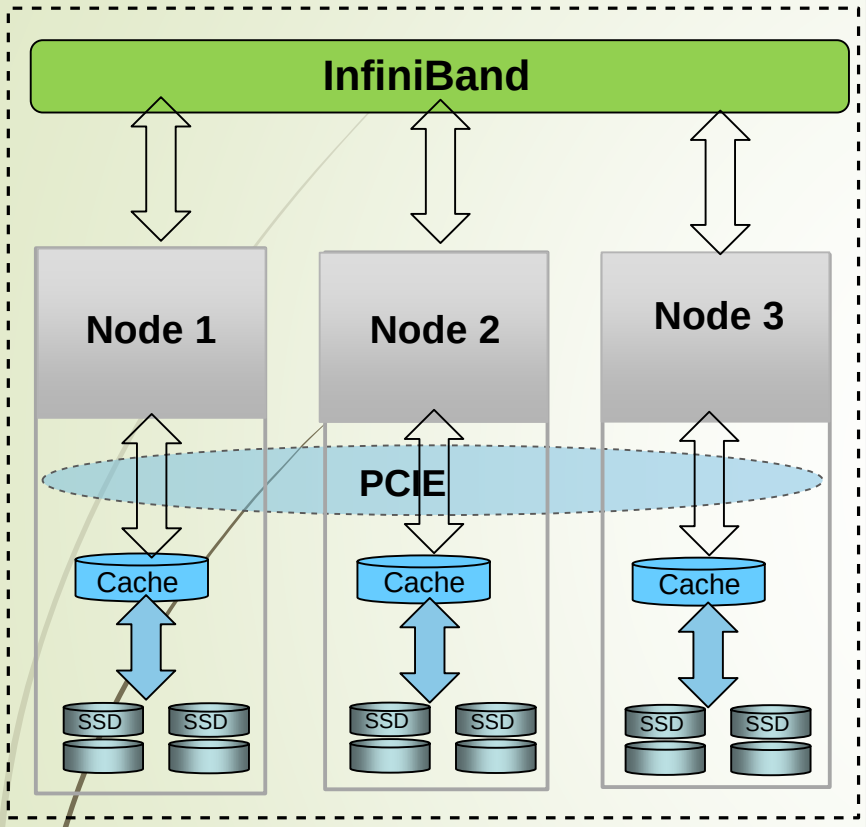
Высокая производительность & Низкая латентность — InfiniBand

Сверхвысокоскоростной обмен данными InfiniBand



- Поддержка 56 Gbit/s FDR IB ипредоставление сверхбыстрого обмена.
- Поддержка remote direct memory access (RDMA) для обеспечения сверхбыстрого обмена между узлами.
- Применение многоуровневой сетевой среды fat-tree для обеспечения гибкого расширения производительности.
- Обеспечение неблокируемой коммутационной среды в которой трудно создавать заторы .
- Быстрая и беспрепятственная передача вычисляемой и хранимой информации, с латентностью измеряемой на уровне десятков- сотен наносекунд.
- Предоставляет сетевую среду без потерь QoS и гарантирует целостность данных при передаче .
- Допускает соединение со множеством путей для активных и ждущих портов и гарантирует избыточность путей коммуникации.

Высокая надежность — Множество механизмов обеспечения безопасности данных

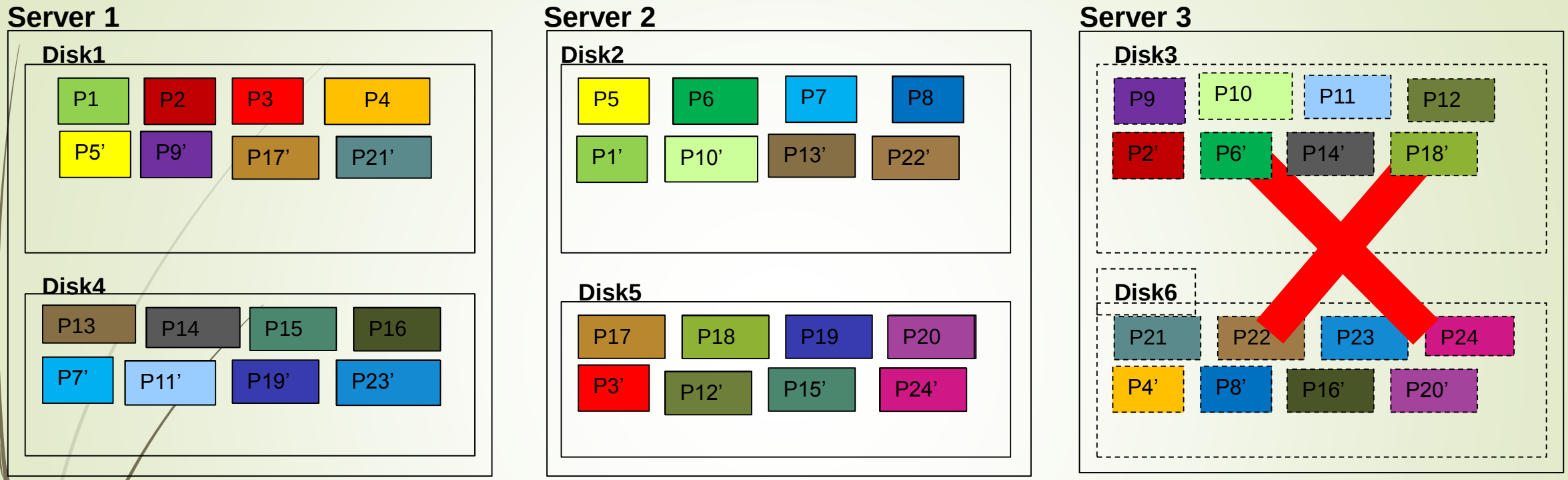


- **Множественное копирование данных:** Данные могут храниться с 1 копией (предоставляя доступность данных, эквивалентную RAID10) или со множеством идентичных копий (3-копии достигают доступность данных 7-nine).
- **Технология PCI-E SSD кеша:** поддерживает быстрые чтение и запись и гарантирует отсутствие потерь данных upon при выключении питания системы.
- **Протокол строго согласованных репликаций:** Если одна часть данных успешно записана в прикладную программу, сохраняются одна или несколько согласованных резервных копий. Затем любая копия может предоставить правильные данные при следующем чтении .

Множество механизмов безопасности данных обеспечивают их защищенность.

HUAWEI

Высокая надежность — Быстрая параллельная реконструкция данных

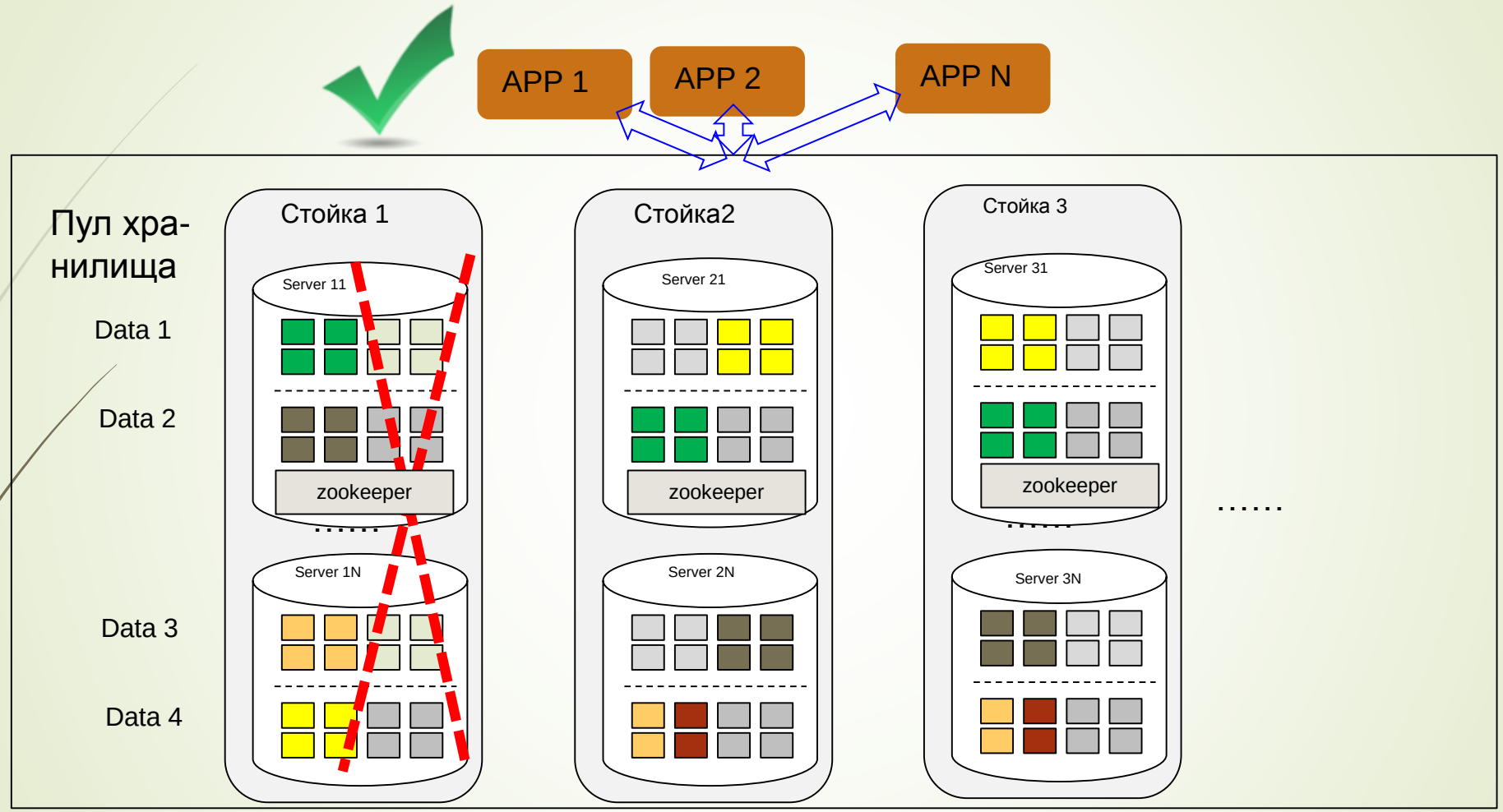


- Данные фрагментированы в пуле ресурсов и многие диски реконструируют эти фрагменты параллельно.
- Если диск отказывает, эти фрагменты автоматически строятся. Если диски в не оптимальном состоянии, происходит упреждающее перестроение.
- Все диски используются не только для резервирования, но и как основные. Более того, в случае отказа диска, данные неутрачены, причем немедленная замена отказавшего диска не требуется. (След. шаг: отложенное восстановление в Ceph!)

Для реконструкции 1 ТВ данных нужно менее 30 минут (12 часов для традиционного IP SAN СХД)

HUAWEI

Надежность на уровне стойки:



Отказ не воздействует на службы и может быть восстановлен автоматически.

Посетите наши веб- страницы с описанием решений масштабируемых СХД:
Huawei FusionStorage. Краткое описание:
http://www.mdl.ru/Solutions/Put.htm?Nme=**FusionStorage**
Изучаем Ceph, Каран Сингх (перевод):
http://onreader.mdl.ru/**LearningCeph**/content/index.html
Lazy Means Smart: Reducing Repair Bandwidth Costs in Erasure-coded Distributed Storage:
http://onreader.mdl.ru/Ceph/Planning/Blueprints/Hammer/**lazy-recovery**.htm
Shingled Erasure Code (SHEC):
http://onreader.mdl.ru/Ceph/Planning/Blueprints/Hammer/**SHEC**.htm#Fujitsu
Ceph. Рекомендации по оборудованию
http://www.mdl.ru/Solutions/Put.htm?Nme=**CephHW**
Книга рецептов Proxmox, Васим Ахмед (перевод, доп. материалы):
http://onreader.mdl.ru/ProxmoxCookbook/content/**Fencing**.html