АЭРОДИСК
мы делаем будущее

Техническое описание продуктов:
СХД ВОСТОК-5 и ENGINE-5

Дата: 18.03.2022
Версия: 2.0

Аппаратные платформы: контроллерные пары

Линейка СХД АЭРОДИСК серии 5 делится на два продукта:

- АЭРОДИСК Восток – СХД на базе процессоров Эльбрус.
- АЭРОДИСК Engine - СХД на базе процессоров x-86

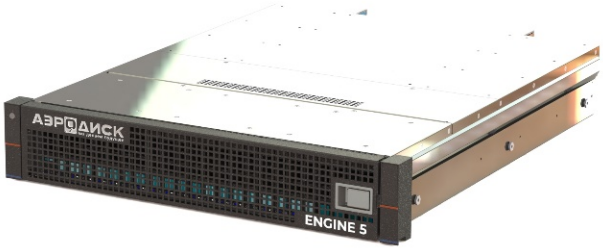
Процессорная архитектура – это единственное отличие данных двух продуктов. Другая начинка СХД, т.е. кодовая база, фронт/бэк-энд адаптеры, корпус, носители информации, модули расширения и т.п. являются идентичными.

Контроллерные пары

Контроллерные пары СХД АЭРОДИСК серии 5 независимо от процессорной архитектуры выпускаются в идентичных корпусах следующих форматов:

- Двухконтроллерное шасси формата SBB (Storage Bridge Bay) – классическое исполнение СХД, узлы контроллерной пары физически расположены в одном корпусе/шасси, соединены внутри корпуса интерконнектом по шине PCI или Ethernet (RDMA) и подключены к общему бэкплейну дисковой корзины на передней панели корпуса. Для дополнительного расширения емкости и/или производительности предусмотрены внешние модули (см. раздел «Модули расширения»).

- Раздельные узлы контрольной пары, то есть контроллеры СХД, расположены в разных корпусах, соединяются между собой с помощью внешнего интерконнекта по оптическому Ethernet (10/25/40/100 Gb/s) с поддержкой RDMA. Дисковые корзины в таком решении могут быть только внешние (см. раздел «Модули расширения»).



На всех доступных вариантах контроллерных пар предустановлено программное обеспечение АЭРОДИСК A-CORE версии 5 или выше.

Все контроллерные пары поддерживают установку следующих Front/Back-end адаптеров:

- Fibre channel 8/16/32 Gb/sec
- Ethernet 1/10/25/40/100 Gb/sec
- Infiniband 40/56/100 Gb/sec

Доступные конфигурации контроллерных пар приведены в документе «Техническая спецификация».

Модули расширения

Модули расширения позволяют решать две задачи: увеличение емкости и увеличение производительности:

- Модули расширения дисковой емкости (дисковые полки)
- Модули увеличения вычислительной мощности (IO-модули)

Модули расширения дисковой емкости – это классические дисковые полки. В системах хранения АЭРОДИСК поддерживаются следующие модели дисковых полок:

- 12 дисков SAS 2,5/3,5’, 2U, 2хБП
- 24 диска SAS 2,5/3,5’, 4U, 2хБП
- 24 диска SAS 2,5’, 2U, 2хБП
- 60 дисков SAS 2,5/3,5’, 4U, 2хБП
- 108 дисков SAS 2,5/3,5’, 4U, 2хБП

Модули увеличения вычислительной мощности (IO-модули) – это дополнительные вычислительные узлы СХД, обеспечивающие операции ввода-вывода для подключенных к ним дисковых корзин / полок. В отличие от контроллеров СХД IO-модули не выполняют каких-либо управляющих функций.

С аппаратной точки зрения IO-модули полностью повторяют контроллерные пары, при этом на программном уровне есть существенные отличия. На IO-модулях установлена ограниченная версия ПО A-CORE (IO-версия), которая не требует дополнительных лицензий как на контроллерные пары. В данную версию ПО включены только Back-end часть IO-движка A-CORE, сервисная консоль (для гарантийного обслуживания) и Restful API.

Все взаимодействие контроллерных пар и IO-модулей реализовано через Restful API в рамках кластера хранения. С точки зрения администратора СХД при этом подход к управлению не меняется, все управление происходит из веб-интерфейса контроллерных пар кластера (более подробно см. функция «Кластер хранения»).

Более подробная информация об аппаратной составляющей, включая лицензионные опции и поддерживаемые носители информации, приведена в «Технической спецификации».

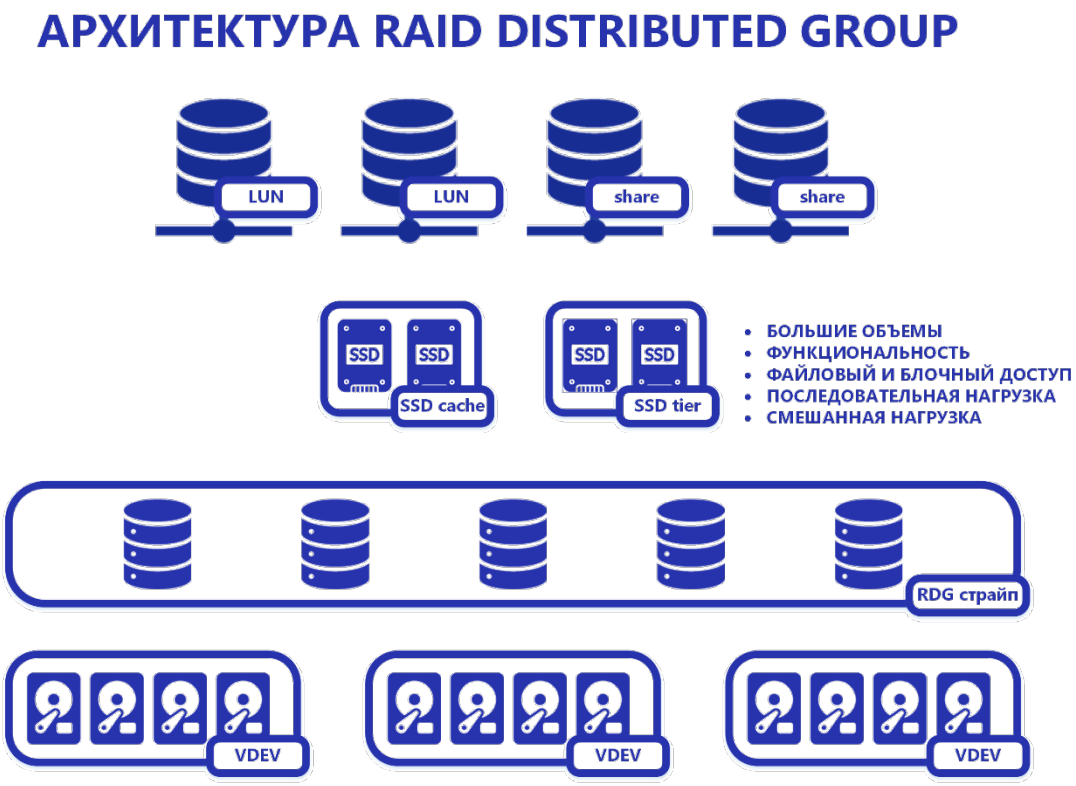
Архитектура хранения RDG

Архитектура хранения данных в СХД АЭРОДИСК поддерживает три метода организации хранения:

- RAID Distributed Group (RDG);
- Dynamic Disk Pool v1 (DDP1).
- Dynamic Disk Pool v2 (DDP2).

Отличительными особенностями реализации RDG в системах АЭРОДИСК являются:

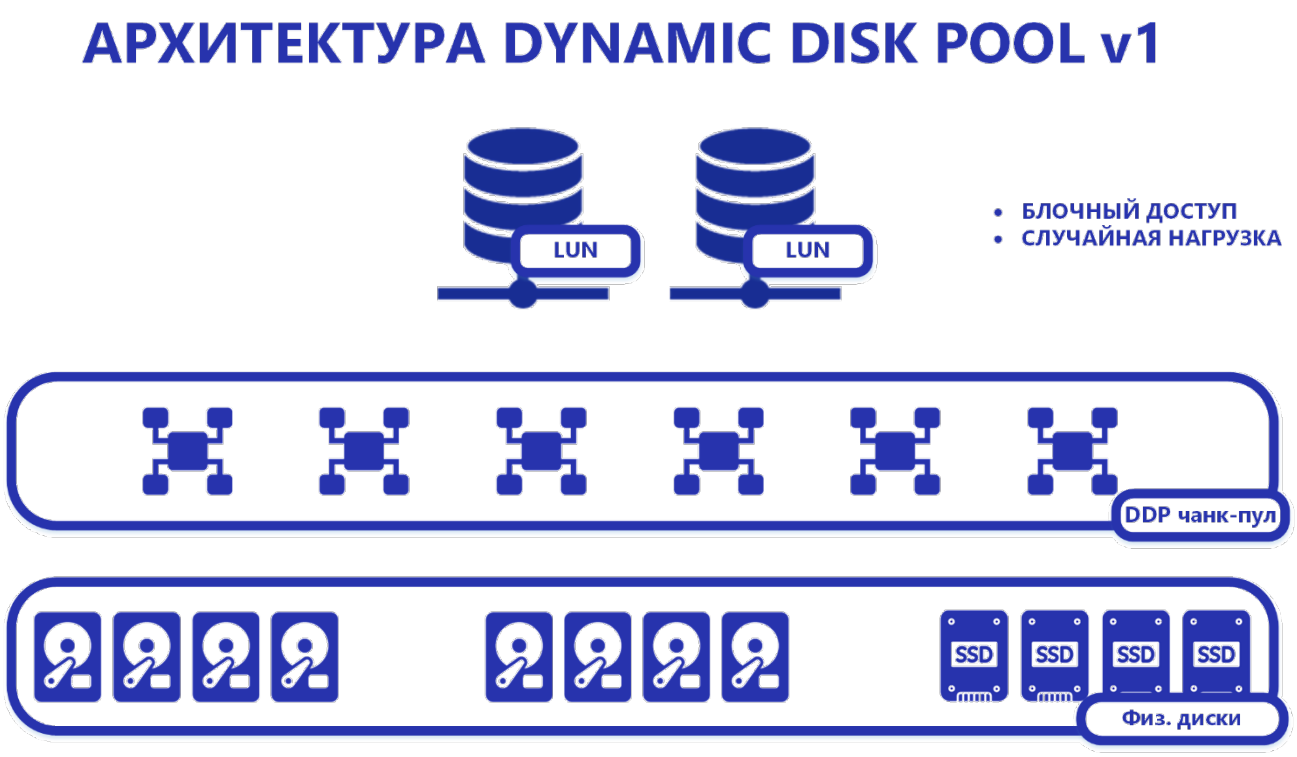
- RDG состоят из виртуальных устройств, каждое из которых имеет заданную структуру RAID (1/10, 5/50, 6/60, 6P/60P) (тройная четность);
- В RDG поддерживается и файловый, и блочный доступ;
- Виртуальные устройства последовательно объединяются в одну виртуальную группу RDG за счет чего количество дисков в группах (и для данных, и для четности) не ограничено;
- Вне зависимости от объема тома или файловой системы все диски в группе участвуют в вводе-выводе для данного тома или файловой системы;
- Диски горячей замены являются глобальными;
- Любая группа может быть, как гибридной, так и стандартной;
- RAM-кэш включен по умолчанию и работает только на чтение;
- SSD кэш на чтение/запись с возможностью тройного и четверного зеркалирования;
- SSD диски для online-tiering;
- Миграция LUN «на лету»;
- Мгновенные снимки, снэпклоны и связанные клоны;
- Скорость перестроения RAID можно регулировать политикой перестроения;
- RDG наилучшим образом подходит для операций последовательного чтения/записи данных, а так же для операций случайного чтения.



Архитектура хранения DDP v1

Отличительными особенностями реализации DDP в системах АЭРОДИСК являются:

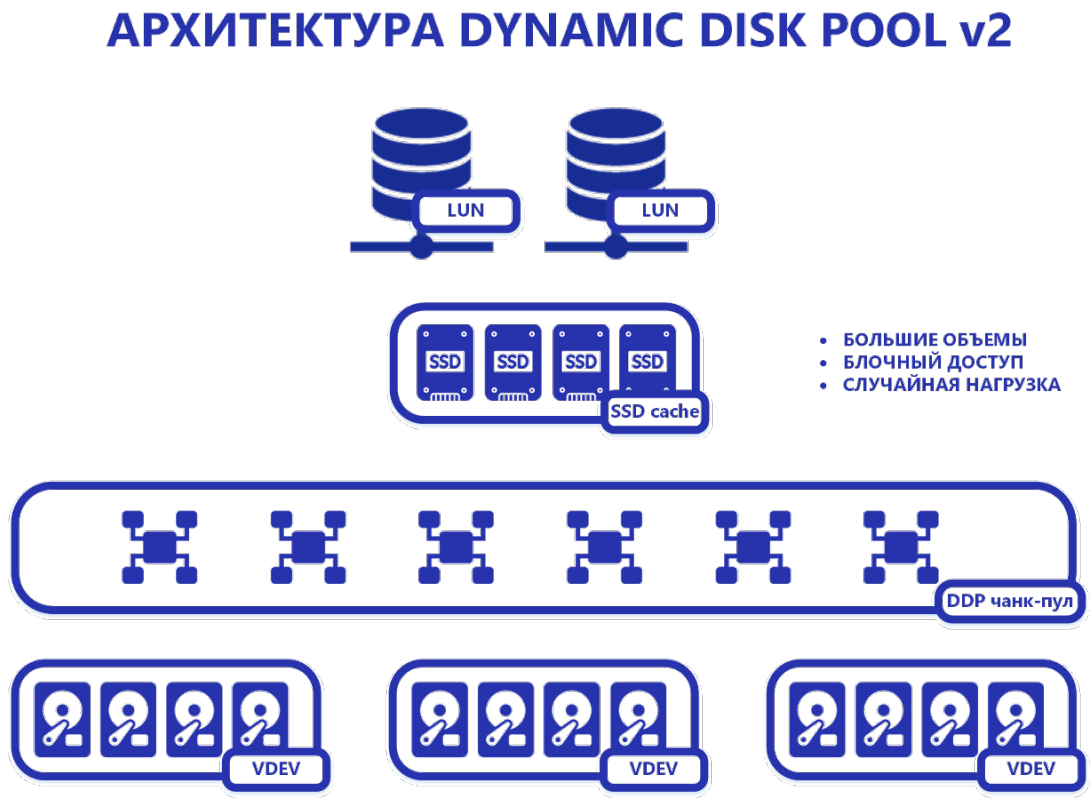
- DDP состоит из произвольного набора дисков – Пул (Pool);
- На каждом пуле можно организовать блочные устройства со следующими уровнями отказоустойчивости: RAID 0, 1, 10, 5, 6;
- В DDP поддерживается только блочный доступ (iSCSI, FC);
- Вне зависимости от объема тома все диски в пуле участвуют в вводе-выводе для данного тома (для RAID5 и RAID6 есть логическое ограничение по количеству дисков для одного блочного устройства, 10 дисков для RAID-5 и 24 диска для RAID-6);
- Диски горячей замены являются глобальными;
- Любая дисковая группа может быть, как гибридной, так и стандартной;
- Миграция LUN «на лету»;
- Компрессия и дедупликация на уровне LUN;
- Клоны и мгновенные снимки;
- SSD-кэш назначается на LUN-ы и работает и на чтение, и на запись;
- При выходе из строя диска происходит частичное перестроение данных (значительно быстрее полного перестроения), так как необходимо восстановить четность данных на уровне чанков только для затронутых LUNов;
- Более высокая производительность по сравнению с RDG для операций случайной записи и чтения особенно при использовании All-Flash конфигураций.



Архитектура хранения DDP v2

Отличительными особенностями реализации DDPv2 в системах АЭРОДИСК являются:

- DDP2 состоит из виртуальных устройств, каждое из которых имеет заданную структуру RAID (1/10, 5/50, 6/60), виртуальные устройства объединяются в страйп (RAID-0), образуя пул;
- На каждом пуле можно организовать блочные устройства со следующими уровнями отказоустойчивости: RAID 0, 1, 10, 5, 6, 50, 60;
- В DDP2 поддерживается только блочный доступ (iSCSI, FC);
- Вне зависимости от объема тома все диски в пуле участвуют в вводе-выводе для данного тома (в отличии от DDP1 RAID5 и RAID6 не имеют ограничений на количестве дисков в блочном устройстве);
- Диски горячей замены являются глобальными;
- Любая дисковая группа может быть, как гибридной, так и стандартной;
- Миграция LUN «на лету»;
- Компрессия и дедупликация на уровне LUN;
- Клоны и мгновенные снимки;
- SSD-кэш назначается на LUN-ы, работая и на чтение, и на запись;
- При выходе из строя диска происходит частичное перестроение данных (значительно быстрее полного перестроения) только на уровне виртуальных устройств (VDEV);
- Более высокая производительность по сравнению с RDG для операций случайной записи и чтения особенно при использовании All-Flash конфигураций.



АЭРОДИСК мы делаем будущее

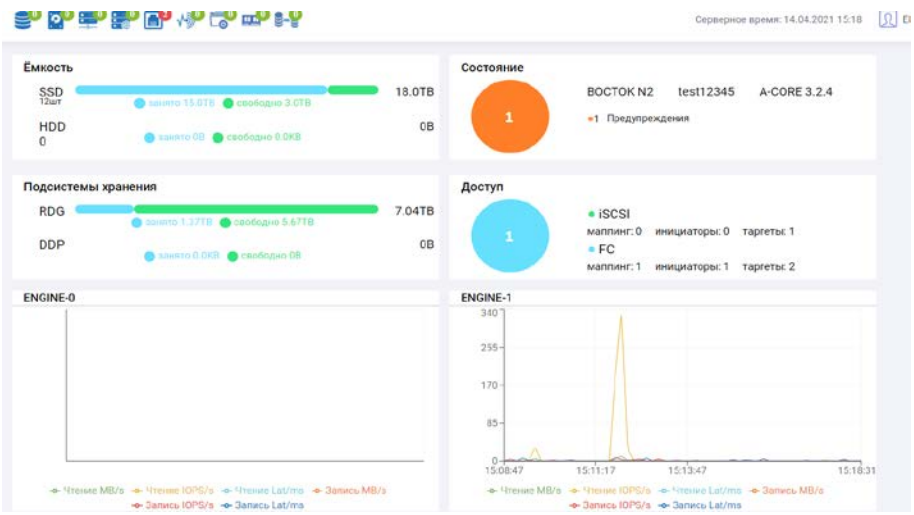
Сравнение функциональности организации хранения

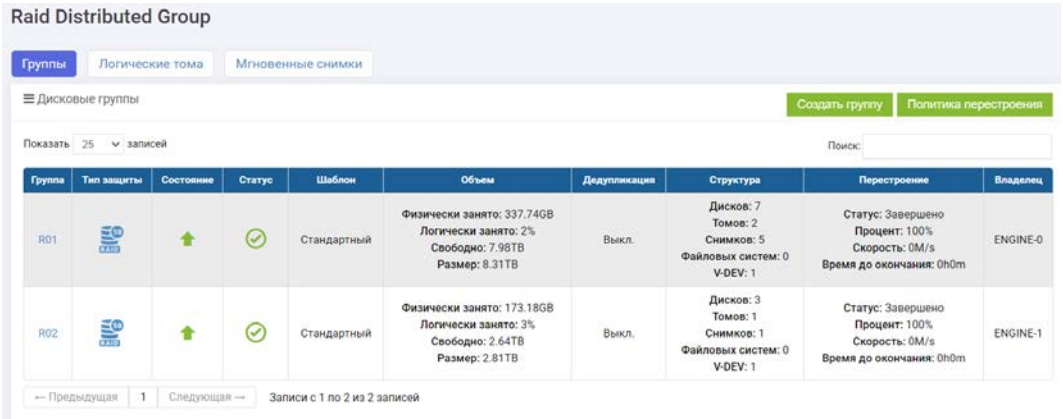
Задачи/функционал	RDG	DDP1	DDP2
Максимальное количество контроллеров	8 в NAS-режиме 2 в SAN-режиме	2 в SAN-режиме	
Отказоустойчивость HA-пар	Active/Active (ALUA)		
Уровни RAID	1/10, 5/50, 6/60, 6/60P (тройная четность)	0, 1, 10 ,5, 6	1/10, 5/50, 6/60
Блочный доступ	Да		
Файловый доступ	Да	Нет	
Протоколы доступа	FC\iSCSI\NFS\SMB	FC\iSCSI	
Гибридные группы (SSD+HDD)	Да		
All Flash группы	Да	Да (предпочтительно)	
Ограничения дисковых групп	Нет	RAID-5 - 10 дисков RAID-6 - 24 диска	Нет
Разные уровни RAID на одной группе	Нет	Да	Нет
Изменение объема дисковой группы	Да		
Встроенная компрессия и дедупликация	Да		
Тонкие тома	Да		
Онлайн-миграция LUN	Да		
SSD-кэш (чтение и запись)	Да		
Онлайн тиринг (SSD+HDD)	Да	Нет	
Снэпшоты	ROW	COW	
Локальная репликация	Да	Нет	
Удаленная репликация (синх/асинх)	Да		
Метрокластер	Да		
Глобальная автозамена дисков	Да		
Политики перестроения RAID	Да	Нет	
Поддержка сетевых меток (VLAN)	Да		

Управление СХД

Все системы хранения АЭРОДИСК используют единый интерактивный интерфейс на русском языке, позволяющий управлять всеми контроллерами СХД, установленными в системе, а также обеспечивает:

- Интерактивный Web-интерфейс на русском языке;
- Визуализация контроллеров, дисков и портов ввода-вывода;
- Визуализация сенсоров и датчиков температуры;
- Мониторинг состояния и нагрузки в реальном времени;
- Логирование действий администратора;
- Возможность выгрузки логов и статистики;
- Командная строка (linux-like) для автоматизации операций;
- Отправка оповещений по SMTP, SNMP, SYSLOG;
- Внешний мониторинг, например, с помощью GRAFANA.





Высокая доступность

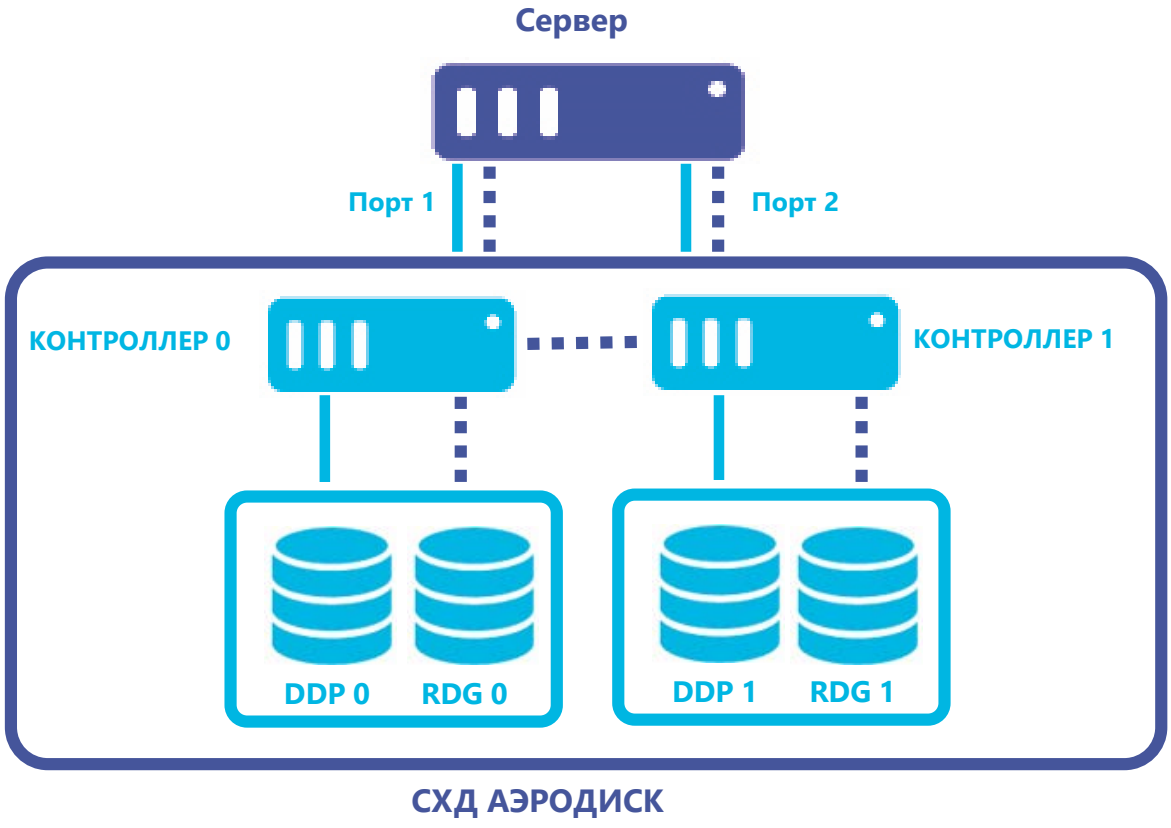
СХД АЭРОДИСК ВОСТОК поддерживает высокую доступность active-active в ассиметричном режиме (ALUA) в конфигурации 2-х контроллеров для SAN режима и до 8-ми контроллеров в NAS-режиме. Это означает, что все системные контроллеры всегда используются при обработке данных. В данном режиме дисковые группы (RDG и DDP) распределяются между всеми активными контроллерами. При этом администратор системы в случае необходимости (например, для обновления) может вручную переключать группы между контроллерами.

Кластерное ПО АЭРОДИСК работает как с блочным, так и с файловым доступом. Heartbeat между нодами выполняется с помощью интерконнекта (внутреннего или внешнего в зависимости от аппаратной реализации контроллерных пар). Кластер автоматически переключает оптимальные и неоптимальные пути, а также автоматически меняет владельца групп хранения в следующих случаях:

- Отказ контроллера (смена владельца);
- Отказ задействованных в воде-выводе портов СХД (смена владельца);
- Отказ порта на хосте (смена путей оптимальный-неоптимальный).

Функциональность: высокая доступность

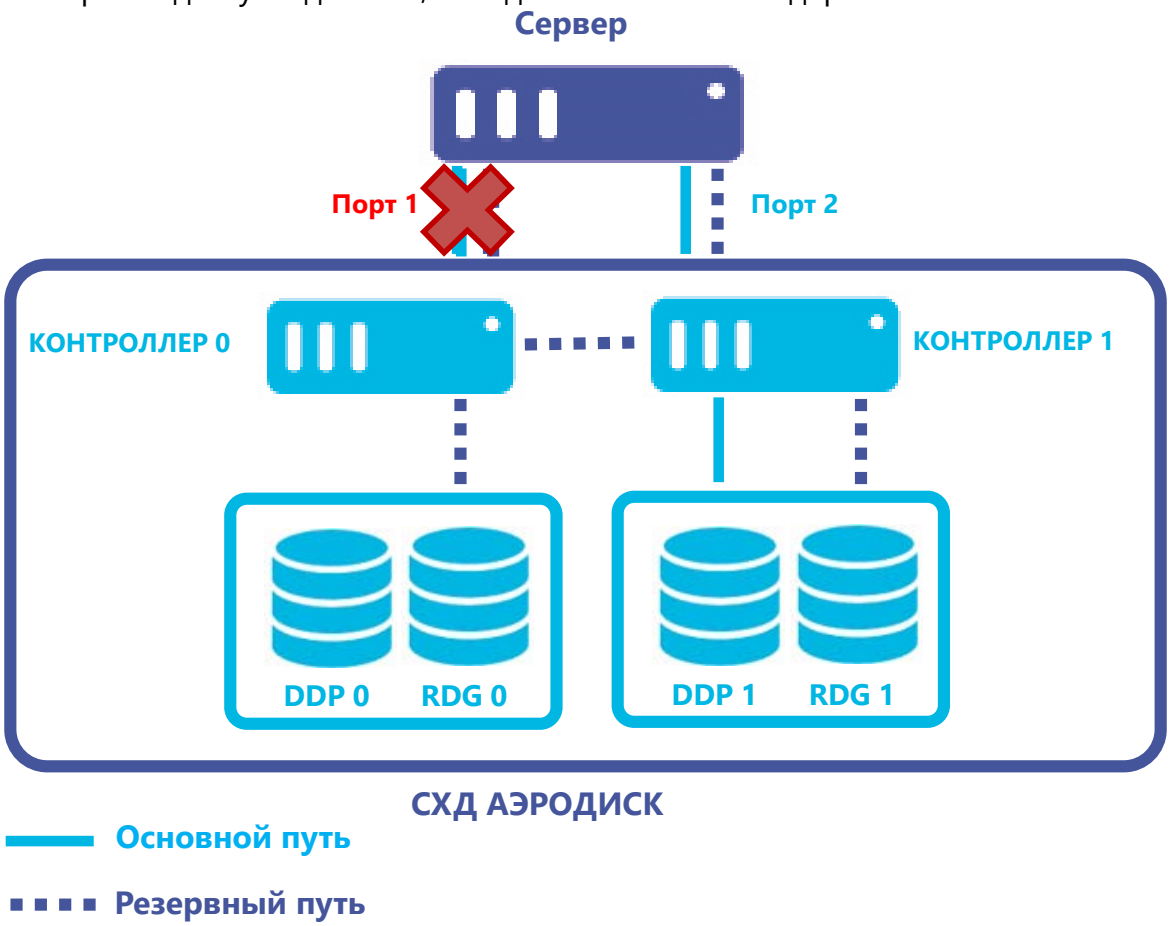
На примере ниже показана 2-х контроллерная конфигурация, которая подключена к 2-м портам хоста, для которых средствами ОС настроен multipath. На СХД созданы 4 группы хранения, для 2-х из них назначен владельцем первый контроллер (Контроллер-0), для 2-х других владельцем назначен второй контроллер (Контроллер-1). Оба контроллера (и 4 группы) видны обоим портам хоста. Для DDP0 и RDG0 владельцем назначен контроллер 0, пути через этот контроллер для данной группы являются оптимальными. При этом существует неоптимальный путь (через интерконнект и Контроллер-1), который задействуется в случае отказа основного порта на хосте. Для DDP1 и RDG1 обратная ситуация: владельцем является Контроллер-1, через него лежит оптимальный путь, а через интерконнект и Контроллер-0 – неоптимальный.



В любой момент администратор СХД может сменить владельца каждой из групп. Процесс смены владельца занимает примерно 5-10 секунд и происходит без прерывания ввода-вывода. Эта же операция выполняется администратором для перевода контроллера в режим обслуживания, например, когда требуется аппаратное или программное обновление СХД.

Отказ порта

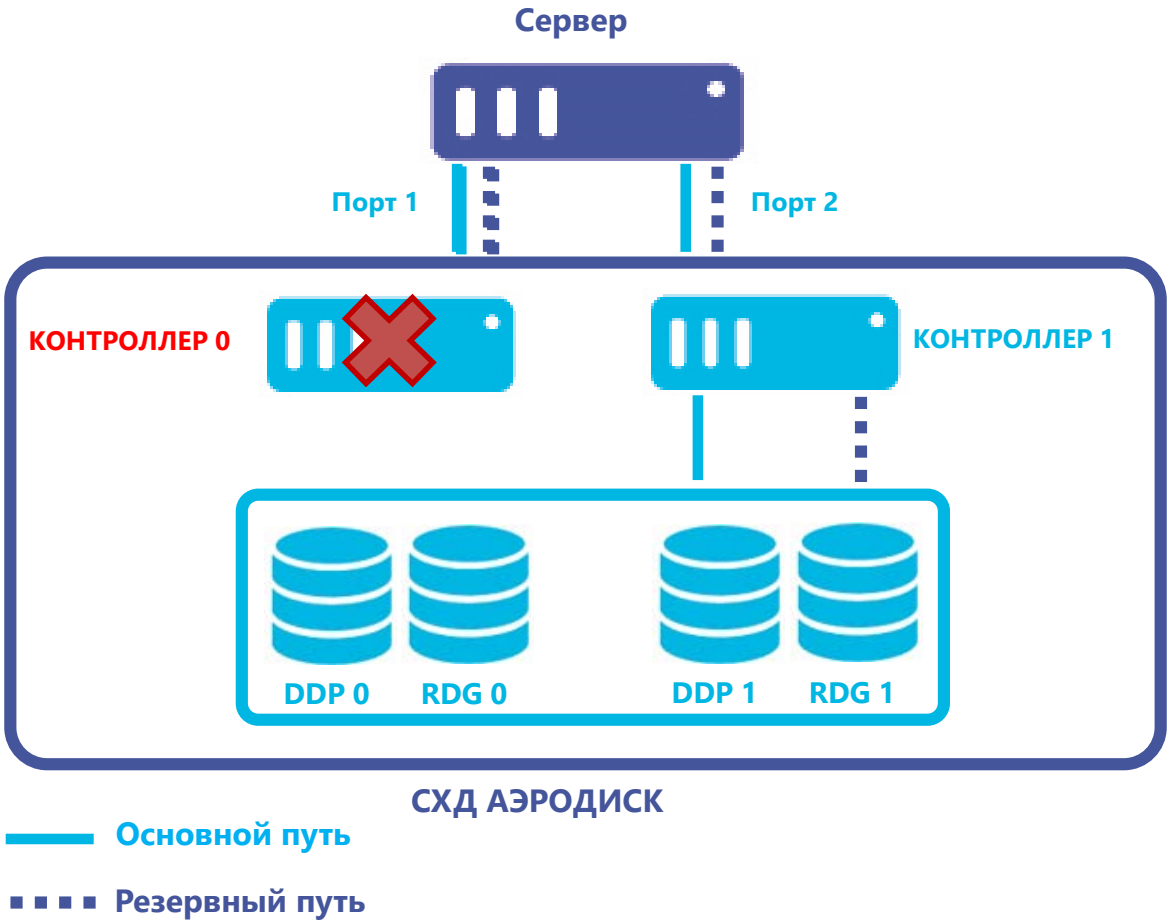
На схеме ниже представлена ситуация отказа порта на хосте, который был оптимальным для групп DDP0 и RDG0 (через Контроллер-0). В этом случае СХД автоматически задействует неоптимальный путь через Контроллер-1 и интерконнект, что сохранит доступ к данным, но с дополнительной задержкой.



Когда порт на хосте будет восстановлен, данные автоматически пойдут по оптимальному пути.

Отказ контроллера

На схеме ниже представлена ситуация отказа контроллера. В случае физической потери контроллера (или 2-х портов ввода-вывода на контроллере) система выполнит принудительную смену владельца всех групп хранения на отказавшем контроллере. Далее произойдёт смена владельца, что происходит без прерывания ввода-вывода.



Когда контроллер контроллер-0 снова вернется в строй, администратору нужно будет вручную сменить владельца на контроллер-0 обратно.

Файловый и блочный доступ

Блочный доступ обеспечивается путем предоставление блочного устройства (LUN) конечному хосту или хостам по протоколам Fibre Channel, iSCSI с поддержкой ISER или IB. Блочный доступ может предоставляться с LUN, созданных как на RDG, так и на DDP группах.

Файловый доступ обеспечивается путем предоставления файловой системы по протоколам NFS и SMB (CIFS) конечному хосту или хостам. Для SMB (CIFS) может использоваться авторизация пользователей с помощью Active Directory. Файловый доступ работает только для RDG групп.

LUN-ы и файловые системы создаются внутри RDG-групп. В рамках одной RDG группы могут функционировать как LUN-ы, так и файловые системы. При этом размер RDG может быть динамично увеличен (т.е. в онлайн режиме) с помощью добавления дополнительных дисков в RDG.

Как для файлового, так и для блочного доступа поддерживаются следующие уровни RAID для RDG:

- RAID 1/10
- RAID 5/50
- RAID 6/60
- RAID 6P/60P (тройная чётность)

Для блочного доступа поддерживаются следующие уровни RAID для DDP:

- RAID 0
- RAID 1/10
- RAID 5
- RAID 6
- RAID 50 (только DDP2)
- RAID 60 (только DDP2)

В системах хранения данных **АЭРОДИСК** файловый и блочный доступ можно обеспечивать с одного и того же контроллера, достаточно лишь наличия соответствующих Front-End адаптеров (FC, Ethernet, IB), установка дополнительных специальных модулей не требуется.

Кластер хранения

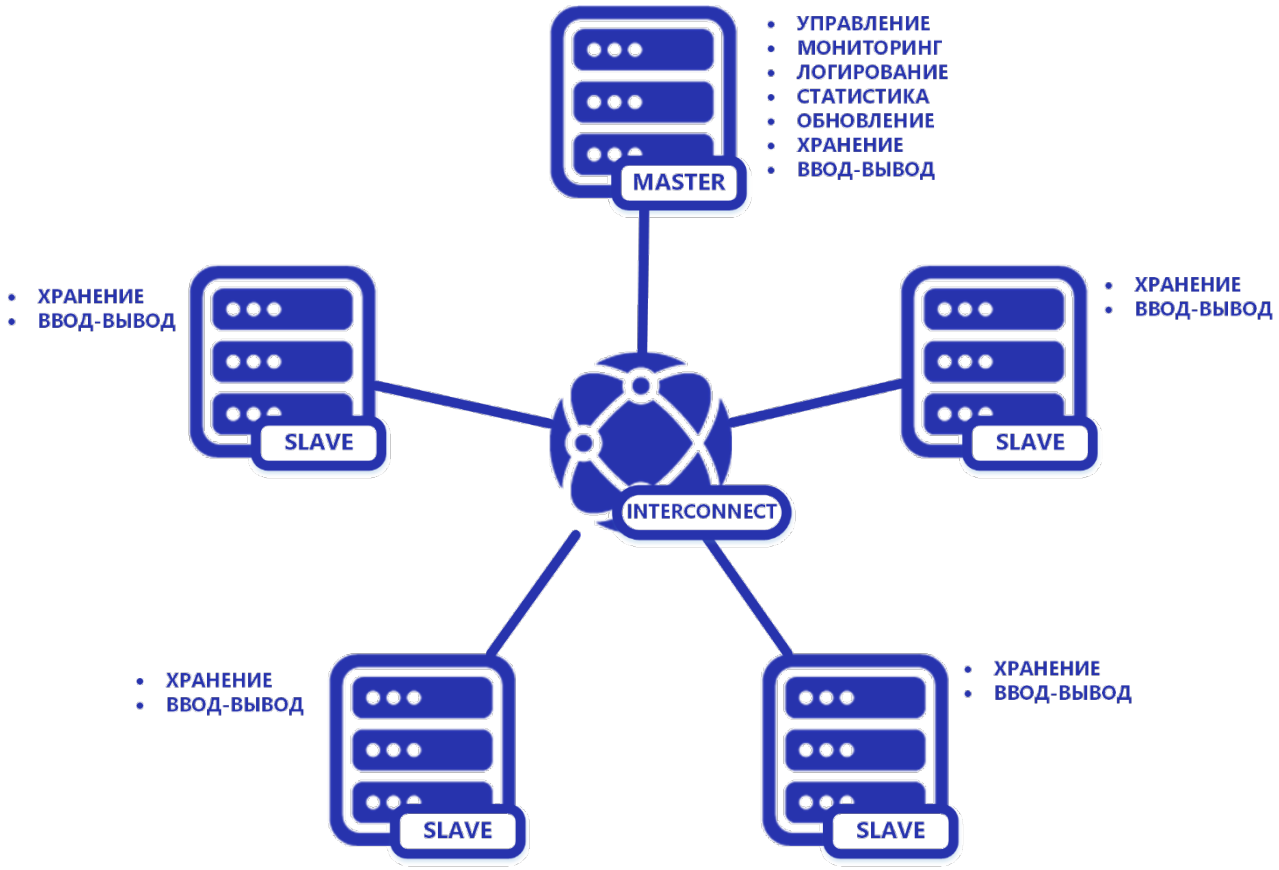
Функция «Кластер хранения» позволяет объединять разные системы хранения данных АЭРОДИСК (до 16 СХД/32 контроллеров) в единую инфраструктуру хранения данных. В кластере хранения для СХД-участников кластера используется ролевая модель, в которой предусмотрено две роли:

- MASTER – контроллерная пара, являющаяся источником управляющих команд для всех остальных СХД в кластере (IO-модулей). MASTER также обеспечивает операции ввода-вывода для подключенных к нему хранилищ (продолжая выполнять роль IO-модуля для своих дисков)
- SLAVE – IO-модуль, обеспечивающий операции ввода-вывода для подключенных к нему хранилищ. Все управляющие команды IO-модуль получает от MASTER-а через Restful API.

Для обмена команд и выполнения других кластерных операций в кластере хранения на физическом уровне используется интерконнект на базе протокола Ethernet (10/25/100 Gb/s). Коммутаторы для интерконнекта при необходимости поставляются в комплекте.

Полезной особенностью является возможность объединения в один кластер хранения СХД разных процессорных архитектур в частности: ВОСТОК-5 на базе процессоров E2K (Эльбрус) и ENGINE-5 на базе процессоров архитектуры x-86.

На рисунке ниже приведена общая логика построения кластера хранения с указанием выполняемых функций в зависимости от роли СХД.



Функциональность: кластер хранения

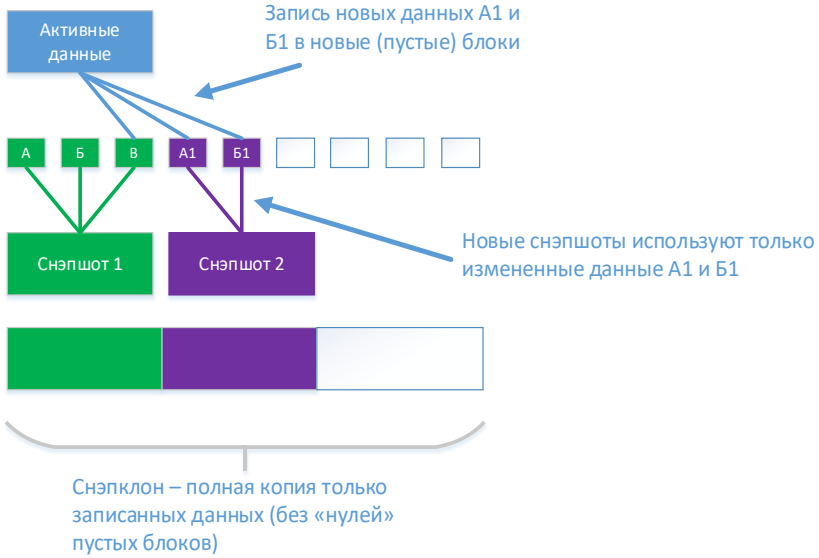
Как видно из схемы, управление кластером осуществляется только с контроллерной пары MASTER-а. MASTER централизованно осуществляет мониторинг, сбор статистики и логов, а также выполняет функции обновления. Кроме того, в рамках кластера хранения предусмотрена миграция данных между всеми СХД - участниками кластера. Локальные операции на конкретных СХД, такие как создание пулов, групп, блочных устройств и т.п., выполняются путем отправки API-команд с MASTER-а на SLAVE-ы. Для управления каждой СХД в отдельности можно выбрать её в интерфейсе кластера хранения и перейти в её «локальный» интерфейс. «Локальный» выделен кавычками, поскольку сам интерфейс функционирует также на MASTER-е, а взаимодействие с управляемой СХД происходит через API.

Роль мастера назначается при инициализации СХД. Для управления кластером предусмотрены соответствующие веб-интерфейс и командная строка, функционирующие в отдельном контейнере контроллерной пары MASTER-а. Подключиться к SLAVE IO-модулю напрямую и осуществлять управление в обход MASTER-а не получится, контейнер с веб-интерфейсом и командной строкой из соображений безопасности там заблокирован.

Задачей кластера хранения является консолидация управления разными СХД АЭРОДИСК. Подключение хостов к СХД не завязано на роль MASTER и выполняется напрямую к HA-парам кластера, поэтому с точки зрения ввода-вывода MASTER не является точкой отказа. Таким образом потенциальная недоступность обоих узлов контроллерной пары MASTER-а не влияет на доступность данных, которые обслуживают IO-модули. При этом на случай выхода из строя обеих нод MASTER-а для восстановления управления SLAVE-ами в спец. версии A-CORE, устанавливаемой на IO-модули, предусмотрена сервисная консоль управления для целей технической поддержки, которая позволяет кроме стандартных функций управления переназначить роль MASTER-а.

Функциональность: Мгновенные снимки, снэпклоны, связанные клоны

Мгновенные снимки (снэпшоты) и связанные клоны в RDG и DDP (в случае тонких лунов) используют модель переадресации при записи (Redirect-on-Write), т.е. СХД всегда пишет новые блоки данных в новое место, переставляя на них указатель, а старые блоки данных (т.е. на которые уже нет указателя) никогда не стираются, а помечаются системой как освобожденные. Этот механизм позволяет создавать любое количество снэпшотов и связанных клонов без какого-либо влияния на производительность СХД.



Снэпшоты создаются мгновенно и изначально не потребляют дисковое пространство, а растут по мере изменения данных.

Связанные клоны создаются мгновенно и сразу могут быть доступны серверу на чтение/запись при наличии маппинга (доступны только для RDG).

Полезной функцией является создание/удаление снэпшотов по расписанию (локальная репликация). Это применимо, если требуется сохранять резервные копии данных очень часто, что невозможно сделать внешними системами резервного копирования, т.к. в силу их специфики (высокая нагрузка на каналы, долгое время записи и пр.) они резервируют данные обычно не чаще чем раз в сутки.

В этом случае есть возможность настроить расписание снэпшотов, например, каждые 3 часа в течение суток со сроком хранения одни сутки. Через сутки снэпшоты начнут перезаписываться заново, а данные старше суток уже будут сохранены внешней системой резервного копирования.

Функциональность: мгновенные снимки, снэпклоны и связанные клоны

Снэпклон – это гибрид клона и снэпшота. Снэпклоны создаются быстрее, чем классические клоны и изначально занимают ровно ту полезную емкость, которую занимает источник. При этом снэпклон, как и классический клон может находиться в любой группе

Восстановление данных из снэпшотов и связанных клонов можно выполнить двумя способами.

- Откатить снэпшот/связанный клон, полностью перезаписав данные LUN/ФС. Это удобно, когда нужно быстро восстановить LUN/ФС полностью.
- Присоединить связанный клон в виде отдельного LUN/ФС к хосту и восстановить данные с этого LUN. Такой способ подходит для ситуаций, когда не нужно восстанавливать LUN/ФС целиком, а нужно восстановить только некоторые объекты (файлы).

Снэпклоны и связанные клоны возможно также подключать к хосту в виде отдельных LUN или файловых систем.

СХД **АЭРОДИСК** не имеет ограничений по количеству созданных снэпшотов и снэпклонов, за исключением физического ограничения используемого оборудования.

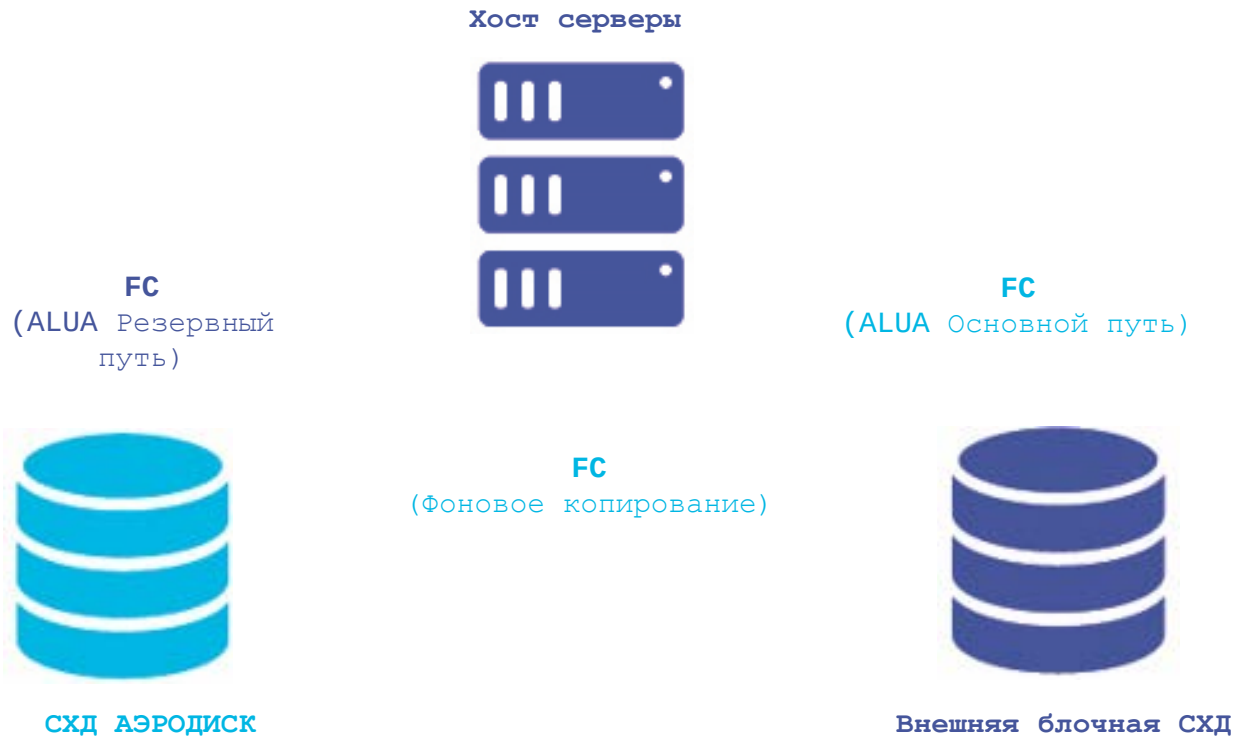
Функциональность: миграция блочных устройств «на лету»

СХД АЭРОДИСК поддерживает функционал миграции блочных устройств на RDG и DDP «на лету» прозрачно для конечного потребителя. Функционал может быть полезен для перемещения данных на другой уровень RAID, например, с RAID5 на RAID10 для увеличения производительности. Так же данные могут быть перемещены на другой тип дисков, например, с NLSAS дисков на SSD диски. Возможны любые направления перемещения блочных устройств между типами дисков и типами рейдов.

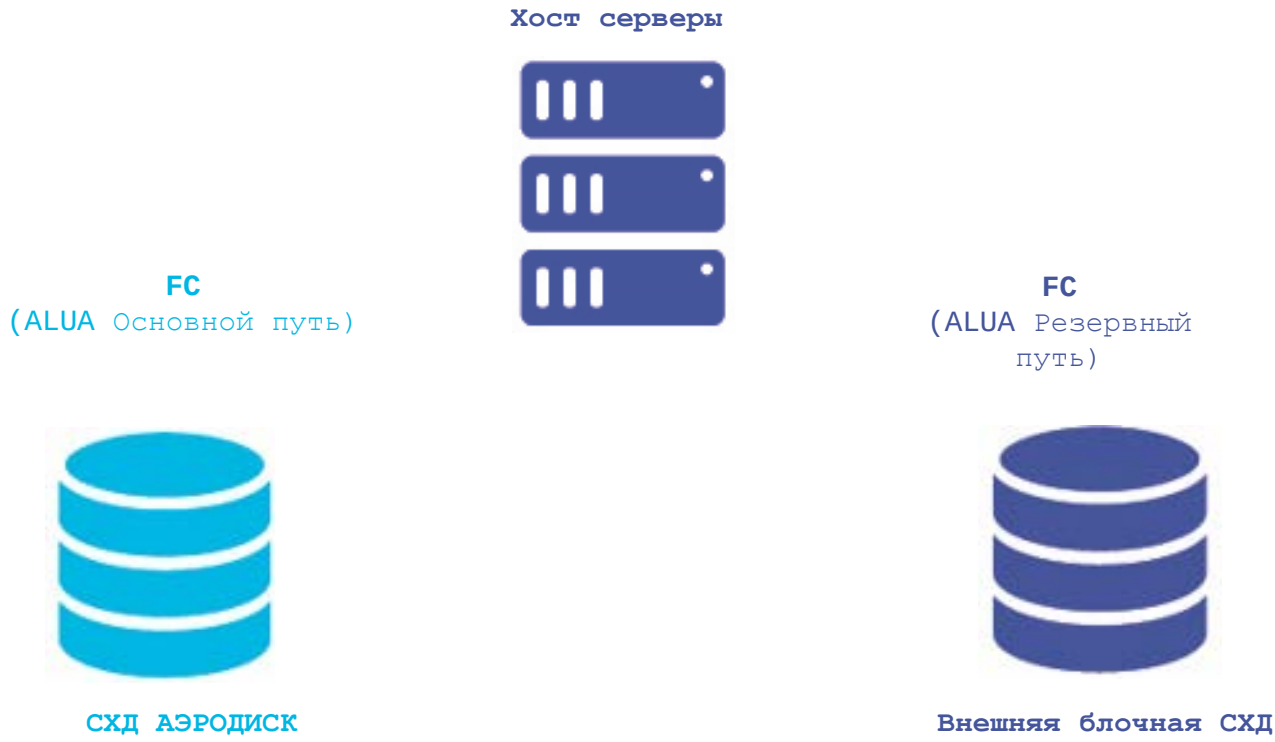
Функциональность: миграция данных с внешних СХД

СХД АЭРОДИСК В позволяет провести онлайн миграцию данных блочных устройств с внешних СХД.

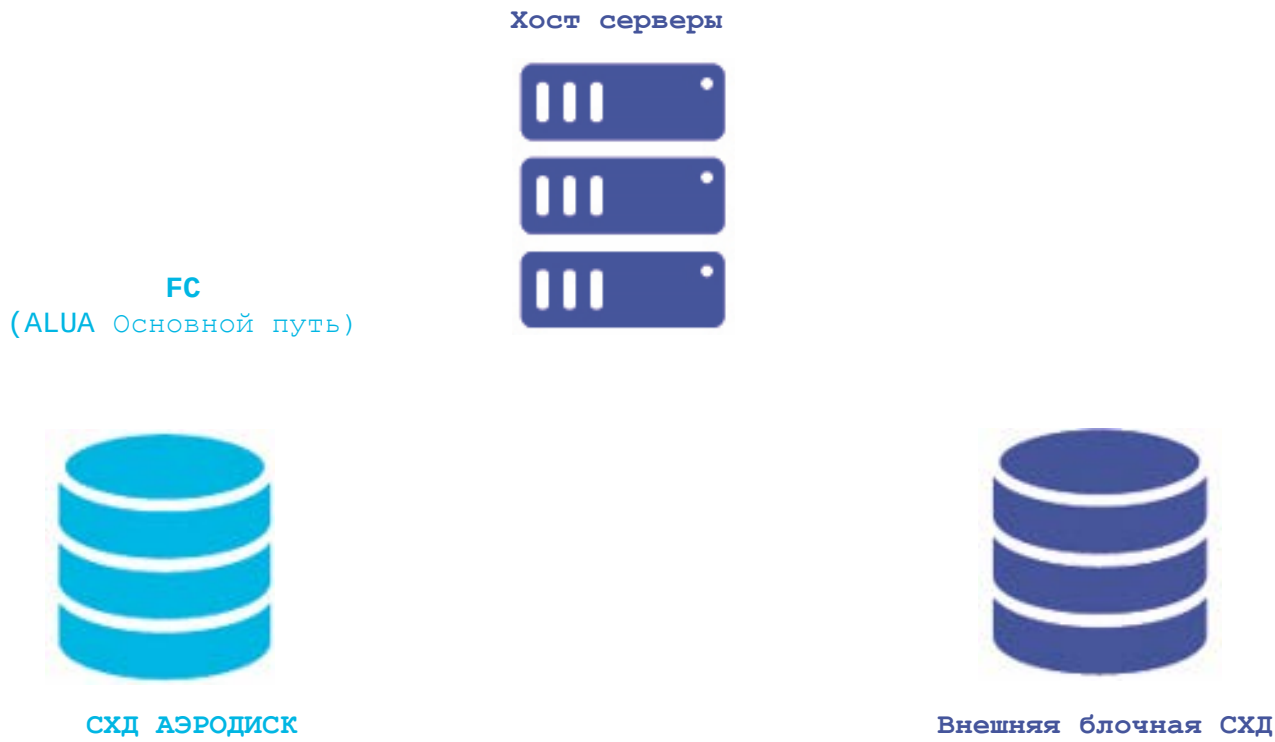
Миграция проходит в несколько этапов. На первом этапе на СХД АЭРОДИСК создается точная копия блочного устройства с внешней СХД. Блочное устройство с внешней СХД презентуется СХД АЭРОДИСК и начинается внутреннее фоновое копирование данных с внешней СХД на блочное устройство на СХД АЭРОДИСК. Хосту презентуется блочное устройство с СХД как еще один путь до существующего блочного устройства с параметром «Резервный».



После завершения фонового копирования данных резервный путь до LUN на СХД АЭРОДИСК становится основным и весь ввод/вывод переключается на него.



Последним шагом является отключение резервного пути старой блочной СХД.



Все манипуляции с путями и фоновым копирование для хоста являются полностью прозрачными.

Функциональность: Репликация

Репликация является функцией, которая обеспечивает защиту данных, используя 2 и более СХД на различных площадках.

Репликация является необходимым компонентом технического решения, если требуется организовать план аварийного восстановления (DRP) на резервной площадке.

В СХД АЭРОДИСК можно использовать 2 режима репликации – **синхронный и асинхронный.** Репликация всегда выполняется через порты Ethernet.

Синхронная репликация обеспечивает абсолютную идентичность данных на обеих или более СХД.

При синхронной репликации транзакции записи применяются только после подтверждения их записи на всех участниках репликации, поэтому для синхронной репликации следует использовать каналы связи с высокой пропускной способностью и низкими задержками.

Синхронная репликация выполняется на уровне блочного устройства LUN. Для файловых систем синхронная репликация не поддерживается.

Асинхронная репликация обеспечивает идентичность данных на СХД с задержкой, которая зависит от качества канала связи. Степень отставания полученных данных от исходных задавать нельзя.

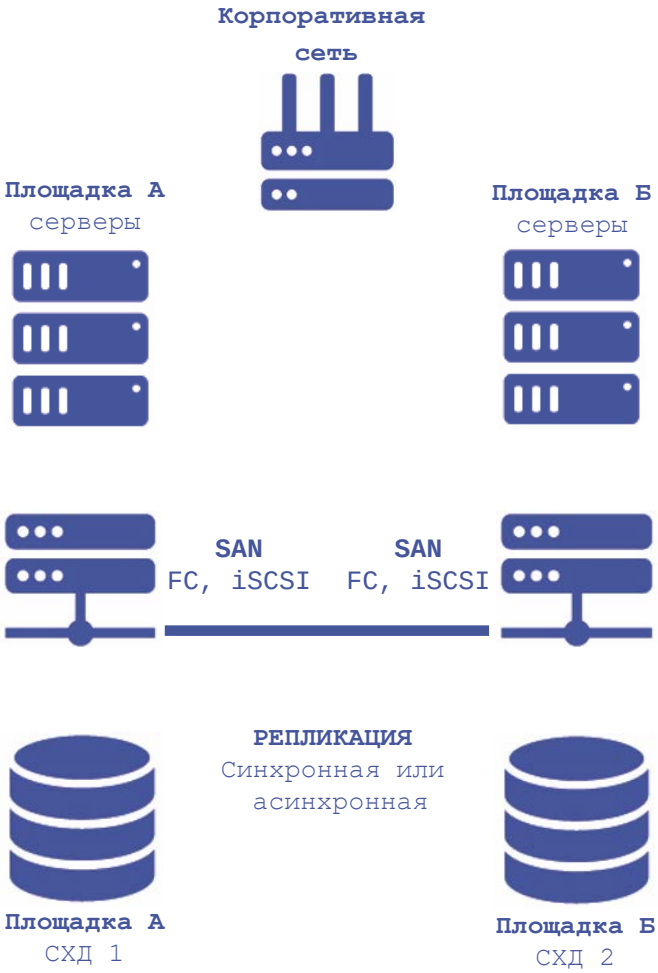
При асинхронной репликации транзакции записи вначале подтверждаются и применяются на исходной СХД, и только после этого происходит передача данных на другие СХД. После получения реплики, получатели подтверждают и применяют транзакции.

Для оптимизации передаваемого трафика используется автоматическая компрессия данных.

Исходя из этого, для асинхронной репликации не требуются каналы связи с высокой пропускной способностью и низкими задержками.

Синхронная и асинхронная репликацию могут использоваться одновременно. Режим репликации задается для каждого блочного устройства по отдельности.

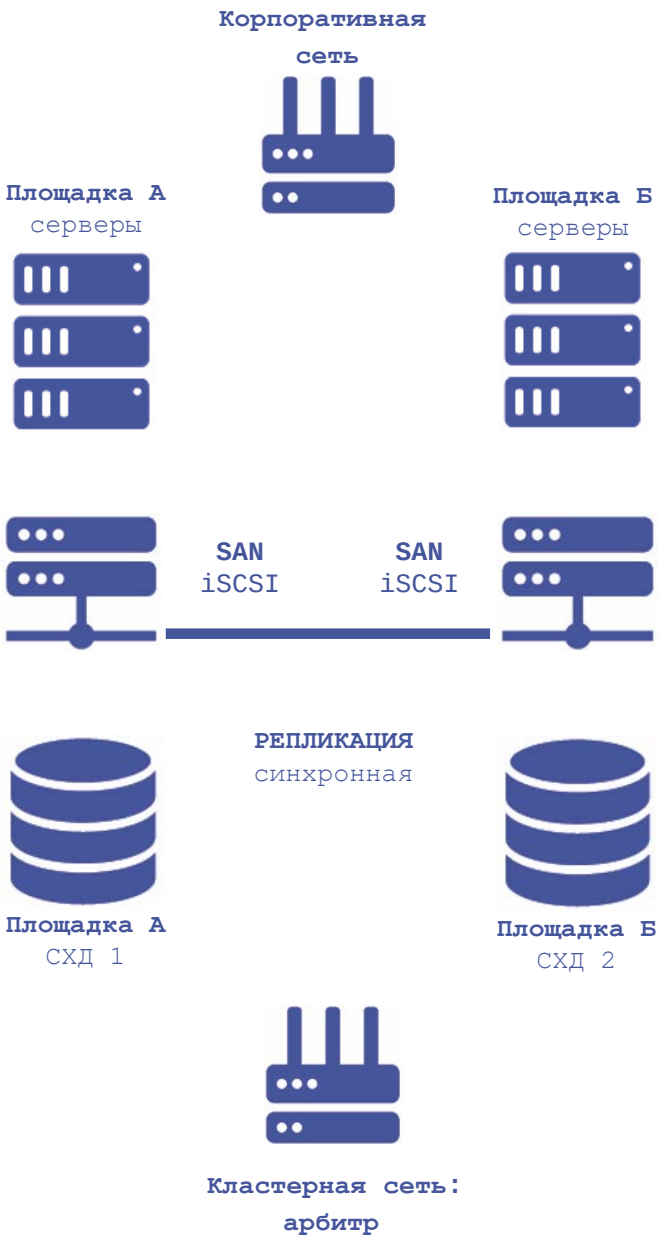
Функциональность: репликация



Для репликации доступны следующие топологии: 1:1, 1:n, n:1, n:m. Хосты могут подключаться к СХД по FC и iSCSI. Каждый хост подключается только к локальной СХД.

Функциональность: Метрокластер

Для автоматизации процессов переключения между площадками можно создать метрокластер. В этой архитектуре присутствуют 2 СХД похожие, но не обязательно однотипные СХД АЭРОДИСК и арбитр, который управляет переключениями между площадками. Арбитр представляет из себя виртуальную машину, которая может работать на любом популярном гипервизоре: АИСТ, KVM, ESXi, Hyper-V. При работе в режиме метрокластера серверы подключаются к СХД только по протоколу iSCSI причем каждый сервер должен иметь доступ к обеим СХД. Пример организации метрокластера представлен на картинке ниже.



Функциональность: Ускорение ввода/вывода для HDD дисков

Для реализации максимальной производительности и гибкости в СХД АЭРОДИСК предусмотрена функция ускорения ввода/вывода для HDD дисков. Данная функция разделяется на три под-функции:

- SSD-кэш чтение и запись (SSD RW);
- SSD-кэш чтение и запись и хранение метаданных (SSD RW +MCACHE);
- Online-tiering (SSD Online-tiering).

SSD-кэширование для RDG

SSD-кэш или **SSD-кэш+MCACHE** логически разделяет RDG на 2 плана производительности:

- Стандартный – где используется один тип дисков и адаптация соответственно выполняется только на уровне оперативной памяти и только для операций чтения;
- Быстрый – где используются SSD диски для кэширования и/или online-tiering.

План производительности назначается автоматически на уровне RDG при добавлении SSD дисков в группу и применяется ко всем LUN-ам и ФС, работающим в данной RDG сразу после добавления.

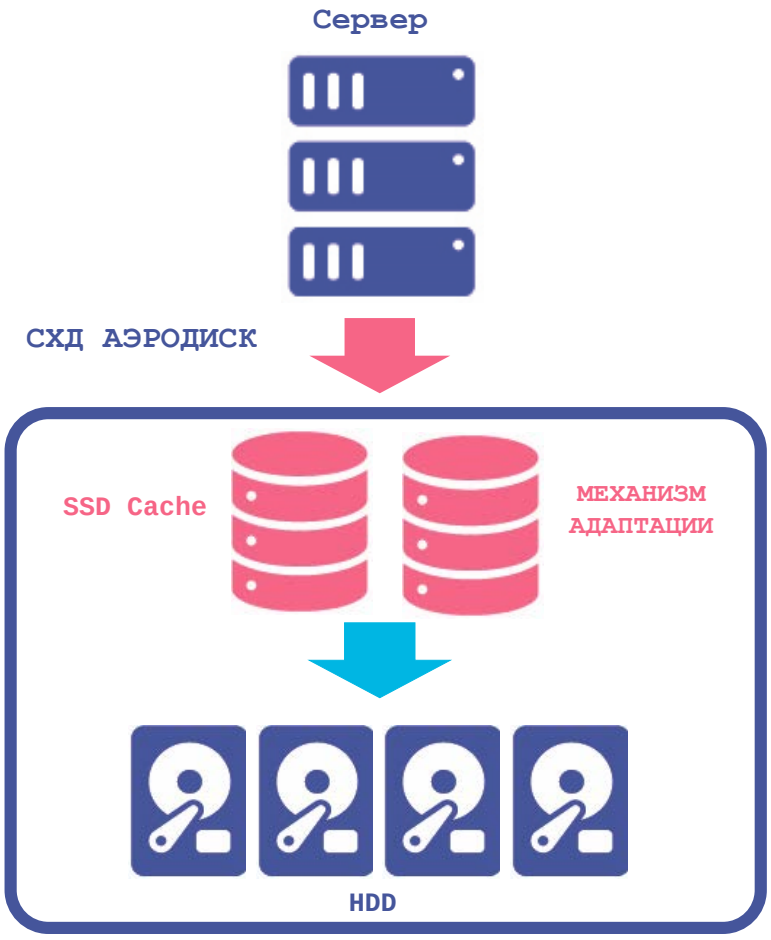
При создании гибридного хранилища SSD диски добавляются в кэш пул на запись/чтение (минимум 2 диска) в RAID1.

SSD-кэш работает во фронтальном режиме и по умолчанию применяется для всех транзакций. При этом чтобы исключить переполнение кэша, применяется механизм циклической адаптации (выталкивания) записей из кэша.

SSD-RW-кэш является достаточно экономичным вариантом повышения производительности СХД, поскольку не требует дисков большого объема (за счет постоянного выталкивания транзакций). При этом, поскольку данный механизм активно использует запись, это утилизирует ресурс надежности SSD-дисков (DWPD) и для данного типа кэша рекомендуется использовать SSD-диски с высоким показателем DPWD (3+).

Функциональность: ускорение ввода-вывода для HDD-дисков

На рисунке ниже приведен пример логики работы SSD-кэша.



Практическая информация о конфигурирование гибридного хранилища приведена в документе «АЭРОДИСК RAID-guide».

Системы **АЭРОДИСК** не имеет ограничения по объему SSD и RAM кэша, за исключением физического ограничения используемого оборудования

Функциональность: ускорение ввода-вывода для HDD-дисков

Online-tiering для RDG

Online-tiering хранение позволяет перемещать блоки данных между различными уровнями в зависимости от нагрузки на них, позволяя тем самым размещать более «горячие» данные (т.е. часто используемые) на быстрых дисках, а более «холодные» данные (т.е. редко используемые) на медленных.

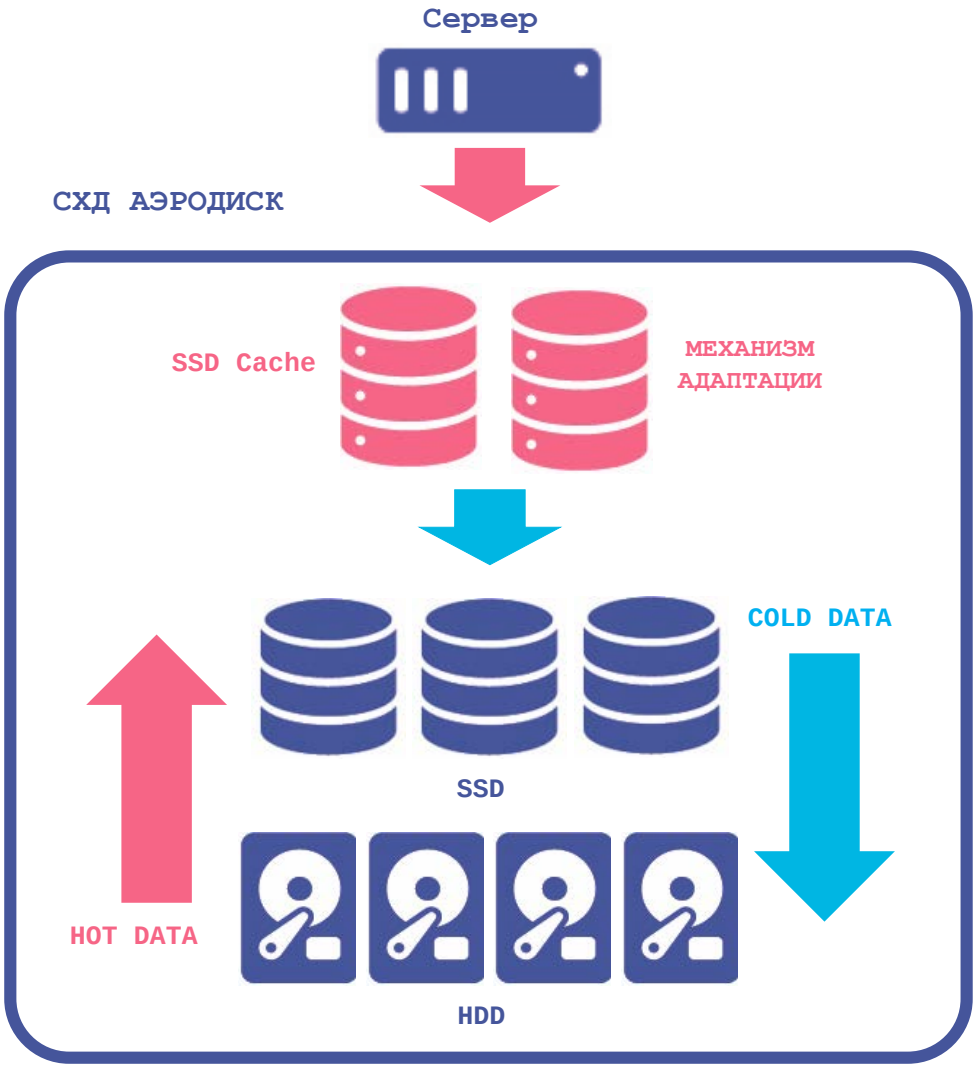
Перемещение блоков данных между уровнями происходит в онлайн-режиме.

Диски для многоуровневого хранения также добавляются на уровне RDG группы, после добавления дисков в online-tier группа меняет статус на «Быстрый».

Минимальное количество дисков на уровень online-tier – 2.

В отличие от механизма SSD-кэширования, данный функционал хранит данные на SSD-дисках пока к ним есть обращения, поэтому для этого механизма рекомендуется использовать не только надежные SSD-диски (DWPD 3+), но и SSD-диски большого объема.

На рисунке ниже приведен пример логики работы многоуровневого хранения.

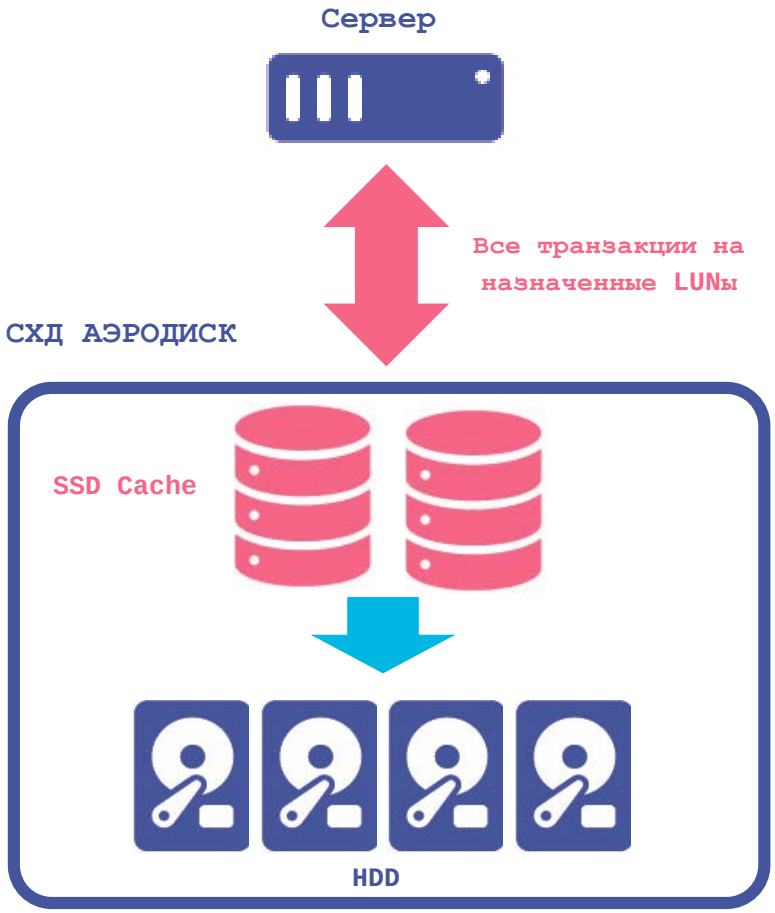


Функциональность: ускорение ввода-вывода для HDD-дисков

SSD-кэширование для DDP применяется на уровне LUN.

При создании гибридного хранилища SSD диски включаются в пул на чтение/запись (RW-CACHE). Одни и те же SSD диски могут быть использованы для кэширования операций ввода/вывода нескольких LUN. Минимальное количество SSD дисков в пуле – 2 штуки.

Через SSD-кэш проходят все операции ввода/вывода не зависимо от размера блока и чем больше кэш, тем выше будет производительность системы.

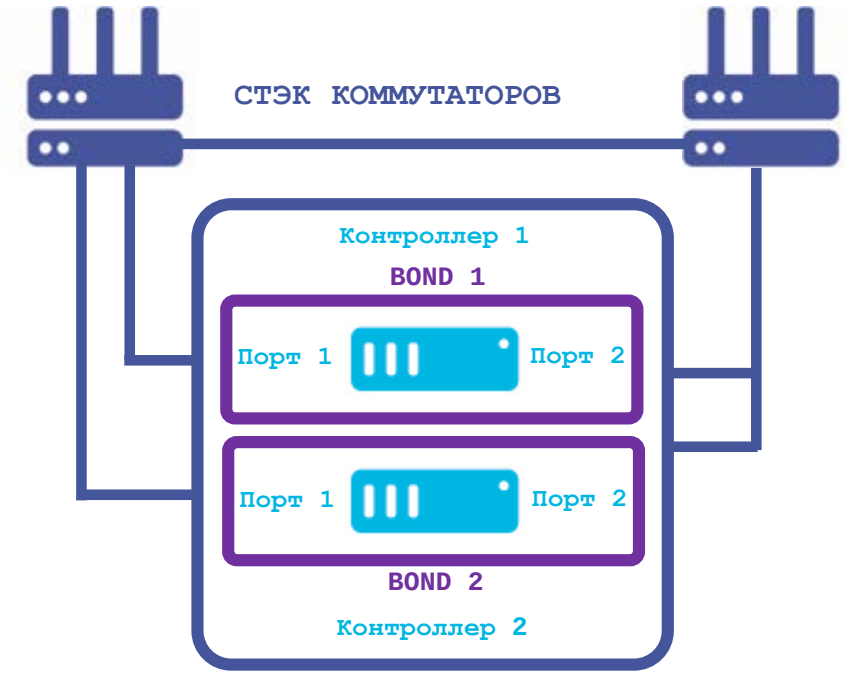


SSD-кэш для DDP активно использует запись, это утилизирует ресурс SSD-дисков (DWPD) и для данного типа кэша рекомендуется использовать SSD-диски с высоким показателем DPWD (3+).

VLAN и BONDING

Для ускорения операций ввода/вывода можно задействовать функционал объединения нескольких физических портов в один логический порт-BOND интерфейс. Поддерживаются как независимые от настроек коммутаторов BOND интерфейсы, так и зависимые от настроек коммутаторов BOND интерфейсы. Объединение нескольких физических интерфейсов дает увеличение пропускной способности, а также повышает уровень отказоустойчивости, так как в рамках BOND интерфейса физический порт может выйти из строя и обмен данными при этом не прекратится. В BOND интерфейс можно объединить до 16 физических однотипных интерфейсов

Для разграничения сетевого доступа, а также для более гибкой настройки СХД под сетевую инфраструктуру заказчика можно задействовать механизм тегирования трафика - VLAN. VLAN могут быть назначены как на физические сетевые интерфейсы, так и на BOND интерфейсы. VLANы могут быть применены как для файловых шар для протоколов NFS/CIFS, так и для блочного доступа по iSCSI.

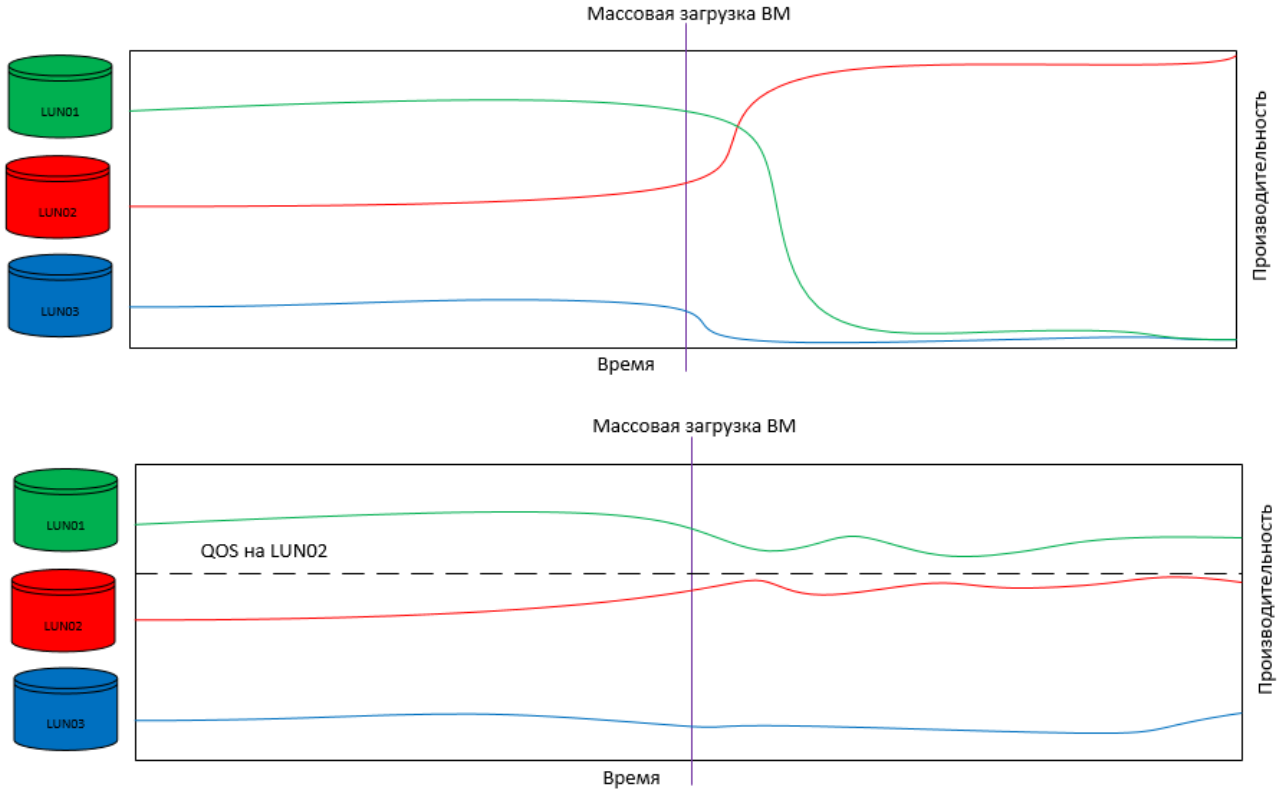


СХД АЭРОДИСК

«ПОРТ 1» и «ПОРТ 2» объединены в BOND 2x10 Гбит/с LACP на Контроллерах 1 и 2. VLAN – 10, 100, 1000 доступны на BOND интерфейсах

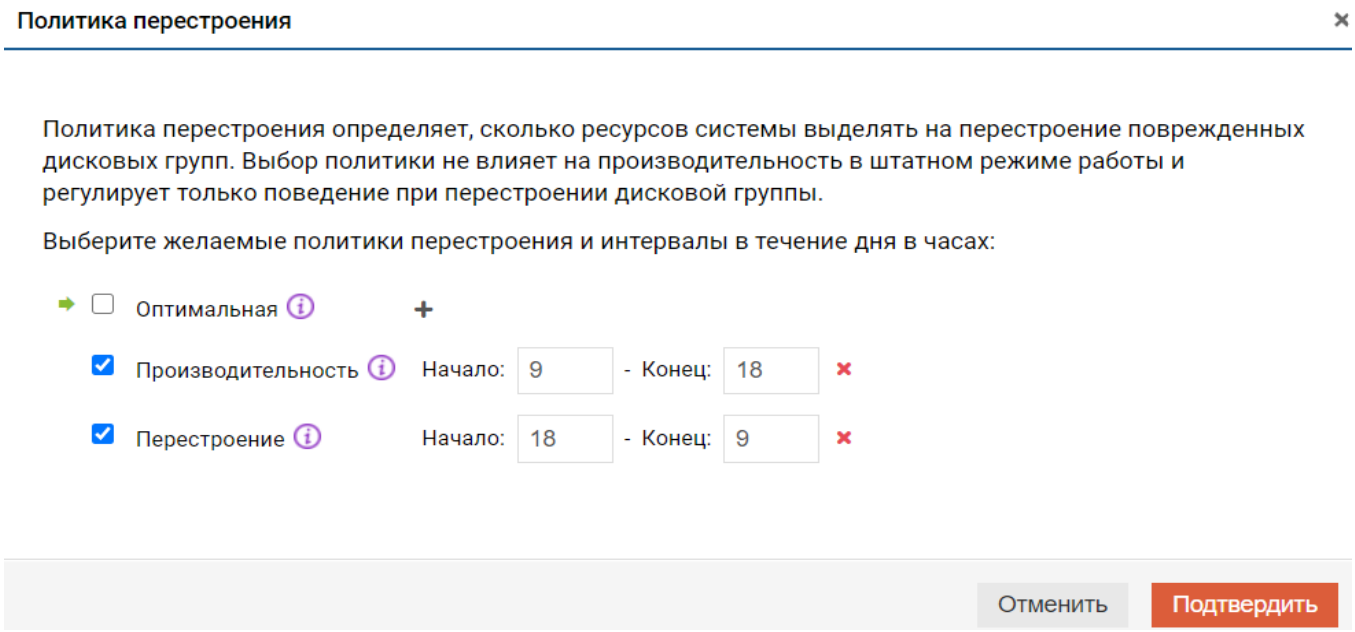
Quality of Service

Настройка качества обслуживание (QoS) позволяет минимизировать эффект «шумного соседа». При правильной настройке этого параметра можно гарантировать, что все потребители ресурсов СХД будут работать так, как ожидает администратор системы. QoS в СХД АЭРОДИСК назначаются на блочные устройства. Параметры качества обслуживания назначаются на уровне каждого конкретного LUN и могут ограничивать его потребление ресурсов в MB/s и IOPS. На картинке ниже приведен пример установки ограничений на LUN02, на котором по расписанию стартует множество ВМ, что является распространённым сценарием при использовании VDI.



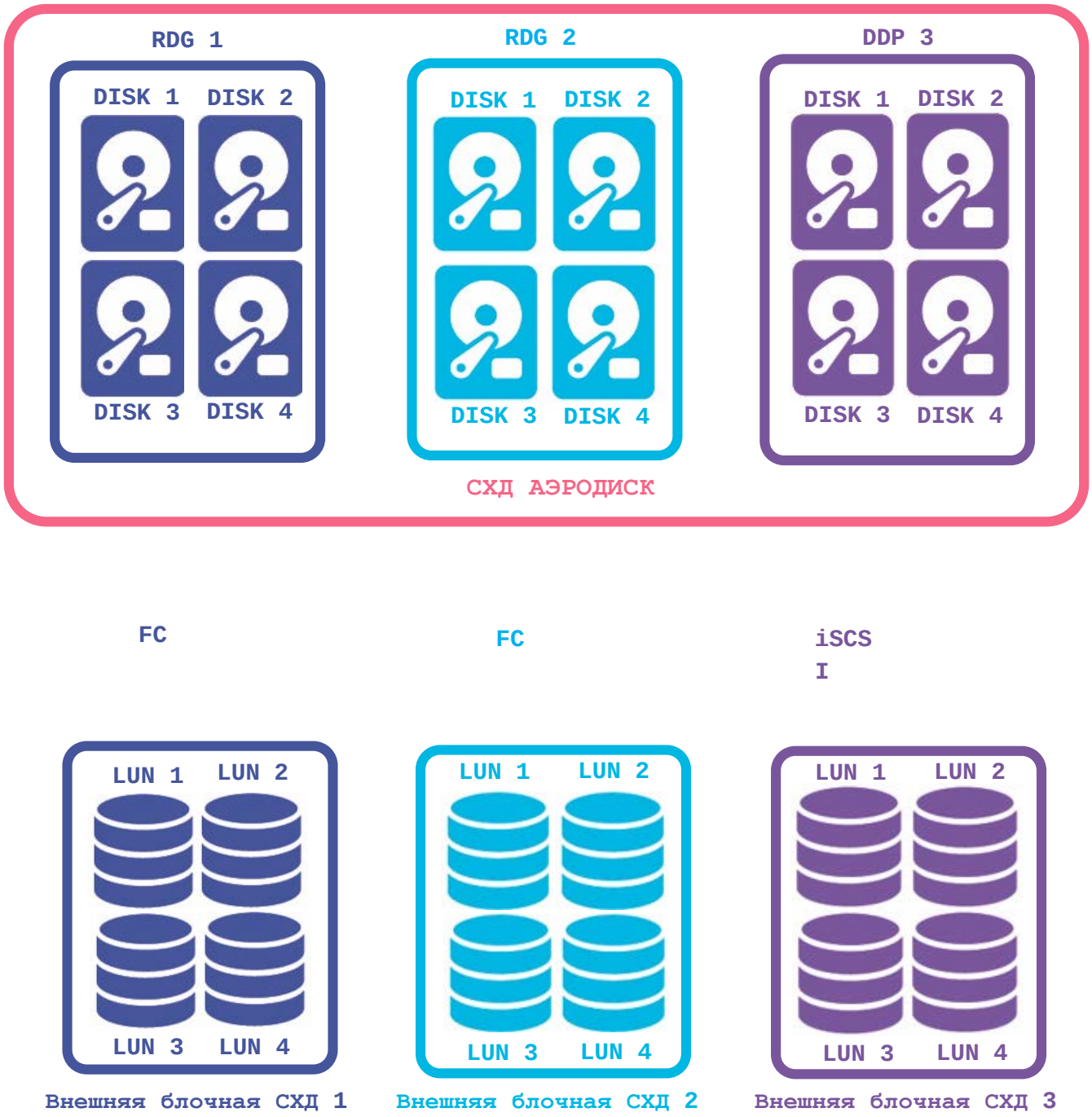
Политика перестроения

При выходе из строя диска в RAID-группе автоматически начинается процесс ее перестроения. При перестроении как правило может страдать общая производительность СХД, то есть могут страдать конечные потребители ресурсов СХД. Чтобы минимизировать эффект от перестроения администратор системы может назначить политику перестроения рейдов, в том числе заданную по расписанию. В системе присутствует 3 варианта политики перестроения, чтобы можно было гранулярно управлять скоростью перестроения рейдов.



Виртуализация сторонних СХД

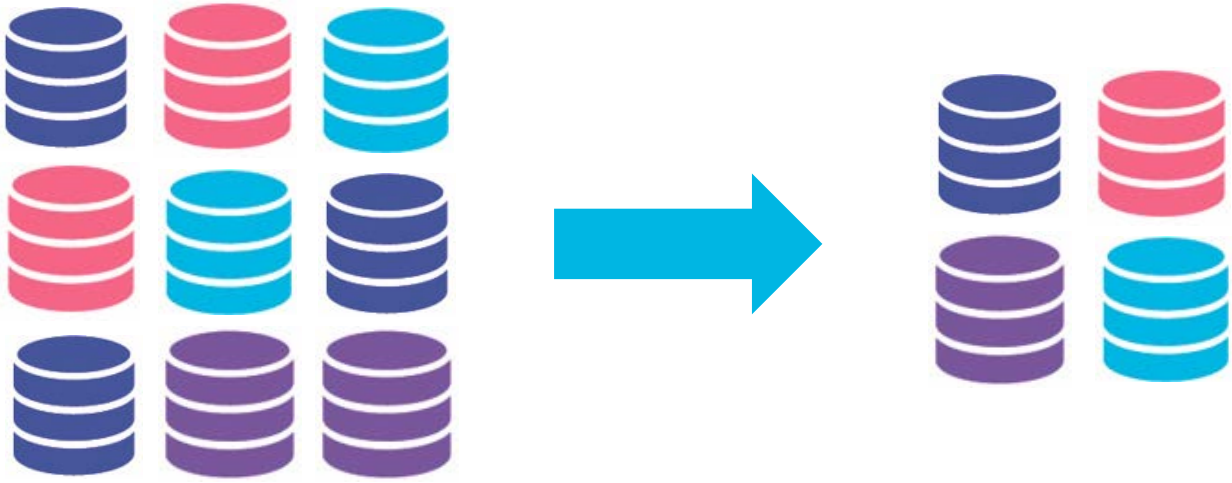
СХД АЭРОДИСК позволяют виртуализовать дисковую емкость сторонних СХД по проколам FC/iSCSI. Для этого на внешних СХД необходимо создать набор блочных устройств, который будет презентован СХД АЭРОДИСК. После этого на презентованных блочных устройствах можно делать стандартные для СХД АЭРОДИСК виртуальные рейды: RDG и/или DDP, на которых впоследствии можно создавать объекты хранения: блочные устройства и файловые шары.



Дедупликация

Дедупликация - это процесс устранения дублей блоков данных при сохранении уникальных блоков для экономии дискового пространства.

На рисунке ниже приведен результат работы дедупликации.



- В системах АЭРОДИСК применяется онлайн дедупликация с фиксированным блоком;
- Дедупликация работает и на RDG, и на DDP;
- Для RDG можно включать на конкретный LUN, для файловых шар только на группу целиком;
- Для DDP включается на конкретный LUN;
- Делит входящие данные на равные блоки;
- Устраняет дубли только когда блоки на 100% совпадают;
- Не создает высокую нагрузку на системные ресурсы;

Функциональность: дедупликация

Процесс дедупликации происходит следующим образом:

- Определение данных для дедупликации;
- Проверка доступности необходимого объема кэш памяти (SSD или RAM);
- Наборы данных сохраняются в таблице дедупликации при сохранении их контрольных сумм;
- при создании дубля данных система вместо выделения нового дискового пространства под дубль добавляет ссылку в таблицу дедупликации, которая указывает на реально существующие данные, вместо того чтобы создавать их дубли.

В зависимости от сферы применения и характера записи дедупликация может снизить потребляемый объем дискового пространства от 20% до 40%

Дедупликация выполняется на блочном уровне, что особенно применимо для больших объемов похожих данных. Например, при дедупликации хранилища виртуальных машин (ВМ) в облаке, уникальными, как правило, являются только некоторые блоки данных, а идентичные данные, такие как гостевые ОС, шаблоны ВМ, клоны ВМ и пр. являются дублируемыми и, соответственно, при дедупликации не потребляют дополнительного объема.

Для 1 ТБ дедуплицируемых данных нужно резервировать 1 GB ОЗУ на хэш таблицу. Для SSD дисков это не существенный объем, а вот для оперативной памяти наоборот. Т.к. объем оперативной памяти СХД ограничен, то рекомендуется использовать дедупликацию при наличии SSD дисков в СХД.

Компрессия

Для экономии места на СХД можно использовать механизм компрессии транзакций. Компрессия транзакций работает в онлайн режиме, то есть данные записываются на диски уже в оптимизированном виде. Так как система оптимизирует размер хранимых данных еще до записи на диски, то в ряде случаев включение этой функции может увеличить количество операций ввода/вывода, так как физических записей/чтений на диски становится меньше.

Для выполнения компрессии транзакций на лету используются выделенные ядра процессора и процесс компрессии никогда не конкурирует за ресурсы. В случае если количество операций ввода/вывода велико и ресурсов выделенного ядра перестает хватать, система на лету автоматически выделяет под процесс компрессии дополнительные выделенные ядра.

Авто-поддержка

Для обеспечения максимальной доступности систем хранения АЭРОДИСК предусмотрена функция автоматической поддержки. Данная опция обеспечивает:

- постоянный проактивный мониторинг всех компонентов СХД;
- автоматическую отправку диагностической информации в АЭРОДИСК в случае сбоя;
- автоматическое открытие обращений (тикетов) в АЭРОДИСК.

Открытие обращений производится путем отправки диагностической информации в виде почтовых уведомлений от контроллеров СХД на серверы поддержки АЭРОДИСК. После прихода данной информации сообщения автоматически преобразуются в тикет и регистрируется, далее специалист АЭРОДИСК, имея необходимую входную информацию, приступает к работе по устранению сбоя.