Архитектура

Программного обеспечения

BlazeX

Версии 3.6 v1

BITBLAZE

ООО «Битблэйз Технологии» (ООО «БитТех»)
ОГРН 1177746396630, ИНН 7731360971 / КПП 773101001
121205, Российская Федерация г. Москва, вн. тер. г. муниципальный округ Можайский, тер. Инновационного центра «Сколково», Большой бульвар, д. 42, стр. 1, 599 р/м 02, этаж 1

https://bitblaze.tech/

ОГЛАВЛЕНИЕ

СОГЛАШЕНИЕ ПО ОФОРМЛЕНИЮ

Для более наглядного представления различных команд, файлов и пр. в документе используется следующее форматирование:

```
Команды и командные утилиты
```

```
Параметры команд и файлов
```

Абзацы в тексте, содержащие важную информацию, выделены следующим образом:

> ПРИМЕЧАНИЕ
>
> Выделенные таким образом указания содержат важную информацию

ВВЕДЕНИЕ

Данная схема архитектуры управляющего ПО BlazeX содержит сведения о компонентах управляющего ПО, и является обязательным документом для ознакомления перед началом работ.

Программное обеспечение BlazeX функционирует в составе ПАК СХД и предназначено для управления распределением дискового пространства и мониторинга рабочих процессов и состояния ПАКа.

BlazeX автоматизирует процесс работы с дисковым пространством путем:

- виртуализации нескольких физических накопителей данных в логическую группу (пул) для повышения отказоустойчивости и (или) производительности;
- организации отдельных областей данных поверх физических групп, отображаемые системой как отдельные блочные устройства (логические тома).

ОБЩЕЕ ОПИСАНИЕ

ПО BlazeX позволяет создавать высокопроизводительные отказоустойчивые СХД и применяется в составе управляющего ПО ПАКов. BlazeX предназначено как для управления, так и для мониторинга СХД в одно- или двухконтроллерном исполнении.

ПО BlazeX поддерживает как одноконтроллерный режим работы, так и двухконтроллерный, при котором оба контроллера активны и имеют постоянный доступ к общей корзине накопителей. Отказоустойчивость системы в таком режиме работы обеспечивается за счет горячего резервирования, позволяющего сохранить доступ ко всем ресурсам (группам дисков, логическим томам) при отказе одного из контроллеров и обеспечить следующие параметры надежности работы:

- защиту от выхода из строя аппаратных компонентов одного узла;
- защиту от отказа интерфейса подключения;
- защиту от сбоев ОС и ПО на отдельном контроллере.

ПО BlazeX позволяет реализовать сетевое хранилище данных (NAS), объединенное с сетью хранения данных (SAN).

Управление ПАК с установленным ПО BlazeX осуществляется через веб-интерфейс с предоставлением инструментария ГИП. Описание процесса работы с ГИП изложено Руководстве пользователя.

Обмен информацией о состоянии между узлами СХД осуществляется через heartbeat, обмен трафиком между контроллерами организован через высокоскоростное соединение (interconnect).

1 СХЕМА АРХИТЕКТУРЫ

Управляющее ПО BlazeX строится на принципах микросервисного подхода, что обеспечивает гибкость, масштабируемость и отказоустойчивость системы. Все взаимодействие пользователя с системой начинается с nginx, который выполняет роль API Gateway - это единая точка входа, отвечающая за маршрутизацию запросов, а также за базовые функции безопасности. Такой подход позволяет централизованно управлять доступом и упростить масштабирование фронтенда.

Когда пользователь обращается к управляющему интерфейсу, nginx определяет, на каком из узлов сейчас активен blazex-control. Если сервис доступен локально, запрос сразу перенаправляется на него, если нет - происходит редирект на нужный узел через механизм heartbeat, что позволяет реализовать прозрачное переключение между узлами и избежать split-brain. Это важно для высокой доступности и отказоустойчивости, пользователь всегда попадает на рабочий экземпляр управляющего сервиса.

blazex-control реализует все API для управления СХД, а также содержит очереди команд. Он работает по принципу REST API, что обеспечивает простоту интеграции и расширяемость. Для изменения конфигурации blazex-control взаимодействует с yaml-config service, который отвечает за хранение и репликацию конфигурации между узлами. Сначала запись производится локально на первом узле, затем изменения реплицируются на второй узел, только после этого пользователю возвращается подтверждение. Такой подход гарантирует консистентность конфигурации и минимизирует риск рассинхронизации между узлами.

Для выполнения пользовательских операций (создание, изменение, удаление ресурсов) blazex-control отправляет команды в blazex-agent. Этот агент запускается на каждом узле и отвечает за непосредственное взаимодействие с инфраструктурой СХД. blazex-control исполняет команды, сканирует состояние ресурсов и через защищенный WebSocket (wss) отправляет результаты обратно на nginx, чтобы пользователь мог видеть актуальное состояние системы в реальном времени. Такой push-подход снижает нагрузку на API и ускоряет отображение изменений пользователю.

Все команды и результаты сканирования, если они отличаются от предыдущих, записываются в событийную базу данных. Для обеспечения целостности и доступности данная БД реплицируется между узлами с помощью DRBD, что позволяет быстро восстановить работу в случае сбоя одного из узлов.

Кластерная отказоустойчивость реализуется через Pacemaker и ocf-agent. На каждом узле работает свой агент, который умеет стартовать, останавливать и мониторить сервисы. В случае сбоя одного из узлов, Pacemaker автоматически переводит все необходимые сервисы на рабочий узел, используя актуальную конфигурацию из yaml-config. Это позволяет обеспечить непрерывность работы и минимизировать простой для пользователя.

Архитектура ПО BlazeX построена на двухконтроллерной кластерной модели, обеспечивающей высокую доступность и отказоустойчивость.

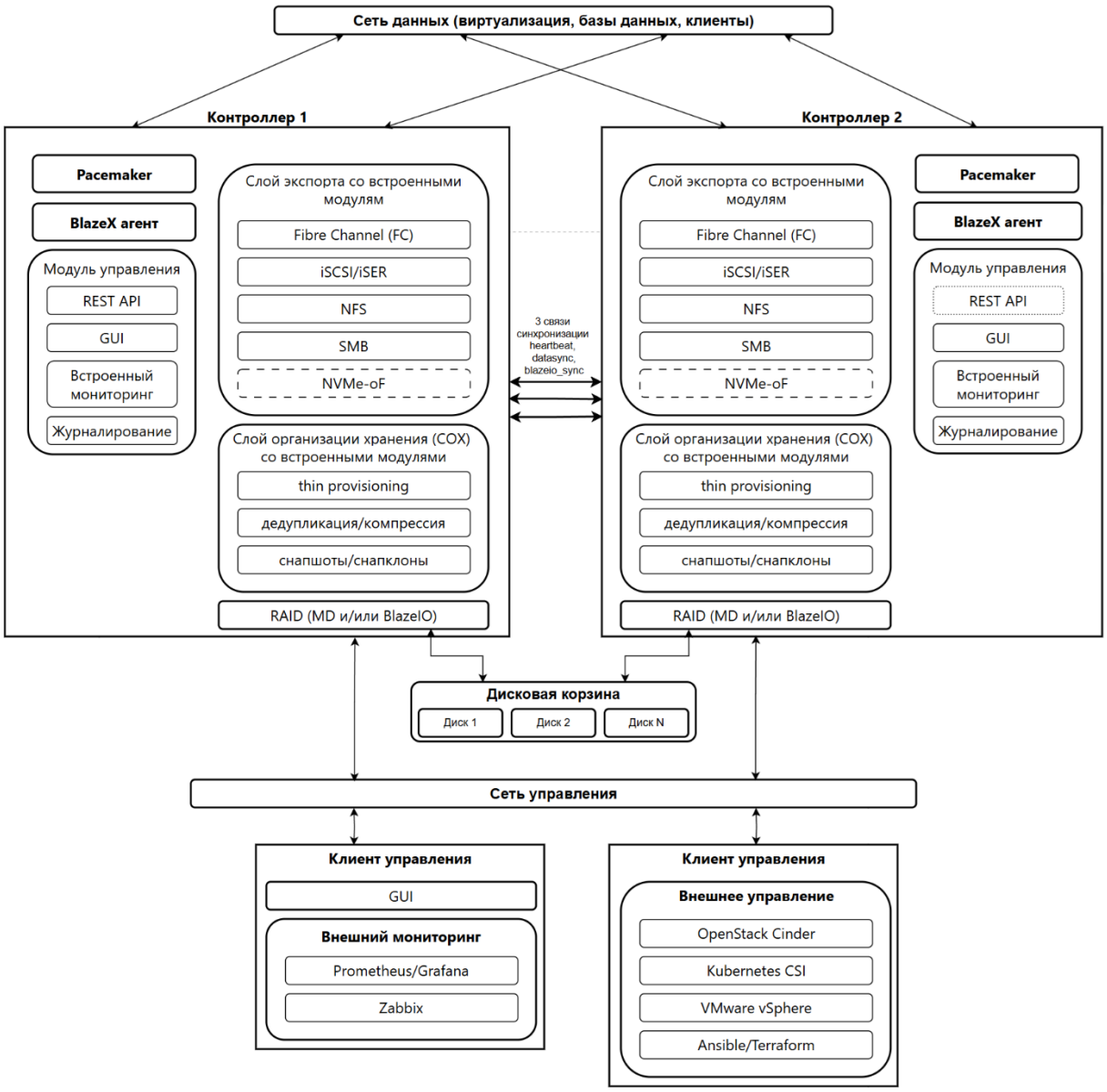


Рисунок 1 - Схема архитектуры управляющего ПО BlazeX

2 КОМПОНЕНТЫ

Каждый контроллер включает следующие ключевые компоненты:

2.1 Кластерный агент на базе Pacemaker

Pacemaker - это кластерный менеджер ресурсов, который выполняет функции кластерного агента: обеспечивает мониторинг состояния контроллера (heartbeat), участие в кластерных механизмах, управление плавающими сервисами и автоматический failover. Включает описанное выше ПО управления с REST API и графическим интерфейсом (GUI), а также подсистемы встроенного мониторинга и журналирования. Между контроллерами настроены два выделенных канала связи для синхронизации состояний (heartbeat) и обмена данными кластерной базы данных (database), что предотвращает возникновение ситуации split-brain и обеспечивает согласованность конфигурации. Pacemaker управляет всеми сервисами слоя экспорта, обеспечивая их высокую доступность.

2.2 Слой экспорта

Предоставляет унифицированный доступ к данным через стандартные протоколы хранения. В его состав входят:

- Fibre Channel (FC) - высокоскоростной протокол для подключения серверов к системам хранения данных на выделенной оптической инфраструктуре с низкими задержками.
- iSCSI/iSER - протоколы для передачи SCSI-команд по IP-сетям (iSCSI) с возможностью ускорения через RDMA (iSER) для снижения нагрузки на процессор.
- NFS - сетевой файловый протокол для UNIX/Linux-систем, позволяющий монтировать удалённые файловые системы локально.
- SMB - протокол файлового доступа, широко используемый в средах Windows для общего доступа к файлам и принтерам.
- NVMe-oF - протокол для высокопроизводительного доступа к NVMe-накопителям через сеть с минимальными задержками (не реализована в 3.6).

2.3 Слой организации хранения (СОХ)

Реализует расширенные функции управления данными поверх физических накопителей. К ним относятся:

- Thin Provisioning (тонкое выделение ресурсов)
- Дедупликация и сжатие данных
- Создание и управление снапшотами и снапклонами

2.5 Уровень RAID

Абстрагирует физические диски, обеспечивая отказоустойчивость и повышение производительности. Поддерживаются как стандартные программные RAID-массивы (MD), так и высокопроизводительное решение BlazeIO. Поверх RAID-групп формируются пулы хранения, которые затем используются COX-слоем для создания логических томов.

2.6 Дисковая корзина

Дисковая корзина представляет собой физический или виртуализированный блок, содержащий массив дисков (Диск 1, Диск 2, …, Диск N), который подключается к обоим контроллерам. Она обеспечивает единое дисковое пространство для уровня RAID, поддерживает горячую замену дисков и подключается к контроллерам через SAS-экспандеры или NVMe-oF (в будущем). Благодаря двухконтроллерному доступу достигается высокий уровень отказоустойчивости: при выходе одного контроллера другой продолжает работать с теми же дисками. Поверх дисковой корзины уровень RAID строит RAID-группы и пулы хранения.

2.7 Связи синхронизации

- **Heartbeat** - канал для постоянной проверки «живучести» соседнего контроллера, позволяющий кластеру мгновенно обнаружить сбой и запустить failover.
- **Datasync** - канал для синхронизации кластерной базы данных и конфигурации между контроллерами, обеспечивающий согласованность настроек и предотвращающий split-brain.
- **blazeio_sync** - высокоскоростной канал для синхронизации кэша, дедуплицированных данных и операций ввода-вывода, специфичный для движка BlazeIO.

2.8 Сеть управления

Управление и интеграция осуществляются через выделенную сеть управления, к которой подключаются:

- Клиенты управления с графическим интерфейсом (GUI) для оперативного администрирования.
- Внешние системы мониторинга (Prometheus/Grafana, Zabbix) для сбора метрик и наблюдения за состоянием системы.
- Внешние платформы оркестрации и управления инфраструктурой, включая OpenStack Cinder, Kubernetes CSI, VMware vSphere, а также инструменты автоматизации Ansible и Terraform, что обеспечивает глубокую интеграцию в современные ИТ-ландшафты и облачные среды.

3 ПРИНЦИП РАБОТЫ КЛАСТЕРА:

В штатном режиме оба контроллера активны и обслуживают входящие запросы на доступ к данным. При отказе одного контроллера происходит автоматический переход (failover) всех его ресурсов и сервисов на второй контроллер. Этот процесс управляется кластерным менеджером (Pacemaker) через агенты (ocf-agent) и использует актуальную конфигурацию, синхронизированную между узлами. Высокоскоростное соединение (interconnect) между контроллерами используется для обмена служебным трафиком и синхронизации данных, обеспечивая минимальное время восстановления и непрерывность доступа к данным.

3.1 Обзор BlazeIO

BlazeIO - программно-определяемая система хранения данных (SDS) с блочным доступом, предназначенная для высокопроизводительных рабочих нагрузок. Разработана компанией «Битблэйз Технологии» (в составе группы «BITBLAZE»).

Функциональность:

- **Блочный доступ** с поддержкой логических томов, отказоустойчивостью и предсказуемыми показателями производительности (IOPS, пропускная способность, задержка).
- **Symmetric Active-Active**
- **Высокая производительность** благодаря проприетарным алгоритмам, неблокирующим трактам передачи данных и оптимизации под NVMe-oF (NVMe over Fabrics).
- **Поддержка тонкого выделения ресурсов (thin provisioning)**, Erasure Coding,
- **Аппаратная независимость** - работает на стандартных серверах с NVMe SSD.

Поддерживаемые типы RAID (оптимизировано для SSD):

BlazeIO использует **Erasure Coding** и **зеркалирование (multi-active)** для обеспечения отказоустойчивости и высокой производительности. Система оптимизирована для работы с NVMe SSD и поддерживает:

- **RAID-подобные схемы на основе Erasure Coding** для быстрого восстановления данных и минимизации накладных расходов.
- **Thin volumes** для эффективного использования дискового пространства.
- **Деградированный режим (healthy degraded)** для минимизации последствий сбоев.

3.2 Обзор MD (Linux Multiple Devices)

MD - встроенный в ядро Linux драйвер программного RAID, предназначенный для объединения нескольких физических дисков в логические устройства.

Функциональность:

- **Создание и управление программными RAID-массивами** (уровни 0, 1, 4, 5, 6, 10).
- **Горячая замена дисков** и восстановление массивов.
- **Поддержка мониторинга** через утилиту mdadm.
- **Работа с блочными устройствами** - может использоваться с файловыми системами, LVM, шифрованием.

Поддерживаемые типы RAID (оптимизировано для HDD):

MD изначально разработан для вращающихся HDD и поддерживает классические уровни RAID:

- **RAID 0** (страйпинг) - для повышения производительности.
- **RAID 1** (зеркалирование) - для отказоустойчивости.
- **RAID 5/6** - для баланса производительности и надежности с контролем четности.
- **RAID 10** - комбинация зеркалирования и страйпинга.
- **Линейный режим** - простое объединение дисков.

Таблица сравнения MD и BlazeIO

Характеристика	MD (Linux MD RAID)	BlazeIO
Оптимизация под тип носителя	HDD	NVMe SSD
Режим доступа	Active-ALUA	Active-ALUA, Symmetric Active-Active
Типы RAID / защиты данных	RAID 0,1,5,6,10, линейный	Erasure Coding, более 2 дисков четности, тонкие тома (2+1), (|4+1), (|4+2), (8+1), (8+2), (8+3), (8+4), (16+1), (16+2), (16+3), (|16+4)
Производительность	Зависит от уровня RAID и дисков	Высокая, предсказуемая, с поддержкой NVMe-oF
Управление	Через mdadm и конфигурационные файлы	Через blazeio-ctl и конфигурационные файлы

- **MD** - классическое решение для программного RAID на HDD, простое и бесплатное, но ограниченное в масштабировании и оптимизации под современные SSD.

- **BlazeIO** - современная SDS-платформа, ориентированная на высокопроизводительные NVMe-накопители, с высокой отказоустойчивостью , подходящая для задач ИИ, аналитики, СУБД и виртуализации.

3.3 Обзор SMB

SMB (Server Message Block) - это сетевой протокол для предоставления общего доступа к файлам, принтерам и другим ресурсам, преимущественно используемый в операционных системах Windows. В версии 3.6 отсутствует интеграция с доменными службами (Active Directory/LDAP). Аутентификация выполняется исключительно через встроенную базу данных контроллера:

- Локальные учётные записи: Пользователи и группы создаются непосредственно на каждом контроллере (или синхронизируются через кластер).
- Управление доступом: Разграничение прав доступа к общим папкам (шарам) и файлам осуществляется на основе SID локальных учётных записей.
- Отсутствие внешних контроллеров домена: Система не может делегировать проверку подлинности внешним доменам.

3.4 Обзор модуля Thin Provisioning (тонкое выделение ресурсов)

Модуль Thin Provisioning позволяет эффективно управлять дисковым пространством за счет его динамического выделения по мере фактической необходимости, а не на основе предварительно зарезервированных объемов. В контексте хранилища, подобного описанному в документации, этот механизм обеспечивает:

1. **Динамическое выделение**: Физическое пространство на дисках выделяется для томов (LUN) или файловых систем не сразу в полном объеме, а постепенно, по мере записи данных. Это позволяет создать логические тома, суммарный размер которых превышает доступную физическую емкость (overcommitment), что повышает гибкость и эффективность использования ресурсов.

2. **Экономия ресурсов**: Значительно снижается потребление дискового пространства, так как неиспользуемые зарезервированные области не занимают физическое место на дисках. Особенно эффективно для сред с неполным заполнением томов (например, для виртуальных машин, тестовых сред).

3. **Простота управления**: Администратору не требуется точно прогнозировать рост данных для каждого тома. Емкость пула хранения может быть расширена физически, а тонкие тома автоматически получат доступ к новому пространству.

4. **Интеграция с функциями хранения**: Механизм тонких томов совместим со снэпшотами (с использованием модели ROW - Redirect-on-Write), что позволяет экономить пространство и при создании моментальных копий данных.

5. **Мониторинг и контроль**: Система предоставляет инструменты для мониторинга фактического и выделенного пространства, позволяя предотвратить исчерпание физической емкости пула.

3.5 Обзор модуля Deduplication/Compression (дедупликация/компрессия)

Модуль Deduplication/Compression предназначен для повышения эффективности использования дискового пространства путем устранения избыточных данных. В системе хранения эта функциональность реализуется следующим образом:

1. **Блочная дедупликация**: Система анализирует входящие данные, разделяя их на блоки фиксированного размера. Для каждого блока вычисляется уникальная контрольная сумма (хэш). Если новый блок имеет хэш, идентичный уже сохраненному, система не записывает повторные данные, а создает ссылку на существующий блок.

2. **Онлайн-обработка**: Процесс дедупликации может происходить в режиме реального времени (online) в момент записи данных, что минимизирует необходимость в последующей постобработке и экономит пространство с самого начала.

3. **Сфера применения**: Наиболее эффективна для данных с высокой степенью повторяемости, таких как:

    - Виртуальные машины (идентичные гостевые ОС, шаблоны, клоны).
    - Резервные копии.
    - Файловые хранилища с множеством похожих документов.

4. **Сжатие данных**: Часто используется в связке с дедупликацией. Алгоритмы сжатия дополнительно уменьшают объем уникальных данных перед записью на диск.

5. **Требования к ресурсам**: Для работы требуется выделение оперативной памяти под хэш-таблицы (примерно 1 ГБ на 1 ТБ обрабатываемых данных) и/или быстрые SSD-диски для кэширования метаданных, что ускоряет процесс поиска дубликатов.

6. **Эффективность**: В зависимости от характера данных, технология может снизить потребление дискового пространства **до 10 раз**.

Эта функциональность критически важна для построения экономичных хранилищ большой емкости, особенно в виртуализированных средах и системах резервного копирования.

3.6 Обзор модуля Snapshots/Snapclones (мгновенные снимки и клоны)

Модуль Snapshots/Snapclones обеспечивает создание мгновенных, пространственно-эффективных копий томов (LUN) или файловых систем для целей резервирования, тестирования или восстановления данных.

Мгновенные снимки (Snapshots):

- **Скорость создания**: Снимки создаются практически мгновенно, независимо от размера исходного тома, так как не происходит физического копирования данных.
- **Модель Copy-on-Write (COW) или Redirect-on-Write (ROW)**. При изменении данных в исходном томе после создания снимка, оригинальные блоки сохраняются для снимка, а новые данные записываются в другое место. Это минимизирует влияние на производительность.
- **Пространственная эффективность**: Изначально снимки не потребляют место. Объем растет пропорционально количеству измененных данных в исходном томе после создания снимка.
- **Планирование**: Возможность создания снимков по расписанию для организации частых точек восстановления (локальная репликация).

Клоны (Clones):

- **Связанные клоны (Thin Clones)**: Создаются на основе снимка. Изначально разделяют с исходным томом общие данные, занимая минимум места. При записи в клон используется механизм ROW/COW. Позволяют быстро развернуть доступную для чтения/записи копию данных.
- **Снэпклон (Snapclone)**: Гибридная технология, объединяющая свойства снимка и клона. Создается быстрее полного клона, изначально занимая емкость, равную используемой данными в источнике.

Функциональность восстановления:

- **Откат (Revert)**: Быстрое восстановление исходного тома или файловой системы до состояния на момент создания снимка.
- **Клонирование**: Присоединение снимка или связанного клона как нового независимого тома к хосту для выборочного восстановления отдельных файлов или данных.

Группы консистентности: Ключевая функция для сложных сред (например, баз данных, кластеров). Позволяет объединить несколько томов в группу и создавать для них **согласованные во времени снимки**, обеспечивая целостность данных между взаимосвязанными приложениями.

- Данный модуль является основой для построения стратегий защиты данных, их репликации, тестирования ПО и разработки без воздействия на производственную среду.

4 ОТКАЗОУСТОЙЧИВОСТЬ BLAZEX

Система BlazeX построена по активно-активной кластерной архитектуре с двумя контроллерами, обеспечивающей непрерывную доступность данных и сервисов.

4.1 Ключевые компоненты отказоустойчивости

- **Кластерный менеджер**
- **Многоуровневая система мониторинга здоровья**
- **Основной канал: Сеть управления**
  - Агенты постоянно обмениваются heartbeat-сообщениями
  - Общий мониторинг через внешние системы (Prometheus/Grafana, Zabbix)
- **Резервные каналы (при отказе сети управления):**
  - 3 специализированные связи синхронизации между контроллерами
    - Heartbeat-соединения (мониторинг жизнеспособности)
    - Database-соединения (синхронизация метаданных)
    - Blazeiosync-соединение (синхронизации кэша, дедуплицированных данных и операций ввода-вывода)
- **Сети данных (FC, iSCSI, NVMe-oF)** - как дополнительный канал диагностики

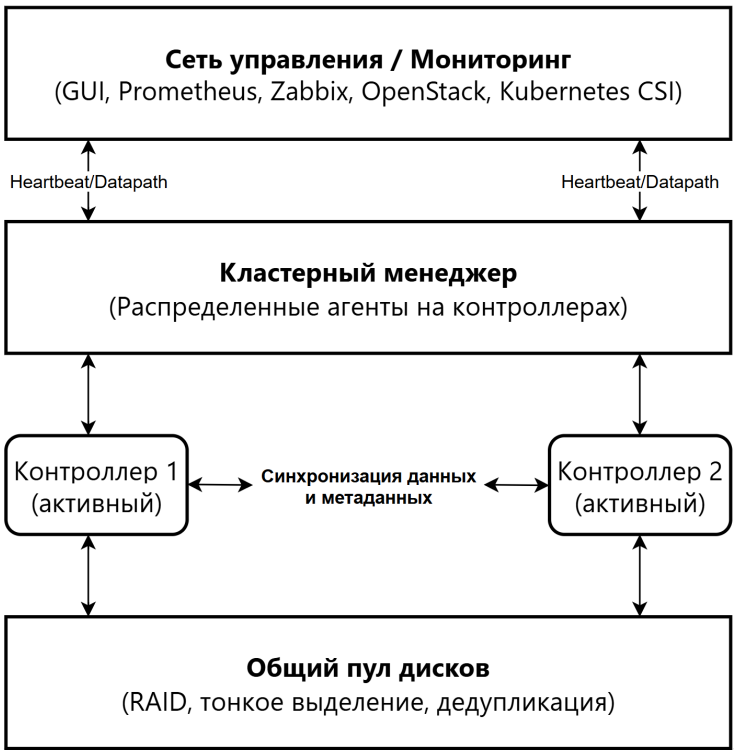
Механизм переключения при отказе:

1. Обнаружение сбоя через множественные каналы
2. Кворумное голосование между активными агентами
3. Автоматическое перераспределение нагрузки на работоспособный контроллер
4. Прозрачное восстановление при возврате узла в строй

Преимущества архитектуры

Компонент	Функция отказоустойчивости
Двойные контроллеры	Активно-активная конфигурация, нет единой точки отказа
Множественные сети	Избыточность каналов связи и мониторинга
Общий пул дисков	Независимый доступ к данным с любого контроллера
Синхронизация метаданных	Консистентность состояния между узлами
Внешний мониторинг	Дополнительный контроль и оповещение

Визуальное представление архитектуры:



4.2 Архитектура ALUA (Asymmetric Logical Unit Access)

Система BlazeX реализует **ALUA архитектуру** с двумя контроллерами, где каждый LUN имеет предпочтительный (оптимизированный) путь доступа и альтернативный (неоптимизированный) путь для обеспечения отказоустойчивости.

Ключевые принципы ALUA

1. Распределение дисков между контроллерами

- **Каждый контроллер** напрямую управляет "своими" дисками

2. Типы путей доступа

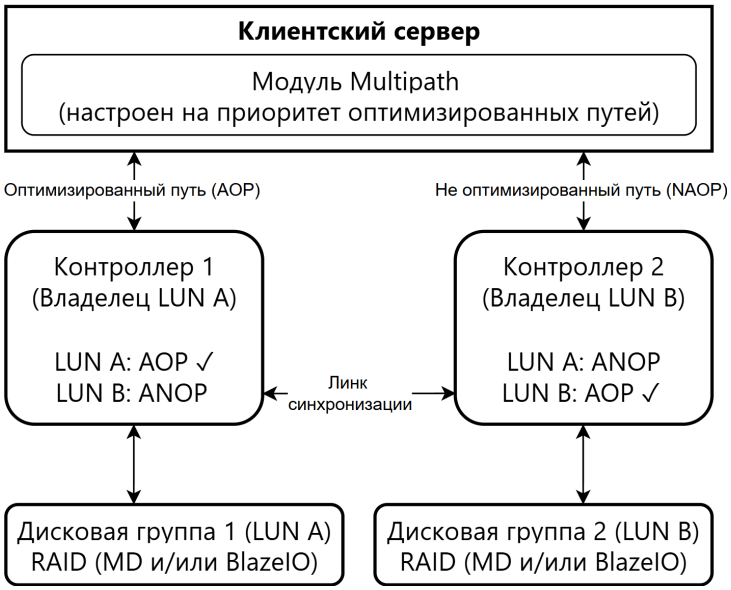
Оптимизированный путь (Active-Optimized)

- **Прямой доступ** к дискам через "владеющий" контроллер
- **Максимальная производительность** (низкая задержка, высокая пропускная способность)
- **Основной рабочий путь** для клиентских операций ввода-вывода

Неоптимизированный путь (Active-Non-Optimized)

- **Косвенный доступ** к дискам через "не владеющий" контроллер
- **Доступ через линк синхронизации** между контроллерами
- **Пониженная производительность** (дополнительная задержка)
- **Резервный путь** на случай отказа оптимизированного пути

Архитектурная схема ALUA



Механизм работы и отказоустойчивости

В нормальном режиме:

- Multipath-модуль на клиенте **обнаруживает все доступные пути**
- **Автоматически выбирает оптимизированный путь** для каждого LUN
- **Направляет I/O-запросы** через контроллер-владелец соответствующего LUN
- **Мониторит доступность** всех путей

При отказе оптимизированного пути (Отказ Контроллера 1):

1. Multipath-модуль обнаруживает недоступность пути к Контроллеру 1
2. Автоматически переключает трафик LUN A на неоптимизированный путь:
3. Клиент → Контроллер 2 → Линк синхронизации → Контроллер 1 → LUN A
4. Производительность LUN A временно снижается (ANOP режим)
5. LUN B продолжает работать через оптимизированный путь (Контроллер 2)

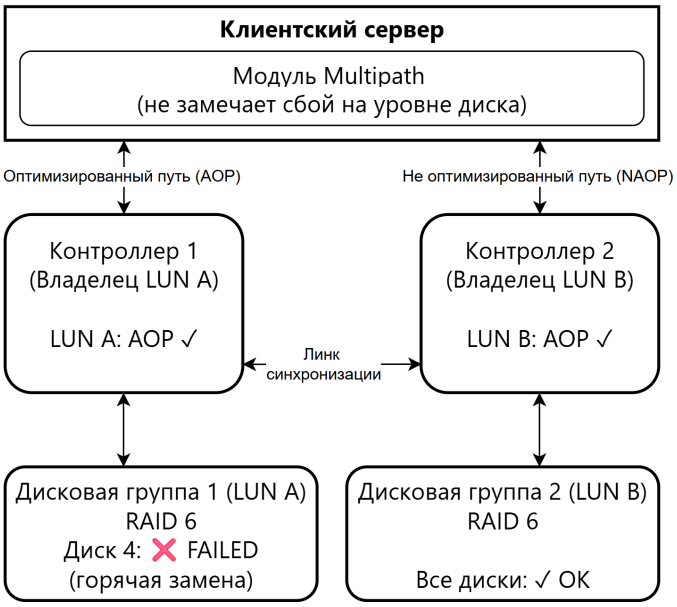
При восстановлении контроллера:

1. Контроллер возвращается в строй
2. Multipath обнаруживает восстановление оптимизированного пути
3. **Автоматически возвращает трафик** на оптимизированный путь
4. **Балансировка нагрузки** восстанавливается

4.3 Отказоустойчивость ALUA

Сценарии отказоустойчивости в Active-ALUA архитектуре

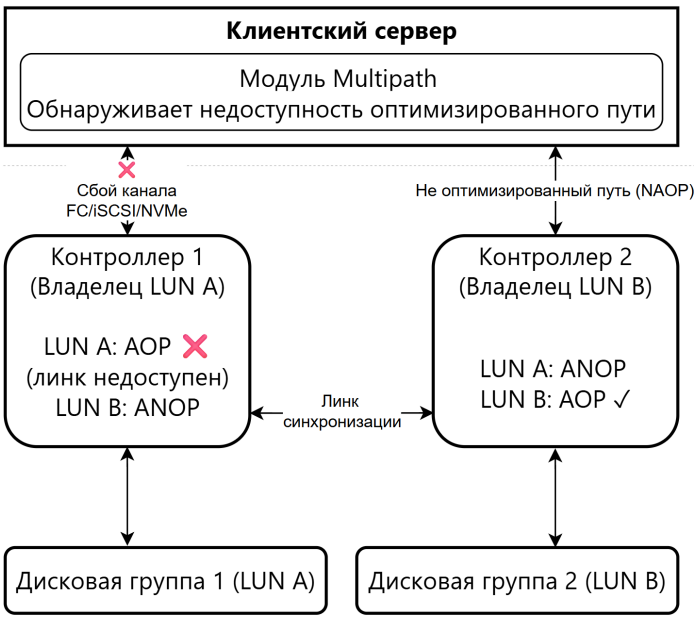
4.3.1. Сбой диска в RAID-массиве



Описание сценария:

1. Один из дисков в RAID-массиве выходит из строя
2. Контроллер-владелец обнаруживает ошибку чтения/записи
3. Запускается восстановление:
   - **RAID продолжает работать** в деградированном режиме
   - **Данные реконструируются** из parity информации оставшихся дисков
   - **Производительность временно снижается** из-за дополнительных вычислений
   - **Администратор уведомляется** о необходимости замены диска
4. Горячая замена диска без остановки системы
5. Автоматическая ребилд после замены диска

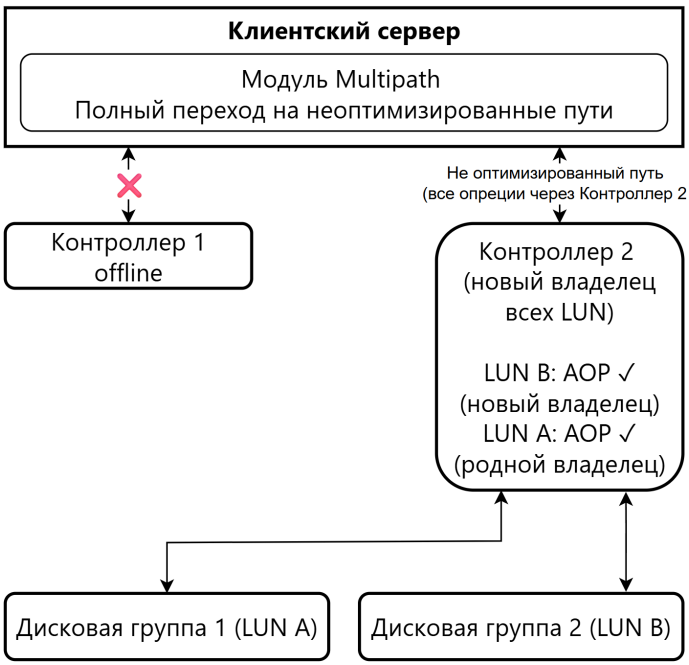
4.3.2. Сбой канала связи (оптимизированного пути)



Описание сценария:

1. Физический канал связи (FC, iSCSI, NVMe) между клиентом и контроллером-владельцем выходит из строя
2. Multipath-модуль получает таймауты/ошибки на оптимизированном пути
3. **Переключение**:
   - **Multipath помечает путь как "failed"**
   - **Автоматическое перенаправление трафика** на неоптимизированный путь через Контроллер 2
   - **LUN A теперь доступен**: Клиент → Контроллер 2 → Связь синхронизации → Контроллер 1 → LUN A
4. **Влияние на производительность**: Временное снижение из-за дополнительной задержки линка оптимизации
5. **Восстановление**: При восстановлении канала - автоматический failback

4.3.3. Сбой контроллера с полным fail-over



Описание сценария:

1. Полный отказ Контроллера 1 (аппаратный сбой, перезагрузка)
2. **Обнаружение**:
   - Контроллер 2 не получает heartbeat через линк синхронизации
   - Multipath на клиентах получает ошибки на всех путях через Контроллер 2
   - Все запросы чтения и записи становятся в очередь
3. **Процесс fail-over**:
   - Контроллер 2 подтверждает отсутствие ответа
   - Контроллер 2 становится владельцем LUN A
   - **Переключение путей**: ANOP → AOP для LUN A

Влияние на клиента:

- **Кратковременная пауза** (секунды) во время переключения
- **Автоматическое продолжение работы** через Контроллер 2
- **Производительность**: LUN B - без изменений, LUN A - полная производительность

Сводная таблица сценариев отказоустойчивости

Сценарий	Обнаружение	Время переключения	Влияние на клиентов	Восстановление
Сбой диска	Контроллером	Нет переключения	Нет	Автоматическая ребилд RAID
Сбой линка	Multipath	< 5 секунд	Снижение производительности	Автоматический fail-back
Сбой контроллера	Heartbeat	10-30 секунд	Краткая пауза, затем нормальная работа	Ручное/автоматическое

Ключевые принципы отказоустойчивости ALUA

1. **Прозрачность для приложений** - все переключения автоматические
2. **Минимальное время недоступности** - от секунд до минут
3. **Сохранение данных** - никаких потерь при корректных сценариях
4. **Автоматическое восстановление** - максимальная автономность
5. **Балансировка нагрузки** - оптимальное использование ресурсов после восстановления

Архитектура обеспечивает **бесперебойную работу** даже в экстремальных сценариях, соответствуя требованиям enterprise-сред с доступностью 99.999%.

ПРИМЕЧАНИЕ

В версии 3.6 **отсутствует автоматический failback**. После восстановления первого контроллера ресурсы остаются на втором. Возвращать их нужно **вручную**. Это сделано намеренно: если контроллер вышел из строя, он мог остаться неисправным, и автоматический возврат ресурсов приведёт к новому падению.

4.4 Архитектура Symmetric Active-Active

Система BlazeX в конфигурации **SAA** (Symmetric Active-Active) представляет собой архитектуру, где оба контроллера обеспечивают **полностью симметричный доступ** ко всем LUN и дискам одновременно, без разделения владения.

Ключевые принципы SAA

1. Полная симметрия доступа

- **Оба контроллера имеют одинаковый доступ** ко всем дискам
- **Нет понятия "владелец LUN"** - каждый LUN доступен через оба контроллера
- **Все пути равнозначны** с точки зрения производительности

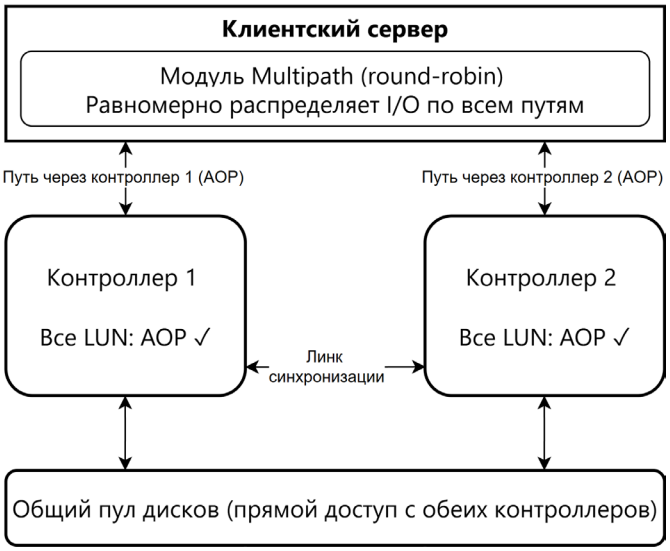
2. Распределенная обработка I/O

- Каждый контроллер обрабатывает запросы независимо
- Согласованность данных обеспечивается через высокоскоростной линк
- Нет необходимости в перенаправлении запросов между контроллерами

3. Упрощенная конфигурация multipath

- Все пути имеют одинаковый приоритет
- Балансировка нагрузки на всех активных путях
- Более быстрая реакция на сбои

Архитектурная схема SAA

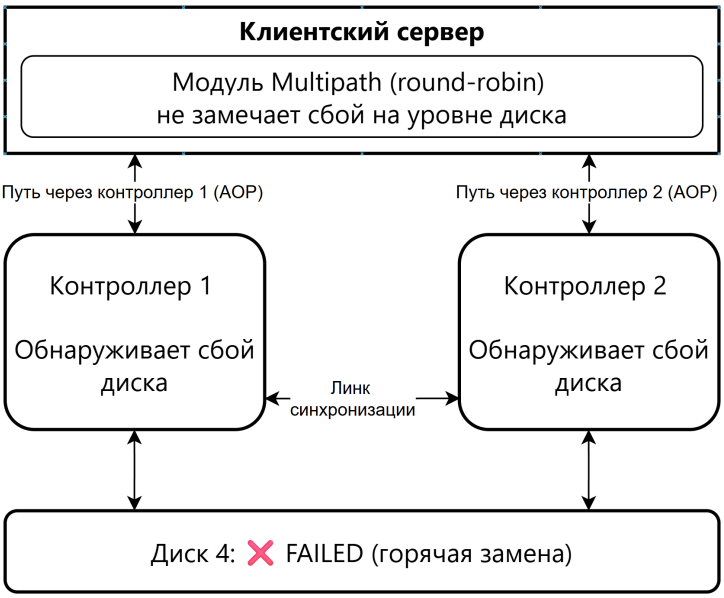


Преимущества SAA архитектуры

Преимущество	Описание
Максимальная производительность	Полное использование ресурсов обоих контроллеров
Мгновенное переключение	Нет задержек при сбое контроллера
Упрощенное управление	Одинаковая конфигурация на всех путях
Лучшая балансировка	Равномерное распределение нагрузки
Высокая доступность	Нет единой точки отказа

4.5 Отказоустойчивость SAA (Symmetric Active-Active)

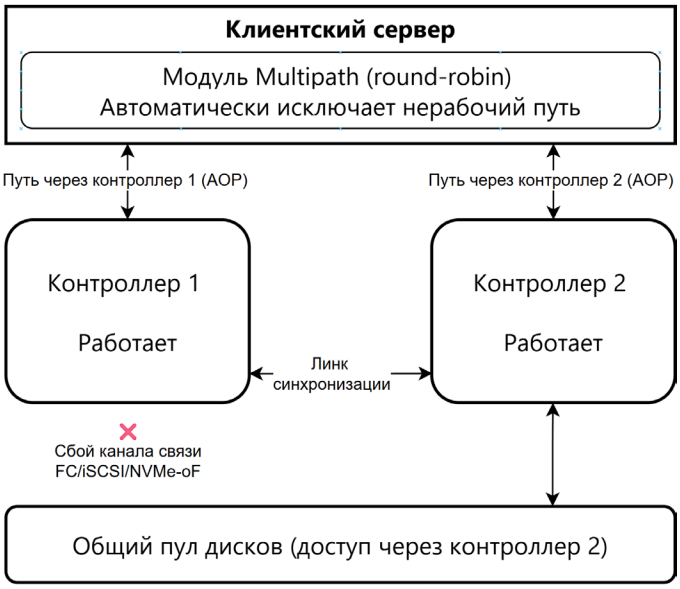
4.5.1. Сбой диска в SAA архитектуре



Описание сценария:

- **Обнаружение**: Оба контроллера одновременно обнаруживают сбой диска
- **Восстановление**: RAID продолжает работу в деградированном режиме
- **Влияние на клиентов**: **Отсутствует** - клиенты не замечают сбой
- **Преимущество SAA**: Оба контроллера могут выполнять ребилд RAID параллельно

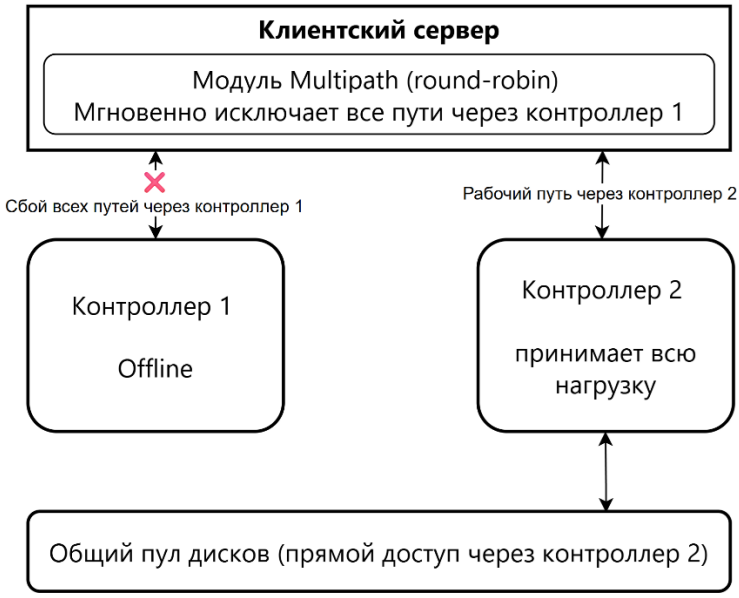
4.5.2. Сбой линка (канала связи) в SAA



Описание сценария:

- **Обнаружение**: Multipath получает таймауты на одном из путей
- **Действие**: Немедленное исключение нерабочего пути
- **Переключение**: Весь трафик перенаправляется на работающий контроллер. Производительность снижается на 50%
- **Восстановление**: Автоматическое возвращение пути при восстановлении
- **Преимущество SAA**: Нет задержек переключения (нет перехвата владения)

4.5.3. Сбой контроллера в SAA архитектуре



Описание сценария:

- **Обнаружение**: Multipath получает ошибки на всех путях через контроллер
- **Действие**: Немедленное переключение на работающий контроллер
- **Время переключения**: **< 2 секунды** (только детектирование сбоя)
- **Восстановление работы**: Контроллер 2 уже имеет полный доступ ко всем данным
- **Преимущество SAA**:
  - Нет передачи владения LUN
  - Нет синхронизации метаданных
  - Минимальное время простоя

Ограничения SAA архитектуры

1. **Сложность координации** - продвинутые функции требуют сложной синхронизации между контроллерами
2. **Производительность** - дедупликация и компрессия в real-time сложны в распределенной среде
3. **Консистентность снапшотов** - создание согласованных моментальных снимков требует глобальных блокировок
4. **Протокольные ограничения NFS** - сложность реализации распределенной файловой системы

4.6 Сравнение SAA vs. ALUA

Параметр	SAA	ALUA
Производительность	Максимальная (оба контроллера активны)	Оптимизированная (по владельцу)
Время переключения	< 2 секунды	10-30 секунд
Сложность конфигурации	Проще	Сложнее
Thin provisioning	✓	✓
RAID	✓	✓
Многоуровневость	✓	✓
Функциональность	Базовое хранилище ❌ Дедупликация ❌ Компрессия ❌ Снапшоты ❌ Снапклоны ❌ Клонирование	Полный функционал ✓ Дедупликация ✓ Компрессия ✓ Снапшоты ✓ Снапклоны ✓ Клонирование
Протоколы	Все протоколы ✓ Fibre Channel (FC) ✓ iSCSI/iSER ✓ NVMe-oF ✓ SMB (CIFS) ❌ NFS	Все протоколы ✓ Fibre Channel (FC) ✓ iSCSI/iSER ✓ NVMe-oF ✓ SMB (CIFS) ✓ NFS
Использование ресурсов	100% обоих контроллеров	~50/50 по LUN

Рекомендуемые сценарии использования SAA

Идеально для:

1. **Высокопроизводительных БД** (Oracle RAC, SQL Server)
2. **Виртуализации** (VMware vSphere, Hyper-V)
3. **Вычислительных кластеров** (HPC)
4. **Транзакционных систем** (банковские операции)
5. **Медиа-обработки** (видео, графика)

Не рекомендуется для:

1. Файловых серверов (нужен NFS)
2. Резервного копирования (нужны снапшоты)

3. Виртуальных десктопов (нужна дедупликация)
4. Облачных хранилищ (нужна компрессия)

Архитектура SAA (Symmetric Active-Active) обеспечивает:

- **Максимальную производительность** - полное использование всех ресурсов
- **Мгновенное переключение** - минимальное время простоя
- **Простоту управления** - одинаковые настройки на всех путях
- **Высокую доступность** - отсутствие единых точек отказа

Однако:

- **Ограниченная функциональность** - без продвинутых функций
- **Нет поддержки NFS** - только блочные протоколы

SAA идеально подходит для приложений, требующих максимальной производительности и минимального времени простоя, но не нуждающихся в продвинутых функциях хранения.

Для полного функционала рекомендуется **ALUA архитектура**.

5 УПРАВЛЕНИЕ СХД BLAZEX

Программное обеспечение BlazeX предоставляет единый интерактивный веб-интерфейс на русском и английском языках, предназначенный для централизованного управления всеми контроллерами СХД в системе. Интерфейс реализует комплексный подход к администрированию, мониторингу и автоматизации операций, обеспечивая высокую доступность, безопасность и удобство эксплуатации системы хранения данных.

5.1 Ключевые возможности интерфейса управления:

Доступ к интерфейсу

- Поддержка протоколов HTTP и защищённого HTTPS для безопасного доступа к веб-интерфейсу.
- Авторизация с использованием логина и пароля, а также возможность интеграции с AD/LDAP.
- Ролевой доступ с предустановленными ролями: Администратор, Оператор, Гость.

Визуализация и управление ресурсами

- Единое представление всех контроллеров, дисков, портов ввода-вывода, групп накопителей (RAID) и логических томов.
- Интерактивная карта оборудования с цветовой индикацией состояния узлов, накопителей и групп.
- Аппаратная индикация дисков для их физической идентификации в стойке.
- Управление группами накопителей с поддержкой драйверов:
    - BlazeIO (ALUA)
    - BlazeIO A/A (Active/Active, Symmetric)
    - MDRAID
- Создание, редактирование, объединение и удаление групп, включая композитные RAID-конфигурации.

Экспорт данных и внешний доступ

- Поддержка протоколов iSCSI, Fibre Channel и NFS.
- Управление таргетами, группами инициаторов и правилами доступа.
- Настройка экспортов NFS с гибкими параметрами доступа для групп клиентов.

Резервирование и отказоустойчивость

- Функция Hot Spare для автоматической замены вышедших из строя дисков.
- Поддержка снимков (snapshots) и клонов (clones).

- Возможность работы в режимах Active/Active и Active/Passive в двухконтроллерном исполнении.

Безопасность и аудит

- Смена пароля и восстановление доступа через консольные скрипты.
- Управление SSH-ключами для безопасного доступа к CLI.
- Интеграция с AD/LDAP для централизованной аутентификации.
- Логирование всех действий администратора с возможностью выгрузки логов.

Автоматизация и API

- REST API для программного управления СХД.
- Командная строка (CLI) с набором команд для автоматизации операций:
    - Управление пулами и томами.
    - Работа со снимками и клонами.
    - Настройка экспортов и пользователей.
- Доступ к CLI через SSH с использованием сгенерированных ключей.

Системные настройки

- Настройка сетевых интерфейсов (DHCP/статический IP, MTU, тип интерфейса).
- Управление временем и датой (часовой пояс, NTP-серверы).
- Конфигурация Syslog для централизованного сбора логов.
- Настройка SMTP для отправки email-уведомлений.

Техническая поддержка и документация

- Раздел «Система» с информацией о версиях ПО, состоянии узлов и лицензировании.
- Встроенная справка и ссылки на сайт поддержки.

5.2 Функциональность мониторинга

Функциональность внутреннего мониторинга

Система BlazeX включает встроенные средства мониторинга, обеспечивающие контроль состояния системы в реальном времени.

Основные компоненты внутреннего мониторинга:

Панель обзора СХД (раздел «Мониторинг» → «Обзор СХД»)

- Графики производительности дисковой полки (IOPS)
- Пропускная способность (МБ/с) по чтению и записи
- Время отклика дисковой полки (мс)
- Загрузка процессоров каждого узла
- Использование оперативной памяти
- Загрузка сетевых портов (входящий/исходящий трафик)

Визуализация состояния оборудования

- Цветовые индикаторы здоровья узлов, дисков и групп
- Карточки оборудования с детальной информацией
- Аппаратная индикация дисков для физической идентификации

Системные уведомления

- Мониторинг состояния NFS-серверов
- Отслеживание здоровья экспортов и групп инициаторов
- Визуализация состояния синхронизации между узлами в двухконтроллерной конфигурации

Журналирование

- **Журнал команд** – фиксирует действия пользователей через Web-интерфейс (кто, когда, какую команду выполнил, статус, результат). Нужен для анализа работы операторов.
- **Журнал событий** – фиксирует системные события (сбои дисков, ошибки, предупреждения, смену статуса компонентов) с указанием важности (Emergency, Critical, Error, Warning, Info). Нужен для диагностики неисправностей.
- **Журнал аудита** – фиксирует события безопасности: вход/выход пользователей, неудачные попытки входа, смена пароля, создание/удаление/изменение прав пользователей. Нужен для расследования инцидентов и соблюдения требований регуляторов.

5.3 Функциональность внешнего мониторинга

BlazeX поддерживает интеграцию с популярными системами мониторинга через стандартные протоколы:

SNMP-мониторинг

- **Поддержка версий SNMP v2c и v3**
- Настройка контроля доступа и строк сообщества
- Конфигурация получателей уведомлений (trap)
- Возможность скачивания MIB-файлов
- Интеграция с Zabbix через snmptrapd
- Тестирование отправки SNMP-трапов

Prometheus + Grafana

- Предоставление метрик по URL
- Поддерживаемые метрики:
    - количество накопителей
    - количество групп накопителей
    - количество логических томов
    - количество файловых экспортов
    - количество блочных экспортов
    - количество ошибок

Zabbix

- Готовые шаблоны мониторинга "Template BlazeX Monitoring"
- Автоматическая настройка через команду blazex-install
- Мониторинг состояния узлов, дисков, пулов и томов
- Настройка триггеров и уведомлений

Syslog

- Централизованный сбор логов на внешний сервер
- Поддержка протоколов UDP и TCP
- Настройка порта и адреса сервера
- Возможность сброса настроек

SMTP-уведомления

- Настройка почтового сервера с поддержкой SSL/TLS
- Конфигурация списка получателей
- Включение/отключение групп уведомлений
- Проверка соединения с SMTP-сервером

ТИПЫ ПОДДЕРЖИВАЕМЫХ ДИСКОВ.

Тип диска	Интерфейс	Макс. ёмк.	Скорость
SAS SSD	SAS 12G	30.72 TB	1800 MB/s 300K IOPS
SATA SSD	SATA 6G	15.36 TB	550 MB/s 100K IOPS
SAS (Nearline)	SAS 12G	20 TB	250 MB/s
SATA HDD	SATA 6G	20 TB	210 MB/s
NVMe (U.2)	PCIe 4.0	15.36 TB	7000 MB/s 1.5M IOPS
NVMe (M.2)	PCIe 4.0	8 TB	6500 MB/s 1M IOPS

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

Узел\Контроллер СХД - сервер в многосерверной конфигурации, обеспечивающий доступ к ресурсам СХД.

Операционная система (ОС) - программное обеспечение, управляющее компьютерами (включая микроконтроллеры) и позволяющее запускать на них прикладные программы.

Программное обеспечение (ПО) - совокупность программ, данных и связанных с ними документов, используемых для управления информационной системой.

Сервер - выделенный вычислительный комплекс, обрабатывающий запросы от других ПК и предоставляющий им необходимую информацию и/или услуги.

Система хранения данных (СХД) - комплекс аппаратного и программного обеспечения, предназначенный для хранения и оперативной обработки информации.

nginx - веб-сервер и обратный прокси, используемый как точка входа для всех пользовательских запросов. В данной архитектуре выполняет функции маршрутизации, балансировки нагрузки.

API Gateway - шлюз, через который проходят все внешние запросы к микросервисам. Позволяет централизованно управлять маршрутизацией, безопасностью и мониторингом.

Split-brain - ситуация, при которой оба узла считают себя активными и пытаются управлять одними и теми же ресурсами, что может привести к потере или повреждению данных.

WebSocket (wss) - протокол для двусторонней передачи данных в реальном времени между клиентом и сервером по защищенному соединению (wss - WebSocket Secure).

DRBD - Distributed Replicated Block Device - технология зеркалирования (репликации) блочных устройств между серверами в реальном времени, что обеспечивает отказоустойчивость данных.

Pacemaker - кластерный менеджер, отвечающий за мониторинг состояния узлов и автоматическое переключение сервисов при сбоях (failover).

ocf-agent - агент управления ресурсами в кластере, реализующий команды запуска, остановки и мониторинга сервисов.

Failover - автоматическое переключение сервисов и ресурсов на резервный узел при сбое основного.

Кластерный агент (Cluster Agent) - программный компонент, работающий на каждом контроллере, который обеспечивает мониторинг состояния узла, участвует в обмене heartbeat-сигналами с другим контроллером и предоставляет интерфейсы управления (REST API, GUI) для локального и удаленного администрирования.

blazex-control - основной микросервис управляющего слоя, реализующий REST API для всех операций управления системой хранения данных (СХД). Отвечает за прием команд от пользователя, их постановку в очередь, взаимодействие с сервисом конфигурации и делегирование задач исполняющим агентам (blazex-agent).

blazex-agent - исполняющий агент, запускаемый на каждом узле кластера. Получает команды от сервиса blazex-control и выполняет непосредственные операции с инфраструктурой хранения: управление дисками, RAID-массивами, логическими томами, а также сбор и отправку данных о состоянии ресурсов.

yaml-config service - сервис конфигурации, отвечающий за хранение, управление и синхронизацию конфигурационных данных системы между узлами кластера. Обеспечивает атомарность изменений и консистентность конфигурации за счет механизма репликации «запись на первом узле - репликация на второй - подтверждение».

Слой экспорта (Export Layer) - программный уровень, предоставляющий данные в виде сетевых ресурсов через стандартные протоколы блочного (FC, iSCSI/iSER, NVMe-oF) и файлового (NFS, SMB) доступа. Преобразует внутренние логические тома в объекты, доступные для клиентов по сети.

Слой организации хранения (COX, Storage Organization Layer) - программный уровень, реализующий расширенные функции управления данными поверх физических или виртуальных дисков. Включает механизмы thin provisioning, дедупликации, сжатия, создания и управления снапшотами и клонами.

Thin Provisioning (Тонкое выделение ресурсов) - технология, позволяющая выделять логическое дисковое пространство приложениям по мере фактической необходимости, а не заранее выделенным физическим ресурсам, что повышает эффективность использования емкости.

Дедупликация (Data Deduplication) - метод устранения избыточных копий данных, при котором одинаковые блоки информации хранятся в единственном экземпляре, что позволяет значительно экономить дисковое пространство.

Снапшот (Snapshot) - моментальный снимок состояния данных в определенный момент времени, позволяющий зафиксировать их целостность и в дальнейшем использовать для восстановления или создания клонов.

Снапклон (Snapshot Clone) - полная, независимая копия данных, созданная на основе снапшота. Может быть смонтирована и использована как отдельный том.

RAID (Redundant Array of Independent Disks) - технология виртуализации данных, объединяющая несколько физических дисков в единый логический элемент для повышения отказоустойчивости и/или производительности.

BlazeIO - высокопроизводительное проприетарное решение для организации RAID-массивов, оптимизированное для работы в составе ПО BlazeX.

Interconnect - выделенное высокоскоростное соединение между контроллерами кластера, используемое для служебного обмена данными, синхронизации состояний и передачи трафика при отработке отказа (failover).

Heartbeat - служебный сигнал (пульс), периодически передаваемый между узлами кластера для подтверждения их работоспособности. Потеря heartbeat-сигналов от одного из узлов является триггером для запуска процедуры восстановления или переключения (failover).

Сеть управления (Management Network) - выделенный сегмент сети, предназначенный для подключения клиентов администрирования, систем мониторинга и платформ оркестрации к системе хранения данных. Обеспечивает изоляцию управляющего трафика от трафика данных (Data Network).

ALUA (Asymmetric Logical Unit Access) - Асимметричный доступ к логическому устройству - механизм, при котором контроллеры СХД предоставляют разные типы путей доступа к одному LUN: оптимизированные (AOP, прямой доступ к «своим» дискам) и неоптимизированные (ANOP, доступ через линк синхронизации). Используется для балансировки нагрузки и отказоустойчивости. В данной архитектуре реализован для работы с драйвером BlazeIO.

SAA (Symmetric Active-Active) - Симметричная активно-активная архитектура - режим работы двухконтроллерной СХД, при котором оба контроллера имеют равноправный прямой доступ ко всем дискам без разделения владения. Все пути доступа являются оптимизированными. Обеспечивает минимальное время переключения при сбое (< 2 с), но не поддерживает NFS и продвинутые функции (дедупликация, сжатие, снапшоты).

ROW (Redirect-on-Write) - Метод организации снапшотов (мгновенных снимков), при котором при изменении данных в исходном томе изменённые блоки записываются в новое место, а снимок продолжает ссылаться на старые блоки. В отличие от COW, не требует чтения исходного блока перед записью, что снижает накладные расходы. Используется в модуле снапшоты/снапклоны.

COW (Copy-on-Write) - Метод организации снапшотов, при котором перед первой записью в изменяемый блок исходные данные копируются в область снимка, и только затем выполняется запись новых данных. Требует операции чтения перед записью, что может влиять на производительность. Альтернатива - ROW.

NVMe-oF (NVMe over Fabrics) - Протокол передачи команд NVMe по сетевым фабрикам (Ethernet, Fibre Channel, InfiniBand). Позволяет обеспечить удалённый доступ к NVMe-накопителям с задержками, сопоставимыми с локальным подключением. В архитектуре BlazeX используется как один из протоколов слоя экспорта (SAN-сценарии).

iSER (iSCSI Extensions for RDMA) - Расширение протокола iSCSI, использующее RDMA (InfiniBand, RoCE, iWARP) для передачи данных в обход сетевого стека ОС. Обеспечивает высокую пропускную способность и низкие задержки по сравнению с классическим iSCSI. Поддерживается в слое экспорта BlazeX наряду с iSCSI.

ТЕХНИЧЕСКАЯ ПОДДЕРЖКА

Техническая поддержка ПО «BlazeX» включает следующий набор услуг:

- предоставление обновлений программного обеспечения по мере выхода новых релизов;
- консультация ИТ-специалистов заказчика по работе управляющего ПО;
- помощь в устранении сбоев, вызванных некорректной работой управляющего ПО;
- помощь в обновлении программного продукта в удаленном режиме.

Контакты службы поддержки и сервиса:

Адрес электронной почты: help@bitblaze.ru

Интернет-сайт: https://bitblaze.tech/

Телефон компании: (3812)-36-11-11

ПРИМЕЧАНИЕ

Техническая поддержка осуществляется в рамках Соглашения об уровне сервиса (SLA).

Все гарантии, касающиеся товаров и услуг, реализуемых ООО «БитТех», изложены в формулировках прямых гарантий, сопровождающих соответствующие товары и услуги.

Никакая информация, приведенная в данном документе, не должна рассматриваться как дополнительная гарантия.