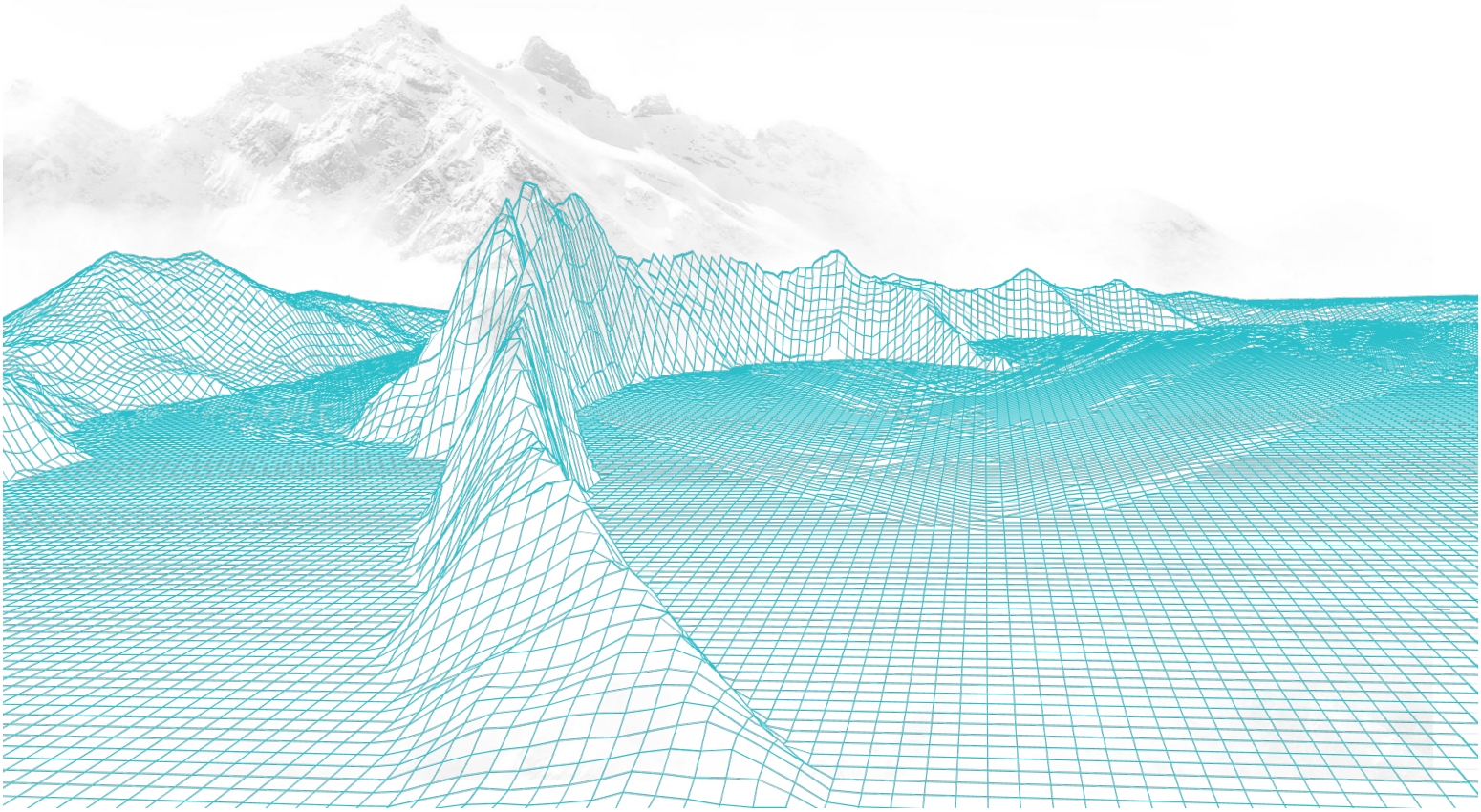
Программное обеспечение системы хранения данных
BAUM STORAGE IN v.2

ТЕХНИЧЕСКОЕ ОПИСАНИЕ

Москва 2024



Описание специального программного обеспечения «BAUM STORAGE IN v.2»

Оглавление

1. Общее описание СПО «BAUM STORAGE IN v.2»

Специальное программное обеспечение “BAUM STORAGE IN v.2” предназначено для работы в составе программно-аппаратного комплекса системы хранения данных (СХД). “BAUM STORAGE IN v.2” обеспечивает централизованное, надежное хранение данных и предоставление доступа к ним по широко распространенным файловым и блочным протоколам. Высокая надежность системы достигается за счет реализации схемы кластера высокой готовности, при которой программное обеспечение, установленное одновременно на двух узлах кластера, отрабатывает запросы клиентов на доступ к единому хранилищу данных. Отказоустойчивость хранилища и надежность хранения данных обеспечивается реализацией технологии RAID и аппаратными решениями управляемого “BAUM STORAGE IN v.2” оборудования.

Задачи:

- Построение отказоустойчивых хранилищ данных;
- Долговременное хранение больших объемов данных;
- Высокая производительность при случайном доступе к данным;
- Использование с системами виртуализации.

Специальное программное обеспечение «BAUM STORAGE IN v.2» состоит из следующих подсистем:

1. Подсистема хранения и обработки данных;
2. Подсистема доступа к данным;
3. Подсистемы управления;
4. Подсистемы мониторинга и журналирования.

Подсистемы логически объединяют в себе программные модули, совместно работающие в рамках выполнения единой задачи. Управление компонентами (сервисами) операционной системы реализовано при помощи разработки специализированных программных модулей, принимающих команды управления и выполняющих определенную последовательность команд управления над подчиненными им системными сервисами, через интерфейс командной строки операционной системы (CLI). Каждый такой модуль реализует

некоторый набор методов - команд внутреннего программного интерфейса приложения (API), которые могут быть вызваны удаленно. Команду на выполнение того или иного метода и её параметры, программный модуль получает в управляющем пакете. После выполнения команды, программный модуль читает стандартный поток вывода и формирует ответный пакет на основании полученных данных о результате работы команды. Программные модули представляют собой событийно-управляемые асинхронные программы, реализованные по единой схеме, в которых для обработки приходящих сообщений реализована «машина состояний».

2. Общее описание подсистемы хранения и обработки данных

Подсистема хранения и обработки данных логически объединяет в себе программные модули, отвечающие за хранение и обработку (оптимизацию) данных, а также за управление накопителями хранилища. В подсистеме реализованы следующие функции:

- Управление и предоставление информации по накопителям;
- Распределение данных, четности и зарезервированного пространства по всем дискам пула;
- Быстрое восстановление целостности пула после сбоя диска;
- Кэширование операции чтения и записи данных;
- Оптимизация данных в момент записи;
- Снепшоты на быстрых пулах;
- Клоны на быстрых пулах;
- Консистентные snapshot PostgreSQL (внешний модуль);
- Компрессия данных на быстрых пулах;
- Дедупликация данных на быстрых пулах.
- QoS;

Подсистема, преобразует физическую емкость всех накопителей пула в блоки заданного размера, формируя виртуальный диск. Далее в эти блоки записываются как сохраняемые данные, так и избыточная информация (четность), требуемая для их восстановления в случае отказа накопителя. Емкость пула может быть разбита на независимые друг от друга разделы – тома данных. На томах могут быть созданы файловые системы. Производительность операций ввода-вывода увеличивается, за счет использования технологий кэширования записи и чтения данных. Для экономии дискового пространства и увеличения производительности операций записи-чтения применяется компрессия записываемых данных, выполняемая на лету, и их последующая декомпрессия при считывании сжатых данных.

Для оптимизации потребления дискового пространства используются алгоритмы дедупликации записываемых данных. Дедупликация – это метод устранения дубликатов данных, при котором повторяющиеся блоки данных заменяются ссылками на ранее записанные блоки.

3. Общее описание подсистемы доступа к данным

Подсистема доступа к данным логически объединяет в себе программные модули управления протоколами доступа к данным и системные компоненты, реализующие эти

протоколы. Подсистема отвечает за управление передачей данных между клиентами и Системой хранения данных (СХД), а также за репликацию данных на тома другой СХД. В подсистеме реализованы следующие функции:

- Запись и чтение данных по протоколам FC и iSCSI;
- Привязка протоколов FC и iSCSI к томам данных;
- Запись и чтение данных по протоколам NFS и SMB (CIFS);
- Привязка протоколов NFS и SMB к файловым системам;
- Управление доступом к данным;
- Репликация данных.

В состав подсистемы также входит модуль управляет списками доступа к ресурсам (томам данных и файловым системам) и управляет подключением к домену Active Directory и серверу LDAP.

4. Общее описание подсистемы управления

Подсистема управления логически объединяет в себе компонент веб-сервера, модуль управления из командной строки (CLI) и модуль мониторинга состояния и управления кластером. Подсистема отвечает за выполнение команд пользователя, переданных посредством веб-интерфейса, а также за наблюдение за работоспособностью узлов кластера и управление миграцией ресурсов. Наблюдение за работоспособностью узлов кластера реализовано за счет регулярного обмена информационными пакетами через прямое сетевое соединение - интерконнект. За запись и чтение модулями конфигурации системы, а также за синхронизацию базы конфигурации между узлами кластера, отвечает программный модуль хранения конфигурации.

В подсистеме реализованы следующие функции:

- Управление конфигурацией системы;
- Отображение компонентов пользовательского веб-интерфейса(WEB2.0);
- Обработка команд интерфейса управления командной строки (CLI);
- Мониторинг работоспособности узлов кластера;
- Управление миграцией ресурсов;
- Управление сетевыми настройками;
- Единый интерфейс для управления парами контроллеров;
- Посервисное обновление;
- Лицензирование;
- REST-API;
- Cinder driver RuStack (внешний модуль).

В общем случае, алгоритм управления системой работает следующим образом:

1. Пользователь, в графическом интерфейсе управления системой, изменяет некоторые параметры формируя командую последовательность, программный модуль управляющий пользовательским интерфейсом инициирует отправку управляющего сообщения одному из модулей, отвечающих за различные параметры аппаратного обеспечения. В сообщении содержится имя вызываемого метода, реализованного в этом программном модуле и

необходимые параметры.

2. Программный модуль получивший сообщение, в свою очередь отправляет некоторую последовательность команд управляемому им системному компоненту.

3. Системный компонент возвращает результат выполнения команды, который определенным образом интерпретируется программным модулем. Далее модуль формирует ответное сообщение и направляет его модулю – отправителю, в данном случае модулю пользовательского интерфейса. Изменение параметров системы также заносится в хранилище конфигурации и регистрируется в системных логах.

4. Получив результат управляющего воздействия, программный модуль пользовательского интерфейса интерпретирует его и выводит в том или ином виде в интерфейс пользователя.

Алгоритм мониторинга работоспособности кластера работает следующим образом:

Модуль мониторинга работоспособности и управления кластером выполняет наблюдение за работоспособностью узлов кластера регулярно обмениваясь с аналогичным модулем соседнего узла информационными пакетами через сетевой интерконнект. Этом алгоритм получил название – сетевой heartbeat. В случае отсутствия ответа, через заданное количество попыток, модуль инициирует процедуру переподключения ресурсов, зарегистрированных на отказавшем узле, на свой узел (миграции ресурсов).

Во избежание ложного срабатывания механизма переподключения ресурсов при пропадании связи через интерконнект, проверка статуса соседнего узла кластера также дублируется через общую дисковую подсистему кластера, путем записи и чтения узлами кластера данных в определенных секторах дисковых накопителей. Этот алгоритм называется «дисковый heartbeat».

При выполнении миграции ресурсов с рабочего узла кластера, например, при проведении планового обслуживания, запускается другой сценарий миграции. Сначала узел, с которого забирают ресурсы, выполняет отключение ресурсов от своих служб в определенной последовательности. Эти ресурсы переподключаются к одноименным службам второго узла. После передачи своих ресурсов узел устанавливает статус «отдал ресурсы», а второй узел устанавливает статус «принял ресурсы». Статусы записываются в базу конфигурации.

При возврате ранее отданных ресурсов, узел выполняет переподключение ресурсов в обратной последовательности.

5. Общее описание подсистемы мониторинга и журналирования

Подсистема мониторинга и журналирования логически объединяет в себе системные компоненты чтения состояния датчиков аппаратного обеспечения, программные модули: мониторинга аппаратного обеспечения, самодиагностики, логирования событий. Подсистема отвечает за мониторинг работы системы и фиксацию событий, возникающих в процессе её работы.

В подсистеме реализованы следующие функции:

- Мониторинг состояния аппаратного и программного обеспечения;
- Отображение состояния портов Fibre Channel и SAS адаптеров;
- Отображение состояния сетевых соединений и статуса сетевых портов;
- Журналирование событий;
- Управление архивированием и выгрузкой журналов событий;
- Отправка статистики о работе системы по протоколу SNMP;

- Алгоритмы машинного обучения. Облачное решение (внешний модуль).
- Алгоритмы машинного обучения. Установка на сайте заказчика (внешний модуль)

Программные модули системы отслеживают свое состояние и состояние ресурсов, которыми они управляют. В случае обнаружения значительных отклонений в работе системы, могущих повлиять на её работоспособность, сервисы отправляют уведомления модулю самодиагностики, который анализирует проблему и отправляет уведомление о ней в интерфейс управления системы, лог-файл, и отправляет уведомление по E-mail.