
СИСТЕМЫ ХРАНЕНИЯ ДАННЫХ

НАУКА И ТЕХНОЛОГИИ

УДК 004.75
ББК 32.32.972,5
С34

Авторы:

А.П. Гантимуров, В.А. Нелюб, А.С. Мельников, М.В. Волков,
А.С. Бородулин, А.С. Июдин

С34 **Системы хранения данных : наука и технологии** / [А. П. Гантимуров
и др.]. — Москва : Издательство МГТУ им. Н. Э. Баумана, 2023. —
108, [2] с. : ил.

ISBN 978-5-7038-6151-6

Рассматриваются теоретические и практические аспекты построения и функционирования развитой системы хранения данных на примере системы хранения данных BAUMSTORAGE, реализующей современные технологии гибкого высокопроизводительного надежного управления инфраструктурой хранения, доступом и откликом при работе с большими объемами данных в распределенной вычислительной среде.

Для инженеров, работающих в сфере информационных технологий.

УДК 004.75
ББК 32.972,5

ISBN 978-5-7038-6151-6

© Оформление. Издательство
МГТУ им. Н.Э. Баумана, 2023

Предисловие

В книге рассматриваются практические аспекты построения и функционирования развитой системы хранения данных (СХД) на примере BAUMSTORAGE UDS (далее BAUMSTORAGE), реализующей современные технологии гибкого высокопроизводительного надежного управления инфраструктурой хранения, доступом и откликом при работе с большими объемами данных в распределенной вычислительной среде.

Глава 1 представляет основные возможности BAUMSTORAGE и реализующие их технологии, положенные в основу функционирования программного обеспечения системы. Рассмотрены особенности организации файловой системы (ФС) BAUM FileSystem (BFS) и используемые протоколы доступа к данным. Описан ряд эффективных способов обеспечения надежного хранения данных/метаданных и их обработки (моментальные снимки, клоны, пулы, кластеризация, дедупликация, компрессия и др.), сетевые интерфейсы, методы репликации. Излагаются сведения об используемых решениях по поддержке оптимальной производительности серверов за счет функций кэша чтения и записи, RDMA-синхронизации. Показано, как выполняются мониторинг оборудования и производительности, логирование событий, самодиагностика работы СХД и сбор статистики. Затронуты вопросы информационной безопасности, возможностей масштабирования и интеграции.

Глава 2 посвящена рассмотрению архитектуры аппаратной части и программного обеспечения BAUMSTORAGE. Описываются принципы построения системной архитектуры, назначение и взаимодействие ее компонент. Детально разбирается работа всех сервисов системы.

Материал главы 3 будет полезен ИТ-специалистам в решении проблем при сопровождении BAUMSTORAGE, по каждой из которых указаны симптомы, вероятная причина и предлагаемые действия по устранению проблемы.

В главе 4 даны ответы на часто возникающие вопросы BAUMSTORAGE. Для удобства вопросы разбиты по темам, что позволяет легко в них ориентироваться и найти ответ на интересующий пользователя вопрос.

В главе 5 уделено внимание подробному рассмотрению методов и средств, применяемых в составе BAUMSTORAGE и обеспечивающих ее высокоэффективную работу.

Глоссарий книги содержит описание терминов BAUMSTORAGE.

1. ВОЗМОЖНОСТИ СХД

Широкий спектр возможностей и особенности развитой СХД можно проиллюстрировать на примере BAUMSTORAGE, структурно-функциональная схема которой представлена на рис. 1.1.

1.1. Файловая система

1.1.1. Обычная функциональная файловая система

Функционал обычных пулов предоставляется обычной функциональной файловой системой BAUMSTORAGE. Ключевой особенностью данной файловой системы является использование технологии двойной записи COW. Суть данного алгоритма заключается в том, что системой выделяется новый блок каждый раз, когда выполняется новая запись или происходит изменение ранее записанных данных. Благодаря этому файловая система всегда находится в консистентном состоянии: блок либо удачно записан и на него установлен указатель в дереве адресации, либо в процессе записи произошел сбой и указатель отправляет к старому блоку. Для защиты данных от сбоя применяется также раздельное хранение данных и метаданных. Целостность данных и метаданных подтверждается контрольными суммами. При чтении любого блока вычисляется его контрольная сумма, которая сравнивается с сохраненной, и в случае расхождения исходный блок восстанавливается из дублирующих блоков.

На уровне дискового массива (пула) файловая система поддерживает практически все распространенные типы (уровни) избыточности — RAID 0, 1, 5, 6. Поддерживается также уровень избыточности с выделением трех дисков под четность. В BAUMSTORAGE он называется B3. Поддерживаются и комбинированные типы: RAID 10, 50, 60, B30. Дисковые пулы собираются из рейд-групп — набора дисков, блоки которых выделяются под хранение данных и контрольных сумм. Комбинированные типы формируются добавлением к первой рейд-группе новых рейд-групп. Путем добавления рейд-групп выполняется также расширение пространства пула. Файловая система позволяет создавать на пулах как блочные (тома), так и файловые (файловые системы) ресурсы.

Кэширование операций записи выполняется с помощью добавления к пулу выделенного производительного SSD-диска, на который файловая система записывает данные, из которых собирает страйп для записи на пул, после чего данные записываются на более медленные диски пула. Данные в кэше записи хранятся в виде журнала транзакций. При отсутствии

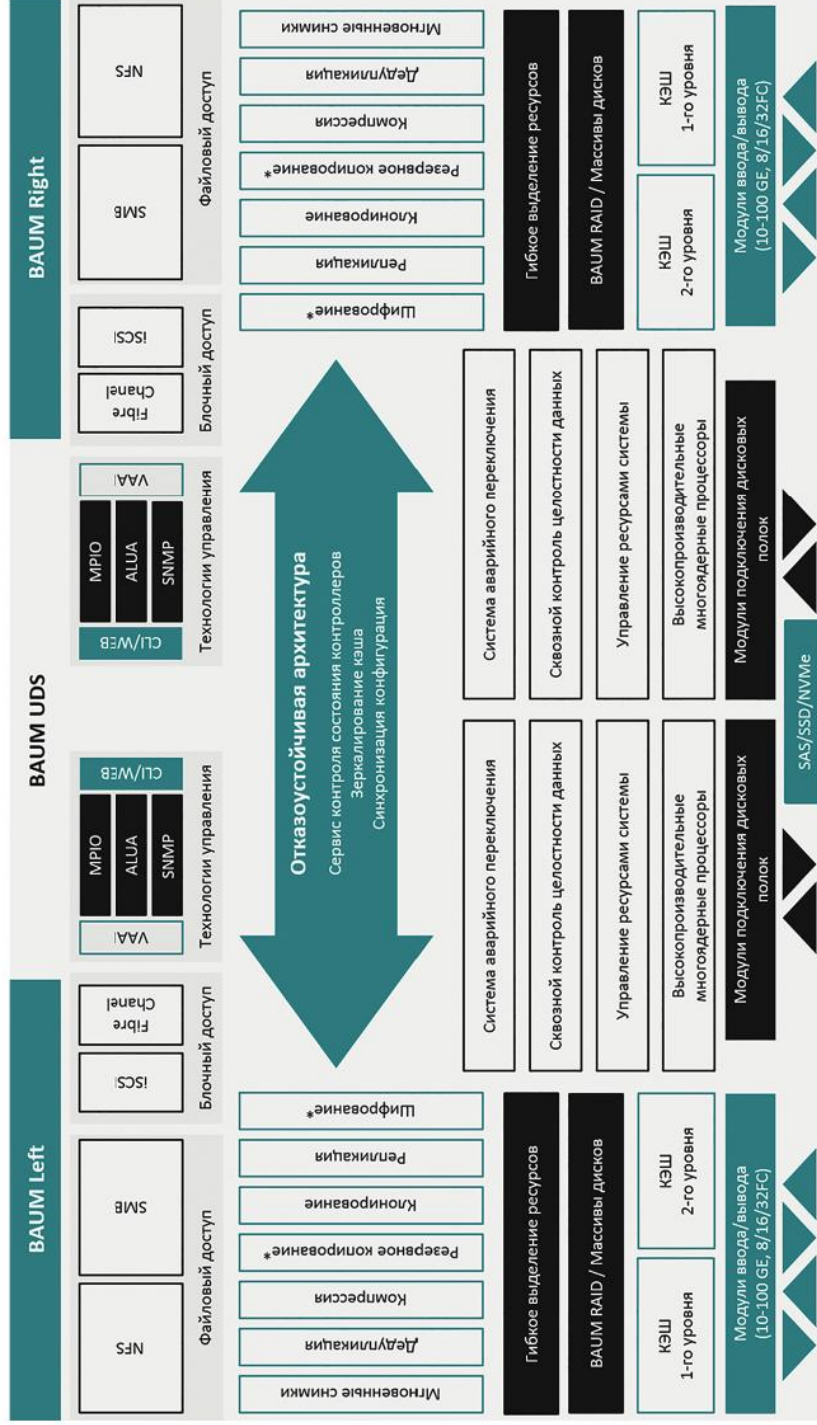


Рис. 1.1. Возможности СХД

выделенного диска под кэш записи, файловая система выделяет для записи журнала транзакций место на основном пуле, что приводит к уменьшению его производительности. Применение дискового кэша записи в данном случае оправдано, поскольку даже при сборке пула из флеш-дисков его производительность будет меньше одного SSD-диска за счет применения технологии двойной записи.

1.1.2. Быстрая файловая система

Быстрая файловая система предоставляет функционал быстрых пулов за счет применения принципа динамического дискового пула (DDP). Из дисков собирается пул, все пространство которого разбивается на множество небольших блоков одинакового размера. Из этих блоков собираются логические массивы (тома) с нужной степенью избыточности. Блоки выбираются таким образом, чтобы при сборке тома обеспечивать равномерное распределение блоков данных и блоков четности по всем дискам пула. Для томов поддерживаются основные типы избыточности — RAID 0, 1, 5, 6. Кроме того, поддерживается комбинированный RAID 10. Помимо блочных ресурсов (томов), быстрая файловая система способна создавать и файловые системы. Для максимальной производительности рекомендуется использовать в пулах быстрой файловой системы (быстрых пулах) только флеш-диски.

Для кэширования операций записи быстрая файловая система не использует дисковый кэш, после сборки страйпа в оперативной памяти выполняется его прямая запись в дисковый пул. Не используется и кэш чтения.

Из-за специфики работы быстрых пулов для них недоступно большинство функций, предоставляемых обычными пулами. В настоящий момент быстрые пулы поддерживают:

- расширение объема пула (таким же количеством дисков, что используется в пуле);
- «толстые» и «тонкие» тома;
- расширение тома;
- создание файловой системы;
- автоматическую замену вышедшего из строя диска на запасной;
- снимки «тонкого» тома, восстановление тома из снимка.

1.2. Протоколы

1.2.1. Блочные протоколы

Семейство протоколов FC

Fibre Channel (FC) — семейство протоколов для высокоскоростной передачи данных. FC используется как стандартный способ подключения к системам хранения данных уровня предприятия. Для соединения с сетью FC

использует одноименные контроллеры и коммутаторы. В настоящий момент используются контроллеры со скоростью портов 8, 16 и 32 Гб. FC обеспечивает блочный доступ к устройствам хранения данных путем передачи команд SCSI по сети TCP/IP.

В терминологии Fibre Channel устройства называются узлами (nodes). Это весьма напоминает узлы в терминологии сетей IP. Как и узел IP, который зачастую получает несколько IP-адресов, узел Fibre Channel может иметь несколько портов. Разница между ними в том, что порт Fibre Channel представляет собой физический элемент, а порт IP — логический. Каждый узел Fibre Channel имеет уникальное 64-разрядное имя WWN (World Wide Name), которое назначается производителем. Это напоминает уникальные MAC-адреса, которые назначаются производителями сетевым адаптерам Ethernet.

Каждый порт сети хранения данных на базе кольца с разделением доступа Fibre Channel имеет 8-битовый адрес, а порт в коммутируемой связанной архитектуре — 24-битовый. При подключении кольца с разделяемым доступом (arbitrated loop) к коммутатору связанной архитектуры (fabric switch) коммутатор представляет 8-битовый адрес в виде 24-битового. Оба идентификатора порта назначаются динамически.

В терминологии Fibre Channel:

Fabric — Fibre Channel Fabric — это просто сеть, состоящая из одного или нескольких коммутаторов Fibre Channel. Это может быть один свитч, но в крупных сетях используется множество коммутаторов, соединенных таким образом, чтобы обеспечить более высокую пропускную способность и доступность в случае отказа одного или нескольких коммутаторов.

Initiator — потребитель хранилища, обычно сервер с картой адаптера, который называется адаптером главной шины (HBA). Инициатор «инициирует» подключение по Fabric к одному или нескольким портам системы хранения пользователя, которые называются целевыми портами.

Target — целевые порты — это порты в системе хранения пользователя, которые доставляют тома хранения (называемые целевыми устройствами или LUN) инициаторам.

Порты адаптера могут работать как в режиме target, так и в режиме initiator. При старте адаптера порт автоматически включается в режим инициатора.

WWN — в Fibre Channel есть специальный термин для идентификатора порта, в частности WWN, или «всемирное имя». Точно так же, как каждая сетевая карта имеет уникальный MAC-адрес для каждого порта сетевой карты, каждый отдельный порт FC имеет уникальный 64-битный порт WWN. Например, WWN порта FC может выглядеть так: 20:00:00:81:23:45:ac:01.

Следует обратить внимание на то, что инициаторы не разговаривают с инициаторами, так же, как и целевые порты. Как правило, считается плохой практикой, когда инициаторы на одном сервере подключаются к инициаторам на другом. Это происходит потому, что при подключении к Fabric инициатор будет обращаться к каждому отдельному целевому порту (и каждому

Оглавление

Предисловие	3
1. Возможности СХД	4
1.1. Файловая система	4
1.1.1. Обычная функциональная файловая система	4
1.1.2. Быстрая файловая система	6
1.2. Протоколы	6
1.2.1. Блочные протоколы	6
Семейство протоколов FC	6
Протокол iSCSI	8
Протокол iSCSI RDMA	9
1.2.2. Файловые протоколы	9
Протокол SMB	9
Протокол NFS	10
Протокол NFS RDMA	11
Протокол FTP	11
1.3. Надежность	11
1.3.1. Мгновенный снимок	11
1.3.2. Кластеризация (работа кластера)	12
1.3.3. Фоновая проверка ошибок в файловой системе	13
1.3.4. Мультипас-подключение дисковых полок (SAS)	14
1.3.5. Мультипас-подключение нагрузки для блочных протоколов доступа (ALUA)	14
1.4. Пулы	15
1.4.1. Обычные пулы	16
1.4.2. Быстрые пулы (DDP)	17
1.4.3. Запасные диски (hot spare)	17
1.5. Тома	17
1.6. Сети (Ethernet)	19
1.6.1. Назначение интерфейсов	19
Интерфейсы для межконтроллерного взаимодействия (интерконнекты)	19
Управляющие интерфейсы	19
Интерфейсы для подключения нагрузки	19
1.6.2. Типы интерфейсов	20
1.7. Репликация	20
1.7.1. Асинхронная репликация	20
1.7.2. Синхронная репликация	21
1.8. Производительность	22
1.8.1. Кэш записи (запись страйпами)	22
1.8.2. Кэш чтения	23
1.8.3. RDMA-синхронизация	23
1.8.4. NVMe Of-поддержка	24

1.9. Мониторинг и логирование	24
1.9.1. Мониторинг работы оборудования	24
1.9.2. Мониторинг производительности	24
1.9.3. Отправка статистики работы СХД по протоколу SNMP v2	24
1.9.4. Логирование событий	25
1.9.5. Самодиагностика работы СХД	27
1.10. Безопасность.	27
1.11. Масштабируемость	28
1.12. Интеграция	28
1.13. Подключение внешних дисковых полок	28
2. Архитектура системы	30
2.1. Сервисы системы	31
2.2. Описание работы сервисов	34
3. Устранение проблем	47
3.1. Дисковая подсистема	47
Восстановление системного рейда	47
Потеря пул (пул в статусе LOST)	48
Контроллер не видит дисков, установленных во внутренней корзине (AIC HA401).	48
Удаление диска, не отображаемого в Web-интерфейсе	49
3.2. Ресурсы	50
После перезагрузки контроллер недоступен через Web-интерфейс, сервисы не подняты	50
Увеличение и уменьшение размера тома на клиенте Linux	51
3.3. Полезная информация	52
Ротация логов	52
Режимы синхронной записи на BAUMSTORAGE	52
Клоны (клон снимка)	52
Использование кодов Рида — Соломона для восстановления данных в RAID 5, 6, B3	53
3.4. Полезные команды	54
Поддерживаемая версия NFS	54
Информация по SAS-адаптерам (утилита lsiutil)	54
Команды встроенного BMC-сервера (IPMI)	55
4. Часто возникающие вопросы	56
4.1. Архитектура	56
4.2. Гарантия и сервис	57
4.3. Диски	58
4.4. Интеграция	59
4.5. Кэширование	59
4.6. Миграция ресурсов	61
4.7. Мониторинг и логирование	62
4.8. Протокол FC	63

4.9. Протокол NFS	64
4.10. Пулы	64
4.11. Разное	65
4.12. Сети	66
4.13. Синхронная репликация	67
4.14. Снапшот	67
4.15. Технологии	68
5. Теория технологий	69
5.1. Влияние включения кэша на производительность	69
5.2. Глубина очереди	69
5.3. Влияние глубины очереди на производительность и задержку ..	69
5.4. Влияние типа RAID-массива на производительность	70
5.5. Стандартные пулы и снапшот	71
5.5.1. Описание механизма создания снапшота	71
5.5.2. Копирование при записи (COW)	72
5.5.3. Перенаправление при записи (ROW)	73
5.5.4. Аппаратная реализация снапшотов	76
5.5.5. Методика проверки возможности отказа системы при заполнении пула данными на 100 %	82
Предложения по решению проблемы заполнения снимками всего свободного пространства пула	84
5.6. Быстрые пулы и снимки	85
Обозначения и понятия	85
Снапшоты	86
5.7. Варианты кэширования записи	88
5.7.1. Кэш записи в BAUMSTORAGE SWARM	88
5.7.2. Кэш записи в UDS (функциональная FS)	90
5.7.3. Кэш записи в Huawei Dorado	91
Глоссарий	93