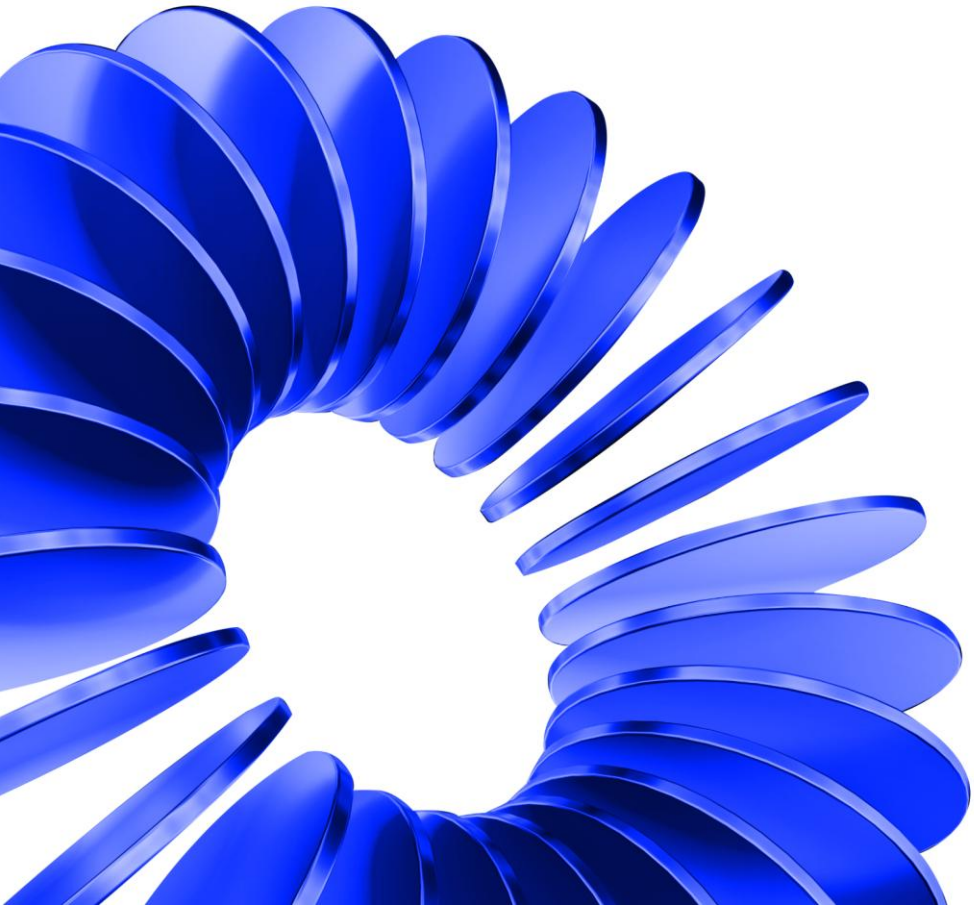
ОБЩЕСТВО С ОГРАНИЧЕННОЙ
ОТВЕТСТВЕННОСТЬЮ «РУСБИТЕХ-АСТРА»

СИСТЕМА ХРАНЕНИЯ ДАННЫХ TROK

АППАРАТНАЯ КОНФИГУРАЦИЯ КЛАСТЕРА



Москва, 2026г.

СОДЕРЖАНИЕ

1. ДОМЕН ОТКАЗА 3
2. ОБЪЕМ ДИСКОВОГО ПРОСТРАНСТВА 5
2.1. Расчет объема хранилища 5
2.2. Адаптивный запас хранилища в опоре на уровень SLA 6
2.3. Пропускная способность в сравнении со скоростью локального диска 7
3. ОСОБЕННОСТИ ОПРЕДЕЛЕНИЯ АРХИТЕКТУРЫ ХРАНИЛИЩА ДЛЯ ФАЙЛОВ БОЛЬШИХ РАЗМЕРОВ 9
4. IOPS В SDS 12
4.1. Расчёт IOPS для DRBD 12
4.2. Расчёт задержки (Latency) для DRBD 13
4.3. Расчёт IOPS по типу рабочей нагрузки 14
4.4. Производительность в деградированном состоянии 15
5. КОЛИЧЕСТВО ЯДЕР ПРОЦЕССОРА И ТРЕБОВАНИЯ SLA 16
6. РЕКОМЕНДУЕМОЕ РАСПРЕДЕЛЕНИЕ ОПЕРАТИВНОЙ ПАМЯТИ 17
7. СЕТЕВЫЕ АСПЕКТЫ ПРИ ИСПОЛЬЗОВАНИИ DRBD 18
7.1. Планирование сетевой инфраструктуры 18
7.2. Чувствительность к задержкам 19
7.3. Избыточность и связывание 20
7.4. RDMA 20
7.5. Аспекты дезагрегированного хранилища 21
7.6. Удалённая репликация и другие методы аварийного восстановления 22
7.7. Рекомендации для определенных рабочих нагрузок 23

1. ДОМЕН ОТКАЗА

Домен отказа определяет, как распределяются копии данных для обеспечения отказоустойчивости. Реплики одного ресурса (тома) должны размещаться на разных узлах хранения.

Максимум 1 копия данных на одном физическом/виртуальном сервере гарантирует, что при выходе узла из строя теряется не более одной реплики.

В качестве иллюстрации, при использовании инфраструктуры, состоящей из трёх узлов с поддержкой трёх копий данных, достигается высокая устойчивость системы к отказам и обеспечивается непрерывность доступа к ресурсам:

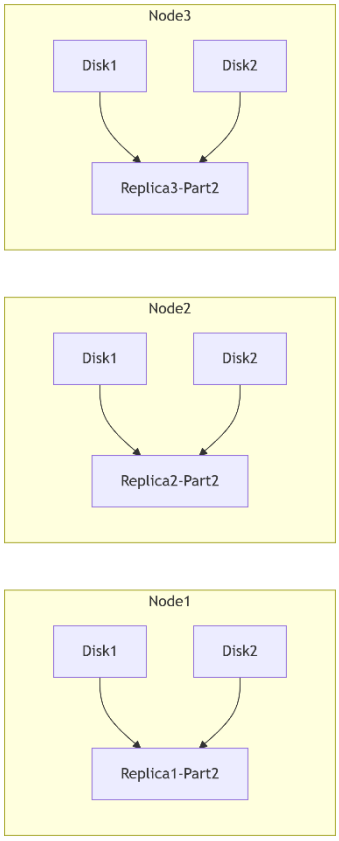


Рисунок 1 – Визуальное представление распределения данных

– В случае выхода из строя диска на узле Node1, данные автоматически восстанавливаются за счёт реплик, размещённых на диске Disk2, что исключает потерю доступа к данным.

– При отказе всего сервера Node2 программное обеспечение ожидает восстановления работы отказавшего узла. Если сервер не удаётся вернуть в строй,

система может создать новую реплику на дополнительном узле (например, Node4), если он будет добавлен в кластер.

2. ОБЪЕМ ДИСКОВОГО ПРОСТРАНСТВА

2.1. Расчет объема хранилища

Расчет общего объема дискового пространства для хранения данных является критически важным этапом при проектировании и эксплуатации систем хранения. Точное определение необходимого объема дискового пространства позволяет обеспечить надежность данных за счет репликации, предотвратить переполнение хранилища и поддерживать оптимальную производительность системы. Кроме того, учет резервного свободного пространства способствует эффективному управлению ресурсами и снижает риски деградации работы при пиковых нагрузках. Такой расчет помогает сбалансировать затраты на оборудование и эксплуатацию с требованиями к доступности и устойчивости данных.

Расчет емкости системы хранения данных (SDS) выполняется на основе трех ключевых параметров: полезного объема данных (D), коэффициента репликации (R) и норматива резервного свободного пространства (F).

Полезный объем данных (D) – это фактические данные без учета избыточности, например, 50 ТБ файлов баз данных.

Коэффициент репликации (R) определяет количество полных копий данных для обеспечения отказоустойчивости; например, при трехкратном зеркалировании R = 3.

Резерв свободного места (F) – это доля пространства, зарезервированная поверх данных с репликацией для поддержания производительности, дефрагментации и оперативного маневра. Выражается десятичной дробью (0.2 для 20%). Для систем с интенсивными операциями ввода-вывода минимальное значение может составлять 15% (F = 0.15).

Формула для расчета общего требуемого дискового пространства выглядит следующим образом:

$$SDS_{size} = D \times R \times (1 + F)$$

Данная модель демонстрирует, что общий объем складывается из емкости под реплицированные данные (D × R) и дополнительного резерва, пропорционального этому объему.

Для наглядности рассмотрим пример. При полезном объеме данных D = 100 ТБ, коэффициенте репликации R = 3 и резерве F = 0.2 (20%) расчет будет таким:

$$SDS_{size} = 100 \times 3 \times (1 + 0.2) = 300 \times 1.2 = 360 \text{ ТБ}$$

Из этого результата видна разбивка:

Данные с репликацией: 300 ТБ.

Резерв свободного места: 60 ТБ, что составляет 20% от объема реплицированных данных.

Важно отметить, что параметр F должен быть неотрицательным (F ≥ 0). Для высоконагруженных систем минимальный рекомендуемый резерв можно определить следующим образом:

$$\mathrm{F}_{min} = max(0.15, 0.5 \times (R - 1))$$

При изменении норматива резерва с F_{old} на F_{new} пересчет общего объема системы выполняется по формуле:

$$SDS_{new} = SDS_{old} \times \frac{(1 + F_{new})}{(1 + F_{old})}$$

2.2. Адаптивный запас хранилища в опоре на уровень SLA

Если требуется включить автоматическое масштабирование, то итоговый размер тома увеличивается в зависимости от уровня SLA. Адаптивный запас места для автоматического масштабирования томов рассчитывается с учётом уровня SLA следующим образом:

– При самом высоком уровне требований (SLA = 1) необходимо увеличить размер тома на 50% (умножить на 1.5), чтобы обеспечить дополнительный запас для стабильной работы и роста нагрузки.

– При среднем уровне требований (SLA = 2) размер тома увеличивается на 20% (умножается на 1.2), что даёт умеренный запас пространства.

– При низком уровне требований (SLA = 3) дополнительный запас не добавляется (умножение на 1.0), поскольку требования к масштабированию минимальны.

Таким образом, чем выше требования к качеству обслуживания, тем больший запас места резервируется для автоматического масштабирования томов, что позволяет системе гибко адаптироваться к изменяющимся нагрузкам.

2.3. Пропускная способность в сравнении со скоростью локального диска

Один высокопроизводительный NVMe-накопитель может поддерживать пропускную способность более 8 ГБ/с. Более подробная информация о каждом типе накопителя указана в таблице ниже.

Таблица 1 – Типы накопителей для ПК и серверов

Тип накопителя	Интерфейс / протокол	Ориентировочная пропускная способность	Комментарии
HDD	SATA III	до 150 МБ/с	7200 об/мин, типичные серверные и настольные диски
SATA SSD	SATA III	до 550–600 МБ/с	Популярный вариант для ПК и бюджетных серверов
NVMe SSD	PCIe 3.0 x4	~3500 МБ/с	Широко распространены в ПК и серверах
	PCIe 4.0 x4	~7000 МБ/с	Новые поколения серверных и игровых ПК
	PCIe 5.0 x4	~14000 МБ/с	Появляются в высокопроизводительных системах

SAS SSD	SAS 12 Гбит/с	~1500 МБ/с	Часто используются в серверных системах и СХД

Помимо типа накопителя необходимо также учитывать особенности протоколов доступа к данным по сетям. Их особенности передачи данных указаны в следующей таблице.

Таблица 2 – Протоколы и интерфейсы доступа в ПК и серверах

Протокол / Интерфейс	Ориентировочная пропускная способность	Комментарии
SATA III	6 Гбит/с (~600 МБ/с)	Стандартный интерфейс для HDD/SSD
SAS 12 Гбит/с	12 Гбит/с (~1,5 ГБ/с)	Серверные накопители и СХД
PCIe 3.0 x4	32 Гбит/с (~4 ГБ/с)	Интерфейс NVMe SSD
PCIe 4.0 x4	64 Гбит/с (~8 ГБ/с)	Новейшие NVMe SSD
PCIe 5.0 x4	128 Гбит/с (~16 ГБ/с)	Перспективный стандарт
iSCSI (1 GbE)	До 1 Гбит/с (~125 МБ/с)	Используется для доступа к SAN
iSCSI (10 GbE)	До 10 Гбит/с (~1,25 ГБ/с)	Более высокоскоростные SAN
NVMe-oF (Ethernet, RDMA)	До 100 Гбит/с (зависит от сети)	Высокопроизводительный доступ к NVMe накопителям по сети
NVMe-oF (Fibre Channel)	До 128 Гбит/с	Используется в SAN для серверов

3. ОСОБЕННОСТИ ОПРЕДЕЛЕНИЯ АРХИТЕКТУРЫ ХРАНИЛИЩА ДЛЯ ФАЙЛОВ БОЛЬШИХ РАЗМЕРОВ

Для эффективного управления файлами большого размера (50+ ТБ) в программно-определяемых хранилищах требуется комплексный подход, объединяющий сегментацию данных, передовые методы обеспечения отказоустойчивости и специализированные технологии хранения.

Механизмы избыточности обеспечивают сохранность данных при сбоях оборудования, создавая дополнительные копии информации или математически рассчитанные корректирующие коды. Традиционные решения включают RAID-массивы (например, RAID 6 с двойной чётностью), где данные распределяются по нескольким дискам с резервированием, но они уязвимы к ошибкам чтения (URE) при восстановлении больших объёмов. Более совершенные файловые системы, такие как ZFS, комбинируют контрольные суммы данных, мгновенные снапшоты и RAID-Z, который динамически проверяет целостность блоков, минимизируя риски повреждения. Более подробная информация о механизмах избыточности представлена в таблице ниже.

Таблица 3 – Сравнение технологий хранения для файла 50 ТБ

Параметр	RAID 5 (HDD)	RAID 6 (HDD)	RAID-Z2 (ZFS)
Нужно дисков	7×8 ТБ HDD	8×8 ТБ HDD	6×10 ТБ HDD
Полезная ёмкость	48 ТБ	48 ТБ	40 ТБ
Отказоустойчивость	1 диск	2 диска	2 диска + контрольные суммы
Время восстановления	20–30 часов	40–60 часов	15–25 часов
Риск URE (Backblaze Stats 2023)	Критичный (33%)	Высокий (15%)	Низкий (<1%)

Влияние различных технологий избыточности на пропускную способность определяется их архитектурой, механизмами расчета четности, распределением нагрузки и требованиями к вычислительным ресурсам. Рассмотрим сравнительный анализ указанных технологий.

RAID 5 на базе HDD обеспечивает чтение, близкое к линейному масштабированию — примерно N-кратную скорость одного диска; например, для массива из четырёх дисков последовательная скорость чтения достигает до 600 МБ/с. Однако запись существенно ограничена из-за необходимости вычисления четности: каждая операция записи требует четырёх I/O операций — чтения данных и старой четности, вычисления новой четности и записи результата. В итоге скорость записи составляет 25–75 % от производительности RAID 0, что для четырёх дисков примерно 150–400 МБ/с. Ключевыми факторами здесь являются «пенальти» записи, обусловленное четырьмя I/O на операцию, аппаратные RAID-контроллеры с кэшированием, снижающие задержки, и значительное замедление случайных записей (до 30 МБ/с для 4K-операций).

RAID 6, также реализуемый на HDD, демонстрирует чтение, аналогичное RAID 5, то есть примерно N-кратное ускорение. Запись при этом на 20–30 % медленнее RAID 5 из-за необходимости вычислять двойную четность; для четырёх дисков скорость записи составляет порядка 100–300 МБ/с. Особенности RAID 6 включают необходимость вычисления двух блоков четности, что приводит к шести I/O операциям на запись, а также более длительное восстановление при сбоях — на 50–100 % дольше, чем в RAID 5, что уменьшает доступную пропускную способность при отказах.

Технология RAID-Z2 в файловой системе ZFS обеспечивает высокую скорость чтения благодаря параллельному доступу к виртуальным устройствам (vdev). Например, для массива из восьми дисков по 10 ТБ скорость чтения достигает до 800 МБ/с. Запись в RAID-Z2 быстрее традиционного RAID 6 благодаря использованию метода Copy-on-Write, исключающего операции «read-modify-write» для мелких записей, а также адаптивным страйпам, динамически

подстраивающим размер полосы под нагрузку. Важными факторами являются распределение данных с чередованием блоков данных и четности, использование кэширования ARC в оперативной памяти и L2ARC на SSD, а также аппаратное ускорение вычислений четности с помощью векторных инструкций AVX.

Таким образом, каждая технология избыточности обладает уникальными характеристиками, влияющими на пропускную способность, и выбор оптимального решения должен учитывать баланс между производительностью, надежностью и затратами ресурсов.

4. IOPS В SDS

IOPS (Input/Output Operations Per Second) — фундаментальная метрика производительности систем хранения, определяющая количество операций чтения/записи в секунду.

Расчёт производительности DRBD (Distributed Replicated Block Device) основывается на учёте ограничений как со стороны дисковой подсистемы, так и сети, а также особенностей синхронной репликации.

4.1. Расчёт IOPS для DRBD

Общая максимальная производительность по операциям ввода-вывода (Total IOPS) определяется минимальным значением между производительностью локального диска и пропускной способностью сети, скорректированной на количество реплик:

$$IOPS_{total} = \min(IOPS_{disk}, \frac{IOPS_{network}}{replica_count})$$

Здесь $IOPS_{network}$ рассчитывается как минимальное значение между максимальной скоростью записи одного диска и пропускной способностью сети, делённой на удвоенный размер операции записи (учитывая необходимость передачи данных в обе стороны репликации):

$$IOPS_{network} = \min(IOPS_{single_disk_write}, \frac{network_bandwith}{write_io_size \times 2})$$

Возьмем в качестве примера следующие параметры для синхронной записи:

- Дисковая подсистема: 100 000 IOPS (SATA SSD);
- Сеть: 10 Gbps (1.25 GB/s);
- Размер операции записи (write IO size): 4 KB;
- Количество реплик (replica_count): 2.

Для вычисления сетевого IOPS подставьте значения в формулу:

$$IOPS_{network} = \min(100\,000, \frac{1{,}25 \times 10^9}{4\,096 \times 2})$$

Вычисления знаменателя:

$$4\,096 \times 2 = 8192 \text{ байта}$$

Вычисления дроби:

$$\frac{1{,}25 \times 10^9}{8\,192} \approx 152\,587\ IOPS$$

Таким образом,

$$IOPS_{network} = \min(100\,000,\ 152\,587) = 100\,000\ IOPS$$

при данных параметрах ограничивающим фактором является дисковая подсистема с 100 000 IOPS, а пропускная способность сети позволяет обрабатывать до ≈152 587 IOPS, что выше дискового ограничения. Следовательно, максимальная производительность по IOPS будет определяться скоростью дисковой подсистемы.

В стандартной конфигурации DRBD все клиентские операции чтения выполняются только на том узле, где ресурс находится в состоянии Primary. Производительность чтения при этом ограничивается производительностью локального диска этого узла и не масштабируется с количеством реплик.

$$IOPS_{effective_read(\text{Primary-Only})} = IOPS_{local_disk_read}$$

Для примера:

- Локальный диск: 1000 IOPS;
- Реплик: 2.

$$IOPS_{effective_read} = 1000$$

(ограничено одним диском primary-узла)

4.2. Расчёт задержки (Latency) для DRBD

При записи в синхронном режиме задержка складывается из максимума между задержкой диска и задержкой сети, умноженного на количество реплик:

$$wrire_{latency} = \max\left(disk_{latency},\ network_{latency}\right) \times replica_count$$

Для чтения, если оно локальное, задержка равна задержке диска:

$$read_{latency} = disk_{latency}$$

Пример расчёта задержки записи:

- Задержка диска (SSD): 2 мс;
- Время передачи пакета по сети (RTT): 0.5 мс.

С учётом того, что синхронная запись подразумевает обмен данными туда и обратно (RTT умножается на 2):

$$wrire_{latency} = 2\,ms + (0{,}5\,ms\ \times\ 2) = 3\,ms$$

Сбор данных для оценки производительности DRBD можно осуществить с помощью инструмента fio, например, следующей командой:

```
   fio --filename=/dev/drbd1000 --rw=randwrite --ioengine=libaio --direct=1 --bs=4k
--numjobs=1 --iodepth=32 --runtime=60 --name=test
```

Данная команда выполняет случайные записи блоками по 4 КБ с асинхронным вводом-выводом, что позволяет получить реальные показатели IOPS и задержек на устройстве DRBD.

4.3. Расчёт IOPS по типу рабочей нагрузки

На значения показателей IOPS оказывает существенное влияние характер и тип данных, обрабатываемых в хранилище. При вычислении итоговой производительности системы целесообразно учитывать специфику различных типов рабочих нагрузок, что позволяет более точно оценить реальные возможности ввода-вывода:

– Если рабочей нагрузкой является база данных (DB), то фактическая производительность увеличивается на 50% (умножается на 1.5). Это связано с тем, что базы данных часто используют мелкие и частые операции ввода-вывода, которые могут лучше использовать возможности дисковой подсистемы.

– Если рабочая нагрузка – виртуальная машина (VM), то производительность увеличивается на 20% (умножается на 1.2). Виртуальные машины создают смешанные нагрузки, которые обычно более требовательны, чем простое файловое хранилище, но менее интенсивны, чем базы данных.

– Для остальных типов нагрузок (например, файловое хранилище — FS) производительность принимается без изменений (умножается на 1.0), то есть без дополнительного коэффициента.

4.4. Производительность в деградированном состоянии

При выходе из строя одного из узлов кластера производительность системы хранения данных снижается в соответствии со следующими практическими корректировками:

– Операции чтения: пропускная способность снижается до 30% от номинальной в конфигурации с трёхкратной репликацией при потере одной копии данных. Снижение обусловлено необходимостью реорганизации процессов доступа к данным.

– Операции записи: производительность ограничивается пропускной способностью сети, выделенной для процесса восстановления данных, что может приводить к увеличению задержек.

Для оценки времени восстановления работоспособности системы применяется формула:

$$\mathrm{T} = \frac{V}{S_{min}} \times K$$

Где:

V – объём данных, подлежащих восстановлению;

S_{min} – минимальная скорость передачи данных, определяемая пропускной способностью сети и значением параметра disk.c-min-rate в конфигурации DRBD;

K – коэффициент, учитывающий дополнительные факторы (например, нагрузку на CPU или параллельные процессы).

Указанная модель позволяет прогнозировать длительность восстановления при проектировании отказоустойчивых конфигураций.

5. КОЛИЧЕСТВО ЯДЕР ПРОЦЕССОРА И ТРЕБОВАНИЯ SLA

Этот принцип расчёта количества ядер процессора (CPU) основан на уровне требуемого соглашения об уровне обслуживания (SLA), который определяет, насколько высокую производительность и надёжность нужно обеспечить для каждого ресурса:

– Если SLA равен 3 (низкий уровень требований), на один ресурс выделяется 2 ядра CPU.

– Если SLA равен 2 (средний уровень требований), на один ресурс выделяется 4 ядра CPU.

– Если SLA равен 1 (высокий уровень требований), на один ресурс выделяется 8 ядер CPU.

То есть, чем выше требования к качеству обслуживания (ниже значение SLA), тем больше ядер процессора необходимо выделять на каждый ресурс для обеспечения нужной производительности и стабильности работы.

При этом минимальное количество ядер, достаточное для эффективной работы DRBD при стандартных нагрузках – 2 ядра CPU на ресурс, с возможностью увеличения в зависимости от конкретных условий эксплуатации и требований к производительности. Это базовые значения, требующие валидации под нагрузкой.

6. РЕКОМЕНДУЕМОЕ РАСПРЕДЕЛЕНИЕ ОПЕРАТИВНОЙ ПАМЯТИ

Практический минимальный объем памяти составляет 8 GB ОЗУ на узел хранения (рекомендуемый не менее 64 GB), особенно для производственных или больших объемов. Общее правило заключается в том, что на 1 TiB хранилища DRBD на одноранговый узел требуется около 32 MiB ОЗУ. Если у вас десять ресурсов DRBD, каждый из которых имеет объем 100 GiB реплицируемых между двумя одноранговыми узлами, вам потребуется не менее 64 MiB памяти только для репликации DRBD. DRBD поддерживает максимальный размер устройства 1 PiB (1024 TiB) на ресурс.

Если вы используете программное обеспечение и DRBD в гиперконвергентной среде, рассмотрите возможность значительного расширения объёма памяти (от 256 ГБ до 1024 ГБ) для комфортного размещения рабочих нагрузок приложений и требований к памяти DRBD.

Это базовые значения, требующие валидации под нагрузкой.

7. СЕТЕВЫЕ АСПЕКТЫ ПРИ ИСПОЛЬЗОВАНИИ DRBD

Производительность синхронной репликации в DRBD существенно определяется характеристиками сетевого канала, в частности его пропускной способностью и задержками передачи данных. Ограниченная пропускная способность или высокая латентность сети создают узкое место при передаче реплицируемых записей и получении соответствующих подтверждений, что негативно сказывается на общей эффективности репликационного процесса.

Целесообразно физически разделить трафик репликации, клиентский доступ и управление для минимизации конкуренции за сетевые ресурсы и снижения задержек.

7.1. Планирование сетевой инфраструктуры

Требования к сетевой топологии и оборудованию определяются типом передаваемых данных и необходимым уровнем сервиса:

– Сеть репликации данных: требуется выделенный физический сетевой интерфейс.

– Рекомендуемая пропускная способность: 25 или 100 Gigabit Ethernet (GbE) для обеспечения высокой скорости синхронизации данных между узлами.

– Размер MTU: 9000 байт (Jumbo Frames) для повышения эффективности передачи крупных блоков данных.

– Задержка (latency): должна составлять менее 100 микросекунд для исключения задержек в процессе репликации.

– Клиентская сеть доступа к данным: требуется выделенный физический интерфейс.

– Для блоковых протоколов (iSCSI, NVMe-oF): рекомендуется использовать линии связи 10 или 25 GbE.

– Для файловых протоколов (NFS, SMB): достаточно пропускной способности от 10 GbE и выше.

– Сеть управления: для передачи служебной информации может использоваться отдельный интерфейс.

– Пропускная способность: 1 GbE является достаточной для большинства задач мониторинга и управления.

Разделение сетей позволяет изолировать виды трафика, что повышает отказоустойчивость и упрощает диагностику проблем производительности.

7.2. Чувствительность к задержкам

10-гигабитное Ethernet-соединение в реальном выражении обеспечивает пропускную способность около 1,25 ГБ/с. Если скорость локального диска превышает пропускную способность сети, сеть становится узким местом для ввода-вывода при использовании DRBD для репликации данных.

Чтобы в полной мере использовать производительность NVMe, рассмотрите сетевые соединения 25, 40 или 100 GbE между узлами репликации DRBD. Однако, поскольку более быстрая сетевая инфраструктура (сетевые карты, коммутаторы и т.д.) обходится дороже, тщательно оцените соотношение затрат и выгод.

Синхронная запись DRBD требует подтверждения от удаленных узлов. Даже при достаточной пропускной способности высокая задержка может снизить производительность. Рекомендуется использовать сетевое оборудование с низкой задержкой, например, оптоволоконное с минимальным количеством переходов. Если это невозможно, рассмотрите асинхронный режим репликации DRBD.

Основные типы интернет-соединения и особенности передачи данных при использовании каждого из них перечислены в таблице ниже.

Таблица 4 – Основные типы интернет-соединений

Параметр	Ethernet (Стандартный)	RoCE (RDMA over Ethernet)
Скорость передачи	1 Гбит/с – 800 Гбит/с	25 – 400 Гбит/с

Задержка (latency)	50 – 200 мкс	1 – 5 мкс
Пропускная способность	До 800 Гбит/с (на порт)	До 400 Гбит/с
Технология доступа	TCP/IP, UDP	RDMA поверх Ethernet
Требования к инфраструктуре	Стандартные коммутаторы	Совместимость с DCB (Data Center Bridging)

7.3. Избыточность и связывание

Сетевое связывание, например, протокол управления агрегацией каналов (LACP), может обеспечить отказоустойчивость и объединить несколько каналов для увеличения пропускной способности. DRBD имеет опцию балансировки нагрузки для трафика репликации TCP/IP. Вы также можете настроить балансировку нагрузки DRBD для использования нескольких каналов. Стабильная доставка пакетов, а также минимальный джиттер и задержка имеют решающее значение при настройке резервирования и объединения сетей.

7.4. RDMA

При использовании TCP/IP вы можете столкнуться с падением производительности системы при достижении скорости передачи данных 10 гигабит в секунду (Гбит/с). После этого производительность системы продолжает существенно снижаться. Если вы столкнулись с таким снижением производительности или ожидаете его, рассмотрите возможность использования альтернативного транспортного протокола — RDMA. RDMA передает данные напрямую между физической памятью двух систем по высокоскоростным сетям, например, InfiniBand или RoCE, минуя процессор и операционную систему. Это снижает задержку и нагрузку на процессор, повышая пропускную способность и эффективность сред хранения или кластеризации, критически важных для производительности.

Реализация RDMA требует выделенного специализированного оборудования. Перед покупкой оборудования необходимо оценить затраты и выгоды.

7.5. Аспекты дезагрегированного хранилища

При использовании дезагрегированной архитектуры хранилища, где хосты подключаются к отдельным узлам хранения, необходимо учитывать следующее.

Каждый ресурс DRBD зеркалирует данные как минимум на два узла хранения с данными. В дезагрегированной архитектуре хранилища ваши вычислительные хосты отправляют каждую операцию записи по сети на каждый узел хранения, содержащий реплику, так называемые полнодисковые узлы. При использовании другого интерфейса хранения, например, NFS, NVMe-oF или iSCSI, операции записи отправляются только на основной узел DRBD. Основной узел затем реплицирует изменения на вторичные узлы. Если хост к сети хранения данных подключен только по одному каналу 10GbE, один и тот же блок данных должен передаваться дважды в типичном трёхузловом кластере (2 полнодисковых узла + 1 бездисковый узел-посредник): по одному разу на каждый полнодисковый узел-реплику. Поскольку канал 10GbE имеет фиксированную максимальную пропускную способность, около 1,25 ГБ/с в реальных условиях, доступная полоса пропускания фактически распределяется или «разделяется» между этими одновременными потоками.

Это означает, что вам может потребоваться рассмотреть вопрос распределения пропускной способности и возможных узких мест в сети.

Совместное использование пропускной способности: при насыщении канала связи каждый путь записи к каждому узлу может использовать только часть общей пропускной способности 10GbE.

Повышенная пропускная способность диска и узкие места сети: даже если ваши высокопроизводительные локальные диски, такие как NVMe или SSD-накопители, способны справиться с более высокой пропускной способностью, трафик репликации DRBD по одному каналу 10GbE может стать ограничивающим фактором.

Для решения этих проблем можно рассмотреть следующие варианты:

Объединение каналов: объединение нескольких каналов 10GbE, например, с помощью LACP или балансировки нагрузки DRBD, для увеличения общей доступной пропускной способности и обеспечения избыточности на хостах.

Более быстрые сети: рассмотрите переход на 25GbE, 40GbE или 100GbE, если производительность диска и требования к рабочей нагрузке оправдывают это, а преимущества оправдывают затраты. Разделение трафика: используйте отдельные физические или логические сети для отделения трафика репликации от обычного трафика данных, снижая конкуренцию за пропускную способность. Короче говоря, поскольку каждая запись должна быть отправлена на два (или более) узла хранения, этот канал 10GbE обрабатывает несколько параллельных потоков данных одной и той же операции записи, что фактически снижает доступную полосу пропускания для каждого потока, если канал близок к насыщению. На практике ваша «используемая» полоса пропускания для каждого копирования данных может быть уменьшена вдвое. Вам следует запланировать достаточную суммарную полосу пропускания, чтобы избежать насыщения при параллельном выполнении нескольких ресурсов DRBD или больших объемов записи.

7.6. Удалённая репликация и другие методы аварийного восстановления

Если не настроено иное, DRBD по умолчанию реплицирует данные синхронно. DRBD также поддерживает асинхронный или полусинхронный режимы репликации между географически разнесёнными локациями, что обычно используется при создании плана аварийного восстановления данных. Если пропускной способности или задержки между сетевыми каналами недостаточно для полностью синхронной записи, у вас есть альтернативы:

Резервное копирование и отправка снимков

Если пропускная способность ограничена или задержка слишком велика для непрерывной репликации, вы можете периодически создавать снимки ресурсов DRBD, управляемых программным обеспечением, и переносить их на удаленную

площадку, используя доставку снимков (также называемую резервной копией). Для доставки снимков требуется, чтобы ресурсы поддерживались томами LVM или ZFS с тонким выделением ресурсов. Этот тип репликации отличается от стандартной репликации DRBD, но может быть эффективным подходом для аварийного восстановления в средах с ограниченными сетевыми ресурсами.

Асинхронные протоколы DRBD (A)

При использовании протокола асинхронной репликации DRBD A операции записи подтверждаются локально до того, как одноранговый узел DRBD подтвердит удаленную запись. Это снижает локальную задержку ввода-вывода, но создает риск потери данных в случае сбоя основного узла до завершения удаленной репликации, поскольку у удаленного узла может не быть актуальной копии данных.

7.7. Рекомендации для определенных рабочих нагрузок

Базы данных

При использовании DRBD для репликации данных баз данных, особенно для данных, связанных с обработкой онлайн-транзакций (OLTP) или онлайн-аналитической обработкой (OLAP), критически важны высокая производительность ввода-вывода (IOPS) и низкая задержка. Для этой задачи рекомендуется использовать NVMe или SSD-хранилища. Также стоит рассмотреть использование сетевых соединений 25, 40 или 100 GbE, чтобы обеспечить соответствие производительности дисков. RAID 10 или локальные SSD для чтения помогают обеспечить локальную отказоустойчивость без немедленного переключения ресурсов в случае сбоя диска в массиве RAID.

Виртуализация и контейнеры

При использовании DRBD для репликации данных хранилища, поддерживающего виртуализированные рабочие нагрузки или контейнеры, случайные паттерны ввода-вывода, обычно встречающиеся в этом сценарии, извлекут выгоду из использования быстрых SSD или NVMe-дисков. Рекомендуется обеспечить наличие 64 ГБ или более оперативной памяти на узел,

или больше для гиперконвергентных сред. Рекомендуется использовать как минимум 10 GbE сетевые соединения.

Архивирование и резервное копирование

При использовании DRBD для создания высокодоступных архивных или резервных данных приоритетом является большая емкость, а не высокая производительность. Для этой задачи можно использовать жесткие диски SAS или SATA как экономически эффективное решение. Использование RAID 6 или RAID 10 может обеспечить баланс между емкостью и целями локальной избыточности.