



скала^р

Аварийное восстановление
Postgres Pro Enterprise в Машинах МБД.П
при помощи pg_probackup

Бурцев Александр – владелец продукта
Власов Алексей – архитектор машины

Машина баз данных Скала^р МБД.П

Программно-аппаратный комплекс с СУБД Postgres Pro Enterprise



Характеристики:

- Высокая надежность и отказоустойчивость
- Высокопроизводительная сеть внутреннего взаимодействия 100 Гбит/с низкими задержками
- Объем баз до 160 ТБ (без сегментирования)
- Встроенная система резервного копирования от 12 ТБ/час

Замещаемые технологии:

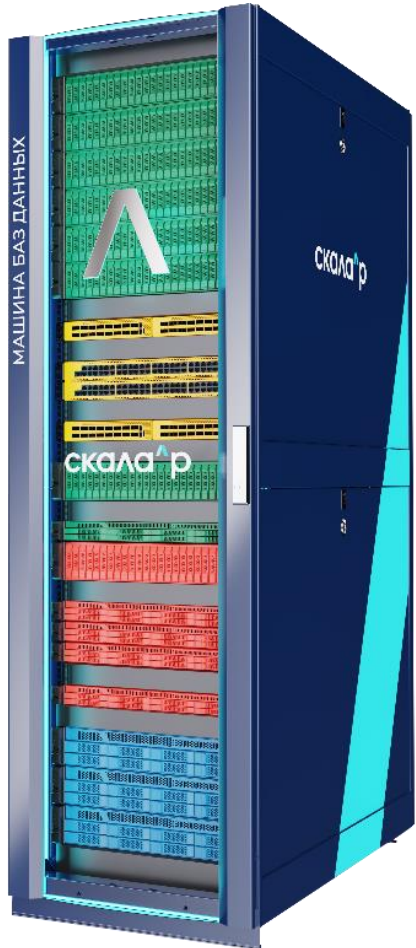
- Oracle Exadata
- Oracle Data Guard, MSSQL AlwaysOn
- Oracle Zero Data Loss Appliance

Рекомендовано
от 12 000
транзакций в секунду

Пиковая
производительность
65 000+
транзакций/сек



Модульная архитектура Машины Скала^р МБД.П



Модуль Баз Данных

- Кластер из трех узлов (мастер, синхронная реплика, асинхронная реплика)
- Повышение производительности за счет записи на мастер, а чтения с синхронной реплики

Блок коммутации и агрегации

- Сети внешнего доступа
- Сети внутреннего взаимодействия до 100 ГБ/с
- Выделенная сеть для управления и мониторинга

Блок мониторинга и управления

- Управление эксплуатацией и автоматизация критических процедур
- Мониторинг состояния всех компонентов Машины

Блок хранения резервных копий

- Хранение резервных копий БД и WAL
- Отказоустойчивый кластер узла хранения
- Окно полного резервного копирования до 10 часов

Командная работа над тестами производительности резервного копирования Скала^р МБД.П и МХД.О



скала^р

производитель
ПАК для
высоконагруженных
IT-решений



системный интегратор,
проектирует и внедряет
проекты в критически
важные ИС



российский разработчик
СУБД Postgres Pro

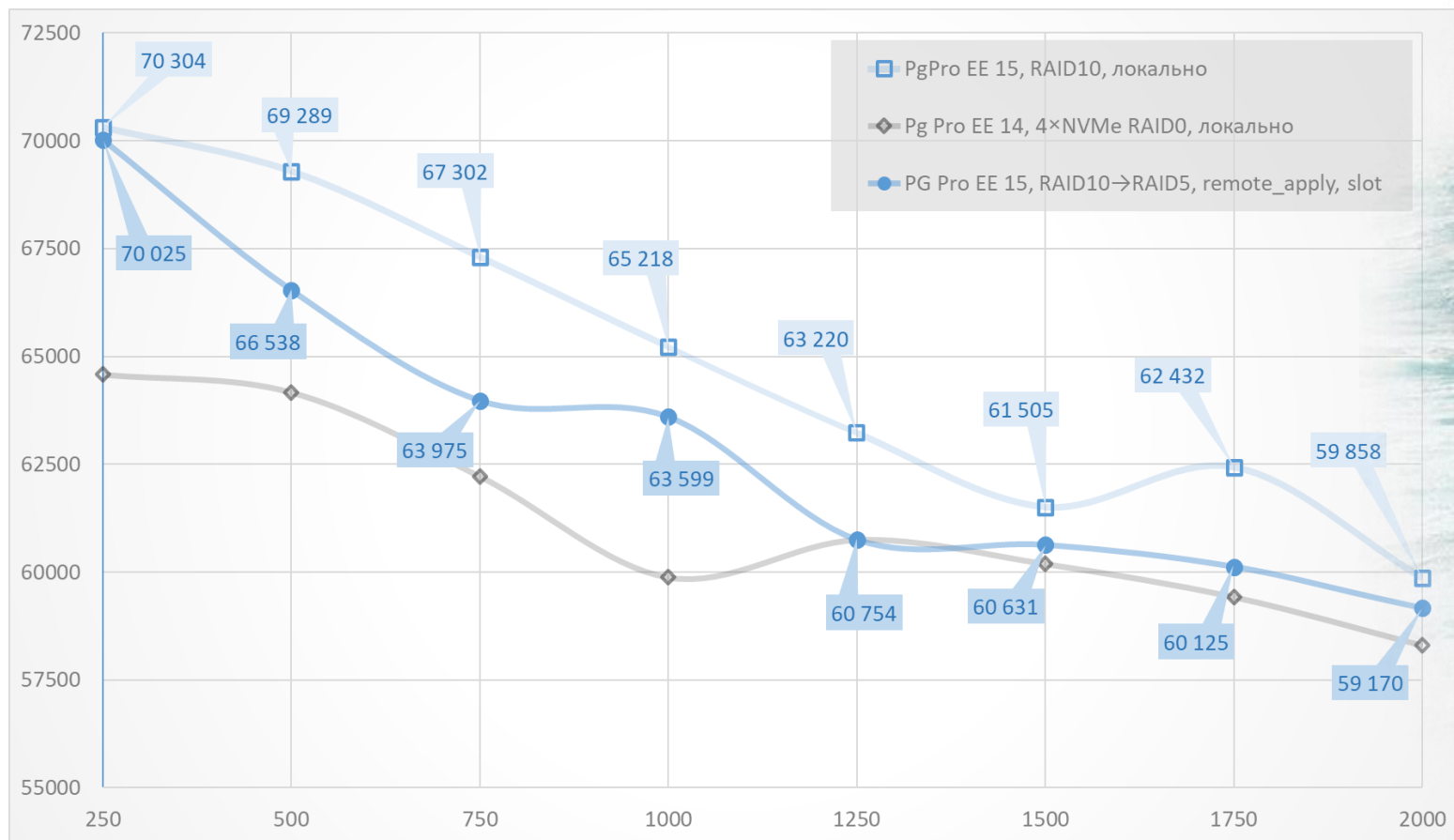




Работа Postgres Pro Enterprise Certified в ПАК Скала^р МБД.П

Производительность Скала[^]р МБД.П на базовых тестах ТСП-В

TPS для СУБД Postgres Pro Enterprise Certified в ПАК



Машина баз данных Скала^Λр МБД.П



Технические показатели

Типовые размеры томов данных
СУБД

Тип 1 до **10** ТБ

Тип 2 до **20** ТБ

Тип 3 до **40** ТБ

Тип 4 до **80** ТБ

Тип 5 до **160** ТБ

Интеграция в инфраструктуру
под ландшафт Заказчика

LDAP
и **OpenIDConnect**

KeyCloak

Время отработки
сетевых отказов
уровня коммутации

не более **500** мс

Производительность*

до **65 000** TPS

Время переключения
сервиса СУБД
в случае отказа

не более **1-2** минуты

Скорость восстановления

6-22 ТБ/час

* На объеме горячих данных 1ТБ (объем с которым постоянно работает приложение) и до 5000 пользователей



Особенности работы СУБД более 40ТБ
в Машинах баз данных Скала^р МБД.П

Необходимость в быстром восстановлении СУБД



Требования к скорости работы:

- Полная резервная копия и работы с ней за 1 ночь (до 10 часов)
- Быстрые инкременты 3-х видов: DELTA, PAGE, PTRACK
- Merge инкрементов в новый полный бэкап
- Восстановление и валидация от 1 до 10 часов

Требования к надежности:

- Несколько полных копий СУБД и инкременты для восстановления на PITR
- Сохранность резервной копии при аварии до 50% дисков МХД.О
- 3-2-1 стратегия хранения резервных копий
- Интеграция с СРКиВД отечественных производителей

Технические характеристики:

- Скорость резервного копирования в Узел хранения от 13 ТБ\час
- Скорость резервного копирования МХД.О от 10 ТБ\час



МБД.П #1

МБД.П #2

МХД.О

Машина хранения данных Скала^р МХД.О

для хранения до 4 трлн объектов на Машину с скоростью до 1 млн объектов в секунду



Сценарии применения:

- Хранение
- Сохранность до 8ПБ на одну Машину данных при отказе до 50% узлов

Ключевые преимущества:

- Совместимость с Amazon S3 API
- Безопасность: журналирование и аудит
- Самый быстрый и надежный SDS с отечественной кодовой базой
- Катастрофоустойчивость «из коробки»
- Сжатие данных до 20 раз
- Глубокая интеграция с МБД.П

Замещаемые продукты и технологии:

Традиционные системы хранения файлов

SAN, NAS, HDFS

Сервисы

Amazon S3



Инкрементальное копирование в pg_probackup



DELTA:

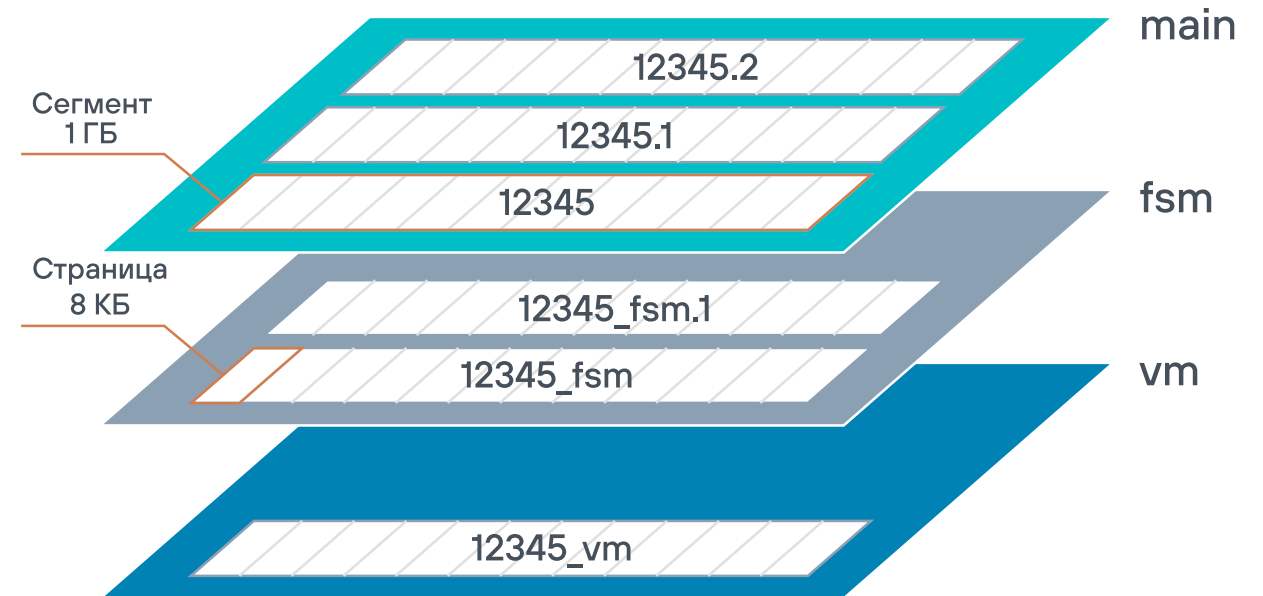
- Сравнение всех файлов и вычисление разницы по сегментам данных

PAGE:

- Сравнение WAL записей и вычисление разницы изменившихся сегментов

PTRACK:

- Сканирование карты изменений блоков данных

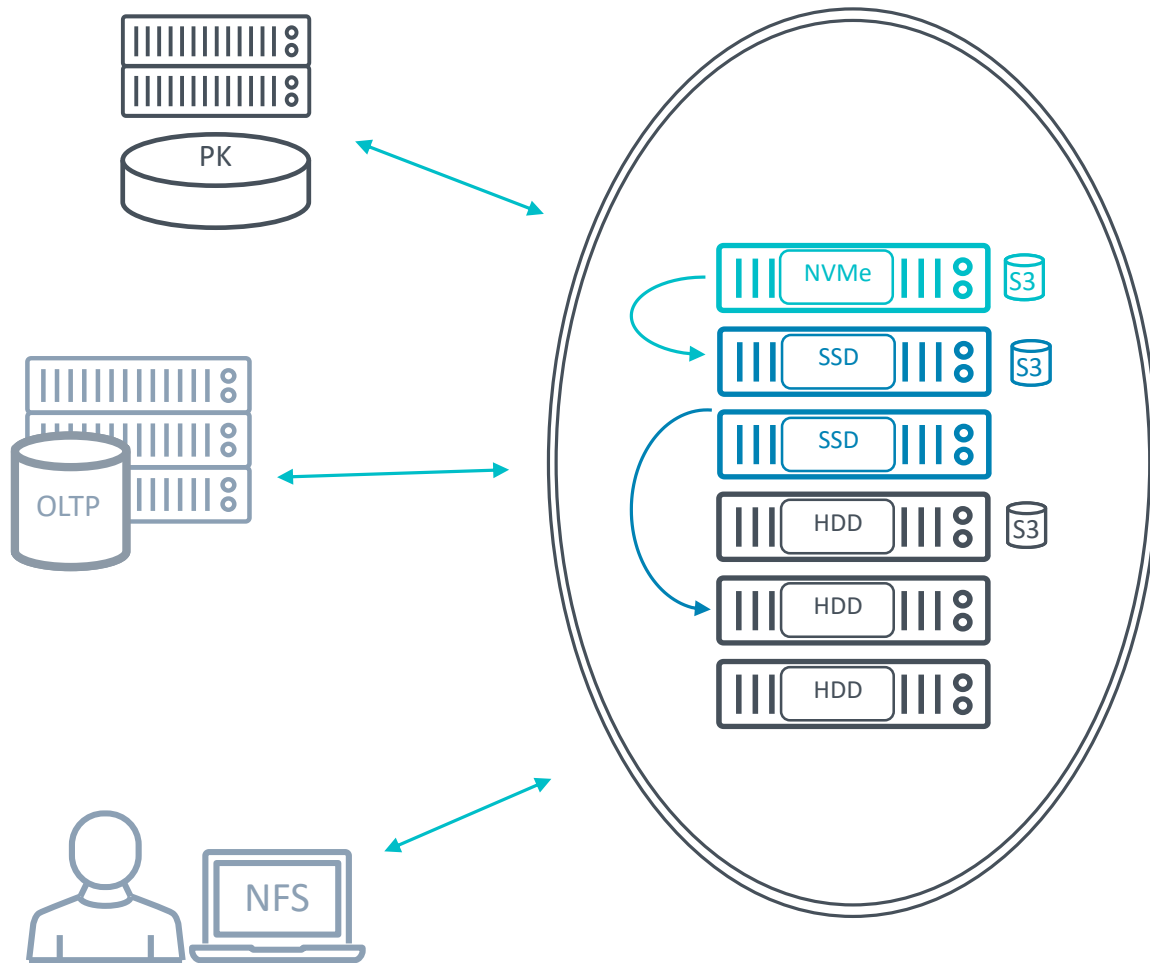


- fsm — free space map
- vm — visibility map

Машина хранения данных Скала^р МХД.О



Архитектура с разными накопителями NVMe, SSD, HDD



- Производительность до 2Гбайт/с на один узел хранения при записи (24 SSD диска)
- Производительность до 5000 объектов в секунду на один узел хранения при записи
- Реализация классов хранения с возможностью переноса архивных резервных копий на носители с высокой плотностью хранения
- **Интеграция с ПО Спектр и pg_probackup Enterprise**
- Платформа управления парком Машин Скала^р

Цели тестирования Машин МБД.П и МХД.О



Задача тестирования:

- Создать сценарий отказа Вычислительного узла СУБД
- Определить границы применимости МХД.О и МБД.П
- Определить границы применимости Узлов хранения в МБД.П
- Определить требования к оборудованию для новых Машины

Платформа для теста:

- Сервер «Аквариус» T50 D224CF

Технические характеристики:

- CPU 2x Xeon Gold 6248R
- RAM 1Tb DDR4
- 8x NVMe 3Tb
- 2x100G Eth



Сетевая конфигурация Машины МБД.П с СРК медиа-сервером



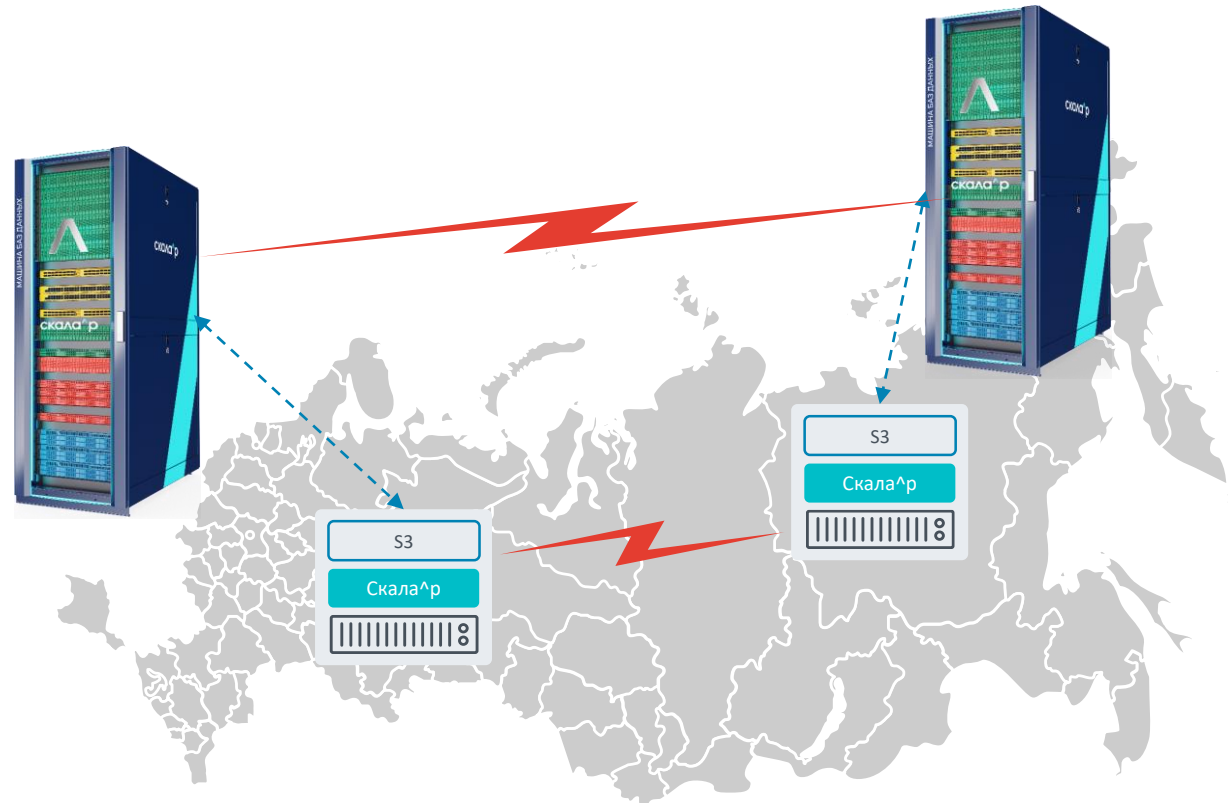
- Прямая запись в МХД.О S3
- Запись через медиа-сервер по API (только мета-данные)
- Запись через медиа-сервер (очень медленно решение)



Аварийная отказоустойчивость в МБД.П и МХД.О



- Гео распределенный кластер МХД.О и МБД.П без единой точки отказа
- Сохранность данных при отказе до 50% узлов хранения в МХД.О





Postgres Pro Backup Enterprise

Восстановление реплики

pg_probackup catchup

Скорость работы pg_probackup catchup



- pg_probackup catchup FULL:
- Объем БД: 3014GB
- Время копирования БД: 7m:52s
- Количество потоков: 96
- Скорость копирования данных: ~ 23TB/h

```
INFO: Source PGDATA size: 3014GB (excluded 0)
INFO: Start LSN (source): 2CE6/A8000028, TLI: 37
INFO: Start transferring data files
INFO: wait for pg_stop_backup()
INFO: pg_stop_backup() successfully executed
INFO: stop_stream_lsn 2CE6/ABF24000 currentpos 2CE6/ABF24000
INFO: Databases synchronized. Transferred datafiles size: 1449MB, transferred wal size: 64MB, time elapsed: 7m:52s
```

Устройство отказоустойчивого кластера в МБД.П



Кластер – 3 экземпляра

- Локальные диски
- Поточная репликация

Отказ мастера

- Автоматическое восстановление сервиса на синхронной реплике

Отказ одной из реплик

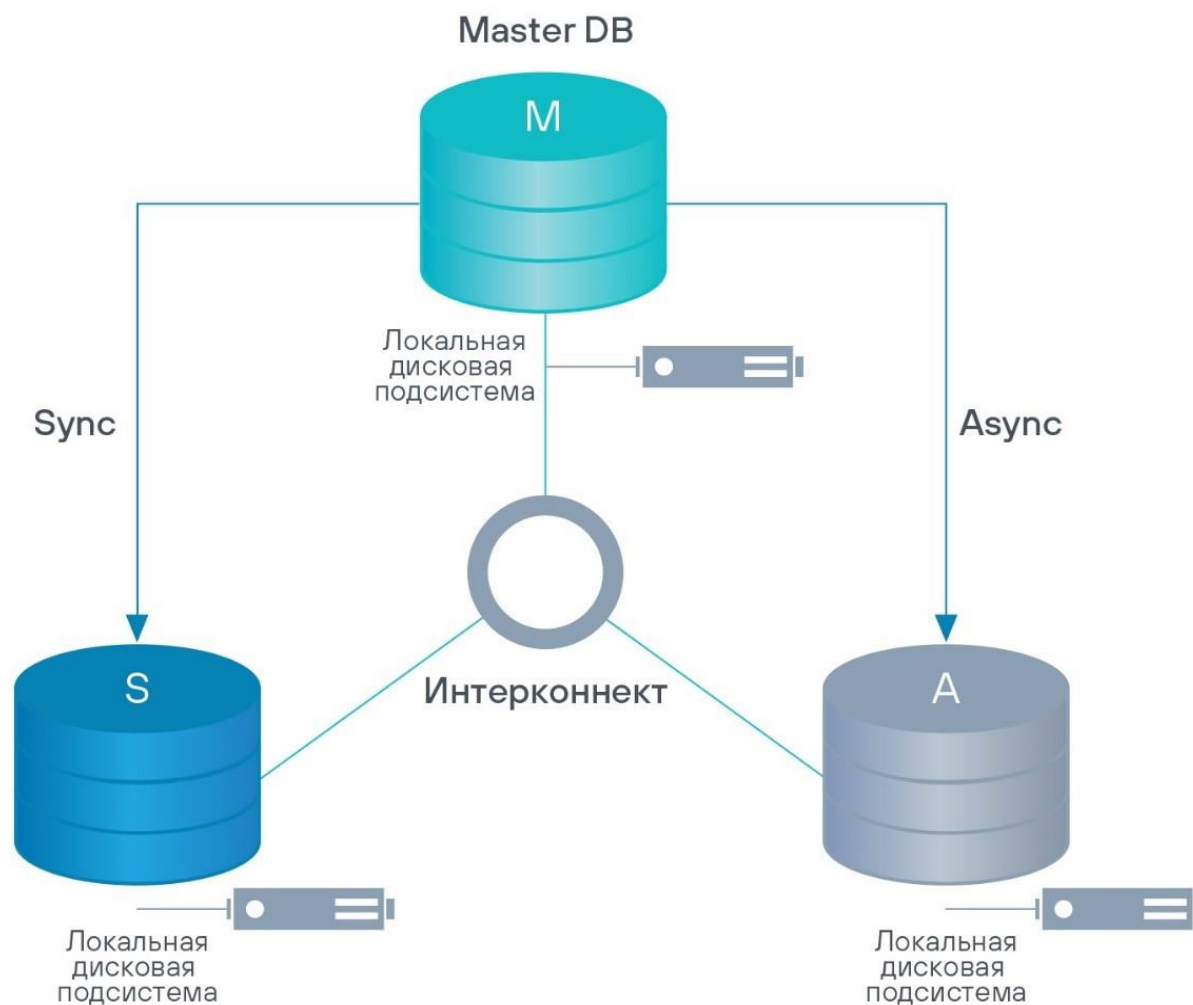
- Не влияет на работу сервиса

Отказ двух экземпляров

- Остановка сервиса БД

Отказ экземпляра

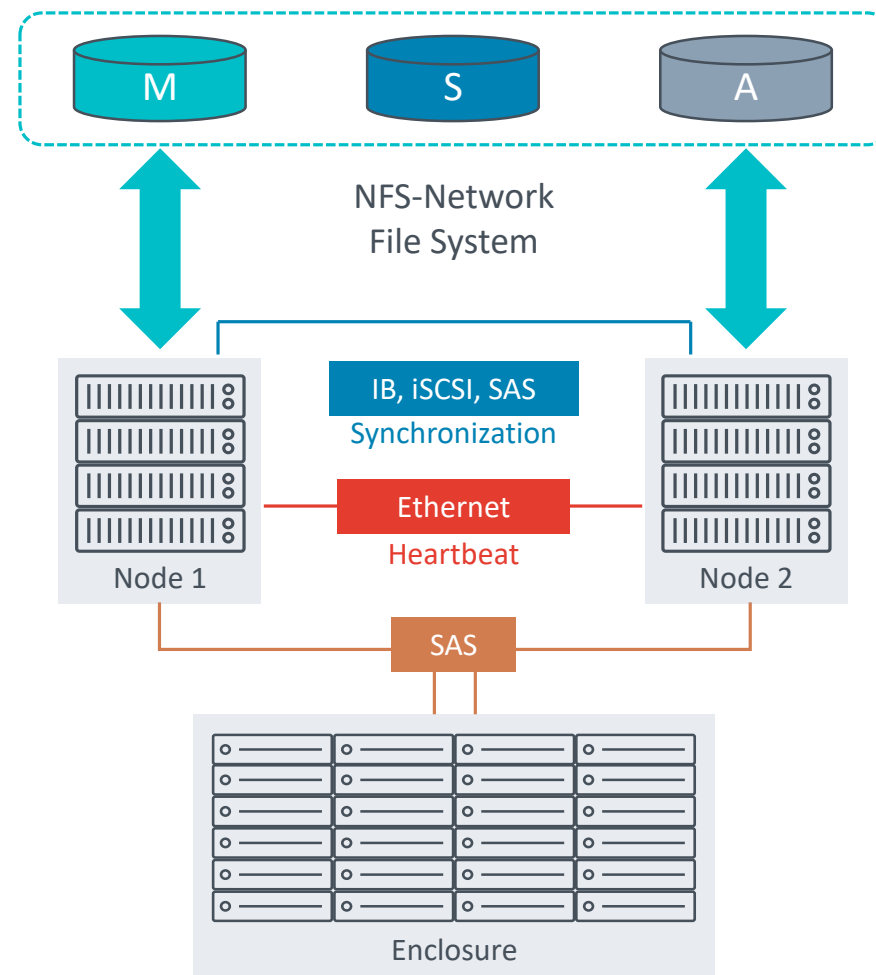
- Время восстановления для 20 ТБ БД до 2 часов



Обзор архитектуры Узла хранения в МБД.П



- Скорость бэкапа/восстановления от **4ТБ/час до 20ТБ/час**
- Двухконтроллерный режим
- Автоматическое переключение узлов (Auto Failover)
- Сохранность БД на внешних носителях
- Полное и инкрементальное резервное копирование
- Архивирование журналов предзаписи
- Доступность архива журнала для восстановления на момент времени (point-in-time recovery)
- Возможность восстановления реплики БД при большом отставании от ведущего сервера
- Использование программного RAID: защита от ошибок записи (write hole)



Спектр использует pg_probackup и pg_basebackup



Сценарии применения:

- Быстрое восстановление реплики, полная копия
- Инкрементальное восстановление
PTRACK / PAGE / DELTA

Ограничения pg_probackup в HA-кластере:

- rate-limit
- io-limit
- S3 merge инкрементов

Идеи для обсуждения:

- Можно сделать хранение с разных узлов в СХД и S3, но как быть с WAL?
- Одновременная копия WAL в S3 и СХД



МБД.П #1

МБД.П #2

МХД.О



Postgres Pro Backup Enterprise

Резервное копирование

pg_probackup backup

Скорость работы pg_probackup backup



- pg_probackup FULL backup (МХД.О):
- Объем БД: 3963GB
- Время бэкапа: 7m:35s
- **Количество потоков: 96**
- Алгоритм сжатия=lz4
- Уровень сжатия =1
- Степень сжатия бэкапа = 6.45
- **Скорость копирования данных: ~ 31TB/h**
- Утилизация IO – 100%, утилизация CPU < 30%

- pg_probackup FULL backup (МХД.О):
- Объем БД: 3014GB
- Время бэкапа: 4m:22s
- **Количество потоков: 200**
- Алгоритм сжатия=lz4
- Уровень сжатия =1
- Степень сжатия бэкапа = 5.47
- **Скорость копирования данных: ~ 42TB/h**
- Утилизация IO – 100%, утилизация CPU < 30%

Mode	WAL Mode	TLI	Time	Data	WAL	Zalg	Zratio
FULL	STREAM	55/0	7m:35s	614GB	64MB	lz4	6.45

Mode	WAL Mode	TLI	Time	Data	WAL	Zalg	Zratio
FULL	STREAM	55/0	4m:22s	551GB	64MB	lz4	5.47

Сценарии тестирования pg_probackup

Создание резервной копии с компрессией ZSTD / LZ4 и CFS



LZ4:

- Оптимальный вариант для гибридных систем (виртуализация, контейнеризация или DBaaS)

ZSTD:

- Максимальное сжатие и скорость снятия полной резервной копии в Машине баз данных Скала^р МБД.П

CFS:

- Оптимизация на уровне архитектуры БД и приложений, максимальная производительность для репликации и резервного копирования

Скорость резервного копирования Shardman СУБД:

- Скорость снятия бэкапа самого объемного сегмента кластера
- Скорость зависит от скорости SYNCPOINT всего кластера



Скорость формирования и передачи резервных копий





Postgres Pro Backup Enterprise

Восстановление данных

pg_probackup restore

Скорость восстановления из МХД.О



- pg_probackup restore (МХД.О):
- Объем БД: 3963Gb
- Время восстановления: 7m:42s
- Количество потоков: 96
- Скорость копирования данных: ~ 29TB/h

```
[postgres@mbdl-004 log]$ while true;do du -sh /pgdata/n5001mbd001;sleep 60;done
1,2T    /pgdata/n5001mbd001
1,8T    /pgdata/n5001mbd001
2,4T    /pgdata/n5001mbd001
2,9T    /pgdata/n5001mbd001
3,4T    /pgdata/n5001mbd001
3,9T    /pgdata/n5001mbd001
```

```
INFO: Start restoring backup files. PGDATA size: 3963GB
INFO: Backup files are restored. Transferred bytes: 3963GB, time elapsed: 7m:42s
INFO: Restore incremental ratio (less is better): 100% (3963GB/3963GB)
INFO: Syncing restored files to disk
INFO: Restored backup files are synced, time elapsed: 0
```



Postgres Pro Backup Enterprise

Объединение инкрементов в последний FULL

pg_probackup merge

Скорость работы Merge

- pg_probackup merge(MXD.O):
- Объем FULL : 551GB
- **Суммарный объем 6 PTRACK: 606GB**
- Количество потоков: 96
- Алгоритм сжатия=lz4
- **Итоговый объем FULL: 559GB**
- **Время слияния: 5m:27s**
- Степень сжатия бэкапа = 5.54

Mode	WAL Mode	TLI	Time	Data	WAL	Zalg	Zratio	S
PTRACK	STREAM	55/55	4m:32s	102GB	64MB	lz4	4.45	2
PTRACK	STREAM	55/55	15m:20s	132GB	60GB	lz4	3.62	2
PTRACK	STREAM	55/55	9m:25s	14GB	44GB	lz4	2.94	2
PTRACK	STREAM	55/55	6m:15s	101GB	7744MB	lz4	4.16	2
PTRACK	STREAM	55/55	21m:50s	179GB	92GB	lz4	4.21	2
PTRACK	STREAM	55/55	14m:0s	39GB	65GB	lz4	3.38	2
FULL	STREAM	55/0	4m:22s	551GB	64MB	lz4	5.47	2



Mode	WAL Mode	TLI	Time	Data	WAL	Zalg	Zratio	S
FULL	STREAM	55/0	5m:27s	559GB	64MB	lz4	5.54	2
FULL	STREAM	55/0	4m:38s	566GB	64MB	lz4	6.86	2

Производительность Скала[^]р МБД.П и МХД.О



Скорость репликации данных

от 10 до 23 ТБ/час

Скорость резервного
копирования

до 30 ТБ/час

Скорость восстановления
данных из резервной копии

до 20ТБ/час

Скорость записи в МХД.О

от 10 до 40 ТБ/час

Скорость восстановления из
МХД.О

от 10 до 29 ТБ/час

КАК ТЕПЕРЬ ЖИТЬ?

Пример Скала^р МБД.П с технологией Shardman



МБД.П

- 70ТБ вся СУБД
- Скорость резервной копии 6-7 часов

Shardman

- 3 сегмента
- 35ТБ — максимальный сегмент
- Скорость резервной копии кластера с валидацией 3-4 часа

Сегмент Центр

35 ТБ

Сегмент Урал

до 15 ТБ

Сегмент Сибирь

20ТБ

Сравнение Машин баз данных Скала^Λр МБД.П



2023

- Время восстановления для 20 ТБ БД до 5 часов
- Скорость резервного копирования в СРК Машины до 4ТБ\час

2024

- Время восстановления для 20ТБ до 2 часов
- Скорость резервного копирования в СРК Машины от 13 ТБ\час
- Скорость резервного копирования в МХД.О от 10ТБ\час
- Скорость резервного копирования 100ТБ БД до 10 часов
- Возможность применения решения в Машине с Shardman



Приглашаем к открытой
дискуссии сценариев
работы МБД.П и Postgres!

Будем благодарны
за обратную связь!





Сессия вопросов и ответов.

PgConf 2024 – МБД.П24.1