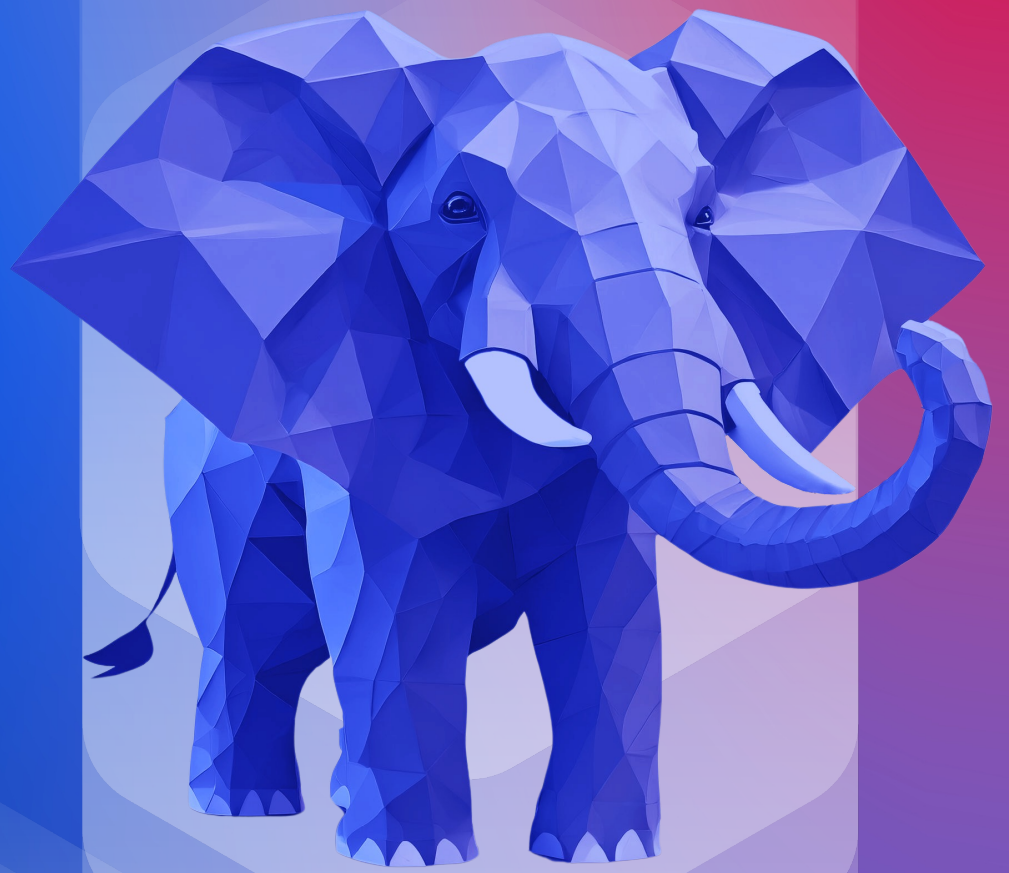
PostgresPro

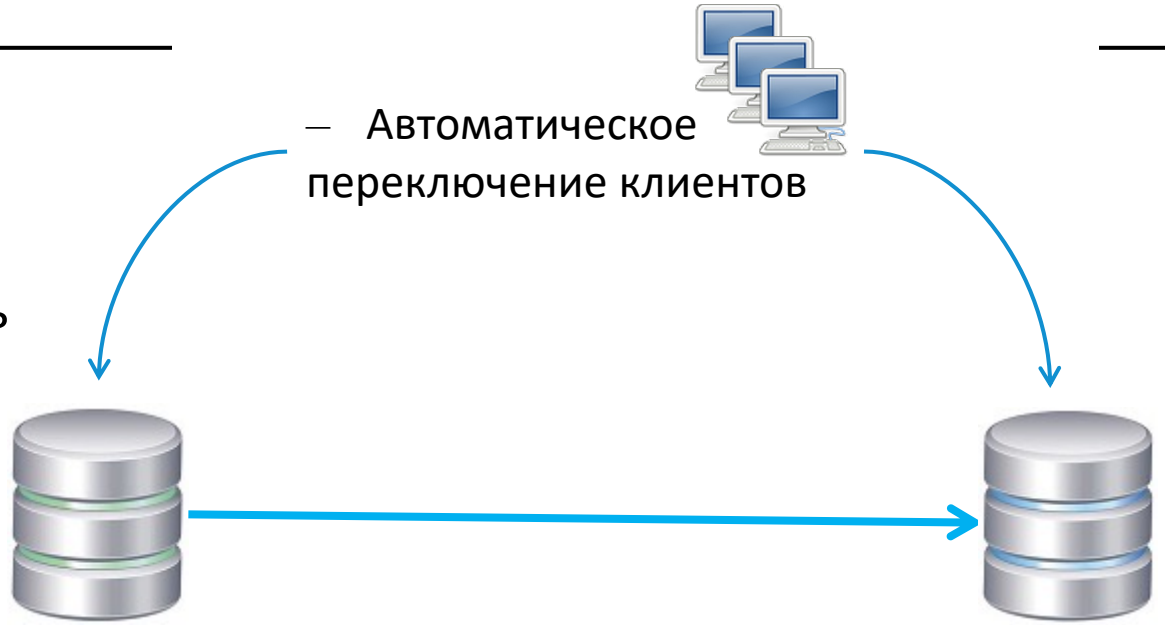
Встроенный отказоустойчивый кластер BiHA (Build-in High Availability)



Забелин Андрей

a.zabelin@postgrespro.ru

Postgres Pro : Технологии высокой доступности



– Автоматическое переключение клиентов

Мультимастер
– Масштабируемость
– Защита от сбоя сервера

Физическая репликация
– Защита от сбоя СХД,
– Защита от катастрофы
– Вынос read-only нагрузки

pg_probackup
– Бэкап на диск или в облако

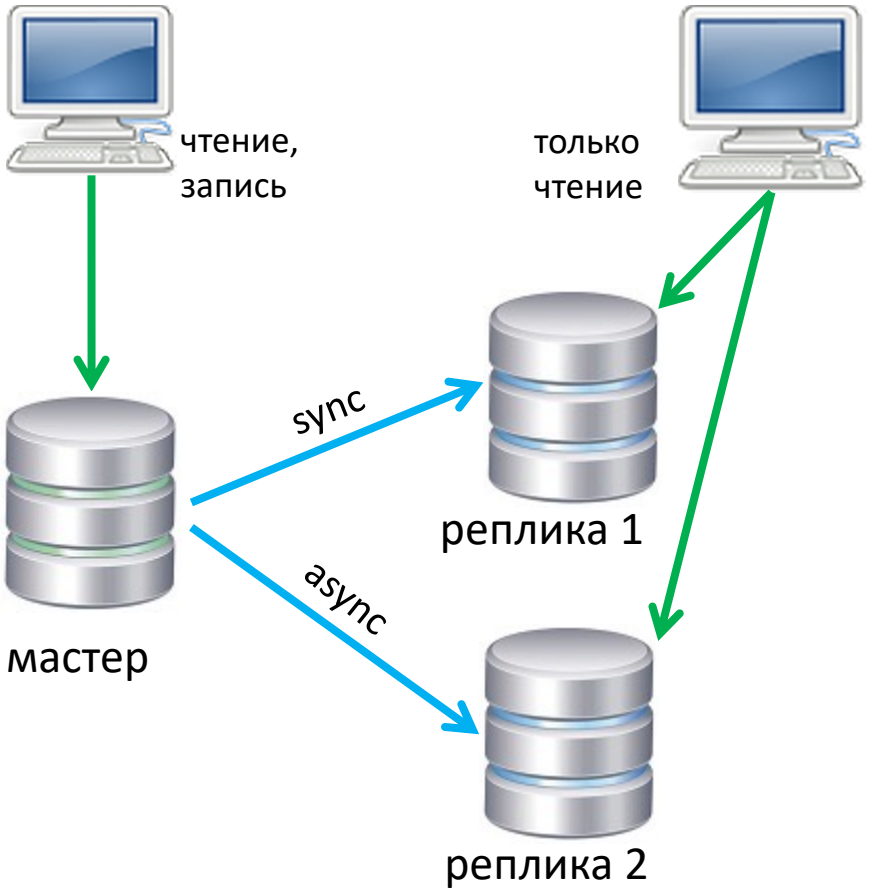
BiHA
– Автоматическое переключение мастера на реплику

shardman
– Горизонтальная масштабируемость

pg_repack
– Снижение плановых простоев

Логическая репликация
в режиме Active-active

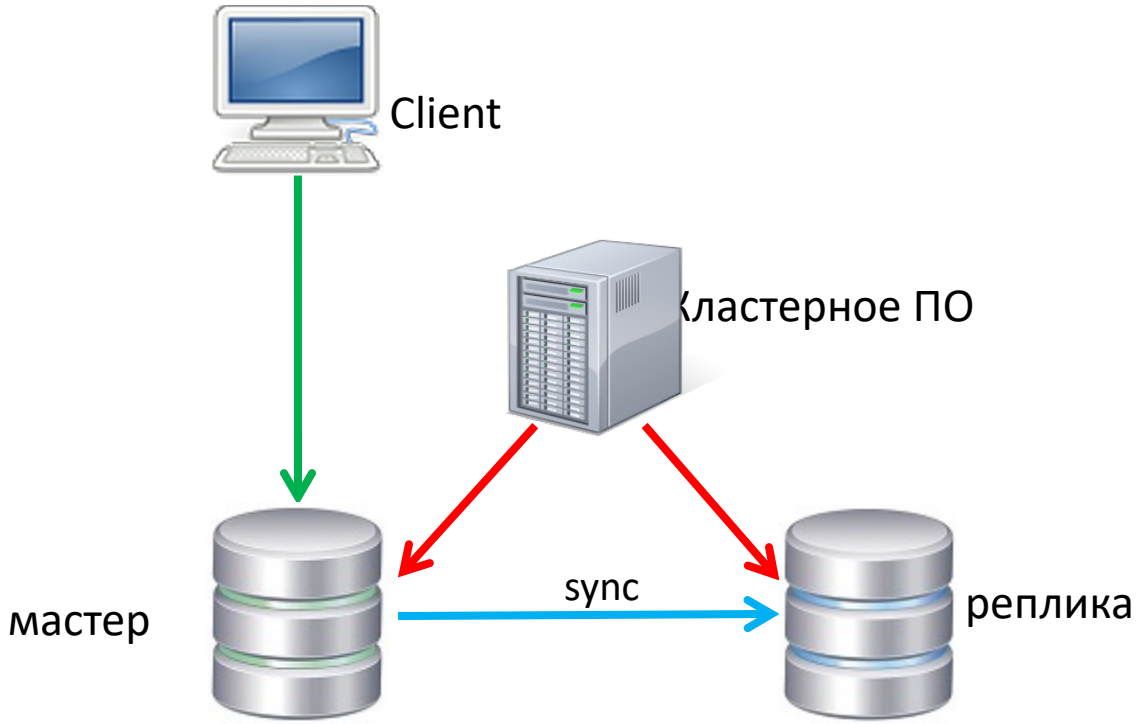
Postgres Pro : Физическая репликация



- Репликация :
  - синхронная/асинхронная,
- Реплика может быть открыта на чтение
  - часть нагрузки переносится с мастера
  - небольшие оперативные in-memory таблицы открыты на запись
  - резервная копия может выполняться на реплике
  - восстановление битых блоков БД из реплики
  - проверка битых записей журналов WAL
- Реплика может быть географически удалена

Автоматическое переключение с мастера на реплику

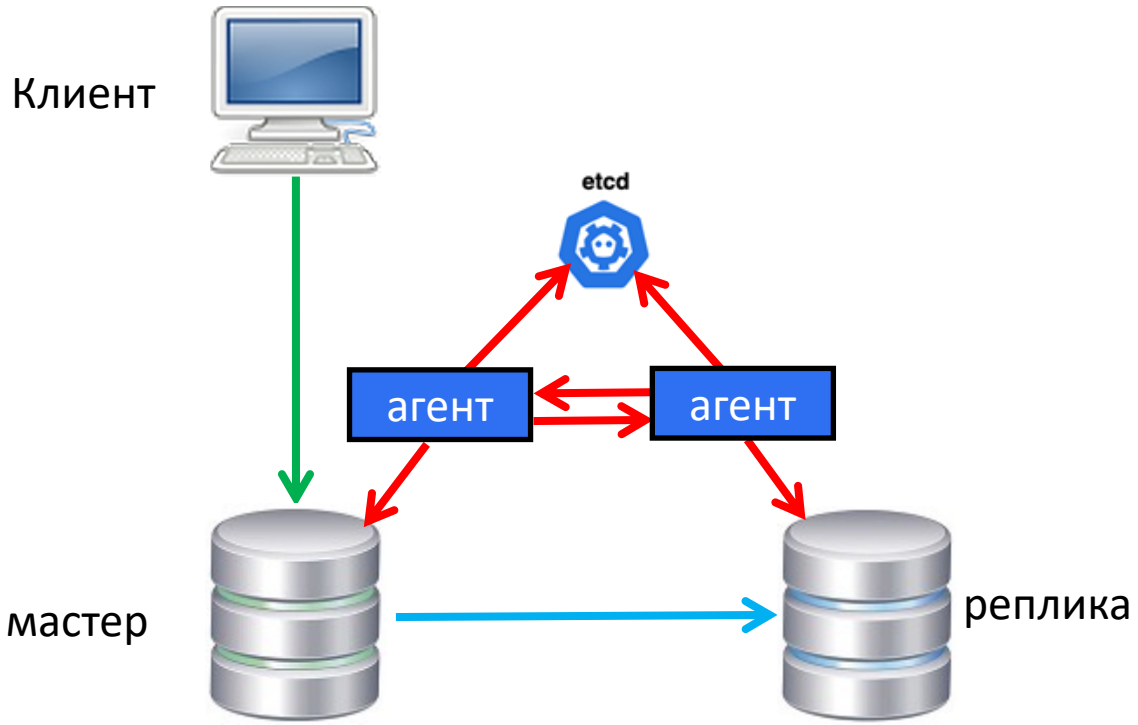
- Решение о смене ролей в отказоустойчивом кластере при сбое мастера может приниматься автоматически
- Необходимо также автоматически переключить на новый мастер и клиентов
- Основная задача кластерного ПО обнаружить сбой, сменить роль реплики на новый мастер, но при этом не допустить работу двух узлов в режиме записи



Примеры кластерного ПО : Patroni, Stolon, Corosync ….

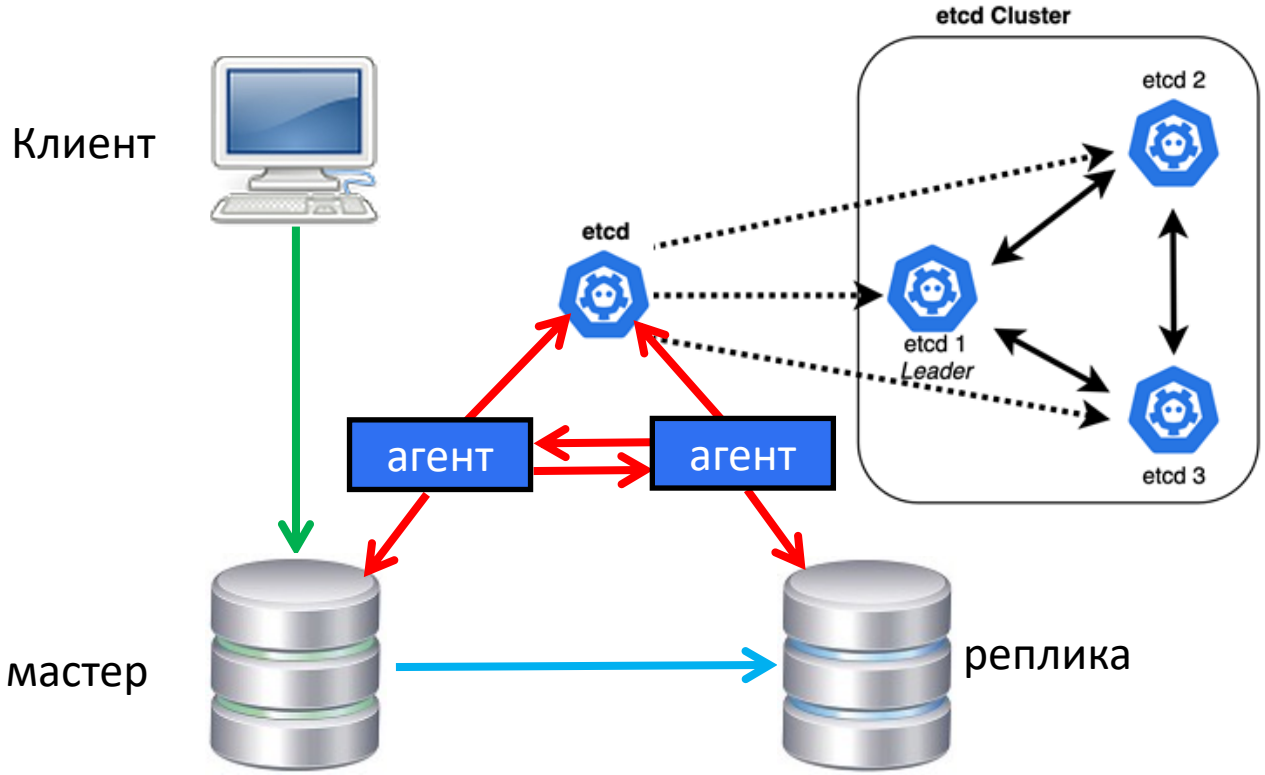
Недостатки внешнего кластерного ПО

- Внешний кластер имеет сложную архитектуру (дополнительные узлы, сетевые каналы и т.п.)



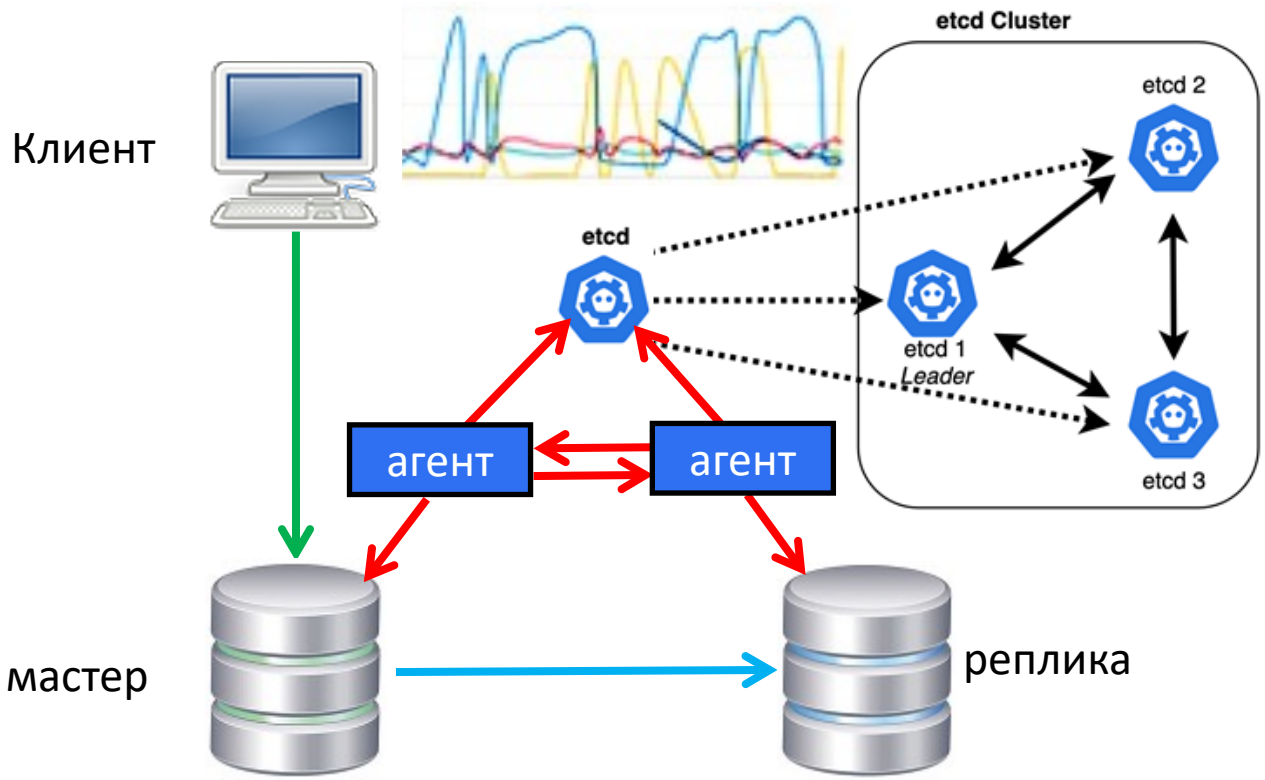
Недостатки внешнего кластерного ПО

- Внешний кластер имеет сложную архитектуру (дополнительные узлы, сетевые каналы и т.п.)
- Для элементов кластерного ПО тоже требуется отказоустойчивость



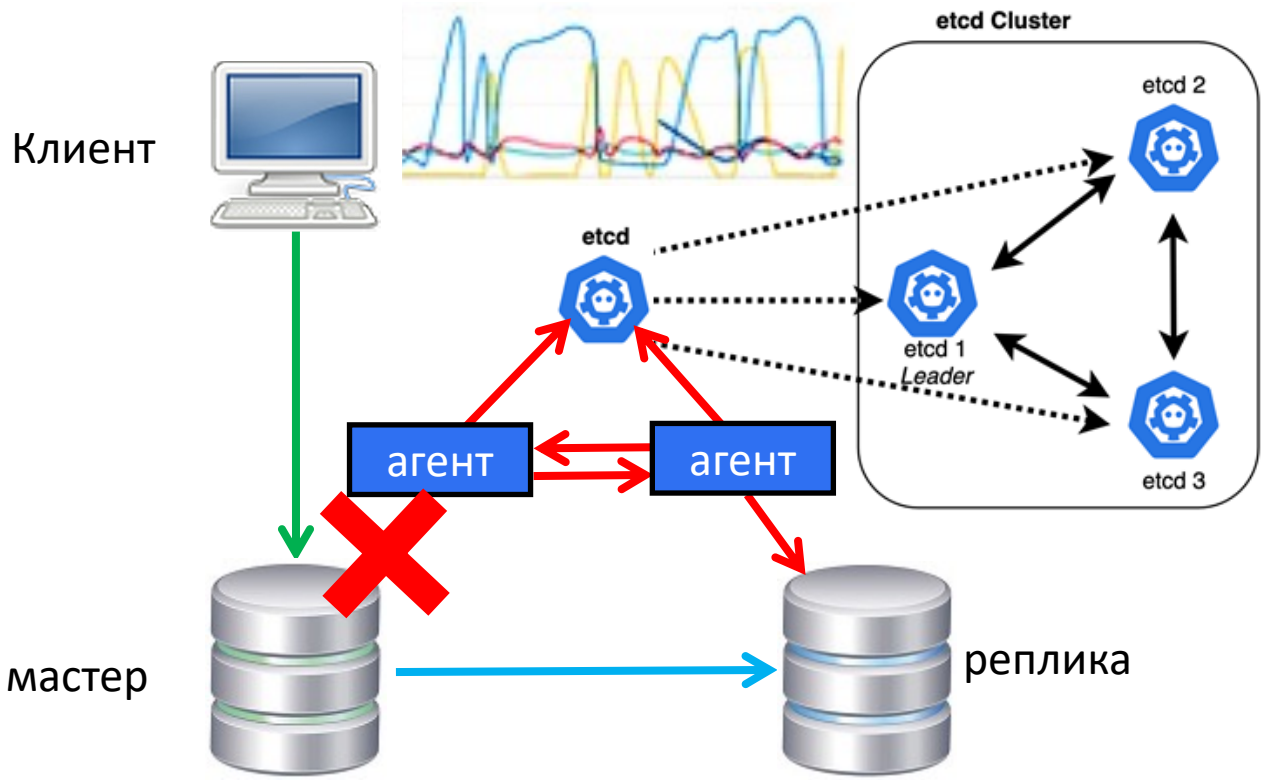
Недостатки внешнего кластерного ПО

- Внешний кластер имеет сложную архитектуру (дополнительные узлы, сетевые каналы и т.п.)
- Для элементов кластерного ПО тоже требуется отказоустойчивость
- Сложность мониторинга



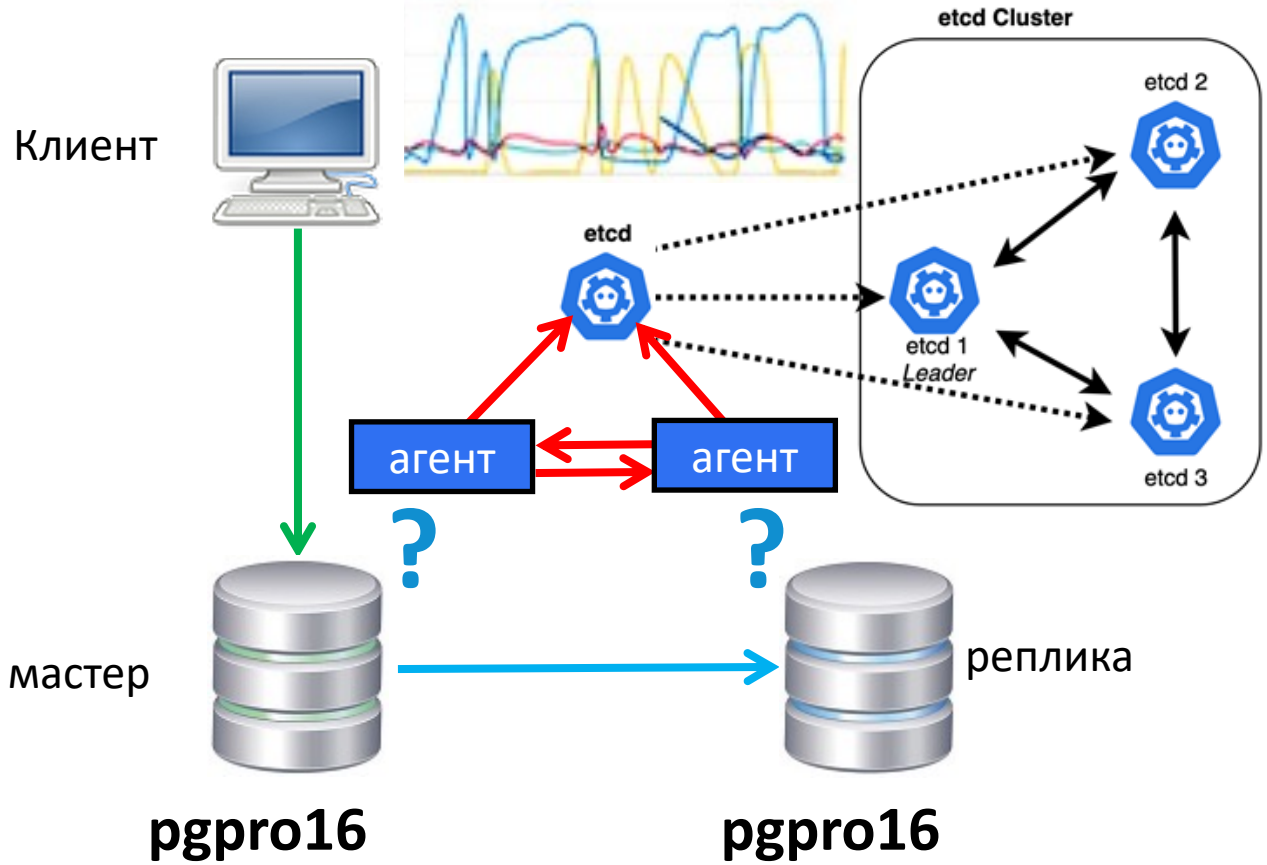
Недостатки внешнего кластерного ПО

- Внешний кластер имеет сложную архитектуру (дополнительные узлы, сетевые каналы и т.п.)
- Для элементов кластерного ПО тоже требуется отказоустойчивость
- Сложность мониторинга
- Большая нагрузка на БД может расцениваться как отказ узла



Недостатки внешнего кластерного ПО

- Внешний кластер имеет сложную архитектуру (дополнительные узлы, сетевые каналы и т.п.)
- Для элементов кластерного ПО тоже требуется отказоустойчивость
- Сложность мониторинга
- Большая нагрузка на БД может расцениваться как отказ узла
- Задержка с обновлениями версий

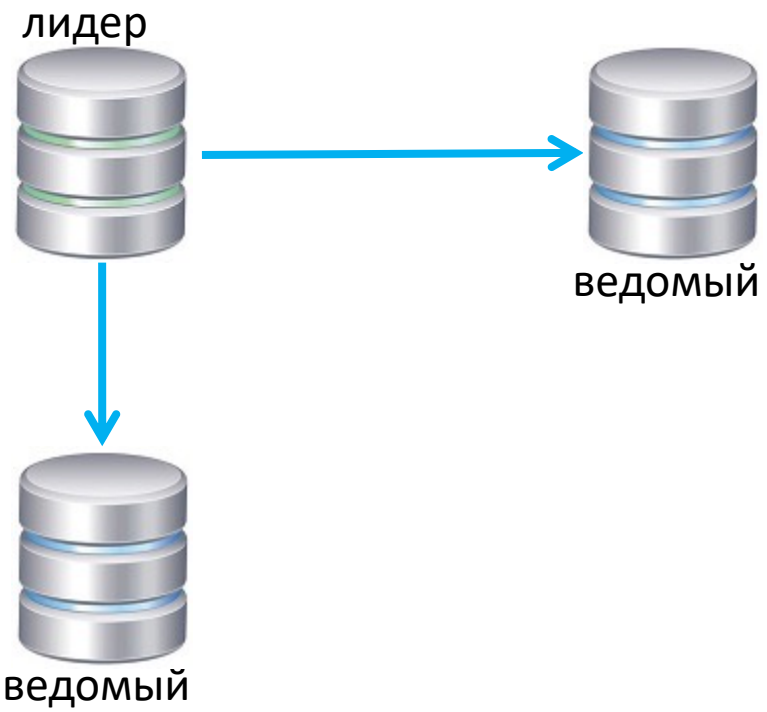


Встроенный отказоустойчивый кластер BiHA

Архитектура

Кластер состоит из нескольких узлов

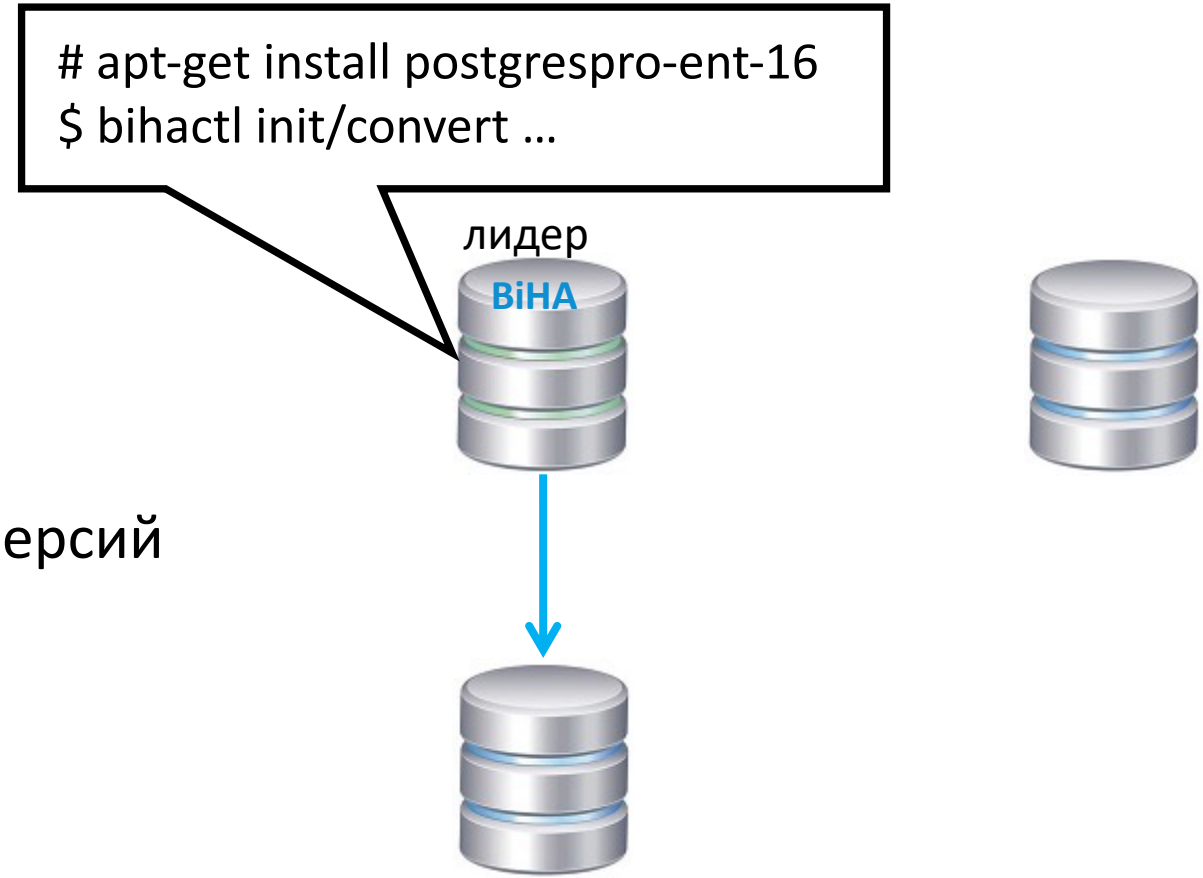
- один является лидером (leader),
- другие являются ведомыми (follower).



Встроенный отказоустойчивый кластер BiHA

Простая установка

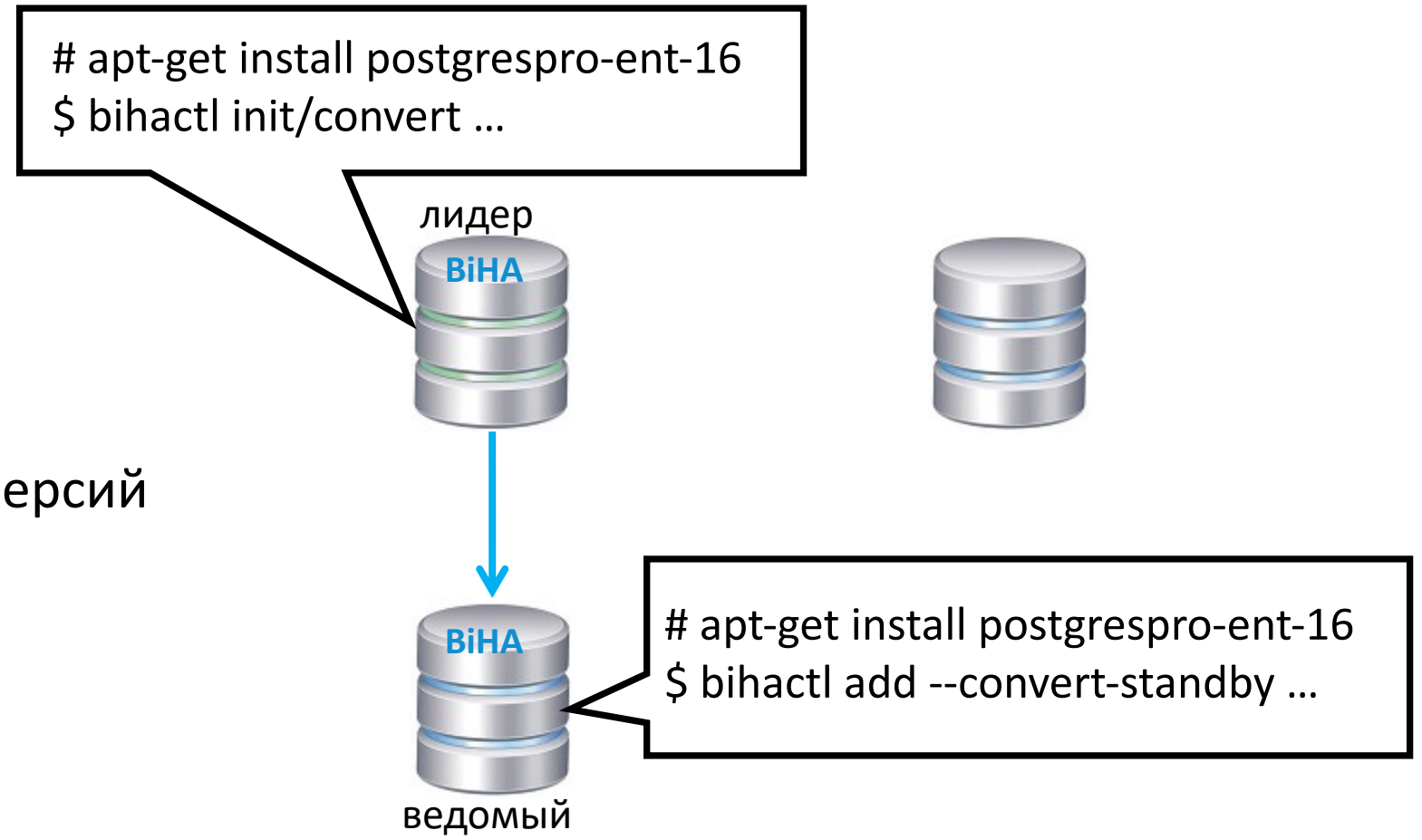
- BiHA кластер встроен в Postgres Pro.
- Простая установка и конфигурирование
- Не требуется установка дополнительного ПО
- Оперативные обновления версий



Встроенный отказоустойчивый кластер BiHA

Простая установка

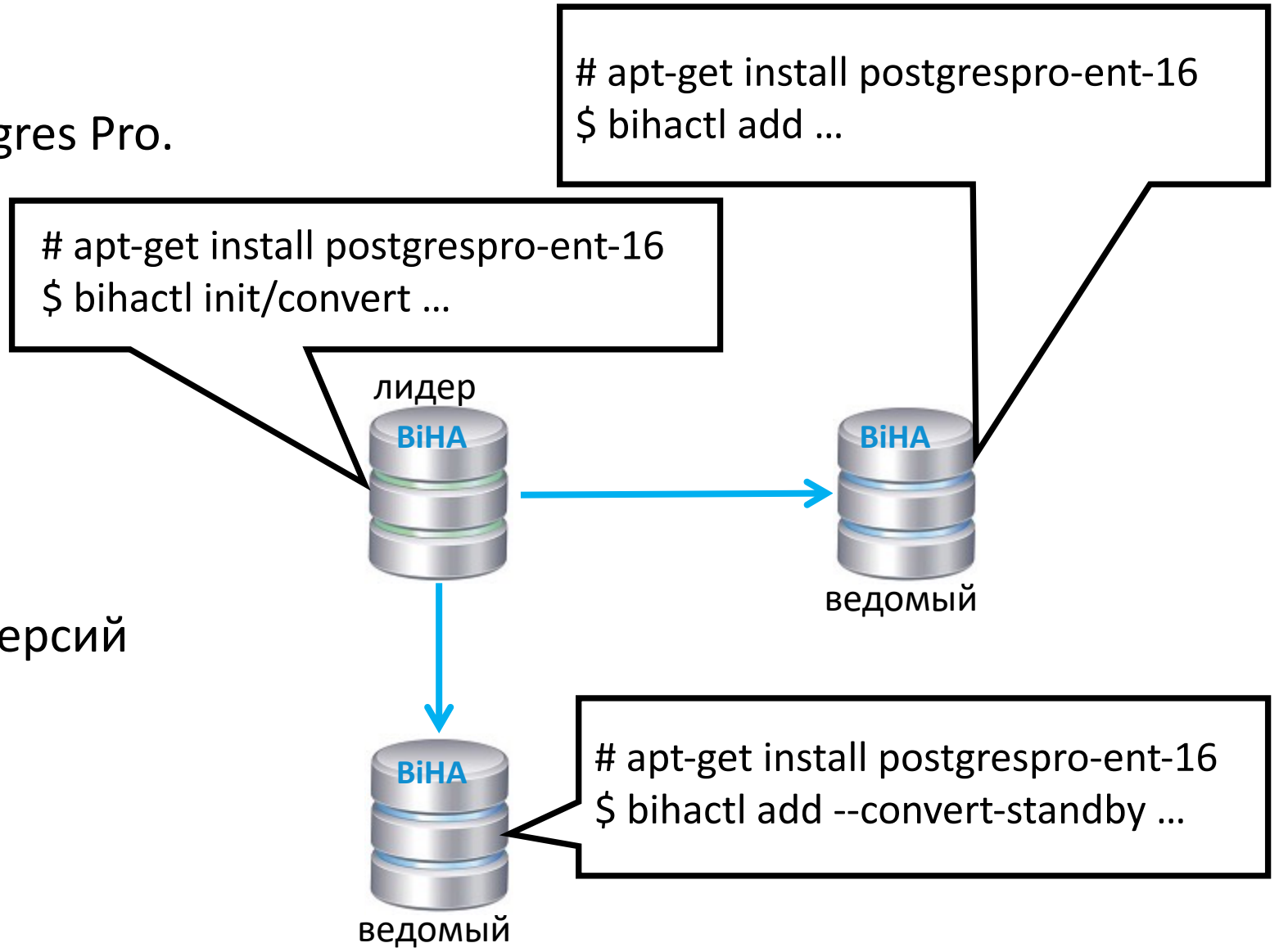
- BiHA кластер встроен в Postgres Pro.
- Простая установка и конфигурирование
- Не требуется установка дополнительного ПО
- Оперативные обновления версий



Встроенный отказоустойчивый кластер BiHA

Простая установка

- BiHA кластер встроен в Postgres Pro.
- Простая установка и конфигурирование
- Не требуется установка дополнительного ПО
- Оперативные обновления версий

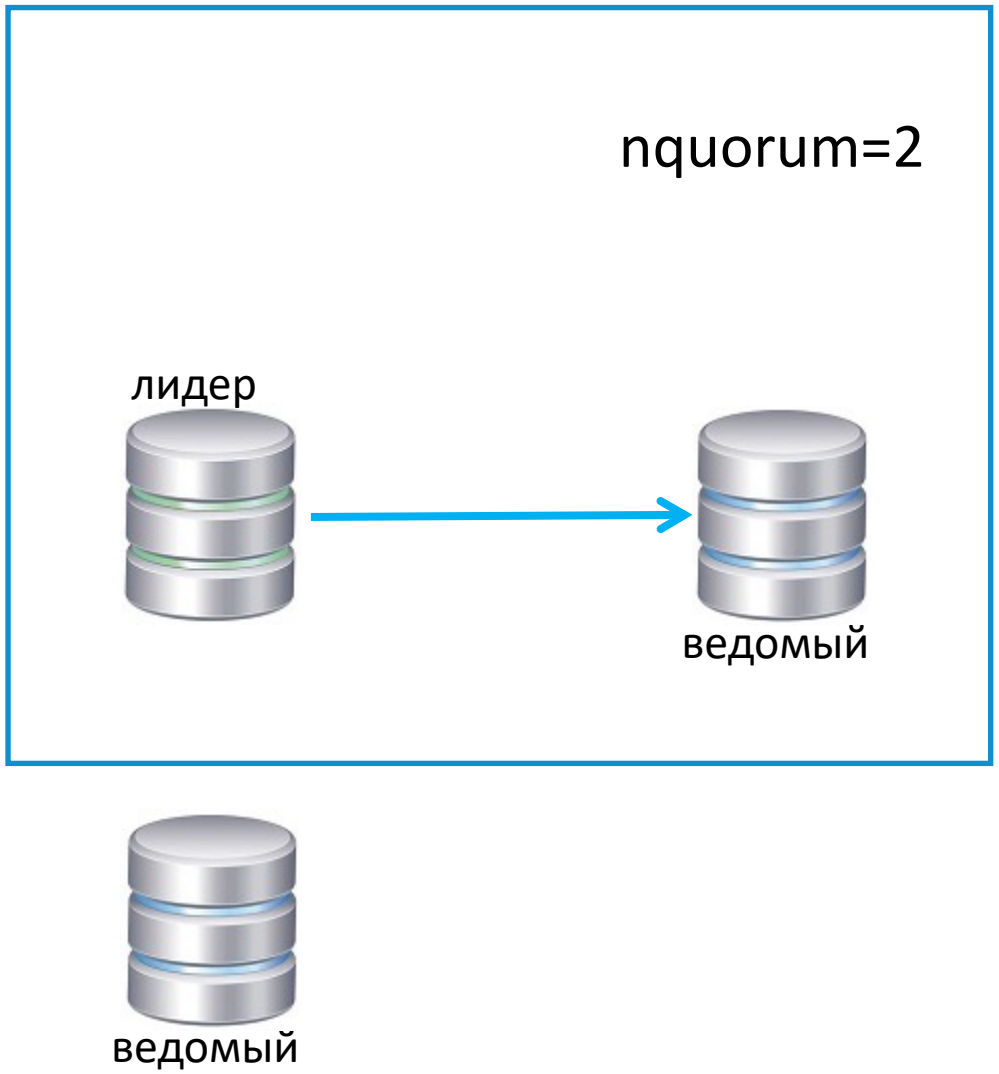


Встроенный отказоустойчивый кластер BiHA

Кластерный кворум

Кворум определяет минимальное количество узлов кластера

Лидер продолжает работать, если соблюдается кворум

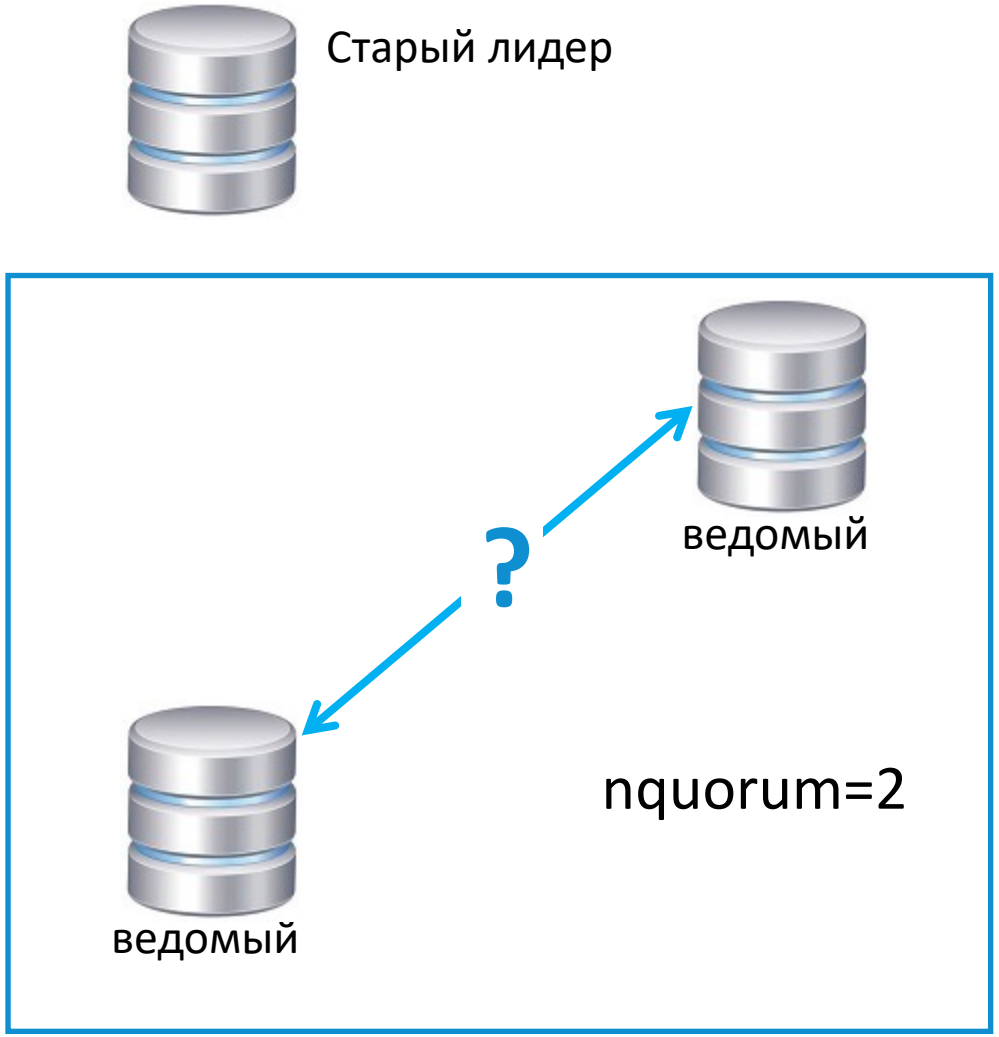


Встроенный отказоустойчивый кластер BiHA

Кластерный кворум

Лидер не может продолжать работу, если не соблюдается кворум

Ведомые организуют выборы нового лидера, если кластер содержит достаточное количество узлов

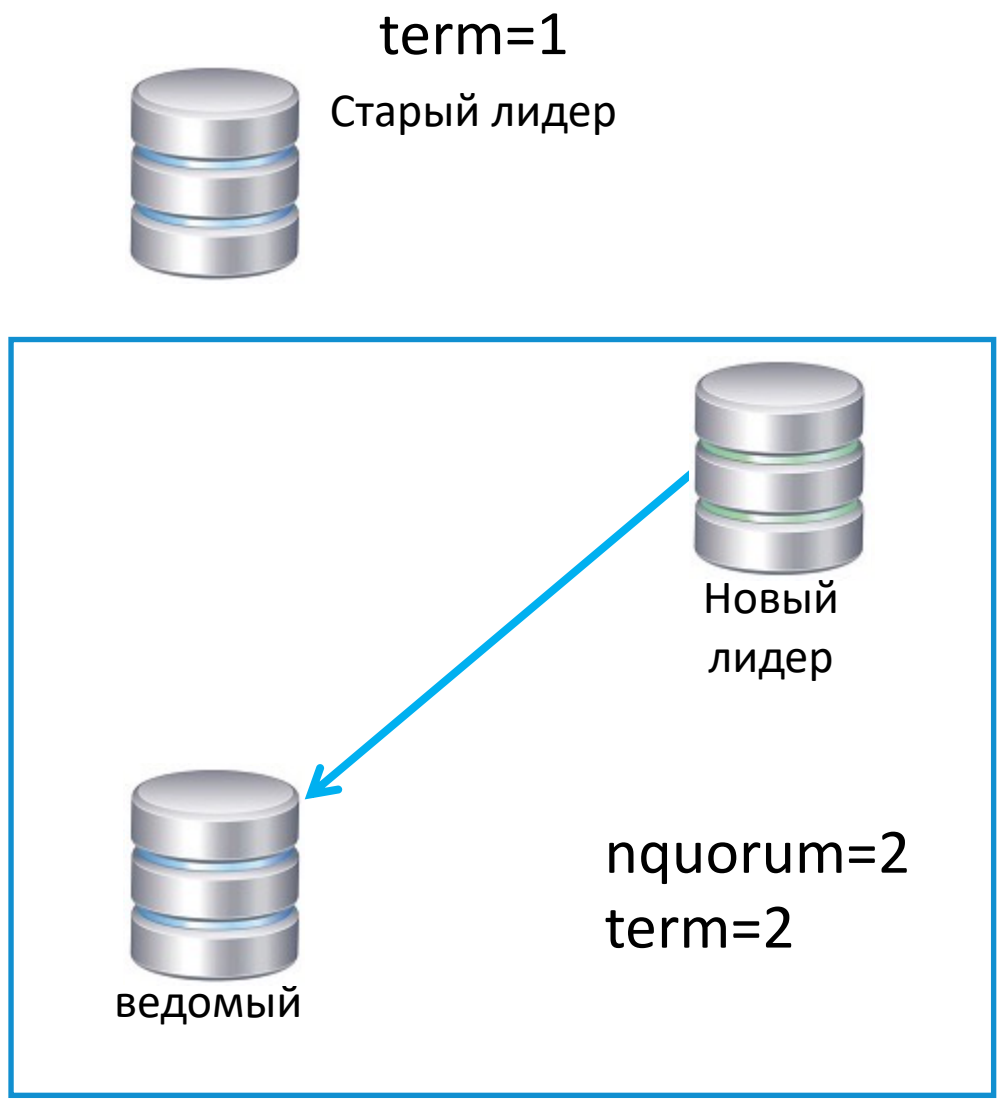


Встроенный отказоустойчивый кластер BiHA

Поколение кластера

После выбора нового лидера
в кластере меняется поколение

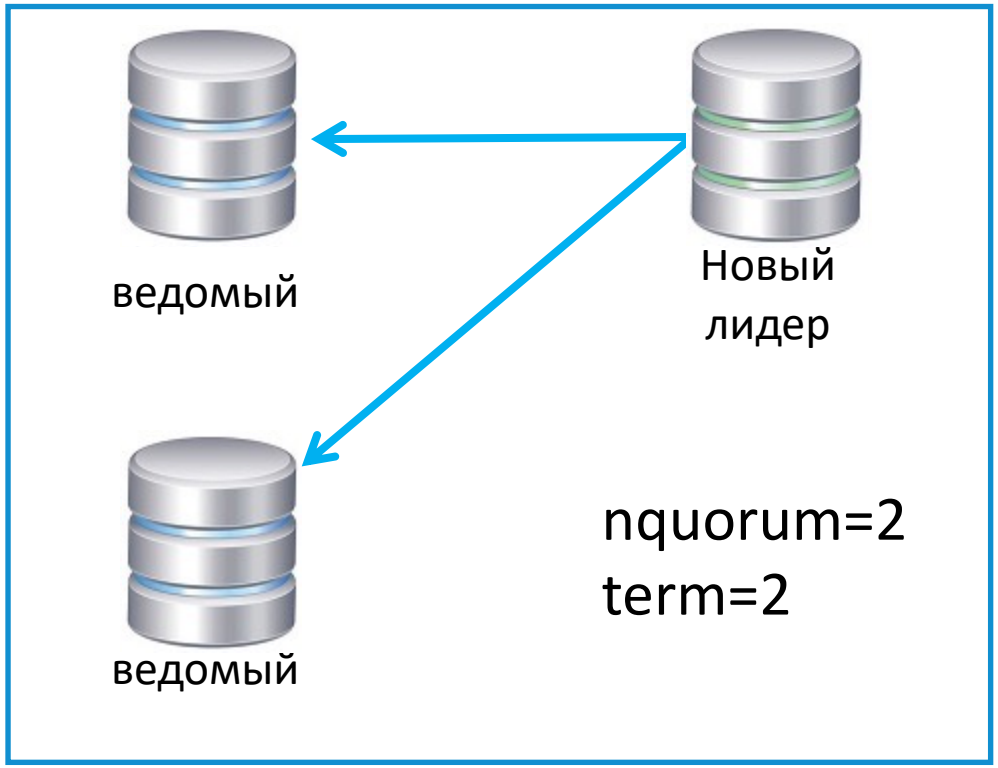
Старый лидер остаётся в старом
поколении



Встроенный отказоустойчивый кластер BiHA

Поколение кластера

При возвращении старого лидера в кластер он не может быть уже лидером и переходит в режим ведомого



Встроенный отказоустойчивый кластер BiHA

Управляющий канал

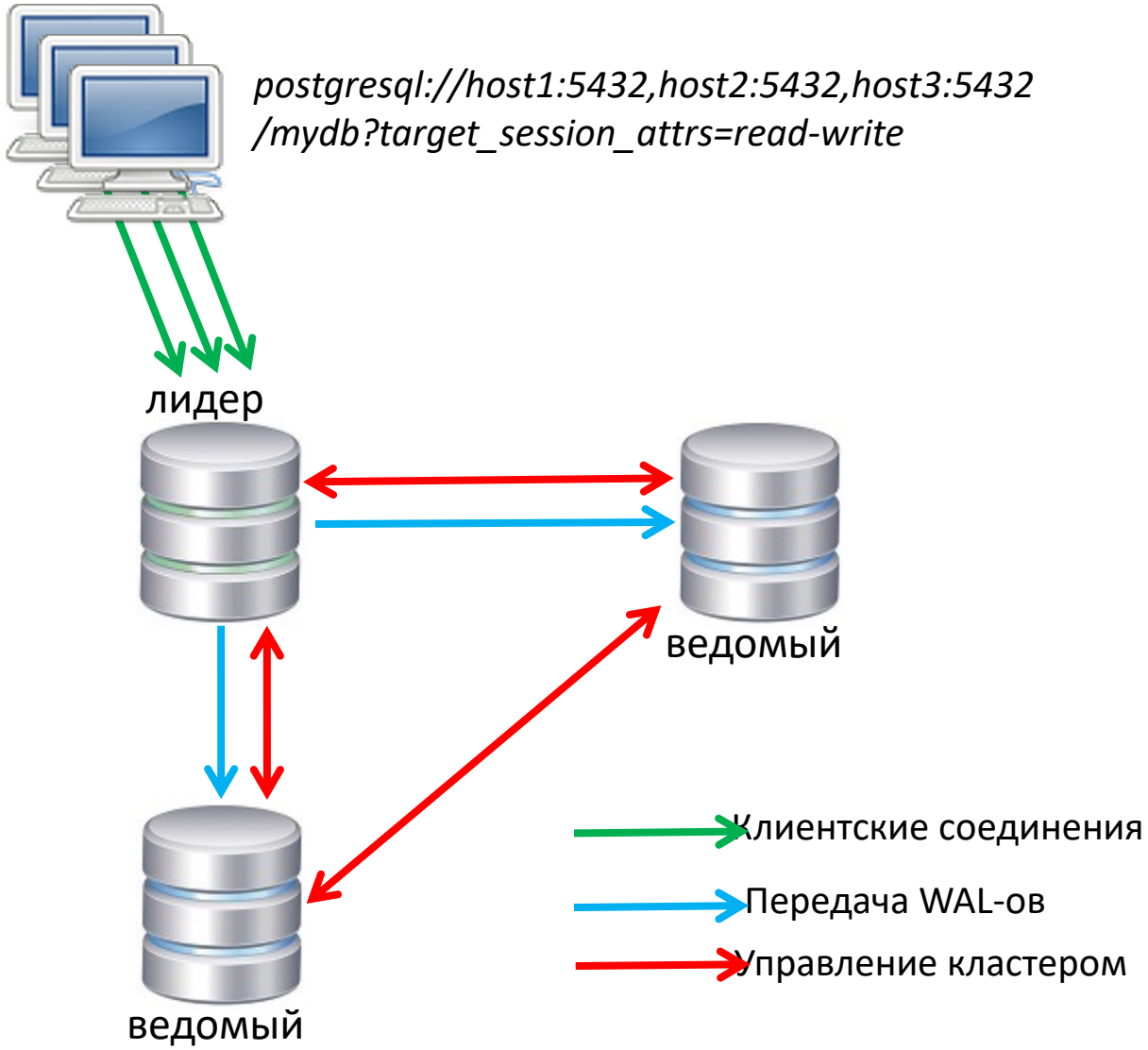
- Взаимодействие узлов друг с другом осуществляется с использованием управляющего канала
- между любыми двумя узлами устанавливается сетевое соединение по протоколу TCP.
- Непрерывный мониторинг состояния узлов кластера.

Автоматическое переключение соединения на стороне клиента на новый мастер

На клиенте (libpq, JDBC) можно перечислить все узлы кластера,

а также указать параметр *target_session_attrs=read-write* .

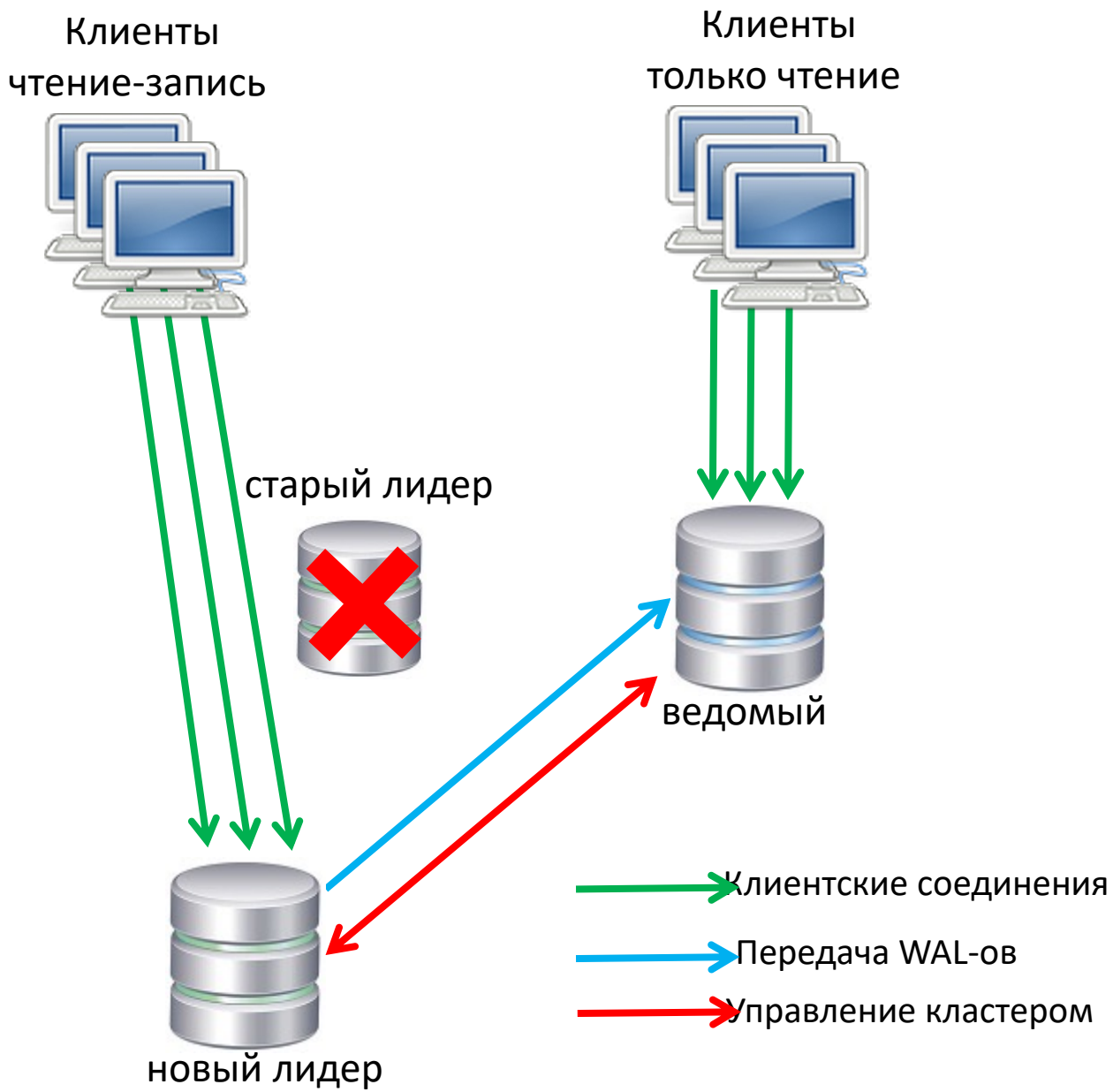
При сбое узла клиент автоматически подключится к новому лидеру



Встроенный отказоустойчивый кластер BiHA

Отказ лидера

- Автоматическая смена лидера происходит в аварийных ситуациях
- При выходе из строя лидера ведомые организуют процесс голосования для выбора нового лидера.
- Новым лидером становится ведомый узел с максимальным WAL (у него минимум потерь)



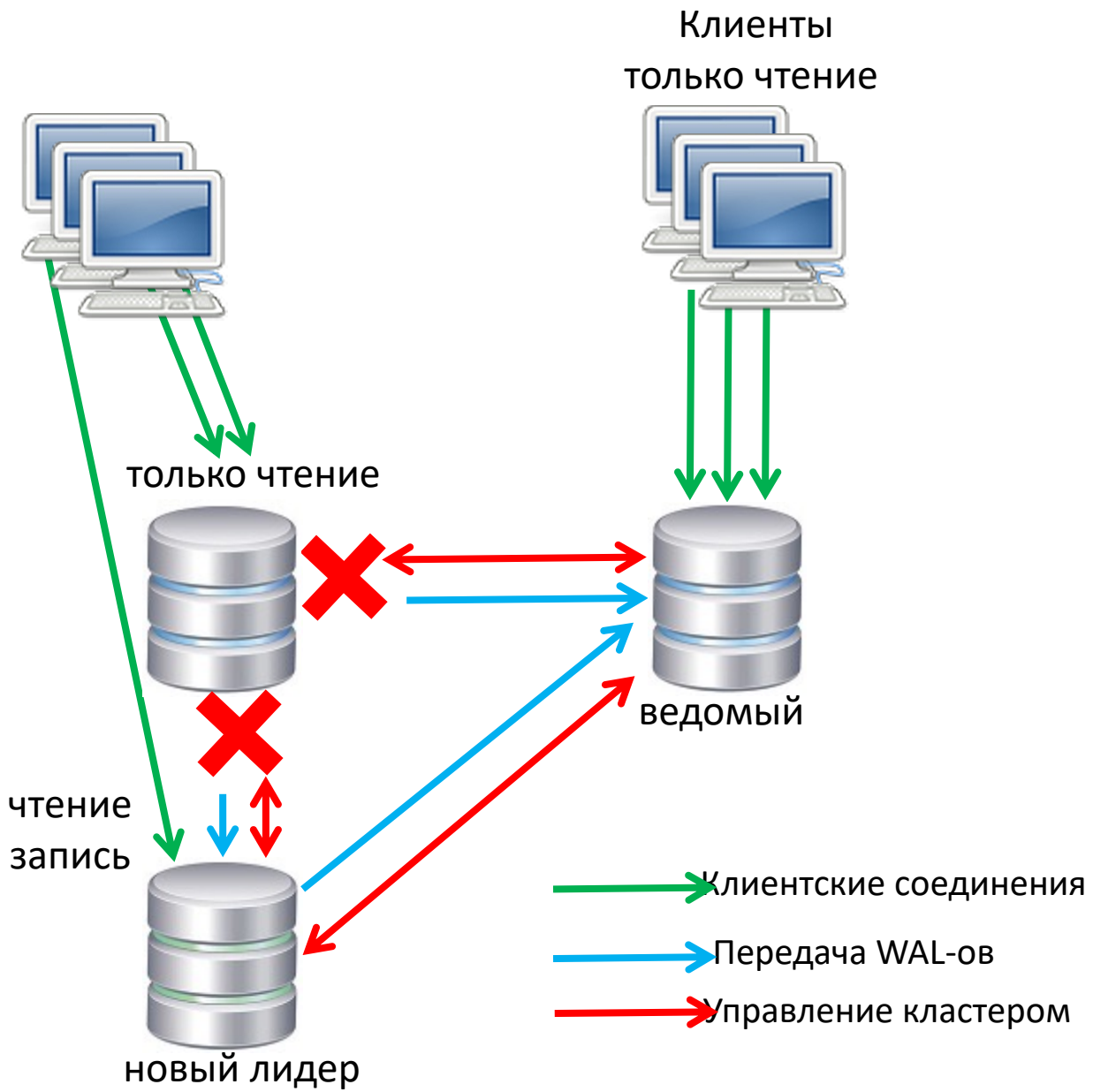
Встроенный отказоустойчивый кластер BiHA

Сетевая изоляция лидера

Когда лидер теряет связь
с необходимым количеством узлов,
лидер переводится в режим только чтение
до разрешения конфликта:

- либо когда восстановится соединение с недостающими узлами,
- либо когда администратор устранит сбой вручную.

Эта защита обеспечивает запрет на выполнение любых операций , модифицирующих WAL,
для предотвращения записи одновременно на несколько лидеров (split-brain).



Встроенный отказоустойчивый кластер BiHA

Сетевая изоляция лидера

Можно разрешить работу лидера

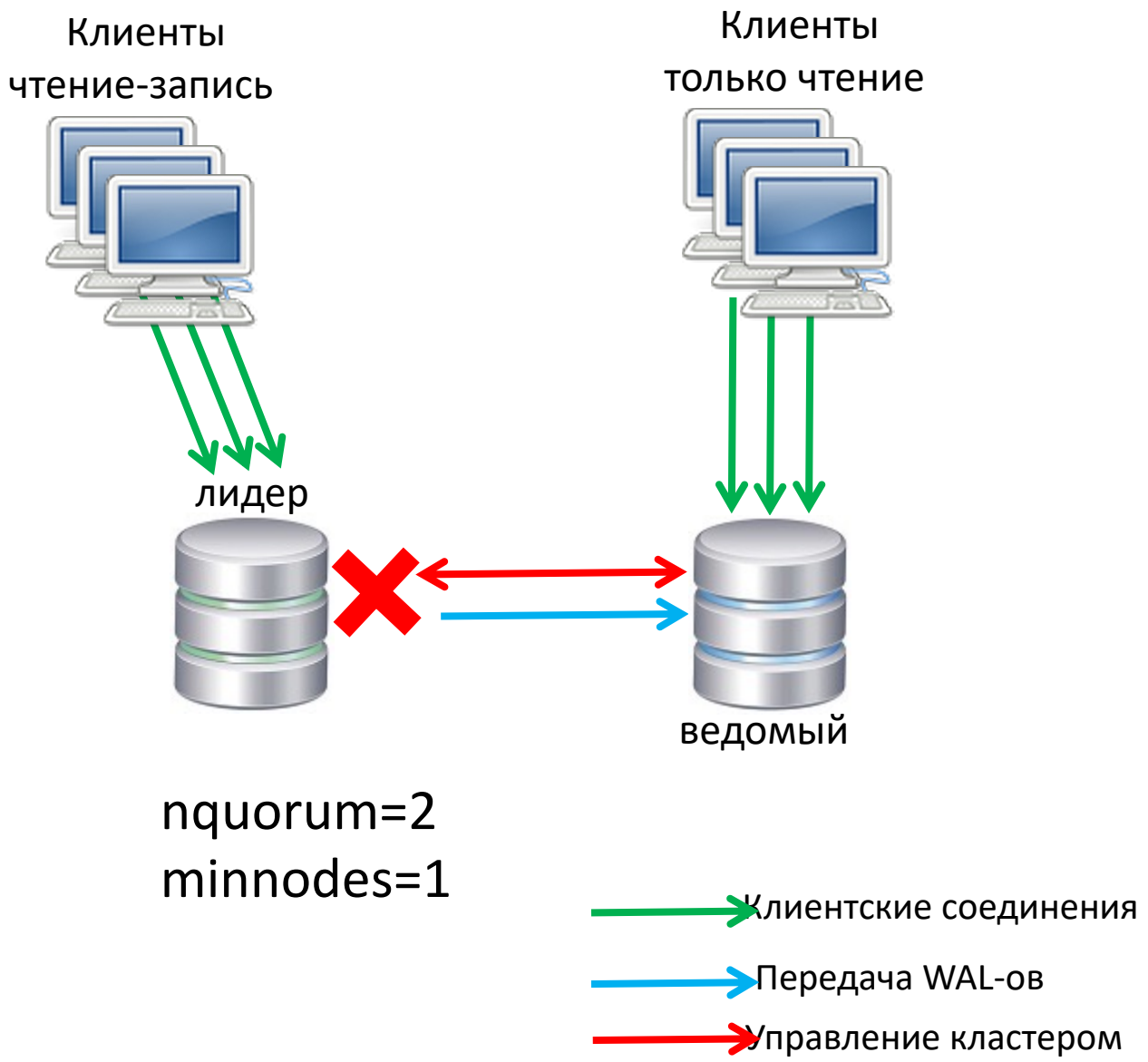
без кворума в режиме записи,

указав минимальное количество

работающих узлов (minnodes)

меньше чем минимальное количество

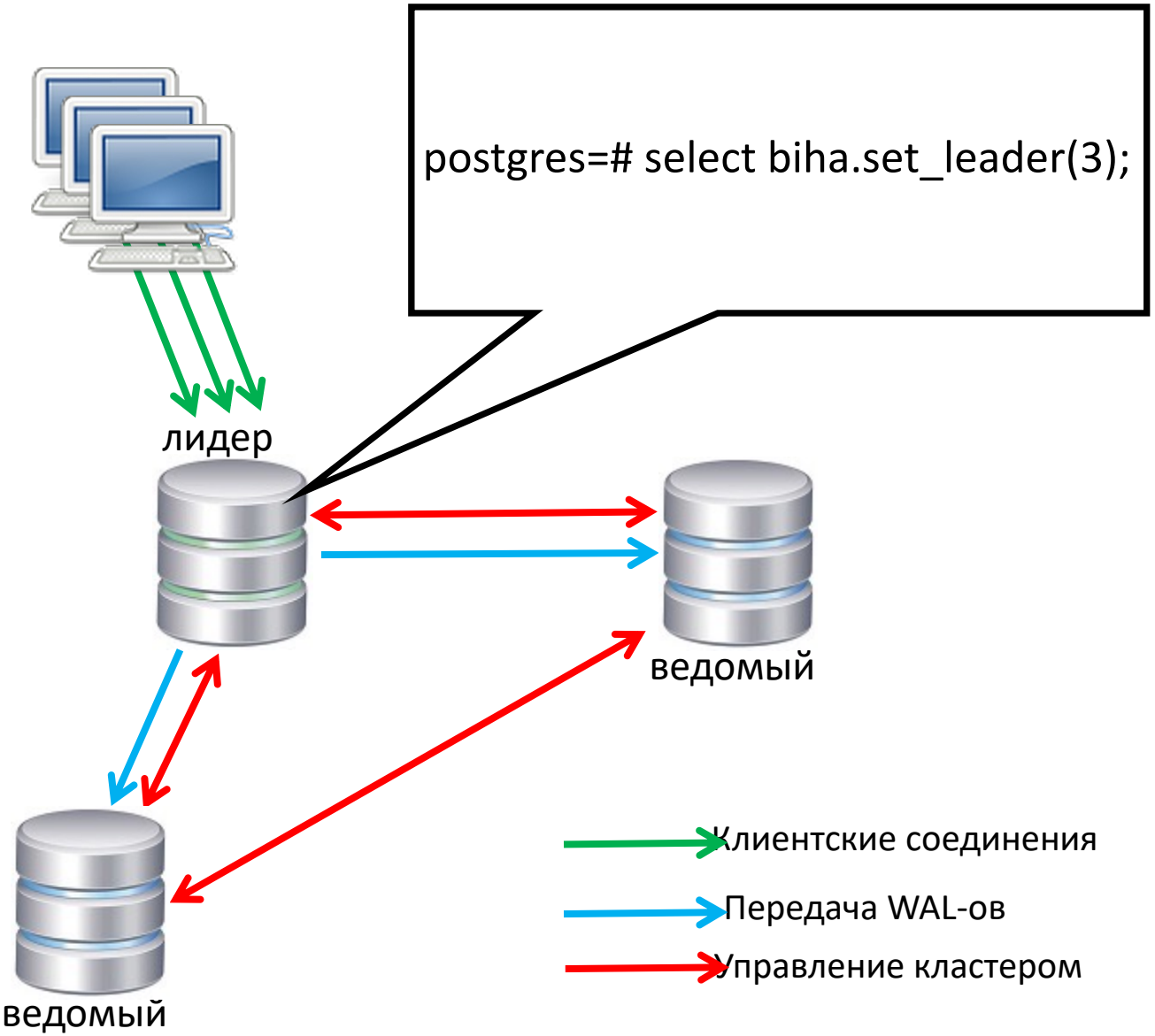
узлов для кворума (nquorum).



Встроенный отказоустойчивый кластер BiHA

Назначение лидера вручную

- для перевода лидера в режим обслуживания
- для назначения лидера на предпочтительный хост
- после возврата старого лидера

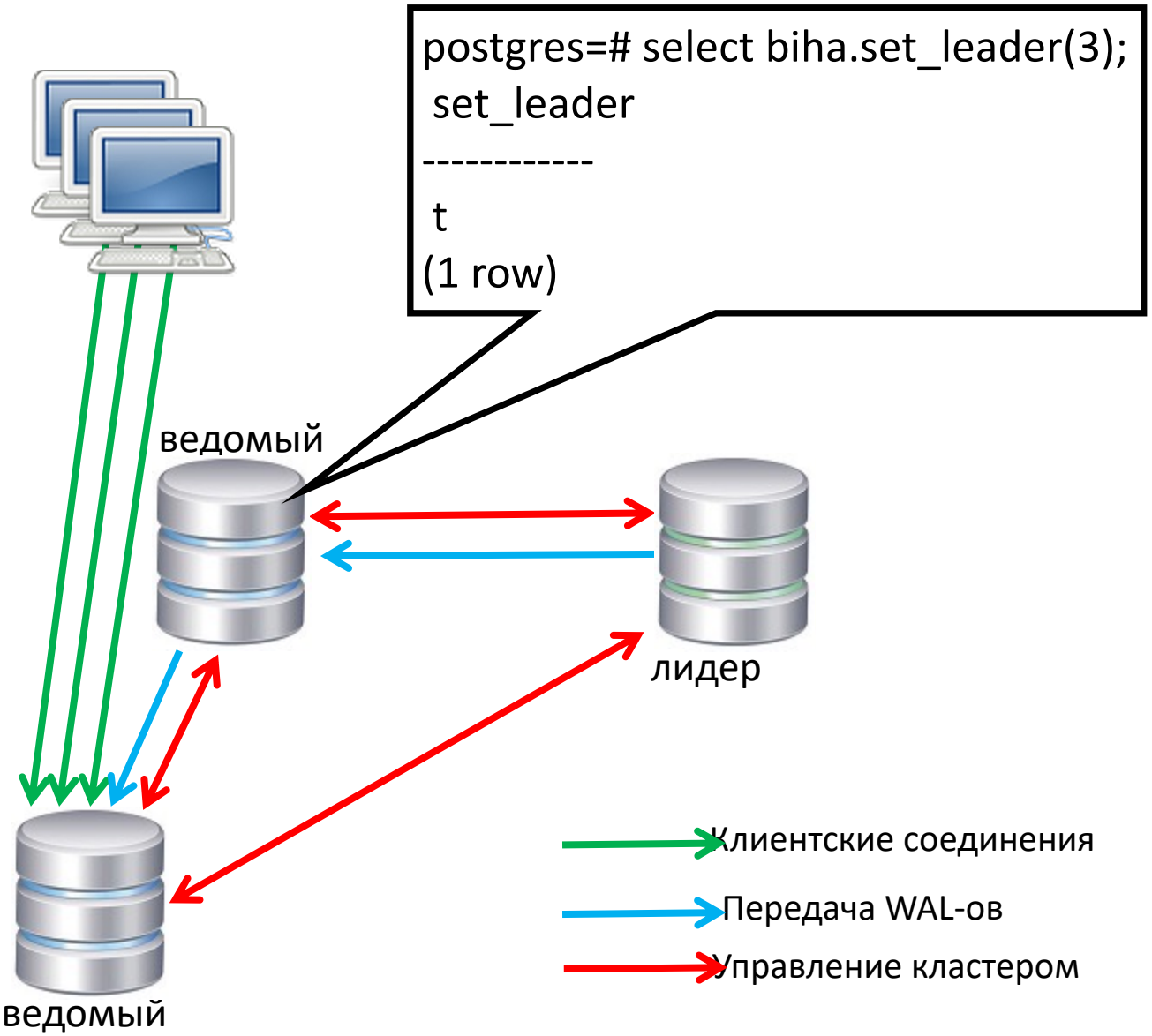


Встроенный отказоустойчивый кластер BiHA

Назначение лидера вручную

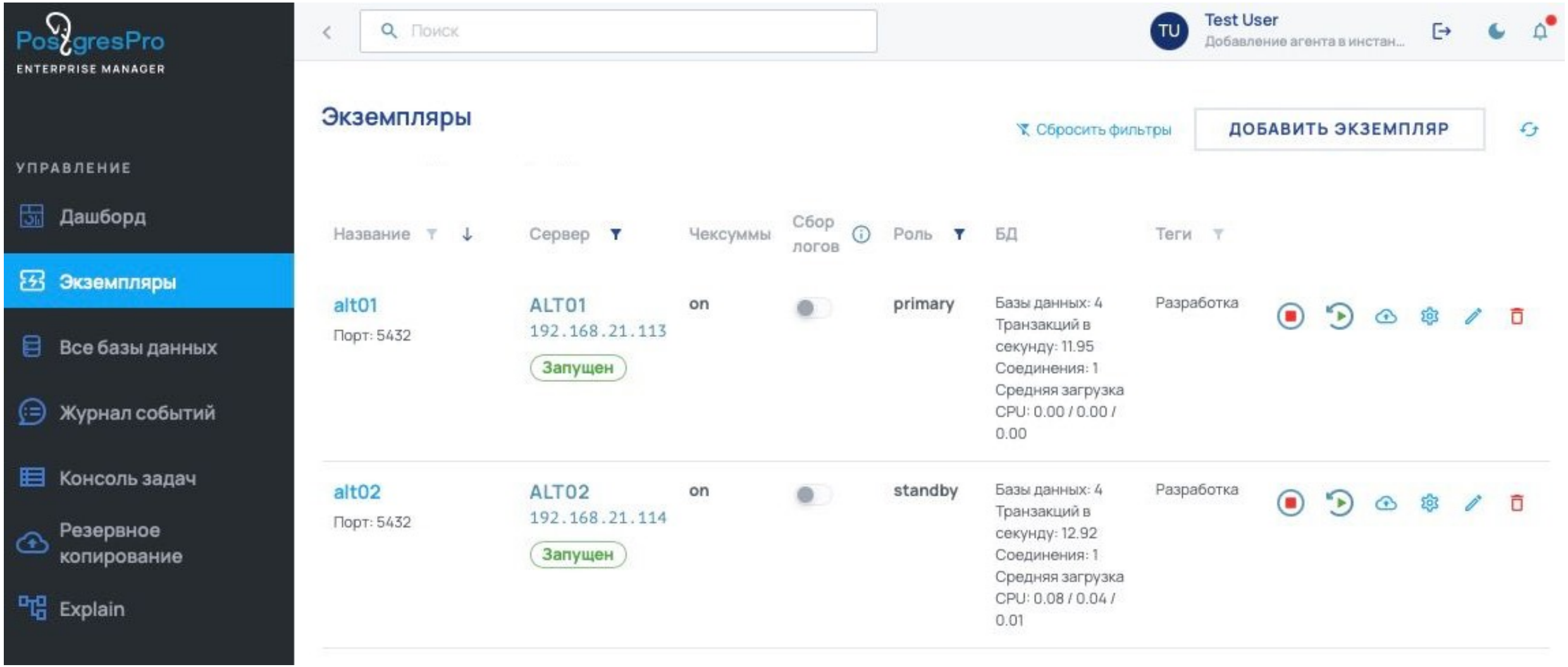
Назначение лидера через SQL-интерфейс используя функцию set_leader(id):

- в кластере блокируются все попытки выборов (устанавливается таймаут)
- текущий лидер переключается в режим ведомого
- выбранный узел становится новым лидером
- Если за выделенный таймаут процедура не завершена, выбранный узел становится ведомым, а нового лидера выбирает голосование



Postgres Pro Enterprise Manager (PPEM)

административная панель управления



PostgresPro

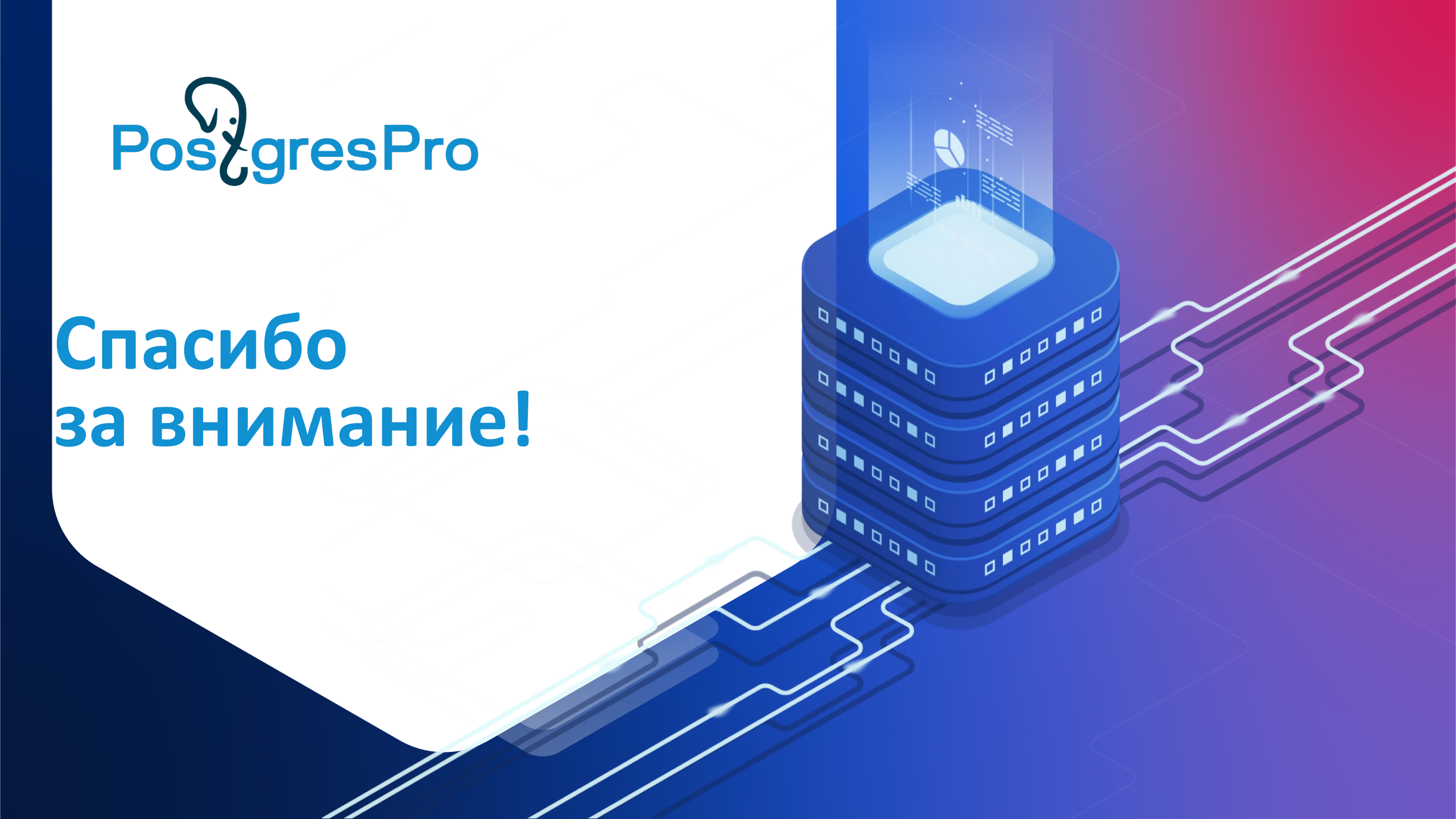
Встроенный отказоустойчивый кластер BiHA :

- Упрощает настройку кластера физической репликации
- Автоматически назначает нового мастера при сбое
- Изолирует узлы вне кластера (режим только чтение)
- Не имеет недостатков внешнего кластерного ПО
- Не требует дополнительного ПО и лицензий

Входит в дистрибутив Postgres Pro Enterprise 16.

PostgresPro

Спасибо за внимание!



PostgresPro

Q&A