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

pg_probackup
- Бэкап на диск или в облако

pg_repack
- Снижение плановых простоев

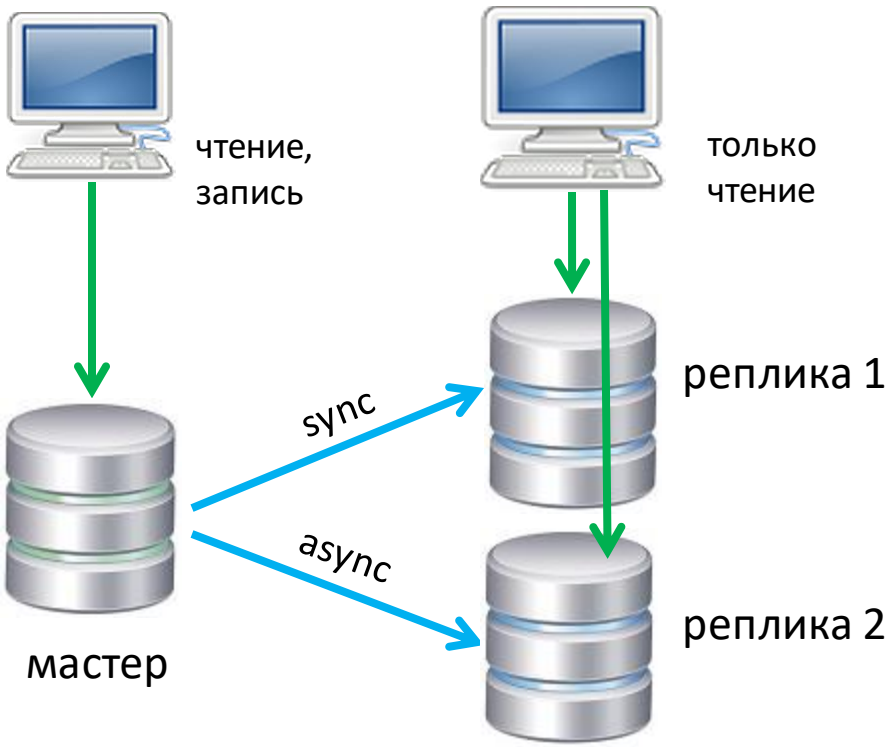
shardman
- Горизонтальная масштабируемость

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

BiHA
- Автоматическое переключение мастера на реплику

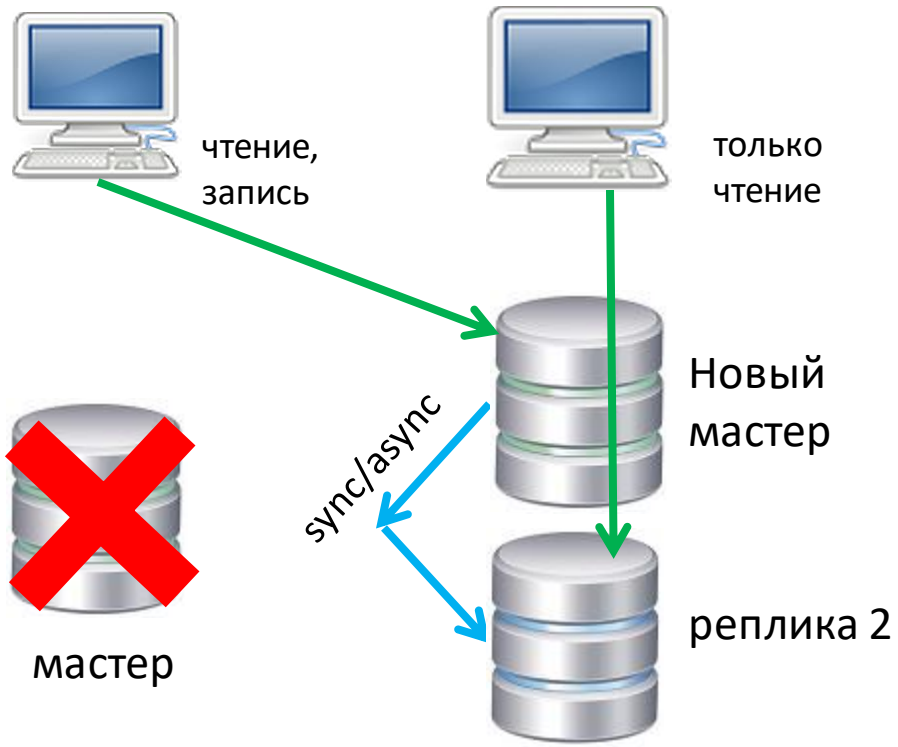
Логическая репликация
в режиме Active-active

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 из реплики
  - проверка битых записей журналов WAL
- Реплика может быть географически удалена

Сбой мастера

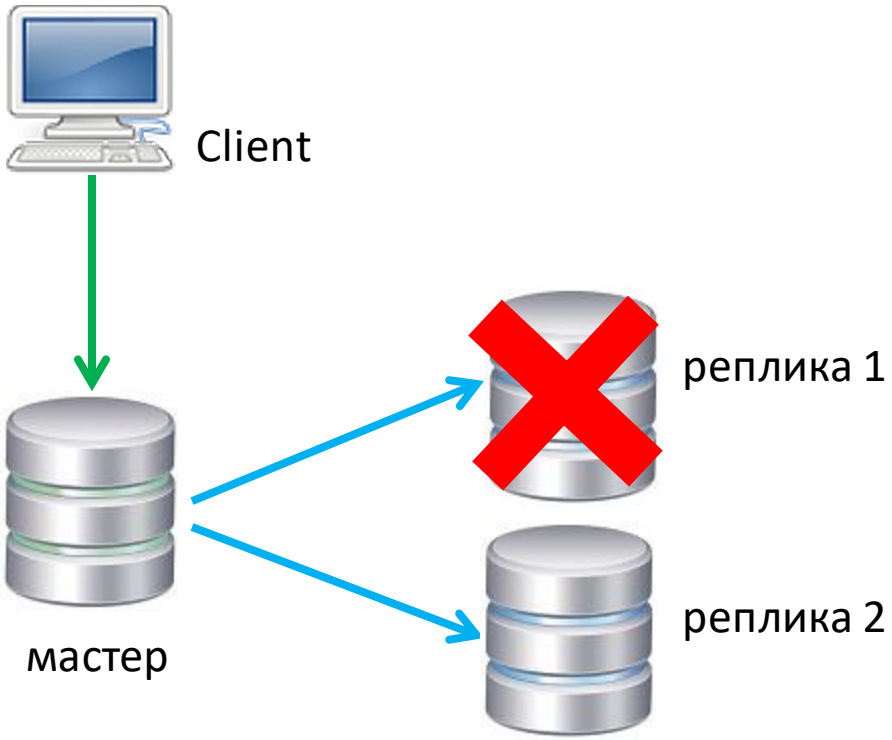


Клиенты могут переключиться на новый мастер без потери завершённых транзакций, если этот кластер баз данных был синхронной репликой.

Нельзя допускать подключения клиентов к старому мастеру после устранения сбоя.

Необходимо либо пересоздать кластер на месте старого мастера, либо старый мастер синхронизировать с новым мастером с помощью pg_rewind

Сбой реплики

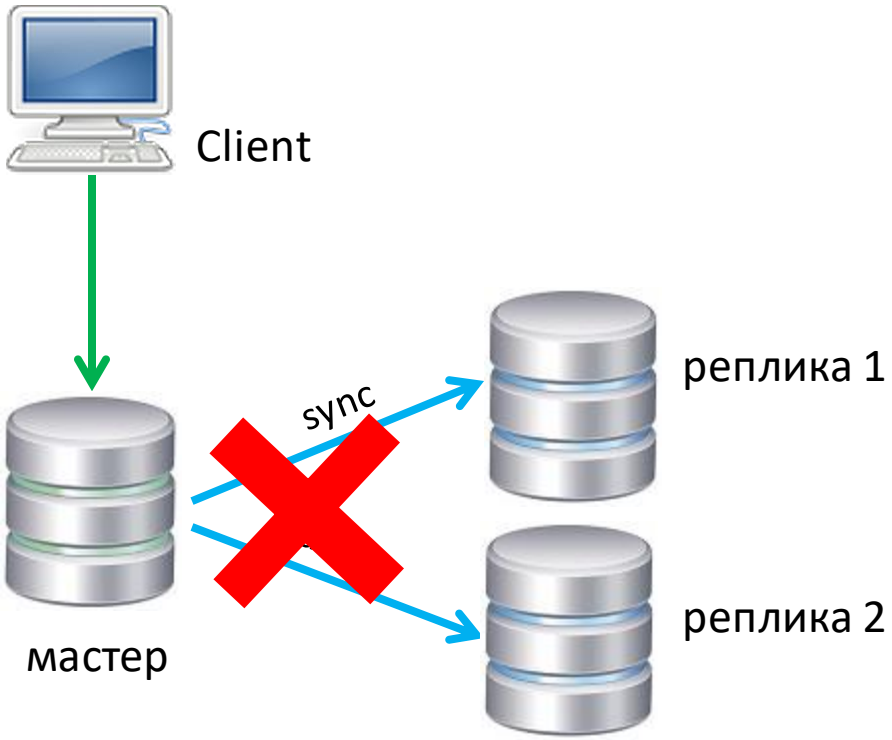


При сбое асинхронной реплики транзакции на мастере продолжаются.

При сбое единственной синхронной реплики транзакция на мастере может не завершиться никогда.

Параметр synchronous_standby_names определяет список ведомых серверов, которые могут поддерживать синхронную репликацию, а также значение MIN - минимальное количество синхронных резервных серверов, которые должны быть подключены к ведущему, чтобы он продолжал работать.

Сбой сети между мастером и репликами

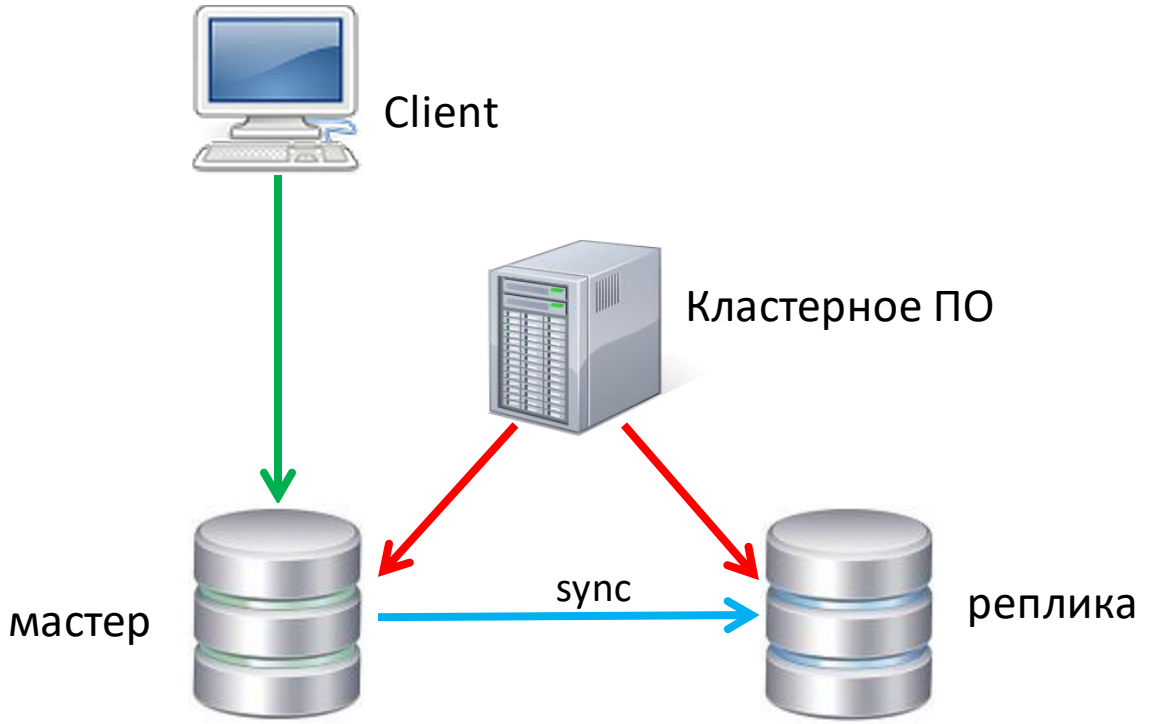


Нельзя переводить реплики в режим записи, если к мастеру подключены клиенты. Изменения, сделанные на реплике нельзя будет восстановить на мастере.

Желательно конфигурировать сеть между мастером и различными репликами по разным каналам связи, устранив единую точку отказа.

Автоматическое переключение с мастера на реплику

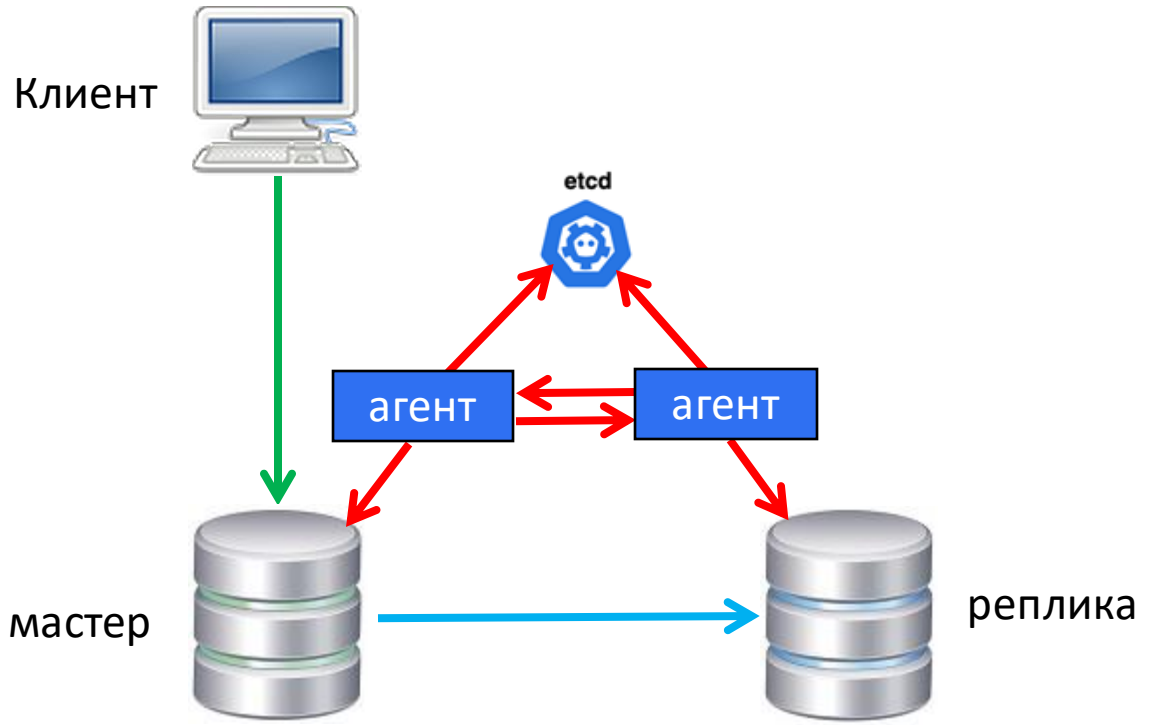
- Решение о смене ролей в отказоустойчивом кластере при сбое мастера может приниматься автоматически
- Необходимо также автоматически переключить на новый мастер и клиентов
- Основная задача кластерного ПО обнаружить сбой, сменить роль реплики на новый мастер, но при этом не допустить работу двух узлов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



Примеры кластерного ПО : Patroni, Stolon, Corosync ….

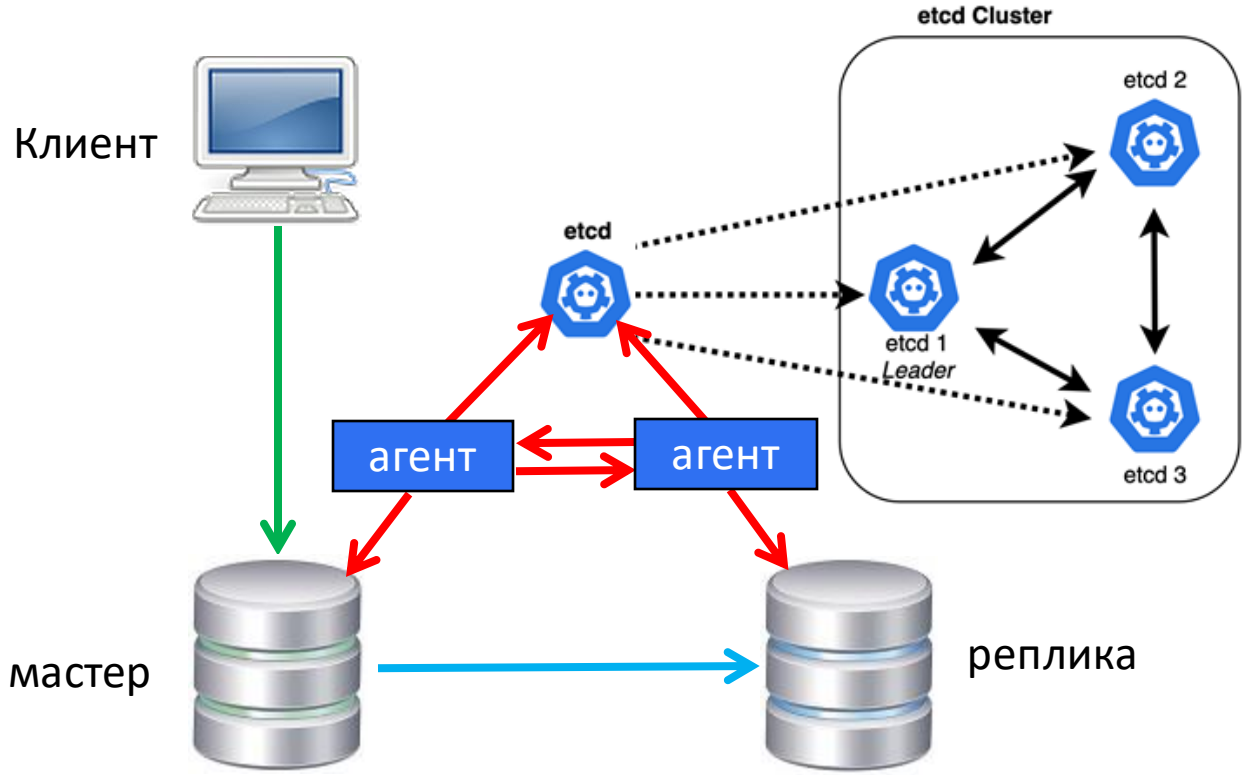
Внешнее кластерное ПО

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



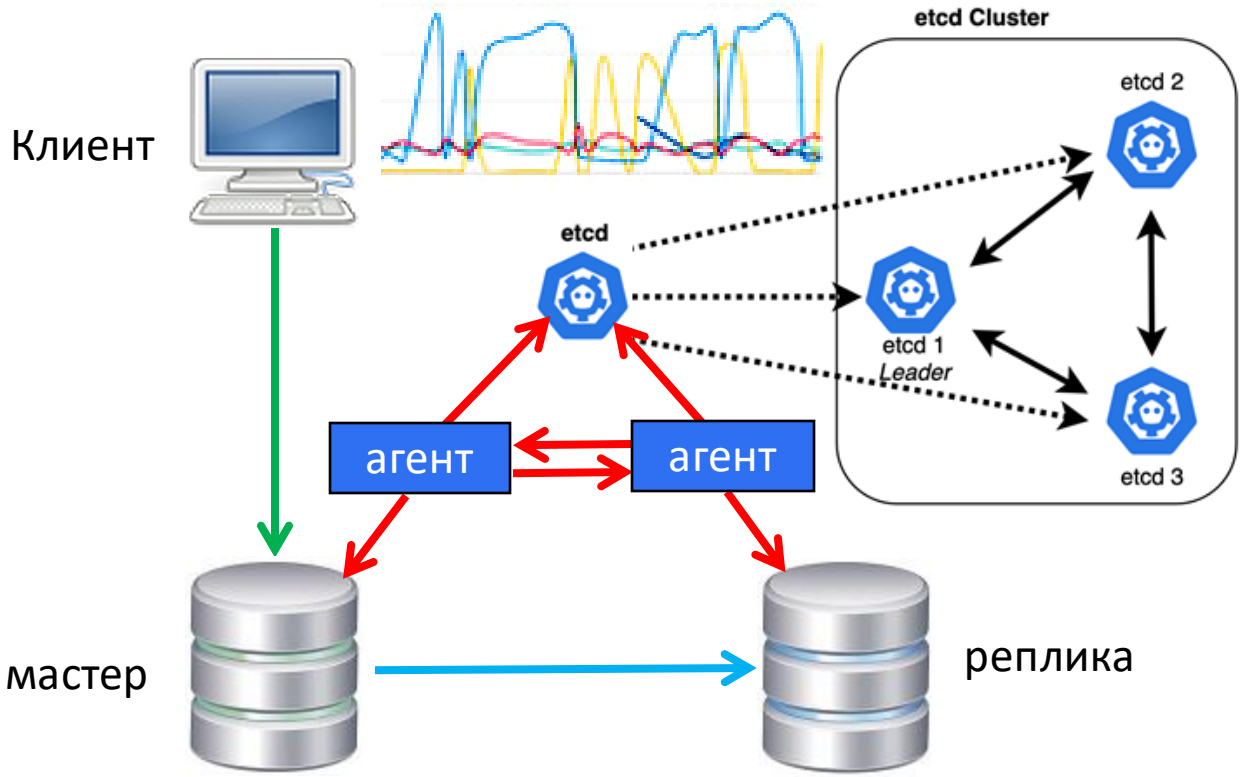
Внешнее кластерное ПО

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



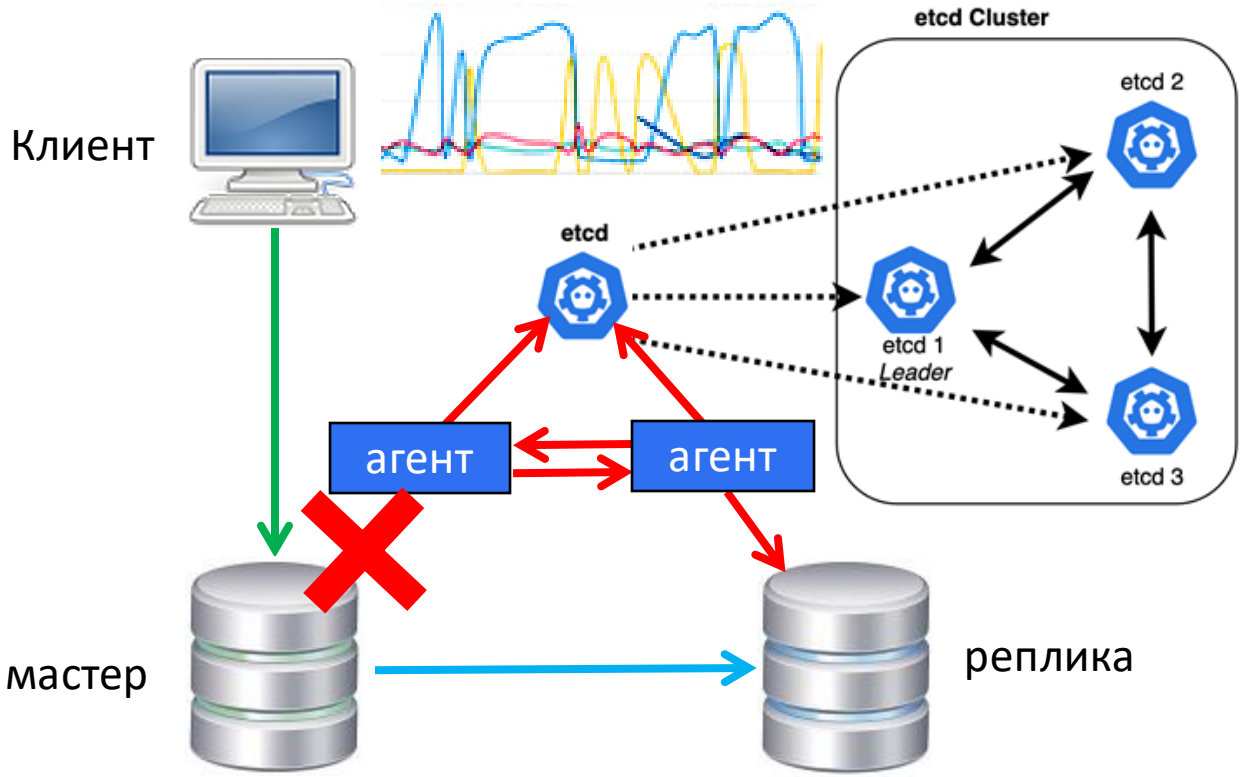
Внешнее кластерное ПО

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



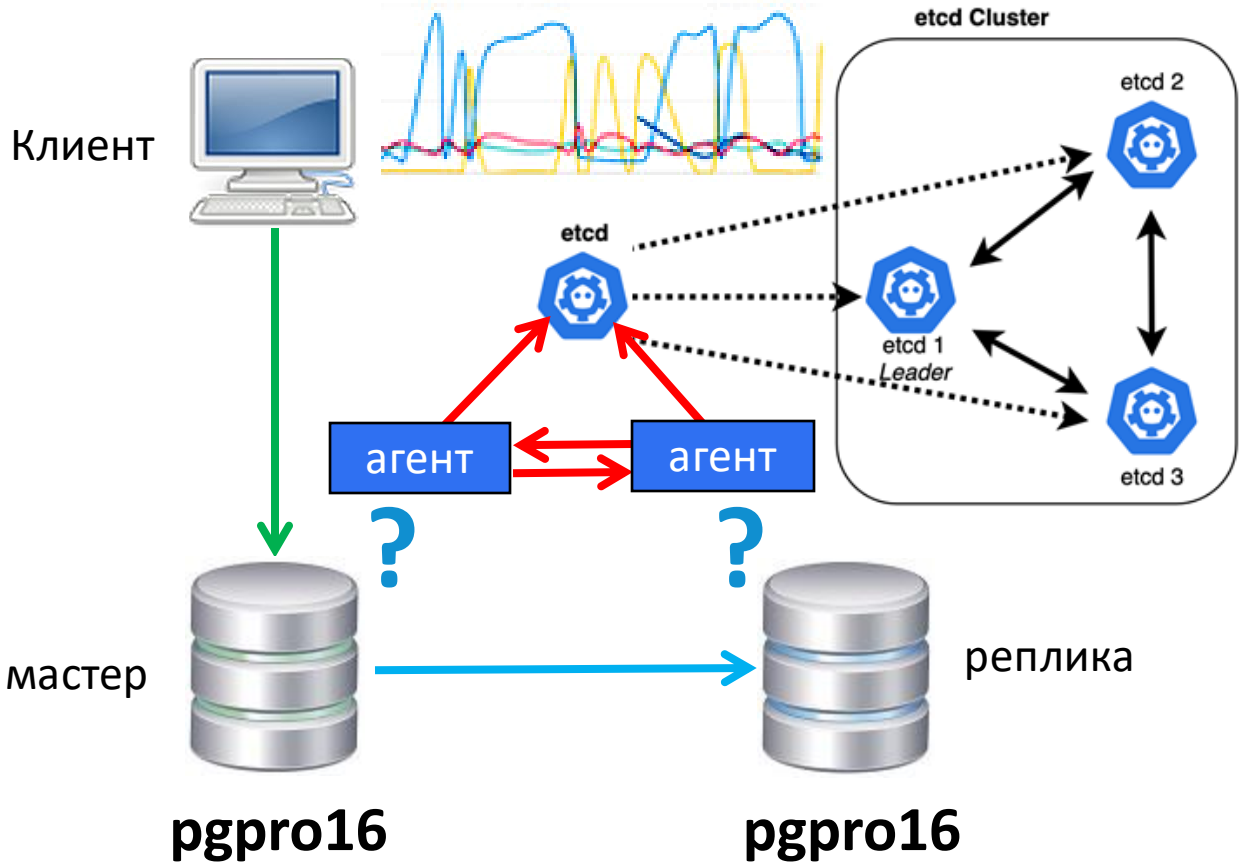
Внешнее кластерное ПО

- 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



Внешнее кластерное ПО

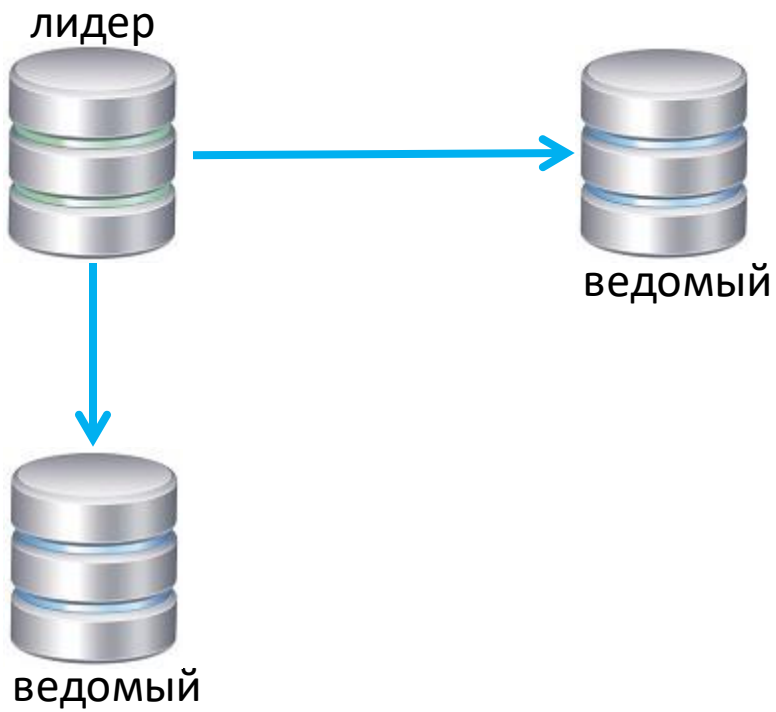
- Внешний кластер имеет сложную архитектуру (дополнительные узлы, сетевые каналы и т.п.)
- Для элементов кластерного ПО тоже требуется отказоустойчивость
- Сложность мониторинга
- Большая нагрузка на БД может расцениваться как отказ узла
- Задержка с обновлениями версий



Встроенный отказоустойчивый кластер BiHA
Архитектура

Кластер состоит из нескольких узлов

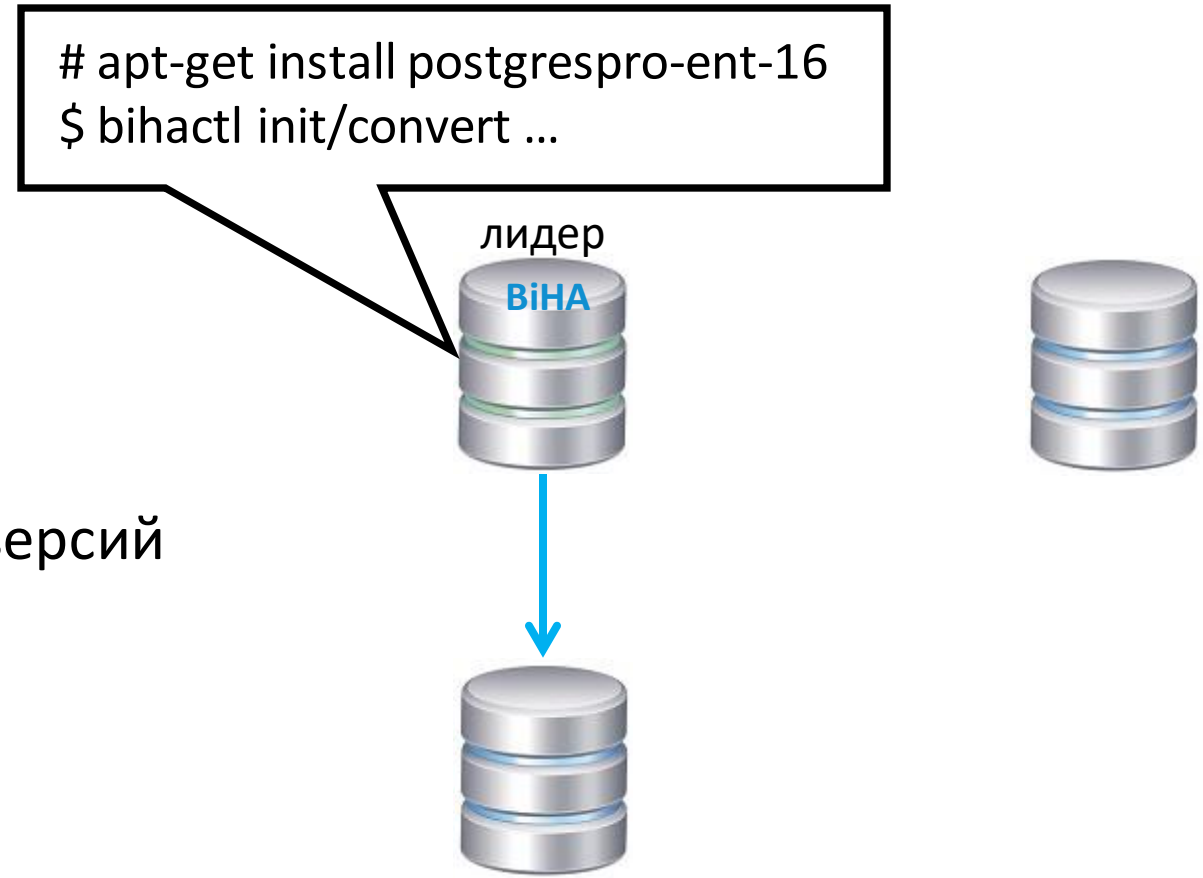
- один является лидером (leader),
- другие являются ведомыми (follower).



Встроенный отказоустойчивый кластер BiHA

Простая установка

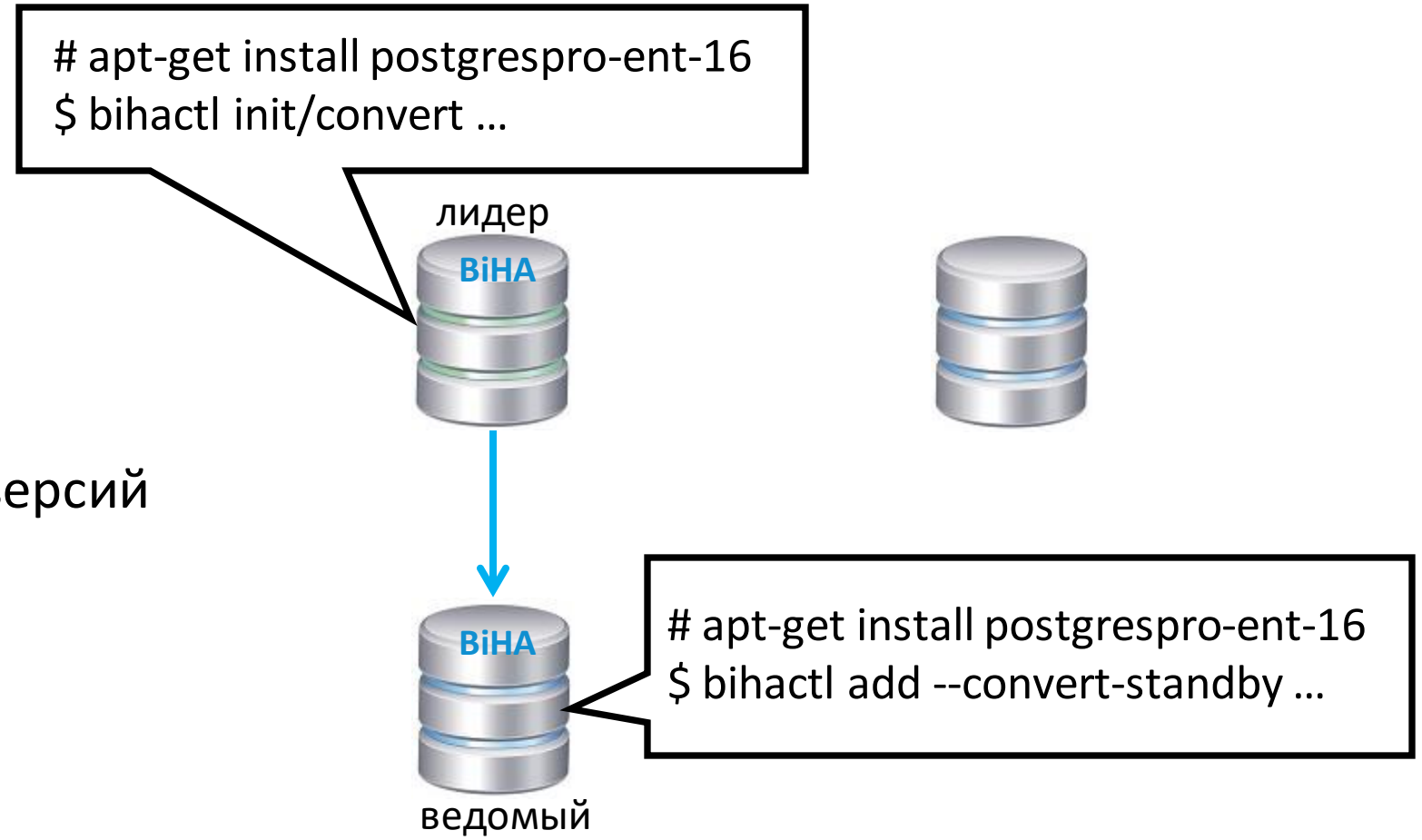
- BiHA кластер встроен в Postgres Pro.
- Простая установка и конфигурирование
- Не требуется установка дополнительного ПО
- Оперативные обновления версий



Встроенный отказоустойчивый кластер BiHA

Простая установка

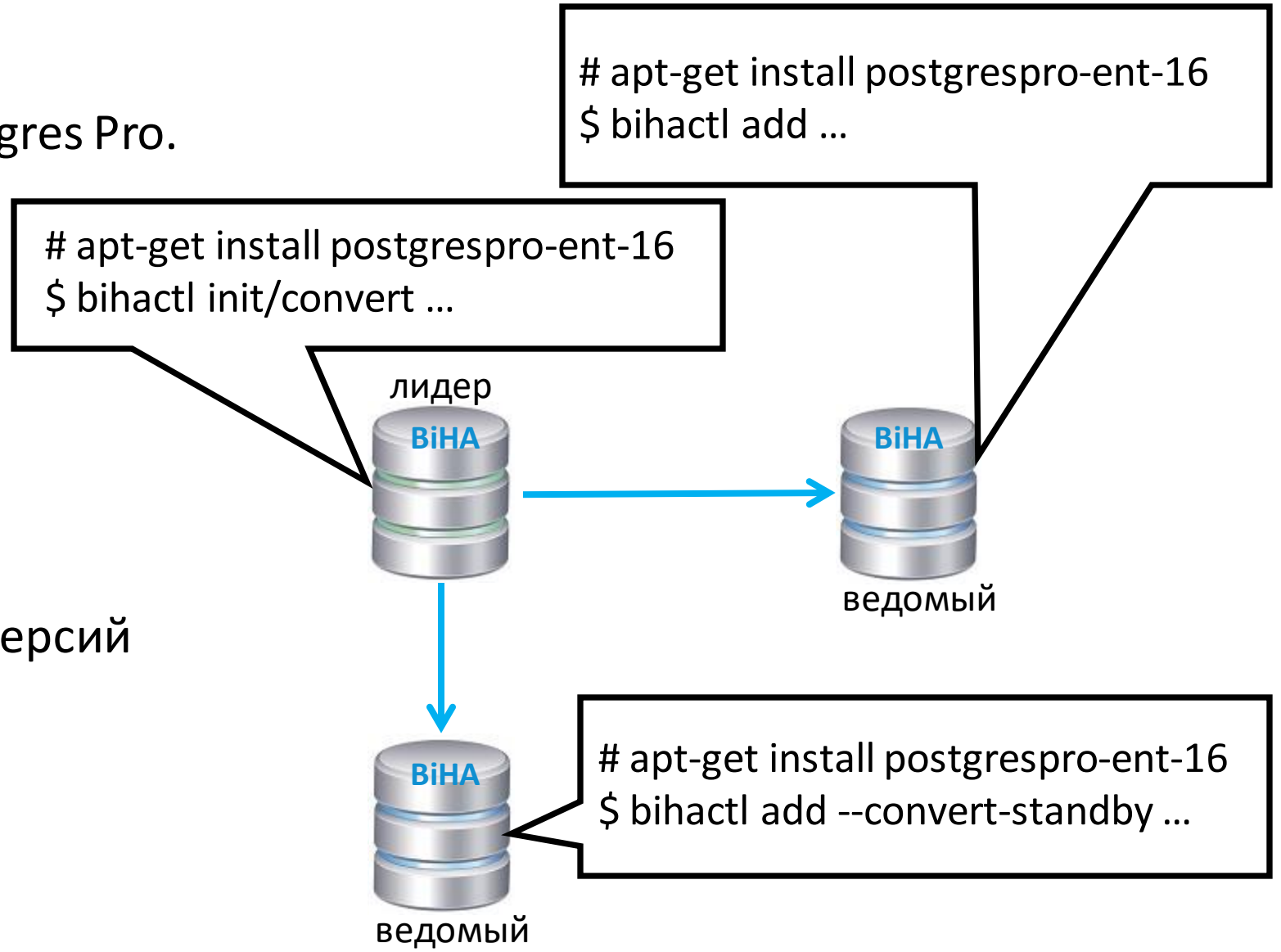
- BiHA кластер встроен в Postgres Pro.
- Простая установка и конфигурирование
- Не требуется установка дополнительного ПО
- Оперативные обновления версий



Встроенный отказоустойчивый кластер BiHA

Простая установка

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

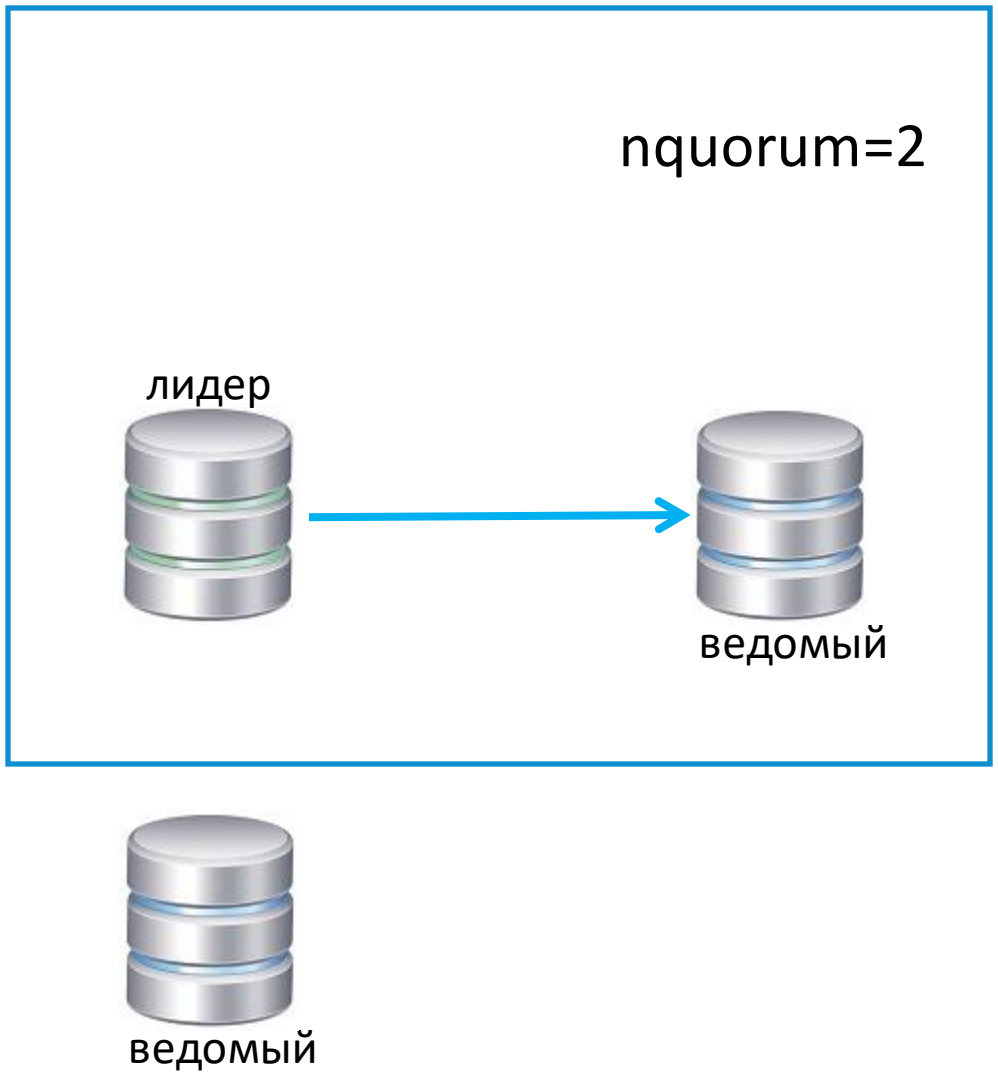


Встроенный отказоустойчивый кластер BiHA

Кластерный кворум

Кворум определяет минимальное количество узлов кластера

Лидер продолжает работать, если соблюдается кворум

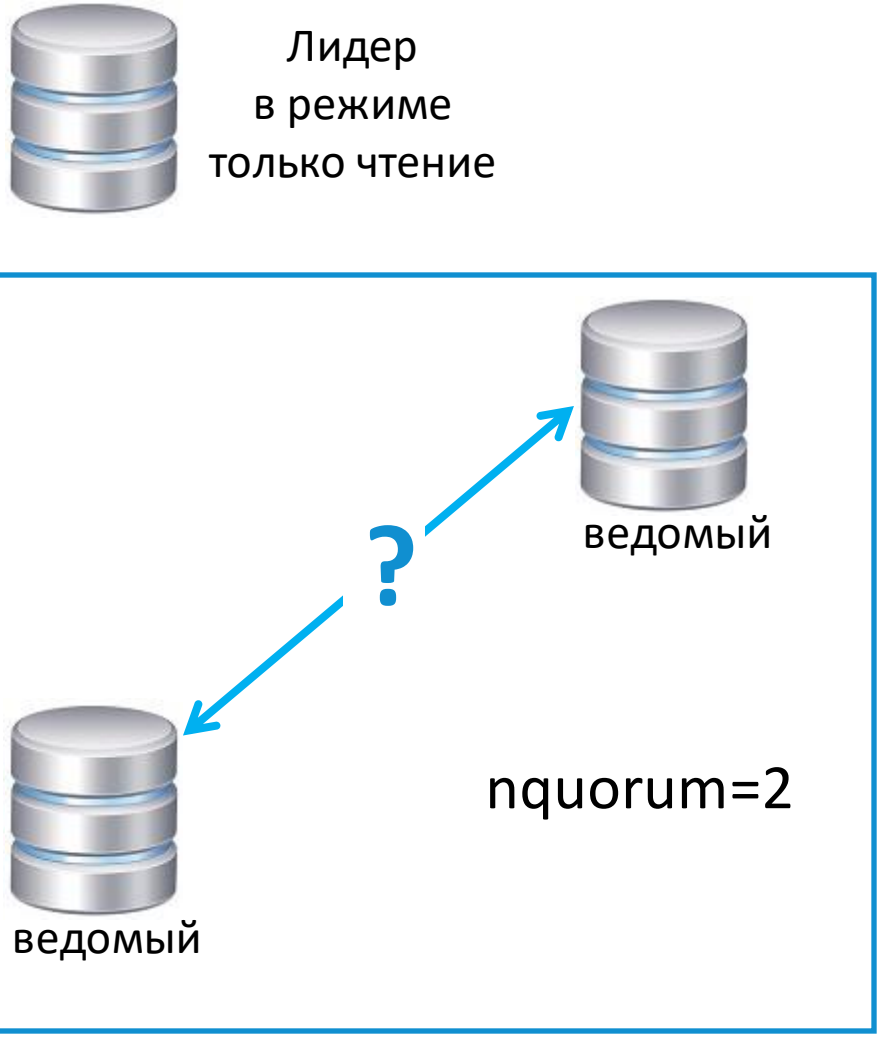


Встроенный отказоустойчивый кластер BiHA

Кластерный кворум

Лидер переходит в режим только чтение, если не соблюдается кворум

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

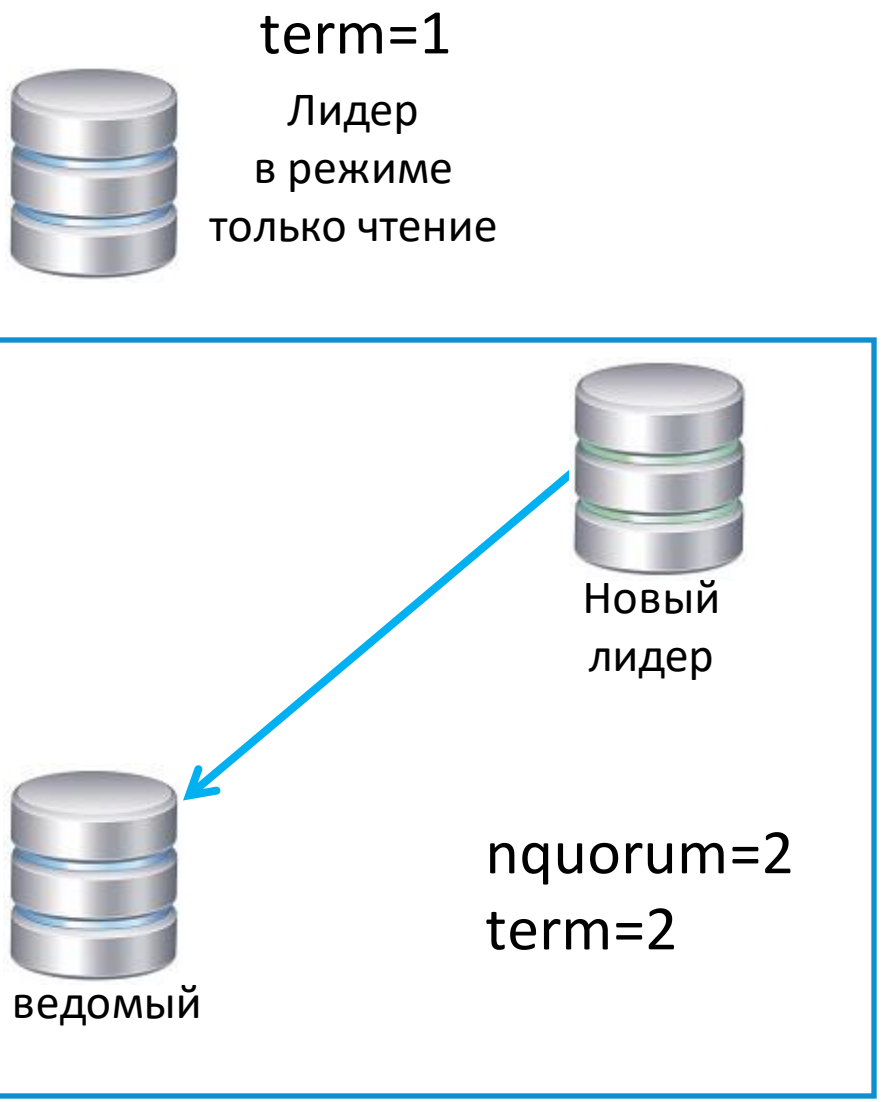


Встроенный отказоустойчивый кластер BiHA

Поколение кластера

После выбора нового лидера
в кластере меняется поколение

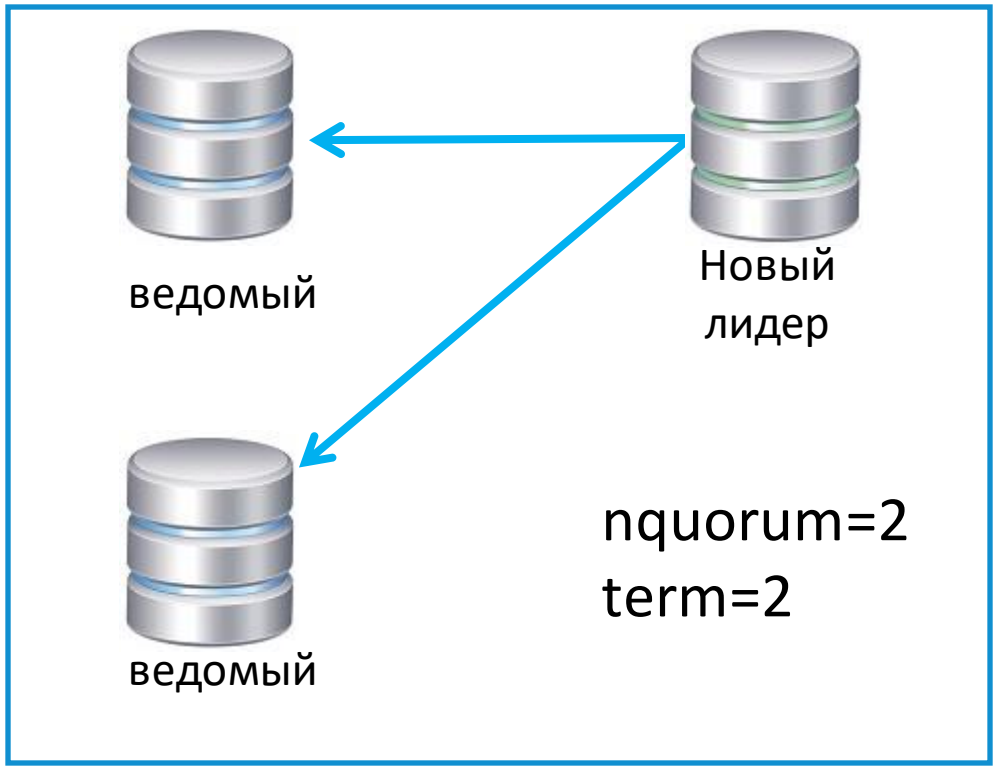
Старый лидер остаётся в старом
поколении



Встроенный отказоустойчивый кластер BiHA

Поколение кластера

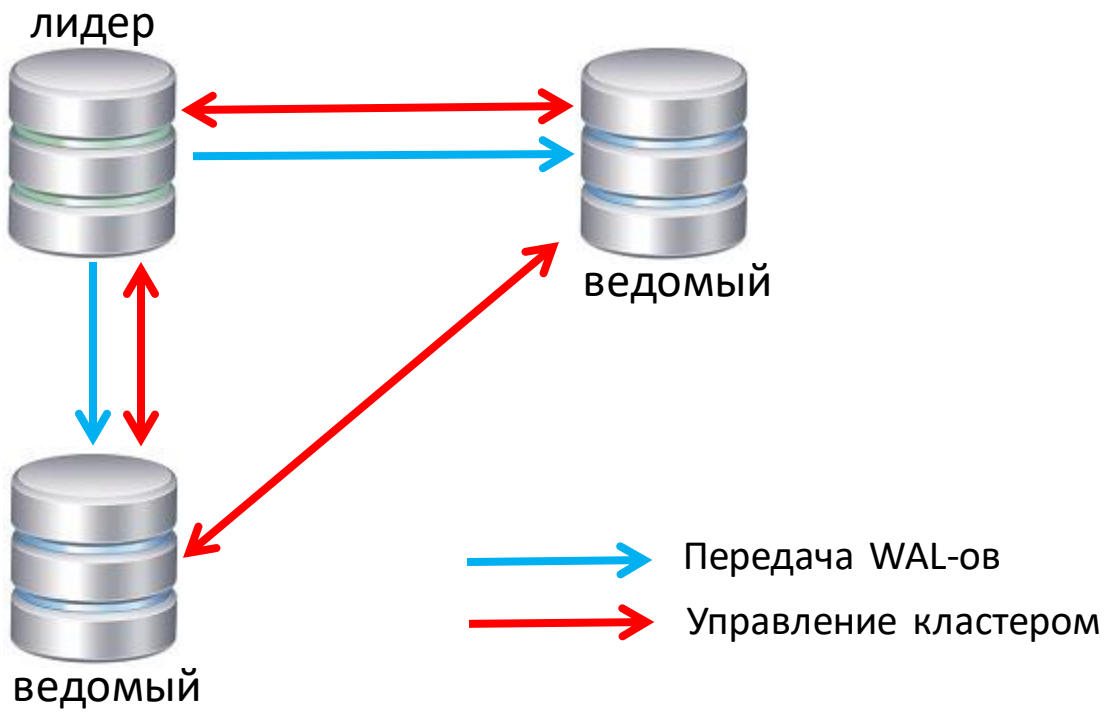
При возвращении старого лидера в кластер он не может быть уже лидером и переходит в режим ведомого



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.

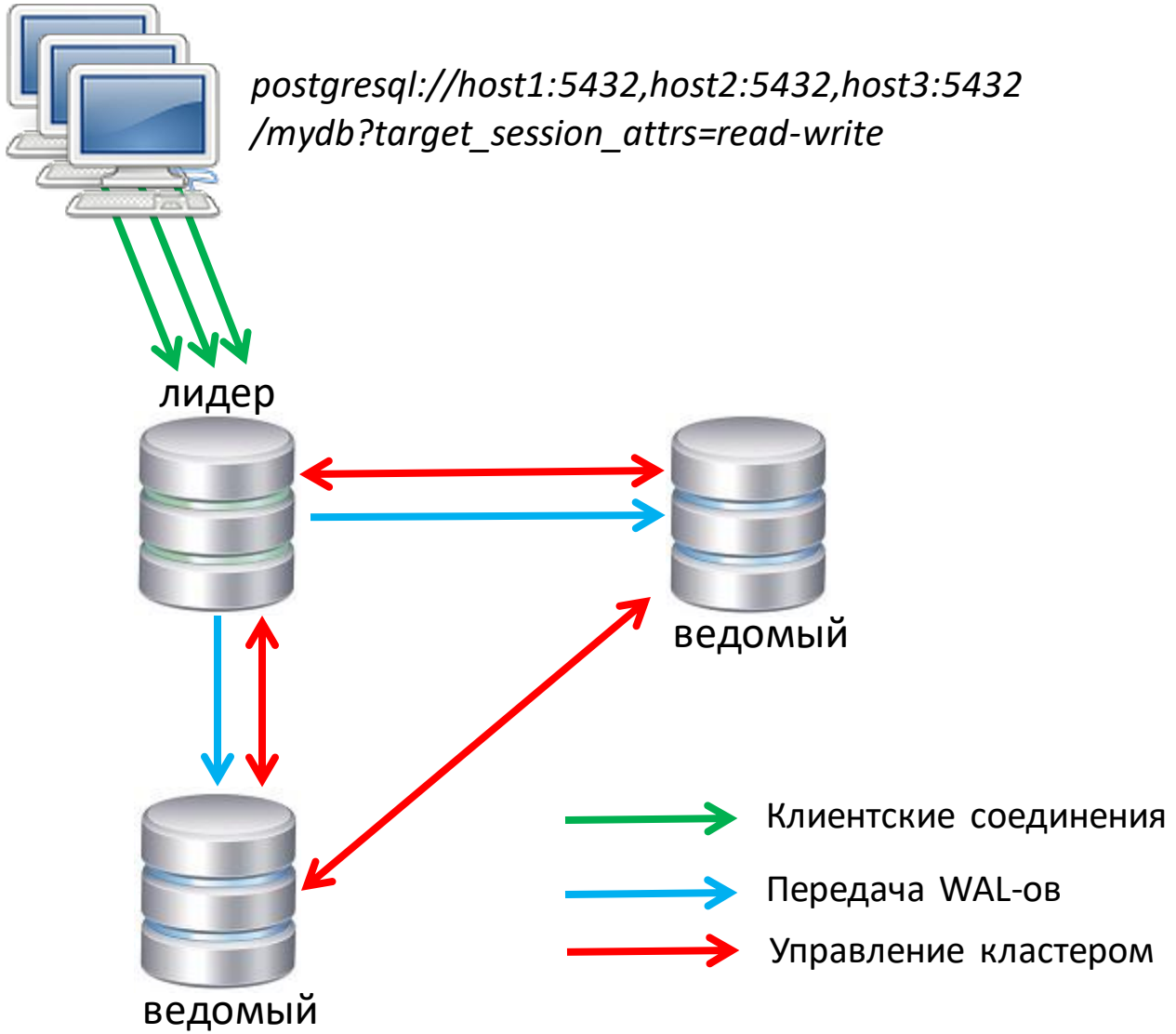


Автоматическое переключение соединения на стороне клиента на новый мастер

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

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

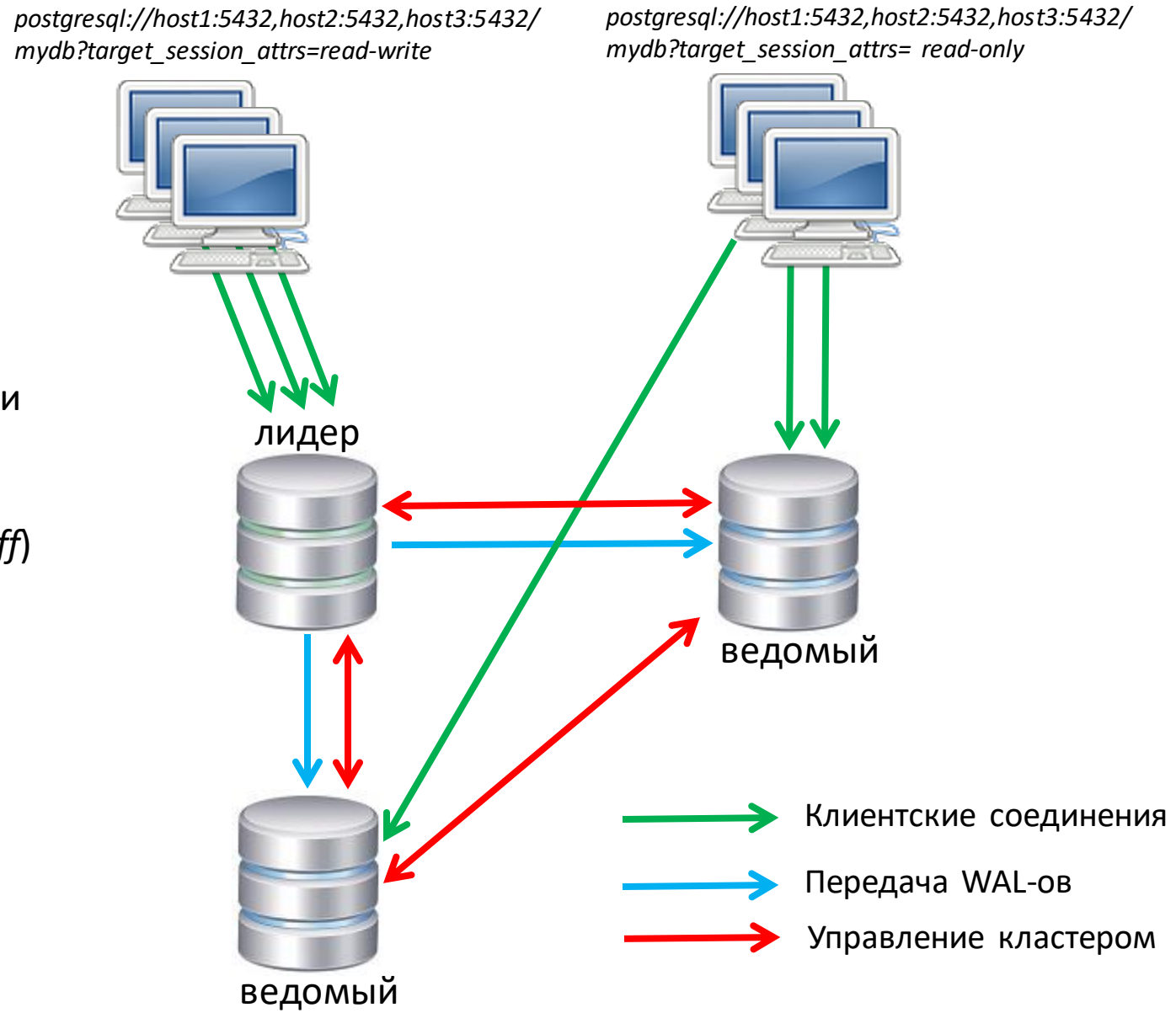
При сбое узла клиент автоматически подключится к новому лидеру



Автоматическое переключение соединения на стороне клиента

Параметр *target_session_attrs* при установке соединения вместе с указанием имён нескольких серверов позволяет выбрать из них первый подходящий вариант:

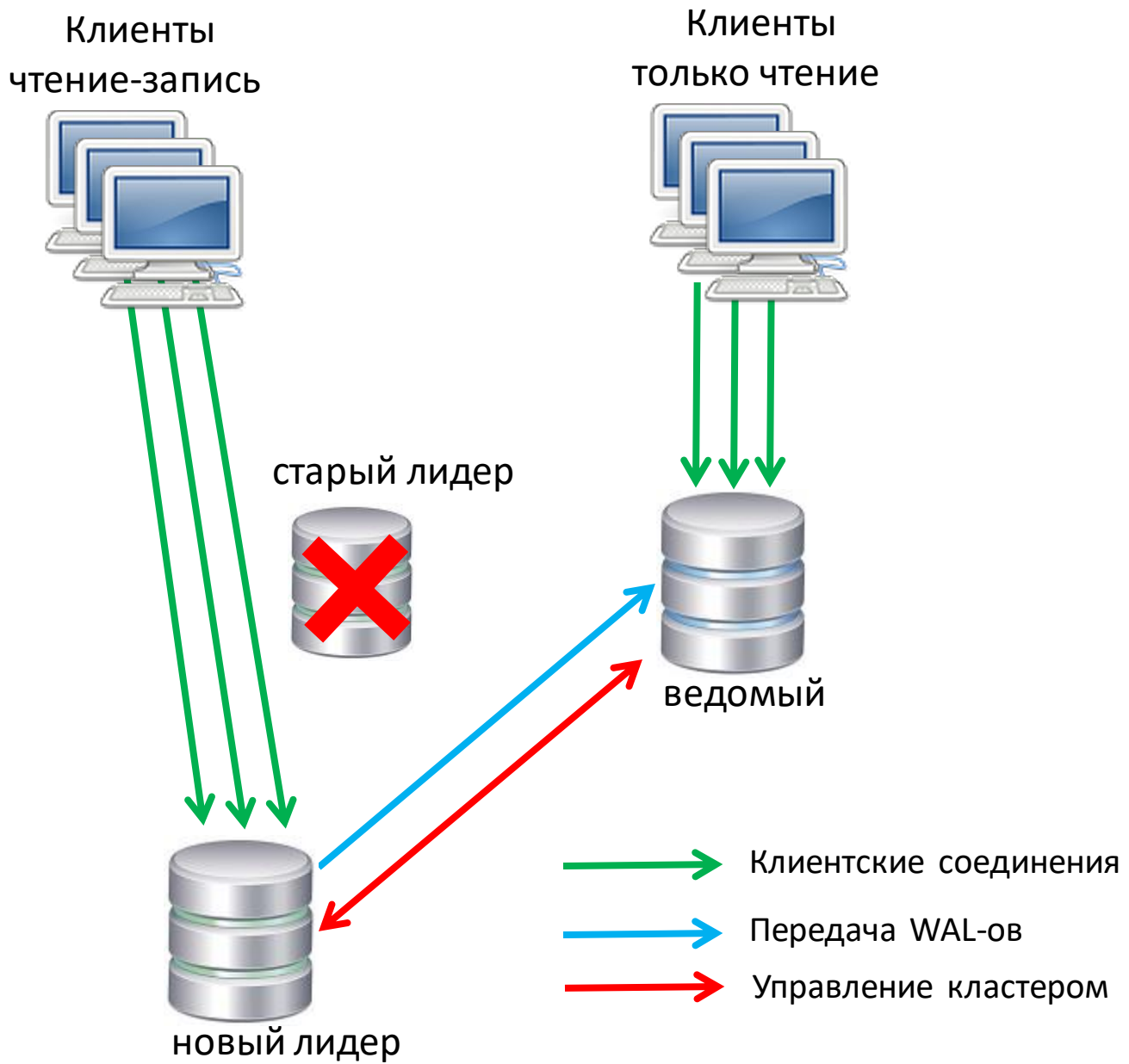
- *any* (по умолчанию) - любое успешное соединение
- *read-write* - сеанс должен принимать транзакции чтения-записи по умолчанию:
  - сервер не в режиме горячего резерва,
  - параметр *default_transaction_read_only = off*)
- *read-only* - сеанс не должен принимать транзакции чтения-записи по умолчанию
- *primary* - сервер не в режиме горячего резерва
- *standby* - сервер в режиме горячего резерва
- *prefer-standby* - сначала пытаться найти резервный сервер, но если ни один не является резервным, попробовать снова в режиме any



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)



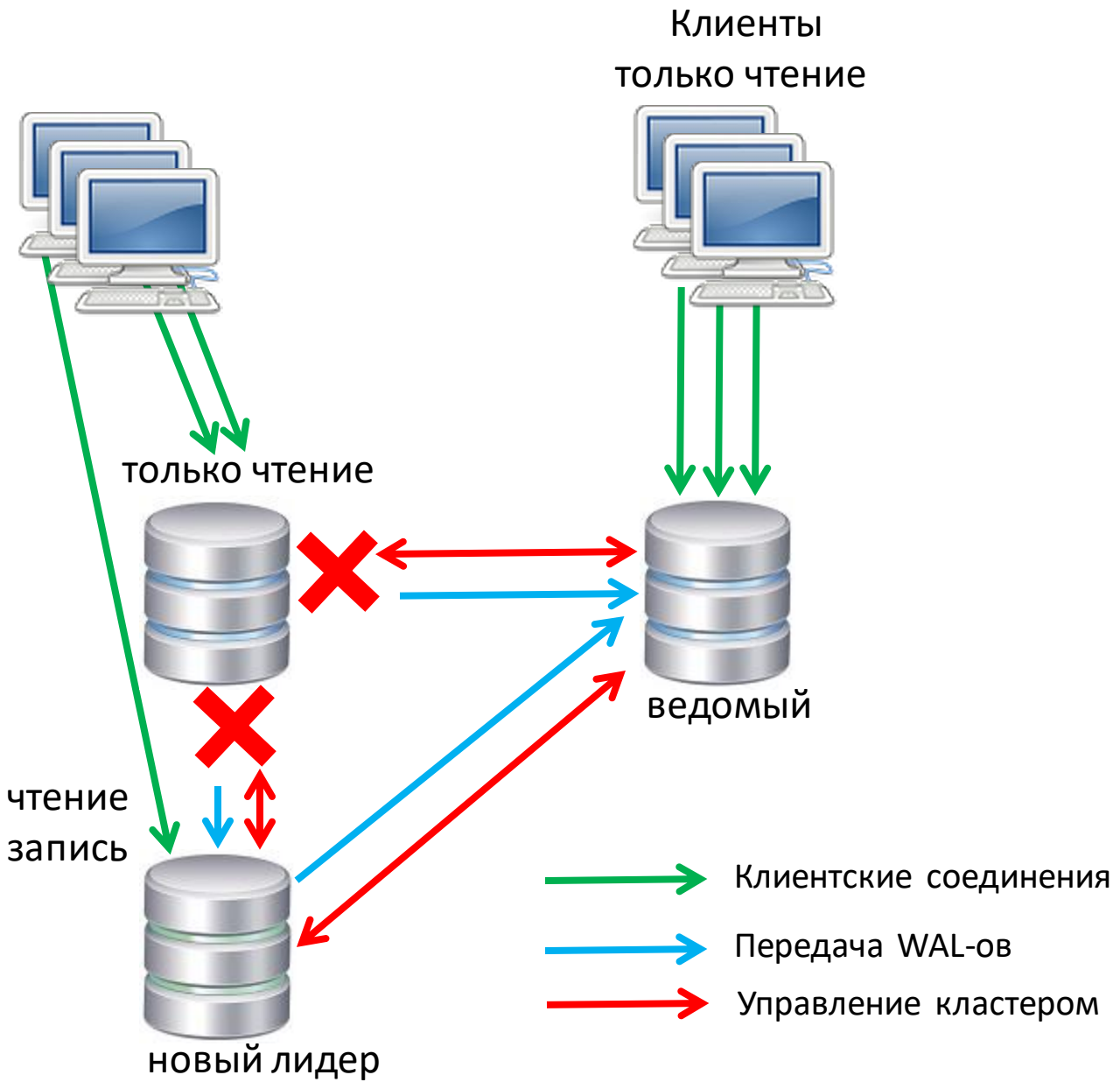
Встроенный отказоустойчивый кластер BiHA

Сетевая изоляция лидера

Когда лидер теряет связь
с необходимым для кворума 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

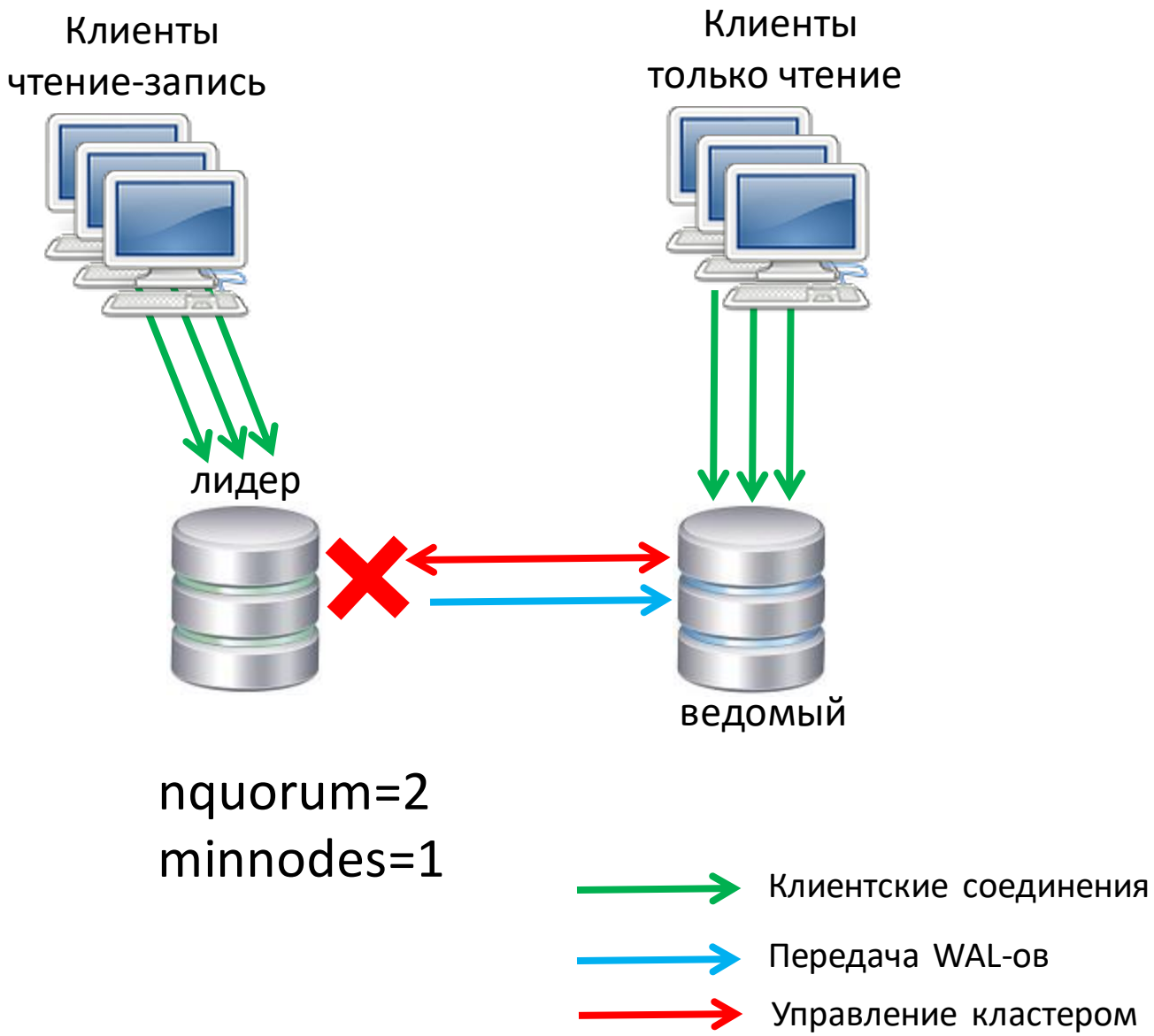
без кворума в режиме записи,

указав минимальное количество

работающих узлов (minnodes)

меньше чем минимальное количество

узлов для кворума (nquorum).

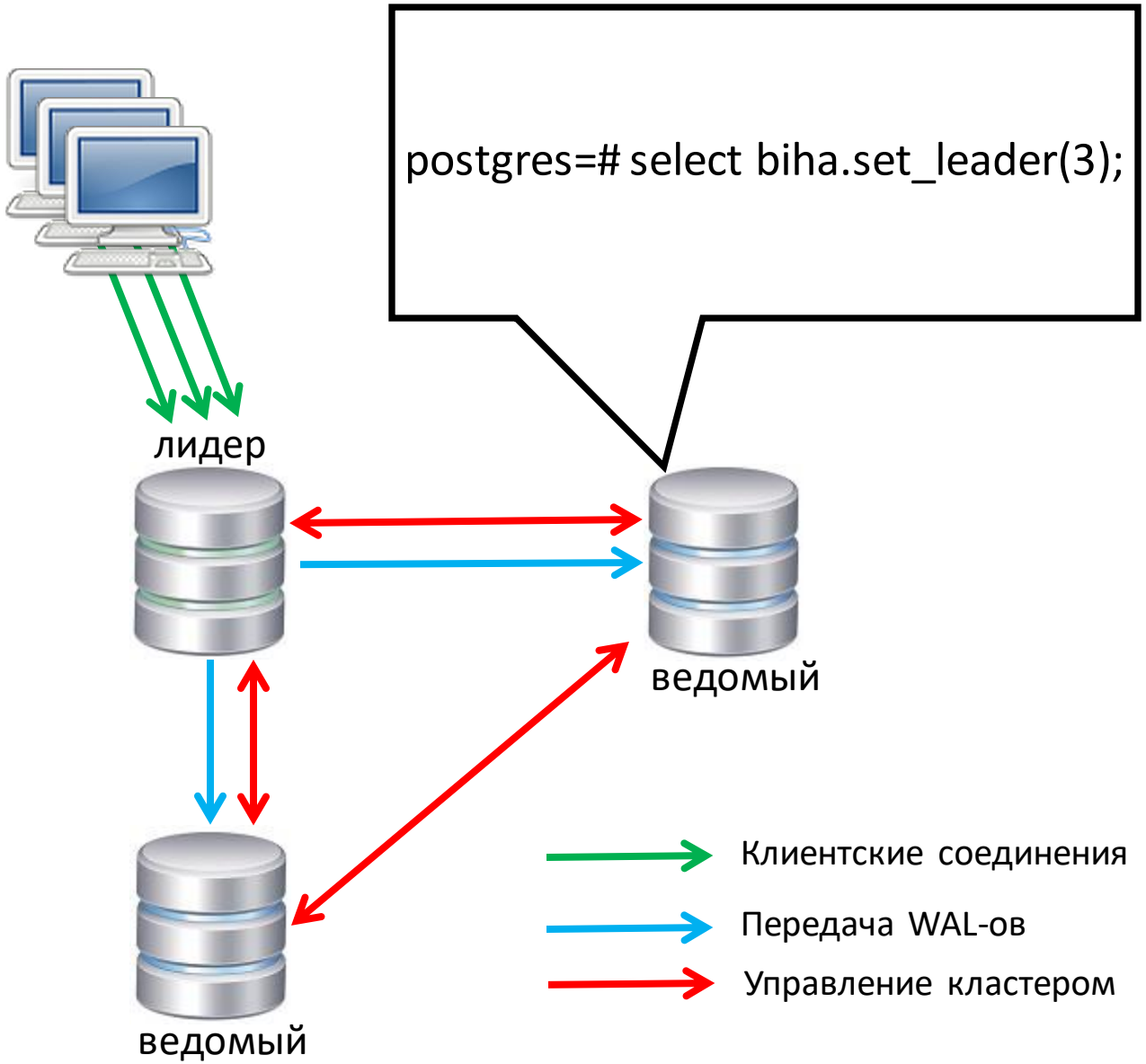


Встроенный отказоустойчивый кластер BiHA

Назначение лидера вручную

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

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

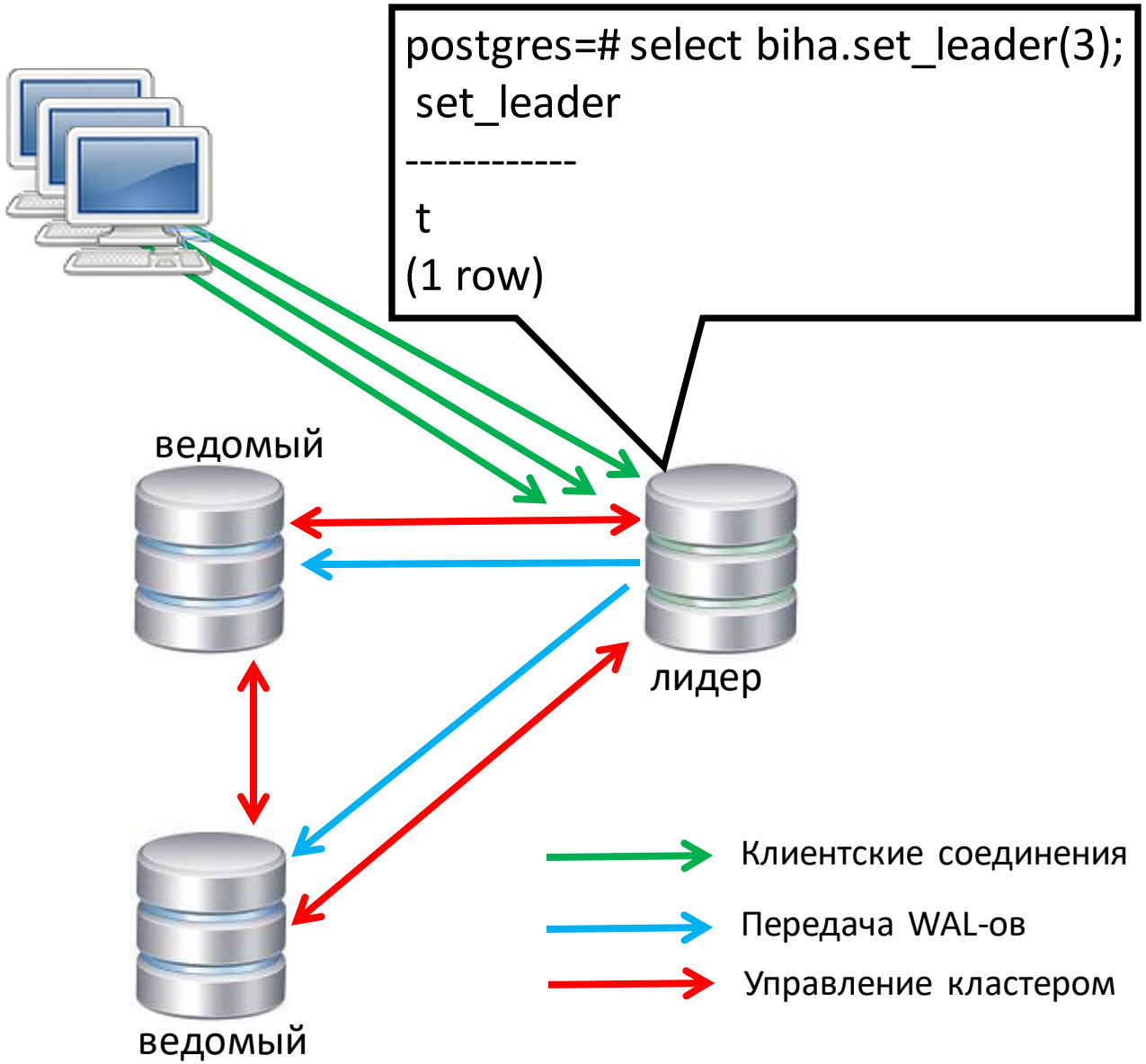


Встроенный отказоустойчивый кластер BiHA

Назначение лидера вручную

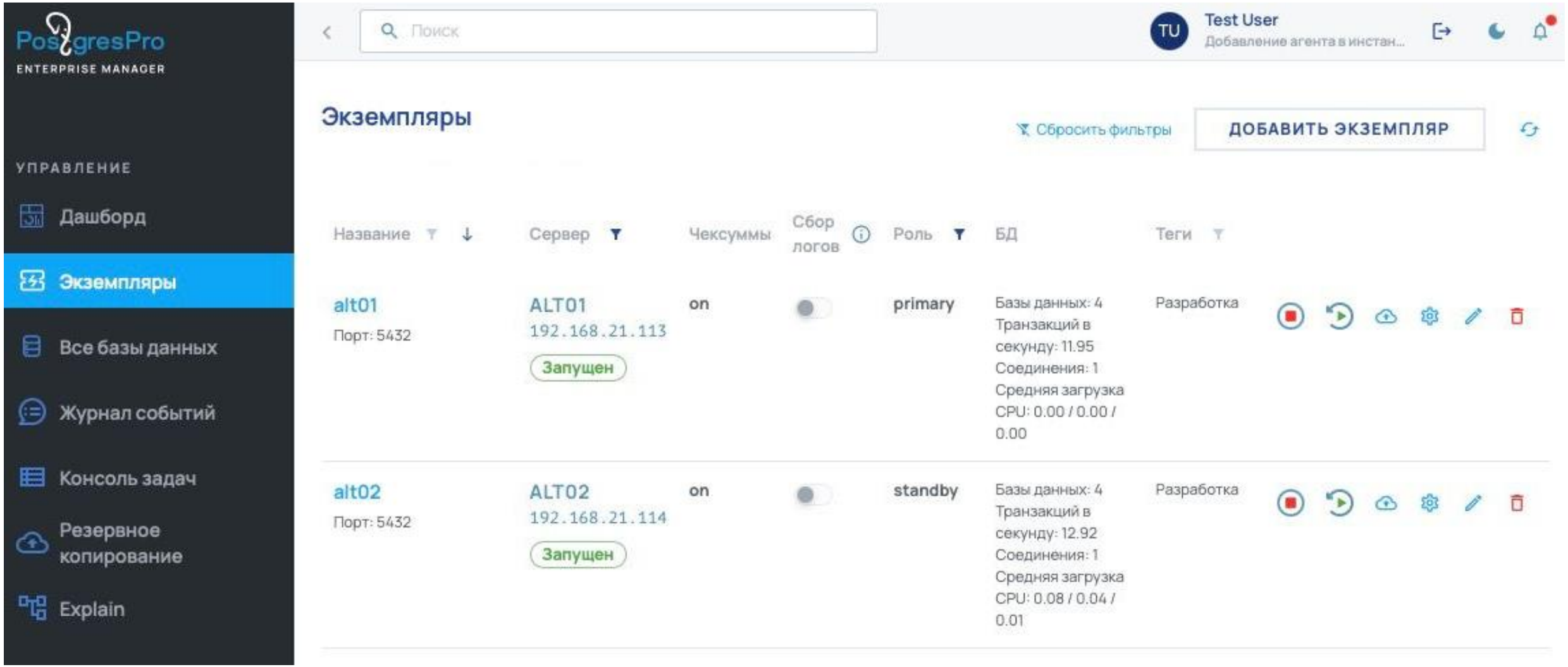
Назначение лидера через SQL-интерфейс используя функцию set_leader(id):

- 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