

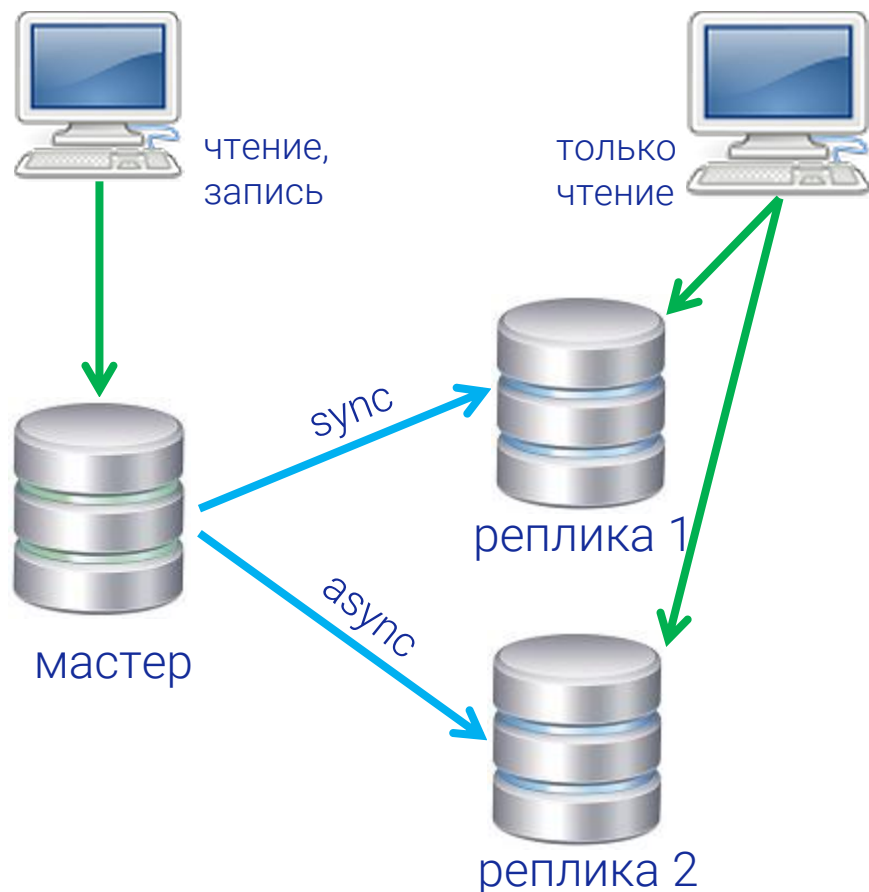


ViHA: встроенный отказоустойчивый кластер

Забелин Андрей
a.zabelin@postgrespro.ru



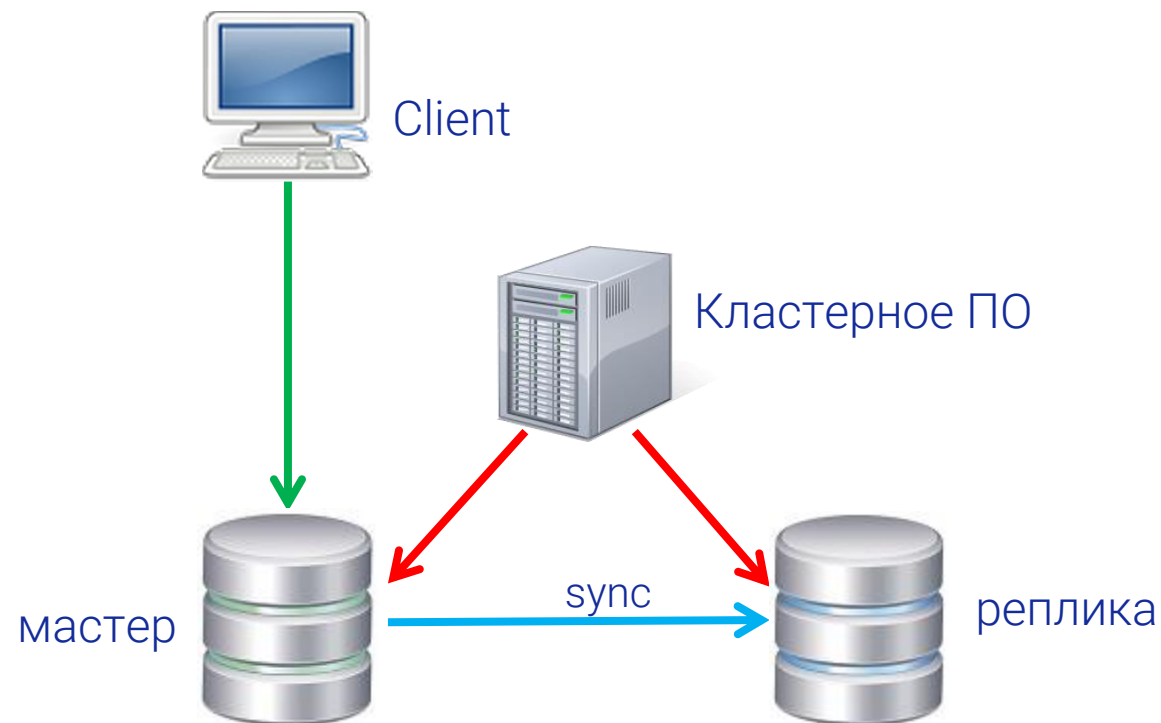
Postgres Pro : Физическая репликация



- 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

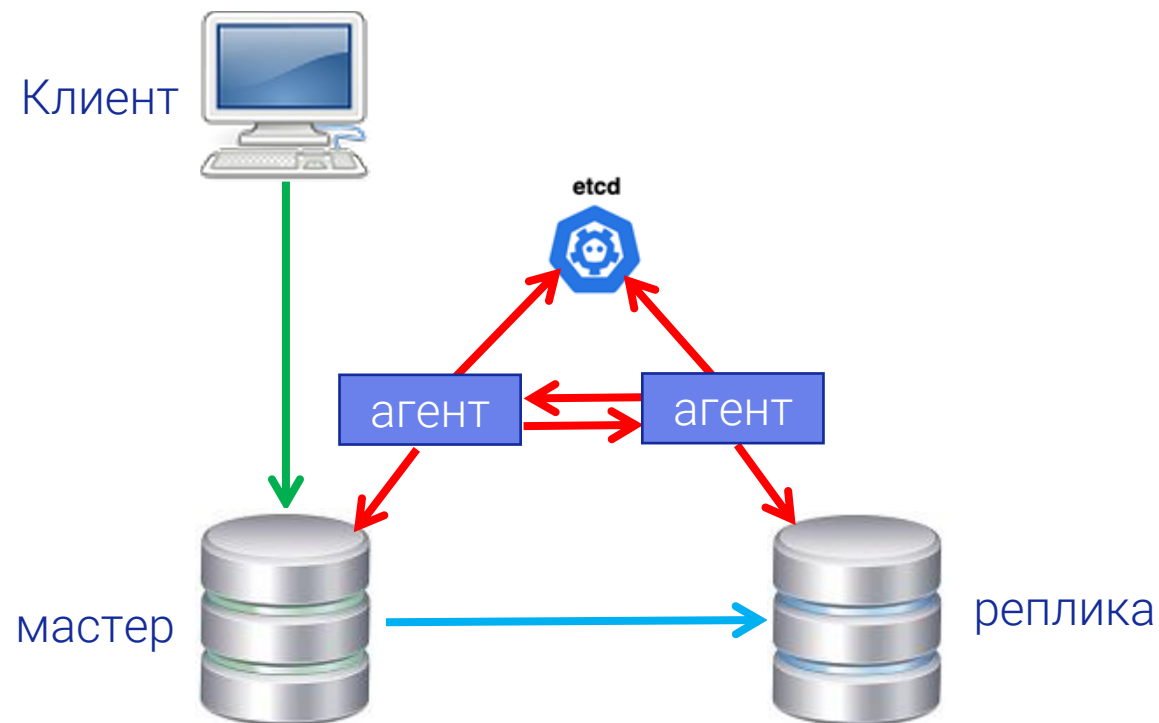
- Решение о смене ролей в отказоустойчивом кластере при сбое мастера может приниматься автоматически
- Необходимо также автоматически переключить на новый мастер и клиентов
- Основная задача кластерного ПО обнаружить сбой, сменить роль реплики на новый мастер, но при этом не допустить работу двух узлов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



Примеры кластерного ПО : Patroni, Stolon, Corosync + Pacemaker

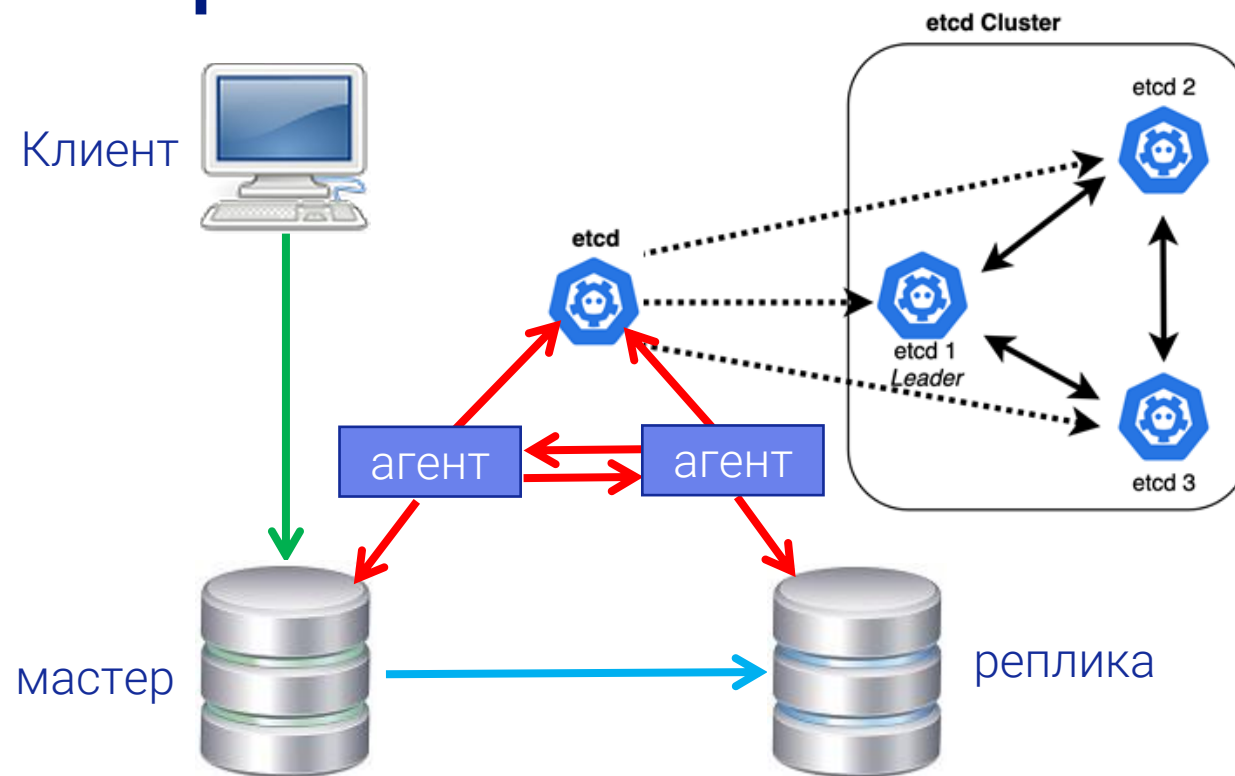
Недостатки внешнего кластерного ПО

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



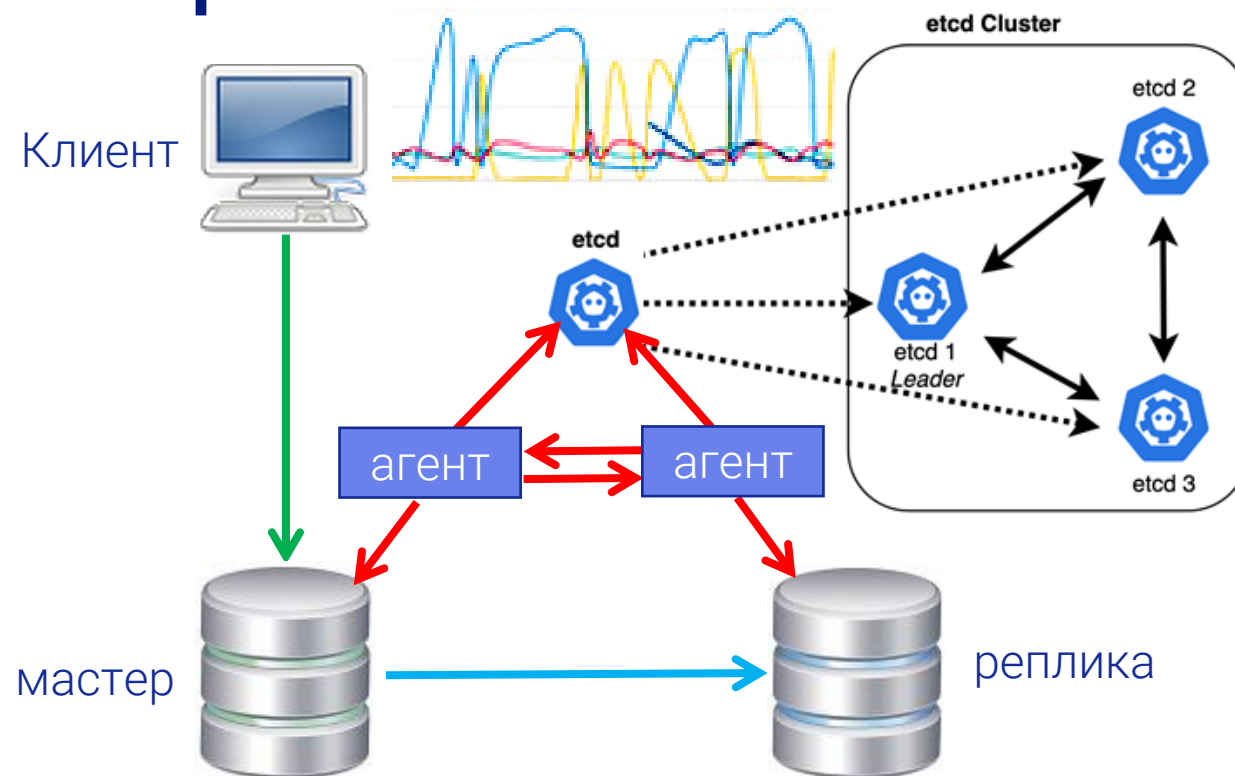
Недостатки внешнего кластерного ПО

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



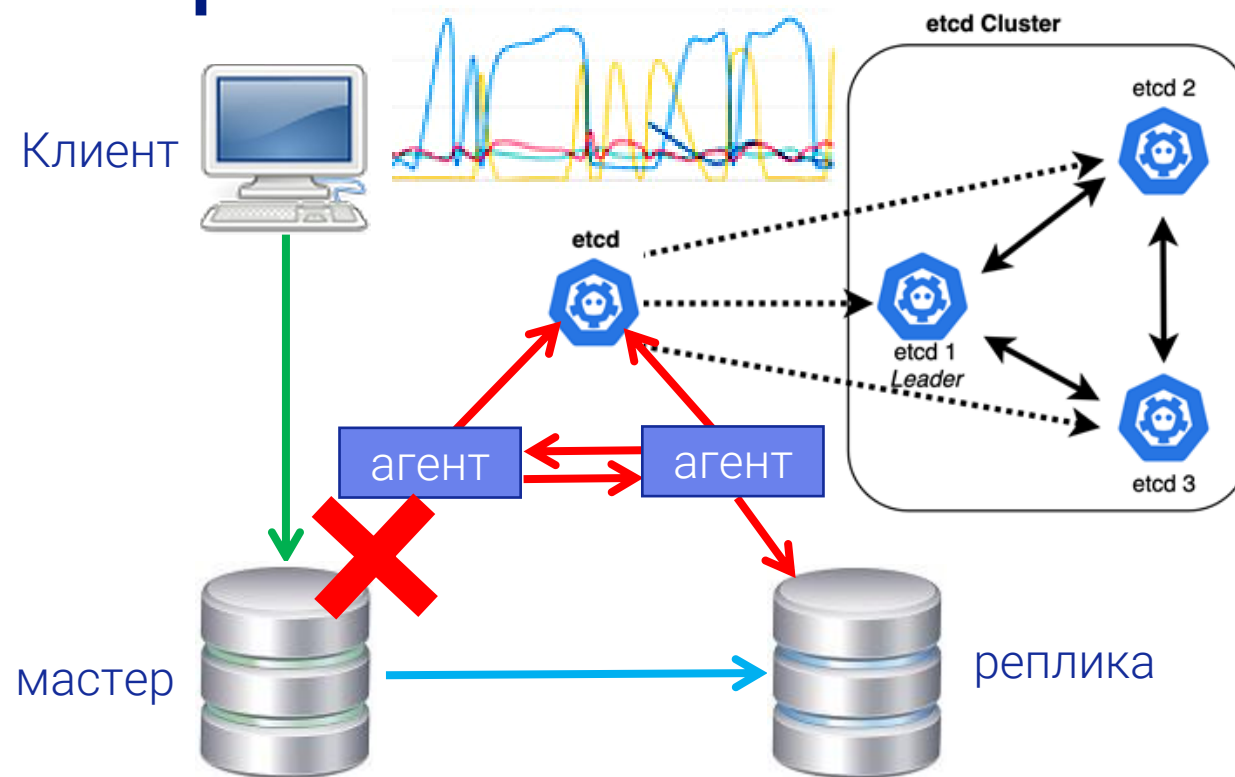
Недостатки внешнего кластерного ПО

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



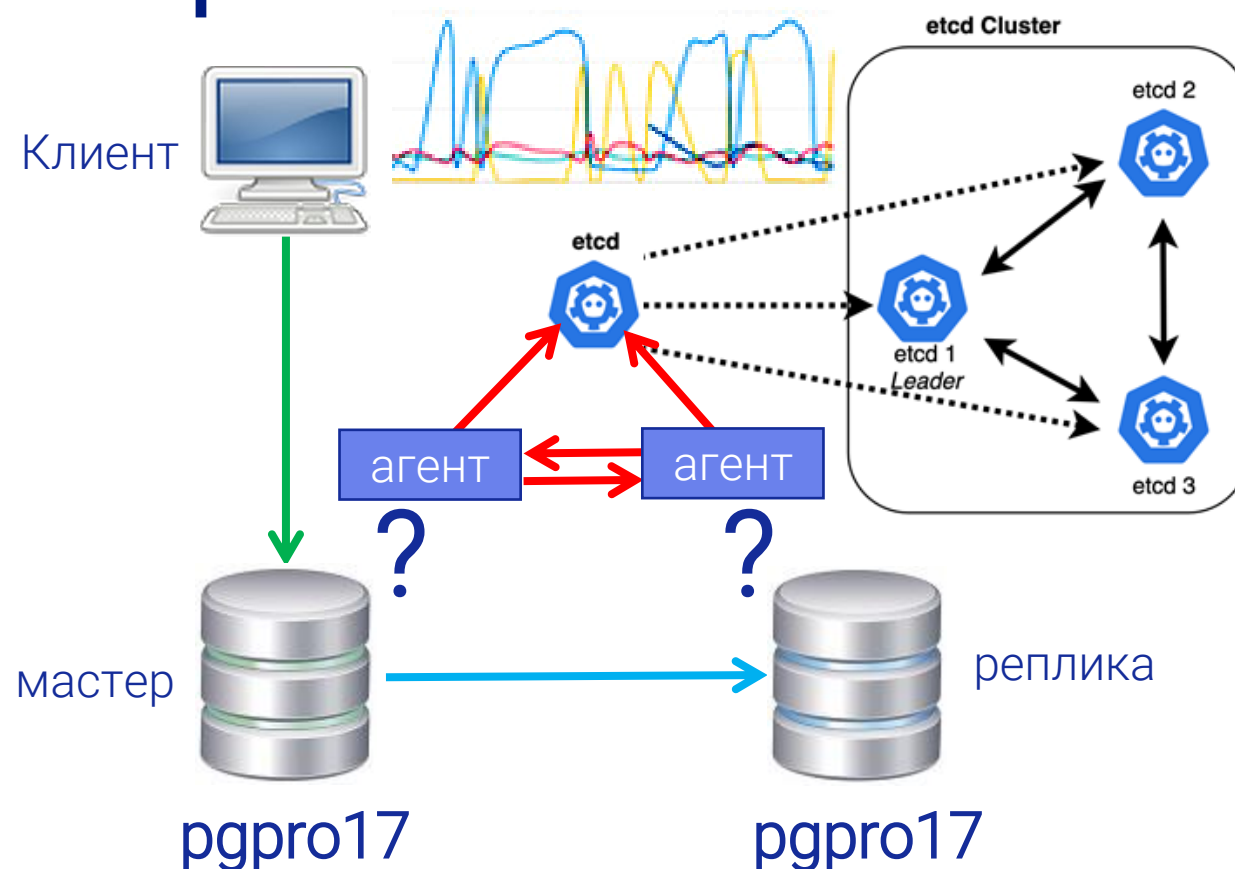
Недостатки внешнего кластерного ПО

- 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

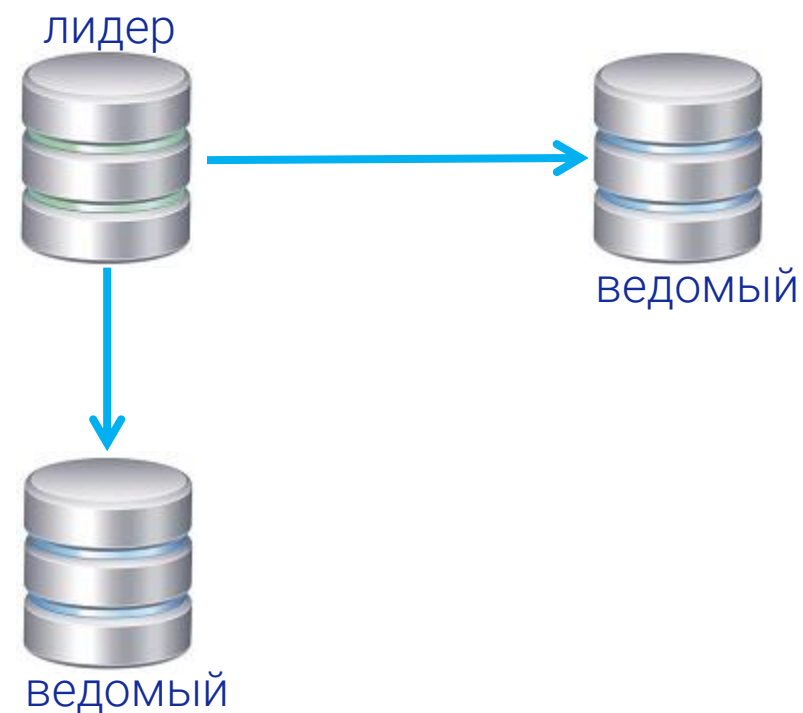
- Внешний кластер имеет сложную архитектуру (дополнительные узлы, сетевые каналы и т.п.)
- Для элементов кластерного ПО тоже требуется отказоустойчивость
- Сложность мониторинга
- Большая нагрузка на БД может расцениваться как отказ узла
- Задержка с обновлениями версий



ВiНA: Архитектура

Кластер состоит из нескольких узлов

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



BiHA: Простая установка

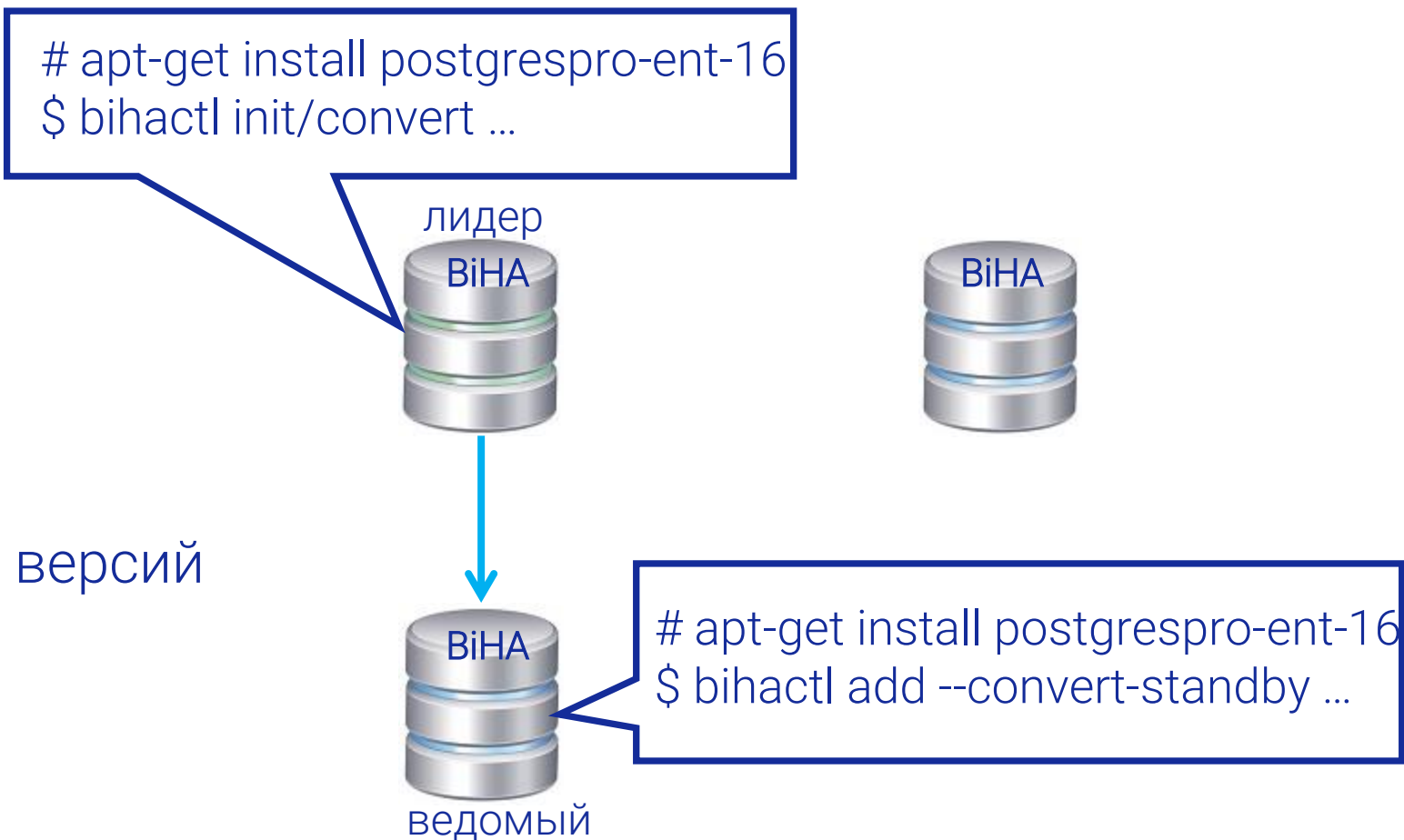
- BiHA кластер встроен в Postgres Pro.
- Простая установка и конфигурирование
- Не требуется установка дополнительного ПО
- Оперативные обновления версий

```
# apt-get install postgrespro-ent-16  
$ bihactl init/convert ...
```



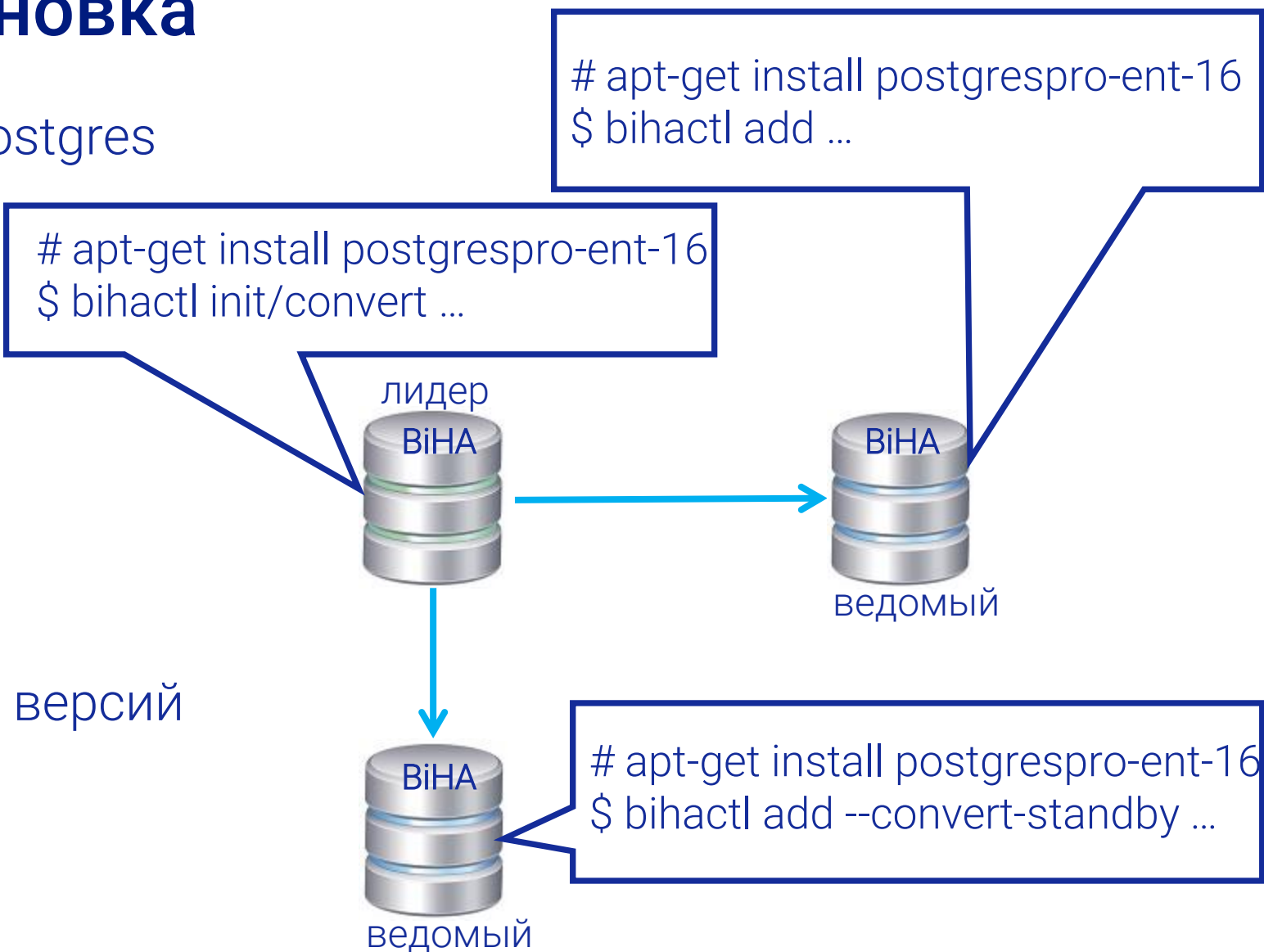
BiHA: Простая установка

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



ViHA: Простая установка

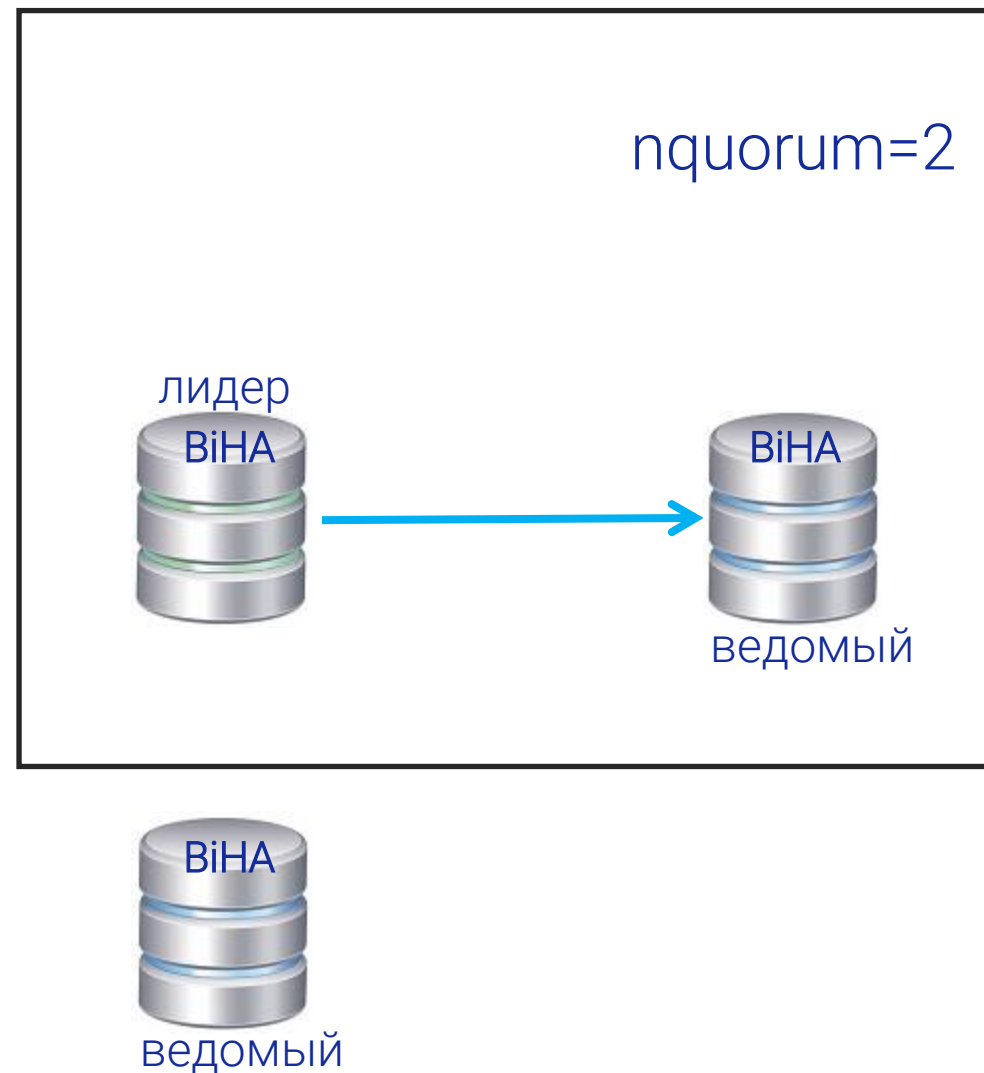
- ViHA кластер встроен в Postgres Pro.
- Простая установка и конфигурирование
- Не требуется установка дополнительного ПО
- Оперативные обновления версий



ViNA: Кластерный кворум

Кворум определяет минимальное количество узлов кластера

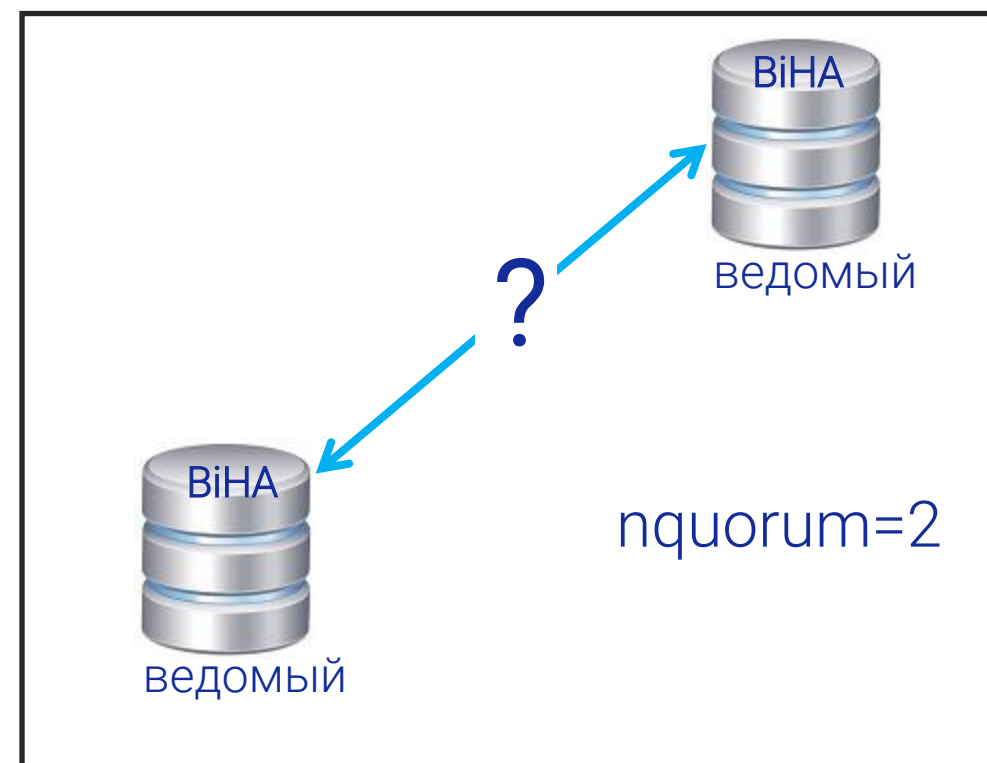
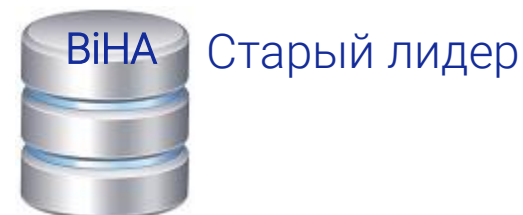
Лидер продолжает работать, если соблюдается кворум



ViHA: Кластерный кворум

Лидер не может продолжать работу, если не соблюдается кворум

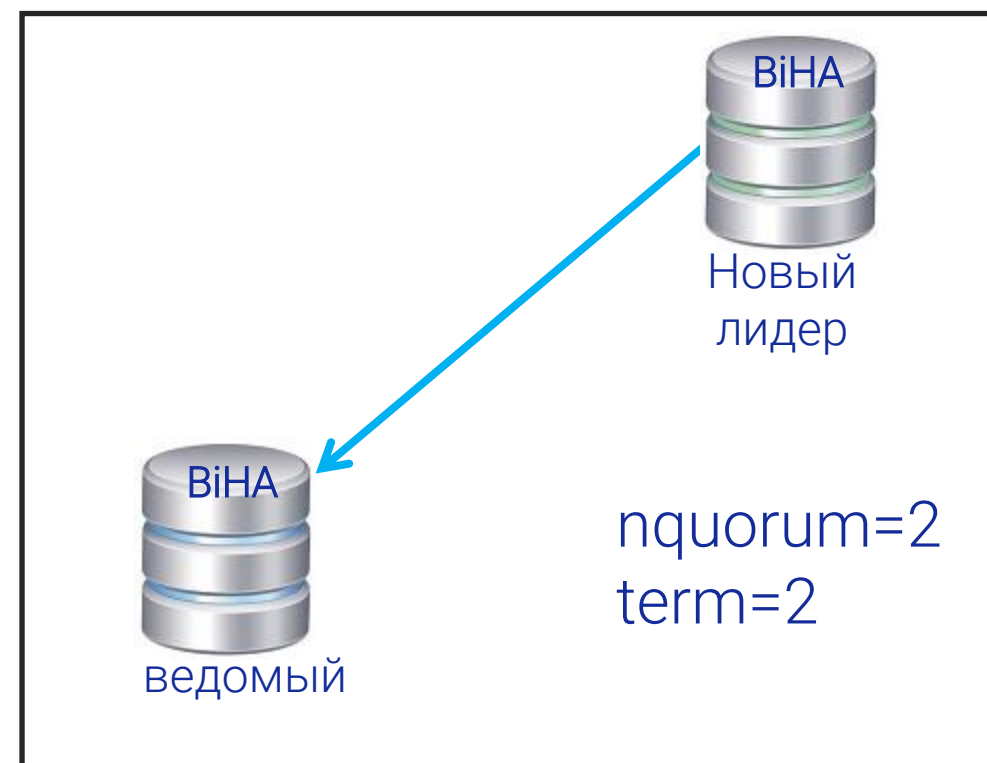
Ведомые организуют выборы нового лидера, если кластер содержит достаточное количество узлов



ВiНА: Поколение кластера

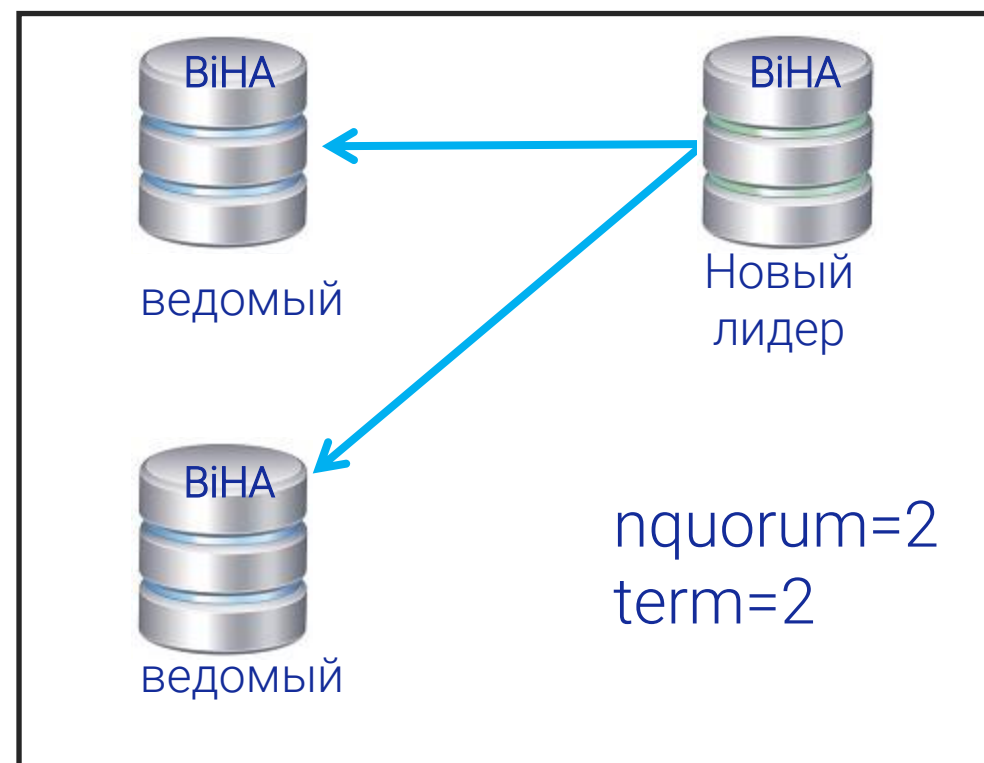
После выбора нового лидера
в кластере меняется поколение

Старый лидер остаётся в старом
поколении



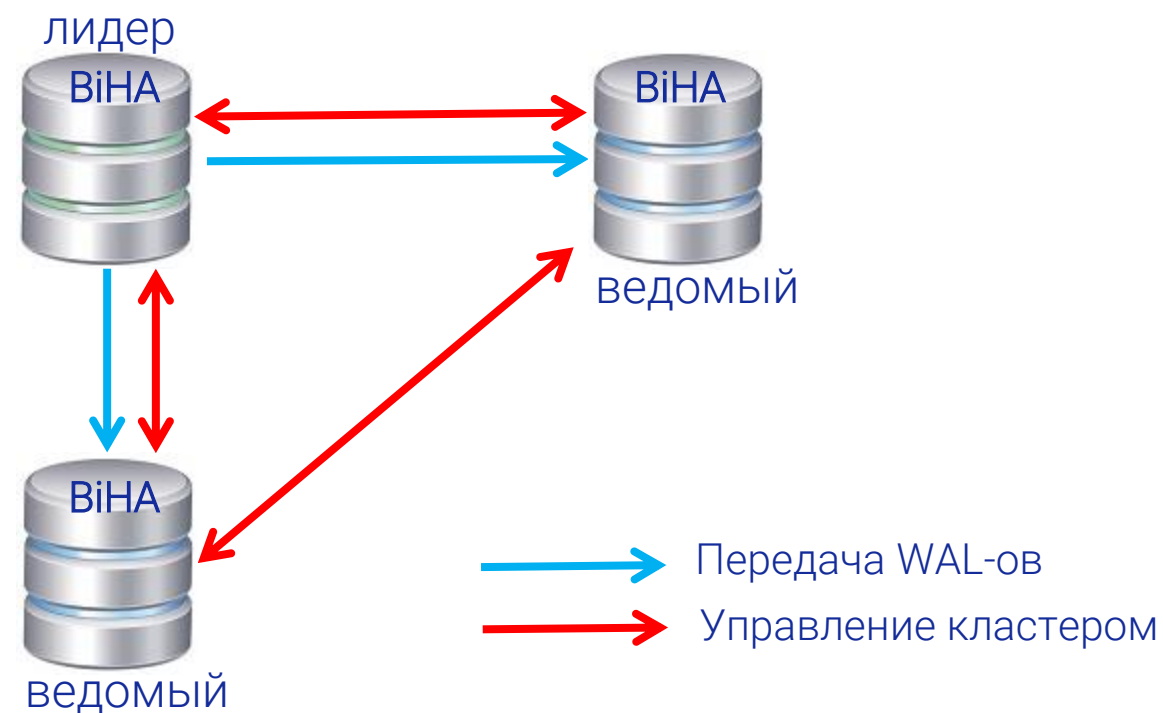
ВiНА: Поколение кластера

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



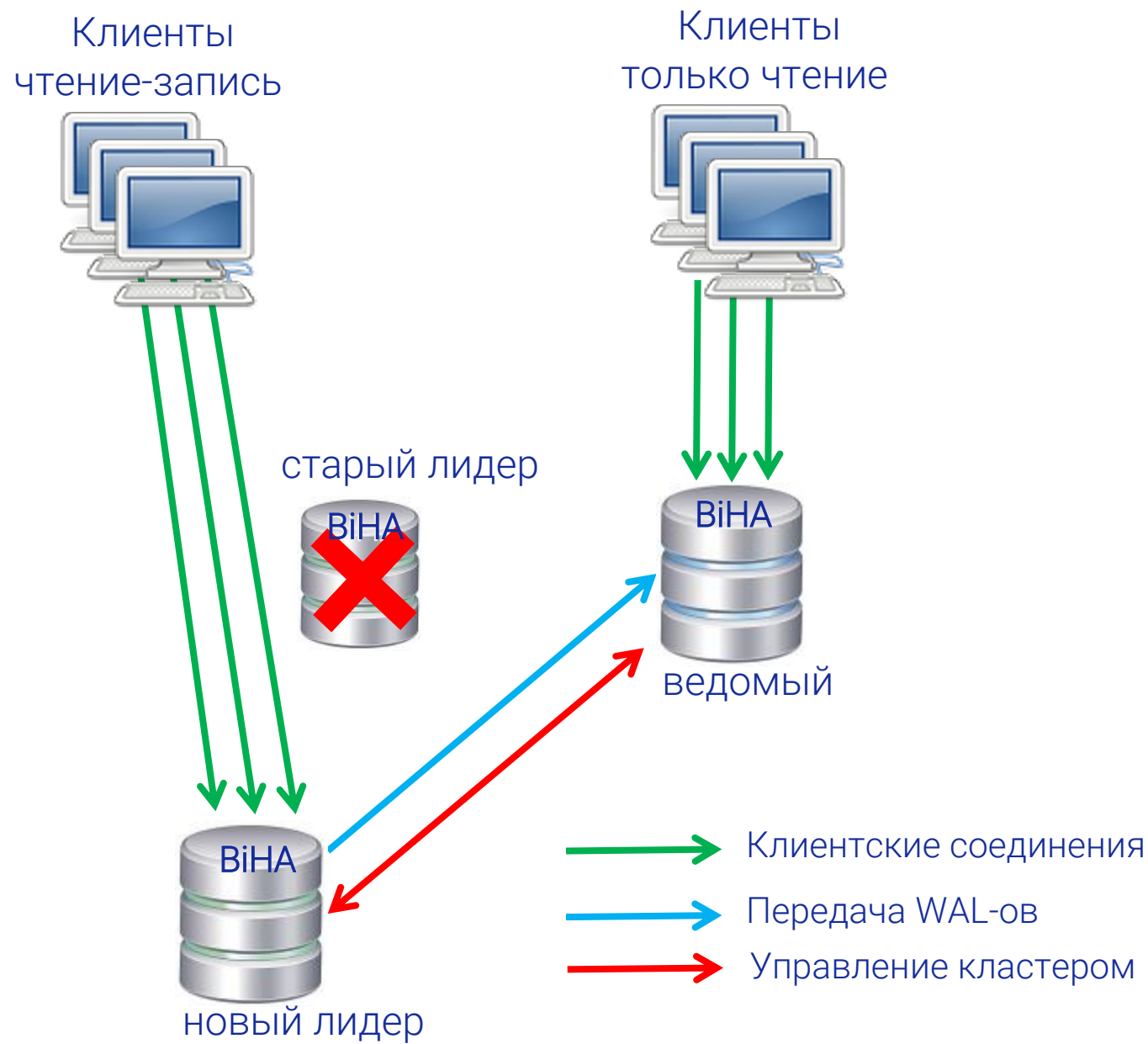
ВiНА: Управляющий канал

- 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.



BiNA: 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)

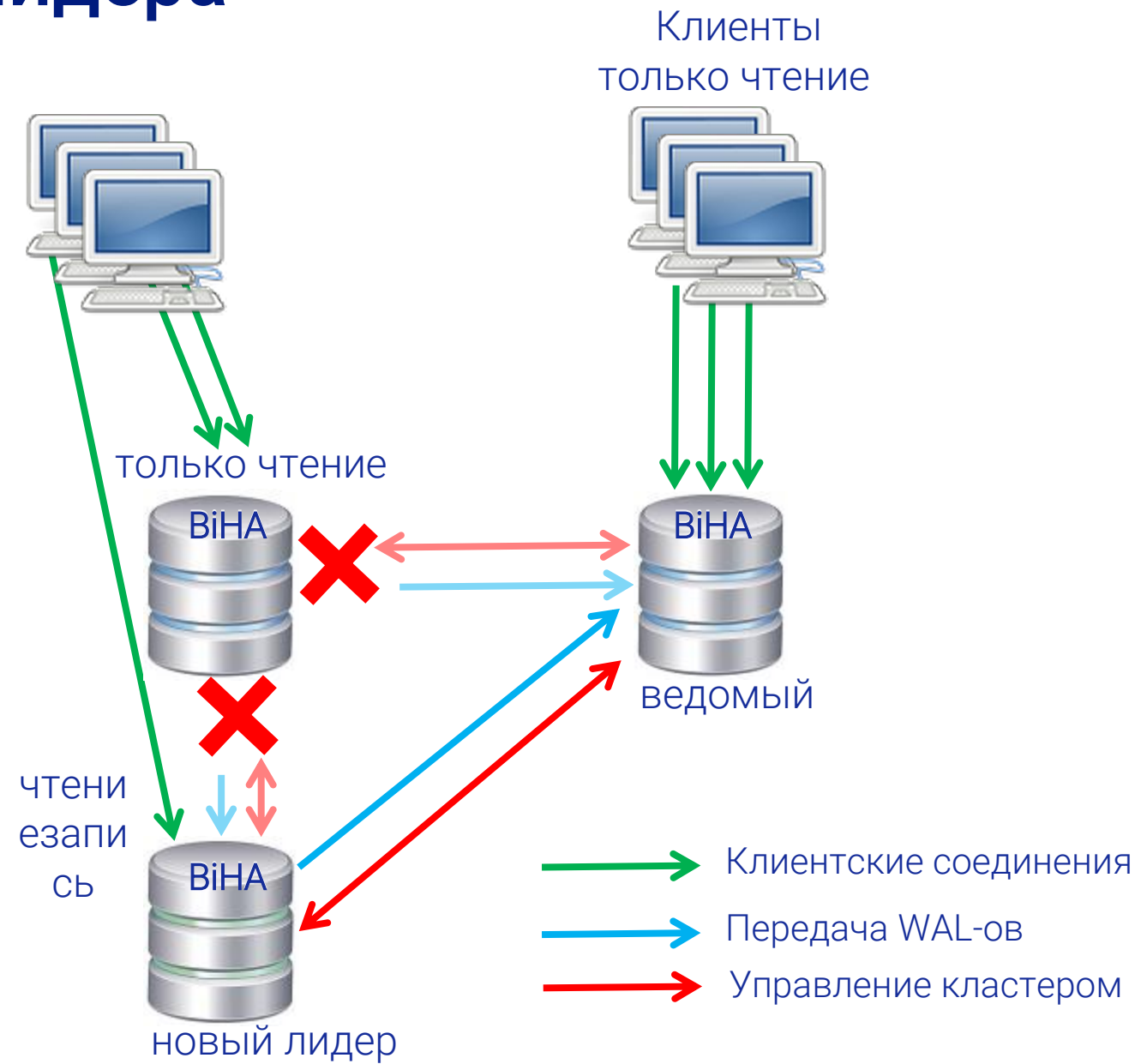


BiNA: Сетевая изоляция лидера

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

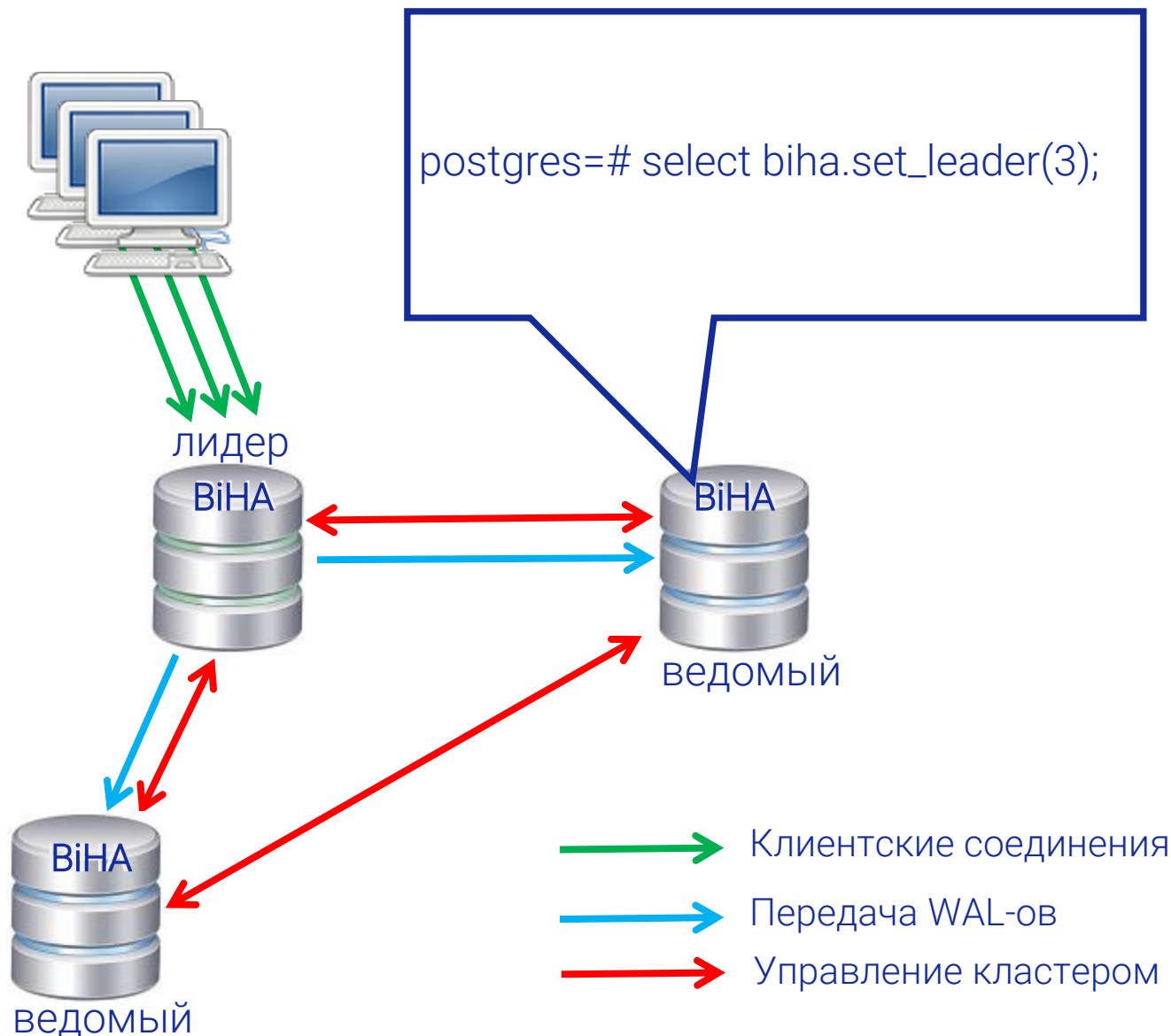
- либо когда восстановится соединение с недостающими узлами,
- либо когда администратор устранил сбой вручную.

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



BiHA: Назначение лидера вручную

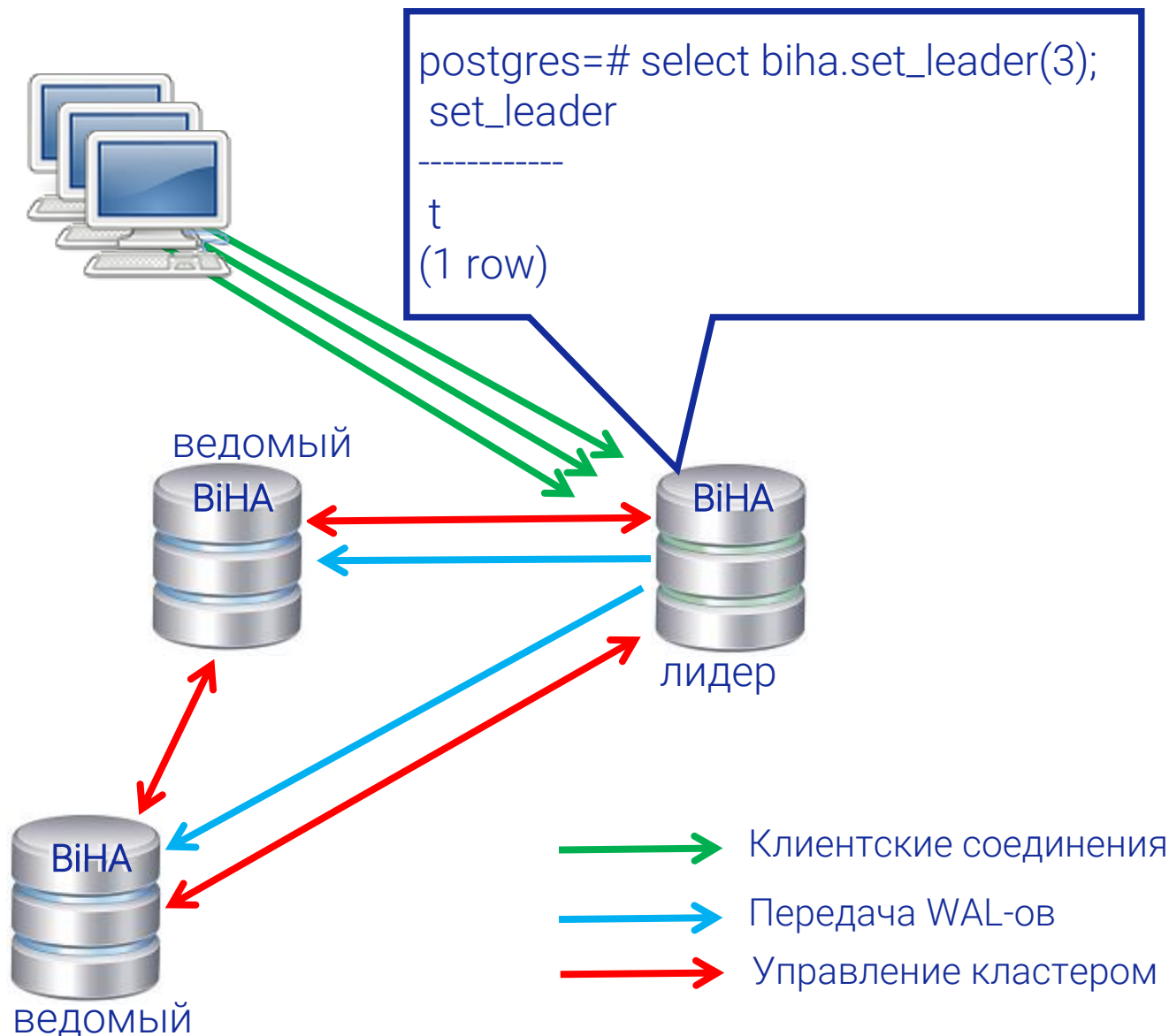
- для перевода лидера в режим обслуживания
- для назначения лидера на предпочтительный хост
- после возврата старого лидера



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



BiHA: встроенный отказоустойчивый кластер

Встроен в ядро Postgres Pro Enterprise начиная с версии 16

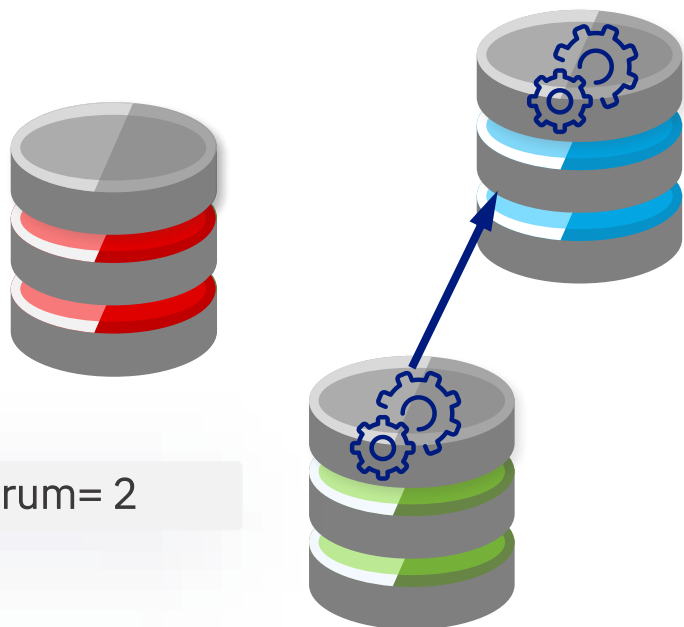
Простая установка и конфигурирование

Не требуется установка дополнительного ПО

Лидер продолжает работать, если соблюдается кворум

В случае сбоя ведомые предлагают себя кандидатами и организуются выборы нового лидера

Защита от split-brain



`biha.nquorum= 2`

В сети

Инфраструктура
Кластеры

Обзор

biha_cluster_1



ИЗМЕНИТЬ ТОПОЛОГИЮ



Менеджер отказоустойчивости Biha
Узлы, шт. 3
Версия PostgreSQL 17.5.1 enterprise
Сетевой адрес bpi-ppem23-biha-1.l.postgrespro.ru
Последнее обновление 24.09.2025 16:41:13

Редактировать



Экземпляр Тип	Узел выше	Статус узла	Режим репликации Статус ▾	Сетевой адрес Сервер	Приложение	Время отставания ▴ ▾		Воспроизвед.	Время ответа	Действия
Запись	Сброс									
instance-1 🌀 Лидер	—	✅ Запущен	— ● Активный	bpi-ppem23-biha-1.	—	—	—	—	—	🗑
instance-2 Последователь	instance-1	✅ Запущен	Асинхронный ● Активный	192.168.21.69	biha_node_2	0 0	0	0	24.09.2025 16:41	☆ 🗑
instance-4 Последователь	instance-1	✅ Запущен	Асинхронный ● Активный	192.168.20.206	biha_node_3	0 0	0	0	24.09.2025 16:41	☆ 🗑

← Назад

Настройка

Область

→ Далее

biha_cluster_1

Менеджер кластеризации PostgreSQL

Узелов: 3

Версия: PostgreSQL 15.1 enterprise

Консоль администратора: bpi-ppem23-biha-1

Имя инстанса	Тип	Статус	Роль	Консоль администратора
Instance 1	Лидер	Запущен	Активный	bpi-ppem23-biha-1
Instance 2	Последователь	Запущен	Активный	192.168.21.69
Instance 3	Последователь	Запущен	Активный	192.168.20.206

biha_cluster_1

Изменение топологии кластера

Лидер

instance-1

▼

Узлов в кластере, шт.

3

Последователи • 3 шт.

▲

Сервер

bpi-ppem23-biha-2.l.postgrespro.ru

Каталог данных /var/lib/pgpro/ent-17/data

Сетевой адрес 192.168.21.69

Порт 5432

🗑

Сервер

bpi-ppem23-biha-3.l.postgrespro.ru

Каталог данных /var/lib/pgpro/ent-17/data

Сетевой адрес 192.168.20.206

Порт 5432

🗑

Сервер*

Выберите

▼

Каталог данных /var/lib/pgpro/ent-17/data

Сетевой адрес —

Порт 5432

🗑

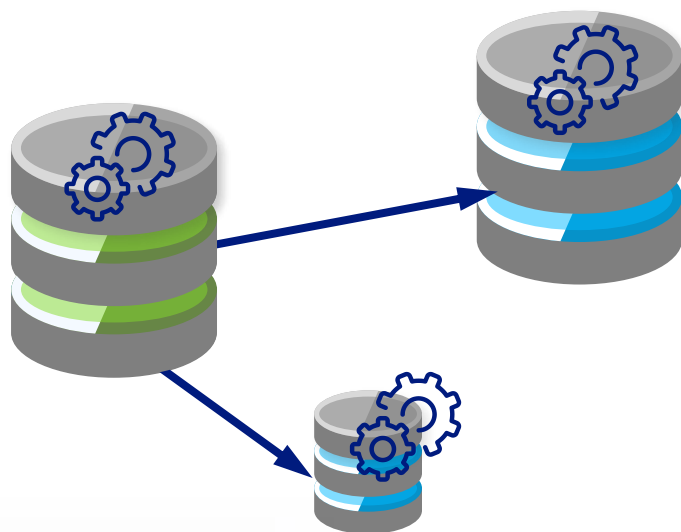
⊕ Добавить сервер

⚠ Для обеспечения кворума рекомендуем составлять кластер из нечетного числа узлов

ОТМЕНИТЬ ИЗМЕНЕНИЯ

СОХРАНИТЬ

ViHA: узел рефери



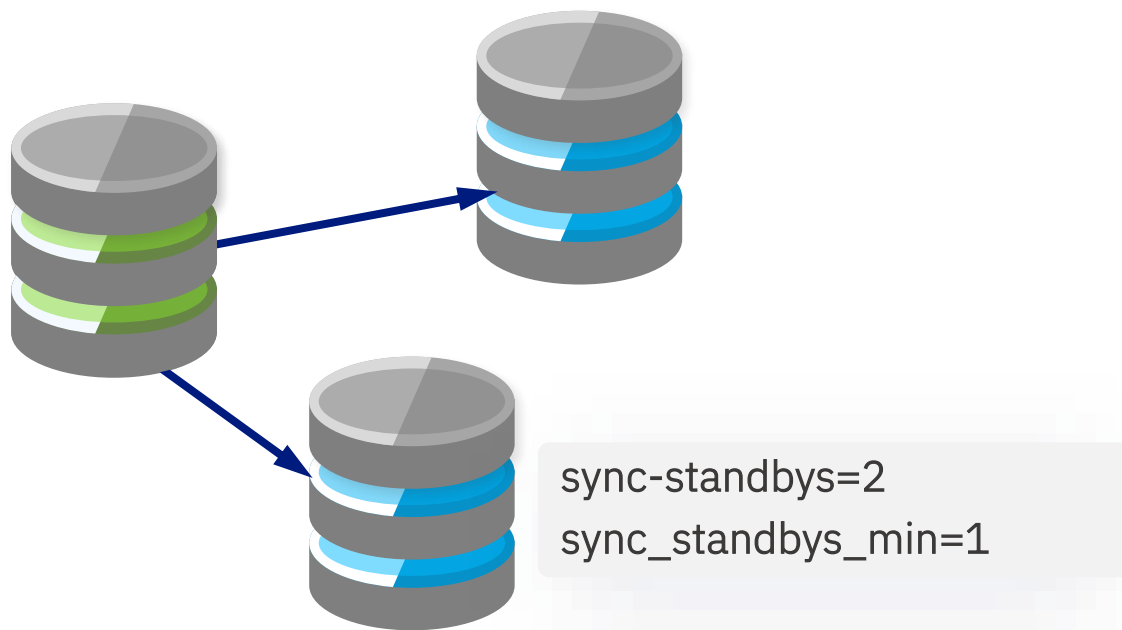
mode=
referee или
referee_with_wal

Рефери – легковесный экземпляр, который не содержит пользовательских данных, но является членом кластера ViHA: сам кандидатом на звание нового лидера не выступает, но участвует в голосовании

Рефери может работать в режиме `referee_with_wal` :

- Получает весь WAL и фильтрует его, применяются только системные записи WAL без пользовательских данных
- При сбое лидера может отправлять WAL кандидату на нового лидера, если тот отстаёт от рефери

BiHA: синхронный режим

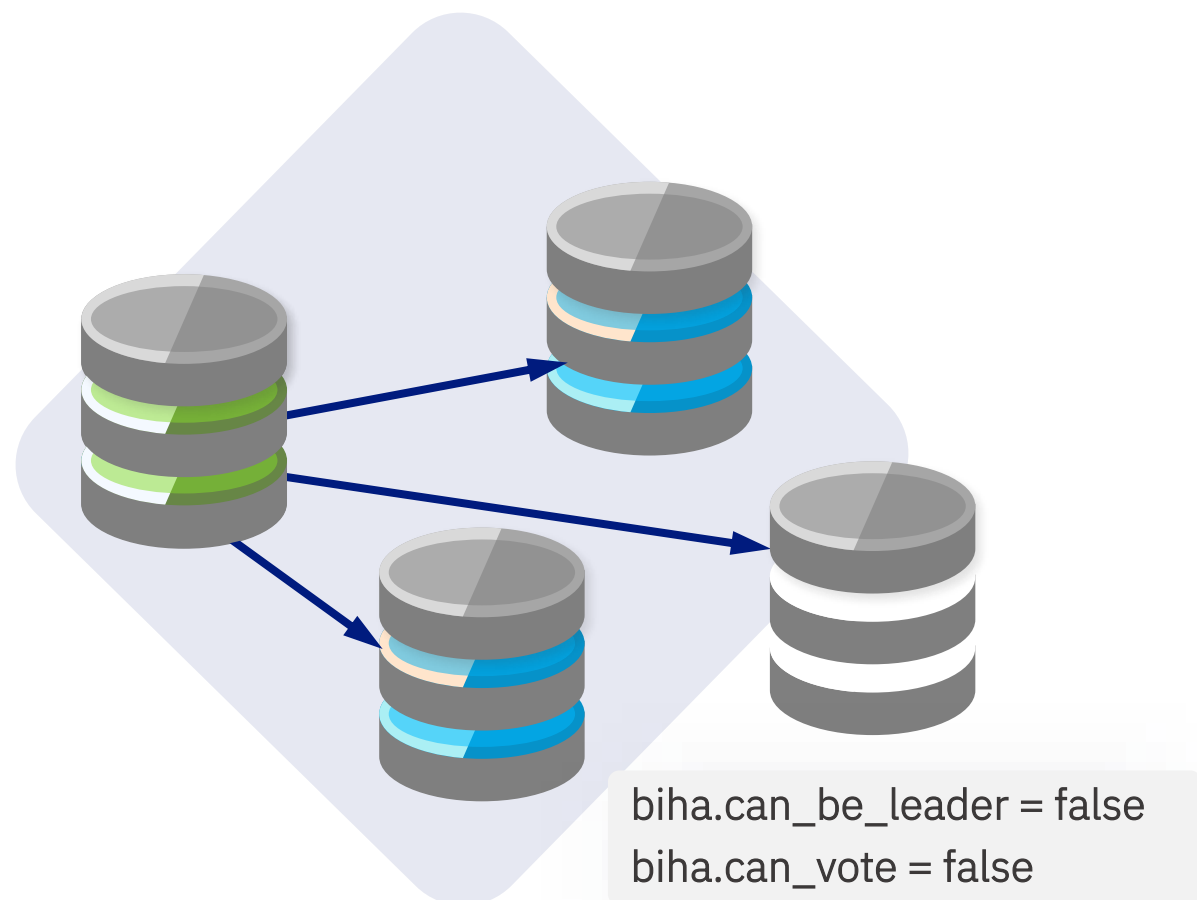


В Shardman отказоустойчивый кластер BiHA работает в синхронном режиме

Под капотом выполняется
`bihactl init ... --sync-standbys=
<число_синхронных_ведомых_серверов>`

Кроме того, настраивается параметр `sync_standbys_min` - минимальное количество синхронных ведомых серверов, которые должны быть доступны, чтобы лидер продолжал работу.

BiHA: узел, который не может стать лидером

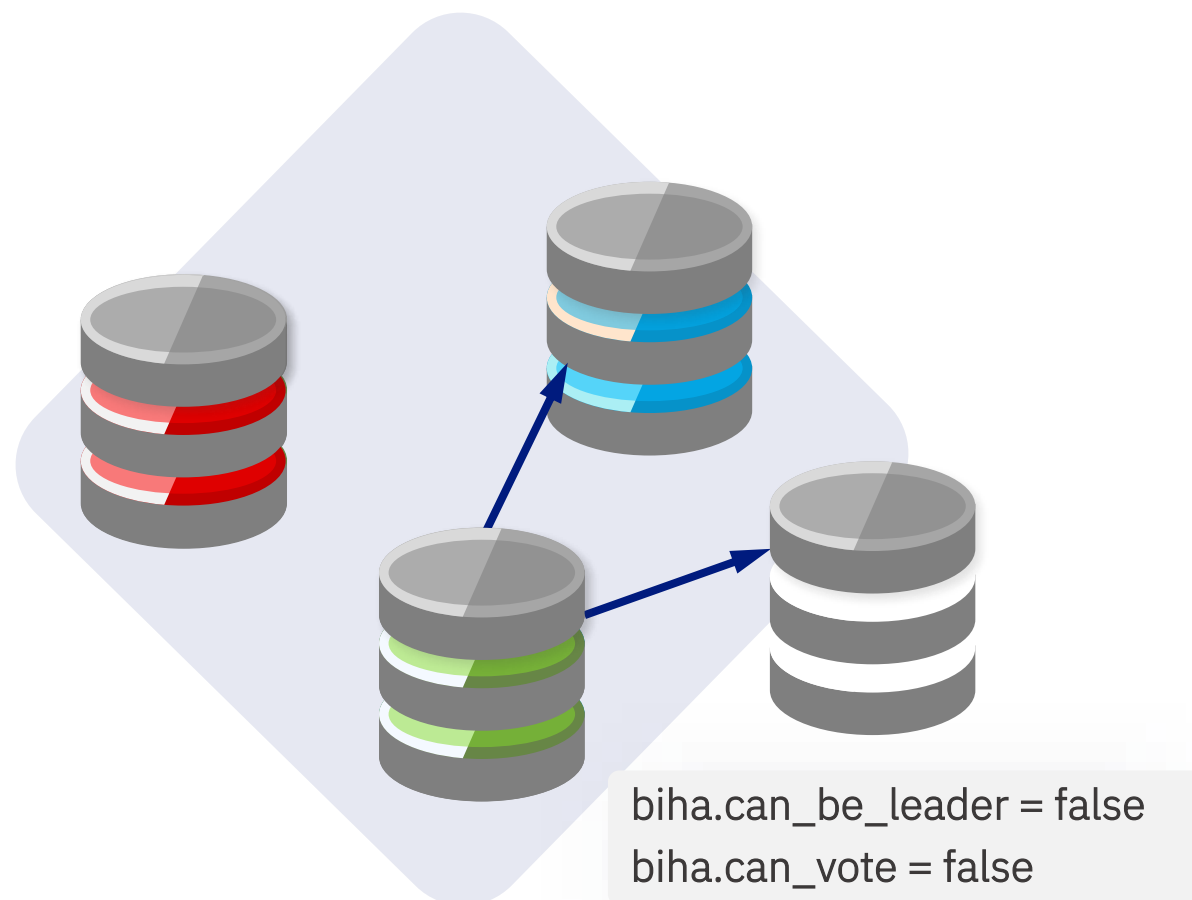


Если вы считаете, что один из серверов по какой-то причине не должен никогда стать лидером, например:

- недостаточно производительный,
- подключён к другой сети,
- включено отложенное применение WAL,

то можно запретить узлу становиться кандидатом и даже голосовать

BiHA: узел, который не может стать лидером



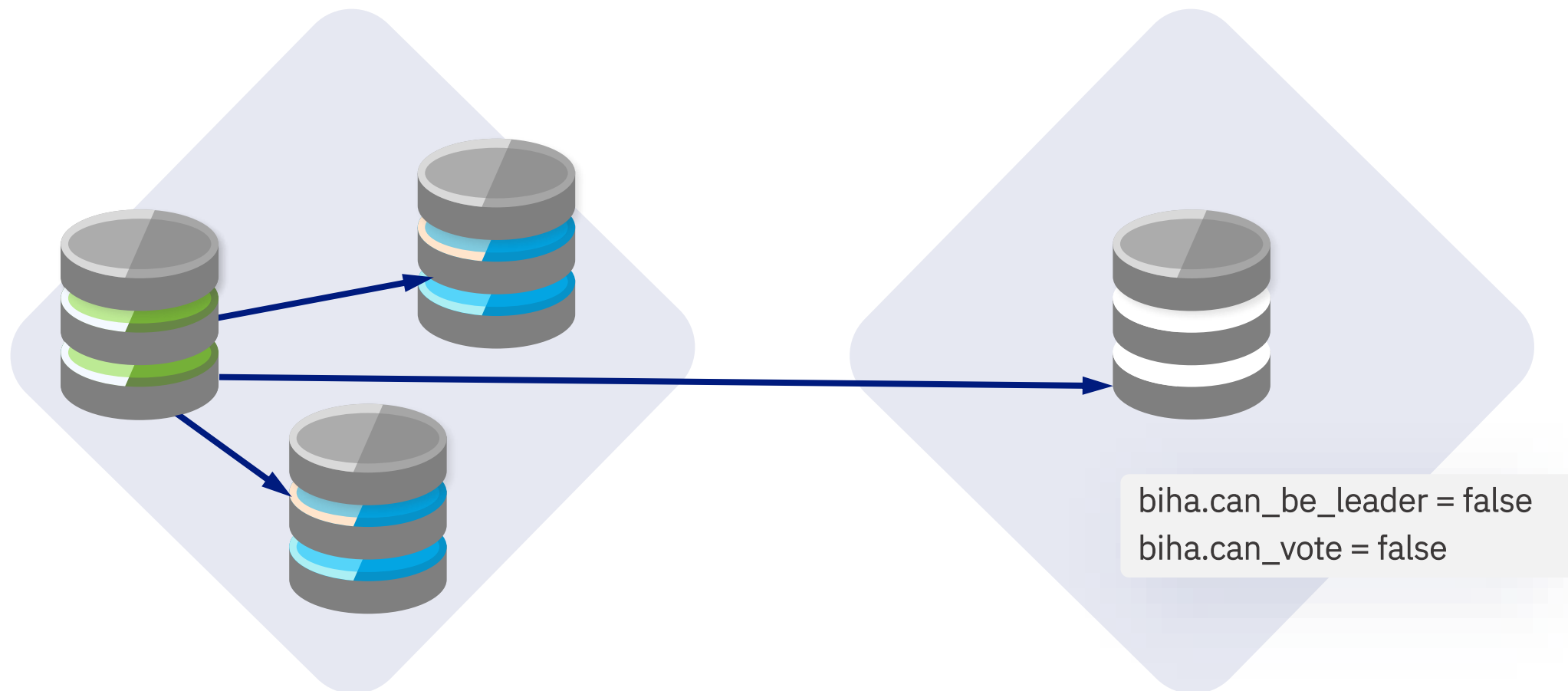
Если вы считаете, что один из серверов по какой-то причине не должен никогда стать лидером, например:

- недостаточно производительный,
- подключён к другой сети,
- включено отложенное применение WAL,

то можно запретить узлу становиться кандидатом и даже голосовать

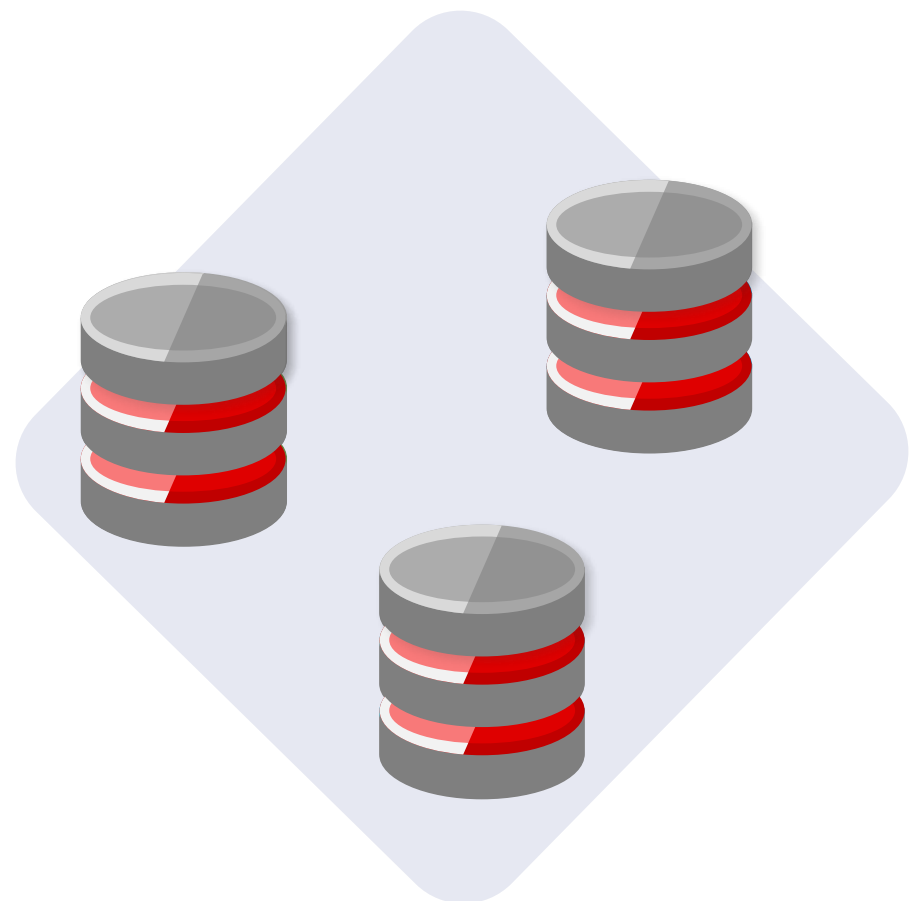
BiHA: узел в отдельном ЦОД

Лидером может стать только узел в основном ЦОД

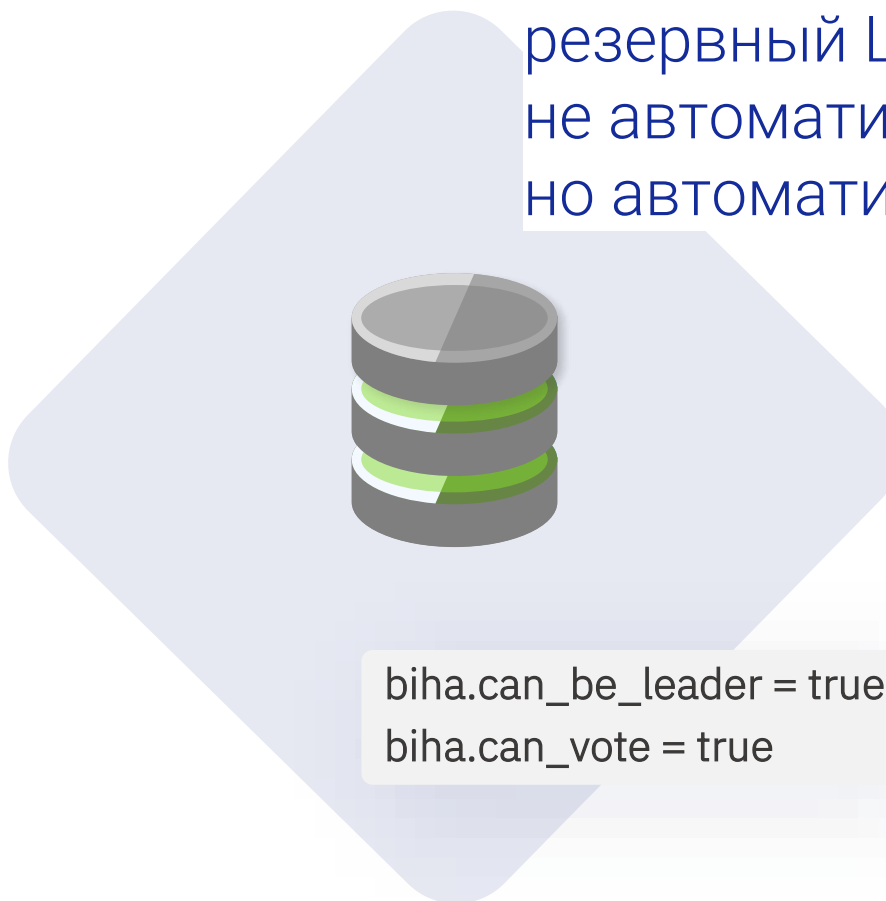


BiHA: узел в отдельном ЦОД

Сбой всего ЦОД не приведёт к потере данных



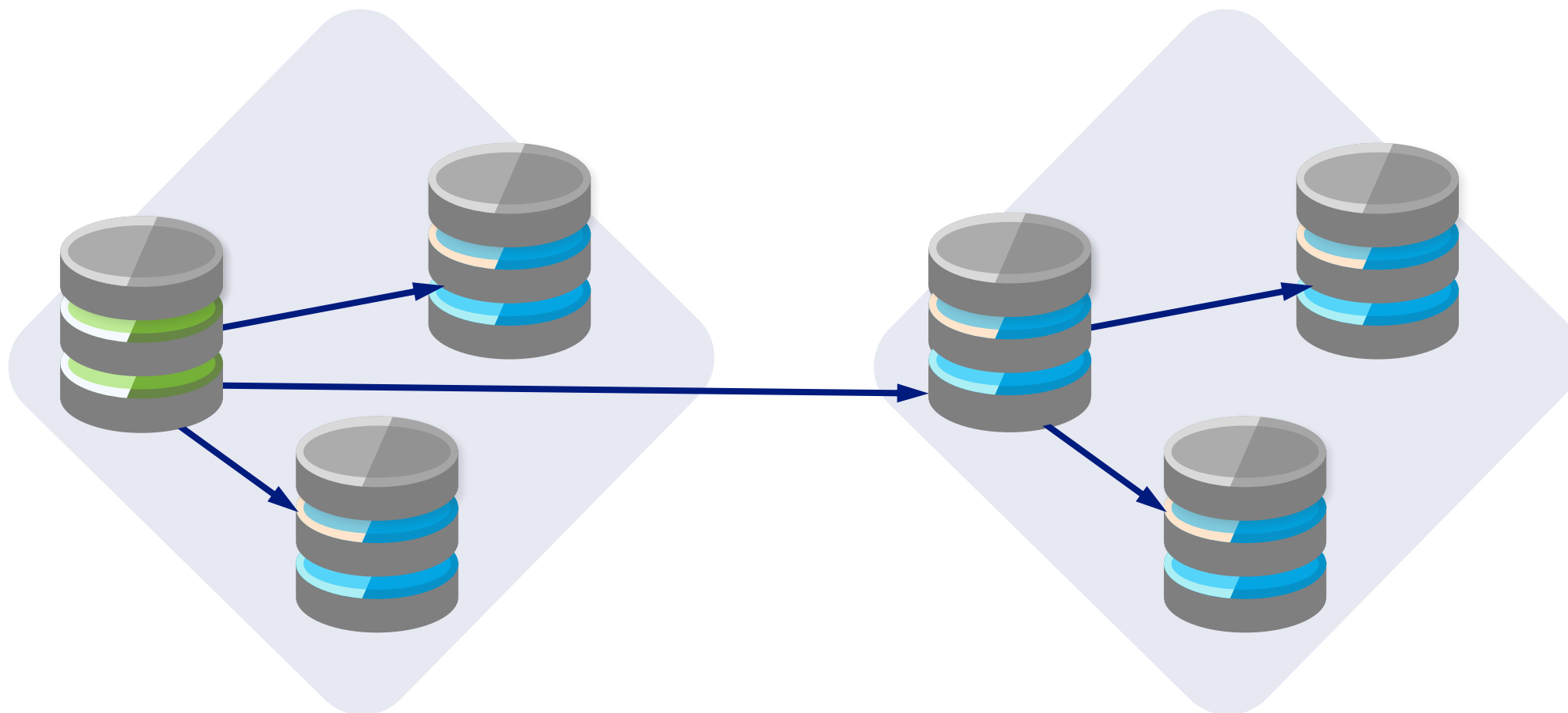
переключение на резервный ЦОД не автоматическое, но автоматизированное



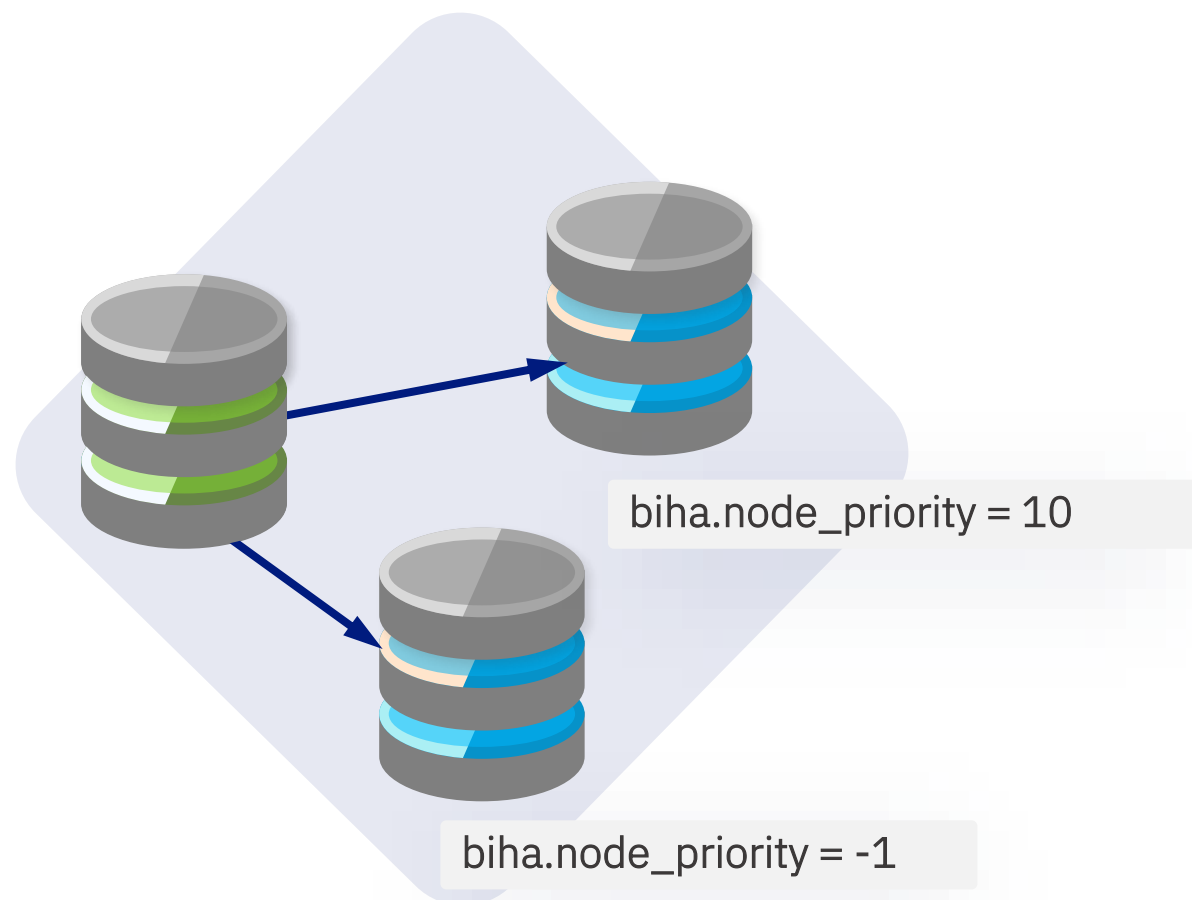
```
biha.can_be_leader = true  
biha.can_vote = true
```

ВiНА: катастрофоустойчивость

Катастрофоустойчивость – основное направление развития в 2025 году



ВiHA: приоритеты кандидатов в лидеры

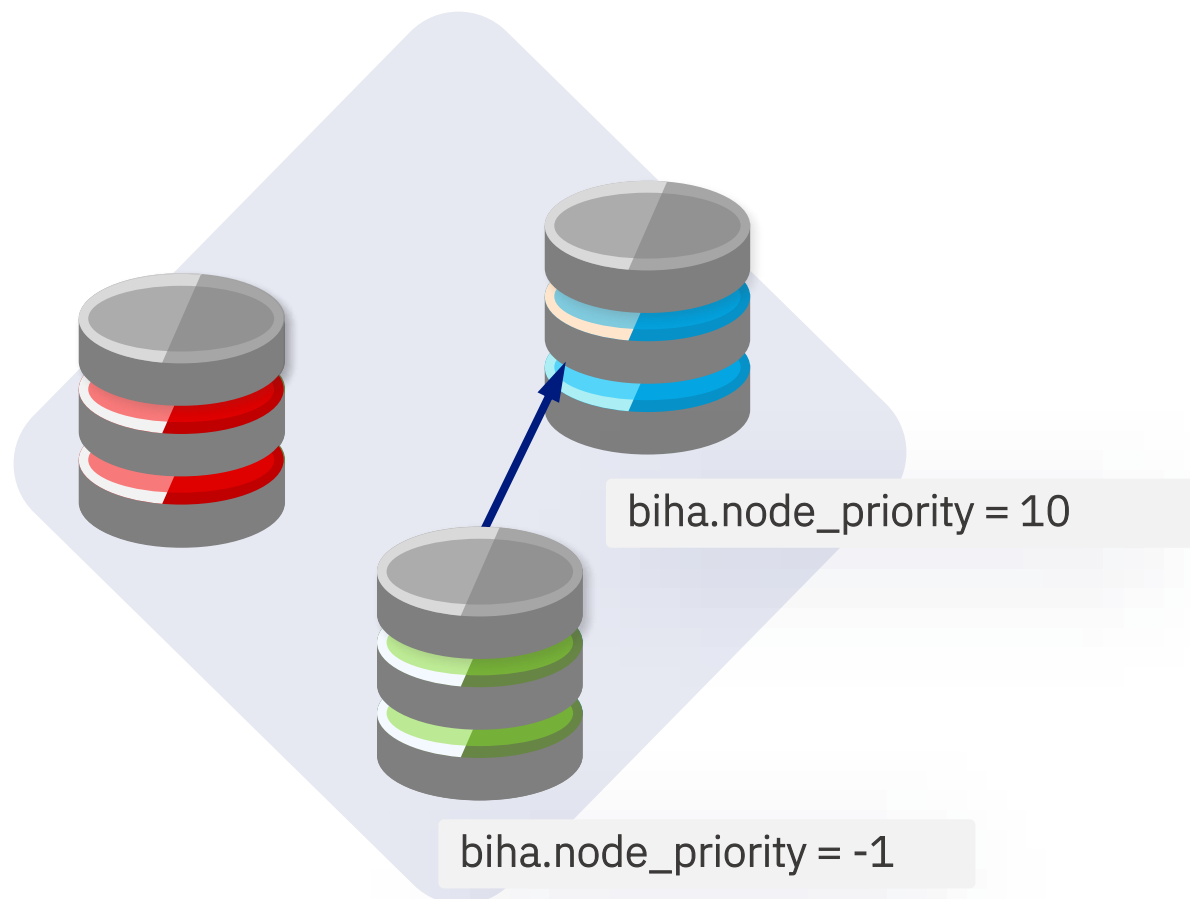


Можно установить приоритеты для выдвижения узла в качестве кандидата

Приоритет узла в секундах определяет тайм-аут, по достижении которого узел предложит себя в качестве кандидата на выборах.

Только в синхронном кластере

ВiНА: приоритеты кандидатов в лидеры

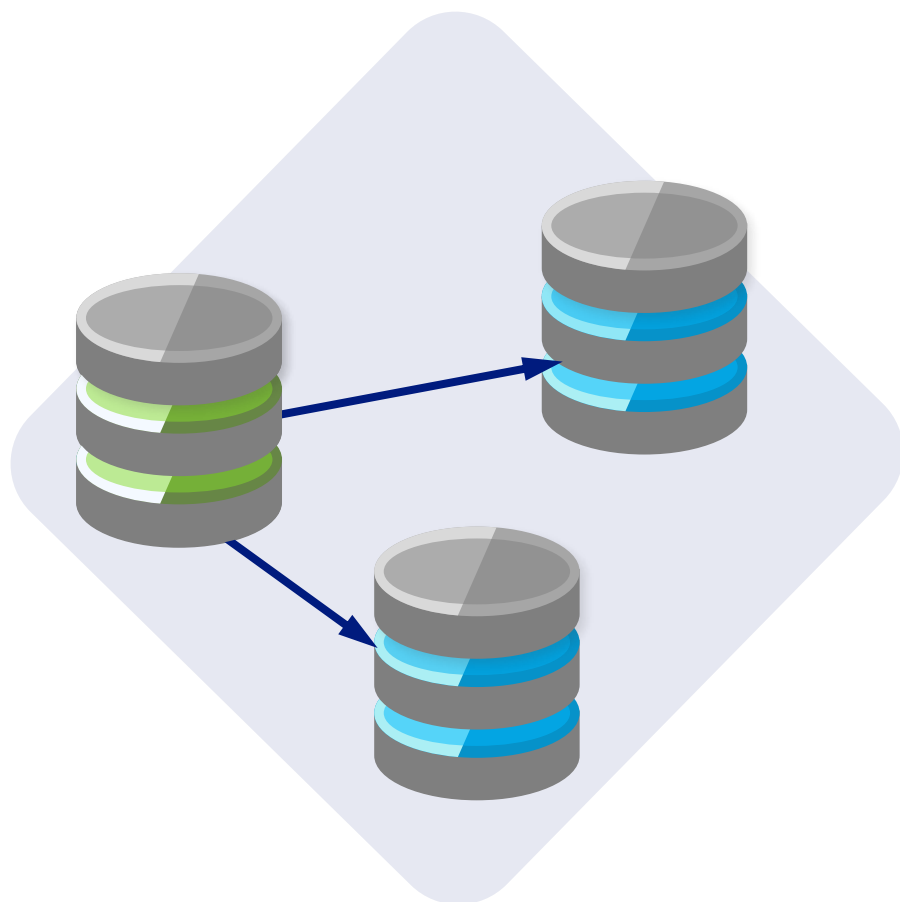


Можно установить приоритеты для выдвижения узла в качестве кандидата

Приоритет узла в секундах определяет тайм-аут, по достижении которого узел предложит себя в качестве кандидата на выборах.

Только в синхронном кластере

ВіНА : Функции-обработчики смены состояний



Вы создаёте SQL-функцию и регистрируете её в качестве обработчика. SQL-функция может уведомлять внешние сервисы о событиях в ВіНА-кластере.

Например:

- `CANDIDATE_TO_LEADER` вызывается на узле, выбранном в качестве нового лидера.
- `LEADER_CHANGED` вызывается на каждом узле ВіНА-кластера при смене лидера.

BiHA : Функции-обработчики смены состояний



Вы создаёте SQL-функцию и регистрируете её в качестве обработчика. SQL-функция может уведомлять внешние сервисы о событиях в BiHA-кластере.

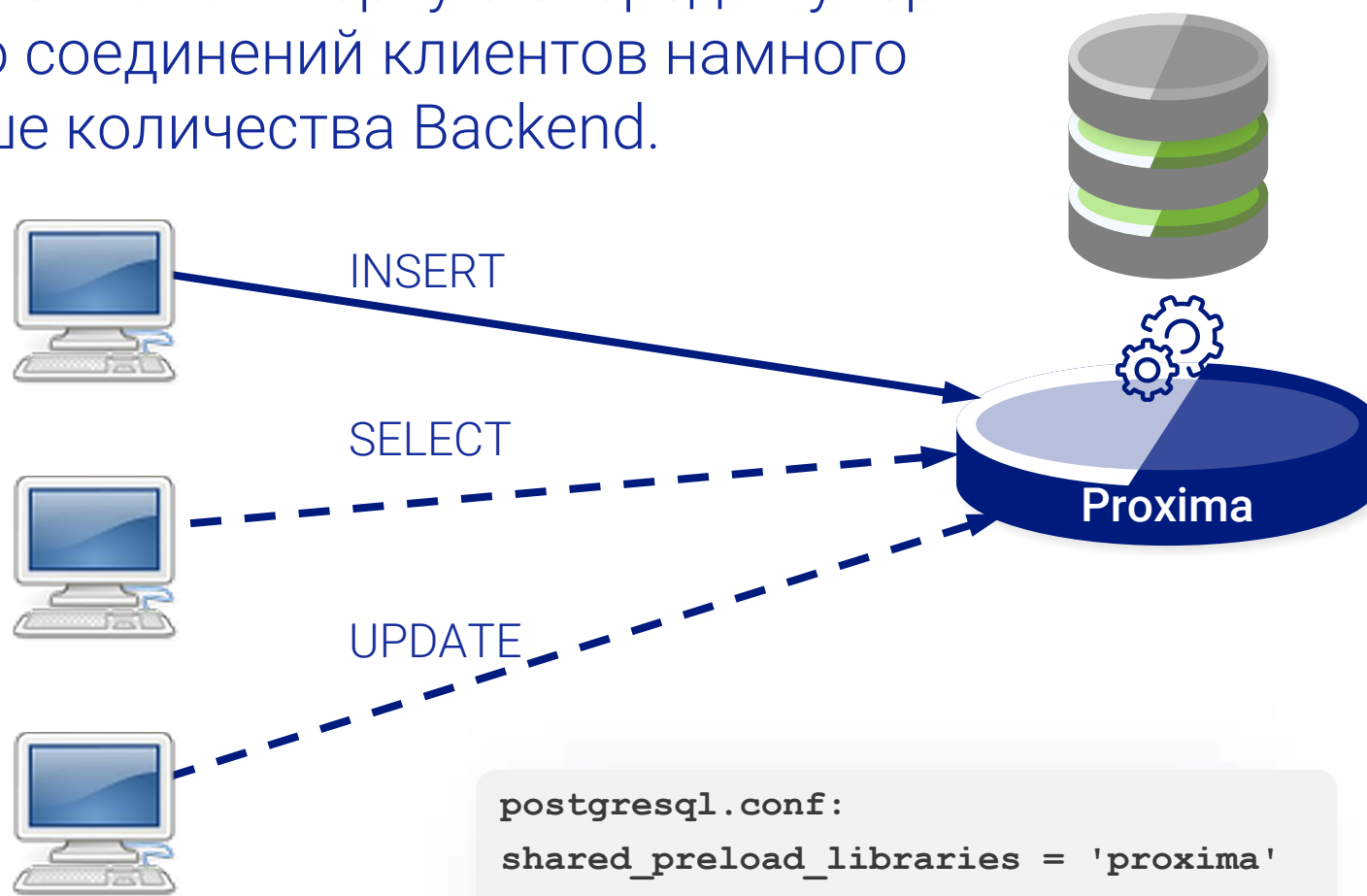
Например:

- `CANDIDATE_TO_LEADER` вызывается на узле, выбранном в качестве нового лидера.
- `LEADER_CHANGED` вызывается на каждом узле BiHA-кластера при смене лидера.

Если исполнение функции-обработчика длится дольше, чем `biha.callbacks_timeout`, BiHA останавливает исполнение и продолжает работать в обычном режиме.

Proxima

Proxima – это в первую очередь пулер.
Число соединений клиентов намного больше количества Backend.



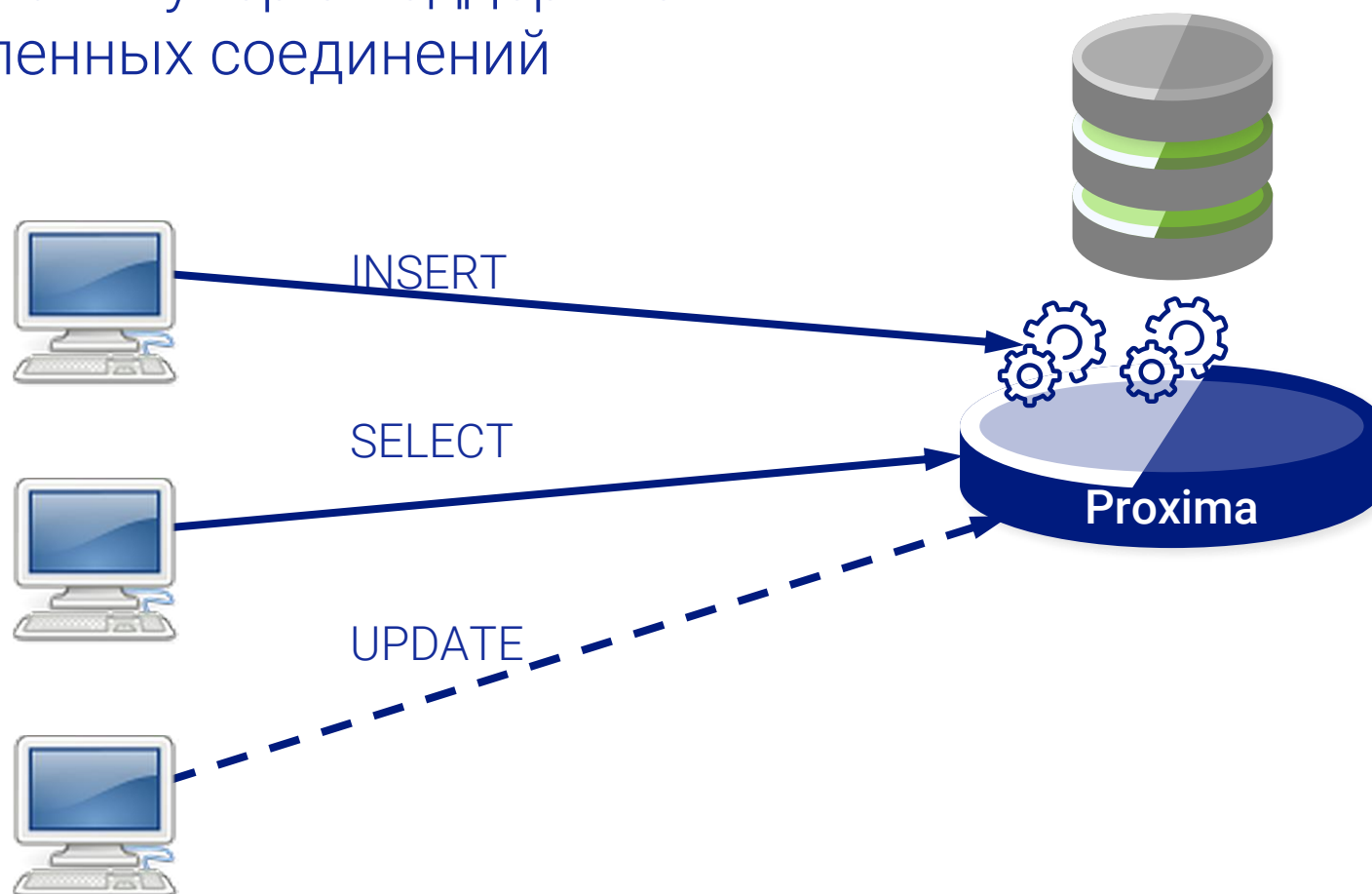
```
postgresql.conf:  
shared_preload_libraries = 'proxima'  
proxima.cluster_mode = 'standalone'
```

Backend пул - компонент Proxima - множество соединений с Backend процессами Postgres.

Каждый свободный Backend может быть использован для выполнения транзакции клиента, после чего соединение возвращается в pool в качестве свободного.

Proxima

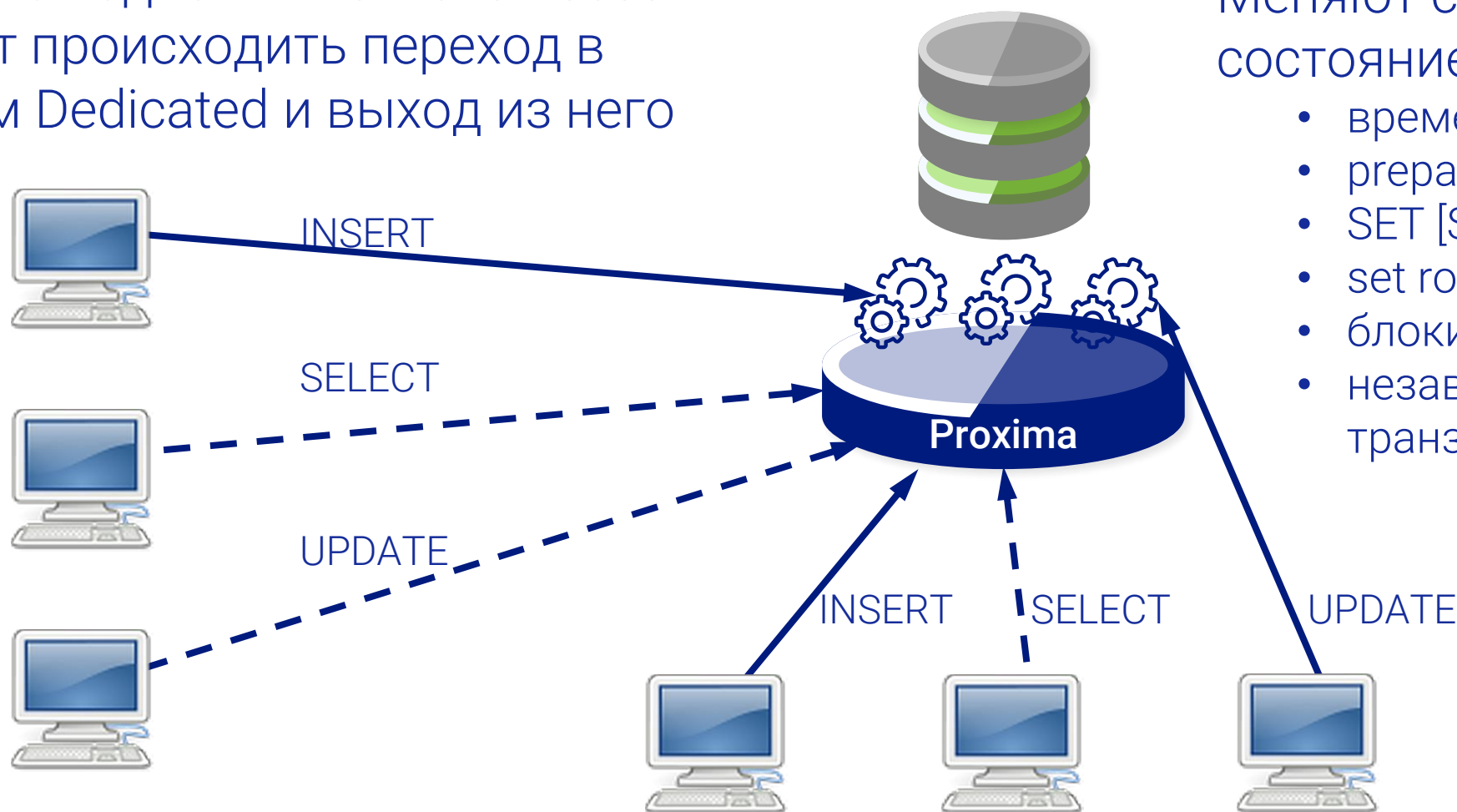
Proxima – пулер с поддержкой выделенных соединений



Dedicated режим:
если клиент создает
в процессе
своей работы
сессионный объект,
то дальнейшую
работу с ним
он проводит
через конкретный
Backend процесс.

Proxima

В рамках одной клиентской сессии может происходить переход в режим Dedicated и выход из него

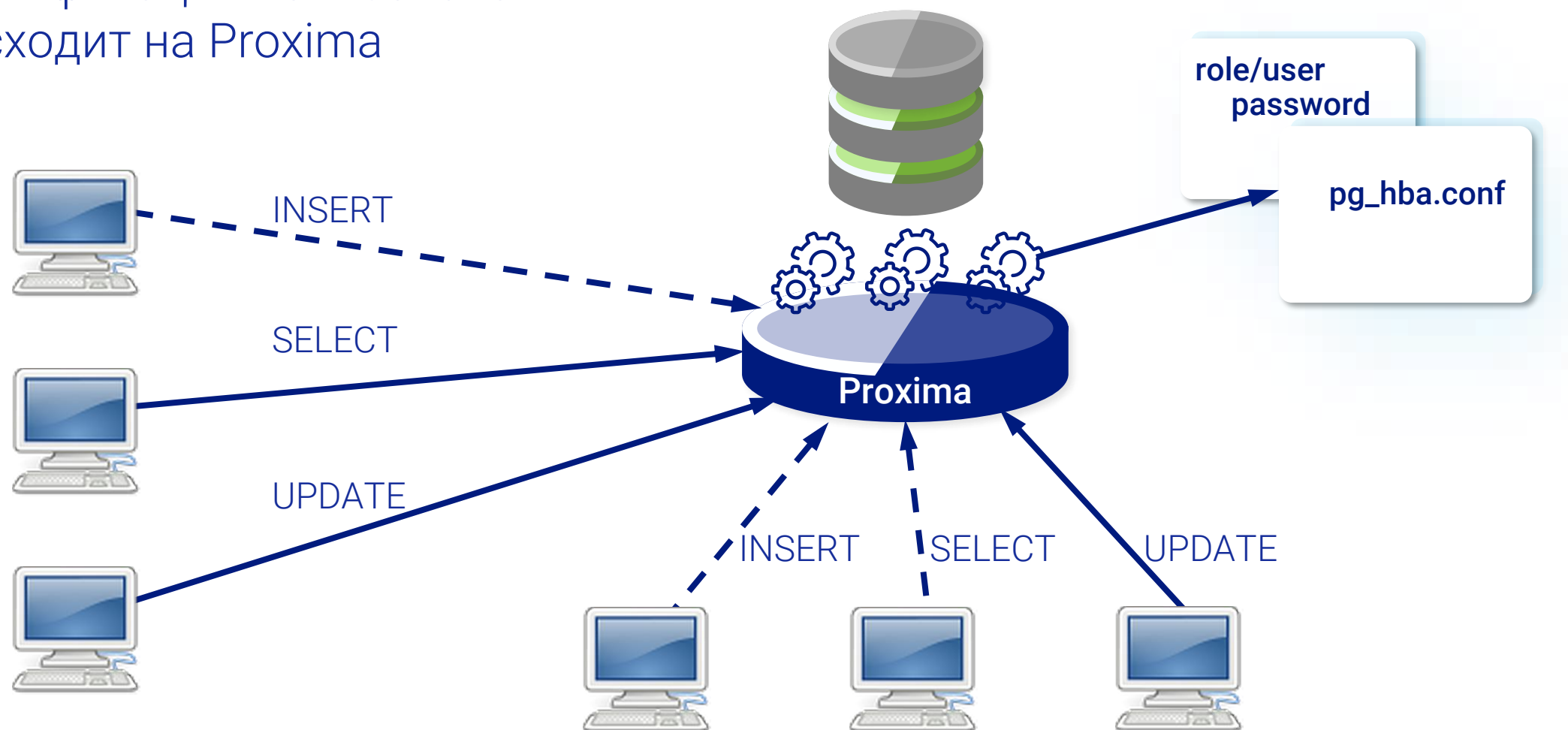


Меняют сессионное состояние:

- временные таблицы,
- prepare statement ,
- SET [SESSION]
- set role,
- блокировки ,
- незавершённые транзакции и т.п.

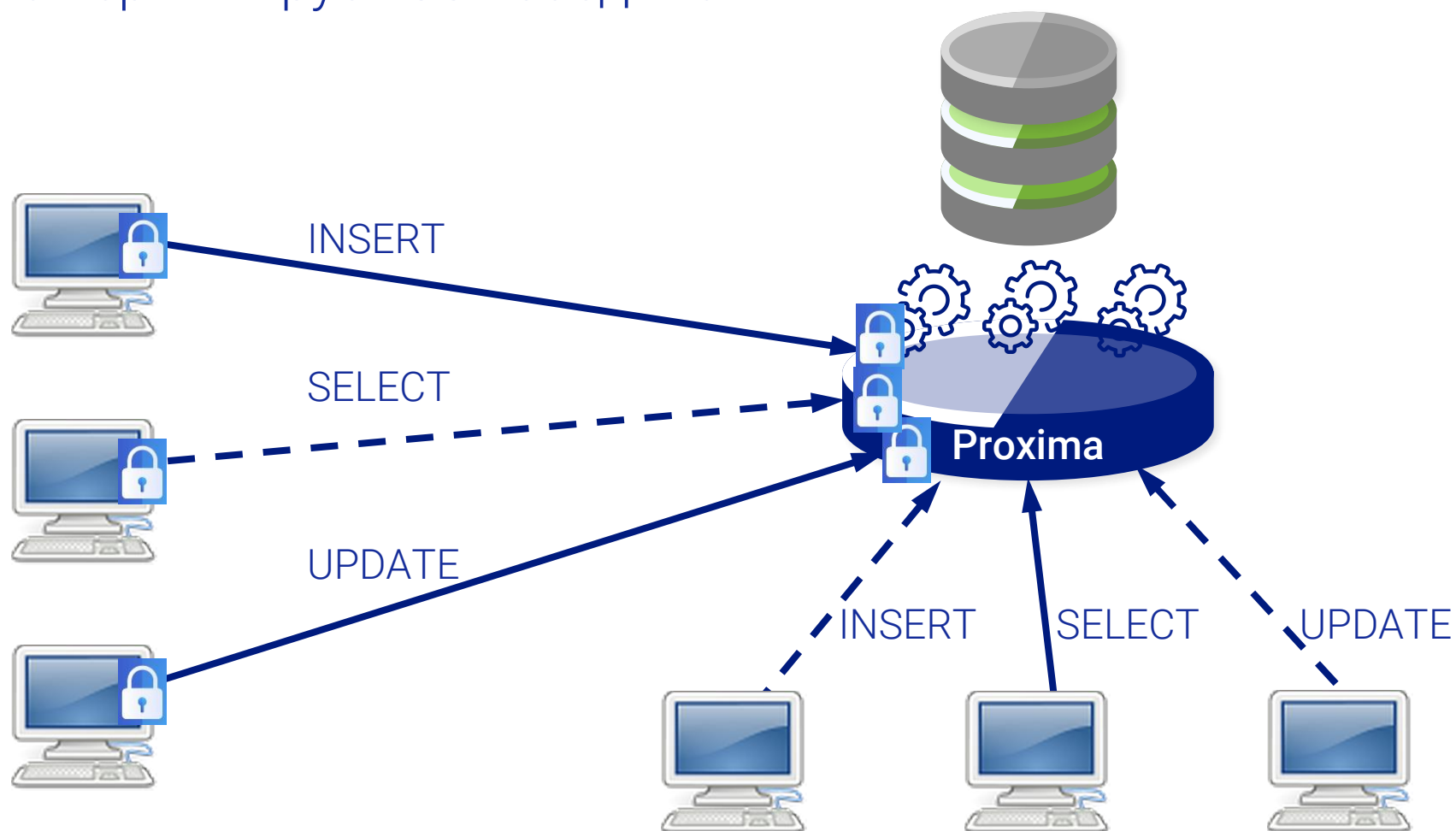
Proxima

Аутентификация пользователя
происходит на Proxima



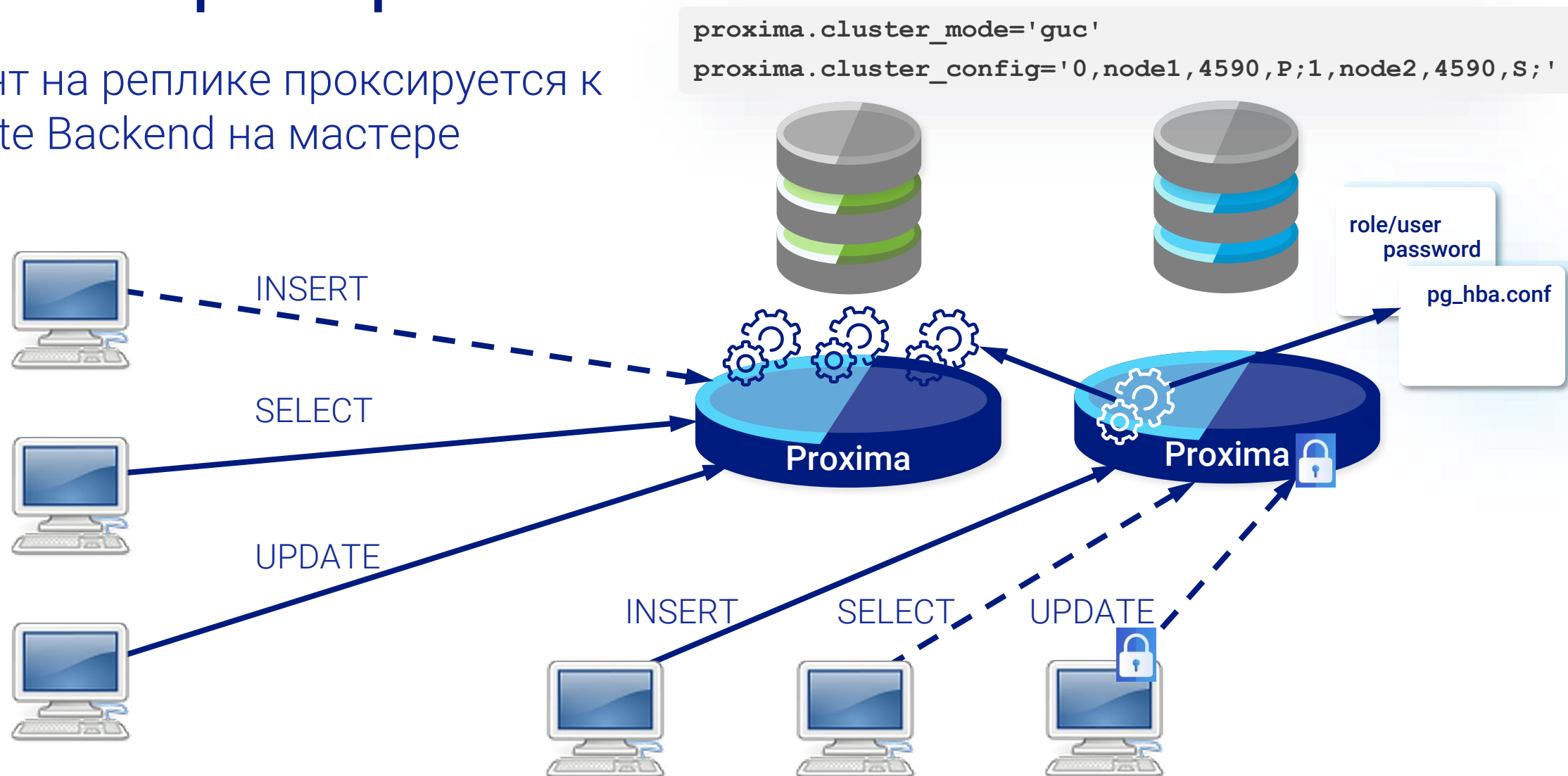
Proxima

Proxima терминирует SSL соединения



Proxima: проксирование

Клиент на реплике проксируется к Remote Backend на мастере



Proxima + BiHA

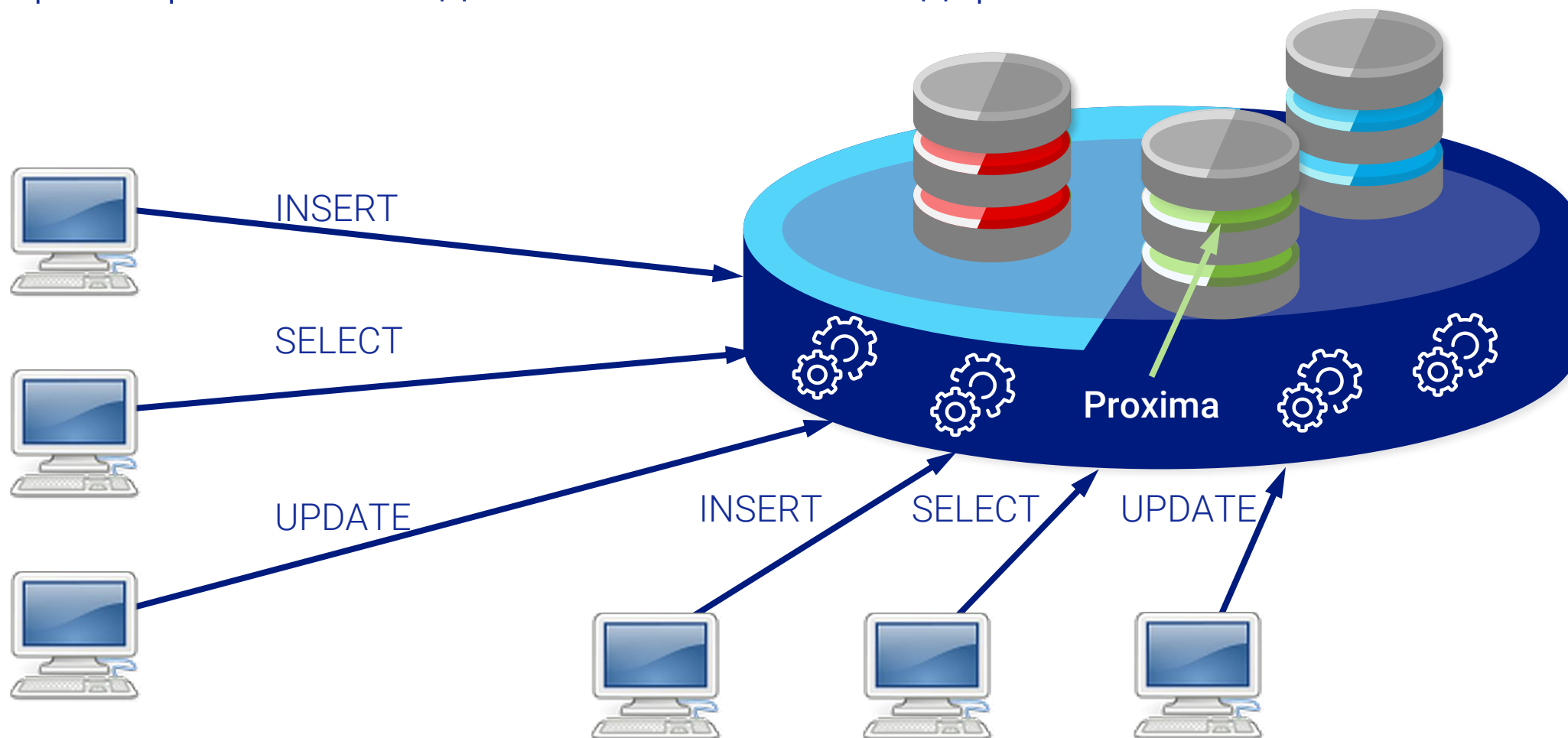
Любой узел может быть точкой входа в кластер
Определение лидера BiHA происходит на лету

```
postgresql.conf:  
shared_preload_libraries = 'proxima'  
proxima.cluster_mode = 'biha'
```



Proxima + BiHA

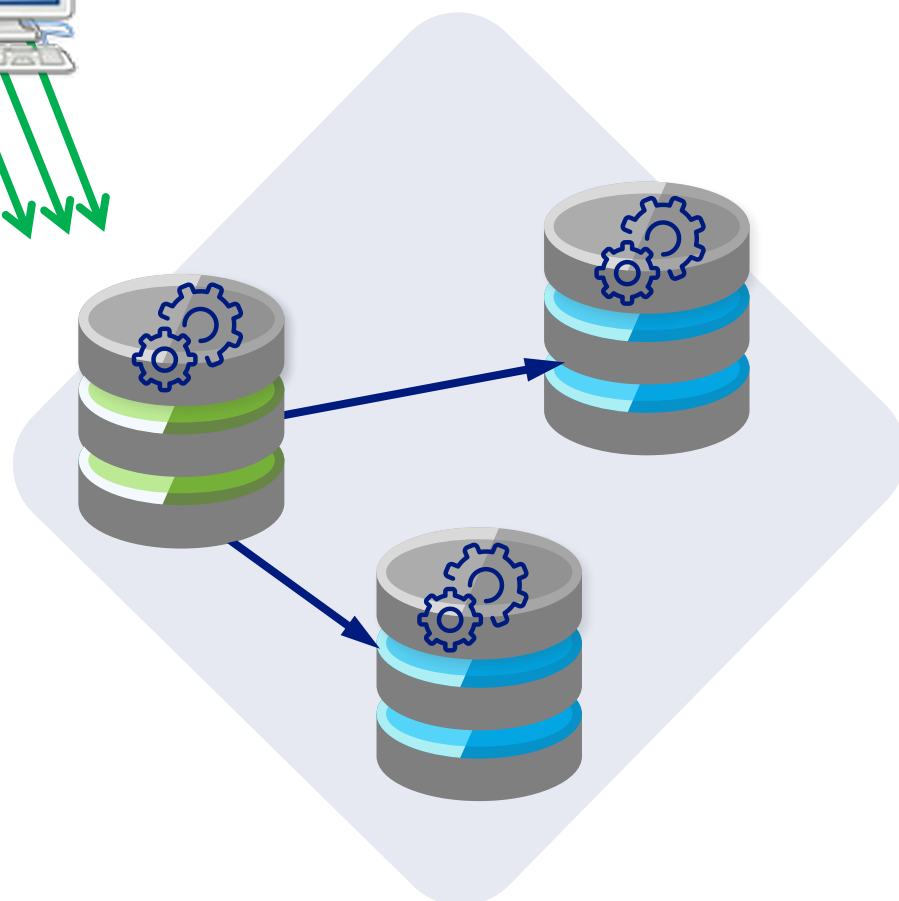
Proxima перенаправляет соединения на новый лидер BiHA



Автоматическое переключение соединения на стороне клиента на новый мастер



*postgresql://host1:5432,host2:5432,host3:5432/mydb?
target_session_attrs=read-write*



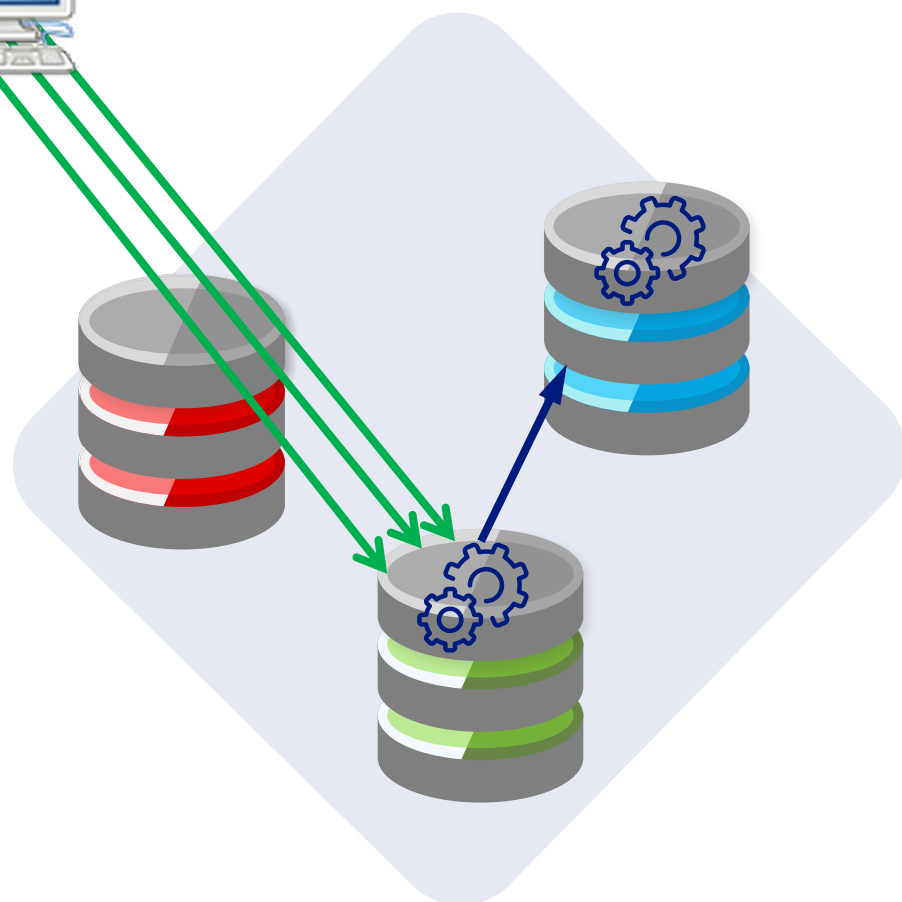
На клиенте (libpq, JDBC) можно перечислить все узлы кластера,

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

Автоматическое переключение соединения на стороне клиента на новый мастер



`postgresql://host1:5432,host2:5432,host3:5432/mydb?
target_session_attrs=read-write`



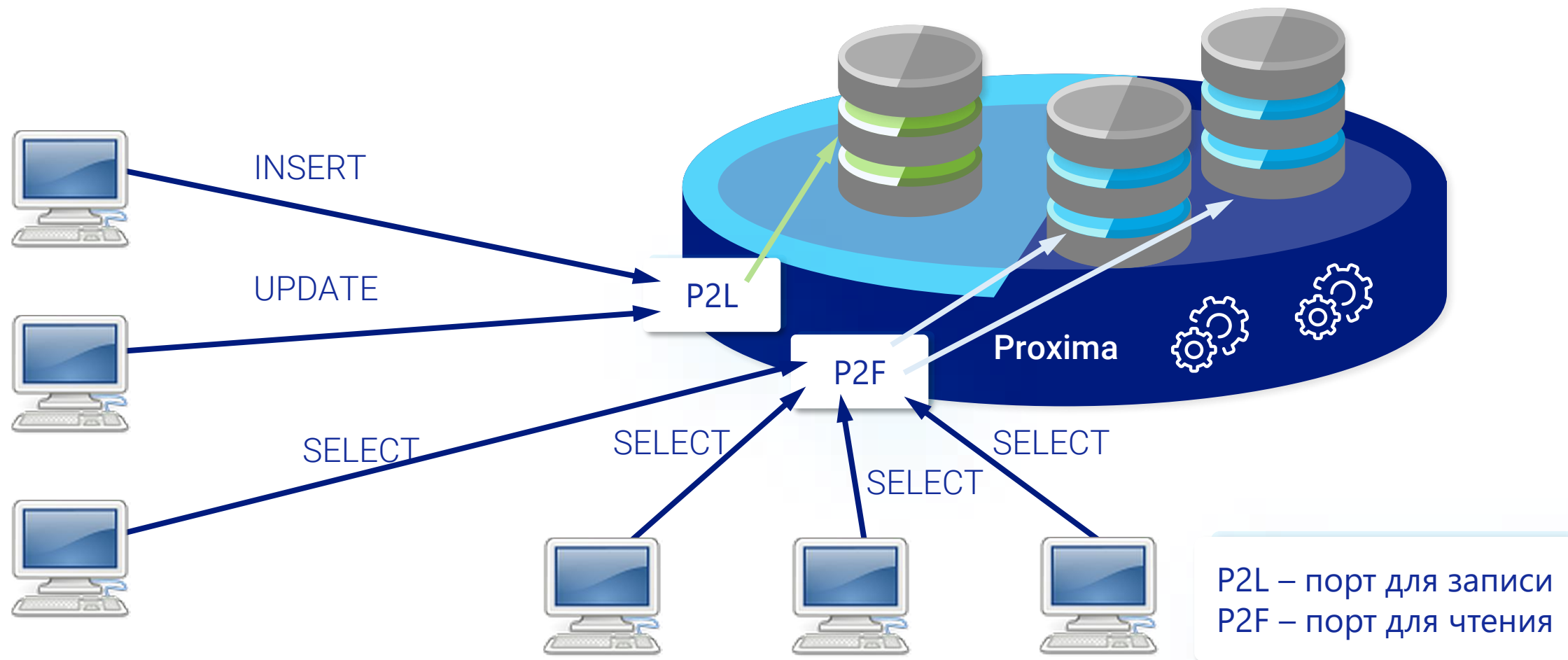
На клиенте (libpq, JDBC) можно перечислить все узлы кластера,

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

При сбое узла клиент автоматически подключится к новому лидеру

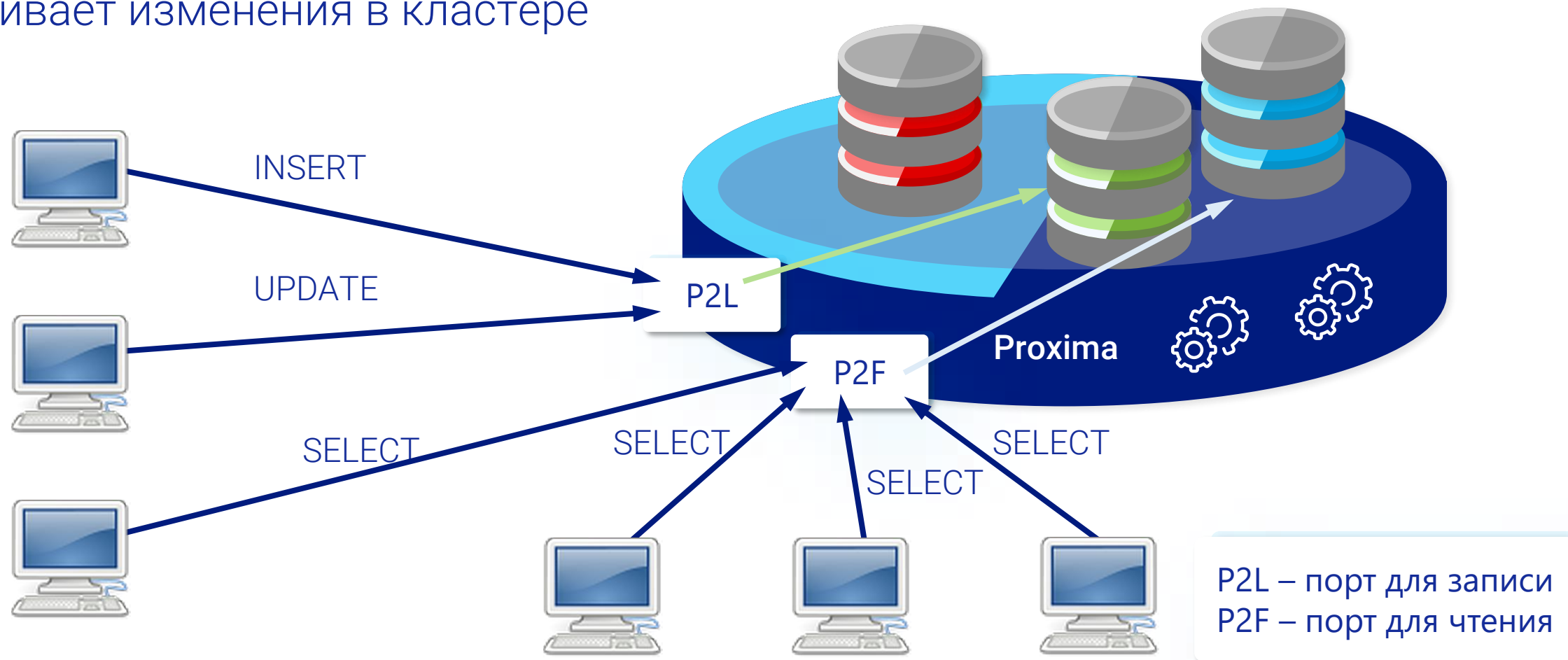
Proxima 17.5 : балансировка читающей нагрузки

Proxima соединения к P2F перенаправляет на реплики

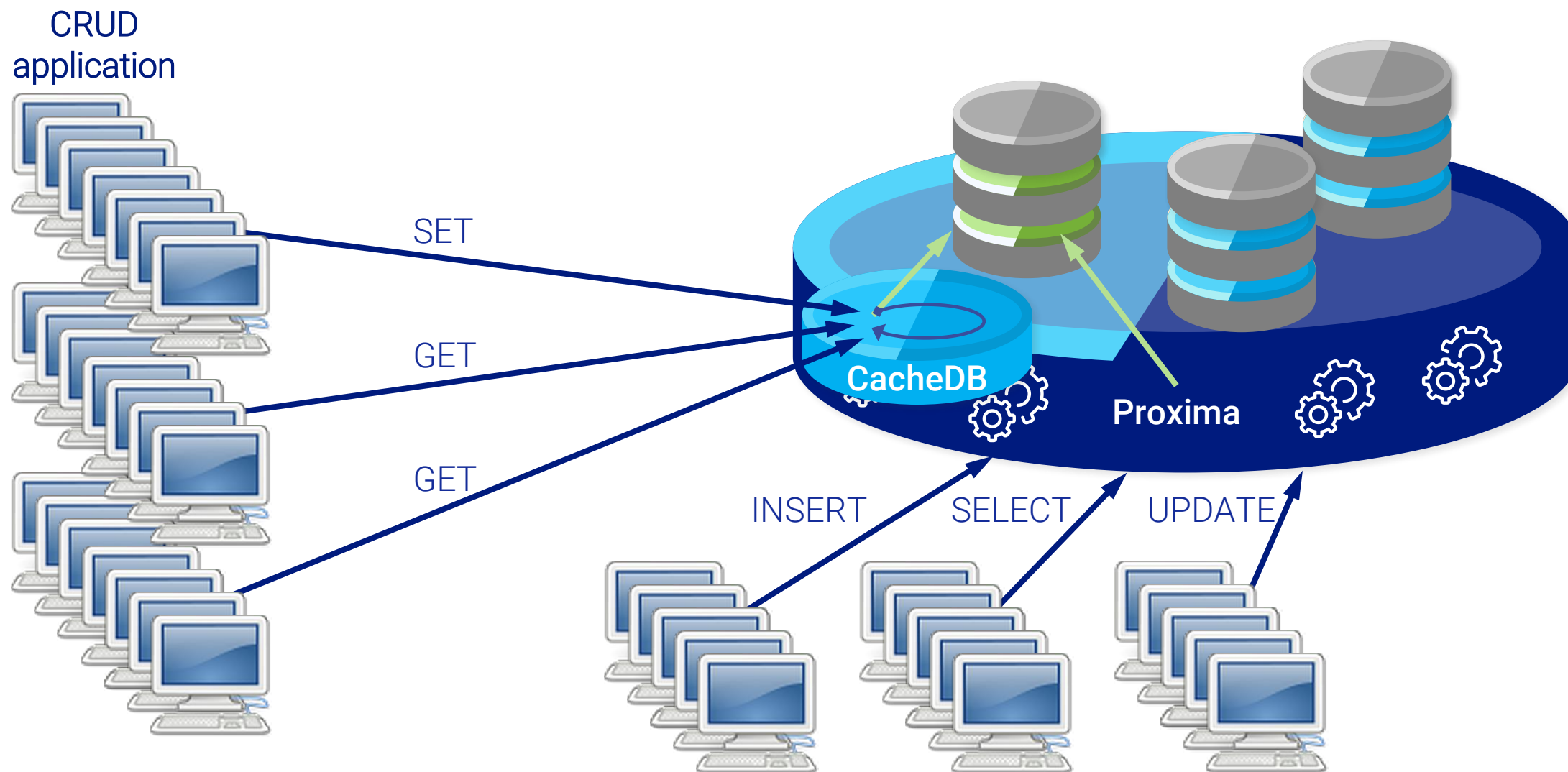


Proxima 17.5 : балансировка читающей нагрузки

Proxima соединения к P2F перенаправляет на реплики и отслеживает изменения в кластере



Разрабатываем новую технологию - CacheDB



Дополнительная информация

- Описание
 - <https://postgrespro.ru/materials/5971621>
- Документация
 - <https://postgrespro.ru/docs/enterprise/17/biha-solution>
 - <https://postgrespro.ru/docs/enterprise/17/proxima>
- Публикации
 - <https://habr.com/ru/companies/postgrespro/articles/898396>
 - <https://habr.com/ru/companies/postgrespro/articles/927040>
 - <https://habr.com/ru/companies/postgrespro/articles/927034>
- Презентации
 - <https://pgconf.ru/talk/2002423>
 - <https://pgconf.ru/talk/1589550>

Спасибо за внимание!

