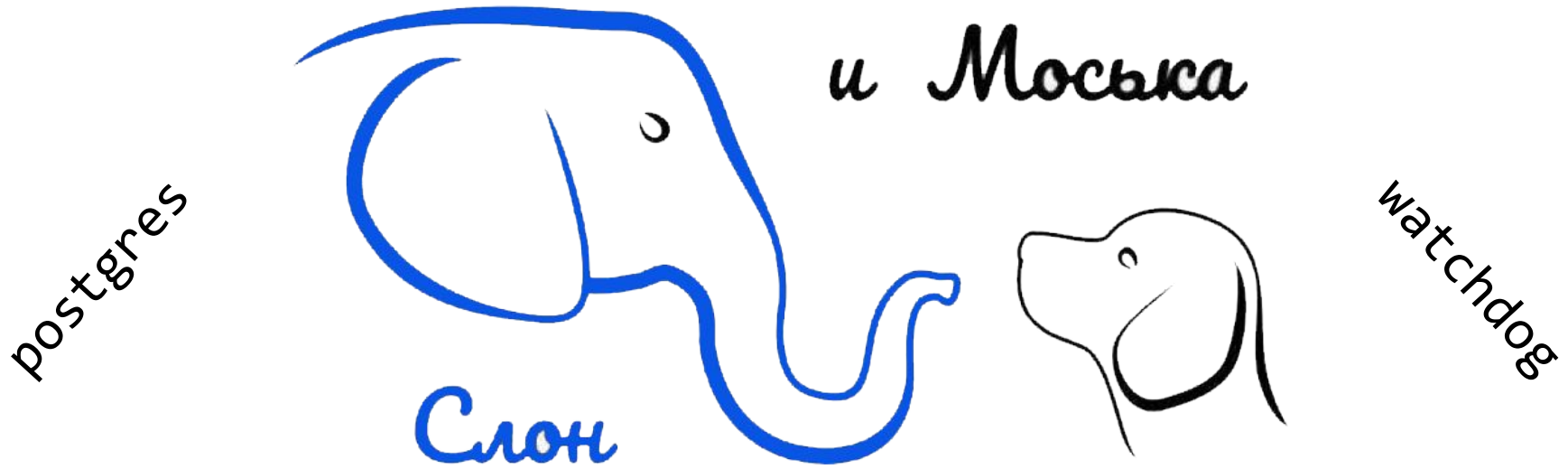
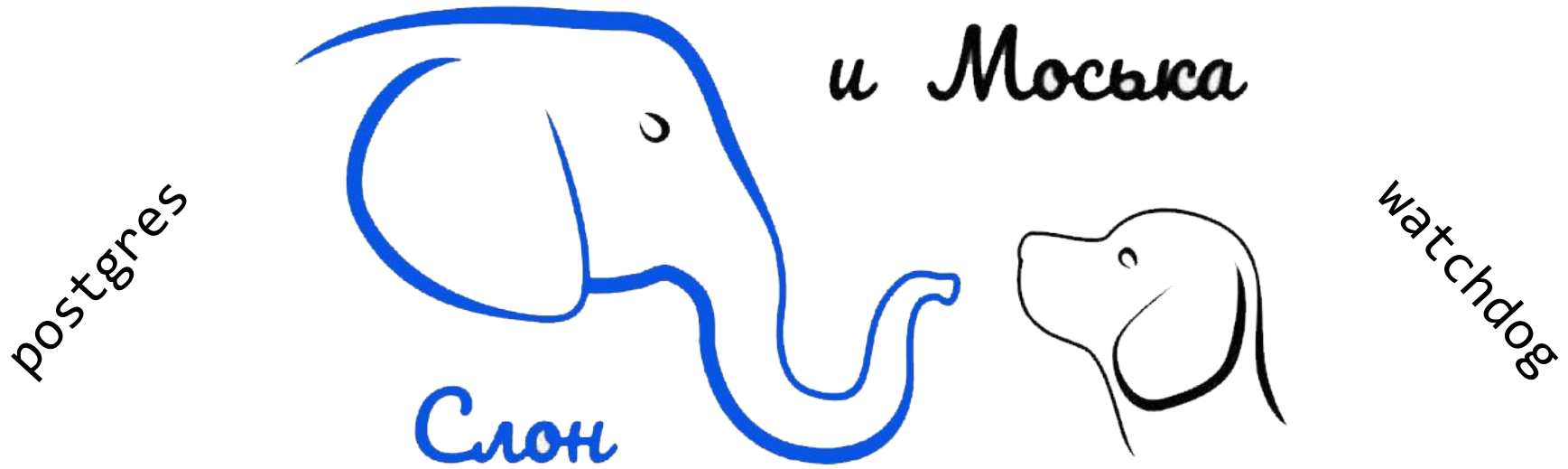
PostgresPro



Проблема фенсинга в кластерах PostgreSQL

Павел Конотопов
руководитель кластерной группы департамента внедрения и технической поддержки

Email: p.konotopov@postgrespro.ru



«The Elephant and the Pug»*
The Fencing Problem in PostgreSQL Clusters.

Pavel Konotopov
Email: p.konotopov@postgrespro.ru

* In my presentation's title, "The Elephant and the Pug," the Pug refers to the audacious character from Ivan Krylov's fable. Despite her small size, the Pug fearlessly barks at the Elephant, believing she has the same might. This serves as a metaphor for the small but critical role of fencing in maintaining the robustness of PostgreSQL clusters.

О чем этот доклад?

- Какую проблему решает фенсинг?
- Как использовать watchdog?
  - Аппаратный
  - Программный
- Фенсинг в кластерным ПО
  - Patroni
  - BiHA
- Что делать когда нет watchdog?
- Итоги

Фенсинг

- **Процесс изоляции узла**
  - разные стратегии
    - STONITH
    - POISON PILL
- **В случае возникновения split-brain**
  - Предотвращает доступ к общим ресурсам
  - Гарантирует целостность данных
- **Критически важный компонент в управлении кластером**

Фенсинг

- **Процесс изоляции узла**
  - разные стратегии
    - **STONITH***
    - POISON PILL

- **В случае возникновения split-brain**
  - Предотвращает доступ к общим ресурсам
  - Гарантирует целостность данных

- **Критически важный компонент в управлении кластером**

***STONITH не всегда можно реализовать,**
см. доклад Игоря Косенкова - https://pgconf.ru/talk/1589499

Фенсинг

- **Процесс изоляции узла**
  - разные стратегии
    - STONITH
    - **POISON PILL**

- **В случае возникновения split-brain**
  - Предотвращает доступ к общим ресурсам
  - Гарантирует целостность данных

- **Критически важный компонент в управлении кластером**

Фенсинг

- **Может быть реализован**
  - на аппаратном уровне
    через управление питанием узлов
  - на программном уровне
    с помощью агентов

- **В кластере PostgreSQL фенсингом управляет кластерное ПО**

Фенсинг

- **Избежание split-brain**
  - Связь между узлами кластера частично нарушена
  - Большинство узлов – продолжает работу
  - Меньшинство узлов
    - «только для чтения»
    - отключаются
- **Гарантия консистентности**
    - только один узел доступен на запись данных

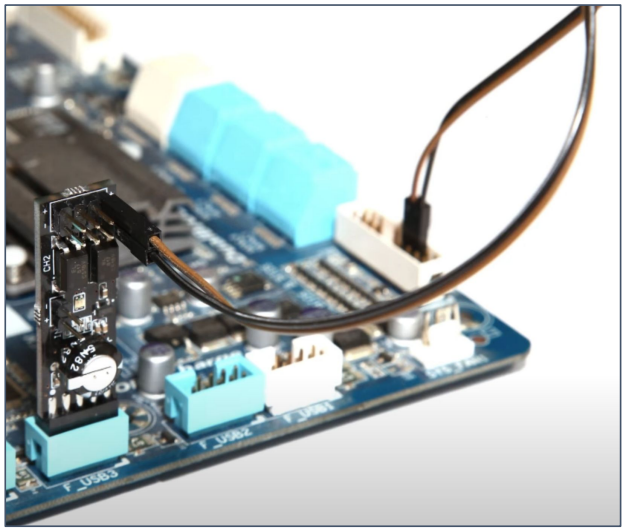
Фенсинг

- **Принудительное отключение узла**
  - Узел кластера перестает адекватно отвечать на запросы
  - Перестает синхронизироваться с остальными узлами

- **Что можно сделать?**
  - Отключить питание
  - Перезагрузить
  - Изолировать узел от остальных
    - например, выключить порт на коммутаторе
  - Другие действия

Фенсинг: реализация **watchdog**

- **Аппаратный**
  - физическое устройство
  - эмуляция физического устройства гипервизором
- **Программный**
  - модуль ядра



Фенсинг: гипервизоры

- **Xen**
- **KVM**
  - oVirt, Proxmox и тд
- **bhyve, VMD (для фанатов *bsd)**
- **VMWare**
  - ESXi, vCenter, vCloudDirector
- **Hyper-V**
- **Импортозамещающие системы управления**
  - Alt PVE, Ред Виртуализация и тд.

Фенсинг: Proxmox, аппаратный watchdog

PostgresPro

- **Конфигурация виртуальный машины**

```
/etc/pve/qemu-server/<vmid>.conf
watchdog: model=i6300esb,action=stop
```

- **В виртуальной машине выполнить установку пакета**

```
apt-get/rpm/dnf install watchdog
cp /lib/modprobe.d/blacklist-watchdog.conf /etc/modprobe.d/
```

- **Добавить строки**

```
modinfo i6300esb
echo i6300esb > /etc/modprobe.d/i6300es.conf
```

- **Сделать модуль загружаемым при старте ВМ**

```
echo i6300esb >> /etc/modules-load.d/modules.conf
```

Фенсинг: Proxmox, аппаратный watchdog

PostgresPro

- **Выполнить остановку и старт виртуальной машины**
- **Назначить права на использование (можно через udev)**

```
chown postgres:postgres /dev/watchdog
```

- **Можно использовать systemd сервис watchdog**
  - добавить в `/etc/watchdog.conf`

```
watchdog-device = /dev/watchdog
realtime = yes
priority = 1
```

- **Проверить статус устройства**

```
wdctl -F /dev/watchdog
Device:         /dev/watchdog
Identity:       i6300ESB timer [version 0]
Timeout:        30 seconds
Pre-timeout:     0 seconds
```

Фенсинг: программный **watchdog**

PostgresPro

- **Выполнить загрузку модуля ядра**

```
modprobe softdog
```

- **Назначить права на использование**

```
chown postgres:postgres /dev/watchdog
```

- **При перезагрузке права придется установить заново**
- **Юзай udev!**

Фенсинг в **Patroni:** настройка **Patroni**

PostgresPro

- **Конфигурация Patroni, секция watchdog**

```
watchdog:
  device: /dev/watchdog
  mode: automatic
```

- **Systemd unit Patroni**

```
[Service]
ExecStartPre=-/usr/bin/sudo /sbin/modprobe softdog
ExecStartPre=-/usr/bin/sudo /bin/chown postgres /dev/watchdog
CPUSchedulingPolicy=fifo
CPUSchedulingPriority=99
CPUSchedulingResetOnFork=true
```

PostgresPro

- **Что делать, когда аппаратный watchdog недоступен?**

- **Watchdog выручает не во всех ситуациях**
  - Зависание виртуальной машины, контролируемое гипервизором
  - Thin provisioning дисков с PGDATA/PGWAL
    - закончилось место
    - нужно расширить диск
    - а на storage гипервизора закончилось место

Patroni: неприятное поведение

PostgresPro

- **Виртуальная машина узла-лидера «зависла»**
- **Реплика стала новым мастером**
- **Клиенты активно пишут в новый мастер**
- **Виртуальная машина узла-лидера «отвисла»**
- **Новый мастер экстренно демоутится**
- **На бывшем новом мастере запускется pg_rewind**

```
use_pg_rewind: true
```

- **Потеря данных!**
- **Как избежать?**

Фенсинг в **Patroni:** тесты

PostgresPro

- **Эмуляция «зависания»**
- **Pause на стороне гипервизора?**
  - Не отрабатывает, как нам необходимо
- **Воспользуемся возможностями ядра Linux**
  - функция `stop-machine()`

Фенсинг в **Patroni:** тесты

PostgresPro

- **Вызов функции**
  - `stop_machine(stop_func,NULL,NULL)`
  - нужно написать `stop_func()`

- **Напишем модуль ядра! It's easy! :)**
  - остановит работу всей системы на некоторое время
    - невозможность запуска всех процессов
    - невозможность запуска всех функций ядра
    - функции ядра – драйверы устройств и обработчики прерываний
    - с точки зрения гипервизора
      все операции ввода-вывода остановятся

Фенсинг в **Patroni:** модуль ядра

PostgresPro

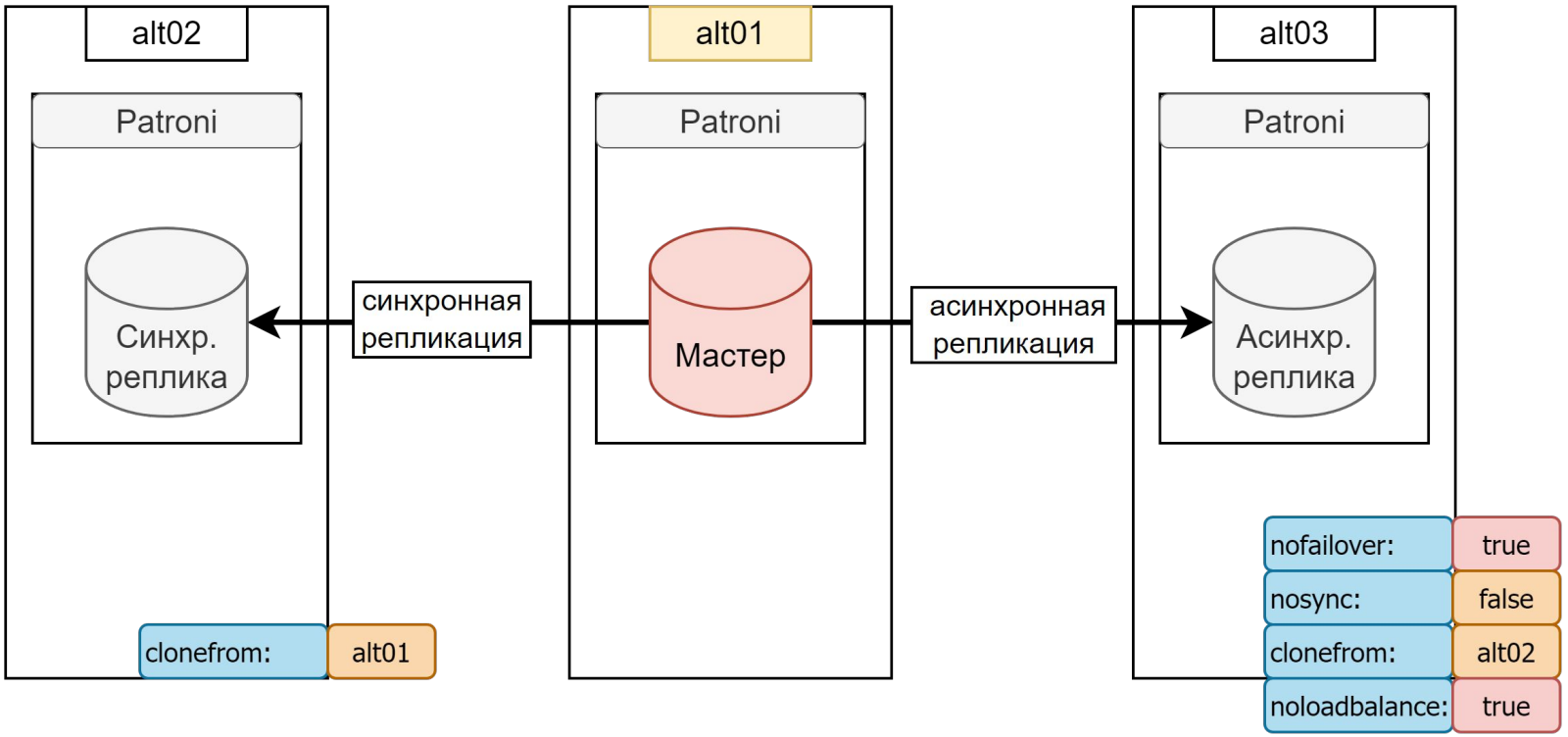
```
#define STOP_MSECS 60000

static int stop_func(void *arg) {
    mdelay(STOP_MSECS);
    return 0;
}

static int mymodule_init(void) {
     int ret = 0;
     ret = stop_machine(stop_func, NULL, NULL);
     return ret;
}

static void mymodule_exit(void) { /* do nothing */ }

module_init(mymodule_init);

module_exit(mymodule_exit);
```

Фенсинг в **Patroni:** тест

PostgresPro

- **Чтобы собрать (для Alt10)**
  - apt-get install
    kernel-image
    kernel-headers
    kernel-headers-modules
    + dev tools
- **На лидере выполним:**
  - `insmod stop-machine`
- **Получим ожидаемое поведение**
- **Аппаратный watchdog**
  - произойдет перезагрузка или выключение виртуальной машины
- **Программный softdog – не отработает**

Фенсинг в **Patroni:** схема кластера

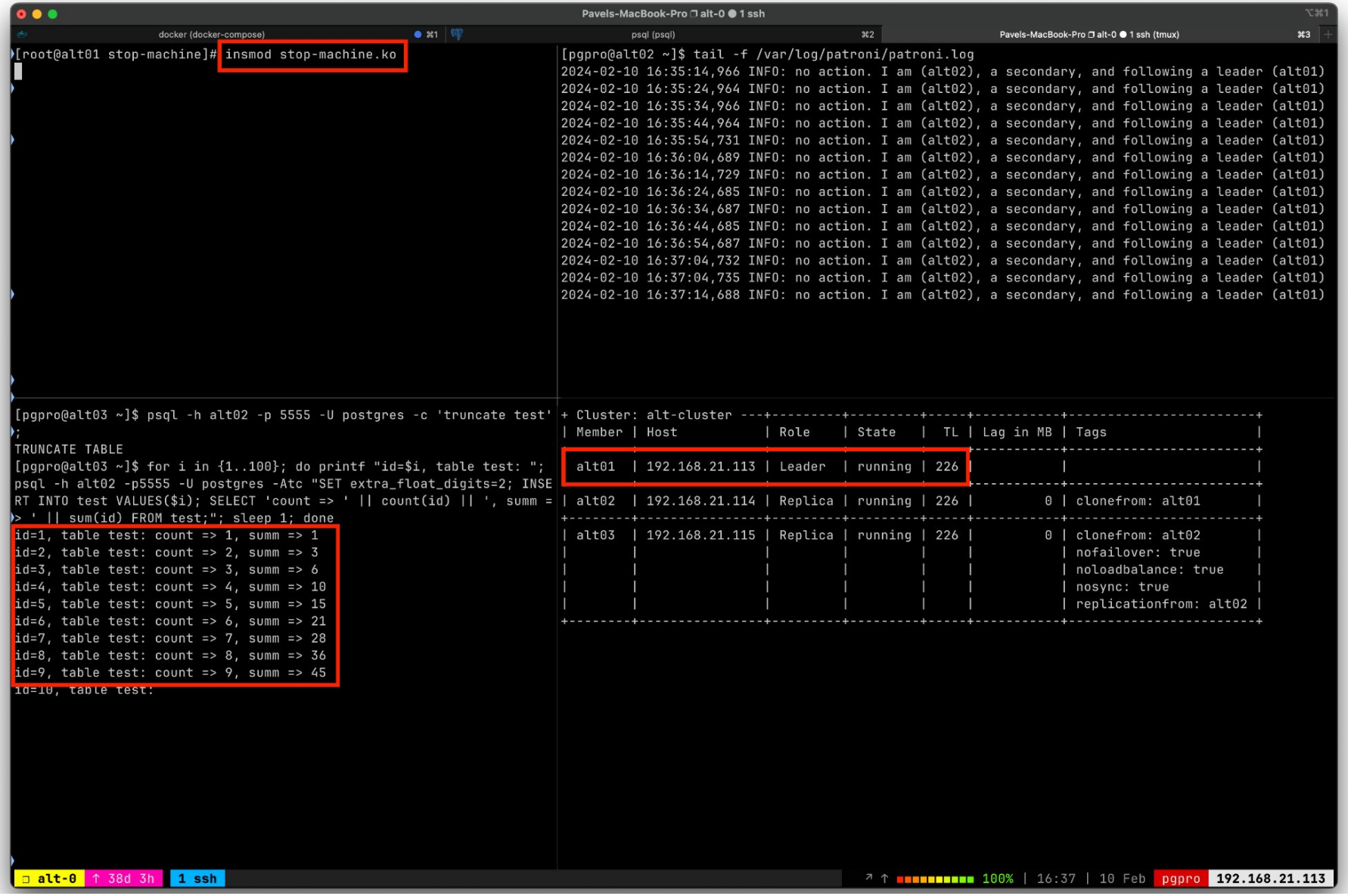


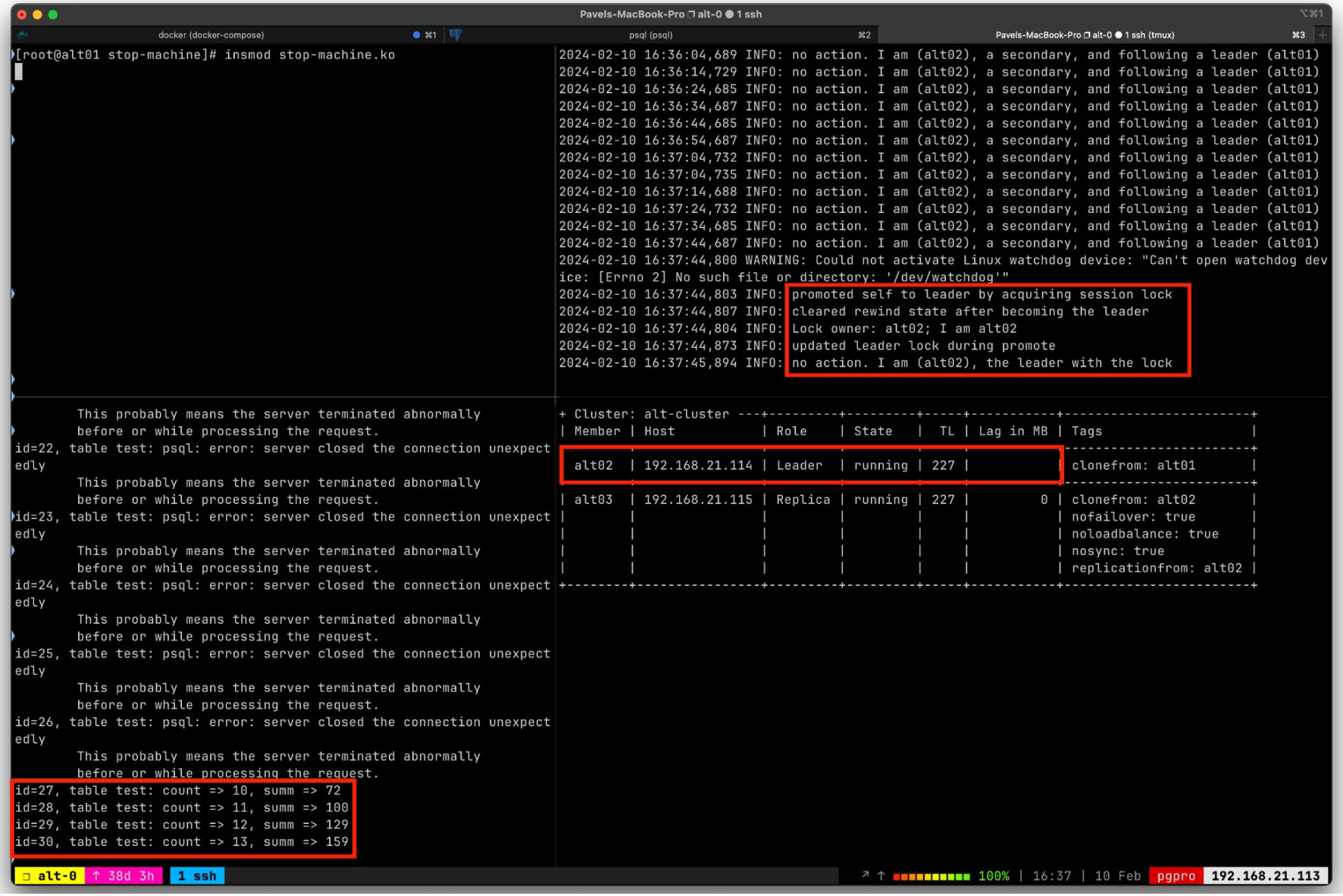
Фенсинг в **Patroni:** тест

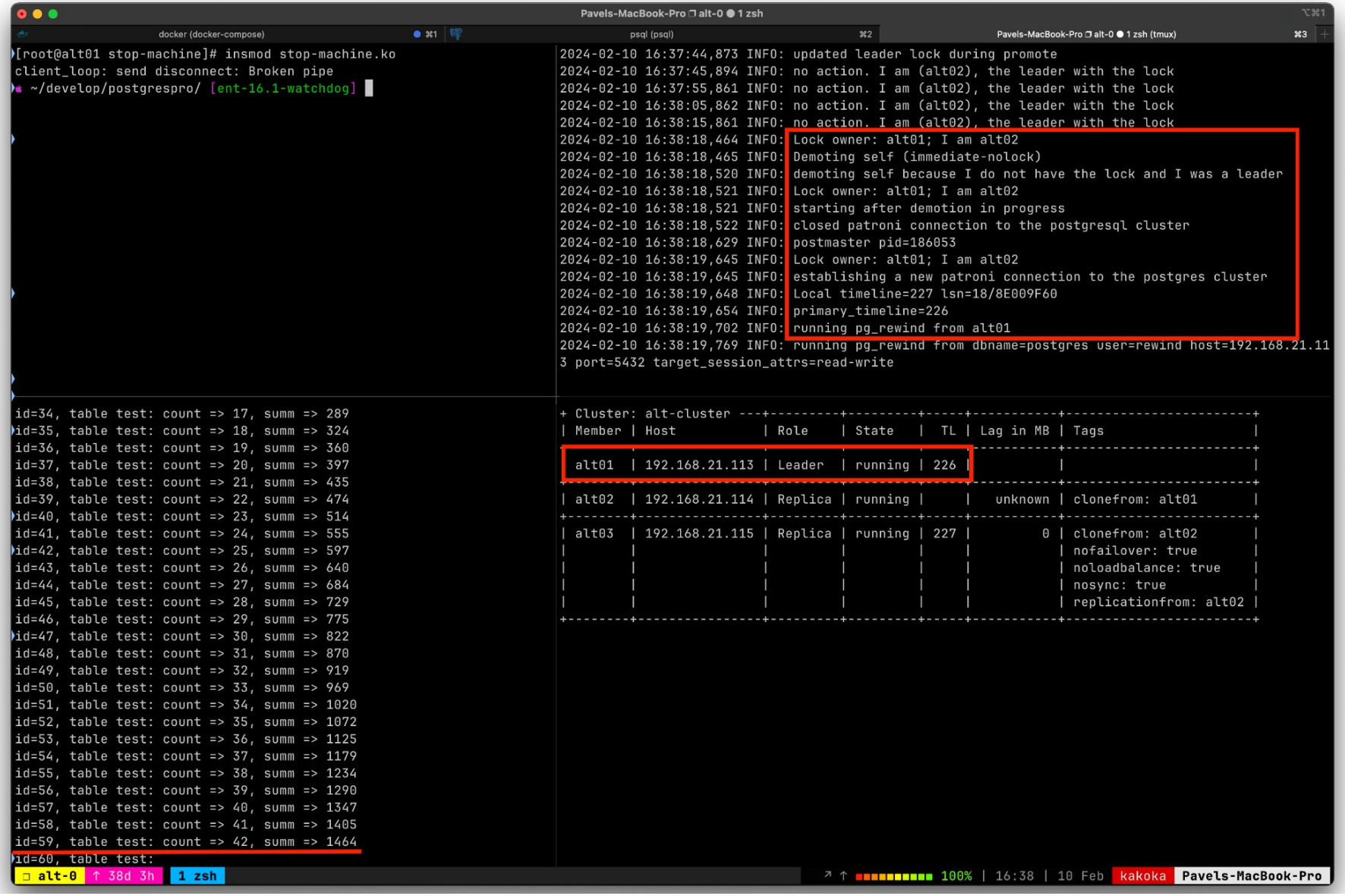
PostgresPro

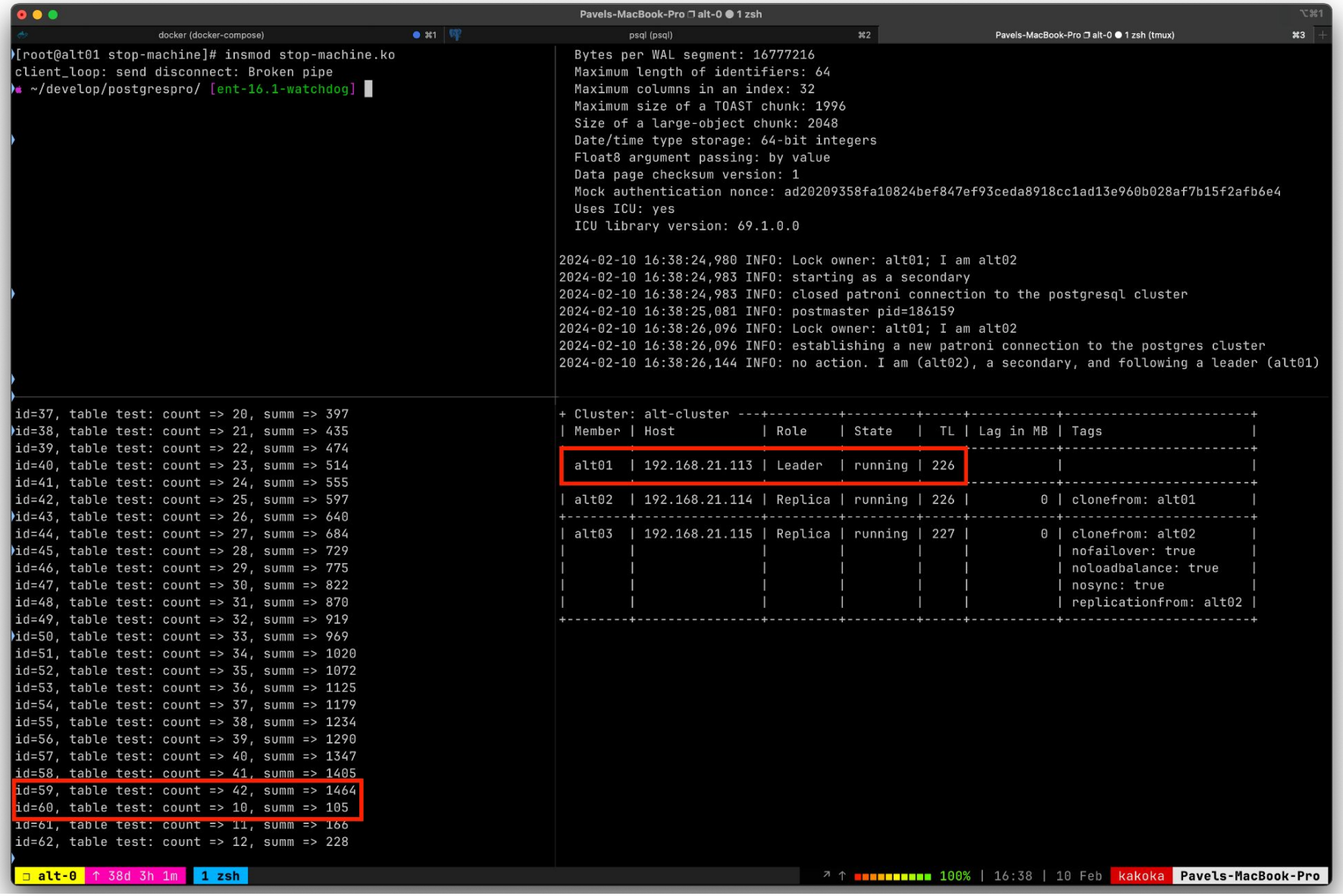
```
use_pg_rewind: true
```

```
# вставляем строку, считаем количество строк и сумму их значений
for i in {1..100}; do
  printf "id=$i, table test: "
  psql -h alt02 -p 5555 -U postgres -Atc \
  "INSERT INTO test VALUES($i); \
  SELECT 'count => ' || count(id) || ', summ => ' || sum(id) FROM test;"
  sleep 1
done
```


Pavels-MacBook-Pro alt-0 1 ssh
docker (docker-compose)
psql (psql)
Pavels-MacBook-Pro alt-0 1 ssh (tmux)
[root@alt01 stop-machine]# insmod stop-machine.ko
[pgpro@alt02 ~]$ tail -f /var/log/patroni/patroni.log
2024-02-10 16:35:14,966 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:35:24,964 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:35:34,966 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:35:44,964 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:35:54,731 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:36:04,689 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:36:14,729 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:36:24,685 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:36:34,687 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:36:44,685 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:36:54,687 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:37:04,732 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:37:04,735 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:37:14,688 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
[pgpro@alt03 ~]$ psql -h alt02 -p 5555 -U postgres -c 'truncate test';
TRUNCATE TABLE
[pgpro@alt03 ~]$ for i in {1..100}; do printf "id=$i, table test: "; psql -h alt02 -p5555 -U postgres -Atc "SET extra_float_digits=2; INSERT INTO test VALUES($i); SELECT 'count => ' || count(id) || ', summ => ' || sum(id) FROM test;"; sleep 1; done
id=1, table test: count => 1, summ => 1
id=2, table test: count => 2, summ => 3
id=3, table test: count => 3, summ => 6
id=4, table test: count => 4, summ => 10
id=5, table test: count => 5, summ => 15
id=6, table test: count => 6, summ => 21
id=7, table test: count => 7, summ => 28
id=8, table test: count => 8, summ => 36
id=9, table test: count => 9, summ => 45
id=10, table test:
+ Cluster: alt-cluster ---+---------+---------+-----+-----------+-----------------------+
| Member | Host | Role | State | TL | Lag in MB | Tags |
+--------+---------------+---------+---------+-----+-----------+-----------------------+
| alt01 | 192.168.21.113 | Leader | running | 226 | | |
+--------+---------------+---------+---------+-----+-----------+-----------------------+
| alt02 | 192.168.21.114 | Replica | running | 226 | 0 | clonefrom: alt01 |
+--------+---------------+---------+---------+-----+-----------+-----------------------+
alt03	192.168.21.115	Replica	running	226	0	clonefrom: alt02
						nofailover: true
						noloadbalance: true
						nosync: true
						replicationfrom: alt02
+--------+---------------+---------+---------+-----+-----------+-----------------------+
alt-0 38d 3h 1 ssh
100% | 16:37 | 10 Feb pgpro 192.168.21.113


Pavels-MacBook-Pro alt-0 1 ssh
docker (docker-compose)
psql (psql)
Pavels-MacBook-Pro alt-0 1 ssh (tmux)
[root@alt01 stop-machine]# insmod stop-machine.ko
2024-02-10 16:36:04,689 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:36:14,729 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:36:24,685 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:36:34,687 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:36:44,685 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:36:54,687 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:37:04,732 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:37:04,735 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:37:14,688 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:37:24,732 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:37:34,685 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:37:44,687 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
2024-02-10 16:37:44,800 WARNING: Could not activate Linux watchdog device: "Can't open watchdog dev
ice: [Errno 2] No such file or directory: '/dev/watchdog'"
2024-02-10 16:37:44,803 INFO: promoted self to leader by acquiring session lock
2024-02-10 16:37:44,807 INFO: cleared rewind state after becoming the leader
2024-02-10 16:37:44,804 INFO: Lock owner: alt02; I am alt02
2024-02-10 16:37:44,873 INFO: updated leader lock during promote
2024-02-10 16:37:45,894 INFO: no action. I am (alt02), the leader with the lock
This probably means the server terminated abnormally
before or while processing the request.
id=22, table test: psql: error: server closed the connection unexpect
edly
This probably means the server terminated abnormally
before or while processing the request.
id=23, table test: psql: error: server closed the connection unexpect
edly
This probably means the server terminated abnormally
before or while processing the request.
id=24, table test: psql: error: server closed the connection unexpect
edly
This probably means the server terminated abnormally
before or while processing the request.
id=25, table test: psql: error: server closed the connection unexpect
edly
This probably means the server terminated abnormally
before or while processing the request.
id=26, table test: psql: error: server closed the connection unexpect
edly
This probably means the server terminated abnormally
before or while processing the request.
id=27, table test: count => 10, summ => 72
id=28, table test: count => 11, summ => 100
id=29, table test: count => 12, summ => 129
id=30, table test: count => 13, summ => 159
+ Cluster: alt-cluster
Member
Host
Role
State
TL
Lag in MB
Tags
alt02
192.168.21.114
Leader
running
227
clonefrom: alt01
alt03
192.168.21.115
Replica
running
227
0
clonefrom: alt02
nofailover: true
noloadbalance: true
nosync: true
replicationfrom: alt02
alt-0
38d 3h
1 ssh
100%
16:37
10 Feb
pgpro
192.168.21.113


Pavels-MacBook-Pro ❐ alt-0 ● 1 zsh
docker (docker-compose) ⌘1
psql (psql) ⌘2
Pavels-MacBook-Pro ❐ alt-0 ● 1 zsh (tmux) ⌘3
[root@alt01 stop-machine]# insmod stop-machine.ko
client_loop: send disconnect: Broken pipe
~/develop/postgrespro/ [ent-16.1-watchdog]
2024-02-10 16:37:44,873 INFO: updated leader lock during promote
2024-02-10 16:37:45,894 INFO: no action. I am (alt02), the leader with the lock
2024-02-10 16:37:55,861 INFO: no action. I am (alt02), the leader with the lock
2024-02-10 16:38:05,862 INFO: no action. I am (alt02), the leader with the lock
2024-02-10 16:38:15,861 INFO: no action. I am (alt02), the leader with the lock
2024-02-10 16:38:18,464 INFO: Lock owner: alt01; I am alt02
2024-02-10 16:38:18,465 INFO: Demoting self (immediate-nolock)
2024-02-10 16:38:18,520 INFO: demoting self because I do not have the lock and I was a leader
2024-02-10 16:38:18,521 INFO: Lock owner: alt01; I am alt02
2024-02-10 16:38:18,521 INFO: starting after demotion in progress
2024-02-10 16:38:18,522 INFO: closed patroni connection to the postgresql cluster
2024-02-10 16:38:18,629 INFO: postmaster pid=186053
2024-02-10 16:38:19,645 INFO: Lock owner: alt01; I am alt02
2024-02-10 16:38:19,645 INFO: establishing a new patroni connection to the postgres cluster
2024-02-10 16:38:19,648 INFO: Local timeline=227 lsn=18/8E009F60
2024-02-10 16:38:19,654 INFO: primary_timeline=226
2024-02-10 16:38:19,702 INFO: running pg_rewind from alt01
2024-02-10 16:38:19,769 INFO: running pg_rewind from dbname=postgres user=rewind host=192.168.21.113 port=5432 target_session_attrs=read-write
id=34, table test: count => 17, summ => 289
id=35, table test: count => 18, summ => 324
id=36, table test: count => 19, summ => 360
id=37, table test: count => 20, summ => 397
id=38, table test: count => 21, summ => 435
id=39, table test: count => 22, summ => 474
id=40, table test: count => 23, summ => 514
id=41, table test: count => 24, summ => 555
id=42, table test: count => 25, summ => 597
id=43, table test: count => 26, summ => 640
id=44, table test: count => 27, summ => 684
id=45, table test: count => 28, summ => 729
id=46, table test: count => 29, summ => 775
id=47, table test: count => 30, summ => 822
id=48, table test: count => 31, summ => 870
id=49, table test: count => 32, summ => 919
id=50, table test: count => 33, summ => 969
id=51, table test: count => 34, summ => 1020
id=52, table test: count => 35, summ => 1072
id=53, table test: count => 36, summ => 1125
id=54, table test: count => 37, summ => 1179
id=55, table test: count => 38, summ => 1234
id=56, table test: count => 39, summ => 1290
id=57, table test: count => 40, summ => 1347
id=58, table test: count => 41, summ => 1405
id=59, table test: count => 42, summ => 1464
id=60, table test:
+ Cluster: alt-cluster ---+---------+---------+-----+-----------+-----------------------+
Member	Host	Role	State	TL	Lag in MB	Tags
alt01	192.168.21.113	Leader	running	226		
alt02	192.168.21.114	Replica	running		unknown	clonefrom: alt01
alt03	192.168.21.115	Replica	running	227	0	clonefrom: alt02
						nofailover: true
						noloadbalance: true
						nosync: true
						replicationfrom: alt02
alt-0 ↑ 38d 3h 1 zsh
100% | 16:38 | 10 Feb kakoka Pavels-MacBook-Pro


Pavels-MacBook-Pro ❐ alt-0 ● 1 zsh
docker (docker-compose) ⌘1
psql (psql) ⌘2
Pavels-MacBook-Pro ❐ alt-0 ● 1 zsh (tmux) ⌘3
[root@alt01 stop-machine]# insmod stop-machine.ko
client_loop: send disconnect: Broken pipe
~/develop/postgrespro/ [ent-16.1-watchdog]
Bytes per WAL segment: 16777216
Maximum length of identifiers: 64
Maximum columns in an index: 32
Maximum size of a TOAST chunk: 1996
Size of a large-object chunk: 2048
Date/time type storage: 64-bit integers
Float8 argument passing: by value
Data page checksum version: 1
Mock authentication nonce: ad20209358fa10824bef847ef93ceda8918cc1ad13e960b028af7b15f2afb6e4
Uses ICU: yes
ICU library version: 69.1.0.0
2024-02-10 16:38:24,980 INFO: Lock owner: alt01; I am alt02
2024-02-10 16:38:24,983 INFO: starting as a secondary
2024-02-10 16:38:24,983 INFO: closed patroni connection to the postgresql cluster
2024-02-10 16:38:25,081 INFO: postmaster pid=186159
2024-02-10 16:38:26,096 INFO: Lock owner: alt01; I am alt02
2024-02-10 16:38:26,096 INFO: establishing a new patroni connection to the postgres cluster
2024-02-10 16:38:26,144 INFO: no action. I am (alt02), a secondary, and following a leader (alt01)
id=37, table test: count => 20, summ => 397
id=38, table test: count => 21, summ => 435
id=39, table test: count => 22, summ => 474
id=40, table test: count => 23, summ => 514
id=41, table test: count => 24, summ => 555
id=42, table test: count => 25, summ => 597
id=43, table test: count => 26, summ => 640
id=44, table test: count => 27, summ => 684
id=45, table test: count => 28, summ => 729
id=46, table test: count => 29, summ => 775
id=47, table test: count => 30, summ => 822
id=48, table test: count => 31, summ => 870
id=49, table test: count => 32, summ => 919
id=50, table test: count => 33, summ => 969
id=51, table test: count => 34, summ => 1020
id=52, table test: count => 35, summ => 1072
id=53, table test: count => 36, summ => 1125
id=54, table test: count => 37, summ => 1179
id=55, table test: count => 38, summ => 1234
id=56, table test: count => 39, summ => 1290
id=57, table test: count => 40, summ => 1347
id=58, table test: count => 41, summ => 1405
id=59, table test: count => 42, summ => 1464
id=60, table test: count => 10, summ => 105
id=61, table test: count => 11, summ => 166
id=62, table test: count => 12, summ => 228
+ Cluster: alt-cluster ---+---------+---------+-----+-----------+-----------------------+
Member	Host	Role	State	TL	Lag in MB	Tags
alt01	192.168.21.113	Leader	running	226		
alt02	192.168.21.114	Replica	running	226	0	clonefrom: alt01
alt03	192.168.21.115	Replica	running	227	0	clonefrom: alt02
						nofailover: true
						noloadbalance: true
						nosync: true
						replicationfrom: alt02
alt-0 ↑ 38d 3h 1m 1 zsh
100% | 16:38 | 10 Feb kakoka Pavels-MacBook-Pro

Фенсинг в **Patroni: pre-promote callback**

PostgresPro

- **доступен с версии 2.0.0**

```
postgresql:
   pre_promote: '/usr/local/bin/stop_vm.sh'
```

- **если вызов скрипта завершается ненулевым кодом**
  - Patroni не продвигает реплику
  - Удаляет ключ лидера из DCS

Настройка **Proxmox**

PostgresPro

- **Добавить роль**
  - `pveum roleadd fencer -privs "VM.PowerMgmt VM.Console VM.Audit"`
- **Добавить токен**
  - `pveum user token add tech-user@pve stop-vm`
- **Назначить ACL на токен**
  - `pveum acl modify /vms/162 -token tech-user@pve!stop-vm -role fencer`

- **Проверить права**
  - `pveum user token permissions username@ldap stop-vm`

```
┌──────────┬──────────────────┐
│ ACL path │ Permissions      │
╞══════════╪══════════════════╡
│ /vms/162 │ VM.Audit (*)     │
│          │ VM.Monitor (*)   │
│          │ VM.PowerMgmt (*) │
└──────────┴──────────────────┘
```

- **Проверить доступ через API**
  - `curl -s -k -H 'Authorization: PVEAPIToken=`*`tech-user@pve!stop-vm=b079627x-xxxx-xxxx-xxxx-bf07c4deaac0`*`' 'https://192.168.20.100:8006/api2/json/access/permissions' | jq '.data."/vms/162"'`

```
{
    "VM.PowerMgmt": 1,
    "VM.Audit": 1,
    "VM.Monitor": 1
}
```

Фенсинг в **Patroni: pre-promote** скрипт

PostgresPro

- **`check_ping()`**
  - вернуть 0, если узел доступен
- **`check_patroni_url()`**
  - вернуть 0, если Patroni REST API доступен
- **`check_vmid_status()`**
  - проверить в каком состоянии виртуальная машина
  - вернуть 0, если остановлена
- **`stop_vmid()`**
  - остановить виртуальную машину
  - вернуть 0, если виртуальная машина остановлена

Фенсинг в **Patroni:** тестируем **pre-promote** скрипт

PostgresPro

- **Остановим systemd сервис patroni**

```
Feb 07 22:45:12 alt02 patroni[27384]: Host alt01 is reachable.
Feb 07 22:45:12 alt02 patroni[27395]: server promoting
```

- **Отключим ответ на ping**

```
echo 1 > /proc/sys/net/ipv4/icmp_echo_ignore_all
```

- **Отключим ответ на ping и остановим сервис patroni**

```
Feb 07 20:10:21 alt02 patroni[22717]: Host alt01 is not reachable.
Feb 07 20:10:21 alt02 patroni[22717]: URL http://alt01:8008 is available.
Feb 07 20:10:21 alt02 patroni[22758]: server promoting
```

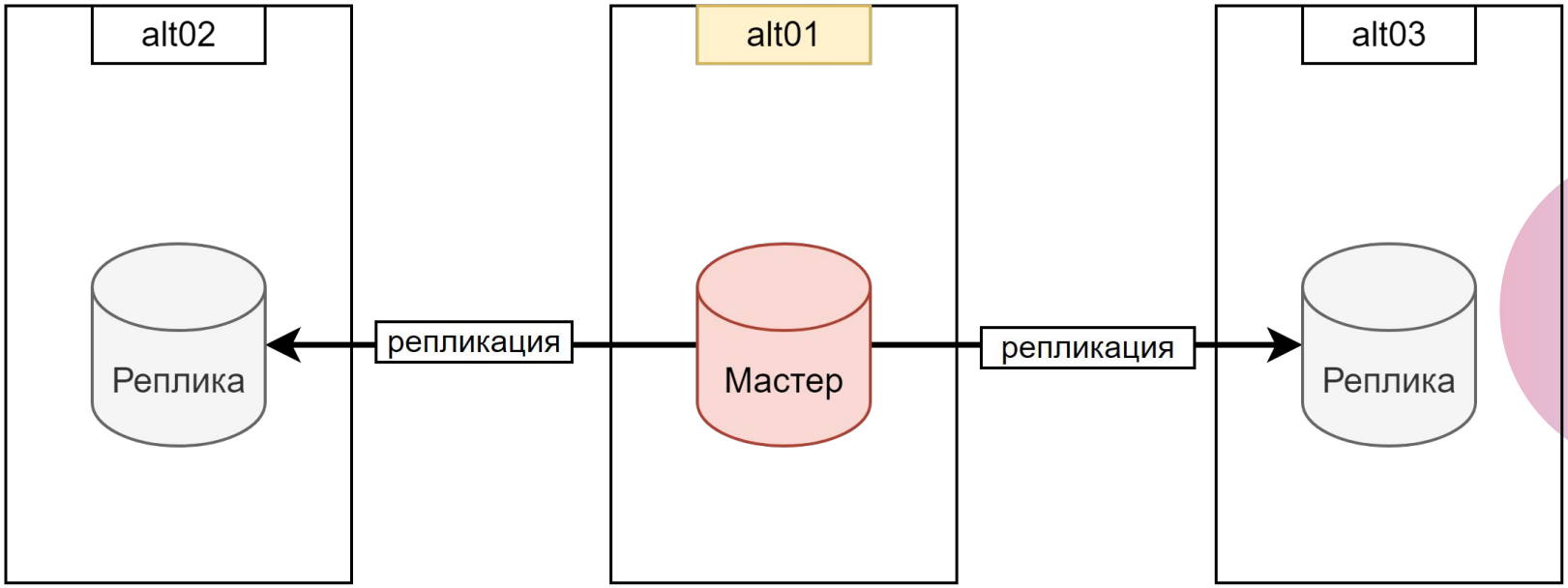
Фенсинг в **Patroni: pre-promote** скрипт

PostgresPro

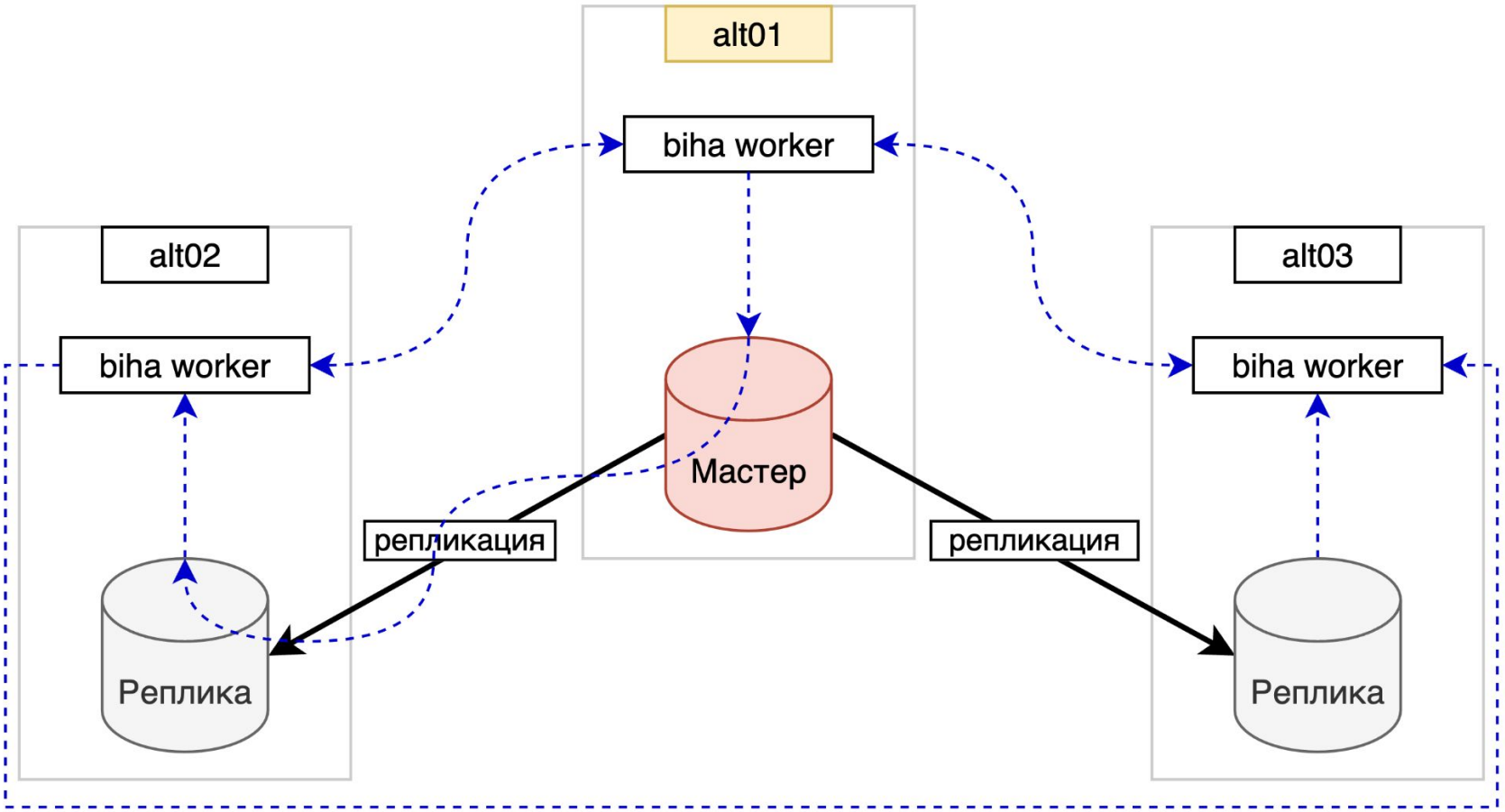
- **«Заморозим» виртуальную машину**

```
Feb 07 20:15:39 alt02 patroni[22937]: Host alt01 is not reachable.
Feb 07 20:15:44 alt02 patroni[22937]: URL http://alt01:8008 is not available.
Feb 07 20:15:44 alt02 patroni[22937]: We should stop vm 162!
Feb 07 20:15:44 alt02 patroni[22937]: VM 162 stopping, waiting...
Feb 07 20:15:45 alt02 patroni[22937]: STATUS running
Feb 07 20:15:45 alt02 patroni[22937]: VM 162 still stopping...
Feb 07 20:15:50 alt02 patroni[22937]: STATUS stopped
Feb 07 20:15:50 alt02 patroni[22937]: VM 162 has been stopped!
Feb 07 20:15:50 alt02 patroni[22980]: server promoting
```

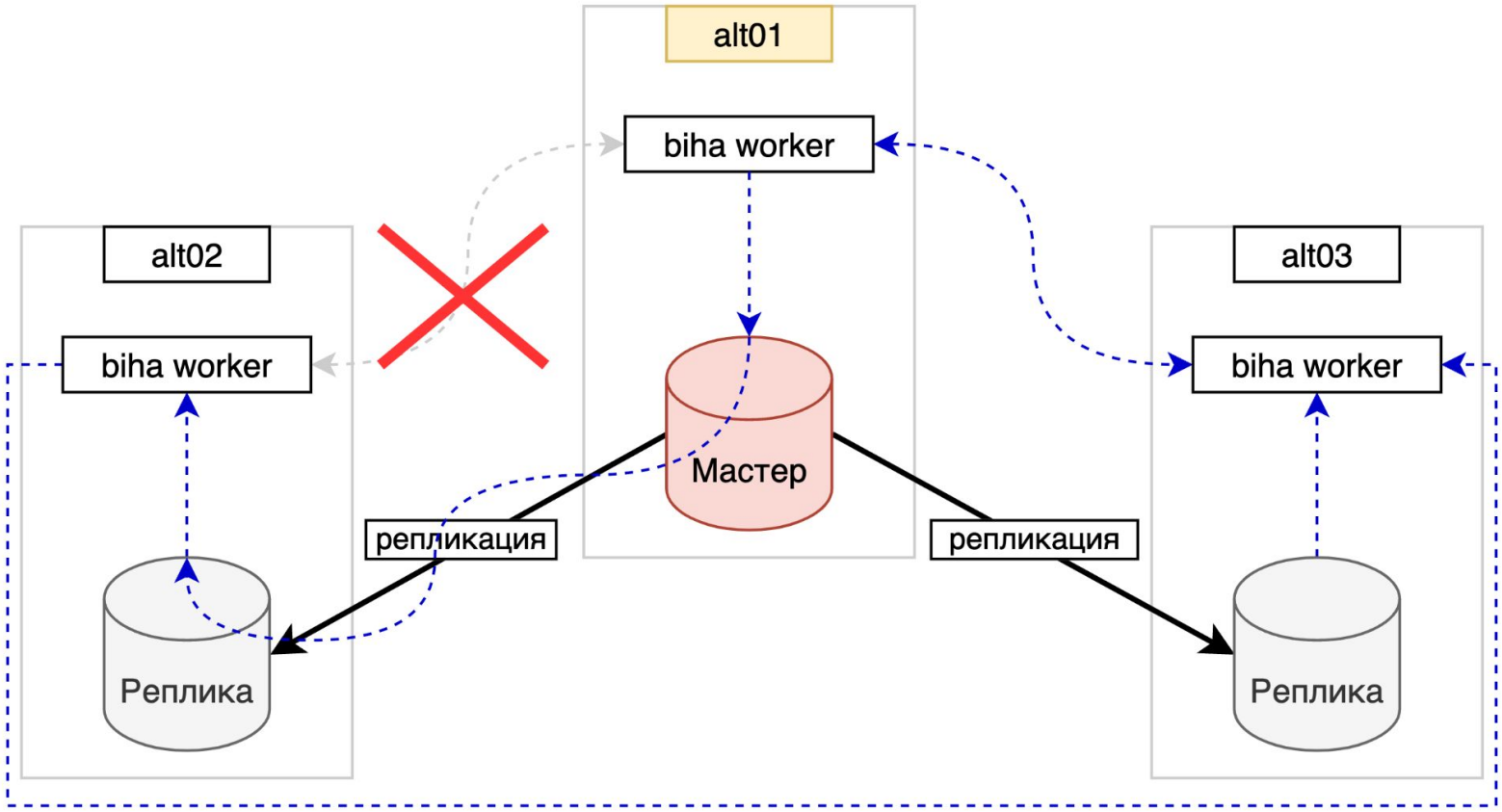
Фенсинг в **BiHA:** схема кластера



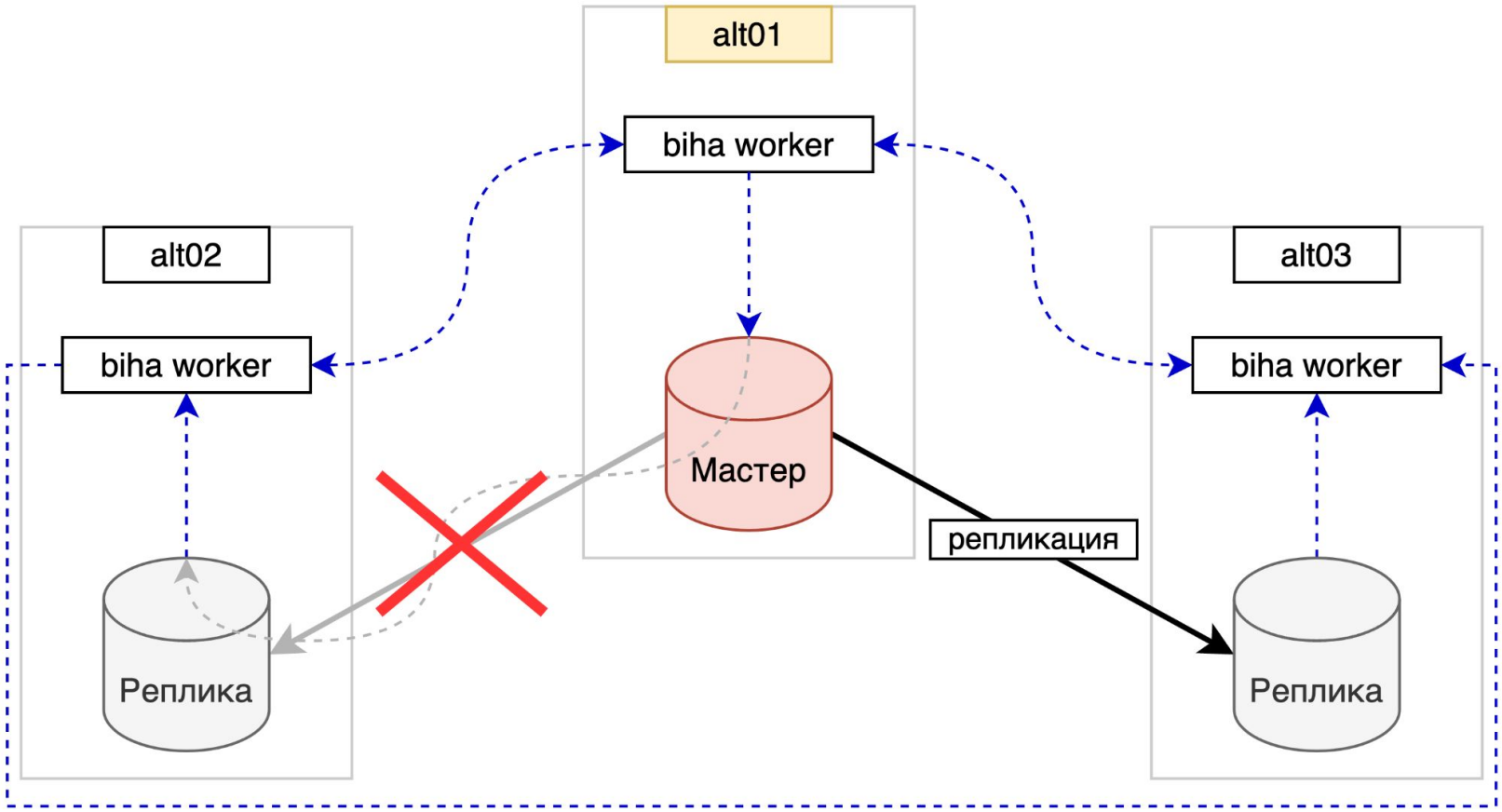
Фенсинг в **BiHA:** схема кластера



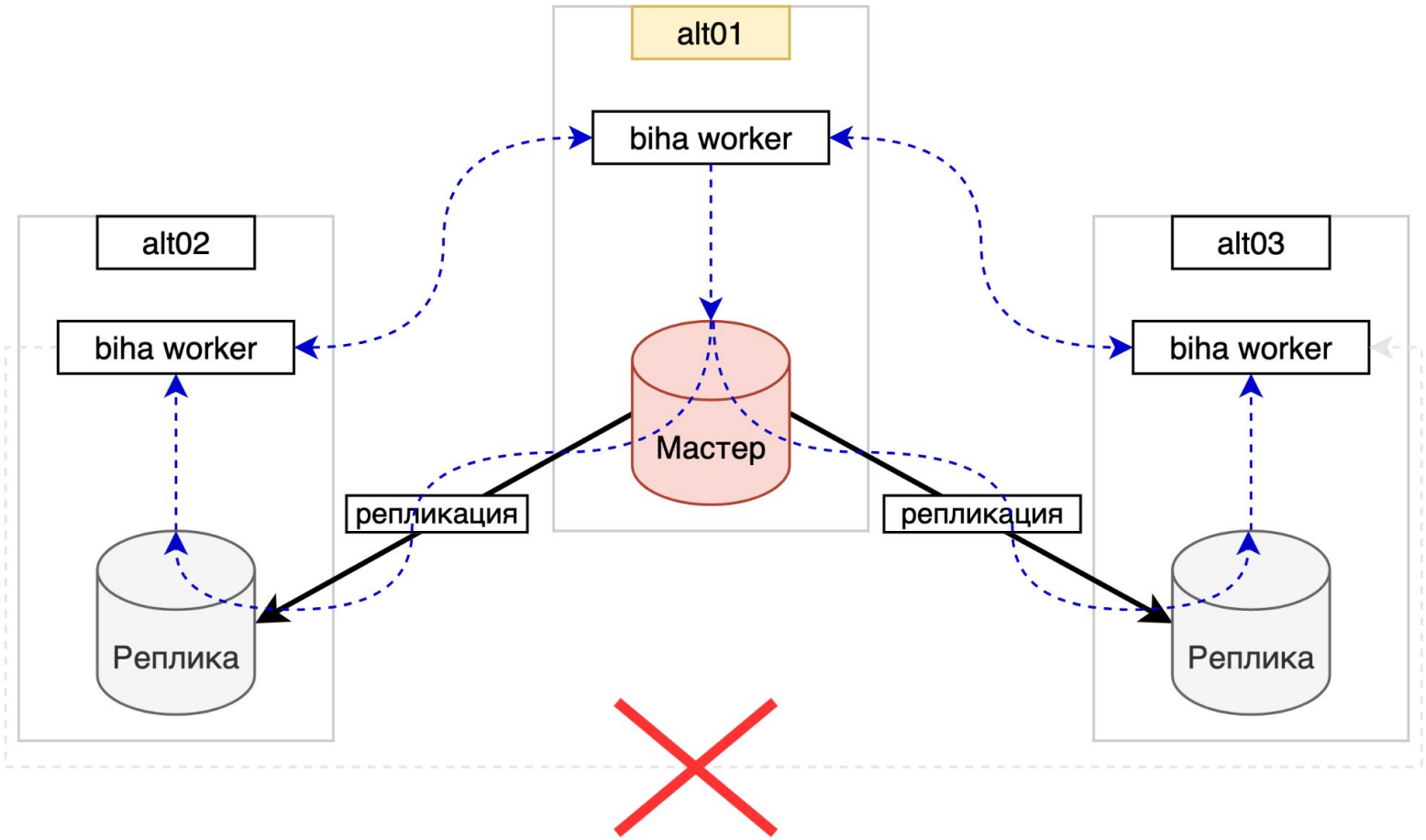
BiHA: отказы


alt01
biha worker
alt02
biha worker
alt03
biha worker
Мастер
репликация
репликация
Реплика
Реплика

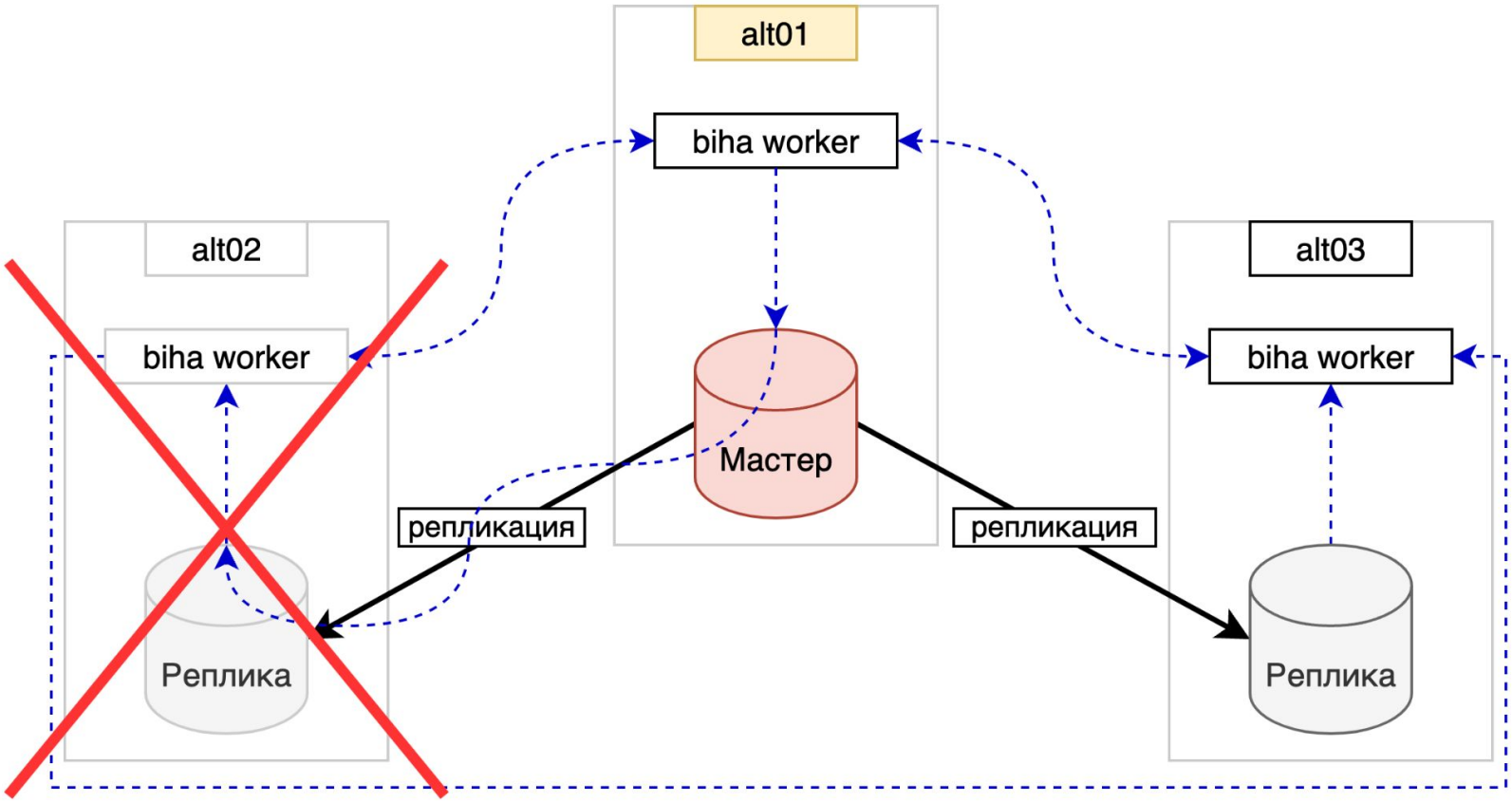
BiHA: отказы


alt01
biha worker
alt02
biha worker
alt03
biha worker
Мастер
репликация
Реплика
Реплика

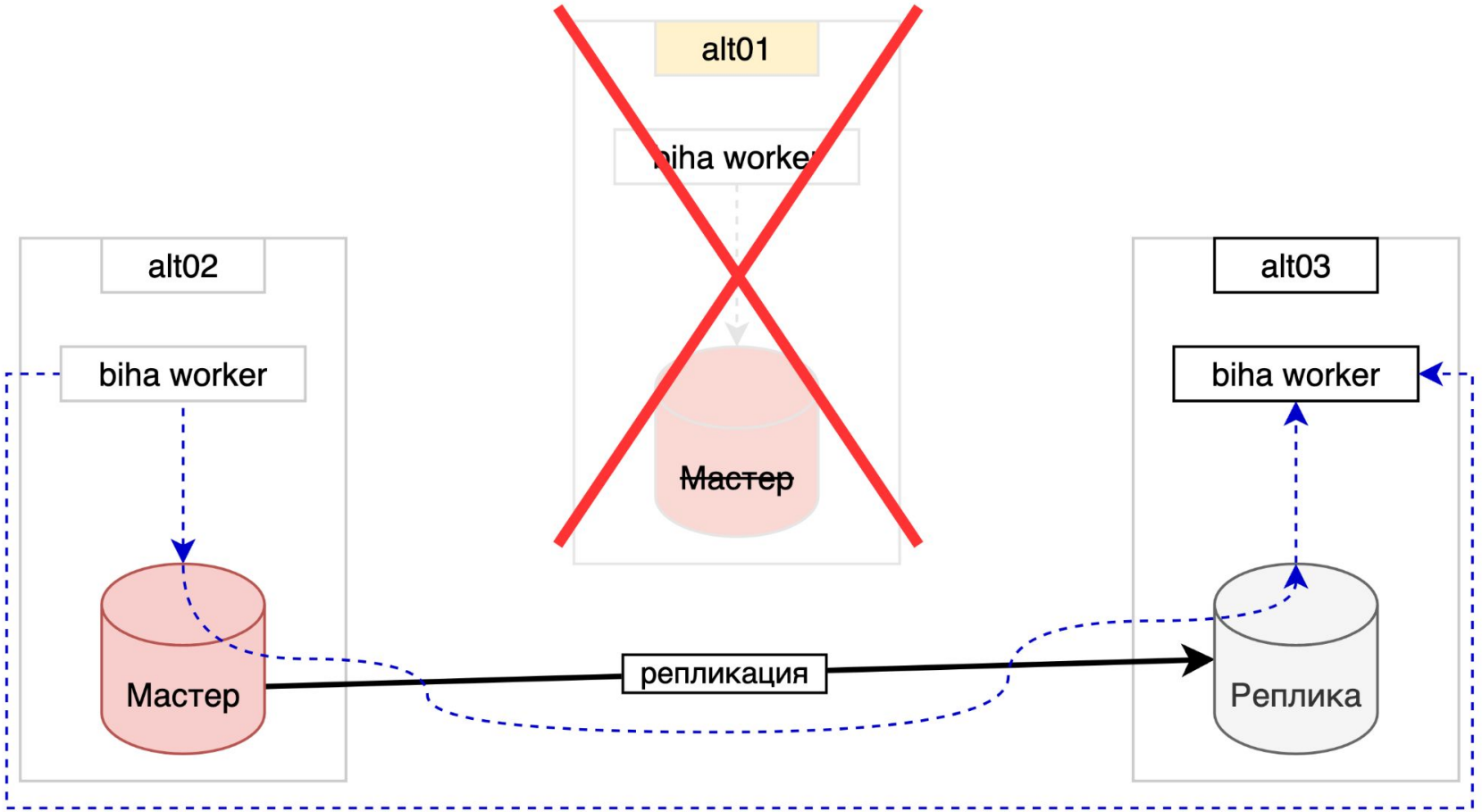
BiHA: отказы


alt01
biha worker
alt02
biha worker
alt03
biha worker
Мастер
репликация
репликация
Реплика
Реплика

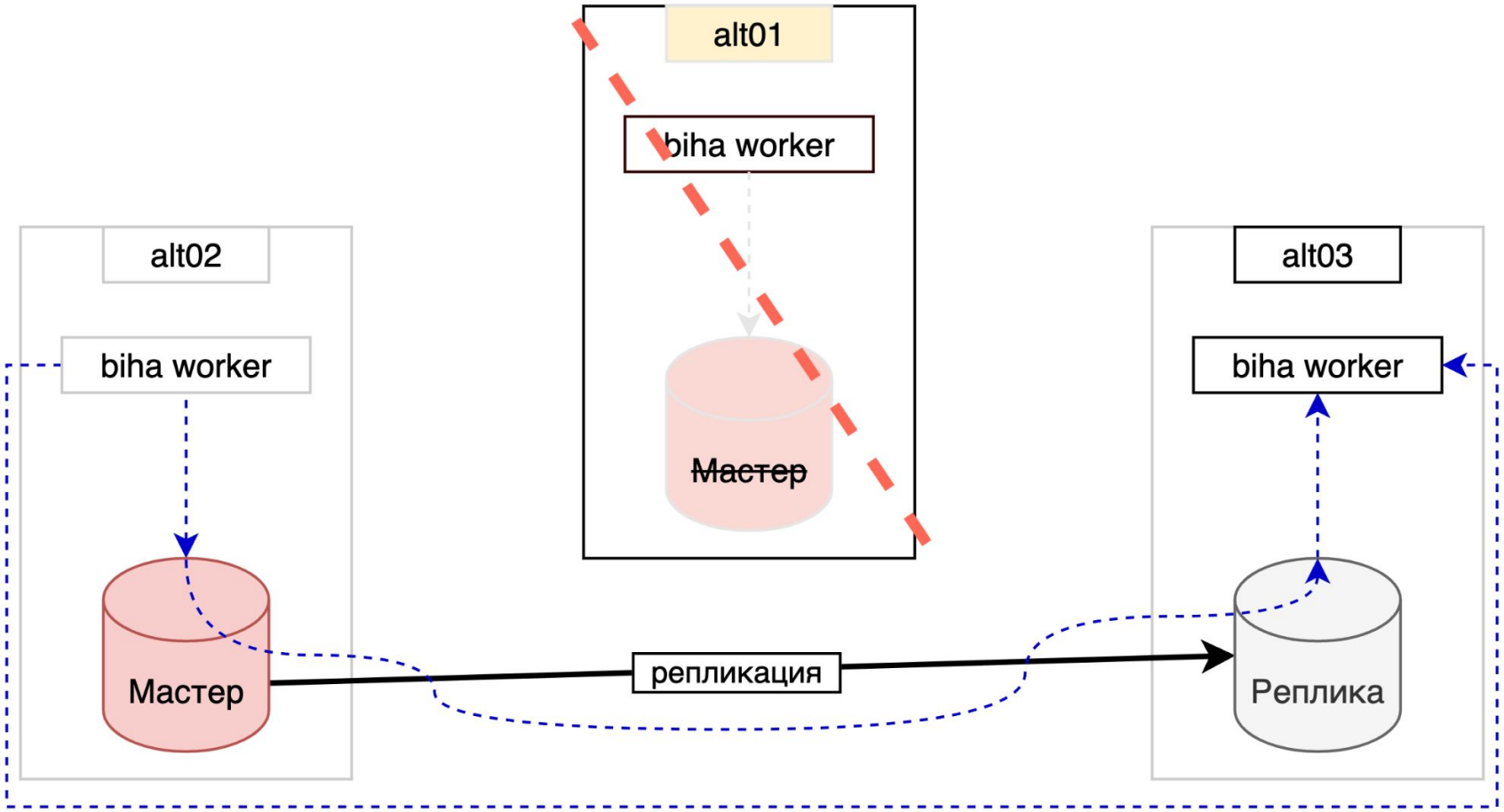
BiHA: отказы


alt01
biha worker
alt02
biha worker
alt03
biha worker
Мастер
репликация
репликация
Реплика
Реплика

BiHA: отказы


alt01
biha worker
Мастер
alt02
biha worker
Мастер
репликация
alt03
biha worker
Реплика

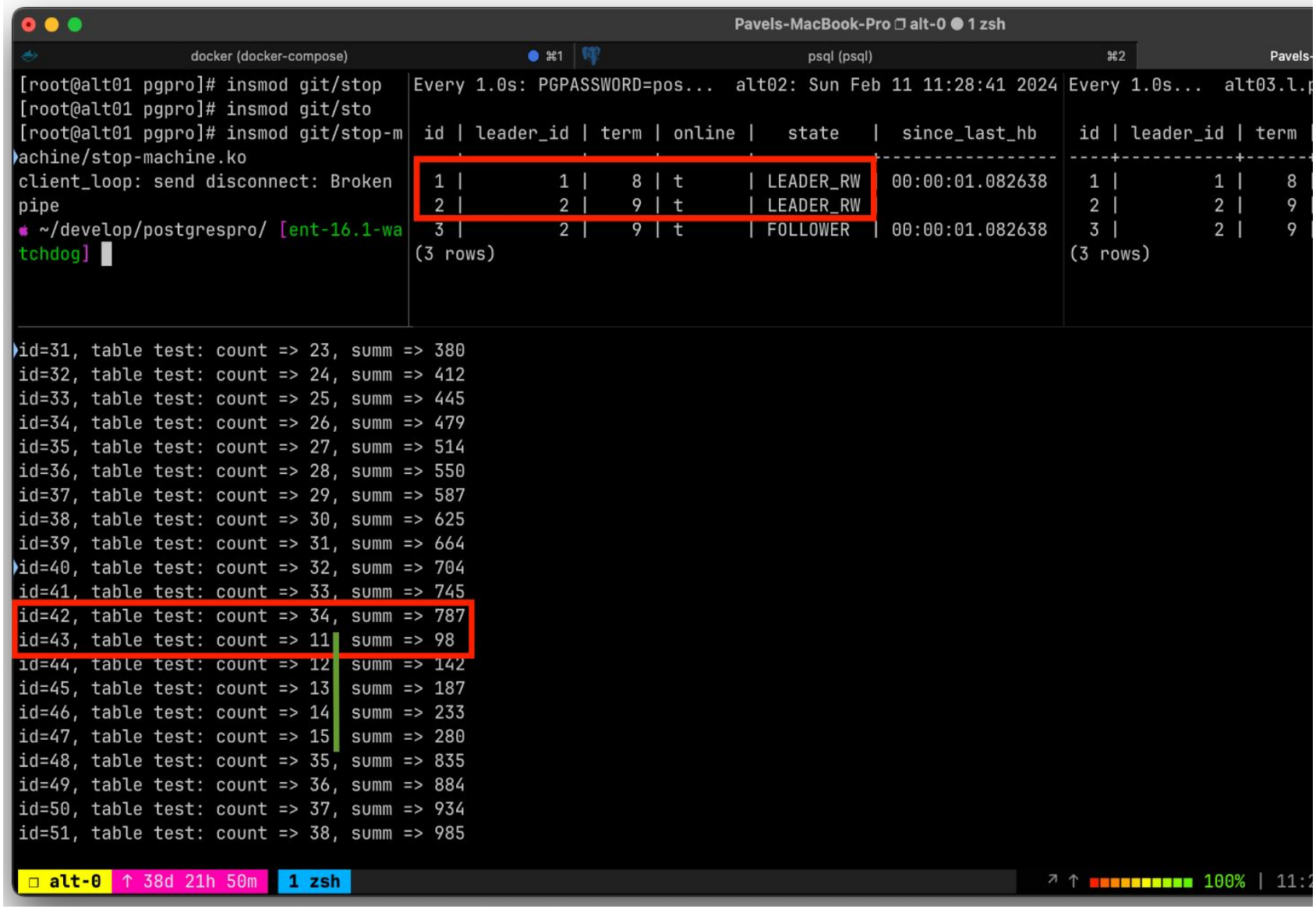
BiHA: «висящий мастер»



Фенсинг в BiHA: тест

```
biha.autorewind = on
```

```
# вставляем строку, считаем количество строк и сумму их значений

PG_URL="postgresql://postgres:postgres@alt01:5444,alt02:5444,alt03:5444/postgres?
target_session_attrs=read-write&connect_timeout=1"

for i in {1..100}; do

  printf "id=$i, table test: "

  /opt/pgpro/ent-16/bin/psql -d $PG_URL -U postgres -Atc \

  "INSERT INTO test VALUES($i); \

  SELECT 'count => ' || count(id) || ', summ => ' || sum(id) FROM test;"

  sleep 0.2

done
```


Pavels-MacBook-Pro alt-0 1 zsh
docker (docker-compose)
psql (psql)
[root@alt01 pgpro]# insmod git/stop
[root@alt01 pgpro]# insmod git/sto
[root@alt01 pgpro]# insmod git/stop-m
achine/stop-machine.ko
client_loop: send disconnect: Broken pipe
~/develop/postgrespro/ [ent-16.1-watchdog]
Every 1.0s: PGPASSWORD=pos... alt02: Sun Feb 11 11:28:41 2024
id | leader_id | term | online | state | since_last_hb
1 | 1 | 8 | t | LEADER_RW | 00:00:01.082638
2 | 2 | 9 | t | LEADER_RW |
3 | 2 | 9 | t | FOLLOWER | 00:00:01.082638
(3 rows)
Every 1.0s... alt03.l.p
id | leader_id | term
1 | 1 | 8
2 | 2 | 9
3 | 2 | 9
(3 rows)
id=31, table test: count => 23, summ => 380
id=32, table test: count => 24, summ => 412
id=33, table test: count => 25, summ => 445
id=34, table test: count => 26, summ => 479
id=35, table test: count => 27, summ => 514
id=36, table test: count => 28, summ => 550
id=37, table test: count => 29, summ => 587
id=38, table test: count => 30, summ => 625
id=39, table test: count => 31, summ => 664
id=40, table test: count => 32, summ => 704
id=41, table test: count => 33, summ => 745
id=42, table test: count => 34, summ => 787
id=43, table test: count => 11 summ => 98
id=44, table test: count => 12 summ => 142
id=45, table test: count => 13 summ => 187
id=46, table test: count => 14 summ => 233
id=47, table test: count => 15 summ => 280
id=48, table test: count => 35, summ => 835
id=49, table test: count => 36, summ => 884
id=50, table test: count => 37, summ => 934
id=51, table test: count => 38, summ => 985
alt-0 38d 21h 50m 1 zsh
100%

Фенсинг в BiHA

PostgresPro

- **Находится в стадии разработки**
- **Но мы же инженеры! :)** 
- **Попробуем найти выход для случаев**
  - «Завис» PostgreSQL/BiHA bgworker
  - «Заморозилась» виртуальная машина

Фенсинг в **BiHA:** СУБД как сервис

PostgresPro

- **Используем systemd-watchdog**
- **Необходимо сделать:**
  - На уровне ОС
    - задействовать механизм нотификации
      - в unit файле включить настройку systemd-watchdog
  - В ядре PostgreSQL
    - использовать системный вызов **sd_notify()**
    - в коде найти удобное место для патча
- **Systemd будет контролировать работу PostgreSQL**
- **Если postmaster теряет отзывчивость, systemd остановит сервис (kill -9)**

Фенсинг в **BiHA:** патч для **systemd watchdog**

PostgresPro

```
#ifdef USE_SYSTEMD
/* check systemd watchdog is enabled, read polling interval from systemd unit */
    isWatchdogEnabled = sd_watchdog_enabled(0, &watchdogIntInUs) > 0;

    if (isWatchdogEnabled) {
        ereport(LOG,
                errmsg("systemd watchdog is enabled, watchdog polling interval %ld",
                       watchdogIntInUs));
    } else {
        ereport(LOG,
                errmsg("systemd watchdog disabled"));
    }
#endif
```

Фенсинг в **BiHA:** патч для **systemd watchdog**

PostgresPro

```
ServerLoop()
#ifdef USE_SYSTEMD
/* sending signal to systemd watchdog when we are cheking postgres.pid file */
if (isWatchdogEnabled)
   {
    sd_notify(0, "WATCHDOG=1");
    ereport(LOG,
    (errmsg("notifying systemd watchdog that postgres is alive, interval %ld sec",
            now - last_lockfile_recheck_time)));
   } else {
         ereport(LOG,(errmsg("can't notify systemd watchdog")));
      }
#endif
```

Фенсинг в **BiHA: unit** файл

```
[Service]
...
WatchdogSec=65s
```

```
pkill -SIGSTOP postgres
```

```
[1098]LOG: [BiHA BCP 2 ACTIVE] id 2, socket 25, state ACTIVE
[1098]LOG: [BiHA BCP 3 ACTIVE] id 3, socket 26, state ACTIVE
[1090]LOG: checking postmaster.pid file here if exists, interval 60 sec
[1090]LOG: notifying systemd watchdog that postgres is alive, interval 60 sec
```

Итоги

PostgresPro

- **Подумайте о том, нужен ли автоматический rewind**
- **Включите синхронную репликацию**
  - **Patroni**
    - `syncronous_mode: true`
    - `syncronous_mode_strict: true`
    - `syncronous_node_count: 1`
  - **BiHA**
    - `synchronous_commit: true`
    - `synchronous_standby_names = 'ANY 1 (node1,node2,node3)'`

Итоги

- **Используйте watchdog**
  - на уровне ОС или на уровне кластерного ПО
  - аппаратный
    - включите эмуляцию в среде виртуализации
  - или используйте программный
  - требуйте доступ к API гипервизора!
  - мониторинг ВМ на стороне гипервизора
- **Инструмент сравнения данных**
- **Резервная копия и архив WAL**

Ссылки

- https://www.freedesktop.org/software/systemd/man/latest/systemd.service.html
- Встроенный отказоустойчивый кластер – https://postgrespro.ru/docs/enterprise/16/biha

PostgresPro

Спасибо за внимание!

PostgresPro

Q&A