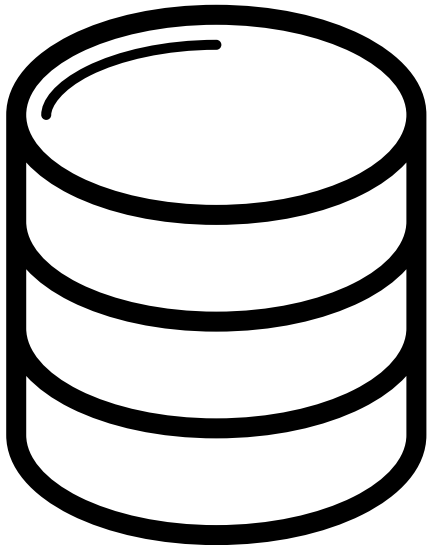
Построение крупных кластеров Tarantool из 100+ узлов

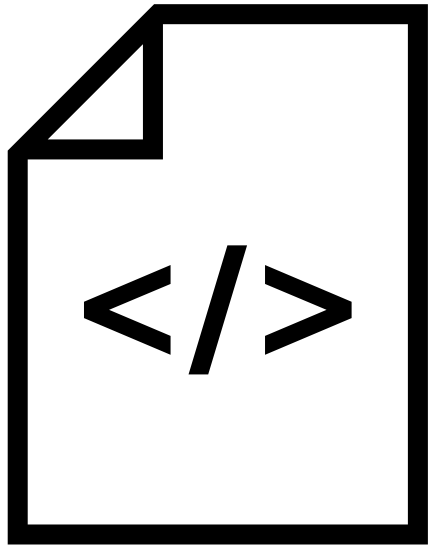
Ярослав Дынников

Tarantool, Mail.ru Group

Tarantool =



+



База данных

(Транзакции, WAL)

Сервер приложений (Lua)

(Бизнес логика, HTTP)

Ядро

- 20 разработчиков C
- Развитие продукта

Команда решений

- 35 разработчиков Lua
- Коммерческие проекты

Ядро

- 20 разработчиков C
- Развитие продукта

Команда решений

- 35 разработчиков Lua
- Коммерческие проекты

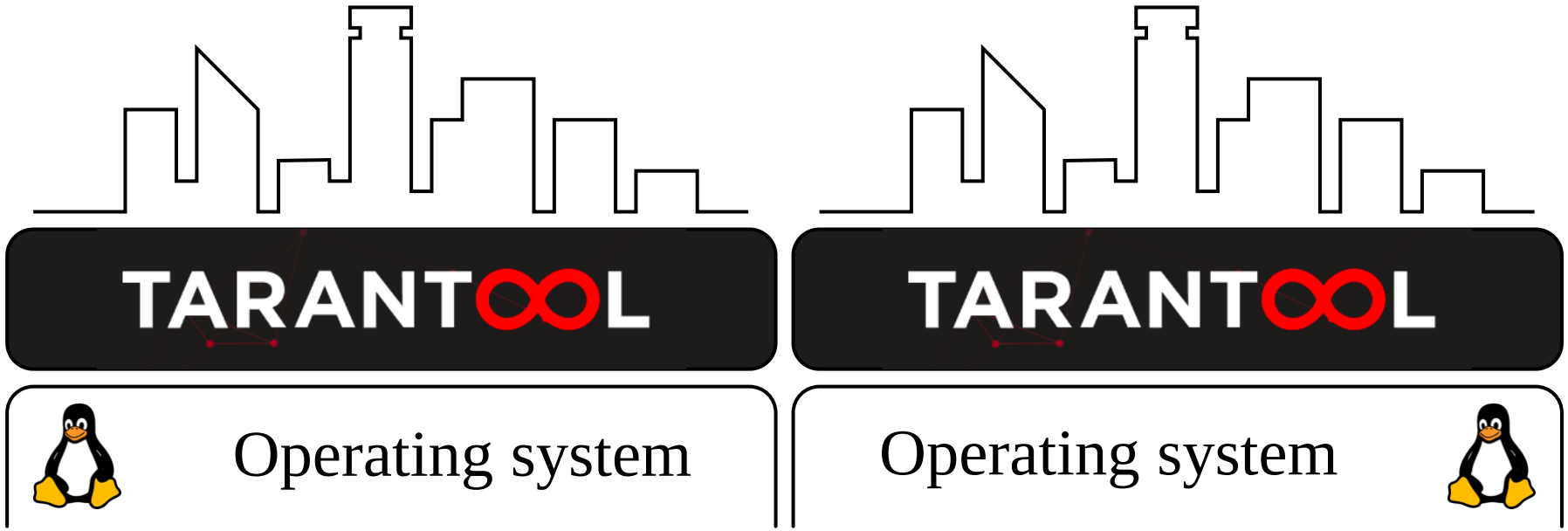
Цели

- Меньше багов, больше продуктивность

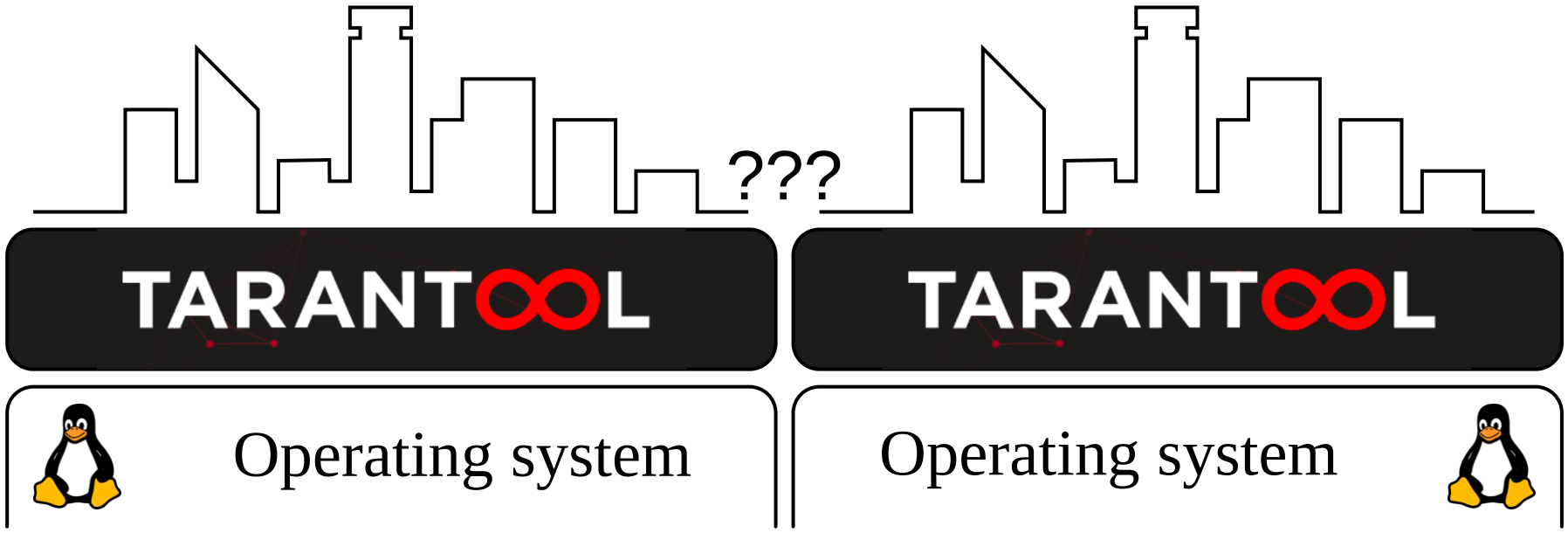
Инструменты разработчика

Инструменты разработчика

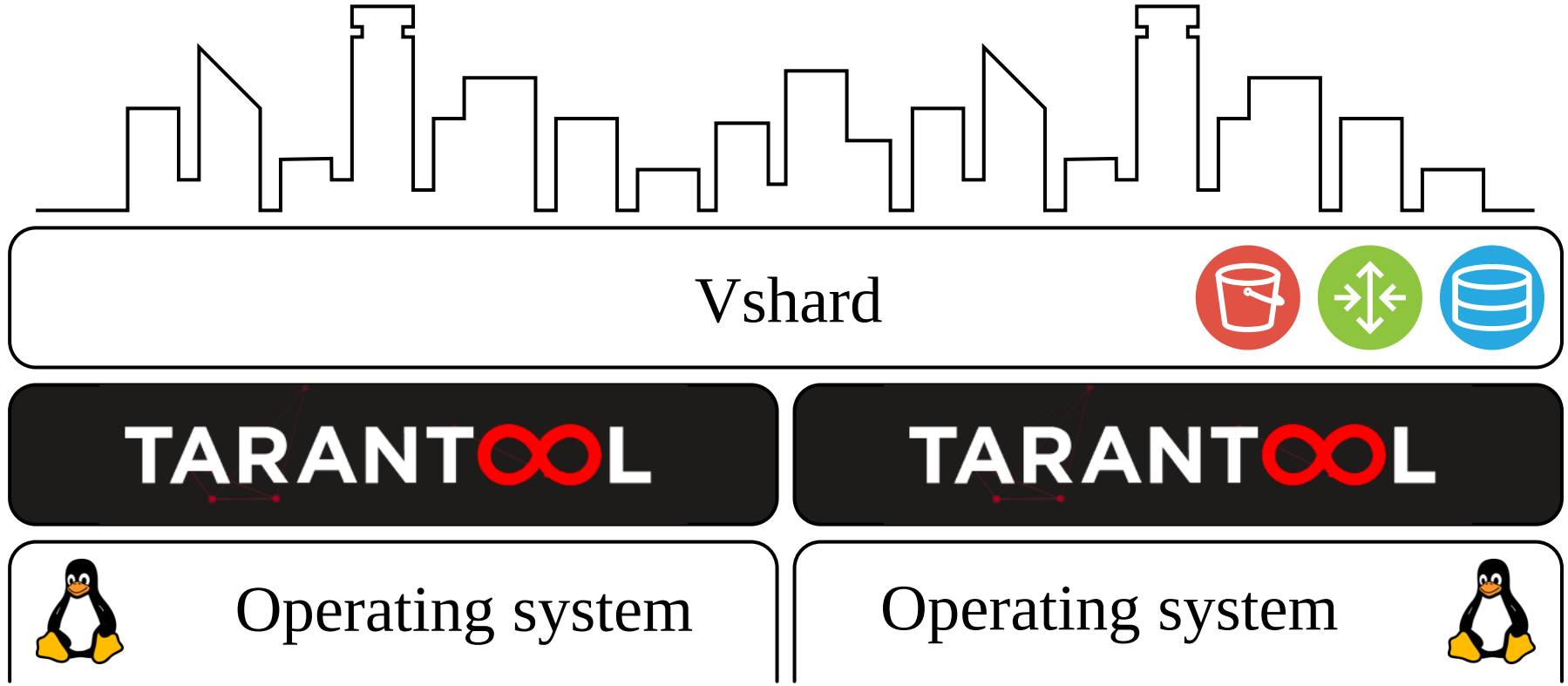
Инструменты разработчика



Инструменты разработчика



Инструменты разработчика



Конфигурация vshard

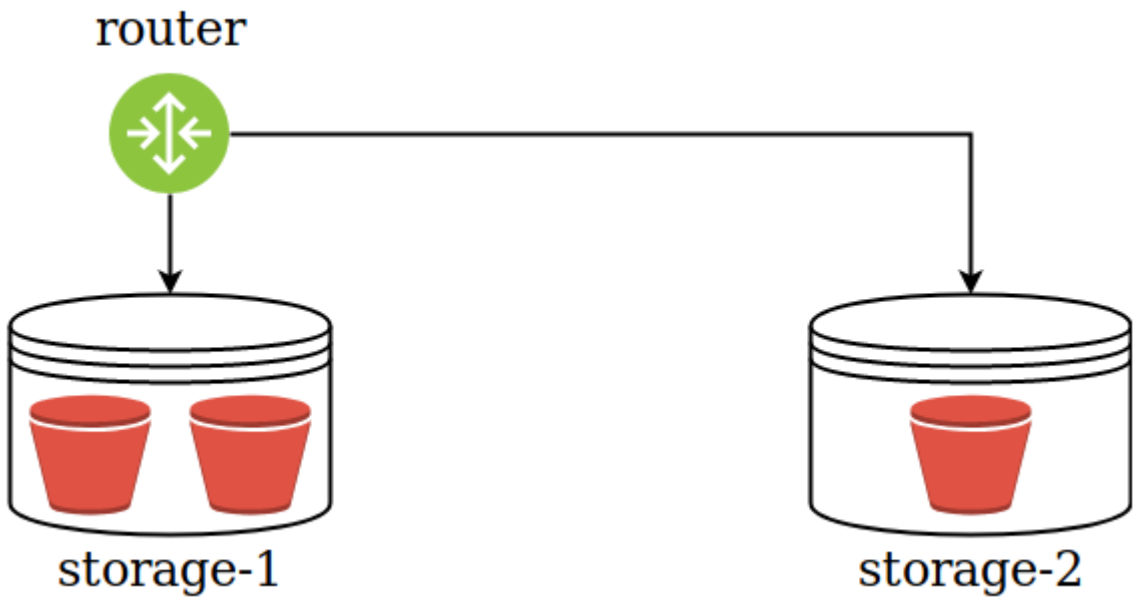
- Vshard управляется программно:

```
sharding_cfg = {
    ['cbf06940-0790-498b-948d-042b62cf3d29'] = {
        replicas = { ... },
    },
    ['ac522f65-aa94-4134-9f64-51ee384f1a54'] = {
        replicas = { ... },
    },
}
```

```
vshard.router.cfg(...)
vshard.storage.cfg(...)
```

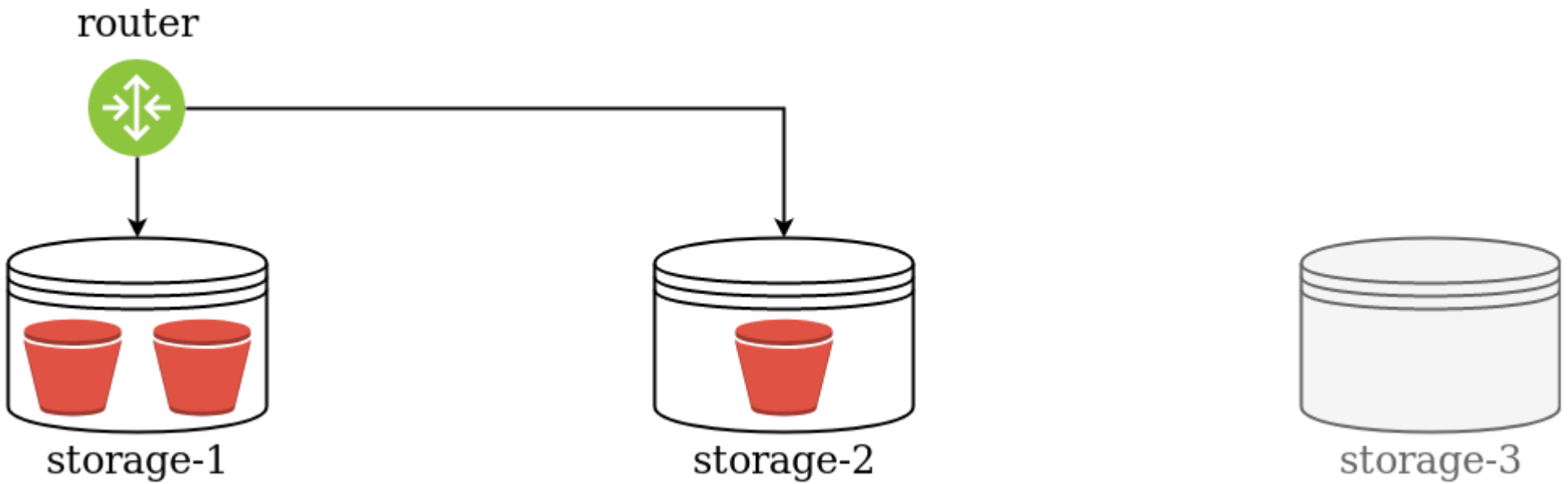
Vshard - горизонтальное масштабирование в Tarantool

- Vshard группирует данные по виртуальным "ведёркам"
- Ведёрок много, они распределены по серверам



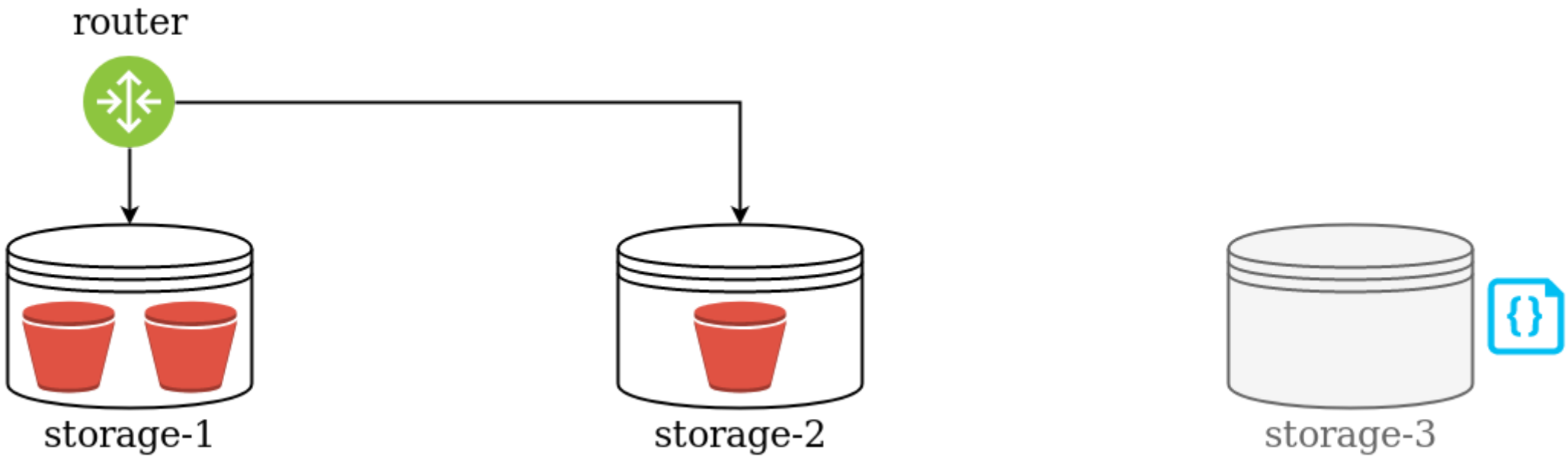
Vshard - горизонтальное масштабирование в Tarantool

- Vshard группирует данные по виртуальным "ведёркам"
- Ведёрок много, они распределены по серверам



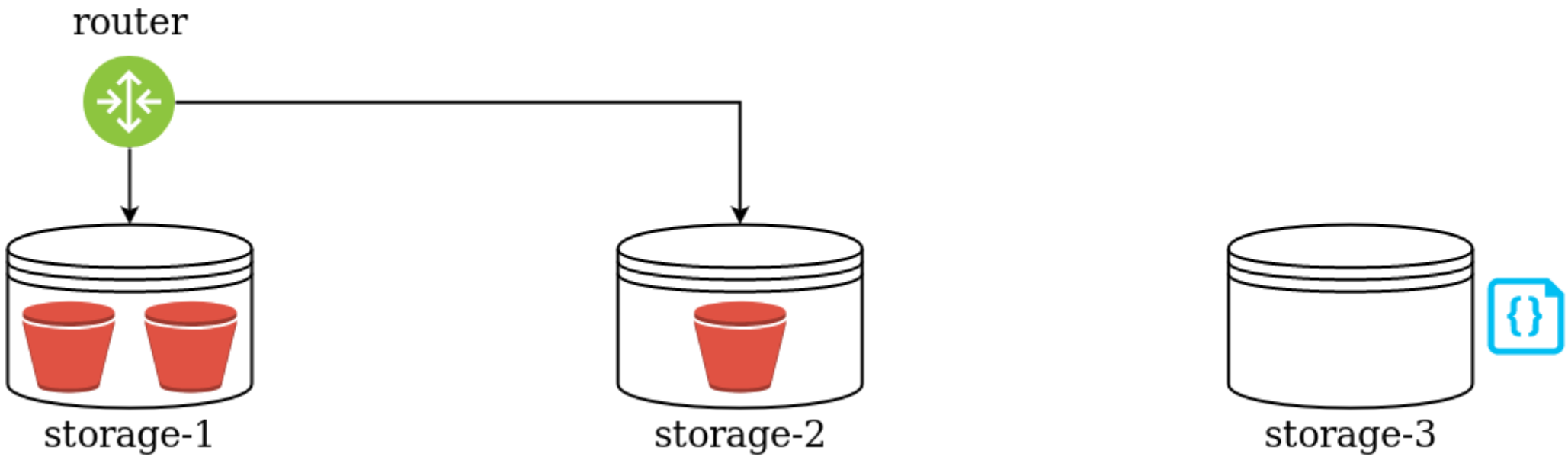
Vshard - горизонтальное масштабирование в Tarantool

- Vshard группирует данные по виртуальным "ведёркам"
- Ведёрок много, они распределены по серверам



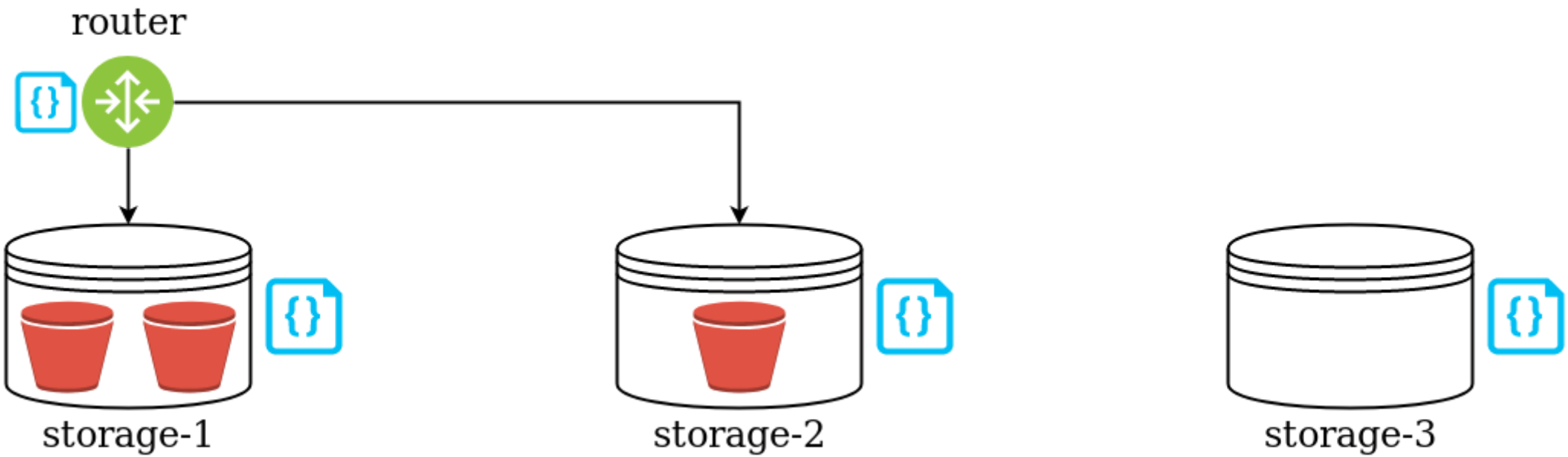
Vshard - горизонтальное масштабирование в Tarantool

- Vshard группирует данные по виртуальным "ведёркам"
- Ведёрок много, они распределены по серверам



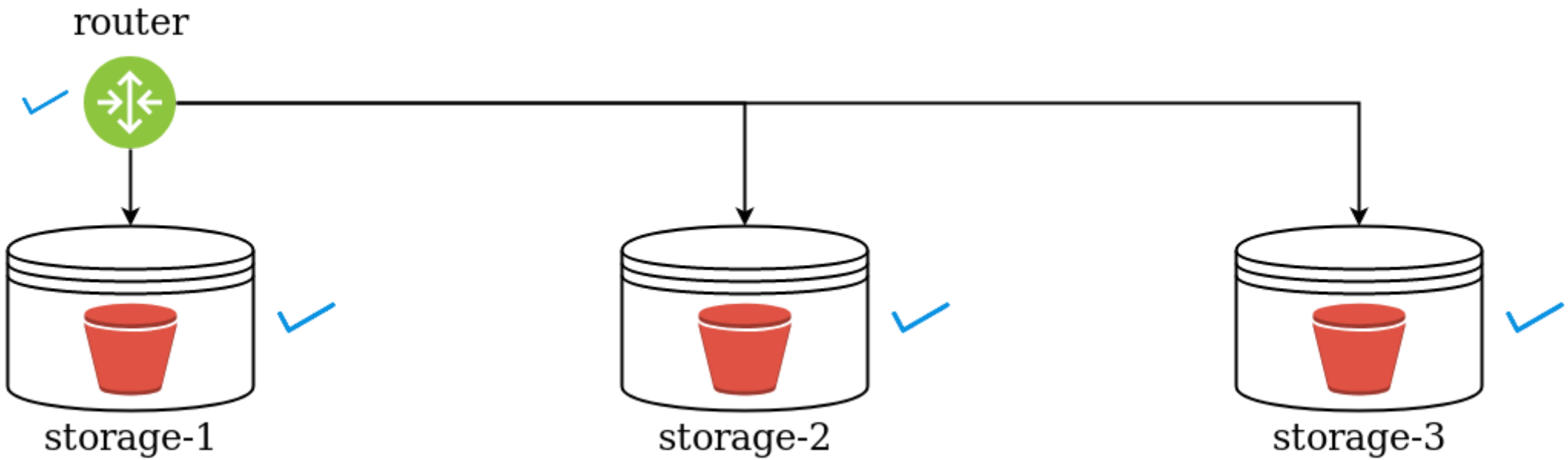
Vshard - горизонтальное масштабирование в Tarantool

- Vshard группирует данные по виртуальным "ведёркам"
- Ведёрок много, они распределены по серверам

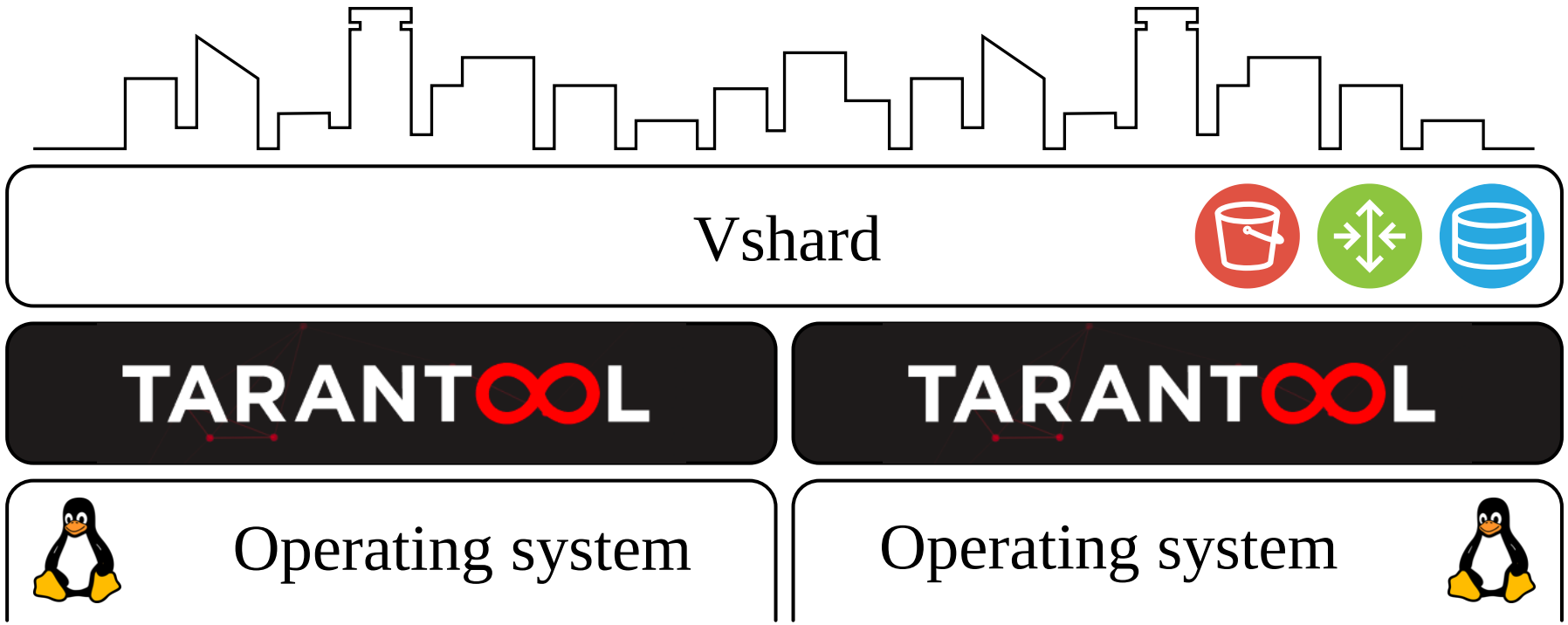


Vshard - горизонтальное масштабирование в Tarantool

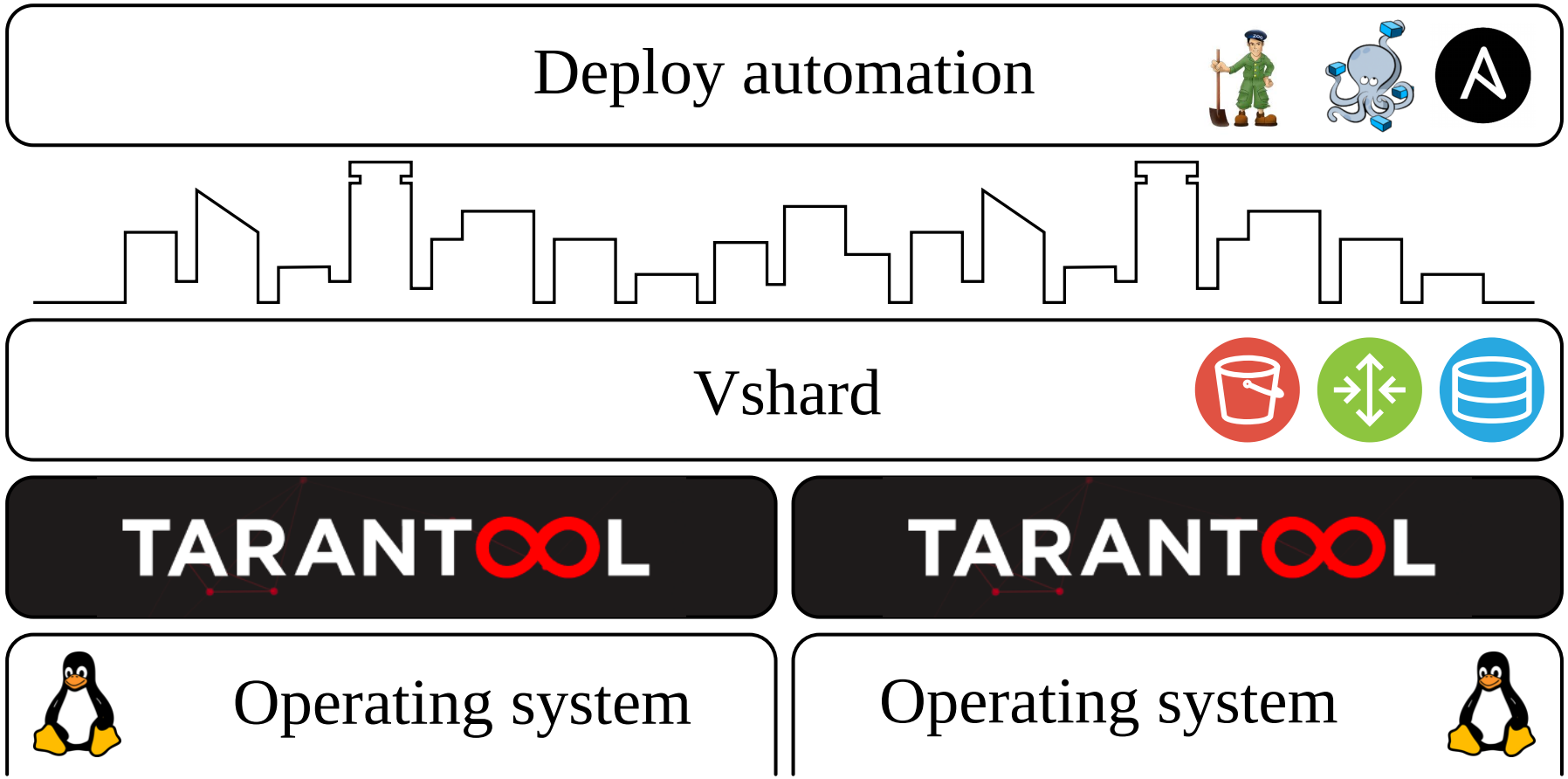
- Vshard группирует данные по виртуальным "ведёркам"
- Ведёрок много, они распределены по серверам



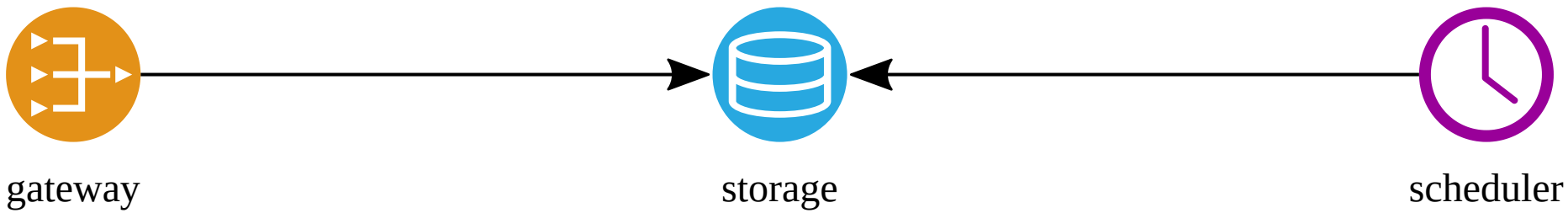
Оркестрация vshard



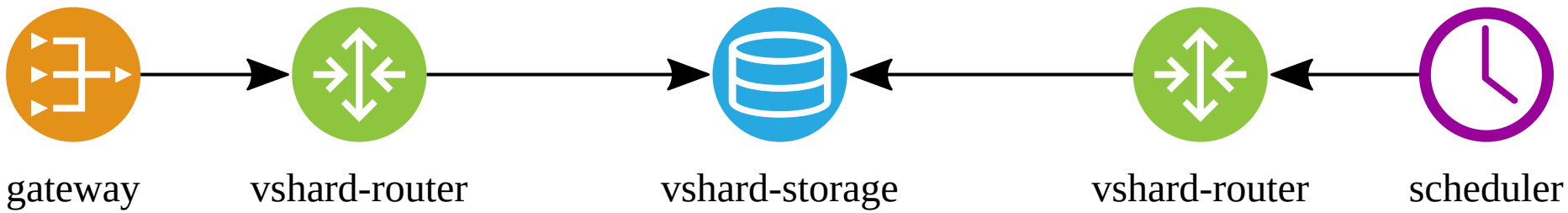
Оркестрация vshard



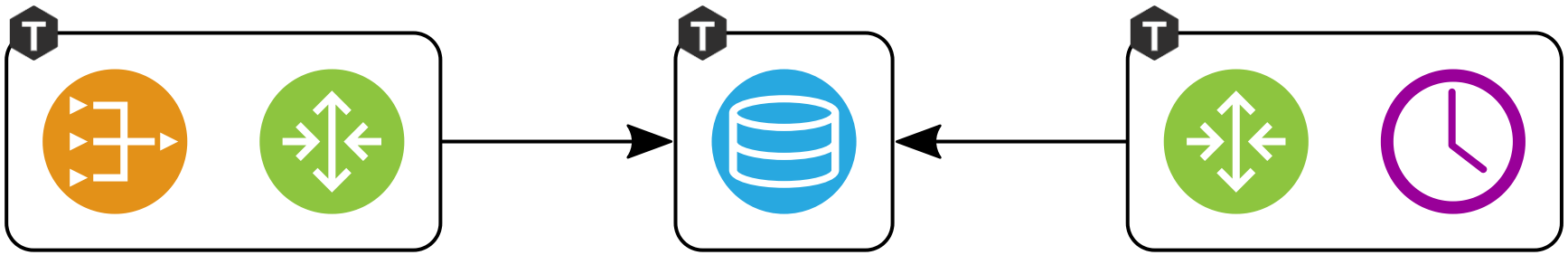
Организация бизнес логики



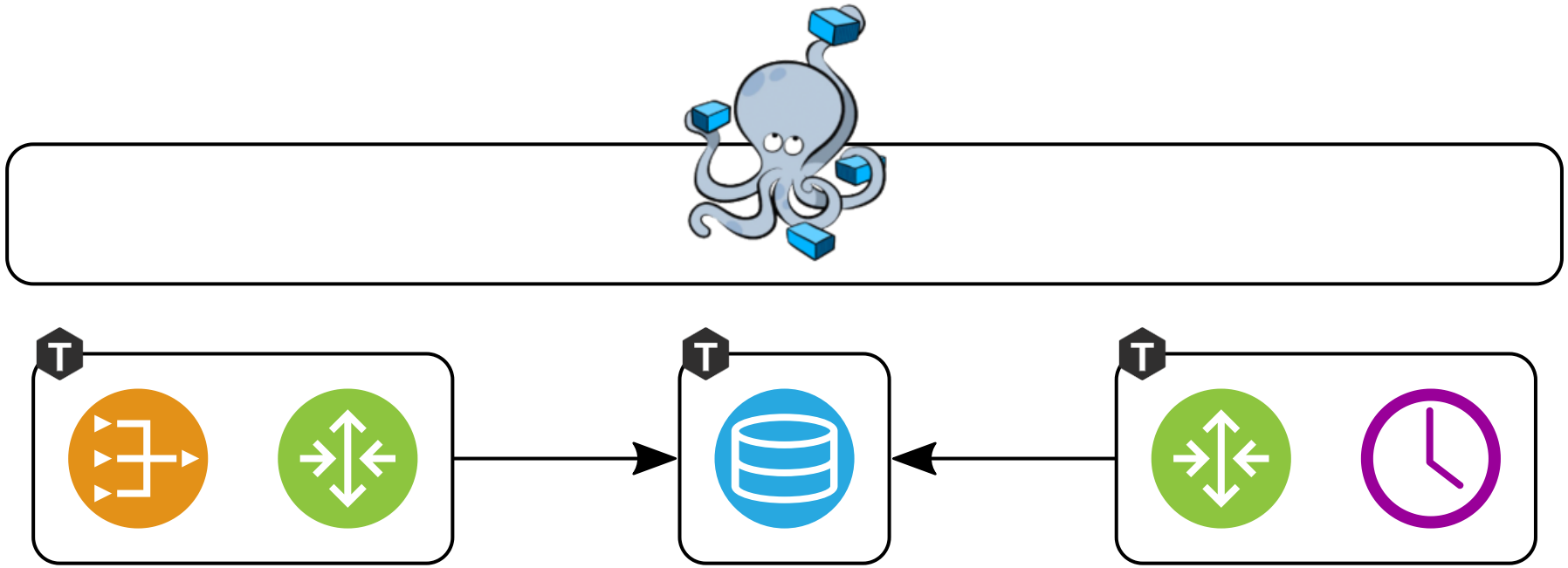
Организация бизнес логики



Организация бизнес логики

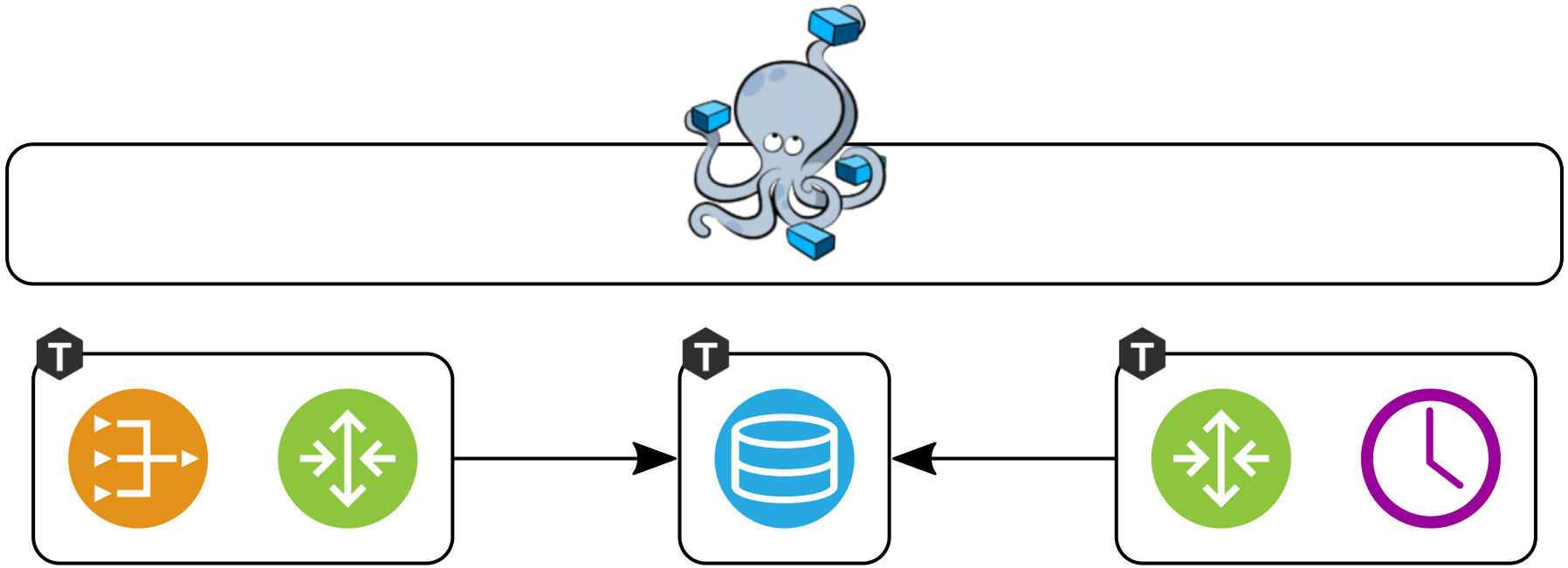


Организация бизнес логики



Организация бизнес логики

Как тестировать?

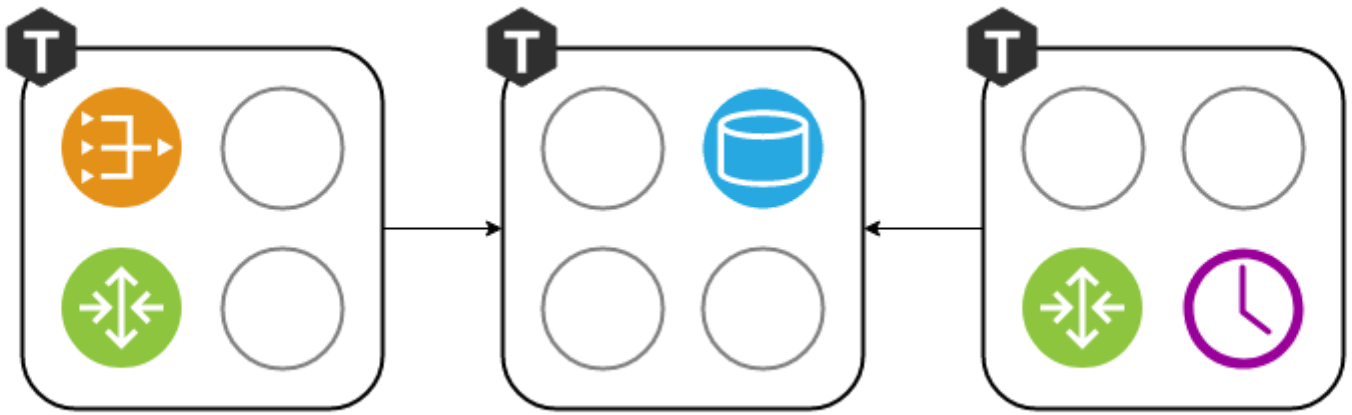


Как должно выглядеть управление кластером?

- Мы запускаем процессы
- Конфигурацией vshard управляет Tarantool

Как должно выглядеть управление кластером?

- Мы запускаем процессы
- Конфигурацией vshard управляет Tarantool

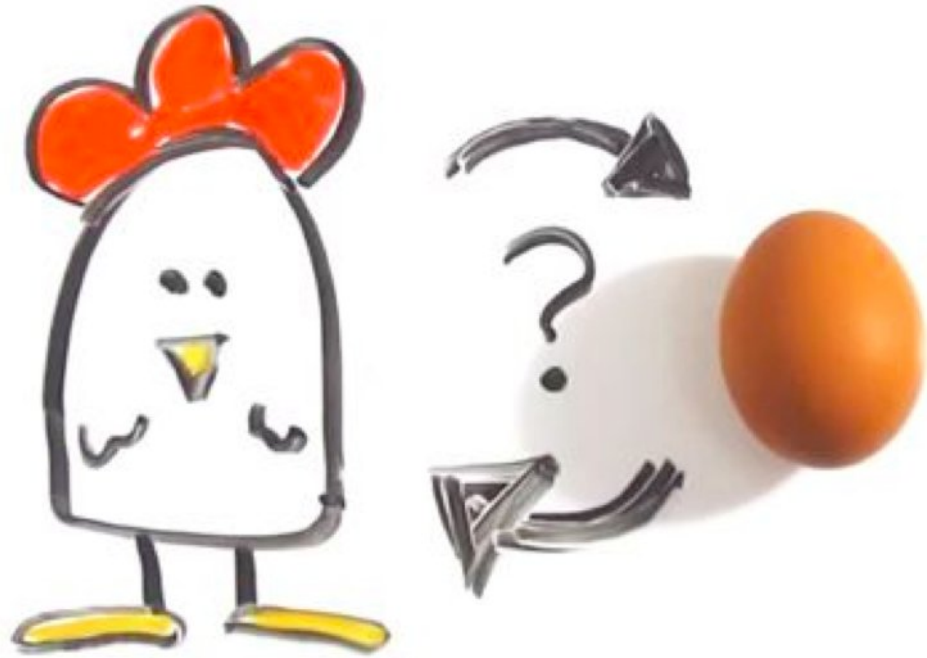


Конфигурация кластера

- Одинаковая **везде**
- Обновляется двухфазным коммитом
- Самое главное **- топология**:

```
topology:
  servers:                # two servers
    s1:
      replicaset_uuid: A
      uri: localhost:3301
    s2:
      replicaset_uuid: A
      uri: localhost:3302
  replicasets:            # one replicaset
    A:
      roles: ...
```

Курица или яйцо?

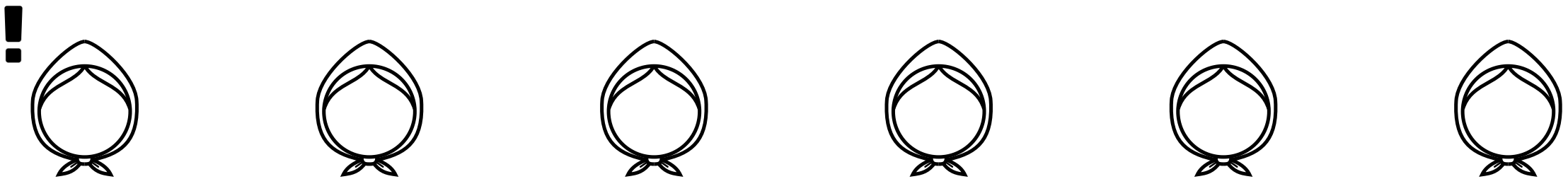


База данных
`net.box call`

Новый процесс
`box.cfg listen`

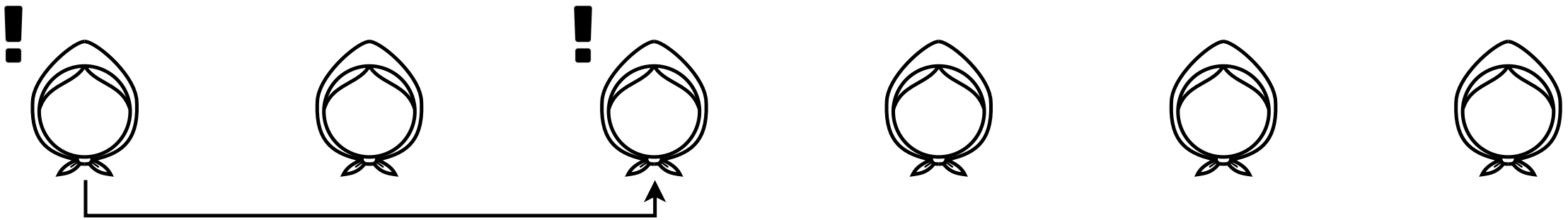
Внутренний мониторинг

- Протокол SWIM - распространение **слухов** по UDP



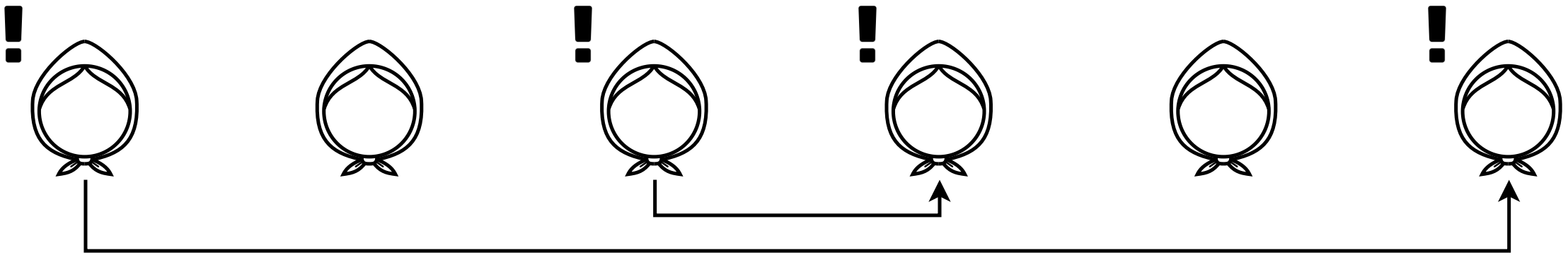
Внутренний мониторинг

- Протокол SWIM - распространение **слухов** по UDP



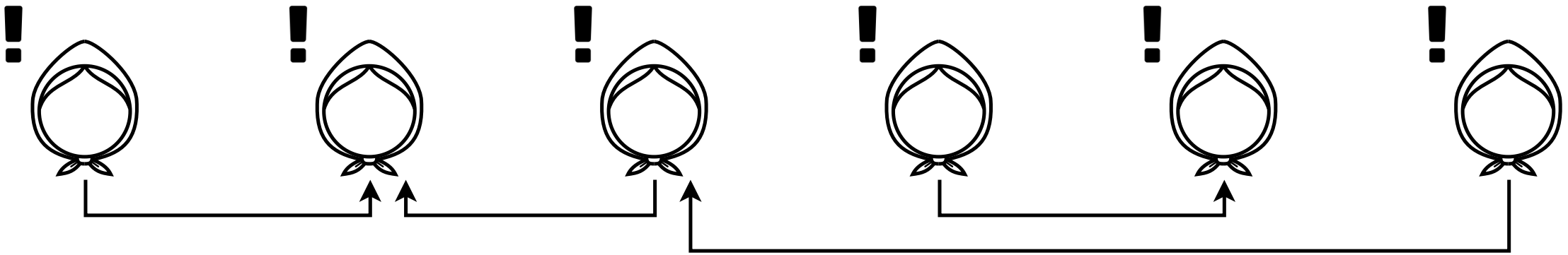
Внутренний мониторинг

- Протокол SWIM - распространение **слухов** по UDP



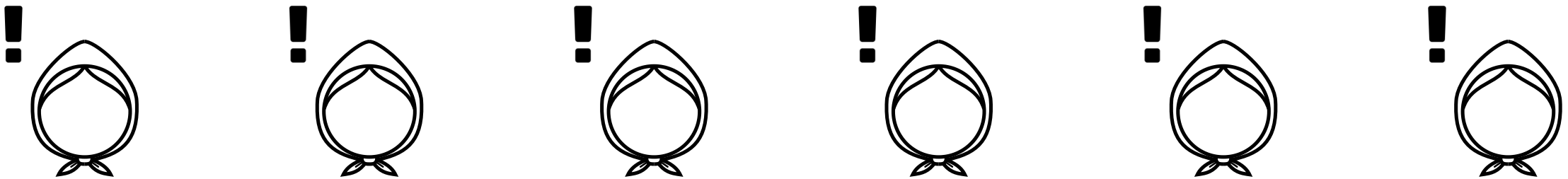
Внутренний мониторинг

- Протокол SWIM - распространение **слухов** по UDP



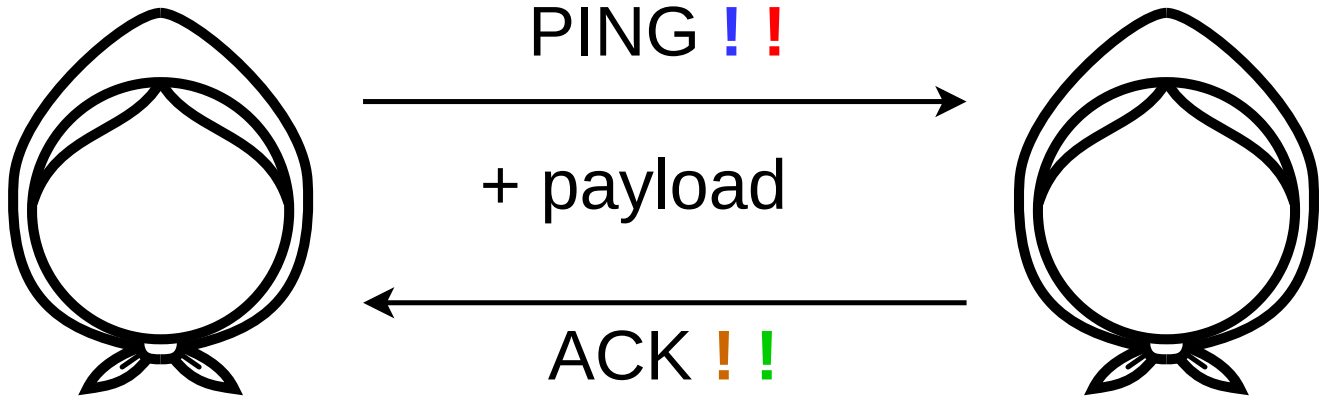
Внутренний мониторинг

- Протокол SWIM - распространение **слухов** по UDP
- Всеобщая осведомлённость: O(logN)
- Нагрузка на сеть: O(N)

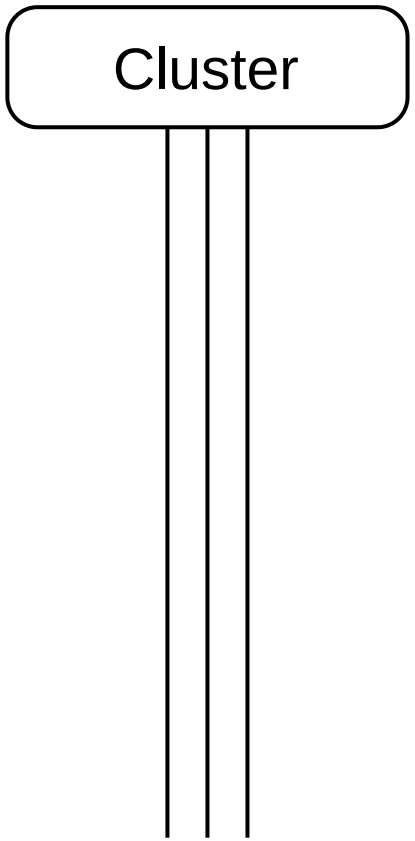


Внутренний мониторинг

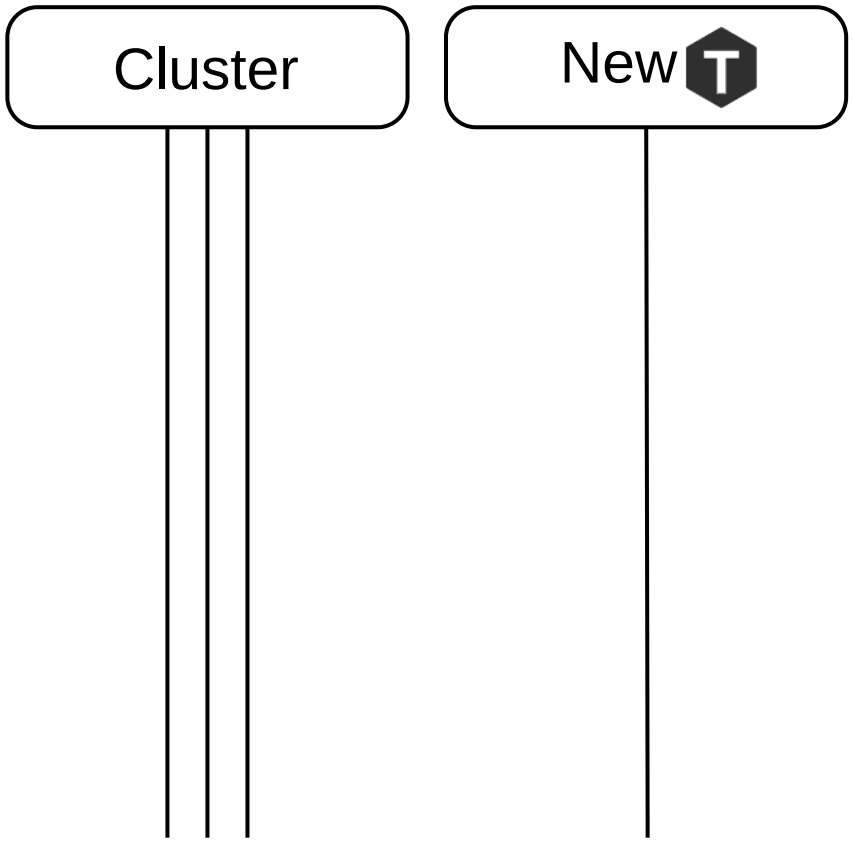
- Протокол SWIM - распространение **слухов** по UDP
- Всеобщая осведомлённость: O(logN)
- Нагрузка на сеть: O(N)



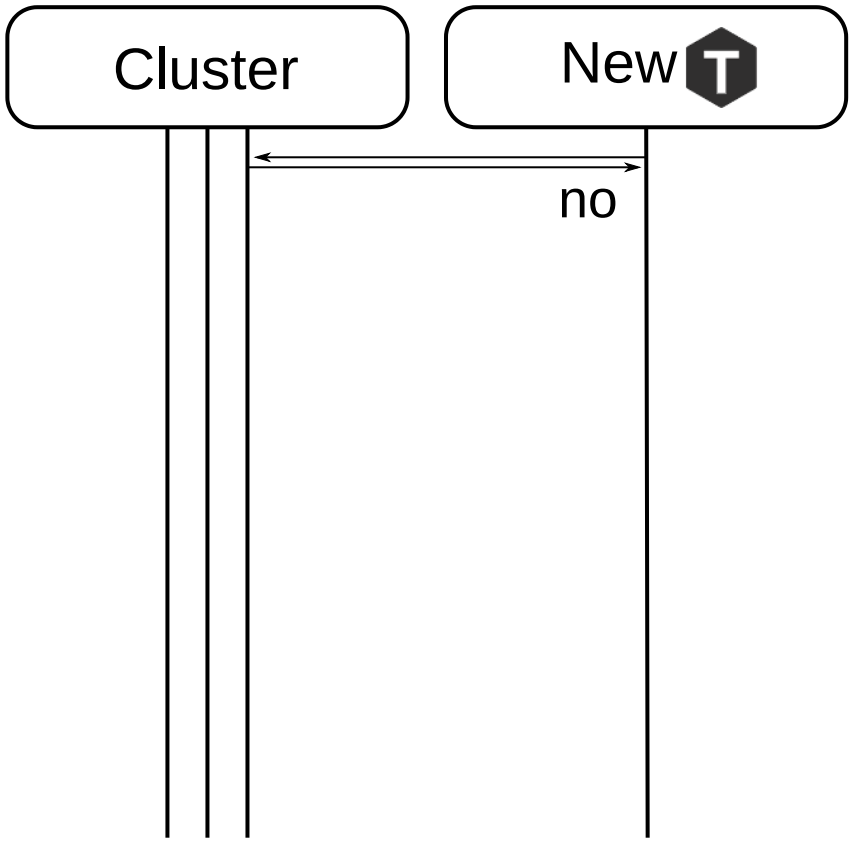
Прототип кластера



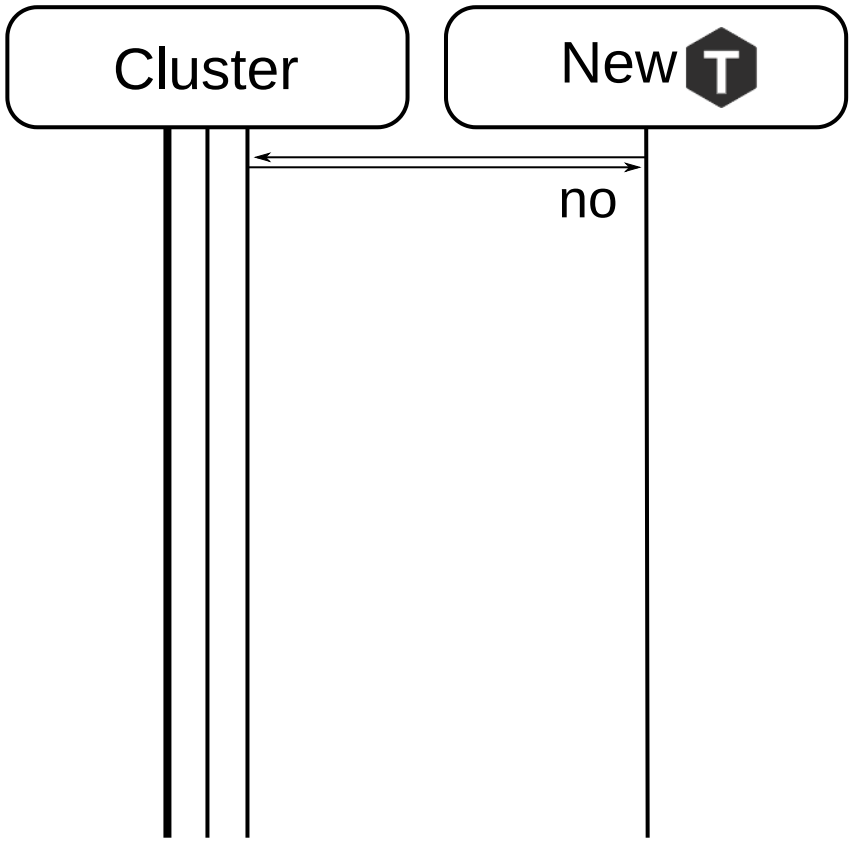
Прототип кластера



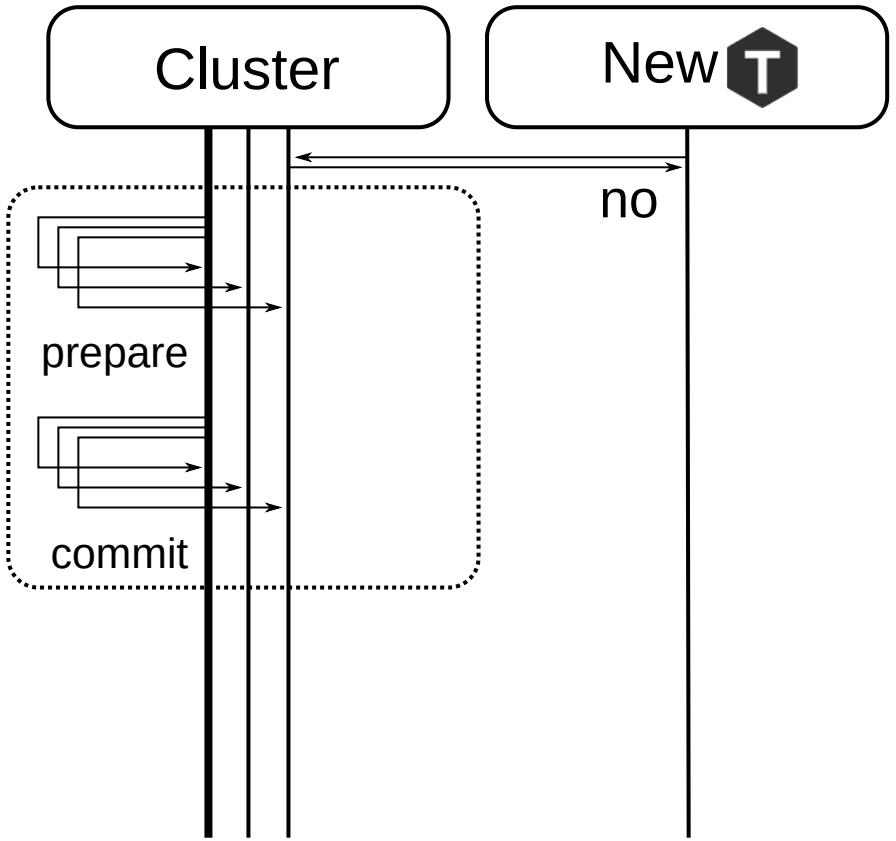
Прототип кластера



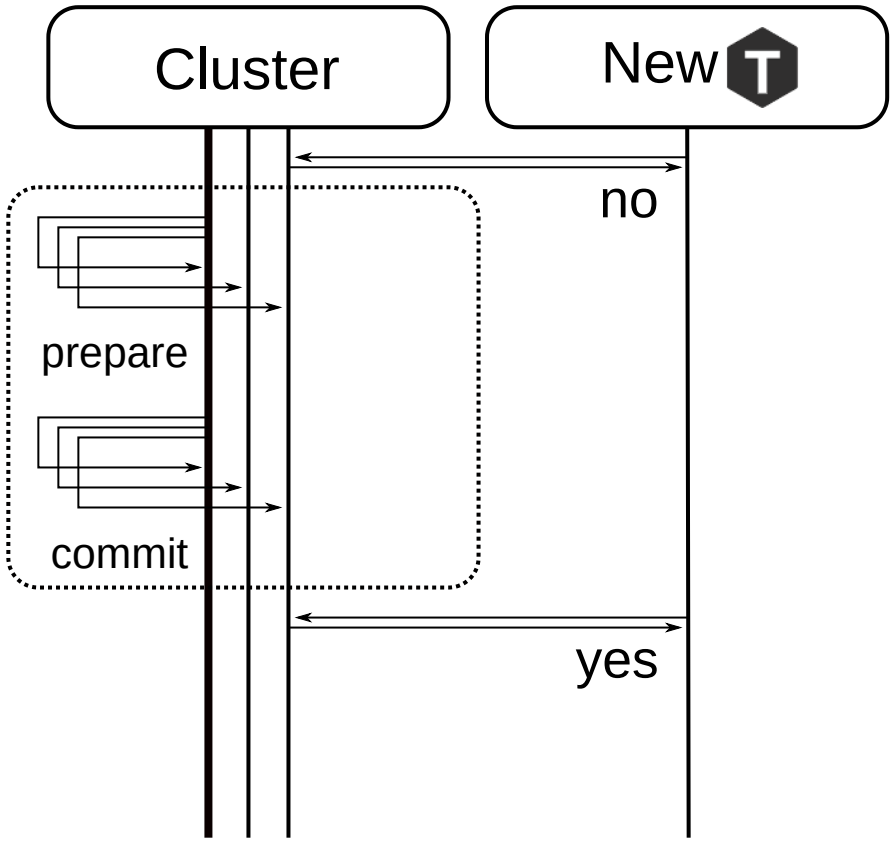
Прототип кластера



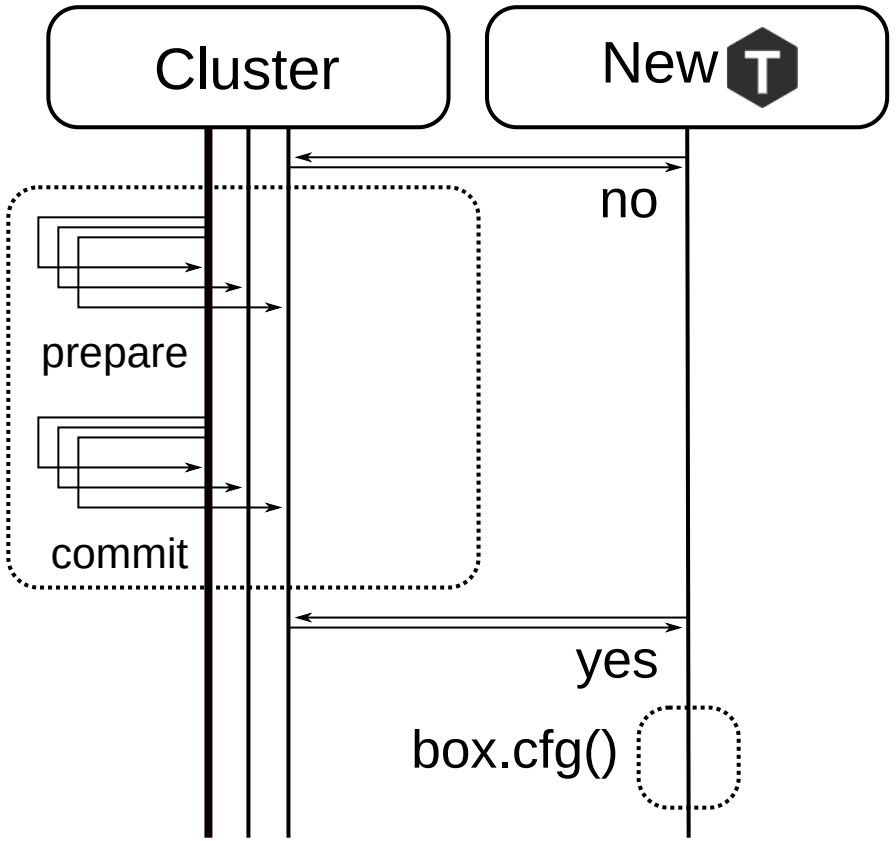
Прототип кластера



Прототип кластера



Прототип кластера

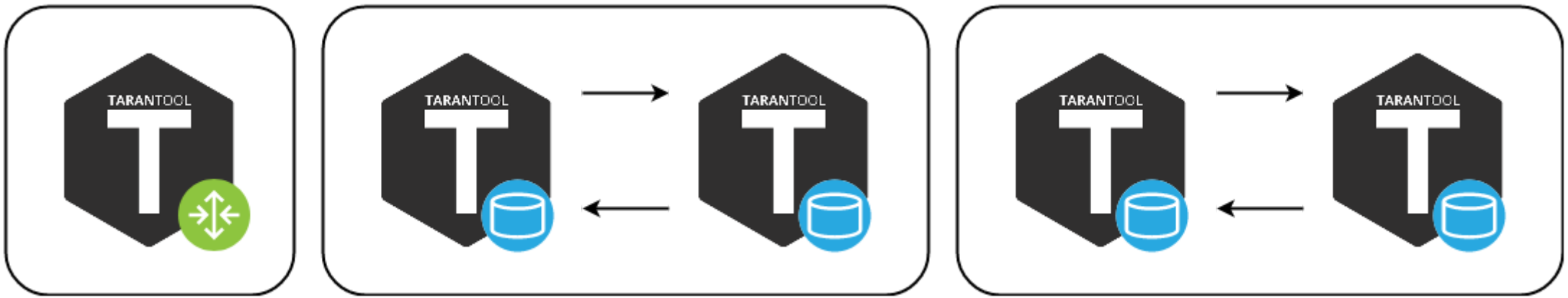


Структура кластера

1. Сервера повсюду

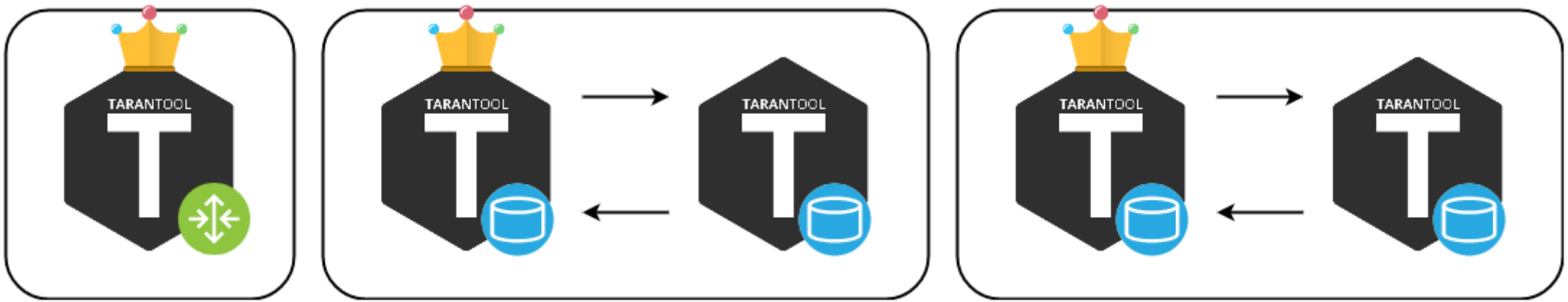
Структура кластера

1. Сервера повсюду (объединены в репликационные группы)



Структура кластера

1. Сервера повсюду (объединены в репликационные группы)
2. В каждой группе есть "лидер"



Структура кластера

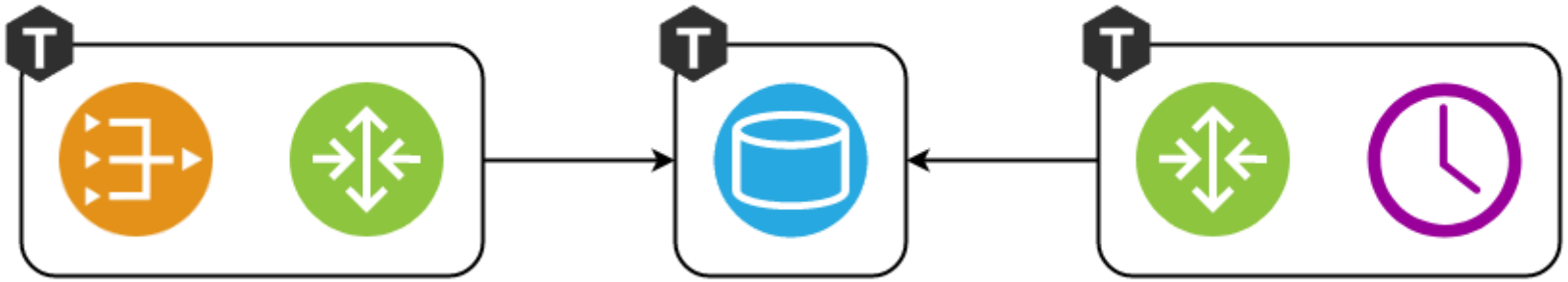
1. Сервера повсюду (объединены в репликационные группы)
2. В каждой группе есть "лидер"
3. Внутренний мониторинг позволяет обрабатывать отказы

Структура кластера

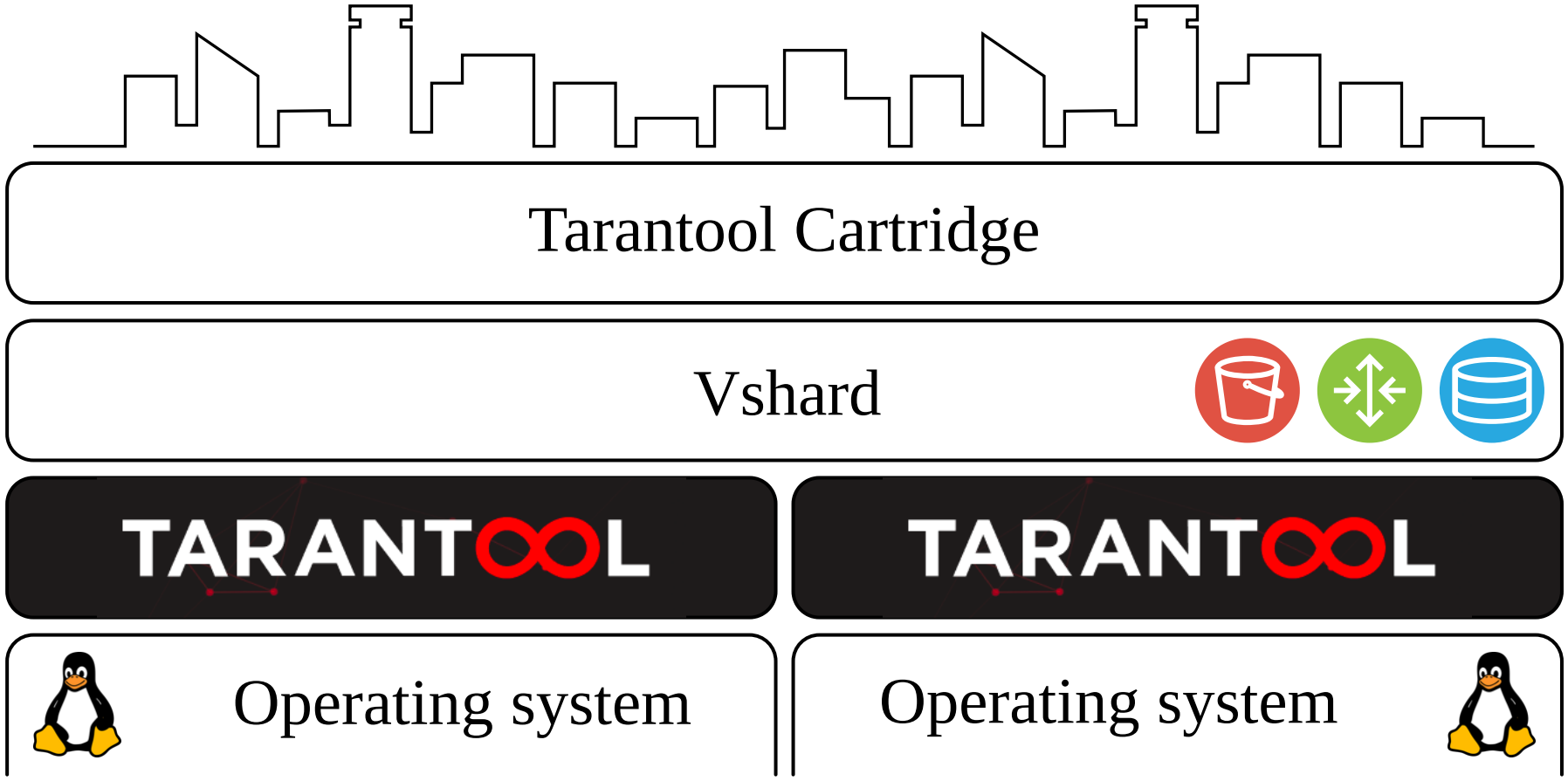
1. Сервера повсюду (объединены в репликационные группы)
2. В каждой группе есть "лидер"
3. Внутренний мониторинг позволяет обрабатывать отказы

Разработка кастомных ролей

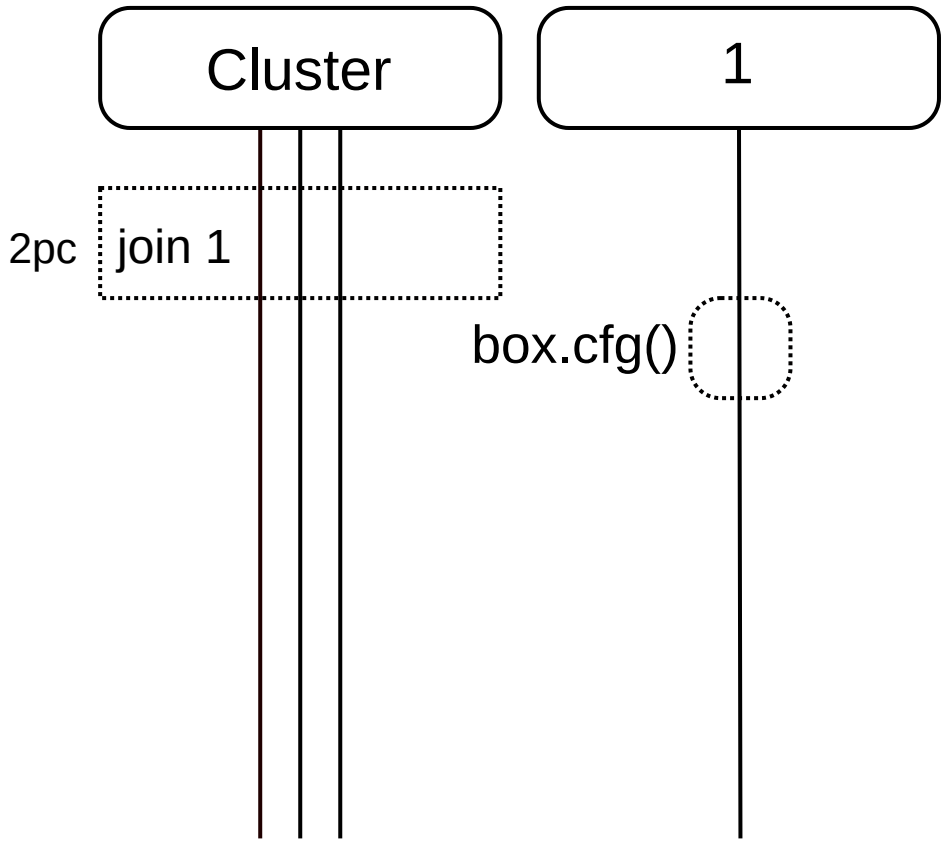
- `function init()`
- `function validate_config()`
- `function apply_config()`
- `function stop()`



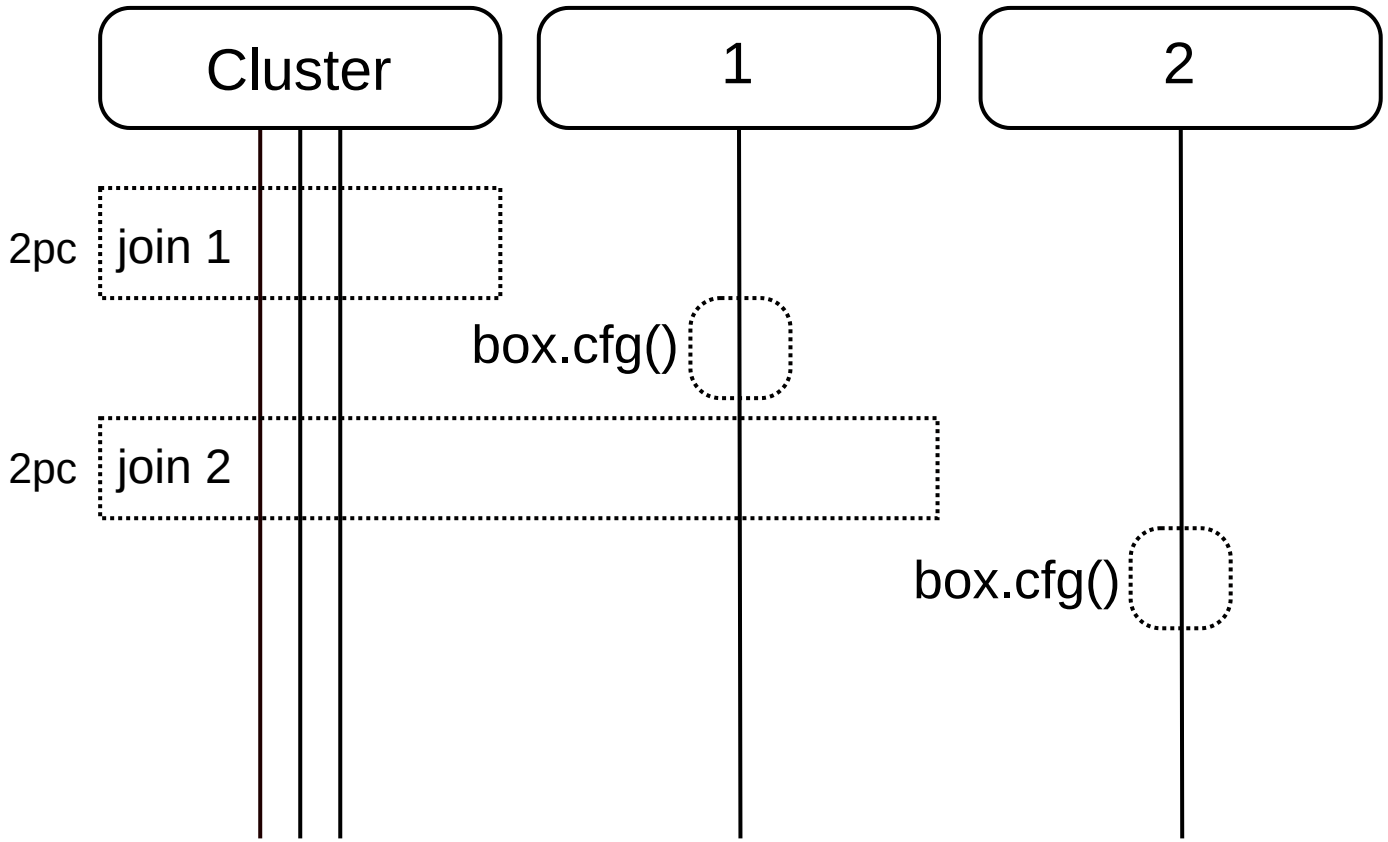
Фреймворк разработки распределённых сервисов



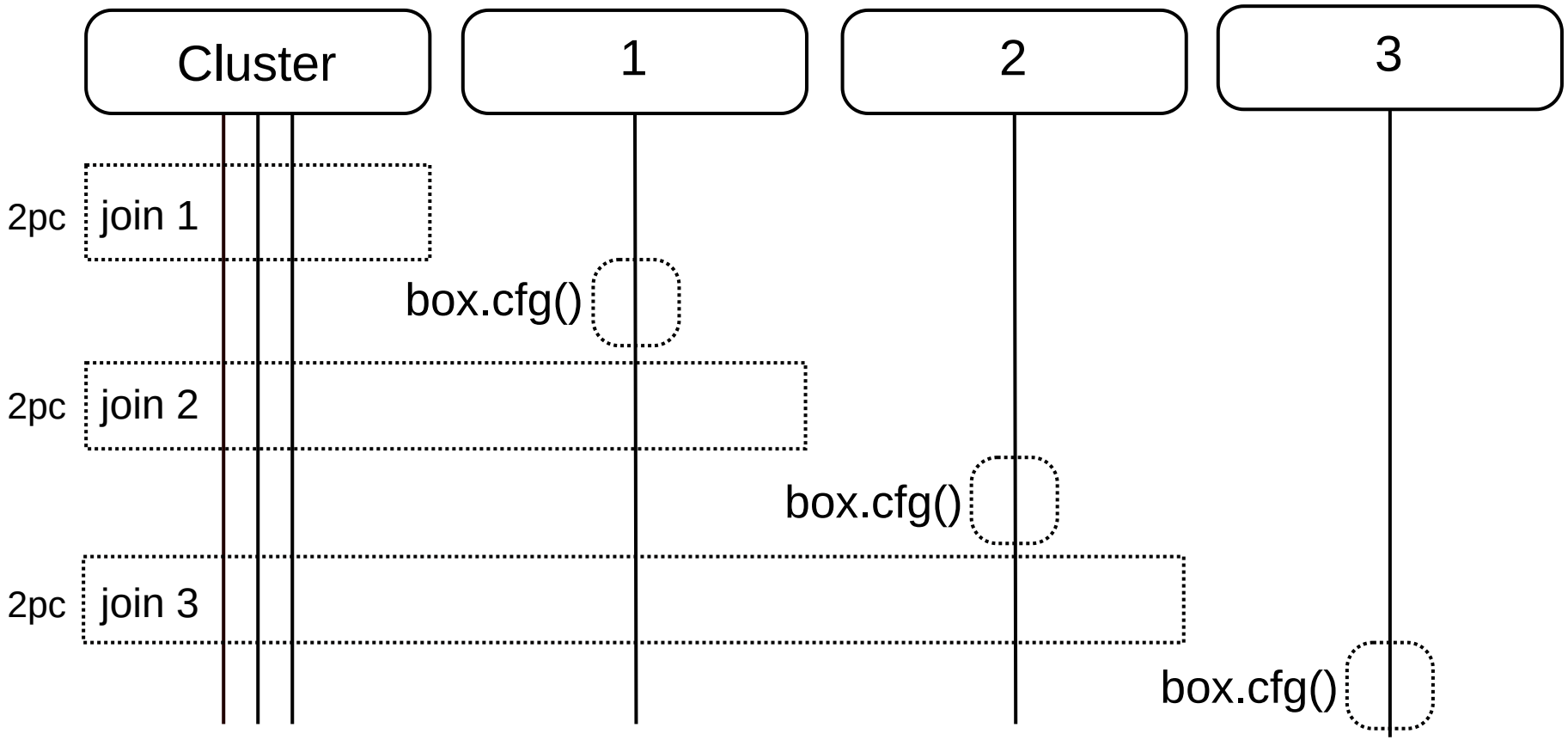
Проблемы масштабирования



Проблемы масштабирования



Проблемы масштабирования



Проблема масштабирования №1

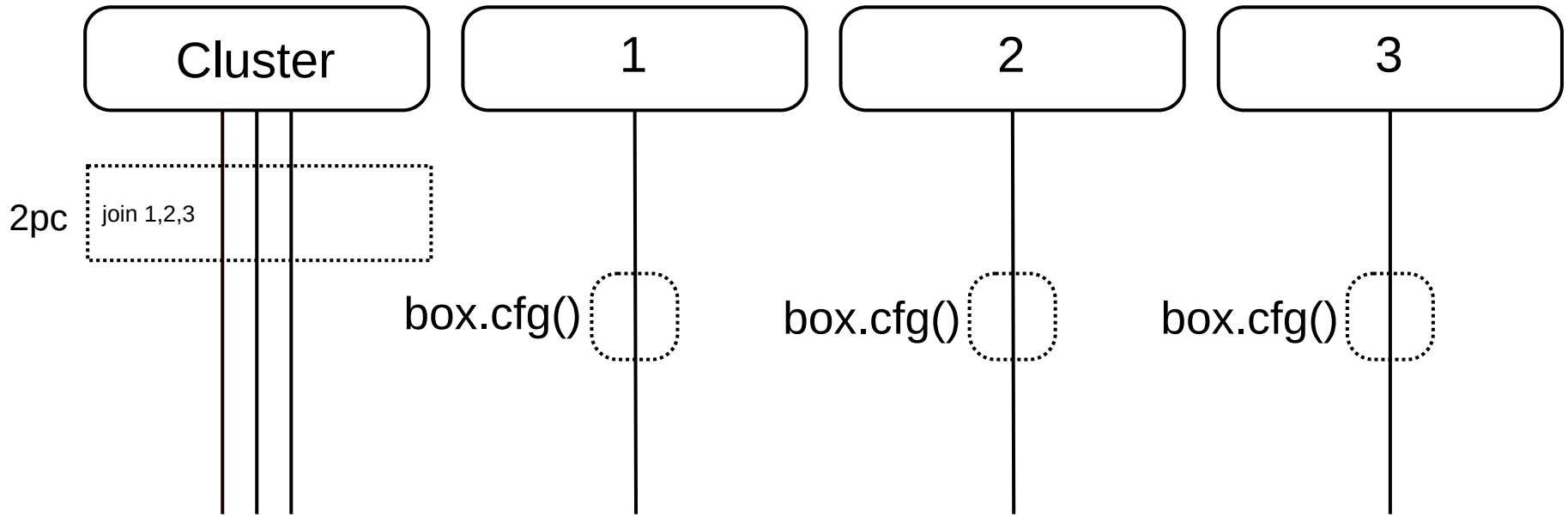
- Присоединение инстансов по-одному не эффективно

Проблема масштабирования №1

- Присоединение инстансов по-одному не эффективно
- Надо делать один двухфазный коммит на всех

Проблема масштабирования №1

- Присоединение инстансов по-одному не эффективно
- Надо делать один двухфазный коммит на всех



Проблема масштабирования №1

- Гонки появляются в другом месте

```
box.cfg({
  listen = ...,
  replication = {
    'user:password@localhost:3301',
    'user:password@localhost:3302',
  }
})

box.schema.user.create('user', 'password')
box.schema.user.grant('user', 'replication')
```

Проблема масштабирования №2

- Поллинг не эффективен
- Проблемы инициализации сложно диагностировать

Проблема масштабирования №2

- Поллинг не эффективен
- Проблемы инициализации сложно диагностировать

```
iproto.listen('0.0.0.0', 3301, {
  username = ...,
  password = ...,
})
```

Проблема масштабирования №2

- Поллинг не эффективен
- Проблемы инициализации сложно диагностировать

```
iproto.listen('0.0.0.0', 3301, {
  username = ...,
  password = ...,
})
```

```
conn.eval('
  iproto.stop()
  box.cfg({listen = 3301})
')
```

Ссылки

- Cartridge framework - github.com/tarantool/cartridge
- Cartridge CLI - github.com/tarantool/cartridge-cli
- This presentation - rosik.github.io/2019-bigdatadays

Вопросы?