Patroni
Отказоустойчивый кластер - это просто

Александр Кукушкин, Алексей Клюкин
Zalando SE
PGDay Russia'16, Санкт-Петербург

zalando

О нас

Alexander Kukushkin

Database Engineer @ZalandoTech
Email: alexander.kukushkin@zalando.de

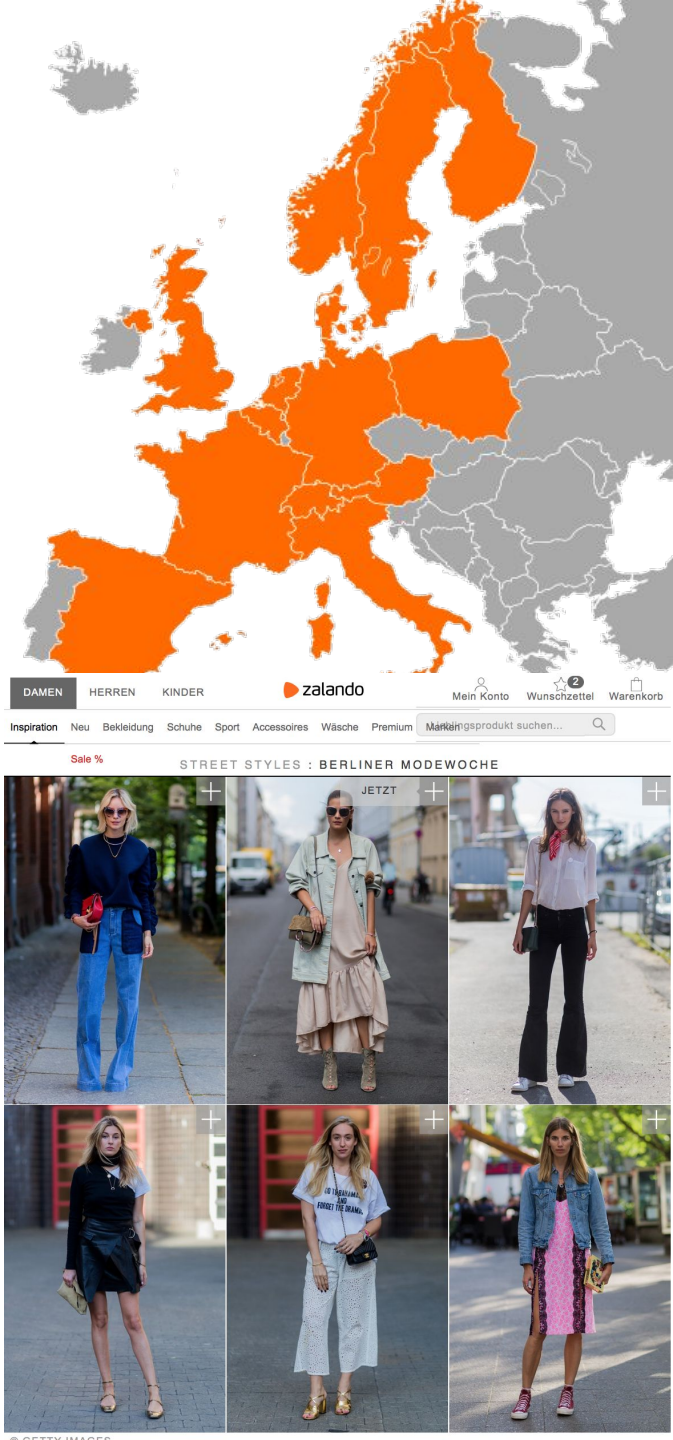
Oleksii Kliukin

Database Engineer @ZalandoTech
Email: oleksii.kliukin@zalando.de
Twitter: @hintbits

zalando

Zalando

- ~ 3 bn EUR revenue
- ~ 160 m visits/month
- 60% visits from mobile devices
- > 170 databases
- > 1000 tech employees
- We are hiring!



Автономные команды и Radical Agility

Организации, проектирующие системы, неизбежно выпускают продукт, структура которого повторяет структуру коммуникаций внутри организации.

Закон Мелвина Конвея

zalando

PostgreSQL в облаках (cattle vs pets)

- Быстрое развертывание новых СУБД
- Взаимозаменяемые сервера
- Максимально-стандартизированная конфигурация
- Автоматическая настройка

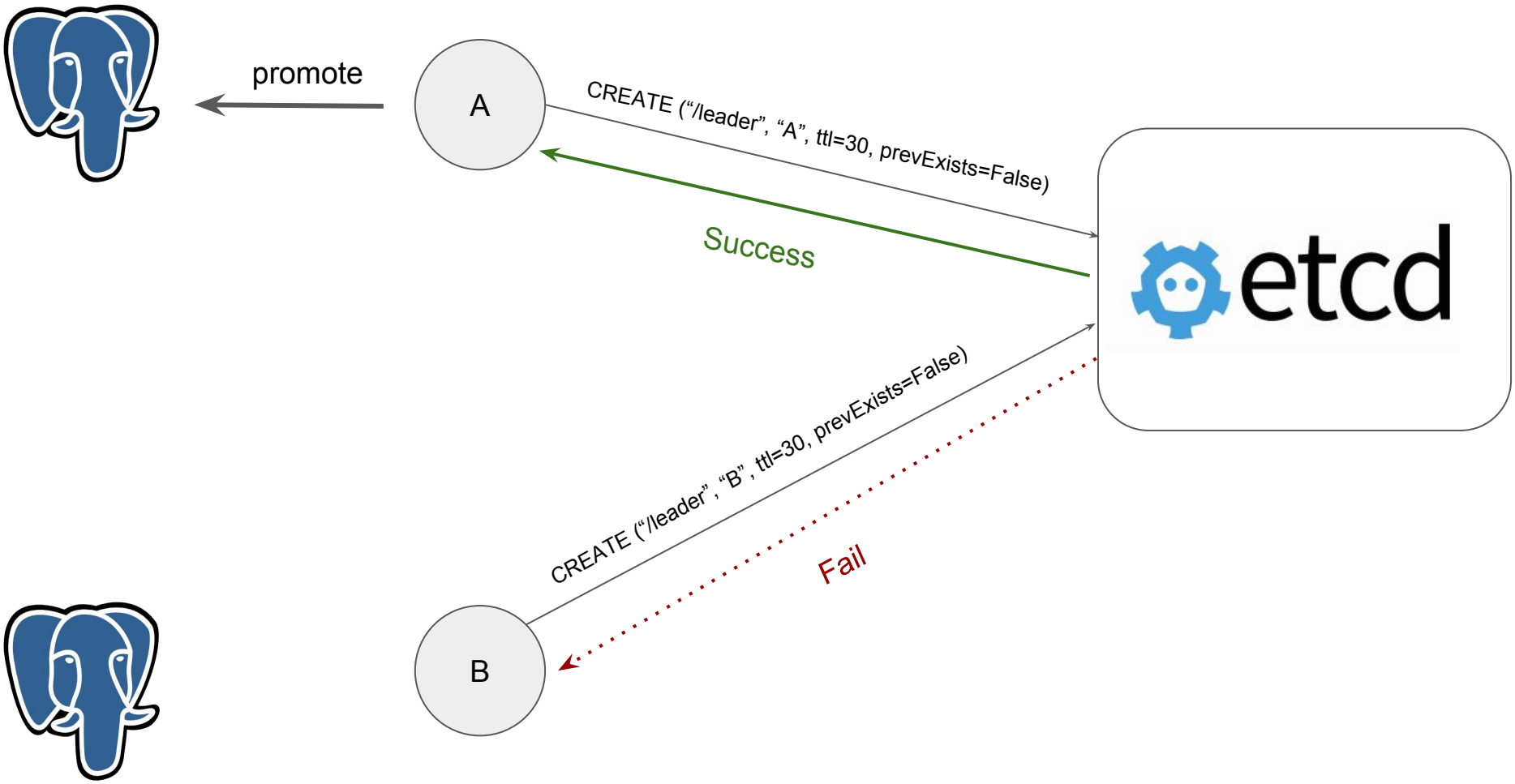
Хранение состояния кластера

- Проблема split-brain
- Авторитетный источник информации о мастере
- Хранение информации о мастере и репликах
- Атомарные операции (CAS)
- Отказоустойчивость

Распределенное хранилище данных (DCS)

- Key - value
- Алгоритмы решения задач консенсуса (RAFT, PAXOS)
- Отказоустойчивость
- Поддержка TTL для ключей или сессий
- Возможность “следить” за изменениями ключей (watches)

Гонка за лидером



Patroni + DCS

- DCS + PostgreSQL = отказоустойчивость
- Patroni управляет PostgreSQL
- TTL для ключа или сессии лидера
- Watch для ключа лидера
- Выборы лидера

Постоянные ключи

- Initialize
  - "key": "/service/testcluster/initialize",

    "value": "6303731710761975832",
- Leader optime
  - "key": "/service/testcluster/optime/leader",

    "value": "67393608",
- Config
  - "key": "/service/testcluster/config",

    "value": "{\"postgresql\":{\"parameters\":{\"synchronous_standby_names\":\"*\"}}}",

zalando

Временные ключи

- Leader
  - "key": "/service/testcluster/leader",

    "value": "dbnode2",

    "ttl": 22
- Members
  - "key": "/service/testcluster/members/dbnode2",

    “value": "{\"role\":\"master\",\"state\":\"running\",

    \"conn_url\":\"postgres://172.17.0.3:5432/postgres\",

    \"api_url\":\"http://172.17.0.3:8008/patroni\",\"xlog_location\":67393608}",

    "ttl": 22

zalando

Основные этапы работы

- Инициализация
- Выборы лидера
- Добавление новой реплики
- Проверка наличия ключа лидера
- Потеря лидера или demote

Возможности Patroni

- Поддержка Consul, Zookeeper и etcd
- Manual and Scheduled Failover
- Cascading replication/pg_basebackup from replica
- Synchronous replication
- Поддержка pg_rewind
- Customizable replica creation methods

Возможности Patroni

- Callbacks
  - on_start, on_stop, on_restart, on_role_change
- REST API
- Динамическая конфигурация
- Tags
  - nofailover, noloadbalance, clonefrom, replicatefrom
- patronictl

Динамическая конфигурация

- Идентичная конфигурация на всех узлах:
  - ttl, loop_wait, retry_timeout, maximum_lag_on_failover
  - wal_level, hot_standby
  - max_connections, max_prepared_transactions, max_locks_per_transaction, max_worker_processes, track_commit_timestamp, wal_log_hints
  - wal_keep_segments, max_replication_slots
- Динамическое изменение конфигурации PostgreSQL/Patroni
- Поддержка restart-only параметров (pending_restart flag)
- Хранение параметров в DCS
- Приоритет: patroni.yaml, DCS

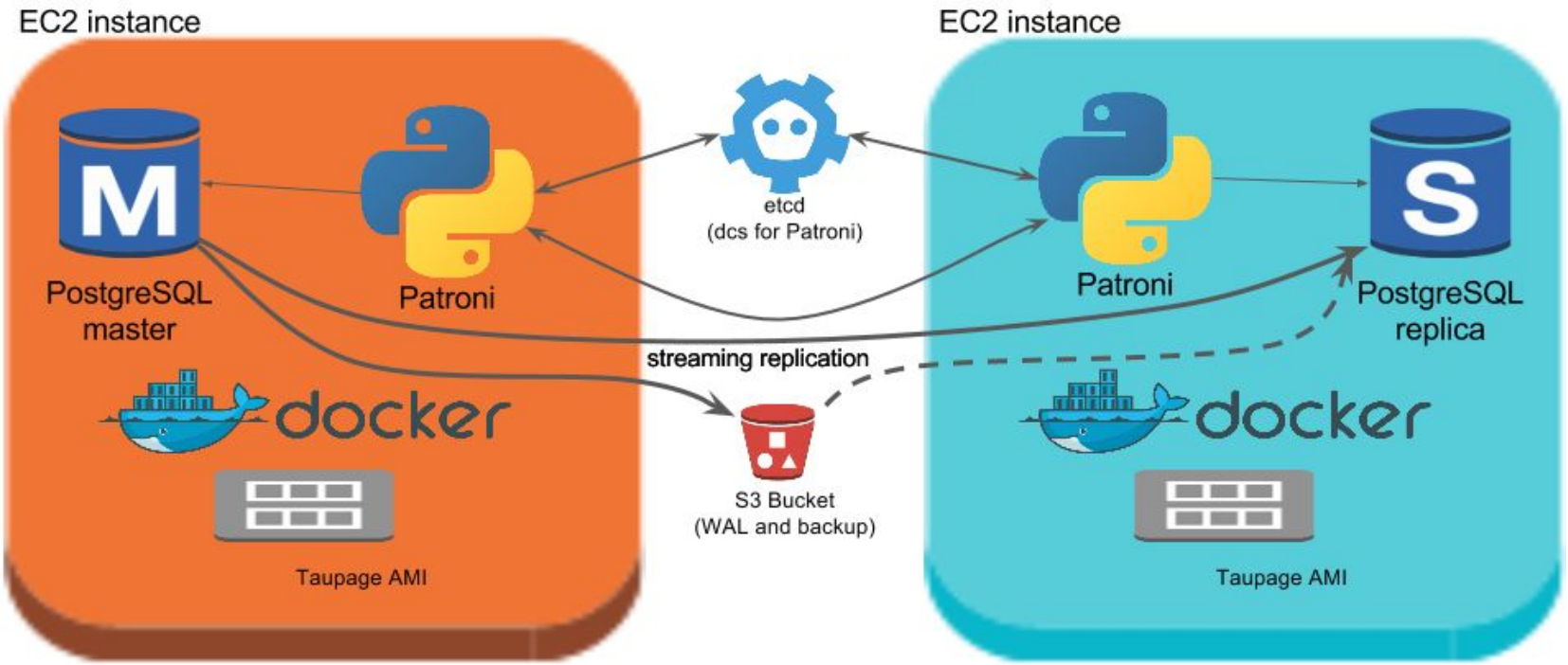
LIVE DEMO

zalando

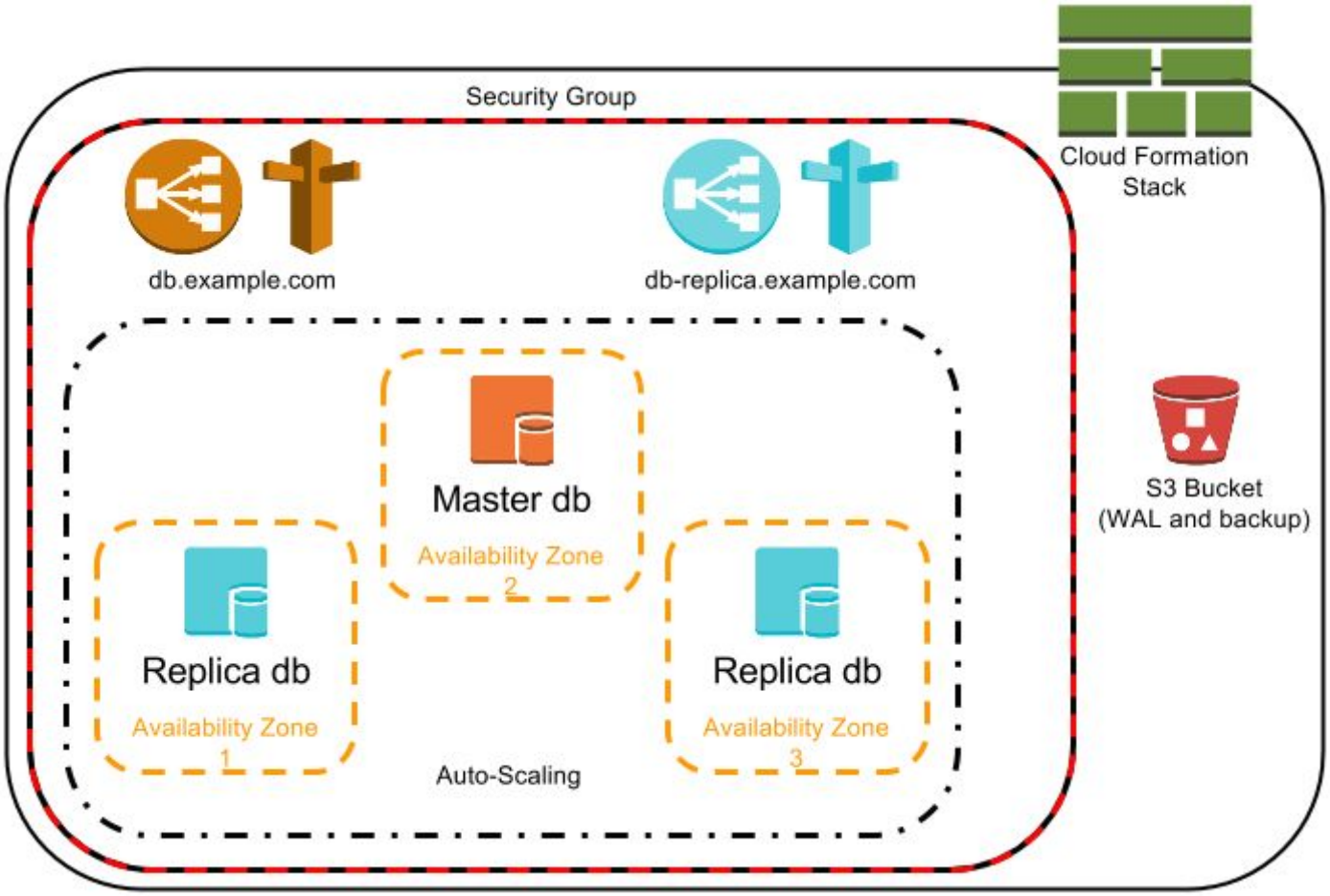
Patroni - это конструктор

- Автоматическое переключение клиентов
  - haproxy, pgbouncer
- Хранилище WAL файлов
- Регулярные резервные копии
- Мониторинг
- DB servers deployment

Spilo: Patroni + Docker + AWS



zalando


Security Group
Cloud Formation Stack
db.example.com
db-replica.example.com
Master db
Availability Zone 2
Replica db
Availability Zone 1
Replica db
Availability Zone 3
Auto-Scaling
S3 Bucket (WAL and backup)

Automatic failover is hard

- В какой момент переключать мастер:
  - Надежность опеределения момента переключения vs доступность системы
  - Проблема слишком частых переключений
  - Конфигурация DCS
- Старый мастер с pg_rewind и потеря данных
  - Нужен pg_rewind = на мастере есть нереплицированные записи
  - Что делать если pg_rewind завершилась безуспешно?
- Как выбрать нового мастера?
  - XLOG position
  - Надежность хоста/соединения
  - Производительность

Настройка параметров Patroni

- Когда переключать мастер
  - ttl, loop_wait, retry_timeout. 2 x retry_timeout + loop_wait < ttl
- Выборы нового мастера
  - maximum_lag_on_failover; < 16MB - нет гарантированного восстановления из base backup
  - nofailover
- pg_rewind
  - use_pg_rewind, remove_data_directory_on_rewind_failure

Спасибо!

https://github.com/zalando/patroni

zalando

Ссылки

- Spilo: https://github.com/zalando/spilo
- Confd: http://www.confd.io
- Etcd: https://github.com/coreos/etcd
- RAFT: http://thesecretlivesofdata.com/raft/