

Географически распределенное приложение внутри кластера Kubernetes

Андрееенко Артём
CIO Prisma Labs

GDG DevFest Астрахань
24 ноября 2018 г



План доклада

- О компании
- О себе
- Описание проблемы
- О Kubernetes
- Про инфраструктуру
- Про приложение



Prisma Labs

<https://prisma-ai.com>

- iPhone и Android приложение
- Искусственный интеллект, который перерисовывает фото в другом стиле
- Создано в 2016 году
- 150М скачиваний



Prisma Photo Editor

Turn photos into works of art

Prisma is a photo-editing app that creates amazing photo effects, transforming photos into paintings. Prisma uses artificial neural networks that enable users to make photos appear like they were painted by Picasso, Munch or even Salvador Dali himself.

Like 138K

Tweet



Available on the
App Store



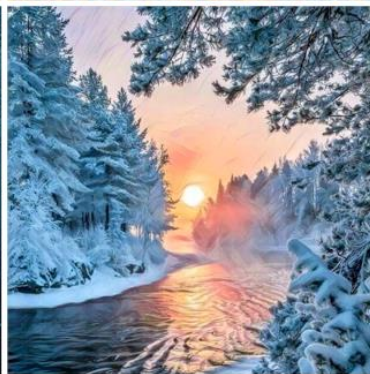
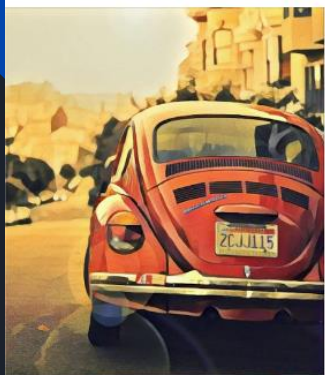
GET IT ON
Google Play



APP STORE
APP OF THE YEAR 2016



GOOGLE PLAY
BEST APP OF 2016





О себе

- Занимаюсь бэкендом и инфраструктурой в Prisma Labs
- Ведущий подкаста GolangShow
<https://golangshow.com>
- Продвигаю язык Go в русскоязычном сообществе
<http://slack.golang-ru.com>



Проблема

- Пользователи приложения расположены по всей планете
- Высокая задержка выполнения обращений к бэкенду
- Высокая частота обрыва соединений из большого количества промежуточных узлов и таймаутов
- При возрастании размеров запросов и ответов API вызовов нелинейно растет время выполнения запроса на клиенте

Задержки прохождения пакетов в Интернете



Время приема-передачи (RTT) между локациями Servers.com* и Google Compute Engine*

[Посмотреть PDF](#)

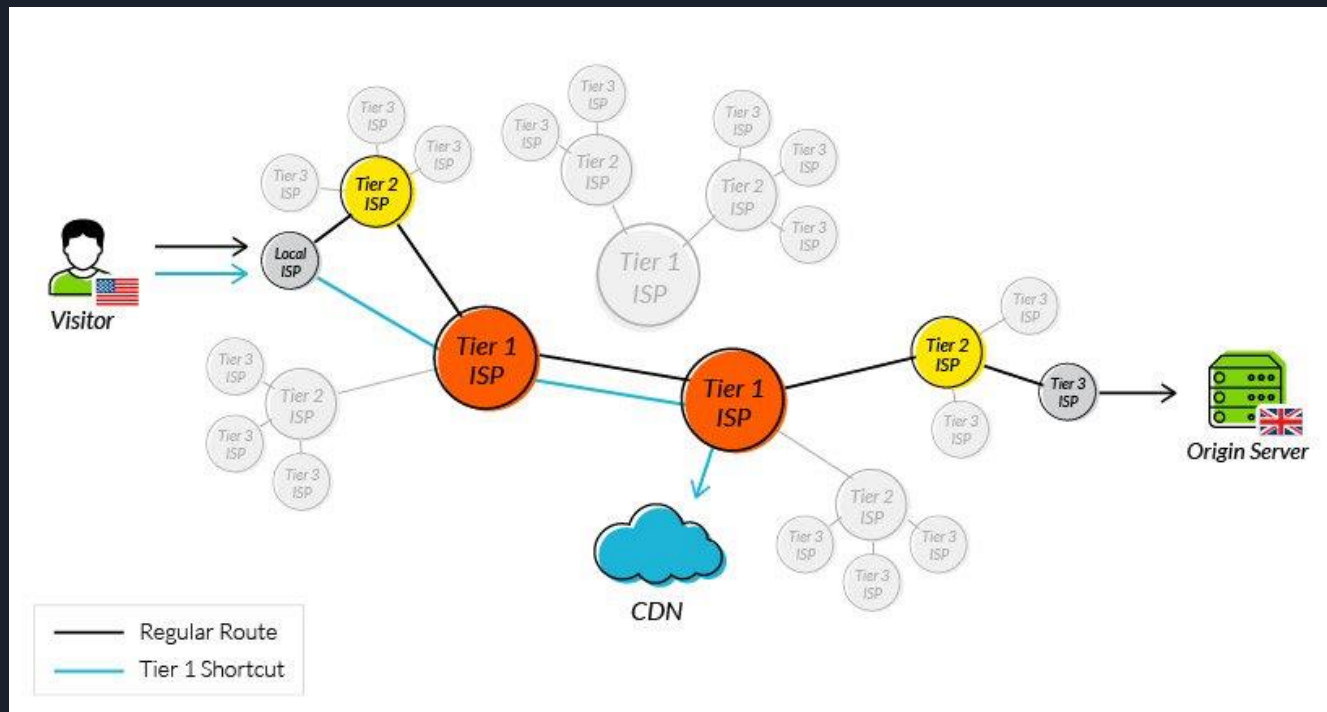
	AMS1	DFW1, DFW2	LUX1	MOW1	SIN1	WAS2	HKG1
Чжанхуа, Тайвань asia-east1-b	257 ms	158 ms	281 ms	299 ms	49 ms	182 ms	15 ms
Токио, Япония asia-northeast1-b	232 ms	133 ms	245 ms	274 ms	74 ms	164 ms	47 ms
Мумбаи, Индия asia-south1-b	356 ms	253 ms	367 ms	394 ms	63 ms	275 ms	240 ms
Джуронг-Уэст, Сингапур asia-southeast1-b	296 ms	193 ms	316 ms	337 ms	2 ms	217 ms	38 ms
Сидней, Австралия australia-southeast1-b	277 ms	170 ms	289 ms	321 ms	171 ms	196 ms	227 ms
Сен-Гиллен, Бельгия europe-west1-b	13 ms	114 ms	17 ms	45 ms	238 ms	89 ms	272 ms



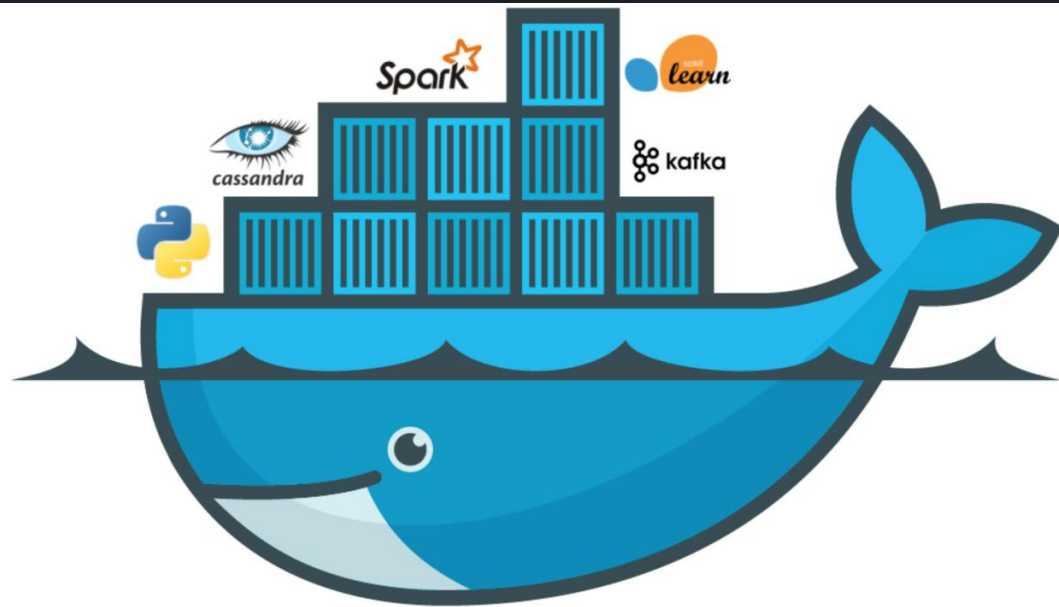
Ограничение скорости передачи данных из-за задержек

- Стандартный размер ip-пакета 1460 байт
- Скорость интернета пользователя 100 Мбит/сек
- Задержка 360мс
- Максимальная скорость 6.55 Мбит/сек или 838 килобайт/сек

CDN решение для статического контента



Что такое Docker?





Что такое Docker?

- контейнеризация вместо виртуализации
- использование механизмов ядра Linux
- изолирование ресурсов: виртуальная память, сри, диск, сеть

Что такое Kubernetes?



- Создан Google (первый релиз 7 июня 2014)
- Open Source (<https://github.com/kubernetes/kubernetes>)
- Планировщик контейнизированных приложений
- Абстракция над железом / облаком / сетью
- Ключевой проект Cloud Native Computing Foundation



Kubernetes Pod

- базовая абстракция
- один или несколько контейнеров в общем namespace
- выделенный ip адрес на каждый Pod
- Pod работает только на одном хосте



ReplicaSet

- группа подов
- параметр `replicas` устанавливает кол-во подов
- перезапуск подов, если под или нода упали



Пример ReplicaSet

```
apiVersion: apps/v1
kind: ReplicaSet
metadata:
  name: frontend
  labels:
    app: guestbook
    tier: frontend
spec:
  # modify replicas according to your case
  replicas: 3
  selector:
    matchLabels:
      tier: frontend
    matchExpressions:
      - {key: tier, operator: In, values: [frontend]}
  template:
```




Deployments

- Расширение ReplicaSet
- Позволяет поддерживать несколько ReplicaSet и переключаться между ними



Пример Deployments

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx-deployment
  labels:
    app: nginx
spec:
  replicas: 3
  selector:
    matchLabels:
      app: nginx
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
        - name: nginx
          image: nginx:1.15.4
          ports:
            - containerPort: 80
```



DaemonSet

- запускает поды из группы на каждой ноде
- hostname каждого пода детерминирован (пример example-pod-01, example-pod-02)
- Подходит для запуска нод Ceph, GlusterFS, локальных сборщиков логов, метрик



Пример DaemonSet

```
apiVersion: apps/v1
kind: DaemonSet
metadata:
  name: fluentd-elasticsearch
  namespace: kube-system
  labels:
    k8s-app: fluentd-logging
spec:
  selector:
    matchLabels:
      name: fluentd-elasticsearch
  template:
    metadata:
      labels:
        name: fluentd-elasticsearch
    spec:
      tolerations:
        - key: node-role.kubernetes.io/master
          effect: NoSchedule
      containers:
```



StatefulSet

- подключает к каждому поду в группе постоянные том (PersistentVolume)
- при падении ноды под запускается на другой ноде с подключением того же существующего с данными приложения тома
- подходит для запуска баз данных и хранилищ данных (MySQL, PostgreSQL, ElasticSearch, ClickHouse, Prometheus)



Пример StatefulSet

```
apiVersion: apps/v1
kind: StatefulSet
metadata:
  name: web
spec:
  serviceName: "nginx"
  replicas: 2
  selector:
    matchLabels:
      app: nginx
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
        - name: nginx
          image: k8s.gcr.io/nginx-slim:0.8
          ports:
            - containerPort: 80
              name: web
          volumeMounts:
            - name: www
              mountPath: /usr/share/nginx/html
      volumeClaimTemplates:
        - metadata:
            name: www
          spec:
            accessModes: [ "ReadWriteOnce" ]
            resources:
              requests:
                storage: 1Gi
```



Ingress

- сетевой сервис, который принимает внешние запросы
- поддержка TCP / HTTP / UDP
- закрытые реализации ingress в качестве GCP Load Balancer / AWS Load Balancer
- софтверная реализация nginx-ingress
- поддержка TLS



Настройка Kubernetes

- Kubespray - набор ansible скриптов
<https://github.com/kubernetes-incubator/kubespray>
- Kubeadm - консольная утилита для запуска кластер
- Minikube - миниинсталяция для тестов и экспериментов (1 нода)

Пример с Kubespray

```
[all]
devfest-au-01 ansible_host=13.238.154.63 ip=13.238.154.63 provider=aws
devfest-br-01 ansible_host=35.199.70.110 ip=35.199.70.110 provider=gcp
devfest-sg-01 ansible_host=167.99.75.69 ip=167.99.75.69 provider=do
devfest-ru-01 ansible_host=130.193.48.78 ip=130.193.48.78 provider=ydc

[kube-master]
devfest-ru-01

[etcd]
devfest-ru-01

[kube-node]
devfest-au-01
devfest-br-01
devfest-sg-01

[k8s-cluster:children]
kube-master
kube-node
```

```
# ansible-playbook -i inventories/devfest/hosts.ini cluster.yml
```



Kubectl

- консольная утилита для управления кластером
- показывает все типы ресурсов в кластере
kubectl get pods
- позволяет запускать yaml описания приложения
kubectl apply -f app.yml
- отображает логи подов
kubectl logs -f <pod name>

Helm (<https://helm.sh>)



- пакетный менеджер
- сторонний проект (не Google)
- содержит репозиторий пакетов, по которым можно делать поиск
helm search <packet name substring>
- одной командой позволяет разворачивать сложные кластера
например, Stolon - отказоустойчивый кластер PostgreSQL
helm install <packet name> -v <packet values yaml file> -n <installation name>



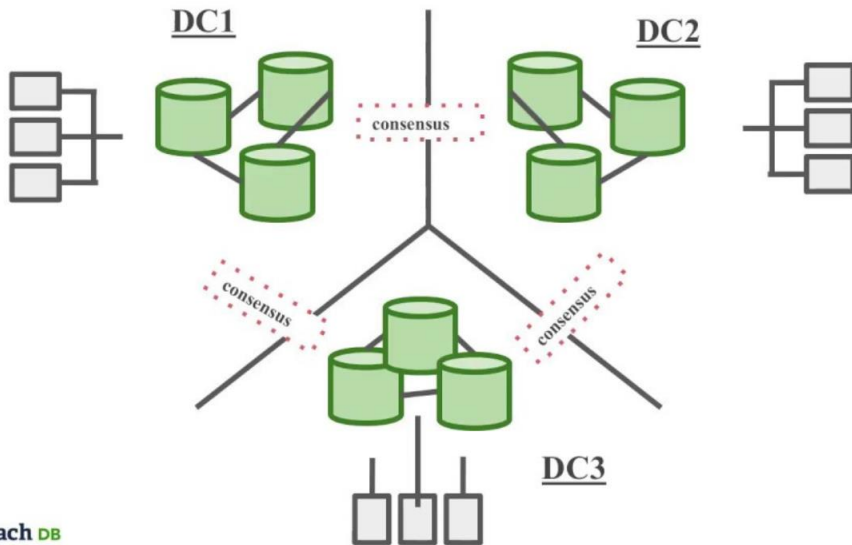
CockroachDB

- реляционная распределенная база данных
- master-master репликация
- PostgreSQL совместимый протокол
- есть джойны
- развитие Google Spanner whitepaper
- основана ex-инженерами Google
- адаптирована для работы с высокими задержками передачи данных между узлами

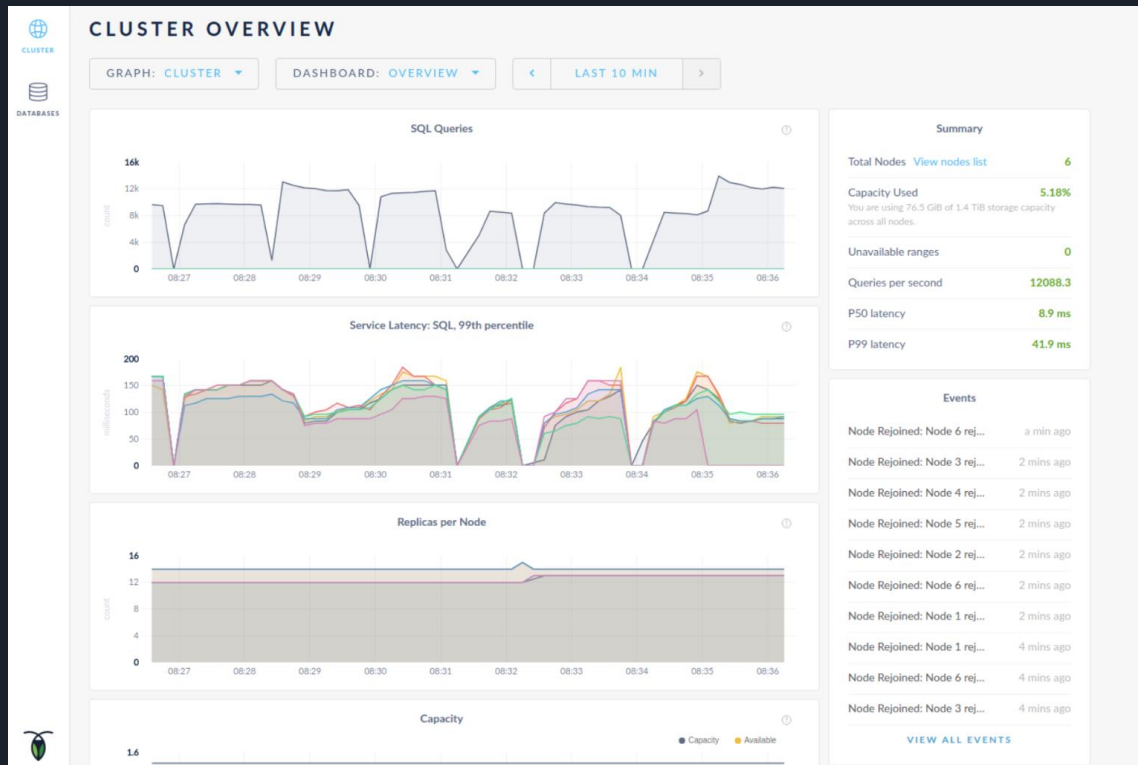


CockroachDB

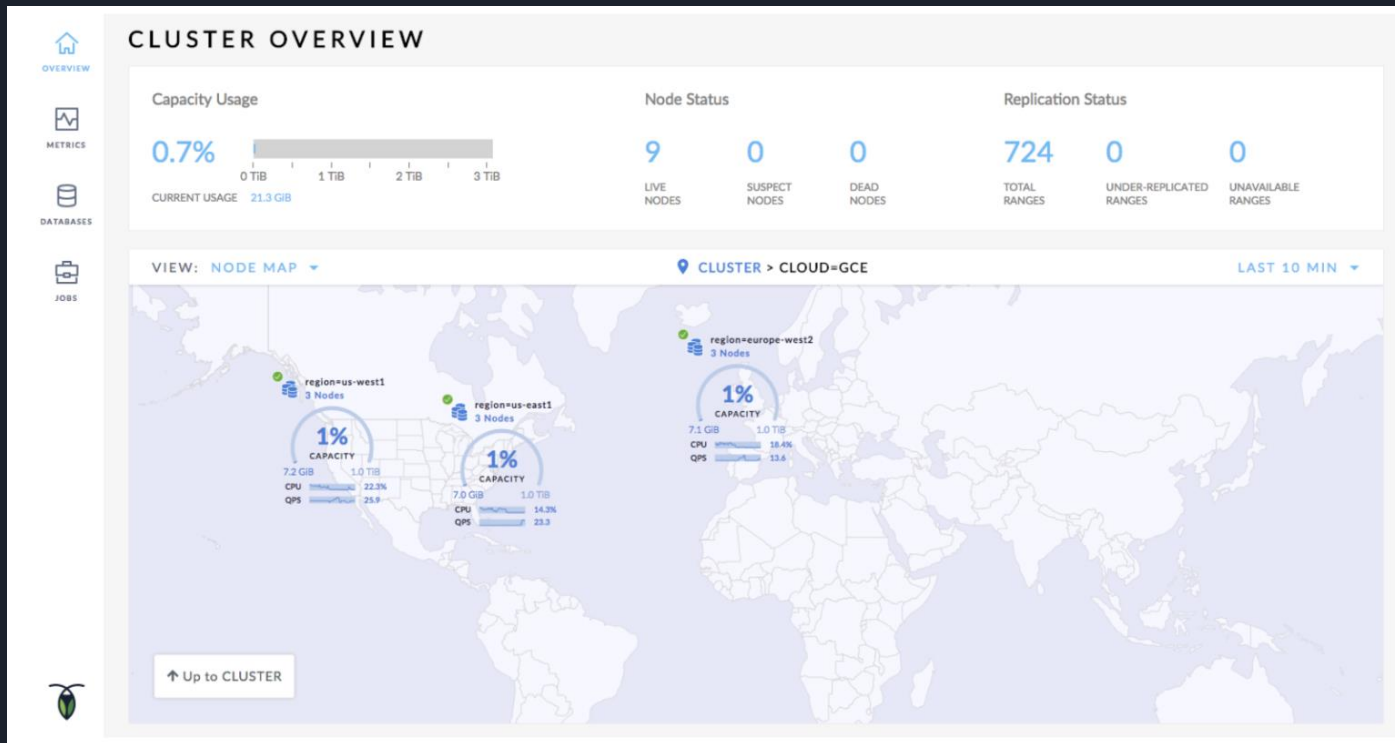
Data Integrity at Scale: Multi-Active Availability



CockroachDB



CockroachDB 2.0





Балансировка трафика с помощью DNS

- просто
- дешево
- низкая точность
- кэш провайдеров снижает отказоустойчивость
- пример провайдера сервиса
AWS Route53



Балансировка трафика с Anycast IP

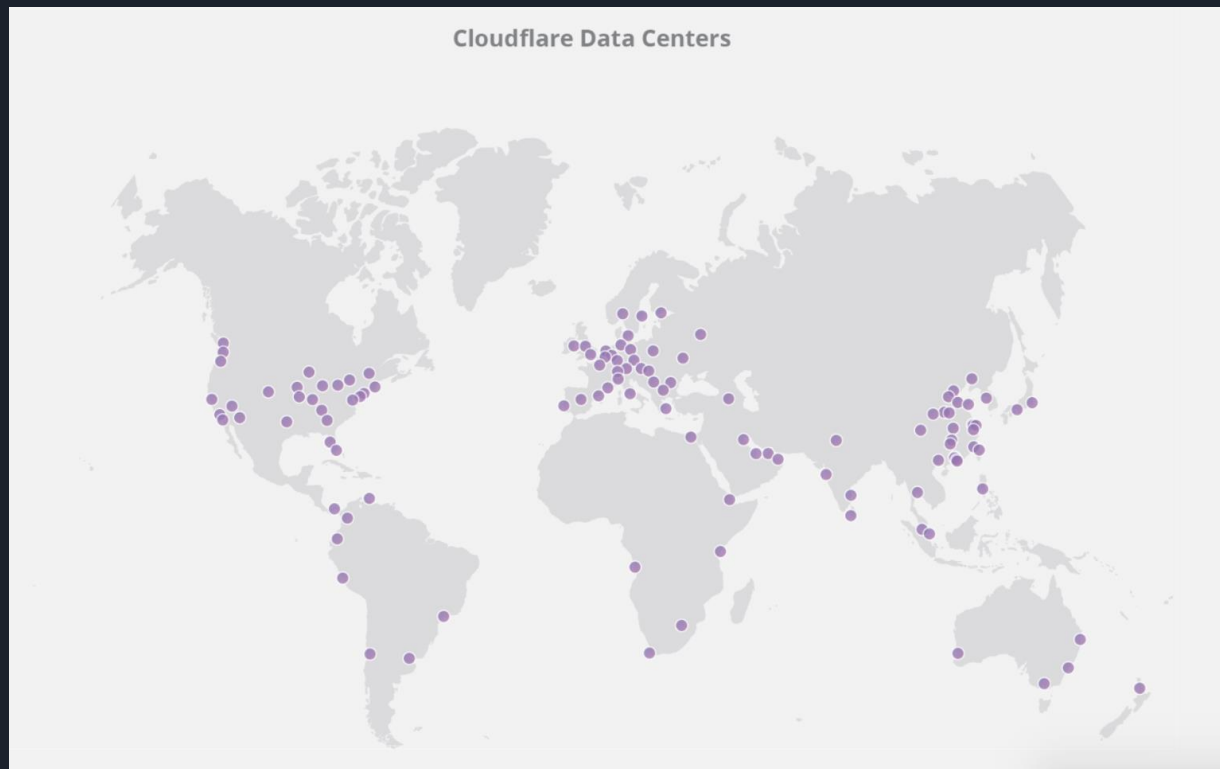
- сложнее настроить
- более точный выбор ближайшего сервера
- дороже, если трафик идет через промежуточный провайдер Anycast IP
- пример провайдера
Cloudflare



Cloudflare

- CDN
- anycast ip
- 150+ PoPs (точек присутвия) по всему миру
- балансировщик трафика с гео распределением

Cloudflare датацентры





Распределенный кластер Kubernetes

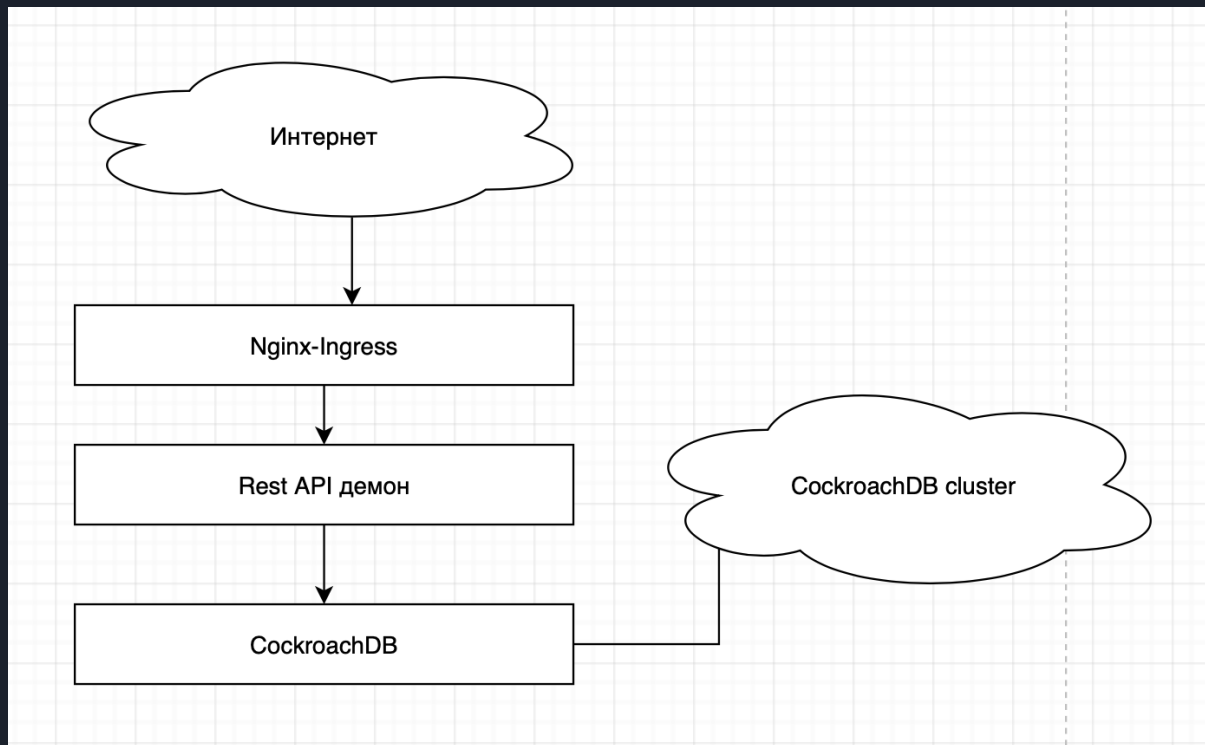
- Москва / Яндекс Облако
- Сингапур, Сингапур / DigitalOcean
- Сидней, Австралия / Amazon AWS
- Сан-Паулу, Бразилия / Google GCP




Задержки между нодами кластера

	Москва	Сидней	Сингапур	Сан-Паулу
Москва		360мс	195мс	260мс
Сидней	360мс		150мс	360мс
Сингапур	195мс	150мс		320мс
Сан-Паулу	260мс	305мс	320мс	

Поды одной ноды :-)



Настройка балансировщика Cloudflare

Health	Hostname	Available Pools	TTL	Proxied	Enabled	
<div><div></div><div>Critical</div></div>	devfest	<div><div></div>0 of 4 Pools ▾</div>	<div>Automa... ▾</div>		<div>On ▹▹</div>	<div><div></div><div></div></div>
Pool Name		Available Origins				
<div><div></div><div>Critical</div></div>	australia	<div><div></div>0 of 1 Origins ►</div>	<div>On ▹▹</div>			
<div><div></div><div>Critical</div></div>	brazil	<div><div></div>0 of 1 Origins ►</div>	<div>On ▹▹</div>			
<div><div></div><div>Critical</div></div>	russia	<div><div></div>0 of 1 Origins ▾</div>	<div>On ▹▹</div>			
Origin Name		Origin Address	Weight			
<div><div></div><div>Critical</div></div>	ru01	<div>130.193.48.78</div>	<div>1</div>	<div>On ▹▹</div>		
<div><div></div><div>Critical</div></div>	singapore	<div><div></div>0 of 1 Origins ►</div>	<div>On ▹▹</div>			

Help ►



Спасибо за внимание!

Вопросы?