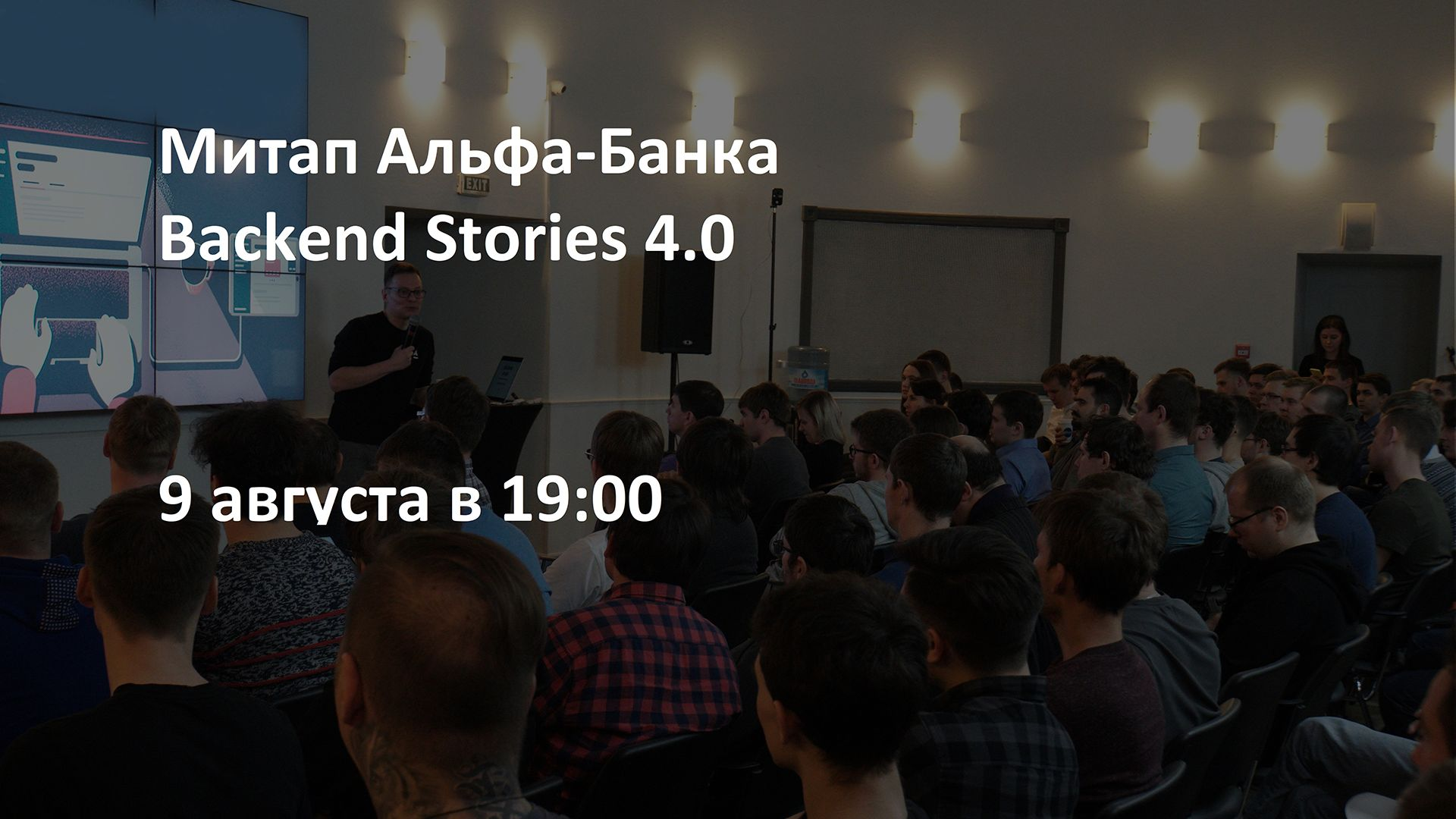

Митап Альфа-Банка
Backend Stories 4.0
9 августа в 19:00

ОТКАЗОУСТОЙЧИВОСТЬ В БОЛЬШОМ ИНТЕРНЕТЕ

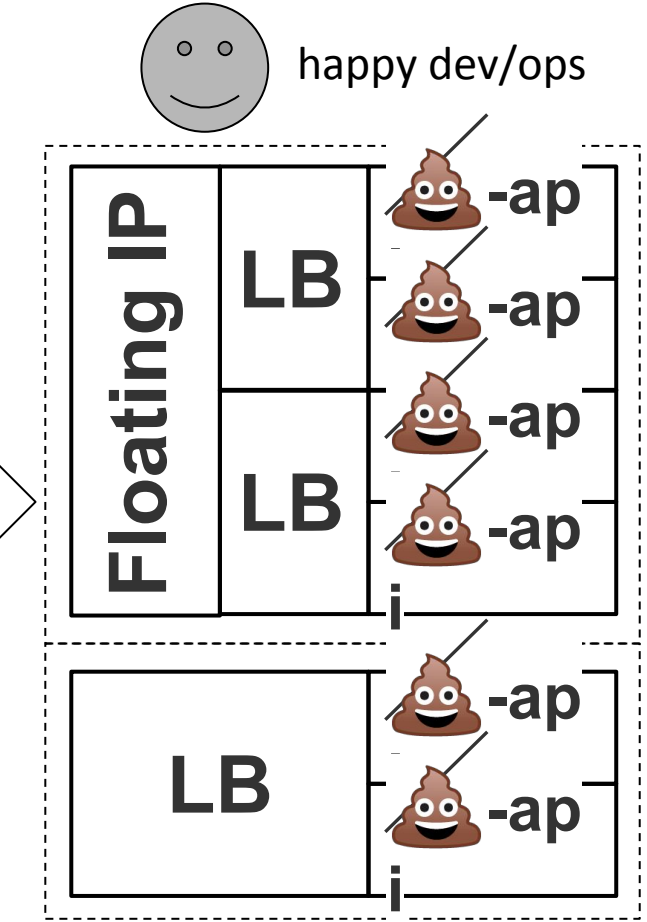
ABOUT ME

- Руководитель Центра компетенций Java в Альфа-Банке
- Был разработчиком и системным администратором
- Очень много занимаюсь наймом
- Считаю, хороший разработчик должен отвечать не только за код (devops и все такое :)

DISCLAIMER

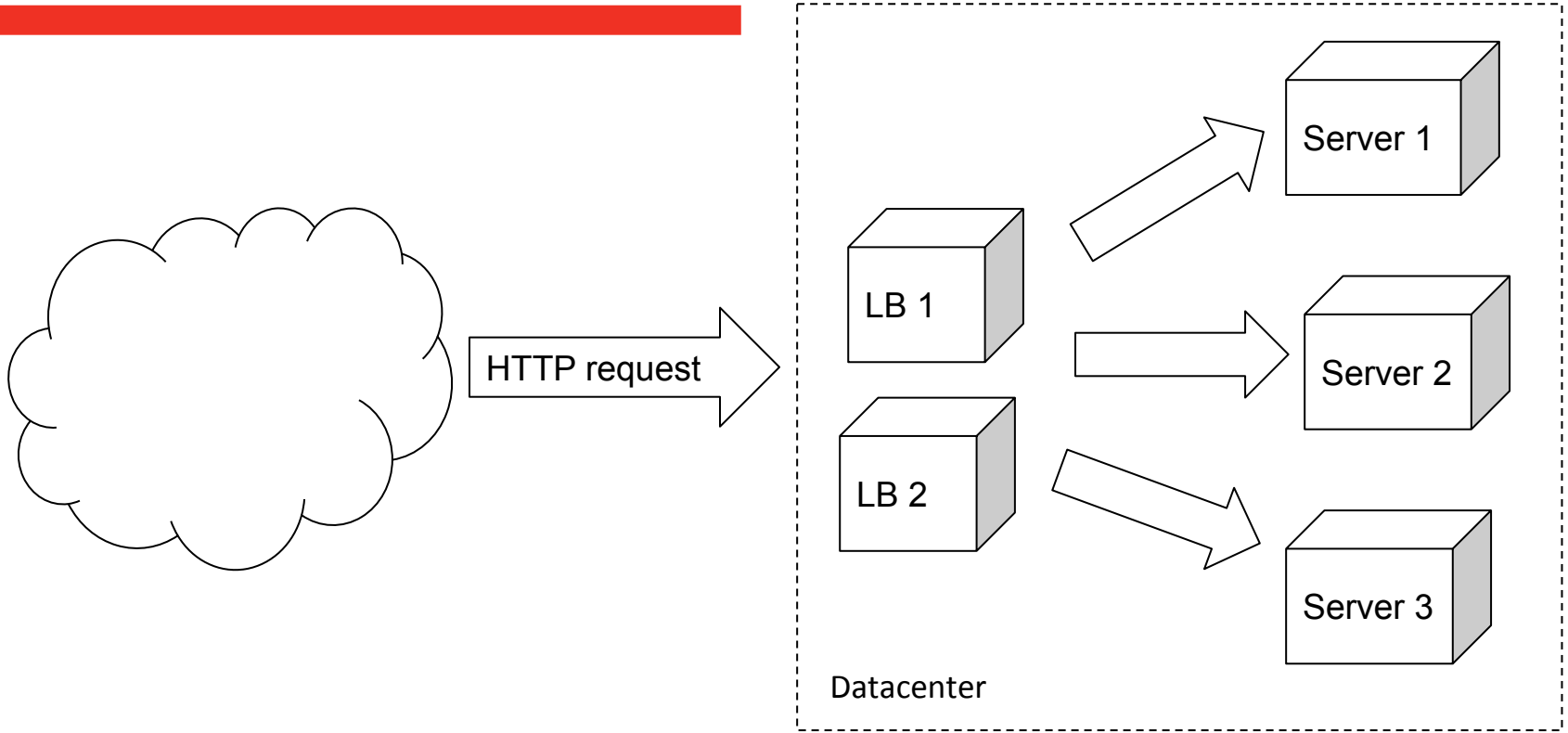
- Доклад совсем не про dev
- И не про формулы, девятки, SLA и т.д.
- Будет много капитанства про сети и протоколы
- Language is смешанный
- Не все термины взяты из RFC
- Надо знать про IP-адреса, маски и сети

THE PROBLEM

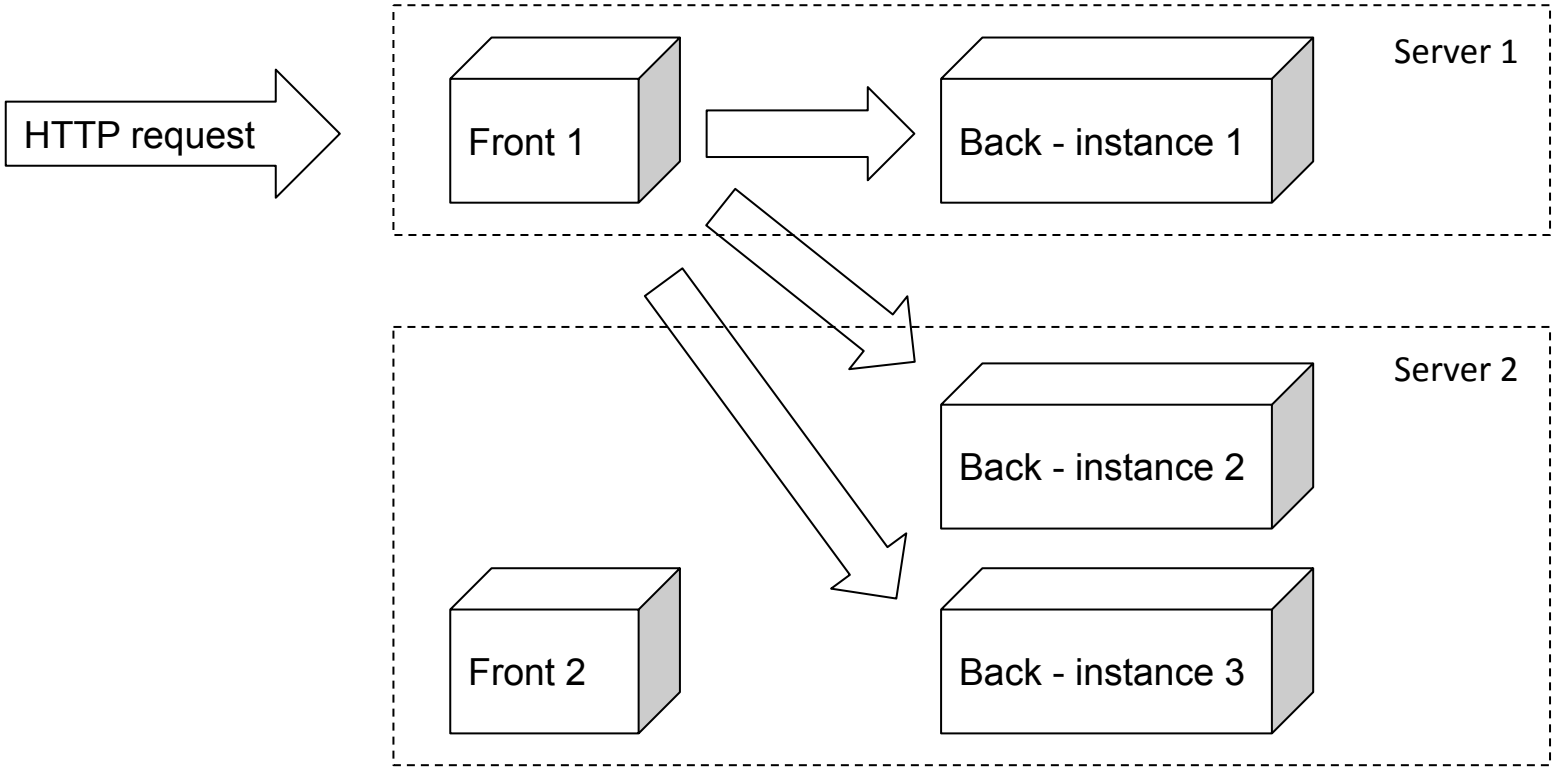


О ЧЕМ МЫ НЕ БУДЕМ ГОВОРИТЬ

SERVER BALANCING

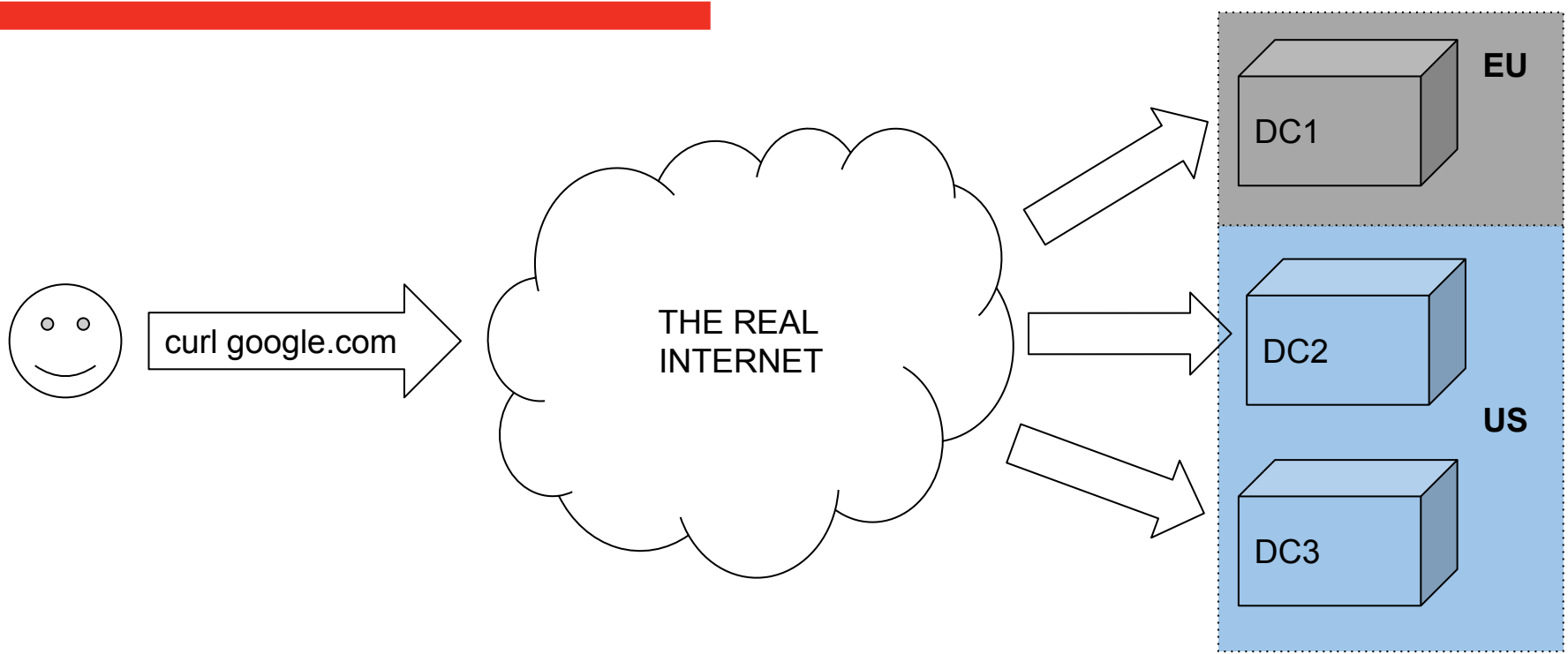


SERVICE DISCOVERY



А БУДЕМ ПРО

INTERNET ROUTING & BALANCING



INTERNET ROUTING & BALANCING

- Round robin DNS
- Geo DNS
- BGP Anycast routing
- Multihome BGP
- Всякое разное

ОГЛАВЛЕНИЕ

Зачем нужен DNS

drive.google.com

- symbol name
- server name
- host name
- domain name
- fully qualified domain name

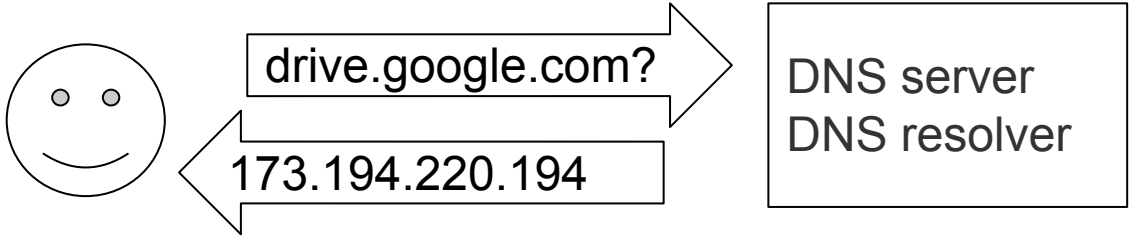
резолвинг

173.194.220.194

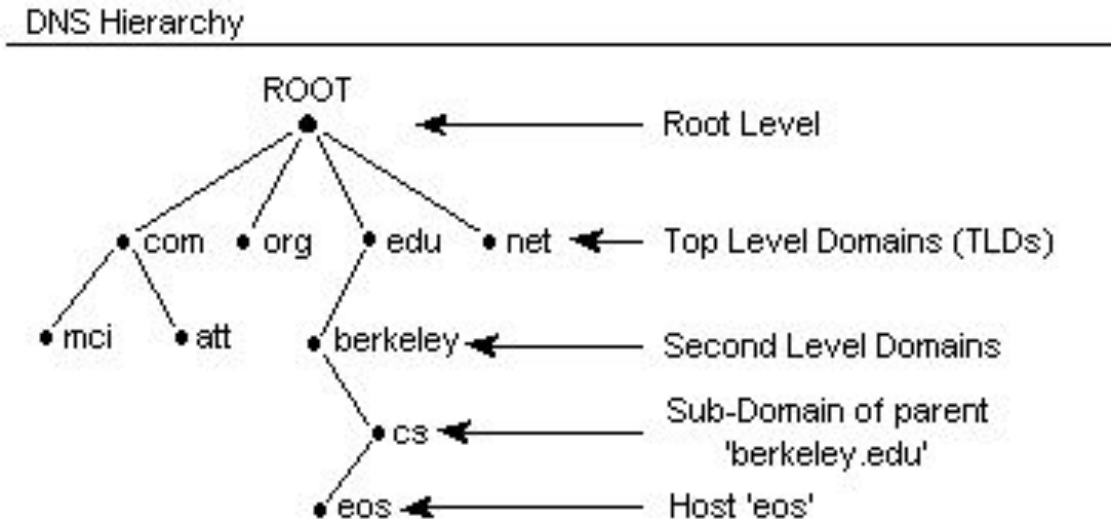
- IP-address

2a00:1450:4010:c09::c2

- IPv6-address



DNS is a tree



https://www.inetdaemon.com/tutorials/internet/dns/operation/hierarchy.shtml

DNS entities

1. domain
2. zone
3. host

Вопрос: **google.com** - это что?

Ответ: все сразу

DNS entities

1. google.com domain = google.com + drive.google.com + …
2. google.com zone = google.com only
3. google.com host = zone record (type A), just IP-address

Вопрос: **www.google.com** - это что?

Ответ: просто хост

DNS zone & records

- DNS-зона, обычно, хранится в файле и состоит из записей
- Файл зоны **google.com -** /etc/named/master/**google.com**
- Запись **drive.google.com** состоит из:

drive	CNAME	300	wide-docs.**l.google.com.**
Name	Type	TTL	Value

wide-docs	A	300	185.143.173.149
Name	Type	TTL	Value

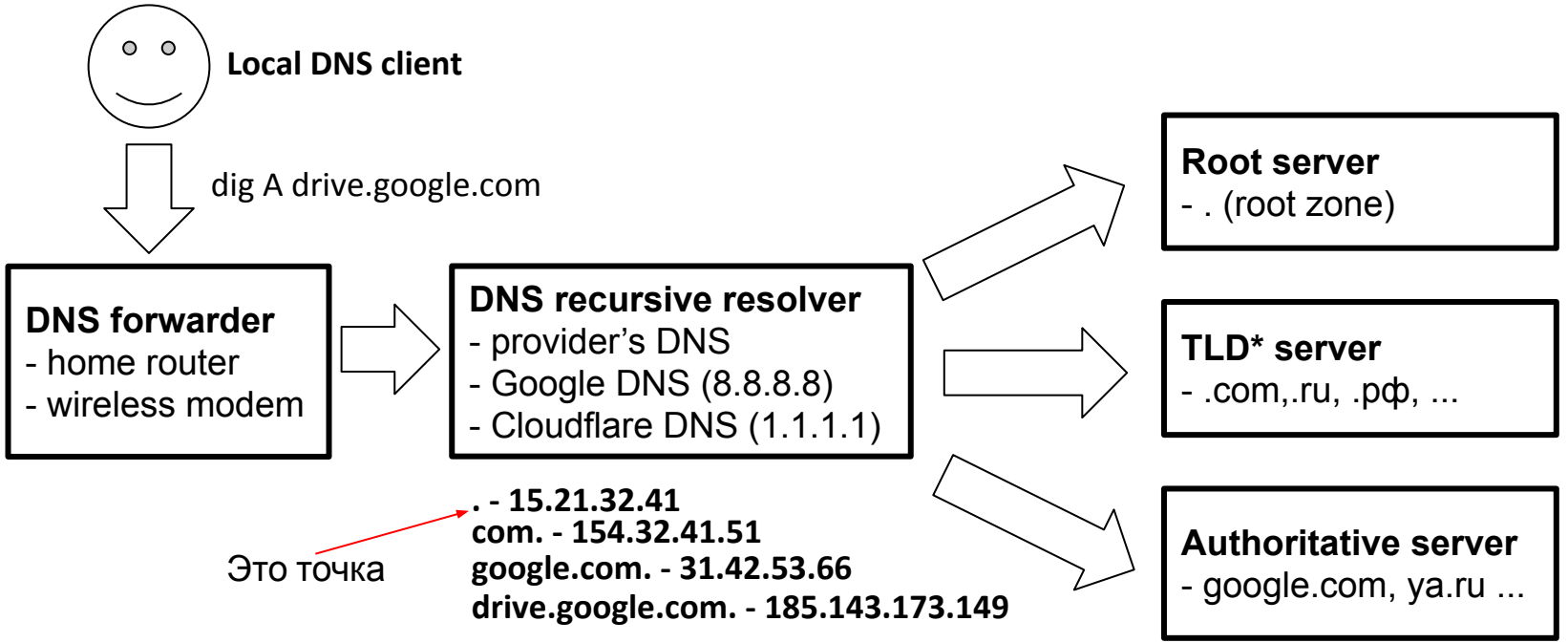
DNS record types

Types:

- **A - IP-address (host)**
- CNAME - Alias
- NS - Authoritative name server
- MX - Mail receiver
- TXT - text comment
- …

https://en.wikipedia.org/wiki/List_of_DNS_record_types

DNS resolving and server types



* TLD - top level domain

https://www.cloudflare.com/learning/dns/dns-server-types/

DNS root servers

- **13** root nameservers total (A-Root … M-Root)
- As of 2019-07-14, the root server system consists of **997** instances operated by the **12** independent root server operators
- Thanks to BGP Anycast :)
- No roots - no Internet
- И в России есть нюансы

https://root-servers.org/
https://en.wikipedia.org/wiki/Root_name_server

Картинка с http://casinoyay.com/kak-uskolznut-ot-vsevidyashhego-oka-roskomnadzora/

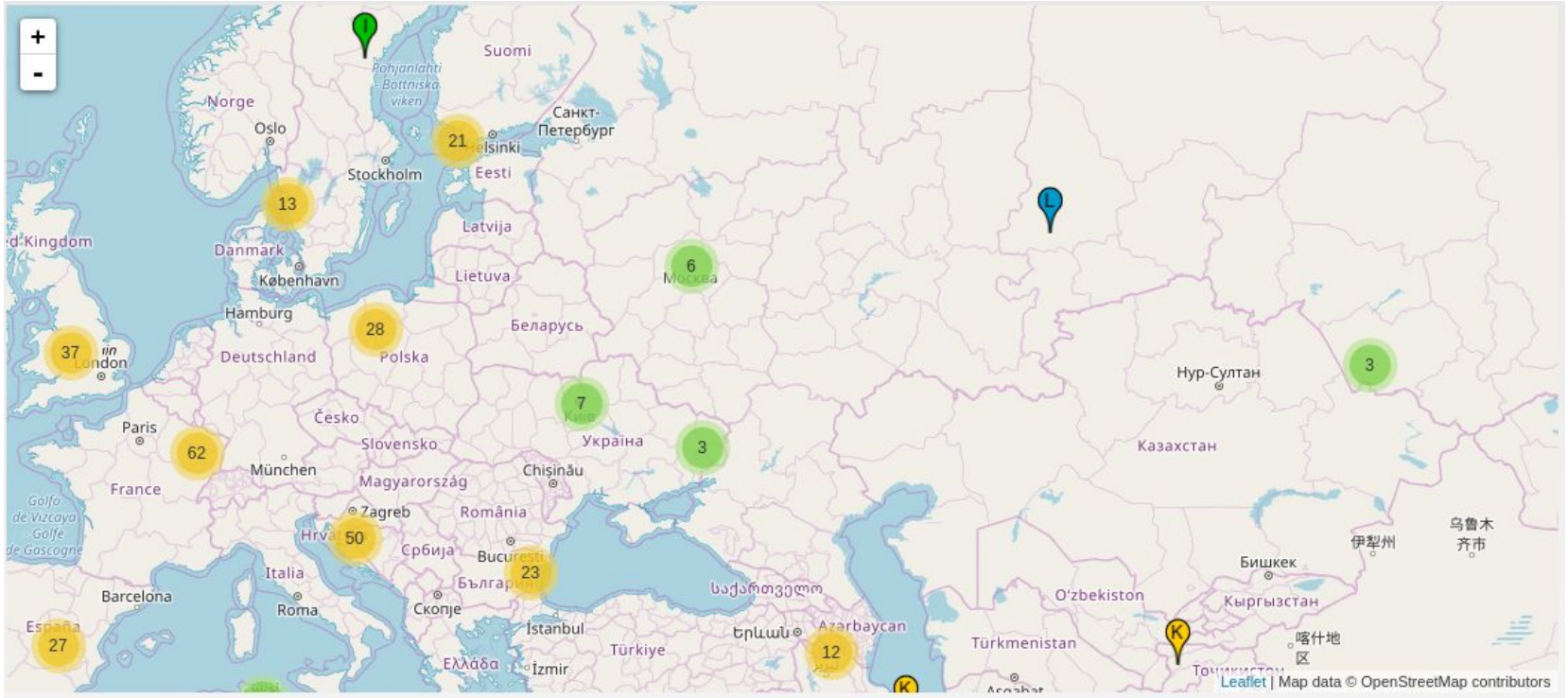
DNS root servers

На 22.05.2018 в России размещено 11 реплик корневых серверов DNS, в том числе:

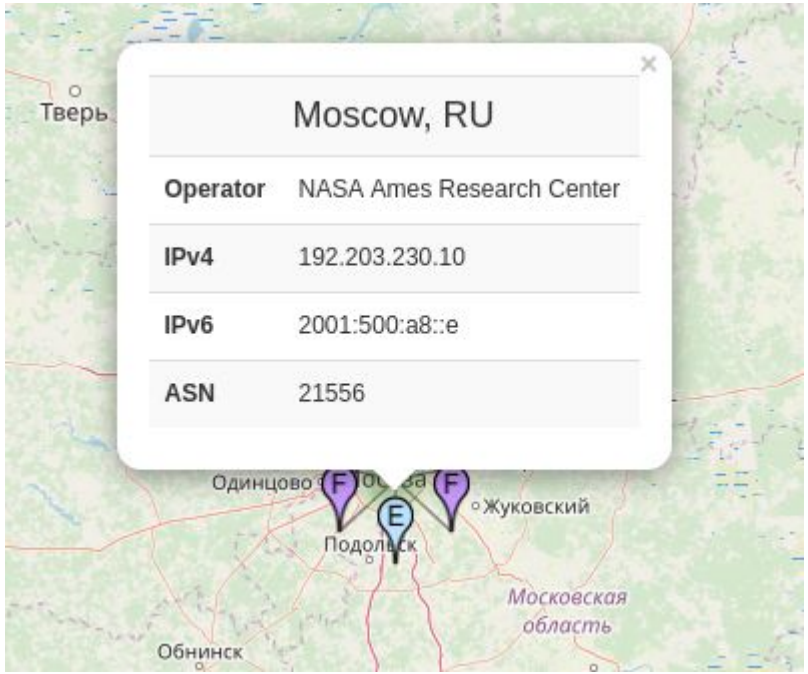
- f.root (Москва — 2 шт.);
- i.root (Санкт-Петербург);
- j.root (Москва, Санкт-Петербург);
- k.root (Москва, Санкт-Петербург, Новосибирск);
- l.root (Москва, Ростов-на-Дону, Екатеринбург).

https://ru.wikipedia.org/wiki/Корневые_серверы_DNS

DNS root servers



DNS root servers



И это тоже влияет на отказоустойчивость

Картинка с https://roskomsvoboda.org/33839/

ОГЛАВЛЕНИЕ

Round robin DNS

alfaconf.pro zone:

- alfaconf.pro A 3600 185.143.173.130
- alfaconf.pro A 3600 185.143.173.149
- alfaconf.pro A 3600 185.143.173.151

Round robin DNS demo

1. Настройка round-robin в DNS-зоне
2. Резолвинг всех хостов, TTL (dig)
3. Выбор одного из хостов (ping)
4. Системный локальный кэш MacOS
5. Chrome & Firefox
6. Особенности кэшей :)

Local DNS resolver caching

DNS-записи кэшируются на клиенте и промежуточных DNS-резолверах

Сброс локального кэша для macOS:

```
sudo killall -HUP mDNSResponder
sudo killall -INFO mDNSResponder
```

Java DNS resolving - работает как надо:

```
InetAddress address = InetAddress.getByName("alfaconf.pro");
```

Browser DNS caching

Настройки в Chrome

- chrome://net-internals/#dns

Настройки в Firefox

- about:config
  - network.dnsCache*

Не забываем про чистку браузерных кэшей

Особенности Round robin DNS

- Пользователь ходит на один IP в течение TTL
- Если этот IP вылетел, то сервис не доступен в течение TTL
- Нет инвалидации DNS кэшей :)
- Нет гарантии решения проблемы при обновлении кэша
- Это в целом проблемы DNS (к Geo DNS тоже относится)
- RR DNS в чистом виде не дает availability

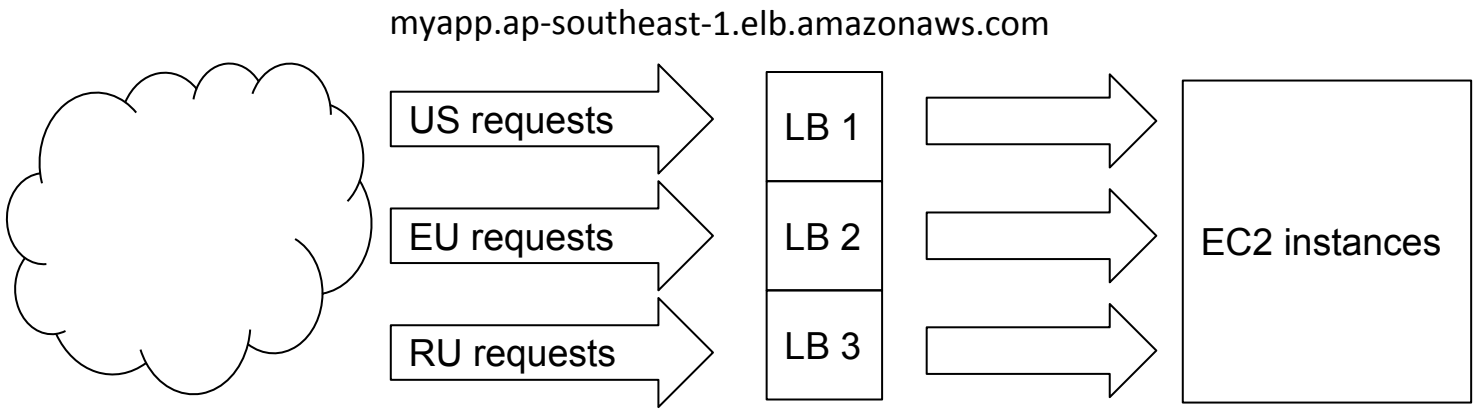
Плюшки RR DNS

- Простота
- RR DNS + алертинг + API у DNS-провайдера = ~availability
- RR DNS + алертинг + быстрый ввод IP = ~availability
- RR DNS хорош для масштабирования нагрузки на LB
- Elastic Load Balancer (ELB - AWS) - использует именно RR DNS

RR DNS - ELB

TTL у ELB = 60 секунд

Scalability, not Availability

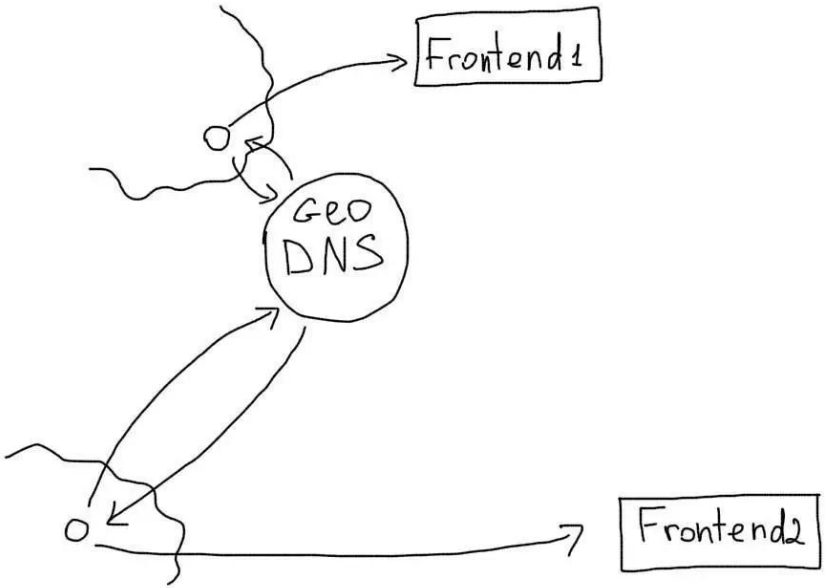


When RR DNS is not enough

- round robin is too simple

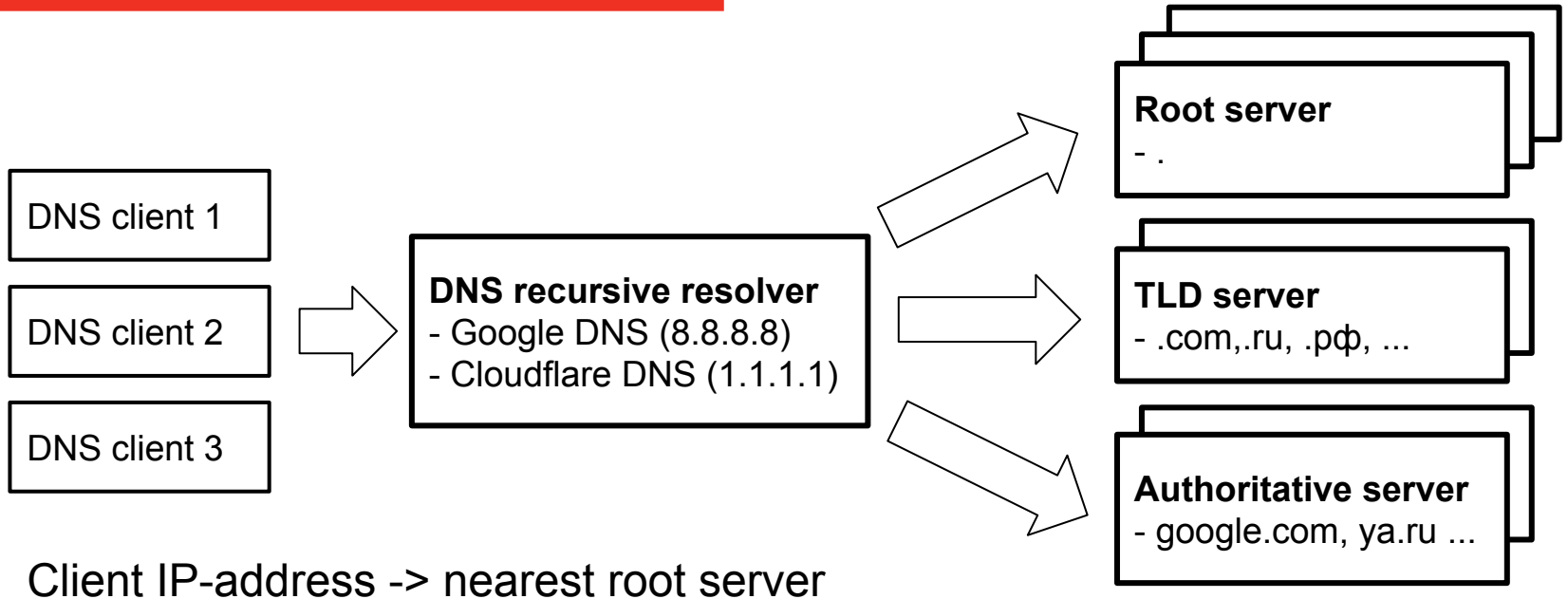
ОГЛАВЛЕНИЕ

GEO DNS



Картинка с https://ruhighload.com/dns+балансировка+

GEO DNS resolving



Client IP-address -> nearest root server

https://en.wikipedia.org/wiki/EDNS_Client_Subnet
https://developers.google.com/speed/public-dns/faq

When DNS is not enough

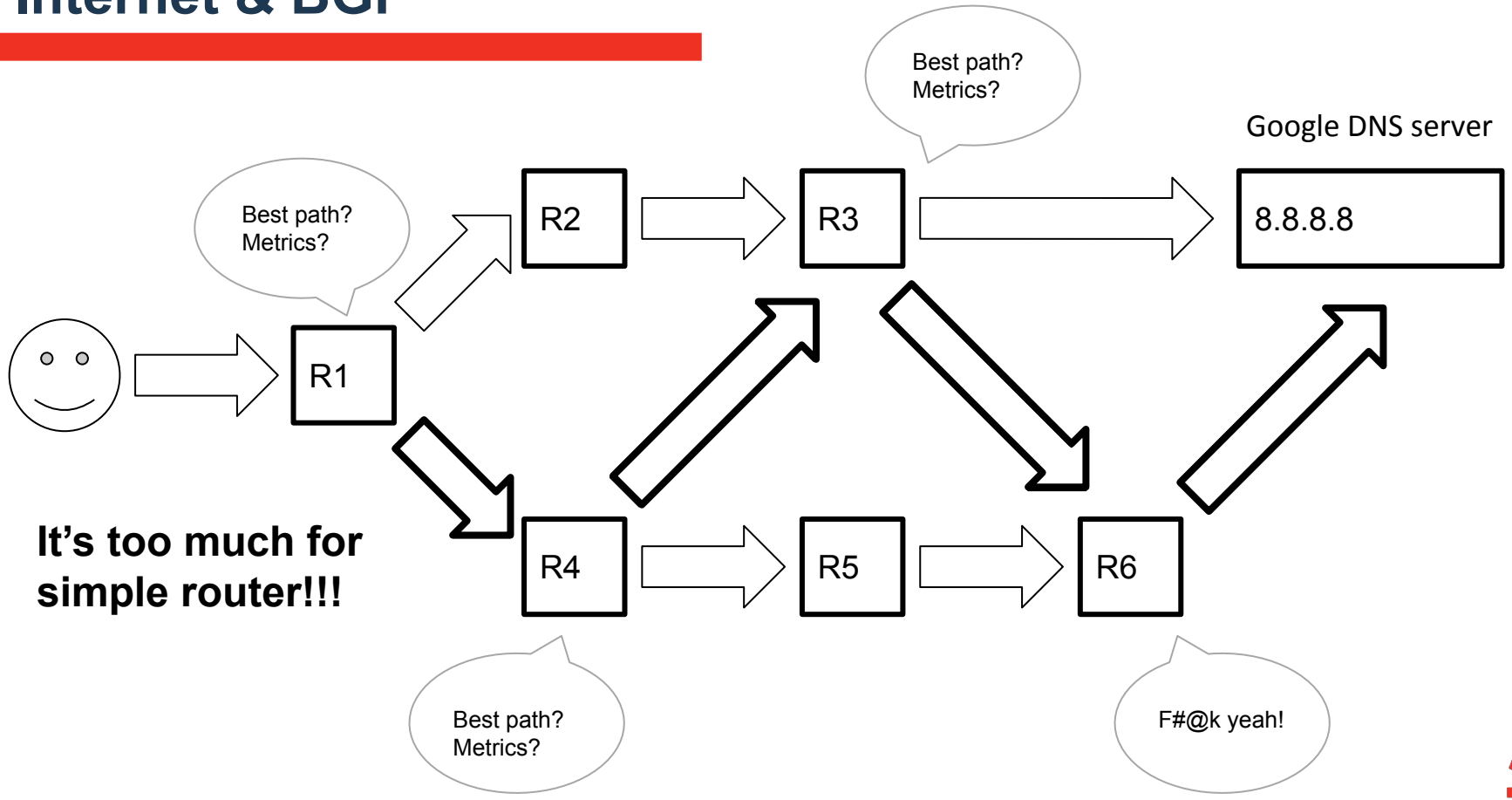
Public DNS servers:

- Google 8.8.8.8, 8.8.4.4
- Cloudflare 1.1.1.1
- OpenDNS
- root servers

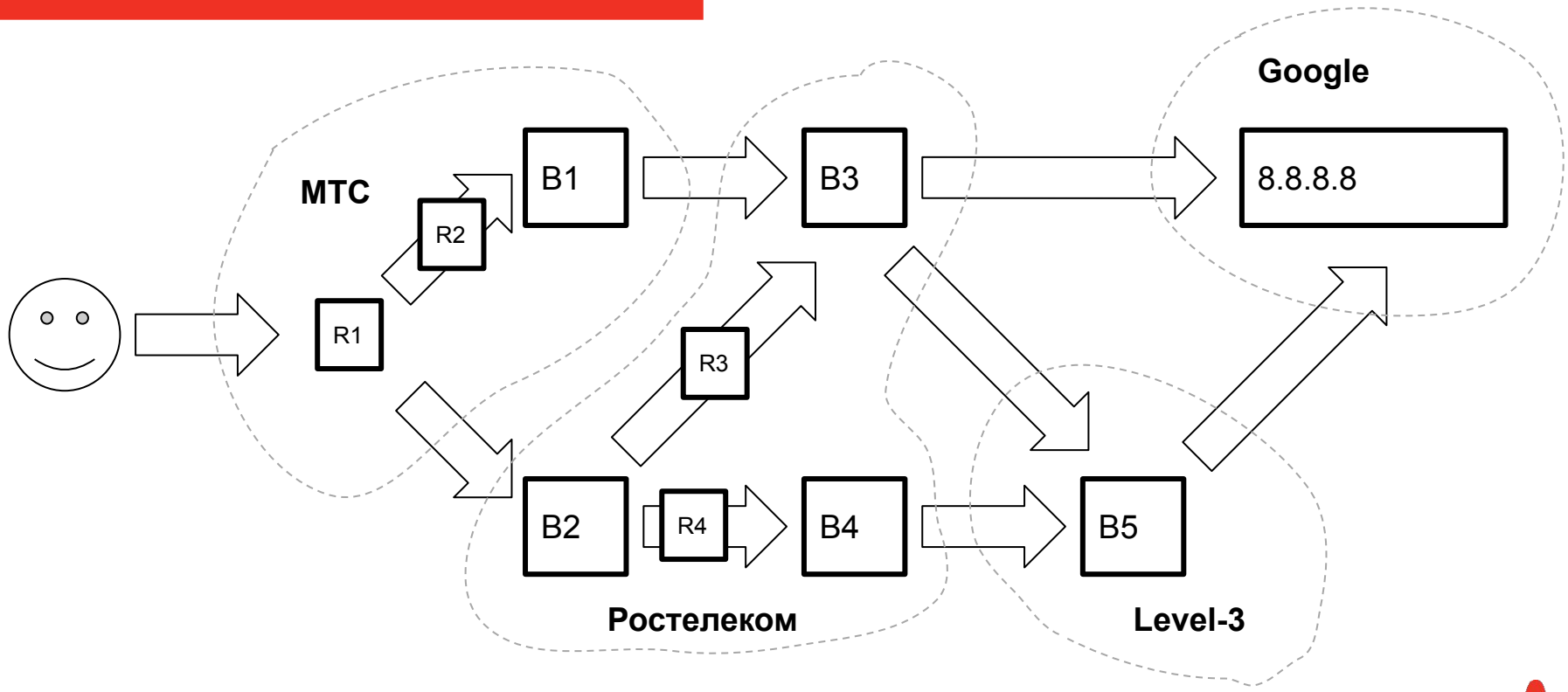
Мы не можем ходить к DNS-серверу по domain name

ОГЛАВЛЕНИЕ

Internet & BGP



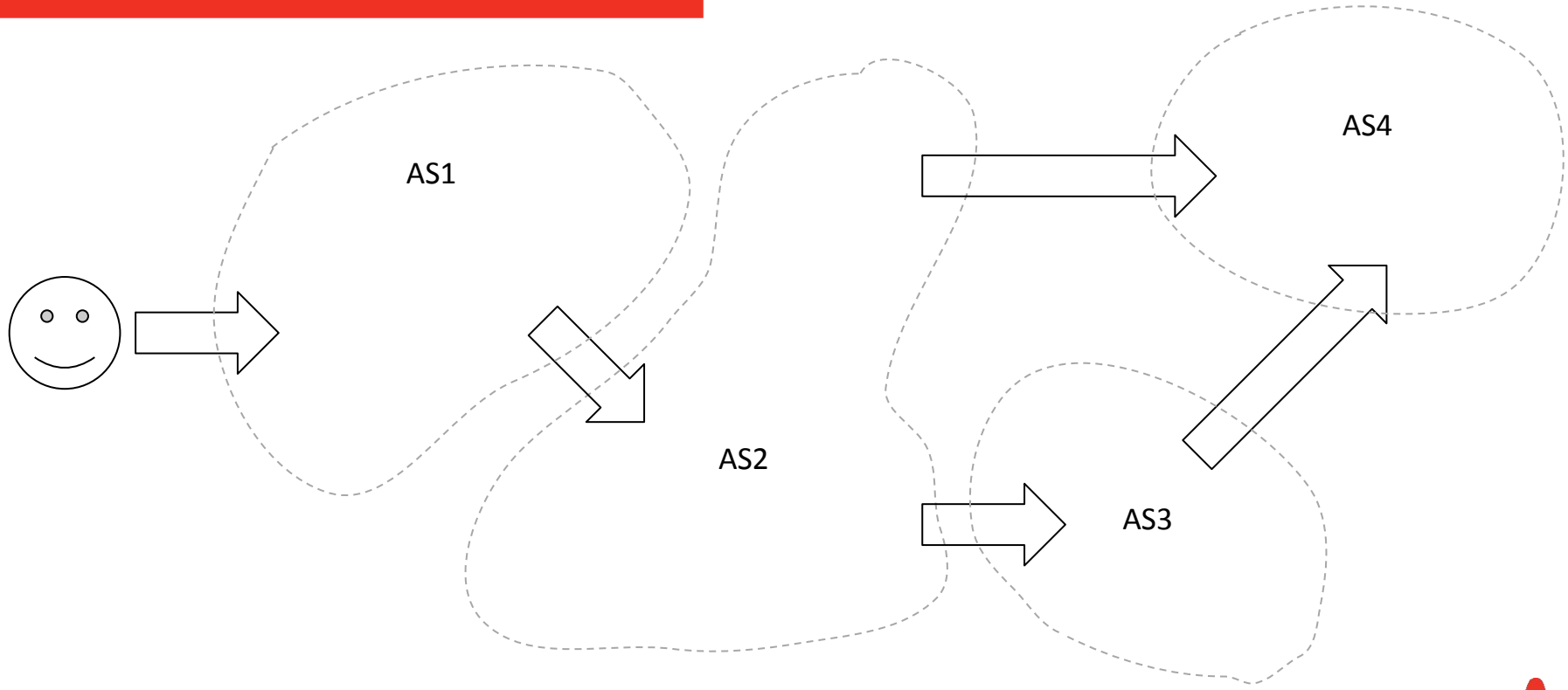
Autonomous systems



Автономная система

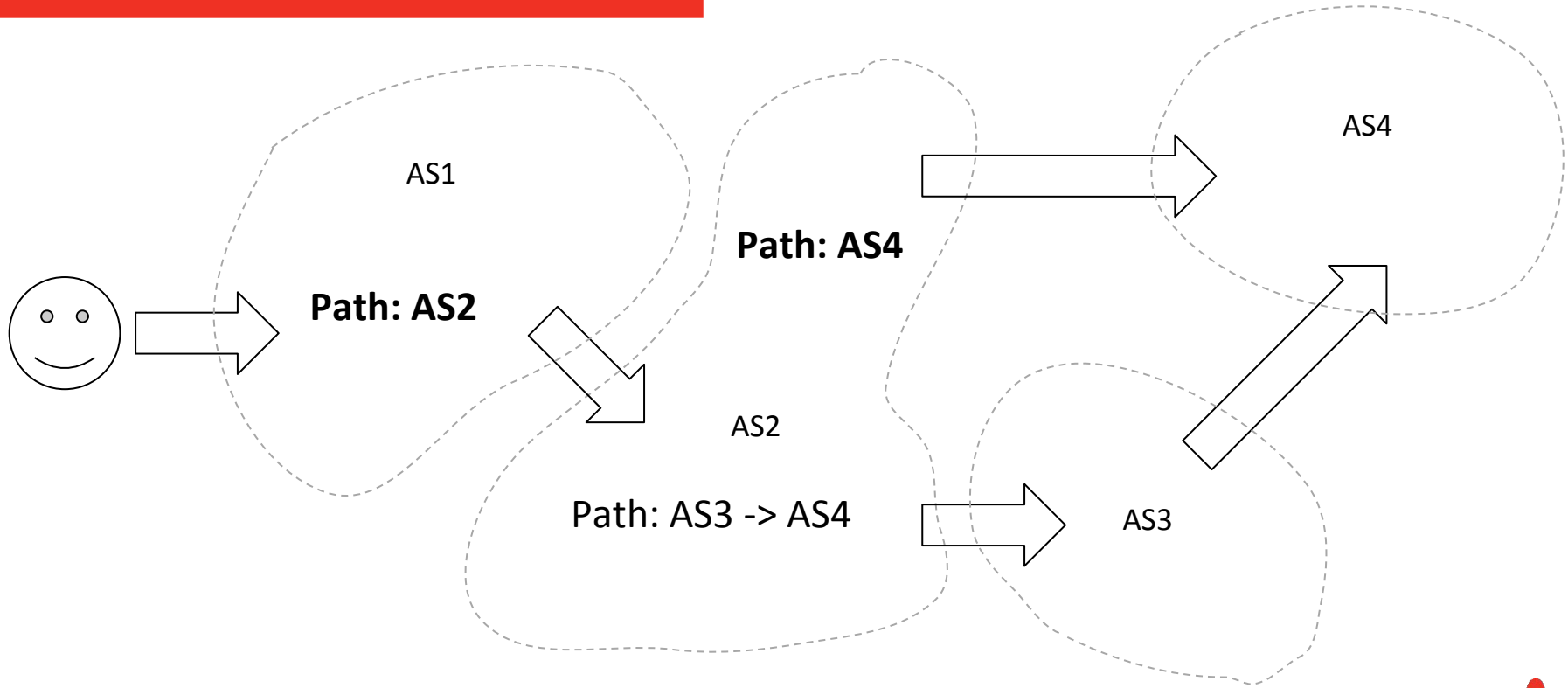
- набор внутренних и внешних (border) маршрутизаторов
- набор IP-сетей (префиксов)
- ограничена одним владельцем (administrative management)
- роутинг внутри AS - её внутреннее дело
- роутинг между AS - BGP
- связанные посредством бордеров AS - соседи (neighbors)

Autonomous systems

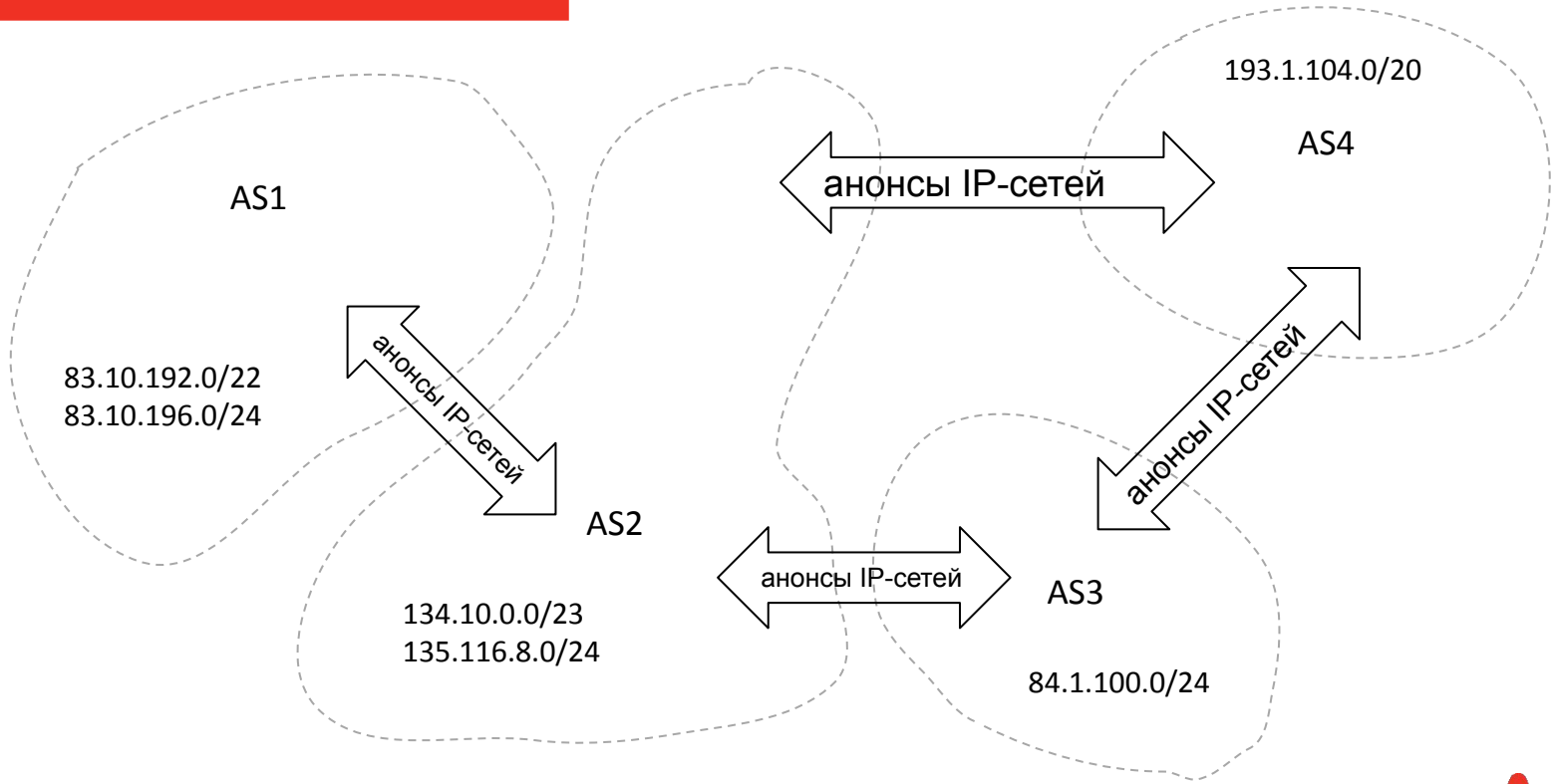


BGP best path

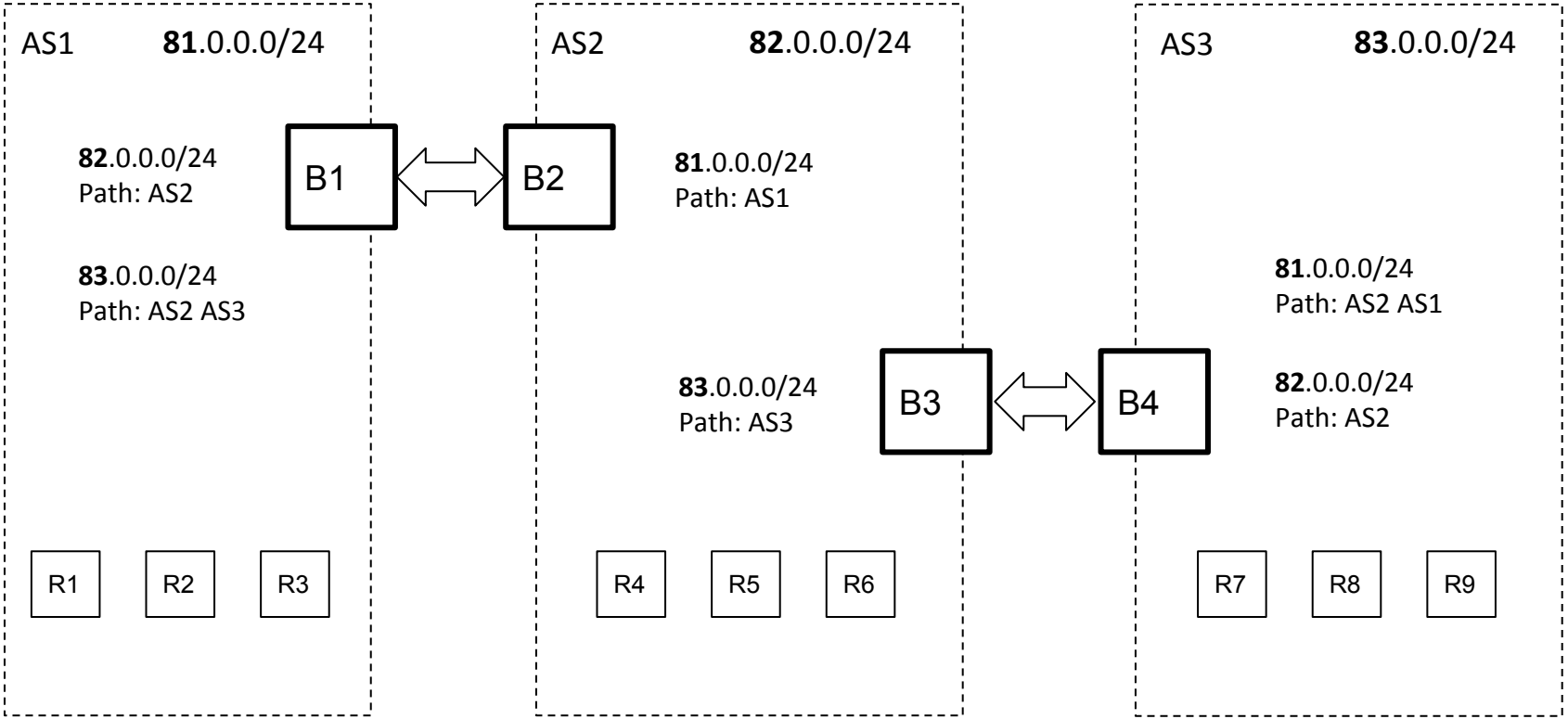
AS path length



BGP peering


193.1.104.0/20
AS4
AS1
анонсы IP-сетей
анонсы IP-сетей
83.10.192.0/22
83.10.196.0/24
анонсы IP-сетей
AS2
анонсы IP-сетей
AS3
134.10.0.0/23
135.116.8.0/24
84.1.100.0/24

BGP peering in details



BGP stats

- ~ 65k AS
- ~ 800k IP-сетей (префиксов)
- ~ 10k изменений анонсов в неделю
- http://as2914.net (galaxy)

https://www.cidr-report.org/as2.0/

BGP - это протокол доверия

- **Нет аутентификации соседа.** Формально есть MD5 :)
- **Нет фильтрации.** У BGP есть фильтры и они описаны, но ими не пользуются, либо пользуются неправильно.
- **Очень просто установить соседство.** Настройка — пара строчек конфига.
- **Не требуются права на управление BGP.** Никто не отберет права за настройку BGP в пьяном виде.

https://habr.com/en/company/oleg-bunin/blog/456582/

Можно анонсить чужие сети

- **Пакистан против YouTube.** В 2008 году ребята из Пакистана решили заблокировать у себя YouTube. Сделали они это настолько хорошо, что без котиков осталось полмира.
- **DV LINK захватил префиксы Google, Apple, Facebook, Microsoft.** В том же 2017 российский провайдер DV LINK начал зачем-то анонсировать сети Google, Apple, Facebook, Microsoft и некоторых других крупных игроков.
- **eNet из США захватил префиксы AWS Route53 и MyEtherwallet.** В 2018 году провайдер из Огайо или кто-то из его клиентов проанонсировал сети Amazon Route53 и криптокошелька MyEtherwallet.
- **Как Verizon и BGP Optimizer устроили большой оффлайн.** 24.06.2019 на небольшую компанию в Пенсильвании хлынул поток трафика из множества маршрутов, проходящих через крупного провайдера Verizon (пострадали Amazon, Linode, Cloudflare)

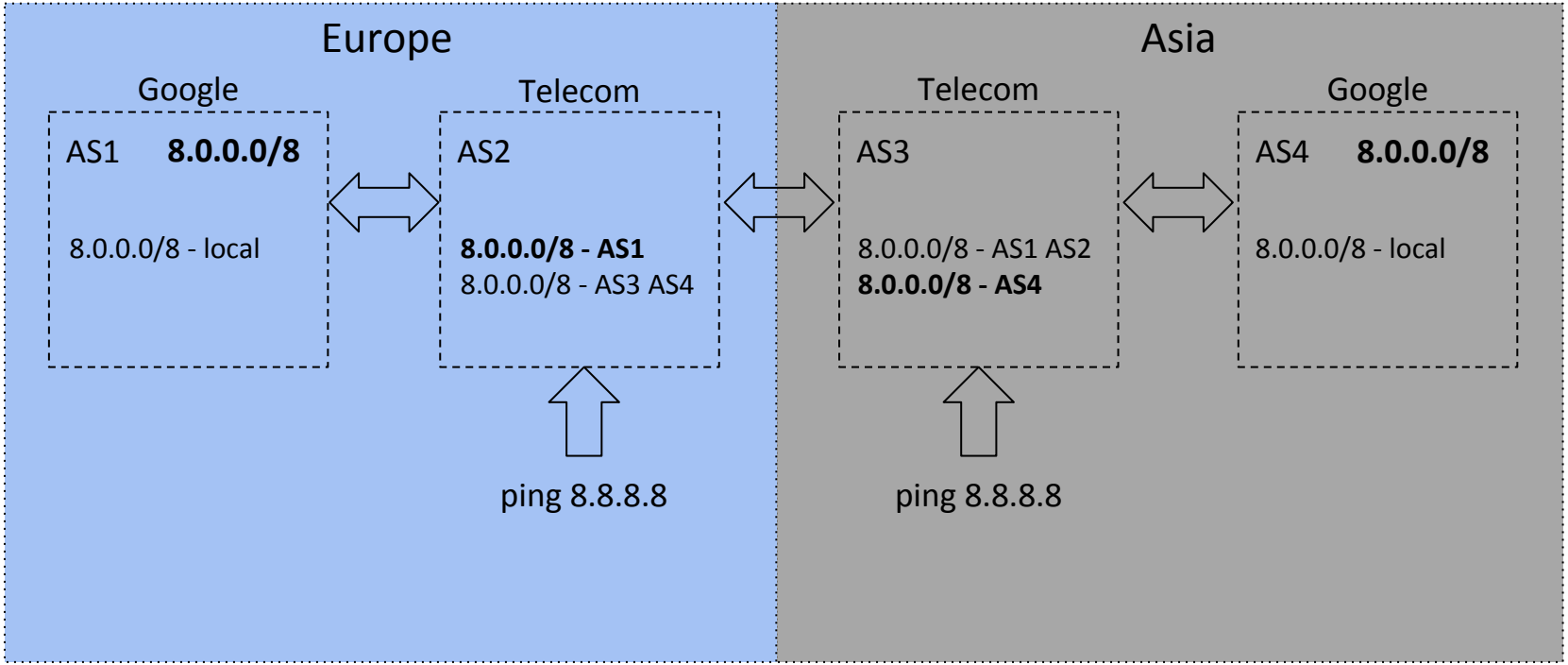
https://habr.com/en/company/oleg-bunin/blog/456582/

ОГЛАВЛЕНИЕ

BGP Anycast

- Но если анонсить свои сети с разных AS ...
- Получим BGP Anycast

BGP Anycast


Europe
Asia
Google
Telecom
Telecom
Google
AS1 8.0.0.0/8
8.0.0.0/8 - local
AS2
8.0.0.0/8 - AS1
8.0.0.0/8 - AS3 AS4
AS3
8.0.0.0/8 - AS1 AS2
8.0.0.0/8 - AS4
AS4 8.0.0.0/8
8.0.0.0/8 - local
ping 8.8.8.8
ping 8.8.8.8

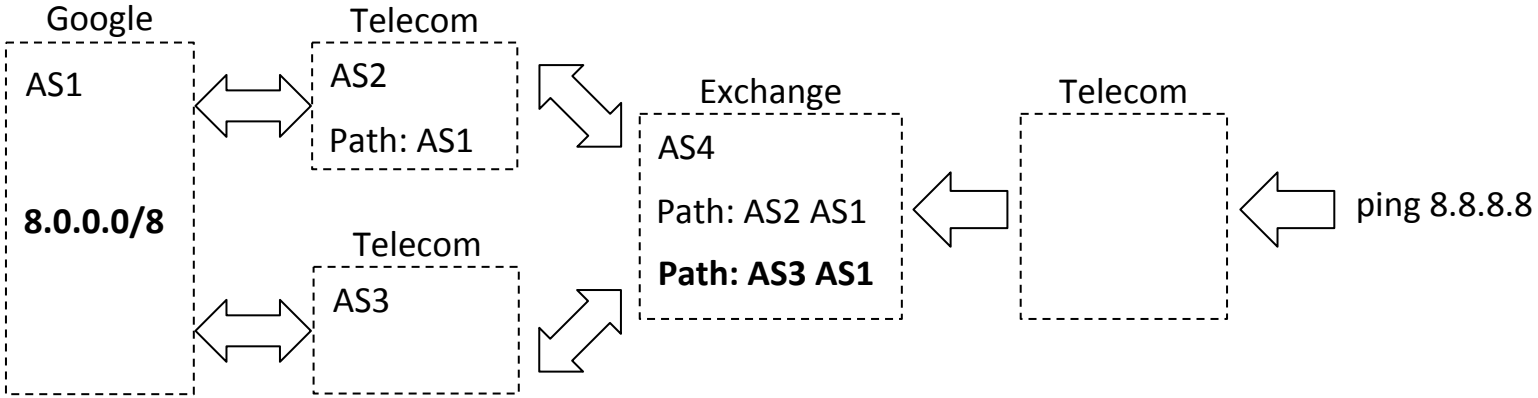
Chicken or egg problem solved

Anycast BGP:

- достаточно просто
- нужны географически разнесенные точки присутствия

ОГЛАВЛЕНИЕ

Multihome BGP



Multihome BGP

- Не нужны разные точки присутствия
- Нужны лишь подключения к разным телеком-провайдерам
- Используют почти все крупные сервисы
- alfabank.ru точно :)

Особенности BGP

- Легко отстрелить себе ногу и пол-интернета
- Во время переключения best path могут быть потери
- Нужны AS, IP-сети, свое железо, каналы связи

Плюшки BGP

- Маршрутная информация обновляется асинхронно, никаких кэшей, время недоступности минимально
- Дает отказоустойчивость на IP-уровне
- Изначально в BGP заложены возможности гибкого управления входящим трафиком (веса, предпочтения, балансировка)

ОГЛАВЛЕНИЕ

ВСЯКОЕ РАЗНОЕ

- HTTP redirects, типа google.com -> google.ru
- да хватит уже

Выводы

- Нельзя ограничиваться приложением / сервером / датацентром
- Даже AWS / GCE / DO / Azure не панацея
- Нужно использовать правильные инструменты

DNS links

- https://www.inetdaemon.com/tutorials/internet/dns/operation/hierarchy.shtml
- https://en.wikipedia.org/wiki/List_of_DNS_record_types
- https://www.cloudflare.com/learning/dns/dns-server-types/
- https://en.wikipedia.org/wiki/Root_name_server
- https://root-servers.org/
- https://ruhighload.com/dns+балансировка+
- https://en.wikipedia.org/wiki/EDNS_Client_Subnet
- https://developers.google.com/speed/public-dns/faq

BGP links

- http://as2914.net
- https://www.cidr-report.org/as2.0/
- https://ru.wikipedia.org/wiki/Anycast
- https://habr.com/en/company/oleg-bunin/blog/456582/
- https://habr.com/ru/company/qrator/blog/457446/
- http://noc.runnet.ru/lg/ or http://lg.rinet.ru/ (looking glasses)

СПАСИБО