МГТУ
им. Н.Э. Баумана
ТЕХНОПАРК
Mail.Ru Group

Проектирование высоконагруженных систем

Лекция №3

Быков Александр

Масштабируемость (англ. scalability) – способность системы справляться с увеличением рабочей нагрузки, при добавлении ресурсов, как правило аппаратных.

Количественно можно оценить как отношение полученного роста производительности к увеличению кол-ва используемых ресурсов. Если отношение близко к единице то масштабирование называют *линейным*.

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Балансировка нагрузки (Load balancing)

Распределение нагрузки на множество серверов бесполезно без возможности распределять нагрузку в нужных пропорциях (балансировать).

Если из-за дисбаланса вся нагрузка попадает на один сервер то никакой масштабируемости быть не может.

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Вертикальное масштабирование

- Доставить процессоров, памяти, дисков
- Купить более мощный сервер
- Купить еще более мощный сервер (очень дорого)
- Соотношение роста производительности на вложенные деньги стремительно падает с выходом за пределы массовых конфигураций
- Наступит момент когда *все равно не хватит*

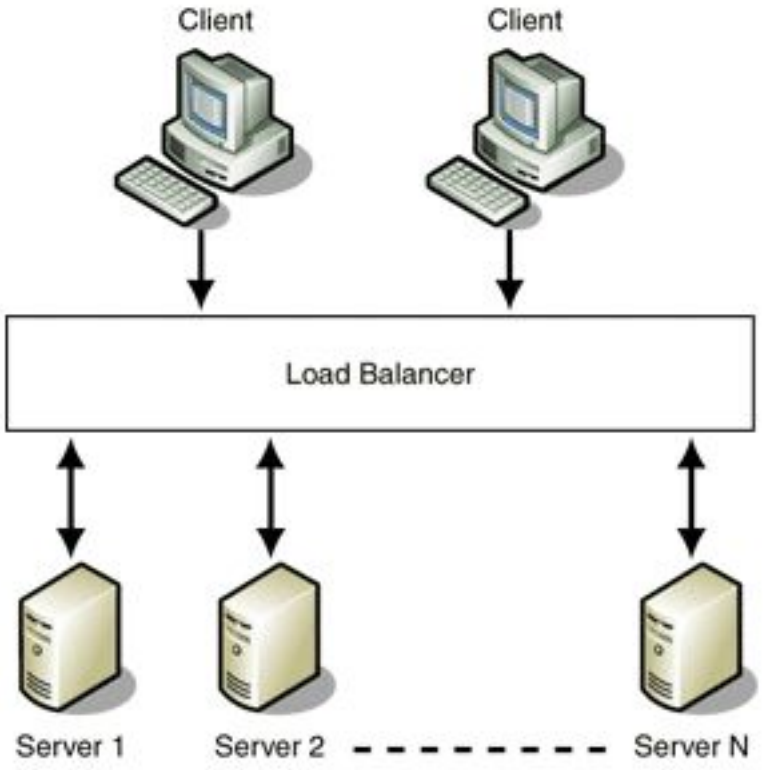
МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Горизонтальное масштабирование

- Разнести нагрузку на несколько серверов
- Возможно там где нет/мало общих ресурсов
- Часто требует изменений в системе
- Возрастает сложность поддержки
- На больших масштабах дает снижение TCO*

*TCO – Total Cost of Ownership
(совокупная стоимость владения)

Балансировка нагрузки



МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Алгоритмы балансировки

Алгоритм	Вес	Состояние	Обратная связь
Random	+		
Round-Robin	+	+	
Hash	(+)		
Least connections		+	+
Least response time		+	+
Least load		+	+
Sticky sessions			

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Round-Robin

Бекенды получают запросы по очереди в одном и том же порядке (циклическое расписание).

Веса:
A – 10
B – 10
C – 10

“ABC” – расписание запросов

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Weighted Round-Robin

Бекенды получают запросы пропорционально весам согласно расписанию.

Веса:
A – 50%
B – 20%
C – 10%

“AAAAABBC” – расписание с пиками

“ABAABAAC” – расписание равномерное

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Weighted Round-Robin

Достоинства:

- Равномерный RPS на бекендах

Подходит:

- Запросы одинаковой стоимости
- Запросы небольшой стоимости

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Hash

Формула:

N (номер сервера) = Hash(IP) % M (количество бекендов)

Возможные варианты:

HASH (src ip)
HASH (src ip + src port)
HASH (src ip + src port + dst ip + dst port) *

* Можно предсоздать много бакетов и балансировать RR

Hash

Достоинства:

- Отсутствие состояния на балансировщике
- Запросы одного пользователя попадают на один бекенд

Недостатки:

- Дисбаланс
- Перераспределение пользователей при смене

Consistent Hash

При изменении состава бекендов мигрирует минимально-возможное количество пользователей.

Пример:

http://nginx.org/en/docs/http/ngx_http_upstream_module.html#hash

Уровни балансировки

- DNS
- Routing
- L4 (IP)
- L7 (HTTP)

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Классификация методов

- GSLB – Global Scale Load Balancing
- LSLB – Local Scale Load Balancing

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Round-Robin DNS

- DNS сервер отвечает несколькими адресами
- При следующем ответе сдвигает список на 1 позицию
- Клиенты обычно берут первый адрес из списка
- Windows Vista когда появилась брала «ближайший»
- Короткий TTL чтобы отключать упавшие сервера

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Round-Robin DNS

Достоинства:

- Простота и дешевизна
- Минимальная нагрузка на DNS сервер
- Отсутствие одной точки входа
- Небольшое добавление отказоустойчивости

Round-Robin DNS

Недостатки:

- Ограниченное кол-во серверов в UDP-ответе
- Игнорирование TTL на кеширующих серверах
- Долгое отключение упавшего сервера
- Сильный дисбаланс между серверами
- Windows Vista берет «ближайший» адрес а не первый
- Невозможен взвешенный Round-Robin

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

mail.ru

```
$ host mail.ru
mail.ru has address 94.100.191.209
mail.ru has address 94.100.191.210
mail.ru has address 94.100.191.241
mail.ru has address 94.100.191.242
mail.ru has address 94.100.191.243
mail.ru has address 94.100.191.244
mail.ru has address 94.100.191.245
mail.ru has address 94.100.191.246
mail.ru has address 94.100.191.247
mail.ru has address 94.100.191.248
mail.ru has address 94.100.191.249
mail.ru has address 94.100.191.250
mail.ru has address 94.100.191.201
mail.ru has address 94.100.191.202
mail.ru has address 94.100.191.203
mail.ru has address 94.100.191.204
mail.ru has address 94.100.191.205
mail.ru has address 94.100.191.206
mail.ru has address 94.100.191.207
mail.ru has address 94.100.191.208
```

```
$ host mail.ru
mail.ru has address 94.100.191.250
mail.ru has address 94.100.191.201
mail.ru has address 94.100.191.202
mail.ru has address 94.100.191.203
mail.ru has address 94.100.191.204
mail.ru has address 94.100.191.205
mail.ru has address 94.100.191.206
mail.ru has address 94.100.191.207
mail.ru has address 94.100.191.208
mail.ru has address 94.100.191.209
mail.ru has address 94.100.191.210
mail.ru has address 94.100.191.241
mail.ru has address 94.100.191.242
mail.ru has address 94.100.191.243
mail.ru has address 94.100.191.244
mail.ru has address 94.100.191.245
mail.ru has address 94.100.191.246
mail.ru has address 94.100.191.247
mail.ru has address 94.100.191.248
mail.ru has address 94.100.191.249
```

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

vk.com

```
$ host vk.com
vk.com has address 87.240.131.100
vk.com has address 87.240.131.101
vk.com has address 87.240.131.102
vk.com has address 87.240.131.103
vk.com has address 87.240.131.104
vk.com has address 87.240.131.117
vk.com has address 87.240.131.118
vk.com has address 87.240.131.119
vk.com has address 87.240.131.120
vk.com has address 87.240.143.241
vk.com has address 87.240.143.242
vk.com has address 87.240.143.243
vk.com has address 87.240.143.244
vk.com has address 87.240.143.245
vk.com has address 87.240.143.246
vk.com has address 87.240.143.247
vk.com has address 87.240.143.248
vk.com has address 93.186.224.244
vk.com has address 93.186.224.245
vk.com has address 93.186.224.246
vk.com has address 87.240.131.97
vk.com has address 87.240.131.98
vk.com has address 87.240.131.99
```

```
$ host vk.com
vk.com has address 87.240.131.101
vk.com has address 87.240.131.102
vk.com has address 87.240.131.103
vk.com has address 87.240.131.104
vk.com has address 87.240.131.117
vk.com has address 87.240.131.118
vk.com has address 87.240.131.119
vk.com has address 87.240.131.120
vk.com has address 87.240.143.241
vk.com has address 87.240.143.242
vk.com has address 87.240.143.243
vk.com has address 87.240.143.244
vk.com has address 87.240.143.245
vk.com has address 87.240.143.246
vk.com has address 87.240.143.247
vk.com has address 87.240.143.248
vk.com has address 93.186.224.244
vk.com has address 93.186.224.245
vk.com has address 93.186.224.246
vk.com has address 87.240.131.97
vk.com has address 87.240.131.98
vk.com has address 87.240.131.99
vk.com has address 87.240.131.100
```

www.yandex.ru

```
$ host www.yandex.ru
www.yandex.ru has address 87.250.250.203
www.yandex.ru has address 87.250.251.3
www.yandex.ru has address 93.158.134.3
www.yandex.ru has address 93.158.134.203
www.yandex.ru has address 213.180.193.3
www.yandex.ru has address 213.180.204.3
www.yandex.ru has address 77.88.21.3
www.yandex.ru has address 87.250.250.3
```

```
$ host www.yandex.ru
www.yandex.ru has address 87.250.251.3
www.yandex.ru has address 93.158.134.3
www.yandex.ru has address 93.158.134.203
www.yandex.ru has address 213.180.193.3
www.yandex.ru has address 213.180.204.3
www.yandex.ru has address 77.88.21.3
www.yandex.ru has address 87.250.250.3
www.yandex.ru has address 87.250.250.203
```

google.com

```
$ host google.com
google.com has address 173.194.32.227
google.com has address 173.194.32.228
google.com has address 173.194.32.229
google.com has address 173.194.32.230
google.com has address 173.194.32.231
google.com has address 173.194.32.232
google.com has address 173.194.32.233
google.com has address 173.194.32.238
google.com has address 173.194.32.224
google.com has address 173.194.32.225
google.com has address 173.194.32.226
```

```
$ host google.com
google.com has address 173.194.32.228
google.com has address 173.194.32.229
google.com has address 173.194.32.230
google.com has address 173.194.32.231
google.com has address 173.194.32.232
google.com has address 173.194.32.233
google.com has address 173.194.32.238
google.com has address 173.194.32.224
google.com has address 173.194.32.225
google.com has address 173.194.32.226
google.com has address 173.194.32.227
```

Xixi DNS

- Выдает один случайный адрес в ответе
- Взвешенный Round-Robin

Geo-based DNS

- Сервер выдает адрес ближайшего к пользователю ДЦ
- Сервер видит адрес Resolver а не конечного клиента

Возможные пути решения проблемы:

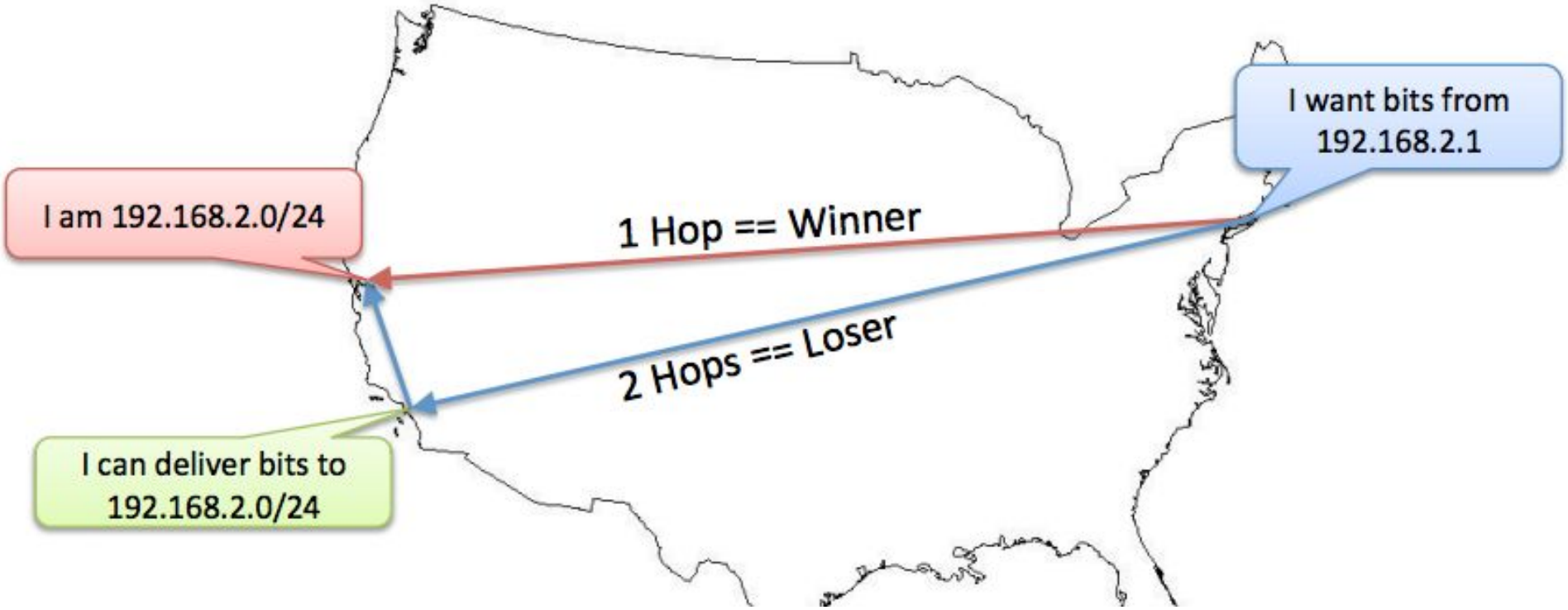
- DNS сервер 8.8.8.8 от Google
- Google предлагает добавить в DNS-запрос IP клиента

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Latency-based DNS

- Сервер выдает адрес ближайшего к пользователю ДЦ с с минимальным RTT
- Сервер видит адрес Resolver а не конечного клиента
- Пример: AWS Route 53

BGP Anycast


I am 192.168.2.0/24
I want bits from 192.168.2.1
1 Hop == Winner
2 Hops == Loser
I can deliver bits to 192.168.2.0/24

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

BGP Anycast

Недостатки:

- Сложность конфигурации и поддержки
- Поломка соединений при перескоке трафика между ДЦ
- Необходима своя AS и сеть /24*

Hardware Load Balancers

- Cisco CSS (L4)
- Cisco ACE (L7)
- F5 BIG-IP
- Citrix NetScaler
- Radware ADC

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Software Load Balancers

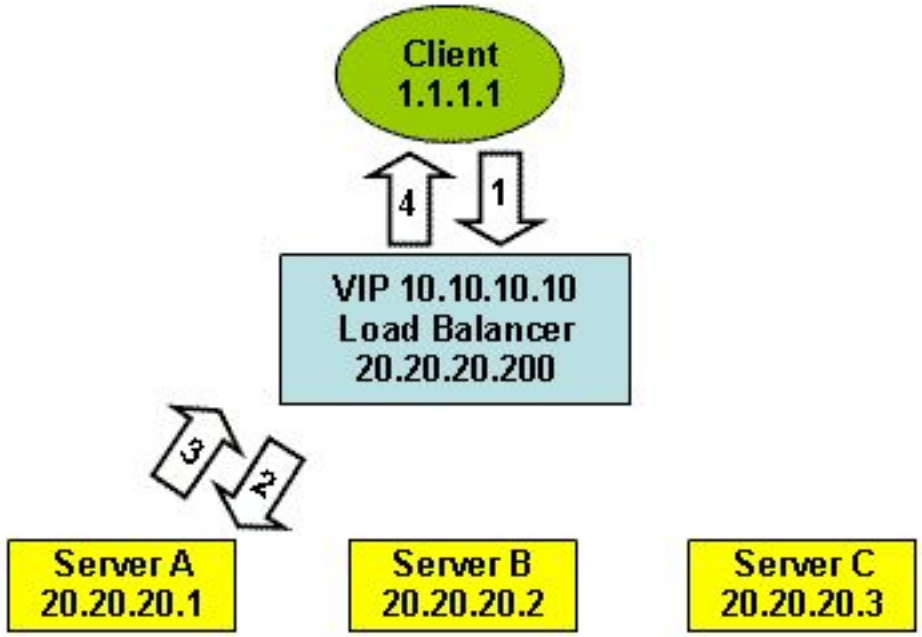
- LVS (Linux Virtual Server)
- Nginx
- HAProxy
- ATS

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Layer 3/4 Load Balancing

- Virtual Server via NAT
- Virtual Server via IP Tunneling
- Virtual Server via Direct Routing

Layer 4: Virtual Server via NAT



МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Layer 4: Virtual Server via NAT

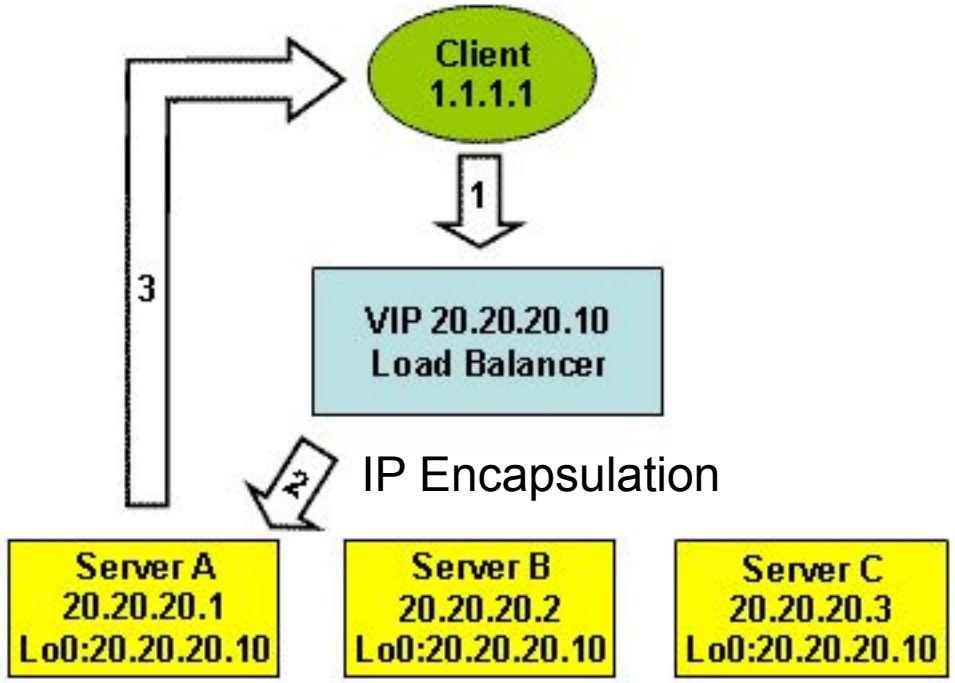
Плюсы:

- Сервера могут быть в разных физических сетях

Минусы:

- Большая нагрузка на процессор
- Весь обратный трафик идет через балансер

Layer 4: Virtual Server via IP Tunneling



Layer 4: Virtual Server via IP Tunneling

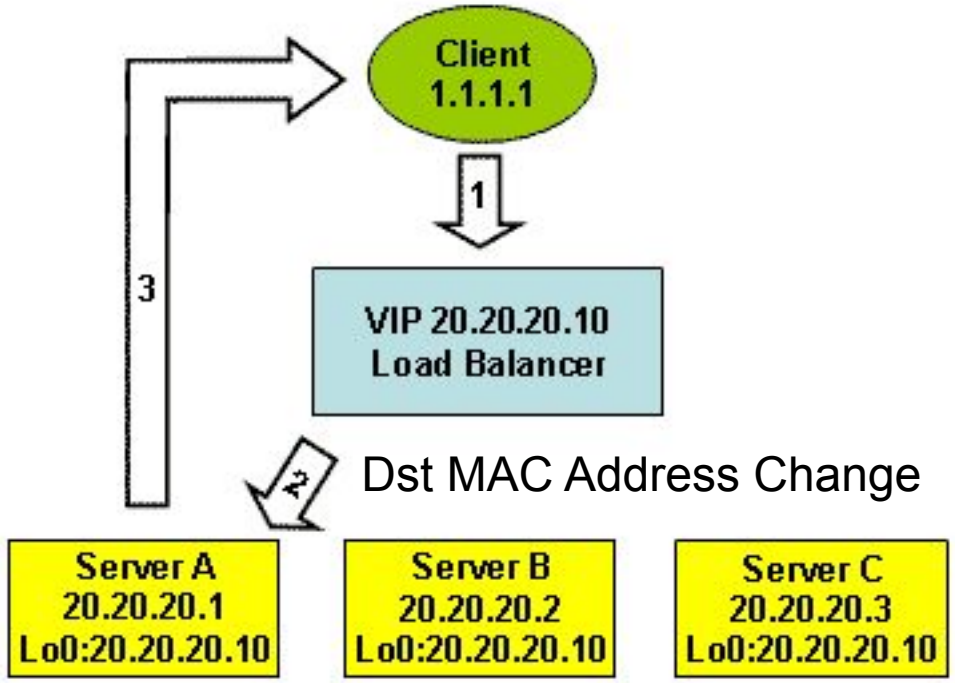
Плюсы:

- Сервера могут быть в разных физических сетях
- Высокая производительность

Минусы:

- Дополнительная нагрузка на процессор
- Сложная настройка инкапсуляции

Layer 3: Virtual Server via Direct Routing



Layer 3: Virtual Server via Direct Routing

Плюсы:

- Высокая производительность

Минусы:

- Сервера должны быть в одной физической сети

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

keepalived

- Программный монитор доступности нод
- Сигнализирует балансеру при падении/подъемы ноды
- Умеет VRRP/CARP резервирование нод между собой

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Проверка работоспособности серверов

- Отвечает на PING
- Принимает соединение на порт
- Отвечает на простой HEAD или GET запрос
- Отвечает на специальный запрос (cgi-bin/ping)

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

BGP/RIP балансировка внутри ДЦ

- То же самое что для для балансинга между ДЦ
- Требует настройки на сетевом оборудовании
- Требует специального размещения оборудования

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Layer 7 Load Balancing

HTTP Reverse Proxy:

- TCP multiplexing
- Persistence / Sticky sessions / Client affinity
- Кеширование
- SSL Termination
- Gzip
- Гибкие настройки для учета бизнес-логики

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Layer 7 Load Balancing

Плюсы:

- Высокая гибкость конфигурации
- Надежное резервирование
- Равномерное распределение нагрузки
- Решение проблемы медленных клиентов

Минусы:

- Относительно низкая производительность
- Необходима модификация ПО для приема X-Real-IP

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Выбор таймаута на запрос к upstream

(Оптимизационная задача)

Проблемы:

- Небольшой timeout: обрежем долгие живые запросы
- Большой timeout: затормозим обычные запросы

Идеи:

- Использование квантилей для выбора значения
- Разные настройки для разных типов запросов

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Политика failover

Алгоритм выбора другого (следующего) бекенда в случае отказа текущего в процессе обработки запроса

На примере nginx:

proxy_next_upstream <список ситуаций>

proxy_next_upstream_timeout

proxy_next_upstream_tries

Балансировка внутри проекта

Варианты:

1. DNS
2. L4
3. L7 на отдельных серверах
4. L7 локальный на клиенте (aka sidecar proxy)

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

SSL Termination

- Session cache – работает в пределах одного IP
- Session tickets – поддерживают не все браузеры
- Perfect Forward Secrecy (PFS)
- Application Transport Security (ATS) в iOS 9
- Let’s Encrypt – бесплатные SSL-сертификаты

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Redirect Based Load Balancing

- Почти также дешев как DNS
- Увеличивает задержку на каждый запрос
- Имеет единую точку отказа
- Хорош для выдачи ближайшего к пользователю сервера

Application Based Load Balancing

- Выдаем в приложении прямые ссылки на сервера
- Обеспечиваем балансировку и отказоустойчивость
- При хорошей реализации одно из самых эффективных решений

Минусы:

- Неудобно программировать
- Раскрываем внутреннее устройство проекта

Client Based Load Balancing

- Выбор сервера в коде страницы в браузере
- Легко обеспечить отказоустойчивость запросив другой сервер

Минусы:

- Плохо совместим с AJAX из-за crossdomain policy

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Функциональное разделение

- Уносим форум на отдельный домен
- Уносим статику на отдельный домен
- И так далее…

Минусы:

- Сложно администрировать
- Легко попасть на неделимый кусок сайта

МГТУ им. Н.Э. Баумана ТЕХНОПАРК Mail.Ru Group

Домашнее задание №2

- Собрать Load Balancer в облачном сервисе на выбор (AWS, DigitalOcean, Windows Azure, Google Cloud etc)
- Распределить нагрузку на несколько серверов (L4: NAT, tunneling, L7: nginx, Haproxy)*
- Продемонстрировать распределение нагрузки (системные и пользовательские метрики)
- Продемонстрировать срабатывание механизмов отказоустойчивости к падению одного сервера

СПАСИБО ЗА ВНИМАНИЕ

Быков Александр

bykov@corp.mail.ru