

Настройка балансировки нагрузки с масштабированием в GCP

Описание

Экземпляры WCS на Google Cloud Platform поддерживают балансировку нагрузки при помощи TCP Network Load Balancer.

При этом WebSocket-соединения будут автоматически распределены между активными серверами в балансировщике нагрузки. В случае применения заданной политики масштабирования (если целевой показатель, например, загрузка процессора на сервере, достиг заданного значения) будут запущены новые экземпляры сервера и автоматически добавлены в балансировщик.

Для настройки необходимо создать следующие компоненты

- Образ диска, который будет использоваться в шаблоне при создании нового экземпляра
- Шаблон, на основе которого будут создаваться новые экземпляры сервера при масштабировании
- Группа масштабирования
- Балансировщик нагрузки
- Настройки контроля активности сервера

Рассмотрим пример развертывания CDN для доставки WebRTC потоков, состоящей из одного Origin и группы масштабирования Edge (от 1 до 3 экземпляров) с масштабированием по загрузке процессора.

Подготовка серверов

1. Разверните Origin и Edge серверы, как описано [здесь](#). Назначьте Origin серверу статический внутренний IP адрес. Зарезервируйте статический внешний IP адрес для балансировщика.
2. Настройте CDN на стороне Origin сервера

```
cdn_enabled           = true
cdn_ip                = <origin_internal_ip>
cdn_role              = origin
cdn_nodes_resolve_ip = false
```

3. Настройте CDN на стороне Edge сервера

```
cdn_enabled           = true
cdn_ip                = <edge_internal_ip>
cdn_point_of_entry   = <origin_internal_ip>
cdn_role              = edge
cdn_nodes_resolve_ip = false
```

4. В настройке Edge сервера укажите параметр

```
http_enable_root_redirect=false
```


5. Подготовьте и [импортируйте](#) SSL сертификаты на Origin и Edge серверы. Не рекомендуется использовать Let'sEncrypt, поскольку это приведет к необходимости обновлять образ диска Edge сервера каждые три месяца.

Создание образа диска Edge сервера

1. Остановите экземпляр Edge сервера
2. Перейдите в раздел **Compute Engine** - **Images**, нажмите **Create image**. Выберите в качестве диска-источника диск экземпляра Edge сервера и нажмите **Create**

← Create an image

Name ?
Name is permanent

test-edge-image-1 

Source ?
Disk

Source disk ?
test-edge-1

Location ?
 Multi-regional
 Regional
eu (European Union) (default)

Family (Optional) ?

Description (Optional)

Labels ? (Optional)


+ Add label

Encryption
Data is encrypted automatically. Select an encryption key management solution.

Google-managed key
No configuration required

Customer-managed key
Manage via Google Cloud Key Management Service

Customer-supplied key
Manage outside of Google Cloud

Your free trial credit will be used for this image. [GCP Free Tier](#) 

Create **Cancel**

Equivalent [REST](#) or [command line](#)

После создания образа диска не удаляйте исходный экземпляр Edge сервера, он потребуется при изменении настроек.

Создание шаблона Edge сервера

1. Перейдите в раздел **Compute Engine** - **Instance templates**, нажмите **Create image**.
Выберите конфигурацию VM

← Create an instance template

Describe a VM instance once and then use that template to create groups of identical instances [Learn more](#)

Name ?
Name is permanent

Machine configuration

Machine family

[General-purpose](#) [Memory-optimized](#) [Compute-optimized](#)

Machine types for common workloads, optimized for cost and flexibility


Series

N1 ▼

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type

n1-standard-1 (1 vCPU, 3.75 GB memory) ▼

	vCPU	Memory
	1	3.75 GB


⌵ CPU platform and GPU

Container ?

Deploy a container image to this VM instance. [Learn more](#)

2. В разделе **Boot disk** нажмите **Change**

Boot disk ?



New 20 GB standard persistent disk
Image
test-edge-image-1

[Change](#)

На вкладке **Custom images** выберите образ диска Edge сервера

Boot disk

Select an image to create a boot disk. The image determines the operating system installed on the instance. Can't find what you're looking for? Explore hundreds of VM solutions in [Marketplace](#).

Show images from
Test GCP LB

Show deprecated images

Image
test-edge-image-1

Created on Jun 25, 2020, 1:53:31 PM

Boot disk type Size (GB)

Standard persistent disk 20

3. На вкладке **Security** добавьте публичный ключ для доступа к серверу по SSH, если у Вас нет ключей, привязанных к проекту, и нажмите **Create**

Management **Security** Disks Networking Sole Tenancy

Shielded VM [?](#)
Turn on all settings for the most secure configuration.

Turn on Secure Boot [?](#)
 Turn on vTPM [?](#)
 Turn on Integrity Monitoring [?](#)

SSH Keys
These keys allow access only to this instance, unlike [project-wide SSH keys](#) [Learn more](#)

Block project-wide SSH keys
When checked, project-wide SSH keys cannot access this instance [Learn more](#)

gcp

```
gTaJ8gvi6x9RQB6niVuTN80cK3H1A4xINxQ29GGxWJ  
wXe4kRKIkM4QnxUTsNNsC6yc/d57Ur773518Tevf3v  
4GcWQ9gCPvoIIHZqE79zB0xbRhggjj4ED1rRbC11ug0  
uGO+2kaChLkxHehJ+Xotz/NW0Az0cwkW1YSZGDditT  
vICrIDvRXFD0nuSuj8EpBU3Jjj54zChTI2k4dUDcPY  
kA/bAgy2tF5Ajc50ZCPIVcOu74R1/7RZ1YqgIJ1g+L  
aB_gcp
```

[+ Add item](#)

[^ Less](#)

You can always create instance templates free of charge. Your free trial credit won't be used.

[Create](#) [Cancel](#)

Equivalent [REST](#) or [command line](#)

Создание группы масштабирования

1. Перейдите в раздел **Compute Engine** - **Instance groups**, нажмите **Create instance group**. Выберите регион и зону расположения группы, укажите шаблон Edge

сервера

← Create an instance group

To create an instance group, select one of the options:

- New managed instance group**
A group of VMs created from a template.
Supports autohealing, autoscaling, auto updating, regional deployments, and load balancing.
- New managed instance group for stateful workloads**
A group of VMs created from a template, with preserved disks and metadata individually for each VM.
Supports autohealing, auto updating, regional deployments, and load balancing for stateful workloads.
- New unmanaged instance group**
A group of existing VMs that you manage.
Supports load balancing.

Organize VM instances in a group to manage them together. [Instance groups](#)

Name ⓘ
Name is permanent
test-edge-instance-group

Description (Optional)

Location
To ensure higher availability, select a multiple zone location for an instance group. [Learn more](#)

Single zone
 Multiple zones

Region ⓘ
Region is permanent
europe-west3 (Frankfurt)

Zone ⓘ
Zone is permanent
europe-west3-c

[Specify port name mapping](#) (Optional)

Instance template ⓘ
test-edge-template

Number of instances
Based on autoscaling configuration

2. Выберите режим **Autoscale** по метрике **CPU utilization**, укажите целевую величину **80%** и максимальное количество экземпляров **3**

Autoscaling
Use autoscaling to allow automatic resizing of this instance group for periods of high and low load. [Autoscaling groups of instances](#) ↗

Autoscaling mode

Autoscale

Autoscaling metrics
Use metrics to determine when to autoscale the group.
[Autoscaling policy and target utilization](#) ↗

New metric ^

Metric type

CPU utilization

Target CPU utilization ?

80 %

Done Cancel

+ Add new metric

Cool down period ?
Specify how long to wait for a new instance before taking its metrics into account.
[Cool down period](#) ↗

60 seconds

Minimum number of instances ? 1

Maximum number of instances ? 3

Scale In Controls ?
Prevent a sudden drop in the number of running VM instances in the group by controlling the process of scaling in. [Learn more](#)

Enable Scale In Controls

Delete autoscaling configuration

3. Включите проверку состояния VM (Auto healing) и создайте настройку проверки сервера. Укажите протокол TCP, порт `8081` и запрос `/health-check`

← Create a health check

Health checking mechanisms determine whether VM instances respond properly to traffic. You cannot create a legacy health check using this page. For more information, refer to the [Health Checks Concepts](#) documentation.

Name
wcs-health-check  

Description
WCS health checking 

Scope

- Global
 Regional

Protocol
TCP 

Port
8081 

Proxy protocol
NONE 

Request
/health-check 

Response


Logs

Turning on Health check logs can increase costs in Stackdriver.

- On
 Off

Настройте критерии проверки и нажмите **Create** для возврата к редактированию

группы

Health criteria

Define how health is determined: how often to check, how long to wait for a response, and how many successful or failed attempts are decisive

Check interval seconds [?](#) **Timeout** seconds [?](#)

Healthy threshold consecutive successes [?](#)

Unhealthy threshold consecutive failures [?](#)

You can create this health check free of charge

[CREATE](#) [CANCEL](#) Equivalent [REST](#) or [command line](#)

4. Разверните пункт **Advanced creation options** и установите переключатель **Do not retry machine creation**, затем нажмите **Create**

Advanced creation options

Advanced configuration controlling how the instance group is created

Do not retry machine creation.
If Compute Engine hits a usage limit or error during instance creation, then reduce the instance group size to create as many instances as possible.

[^ Hide advanced creation options](#)

Your free trial credit will be used for VM instances in this group. [GCP Free Tier](#) [↗](#)

[Create](#) [Cancel](#)

Equivalent [REST](#) or [command line](#)

Группа масштабирования будет создана, и один экземпляр будет запущен

Instance groups									
CREATE INSTANCE GROUP REFRESH DELETE									
Instance groups are collections of VM instances that use load balancing and automated services, like autoscaling and autohealing. Learn more									
Filter resources Columns									
Name	Zone	Instances	Template	Group type	Creation time	Recommendation	Autoscaling	In use by	
<input checked="" type="checkbox"/> test-edge-instance-group	europa-west3-c	1	test-edge-template	Managed	Jun 25, 2020, 2:09:10 PM		On: Target CPU utilization 80%		

Создание балансировщика нагрузки

1. Перейдите в раздел **Network** – **Load balancers** и нажмите **Create load balancer**. Выберите **TCP Load Balancing**

← Create a load balancer

HTTP(S) Load Balancing
Layer 7 load balancing for HTTP and HTTPS applications [Learn more](#)
Configure
HTTP LB
HTTPS LB (includes HTTP/2 LB)
Options
Internet-facing or internal
Single or multi-region
[Start configuration](#)

TCP Load Balancing
Layer 4 load balancing or proxy for applications that rely on TCP/SSL protocol [Learn more](#)
Configure
TCP LB
SSL Proxy
TCP Proxy
Options
Internet-facing or internal
Single or multi-region
[Start configuration](#)

UDP Load Balancing
Layer 4 load balancing for applications that rely on UDP protocol [Learn more](#)
Configure
UDP LB
Options
Internet-facing or internal
Single-region
[Start configuration](#)

2. Выберите внешний балансировщик **From internet to my VMs** и регион расположения балансировщика

← Create a load balancer

Please answer a few questions to help us select the right load balancing type for your application

Internet facing or internal only
Do you want to load balance traffic from the Internet to your VMs or only between VMs in your network?

From Internet to my VMs
 Only between my VMs

Multiple regions or single region
Do you want to place the backends for your load balancer in a single region or across multiple regions?

Multiple regions (or not sure yet)
 Single region only

[Continue](#)

3. В разделе **Backend configuration**, на вкладке **Select existing instance groups** выберите группу масштабирования Edge серверов и укажите привязку сессии к IP и протоколу клиента

New TCP load balancer

Name [?]
Name is permanent
test-lb

Backend configuration
Your backend is configured →

Frontend configuration
You have not configured your frontend yet

Review and finalize
Optional

Create **Cancel**

Backend configuration

Name [?]
test-lb

Region [?]
europe-west3

Backends [?]
Select existing instance groups **Select existing instances**

test-edge-instance-group ×

No more instance groups available in this region

Backup pool [?] (Optional)
None

Failover ratio [?]
10 %

Health check [?]
wcs-lb-health-check (HTTP)
port: 8081, timeout: 5s, check interval: 10s, unhealthy threshold: 3 attempts

Session affinity [?]
Client IP and protocol

4. Выберите **Create health check**. Создайте настройку проверки сервера, укажите порт **8081** и запрос **/**

Create a health check

Autohealing instance groups and load balancing use health checks to detect when an instance is unresponsive [Learn more](#)

Name [?]
Name is permanent
wcs-lb-health-check

Description (Optional)
WCS health check for load balancer

Protocol
HTTP

Port [?]
8081

Request path [?]
/

More

Health criteria

Define how health is determined: how often to check, how long to wait for a response, and how many successful or failed attempts are decisive

Check interval ?	Timeout ?
10 seconds	5 seconds
Healthy threshold ?	Unhealthy threshold ?
2 consecutive successes	3 consecutive failures

Save and continue Cancel

5. В разделе **Frontend configuration** создайте конфигурации для TCP портов **8081**, **8080**, **8443**, **8444** для HTTP(S) и WS(S). Укажите статический внешний IP адрес для балансировщика

[←](#) New TCP load balancer

Name ⓘ
Name is permanent

Backend configuration
Your backend is configured

Frontend configuration
Your frontend is configured →

ⓘ **Review and finalize**
Optional

Frontend configuration

Specify an IP address, port and protocol. This IP address is the frontend IP for your clients requests.

New Frontend IP and port 🗑️ ⬆️

Name (Optional) ⓘ
Name is permanent

[Add a description](#)

Protocol
TCP

Network Service Tier ⓘ
 Premium (Current project-level tier, [change](#)) ⓘ
 Standard ⓘ

IP
test-lb-entry-point (34.107.5.128) ▾

Port

[+ Add Frontend IP and port](#)

[←](#) New TCP load balancer

Name ⓘ
Name is permanent

Backend configuration
Your backend is configured

Frontend configuration
Your frontend is configured →

ⓘ **Review and finalize**
Optional

Frontend configuration

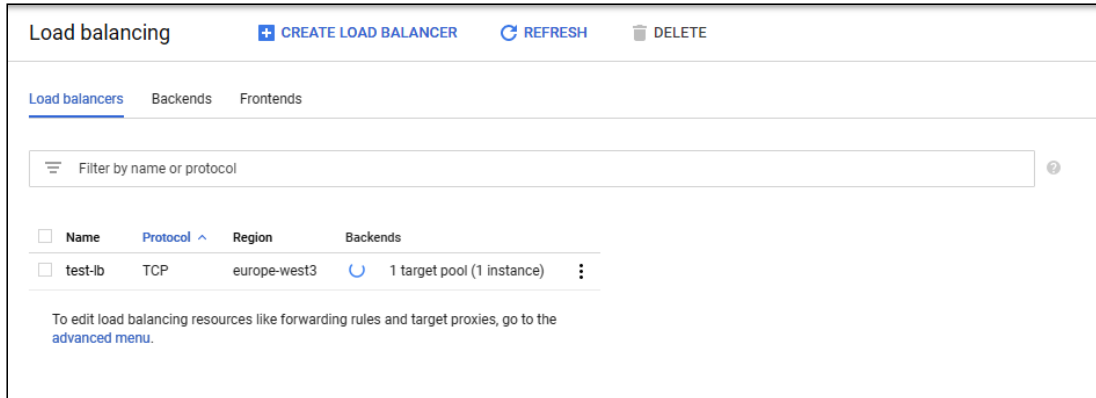
Specify an IP address, port and protocol. This IP address is the frontend IP for your clients requests.

Protocol:TCP, IP:34.107.5.128, Port:8081	<i>Not saved</i> ✎
Protocol:TCP, IP:34.107.5.128, Port:8080	<i>Not saved</i> ✎
Protocol:TCP, IP:34.107.5.128, Port:8444	<i>Not saved</i> ✎
Protocol:TCP, IP:34.107.5.128, Port:8443	<i>Not saved</i> ✎

[+ Add Frontend IP and port](#)

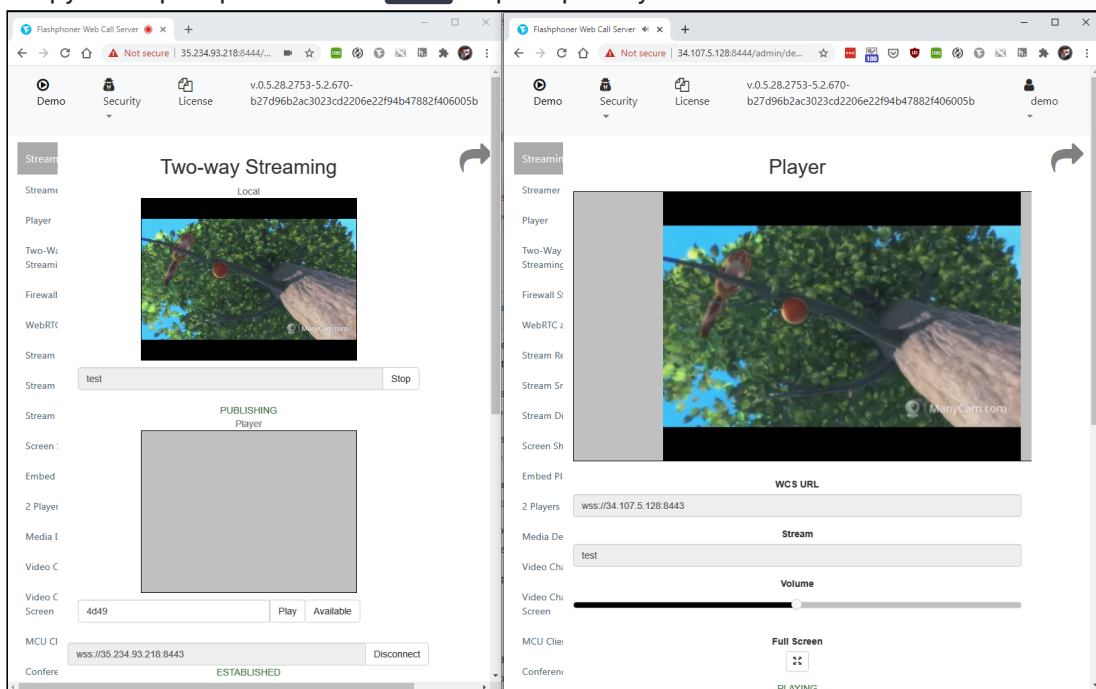
Вы можете добавить другие необходимые порты (**1935** для RTMP подписчиков, **8082**, **8445** для HLS и т.д в зависимости от сценария использования Edge серверов)

6. Нажмите **Create**. Балансировщик нагрузки запустится



Тестирование балансировщика нагрузки

1. Войдите в веб интерфейс Origin сервера, опубликуйте поток **test** в примере Two Way Streaming
2. Войдите в веб-интерфейс Edge сервера, используя IP адрес балансировщика нагрузки. Проиграйте поток **test** в примере Player



Изменение настроек Edge серверов

Для того, чтобы изменить настройки Edge серверов в группе масштабирования, например, обновить SSL сертификаты, необходимо обновить образ диска Edge сервера следующим образом:

1. Отключите масштабирование и удалите все экземпляры Edge серверов в группе
2. Запустите исходных экземпляр Edge сервера
3. Внесите необходимые изменения в настройки (например, обновите SSL сертификаты)
4. Остановите исходный экземпляр Edge сервера
5. Удалите образ диска Edge сервера
6. Создайте новый образ диска Edge сервера с тем же именем образа (например `test-edge-image-1`)
7. Включите масштабирование в группе (настройки масштабирования при этом сохраняются)