Настройка балансировки нагрузки с масштабированием в GCP

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. Зарезервируйтестатически й 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 серверраи 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

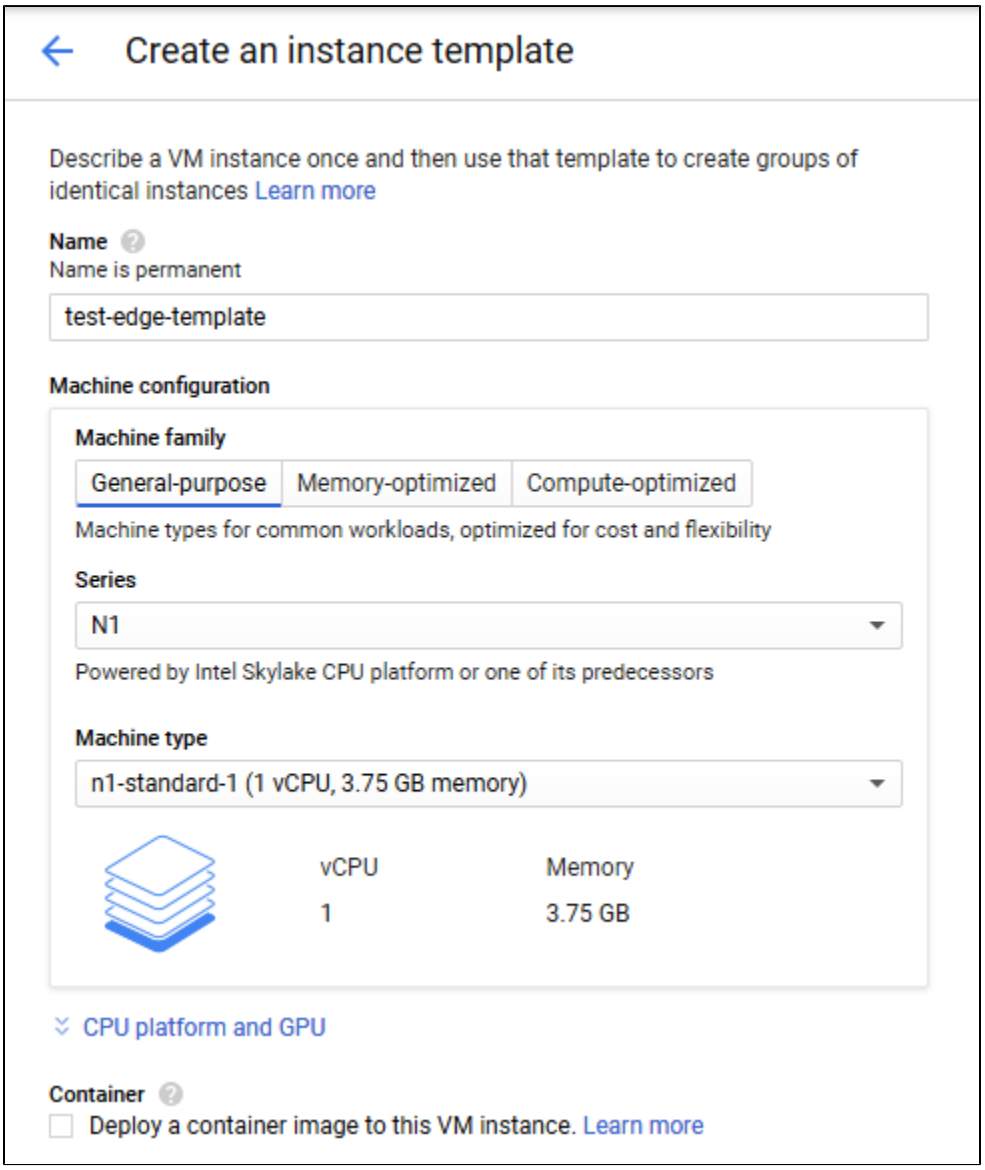
Create Cancel

Equivalent REST or command line

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

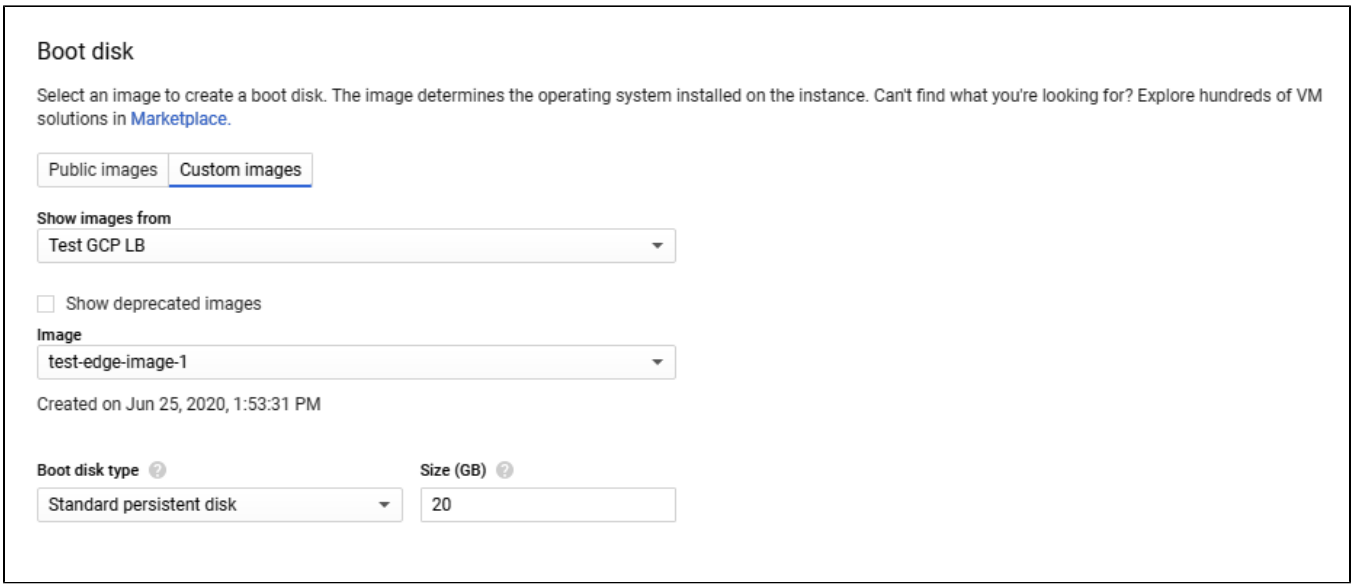
Создание шаблона Edge сервера

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

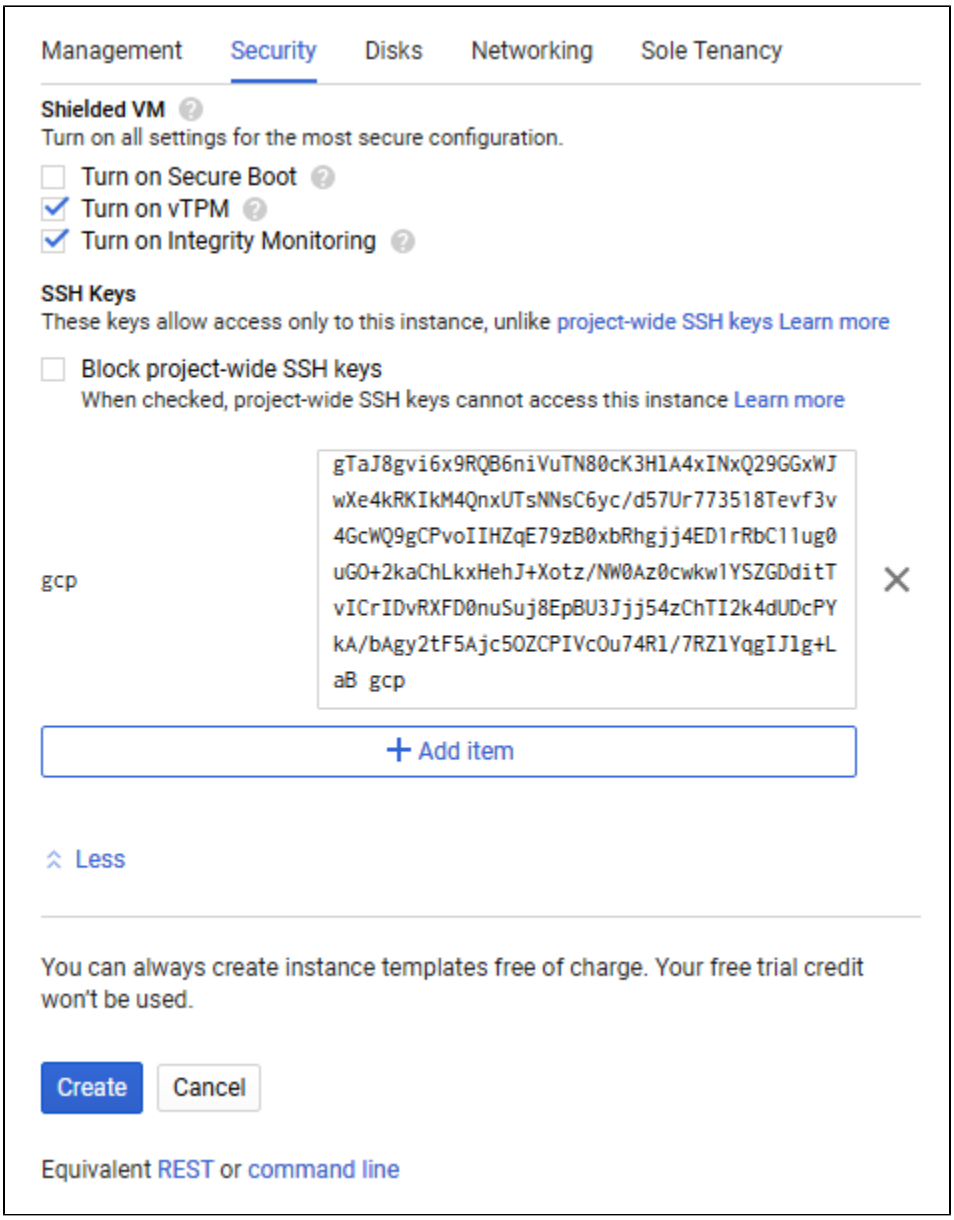


2.В разделе "Boot disk" нажмите "Change"

На вкладке "Custom images" выберите образ диска Edge сервера

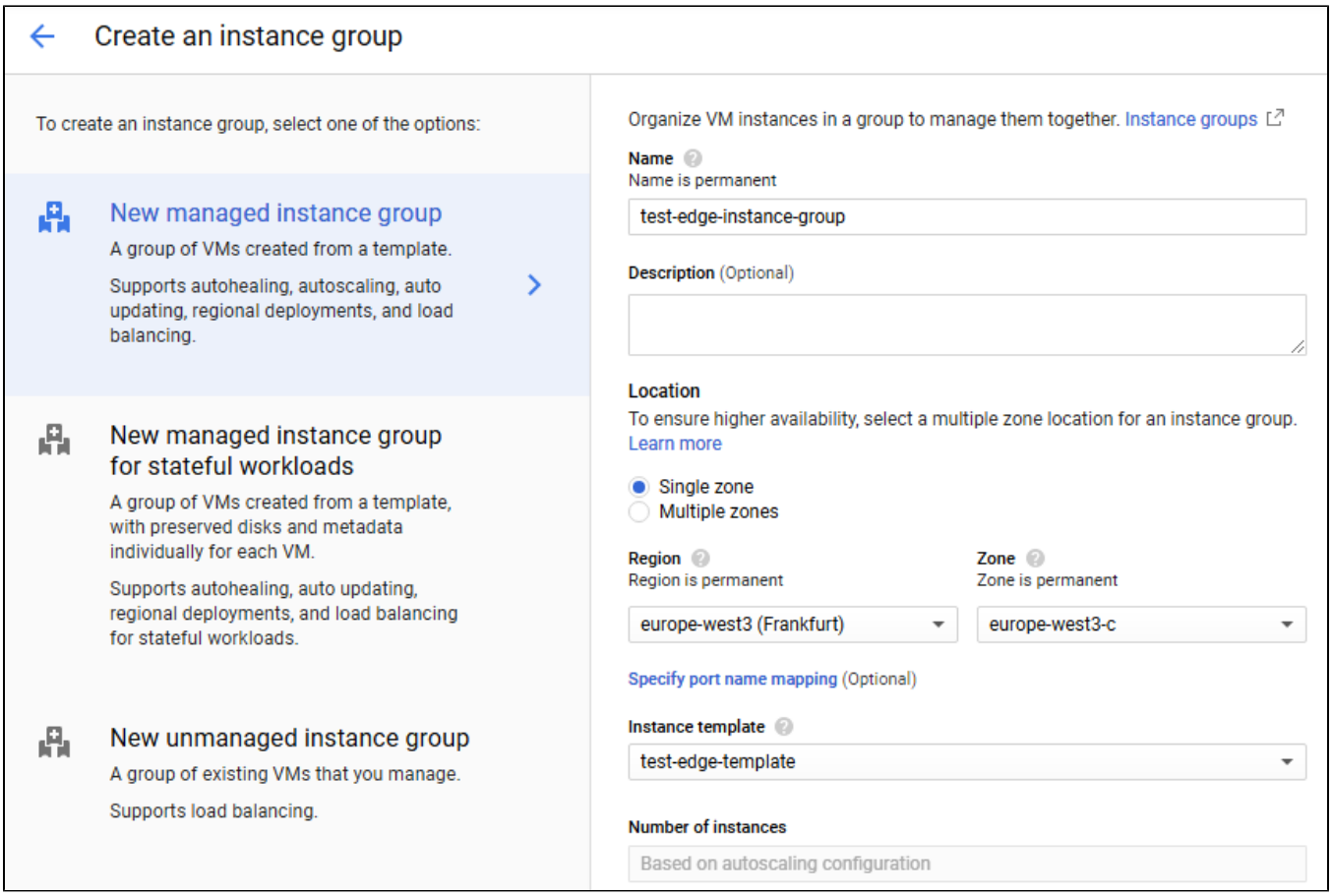


3.На вкладке "Security" добавьте публичный ключ для доступа к серверу по SSH, если у Вас нет ключей, привязанных к проекту, и нажмите "Create"



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



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

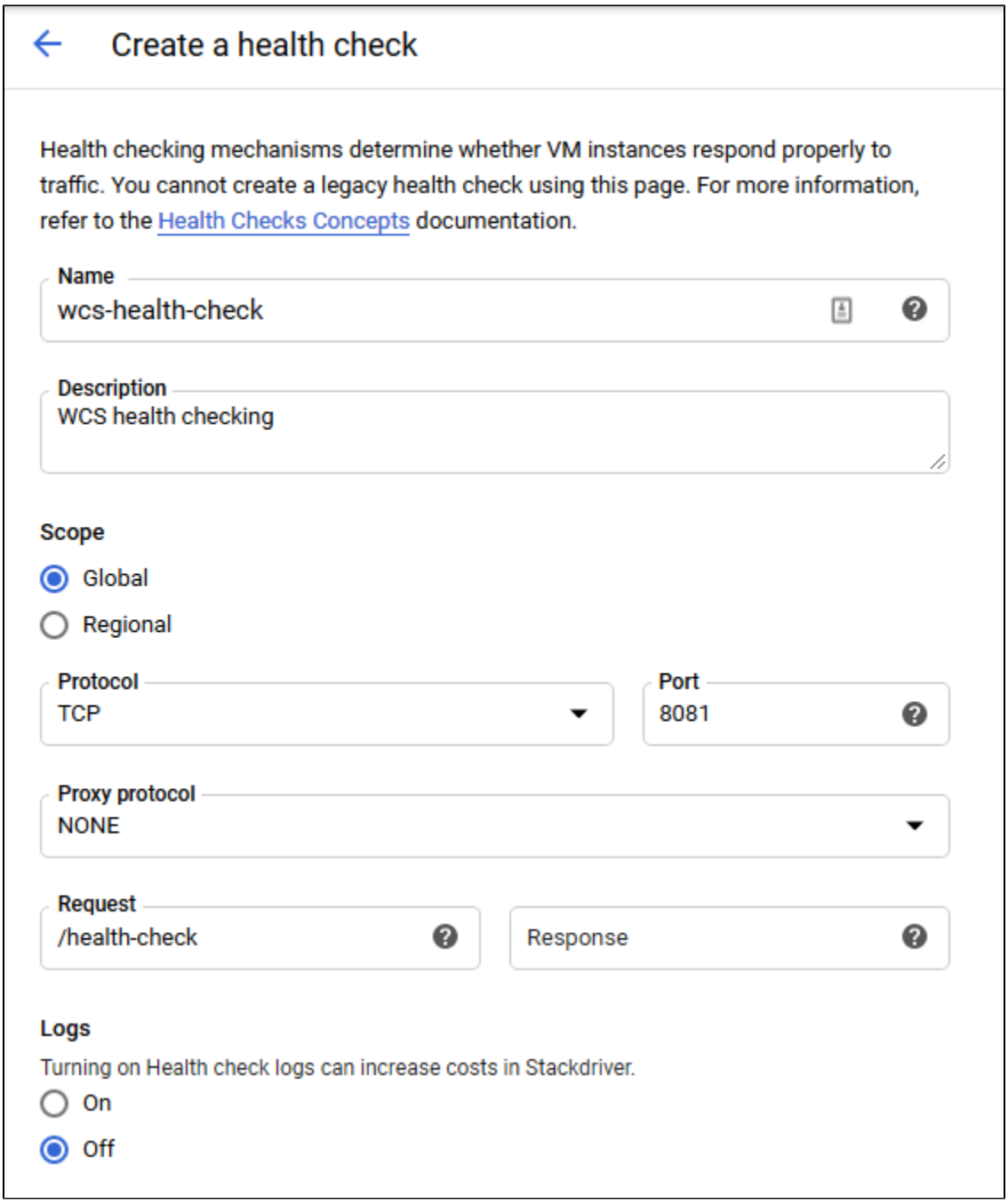
Scale In Controls

Prevent a sudden drop in the number of running VM instances in the group by controlling the process of scaling in. Learn more

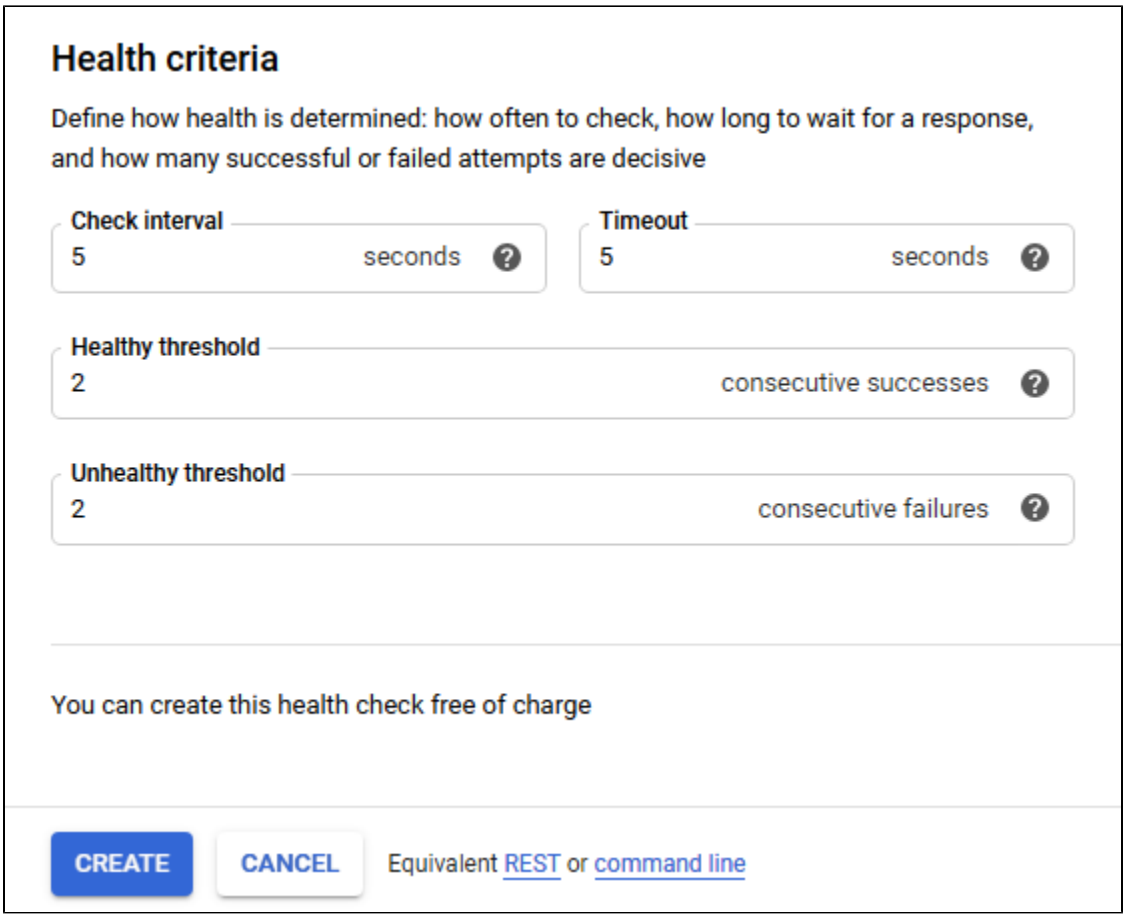
Enable Scale In Controls

Delete autoscaling configuration

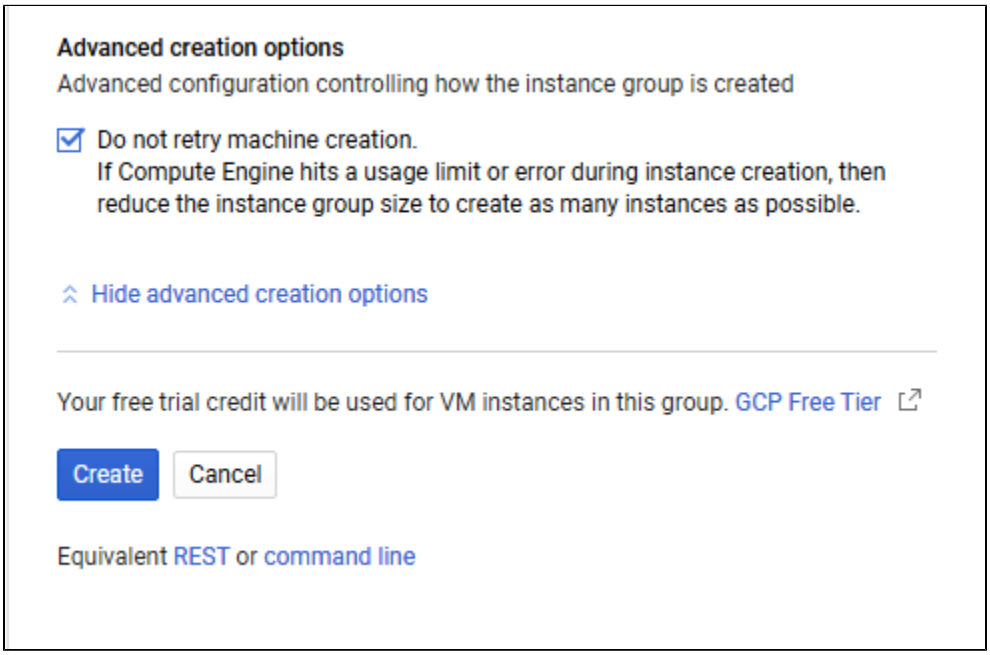
3. Включите проверку состояния ВМ (Auto healing) и создайте настройку проверки сервера. Укажите протокол TCP, порт 8081 и запрос /health-check



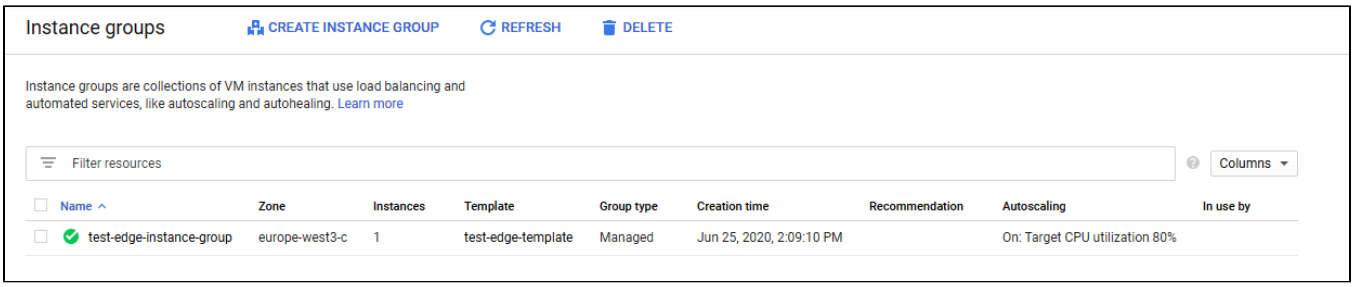
Настройте критерии проверки и нажмите "Create" для возврата к редактированию группы



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

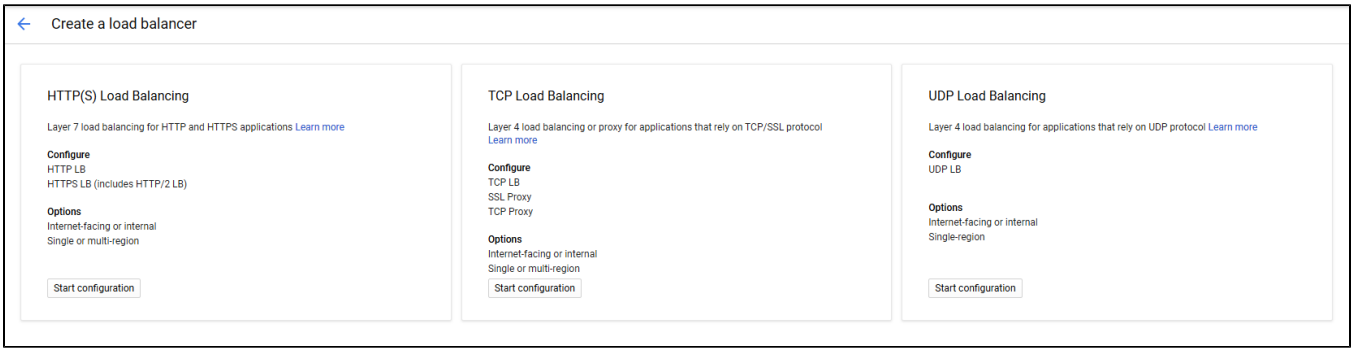


Группа масштабирования будет создана, и один экземпляр будет запущен

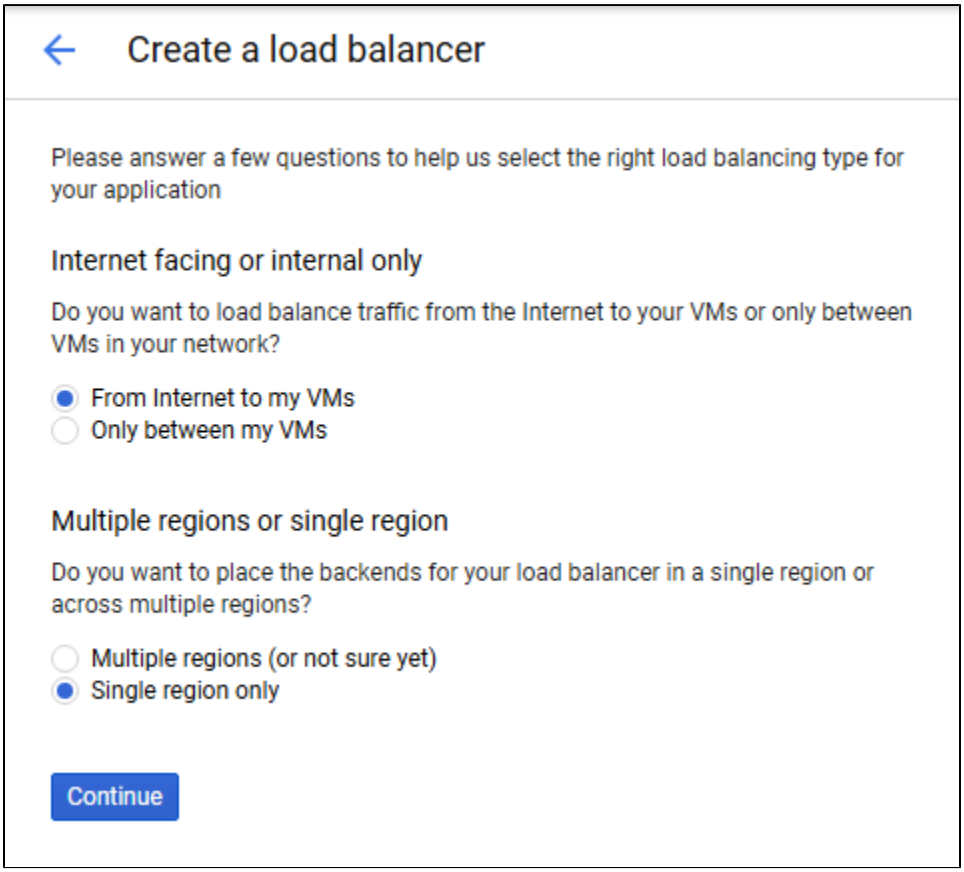


Создание балансировщика нагрузки

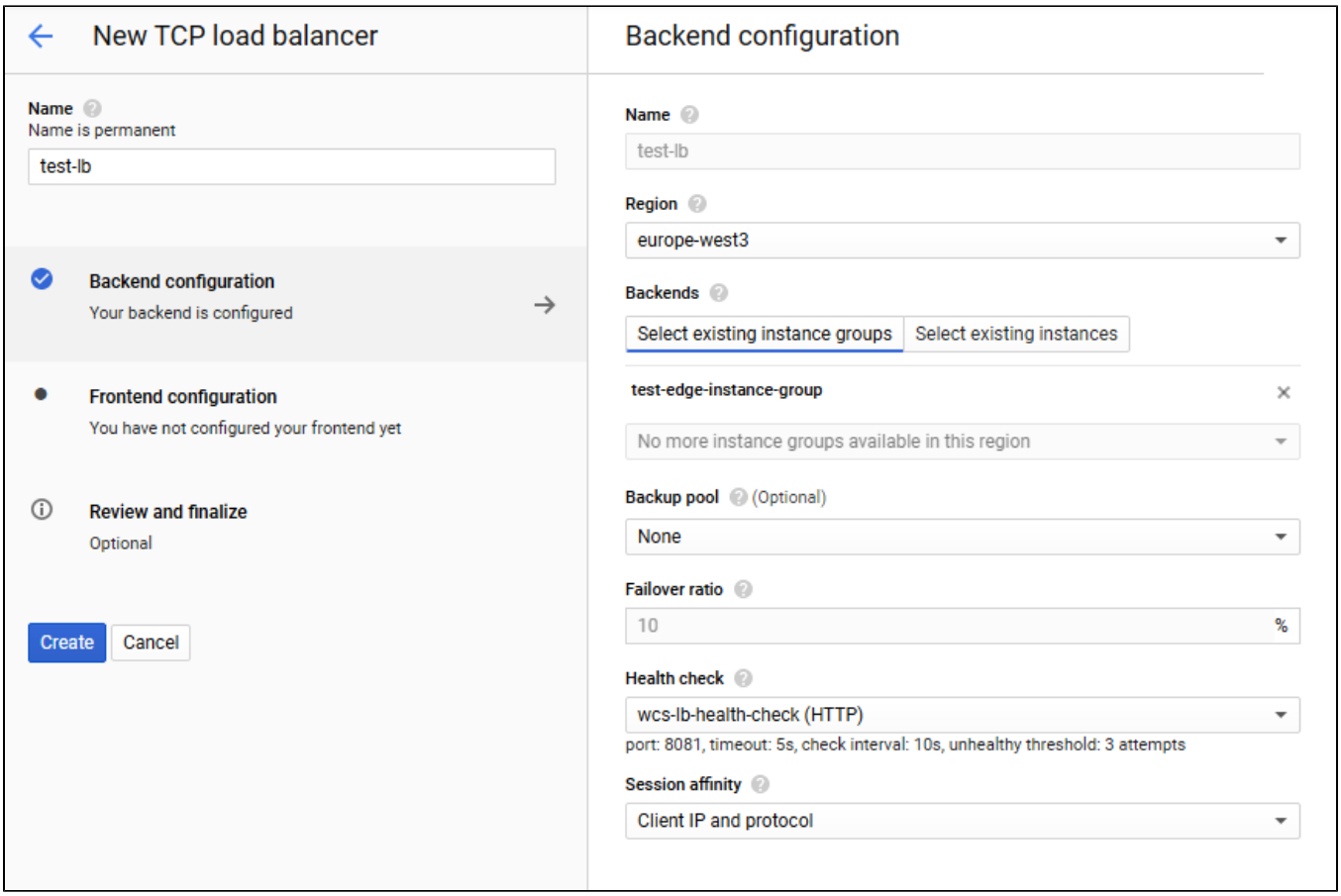
1. Перейдите в раздел "Network – Load balancers" и нажмите "Create load balancer". Выберите "TCP Load Balancing"



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



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



4. Выберите "Create health check". Создайтенастройку 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

10 seconds

Timeout

5 seconds

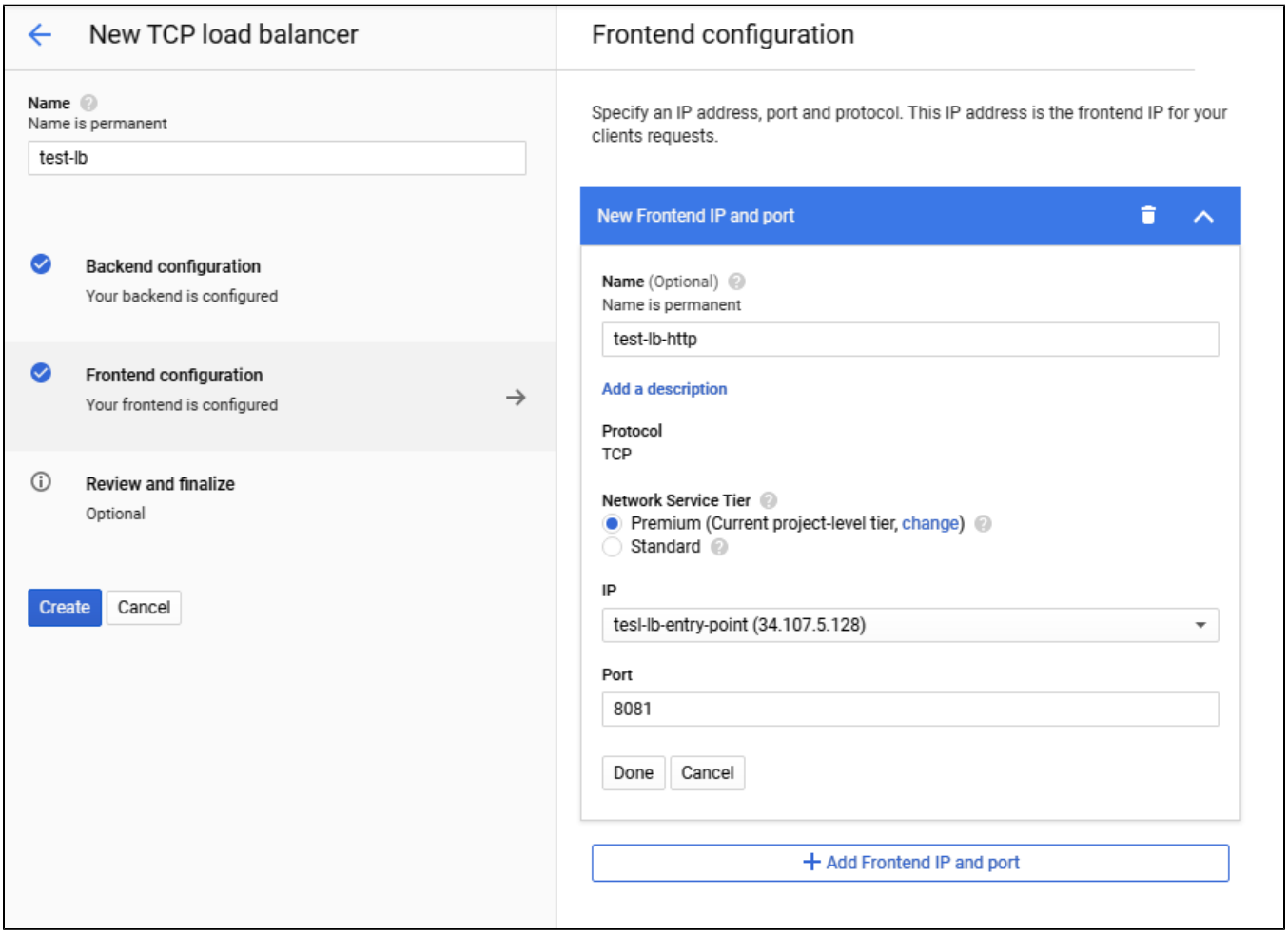
Healthy threshold

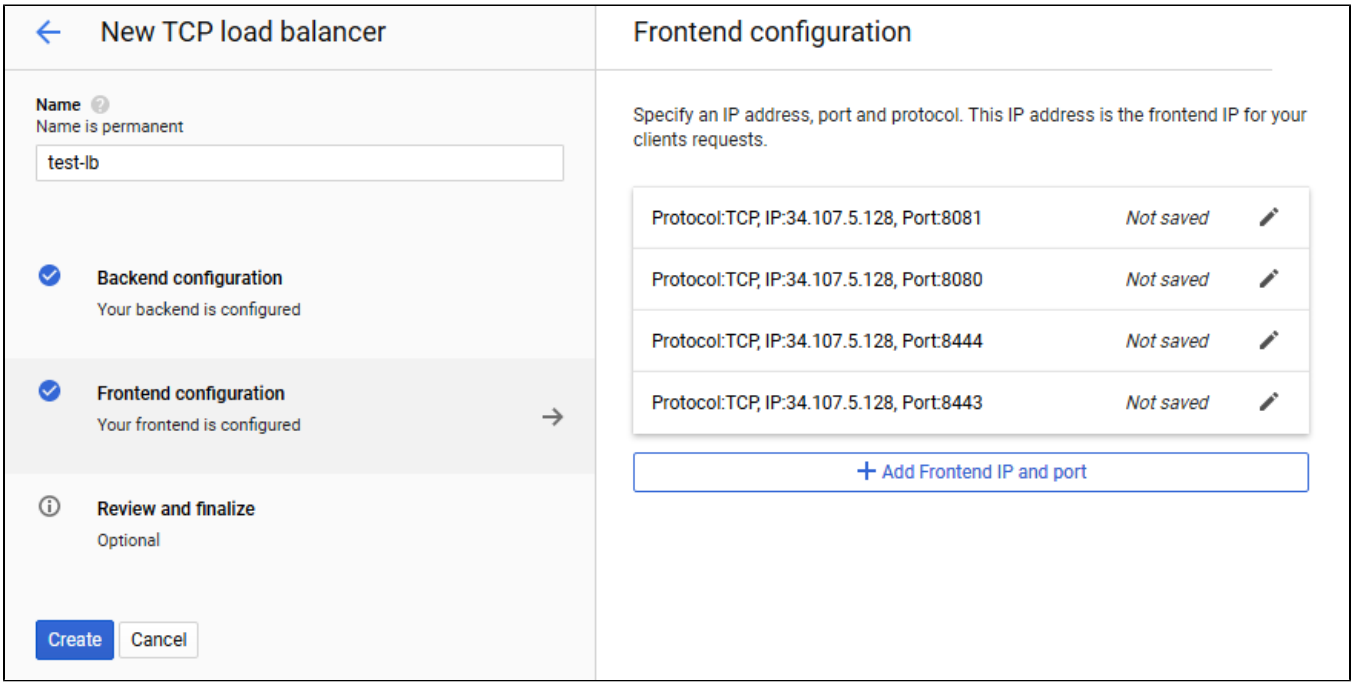
2 consecutive successes

Unhealthy threshold

3 consecutive failures

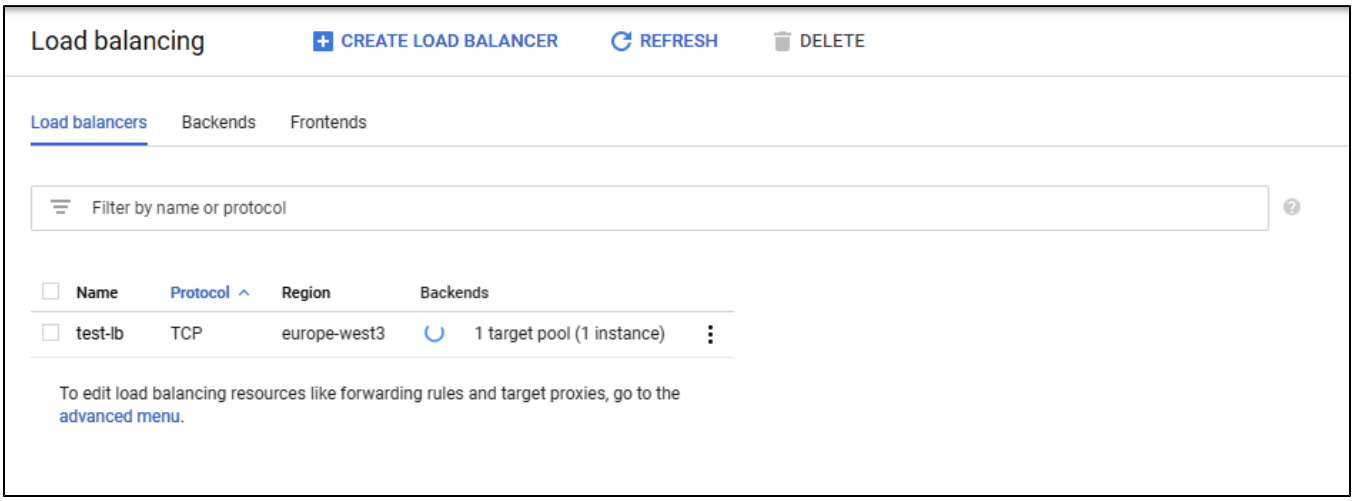
5. В разделе "Frontend configuration" создайте конфигурации для TCP портов 8081, 8080, 8443, 8444 для HTTP(S) и WS(S). Укажите статический внешний IP адрес для балансировщика





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

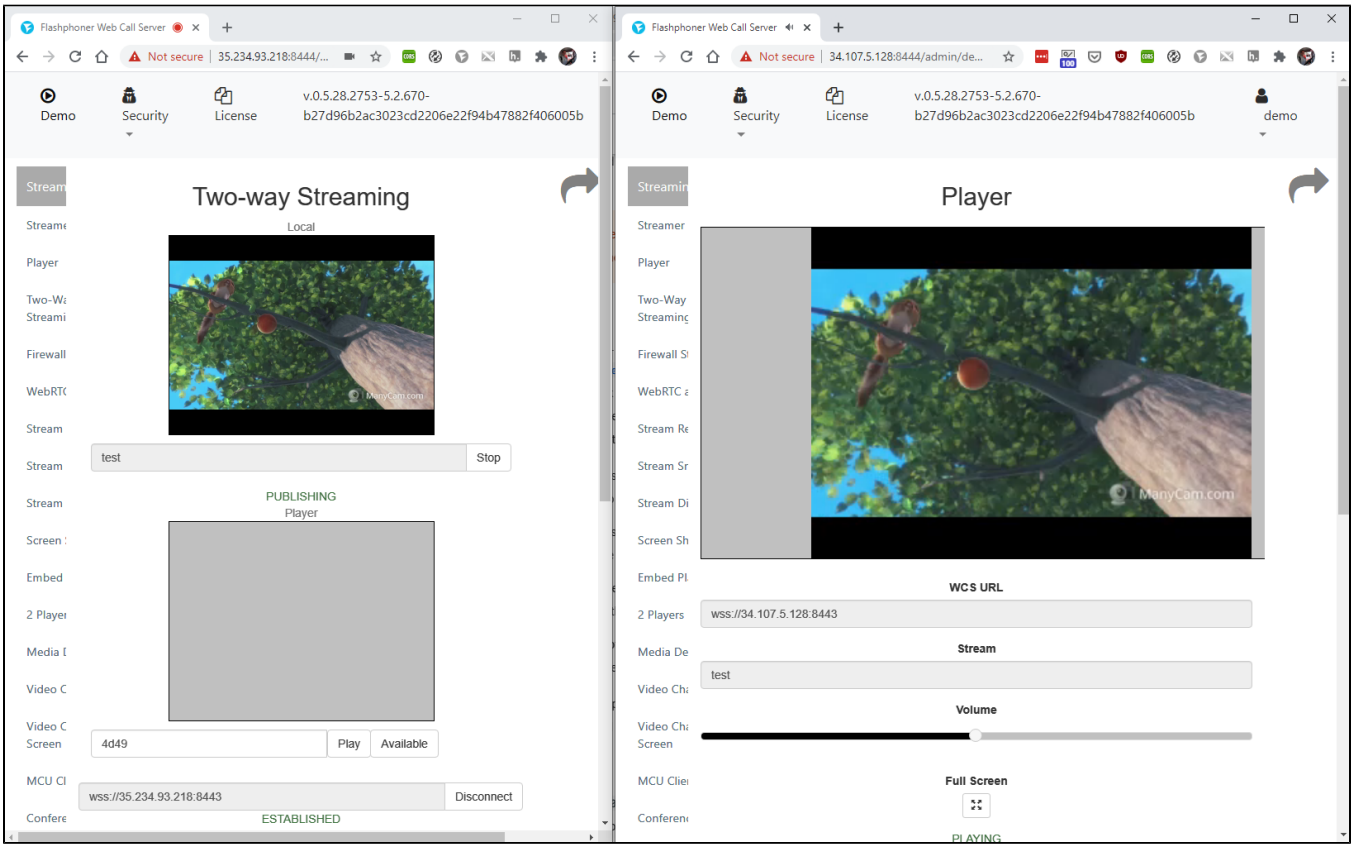
6. Нажмите "Create". Балансировщик нагрузки запустится



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 сохранятся)