Настройка балансировки нагрузки с масштабированием при помощи AWS ELB

Описание

Экземпляры WCS на Amazon поддерживают балансировку нагрузки при помощи AWS elastic load balancer (ELB).

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 потребуются следующие компоненты

- Образ, на основе которого будут создаваться новые экземпляры сервера при масштабировании
- Балансировщик нагрузки
- Конфигурация для запуска
- Группа масштабирования

Запуск классического балансировщика нагрузки на базе собственного образа с автоматическим масштабированием

Создание балансировщика нагрузки с автомасштабированием на базе собственного образа полезно для длительной эксплуатации (месяцы, годы). В этом случае базовый образ из AWS Marketplace обойдется дороже из-за почасовой оплаты, поэтому рекомендуется приобрести и активировать собственную лицензию WCS.

Attention

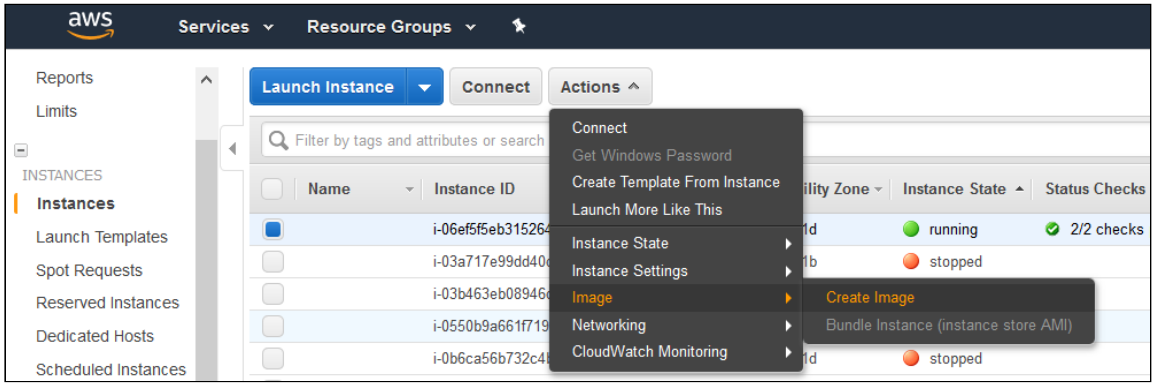
Классический балансировщик нагрузки перестанет поддерживаться в августе 2022 года

1. Создание образа

1.1. Создайте экземпляр из готового образа FlashphonerWebCallServer AMI и настройте WCS

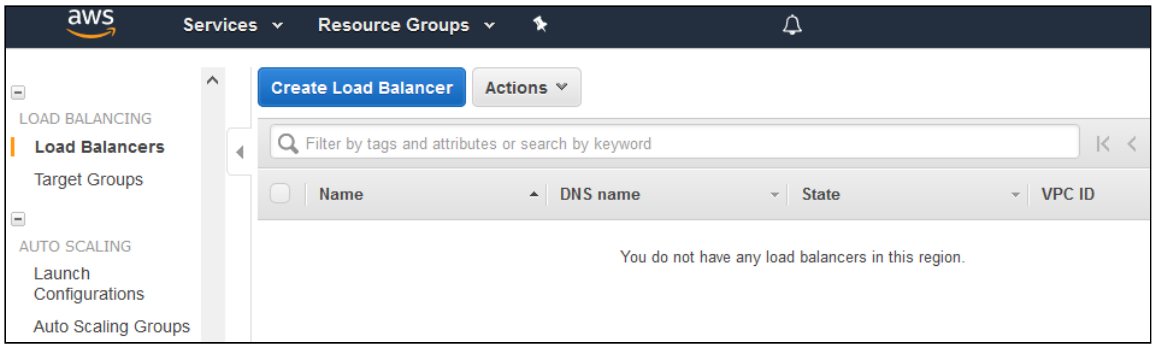
- активируйте лицензию
- импортируйте SSL сертификат
- настройте сервер

1.2. В консоли AWS, выберите созданный экземпляр, затем в меню `Actions` - `Image` - `Create Image` создайте новый образ:

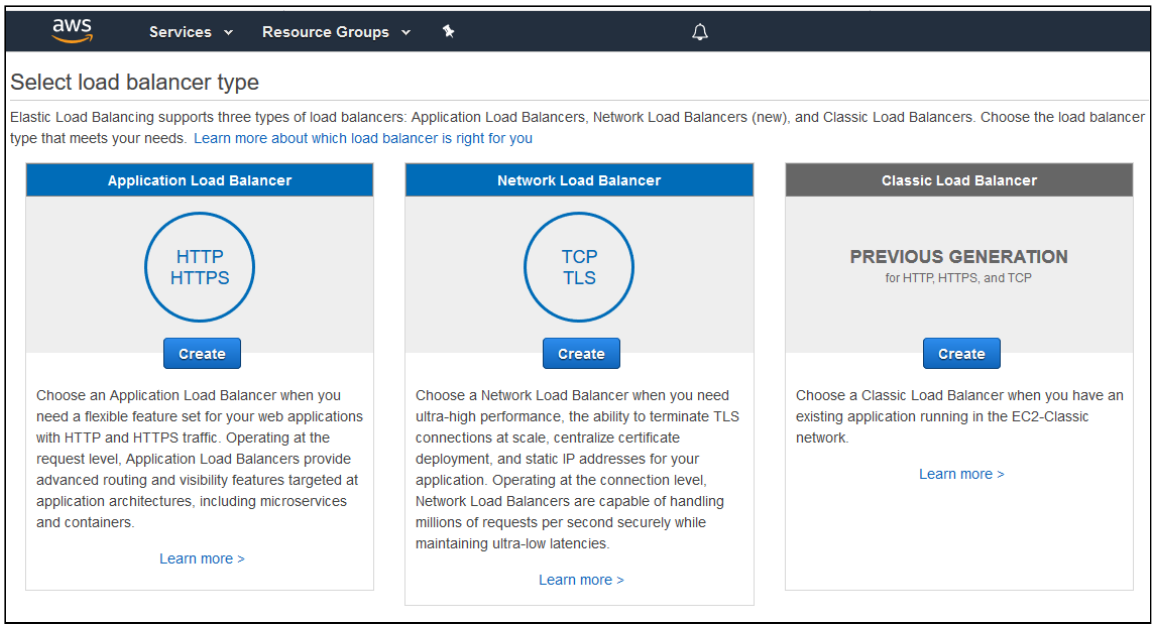


2. Создание балансировщика нагрузки

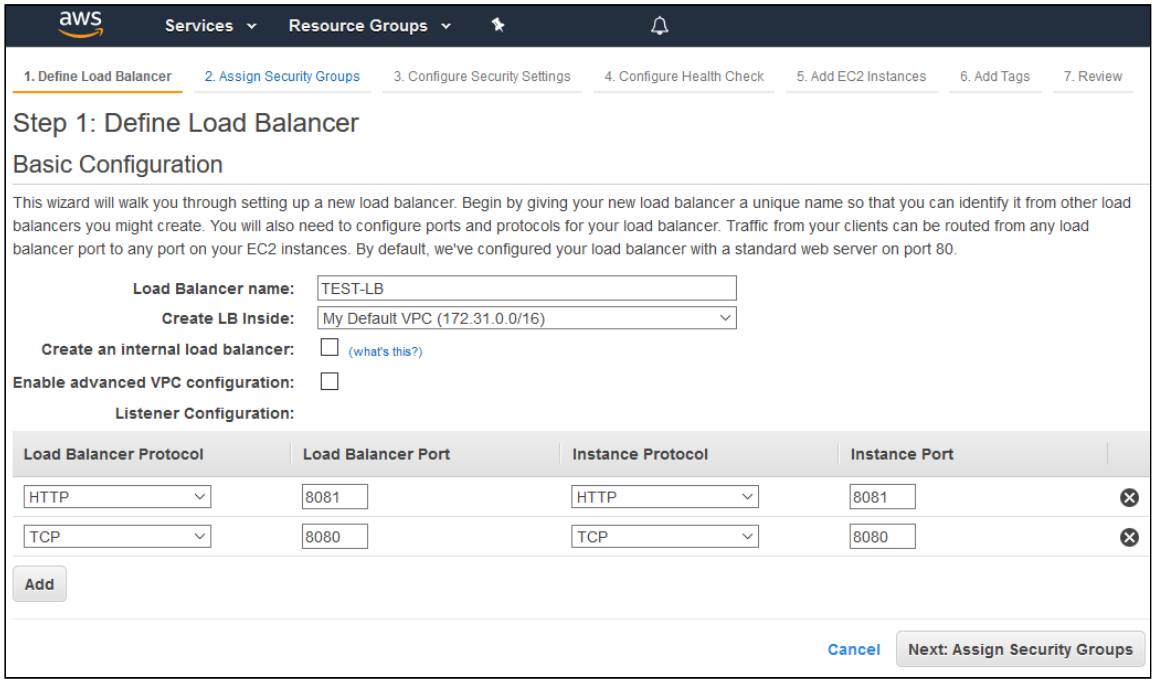
2.1. В консоли AWS, перейдите в пункт `EC2` - `Load Balancers` и нажмите `Create Load Balancer`



2.2 Выберите тип `Classic Load Balancer` (Данный тип позволяет выбрать порты для контроля состояния сервера)



2.3. На первом шаге настройки, добавьте требуемые протоколы. Например TCP, port 8080 для WebSocket соединений (`ws:host:8080`).

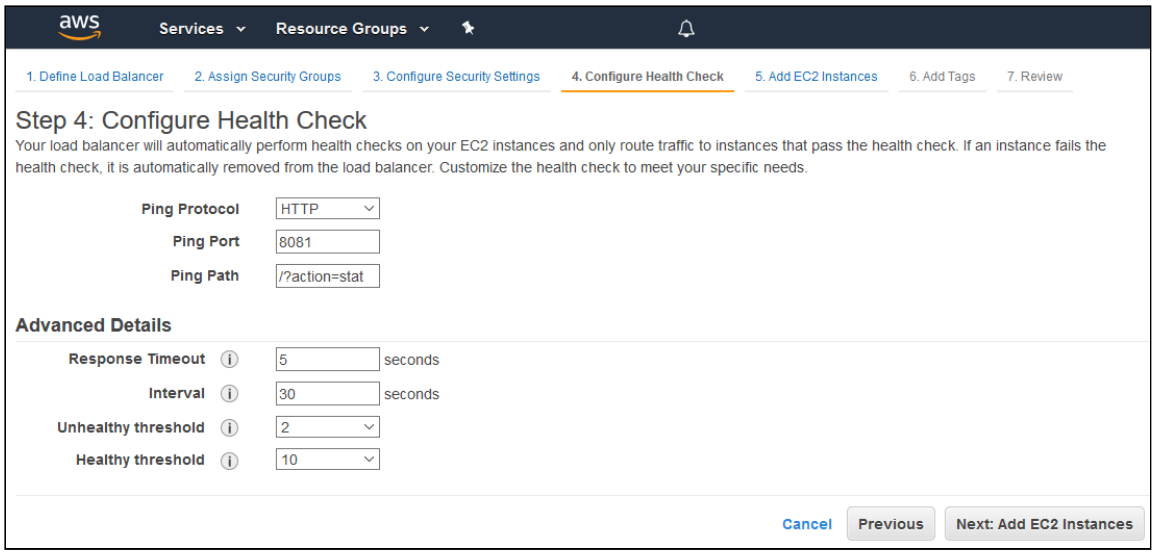


2.4. Определите группу безопасности.

2.5. Настройте проверку состояния сервера

Используйте следующие URL для контроля состояния

- по HTTP: `http://WCS_ADDRESS:8081/?action=stat`
- по HTTPS: `https://WCS_ADDRESS:8444/?action=stat`

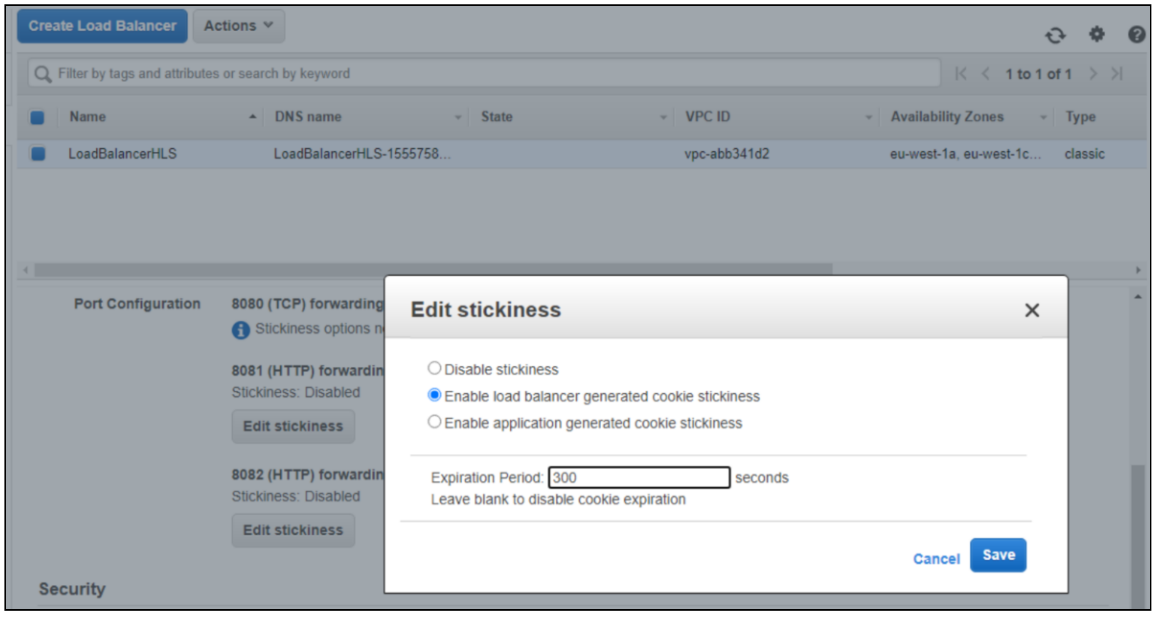


2.6. Добавьте существующие экземпляры сервера, по необходимости

По умолчанию, балансировщик будет распределять трафик между всеми доступными зонами в Вашем регионе.

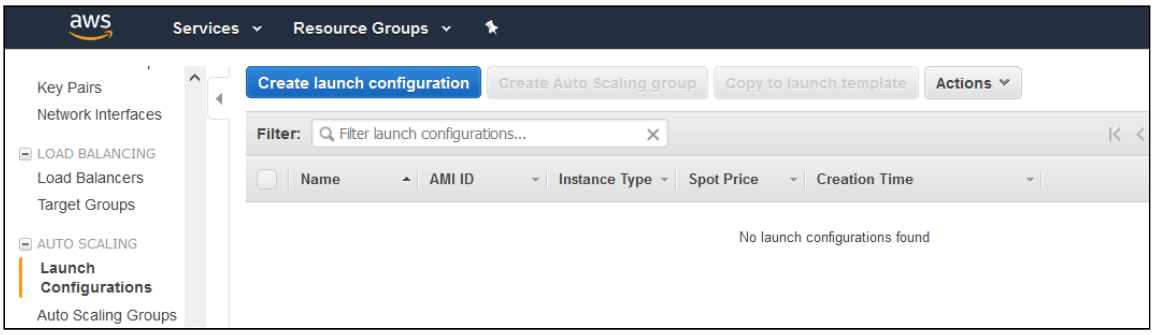
2.7. Завершите создание балансировщика нагрузки

2.8. Включите залипание (stickiness) для HTTP/HTTPS портов балансировщика

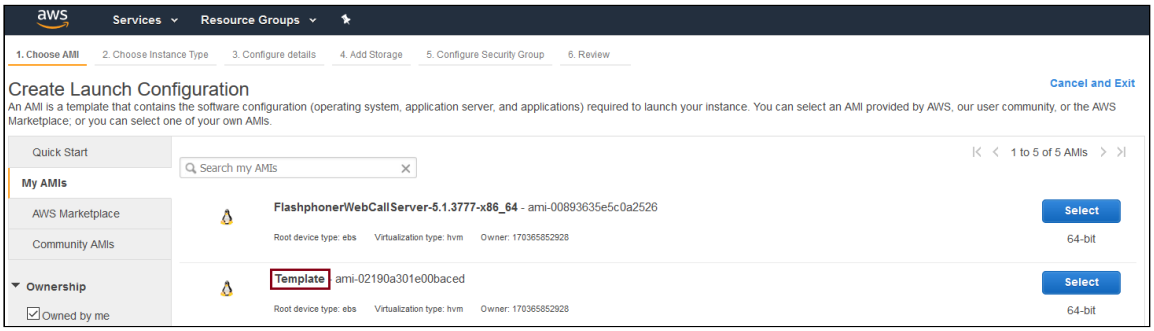


3. Создание новой конфигурации запуска

3.1. В консоли AWS, выберите пункт `EC2` - `Launch Configurations` и нажмите `Create launch configuration`

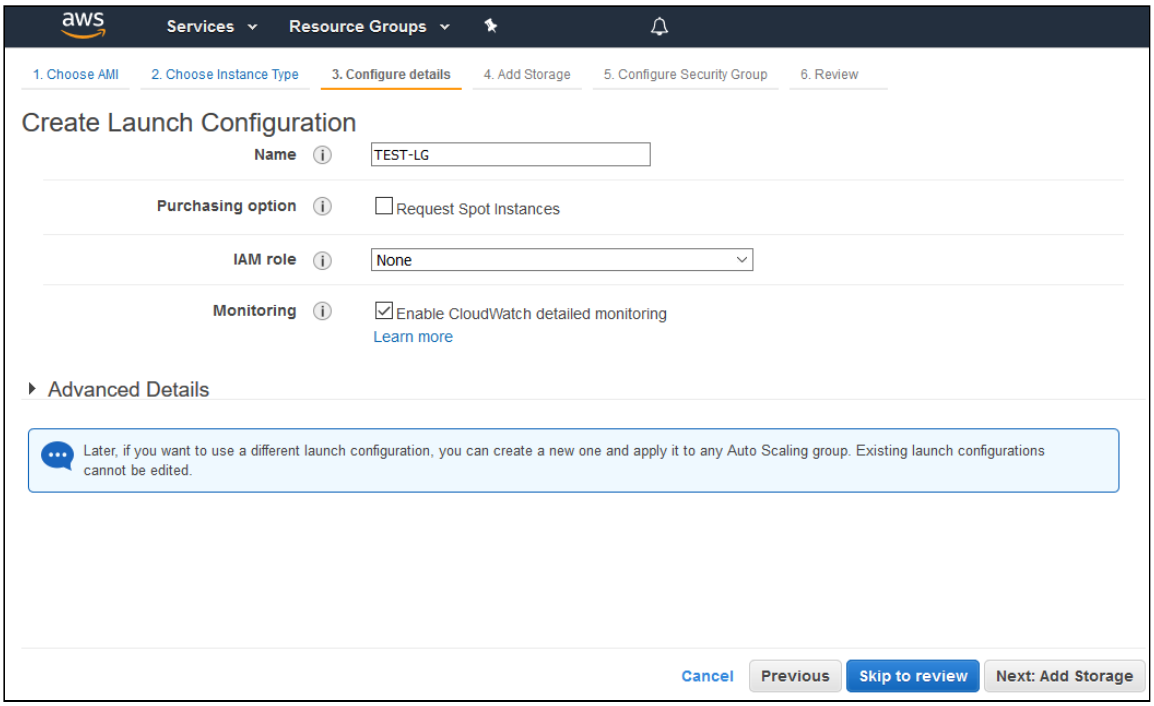


3.2. Выберите образ, созданный на шаге 1, с требуемой настройкой WCS



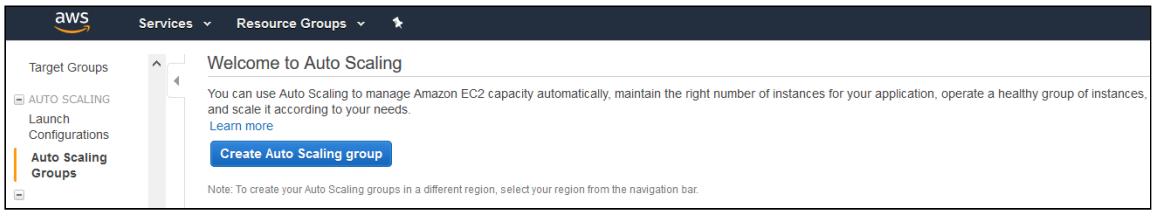
3.3. Завершите создание конфигурации запуска

В разделе `Advanced details` может быть настроен мониторинг с периодичностью до 1 минуты.

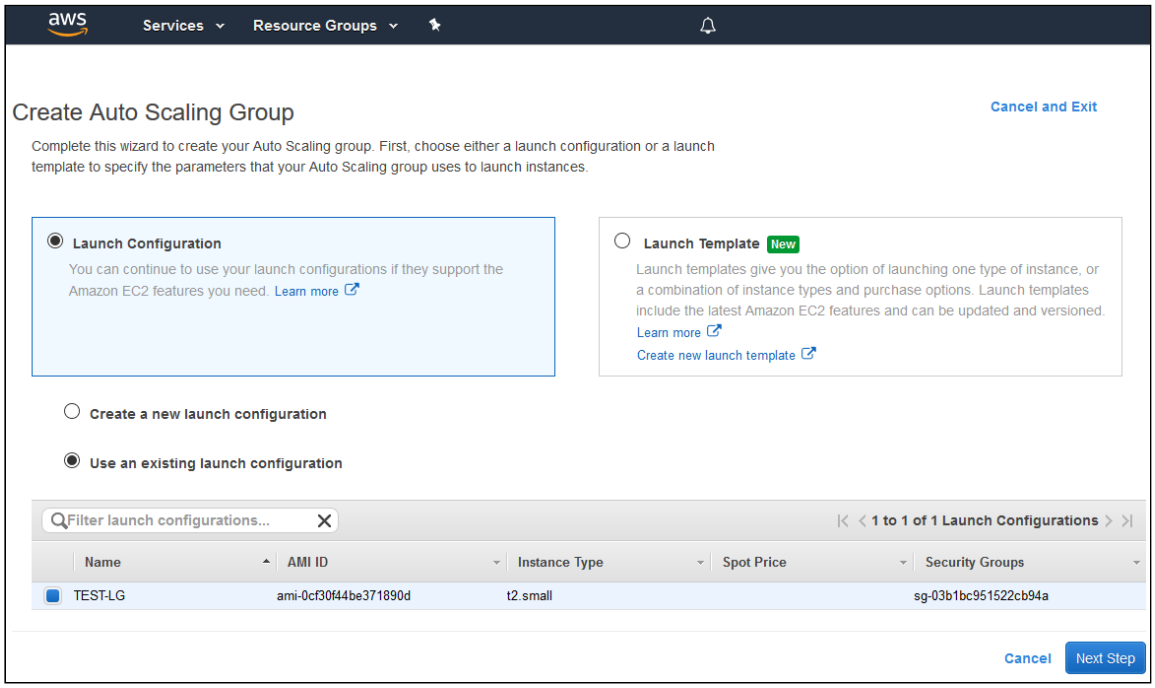


4. Создание новой группы масштабирования

4.1. В консоли AWS, выберите пункт `EC2` - `Auto Scaling Groups` и нажмите `Create Auto Scaling group`

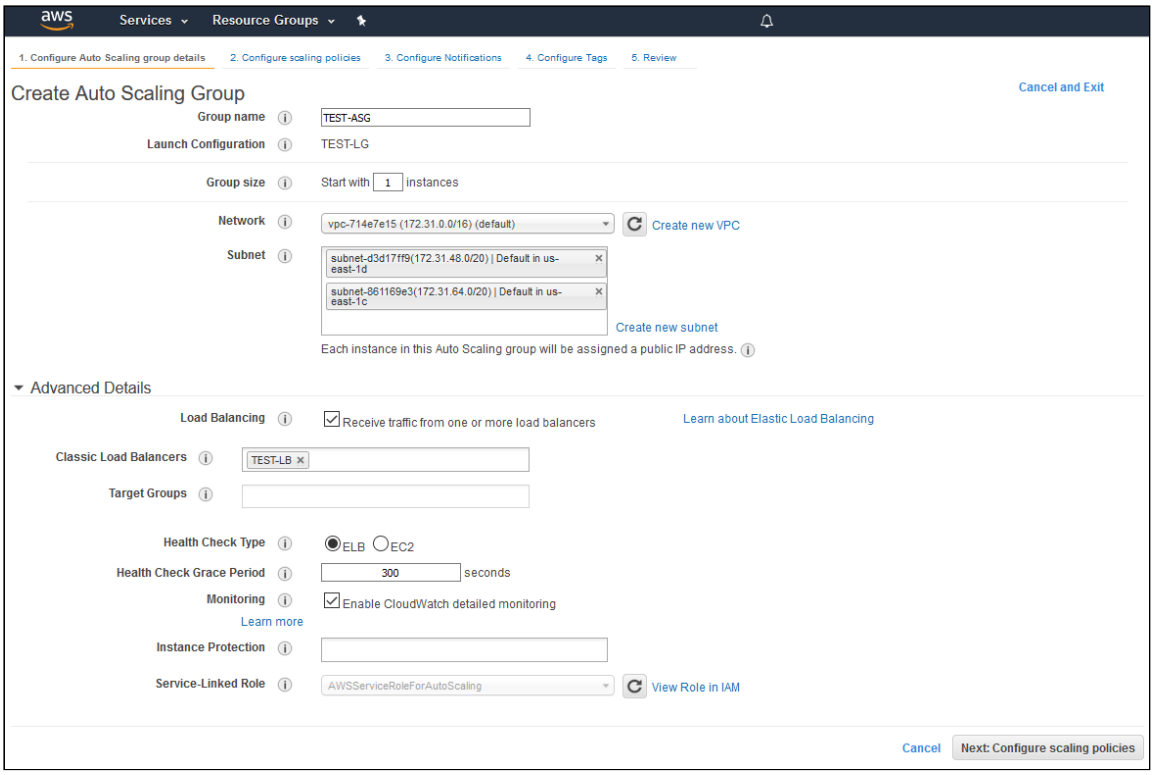


4.2. Выберите требуемую конфигурацию запуска, или выберите `Create a new launch configuration` для создания новой

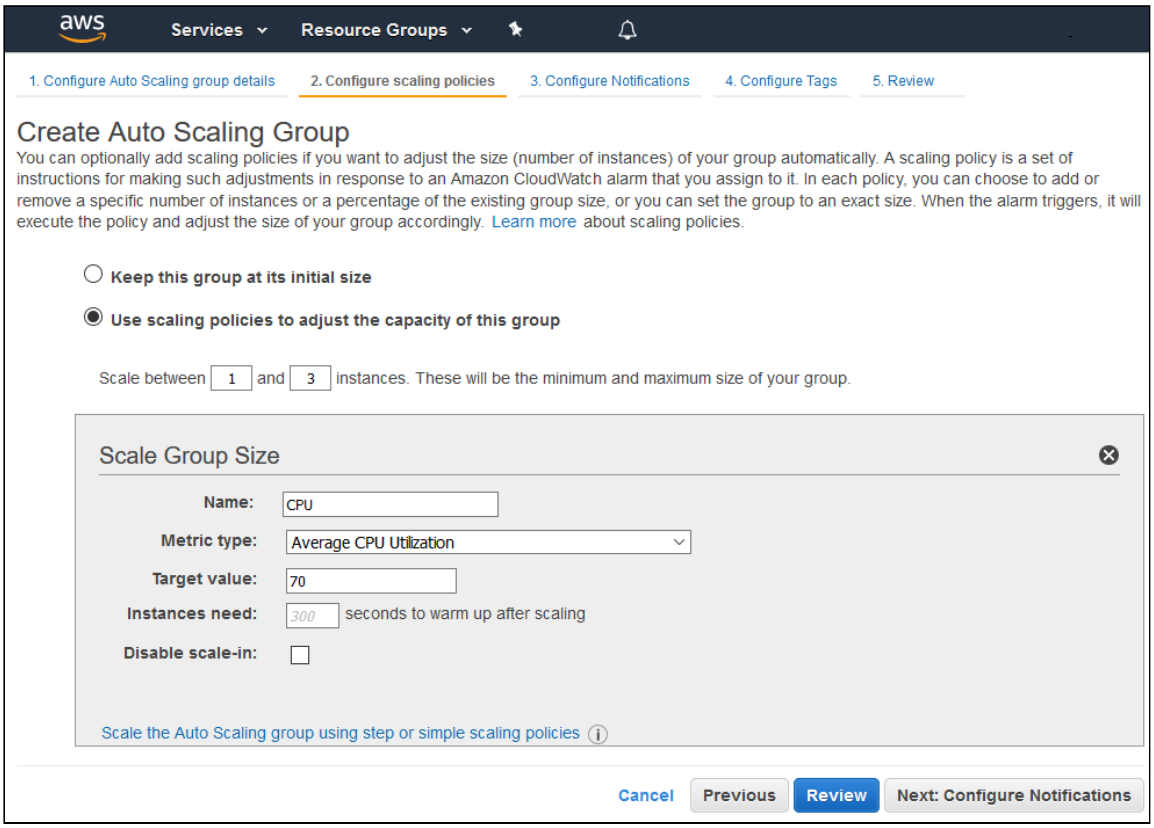


4.3. Настройте детали группы масштабирования

- добавьте требуемые подсети
- добавьте требуемый балансировщик нагрузки



4.4. Настройте политики масштабирования



4.5. Завершите создание группы масштабирования

Запуск балансировщика нагрузки для приложений с использованием существующих серверов

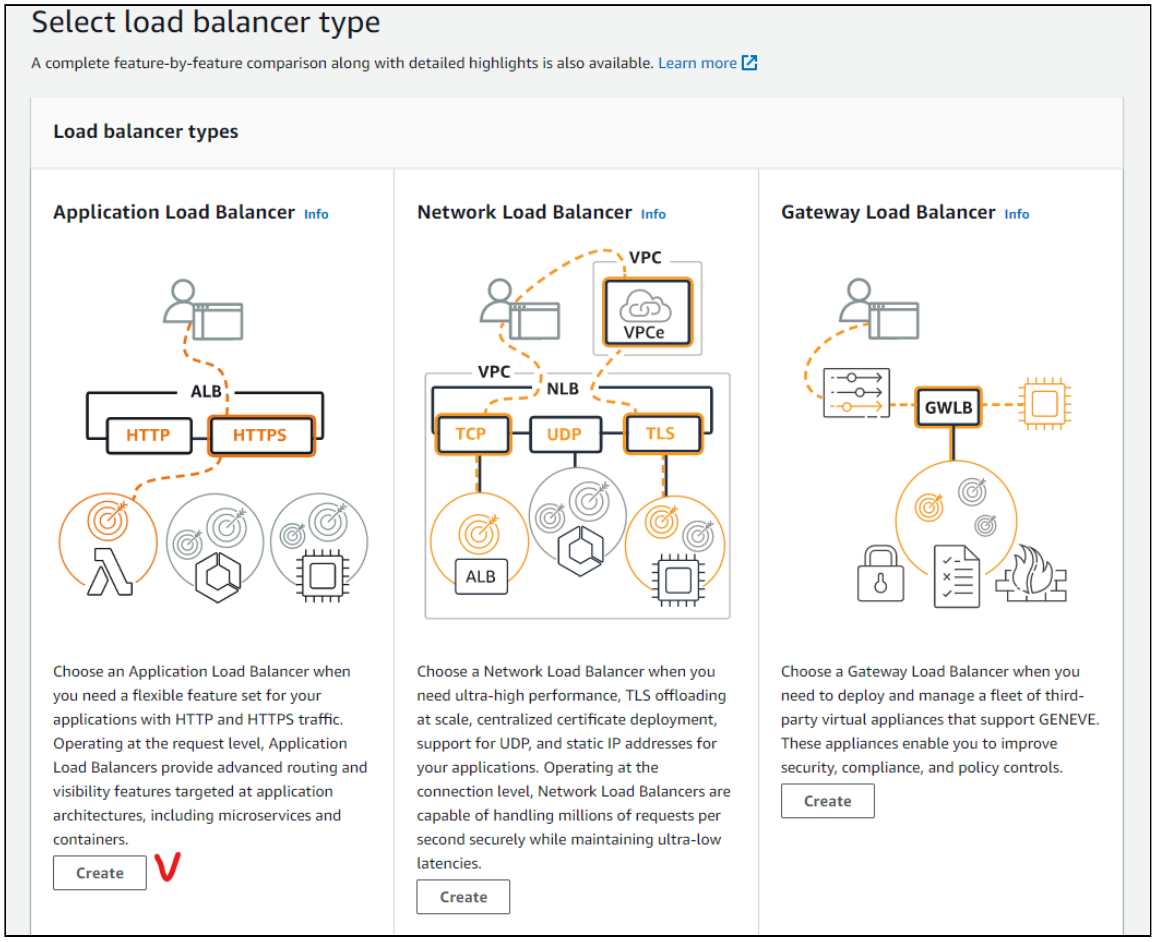
В некоторых случаях набор серверов уже запущен и настроен (например, группа Origin серверов в CDN), и требуется организовать балансировку нагрузки между этими серверами. Для решения этой задачи предназначен Application Load Balancer.

1. Запуск серверов

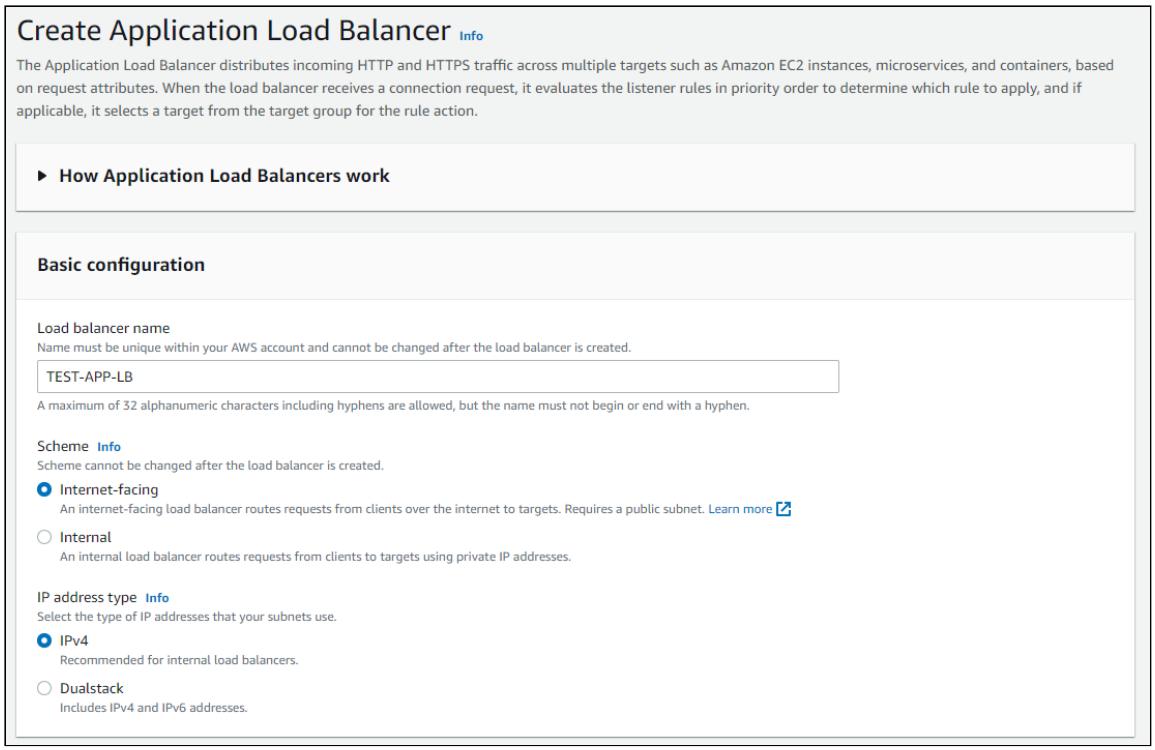
Запустите и настройте необходимое количество серверов, как описано здесь.

2. Создание балансировщика нагрузки

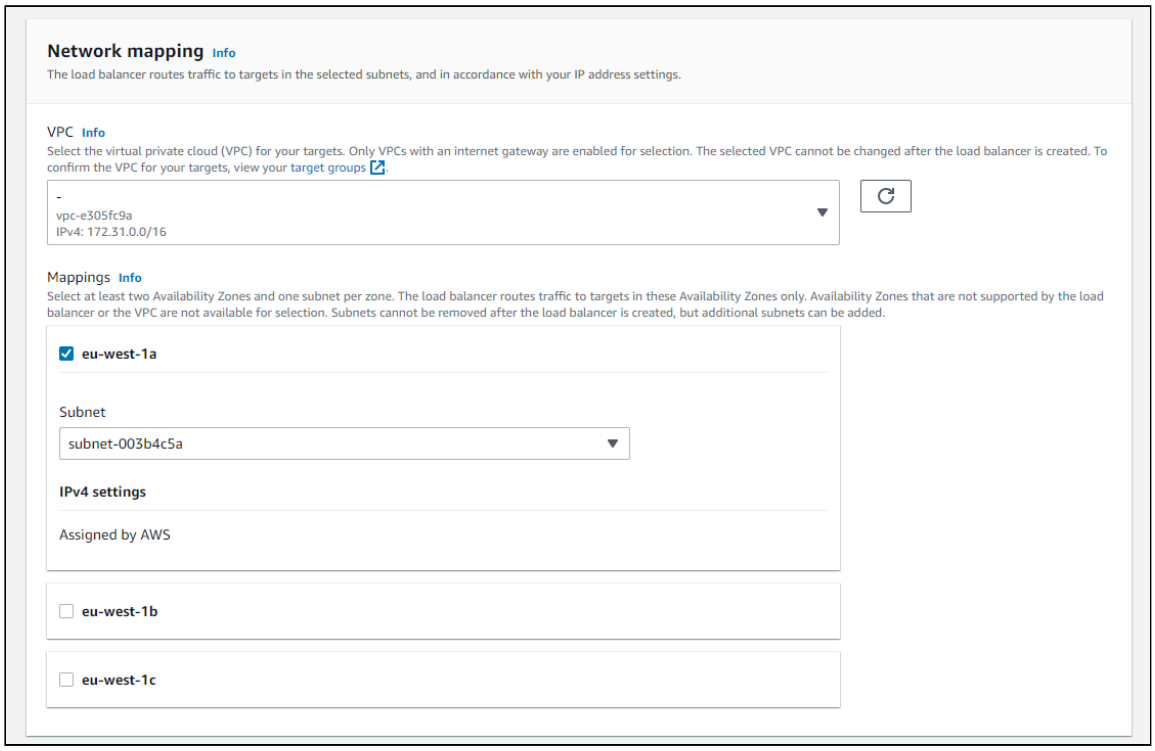
2.1. Перейдите в боковом меню EC2 Console в раздел `Load balancers` - `Load balancers` и нажмите `Create load balancer`. В окне мастера создания балансировщика нажмите `Create` для Application Load Balancer



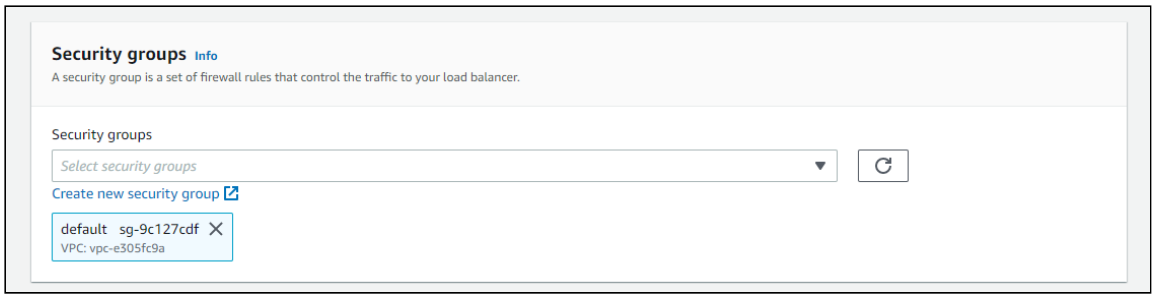
2.2. Введите имя балансировщика, выберите тип `Internet-facing` (предлагается по умолчанию)



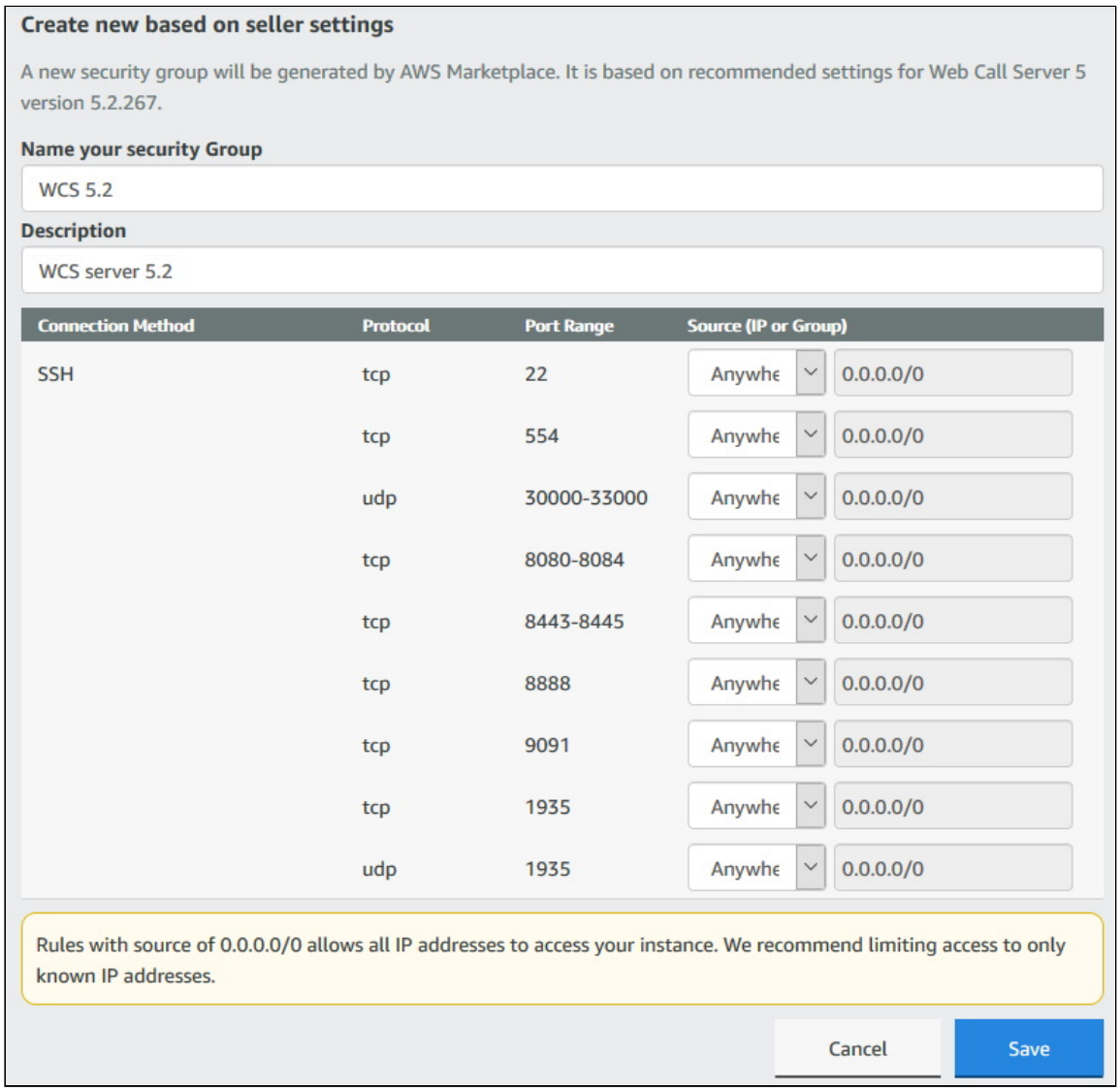
2.3. В разделе `Network mapping` выберите необходимые подсети



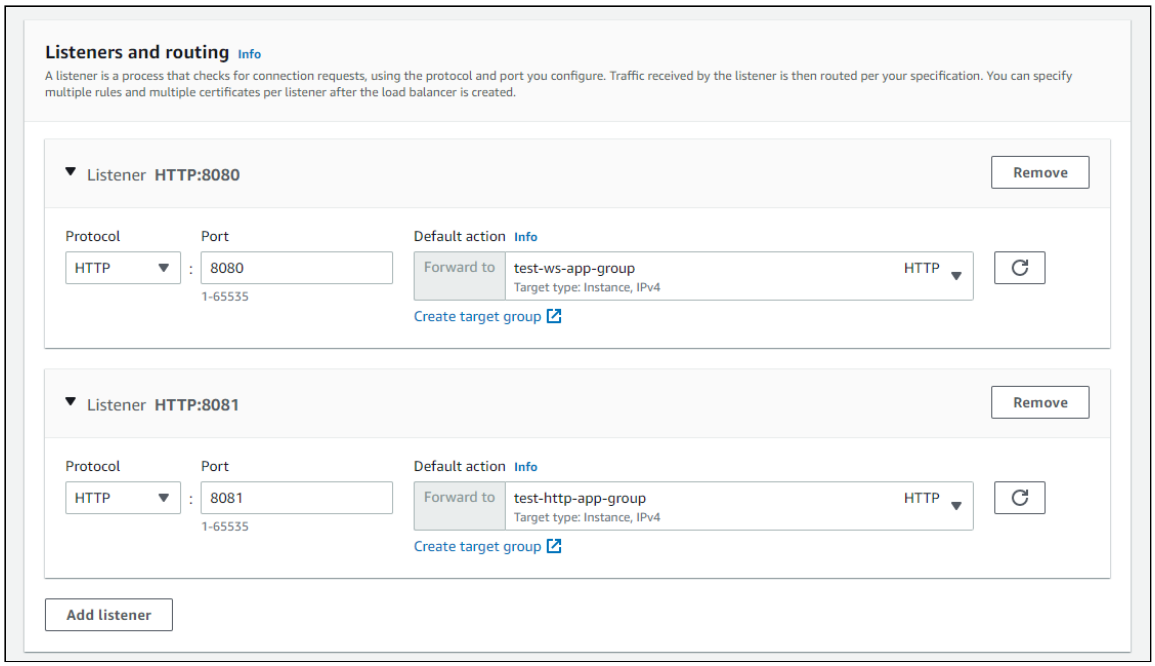
2.4. Выберите или создайте необходимые security groups



Обратите внимание, что в security group для балансировщика должен быть открыт стандартный набор портов WCS (пример)

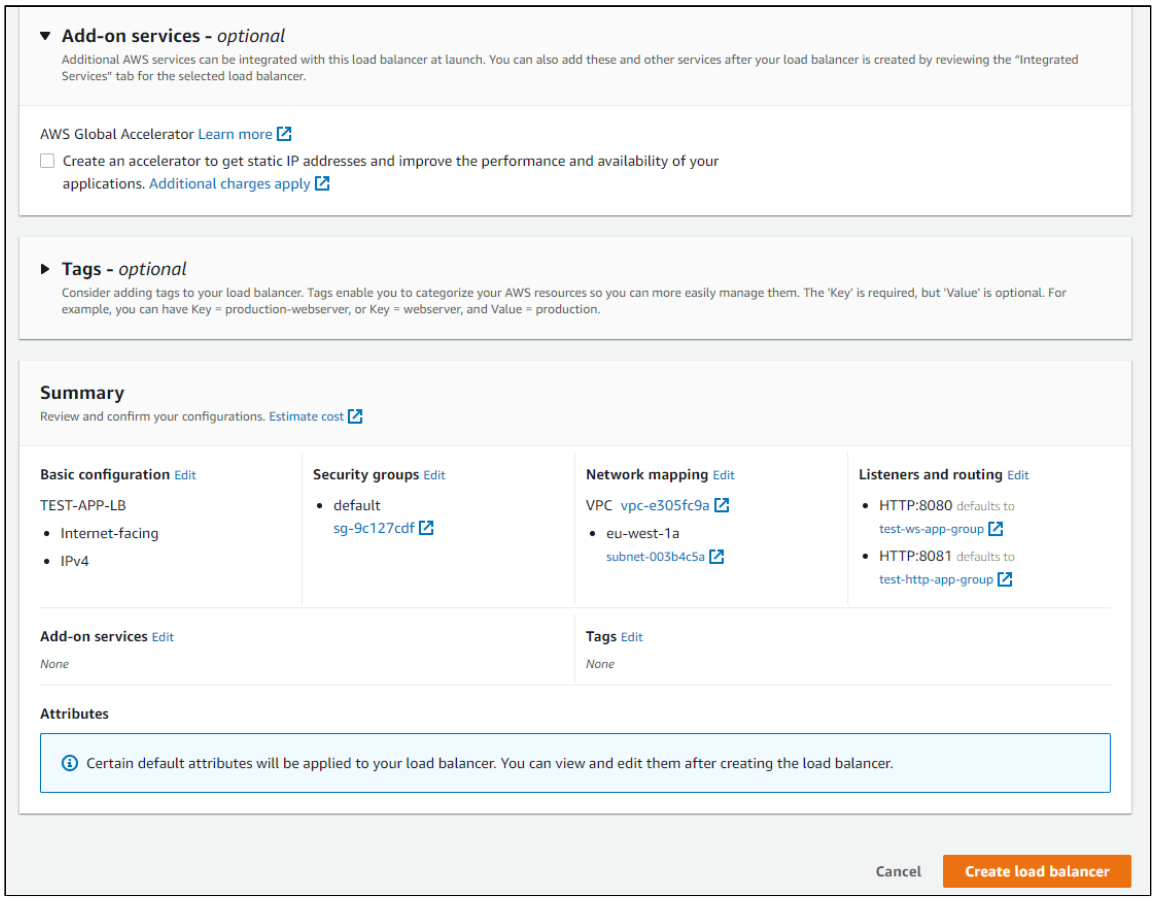


2.5. В разделе Listeners and routing добавьте Websocket порт (обязательно) и HTTP порт (по необходимости)

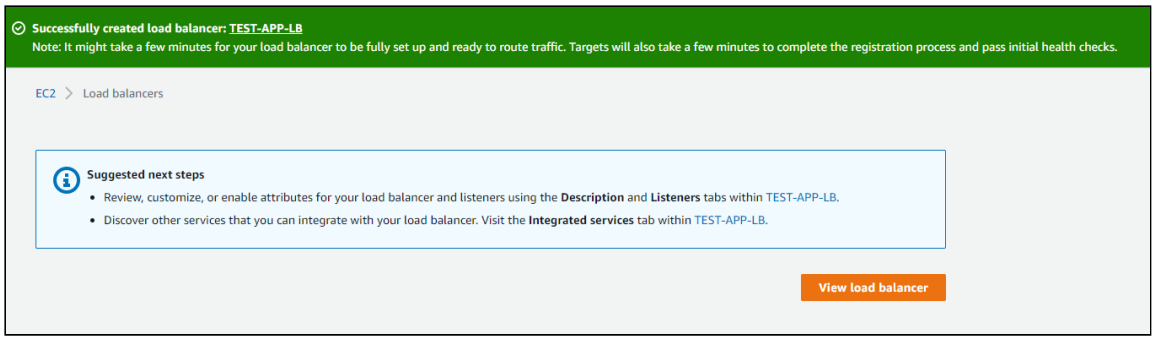


Для каждого из портов создается целевая группа (target group), см ниже.

2.6. Внизу страницы нажмите `Create load balancer`

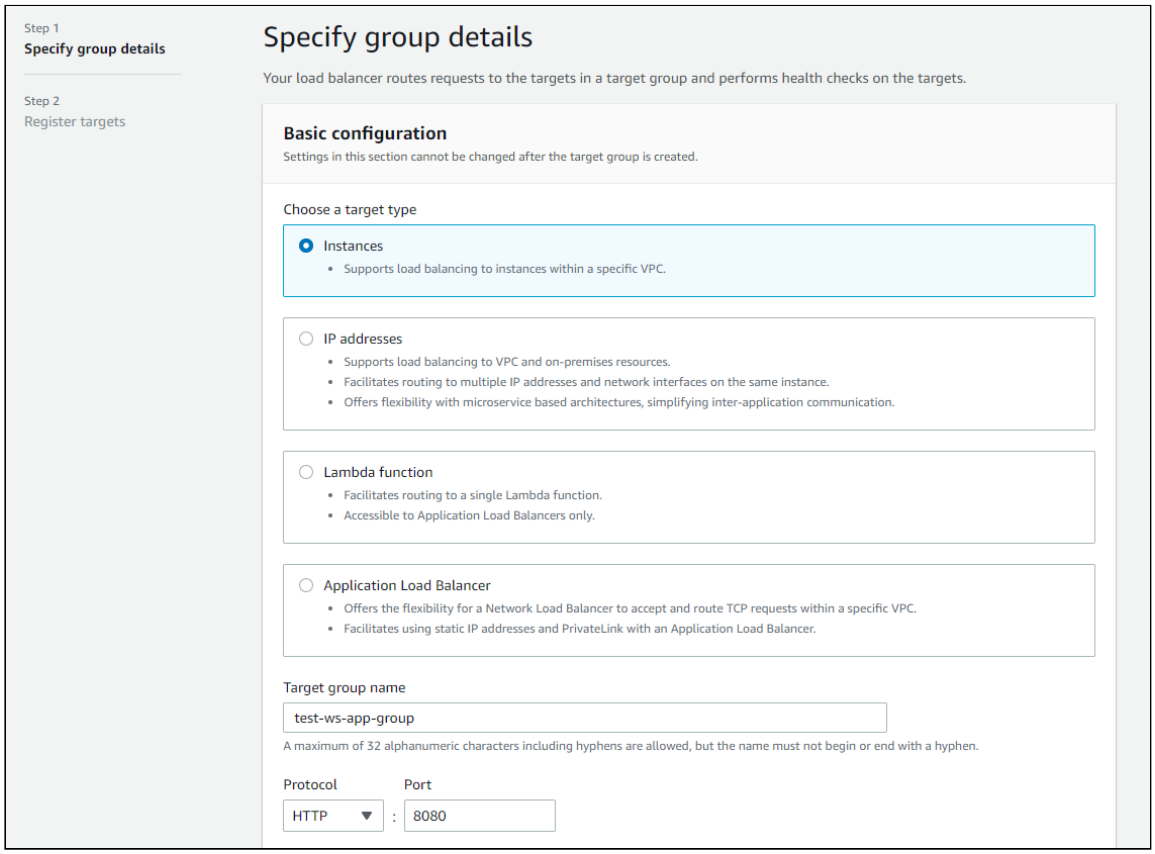


Создание балансировщика завершено

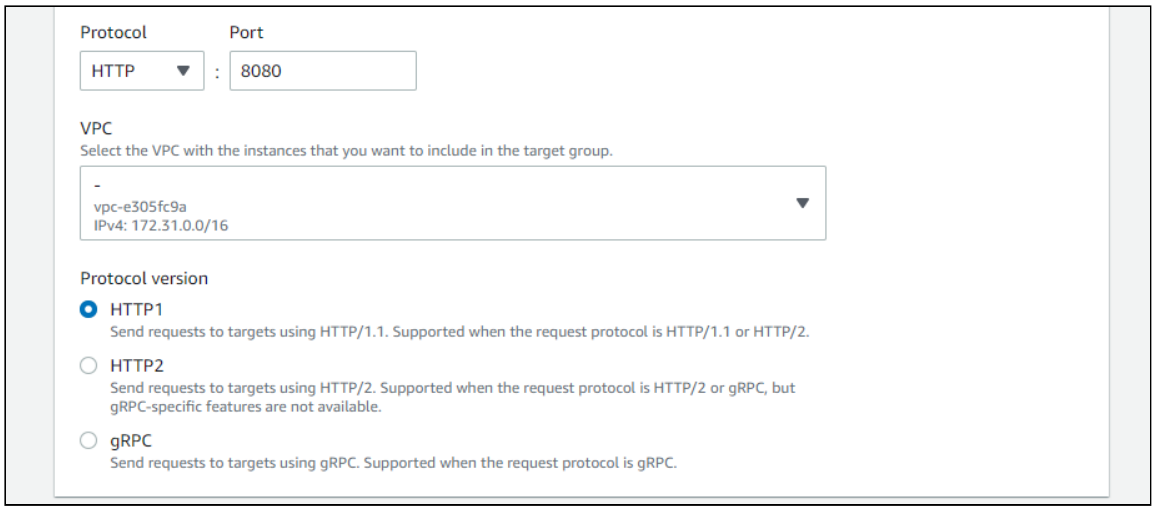


3. Создание целевой группы для Websocket порта

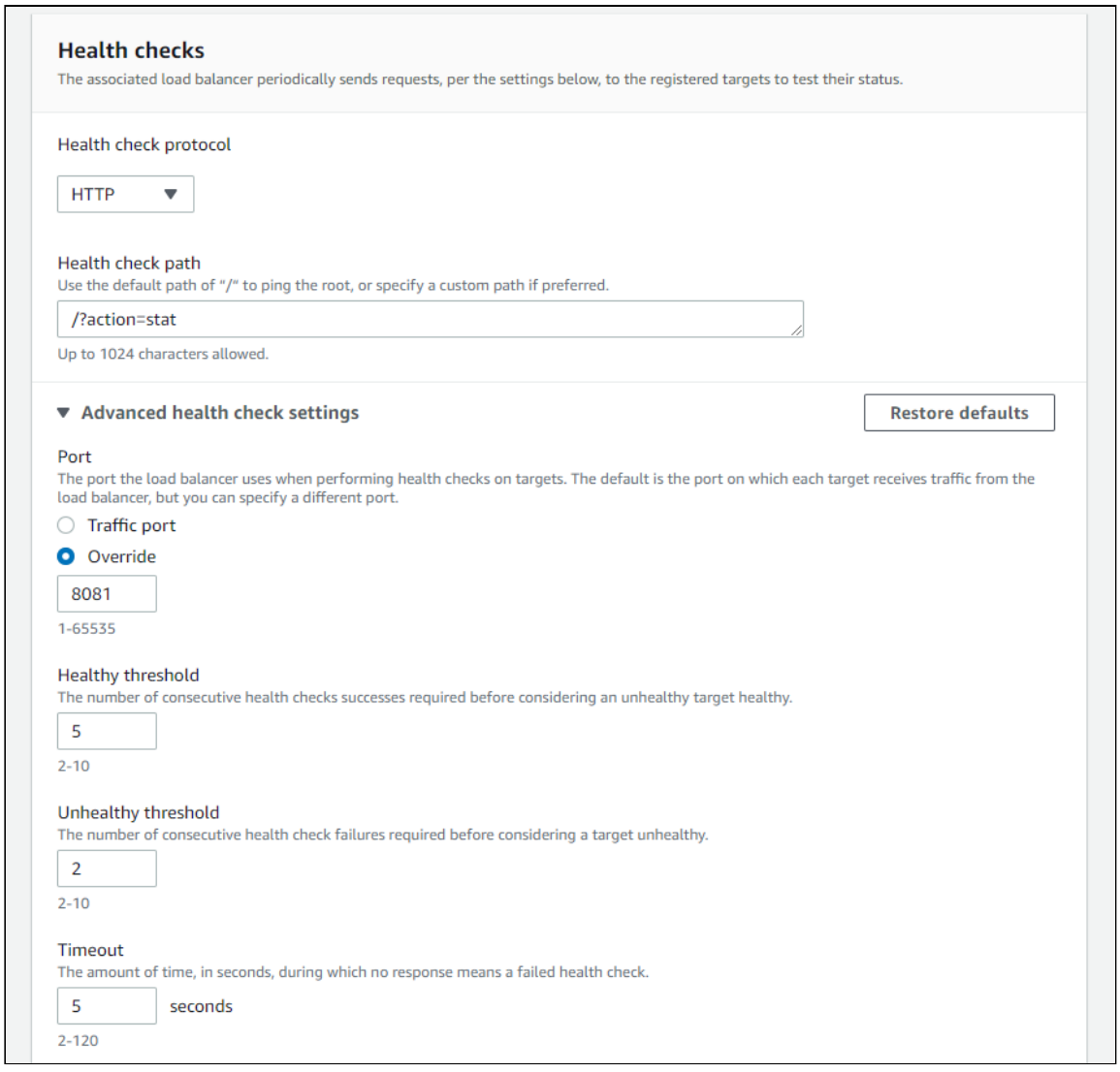
3.1. Выберите тип целей `Instances` (предлагается по умолчанию), задайте имя группы



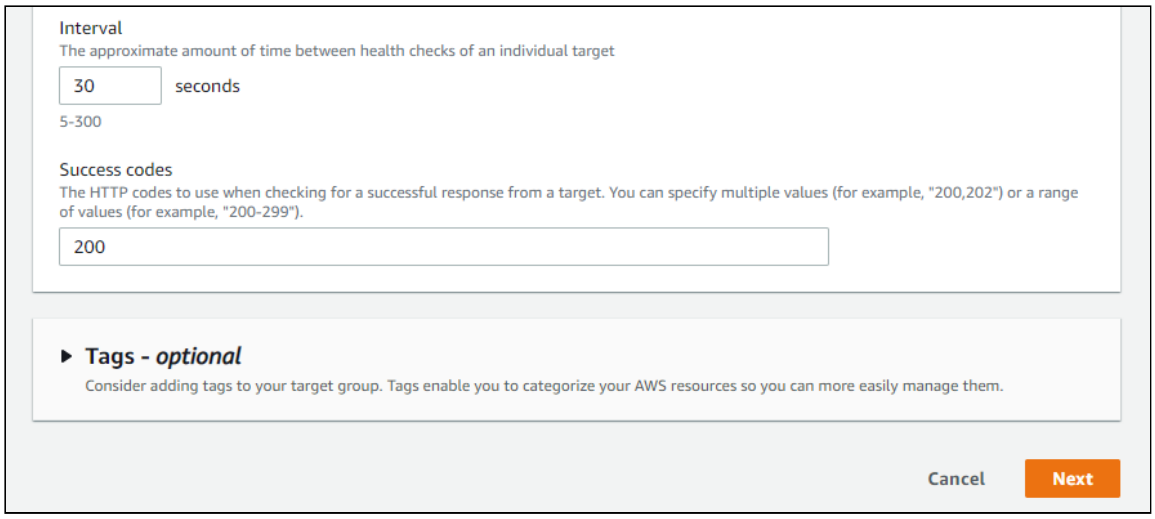
3.2. Укажите websocket порт WCS сервера `8080`, выберите подсеть и версию протокола `HTTP1`



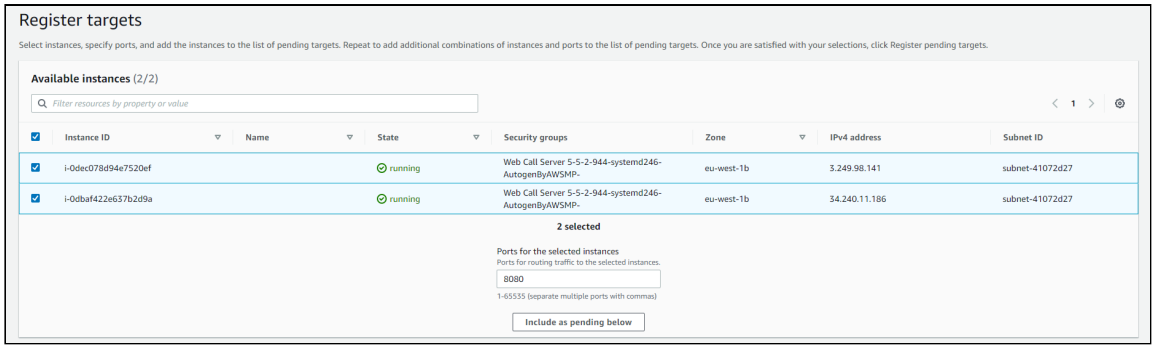
3.3. В разделе Health check настройте проверку работоспособности сервера, используя HTTP порт (8081) и запрос страницы статистики `/?action=stat`



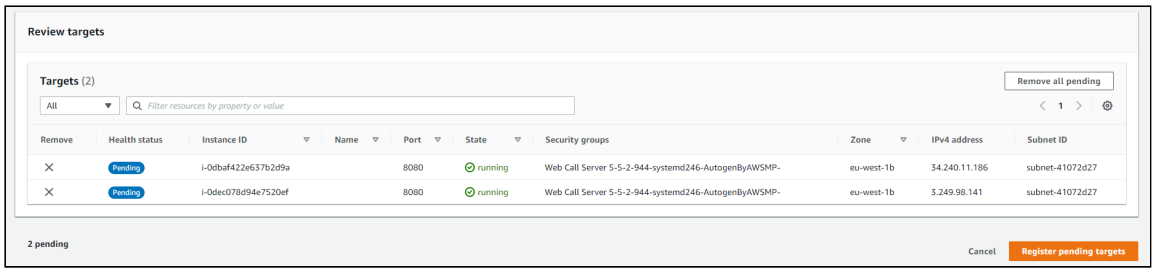
и нажмите `Next`



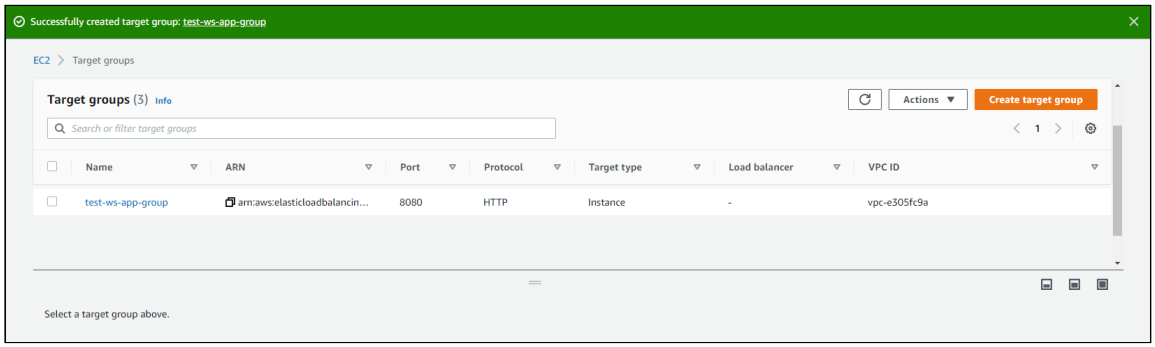
3.4. На странице `Register targets` выделите нужные серверы и нажмите `Include as pending below`



Затем нажмите `Register pending targets`



На этом создание целевой группы завершено

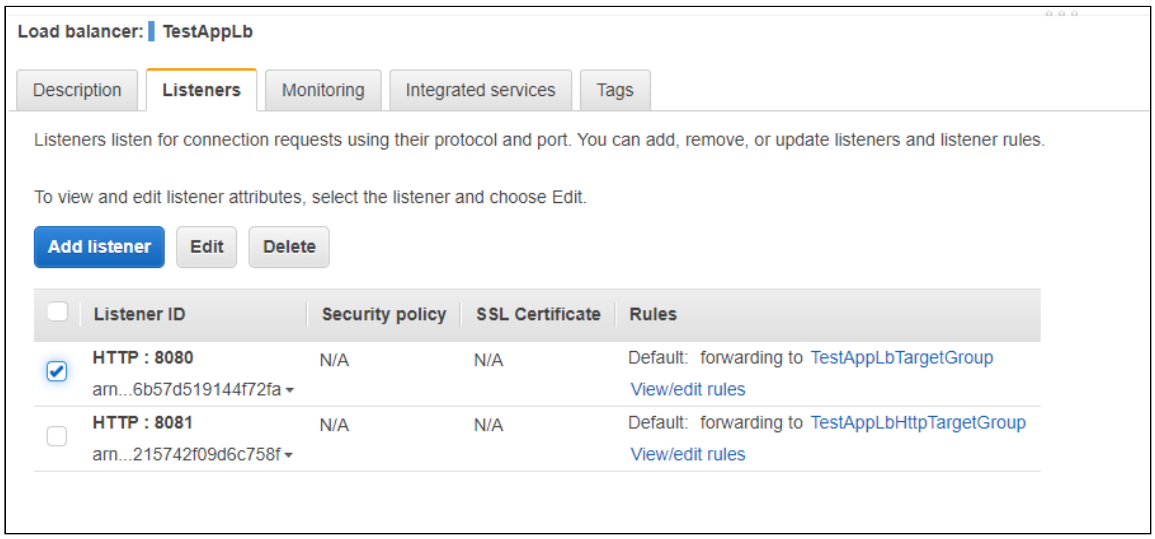


Балансировщик нагрузки, использующий данную группу, начнет работать после того, как хотя бы один сервер в группе успешно пройдет проверку работоспособности.

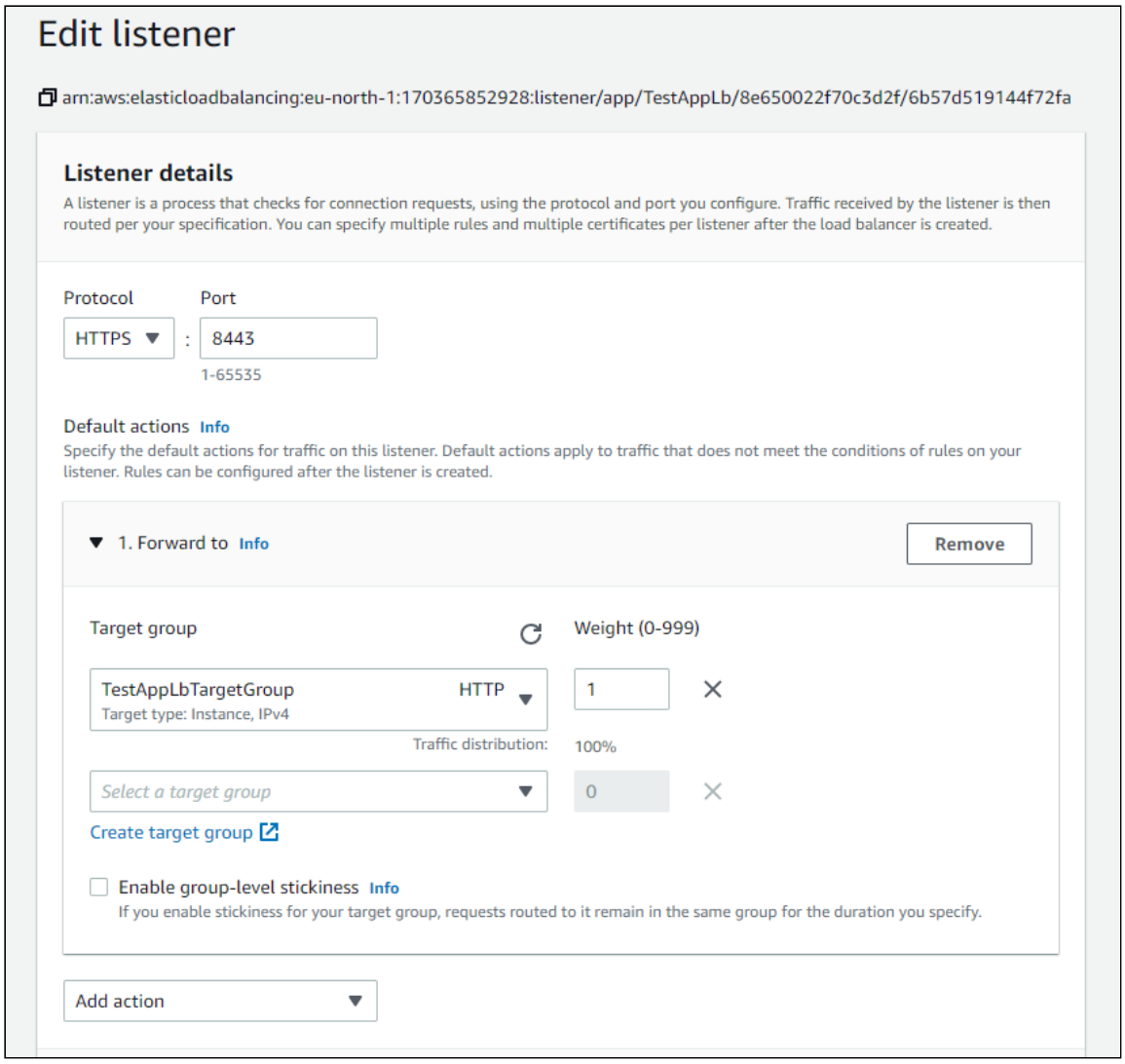
4. Редактирование порта балансировщика

Если балансировщик создается для группы автоматического масштабирования (см ниже), при его создании нельзя выбрать порт HTTPS, только HTTP. В этом случае настройки порта следует изменить.

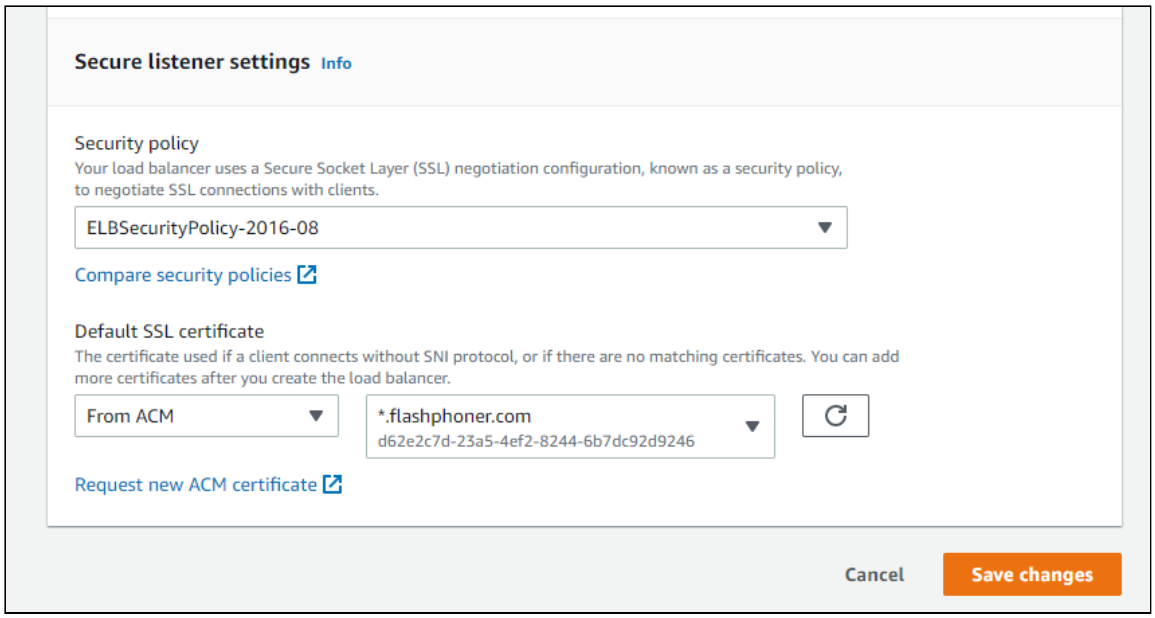
4.1. В разделе `Load balancers` - `Load balancers` выберите вкладку `Listeners` для нужного балансирощика. Выберите Websocket порт и нажмите `Edit`



4.2. Выберите протокол HTTPS и задайте Secure Websocket порт (например, `8443`)



4.3. В разделе `Secure listener settings` выберите SSL-сертификат из имеющихся для использования для домена, который будет настроен для точки входа балансировщика, или создайте новый. Затем нажмите Save changes



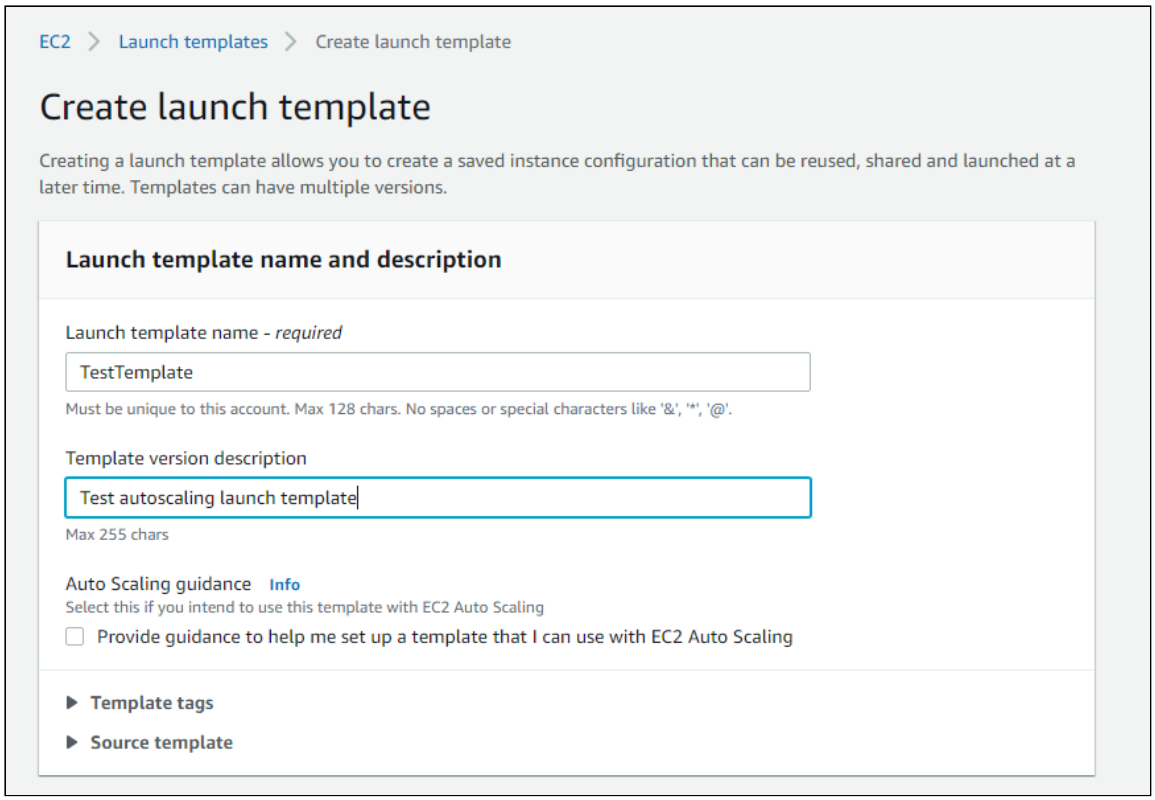
Настройки порта балансировщика изменены и применяются немедленно

Запуск балансировщика нагрузки для приложений на базе образа из AWS Marketplace с автоматическим масштабированием

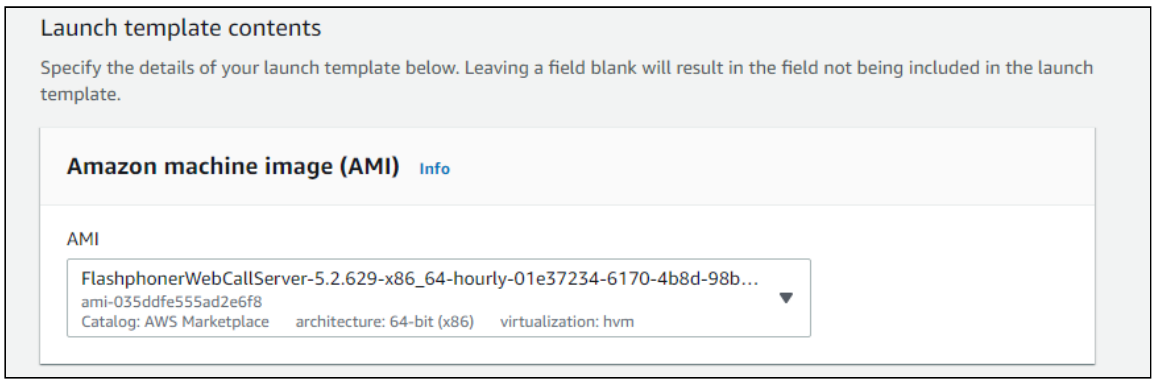
Создание балансировщика нагрузки с автомасштабированием на базе образа из AWS Marketplace полезно для периодического запуска группы серверов, например, при проведении мероприятия (длительностью в часы, дни, недели). В этом случае собственная лицензия с помесячной оплатой обойдется дороже, поэтому на короткие промежутки времени рекомендуется использовать образ из Marketplace.

1. Создание шаблона запуска (Launch Template)

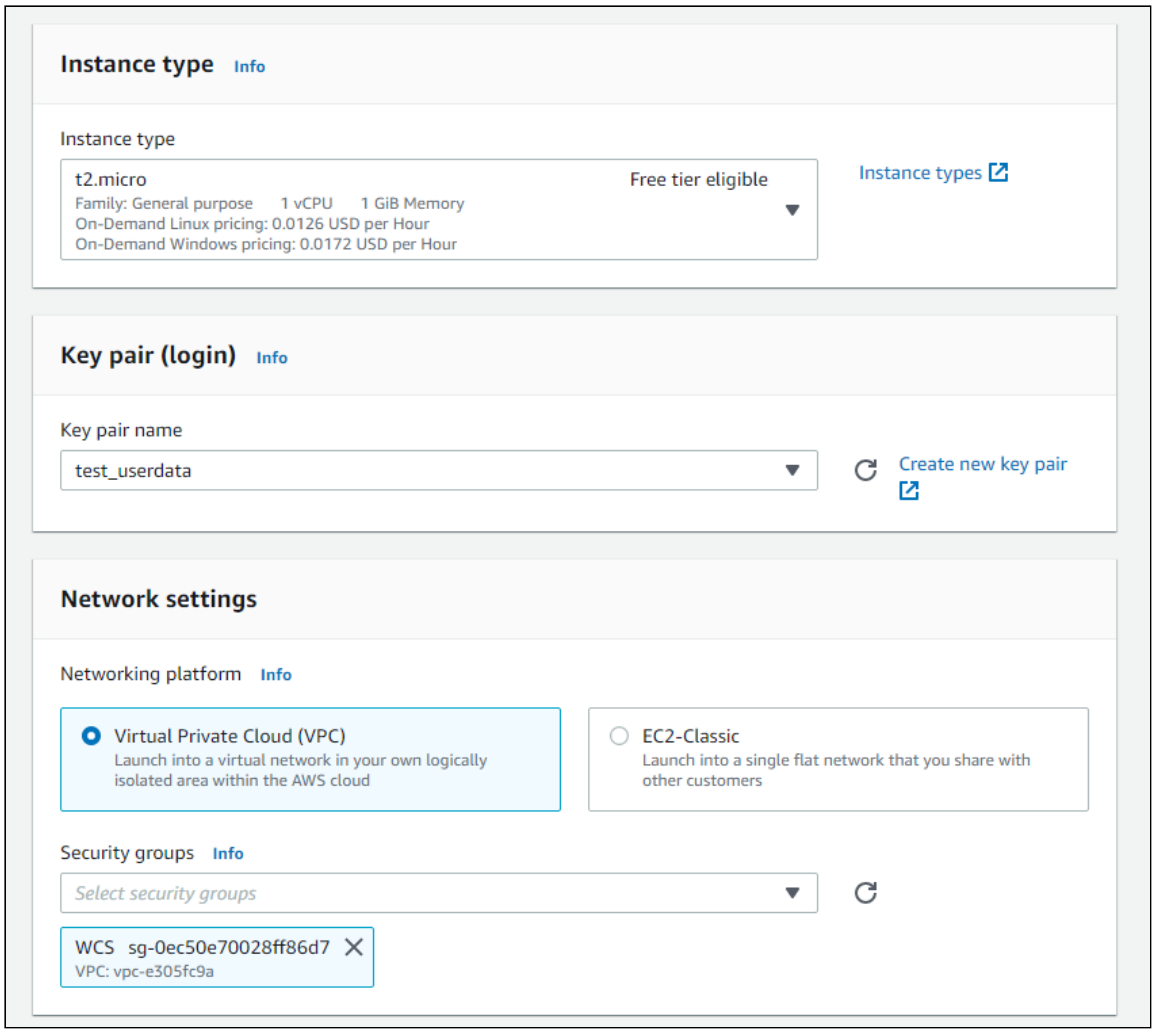
1.1. Перейдите в боковом меню EC2 Console в раздел `Instances` - `Launch Templates` и нажмите `Create launch template`. Откроется мастер создания шаблонов запуска. Введите имя и описание шаблона



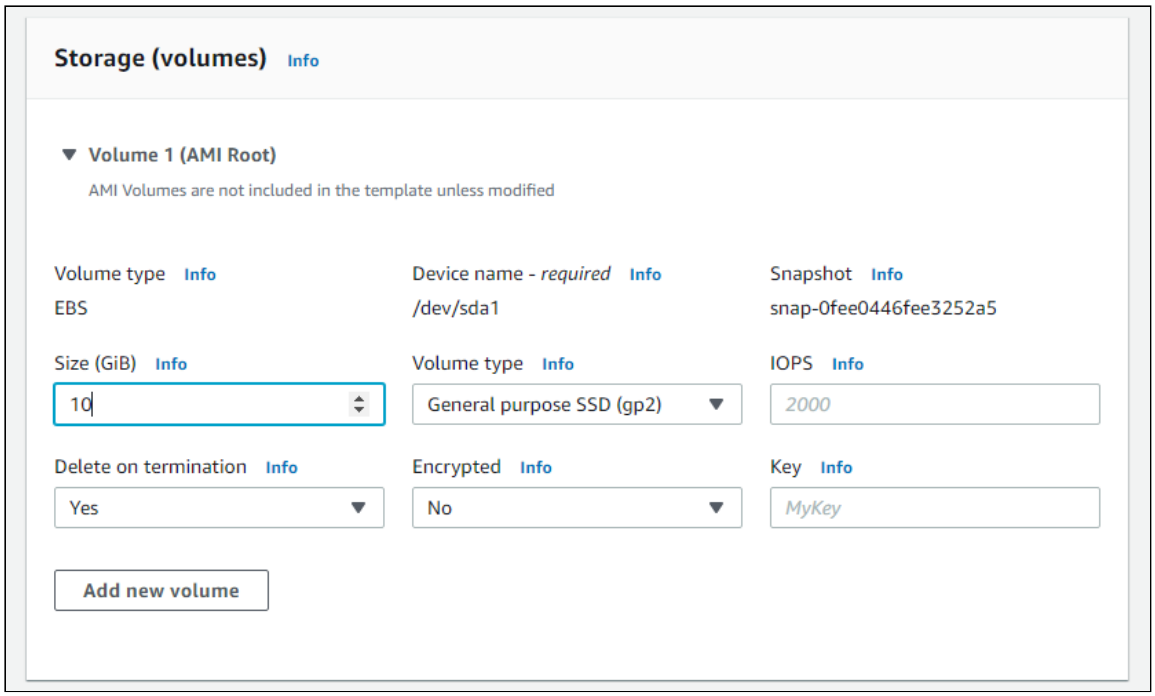
1.2. Выберите последний образ FlashphonerWebCallServer



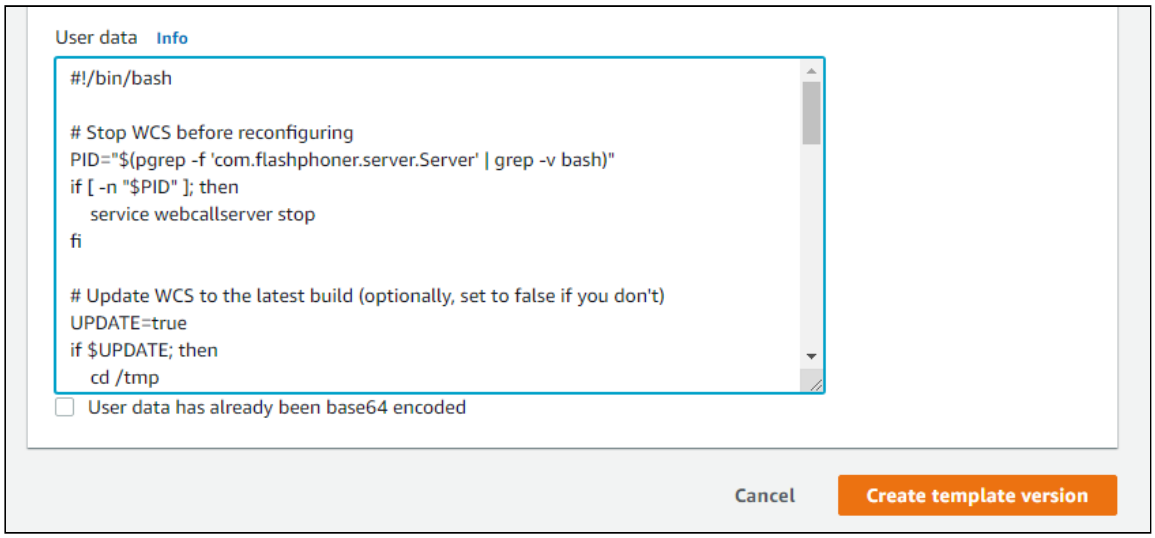
1.3. Выберите тип ВМ, пару ключей для доступа к экземпляру по SSH, группу безопасности



1.4. Задайте размер и параметры диска для экземпляров, создаваемых на базе этого шаблона



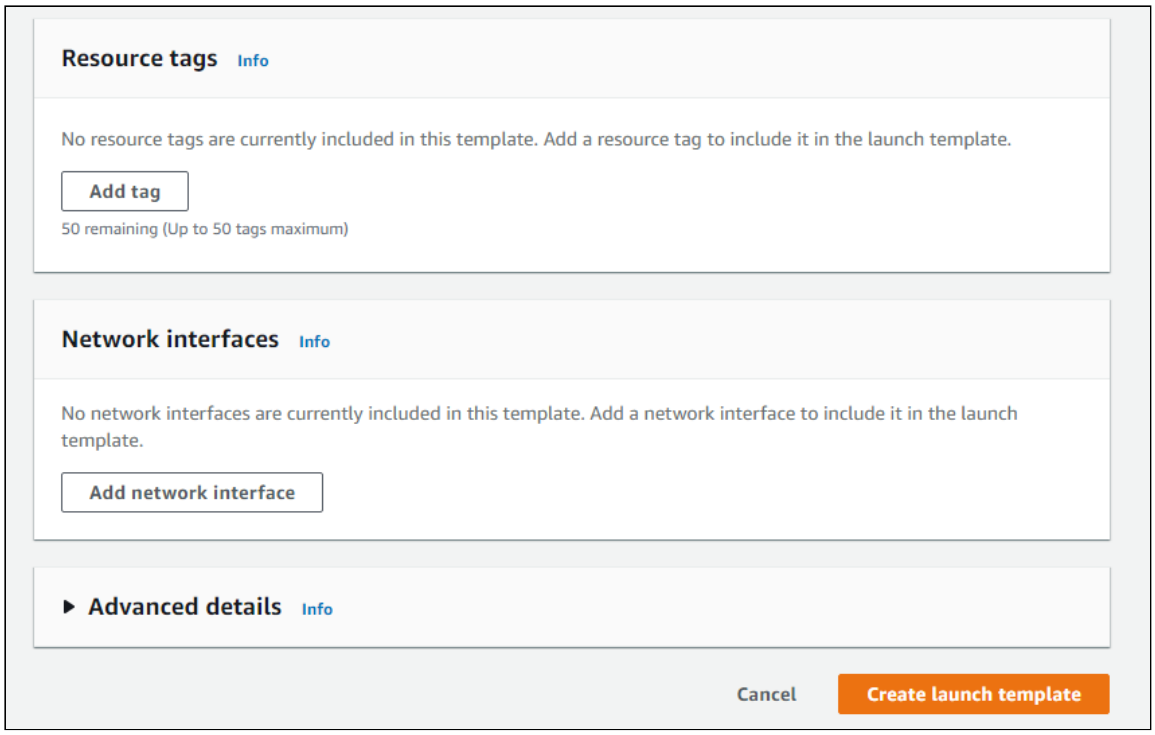
1.5. Разверните раздел формы `Advanced details`. В поле `User data` введите скрипт для обновления и настройки WCS



Пример скрипта обновления и настройки Edge сервера для масштабируемой CDN, предназначенной для доставки WebRTC

Edge setup script

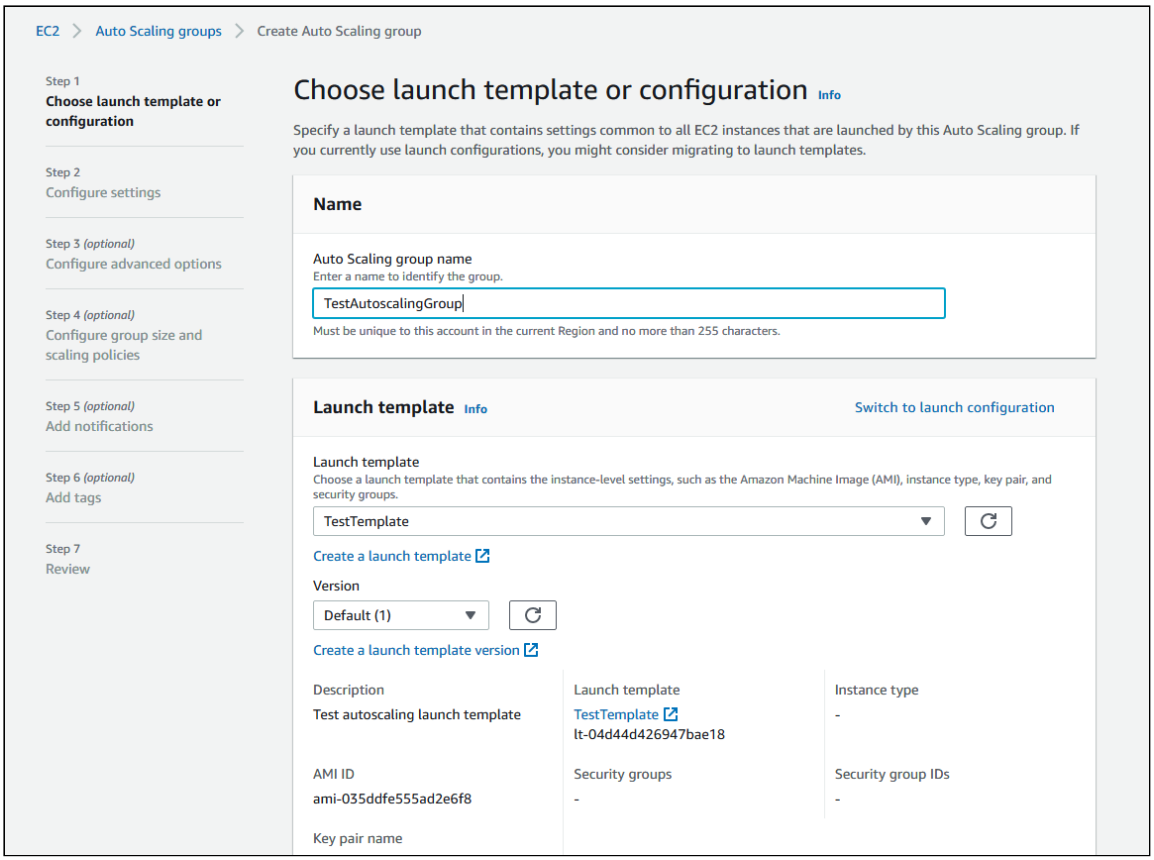
1.6. Нажмите `Create launch template`



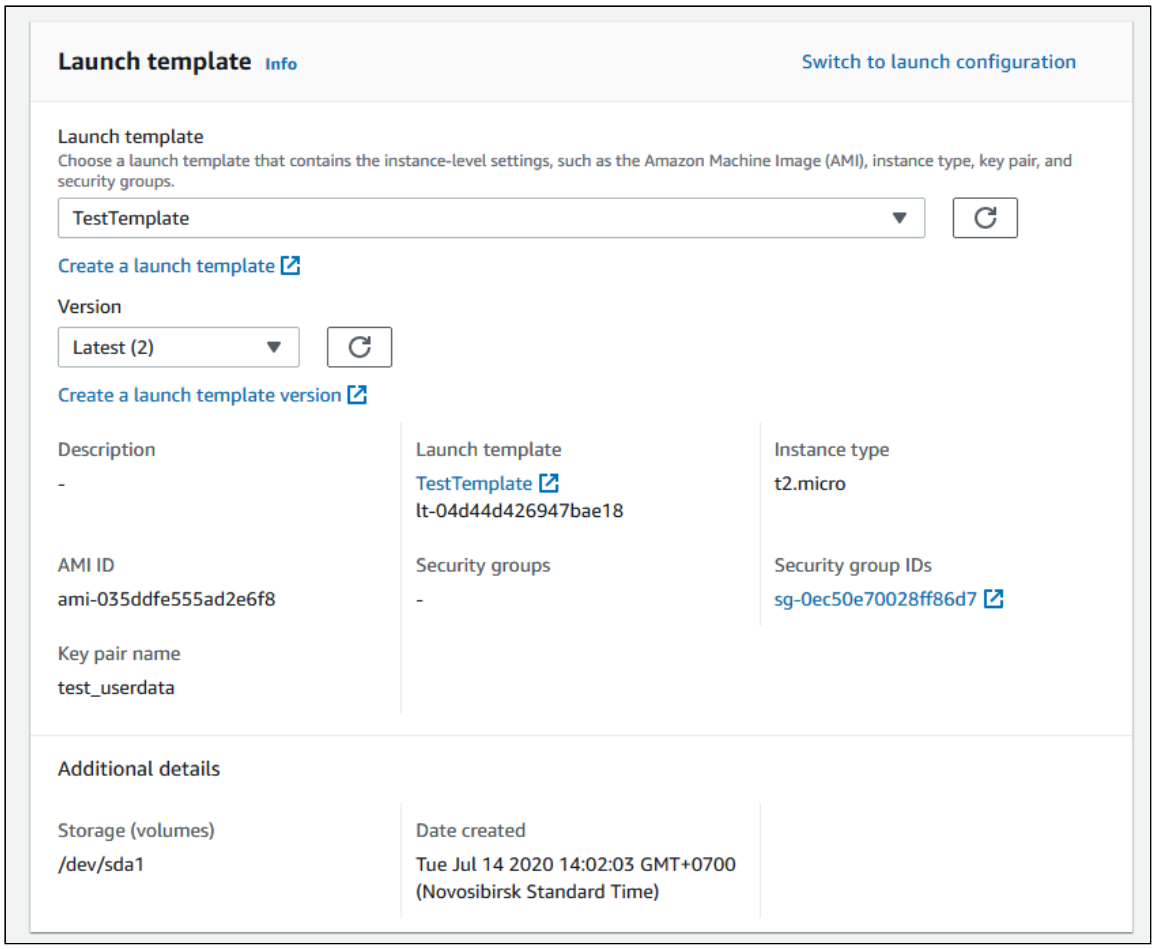
Шаблон запуска будет создан

2. Создание группы масштабирования

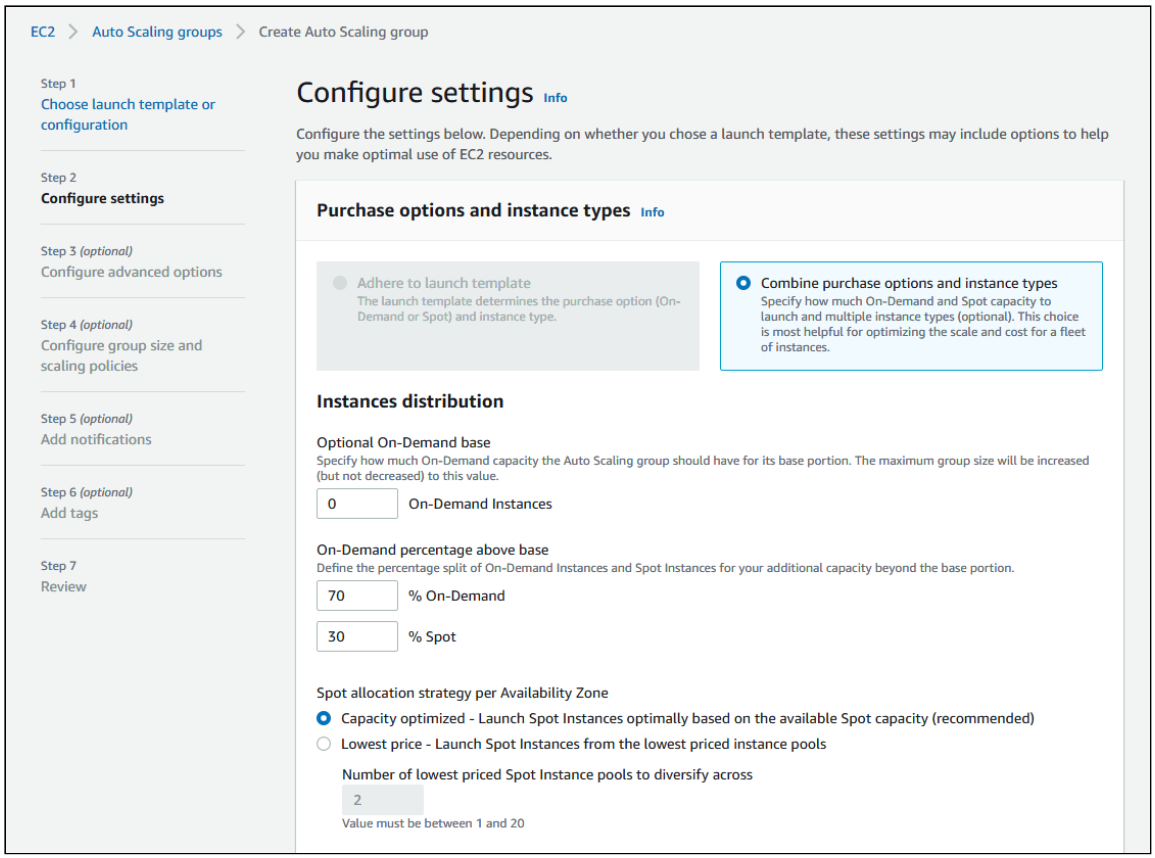
2.1. Перейдите в боковом меню EC2 Console в раздел `Instances` - `Auto Scaling Groups` и нажмите `Create an Auto Scaling Group`. Откроется мастер создания группы. Введите имя группы



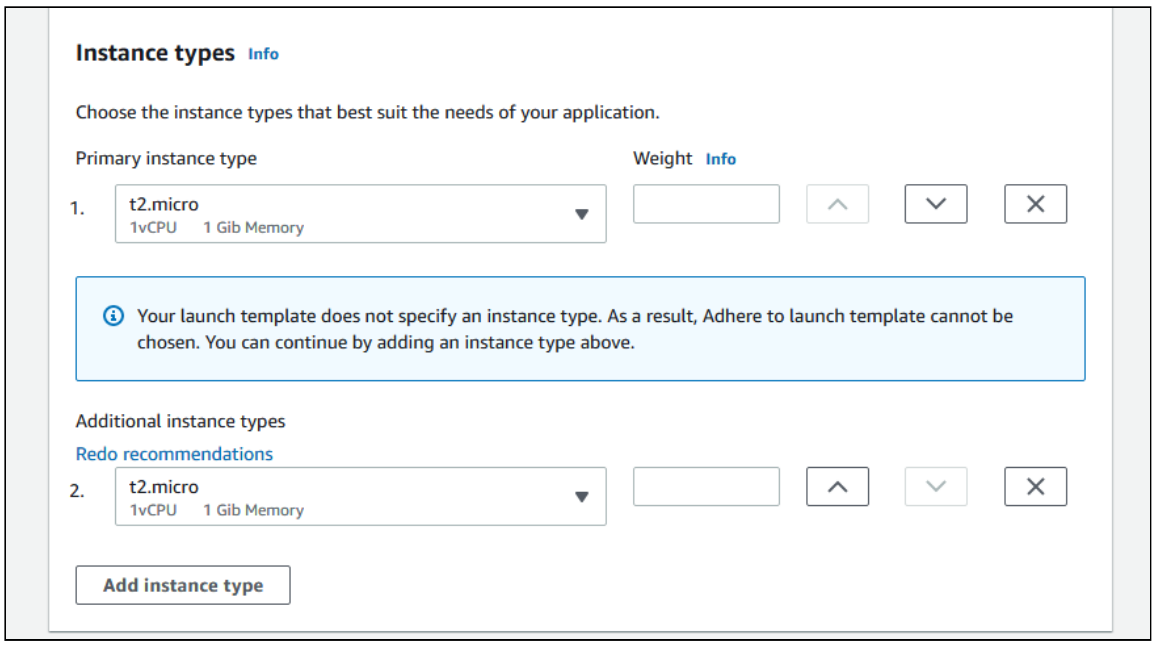
2.2. Выберите шаблон запуска, укажите версию `Latest`



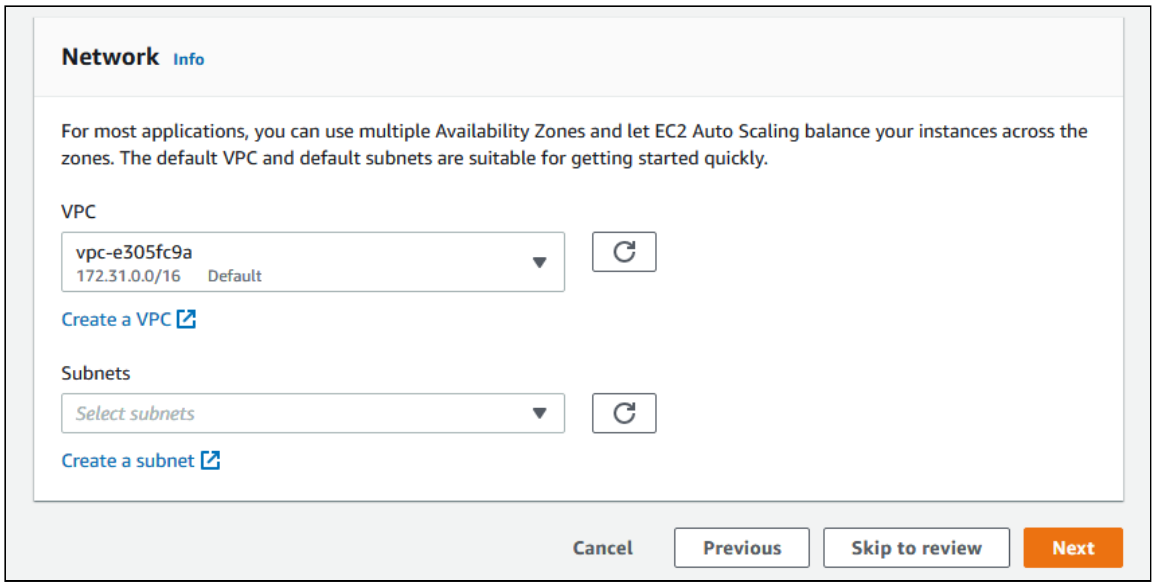
2.3. Укажите соотношение между используемыми ВМ (по требованию/с аукциона). По умолчанию предлагается `70 %` по требованию, рекомендуется указать `100 %`



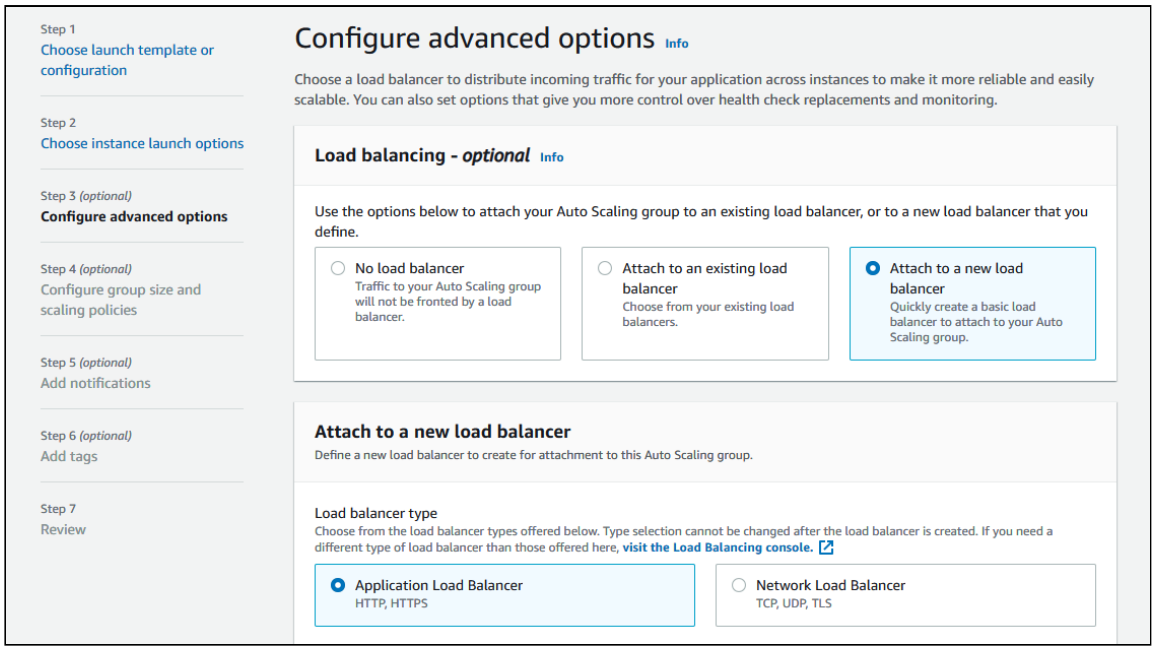
2.4 Выберите типы создаваемых ВМ



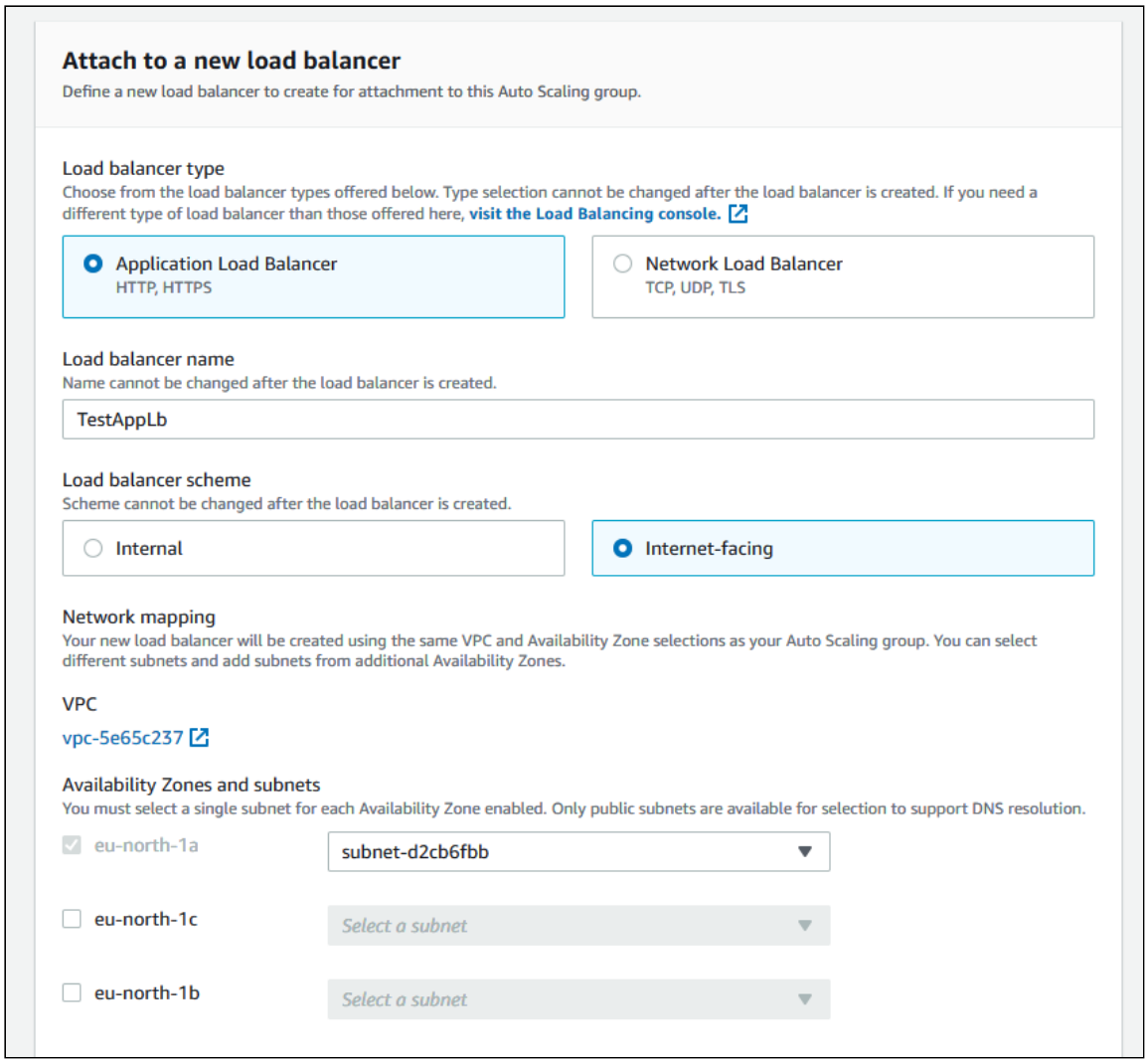
2,5. Выберите сеть и подсети для создаваемых ВМ



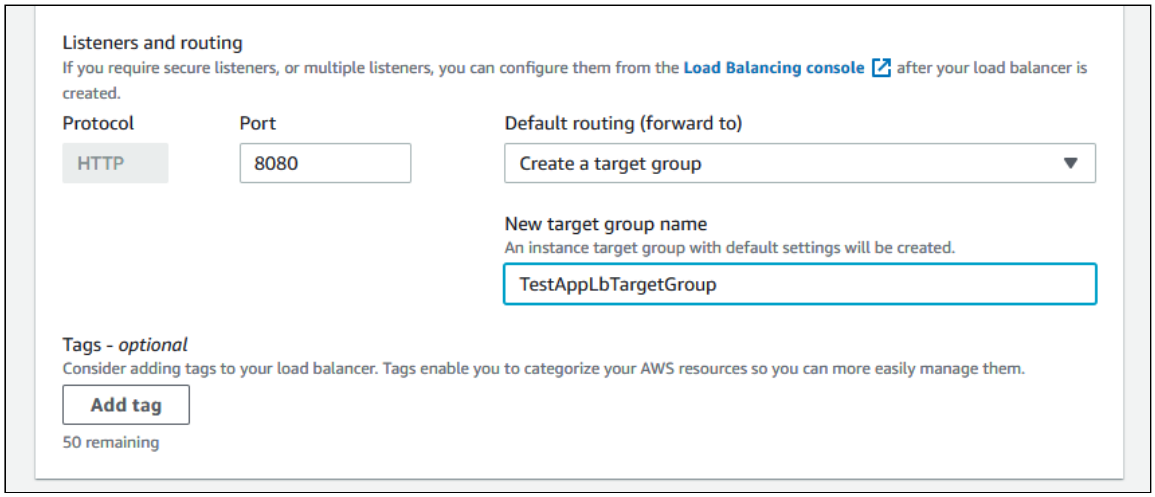
2.6. Выберите `Attach to a new load balancer`



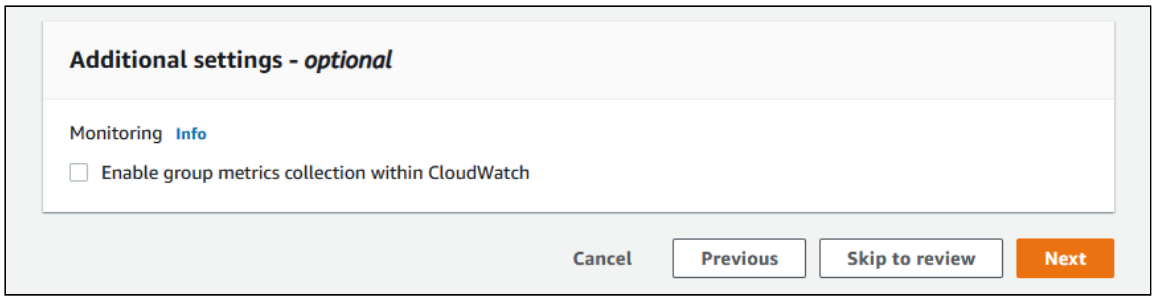
2.7. Выберите тип балансировщика `Application Load balancer`, задайте имя, выберите `Internet-facing`, укажите зоны доступности и соответствующие подсети



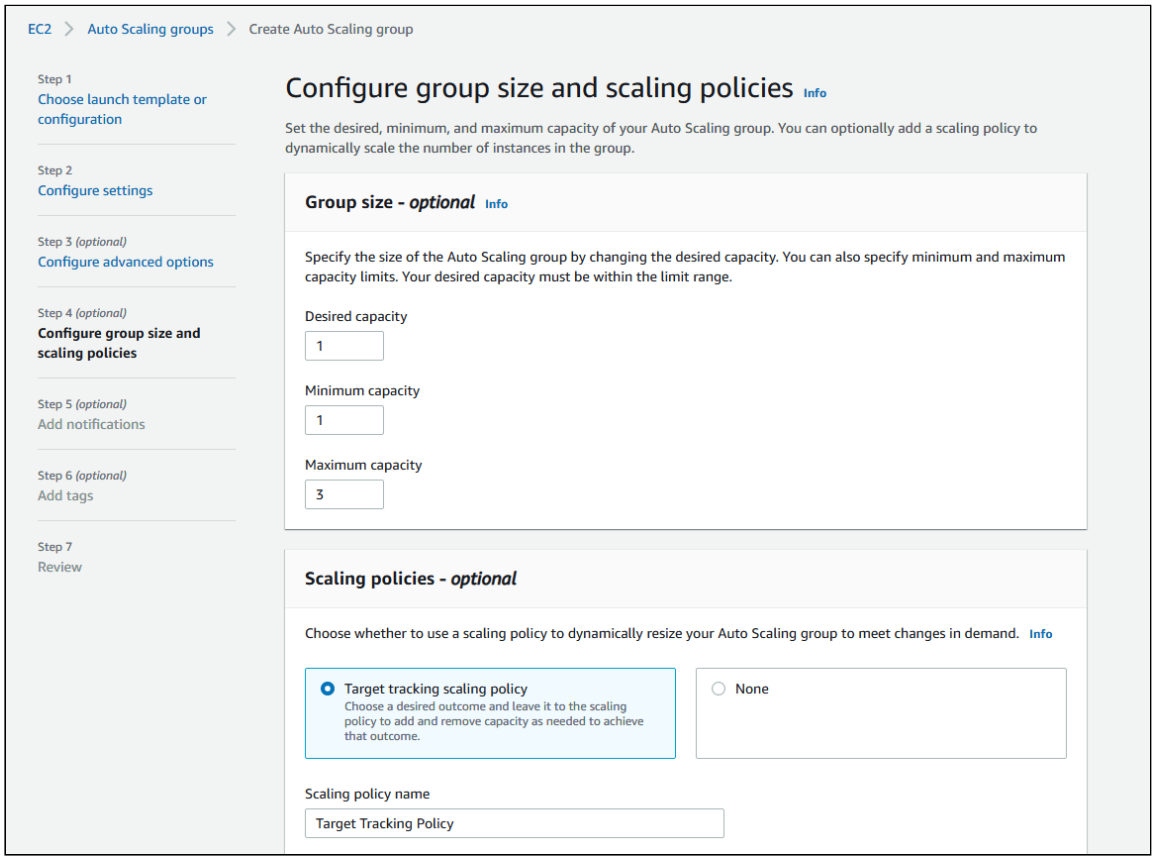
2.8. В разделе `Listeners and routing` укажите Websocket порт (8080), выберите `Create a target group` и укажите имя целевой группы. которая будет создана



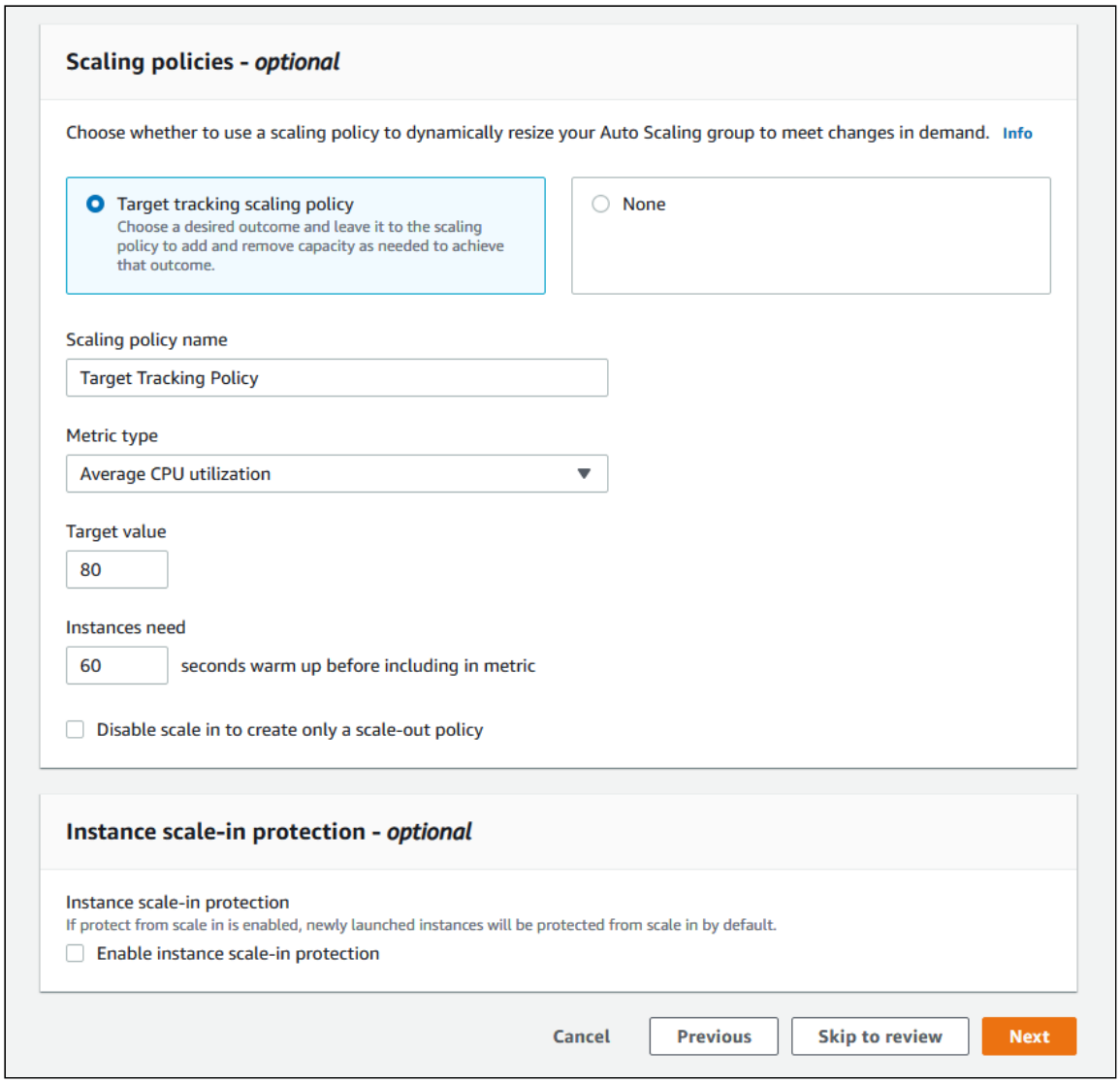
Затем нажмите `Next`



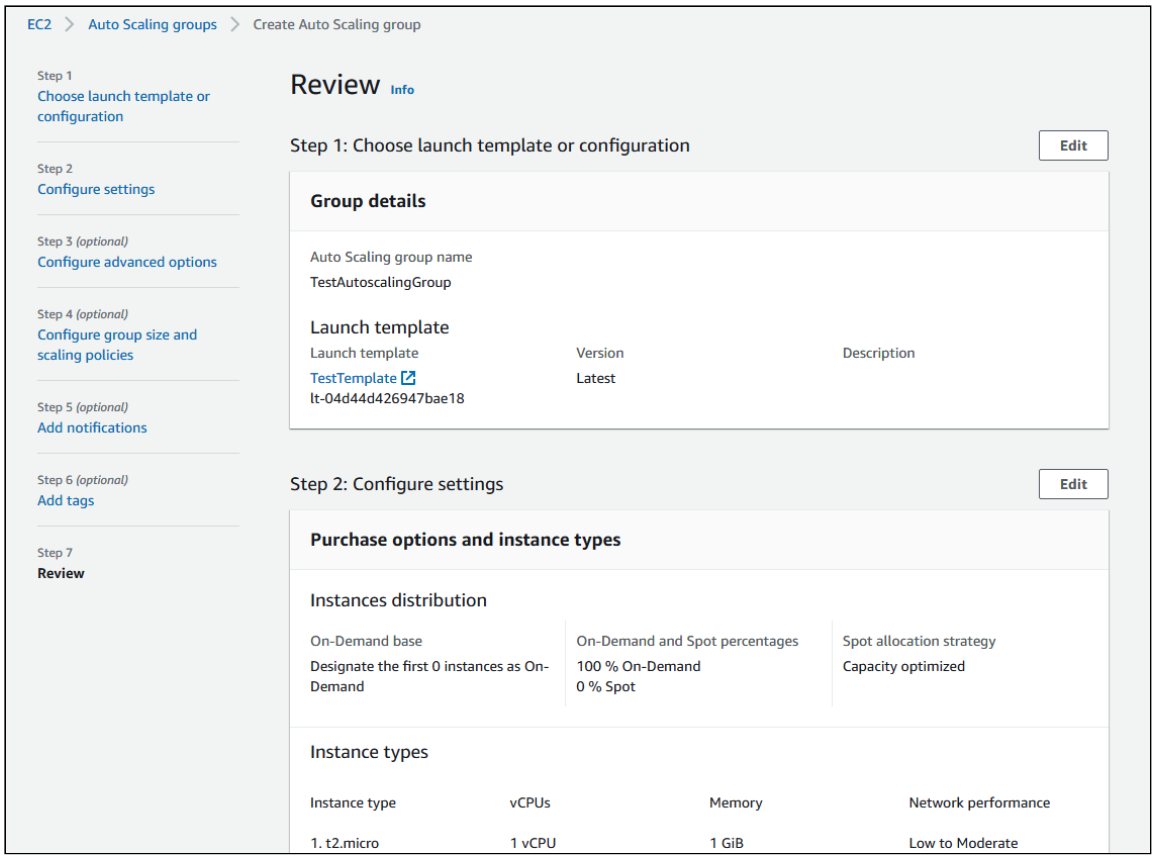
2.9. Задайте максимальный размер группы



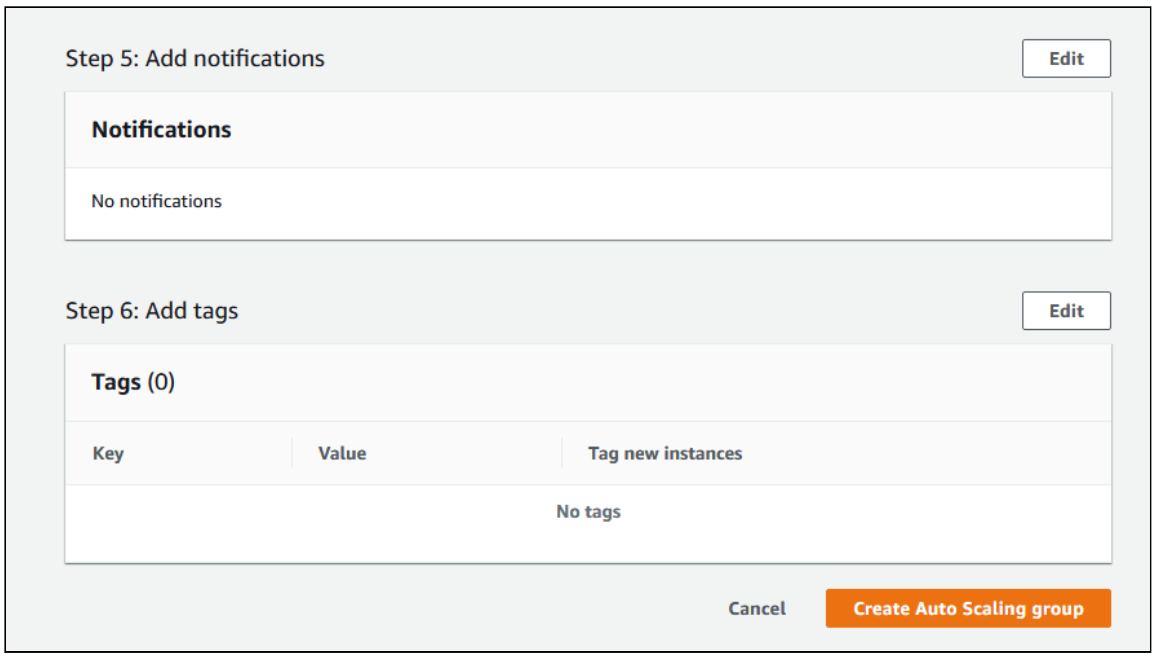
2.10. Укажите политику масштабирования по загрузке процессора, задайте время запуска экземпляра



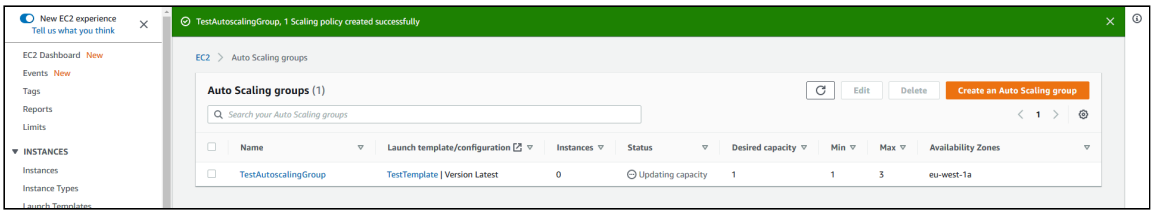
2.11. Проверьте параметры группы



2.12. Нажмите `Create Auto Scaling group`



Будет создана группа масштабирования, и запущен один экземпляр в группе



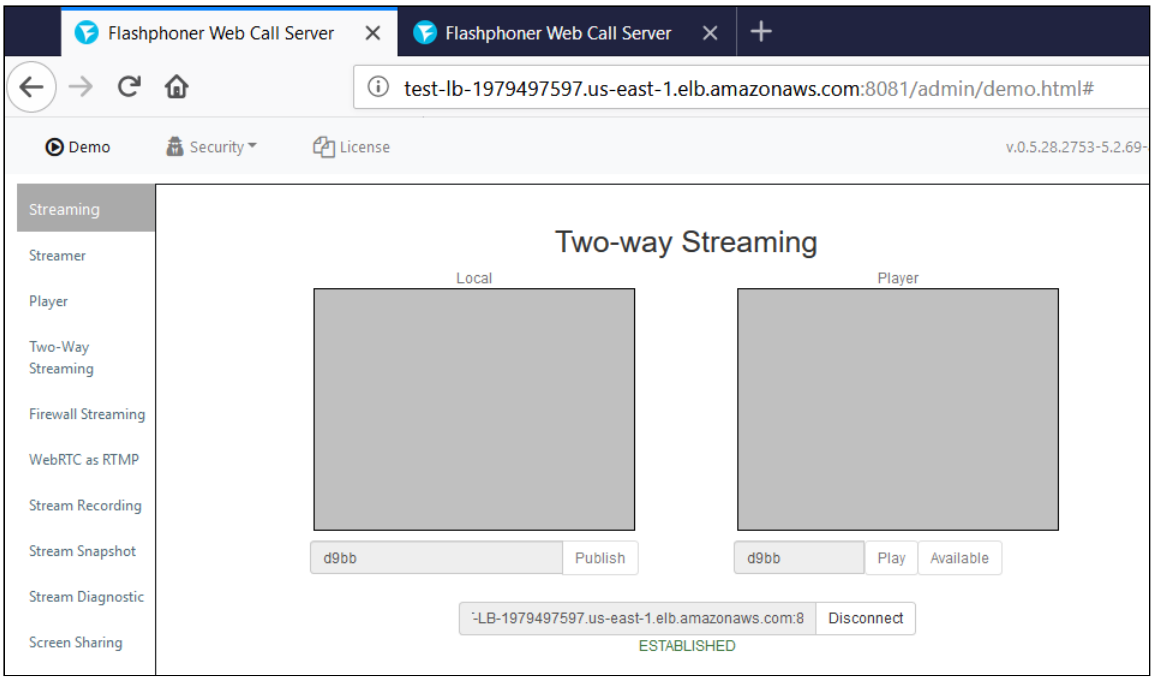
2.13. Измените параметры порта созданного балансировщика нагрузки, как описано выше

Тестирование

Если в балансировщике еще не запущено ни одного экземпляра сервера, новый экземпляр запустится в момент, когда будет создана группа масштабирования, получающая трафик от этого балансировщика. При выполнении условий масштабирования, будут созданы следующие экземпляры (Для загрузки процессора в ходе тестирования, может использоваться публикация потока с транскодингом, например, публикация RTMP потока в автоматически создаваемый микшер). Все запущенные экземпляры добавляются в соответствующий балансировщик.

Если экземпляр сервера (один или более из добавленных в балансировщик) работает, с ним может быть установлено WebSocket соединение, например `ws://Load balancer DNS name:8080`.

Пример Two-way Streaming, открытый по адресу балансировщика или конкретного экземпляра, может быть использован для установки WebSocket соединения:



Чтобы убедиться, что соединения распределяются между активными экземплярами в балансировщике нагрузки, используйте страницу статистики:

`http://WCS_ADDRESS:8081/?action=stat`

Откройте страницу каждого экземпляра для контроля значения `connections_websocket`: