Set Up Network and HTTP Load Balancers

네트워크 부하 분산기와 HTTP 부하 분산기의 차이점, 그리고 Compute Engine VM에서 실행되는 애플리케이션용으로 부하 분산기 설정하는 방법 제시
*Reference: https://cloud.google.com/load-balancing/docs/network?hl=ko

모든 리소스에 대한 기본 리전 및 영역 설정

1. 기본 리전 설정
   - gcloud config set compute/region us-east4
2. 기본 영역 설정
   - gcloud config set compute/zone us-east4-a

다중 웹 서버 인스턴스 만들기

*시나리오: Compute Engine VM 인스턴스 3개 생성 후 Apache 설치 > HTTP 트래픽이 인스턴스에 도달할 수 있도록 방화벽 규칙 추가

1. 기본 영역에 가상 머신 www1, www2, www3 생성
   example)

```
[www1]
  gcloud compute instances create www1 \
--zone=us-east4-a \
--tags=network-lb-tag \
--machine-type=e2-medium \
--image-family=debian-11 \
--image-project=debian-cloud \
--metadata=startup-script='#!/bin/bash
  apt-get update
  apt-get install apache2 -y
  service apache2 restart
  echo "
<h3>웹 서버: www1</h3>" | tee /var/www/html/index.html'

[www2]
  gcloud compute instances create www2 \
--zone=us-east4-a \
--tags=network-lb-tag \
```

```
--machine-type=e2-medium \
--image-family=debian-11 \
--image-project=debian-cloud \
--metadata=startup-script='#!/bin/bash
  apt-get update
  apt-get install apache2 -y
  service apache2 restart
  echo "
<h3>웹 서버: www2</h3>" | tee /var/www/html/index.html'

[www3]
  gcloud compute instances create www3 \
--zone=us-east4-a \
--tags=network-lb-tag \
--machine-type=e2-medium \
--image-family=debian-11 \
--image-project=debian-cloud \
--metadata=startup-script='#!/bin/bash
  apt-get update
  apt-get install apache2 -y
  service apache2 restart
  echo "
<h3>웹 서버: www3</h3>" | tee /var/www/html/index.html'
```

2. VM 인스턴스에 외부 트래픽을 허용하는 방화벽 규칙 만들기
   - gcloud compute firewall-rules create www-firewall-network-lb
     --target-tags network-lb-tag --allow tcp:80
3. 인스턴스의 외부 IP 주소를 가져와 주소가 정상적으로 작동되는지 확인
   *curl을 사용하여 각 인스턴스가 실행 중인지 확인하고 이때 외부 IP는 list에서 출력되는 외부 IP값을 이용할 것
   - gcloud compute instances list
   - curl http:// `external IP address`

부하 분산 서비스 구성

부하 분산 서비스를 구성하면 가상 머신 인스턴스는 사용자가 구성한 고정 IP 주소로 전송되는 패킷을 수신한다.
Compute Engine 이미지로 만든 인스턴스는 해당 IP 주소를 처리하도록 자동 구성된다.

1. 부하 분산기의 고정 외부 IP 주소 생성

- gcloud compute addresses create network-lb-ip-1 \ --region us-east4

2. 기존 HTTP 상태 점검 리소스 추가
   - gcloud compute http-health-checks create basic-check
3. 인스턴스와 같은 리전에 타켓 풀 추가
   *타켓 풀을 생성하고 서비스가 작동하려면 상태 확인이 필요하므로 위에서 추가한 상태 점검 리소스를 사용한다.
   - gcloud compute target-pools create www-pool \ --region us-east4 -http-health-check basic-check
4. 풀에 인스턴스를 추가
   - gcloud compute target-pools add-instances www-pool \ --instances www1, ww2, ww3
5. 전달 규칙 추가
   - gcloud compute forwarding-rules create www-rule \ --region us-east4 \ --ports 80 \ --address network-lb-ip-1 \ --target-pool www-pool

인스턴스에 트래픽 보내기

부하 분산 서비스를 위에서 구성했으므로 전달 규칙에 트래픽을 보내고 트래픽이 여러 인스턴스에 분산되는 것을 확인할 수 있다.

1. 부하 분산기에서 사용하는 www-rule 전달 규칙의 외부 IP 주소 보기
   - gcloud compute forwarding-rules describe www-rule --region us-east4
2. 외부 IP 주소 액세스
   - IPADDRESS=$(gclooud compute forwarding-rules describe www-rule --region us-east4 --format="json" | jq -r .IPAddress)
3. 외부 IP 주소 표시
   - echo $IPADDRESS
4. 외부 IP 주소에 액세스 및 이전 명령어에서 사용한 외부 IP 주소로 변경
   - while true; do curl -m1 $IPADDRESS; done
     *curl 명령어가 실행되면 인스턴스 세 개에서 무작위로 응답한다. 처음에 응답하지 않는다면 구성이 완전히 로드되어 인스턴스가 정상으로 표시될 때까지 30초 정도 기다린 후 다시 시도하고, Ctrl + C를 통해 명령어 실행을 중지할 수 있다.

(L7) HTTP 부하 분산기 만들기

HTTP(S) 부하 분산은 GFE(Google Front-end)에서 구현된다.
GFE는 Google의 글로벌 네트워크 및 제어 영역을 사용하여 전 세계로 분산되고 운영된다.
URL이 각기 적절한 인스턴스 집합으로 라우팅되도록 URL 규칙을 구성할 수 있다.

요청은 항상 사용자와 가장 가까운 인스턴스 그룹으로 라우팅된다.
*그룹의 용량이 충분하며 요청에 적합한 그룹의 경우

가장 가까운 그룹에 용량이 충분하지 않으면 용량이 있는 가장 가까운 그룹으로 요청이 전송된다.

Compute Engine 백엔드로 부하 분산기를 설정하려면 VM이 인스턴스 그룹에 있어야 한다.
관리형 인스턴스 그룹은 외부 HTTP 부하 분산기의 백엔드 서버를 실행하는 VM을 제공한다.

1. 부하 분산기 템플릿 생성

```
gcloud compute instance-templates create lb-backend-template \
--region= \
--network=default \
--subnet=default \
--tags=allow-health-check \
--machine-type=e2-medium \
--image-family=debian-11 \
--image-project=debian-cloud \
--metadata=startup-script='#!/bin/bash
  apt-get update
  apt-get install apache2 -y
  a2ensite default-ssl
  a2enmod ssl
  vm_hostname="$(curl -H "Metadata-Flavor:Google" \
  http://169.254.169.254/computeMetadata/v1/instance/name)"
  echo "Page served from: $vm_hostname" | \
  tee /var/www/html/index.html
  systemctl restart apache2'
```

   *관리형 인스턴스 그룹(MIG)을 사용하면 동일한 여러 VM에서 앱을 운영할 수 있다.
   자동 확장, 자동 복구, 리전(멀티 영역) 배포, 자동 업데이트 등 자동화된 MIG 서비스를 활용하여 워크로드의 확장성 및 가용성을 높일 수 있다.
2. 템플릿을 기반으로 관리형 인스턴스 그룹 생성
   - gcloud compute instance-groups managed create lb-backend-group --template=lb-backend-template --size=2 --zone=
3. 방화벽 규칙 만들기
   - gcloud compute firewall-rules create fw-allow-health-check \ --network=default \ --action=allow \ --direction=ingress \ --source-ranges=130.211.0.0/22,35.191.0.0/16 \ --target-tags=allow-health-check \ --rules=tcp:80
     *위 규칙은 Google Cloud 상태 점검 시스템의 트래픽을 허용하는 인그레스 규칙이다.
4. 인스턴스가 실행 중이므로 클라이언트가 부하 분산기에 연결하는 데 사용하는 전역 고정 외부 IP 주소를 설정한다.
   - gcloud compute addresses create lb-ipv4-1 \ --ip-version=IPV4 \ --global

5. 예약된 IPv4 주소 확인
    - gcloud compute addresses describe lb-ipv4-1 \ --format="get(address)" \ --global
6. 부하 분산기용 상태 점검 만들기
    - gcloud compute health-checks create http http-basic-check \ --port 80
      *Google Cloud는 백엔드 인스턴스가 트래픽에 제대로 응답하는지 확인하는 상태 점검 매커니즘을 제공한다.
7. 백엔드 서비스 만들기
    - gcloud compute backend-services create web-backend-service \ --protocol=HTTP \ --port-name=http \ --health-checks=http-basic-check \ --global
8. 백엔드 서비스에 인스턴스 그룹을 백엔드에 추가하기
    - gcloud compute backend-services add-backend web-backend-service \ --instance-group=lb-backend-group \ --instance-group-zone= \ --global
9. URL 맵을 만들어 들어오는 요청을 기본 백엔드 서비스로 라우팅
    - gcloud compute url-maps create web-map-http \ --default-service web-backend-service
      *URL 맵은 백엔드 서비스 또는 백엔드 버킷으로 요청을 라우팅하는데 사용되는 Google Cloud 구성 리소스이다. 외부 HTTP(S) 부하 분산기를 사용하면 단일 URL 맵을 사용하여 URL 맵에 구성된 규칙에 따라 요청을 다른 대상으로 라우팅할 수 있다.
      1) http://exmaple.com/video
      - 위 요청은 하나의 백엔드 서비스로 이동한다.
      2) http://example.com/audio
      - 위 요청은 다른 백엔드 서비스로 이동한다.
      3) http://example.com/images
      - 위 요청은 Cloud Storage 백엔드 버킷으로 이동한다.
      4) 다른 호스트 및 경로 조합에 대한 요청은 기본 백엔드 서비스로 이동한다.
10. 대상 HTTP 프록시를 만들어 URL 맵에 요청을 라우팅한다.
    - gcloud compute target-http-proxies create http-lb-proxy \ --url-map web-map-http
11. 들어오는 요청을 프록시로 라우팅하는 전역 전달 규칙을 만든다.
    - gcloud compute forwarding-rules create http-content-rule \ --address=lb-ipv4-1\ --global \ --target-http-proxy=http-lb-proxy \ --ports=80
      *전달 규칙 및 위 IP 주소는 Google Cloud 부하 분산기의 프런트엔드 구성을 나타낸다.

인스턴스로 전송되는 트래픽 테스트

1. Google Cloud Console > Menu > Network Service > Load Balancing

2. 위에서 만든 부하 분산기 클릭 후 백엔드 섹션에서 백엔드 이름을 클릭하여 VM이 정상 상태인지 확인
   *VM이 정상이면 웹브라우저에서 http:// `load balancer IP address` 로 이동하여 부하 분산기를 테스트한다.