**Họ Tên :** Phạm Thị Hồng Linh

**MSSV:** B1809365

**CHỦ ĐỀ: CÁC CƠ CHẾ HÌNH THÀNH CHỨC NĂNG ĐÁM MÂY**

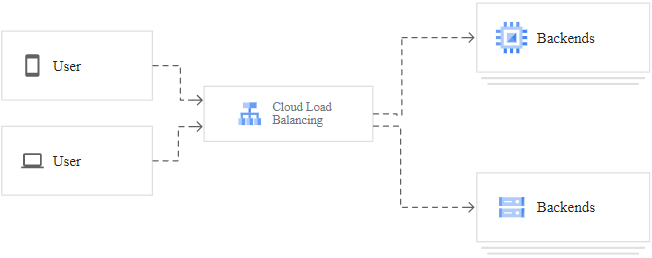
*Bài tập 1: Hãy mô tả cơ chế hoạt động của cơ chế cân bằng tải và co giãn tài nguyên của đám mây Google.*

***1/ Cơ chế cân bằng tải của đám mây Google:***

* **Giới thiệu: (**[**https://cloud.google.com/compute/docs/load-balancing-and-autoscaling**](https://cloud.google.com/compute/docs/load-balancing-and-autoscaling)**)**
* Google Cloud cung cấp tính năng cân bằng tải phía máy chủ để có thể phân phối lưu lượng truy cập đến trên nhiều phiên bản máy ảo (VM)
* Cân bằng tải Google Cloud sử dụng các tài nguyên quy tắc chuyển tiếp để phù hợp với một số loại lưu lượng nhất định và chuyển tiếp nó tới trình cân bằng tải

(Ví dụ: một quy tắc chuyển tiếp có thể khớp với lưu lượng TCP dành cho cổng 80 trên địa chỉ IP 192.0.2.1, sau đó chuyển tiếp nó đến một bộ cân bằng tải, sau đó chuyển tiếp nó đến các phiên bản VM tốt)

* Cân bằng tải Google Cloud là một dịch vụ được quản lý, có nghĩa là các thành phần của nó là dư thừa và có tính khả dụng cao. Nếu một thành phần cân bằng tải bị lỗi, nó sẽ được khởi động lại hoặc thay thế tự động và ngay lập tức
* Bộ cân bằng tải phân phối lưu lượng người dùng trên nhiều phiên bản ứng dụng của bạn. Bằng cách phân tán tải, cân bằng tải làm giảm nguy cơ ứng dụng của bạn gặp sự cố về hiệu suất.(<https://cloud.google.com/load-balancing/docs/load-balancing-overview>)



*Hình 1: Tổng quan đơn giản về cân bằng tải*

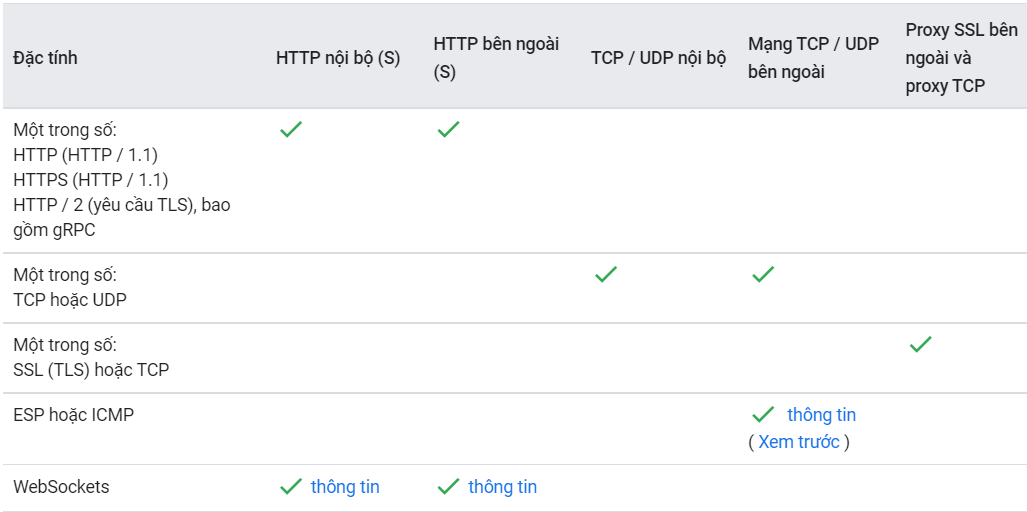
* **Các tính năng cân bằng tải được cung cấp bởi google cloud:**(<https://cloud.google.com/load-balancing/docs/load-balancing-overview>)
* Địa chỉ IP duy nhất để phục vụ như giao diện người dùng
* Tính năng tự động tính toán thông minh tự động cho các phụ trợ của bạn
* Cân bằng tải bên ngoài khi người dùng của bạn truy cập các ứng dụng của bạn từ internet
* Cân bằng tải nội bộ khi khách hàng của bạn ở bên trong Google Cloud
* Cân bằng tải theo khu vực để biết khi nào các ứng dụng của bạn khả dụng trong một khu vực
* Cân bằng tải toàn cầu để biết khi nào các ứng dụng của bạn khả dụng trên toàn thế giới
* Cân bằng tải chuyển qua (xem thêm trả về máy chủ trực tiếp (DSR) hoặc định tuyến trực tiếp )
* Cân bằng tải dựa trên proxy (như một giải pháp thay thế cho truyền qua)
* Cân bằng tải dựa trên lớp 4 để hướng lưu lượng truy cập dựa trên dữ liệu từ mạng và các giao thức lớp truyền tải như [TCP, UDP, ESP hoặc ICMP](https://cloud.google.com/load-balancing/docs/network/networklb-backend-service)
* Cân bằng tải dựa trên lớp 7 để thêm các quyết định định tuyến yêu cầu dựa trên các thuộc tính, chẳng hạn như tiêu đề HTTP và mã định danh tài nguyên thống nhất
* Tích hợp với [Cloud CDN](https://cloud.google.com/cdn) để phân phối nội dung được lưu trong bộ nhớ cache
* **Các loại cân bằng tải trên google cloud:**
* *Cân bằng tải toàn cầu so với khu vực:*
* Sử dụng **cân bằng tải toàn cầu** khi phần phụ trợ của bạn được phân phối trên nhiều khu vực, người dùng của bạn cần quyền truy cập vào các ứng dụng và nội dung giống nhau và bạn muốn cung cấp quyền truy cập bằng cách sử dụng một địa chỉ IP anycast duy nhất. Cân bằng tải toàn cầu cũng có thể cung cấp kết thúc IPv6.
* Sử dụng **cân bằng tải khu vực** khi phần phụ trợ của bạn ở một khu vực và bạn chỉ yêu cầu kết thúc IPv4.
* *Cân bằng tải bên ngoài so với bên trong:*
* **Bộ cân bằng tải bên ngoài** phân phối lưu lượng truy cập từ internet đến mạng Google Cloud Virtual Private Cloud (VPC)
* **Bộ cân bằng tải nội bộ** phân phối lưu lượng truy cập đến các phiên bản bên trong Google Cloud.

Xem thêm tại: <https://cloud.google.com/load-balancing/docs/load-balancing-overview>

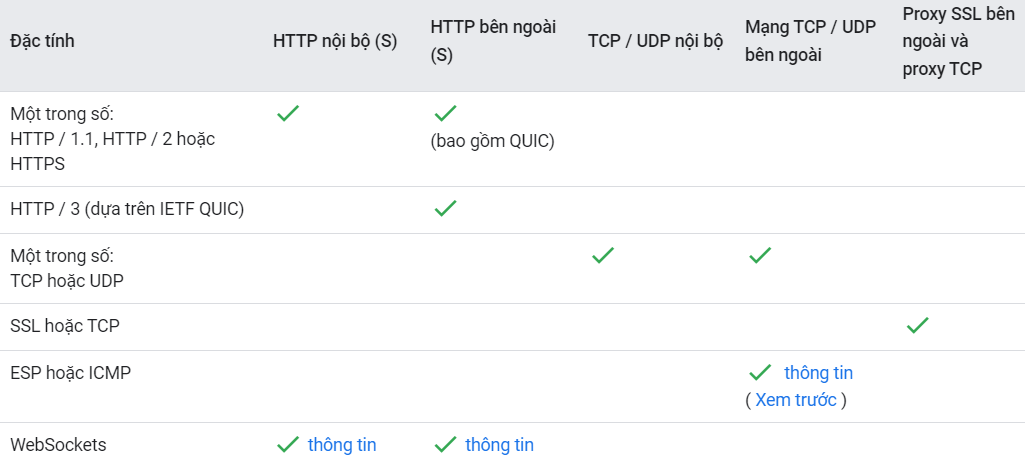
* **Các đặc trưng cân bằng tải được cung cấp bởi google cloud: (**[**https://cloud.google.com/load-balancing**](https://cloud.google.com/load-balancing)**)**
* Cân bằng tải HTTP (S)
* Ghi nhật ký đám mây
* Cân bằng tải TCP / SSL
* Tính năng tự động thay đổi tỷ lệ liền mạch
* Giảm tải SSL
* Kiểm tra sức khỏe có độ trung thực cao
* Hỗ trợ tính năng nâng cao
* Sự giống nhau
* Tích hợp CDN đám mây
* Cân bằng tải UDP
* **Các loại cân bằng tải:(**[**https://cloud.google.com/load-balancing/docs/features**](https://cloud.google.com/load-balancing/docs/features)**)**
* HTTP nội bộ (S)
* HTTP bên ngoài (S)
* TCP / UDP nội bộ
* Mạng TCP / UDP bên ngoài
* Proxy SSL bên ngoài và proxy TCP

**Giao thức từ bộ cân bằng tải đến phần phụ trợ:**

* + Phần này không áp dụng cho các bộ cân bằng tải HTTP (S) bên ngoài có phụ trợ NEG không máy chủ. Cài đặt giao thức dịch vụ phụ trợ bị bỏ qua đối với các bộ cân bằng tải này.



**Giao thức từ máy khách đến bộ cân bằng tải:**



***2/ Cơ chế co giãn tài nguyên đám mây:  (***[***https://cloud.google.com/compute/docs/load-balancing-and-autoscaling***](https://cloud.google.com/compute/docs/load-balancing-and-autoscaling)***)***

* Compute Engine cung cấp tính năng tự động thay đổi tỷ lệ để tự động thêm hoặc xóa các phiên bản VM khỏi một [nhóm phiên bản được quản lý](https://cloud.google.com/compute/docs/instance-groups#managed_instance_groups) dựa trên việc tăng hoặc giảm tải
* Tính năng tự động thay đổi tỷ lệ cho phép ứng dụng của bạn xử lý lưu lượng truy cập gia tăng một cách linh hoạt và giảm chi phí khi nhu cầu về tài nguyên thấp hơn. Sau khi bạn xác định chính sách tự động thay đổi tỷ lệ, bộ đo tỷ lệ tự động thực hiện chia tỷ lệ tự động dựa trên tải đo được
* Khi tạo một bộ đo [tự động](https://cloud.google.com/compute/docs/autoscaler) , phải chỉ định ít nhất một tín hiệu tự động định tỷ lệ. Mỗi tín hiệu có thể dựa trên việc sử dụng CPU, khả năng phục vụ cân bằng tải, số liệu giám sát đám mây hoặc lịch trình.
* Nếu [sử dụng chính sách](https://cloud.google.com/compute/docs/autoscaler/multiple-policies) tự động thay đổi tỷ lệ [với nhiều tín hiệu](https://cloud.google.com/compute/docs/autoscaler/multiple-policies) , trình tự động phân tỷ lệ một nhóm cá thể dựa trên tín hiệu cung cấp số lượng phiên bản VM lớn nhất trong nhóm