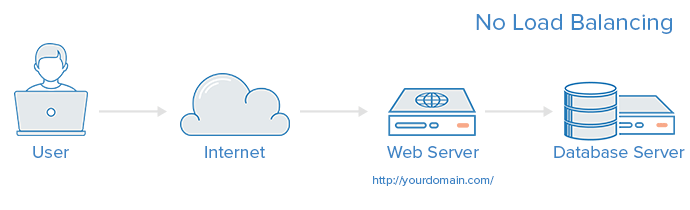
**Load Balancing – Cân bằng tải là gì và cách thức hoạt động**

**Định nghĩa:**

Load balancing cân bằng tải là một thành phần đóng vai trò quan trọng trong các cơ sở hạ tầng mạng ngày nay. Nó được dùng để cải thiện hiệu suất và độ tin cậy của các website, ứng dụng, cơ sở dữ liệu và các dịch vụ khác bằng phương thức phân phối khối lượng công việc ra nhiều cloud server cùng lúc.

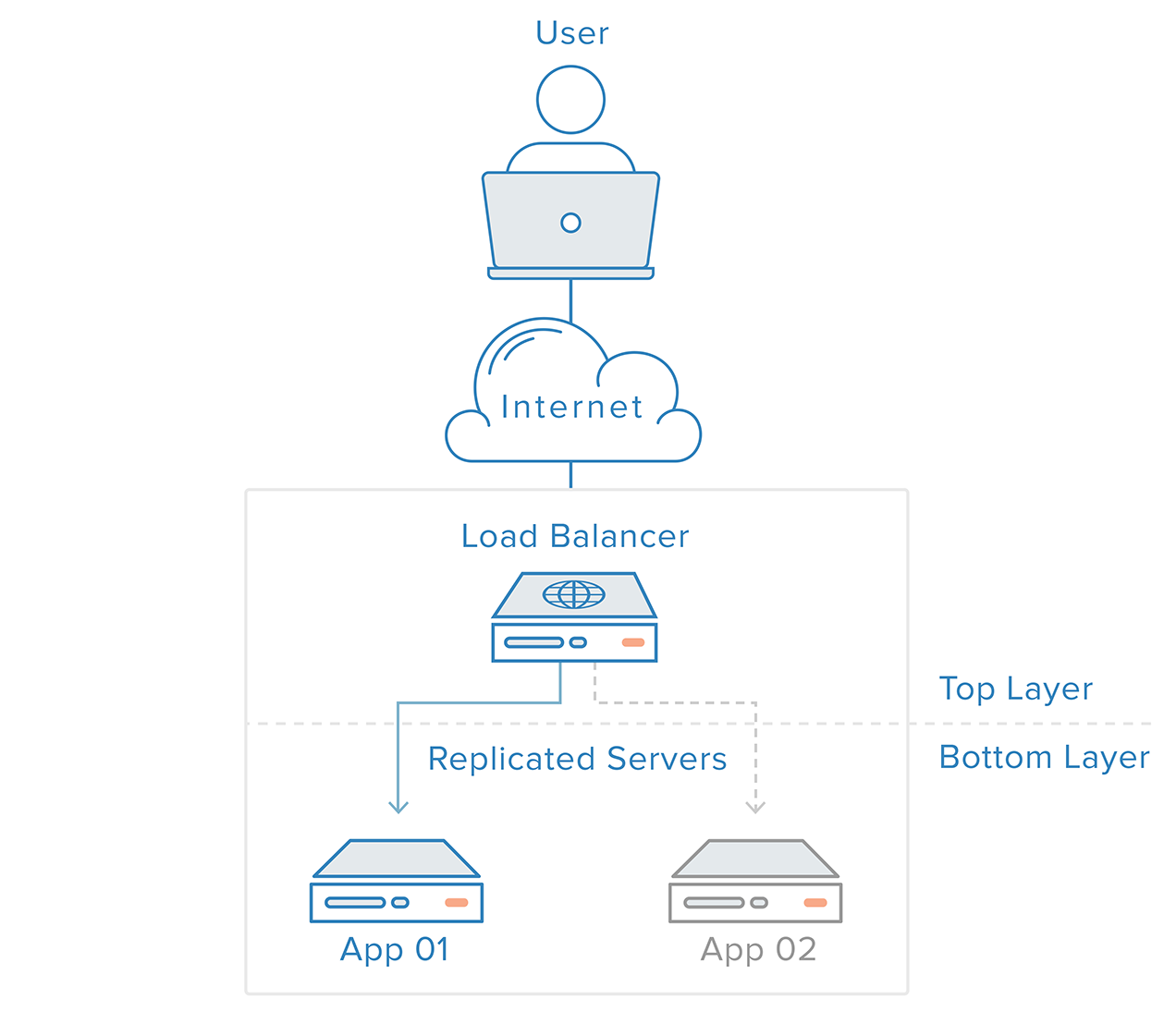
**Giới thiệu (cloud server)**

Một cơ sở hạ tầng website không có cân bằng tải sẽ trông như thế này:



Trong ví dụ này, người dùng kết nối trực tiếp đến[web server](https://vinadata.vn/tape-backup-dich-vu-luu-tru-du-phong/) (tại yourdomain.com). Nếu web server duy nhất này gặp vấn đề, người dùng sẽ không thể truy cập vào website được nữa. Ngoài ra, nếu có nhiều người dùng cố gắng truy cập vào [server](https://vinadata.vn/cdn-tang-toc-truy-cap-phan-phoi-noi-dung/) cùng một lúc thì sẽ xảy ra quá tải, khiến thời gian tải website chậm đi hoặc kết nối bị gián đoạn.

Điều này có thể khắc phục bằng cách thêm vào hệ thống cân bằng tải và ít nhất một web server hỗ trợ bổ sung trên backend. Thông thường, tất cả các nội dung mà người dùng nhận được là như nhau bất kể họ đang kết nối với server nào.



Trong ví dụ ở trên, hệ thống cân bằng tải sẽ chuyển tiếp yêu cầu của người dùng đến một server phụ trợ và đáp ứng trực tiếp yêu cầu đó. Vị trí nếu trục trặc có thể gây ảnh hưởng cho cả hệ thống bây giờ chính là vị trí cân bằng tải. Điều này có thể khắc phục bằng cách đưa vào một hệ thống cân bằng tải thứ hai, nhưng trước khi nói về điều này hãy tìm hiểu cách thức hoạt động của hệ thống cân bằng tải.

Loại giao thức dữ liệu nào cân bằng tải có thể xử lý?

Quản trị viên của hệ thống cân bằng tải có thể tạo các quy định chuyển tiếp đối với bốn loại giao thức chính:

* **HTTP** – Chuẩn cân bằng HTTP trực tiếp yêu cầu dựa trên cơ chế HTTP chuẩn. Bộ cân bằng tải đặt các tiêu đề X-Forwarded-For, X-Forwarded-Proto, và X-Forwarded-Port để cung cấp các thông tin về các yêu cầu gốc tới hệ thống backend.
* **HTTPS** – cân bằng HTTPS hoạt động tương tự như HTTP, với sự bổ sung mã hóa. Mã hóa được xử lý theo một trong hai cách: một là duy trì mã hóa với việc cho phép SSL đi từ đầu đến cuối hệ thống, hai là đặt bộ giải mã trên hệ thống cân bằng tải và gửi dữ liệu không mã hóa đến cuối hệ thống.
* **TCP** – Đối với các ứng dụng không sử dụng HTTP hoặc HTTPS, lưu lượng TCP cũng có thể được cân bằng. Chẳng hạn như lượng truy cập vào một cụm cơ sở dữ liệu có thể được mở rộng trên tất cả các máy chủ.
* **UDP** – Gần đây, một số hệ thống đã hỗ trợ cân bằng tải cho giao thức Internet cốt lõi như DNS và syslogd sử dụng UDP.

Các quy định chuyển tiếp này sẽ xác định giao thức và cổng vào trên bộ cân bằng tải, sau đó bản đồ hóa giao thức và cổng mà bộ cân bằng tải dùng để chuyển dữ liệu đến hệ thống backend.

Bộ cân bằng tải chọn server backend như thế nào?

Các hệ thống cân bằng tải sẽ chọn cloud server để chuyển tiếp yêu cầu dựa trên sự kết hợp của hai yếu tố. Đầu tiên là đảm bảo các cloud server đã chọn có thể phản hồi chính xác các yêu cầu, và sau đó là sử dụng một quy tắc được cấu hình sẵn để “nhặt” một hay nhiều cloud server trong số này.

**Kiểm tra “sức khỏe”**

Hệ thống cân bằng tải chỉ chuyển tiếp dữ liệu đến các server backend “khỏe mạnh”. Để kiểm tra sức khỏe của server backend, hệ thống sẽ thường xuyên kết nối đến backend server thông qua các giao thức và cổng thiết lập bởi quy tắc chuyển tiếp, từ đó đảm bảo server có phản hồi. Nếu server không vượt qua được bước kiểm tra này, chúng sẽ tự động bị loại bỏ khỏi hệ thống cho đến khi có thể phản hồi lại bước kiểm tra sức khỏe đó.

**Các thuật toán cân bằng tải**

Các thuật toán phổ biến để xác định cloud server nào trên backend khỏe mạnh bao gồm:

**Round Robin:** phương thức này lựa chọn server theo tuần tự. Bộ cân bằng tải sẽ chọn server đầu tiên trong danh sách của mình cho yêu cầu đầu tiên, sau đó di chuyển xuống server tiếp theo trong danh sách theo thứ tự và bắt đầu lại từ đầu khi hết danh sách.

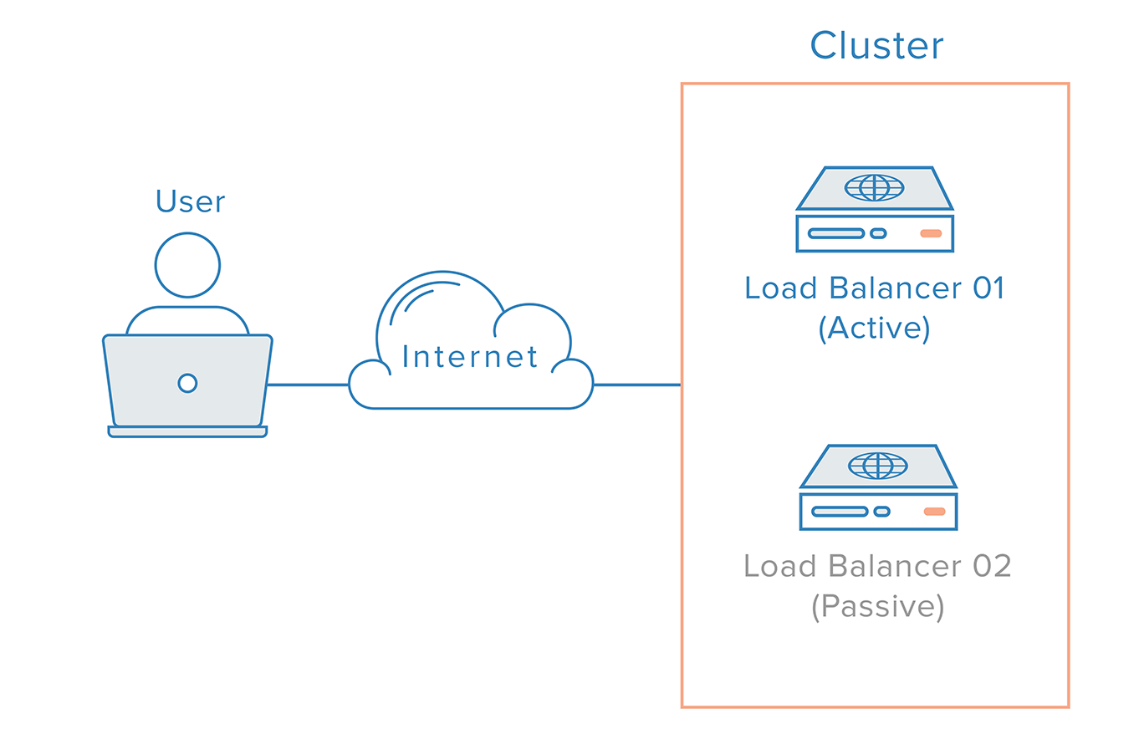
**Least Connections**: hệ thống cân bằng tải sẽ chọn server có ít kết nối nhất và cách này được khuyên dùng khi tốc độ truy cập bị chậm.

**Source**: với thuật toán Source, bộ cân bằng tải sẽ chọn server dựa trên một chuỗi các IP gốc của yêu cầu, chẳng hạn như IP của người truy cập. Phương thức này đảm bảo một người dùng cụ thể sẽ luôn kết nối với cùng một server.

Các thuật toán được cung cấp cho quản trị viên sẽ tùy thuộc vào công nghệ mà bộ cân bằng tải sử dụng.

**Cân bằng tải dự phòng**

Để ngăn chặn khả năng ảnh hưởng đến toàn hệ thống nếu bộ cân bằng tải gặp trục trặc, một bộ cân bằng tải dự phòng có thể kết nối vào để tạo thành một cụm cân bằng tải. Mỗi bộ cân bằng tải đều có thể phát hiện lỗi và phục hồi.



Trong trường hợp bộ cân bằng tải chính bị lỗi, DNS sẽ đưa người dùng đến với bộ cân bằng tải dự phòng. Để tiết kiệm thời gian khi quá trình thay đổi DNS có thể khá lâu mới thông báo được lên Internet, ngoài ra để việc chống lỗi này hoạt động tự động thì nhiều quản trị viên sử dụng hệ thống cho phép linh hoạt thay đổi IP, chẳng hạn như Floating IP.

Bằng cách cung cấp một địa chỉ IP tĩnh tùy chỉnh được khi cần thiết, các vấn đề về mất thời gian thông báo lên Internet và lưu bộ nhớ đệm khi thay đổi DNS có thể loại bỏ. Tên miền có thể duy trì liên kết với cùng một địa chỉ IP, trong khi địa chỉ IP này được di chuyển giữa các cloud server.

Một hệ thống cơ sở hạ tầng sử dụng Floating IP cao cấp sẽ giống như sau:

