fastly

Fastly Load Balancer

Content-aware load balancing for any infrastructure environment

Load balancing is critical for application availability and resiliency, yet existing solutions have been outpaced by advancements in infrastructure and service architecture. The load balancer market is crowded with a mix of appliance-based application delivery controllers (ADCs) and cloud-based solutions. ADCs are an evolution of early load balancers and are still the most prevalent despite an inability to scale elastically in real time and high maintenance and support costs. Cloud-based load balancers can offer improved performance and cost savings, but there are still a number of significant shortcomings.

Most cloud-based load balancers are built on top of DNS, which limits their ability to route traffic only by IP address. These solutions are unable to see anything in the request, so they cannot provide a single unified service for a microservices architecture. Furthermore, DNS-based solutions rely on time to live (TTL), a mechanism whereby responses from a DNS lookup are cached for a time period designated by the server. This removes immediacy and control. This lack of instant convergence is perhaps most apparent in the event of origin failure, when users could get errors waiting for TTLs expire at the DNS resolver before their requests are rerouted.

Why Fastly

Fastly's cloud-based load balancer was designed to overcome these challenges. We make load balancing decisions at Layer 7 rather than at the DNS layer, allowing us to make application-specific decisions on every request. Failover decisions are also made on every request, not just when the DNS cache expires. This facilitates immediate automated failover to a fallback server if the primary server is unavailable.

Our load balancer is built on top of the Fastly edge cloud platform, so you also get the benefits of granular control, immediate scalability, and real-time visibility. You can easily add other Fastly services to provide a unified architecture across your entire application, including core delivery, DDoS, and WAF. Our platform supports client requests over IPv6 and HTTP/2.

Content-aware routing

Unlike DNS-based solutions, Fastly balances HTTP and HTTPS requests to your servers using granular content-aware routing decisions. You can create any number of custom rules to intelligently route traffic using various request aspects such as client location, user logged-in status, device type, cookies, URL path, and HTTP headers. This allows you to better support your application architecture and optimize client responses before delivery.

Key Differentiators:

- **Content-aware routing with any number of custom rules for granular control**
- **Multi-cloud and hybrid-cloud for high availability and redundancy**
- **Direct traffic to or from servers instantly and programmatically**
- **Instantly scale to multiple Tbps to mitigate thundering herd problem**
- **Ready for use with containers and virtual instances**

Our load balancer spreads load to your servers using distribution methodologies including random, round robin, weighted round robin, and hash for sticky sessions.

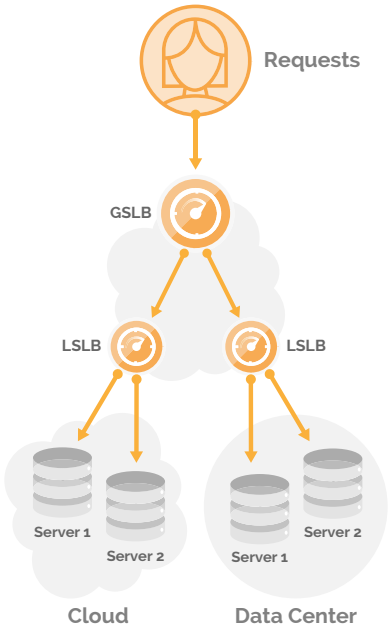
Infrastructure-agnostic distribution
Fastly efficiently manages traffic across multiple infrastructure-as-a-service providers, data centers, and hybrid-clouds. You can use Fastly as a global server load balancer (GSLB) to route traffic across any geographically distributed infrastructure deployments. We also act as a local server load balancer (LSLB) within each data center or cloud region.

Real-time control
Our dynamic server functionality allows you to programmatically add, delete, or modify your servers without having to version your VCL. You can also add, delete, or modify your custom routing rules via API. Any changes made to your routing configurations are applied globally within five seconds. This enables you to make programmatic changes to your load balancer server configuration, allowing you to integrate load balancing into your continuous integration and delivery workflow.

Instant convergence and failover ensure that requests are sent or drained immediately from your servers without waiting for TTLs to expire. While automatic HTTP-based health checks ensure that requests are only sent to servers that are healthy and responsive, you can also define whether to failover to another available server in case the primary server becomes unavailable during request handling.

Traffic scalability
Fastly's architecture enables our load balancer to instantly scale to multiple terabits per second (Tbps) for cost-effectiveness and transparency. We have no time-based scaling limitations or capacity constraints, unlike ADCs or many elastic load balancers which are susceptible to the thundering herd problem. That means that unexpectedly high request levels won't result in availability issues or performance degradation at the load balancer.

Use Fastly for global and/or local load balancing across your infrastructure.



Customer use cases

1) **Geo-based load balancing with instant auto-failover.** We allow you to implement routing rules based on geolocation. If one of your origin servers becomes unresponsive, we automatically reroute requests to any of your configured fallback servers.
2) **Data migration across infrastructure with ease.** We enable you to quickly achieve a seamless migration by simply configuring Fastly to look for content in multiple locations, thereby maintaining a high quality of experience for your users.
3) **Canary software in production with confidence.** With Fastly you can test software with a small percentage of traffic based on any aspect of the request and/or random selection. Our real-time logs provide full visibility to your traffic, and you have the flexibility to increase or decrease the percentage of traffic to the server based on your confidence level.