AWS Global Accelerator

AWS Global Accelerator is a service that improves the availability and performance of applications with local or global users.

It provides static IP addresses that act as a fixed entry point to application endpoints in a single or multiple AWS Regions, such as Application Load Balancers, Network Load Balancers or EC2 instances.

Uses the AWS global network to optimize the path from users to applications, improving the performance of TCP and UDP traffic.

AWS Global Accelerator continually monitors the health of application endpoints and will detect an unhealthy endpoint and redirect traffic to healthy endpoints in less than 1 minute.

This YouTube video from the New York summit in 2019 provides a great overview:

Details and Benefits

Uses redundant (two) static anycast IP addresses in different network zones (A and B).

The redundant pair are globally advertized.

Uses AWS Edge Locations – addresses are announced from multiple edge locations at the same time.

Addresses are associated to regional AWS resources or endpoints.

AWS Global Accelerator’s IP addresses serve as the frontend interface of applications.

Intelligent traffic distribution: Routes connections to the closest point of presence for applications.

Targets can be Amazon EC2 instances or Elastic Load Balancers (ALB and NLB).

By using the static IP addresses, you don’t need to make any client-facing changes or update DNS records as you modify or replace endpoints.

The addresses are assigned to your accelerator for as long as it exists, even if you disable the accelerator and it no longer accepts or routes traffic.

Does health checks for TCP only – not UDP.

Can assign target weight within a region to control routing and also “dial” up or down traffic to a region.

Fault tolerance:

* Has a fault-isolating design that increases the availability of your applications.
* AWS Global Accelerator allocates two IPv4 static addresses that are serviced by independent network zones.
* Similar to Availability Zones, these network zones are isolated units with their own set of physical infrastructure and service IP addresses from a unique IP subnet.
* If one IP address from a network zone becomes unavailable, due to network disruptions or IP address blocking by certain client networks, client applications can retry using the healthy static IP address from the other isolated network zone.

Global performance-based routing:

* AWS Global Accelerator uses the vast, congestion-free AWS global network to route TCP and UDP traffic to a healthy application endpoint in the closest AWS Region to the user.
* If there’s an application failure, AWS Global Accelerator provides instant failover to the next best endpoint.

Fine-grained traffic control:

* AWS Global Accelerator gives you the option to dial up or dial down traffic to a specific AWS Region by using traffic dials.
* The traffic dial lets you easily do performance testing or blue/green deployment testing for new releases across different AWS Regions, for example.
* If an endpoint fails, AWS Global Accelerator assigns user traffic to the other endpoints, to maintain high availability.
* By default, traffic dials are set to 100% across all endpoint groups so that AWS Global Accelerator can select the best endpoint for applications.

Continuous availability monitoring:

* AWS Global Accelerator continuously monitors the health of application endpoints by using TCP, HTTP, and HTTPS health checks.
* It instantly reacts to changes in the health or configuration of application endpoints, and redirects user traffic to healthy endpoints that deliver the best performance and availability to end users.

Client affinity:

* AWS Global Accelerator enables you to build applications that require maintaining state.
* For stateful applications where you need to consistently route users to the same endpoint, you can choose to direct all requests from a user to the same endpoint, regardless of the port and protocol.

Distributed denial of service (DDoS) resiliency at the edge:

* By default, AWS Global Accelerator is protected by AWS Shield Standard, which minimizes application downtime and latency from denial of service attacks by using always-on network flow monitoring and automated in-line mitigation.

You can also enable AWS Shield Advanced for automated resource-specific enhanced detection and mitigation, as well as 24×7 access to the AWS DDoS Response Team (DRT) for manual mitigations of sophisticated DDoS attacks.