Bruno Schaatsbergen

Latency Based Routing in Aws

in blog

date 11/14/2021

***

Distance affects performance. AWS offers two services to optimize routing: Global Accelerator and Route 53 latency-based routing. I'm trying to understand the differences between them and when to use each.

Global Accelerator

Global Accelerator addresses common DNS challenges[1] by bypassing reliance on IP address caches. With two static IPv4 addresses serving as single entry points for user connections, there's no need for you to manage DNS configurations.

These IPv4 addresses are hosted in separate network zones to ensure fault tolerance. Similar to Availability Zones (AZs), network zones are isolated units with distinct sets of physical infrastructure. Upon configuration, if one IPv4 address becomes unavailable due to IP blocking or network disruptions, client applications can simplfy retry using the healthy static IP address from the alternate network zone.

By leveraging AWS's globally distributed edge locations, traffic directed through Global Accelerator enters the AWS Backbone network , offering faster routing compared to traversing extensive sections of the internet.

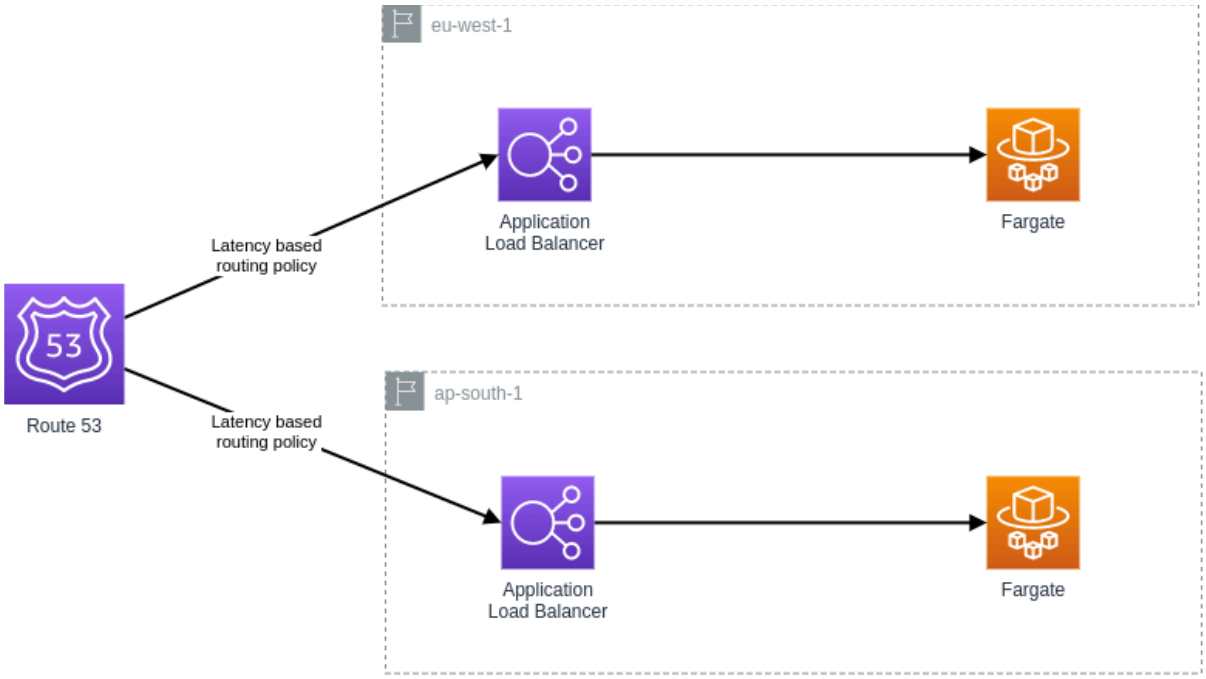
Global Accelerator utilizes Border Gateway Protocol (BGP) Anycast to efficiently route traffic across multiple paths (edge locations) towards its destination. To simplify, a BGP announcement serves as a means for routers to signal their capability and readiness to receive traffic for specific IP addresses. As traffic traverses routers, known as hops, minimizing these hops reduces latency.

To simplify, when you call the Global Accelerator endpoint, your ISP routes the traffic to the nearest AWS edge location. The edge location then routes the traffic to the nearest AWS region. This is a more efficient way to route traffic compared to the traditional method of routing traffic through the internet to the AWS region.

Route 53 latency-based routing

Route 53's latency-based routing directs users to the lowest-latency AWS endpoint available. This is commonly used in active-active architectures to balance traffic between resources in different regions. Route 53 uses DNS telemetry and network latency to return the best latency record for a given query. These latency records are continuously measured over a period of time[2].

Because of this, someone in Amsterdam would be routed to eu-west-1 , while someone in Mumbai would be routed to ap-south-1 .



How does this differ from geolocation-based routing policies?

Although both policies might direct a user in Amsterdam to eu-west-1, they prioritize different criteria. Consider a scenario where compliance mandates data storage in specific regions. Geolocation-based routing ensures adherence to such requirements, unlike latency-based routing, which relies solely on past latency measurements.

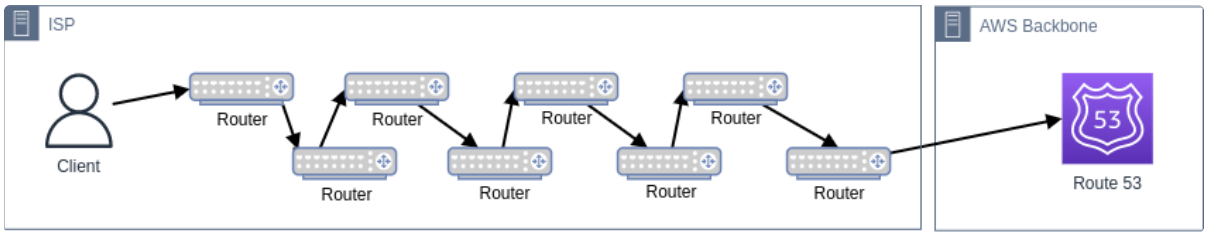
Latency-based routing dynamically adjusts routing based on recent latency measurements. Consequently, a request routed to us-east-1 this week might be directed to eu-west-1 the next.

Choosing Between Them

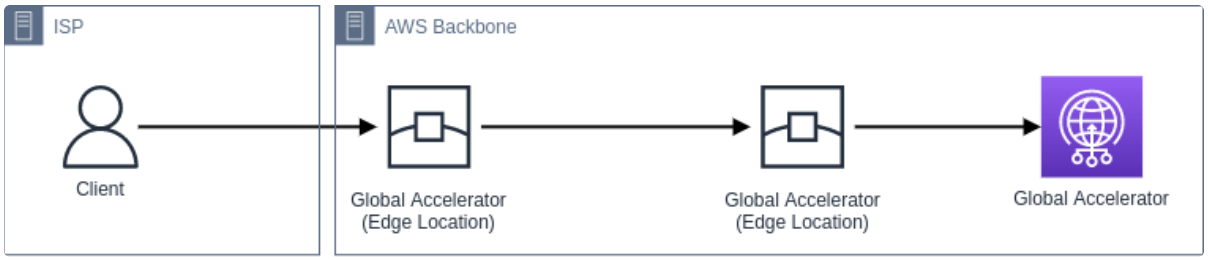
Determining the appropriate routing policy depends on various factors. Route 53 latency-based routing utilizes DNS telemetry and network latency, spending more time traversing ISP networks and the internet compared to Global Accelerator.

A mnemonic might aid in remembering the distinction:

Latency-based routing optimizes Backbone entry closest to the target backend.



Global Accelerator prioritizes Backbone entry closest to the client.



Anycast[3], used by Global Accelerator, can be slow to respond to network events such as link failures. Networks have to propagate BGP updates when conditions change. Luckily, Global Accelerator provides us with 2 static IPv4 addresses, and these are hosted in separate network zones to ensure fault tolerance. This means that if one IP address becomes unavailable, the client can retry using the healthy static IP address from the alternate network zone.

Global Accelerator performs TLS termination at edge. This is very useful as now the 3-way handshake happens at the edge location instead of the DNS server. This reduces the latency for the client.

What I like about Global Accelerator is that when you query the accelerator endpoint, you get an anycast IP address. This reduces the amount of firewall rules you would need to manage if you were to use Route 53 latency-based routing.

In contrast, Route 53 latency-based routing is easier to set up. You can combine it with other routing policies like geolocation-based routing. This is useful if you also want to route traffic based on the location of the client, for compliance reasons.

Pricing

Global Accelerator charges per GB of data transferred and the number of accelerators provisioned. Route 53 charges per query and the number of hosted zones.

Conclusion

If you serve a global audience and want to reduce latency, Global Accelerator might be the better choice. It will provide you with a consistent low latency experience for your clients.

If you have a simpler architecture with fewer regions and want more control over routing based on latency between the user and the region, Latency-Based Routing in Route 53 could be better. It's also easier to set up and manage.

Resources

ა. AWS created a tool to compare Global Accelerator to the public internet .
ბ. A nice session to learn more about Global Accelerator is this Re:Invent 2020 NET311 session .

***

🥿 Footnotes

1. DNS isn't perfect, long TTLs reduce traffic on the Internet and prevent domain name servers from overloading, while short TTLs often cause heavy loads on authoritative name servers but are useful for swift disaster recovery. Too many devices cache TTL in their own way regardless of the value set in the DNS record. ↩

2. Latency record measurements are continuously ran and usually measured over the span of a few weeks. ↩
3. What is Anycast? | How does Anycast work? . ↩