

Service Name

AWS Global Accelerator

Service Overview

AWS Global Accelerator is a networking service that improves the availability and performance of applications for both local and global users by routing traffic over the AWS global network. It provides static IP addresses that serve as fixed entry points to applications and automatically routes user traffic to optimal endpoints based on performance, health, and policies. The service utilizes a global network of 130 Points of Presence (PoPs) in 95 cities across 53 countries to accelerate application delivery and reduce latency by up to 49% and jitter by up to 58%.

Key Use Cases

- **Latency-sensitive applications:** Gaming, media, mobile apps, ad-tech, and financial applications that require reduced internet latency and jitter for optimal user experience
- **Multi-Region disaster recovery:** Detecting application endpoint failures in the primary AWS Region and instantly rerouting traffic to the next available, closest AWS Region for business continuity
- **Global application scaling:** Industries like IoT, retail, media, automotive, and healthcare where frequent client application updates are not feasible but global reach is required
- **Custom routing for specialized applications:** VoIP, online gaming, and EdTech applications that require custom logic to map users to specific endpoints while maintaining performance benefits

Features

- **Static Anycast IP Addresses:** Provides two static IPv4 addresses and two static IPv6 addresses (for dual-stack) that serve as fixed entry points to applications, with support for Bring Your Own IP (BYOIP)
- **Global Performance-Based Routing:** Routes TCP and UDP traffic to healthy application endpoints in the closest AWS Region based on user location, endpoint health, and policies
- **Health Monitoring:** Continuously monitors endpoint health using TCP, HTTP, and HTTPS health checks and automatically redirects traffic to healthy endpoints
- **Traffic Control:** Fine-grained traffic control with traffic dials to adjust percentage of traffic to specific AWS Regions and endpoint weights for load distribution

- **DDoS Protection:** Protected by AWS Shield Standard with option to enable AWS Shield Advanced for enhanced DDoS protection and 24x7 access to DDoS Response Team
- **Custom Routing Accelerators:** Allows custom application logic to route user traffic to specific Amazon EC2 instance destinations in single or multiple AWS Regions
- **Client Affinity:** Enables building applications that require maintaining state by consistently routing users to the same endpoint
- **Fault-Tolerant Design:** Two static IP addresses serviced by independent network zones with automatic failover capabilities

Value Proposition

AWS Global Accelerator provides unique advantages over alternatives by combining global network acceleration with simplified management. Unlike DNS-based solutions, it provides static IP addresses that eliminate the complexity of managing region-specific IPs and DNS updates:

- Compared to Amazon CloudFront, Global Accelerator works with non-HTTP protocols and provides deterministic fast regional failover
- It offers up to 60% performance improvement through the AWS global network, reduces first byte latency by up to 49%, and provides instant failover capabilities
- The service integrates seamlessly with existing AWS infrastructure including ELB, EC2, and VPC resources while providing built-in DDoS protection and requiring no long-term commitments

Service Integration

AWS Global Accelerator integrates deeply with core AWS services to provide comprehensive application acceleration. It works with Elastic Load Balancing (Application Load Balancers and Network Load Balancers) for regional load distribution, Amazon EC2 instances and Elastic IP addresses as direct endpoints, and Amazon VPC subnets for custom routing scenarios:

- The service integrates with Amazon Route 53 for DNS management and custom domain routing, AWS Shield for DDoS protection, Amazon CloudWatch for monitoring and logging, and AWS IAM for access control and cross-account resource sharing
- Additional integrations include Amazon S3 Multi-Region Access Points for object storage acceleration and AWS Application Recovery Controller for coordinated failover management

Onboarding and Offboarding

Getting started with AWS Global Accelerator involves three simple steps through the AWS Management Console or API. First, create an accelerator by providing a name, choosing between standard or custom routing types, and selecting IPv4 or dual-stack address types:

- Second, configure listeners to process inbound connections based on protocol and port specifications, then set up endpoint groups associated with specific AWS Regions
- Third, register endpoints such as Application Load Balancers, Network Load Balancers, EC2 instances, or Elastic IP addresses, configuring weights and health check settings
- The service must be managed from the US West (Oregon) Region, and users can optionally bring their own IP addresses (BYOIP) for custom requirements

Getting Started Guide: <https://docs.aws.amazon.com/global-accelerator/latest/dg/getting-started.html>

Data Removal

To offboard from AWS Global Accelerator, customers must first disable the accelerator through the console or API, then remove all associated listeners and endpoint groups before deletion. When an accelerator is deleted, the static IP addresses assigned to it are permanently lost and cannot be recovered. Cross-account attachments can be deleted separately, which revokes permissions for shared resources. The deletion process is irreversible, so customers should ensure proper planning before removing accelerators to avoid losing static IP configurations.

Backup/Restore and Disaster Recovery

AWS Global Accelerator itself is a disaster recovery enablement service rather than providing traditional backup capabilities. It supports disaster recovery by instantly detecting application endpoint failures and automatically rerouting traffic to healthy endpoints in alternate AWS Regions. The service enables multi-Region resiliency with automatic failover capabilities, helping customers achieve rapid recovery objectives. Configuration settings and accelerator definitions can be managed through Infrastructure as Code tools like AWS CloudFormation for consistent deployment and recovery of the service configuration itself.

Backup Capabilities

AWS Global Accelerator does not integrate with AWS Backup as it is a network routing service rather than a data storage service. The service configuration can be backed up through Infrastructure as Code practices using AWS CloudFormation templates or API scripts:

- Static IP addresses are maintained by the service as long as the accelerator exists but are lost upon deletion
- Customers should maintain configuration backups through code repositories and automation scripts to quickly recreate accelerators if needed

Disaster Recovery

AWS Global Accelerator enhances disaster recovery capabilities but does not directly integrate with AWS Elastic Disaster Recovery. Instead, it serves as a traffic management layer that works in conjunction with disaster recovery solutions by providing instant traffic rerouting to healthy endpoints:

- The service supports automated failover at the network level, complementing application-level disaster recovery solutions
- It can be used with AWS Application Recovery Controller for coordinated failover management across multiple Regions and services

Recovery Objectives

AWS Global Accelerator significantly improves Recovery Time Objectives (RTO) by providing instant traffic failover to healthy endpoints within seconds of detecting failures. The service continuously monitors endpoint health and immediately redirects traffic when issues are detected, enabling near-zero RTO for network-level failovers. For Recovery Point Objectives (RPO), Global Accelerator works with underlying services and applications to minimize data loss during failover scenarios. The service's anycast IP addresses and global routing capabilities ensure users are always directed to the nearest available healthy endpoint.

Service Constraints

Service Quotas: <https://docs.aws.amazon.com/global-accelerator/latest/dg/limits-global-accelerator.html>

FAQ: <https://aws.amazon.com/global-accelerator/faqs/>

Technical Requirements

Service Documentation: <https://docs.aws.amazon.com/global-accelerator/latest/dg/what-is-global-accelerator.html>

User Guide: <https://docs.aws.amazon.com/global-accelerator/latest/dg/what-is-global-accelerator.html>

API Reference: <https://docs.aws.amazon.com/global-accelerator/latest/api/Welcome.html>

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: <https://aws.amazon.com/global-accelerator/pricing/>

Cost Optimization

AWS Global Accelerator offers several cost optimization strategies including traffic dials to control the percentage of traffic directed to specific AWS Regions, enabling gradual traffic shifting and cost management. Customers can optimize costs by strategically placing endpoints in regions with lower data transfer costs and using traffic weights to distribute load efficiently:

- The service supports Bring Your Own IP (BYOIP) to avoid additional IP address costs
- Traffic routing optimization can reduce overall data transfer costs by directing users to the nearest endpoints, and the pay-as-you-go model with no long-term commitments allows for flexible cost management based on actual usage patterns

Savings Considerations

Cost savings with AWS Global Accelerator come from improved application performance leading to better user experience and potentially reduced infrastructure needs in multiple regions. The service eliminates the need for complex DNS-based traffic management solutions and reduces operational overhead:

- By using traffic dials for blue/green deployments and A/B testing, customers can optimize resource utilization and avoid over-provisioning
- The fixed hourly fee model (\$0.025 per hour) combined with usage-based Data Transfer-Premium charges allows predictable cost planning
- Performance improvements can lead to reduced compute requirements and better resource utilization across endpoints, potentially offsetting the service costs through overall infrastructure efficiency gains