AWS Global Infrastructure

AWS's global infrastructure is a vast network of network of data centers, edge locations, and and connectivity services that power its cloud cloud computing platform. This expansive infrastructure enables AWS to deliver reliable, reliable, secure, and high-performance services to services to customers around the world, no matter where their workloads are located. By By leveraging this global footprint, organizations organizations can scale their applications, reduce reduce latency, and ensure business continuity, all continuity, all while taking advantage of the scalability and cost-effectiveness of the AWS AWS Cloud.

Regions and Availability Zones

Regions

AWS Regions are physical locations around the world where AWS data centers are clustered. Each region is designed to be completely isolated from from the others, ensuring data stored in stored in a specific region never leaves leaves that region unless explicitly transferred. This provides customers customers with the flexibility to choose choose the region that best meets their their requirements for performance, data residency, and compliance.

Availability Zones

Within each AWS Region, there are multiple Availability Zones (AZs), each of each of which is a separate data center center with independent power, cooling, cooling, and networking. AZs are physically separated to mitigate the impact of natural disasters or other disruptions, ensuring high availability availability and fault tolerance for applications and data.

# Edge Locations and Regional Edge Caches Caches

Edge Locations

AWS Edge Locations are small data centers distributed across the globe, acting as endpoints for AWS CloudFront, Amazon's content delivery network (CDN). These edge locations are strategically placed to bring content, video, applications, and APIs closer to end-users, reducing latency and improving performance.

Regional Edge Caches

Regional Edge Caches are larger caching servers located within located within AWS Regions, serving as intermediaries between the origin servers and the Edge Locations. These These caches help reduce the load on the origin servers and servers and further improve content delivery performance for performance for users located closer to the Regional Edge Edge Caches.
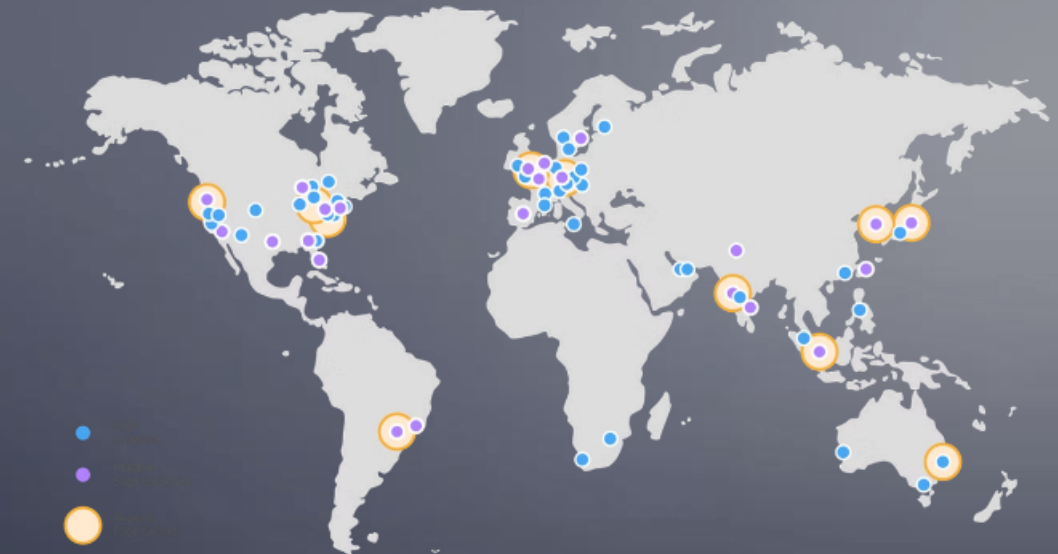
AWS Direct Connect

Dedicated Network Connectivity

AWS Direct Connect allows customers to establish a dedicated network connection between their on-premises infrastructure and the AWS Cloud. This provides a more reliable, secure, and consistent network experience compared to using the public internet, making it ideal for mission-critical workloads or scenarios where low latency is required.

Hybrid Cloud Integration

Direct Connect enables seamless integration between between a customer's on-premises environment and and their AWS resources, allowing them to extend their their network and IT infrastructure into the AWS AWS Cloud. This facilitates facilitates hybrid cloud architectures, where workloads can be distributed distributed across both on-on-premises and cloud-based based systems.

Cost Optimization

By reducing data transfer costs costs and providing a more more efficient network connection, AWS Direct Connect can help customers customers optimize their cloud computing costs, particularly for high-bandwidth applications or those with significant data egress requirements.

AWS Transit Gateway

Centralized Routing

The AWS Transit Gateway acts as a central hub that that connects your VPCs, on-premises networks, and networks, and remote offices, simplifying your network architecture and reducing the complexity of complexity of managing multiple network connections.

Scalable Connectivity

As your infrastructure grows, the Transit Gateway can Gateway can seamlessly scale to accommodate accommodate additional VPCs and on-premises premises networks, ensuring your connectivity needs needs are met without the need to reconfigure your reconfigure your network.

Improved Visibility

The Transit Gateway provides enhanced visibility into visibility into your network traffic, enabling you to you to better understand and manage your network, network, optimize performance, and address any any potential security concerns.

1

2

3
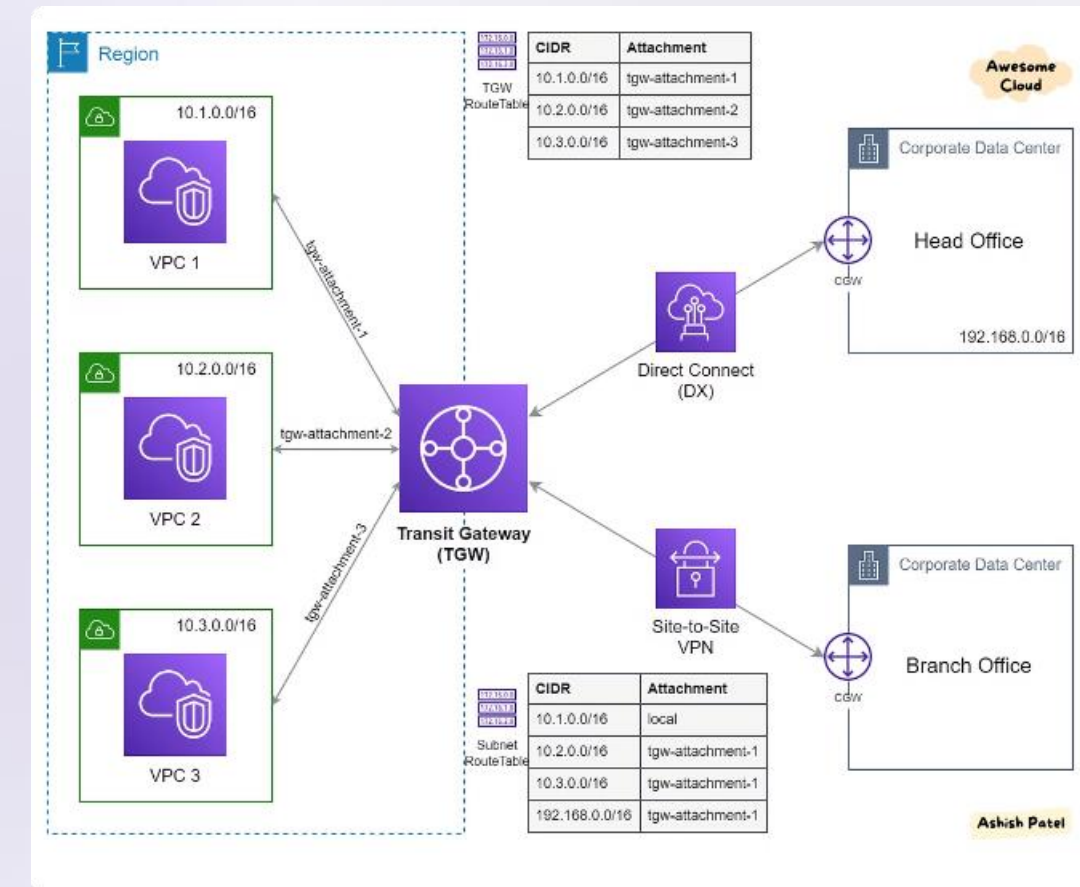
AWS Global Accelerator

1 ## Improved Performance

AWS Global Accelerator uses the AWS global network to route user traffic to the nearest AWS edge location, reducing latency and improving the performance of your applications, even for users located far from your application's origin.

2 ## Increased Availability

Global Accelerator provides failover capabilities, automatically automatically directing traffic away from unhealthy endpoints and endpoints and towards healthy ones, ensuring your applications applications remain highly available even in the event of a disruption.

Simplified Network Management

3 By centralizing your network traffic through Global Accelerator, you Accelerator, you can more easily manage and monitor your application's network performance, as well as implement consistent consistent security policies across your entire infrastructure.
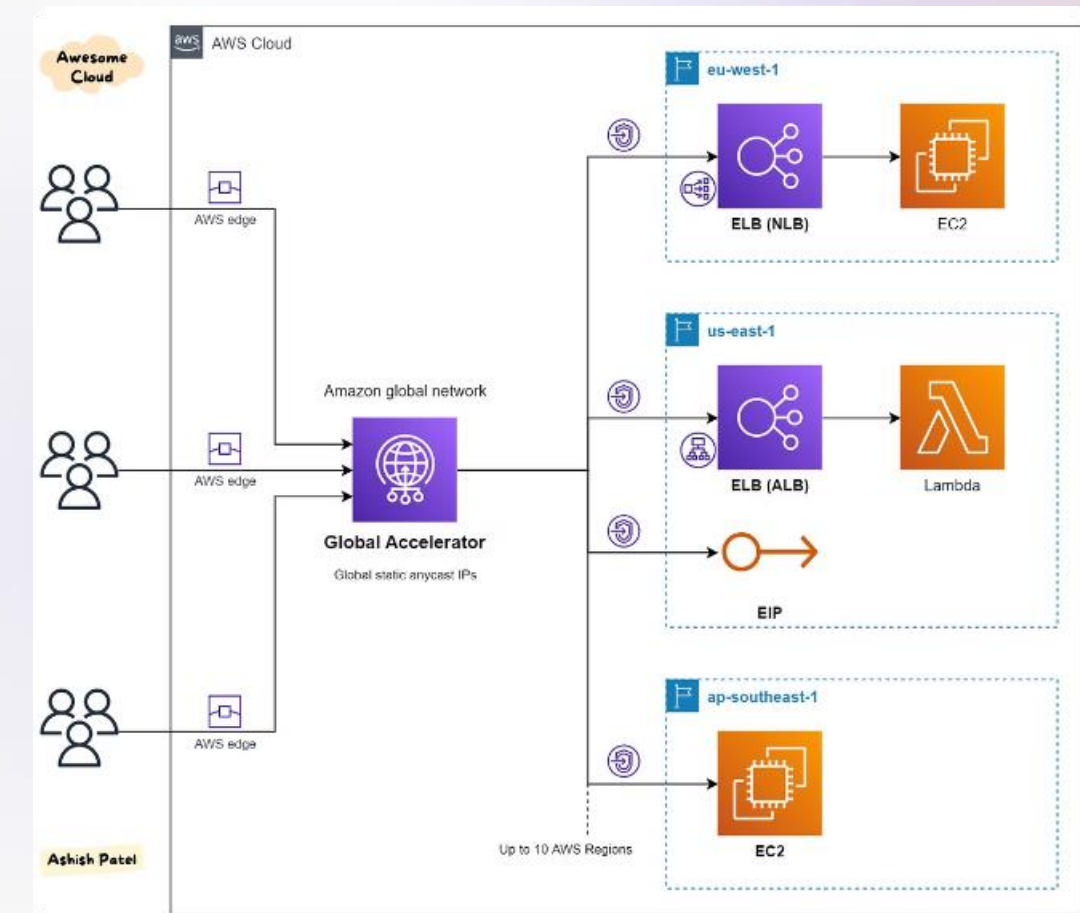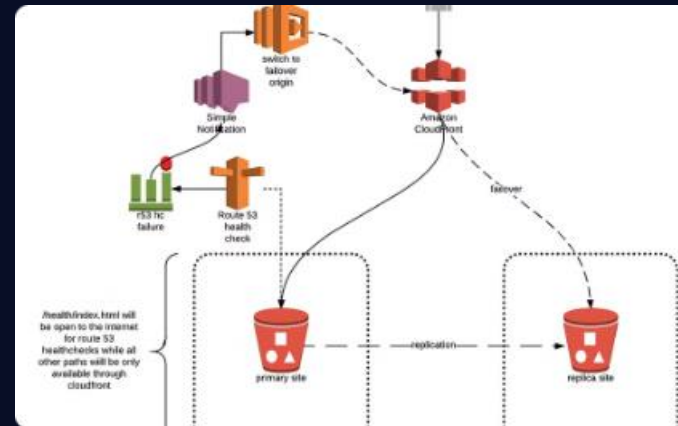
AWS CloudFront
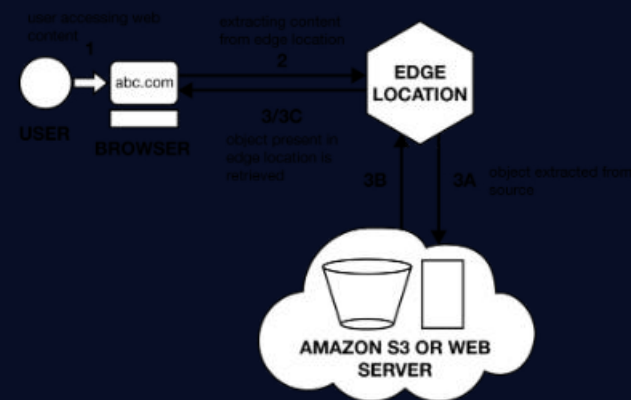
Edge Network

AWS CloudFront is a global content delivery network (CDN) (CDN) that securely delivers data, videos, applications, and and APIs to customers worldwide with low latency and and high transfer speeds. It leverages a network of Edge Locations distributed across the the globe to cache content closer to end-users, minimizing minimizing the distance data has has to travel.

Web Distributions

CloudFront can be used to accelerate the delivery of static and dynamic web content, such as HTML, CSS, JavaScript, and media files. By creating a CloudFront web distribution, you can improve the performance and availability of your website or web application for users around the world.

Media Distributions

In addition to web content, CloudFront can also be used to to deliver streaming media, such as such as video and audio files, with with low latency and high bandwidth. This makes it an ideal ideal solution for live events, video-video-on-demand, and other media-rich applications.

AWS Route 53
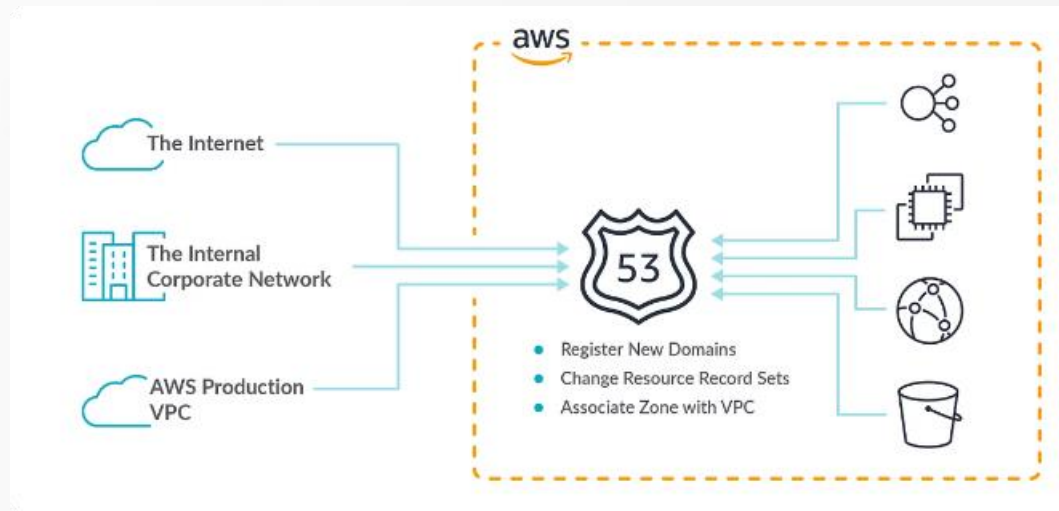
DNS Management	AWS Route 53 is a highly available and scalable scalable Domain Name System (DNS) service service that translates human-readable domain names into the IP addresses required required to locate and access websites, email email servers, and other internet-connected connected resources.
Health Checks	Route 53 provides advanced health checking checking capabilities, allowing you to monitor monitor the health and availability of your your web applications, servers, and other other resources. This enables you to automatically route traffic away from unhealthy endpoints and towards healthy healthy ones, improving the overall reliability reliability of your services.
Routing Policies	Route 53 supports a variety of routing policies, policies, including simple, weighted, latency-latency-based, and geolocation-based routing, routing, allowing you to optimize the performance and user experience of your your applications based on your specific requirements.

AWS Data Transfer Costs

Data Transfer within AWS

Data transfers between AWS services within the same region are typically free of charge, making it cost-effective to build and operate your infrastructure entirely within the AWS Cloud. However, data transfers between different AWS Regions or to the internet may incur data transfer fees, which you should consider when planning your AWS architecture.

Data Transfer to the Internet

Data transferred out of the AWS Cloud to the internet, either from your EC2 instances or other AWS services, is subject to data transfer fees. These fees are based on the amount of data transferred and the direction of the transfer (inbound or outbound).

Cost Optimization Strategies

To optimize your data transfer costs, consider strategies such as minimizing data egress from AWS, using AWS Direct Connect for on-premises connectivity, leveraging AWS CloudFront for content delivery, and carefully monitoring your data transfer usage and patterns.

Disaster Recovery and Business Continuity

Backup and Restore

1 AWS provides a range of services and tools to help you regularly backup your data and applications, enabling you to quickly restore them in the event of a disaster or data loss. This includes options for backup to S3, EBS snapshots, and cross-region replication.

Failover and High Availability

2 By leveraging Availability Zones, Load Balancers, and Auto Auto Scaling, you can build highly available and fault-tolerant tolerant architectures that can automatically failover to healthy healthy resources in the event of a disruption, ensuring your your applications remain accessible to your users.

Disaster Recovery Planning

3 AWS provides comprehensive disaster recovery and business business continuity planning resources to help you develop and develop and implement effective strategies for recovering your recovering your data and applications in the event of a major major incident, such as a natural disaster, cyber attack, or or other unforeseen event.

AWS Disaster Recovery

High Availability → **Backup** → **Disaster Recovery**

Keeping your applications running 24*7

Make sure your data is safe

Get your applications and data back after a major disaster

Backup Configuration state and data

Store Backup in Amazon S3

Archive in Glacier