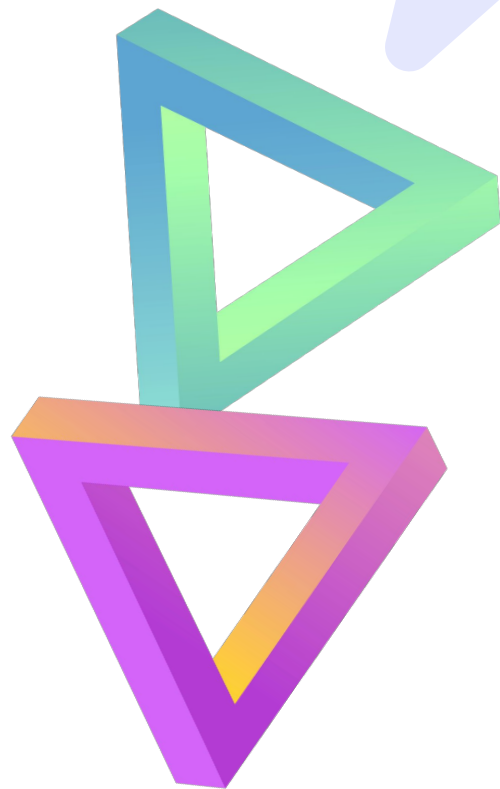




AWS

LOAD BALANCING


By Vignesh S





Three-Tier Architecture

It's an architectural pattern that divides an application into three separate layers or tiers: presentation, business logic, and data storage. This separation helps in achieving modularity, scalability, and maintainability.

- **Presentation Tier**
 - **Application (Business Logic) Tier**
 - **Data Tier**
- 



Presentation Tier

Also known as the user interface (UI) tier, this is where user interactions occur.

It includes components that handle user input, display data, and provide a visual interface for users to interact with the application.

Examples include web browsers, mobile apps, and desktop interfaces.






Application (Business Logic) Tier

This tier contains the core functionality of the application.

It processes user requests, performs business logic, and manages data flow between the presentation and data tiers.

Here, you'd find components responsible for authentication, authorization, data validation, and other application-specific logic. "Simply a program for your application"





Data Tier

The data tier, also referred to as the persistence tier, is responsible for storing and managing data.

It includes databases or any other storage systems that hold the application's data.

This tier handles data retrieval, storage, and management operations.






Cluster

A cluster is a group of interconnected computers or servers that work together to provide a unified computing resource. Clusters are designed to enhance performance, availability, and scalability of applications and services.

Types of Clusters

High-Performance Computing (HPC) Clusters: These clusters are optimized for performing complex computations that require significant processing power, such as scientific simulations and research.






Types of Clusters

High-Availability (HA) Clusters: HA clusters are focused on minimizing downtime and ensuring uninterrupted service by duplicating resources across multiple servers. If one server fails, another takes over seamlessly.

Load Balancing Clusters: These clusters distribute incoming network traffic across multiple servers to prevent overload on a single server, thereby improving response times and overall performance.



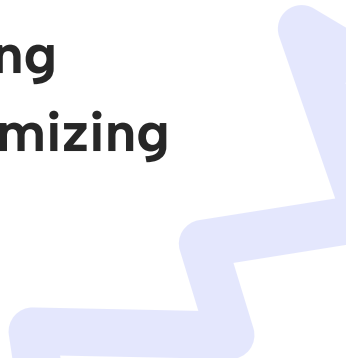


Load Balancing:

Optimizing Resource Utilization

Load balancing is a critical technique used to evenly distribute incoming network traffic across multiple servers or resources.

The main goal is to prevent any single server from being overwhelmed, ensuring optimal performance and minimizing the risk of downtime.





Types of Load Balancing

Layer 4 Load Balancing (Transport Layer):

This type of load balancing operates at the transport layer (TCP/UDP) and makes routing decisions based on IP addresses and port numbers.

It's efficient but doesn't inspect the content of the traffic.






Types of Load Balancing

Layer 7 Load Balancing (Application Layer):

Operating at the application layer, this type can make routing decisions based on content, such as HTTP headers, providing more advanced load balancing and even optimization for specific applications allowing for more sophisticated decision-making compared to lower-layer load balancing.





WHICH ?

The choice between application and transport layer load balancing depends on your specific use case and requirements.

If you need to route traffic based on application-specific attributes, manage session persistence, and perform advanced content-based routing, application load balancing is the better choice.

On the other hand, if you simply need to distribute traffic efficiently across servers without inspecting content, transport layer load balancing is more suitable.





Load Balancing Algorithms

Round Robin: Traffic is distributed evenly in a cyclic manner to each server in the cluster.

Least Connections: Traffic is directed to the server with the fewest active connections, ensuring even distribution of load.

IP Hash: The server is chosen based on a hash of the client's IP address, ensuring the same client is consistently directed to the same server.

Weighted Round Robin/Weighted Least Connections: Servers are assigned different weights, impacting how much traffic each server receives.






WHY ?

Improved Performance:

Load balancers prevent a single server from becoming overwhelmed with incoming traffic, leading to better response times and reduced latency for users.

High Availability:

By automatically redirecting traffic from failed or problematic servers to healthy ones, load balancers minimize downtime and ensure consistent availability of applications.





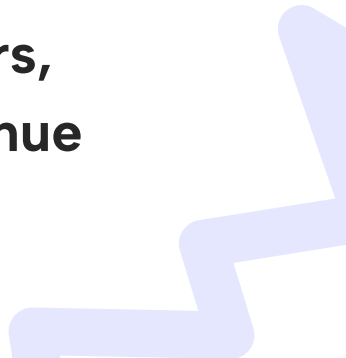
WHY ?

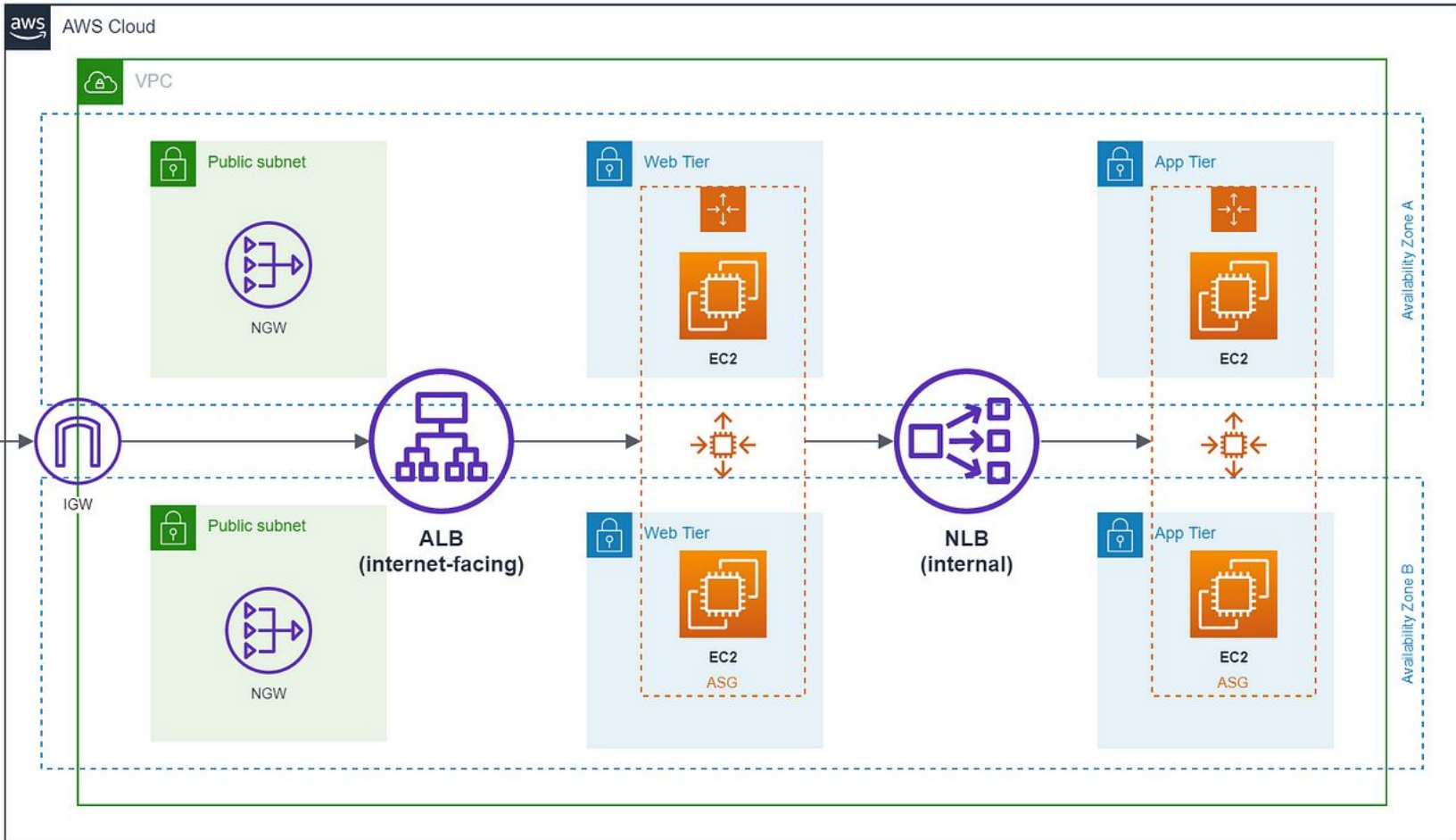
Scalability:

As applications grow, load balancers distribute increased traffic to new instances or resources, facilitating horizontal scaling and accommodating higher demand.

Redundancy:

Load balancers distribute traffic across multiple servers, ensuring that even if one server fails, others can continue handling traffic, maintaining uninterrupted service.







AWS ELB

Classic Load Balancer (CLB)

The Classic Load Balancer is the legacy option that provides basic load balancing across multiple EC2 instances.

While it lacks some advanced features, it remains a straightforward choice for distributing traffic.





AWS ELB

Application Load Balancer (ALB)

ALB operates at Layer 7 (application layer) and excels at routing HTTP/HTTPS traffic. It supports features like content-based routing, path-based routing, and host-based routing.

ALB is ideal for applications with multiple microservices.






AWS ELB

Network Load Balancer (NLB)

NLB operates at Layer 4 (transport layer) and is optimized for handling TCP/UDP traffic.

It's suitable for scenarios requiring low latency and high performance, such as gaming or media streaming.





Conclusion

load balancing ensures high performance, availability, and scalability for applications and services by distributing incoming traffic across multiple servers.

Different load balancing algorithms provide flexibility and customization options, allowing organizations to optimize resource utilization according to their specific needs.

By incorporating load balancing into clusters, businesses can deliver reliable and efficient services to their users while maintaining a seamless user experience.

