# AWS NAT

So far we have learned the fundamentals of VPC and accessing an EC2 instance deployed on VPC. NAT Gateway is a AWS service that helps you to access the internet which instances are configured in private subnet. Without proper routing, no one can access the instance from outside.

NAT is a process in which one or more local IP address are translated into Global IP address. Therefore, it is called Network address translation. It also does vice versa. It also does the translation of port numbers i.e, masks the port number of the host with another port number in the packet, that will be routed to the destination. It then makes the corresponding entries of IP address and port number in the NAT table.

There are two types of AWS NAT

① Public NAT :→ Gateway that resides in public subnet. → To access internet from instance residing in private subnet of VPC

② Private NAT :

→ used for communication between VPCs, or VPCs and transit gateway.

NAT is used for Outbound Internet access.
For inbound Internet access AWS API is used.

. Inbound Traffic:

- A customer accessing a company's online store.
- A remote employee logging into the company's internal network via VPN.
- A third-party service sending data to an API hosted within the network.

Outbound Traffic:

- An employee checking their personal email on a webmail service.
- An internal server fetching updates from a software vendor's site.
- An application within the network making a request to an external API.

-

# NAT Usecases

① Outbound Internet Access:
Instances in VPC's subnet can access internet by requesting NAT Gateway which will route traffic to internet.

② Secure connection among external repositories and instances in private subnets.

③ SaaS application resides on public internet with help of NAT Gateway.

④ Hybrid cloud deployment: on premise data centers connect to nearby AZs or LZs.

# Load balancing in Cloud Computing

Last Updated : 19 Sep, 2023

Load balancing is an essential technique used in cloud computing to optimize resource utilization and ensure that no single resource is overburdened with traffic. It is a process of distributing workloads across multiple computing resources, such as servers, virtual machines, or containers, to achieve better performance, availability, and scalability.

1. In cloud computing, load balancing can be implemented at various levels, including the network layer, application layer, and database layer. The most common load balancing techniques used in cloud computing are:
2. Network Load Balancing: This technique is used to balance the network traffic across multiple servers or instances. It is implemented at the network layer and ensures that the incoming traffic is distributed evenly across the available servers.
3. Application Load Balancing: This technique is used to balance the workload across multiple instances of an application. It is implemented at the application layer and ensures that each instance receives an equal share of the incoming requests.
4. Database Load Balancing: This technique is used to balance the workload across multiple database servers. It is implemented at the database layer and ensures that the incoming queries are distributed evenly across the available database servers.

Load balancing helps to improve the overall performance and reliability of cloud-based applications by ensuring that resources are used efficiently and that there is no single point of failure. It also helps to scale applications on demand and provides high availability and fault tolerance to handle spikes in traffic or server failures.

Sure, here are some advantages and disadvantages of load balancing in cloud computing:

**Advantages:**

1. Improved Performance: Load balancing helps to distribute the workload across multiple resources, which reduces the load on each resource and improves the overall performance of the system.
2. High Availability: Load balancing ensures that there is no single point of failure in the system, which provides high availability and fault tolerance to handle server failures.
3. Scalability: Load balancing makes it easier to scale resources up or down as needed, which helps to handle spikes in traffic or changes in demand.
4. Efficient Resource Utilization: Load balancing ensures that resources are used efficiently, which reduces wastage and helps to optimize costs.

**Disadvantages:**

1. Complexity: Implementing load balancing in cloud computing can be complex, especially when dealing with large-scale systems. It requires careful planning and configuration to ensure that it works effectively.
2. Cost: Implementing load balancing can add to the overall cost of cloud computing, especially when using specialized hardware or software.
3. Single Point of Failure: While load balancing helps to reduce the risk of a single point of failure, it can also become a single point of failure if not implemented correctly.
4. Security: Load balancing can introduce security risks if not implemented correctly, such as allowing unauthorized access or exposing sensitive data.

Overall, the benefits of load balancing in cloud computing outweigh the disadvantages, as it helps to improve performance, availability, scalability, and resource utilization. However, it is important to carefully plan and implement load balancing to ensure that it works effectively and does not introduce additional risks.

Cloud load balancing is defined as the method of splitting workloads and computing properties in a cloud computing. It enables enterprise to manage workload demands or application demands by distributing resources among numerous computers, networks or servers. Cloud load balancing includes holding the circulation of workload traffic and demands that exist over the Internet. As the traffic on the internet growing rapidly, which is about 100% annually of the present traffic. Hence, the workload on the server growing so fast which leads to the overloading of servers mainly for popular web server. There are two elementary solutions to overcome the problem of overloading on the servers-

- First is a single-server solution in which the server is upgraded to a higher performance server. However, the new server may also be overloaded soon, demanding another upgrade. Moreover, the upgrading process is arduous and expensive.
- Second is a multiple-server solution in which a scalable service system on a cluster of servers is built. That's why it is more cost effective as well as more scalable to build a server cluster system for network services.

Load balancing is beneficial with almost any type of service, like HTTP, SMTP, DNS, FTP, and POP/IMAP. It also rises reliability through redundancy. The balancing service is provided by a dedicated hardware device or program. Cloud-based servers farms can attain more precise scalability and availability using server load balancing. **Load balancing solutions can be categorized into two types –**

1. **Software-based load balancers:** Software-based load balancers run on standard hardware (desktop, PCs) and standard operating systems.
2. **Hardware-based load balancer:** Hardware-based load balancers are dedicated boxes which include Application Specific Integrated Circuits (ASICs) adapted for a particular use. ASICs allows high speed promoting of network traffic and are frequently used for transport-level load balancing because hardware-based load balancing is faster in comparison to software solution.

**Major Examples of Load Balancers –**

1. **Direct Routing Requesting Dispatching Technique:** This approach of request dispatching is like to the one implemented in IBM's Net Dispatcher. A real server and load balancer share the virtual IP address. In this, load balancer takes an interface constructed with the virtual IP address that accepts request packets and it directly routes the packet to the selected servers.
2. **Dispatcher-Based Load Balancing Cluster:** A dispatcher does smart load balancing by utilizing server availability, workload, capability and other user-defined criteria to regulate where to send a TCP/IP request. The dispatcher module of a load balancer can split HTTP requests among various nodes in a cluster. The dispatcher splits the load among many servers in a cluster so the services of various nodes seem like a virtual service on an only IP address; consumers interrelate as if it were a solo server, without having an information about the back-end infrastructure.
3. **Linux Virtual Load Balancer:** It is an opensource enhanced load balancing solution used to build extremely scalable and extremely available network services such as HTTP, POP3, FTP, SMTP, media and caching and Voice Over Internet Protocol (VoIP). It is simple and powerful product made for load balancing and fail-over. The load balancer itself is the primary entry point of server cluster systems and can execute Internet Protocol Virtual Server (IPVS), which implements transport-layer load balancing in the Linux kernel also known as Layer-4 switching.