Auto Scaling and Load Balancing

Auto Scaling

Auto scaling is a feature that allows the application to increase or decrease the resources, like servers, compute instances, as per need. It provides sufficient resources for optimal performance of a task, thus meeting performance goals as well as cost efficiency.

Benefits

- **Dynamic scaling**: it does not require any manual intervention.
- **Pay for use**: as you only pay for the resources that are in use, auto-scaling increases cost efficiency.
- **Automatic performance maintenance**: as auto-scaling increases or decreases resources based on workload, resources available for high performance is always the consideration.

Components

The key components of auto-scaling are as follows:

- **Groups:** Multiple EC2 instances are grouped together as a single logical entity in order to scale and manage the resources. The maximum and minimum number of EC2 can be predefined based on the incoming workload.
- **Configuration templates:** also known as launch template. It allows specifying Amazon Machine ID, key-pair, security group etc.
- **Scaling options:**
    - **Dynamic scaling:** adapts to changing environment and scales resources automatically.
    - **Predictive scaling:** pre-defines the number of EC2 resources that might be needed based on the predictive incoming traffic.
    - **Scheduled scaling:** increase or decrease resources based on scheduled time.

Limitations

- Maximum number of instances = 500 per auto-scaling group.
- Instant health checks: if an instance health check fails, auto-scaling will terminate it and launch another one which is time consuming and effects application availability.
- Scaling policy: it allows setting of scaling policies based on cloudwatch metrics but they are complicated to understand and might not always lead to optimal scaling.
- Application dependencies affect auto-scaling.

Load Balancing

Load balancing is an essential technique that distributes incoming traffic to the multiple computing instances. It ensures optimal use of resources resulting in high performance. It diverts the traffic to healthy instances thus providing reliability to cloud computing. Based on implementation, it is of following types:

- **Classic Load Balancers:** It directs traffic between instances but does not support path-based routing and host-based routing. It might reduce efficiency. It is situated between transport layer and application layer.
- **Application Load Balancer:** It is situated in application layer. It supports both path-based routing and host-based routing. It also supports dynamic host port mapping.
- **Network Load Balancer:** It is located in transport layer. It is mainly used for maintaining TCP traffic.
- **Gateway Load Balancer:** It provides the facility to deploy, scale and manage virtual applications like firewall. It combines the transport level gateway and then distributes the traffic.