# Automatic server scaling through Auto-Scaling

## Overview

For system implementation, the design needs to take consideration of the system performance, including the maximum number of concurrent users and peak resource usage.

In some cases, however, it is difficult to predict a sufficient system capacity for the service, or to reflect the capacity estimation for a sudden usage spike in advance due to high cost.

In such cases, by using the **VM Auto-Scaling** function of SDS Cloud, you can operate a cost-effective infrastructure service environment while preparing for unexpected loads or temporary increases in requests. Connect with a self-monitoring function, flexibly expand servers in response to demand change, and enjoy an uninterrupted expansion and smooth load balancing effect with **Load Balancer**.

## Architecture Diagram

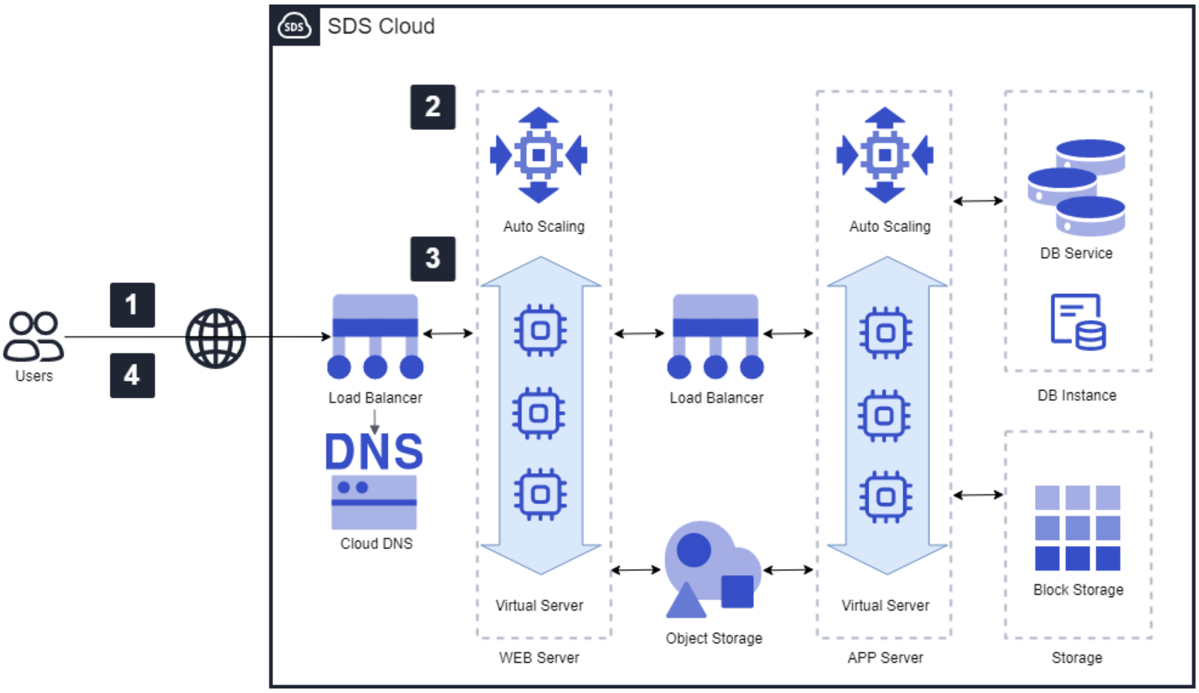

Figure 1. Automatic server scaling through VM Auto-Scaling

1. An auto-scaling group is created and operated only as much as the initial resources for a small load.
2. When the service request volume increases and the system utilization reaches the thresholds in terms of CPU utilization, memory utilization, network throughput, and Disk IOPS, which are set by the operator, the Auto-Scaling instance provisions additional virtual servers according to the predefined unit.
3. The IP address of the newly launched virtual server instance is added to Load Balancer.
4. Users can continue to use service regardless of system scaling in/out.

## Use Cases

A. Building an online shopping site

In a system with large-scale events/traffic predicted during specific times, such as year-end or holidays, real-time usage monitoring and expansion can be applied to flexibly respond to the changing demand.

B. Front-end system in media service

Social media service is sensitive to traffic increases and decreases due to a surge in rush hour traffic or irregular traffic spikes caused by certain events. In this case, stable service and performance can be ensured by setting the appropriate instance type and scale-in/out policy with auto-scaling.

## Pre-requisites

When creating an auto-scaled instance, a **Load Balancer** must be created in advance.

## Limitations

The targets of **VM Auto-Scaling** are limited to **Virtual Server** resources and do not extend to other offerings such as Bare Metal Server. For container products, you can set up your own Auto-Scaling service.

## Considerations

### A. Select instance type

The potential demand for auto-scaling needs to be considered from the initial stage of designing a system as the product created with a general Virtual Server cannot later be converted into a VM Auto-Scaling service.

The threshold value and expansion unit also need to be properly adjusted by monitoring the load situation for optimal resource use and billing plan, taken into account the default instance types.

### B. Server/application environment

The IP address added during the expansion of an instance is added to and serviced through Load Balancer. In this regard, it is necessary to ensure that the application to be used in the Auto-Scaling Group or the front end provides a communication method through the representative VIP (Virtual IP).

The monitoring agent is installed when the server is provisioned for monitoring auto-scaling threshold resource utilization. In order to collect performance data, the monitoring process must be kept up and running.

File Storage is not supported for virtual servers created by VM Auto-Scaling. If files need to be shared with Auto-Scaling Group servers, a file sharing method using Object Storage is recommended.

### C. Charging

The product itself is free of charge, but the resources such as Virtual Server and Block Storage, which are created through VM Auto-Scaling, are charged at a pre-determined rate for each resource.

You can reserve the IP address you want to use in advance and use it when scaling out but there may be an additional charge for using a reserved IP address.

### D. Compliance

When a commercial application is started and used on the Virtual Server in the Auto-Scaling Group, a license issue may occur due to the flexible change of the number of servers and the corresponding core. It is necessary to check the matters related to the license expansion of the application in advance.

## Related Products

- Virtual Server
- DB Service
- Load Balancer
- DNS
- Object Storage
- Block Storage