Alibaba Cloud

Auto Scaling

High application availability with automatic
scaling of servers to keep up with changing traffic needs

Background

In traditional hosting models, there are a fixed number of servers needed to be provisioned. A few servers are kept in standby mode, which are added or removed manually with changing traffic. To handle unpredictable traffic, resources are pre-provisioned based on the traffic expectations to address a possible surge. This provisioning is done on the basis of unreliable capacity planning methods, which can lead to over provisioning due to unutilized server capacity or under-provisioning due to unavailability of required resources.

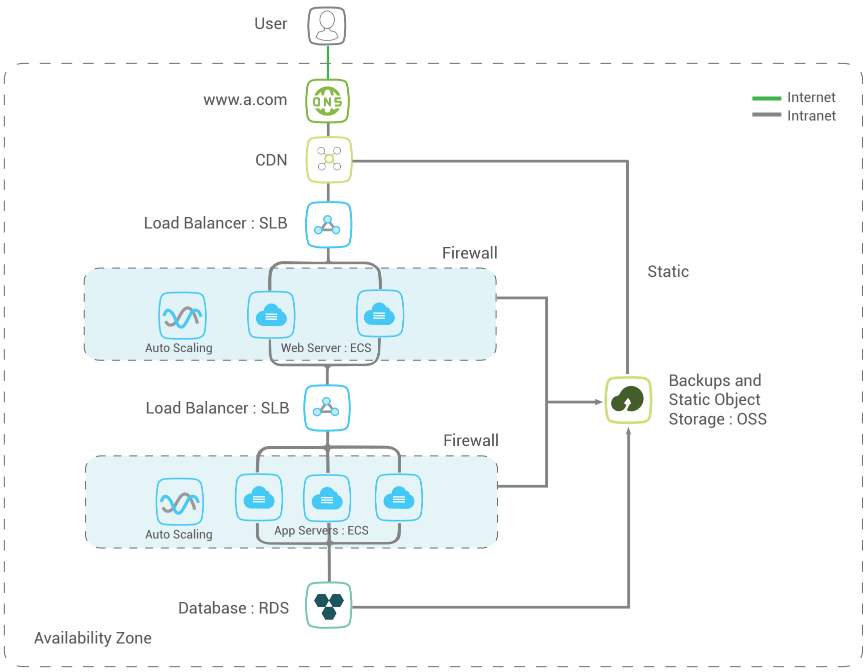
Highlights

- High uptime of your application
- Automatic provisioning of servers
- Better cost management

Benefits

- ✓ Dynamically handles fluctuating traffic peaks by automatic ECS provisioning
- ✓ Let you define multiple launch configurations for different auto scaling groups
- ✓ Configurable user-defined metrics and threshold of triggers to scale the fleet of servers
- ✓ Reduced cost with efficient usage of resources

Recommended Solution Architecture



This architecture diagram illustrates a typical web application hosting architecture with additional auto scaling capabilities:

1. User request is received and served by the nearest DNS server, and automatically routed to the CDN for accelerated content delivery.
2. It is then sent to the mapped Server Load Balancer, which distributes incoming application traffic among multiple ECS instances in a round robin manner.
3. To scale servers based on real-time traffic demands, auto scaling service is configured on web servers and application servers. These servers are automatically added or removed from SLB and RDS whitelists.
4. To store and manage relational data, application servers are connected to ApsaraDB for RDS databases.
5. All database backup archive files, root location backup and log files of the web servers are stored in scalable OSS, which scales up or down automatically ensuring no disruption of services.

intl.aliyun.com