HUAWEI

Elastic Cloud Services for Distributed Deployment

HUAWEI CLOUD
EVERYTHING AS A SERVICE

Foreword

- In the course of business development, service systems can become overloaded, slow down, or even face interruptions due to demand fluctuations and market changes. Building scalable systems helps enhance service stability, improve response speed, and optimize resource utilization.
- This chapter describes what scalability is and how to build scalable systems on Huawei Cloud.

Objectives

- Upon completion of this lesson, you will:
  - Learn the advantages of load balancing on the cloud.
  - Learn how to schedule distributed applications.
  - Learn how to adjust resources for distributed applications.

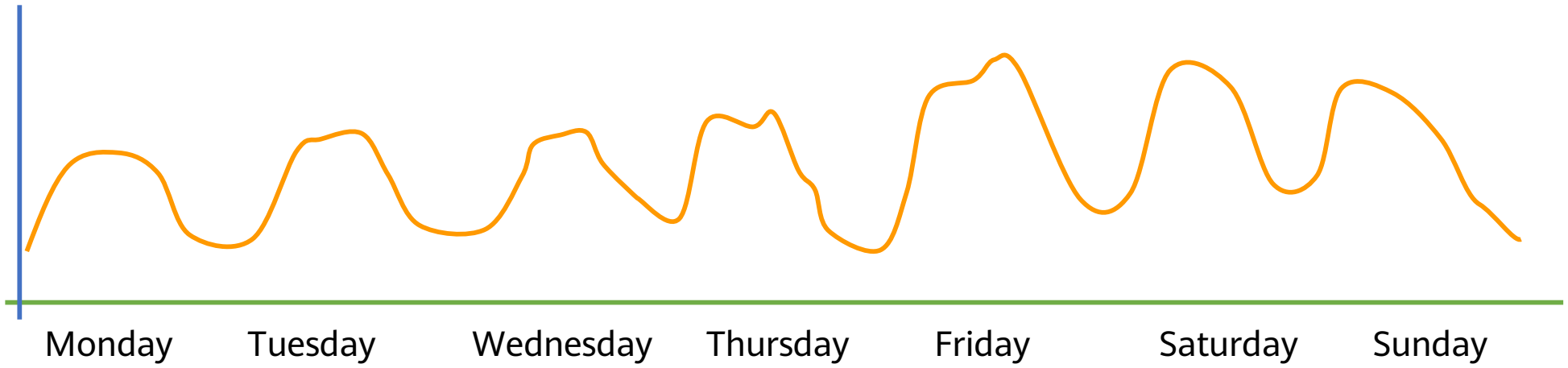
Contents

HUAWEI

Why Is Scalability Important?

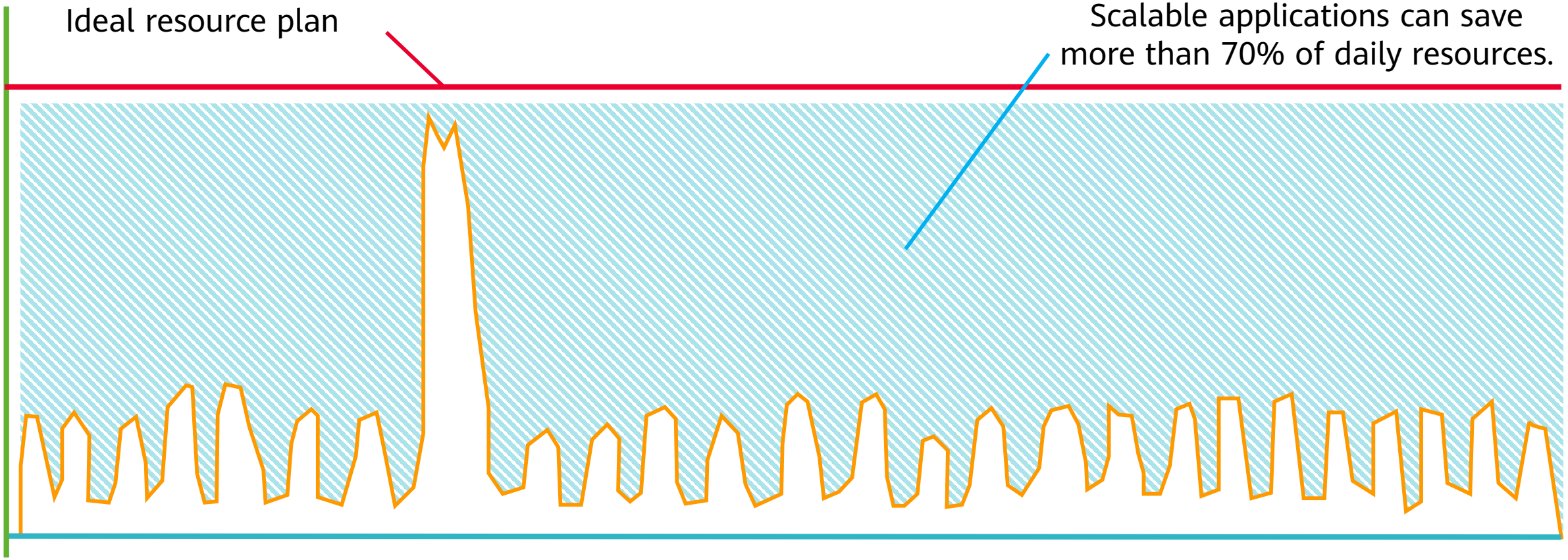
The amount of traffic an application has to handle may fluctuate.

Resources should be adjusted as traffic changes.



HUAWEI

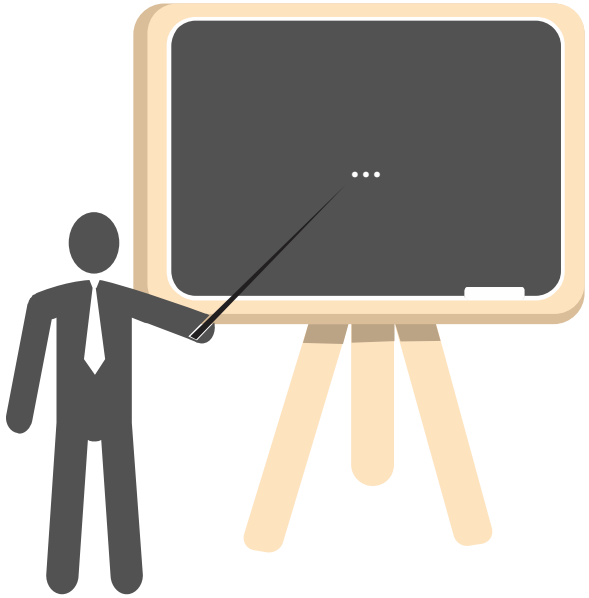
Traffic Surges During Double 11 or Other Big Promotions



It is not just a waste of resources.

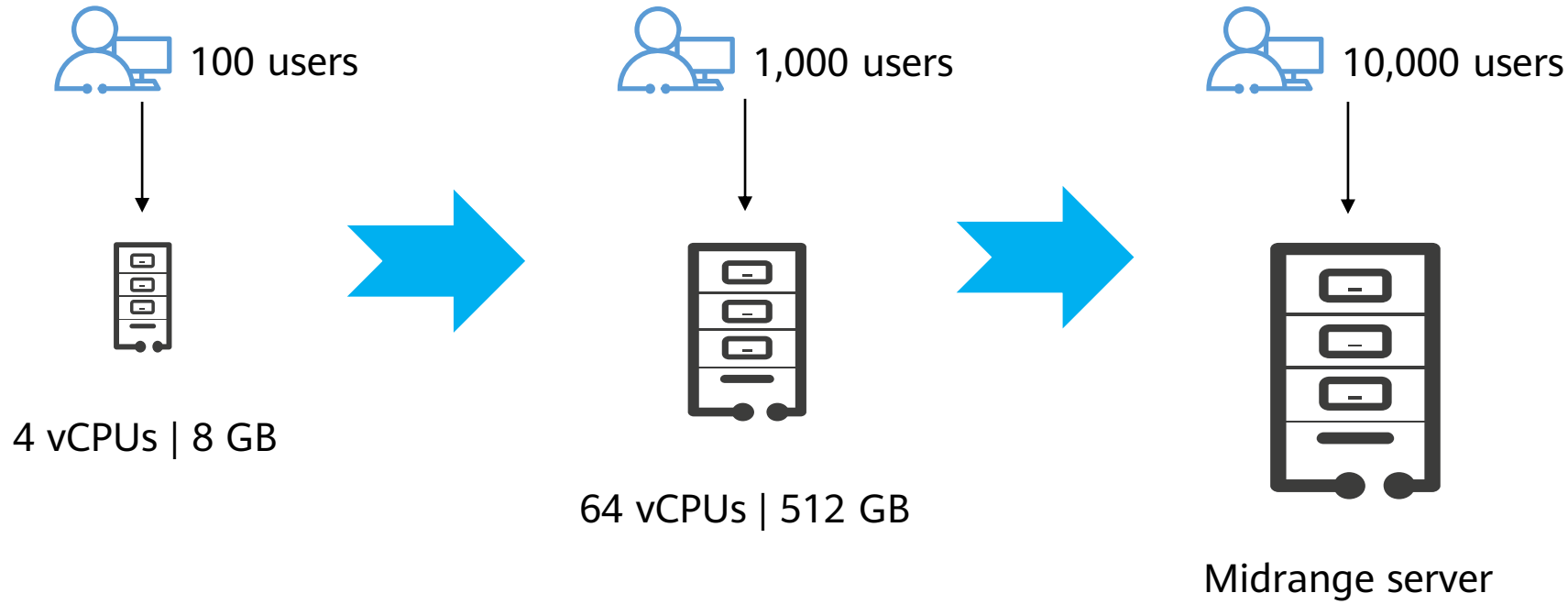
HUAWEI

How Can We Manage Sudden Traffic Changes?



HUAWEI

Method 1: Scaling Up

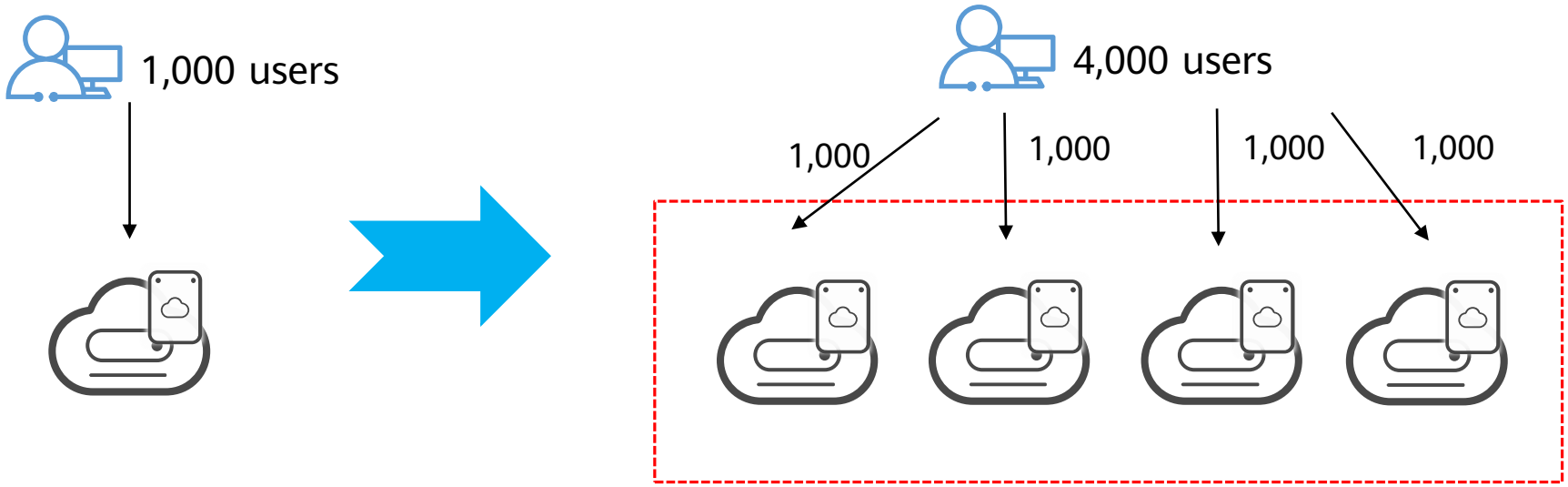


Advantages: simple and independent

Disadvantages: more bottlenecks

HUAWEI

Method 2: Scaling Out



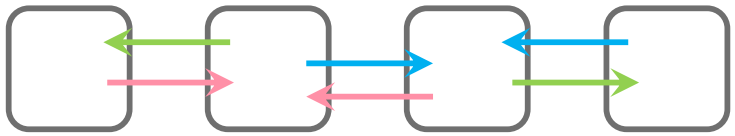
Advantages: You can use simple devices for unlimited scalability and higher availability.

Disadvantages: Technology coordination is required, and this method only works for stateless applications

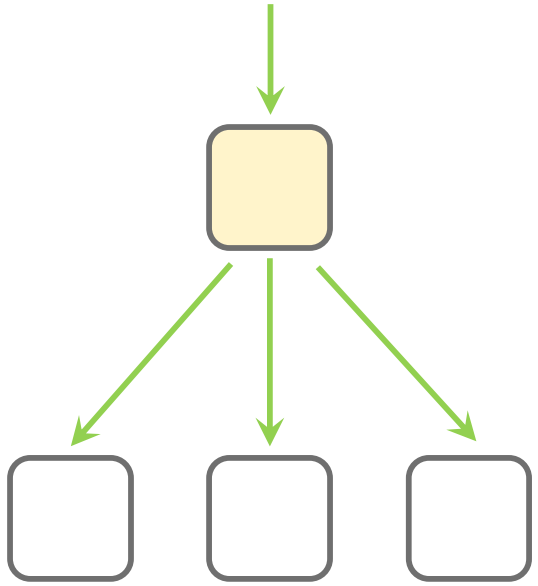
Distributed deployment seems better.

HUAWEI

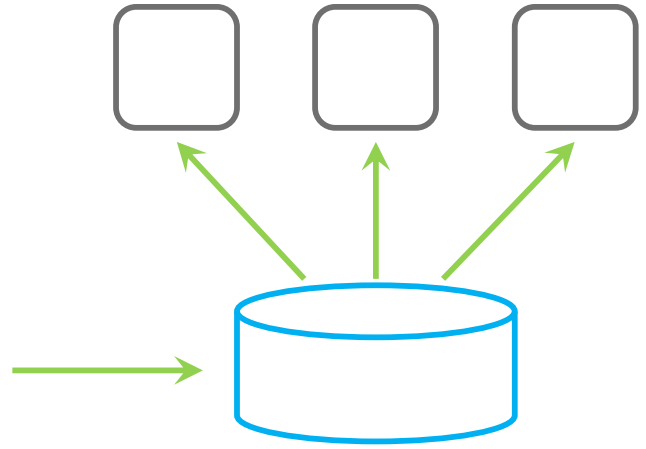
Typical Distributed Models



Synchronous communication and coordination

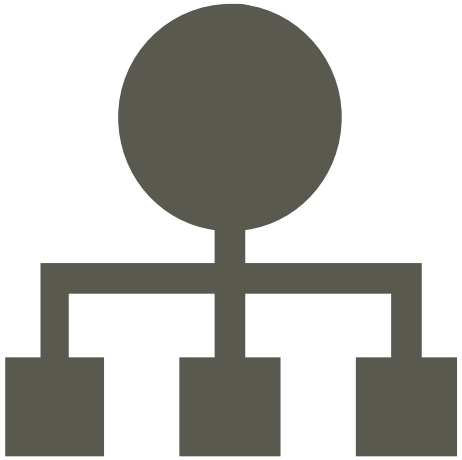


Request distribution

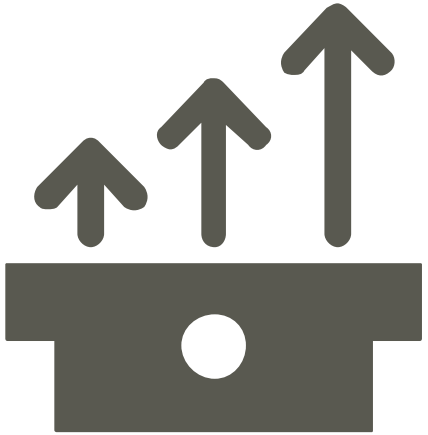


Task polling

Benefits of Distributed Systems



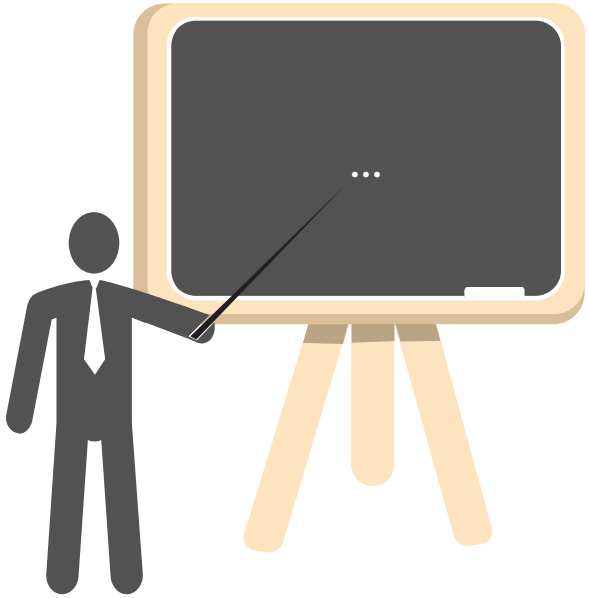
Huge resource pool for required performance



Robust load balancing for stable performance

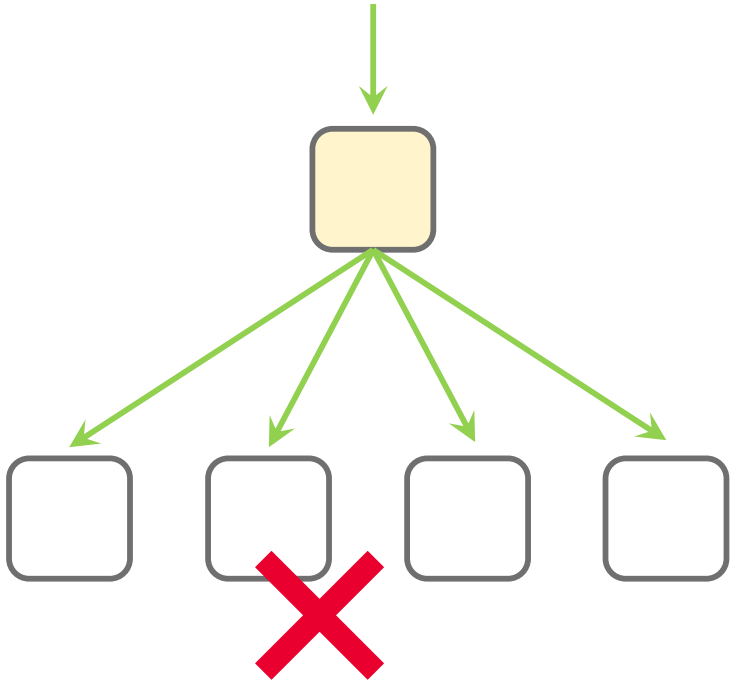
Better fault recovery

What Else Can We Get from Scalable Systems?



HUAWEI

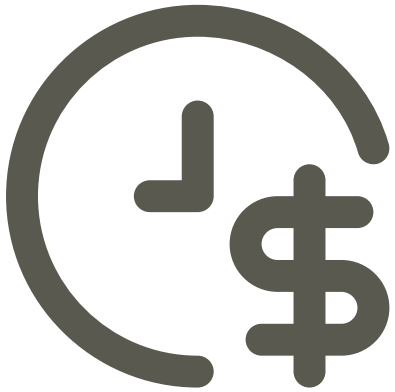
A Good Scalability Design Means Improved Reliability



- If a node is unhealthy, the overall processing capacity is reduced.
- If this happens, the system will automatically add a healthy backend server to ensure stable performance.

HUAWEI

The Value of Scalability



Adjust resources as demand changes to reduce costs.

Replace faulty nodes with healthy ones for higher reliability.

HUAWEI

Contents

HUAWEI

What Are Needed to Enable Scalability?

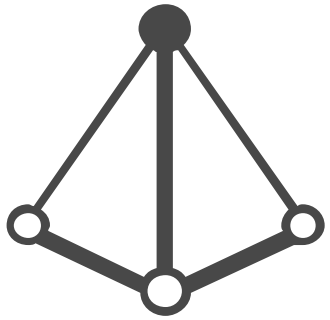
Step 1 Distribute traffic to multiple servers.

Step 2 Get to know when it is the time to scale resources.

Step 3 Scale resources immediately as requirements change.

Elastic Load Balance (ELB)

- ELB automatically distributes incoming traffic across multiple backend servers based on the listening rules you configure. It expands the service capabilities of your applications and improves their availability by eliminating single points of failure (SPOFs).



ELB

- Distributes requests across different backend servers.
- Supports built-in HA.
- Works with Auto Scaling to process a massive number of concurrent requests.
- Routes traffic across backend server groups based on the forwarding policies you have configured.
- Checks the health of backend servers to ensure that requests are routed to healthy servers.
- Supports load balancing at both Layer 4 and Layer 7.

ELB Components

Load balancer

Listener

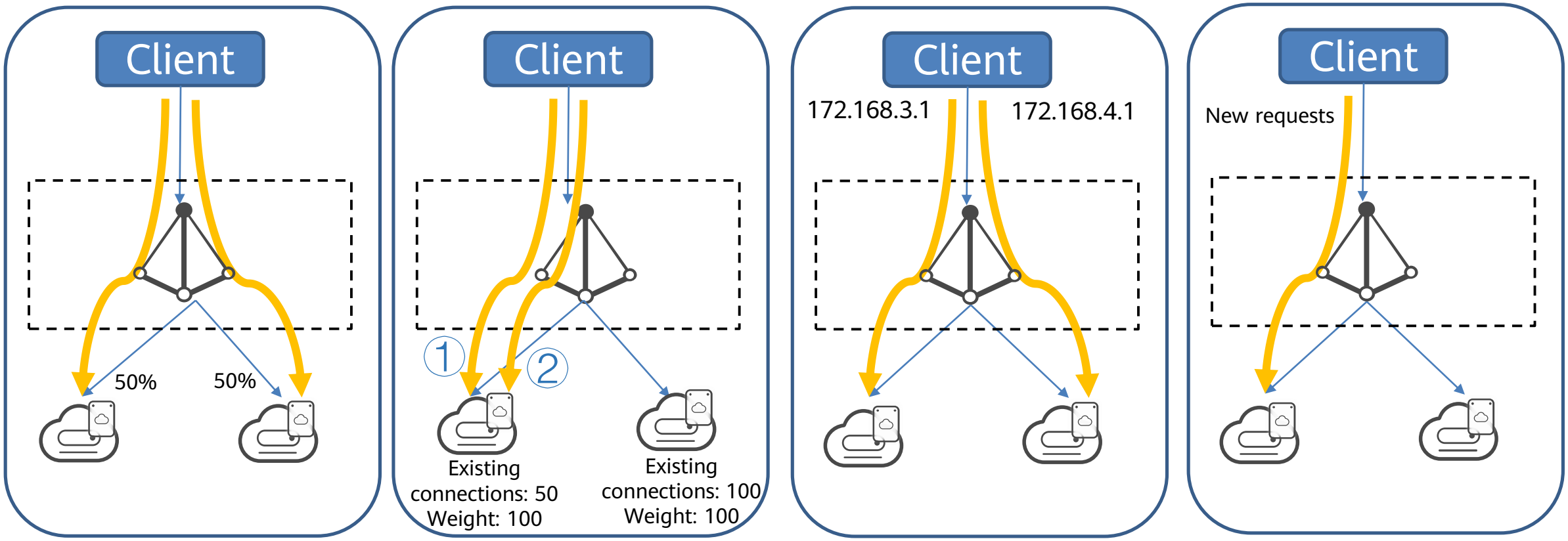
- Routes requests based on the protocol and port you have configured.
- Matches the requests based on the forwarding rules you have configured.
  - Domain name and URL (regular expression match)
  - Advanced forwarding rules such as HTTP request method, HTTP header, and more
- Forwards the requests or returns a fixed response to the clients based on the action you have configured.
  - Forward to a backend server group.
  - Redirect to another listener.
  - Redirect to another URL.

Backend server group

- One or more backend servers
  You can call APIs to add or remove a backend server.
- Multiple load balancing algorithms
  - Weighted round robin
  - Weighted least connections
- Advanced settings
  - Sticky session
  - Health check

HUAWEI

Load Balancing Algorithms



Weighted round robin

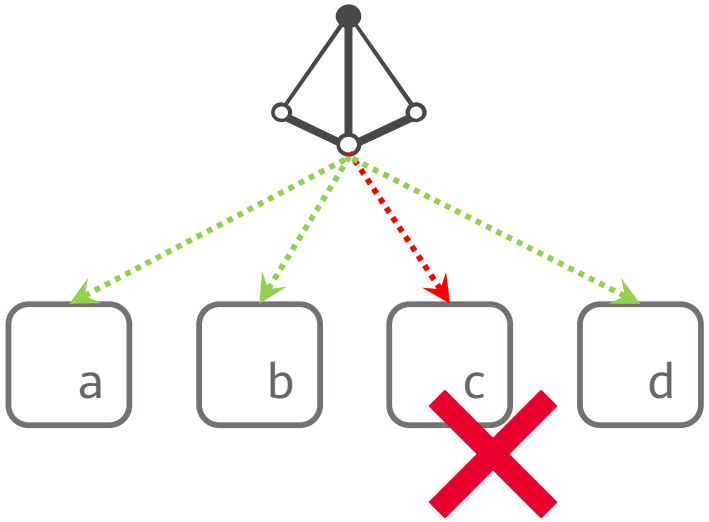
Weighted least connections

Source IP hash

Connection ID

HUAWEI

Health Check



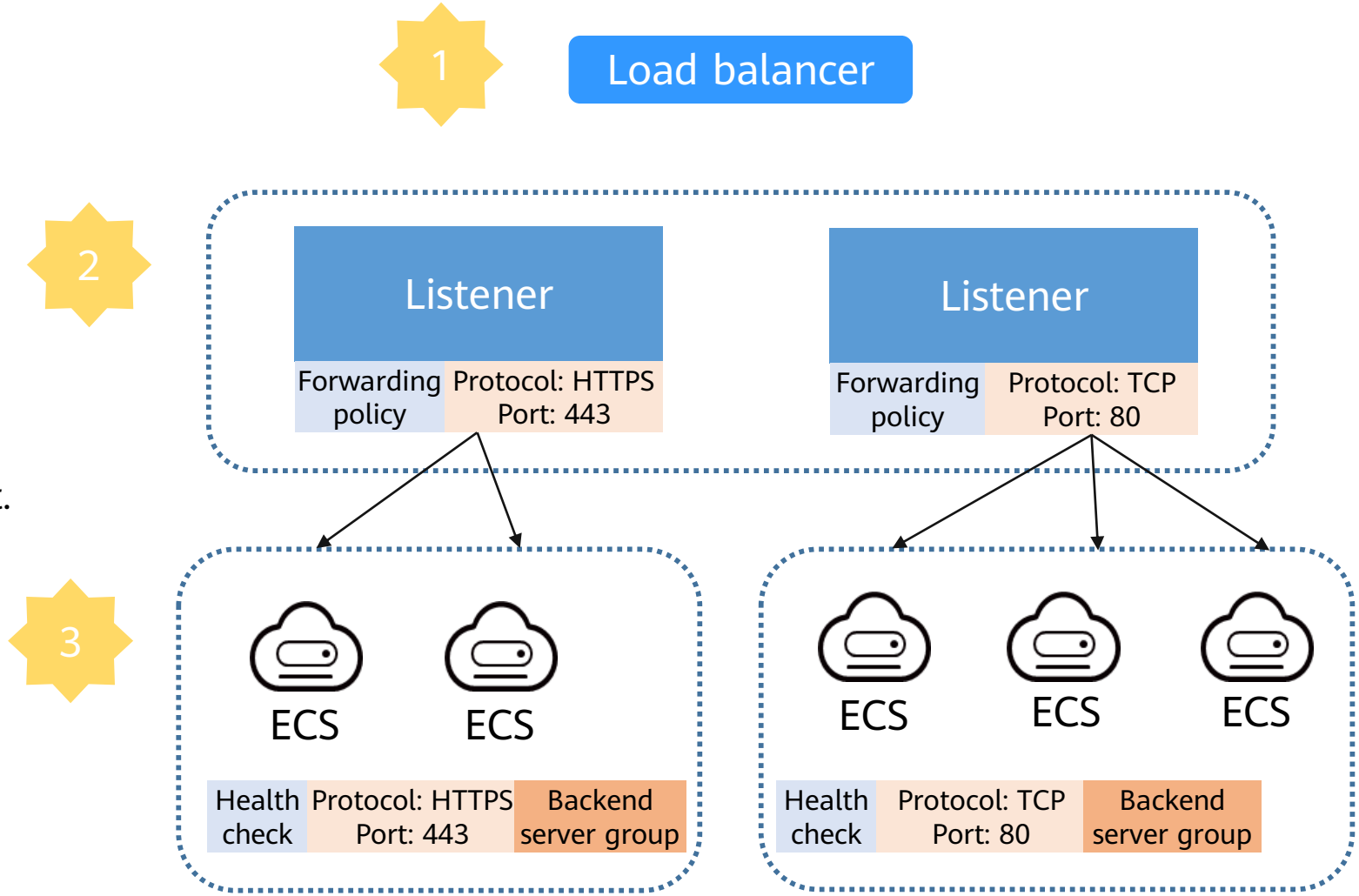
- Configure a health check to avoid routing requests to unhealthy backend servers.
- How does a health check work?
  - ELB connects to a specific TCP port used by the backend server.

    A connection is established with the port.
  - ELB accesses a specific HTTP page, generally the root directory.

    A 2*xx* code is returned.
- You should select a dedicated page for health checks.

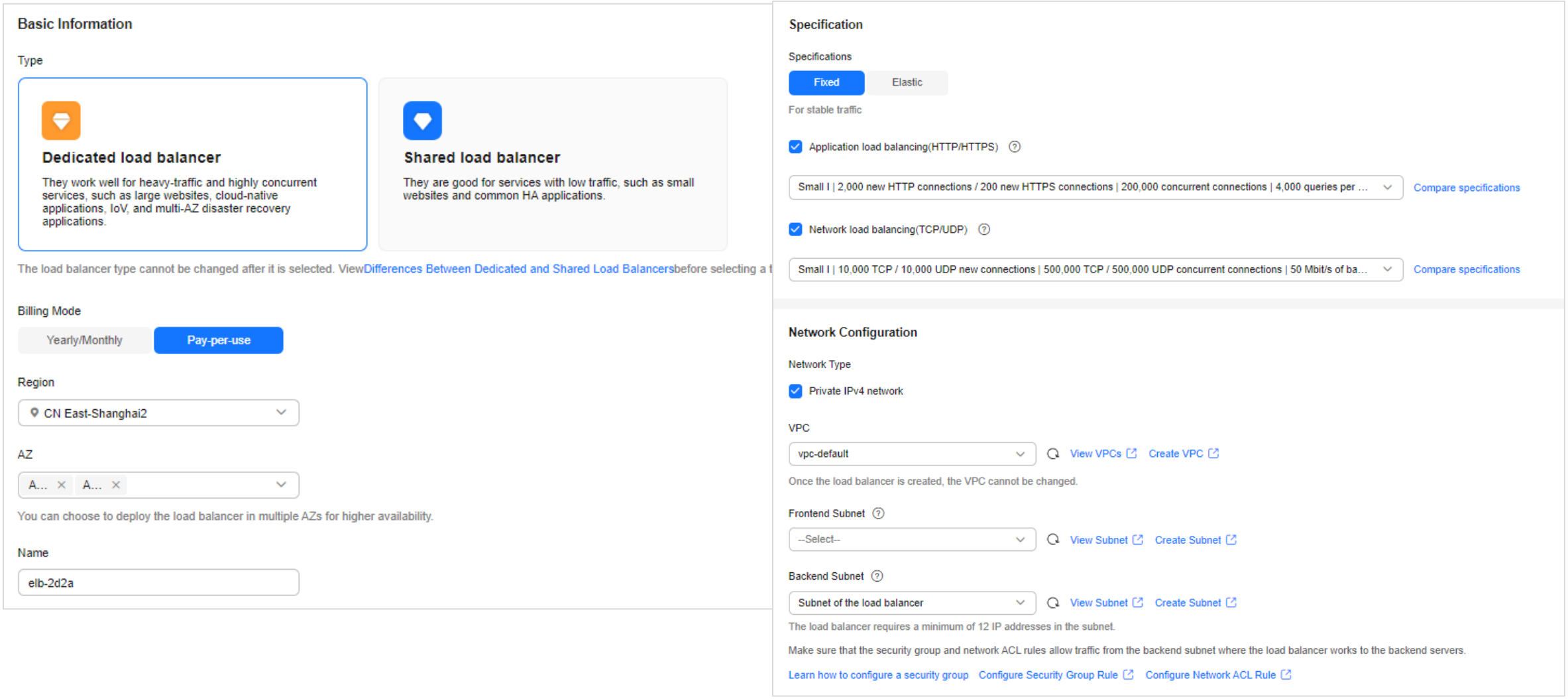
ELB Configuration Process

1. **Creating a Load Balancer**
   - Create a load balancer.
   - Select the load balancer type.
   - Configure the network parameters.
2. **Adding a Listener**
   - Locate the created load balancer.
   - Configure the protocol and port.
3. **Adding a Backend Server Group**
   - Select a load balancing algorithm.
   - Configure a health check.



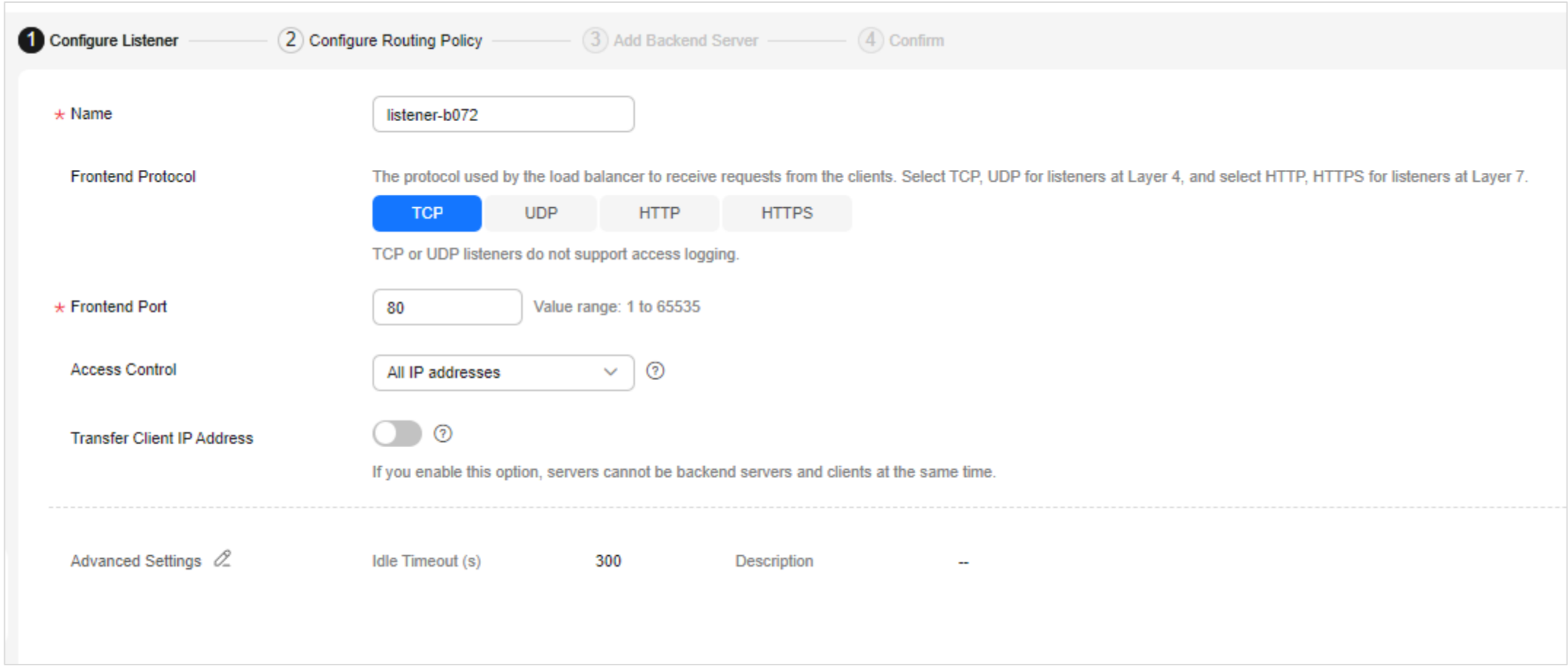
HUAWEI

Creating a Load Balancer



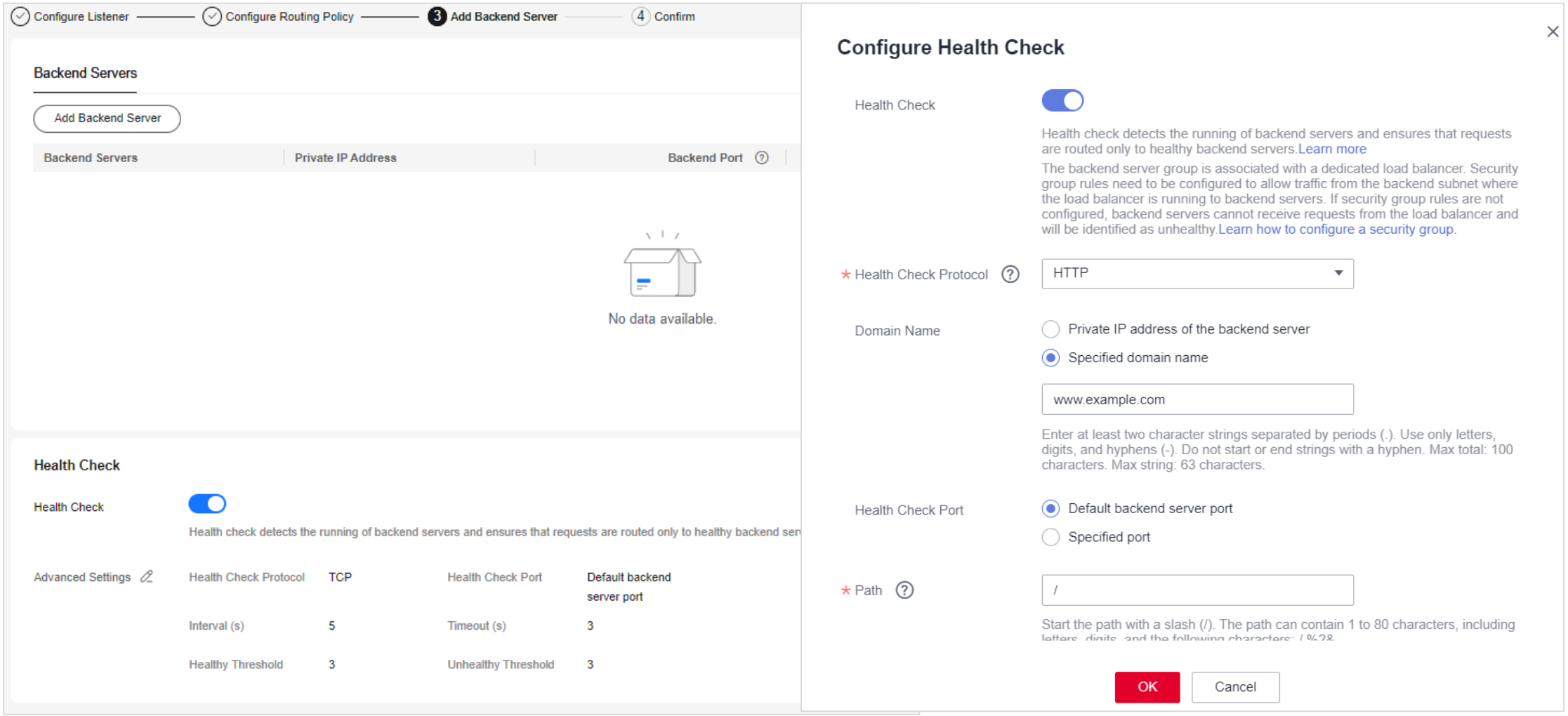
HUAWEI

Adding a Listener



HUAWEI

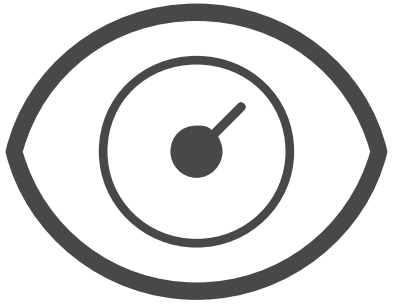
Adding a Backend Server Group and Enabling Health Check



HUAWEI

Cloud Eye

- Cloud Eye is a multi-dimensional resource monitoring service. Users can use Cloud Eye to monitor resources, configure alarm rules, identify resource exceptions, and respond to resource changes.

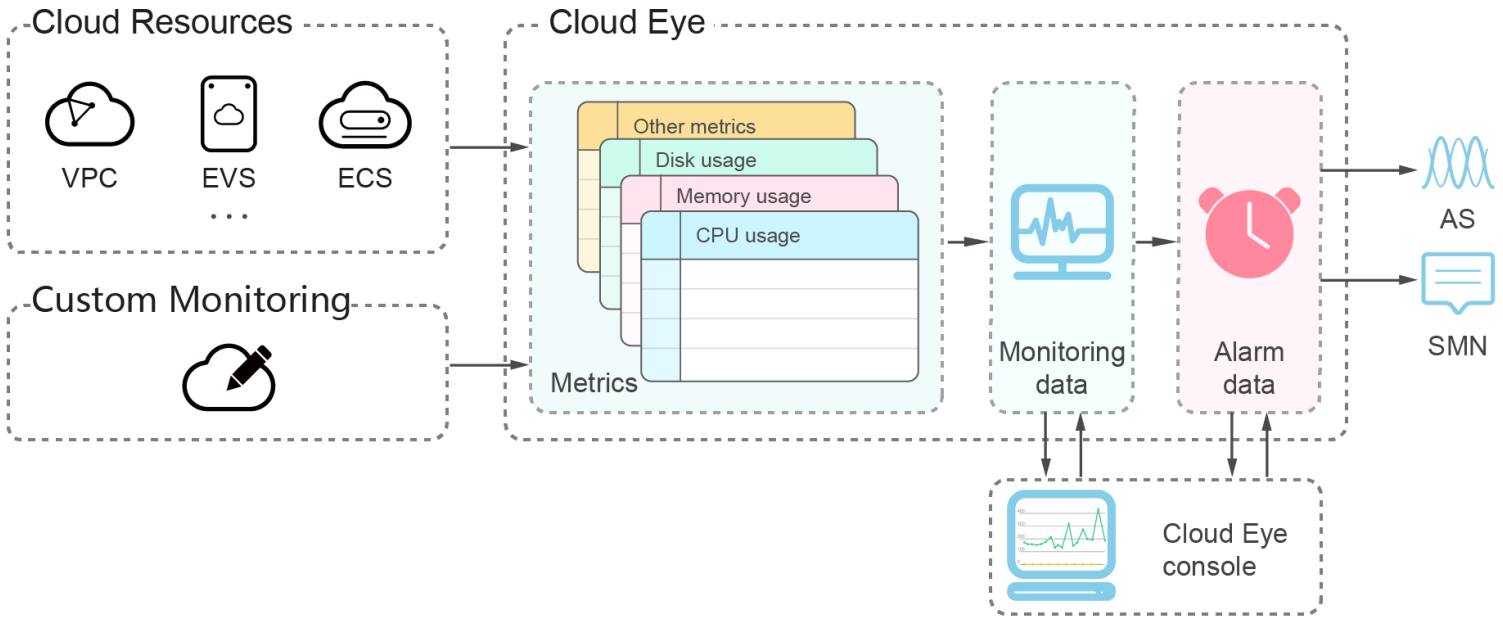


Cloud Eye

- O&M personnel cannot operate and maintain a system just based on their feelings.
- Cloud Eye is not a console where users can control different cloud services.
  - It is essentially a data collection system.
  - It is not tightly coupled with other services.
- Cloud Eye APIs can be integrated into third-party applications.
- Cloud Eye provides dashboards to visualize metrics and supports comprehensive alarm management.

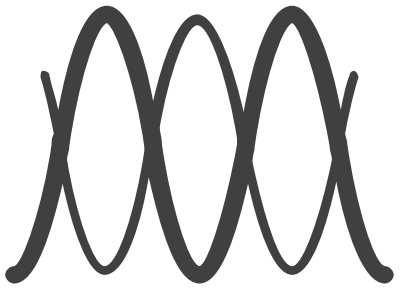
Cloud Eye Architecture

- Cloud Eye collects cloud resource or custom metrics in real time. It allows users to flexibly configure alarm rules based on the collected data. When alarms are triggered, Cloud Eye notifies users using methods such as email or SMS message, and some service systems can respond to alarms automatically to ensure that services run smoothly.



Auto Scaling

- Auto Scaling helps you automatically adjust Elastic Cloud Server (ECS) and bandwidth resources to keep up with changes in demand based on pre-configured scaling policies. It allows you to add ECS instances or bandwidth resources to handle increases in load and also save money by removing resources that are sitting idle.



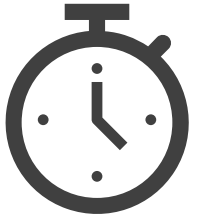
Auto Scaling

- Can create or delete ECS instances across AZs.
- Supports a broad range of conditions and policies to create or delete instances.
- Automatically identifies and replaces unhealthy instances.
- Works together with ELB to distribute incoming traffic across healthy instances.

Scaling Policies and Scenarios

- Real-time monitoring of performance
- Automatic, dynamic scaling
- Reactive scaling

Dynamic scaling based on performance



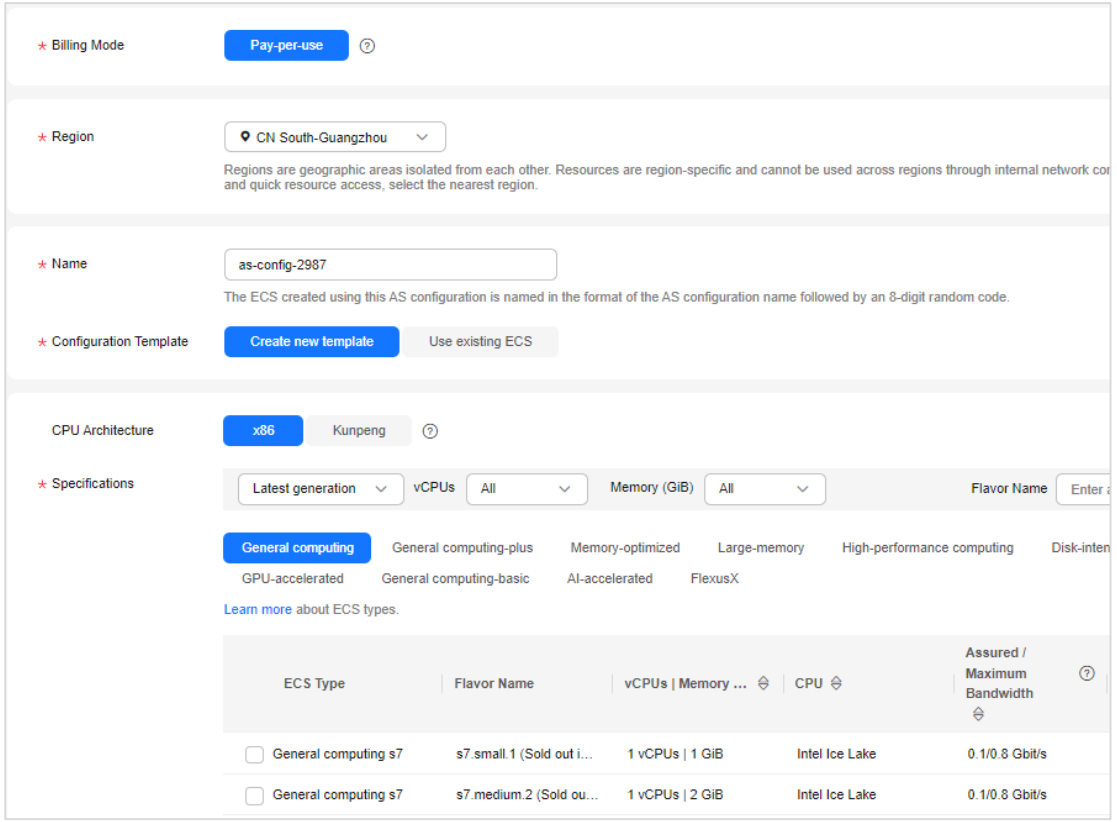
- Regular workload changes
- Scaling by schedule
- Proactive scaling

Scheduled scaling based on predictable workload changes

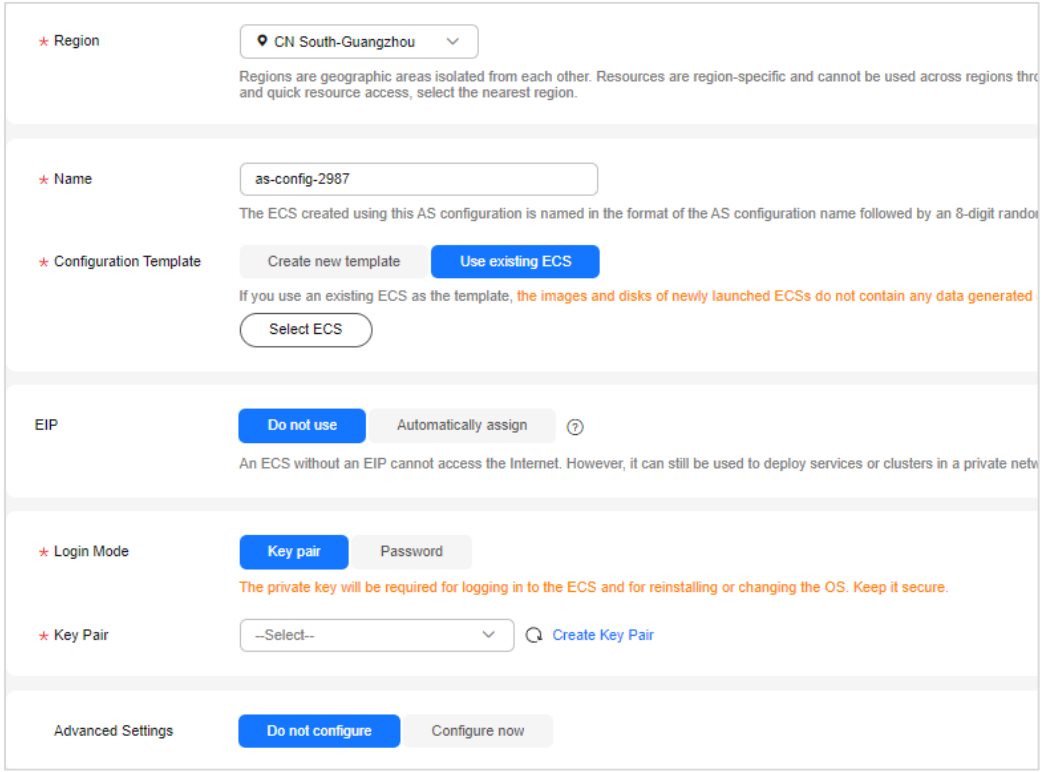
HUAWEI

Using Auto Scaling - Creating a Scaling Template

Create a scaling template. → Create a scaling group. → Create a scaling policy.



Creating a template from scratch

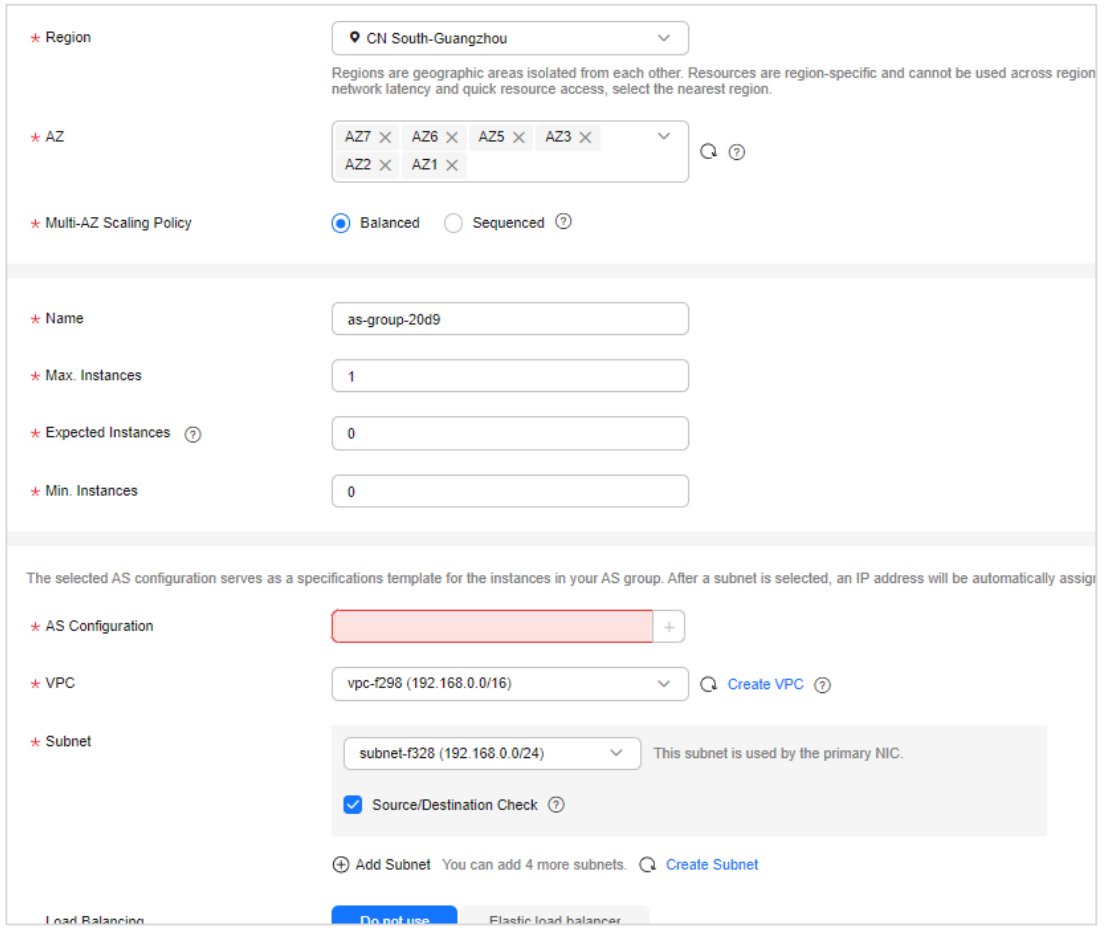


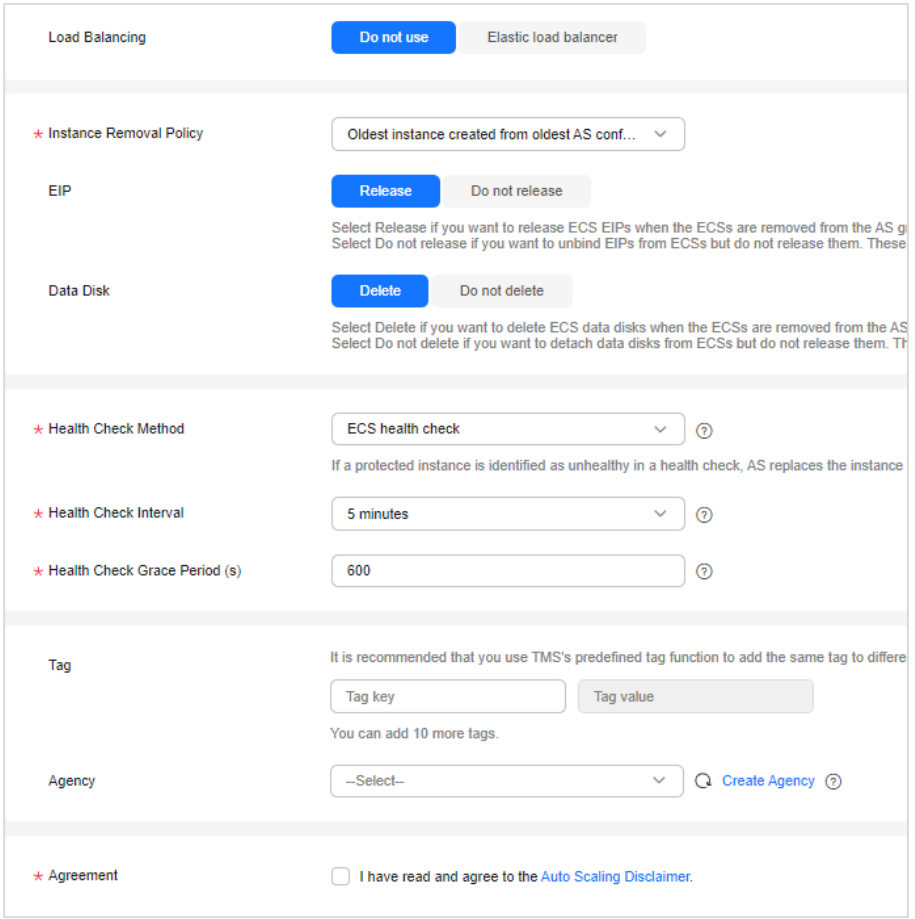
Creating a template from an existing ECS

HUAWEI

Using Auto Scaling - Creating a Scaling Group

Create a scaling template. → Create a scaling group. → Create a scaling policy.

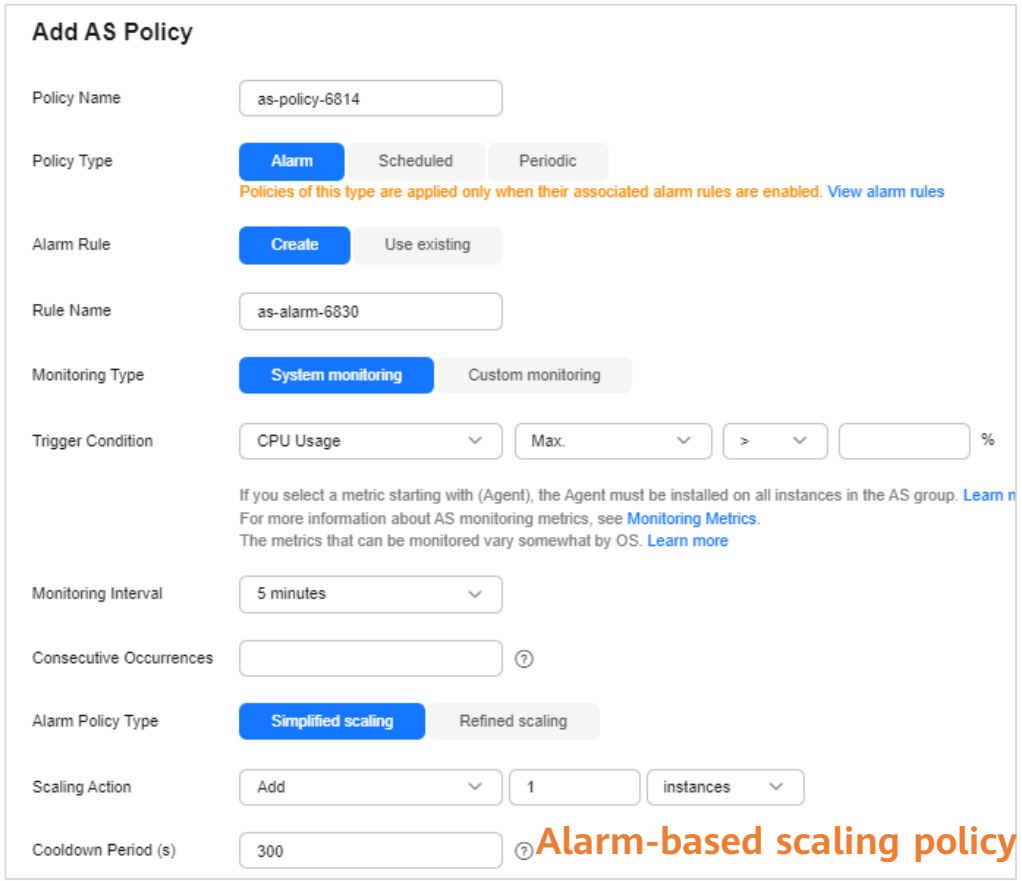


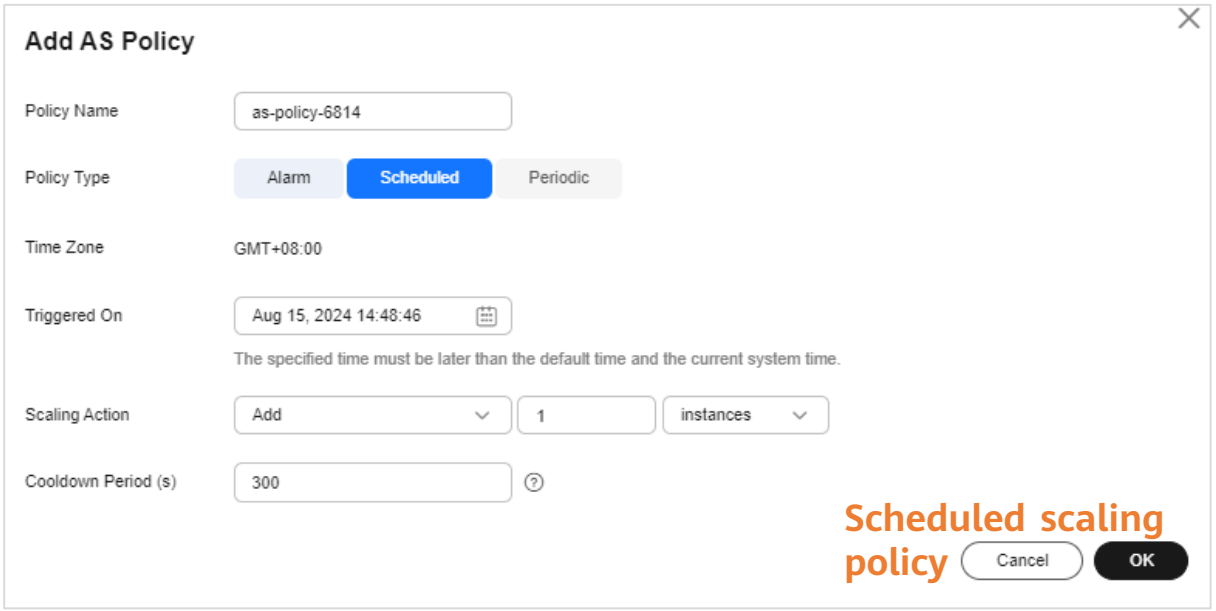


HUAWEI

Using Auto Scaling - Creating a Scaling Policy

Create a scaling template. → Create a scaling group. → Create a scaling policy.

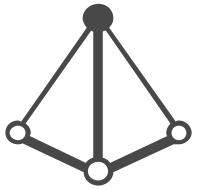




HUAWEI

Cloud Services for Creating Scalable Systems

Step 1 Distribute traffic to multiple servers.

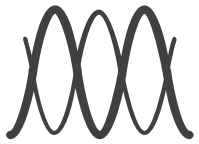


ELB

Step 2 Get to know when it is the time to scale resources.

Cloud Eye

Step 3 Quickly scale in or out resources.

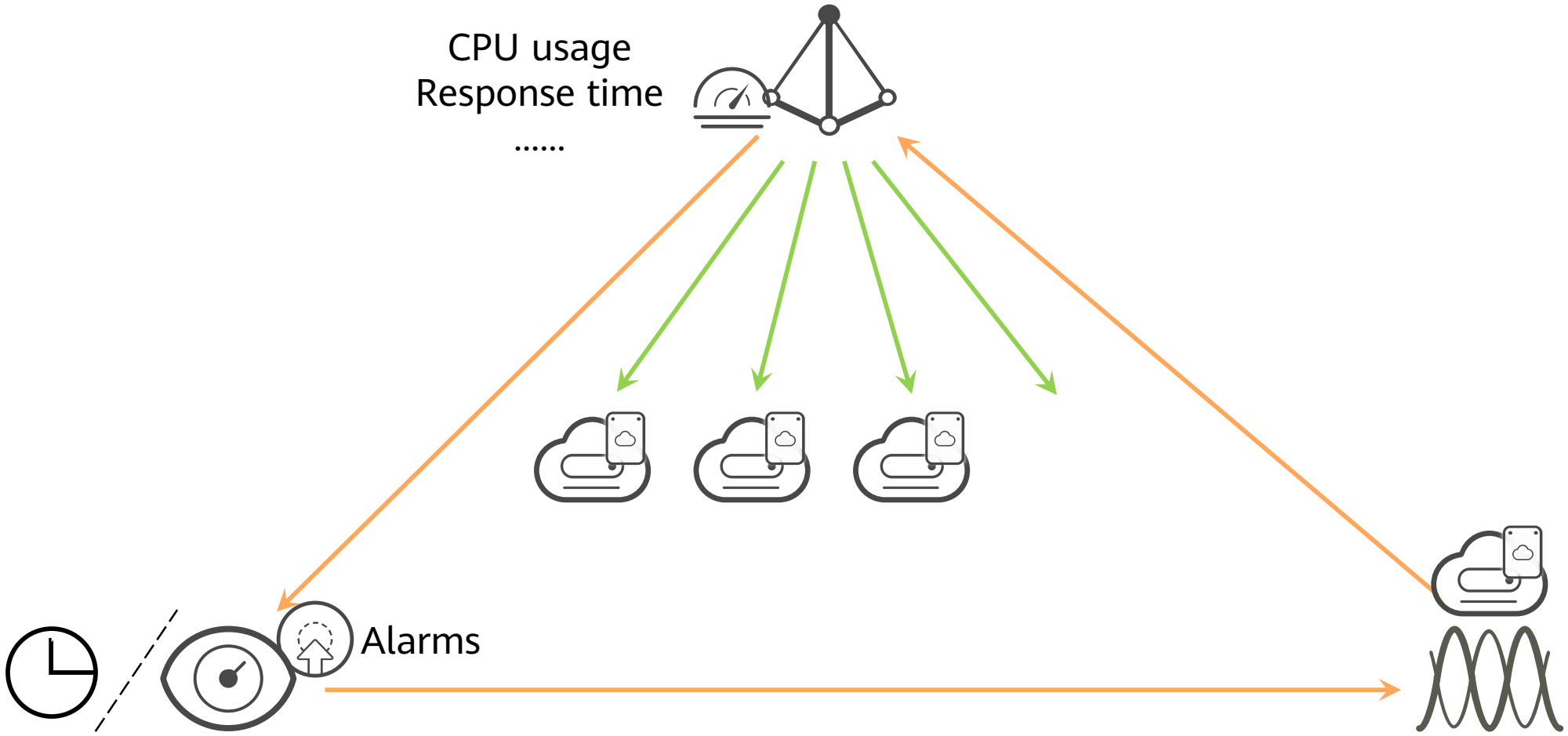


Auto Scaling

Contents

HUAWEI

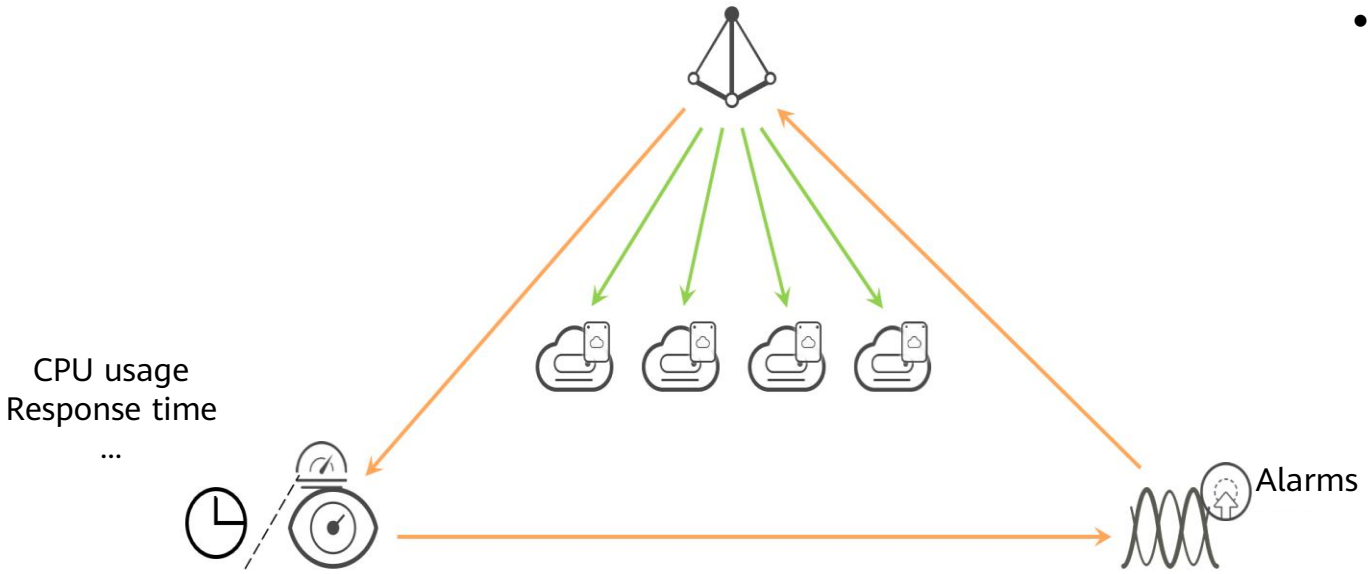
Typical Scalable Website Architecture



HUAWEI

Application Scenarios

- **Heavy-traffic forums**: Service load changes are difficult to predict for heavy-traffic forum websites. AS dynamically adjusts the number of ECSs based on monitored ECS metrics, such as CPU usage and memory usage.
    - **Livestreaming**: A livestreaming website may broadcast popular programs from 14:00 to 16:00 every day. Auto Scaling automatically scales out ECS and bandwidth resources during this period to ensure a smooth viewer experience.



- **E-commerce**: During big promotions, E-commerce websites need more resources. Auto Scaling automatically add resources within minutes to ensure that promotions go smoothly.

Summary

- Learned the advantages of load balancing on the cloud.
- Learned how to schedule distributed applications.
- Learned how to adjust resources for distributed applications.

Quiz

For a scalable system on Huawei Cloud, which of the following can be used to create or delete servers? ()

A. Elastic Load Balance (ELB)

B. Cloud Eye

C. Auto Scaling

D. Elastic Cloud Server (ECS)

Acronyms and Abbreviations

- ECS: Elastic Cloud Server
- EVS: Elastic Volume Service

Recommendations

- Huawei Cloud websites
  - Huawei Cloud: https://www.huaweicloud.com/intl/en-us/
  - Huawei Cloud Developer Institute: https://edu.huaweicloud.com/intl/en-us/

Huawei Cloud
Developer Institute

Thank You.

HUAWEI CLOUD
EVERYTHING AS A SERVICE

www.huaweicloud.com