Auto Scaling

Visão geral de serviço

Edição 01

Data 30-10-2021

HUAWEI TECHNOLOGIES CO., LTD.

Marcas registadas e permissões


e outras marcas registadas da Huawei são marcas registadas da Huawei Technologies Co., Ltd. Todos as outras marcas registadas e os nomes registados mencionados neste documento são propriedade dos seus respectivos detentores.

Aviso

Os produtos, serviços e funcionalidades adquiridos são estipulados pelo contrato feito entre a Huawei e o cliente. Todos ou parte dos produtos, serviços e funcionalidades descritos neste documento pode não estar dentro do âmbito de aquisição ou do âmbito de uso. Salvo especificação em contrário no contrato, todas as declarações, informações e recomendações neste documento são fornecidas "TAL COMO ESTÁ" sem garantias, ou representações de qualquer tipo, seja expressa ou implícita.

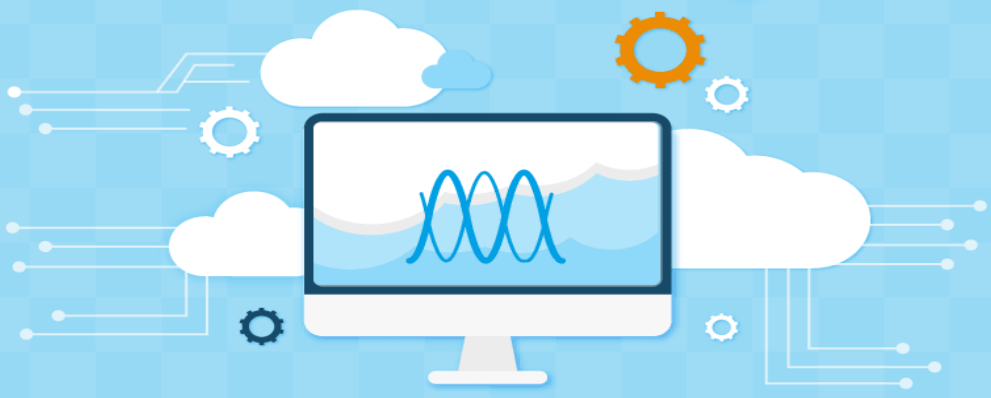
As informações contidas neste documento estão sujeitas a alterações sem aviso prévio. Foram feitos todos os esforços na preparação deste documento para assegurar a exatidão do conteúdo, mas todas as declarações, informações e recomendações contidas neste documento não constituem uma garantia de qualquer tipo, expressa ou implícita.

Índice

1 Infográficos do AS....1
2 O que é o Auto Scaling?....3
3 Vantagens do AS....5
4 Ciclo de vida da instância....10
5 Restrições....15
6 Região e AZ....17
7 Cobrança....19
8 O AS e outros serviços....20
9 Gerenciamento de permissões....23
10 Conceitos básicos....26
11 História de mudanças....28

1 Infográficos do AS

Auto Scaling



01 What Is AS?

Auto Scaling (AS) automatically adjusts resources to keep up with changes in demand based on pre-configured AS policies.

You can specify AS configurations and policies based on service requirements. These configurations and policies free you from having to repeatedly adjust resources to keep up with service changes and spikes in demand. In this way, AS helps you reduce the resources and manpower required.

Create an AS group.

Create an AS configuration and policy.

AS automatically scales out ECSs.

02 Advantages of AS

During the 11.11 Shopping Festival in China on November 11, it is extremely hard to keep up with the massive increase in demand. Traditional solutions are not up to the task.

Provision as many as servers as required for peak demand. All that capacity will be wasted during off-peak hours.

Provision servers based on average loads of applications. Capacity will be insufficient during peak hours.

AS perfectly resolves this issue.

2 O que é o Auto Scaling?

Introdução sobre o AS

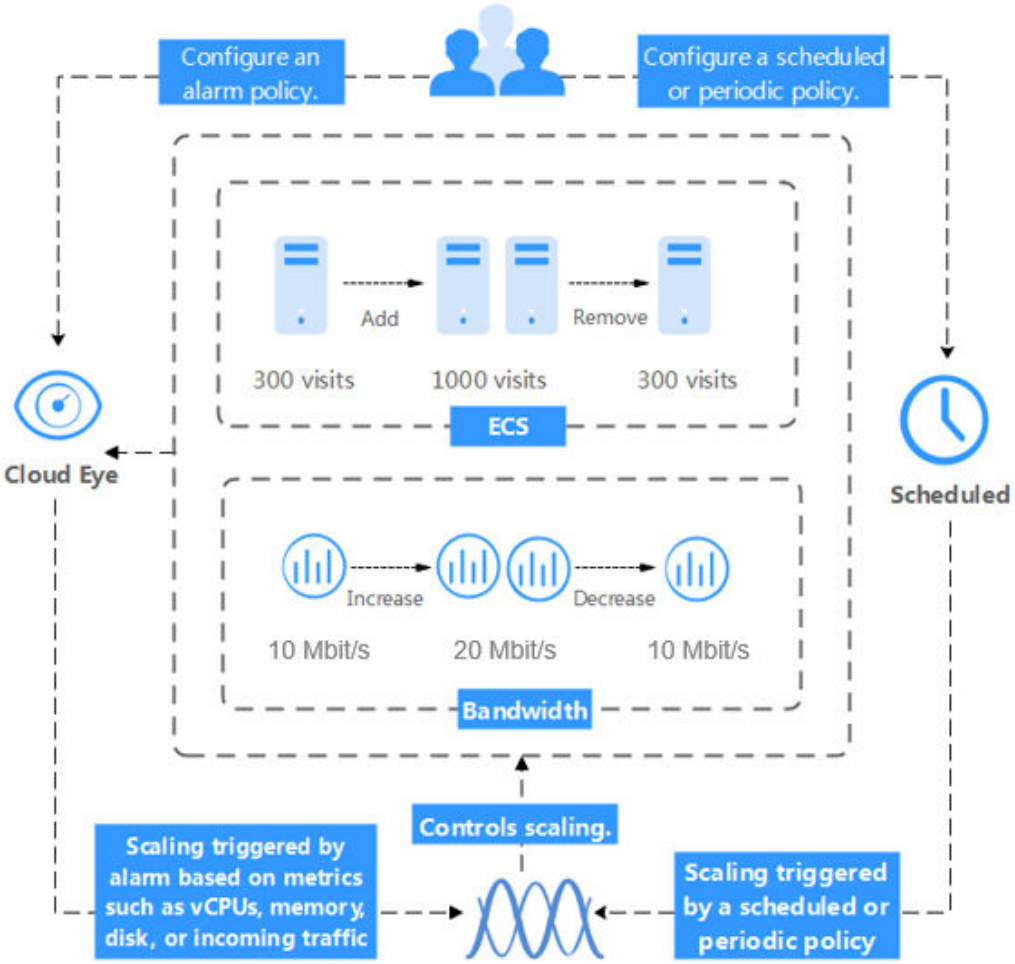
O Auto Scaling (AS) ajuda você a dimensionar automaticamente o Elastic Cloud Server (ECS) e os recursos de largura de banda para acompanhar as mudanças na demanda com base nas políticas de AS pré-configuradas. Ele permite que você adicione instâncias do ECS ou aumente as larguras de banda para lidar com aumentos de carga e também economizar dinheiro removendo recursos que estão ociosos.

Arquitetura

O AS permite dimensionar instâncias e larguras de banda do ECS.

- Controle de dimensionamento: você pode configurar políticas do AS, configurar limites de métricas e agendar quando diferentes ações de dimensionamento são tomadas. O AS acionará ações de dimensionamento em uma programação repetida, em um horário específico ou quando os limites configurados forem atingidos.
- Configuração de política: você pode configurar políticas baseadas em alarme, agendadas e periódicas, conforme necessário.
- Políticas baseadas em alarmes: você pode configurar ações de dimensionamento a serem tomadas quando métricas de alarme, como vCPU, memória, disco e tráfego de entrada, atingirem os limites.
- Políticas agendadas: você pode agendar ações de dimensionamento a serem tomadas em um horário específico.
- Políticas periódicas: você pode configurar ações de dimensionamento a serem executadas em intervalos programados, em um horário específico ou dentro de um intervalo de tempo específico.
- Quando o Cloud Eye gera um alarme para uma métrica de monitoramento, por exemplo, uso da CPU, o AS aumenta ou diminui automaticamente o número de instâncias no grupo de AS ou as larguras de banda.
- Quando o tempo de disparo configurado chega, uma ação de dimensionamento é acionada para aumentar ou diminuir o número de instâncias do ECS ou as larguras de banda.

Figura 2-1 Arquitetura do AS



Acessar o AS

A nuvem pública fornece uma plataforma de gerenciamento de serviços baseada na Web. Você pode acessar o AS usando interfaces de programação de aplicações (APIs) compatíveis com HTTPS ou o console de gerenciamento.

- Chamada das APIs

  Use esse método se for necessário integrar o AS na nuvem pública em um sistema de terceiros para desenvolvimento secundário. Para obter detalhes, consulte ***Referência de API do Auto Scaling***

  .

- Console de gerenciamento

  Use esse método se não precisar integrar o AS a um sistema de terceiros.

  Depois de registrar na nuvem pública, faça logon no console de gerenciamento e selecione **Auto Scaling** na lista de serviços na página inicial.

3 Vantagens do AS

O AS dimensiona automaticamente os recursos para acompanhar as demandas de serviço com base em políticas de AS pré-configuradas. Com o dimensionamento automático de recursos, você pode aproveitar custos reduzidos, disponibilidade aprimorada e alta tolerância a falhas. O AS é usado para os seguintes cenários:

- Fóruns de tráfego pesado: o tráfego em um fórum popular é difícil de prever. O AS ajusta dinamicamente o número de instâncias do ECS com base em métricas do ECS monitoradas, como uso de vCPU e memória.
- Comércio eletrônico: durante grandes promoções, os sites de comércio eletrônico precisam de mais recursos. O AS aumenta automaticamente as instâncias e larguras de banda do ECS em minutos para garantir que as promoções ocorram sem problemas.
- Transmissão ao vivo: um site de transmissão ao vivo pode transmitir programas populares das 14:00 às 16:00 todos os dias. O AS dimensiona automaticamente os recursos de ECS e largura de banda durante esse período para garantir uma experiência de visualização tranquila.

Dimensionamento automático de recursos

O AS adiciona instâncias do ECS e aumenta a largura de banda para suas aplicações quando o volume de acesso aumenta e remove recursos desnecessários quando o volume de acesso cai, garantindo a estabilidade e a disponibilidade do sistema.

- Dimensionamento de instâncias de ECS sob demanda

  O AS dimensiona instâncias do ECS para aplicações com base na demanda, melhorando o gerenciamento de custos. As instâncias do ECS podem ser dimensionadas dinamicamente, de acordo com uma programação ou manualmente:

  - Dimensionamento dinâmico

    O dimensionamento dinâmico permite dimensionar recursos em resposta a mudanças na demanda usando políticas baseadas em alarmes.

  - Dimensionamento agendado

    O dimensionamento agendado ajuda você a configurar seu próprio cronograma de dimensionamento de acordo com alterações de carga previsíveis, criando políticas periódicas ou agendadas.

  - Dimensionamento manual

    Você pode alterar manualmente o número esperado de instâncias do grupo de AS ou adicionar ou remover instâncias de ou para o grupo de AS.

Considere um aplicação de reserva de bilhetes de trem em execução na nuvem pública. A carga da aplicação pode ser relativamente baixa durante Q2 e Q3 porque não há muitos viajantes, mas relativamente alta durante Q1 e Q4. Tradicionalmente, existem duas maneiras de planejar essas mudanças na carga. A primeira opção é provisionar servidores suficientes para que a aplicação sempre tenha capacidade suficiente para atender à demanda, conforme mostrado em **Figura 3-1**. A segunda opção é provisionar servidores de acordo com a carga média da aplicação, conforme mostrado em **Figura 3-2**. No entanto, essas duas opções podem desperdiçar recursos ou não conseguir atender à demanda durante as altas temporadas. Ao ativar o AS para esta aplicação, você tem uma terceira opção disponível. O AS ajuda você a dimensionar os servidores para acompanhar as mudanças na demanda. Isso permite que a aplicação mantenha um desempenho estável e previsível sem desperdiçar dinheiro com recursos desnecessários, conforme mostrado em **Figura 3-3**.

Figura 3-1 Capacidade superprovisionada

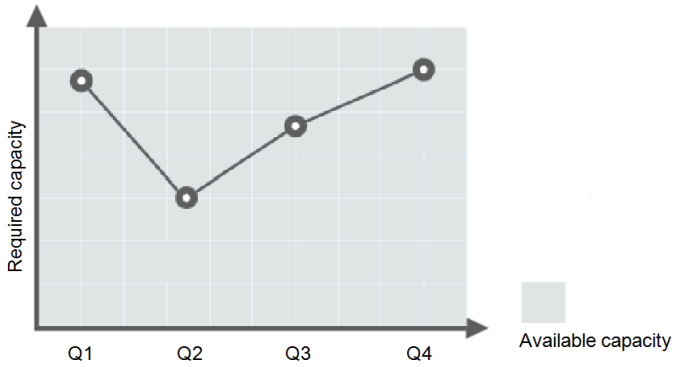


Figura 3-2 Capacidade insuficiente

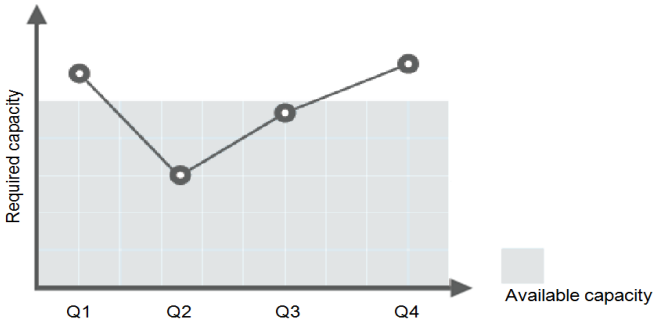
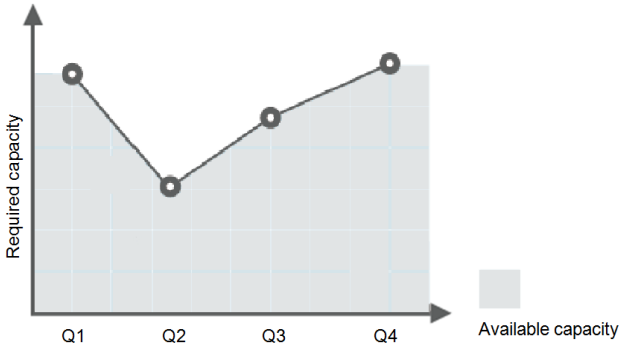


Figura 3-3 Capacidade auto-dimensionada



- Dimensionamento de largura de banda sob demanda

  O AS ajusta a largura de banda para uma aplicação com base na demanda, reduzindo os custos de largura de banda.

  Existem três tipos de políticas de dimensionamento que você pode usar para ajustar a largura de banda IP sob demanda:

  - Políticas baseadas em alarme

    Você pode configurar gatilhos com base em métricas como tráfego de saída e largura de banda. Quando o sistema detecta que as condições de disparo são atendidas, o sistema ajusta automaticamente a largura de banda.

  - Políticas agendadas

    O sistema aumenta, diminui ou ajusta automaticamente a largura de banda para um valor fixo em um cronograma fixo.

  - Políticas periódicas

    O sistema ajusta periodicamente a largura de banda com base em um ciclo periódico configurado.

  Por exemplo, você pode usar uma política baseada em alarme para regular a largura de banda de um site de transmissão ao vivo.

  Para um site de transmissão ao vivo, a carga do serviço é difícil de prever. Neste exemplo, a largura de banda precisa ser ajustada dinamicamente entre 10 Mbit/s e 30 Mbit/s com base em métricas como tráfego de saída e tráfego de entrada. O AS pode ajustar automaticamente a largura de banda para atender aos requisitos. Você só precisa selecionar o EIP relevante e criar duas políticas de alarme. Uma política é aumentar a largura de banda em 2 Mbit/s quando o tráfego de saída for maior que *X* bytes, com o limite definido para 30 Mbit/s. A outra política é diminuir a largura de banda em 2 Mbit/s quando o tráfego de saída for menor que *X* bytes, com o limite definido para 10 Mbit/s.

- Instâncias distribuídas uniformemente pelo AZ

  Para reduzir o impacto da falta de energia ou da rede na estabilidade do sistema, o AS tenta distribuir instâncias do ECS uniformemente entre as AZs usadas por um grupo de AS.

  Uma região é uma área geográfica onde os recursos usados pelas instâncias do ECS estão localizados. Cada região contém várias zonas de disponibilidade (AZs) em que os recursos usam fontes de alimentação e redes independentes. As AZs são fisicamente isoladas umas das outras, mas interconectadas através de uma intranet. As AZs são projetadas para serem isoladas de falhas em outras AZs. Eles fornecem conexões de rede econômicas e de baixa latência para outras AZs na mesma região.

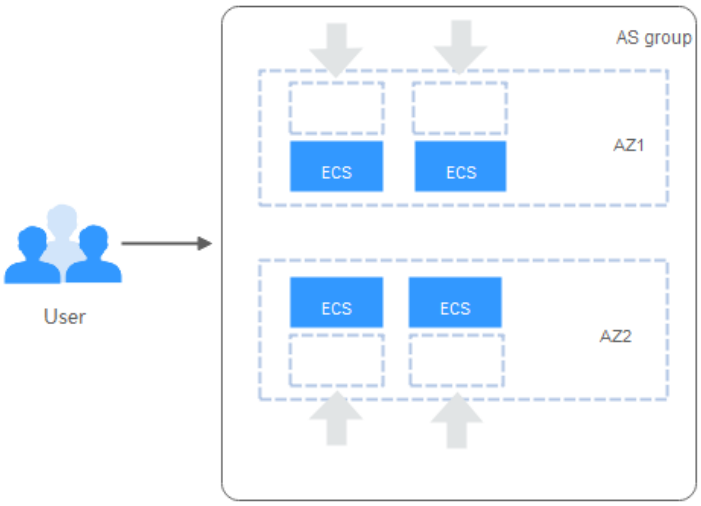
  Um grupo de AS pode conter instâncias do ECS em uma ou mais AZs em uma região. Durante o dimensionamento da capacidade de um grupo de AS, o AS tenta distribuir uniformemente instâncias do ECS entre as AZs usadas pelo grupo de AS com base nas seguintes regras:

  Distribuição uniforme de novas instâncias para AZs balanceadas

  O AS tenta distribuir uniformemente instâncias do ECS entre as AZs usadas por um grupo de AS. Para fazer isso, o AS adiciona novas instâncias à AZ com o menor número de instâncias.

  Considere um grupo de AS contendo quatro instâncias distribuídas uniformemente nas duas AZs usadas pelo grupo de AS. Se uma ação de dimensionamento for acionada para adicionar mais quatro instâncias ao grupo do AS, o AS adicionará duas a cada AZ.

Figura 3-4 Distribuir instâncias uniformemente

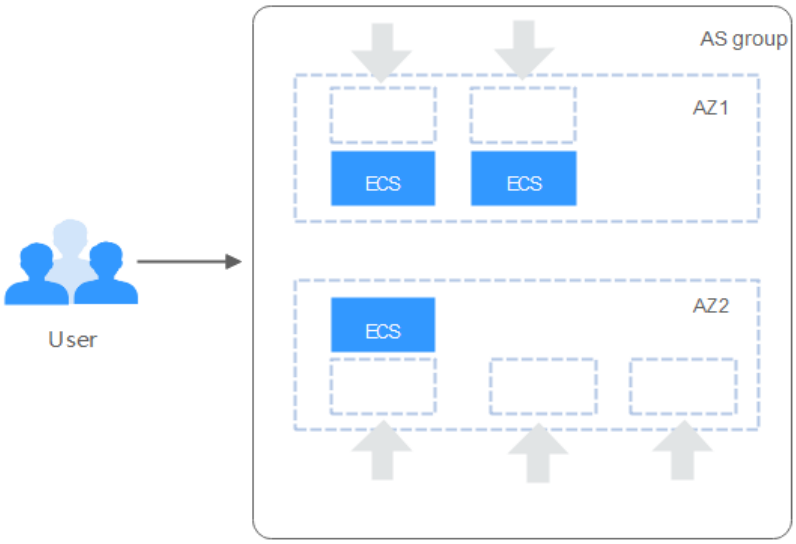


Rebalanceamento de instâncias entre AZs

Depois de adicionar ou remover manualmente instâncias de ou para um grupo de AS, o grupo de AS pode ficar desequilibrado entre as AZs. O AS compensa reequilibrando as AZs durante a próxima ação de dimensionamento.

Considere um grupo de AS contendo três instâncias distribuídas em AZ 1 e AZ 2, com duas em AZ 1 e uma em AZ 2. Se uma ação de dimensionamento for acionada para adicionar mais cinco instâncias ao grupo de AS, o AS adicionará duas a AZ 1 e três a AZ 2.

Figura 3-5 Reequilíbrio de instâncias



Gerenciamento de custos aprimorado

O AS permite que você use instâncias e larguras de banda do ECS sob demanda, dimensionando recursos automaticamente para suas aplicações, eliminando o desperdício de recursos e reduzindo custos.

Maior disponibilidade

O AS garante que você sempre tenha a quantidade certa de recursos disponíveis para lidar com a carga flutuante de suas aplicações.

Usar o ELB com o AS

Trabalhando com o ELB, o AS dimensiona automaticamente as instâncias do ECS com base nas alterações na demanda, garantindo que a carga de todas as instâncias em um grupo de AS permaneça equilibrada.

Depois que o ELB é ativado para um grupo de AS, o AS associa automaticamente um ouvinte de balanceamento de carga a quaisquer instâncias adicionadas ao grupo de AS. Em seguida, o ELB distribui automaticamente o tráfego para todas as instâncias saudáveis no grupo de AS por meio do ouvinte, o que melhora a disponibilidade do sistema. Se as instâncias no grupo de AS estiverem executando uma variedade de tipos diferentes de aplicações, você poderá vincular vários ouvintes de balanceamento de carga ao grupo de AS para ouvir cada um dessas aplicações, melhorando a escalabilidade do serviço.

Alta tolerância a falhas

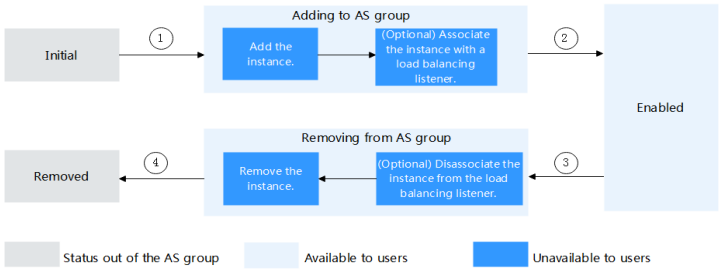
O AS monitora instâncias em um grupo de AS e substitui quaisquer instâncias não saudáveis que detecta por novas.

4 Ciclo de vida da instância

Uma instância do ECS em um grupo de AS passa por diferentes estátuas desde sua criação até sua remoção.

O status da instância será alterado conforme mostrado em **Figura 4-1** se você não tiver adicionado um gancho de ciclo de vida ao grupo de AS.

Figura 4-1 Ciclo de vida da instância



Quando a condição de gatilho 2 ou 4 é atendida, o sistema coloca as instâncias de forma autônoma no próximo status.

Tabela 4-1 Status da instância

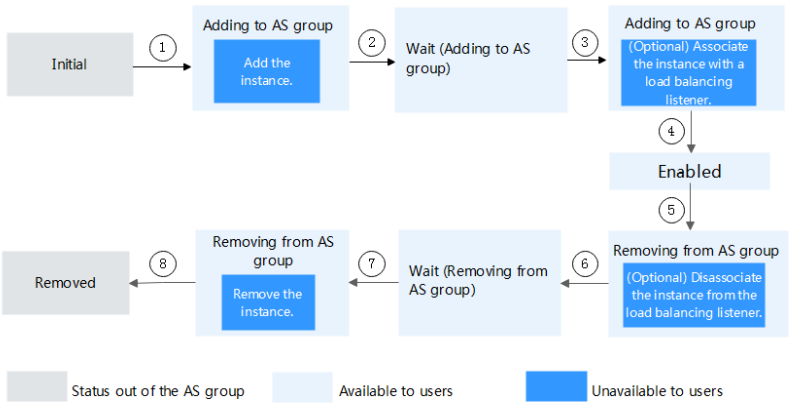
Status	Substatus	Descrição de status	Condição de gatilho
Initial	Nenhum	A instância não foi adicionada ao grupo de AS.	O status da instância será alterado para **Adding to AS group** quando qualquer uma das seguintes condições for atendida: ● Aumentar manualmente o número esperado de instâncias do grupo de AS.
Adding to AS group	Add the instance.	Quando a condição de gatilho 1 é atendida, o AS adiciona a instância para expandir a capacidade do grupo de AS.	

Status	Substatus	Descrição de status	Condição de gatilho
	(Optional) Associate the instance with a load balancing listener.	Quando a condição de gatilho 1 é atendida, o AS associa a instância criada ao ouvinte de balanceamento de carga.	● O sistema expande automaticamente a capacidade do grupo de AS. ● Você adiciona manualmente instâncias ao grupo de AS.
Enabled	Nenhum	A instância é adicionada ao grupo de AS e começa a processar o tráfego de serviço.	O status da instância é alterado de **Enabled** para **Removing from AS group** quando qualquer uma das seguintes condições é atendida: ● Diminuir manualmente o número esperado de instâncias do grupo de AS. ● O sistema remove automaticamente instâncias em uma ação de dimensionamento. ● Uma verificação de integridade mostra que uma instância ativada não é saudável e o sistema a remove do grupo de AS. ● Você remove manualmente instâncias do grupo de AS.
Removing from AS group	(Optional) Disassociate the instance from the load balancing listener.	Quando a condição de gatilho 3 é atendida, o grupo de AS começa a reduzir recursos e desassociar a instância do ouvinte de balanceamento de carga.	
	Remove the instance.	Depois que as instâncias são desvinculadas do ouvinte de balanceamento de carga, elas são removidas do grupo de AS.	
Removed	Nenhum	O ciclo de vida da instância no grupo de AS termino.	Nenhum

Quando uma instância do ECS é adicionada a um grupo AS manualmente ou por meio de uma ação de dimensionamento, ela passa pelos status **Adding to AS group**, **Enabled** e **Removing from AS group**. Em seguida, é finalmente removido do grupo de AS.

Se você adicionou um gancho de ciclo de vida ao grupo de AS, as estátuas de instância são alteradas conforme mostrado na **Figura 4-2**. Quando um evento de aumentar ou diminuir o dimensionamento ocorre no grupo de AS, as instâncias necessárias são suspensas pelo gancho do ciclo de vida e permanecem no status de espera até que o período de tempo limite termine ou você as chame manualmente de volta. Você pode executar operações personalizadas nas instâncias quando elas estiverem no status de espera. Por exemplo, você pode instalar ou configurar o software em uma instância antes de ele ser adicionado ao grupo de AS ou fazer download de arquivos de registros de uma instância antes de ser removido.

Figura 4-2 Ciclo de vida da instância



Na condição de gatilho 2, 4, 6 ou 8, o sistema altera automaticamente o status da instância.

Tabela 4-2 Status da instância

Status	Substatus	Descrição de status	Descrição do gatilho
Initial	Nenhum	A instância não foi adicionada ao grupo AS.	O status da instância é alterado para **Adding to AS group** quando ocorre uma das seguintes situações: ● Aumentar manualmente o número esperado de instâncias de um grupo de AS. ● O sistema adiciona instâncias automaticamente ao grupo de AS em uma ação de dimensionamento. ● Você adiciona manualmente instâncias ao grupo de AS.
Adding to AS group	Add an instance.	Quando a condição de gatilho 1 é atendida, AS adiciona a instância para expandir a capacidade do grupo de AS.	
Wait (Adding to AS group)	Nenhum	O gancho do ciclo de vida suspende a instância que está sendo adicionada ao grupo de AS e coloca a instância no estado de espera.	O status da instância é alterado de **Wait (Adding to AS group)** para **Adding to AS group** quando uma das seguintes operações é executada: ● A ação de retorno de chamada padrão é executada. ● Você executa manualmente a ação de retorno de chamada.

Status	Substatus	Descrição de status	Descrição do gatilho
Adding to AS group	(Optional) Associate the instance with a load balancing listener.	Quando a condição de gatilho 3 é atendida, o AS associa a instância ao ouvinte de balanceamento de carga.	
Enabled	Nenhum	A instância é adicionada ao grupo de AS e começa a processar o tráfego de serviço.	O status da instância é alterado de **Enabled** para **Removing from AS group** quando ocorre uma das seguintes situações: ● Reduzir manualmente o número esperado de instâncias de um grupo de AS. ● O sistema remove automaticamente instâncias em uma ação de dimensionamento. ● Uma verificação de integridade mostra que a instância não está saudável após ser ativada e o sistema a remove do grupo de AS. ● Você remove manualmente uma instância de um grupo de AS.
Removing from AS group	(Optional) Disassociate the instance from the load balancing listener.	Quando a condição de gatilho 5 é atendida, o grupo de AS começa a reduzir recursos e desassociar a instância do ouvinte de balanceamento de carga.	
Wait (Removing from AS group)	Nenhum	O gancho do ciclo de vida suspende a instância que está sendo removida do grupo de AS e define a instância como em estado de espera.	O status da instância é alterado de **Wait (Removing from AS group)** para **Removing from AS group** quando ocorre uma das seguintes situações: ● A ação de retorno de chamada padrão é executada. ● Você executa manualmente a ação de retorno de chamada.
Removing from AS group	Remove the instance.	Quando a condição de gatilho 7 é atendida, AS remove a instância do grupo de AS.	
Removed	Nenhum	O ciclo de vida da instância no grupo de AS termina.	Nenhum

As instâncias são adicionadas a um grupo de AS manualmente ou por meio de uma ação de dimensionamento. Em seguida, eles passam por **Adding to AS group**, **Wait (Adding to AS**

group), **Adding to AS group**, **Enabled**, **Removing from AS group**, **Wait (Removing from the AS group)** e **Removing from AS group** e são finalmente removidos do grupo AS.

5 Restrições

O AS tem as seguintes restrições:

- Somente aplicações sem estado e que podem ser dimensionados horizontalmente podem ser executados em instâncias em um grupo de AS.

  NOTA

  - Um processo ou uma aplicação sem estado pode ser entendido isoladamente. Não há conhecimento armazenado ou referência a transações passadas. Cada transação é feita como se fosse do zero pela primeira vez.

    As instâncias do ECS em que aplicações sem estado estão sendo executados não armazenam dados que precisam ser persistidos localmente.

    Pense em transações sem estado como uma máquina de venda automática: uma única solicitação e uma resposta.
  - Aplicações e processos com estado, no entanto, são aqueles que podem ser retornados repetidamente. Eles são realizados com o contexto de transações anteriores e a transação corrente pode ser afetada pelo que aconteceu durante as transações anteriores.

    As instâncias do ECS nas quais as aplicações com estado estão sendo executados armazenam dados que precisam ser persistidos localmente.

    Transações com estado são realizadas repetidamente, como banco on-line ou e-mail, que são realizadas com o contexto de transações anteriores.
- O AS pode liberar instâncias do ECS em um grupo de AS automaticamente, portanto, as instâncias não podem ser usadas para salvar informações de status da aplicação (como status da sessão) ou dados relacionados (como dados do banco de dados e registros). Se o status da aplicação ou dados relacionados precisarem ser salvos, você poderá armazenar as informações em servidores separados.
- O AS não suporta expansão de capacidade ou dedução de vCPUs e memória de instância.
- Os recursos do AS devem atender aos requisitos de cota listados na **Tabela 5-1**.

Tabela 5-1 Cotas

Item	Descrição	Padrão
Grupo de AS	Número máximo de grupos de AS por região por conta	10

Item	Descrição	Padrão
Configuração de AS	Número máximo de configurações de AS por região por conta	100
Política de AS	Número máximo de políticas de AS por grupo de AS	10
Instância	Número máximo de instâncias por grupo de AS	300
Política de dimensionamento de largura de banda	Número máximo de políticas de dimensionamento de largura de banda por região por conta	10

6 Região e AZ

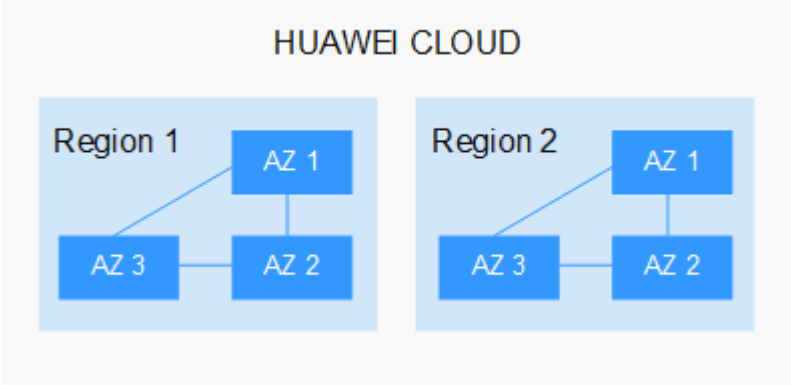
Conceito

Uma região e uma zona de disponibilidade (AZ) identificam a localização de um centro de dados. Você pode criar recursos em uma região e AZ específicas.

- As regiões são divididas com base na localização geográfica e na latência da rede. Serviços públicos, como Elastic Cloud Server (ECS), Elastic Volume Service (EVS), Object Storage Service (OBS), Virtual Private Cloud (VPC), Elastic IP (EIP) e Image Management Service (IMS), são compartilhados na mesma região. As regiões são classificadas em regiões universais e regiões dedicadas. Uma região universal fornece serviços de nuvem universal para locatários comuns. Uma região dedicada fornece serviços específicos para locatários específicos.
- Uma AZ contém um ou mais centros de data físicos. Cada AZ possui resfriamento, sistema de extinção de incêndio, proteção contra umidade e instalações elétricas independentes. Dentro de uma AZ, computação, rede, armazenamento e outros recursos são logicamente divididos em vários clusters. As AZs dentro de uma região são interconectadas usando fibras ópticas de alta velocidade, para suportar sistemas de alta disponibilidade entre AZs.

Figura 6-1 mostra a relação entre regiões e AZs.

Figura 6-1 Regiões e as AZs



HUAWEI CLOUD fornece serviços em muitas regiões do mundo. Selecione uma região e uma AZ com base nos requisitos. Para obter mais informações, consulte **Regiões globais do Huawei Cloud**.

Selecionar uma região

Ao selecionar uma região, considere os seguintes fatores:

- Localização

  É recomendável selecionar a região mais próxima para menor latência de rede e acesso rápido. As regiões dentro do continente chinês fornecem a mesma infraestrutura, qualidade de rede BGP, bem como operações e configurações de recursos. Portanto, se seus usuários-alvo estiverem no continente chinês, você não precisará considerar as diferenças de latência da rede ao selecionar uma região.

  - Se seus usuários-alvo estiverem na Ásia-Pacífico (excluindo o continente chinês), selecione a região **CN-Hong Kong**, **AP-Bangkok**, ou **AP-Singapore**.
  - Se seus usuários-alvo estão na África, selecione a região **AF-Johannesburg**.
  - Se seus usuários de destino estiverem na América Latina, selecione a região **LA-Santiago**.

    NOTA

    A região **LA-Santiago** está localizada no Chile.

- Preço do recurso

  Os preços dos recursos podem variar em diferentes regiões. Para obter detalhes, consulte **Detalhes de preço do produto**.

Selecionar uma AZ

Ao implantar recursos, considere os requisitos de recuperação de desastres (DR) e latência de rede de seus aplicativos.

- Para alta capacidade de DR, implante recursos nas diferentes AZs dentro da mesma região.
- Para menor latência de rede, implante recursos na mesma AZ.

Regiões e endpoints

Antes de usar uma API para chamar recursos, especifique sua região e endpoint. Para obter mais detalhes, consulte **Regiões e endpoints**.

7 Cobrança

Você pode usar o AS gratuitamente, mas as instâncias do ECS criadas automaticamente em um grupo de AS são cobradas com base em pagamento por uso. Para detalhes de preços, consulte Para detalhes de preços, consulte **Cobrança do ECS**. Os EIPs usados pelas instâncias também são cobrados. Para obter detalhes sobre preços, consulte **Cobrança do EIP**. Quando o grupo de AS é dimensionado, as instâncias criadas automaticamente serão removidas do grupo de AS e excluídas. Após a exclusão, essas instâncias não são mais cobradas. As instâncias adicionadas manualmente ainda são cobradas após serem removidas do grupo de AS. Se você não precisar dessas instâncias, cancele a assinatura delas no console do ECS.

Por exemplo, se duas instâncias forem criadas quando um grupo de AS for dimensionado, mas uma hora depois o grupo de AS for dimensionado novamente, as duas instâncias serão removidas do grupo de AS e você será cobrado por essa hora de uso.

8 O AS e outros serviços

O AS pode trabalhar com outros serviços em nuvem para atender às suas necessidades para diferentes cenários.

Figura 8-1 mostra as relações entre o AS e outros serviços.

Figura 8-1 Relações entre o AS e outros serviços

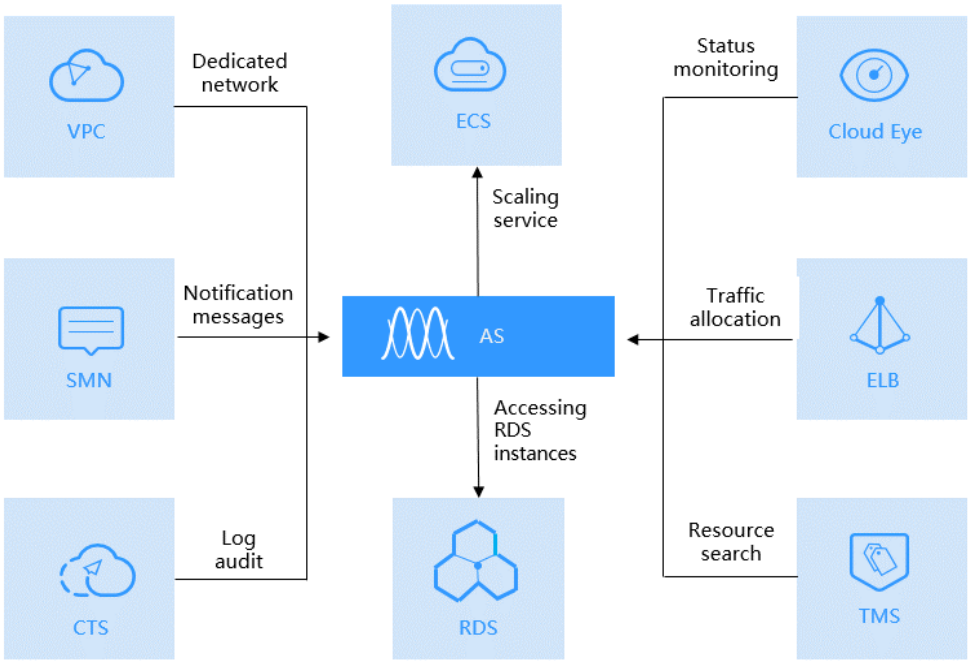


Tabela 8-1 Serviços relacionados

Serviço	Descrição	Interação	Referência
Elastic Load Balance (ELB)	Depois que o ELB é configurado, o AS associa automaticamente as instâncias do ECS a um ouvinte do balanceador de carga ao adicionar os ECSs e desvincula-as ao remover as instâncias. Para que o AS funcione com o ELB, o grupo de AS e o balanceador de carga devem estar na mesma VPC.	O AS distribui o tráfego para todos os ECSs em um grupo de AS.	**Adição de um balanceador de carga a um grupo de AS**
Cloud Eye	Se uma política acionada por alarme for configurada, o AS acionará ações de dimensionamento quando uma condição de alarme especificada no Cloud Eye for atendida.	O AS dimensiona recursos com base no status da instância do ECS monitorada pelo Cloud Eye.	**Métricas de AS**
ECS	As instâncias do ECS adicionadas em uma ação de dimensionamento podem ser gerenciadas e mantidas no console do ECS.	O AS ajusta automaticamente o número de instâncias do ECS.	**Recursos de expansão dinâmica**
Virtual Private Cloud (VPC)	O AS ajusta automaticamente as larguras de banda dos EIPs atribuídos em VPCs e também as larguras de banda compartilhadas.	O AS ajusta automaticamente a largura de banda.	**Criação de uma política de dimensionamento de largura de banda**

Serviço	Descrição	Interação	Referência
Simple Message Notification (SMN)	Se você ativar o serviço SMN, o sistema enviará notificações sobre o status do seu grupo de AS em tempo hábil.	Notificação de mensagem	**Configuração da notificação para um grupo de AS**
Cloud Trace Service (CTS)	Com o CTS, você pode gravar registros de operação de AS para exibição, auditoria e retrocesso.	Auditoria de registros	**Gravação de operações de recursos de AS**
Tag Management Service (TMS)	Se você tiver vários recursos do mesmo tipo, o TMS permite que você gerencie esses recursos com mais facilidade.	Tags	**Adição de tags a grupos e instâncias de AS**
Relational Database Service (RDS)	Os pré-requisitos para acessar diretamente uma instância de banco de dados do RDS a partir de uma instância dimensionada são os seguintes: ● A instância dimensionada e a instância de banco de dados do RDS de destino devem estar na mesma VPC. ● A instância dimensionada deve ser permitida pelo grupo de segurança para acessar instâncias de banco de dados do RDS.	As instâncias dimensionadas podem acessar instâncias de banco de dados do RDS.	**Conectar-se a uma instância de banco de dados do RDS para MySQL por meio de uma rede privada**

9 Gerenciamento de permissões

Se você precisar atribuir diferentes permissões aos funcionários de sua empresa para acessar seus recursos de AS, Identity and Access Management (IAM) é uma boa opção para gerenciamento de permissões refinado. O IAM fornece autenticação de identidade, gerenciamento de permissões e controle de acesso, ajudando você a acessar com segurança seus recursos da Huawei Cloud.

Com o IAM, você pode criar usuários do IAM e atribuir permissões aos usuários para controlar seu acesso a recursos específicos. Por exemplo, você pode atribuir permissões para permitir que alguns desenvolvedores de software usem recursos de AS, mas não permitir que eles excluam ou executem operações de alto risco nos recursos.

Se sua conta da Huawei Cloud não precisar de usuários individuais do IAM para gerenciamento de permissões, pule esta seção.

O IAM pode ser usado gratuitamente. Você paga apenas pelos recursos em sua conta. Para obter mais informações sobre o IAM, consulte **Visão geral de serviço do IAM**.

Permissões do AS

Por padrão, os novos usuários do IAM não têm nenhuma permissão atribuída. Você precisa adicioná-los a um ou mais grupos e anexar políticas ou funções a esses grupos para que esses usuários possam herdar permissões dos grupos e executar operações especificadas em serviços em nuvem.

Ao conceder permissões de AS a um grupo de usuários, defina **Scope** como **Region-specific projects** e selecione projetos (por exemplo, **ap-southeast-2** na região **AP-Bangkok** para que as permissões entrem em vigor. Se você selecionar **All projects**, as permissões entrarão em vigor para o grupo de usuários em todos os projetos específicos da região. Ao acessar o AS, os usuários precisam mudar para uma região onde tenham autorização para usar este serviço.

Você pode conceder permissões aos usuários usando funções e políticas.

- Funções: um tipo de mecanismo de autorização de granulação grosseira que define permissões relacionadas às responsabilidades do usuário. Esse mecanismo fornece apenas um número limitado de funções de nível de serviço para autorização. Ao usar funções para conceder permissões, você também precisa atribuir outras funções das quais as permissões dependem para entrar em vigor. No entanto, as funções não são uma escolha adequada para autorização refinada e controle de acesso seguro.
- Políticas: um tipo de mecanismo de autorização refinado que define as permissões necessárias para realizar operações em recursos de nuvem específicos sob determinadas

condições. Esse mecanismo permite uma autorização baseada em políticas mais flexível, atendendo aos requisitos de controle de acesso seguro. Por exemplo, você pode conceder aos usuários de AS apenas as permissões para gerenciar um determinado tipo de ECSs. A maioria das políticas define permissões com base em APIs. Para as ações de API suportadas pelo AS, consulte **Políticas de permissões e ações suportadas**.

Tabela 9-1 lista todas as políticas do sistema suportadas pelo AS.

Tabela 9-1 Permissões definidas pelo sistema suportadas pelo AS

Nome da política	Descrição	Categoria	Dependência
AutoScaling FullAccess	Todas as permissões de operação em todos os recursos de AS	Política definida pelo sistema	Nenhum
AutoScaling ReadOnlyAccess	Permissões somente leitura em todos os recursos de AS	Política definida pelo sistema	Nenhum
AutoScaling Administrator	Todas as permissões de operação em todos os recursos de AS	Função do sistema	As funções **ELB Administrator**, **CES Administrator**, **Server Administrator** e **Tenant Administrator** precisam ser atribuídas no mesmo projeto.

Tabela 9-2 lista as operações comuns suportadas por cada política de AS definida pelo sistema. Selecione as políticas conforme necessário.

Tabela 9-2 Operações comuns suportadas por cada política definida pelo sistema do AS

Operação	AutoScaling FullAccess	AutoScaling ReadOnlyAccess	AutoScaling Administrator
Criação de um grupo de AS	√	x	√
Modificação de um grupo de AS	√	x	√
Consulta de detalhes sobre um grupo de AS	√	√	√

Operação	AutoScaling FullAccess	AutoScaling ReadOnlyAccess	AutoScaling Administrator
Exclusão de um grupo de AS	√	x	√
Criação de uma configuração de AS	√	x	√
Criação de uma política de AS	√	x	√
Criação de uma política de dimensionamento de largura de banda	√	x	√

Links úteis

- **O que é o IAM?**
- **Criação de um usuário e concessão de permissões de AS**
- **Políticas de permissões e ações suportadas**

10 Conceitos básicos

Grupo de AS

Um grupo de AS consiste em uma coleção de instâncias do ECS que se aplicam ao mesmo cenário. É a base para habilitar ou desabilitar as políticas de AS e realizar ações de dimensionamento.

Configuração de AS

Uma configuração de AS é um modelo que especifica especificações para as instâncias do ECS a serem adicionadas a um grupo de AS. As especificações incluem o tipo de ECS, vCPUs, memória, imagem, modo de logon e disco.

Política de AS

As políticas de AS podem acionar ações de dimensionamento para ajustar o número de instâncias em um grupo de AS. Uma política de AS define a condição para acionar uma ação de dimensionamento e a operação a ser executada em uma ação de dimensionamento. Quando a condição de disparo é atendida, o sistema aciona automaticamente uma ação de dimensionamento.

Ação de dimensionamento

Uma ação de dimensionamento adiciona ou remove instâncias de um grupo de AS. Ele garante que o número esperado de instâncias esteja em execução no grupo de AS, adicionando ou removendo instâncias quando a condição de disparo for atendida, o que melhora a estabilidade do sistema.

Período de resfriamento

Para evitar que uma política baseada em alarmes seja acionada repetidamente pelo mesmo evento, configure um período de arrefecimento. Um período de arrefecimento especifica por quanto tempo qualquer ação de dimensionamento desencadeada por alarme será desativada após uma ação de dimensionamento anterior ser concluída. Este período de resfriamento não se aplica a ações de dimensionamento programadas ou periódicas.

Por exemplo, se você definir o período de resfriamento para 300 segundos (5 minutos) e houver uma ação de dimensionamento agendada para 10:32, mas uma ação de dimensionamento anterior tiver sido concluída às 10:30, Quaisquer ações de

dimensionamento desencadeadas por alarmes serão negadas durante o período de resfriamento das 10:30 às 10:35, mas a ação de dimensionamento programada ainda será acionada às 10:32. Se a ação de dimensionamento agendada terminar às 10:36, um novo período de resfriamento começa às 10:36 e termina às 10:41.

Dimensionamento da largura de banda

O AS ajusta automaticamente uma largura de banda com base nas políticas de dimensionamento que você configurou. O AS só pode ajustar as larguras de banda dos EIPs e compartilhar larguras de banda que são cobradas em uma base de pagamento por uso.

11 História de mudanças

Lançado em	Alterações
30/10/2021	Modificação do seguinte conteúdo: adição da seção "Gerenciamento de permissões."
19/10/2020	Modificação do seguinte conteúdo: adição da seção "Métodos de acesso."
30/09/2019	Esta edição é o segundo lançamento oficial.
19/11/2018	Esta edição é o primeiro lançamento oficial.

