Cloud Container Engine

Visão geral do serviço

Edição	01
Data	2024-09-10

HUAWEI CLOUD COMPUTING 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, os serviços e as funcionalidades adquiridos são estipulados pelo contrato estabelecido entre a Huawei Cloud e o cliente. Os produtos, os serviços e as funcionalidades descritos neste documento, no todo ou em parte, podem 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 ESTÃO" sem garantias ou representações de qualquer tipo, sejam expressas ou implícitas.

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.

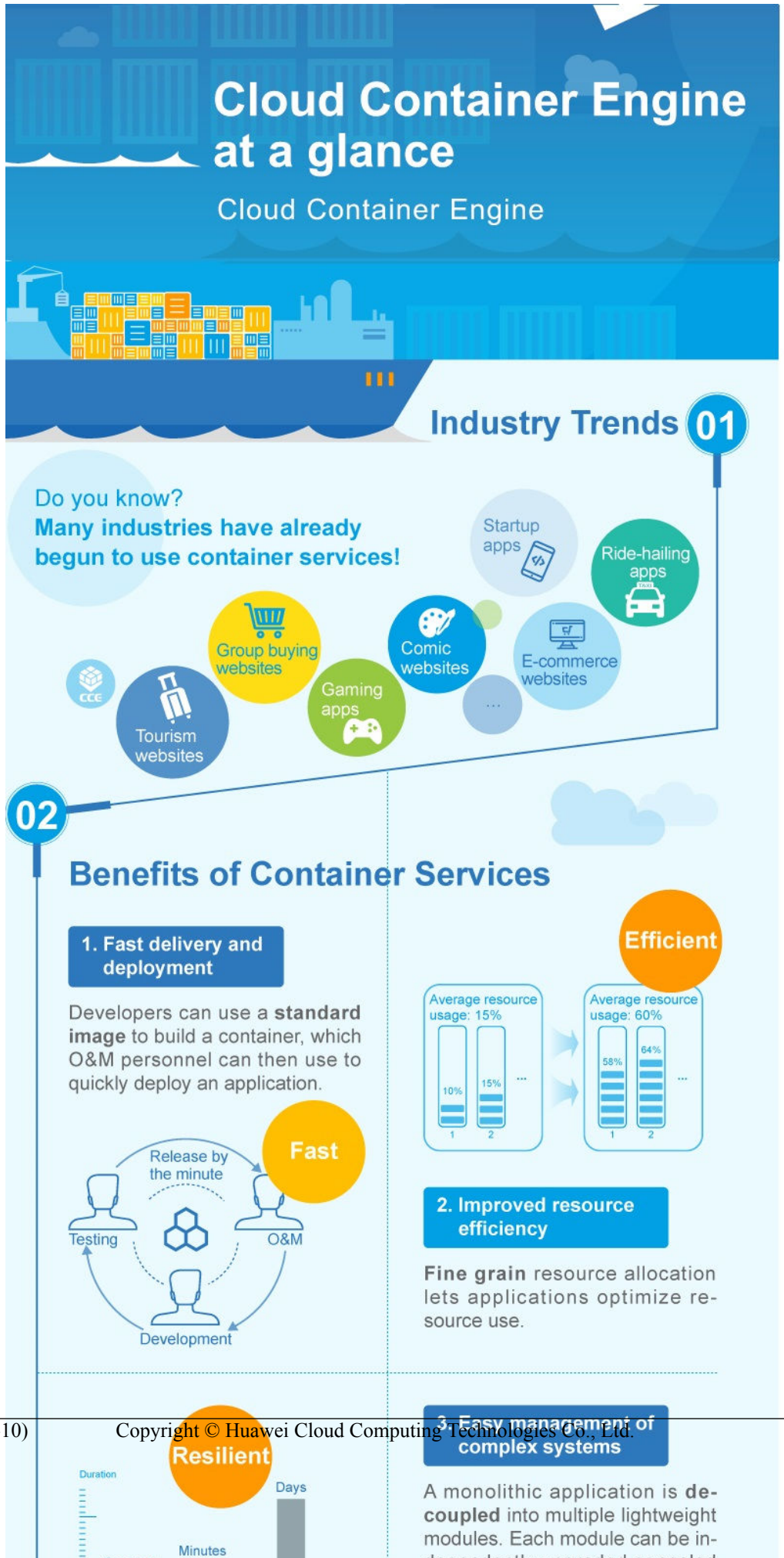
Huawei Cloud Computing Technologies Co., Ltd.

Endereço: Huawei Cloud Data Center, Rua Jiaoxinggong
Avenida Qianzhong
Novo Distrito de Gui'an
Guizhou 550029
República Popular da China

Site: https://www.huaweicloud.com/intl/pt-br/

Índice

1 Infográfico do CCE



2 O que é o Cloud Container Engine?

O Cloud Container Engine (CCE) é um serviço Kubernetes hospedado de classe empresarial altamente escalável para você executar containers e aplicações. Com o CCE, você pode implementar, gerenciar e dimensionar facilmente aplicações conteinerizadas em nuvem.

Por que o CCE?

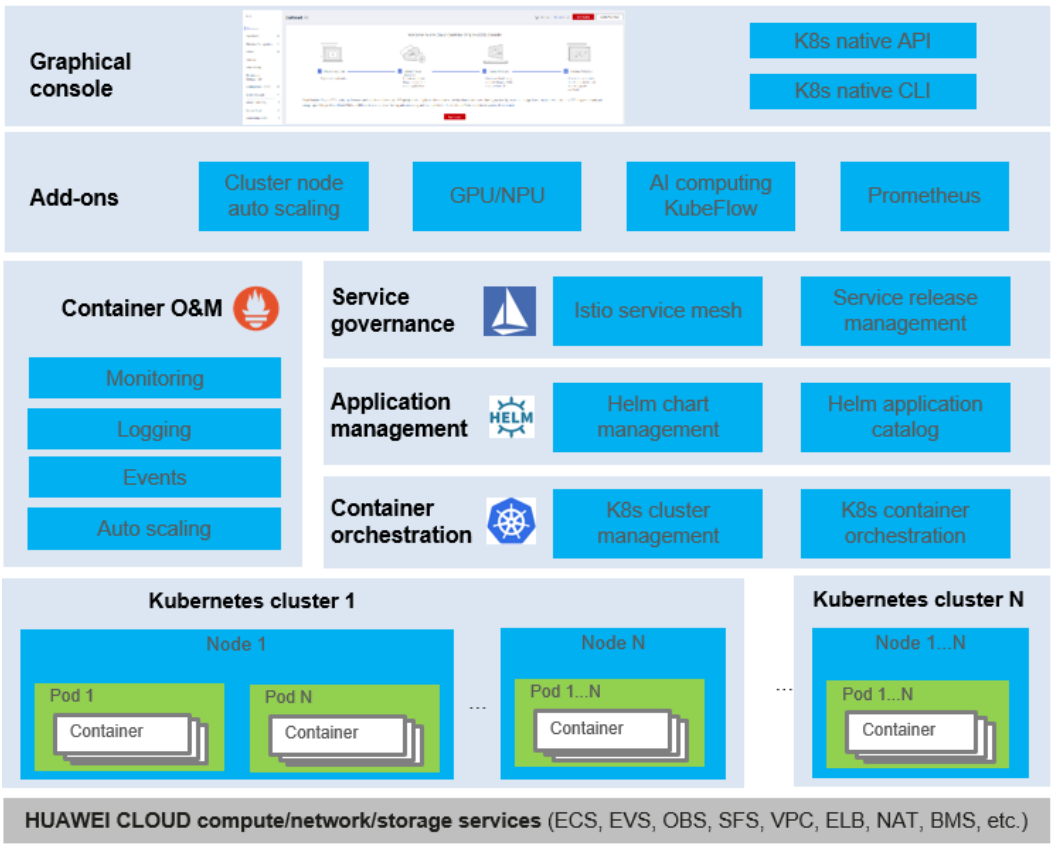
O CCE está profundamente integrada com serviços em nuvem, incluindo serviços de computação de alto desempenho (ECS//BMS), rede (VPC/EIP/ELB) e armazenamento (EVS/OBS/SFS). It supports heterogeneous computing architectures such as GPU, NPU, and Arm. Suportando recuperação de desastres multi-AZ e multi-região, o CCE garante alta disponibilidade de clusters do **Kubernetes**.

A Huawei Cloud é um dos primeiros Kubernetes Certified Service Providers (KCSPs) do mundo e o primeiro participante da China na comunidade Kubernetes. Há muito tempo que contribui para comunidades de container de código aberto e assume a liderança no ecossistema de container. A Huawei Cloud também é fundadora e membro platina da Cloud Native Computing Foundation (CNCF). O CCE é um dos serviços de container do mundo a ser o primeiro a passar no Certified Kubernetes Conformance Program.

Para mais informações, consulte **Vantagens do produto** e **Cenários de aplicação**.

Arquitetura do produto

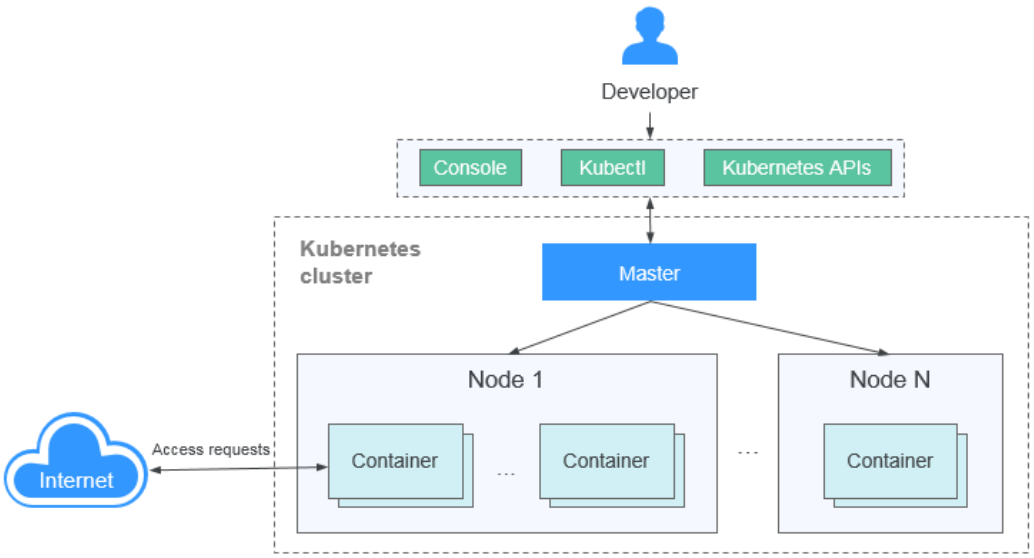
Figura 2-1 Arquitetura do CCE



Acessar ao CCE

Você pode usar o CCE por meio do console do CCE, do kubectl ou das APIs do Kubernetes. **Figura 2-2** mostra o processo.

Figura 2-2 Acessar ao CCE



Caminho de aprendizagem do CCE

Você pode clicar em **aqui** para aprender sobre os fundamentos sobre CCE para que você possa usar CCE e executar O&M com facilidade.

3 Vantagens do produto

Por que o CCE?

O CCE é um serviço de container baseado nas populares tecnologias Docker e Kubernetes e oferece uma variedade de recursos mais adequados à demanda das empresas por executar clusters de container em dimensionamento. Com vantagens exclusivas na confiabilidade do sistema, desempenho e compatibilidade com comunidades de código aberto, o CCE pode atender às particularidades de empresas interessadas em construir nuvens de container.

Facilidade de uso

- Criar um cluster do Kubernetes é tão fácil quanto alguns cliques na interface do usuário da Web (WebUI). O cluster do Kubernetes suporta o gerenciamento de nós de VM ou nós bare-metal e aplica-se ao cenário em que VMs e máquinas físicas são usadas juntas.
- A implementação automática e o O&M de aplicações conteinerizadas podem ser executados em um só lugar durante todo o ciclo de vida da aplicação.
- Clusters e as cargas de trabalho podem ser redimensionadas em apenas alguns cliques na WebUI. Qualquer política de dimensionamento automático pode ser combinada de forma flexível para lidar com picos de carga no momento.
- A WebUI orienta você pelas etapas necessárias para atualizar os clusters do Kubernetes.
- Suporte para Application Service Mesh (ASM) e gráficos de Helm oferece usabilidade pronto para usar.

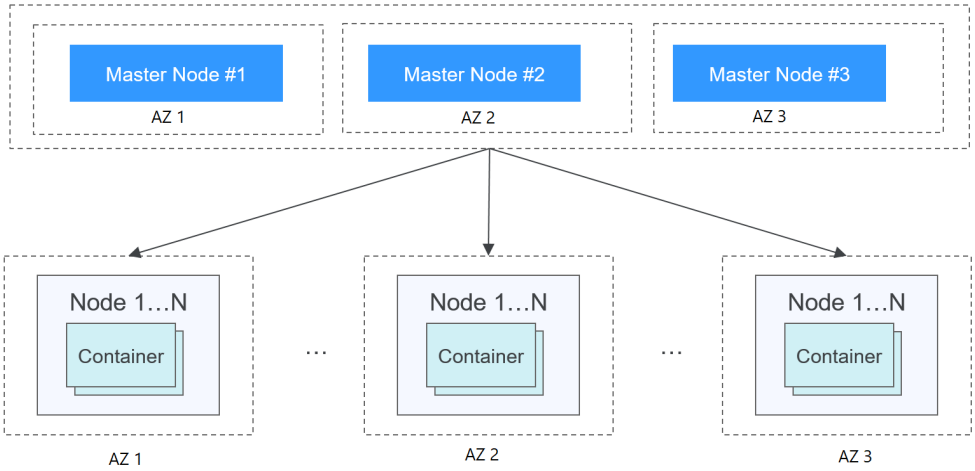
Alto desempenho

- O CCE se baseia em anos de experiência de campo em computação, rede, armazenamento e infraestrutura heterogênea. Você pode lançar containers em escala.
- A arquitetura bare-metal NUMA e as placas de rede InfiniBand de alta velocidade melhoram de três a cinco vezes o desempenho da computação.

Alta disponibilidade e seguro

- Alta confiabilidade: você pode implantar três nós principais em diferentes AZs para o plano de controle de cluster para garantir a alta disponibilidade dos seus serviços. Os nós e as cargas de trabalho em um cluster podem ser balanceadas entre as AZs para formar uma arquitetura multi ativa que garante a continuidade do serviço mesmo quando um dos hosts ou salas de equipamentos está inoperante, ou uma AZ é atingida por desastres naturais.

Figura 3-1 Configuração de alta disponibilidade de clusters



- Seguro: os clusters são privados e totalmente controlados por usuários com IAM e RBAC do Kubernetes profundamente integrados. Você pode definir permissões do RBAC diferentes para usuários do IAM no console.

Aberto e compatível

- CCE é baseado na tecnologia Docker de código aberto que automatiza a implementação, o agendamento de recursos, a descoberta de serviços e o dimensionamento dinâmico de aplicações conteinerizadas.
- CCE é baseado no Kubernetes e compatível com as APIs nativas do Kubernetes, kubectl (uma interface de linha de comando) e versões nativas do Kubernetes/Docker. As atualizações das comunidades do Kubernetes e do Docker são incorporadas regularmente ao CCE.

Análise comparativa do CCE e Sistemas de gerenciamento de cluster do Kubernetes no local

Tabela 3-1 Clusters do CCE versus clusters locais do Kubernetes

Área de foco	Sistemas de gerenciamento de cluster do Kubernetes no local	CCE
Facilidade de uso	O gerenciamento de cluster é complexo. Você precisa lidar com toda a complexidade na instalação, operação, escalabilidade, configuração e monitoramento da infraestrutura de gerenciamento de cluster do Kubernetes. Cada atualização de cluster exige um tremendo ajuste manual, impondo um fardo pesado ao pessoal de O&M.	**Fácil de gerenciar e usar clusters** Você pode criar e atualizar clusters de container do Kubernetes com apenas alguns cliques, sem precisar configurar ambientes Docker ou Kubernetes. A implementação automática e o O&M de aplicações conteinerizadas podem ser realizados no console em um só lugar durante todo o ciclo de vida da aplicação. O suporte para gráficos de Helm oferece usabilidade pronto para uso. Usar clusters do CCE é tão simples quanto escolher um cluster de container e os trabalhos que você deseja executar no cluster. Em seguida, o CCE completa o gerenciamento de clusters para que você possa se concentrar no desenvolvimento de aplicações conteinerizadas.
Escalabilidade	Você precisa avaliar manualmente a carga de serviço e a integridade do cluster antes de decidir redimensionar um cluster.	**Serviço de dimensionamento gerenciado** O CCE pode redimensionar automaticamente clusters e cargas de trabalho à medida que o uso de recursos muda. O uso combinado de políticas de dimensionamento automático pode dimensionar clusters e cargas de trabalho de forma flexível para atender às demandas flutuantes.
Confiabilidade	Apenas um nó principal está disponível em um cluster. Quando o nó principal estiver inativo, todo o cluster, bem como todas as aplicações no cluster, ficarão fora de serviço.	**Alta disponibilidade** Se **High Availability** estiver definida como **Yes** quando você criar um cluster, três nós principais serão criados no cluster, evitando pontos únicos de falha no plano de controle do cluster.

Área de foco	Sistemas de gerenciamento de cluster do Kubernetes no local	CCE
Eficiência	Você precisa criar repositórios de imagens ou reverter para repositórios de imagens de terceiros. As imagens são extraídas dos repositórios em série.	**Implementação rápida de imagens e integração contínua** O CCE funciona com o SoftWare Repository for Container (SWR) para suportar pipelines do DevOps e eliminar a necessidade de escrever manualmente Dockerfiles ou manifestos do Kubernetes. Com os modelos de pipeline do ContainerOps, você pode definir como criar imagens de container, enviá-las para repositórios e implantá-las. As imagens são extraídas dos repositórios em paralelo.
Custo	É necessário um investimento inicial pesado na instalação, gerenciamento e dimensionamento da infraestrutura de gerenciamento de clusters.	**Custo efetivo** Você paga apenas pelos recursos de infraestrutura necessários para armazenar e executar aplicações, bem como pelos nós principais no cluster.

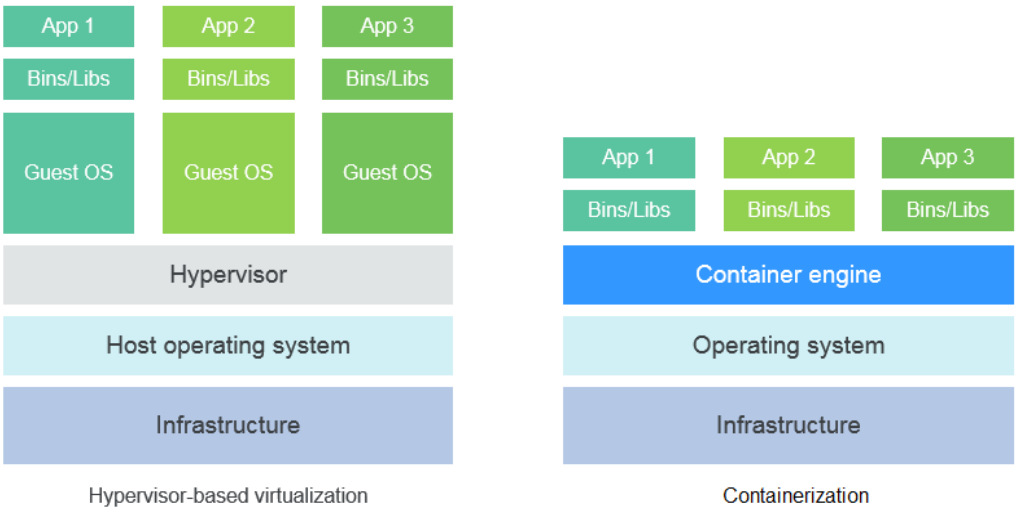
Por que Containers?

O Docker é escrito na linguagem de programação Go projetada pelo Google. Ele fornece virtualização no nível do sistema operacional: os processos de software são isolados uns dos outros usando grupos de controle do Linux (cgroups), namespaces e tecnologias Union FS (por exemplo, AUFS). Tudo o que é necessário para executar um processo de software é embalado em um container. Os containers são isolados uns dos outros e do host.

O Docker avançou para melhorar o isolamento de containers: os containers têm seus próprios sistemas de arquivos e não podem ver os processos ou interfaces de rede uns dos outros. Isso simplifica a criação e o gerenciamento de containers.

A tecnologia de virtualização tradicional fornece virtualização em nível de hardware. Ele cria um conjunto de máquinas virtuais, cada uma com um sistema operacional completo e aplicação dentro. Os containers, por outro lado, não têm seu próprio kernel e todos chamam para o mesmo kernel do sistema operacional host. Além disso, é desnecessário fazer qualquer tipo de virtualização da mesma forma que faz com as VMs. Portanto, os containers do Docker são menores e mais rápidos que as VMs.

Figura 3-2 Comparação entre containers do Docker e VMs



Para resumir, os containers do Docker têm muitas vantagens sobre as VMs.

Utilização de recurso

Sem sobrecarga para virtualizar hardware e executar um sistema operacional completo, os containers podem superar as VMs em velocidade de execução de aplicações, perda de memória e velocidade de armazenamento de arquivos.

Velocidade de iniciar

Leva vários minutos para iniciar uma aplicação em uma VM. As aplicações conteinerizadas do Docker são executados diretamente no kernel do host e não há necessidade de iniciar um sistema operacional completo junto com as aplicações. O tempo de inicialização pode ser reduzido para segundos ou até milissegundos, economizando muito seu tempo em desenvolvimento, teste e implementação.

Ambiente consistente

Um dos maiores problemas que os desenvolvedores sempre têm que lidar é a diferença nos ambientes onde executam suas aplicações. A diferença entre os ambientes de desenvolvimento, teste e produção impede que alguns bugs sejam descobertos antes do lançamento. Uma imagem de container do Docker inclui tudo o que é necessário para executar uma aplicação e isola a aplicação de seu ambiente. Portanto, as aplicações conteinerizadas sempre serão executadas da mesma forma em ambientes de desenvolvimento, teste e produção.

Entrega e implementação contínuas

Para o pessoal de DevOps, seria ideal se as aplicações pudessem ser executados em qualquer lugar após criação ou configuração única.

O Docker fornece criação e implementação confiáveis e frequentes de imagens de container com reversões rápidas e fáceis (devido à imutabilidade da imagem). Os desenvolvedores escrevem Dockerfiles que contêm todas as instruções necessárias para construir imagens de container e mesclar instruções atualizadas regularmente em Dockerfiles, uma prática conhecida como Integração contínua (CI). A equipe de operações pode implantar rapidamente imagens no ambiente de produção, permitindo que o Docker leia instruções do Dockerfiles. A equipe de Ops pode até mesmo seguir a prática de Entrega/Implementação contínua (CD), na

qual cada alteração de instrução é automaticamente criada, testada e, em seguida, enviada para um ambiente de teste que não seja de produção.

O uso de Dockerfiles torna o processo do DevOps visível para todos em uma equipe de DevOps. Desta forma, a equipe de desenvolvedores pode entender melhor as necessidades dos usuários e os problemas enfrentados pela equipe de Ops enquanto mantém a aplicação. Por outro lado, a equipe de Ops pode ter algum conhecimento das condições que devem ser atendidas para executar a aplicação. O conhecimento é útil quando o pessoal do Ops implementa imagens de container no ambiente de produção.

Portabilidade

O Docker garante a consistência ambiental no desenvolvimento, teste e produção, e assim os containers do Docker podem ser portáteis em qualquer lugar. Eles trabalham uniformemente, independentemente de serem executados em máquinas físicas, máquinas virtuais, nuvens públicas, nuvens privadas ou até mesmo laptops. Você pode migrar aplicações de uma plataforma para outra sem se preocupar que a mudança de ambiente fará com que as aplicações não funcionem.

Atualizações de aplicação

As imagens do Docker são compostas por camadas. Cada camada é armazenada apenas uma vez e imagens diferentes podem conter exatamente as mesmas camadas. Isso torna a distribuição eficiente porque as camadas que já foram transferidas como parte da primeira imagem não precisam ser transferidas novamente ao transferir a outra imagem que também possui essas camadas. Para atualizar uma aplicação conteinerizadada, você pode editar a camada superior mais gravável na imagem final ou adicionar camadas à imagem base. Além disso, o Docker colabora com equipes de projetos de código aberto para manter um grande número de imagens oficiais de alta qualidade. Você pode usá-los diretamente no ambiente de produção ou facilmente criar novas imagens com base neles.

Tabela 3-2 Containers versus VMs tradicionais

Característica	Containers	VMs
Velocidade de iniciar	Em segundos	Em minutos
Capacidade do disco	MB	GB
Desempenho	Desempenho quase nativo	Fraca
Capacidade por máquina	Milhares de containers	Dezenas de VMs

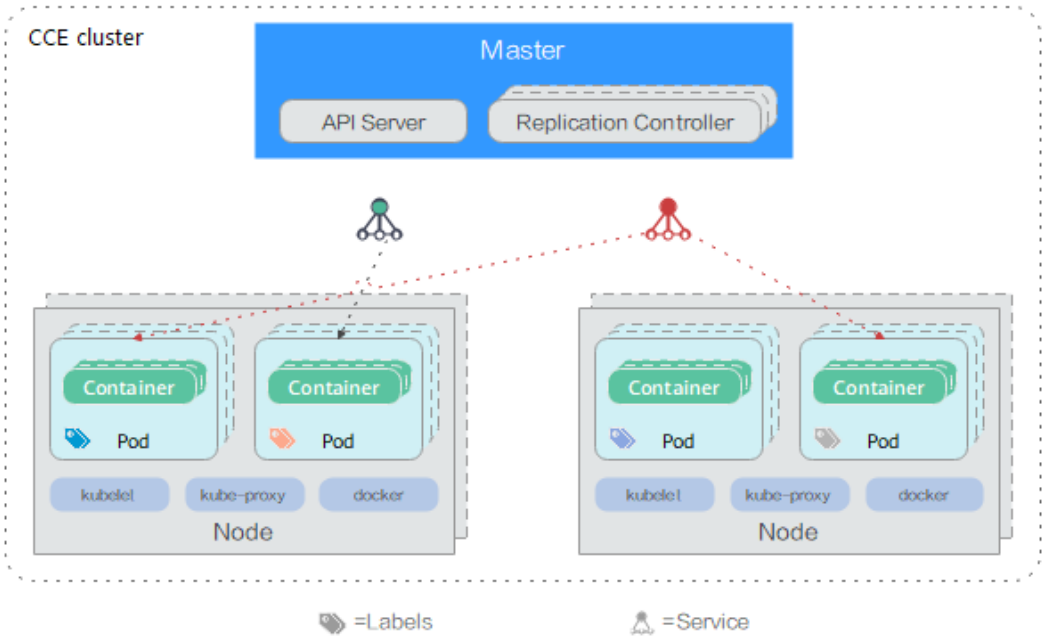
4 Cenários de aplicação

4.1 Gerenciamento de infraestrutura e da aplicação conteinerizada

Cenário de aplicação

Os clusters do CCE oferecem suporte ao gerenciamento de pools de recursos x86 e Arm. Você pode criar clusters de Kubernetes, implementar aplicações conteinerizadas e gerenciar e manter os clusters.

Figura 4-1 Cluster do CCE



Benefícios

A conteinerização reduz os custos de recursos de implementação de aplicações, agiliza a implementação e a atualização e realiza serviços ininterruptos durante as atualizações.

Vantagens

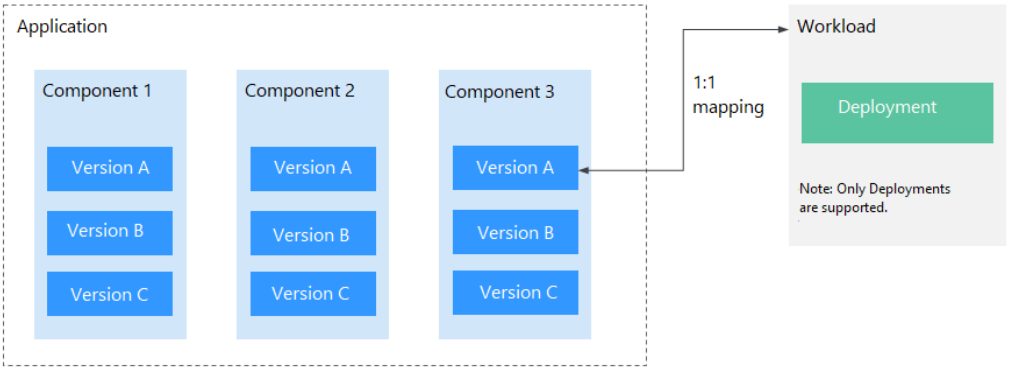
- Implementação de vários tipos de cargas de trabalho

  Suporta Implementações, StatefulSets, DaemonSets, tarefas e tarefas cronometradas.
- Atualização da aplicação

  Suporta atualização em substituição, atualização contínua por proporção ou por número de pods e reversão de atualização.
- Dimensionamento automático

  Suporta dimensionamento automático de nós e cargas de trabalho.

Figura 4-2 Carga de trabalho



4.2 Dimensionamento automático em segundos

Cenários de aplicação

- Surtos de tráfego trazidos por promoções e vendas flash em aplicativos de compras on-line e sites
- Flutuando cargas de serviço de transmissão ao vivo
- Aumento no número de jogadores de jogos que ficam on-line em determinados períodos de tempo

Benefícios

O CCE adapta automaticamente a quantidade de recursos de computação às cargas de serviço flutuantes de acordo com as políticas de dimensionamento automático configuradas. Para dimensionar recursos de computação no nível do cluster, o CCE adiciona ou reduz servidores em nuvem. Para dimensionar os recursos de computação no nível da carga de trabalho, o CCE adiciona ou reduz containers.

Vantagens

- Flexível

  Permite diversas políticas de dimensionamento e dimensiona containers em segundos quando as condições especificadas forem atendidas.
- Altamente disponível

  Detecta automaticamente o status de execução do pod em grupos de dimensionamento automático e substitui pods insalubres por novos.

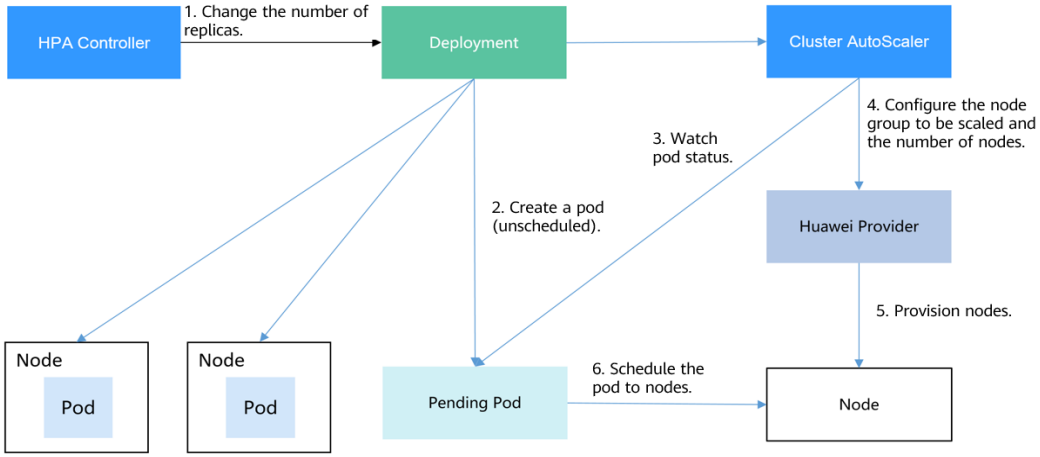
- Custos mais baixos

  Cobra apenas pelos servidores de nuvem que você usa.

Serviços relacionados

HPA (Autodimensionamento Horizontal de pod) + CA (Autodimensionamento de cluster)

Figura 4-3 Como funciona o dimensionamento automático



4.3 Gerenciamento de tráfego de micros serviços

Cenários de aplicação

Grandes sistemas corporativos estão se tornando mais complexos, além do que as arquiteturas de sistemas tradicionais podem lidar. Uma solução popular é o micro serviço. Aplicações complexas são divididas em componentes menores chamados micros serviços. Os micros serviços são desenvolvidos, implementados e dimensionados de forma independente. O uso combinado de micros serviços e containers agiliza a entrega de micros serviços, melhorando a confiabilidade e a escalabilidade das aplicações.

Os micros serviços possibilitam arquiteturas distribuídas. No entanto, mais micros serviços indicam mais complexidade em O&M, comissionamento e gerenciamento de segurança dessas arquiteturas. Os desenvolvedores são frequentemente incomodados por escrever código adicional para governança de micros serviços e integrar o código em seus sistemas de serviço. A este respeito, o CCE fornece uma solução eficiente para liberá-lo da carga de trabalho de gerenciamento.

Benefícios

O CCE está profundamente integrado ao Application Service Mesh (ASM), que permite concluir a liberação em tons de cinza, observar seu tráfego e controlar o fluxo de tráfego sem alterar seu código.

Vantagens

- Usabilidade pronto para usar

O ASM pode ser iniciado em apenas alguns cliques e funciona perfeitamente com o CCE para controlar o fluxo de tráfego de forma inteligente.

- Roteamento inteligente

  As políticas de conexão HTTP/TCP e as políticas de segurança podem ser aplicadas sem modificar o código.

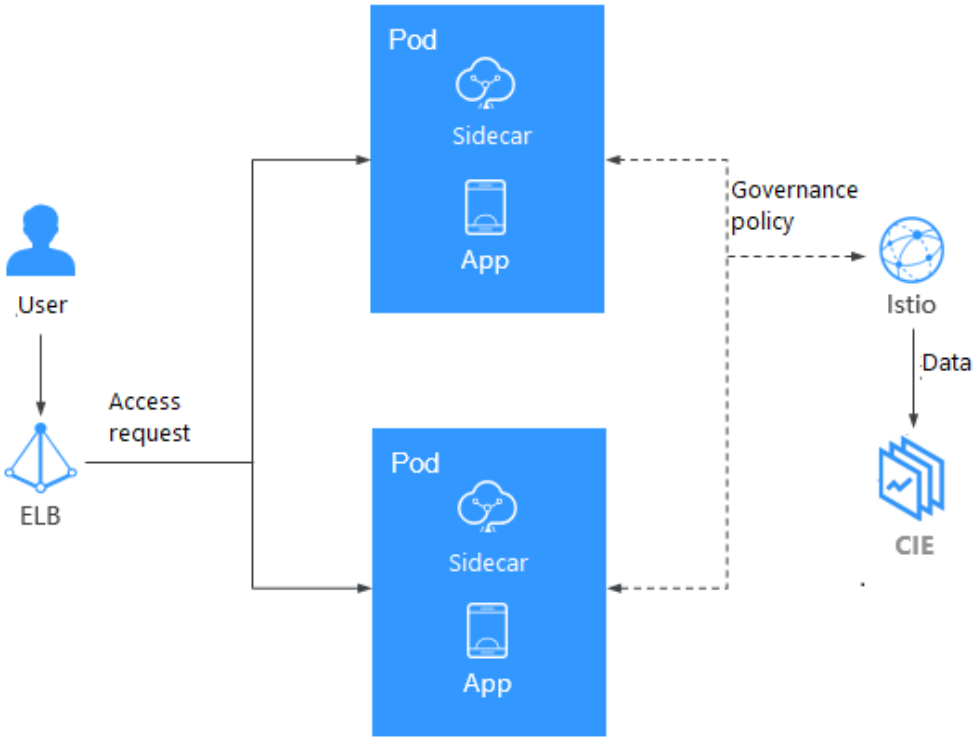
- Visibilidade no tráfego

  Com base nos dados de monitoramento que são coletados de forma não intrusiva, o ASM trabalha em conjunto com o Application Performance Management (APM) para fornecer uma visão panorâmica de seus serviços, incluindo topologia de tráfego em tempo real, rastreamento de chamadas, monitoramento de desempenho e diagnóstico de tempo de execução.

Serviços relacionados

Elastic Load Balance (ELB), Application Performance Management (APM) e Application Operations Management (AOM)

Figura 4-4 Governança de micros serviços



4.4 DevOps e CI/CD

Cenário de aplicação

Seus aplicativos e serviços podem receber muitos comentários e requisitos. Para lançar novos recursos e melhorar a experiência do usuário, você precisa de integração contínua (CI) rápida. Uma ferramenta eficiente para dar suporte a CI é o container. Ao implementar containers,

você pode agilizar o processo desde o desenvolvimento, o teste até o lançamento e a realização da entrega contínua (CD).

Benefícios

O CCE funciona com o SWR para suportar DevOps que completará automaticamente a compilação de código, a criação de imagem, a liberação em tons de cinza e a implementação com base no código-fonte. Os sistemas tradicionais de CI/CD podem ser conectados para conteinerizar aplicações legadas.

Vantagens

- Gerenciamento eficiente de processos

  Reduz a carga de trabalho de scripts em mais de 80% por meio da interação simplificada de processo.
- Integração flexível

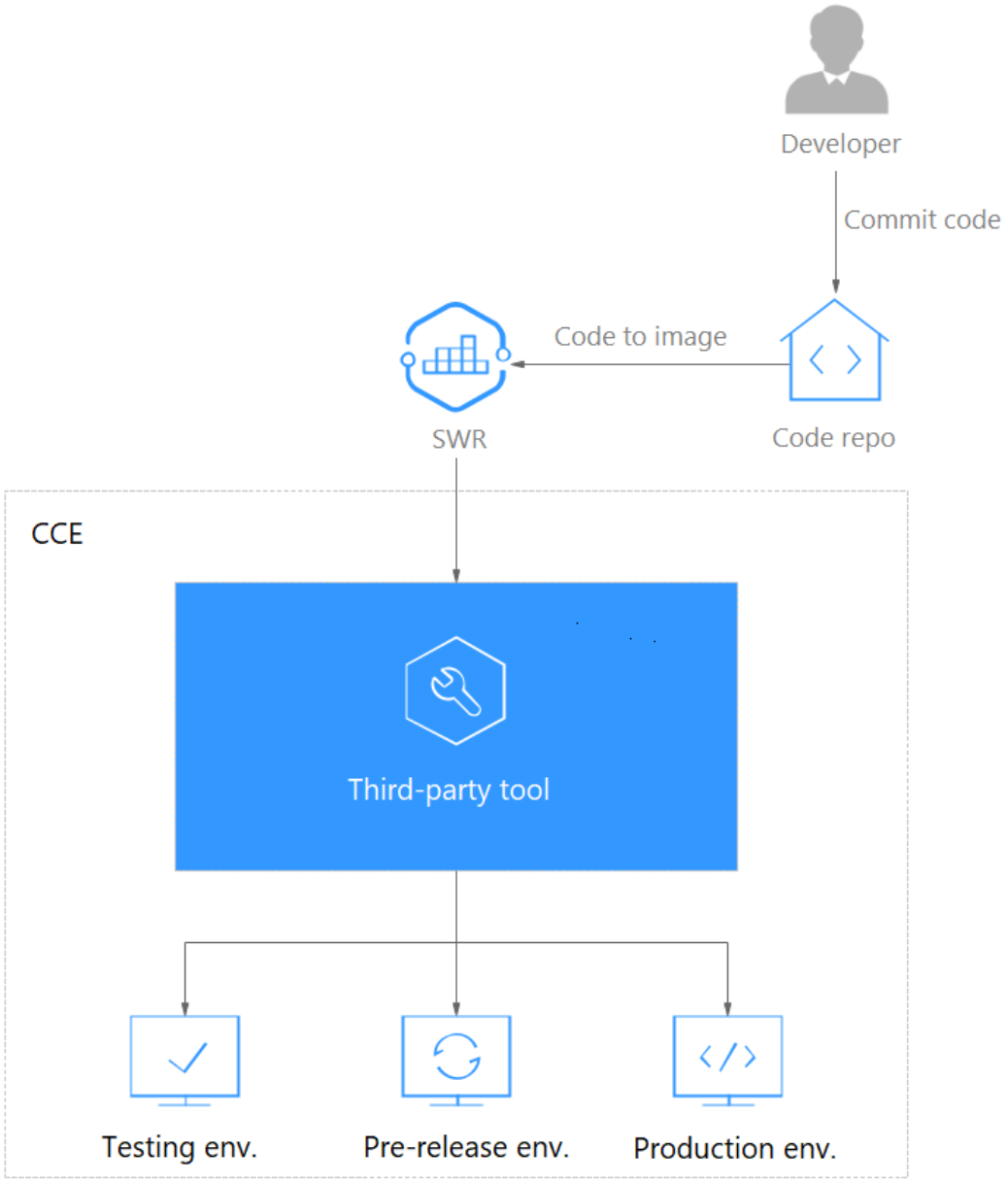
  Fornece várias APIs para integrar com sistemas de CI/CD existentes para personalização detalhada.
- Alto desempenho

  Agenda tarefas de forma flexível com uma arquitetura totalmente conteinerizada.

Serviços relacionados

Software Repository for Container (SWR), Object Storage Service (OBS) e Virtual Private Network (VPN)

Figura 4-5 Como funciona o DevOps



4.5 Arquitetura de nuvem híbrida

Cenários de aplicação

- Implementação em várias nuvens e recuperação de desastres

  Para obter alta disponibilidade do serviço, você pode implementar aplicações em serviços de container de vários provedores de nuvem. Quando uma nuvem está inativa, a carga de aplicações será distribuída automaticamente para outras nuvens.

- Distribuição de tráfego e dimensionamento automático

  Grandes sistemas corporativos precisam abranger instalações de nuvem em diferentes regiões. Eles também precisam ser redimensionáveis automaticamente — eles podem começar pequenos e depois aumentar o dimensionamento à medida que a carga do

sistema cresce. Isso libera as empresas dos custos de planejamento, compra e manutenção de mais instalações de nuvem do que o necessário e transforma grandes custos fixos em custos variáveis muito menores.

- Migração para a nuvem e hospedagem de banco de dados

  Finanças, segurança e outros setores com uma grande preocupação com a confidencialidade dos dados querem manter sistemas críticos em IDCs locais enquanto movem outros sistemas para a nuvem. Espera-se que todos os sistemas, independentemente dos IDCs locais ou da nuvem, sejam gerenciados usando um painel unificado.

- Separação de desenvolvimento de implementação

  Para garantir a segurança do IP, você pode configurar o ambiente de produção em uma nuvem pública e o ambiente de desenvolvimento em um IDC local.

Benefícios

Aplicações e dados podem ser migrados perfeitamente da rede local e para a nuvem, facilitando o agendamento de recursos e a recuperação de desastres (DR). Isso tornou-se possível graças aos containers independentes do ambiente, à conectividade de rede entre nuvens privadas e públicas e à capacidade de gerenciar containers coletivamente no CCE e em sua nuvem privada.

Vantagens

- DR na nuvem

  A multi-nuvem ajuda a proteger os sistemas contra interrupções. Quando uma nuvem está defeituosa, as cargas do sistema são automaticamente desviadas para outras nuvens para garantir a continuidade do serviço.

- Distribuição automática de tráfego

  A latência de acesso é reduzida ao direcionar as solicitações dos usuários para a nuvem regional mais próxima de onde os usuários estão. Uma vez que os aplicativos em IDCs locais são sobrecarregados, algumas das solicitações de acesso de aplicativos podem ser distribuídas para a nuvem com nós e containers dimensionados automaticamente.

- Implementações de serviços separadas e recursos compartilhados

  O CCE permite armazenamento separado para dados de serviço confidenciais e gerais, implantações separadas no ambiente de desenvolvimento e no ambiente de produção e execução separada de serviços gerais e de computação intensiva. Por meio do dimensionamento automático e do gerenciamento unificado de clusters, seus recursos locais e de nuvem podem trabalhar juntos com eficiência.

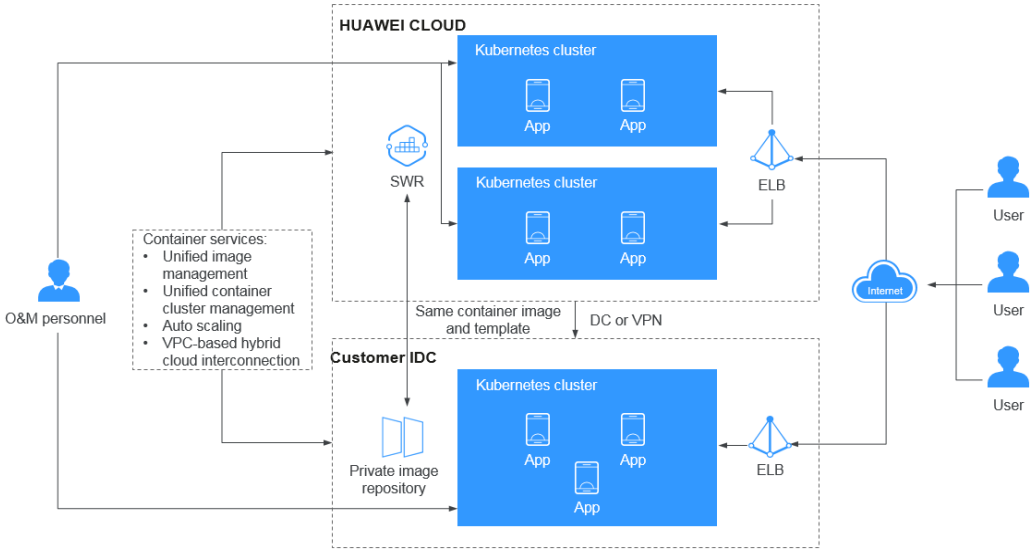
- Custos mais baixos

  Os pools de recursos de nuvem pública podem responder rapidamente a picos de carga por meio do provisionamento automático de recursos. Operações manuais e manutenção não são mais necessárias e você pode economizar muito.

Serviços relacionados

Elastic Cloud Server (ECS), Direct Connect (DC), Virtual Private Network (VPN) e Software Repository for Container (SWR)

Figura 4-6 Como funciona a nuvem híbrida



4.6 Agendamento de alto desempenho

CCE integra Volcano para suportar computação de alto desempenho.

Volcano é um sistema de processamento em lote nativo do Kubernetes. Volcano fornece uma plataforma universal, escalável e estável para executar trabalhos de Big Data e IA. É compatível com estruturas de computação gerais para IA, Big data, sequenciamento de genes e tarefas de renderização. A excelência de Volcano no agendamento de tarefas e no gerenciamento heterogêneo de chips torna a execução e o gerenciamento de tarefas mais eficientes.

Cenário de aplicação 1: implementação híbrida de vários tipos de trabalhos

Vários tipos de estruturas de domínio são desenvolvidos para suportar negócios em diferentes setores. Essas estruturas, como Spark, TensorFlow e Flink, funcionam de forma insubstituível em seus domínios de serviço. Elas não estão trabalhando sozinhos, pois os serviços e as empresas estão se tornando cada vez mais complexos. No entanto, o agendamento de recursos se torna uma dor de cabeça à medida que os clusters dessas estruturas crescem e um único serviço pode ter cargas flutuantes. Portanto, um sistema de agendamento unificado está em grande demanda.

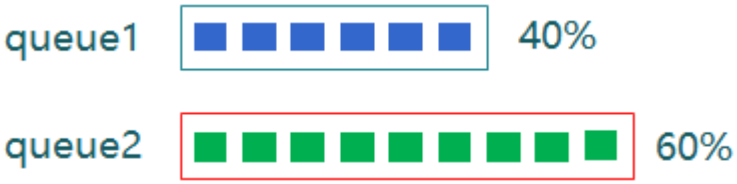
Volcano abstrai uma camada básica comum para computação em lote baseado em Kubernetes. Ele complementa o Kubernetes no agendamento e fornece abstrações de job flexíveis e universais para estruturas de computação. Essas abstrações (Volcano Jobs) são implementadas por meio de modelos de multitarefa para descrever vários tipos de trabalhos (como TensorFlow, Spark, MPI e PyTorch). Diferentes tipos de trabalhos podem ser executados juntos, e o Volcano usa seu sistema de agendamento unificado para realizar o compartilhamento de recursos do cluster.

Cenário de aplicação 2: otimização de agendamento em cenários de multi filas

O isolamento e o compartilhamento de recursos geralmente são necessários quando você usa um cluster do Kubernetes. No entanto, o Kubernetes não suporta filas. Ele não pode compartilhar recursos quando vários usuários ou departamentos compartilham uma máquina. Sem o compartilhamento de recursos baseado em fila, os trabalhos de HPC e Big data não podem ser executados.

Volcano suporta múltiplos mecanismos de compartilhamento de recursos com filas. Você pode definir o **weight** de uma fila. O cluster aloca recursos para a fila calculando a proporção entre o peso da fila e o peso total de todas as filas. Você também pode definir a **capability** do recurso de uma fila para determinar o limite superior de recursos que podem ser usados pela fila.

Por exemplo, na figura a seguir, a fila 1 é alocada 40% dos recursos de cluster e 60% para a fila 2. Desta forma, duas filas podem ser mapeadas para diferentes departamentos ou projetos para usar recursos no mesmo cluster. Se uma fila tiver recursos ociosos, eles poderão ser alocados para trabalhos em outra fila.

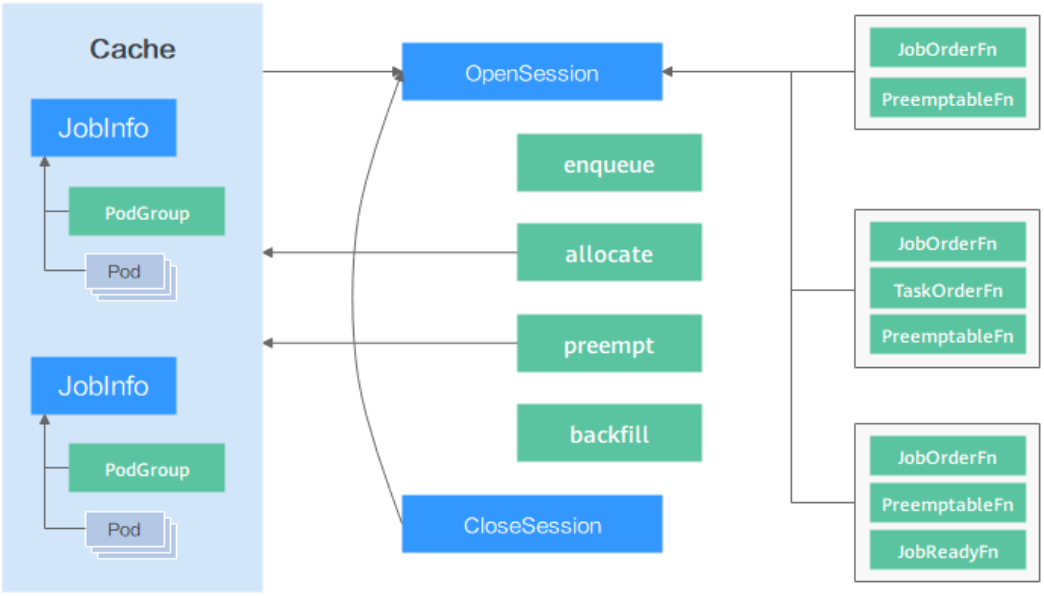


Cenário de aplicação 3: várias políticas de agendamento avançadas

Os containers são programados para nós que atendem aos requisitos de recursos de computação, como CPU, memória e GPU. Normalmente, haverá mais de um nó qualificado. Cada um pode ter um volume diferente de recursos disponíveis para novas cargas de trabalho. O Volcano analisa automaticamente a utilização de recursos de cada plano de programação e ajuda você a alcançar os melhores resultados de implementação com grande facilidade.

A figura a seguir mostra como o Volcano scheduler agenda recursos. Primeiro, o agendador carrega as informações do pod e do PodGroup no servidor de API para cache de agendador. Em uma sessão do agendador, Volcano passa por três fases: OpenSession, chamada de action e CloseSession. Na OpenSession, a política de agendamento que você configurou no plug-in

de agendador é carregada. Na chamada de action, as ações configuradas são chamadas uma a uma e a política de agendamento carregada é usada. Na CloseSession, as operações finais são executadas para concluir o agendamento.



O Volcano scheduler fornece plugins para suportar várias actions de agendamento (como enqueue, allocate, preempt, reclaim e backfill) e políticas de agendamento (como gang, priority, drf, proportion e binpack). Você pode configurá-las conforme necessário. As APIs fornecidas pelo agendador também podem ser usadas para desenvolvimento personalizado.

Cenário de aplicação 4: agendamento de recursos de alta precisão

O Volcano fornece políticas de agendamento de recursos de alta precisão para trabalhos de IA e Big data para melhorar a eficiência da computação. Tomemos TensorFlow como exemplo. Configurar afinidade entre ps e worker e anti-afinidade entre ps e ps, de modo que ps e worker para o mesmo nó. Isso melhora o desempenho de interação de rede e dados entre ps e worker, melhorando assim a eficiência de computação. No entanto, ao agendar pods, o agendador padrão do Kubernetes verifica apenas se as configurações de afinidade e anti-afinidade desses pods entram em conflito com as de todos os pods em execução no cluster, e não considera pods subsequentes que também podem precisar de agendamento.

O algoritmo task-topology fornecido pelo Volcano calcula as prioridades de task e node com base nas configurações de afinidade e anti afinidade entre tasks em um job. As políticas de afinidade e anti-afinidade de task em um job e o algoritmo task-topology garantem que as tasks com configurações de afinidade sejam preferencialmente agendadas para o mesmo node, e os pods com configurações anti afinidade sejam agendados para diferentes nós. A diferença entre o algoritmo task-topology e o agendador padrão do Kubernetes é que o algoritmo task-topology considera os pods como agendados como um todo. Quando os pods são agendados em lotes, as configurações de afinidade e anti afinidade entre os pods não agendados são consideradas e aplicadas aos processos de agendamento dos pods com base nas prioridades.

Benefícios

A execução de containers em servidores de nuvem acelerados por GPU de alto desempenho melhora significativamente o desempenho da computação de IA em três a cinco vezes. As GPUs podem custar muito e compartilhar uma GPU entre containers reduz muito os custos de

computação da IA. Além das vantagens de desempenho e custo, o CCE também oferece clusters totalmente gerenciados que ocultarão toda a complexidade na implementação e gerenciamento de suas aplicações de IA para que você possa se concentrar no desenvolvimento de alto valor.

Vantagens

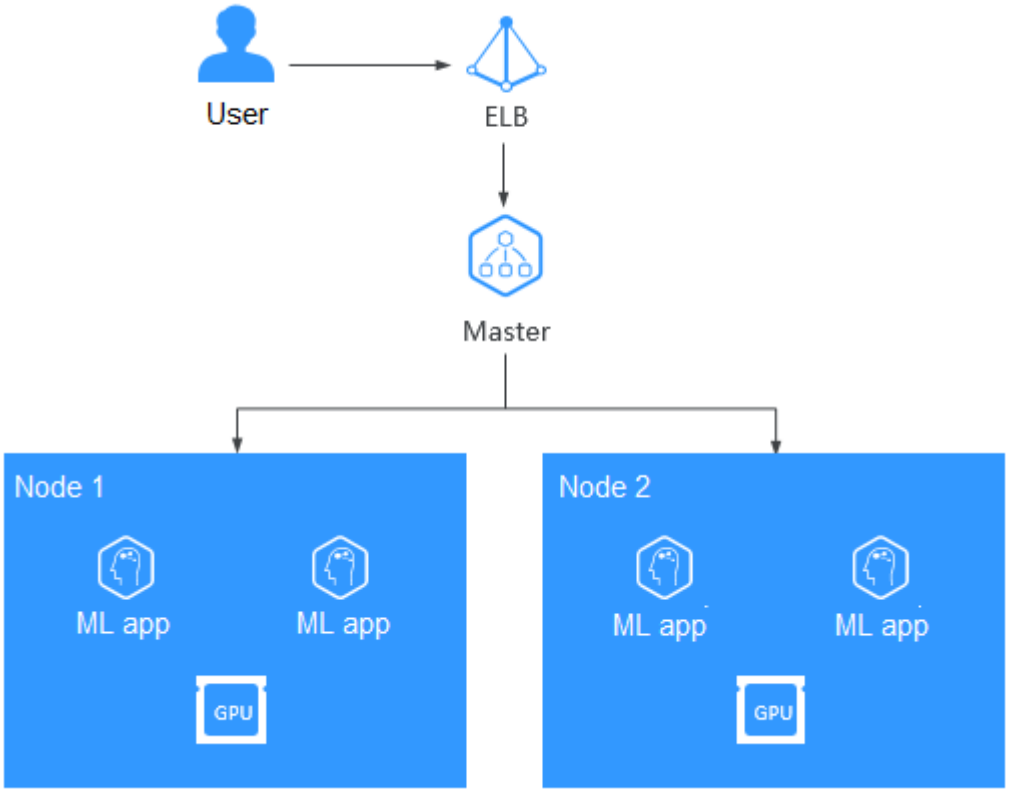
Ao integrar o Volcano, o CCE tem as seguintes vantagens na execução de tarefas de computação de alto desempenho, Big data e IA:

- **Implementação híbrida** de tarefas de HPC, Big data e IA
- **Agendamento otimizado de multi filas**: várias filas podem ser usadas para compartilhamento de recursos de vários locatários e planejamento de grupo com base em prioridades e períodos de tempo.
- **Políticas de agendamento avançadas**: gang scheduling, agendamento justo, preempção de recursos e topologia de GPU
- **Modelo de multitarefa**: você pode usar um modelo para definir várias tarefas em um único Volcano Job, além do limite de recursos nativos do Kubernetes. Volcano Jobs pode descrever vários tipos de job, como TensorFlow, MPI e PyTorch.
- **Plug-ins de extensão de tarefa**: o Volcano Controller permite configurar plug-ins para personalizar a preparação e limpeza do ambiente em etapas como envio de trabalhos e criação de pods. Por exemplo, antes de enviar uma tarefa MPI comum, você pode configurar o plug-in SSH para fornecer as informações SSH dos recursos do pod.

Serviços relacionados

GPU-accelerated Cloud Server (GACS), Elastic Load Balance (ELB) e Object Storage Service (OBS)

Figura 4-7 Como funciona a computação de IA



5 Observações e restrições

Esta seção descreve as observações e restrições sobre o uso de CCE.

Clusters e nós

- Depois que um cluster é criado, os seguintes itens não podem ser alterados:
  - Tipo do cluster. Por exemplo, altere um **cluster de Kunpeng** para um **cluster do CCE**.
  - Número de nós principais no cluster.
  - AZ de um nó principal.
  - Configuração de rede do cluster, como a VPC, sub-rede, bloco CIDR de container, bloco CIDR de serviço, configurações IPv6 e configurações kube-proxy (encaminhamento).
  - Modelo de rede. Por exemplo, altere a **rede de túneis** para a **rede da VPC**.
- As aplicações não podem ser migrados entre namespaces diferentes.
- Atualmente, as instâncias (nós) do ECS criadas são compatíveis com os modos de **pagamento por uso** e modo de cobrança **anual/mensal**. Outros recursos (como balanceadores de carga) são compatíveis com o modo de cobrança de pagamento por uso. Você pode alterar o modo de cobrança de pagamento por uso para anual/mensal no console de gerenciamento para as instâncias do ECS criadas.
- Os nós criados durante a criação do cluster oferecem suporte aos modos de **pagamento por uso** e modos de cobrança **anual/mensal**, mas com as seguintes restrições:
  - Se o cluster a ser criado for pay-per-use, os nós criados no cluster também deverão ser pay-per-use.
  - Se o cluster a ser criado for cobrado anualmente/mensalmente, os nós no cluster são pagos por uso ou cobrados anualmente/mensalmente.
  - Se os nós adicionados após a criação do cluster forem cobrados anualmente/mensalmente, eles precisarão ser renovados separadamente do cluster.

  Observação: se você comprar um nó depois que um cluster for criado, o modo de cobrança do nó não será restringido pelo do cluster.
- Os recursos subjacentes, como ECSs (nós), são limitados por cotas e seu inventário. Portanto, apenas alguns nós podem ser criados com êxito durante a criação do cluster, o escalonamento do cluster ou o dimensionamento automático.
- As especificações do ECS (nó) devem ter mais de 2 núcleos e 4 GB de memória.

- Para acessar um cluster do CCE por meio de uma VPN, certifique-se de que o bloco CIDR da VPN não entre em conflito com o bloco CIDR da VPC em que o cluster reside e o bloco CIDR do container.

Redes

- Por padrão, um serviço NodePort é acessado dentro de uma VPC. Se você precisar usar um EIP para acessar um serviço NodePort por meio de redes públicas, vincule um EIP ao nó do cluster com antecedência:
- Os Serviços LoadBalancer permitem que as cargas de trabalho sejam acessadas de redes públicas por meio do ELB. Este modo de acesso tem as seguintes restrições:
  - Balanceadores de carga criados automaticamente não devem ser usados por outros recursos. Caso contrário, esses balanceadores de carga não poderão ser completamente excluídos.
  - Não altere o nome do ouvinte do balanceador de carga em clusters v1.15 e anteriores. Caso contrário, o balanceador de carga não poderá ser acessado.
- Restrições nas políticas de rede:
  - Apenas os clusters que usam o modelo de rede de túnel suportam políticas de rede. As políticas de rede são classificadas nos seguintes tipos:
    - Ingress: todas as versões suportam este tipo.
    - Egress: este tipo de regra não pode ser definido atualmente.
  - O isolamento de rede não é suportado para endereços IPv6.

Volumes

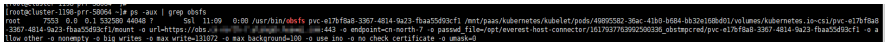
- Restrições em volumes do EVS:
  - Os discos EVS não podem ser conectados em AZs e não podem ser usados por várias cargas de trabalho, vários pods da mesma carga de trabalho ou várias tarefas. O compartilhamento de dados de um disco compartilhado não é suportado entre nós em um cluster do CCE. Se um disco EVS for atacado a vários nós, podem ocorrer conflitos de I/O e conflitos de cache de dados. Portanto, crie apenas um pod ao criar uma Implementação que use discos EVS.
  - Para clusters anteriores à v1.19.10, se uma política HPA for usada para expandir uma carga de trabalho com discos EVS anexados, os pods existentes não poderão ser lidos ou gravados quando um novo pod for agendado para outro nó.

    Para clusters de v1.19.10 e posterior, se uma política HPA for usada para expandir uma carga de trabalho com discos EVS anexados, um novo pod não poderá ser iniciado porque os discos EVS não podem ser anexados.
- Restrições em volumes do SFS:
  - Vários PVs podem usar o mesmo sistema de arquivos do SFS ou SFS Turbo com as seguintes restrições:
    - Não monte todos as PVCs/PVs que usam o mesmo sistema de arquivos do SFS ou SFS Turbo subjacente em um pod. Isso leva a uma falha de inicialização do pod porque nem todos as PVCs podem ser montados no pod devido aos mesmos valores de **volumeHandle** desses PVs.
    - Sugere-se que o parâmetro **persistentVolumeReclaimPolicy** nos PVs seja definido como **Retain**. Caso contrário, quando um PV for excluído, o volume subjacente associado poderá ser excluído. Neste caso, outros PVs associados ao mau funcionamento do volume subjacente.

  - Quando o volume subjacente é usado repetidamente, ative o isolamento e a proteção de ReadWriteMany na camada da aplicação para evitar a substituição e a perda de dados.

- Restrições em volumes do OBS:
  - Se forem utilizados volumes do OBS, o grupo proprietário e a permissão do ponto de montagem não podem ser modificados.
  - O CCE permite que sistemas de arquivos paralelos sejam montados usando SDKs ou PVCs do OBS. Se for usada montagem em PVC, a ferramenta obsfs fornecida pelo OBS deve ser usada. Um processo residente de obsfs é gerado cada vez que um volume de armazenamento de objetos gerado a partir de um sistema de arquivos paralelo é montado em um nó.

    Figura 5-1 Processo residente do obsfs

    Reserve 1 GiB de memória para cada processo de obsfs. Por exemplo, para um nó com 4 vCPUs e 8 GiB de memória, um sistema de arquivos paralelo obsfs deve ser montado em **no máximo** oito pods.

    NOTA

    - Um processo residente do obsfs é executado em um nó. Se a memória consumida exceder o limite superior do nó, o nó funciona mal. Em um nó com 4 vCPUs e 8 GiB de memória, se mais de 100 pods forem montados em um sistema de arquivos paralelo, o nó não estará disponível. Controle o número de pods montados em um sistema de arquivos paralelo em um único nó.
    - Ao usar obsfs, cumpra com as **Restrições de obsfs**.
  - Os contêineres de Kata não suportam volumes do OBS.
- Restrições em snapshots e backups:
  - A função snapshot está disponível **apenas para clusters v1.15 ou posterior** e requer o complemento everest baseado em CSI.
  - O subtipo (I/O comum, I/O alta ou I/O ultra-alta), modo de disco (SCSI ou VBD), criptografia de dados, estado de partilha, e a capacidade de um disco EVS criado a partir de um snapshot deve ser a mesma do disco associado ao snapshot. Esses atributos não podem ser modificados após serem consultados ou definidos.
  - Snapshots podem ser criados apenas para discos EVS disponíveis ou em uso, e um máximo de sete snapshots podem ser criados para um único disco EVS.
  - Snapshots só podem ser criados para PVCs criadas usando a classe de armazenamento (cujo nome começa com csi) fornecida pelo complemento everest. Não é possível criar instantâneos para PVCs criadas usando a classe de armazenamento Flexvolume cujo nome é ssd, sas ou sata.
  - Os dados de snapshot de discos criptografados são armazenados criptografados, e os de discos não criptografados são armazenados não criptografados.

Serviços

Um Serviço é um objeto de recurso do Kubernetes que define um conjunto lógico de pods e uma política pela qual acessá-los.

Um máximo de 6.000 Serviços podem ser criados em cada namespace.

Recursos do cluster do CCE

Há cotas de recursos para seus clusters do CCE em cada região.

Item	Restrições em usuários comuns
Número total de clusters em uma região	50
Número de nós em um cluster (escala de gerenciamento de cluster)	Você pode selecionar 50, 200, 1.000 ou 2.000 nós. Um máximo de 5.000 nós são suportados.
Número máximo de pods de container criados em cada nó de trabalho	Esse número pode ser definido no console quando você estiver criando um cluster. No modelo de rede da VPC, é possível criar no máximo 256 pods.

Recursos dependentes da nuvem subjacente

Categoria	Item	Restrições em usuários comuns
Computação	Pods	1.000
	Núcleos	8.000
	Capacidade de RAM (MB)	16.384.000
Redes	VPCs por conta	5
	Sub-redes por conta	100
	Grupos de segurança por conta	100
	Regras de grupo de segurança por conta	5.000
	Rotas por tabela de rotas	100
	Rotas por VPC	100
	Conexões de emparelhamento de VPC por região	50
	ACLs de rede por conta	200
	Gateways de conexão de camada 2 por conta	5
Balanceamento de carga	Balanceadores de carga elástico	50
	Ouvinte do balanceador de carga	100
	Certificados do balanceador de carga	120
	Políticas de encaminhamento do balanceador de carga	500

Categoria	Item	Restrições em usuários comuns
	Grupo de hosts de back-end do balanceador de carga	500
	Servidor back-end do balanceador de carga	500

6 Detalhes de preço

Itens cobrados

O Cloud Container Engine (CCE) é gratuito. Você só paga pelos recursos (como nós) criados quando estiver usando o CCE. Existem dois tipos de itens de faturamento:

1. **Clusters**: a taxa de cluster é o custo dos recursos usados pelos nós principais. A taxa varia de acordo com o tipo de cluster e o tamanho do cluster. Os tipos de cluster incluem cluster de VM e cluster de BMS (o número de nós principais determina se um cluster é altamente disponível). O tamanho do cluster (também chamado de escala de gerenciamento) indica o número máximo de nós permitidos em um cluster.

   NOTA

   A escala de gerenciamento indica o número de ECSs ou BMSs em um cluster.

   Para obter mais detalhes, consulte **Detalhes de definição de preço do CCE**.
2. **Recursos de IaaS**: o custo dos recursos de IaaS criados para executar nós de trabalho no cluster é cobrado. Os recursos de IaaS, que são criados manualmente ou automaticamente, incluem ECSs, discos do EVS, EIPs, largura de banda e balanceadores de carga.

   Para obter detalhes de preços, consulte **Detalhes de preços do produto**.

Modos de cobrança

O CCE é cobrado em um pagamento por uso ou anualmente/mensalmente.

- Pagamento por uso: é um modo de pagamento por uso. A cobrança começa quando um recurso é provisionado e pára quando o recurso é excluído. Você pode usar os recursos da nuvem conforme necessário e parar de pagar por eles quando não precisar mais deles. Não há pagamento adiantado por excesso de capacidade.

NOTA

A seguir estão os princípios de definição de preço no caso de hibernação de cluster do CCE ou desligamento de nó. Observe que há muitos tipos de nós de cluster e o ECS é usado como exemplo.

- **Hibernação do cluster**: depois que um cluster é hibernado, a cobrança de recursos usados pelos nós principais será interrompido.
- **Desligamento do nó**: a cobrança do nó de trabalho é interrompido quando o nó é interrompido. Observe que hibernar um cluster não interromperá os nós de trabalho no cluster. Para interromper um ECS, efetue logon no console do ECS. Para obter detalhes, consulte **Interrupção de um nó**.

  ECSs interrompidos não são cobrados. Para obter detalhes, consulte **Cobrança do ECS**.

- Anual/mensal: é um modo de pagamento antes de usar. A cobrança anual/mensal oferece um desconto mais significativo do que o pagamento por uso e é recomendado para o uso a longo prazo de serviços em nuvem. Quando você compra um pacote anual/mensal, o sistema deduzirá o custo do pacote da sua conta na nuvem com base nas especificações escolhidas.
- Alteração do modo de cobrança: o modo de cobrança não pode ser alterado dentro do ciclo de cobrança.

AVISO

- Os clusters seguem um plano de preços em camadas. O preço de cada camada varia de acordo com o tamanho e o tipo do cluster.
- Depois que uma assinatura mensal/anual expirar ou um recurso de pagamento por uso ficar em atraso, a HUAWEI CLOUD fornece um período de tempo durante o qual você pode renovar o recurso ou recarregar sua conta. Dentro do período de carência, você ainda pode acessar e usar seu serviço de nuvem. Para obter detalhes, consulte **O que é um período de carência? Quanto tempo dura o período de carência da HUAWEI CLOUD O que é um período de retenção? Qual é a duração do período de retenção da HUAWEI CLOUD**.

Alterações de configuração

Do pagamento por uso a cobrança anual/mensal: você pode alterar o modo de cobrança do cluster de pagamento por uso para cobrança anual/mensal. Após a alteração, os nós principais, os nós de trabalho e os recursos de nuvem (como discos do EVS e EIPs) usados pelo cluster serão cobrados anualmente/mensalmente e um novo pedido será gerado. Os nós e os recursos de nuvem estarão prontos para uso imediatamente após o pagamento do novo pedido.

Da cobrança anual/mensal ao pagamento por uso: os clusters cobrados anualmente/mensalmente não podem mudar para pagamento por uso dentro do ciclo de cobrança. Observe que os clusters de pagamento por uso podem ser excluídos diretamente, mas os clusters cobrados anualmente/mensalmente não podem ser excluídos. Para deixar de utilizar os clusters faturados anualmente/mensalmente, aceda ao Central de Cobrança e **cancele a subscrição deles**.

Observações

- Os cupons de dinheiro não serão devolvidos após você fazer downgrade das especificações dos servidores em nuvem que são comprados usando cupons de dinheiro.

- Você precisará pagar a diferença de preço entre as especificações originais e novas depois de atualizar as especificações do servidor em nuvem.
- O downgrade das especificações do servidor em nuvem (a quantidade de recursos de CPU ou memória) prejudicará o desempenho do servidor em nuvem.
- Se você fizer downgrade das especificações do servidor em nuvem e, em seguida, atualizá-lo para as especificações originais, ainda precisará pagar a diferença de preço incorrida pela atualização.

7 Gerenciamento de permissões

O CCE permite que você atribua permissões a usuários do IAM e grupos de usuários em suas contas de locatário. O CCE combina as vantagens do Identity and Access Management (IAM) e do Controle de Acesso Baseado em Função (RBAC) do Kubernetes para fornecer uma variedade de métodos de autorização, incluindo autorização refinada/token do IAM e autorização com escopo de cluster/namespace.

As permissões do CCE são descritas da seguinte forma:

- **Permissões em nível de cluster**: o gerenciamento de permissões em nível de cluster evolui a partir do recurso de autorização de política do sistema do IAM. Os usuários do IAM no mesmo grupo de usuários têm as mesmas permissões. No IAM, você pode configurar políticas do sistema para descrever quais grupos de usuários do IAM podem executar quais operações em recursos de cluster. Por exemplo, você pode conceder ao grupo de usuários A para criar e excluir o cluster X, adicionar um nó ou instalar um complemento, enquanto concede ao grupo de usuários B para exibir informações sobre o cluster X.

  As permissões em nível de cluster envolvem APIs do CCE que não são do Kubernetes e oferecem suporte a políticas de IAM refinadas e recursos de gerenciamento de projetos corporativos.
- **Permissões em nível de namespace**: você pode regular o acesso de usuários ou grupos de usuários a **recursos do Kubernetes**, como cargas de trabalho, trabalhos e serviços, em um único namespace com base em suas funções do RBAC do Kubernetes. O CCE também foi aprimorado com base em recursos de código aberto. Ele suporta autorização do RBAC com base no usuário ou grupo de usuários do IAM e autenticação do RBAC no acesso a APIs usando tokens do IAM.

  As permissões em nível de namespace envolvem as APIs do Kubernetes do CCE e são aprimoradas com base nos recursos do RBAC do Kubernetes. As permissões em nível de namespace podem ser concedidas a usuários ou grupos de usuários do IAM para autenticação e autorização, mas são independentes de políticas do IAM refinadas. Para obter detalhes, consulte **Uso da autorização do RBAC**.

⚠ CUIDADO

- **Permissões em nível de cluster** são configuradas apenas para recursos relacionados ao cluster (como clusters e nós). Você também deve configurar **permissões de namespace** para operar recursos do Kubernetes (como cargas de trabalho, jobs e Services).
- Depois de criar um cluster v1.11.7-r2 ou posterior, o CCE atribui automaticamente as permissões de administração de cluster de todos os namespaces no cluster para você, o que significa que você tem controle total sobre o cluster e todos os recursos em todos os namespaces.

Permissões no nível do cluster (atribuídas usando políticas do sistema do IAM)

Por padrão, os novos usuários do IAM não têm permissões atribuídas. Você precisa adicionar um usuário a um ou mais grupos e anexar políticas de permissões ou funções a esses grupos. Os usuários herdam permissões dos grupos aos quais são adicionados e podem executar operações especificadas em serviços de nuvem com base nas permissões.

O CCE é um serviço de nível de projeto implementado e acessado em regiões físicas específicas. Para atribuir permissões do CCE a um grupo de usuários, especifique o escopo como projetos específicos da região e selecione os projetos para que as permissões entrem em vigor. Se **All projects** estiver selecionado, as permissões entrarão em vigor para o grupo de usuários em todos os projetos específicos da região. Ao acessar o CCE, os usuários precisam mudar para uma região onde foram autorizados a usar o serviço CCE.

Você pode conceder permissões dos 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.
- Polícias: um tipo de mecanismo de autorização refinado que define as permissões necessárias para realizar operações em recursos em 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 atribuir aos usuários apenas as permissões para gerenciar um determinado tipo de clusters e nós.

Tabela 7-1 lista todas as permissões do sistema suportadas pelo CCE.

Tabela 7-1 Permissões do sistema suportadas pelo CCE

Nome da função/ política	Descrição	Tipo	Dependências
CCE Administrator	Permissões de leitura e gravação para clusters do CCE e todos os recursos (incluindo cargas de trabalho, nós, trabalhos e serviços) nos clusters	Função	Os usuários que recebem permissões desta política também devem receber permissões das seguintes políticas: **Projeto de serviço global**: OBS Buckets Viewer e OBS Administrator **Projetos específicos por região**: Tenant Guest, Server Administrator, ELB Administrator, SFS Administrator, SWR Admin e APM FullAccess **NOTA** Os usuários com políticas de **CCE Administrator** e **NAT Gateway Administrator** podem usar as funções de gateway NAT para clusters.
CCE FullAccess	Permissões de operação comuns em recursos de cluster do CCE, excluindo as permissões em nível de namespace para os clusters (com o RBAC do Kubernetes ativado) e as operações de administrador privilegiado, como configuração de agência e geração de certificados de cluster	Política	Nenhuma.
CCE ReadOnly Access	Permissões para visualizar recursos de cluster do CCE, excluindo as permissões de nível de namespace dos clusters (com o RBAC do Kubernetes ativado)	Política	Nenhuma.

Tabela 7-2 Operações comuns suportadas por políticas do sistema do CCE

Operação	CCE ReadOnlyAccess	CCE FullAccess	CCE Administrator
Criar um cluster	x	√	√
Excluir um cluster	x	√	√
Atualizar um cluster, por exemplo, atualizar parâmetros de programação de nó de cluster e fornecer suporte do RBAC para clusters	x	√	√
Atualizar um cluster	x	√	√
Acordar um cluster	x	√	√
Hibernar um cluster	x	√	√
Listar todos os clusters	√	√	√
Consultar detalhes do cluster	√	√	√
Adicionar um nó	x	√	√
Excluir um ou mais nós	x	√	√
Atualizar um nó de cluster, por exemplo, atualizar o nome do nó	x	√	√
Consultar detalhes do nó	√	√	√
Listar todos os nós	√	√	√
Listar todas as tarefas	√	√	√
Excluir uma ou mais tarefas de cluster	x	√	√
Consultar detalhes da tarefa	√	√	√
Criar um volume de armazenamento	x	√	√
Excluir um volume de armazenamento	x	√	√
Executar operações em todos os recursos do Kubernetes	√	√	√

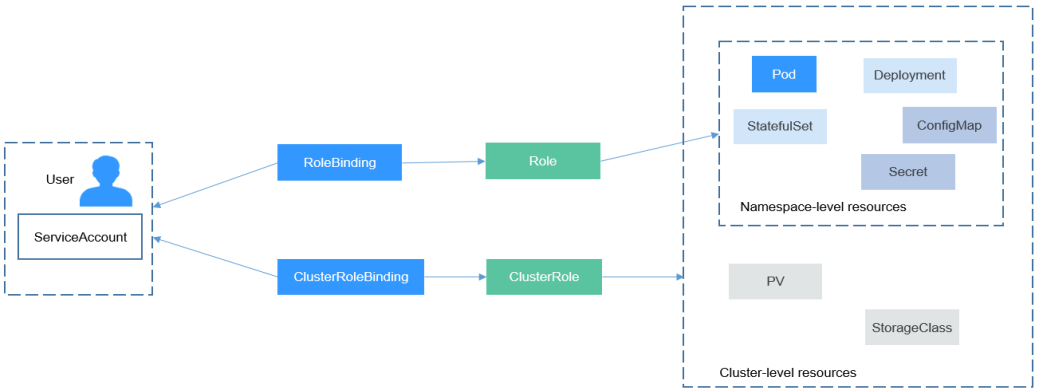
Operação	CCE ReadOnlyAccess	CCE FullAccess	CCE Administrator
Executar todas as operações em um Elastic Cloud Server (ECS)	x	√	√
Executar todas as operações em discos do EVS (Elastic Volume Service) Os discos do EVS podem ser conectados a servidores em nuvem e dimensionados para uma capacidade maior sempre que necessário.	x	√	√
Executar de todas as operações na VPC Um cluster deve ser executado em uma VPC. Ao criar um namespace, você precisa criar ou associar uma VPC para o namespace para que todos os containers no namespace sejam executados na VPC.	x	√	√
Visualizar detalhes de todos os recursos em um ECS No CCE, um nó é um ECS com vários discos do EVS.	√	√	√
Listar todos os recursos em um ECS	√	√	√
Visualizar detalhes sobre todos os recursos de disco do EVS. Os discos do EVS podem ser conectados a servidores em nuvem e dimensionados para uma capacidade maior sempre que necessário.	√	√	√
Listar todos os recursos do EVS	√	√	√

Operação	CCE ReadOnlyAccess	CCE FullAccess	CCE Administrator
Exibir detalhes sobre todos os recursos da VPC Um cluster deve ser executado em uma VPC. Ao criar um namespace, você precisa criar ou associar uma VPC para o namespace para que todos os containers no namespace sejam executados na VPC.	√	√	√
Listar todos os recursos da VPC	√	√	√
Exibindo detalhes sobre todos os recursos do Elastic Load Balance (ELB)	x	x	√
Listar todos os recursos do ELB	x	x	√
Visualizar detalhes do recurso do Scalable File Service (SFS)	√	√	√
Listar todos os recursos do SFS	√	√	√
Exibir detalhes do recurso do Application Operations Management (AOM)	√	√	√
Listar recursos do AOM	√	√	√
Executar todas as operações nas regras de escala automática do AOM	√	√	√

Permissões em nível de namespace (atribuídas usando o RBAC do Kubernetes)

Você pode regular o acesso de usuários ou grupos de usuários aos recursos do Kubernetes em um único namespace com base em suas funções do RBAC do Kubernetes. A API do RBAC declara quatro tipos de objetos do Kubernetes: Role, ClusterRole, RoleBinding e ClusterRoleBinding, que são descritos a seguir:

Figura 7-1 Role binding



No console do CCE, você pode atribuir permissões a um usuário ou grupo de usuários para acessar recursos em um ou vários namespaces. Por padrão, o console do CCE fornece os cinco ClusterRoles a seguir:

- view: tem permissão para exibir recursos do namespace.
- edit: tem permissão para modificar recursos do namespace.
- admin: tem todas as permissões no namespace.
- cluster-admin: tem todas as permissões no cluster.
- psp-global: controla aspectos de segurança sensíveis da especificação do pod.

Além de cluster-admin, admin, edit e view, você pode definir Roles e RoleBindings para configurar as permissões para adicionar, excluir, modificar e consultar recursos, como pods, Implementações e Serviços, no namespace.

Links úteis

- **Visão geral de serviço do IAM**
- **Concessão de permissões em nível de cluster**
- **Políticas de permissões e ações suportadas**

8 Regiões e as AZs

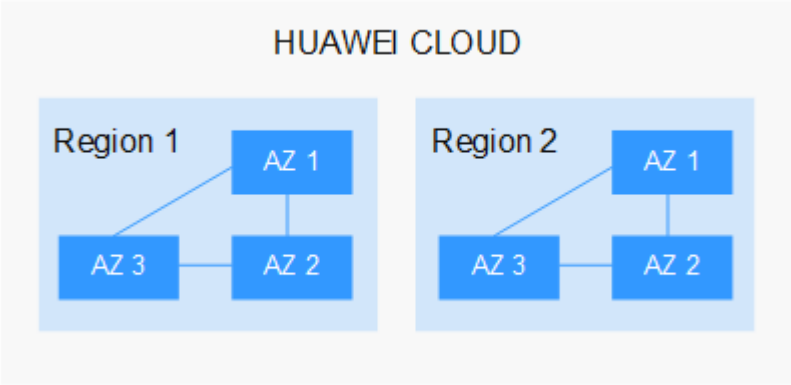
Definição

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

- As regiões são divididas com base na localização geográfica e na latência da rede. Os 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 como regiões universais e regiões dedicadas. Uma região universal fornece serviços de nuvem universal para domínios comuns. Uma região dedicada fornece serviços do mesmo tipo apenas ou para domínios 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 em uma região são interconectadas através de fibras ópticas de alta velocidade. Isso é útil se você implementar sistemas em AZs para obter maior disponibilidade.

Figura 8-1 mostra a relação entre as regiões e as AZ.

Figura 8-1 Regiões e as AZs



Huawei Cloud fornece serviços em muitas regiões do mundo. Você pode selecionar uma região e uma AZ conforme necessário. Para obter mais informações, consulte **Produtos globais e serviços**.

Como escolher uma região?

Ao selecionar uma região, considere os seguintes fatores:

- Localização

  Selecione uma região próxima de você ou de seus usuários-alvo para reduzir a latência da rede e melhorar a taxa de acesso. As regiões da China continental fornecem basicamente a mesma infraestrutura, qualidade de rede BGP, bem como operações e configurações em recursos. Se você ou seus usuários-alvo estiverem na China continental, não será necessário considerar as diferenças de latência da rede ao selecionar uma região.

  – Se você ou seus usuários-alvo estiverem na região Ásia-Pacífico, exceto na China continental, selecione a região **CN-Hong Kong**, **AP-Bangkok** ou **AP-Singapore**.
  – Se você ou seus usuários-alvo estiverem na África do Sul, selecione a região **AF-Johannesburg**.
  – Se você ou seus usuários-alvo estiverem na Europa, selecione a região **EU-Paris**.
  – Se seus usuários-alvo estiverem na América Latina, selecione a região **LA-Santiago**.

    NOTA

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

- Preço do recurso

  Os preços de recurso 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.

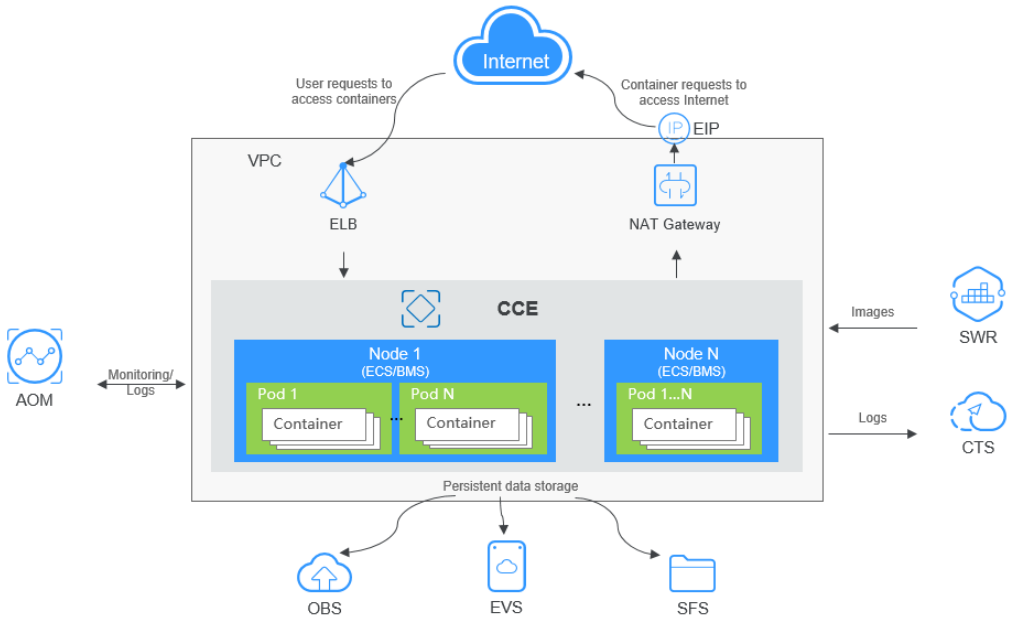
Regiões e pontos de extremidade da VPC

Ao usar uma API para acessar recursos, você deve especificar uma região e um pontos de extremidade. Para obter detalhes, consulte **Regiões e pontos de extremidade**.

9 Serviços relacionados

O CCE funciona com os seguintes serviços de nuvem e requer permissões para acessá-los.

Figura 9-1 Relações entre CCE e outros serviços



Relações entre CCE e outros serviços

Tabela 9-1 Relações entre CCE e outros serviços

Serviço	Relação	Funções relacionadas
Elastic Cloud Server (ECS)	Um ECS com vários discos do EVS é um nó no CCE. Você pode escolher especificações do ECS durante a criação do nó.	• **Compra de um nó** • **Aceitação de nós existentes em um cluster**

Serviço	Relação	Funções relacionadas
Virtual Private Cloud (VPC)	Por motivos de segurança, todos os clusters criados pelo CCE devem ser executados em **VPCs**. Ao criar um namespace, você precisa criar uma VPC ou vincular uma VPC existente ao namespace para que todos os containers no namespace sejam executados nessa VPC.	**Compra de um cluster do CCE**
Elastic Load Balance (ELB)	O CCE trabalha com o ELB para balancear as solicitações de acesso de uma carga de trabalho em vários pods. Quando **ELB** é usado, o endereço do balanceador de carga, em vez do endereço da carga de trabalho, é exposto aos usuários. As solicitações do usuário primeiro chegam ao ELB por meio de uma rede pública e depois são roteadas pelo ELB para diferentes pods da carga de trabalho.	● **Criação de uma Implementação** ● **Criação de um StatefulSet** ● **LoadBalancer**
Gateway NAT	O serviço de gateway NAT fornece tradução de endereço de rede de origem (SNAT) para instâncias de container em uma VPC. O recurso da SNAT converte endereços IP privados dessas instâncias de container para o mesmo EIP, que é um endereço IP público acessível na Internet. Você pode definir regras da SNAT no **NAT gateway** para permitir que os containers acessem a Internet	● **Criação de uma Implementação** ● **Criação de um StatefulSet** ● **DNAT**
Software Repository for Container (SWR)	Um repositório de imagens é usado para armazenar e gerenciar imagens do Docker. Você pode criar cargas de trabalho a partir de imagens no **SWR**.	● **Criação de uma Implementação** ● **Criação de um StatefulSet**
Elastic Volume Service (EVS)	Os discos do EVS podem ser conectados a servidores em nuvem e dimensionados para uma capacidade maior sempre que necessário. Um ECS com vários discos do EVS é um nó no CCE. Você pode escolher especificações do ECS durante a criação do nó.	**Uso de volumes do EVS**

Serviço	Relação	Funções relacionadas
Object Storage Service (OBS)	O OBS fornece armazenamento em nuvem para dados de qualquer tamanho estável, seguro, econômico e baseado em objetos. Com o OBS, você pode criar, modificar e excluir intervalos, bem como fazer upload, baixar e excluir objetos. O CCE permite criar um volume do OBS e anexá-lo a um caminho dentro de um container.	**Uso de volumes do OBS**
Scalable File Service (SFS)	O SFS é um serviço de armazenamento de arquivos compartilhado e totalmente gerenciado. Compatível com o protocolo Network File System, os sistemas de arquivos SFS podem escalar elasticamente até petabytes, garantindo assim o melhor desempenho de aplicações com uso intenso de dados e largura de banda. Você pode usar sistemas de arquivos do SFS como armazenamento persistente para containers e anexar os sistemas de arquivos a containers ao criar uma carga de trabalho.	**Uso de volumes do SFS**
Application Operations Management (AOM)	O AOM coleta arquivos de log de container em formatos como .log do CCE e os despeja no AOM. No console do AOM, você pode facilmente consultar e visualizar arquivos de log. Além disso, o AOM monitora o uso de recursos do CCE. Você pode definir limites de métrica para o uso de recursos do CCE para acionar o dimensionamento automático.	**Coleção de logs de saída padrão de containers**
Cloud Trace Service (CTS)	O CTS registra operações em seus recursos de nuvem, permitindo que você consulte, audite e rastreie solicitações de operação de recursos iniciadas no console de gerenciamento ou em APIs abertas, bem como respostas a essas solicitações.	**Operações do CCE suportadas por CTS**

