Cloud Container Engine

Descripción general del servicio

Edición 01

Fecha 2023-01-09

HUAWEI TECHNOLOGIES CO., LTD.

Marcas y permisos


y otras marcas registradas de Huawei pertenecen a Huawei Technologies Co., Ltd.
Todas las demás marcas registradas y los otros nombres comerciales mencionados en este documento son propiedad de sus respectivos titulares.

Aviso

Las funciones, los productos y los servicios adquiridos están estipulados en el contrato celebrado entre Huawei y el cliente. Es posible que la totalidad o parte de los productos, las funciones y los servicios descritos en el presente documento no se encuentren dentro del alcance de compra o de uso. A menos que el contrato especifique lo contrario, ninguna de las afirmaciones, informaciones ni recomendaciones contenidas en este documento constituye garantía alguna, ni expresa ni implícita.

La información contenida en este documento se encuentra sujeta a cambios sin previo aviso. En la preparación de este documento se realizaron todos los esfuerzos para garantizar la precisión de sus contenidos. Sin embargo, ninguna declaración, información ni recomendación contenida en el presente constituye garantía alguna, ni expresa ni implícita.

Huawei Technologies Co., Ltd.

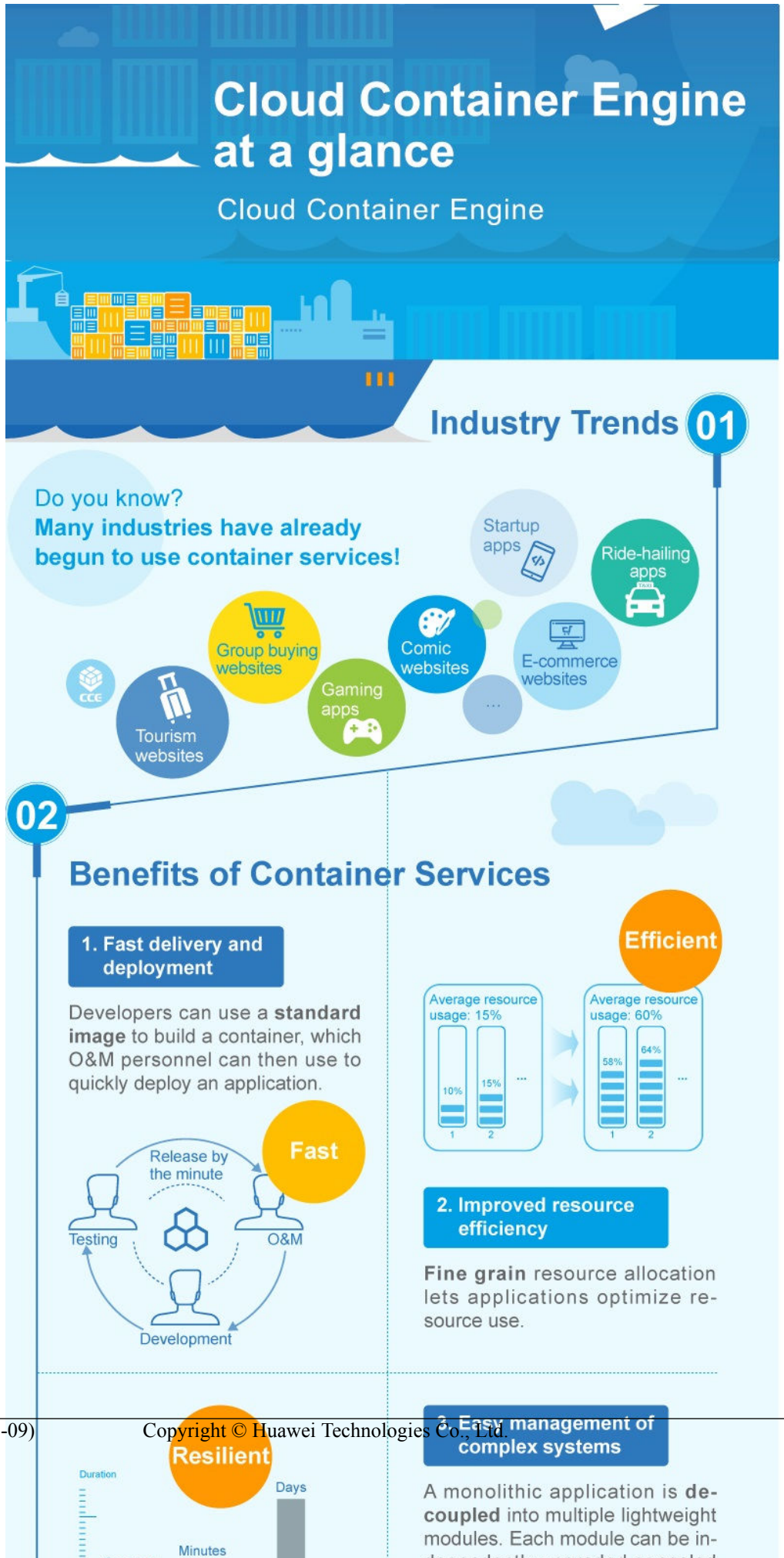
Dirección: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Sitio web: https://www.huawei.com

Email: support@huawei.com

Índice

1 Infografía de CCE



2 ¿Qué es Cloud Container Engine?

Cloud Container Engine (CCE) es un servicio de Kubernetes alojado de clase empresarial altamente escalable para que ejecute los contenedores y aplicaciones. Con CCE, puede implementar, gestionar y escalar fácilmente aplicaciones en contenedores en la nube.

¿Por qué CCE?

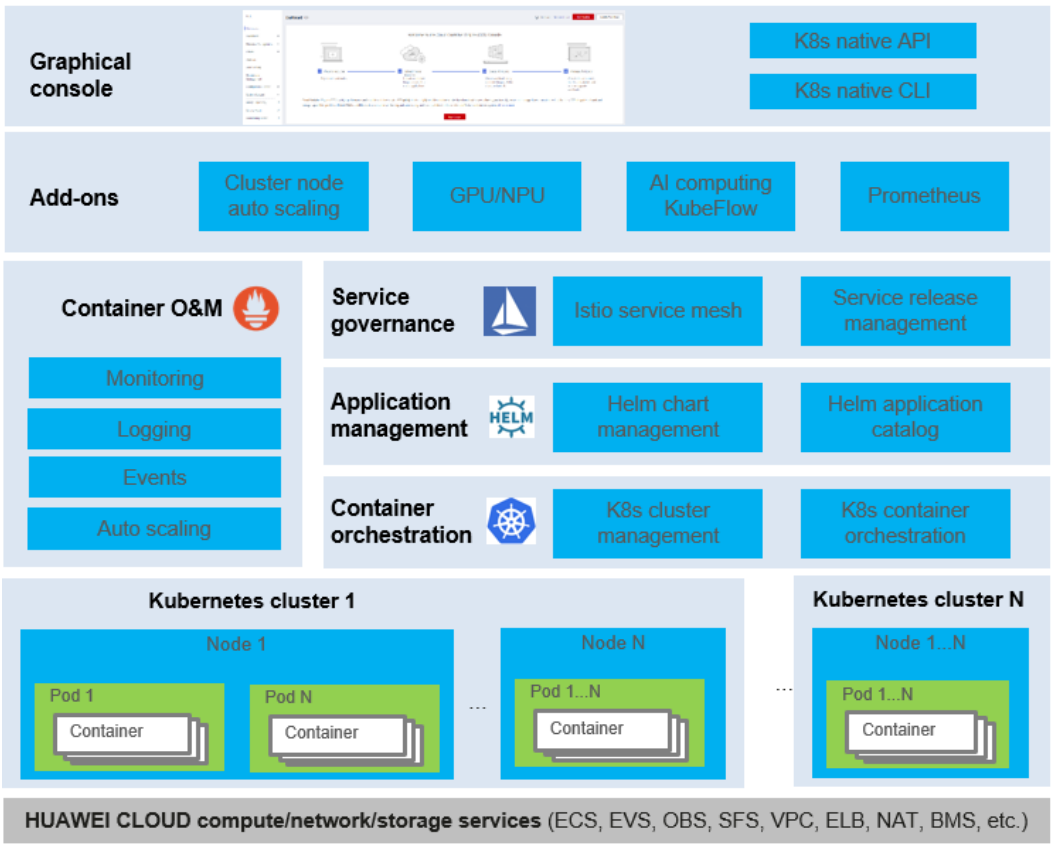
CCE está profundamente integrado con los servicios en la nube, incluidos los servicios de computación de alto rendimiento (ECS/BMS), redes (VPC/EIP/ELB) y almacenamiento (EVS/OBS/SFS). Es compatible con arquitecturas informáticas heterogéneas como GPU, NPU y Arm. Al admitir la recuperación ante desastres multi-AZ y multirregión, CCE garantiza una alta disponibilidad de los clústeres de **Kubernetes**.

Huawei Cloud es uno de los primeros Kubernetes Certified Service Providers (KCSP) del mundo y el primer participante de China en la comunidad de Kubernetes. Durante mucho tiempo ha estado contribuyendo a las comunidades de contenedores de código abierto y tomando el liderazgo en el ecosistema de contenedores. Huawei Cloud también es fundador y miembro platino de Cloud Native Computing Foundation (CNCF). CCE es uno de los servicios de contenedores del mundo en aprobar por primera vez el Certified Kubernetes Conformance Program.

Para obtener más información, consulte **Ventajas del producto** y **Escenarios de aplicación**.

Arquitectura del producto

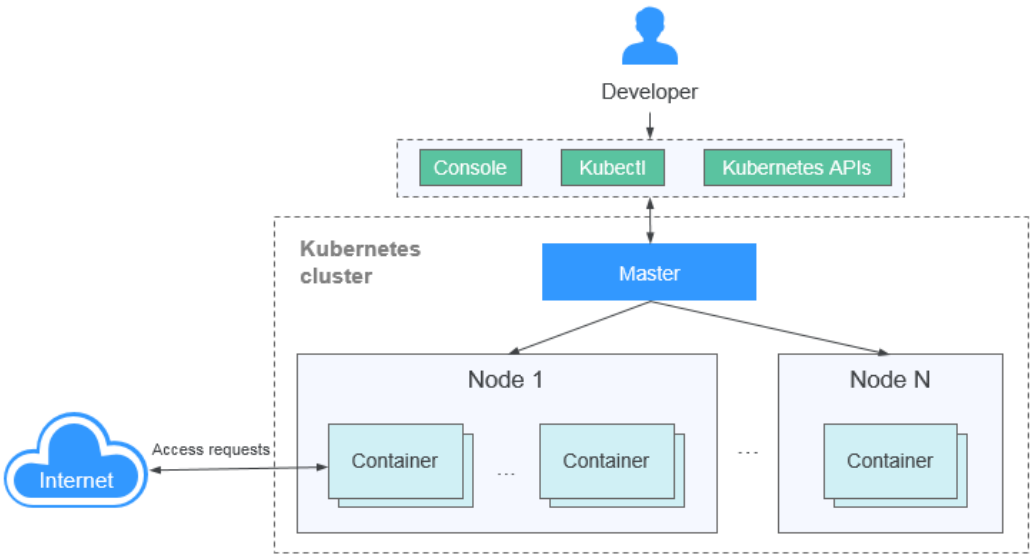
Figura 2-1 Arquitectura de CCE



Acceso a CCE

Puede usar CCE a través de la consola de CCE, kubectl o las API de Kubernetes. **Figura 2-2** muestra el proceso.

Figura 2-2 Acceso a CCE



Ruta de aprendizaje de CCE

Puede hacer clic **aquí** para aprender sobre los fundamentos de CCE para que pueda usar CCE y realizar O&M con facilidad.

3 Descripción general de la función

CCE le permite gestionar varios objetos de recursos, incluidos clústeres, nodos, grupos de nodos, cargas de trabajo, complementos y gráficos. Otras características avanzadas incluyen programación basada en afinidad, ajuste automático, almacenamiento y redes de contenedores, gestión de permisos y administrador del sistema.

Clúster

CCE es un servicio alojado de Kubernetes que simplifica la implementación y gestión de aplicaciones en contenedores. Con CCE, puede crear fácilmente los clústeres de Kubernetes, implementar las aplicaciones en contenedores y gestionarlas y mantenerlas.

- **Implementación integral y O&M**: Puede crear un clúster de contenedores de Kubernetes con solo unos clics, sin necesidad de configurar entornos Docker o Kubernetes. La implementación automática y O&M de aplicaciones en contenedores se pueden realizar en un solo lugar durante todo el ciclo de vida de la aplicación.
- **Múltiples tipos de clústeres**: CCE trabaja estrechamente con servicios de infraestructura heterogénea, incluidos Elastic Cloud Server (ECS), Bare Metal Server (BMS), y GPU-Acceleration Cloud Server (GACS). Puede elegir el tipo de clúster que mejor se adapte a sus necesidades y crear clústeres rápidamente mientras CCE maneja toda la complejidad de la gestión de clústeres.

Tabla 3-1 Funciones de gestión de clústeres

Función	Descripción
Clúster de Turbo de CCE	El clúster de Turbo de CCE admite la implementación híbrida de máquinas virtuales y bare metal servers (BMS) para ofrecer un alto rendimiento basado en servicios de infraestructura.
Clúster de CCE	El clúster de CCE admite la implementación híbrida de máquinas virtuales y bare-metal servers (BMS), y nodos heterogéneos como los nodos de GPU y NPU. Puede ejecutar sus contenedores en un entorno de tiempo de ejecución de contenedores seguro y estable basado en un modelo de red de alto rendimiento.

Función	Descripción
Clúster de Kunpeng	El clúster de Kunpeng (con conjuntos de instrucciones basados en Arm) permite que los contenedores se ejecuten en servidores en la nube que utilizan la arquitectura Arm y los procesadores Kunpeng. Los servidores en la nube acelerados de Kunpeng son fáciles de implementar y proporcionan un rendimiento de ajuste y programación comparable al de los servidores en la nube basados en x86 a solo una fracción de lo que costaría los servidores en la nube basados en x86.
Ajuste automático de clústeres	CCE escala automáticamente un clúster agregando o liberando nodos de trabajo. Por ejemplo, cuando las cargas de trabajo no se pueden programar en el clúster debido a que los recursos del clúster son insuficientes, el escalamiento horizontal se activará automáticamente.
Actualización de clúster	Puede actualizar su clúster a la última versión de Kubernetes o a una versión corregida de errores en la consola de CCE.
Monitoreo de clústeres	Puede ver las métricas de supervisión para conocer el uso de recursos de cada nodo principal de clúster en tiempo real, y recibir y responder a las alarmas de manera oportuna.

Nodo

Un clúster de contenedores consta de un conjunto de máquinas de trabajo, denominadas nodos, que ejecutan aplicaciones en contenedores. Un nodo puede ser una máquina virtual (VM) o una máquina física (PM), dependiendo de sus requisitos de servicio. Los componentes de un nodo incluyen kubelet, tiempo de ejecución del contenedor y kube-proxy. CCE utiliza Elastic Cloud Servers (ECS) de alto rendimiento o Bare Metal Servers (BMS) como nodos para crear clústeres de Kubernetes de alta disponibilidad.

Tabla 3-2 Funciones de gestión de nodos

Función	Descripción
Adición de un nodo	Puede agregar nodos a un clúster de CCE comprando o aceptando nodos. La aceptación de nodos consiste en agregar ECS comprados a un clúster de CCE. Se pueden comprar y agregar nodos heterogéneos, como máquinas virtuales, bare metal servers, nodos habilitados para GPU y nodos habilitados para NPU.
Supervisión de un nodo	CCE utiliza el servicio Cloud Eye para monitorear nodos. Cada nodo corresponde a un ECS.
Restablecimiento de un nodo	Cuando se restablece un nodo de un clúster de CCE, también se eliminarán los servicios que se ejecuten en el nodo. Tenga cuidado cuando realice esta acción. Esta función es compatible con clústeres de v1.13 y posteriores.

Función	Descripción
Eliminación de un nodo	Cuando se elimina un nodo de un clúster CCE, también se eliminan los servicios que se ejecutan en el nodo. Tenga cuidado cuando realice esta acción.

Grupo de nodos

Puede crear un grupo de nodos para que un clúster de CCE cree, administre y destruya rápidamente nodos sin afectar a todo el clúster. Todos los nodos de un grupo de nodos personalizado tienen parámetros y tipo de nodo idénticos. No se puede configurar un solo nodo en un grupo de nodos; cualquier cambio de configuración afecta a todos los nodos del grupo de nodos.

Tabla 3-3 Funciones de gestión del grupo de nodos

Función	Descripción
Creación de un grupo de nodos	Puede crear y ver grupos de nodos.
Gestión de un grupo de nodos	Puede modificar los parámetros de un grupo de nodos, eliminar o clonar un grupo de nodos y migrar nodos.

Carga de trabajo

Una carga de trabajo es una aplicación que se ejecuta en Kubernetes. No importa cuántos componentes haya en su carga de trabajo, puede ejecutarlo en un grupo de pods de Kubernetes. Una carga de trabajo es un modelo abstracto de un grupo de pods en Kubernetes. Las cargas de trabajo clasificadas en Kubernetes incluyen Deployments, StatefulSets, DaemonSets, trabajos y trabajos cron.

CCE proporciona implementación y gestión de contenedores nativos de Kubernetes y admite la gestión del ciclo de vida de las cargas de trabajo de contenedores, incluidas la creación, configuración, supervisión, ajuste automático, actualización, desinstalación, descubrimiento de servicios y equilibrio de carga.

Tabla 3-4 Funciones de gestión de cargas de trabajo

Función	Descripción
Establecimiento de las especificaciones del contenedor	CCE le permite establecer límites de recursos para contenedores agregados durante la creación de cargas de trabajo. Puede solicitar y limitar las cuotas de CPU y memoria utilizadas por cada pod en la carga de trabajo, y establecer si usar GPU y Ascend 310 para cada pod.

Función	Descripción
Configuración de los ganchos del ciclo de vida del contenedor	CCE proporciona funciones de devolución de llamada (ganchos) para la gestión del ciclo de vida de las aplicaciones en contenedores. Por ejemplo, si desea que un contenedor realice una determinada operación antes de detenerse, puede registrar un gancho.
Configuración del comando de inicio del contenedor	Al crear una carga de trabajo o un trabajo, puede utilizar una imagen para especificar los procesos que se ejecutan en el contenedor. Se ejecuta el comando predeterminado de la imagen. Para ejecutar un comando específico o utilizar nuevos valores, configure los siguientes valores: ● **Directorio de trabajo**: directorio de trabajo del comando. Si el directorio de trabajo no se especifica en la imagen o en la consola, el valor predeterminado es /. ● **Comando**: comando que controla la ejecución de una imagen. ● **Args**: parámetros pasados al comando en ejecución.
Configuración de la comprobación de estado del contenedor	CCE puede comprobar regularmente el estado de salud de los contenedores durante el funcionamiento del contenedor. Si la función de comprobación de estado no está configurada, un pod no puede detectar excepciones de servicio ni reiniciar automáticamente el servicio para restaurarlo. Esto dará como resultado una situación en la que el estado del pod es normal pero el servicio en el pod es anormal. CCE proporciona las siguientes sondas de comprobación de estado: 1. **Liveness probe** comprueba si un contenedor todavía está vivo. Es similar al comando **ps** que comprueba si existe un proceso. Si el sondeo de vida útil de un contenedor falla, el clúster reinicia el contenedor. Si la sonda de vida se realiza con éxito, no se ejecuta ninguna operación. 2. **Readiness probe** comprueba si un contenedor está listo para procesar solicitudes de usuario. Al detectarse que el contenedor no está listo, el tráfico de servicio no se dirigirá al contenedor. Algunas aplicaciones pueden tardar mucho tiempo en iniciarse antes de que puedan proporcionar servicios. Esto se debe a que necesitan cargar datos de disco o confiar en el inicio de un módulo externo. En este caso, los procesos de aplicación se están ejecutando, pero las aplicaciones no están listas para proporcionar los servicios. Aquí es donde entra la sonda de preparación. Si la sonda de preparación del contenedor falla, el clúster enmascara todas las solicitudes enviadas al contenedor. Si la sonda de preparación del recipiente se realiza con éxito, se puede acceder al recipiente.
Definición de variables de entorno	Una variable de entorno es una variable cuyo valor puede afectar a la forma en que se comportará un contenedor en ejecución. Puede modificar las variables de entorno incluso después de implementar las cargas de trabajo, lo que aumenta la flexibilidad en la configuración de la carga de trabajo.

Función	Descripción
Recopilando registros	CCE le permite configurar las políticas para recopilar, gestionar y analizar registros de carga de trabajo periódicamente. Estas políticas pueden evitar que los registros se sobredimensionen.

Programación de afinidad y antiafinidad

Puede restringir en qué AZ y nodos sus cargas de trabajo son elegibles o están prohibidos para ser programados. También puede definir reglas para describir qué cargas de trabajo se ubicarán o no con sus cargas de trabajo. La programación de afinidad permite que las cargas de trabajo estén físicamente más cerca de la ubicación del usuario y hace que las rutas de enrutamiento entre contenedores sean lo más cortas posible, lo que a su vez reduce la sobrecarga de red. La programación antiafinidad evita un único punto de fallo al prohibir la ubicación conjunta de instancias que pertenezcan a la misma carga de trabajo. También evita que las cargas de trabajo de interferencia se afecten entre sí al no permitir que se ejecuten en el mismo nodo o AZ.

Tabla 3-5 Políticas de planificación

Función	Descripción
Política de planificación personalizada	Una política de programación personalizada le permite personalizar la afinidad de nodos, la afinidad de la carga de trabajo o la antiafinidad de la carga de trabajo. Una combinación de políticas de programación personalizadas puede satisfacer mejor sus requisitos de servicio.
Política de programación sencilla	Una política de programación simple proporciona las funciones básicas de programación únicas. Le permite configurar la afinidad entre cargas de trabajo y AZ, entre cargas de trabajo y nodos, o entre cargas de trabajo.

Redes

Al integrar profundamente las capacidades de red de Kubernetes con VPC, CCE proporciona redes estables y de alto rendimiento para el acceso mutuo de cargas de trabajo en escenarios complejos.

Tabla 3-6 Funciones relacionadas con la conexión en red

Función	Descripción
Servicio	Los servicios le permiten acceder a una o varias aplicaciones en contenedores. Cada servicio tiene una dirección IP y un puerto fijos durante su ciclo de vida y se dirige a uno o más pods backend. De esta manera, los clientes frontend no necesitan realizar un seguimiento de estos pods, lo que permite agregar o reducir los pods sin preocuparse de los cambios de dirección IP. CCE admite los siguientes tipos de Servicios: ● ClusterIP: El servicio solo es accesible desde dentro del clúster. ● NodePort: Se accede al Servicio utilizando la dirección IP privada o EIP del nodo. ● LoadBalancer: Se accede al Servicio mediante un balanceador de carga. ● DNAT: Se accede al Servicio mediante un gateway de DNAT.
Equilibrio de carga de capa 7 (entrada)	Los ingresos utilizan balanceadores de carga compartidos y dedicados. En comparación con el balanceo de carga de capa 4, el balanceo de carga de capa 7 también admite configuraciones de Localizador uniforme de recursos (URL) y enruta el tráfico de acceso a los servicios en función de las URL. Los servicios también actuarán de acuerdo con los URL.
Política de red	CCE ha mejorado la función de política de red basada en Kubernetes, permitiendo el aislamiento de red en un clúster mediante la configuración de políticas de red. Esto significa que se puede establecer un firewall entre pods. Por ejemplo, para hacer que un sistema de pago solo sea accesible a los componentes especificados por motivos de seguridad, puede configurar políticas de red.
Definición de datos adjuntos de red	Una definición de adjunto de red es un tipo de definiciones de Custom Resource Definitions (CRD) en un clúster. Proporciona elementos de configuración, como VPC y subred, para que los contenedores se conecten a la interfaz de red elástica (ENI). Las cargas de trabajo asociadas con las definiciones de datos adjuntos de red pueden conectarse al ENI, de modo que los contenedores pueden enlazarse directamente con los ENI para exponer los servicios de forma externa.

Volumen persistente (PV)

Además de utilizar discos locales para el almacenamiento, CCE puede almacenar datos de carga de trabajo mediante servicios de almacenamiento en la nube. Actualmente, se admiten los siguientes tipos de almacenamiento en la nube: Elastic Volume Service (EVS), Object Storage Service (OBS), Scalable File Service (SFS) y SFS Turbo.

Tabla 3-7 Funciones relacionadas con el almacenamiento de información

Función	Descripción
Uso de discos locales como volúmenes de almacenamiento	Puede montar el directorio de archivos del host donde se encuentra un contenedor en una ruta de contenedor especificada (correspondiente a hostPath en Kubernetes). Alternativamente, puede dejar la ruta de origen vacía (correspondiente a emptyDir en Kubernetes). Si la ruta de origen se deja vacía, se montará un directorio temporal del host en el punto de montaje del contenedor. Se utiliza una ruta de origen especificada cuando los datos deben almacenarse de forma persistente en el host, mientras que se utiliza emptyDir cuando se necesita almacenamiento temporal.
Uso de discos de EVS como volúmenes de almacenamiento	Los discos de EVS se pueden unir a los contenedores. Mediante el uso de volúmenes de EVS, puede adjuntar el directorio de archivos remoto de un sistema de almacenamiento en un contenedor para que los datos del volumen de datos se conserven permanentemente. Incluso si se elimina el contenedor, los datos en el volumen de datos todavía se almacenan en el sistema de almacenamiento.
Uso de sistemas de archivos de SFS como los volúmenes de almacenamiento	Puede crear los volúmenes de SFS y montarlos en las rutas de contenedores específicas. También se pueden utilizar los volúmenes creados por el servicio SFS subyacente. Los volúmenes de SFS son adecuados para escenarios en los que los datos deben persistir y leerse y escribirse en múltiples nodos. Tales escenarios incluyen el procesamiento de medios, gestión de contenido, análisis de big data y análisis de carga de trabajo.
Uso de bucket de OBS como los volúmenes de almacenamiento	Puede crear volúmenes OBS y montarlos en una ruta de contenedor. OBS se aplica a escenarios como cargas de trabajo en la nube, análisis de datos, análisis de contenido y objetos de punto de acceso.
Uso de sistemas de archivos de Turbo de SFS como volúmenes de almacenamiento	Puede crear los volúmenes de Turbo de SFS y montarlos en una ruta de contenedor. Los sistemas de archivos de Turbo de SFS son rápidos, bajo demanda y escalables, que son adecuados para DevOps y aplicaciones de oficina empresarial.
Creación de instantáneas y copias de seguridad	CCE trabaja con EVS para proporcionar la función de instantánea. Una instantánea es una copia o imagen completa de los datos del disco de EVS en un determinado momento, lo que es de gran ayuda para la recuperación ante desastres de datos.

Complementos

CCE proporciona múltiples tipos de complementos para ampliar las funciones de clúster y satisfacer diversos requisitos.

- CCE admite las API abiertas y las API nativas de la comunidad.
- CCE proporciona un complemento relacionado con kubectl y un kubectl nativo de la comunidad.

Herramientas de ecosistema

CCE funciona sin problemas con Application Service Mesh (ASM) y Helm.

Tabla 3-8 Ecosistema de Kubernetes

Función	Descripción
Application Service Mesh (ASM)	ASM proporciona un enfoque no intrusivo para la gobernanza de los microservicios. Es compatible con la gestión completa del ciclo de vida y del tráfico y es compatible con los ecosistemas de Kubernetes e Istio. La facilidad de uso lista para usar le permite usar mallas de servicio sin necesidad de reescritura de código o instalación manual de proxy.
Gestión de gráficos	Helm es un gestor de paquetes de Kubernetes que facilita la implementación y gestión de paquetes (también llamados gráficos). Un gráfico es una colección de archivos que describen un conjunto relacionado de recursos de Kubernetes. El uso de gráficos controla toda la complejidad en la instalación y gestión de recursos de Kubernetes. En CCE, puede cargar los gráficos para implementar aplicaciones. ● Los gráficos cargados están definidos por el usuario, lo que simplifica la implementación de la carga de trabajo.

Ajuste automático

CCE le permite escalar sus clústeres y cargas de trabajo de forma manual y automática. Cualquier política de ajuste automático se puede combinar de manera flexible para hacer frente a picos de carga en el momento.

Tabla 3-9 Funciones de ajuste automático

Función	Descripción
Ajuste de la carga de trabajo	CCE es compatible con las políticas de HPA y de CustomedHPA. ● HPA: una política que implementa el ajuste horizontal de pods en Kubernetes. En una política de HPA de CCE, puede configurar la ventana de tiempo de enfriamiento y los umbrales de ajuste basados en el HPA de Kubernetes. ● CustomedHPA: una política que escala las implementaciones en función de métricas (como el uso de la CPU y el uso de la memoria) o en un intervalo periódico (cada día/semana/mes o en un punto de tiempo específico). Este tipo de política es una capacidad de ajuste automático mejorada desarrollada por Huawei Cloud.
Ajuste de nodos	CCE proporciona ajuste de nodos a través del complemento del ajuster automático. Los nodos con diferentes especificaciones se pueden agregar automáticamente a través de AZ bajo demanda.

Permisos

La gestión de permisos de CCE le permite asignar permisos a los usuarios y grupos de usuarios de IAM en sus cuentas de tenant. CCE combina las ventajas de Identity and Access Management (IAM) y Kubernetes Role-based Access Control (RBAC) para proporcionar una variedad de métodos de autorización, incluida la autorización de grano fino de IAM, la autorización de token de IAM, autorización en el ámbito del clúster y autorización en todo el espacio de nombres.

Tabla 3-10 Funciones relacionadas con permisos

Función	Descripción
Permisos a nivel de clúster	Los permisos a nivel de clúster de CCE se asignan según las **IAM system policies** y **custom policies**. Puede utilizar grupos de usuarios para asignar permisos a los usuarios de IAM.
Permisos a nivel del espacio de nombre	Puede regular el acceso de los usuarios o los grupos de usuarios a los recursos de Kubernetes en un único espacio de nombres basado en sus roles de Kubernetes RBAC. El CCE también se ha mejorado sobre la base de las capacidades de código abierto. Admite la autorización de RBAC basada en el usuario o grupo de usuarios de IAM, y la autenticación de RBAC en el acceso a las API mediante tokens de IAM.

4 Ventajas del producto

¿Por qué CCE?

CCE es un servicio de contenedores basado en las tecnologías populares de Docker y Kubernetes y ofrece una gran cantidad de características que se adaptan mejor a la demanda de las empresas para ejecutar clústeres de contenedores a escala. Con las ventajas únicas en confiabilidad del sistema, rendimiento y compatibilidad con comunidades de código abierto, CCE puede adaptarse a los detalles de las empresas interesadas en construir nubes de contenedores.

Fácil para el uso

- Crear un clúster de Kubernetes es tan fácil como hacer unos pocos clics en la interfaz de usuario web (WebUI). El clúster de Kubernetes admite la gestión de nodos de máquinas virtuales o nodos de metal desnudo y se aplica al escenario en el que las máquinas virtuales y las máquinas físicas se usan juntas.
- La implementación automática y O&M de aplicaciones en contenedores se pueden realizar en un solo lugar durante todo el ciclo de vida de la aplicación.
- Los clústeres y las cargas de trabajo se pueden cambiar de tamaño con solo unos clics en la WebUI. Cualquier política de ajuste automático se puede combinar de manera flexible para hacer frente a picos de carga en el momento.
- El WebUI le guiará a través de los pasos necesarios para actualizar los clústeres de Kubernetes.
- Soporte para gráficos de Application Service Mesh (ASM) y Helm ofrece facilidad de uso lista para usar.

Alto rendimiento

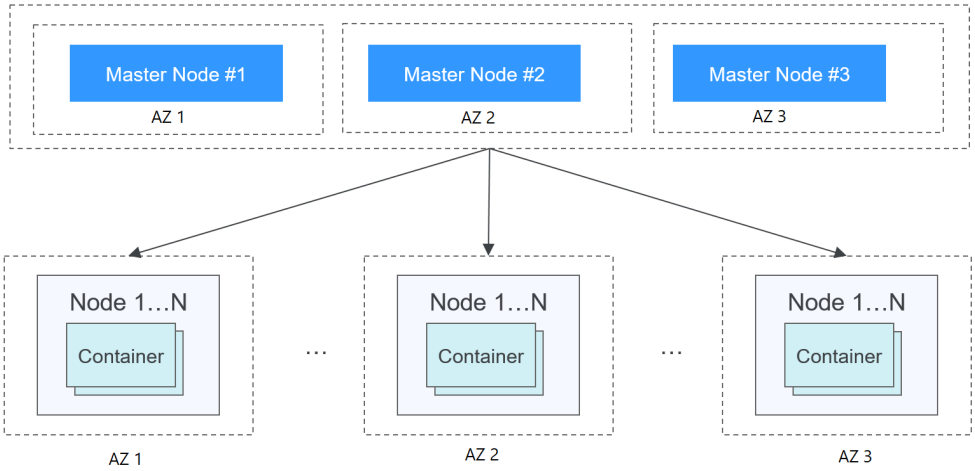
- CCE se basa en años de experiencia de campo en cómputo, redes, almacenamiento e infraestructura heterogénea. Puede lanzar contenedores simultáneamente a escala.
- La arquitectura NUMA de metal puro y las tarjetas de red InfiniBand con alta velocidad ofrecen una mejora de tres a cinco veces en el rendimiento informático.

Altamente disponible y seguro

- Alta confiabilidad: puede implementar tres nodos principales en diferentes AZ para el plano de control del clúster para garantizar una alta disponibilidad de sus servicios. Los nodos y las cargas de trabajo de un clúster pueden equilibrarse con la carga entre AZ para formar una arquitectura multiactiva que garantice la continuidad del servicio incluso

cuando uno de los hosts o salas de equipos está inactivo o una AZ se ve afectada por desastres naturales.

Figura 4-1 Configuración de alta disponibilidad de clústeres



- Seguro: los clústeres son privados y están completamente controlados por los usuarios con IAM y Kubernetes RBAC profundamente integrados. Puede establecer diferentes permisos de RBAC para los usuarios de IAM en la consola.

Abierto y compatible

- CCE se basa en la tecnología de Docker de código abierto que automatiza la implementación, la programación de recursos, el descubrimiento de servicios y el ajuste dinámico de aplicaciones en contenedores.
- CCE se basa en Kubernetes y es compatible con las API nativas de Kubernetes, kubectl (una interfaz de línea de comandos) y las versiones nativas de Kubernetes/Docker. Las actualizaciones de las comunidades de Kubernetes y de Docker se incorporan regularmente en CCE.

Análisis comparativo de los sistemas de gestión de clústeres de Kubernetes locales y CCE

Tabla 4-1 Clústeres de CCE frente a clústeres de Kubernetes locales

Área de enfoque	Sistemas de gestión de clústeres locales de Kubernetes	CCE
Facilidad de uso	La gestión de clústeres es compleja. Debe manejar toda la complejidad de instalar, operar, escalar, configurar y monitorear la infraestructura de gestión de clústeres de Kubernetes. Cada actualización de clúster requiere un tremendo ajuste manual, lo que supone una pesada carga para el personal de O&M.	**Fácil de gestionar y usar clústeres** Puede crear y actualizar clústeres de contenedores de Kubernetes con solo unos clics, sin necesidad de configurar entornos Docker o Kubernetes. La implementación automática y O&M de aplicaciones en contenedores se pueden realizar en la consola, todo en un solo lugar durante todo el ciclo de vida de la aplicación. Soporte para gráficos de Helm ofrece facilidad de uso lista para usar. El uso de clústeres de CCE es tan sencillo como elegir un clúster de contenedores y los trabajos que desea ejecutar en el clúster. A continuación, CCE completa la gestión de clústeres para que pueda centrarse en el desarrollo de aplicaciones en contenedores.
Escalabilidad	Debe evaluar manualmente la carga de servicio y el estado del clúster antes de decidir cambiar el tamaño de un clúster.	**Servicio de ajuste administrado** CCE puede cambiar automáticamente el tamaño de clústeres y cargas de trabajo a medida que cambia el uso de recursos. El uso combinado de políticas de ajuste automático permite escalar de forma flexible clústeres y cargas de trabajo para satisfacer las demandas fluctuantes.
Confiabilidad	Solo hay un nodo principal disponible en un clúster. Una vez que el nodo principal esté inactivo, todo el clúster, así como todas las aplicaciones del clúster quedarán fuera de servicio.	**Alta disponibilidad** Si **High Availability** se establece en **Yes** al crear un clúster, se crearán tres nodos principales en el clúster, evitando puntos únicos de error en el plano de control del clúster.

Área de enfoque	Sistemas de gestión de clústeres locales de Kubernetes	CCE
Eficiencia	Debe crear repositorios de imágenes o volver a repositorios de imágenes de terceros. Las imágenes se extraen de los repositorios en serie.	**Rápida implementación de imágenes e integración continua** CCE trabaja con SoftWare Repository for Container (SWR) para soportar las canalizaciones de DevOps y eliminar la necesidad de escribir manualmente Dockerfiles o manifiestos de Kubernetes. Con las plantillas de canalización de ContainerOps puede definir cómo crear imágenes de contenedores, enviarlas a repositorios e implementarlas. Las imágenes se extraen de repositorios en paralelo.
Costo	Se requiere una gran inversión inicial para instalar, gestionar y escalar la infraestructura de gestión de clústeres.	**Rentable** Solo paga por los recursos de infraestructura necesarios para almacenar y ejecutar aplicaciones, así como por los nodos principales del clúster.

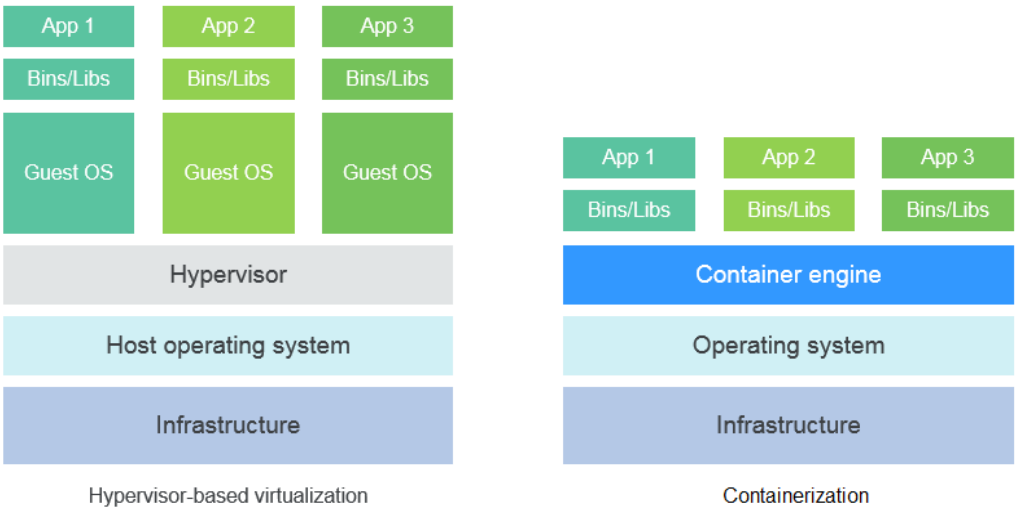
¿Por qué Containers?

Docker está escrito en el lenguaje de programación Go diseñado por Google. Proporciona la virtualización a nivel del sistema operativo: los procesos de software se aíslan entre sí mediante el uso de Linux Control Groups (cgroups), espacios de nombres y tecnologías de Union FS (por ejemplo, AUFS). Todo lo necesario para ejecutar un proceso de software se empaqueta en un contenedor. Los contenedores se aíslan entre sí y del host.

Docker ha avanzado para mejorar el aislamiento de contenedores: los contenedores tienen sus propios sistemas de archivos, y no pueden ver los procesos o las interfaces de red del otro. Esto simplifica la creación y gestión de contenedores.

La tecnología de virtualización tradicional proporciona virtualización a nivel de hardware. Crea un conjunto de máquinas virtuales, cada una con un sistema operativo completo y una aplicación en su interior. Los contenedores, por otro lado, no tienen su propio núcleo y todos llaman al mismo núcleo del host SO. Además, no es necesario hacer ningún tipo de virtualización como lo hace con las máquinas virtuales. Por lo tanto, los contenedores de Docker son más pequeños y más rápidos que las VM.

Figura 4-2 Comparación entre los contenedores de Docker y las máquinas virtuales



En resumen, los contenedores de Docker tienen muchas ventajas sobre las máquinas virtuales.

Utilización de recursos

Sin sobrecarga para virtualizar hardware y ejecutar un SO completo, los contenedores pueden superar a las máquinas virtuales en velocidad de ejecución de aplicaciones, pérdida de memoria y velocidad de almacenamiento de archivos.

Velocidad de arranque

Se tarda varios minutos en iniciar una aplicación en una máquina virtual. Las aplicaciones en contenedores de Docker se ejecutan directamente en el núcleo host y no hay necesidad de iniciar un sistema operativo completo junto con las aplicaciones. El tiempo de inicio se puede reducir a segundos o incluso milisegundos, lo que ahorra mucho tiempo en desarrollo, pruebas e implementación.

Entorno coherente

Uno de los mayores problemas que los desarrolladores siempre tienen que lidiar es la diferencia en los entornos donde ejecutan sus aplicaciones. La diferencia entre los entornos de desarrollo, pruebas y producción impide que se descubran algunos errores antes de la implementación. Una imagen de contenedor de Docker incluye todo lo necesario para ejecutar una aplicación y aísla la aplicación de su entorno. Por lo tanto, las aplicaciones en contenedores siempre funcionarán de la misma manera en entornos de desarrollo, pruebas y producción.

Entrega e implementación continuas

Para el personal de DevOps sería ideal que las aplicaciones pudieran ejecutarse en cualquier lugar después de una única creación o configuración.

Docker proporciona una compilación e implementación confiable y frecuente de imágenes de contenedores con rollbacks rápidos y fáciles (debido a la inmutabilidad de la imagen). Los desarrolladores escriben Dockerfiles que contienen todas las instrucciones necesarias para crear imágenes de contenedores y combinar instrucciones actualizadas regularmente en Dockerfiles, una práctica conocida como Continuous Integration (CI). El equipo de operaciones puede implementar imágenes rápidamente en el entorno de producción al permitir que Docker lea las instrucciones de Dockerfiles. El equipo de operaciones puede incluso

seguir la práctica de Continuous Delivery/Deployment (CD) en la que cada cambio de instrucción se compila, prueba y luego se envía automáticamente a un entorno de pruebas de no producción.

El uso de Dockerfiles hace que el proceso de DevOps sea visible para todos en un equipo de DevOps. De esta manera, el equipo de desarrolladores puede comprender mejor las necesidades de los usuarios y los problemas que enfrenta el equipo de operaciones mientras mantiene la aplicación. Por otro lado, el equipo de operaciones puede tener algún conocimiento de las condiciones que deben cumplirse para ejecutar la aplicación. El conocimiento es útil cuando el personal de operaciones implementa imágenes de contenedores en el entorno de producción.

Portabilidad

Docker garantiza la consistencia ambiental en todo el desarrollo, las pruebas y la producción, por lo que los contenedores de Docker pueden ser portátiles en cualquier lugar. Funcionan de manera uniforme, independientemente de si se ejecutan en máquinas físicas, máquinas virtuales, nubes públicas, nubes privadas o incluso portátiles. Puede migrar aplicaciones de una plataforma a otra sin preocuparse de que el cambio del entorno haga que las aplicaciones no puedan funcionar.

Actualización de aplicaciones

Las imágenes de Docker están compuestas por capas. Cada capa solo se almacena una vez y diferentes imágenes pueden contener exactamente las mismas capas. Esto hace que la distribucion sea eficiente debido a que las capas que ya se han transferido como parte de la primera imagen no necesitan ser transferidas de nuevo cuando se transfiere la otra imagen que tambien tiene estas capas. Para actualizar una aplicación en contenedor, puede editar la capa de escritura superior de la imagen final o agregar capas a la imagen base. Además, Docker colabora con equipos de proyectos de código abierto para mantener un gran número de imágenes oficiales de alta calidad. Puede utilizarlos directamente en el entorno de producción o crear fácilmente nuevas imágenes basadas en ellos.

Tabla 4-2 Contenedores frente a las máquinas virtuales tradicionales

Características	Contenedores	VMs
Velocidad de arranque	En segundos	En minutos
Capacidad de disco	MB	GB
Rendimiento	Rendimiento casi nativo	Débil
Capacidad por máquina	Miles de contenedores	Decenas de máquinas virtuales

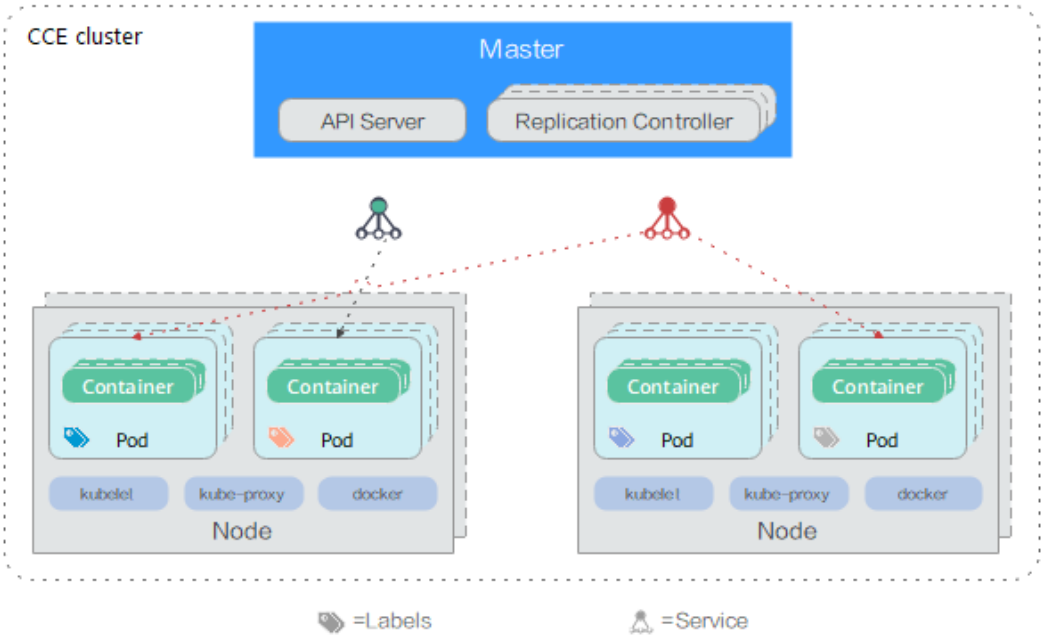
5 Escenarios de aplicación

5.1 Infraestructura y gestión de aplicaciones en contenedores

Escenario de la aplicación

Los clústeres de CCE admiten la gestión de grupos de recursos de x86 y de Arm. Puede crear clústeres de Kubernetes, implementar aplicaciones en contenedores y gestionar y mantener los clústeres.

Figura 5-1 Clúster de CCE



Beneficios

La contenedorización reduce los costos de recursos de implementación de aplicaciones, optimiza la implementación y la actualización y logra servicios ininterrumpidos durante las actualizaciones.

Ventajas

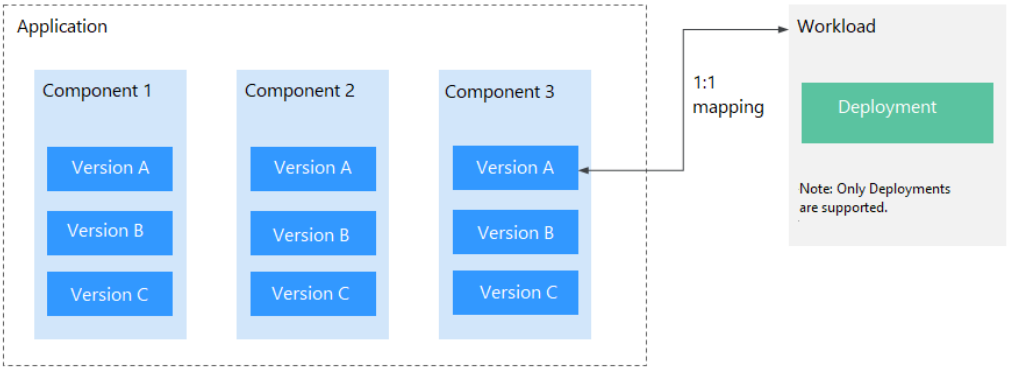
- Implementación de múltiples tipos de cargas de trabajo

  Soporta Deployments, StatefulSets, DaemonSets, trabajos, y trabajos cron.
- Actualización de aplicaciones

  Admite las actualizaciones en reemplazo, actualizaciones continuas por proporción o por número de pods y reversión de actualizaciones.
- Ajuste automático

  Admite el ajuste automático de nodos y cargas de trabajo.

Figura 5-2 Carga de trabajo



5.2 Ajuste automático en segundos

Escenarios de aplicación

- Aumento de tráfico provocado por promociones y ventas flash en aplicaciones y sitios web de compras en línea
- Fluctuación de cargas de servicio de transmisión en vivo
- Aumento en el número de jugadores de juego que se conectan en ciertos períodos de tiempo

Beneficios

CCE adapta automáticamente la cantidad de recursos informáticos a cargas de servicio fluctuantes de acuerdo con las políticas de ajuste automático que haya configurado. Para escalar los recursos informáticos a nivel de clúster, CCE agrega o reduce servidores en la nube. Para escalar los recursos informáticos a nivel de carga de trabajo, CCE agrega o reduce contenedores.

Ventajas

- Flexible

  Permite varias políticas de ajuste y escala contenedores en cuestión de segundos cuando se cumplen las condiciones especificadas.
- Alta disponibilidad:

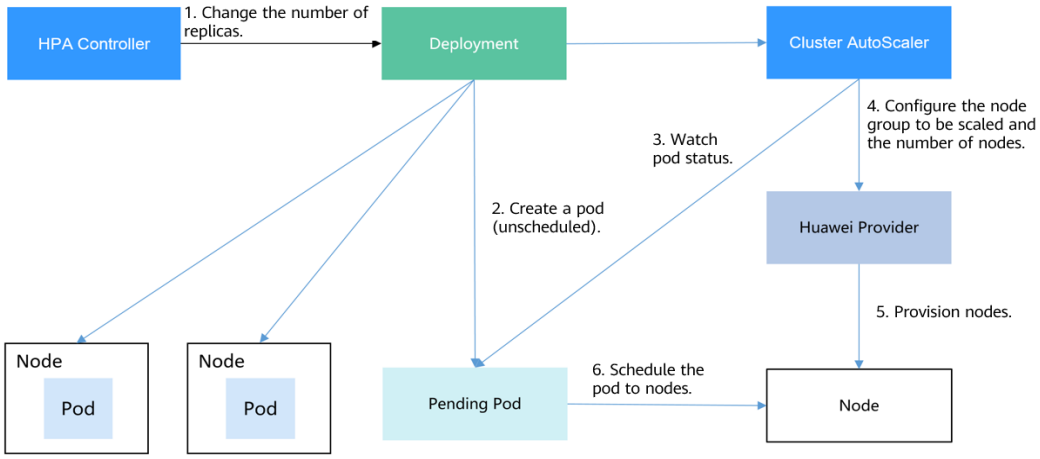
  Detecta automáticamente el estado de funcionamiento del pod en grupos de ajuste automático y reemplaza los pods no saludables por otros nuevos.
- Costos más bajos

  Le cobra solo por los servidores en la nube que utiliza.

Servicios relacionados

HPA (Horizontal Pod Autoscaling) + CA (Cluster AutoScaling)

Figura 5-3 Cómo funciona el ajuste automático



5.3 Gestión de tráfico de microservicios

Escenarios de aplicación

Los sistemas empresariales grandes son cada vez más complejos, más allá de lo que las arquitecturas de sistemas tradicionales pueden manejar. Una solución popular es el microservicio. Las aplicaciones complejas se dividen en componentes más pequeños llamados microservicios. Los microservicios se desarrollan, implementan y escalan de forma independiente. El uso combinado de microservicios y contenedores optimiza la entrega de microservicios al tiempo que mejora la confiabilidad y escalabilidad de las aplicaciones.

Los microservicios hacen posible las arquitecturas distribuidas. Sin embargo, más microservicios indican más complejidad en O&M, puesta en marcha y gestión de seguridad de estas arquitecturas. Los desarrolladores a menudo tienen problemas al escribir código adicional para el gobierno de microservicios e integrar el código en sus sistemas de servicio. En este sentido, CCE proporciona una solución eficiente para liberarle de la carga de trabajo de gestión.

Beneficios

CCE está profundamente integrado con Application Service Mesh (ASM), que le permite completar la liberación en escala de grises, observar su tráfico y controlar el flujo de tráfico sin cambiar su código.

Ventajas

- Uso inmediato

  ASM se puede iniciar con solo unos pocos clics y funciona sin problemas con CCE para controlar inteligentemente el flujo de tráfico.

- Enrutamiento inteligente

  Las políticas de conexión de HTTP/TCP y las políticas de seguridad se pueden aplicar sin modificar el código.

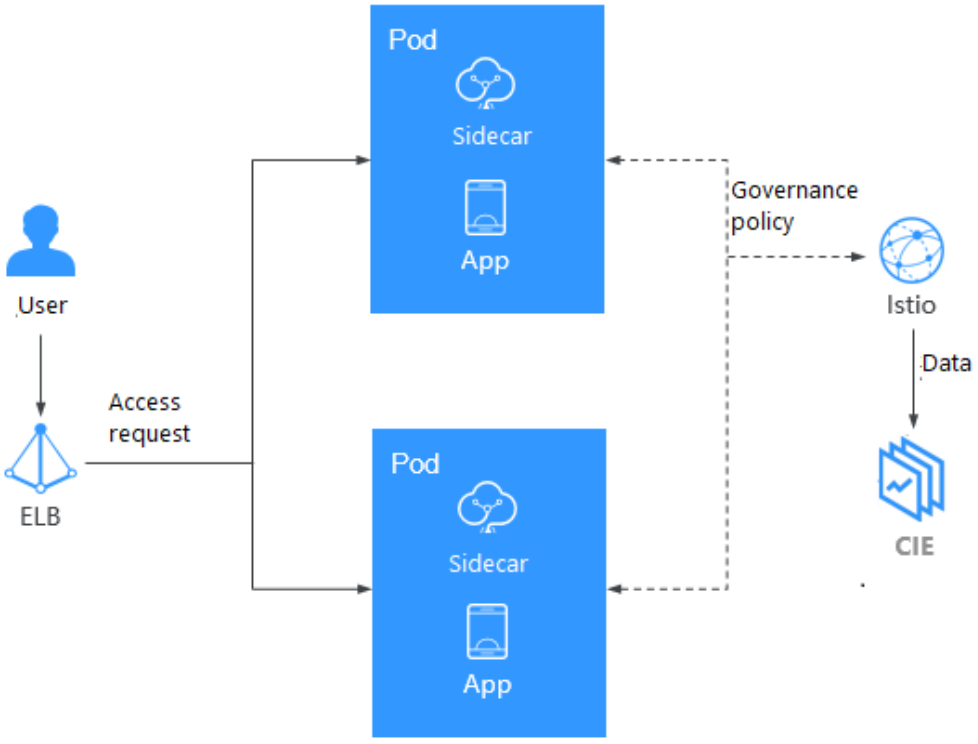
- Visualización del tráfico

  Basándose en los datos de supervisión que se recopilan de forma no intrusiva, ASM trabaja estrechamente con Application Performance Management (APM) para proporcionar una vista panorámica de sus servicios, incluida la topología del tráfico en tiempo real, el seguimiento de llamadas, la supervisión del rendimiento y el diagnóstico en tiempo de ejecución.

Servicios relacionados

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

Figura 5-4 Gobernanza de microservicios



5.4 DevOps y CI/CD

Escenario de la aplicación

Sus aplicaciones y servicios pueden recibir una gran cantidad de comentarios y requisitos. Para lanzar nuevas funciones y mejorar la experiencia del usuario, necesita una integración continua (CI) rápida. Una herramienta eficiente para soportar CI es el contenedor. Mediante la implementación de contenedores, puede optimizar el proceso desde el desarrollo, las pruebas hasta la liberación y la realización de entrega continua (CD).

Beneficios

CCE trabaja con SWR para admitir DevOps que completará automáticamente la compilación de código, la creación de imágenes, la versión en escala de grises y la implementación basada en código fuente. Los sistemas CI/CD tradicionales se pueden conectar para contener aplicaciones heredadas.

Ventajas

- Gestión eficiente de procesos

  Reduce la carga de trabajo de scripting en más de un 80% a través de la interacción de procesos optimizada.
- Integración flexible

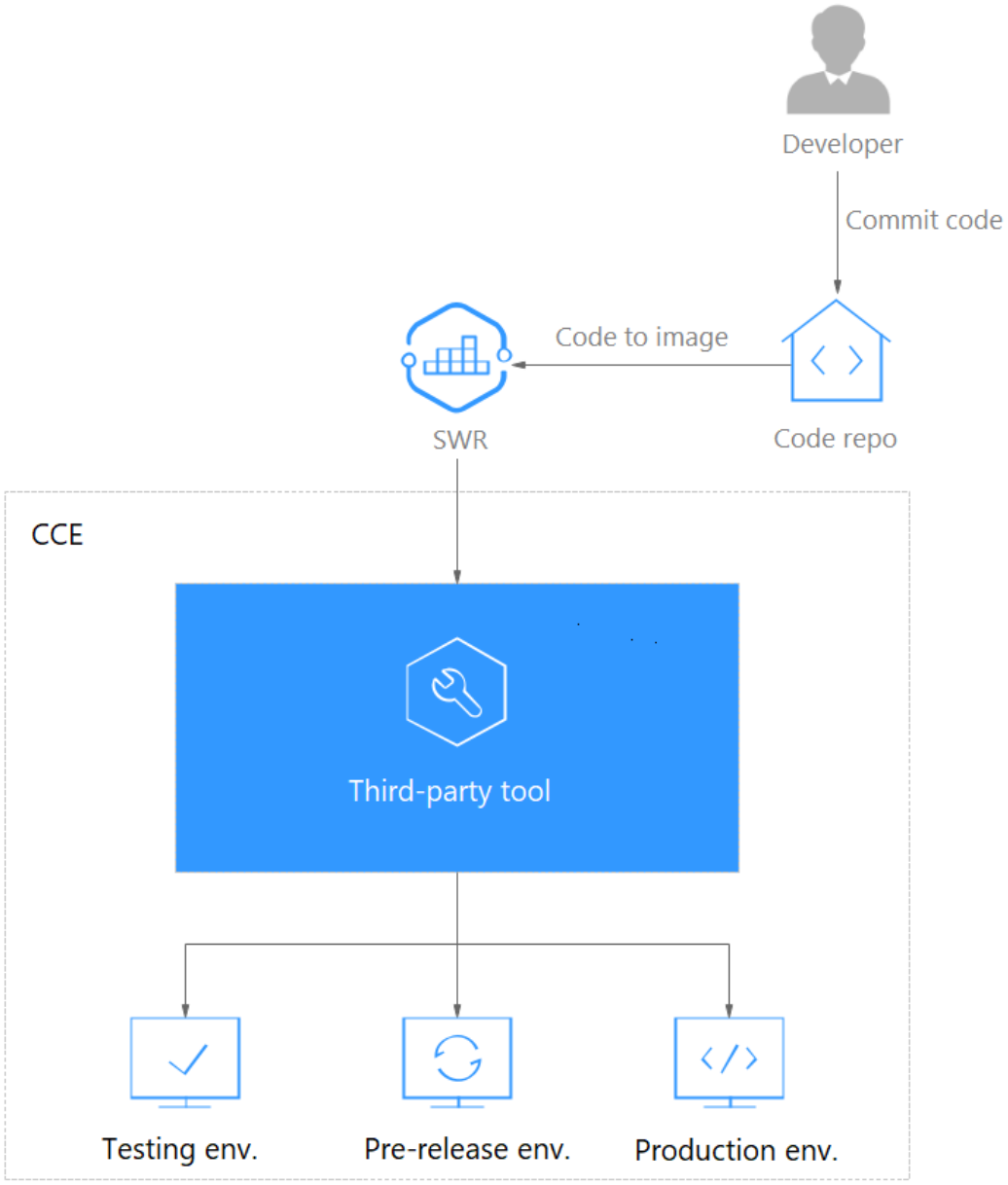
  Proporciona varias API para integrarse con los sistemas CI/CD existentes para una personalización en profundidad.
- Alto rendimiento

  Programa las tareas de forma flexible con una arquitectura completamente en contenedores.

Servicios relacionados

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

Figura 5-5 Cómo funciona DevOps



5.5 Arquitectura de nube híbrida

Escenarios de aplicación

- Implementación multinube y recuperación ante desastres

  Para lograr una alta disponibilidad del servicio, puede implementar aplicaciones en servicios de contenedores desde múltiples proveedores de nube. Cuando una nube está inactiva, la carga de la aplicación se distribuirá automáticamente a otras nubes.

- Distribución del tráfico y ajuste automático

  Los sistemas empresariales grandes deben abarcar instalaciones en la nube en diferentes regiones. También necesitan ser redimensionables automáticamente, pueden comenzar con poco y luego escalar a medida que aumenta la carga del sistema. Esto libera a las

empresas de los costos de planificación, compra y mantenimiento de más instalaciones en la nube de las necesarias y transforma los grandes costos fijos en costos variables mucho más pequeños.

- Migración a la nube y alojamiento de bases de datos

  Las finanzas, la seguridad y otras industrias con una gran preocupación por la confidencialidad de los datos quieren mantener los sistemas críticos en los IDC locales mientras mueven otros sistemas a la nube. Se espera que todos los sistemas, independientemente de los IDC locales o en la nube, se gestionen mediante un panel de control unificado.

- Separación del desarrollo del despliegue

  Para garantizar la seguridad de IP, puede configurar el entorno de producción en una nube pública y el entorno de desarrollo en un IDC local.

Beneficios

Las aplicaciones y los datos pueden migrarse sin inconvenientes entre la red local y la nube, lo que facilita la planificación de recursos y la recuperación ante desastres (DR). Esto es posible a través de contenedores independientes del entorno, conectividad de red entre nubes privadas y públicas, y la capacidad de gestionar contenedores de forma centralizada en CCE y su nube privada.

Ventajas

- DR en la nube

  Multicloud ayuda a proteger los sistemas contra interrupciones. Cuando una nube está defectuosa, las cargas del sistema se desvían automáticamente a otras nubes para garantizar la continuidad del servicio.

- Distribución automática de tráfico

  La latencia de acceso se reduce dirigiendo las solicitudes de los usuarios a la nube regional que está más cerca de donde están los usuarios. Una vez que las aplicaciones en los IDC locales están sobrecargadas, algunas de las solicitudes de acceso a la aplicación se pueden distribuir a la nube con nodos y contenedores escalados automáticamente.

- Implementaciones de servicios separadas y recursos compartidos

  CCE permite el almacenamiento separado de datos de servicios confidenciales y generales, implementaciones separadas en el entorno de desarrollo y el entorno de producción, y ejecución separada de cómputo intensivo y servicios generales. A través del ajuste automático y la gestión unificada de clústeres, sus recursos locales y en la nube pueden trabajar juntos de la manera eficiente.

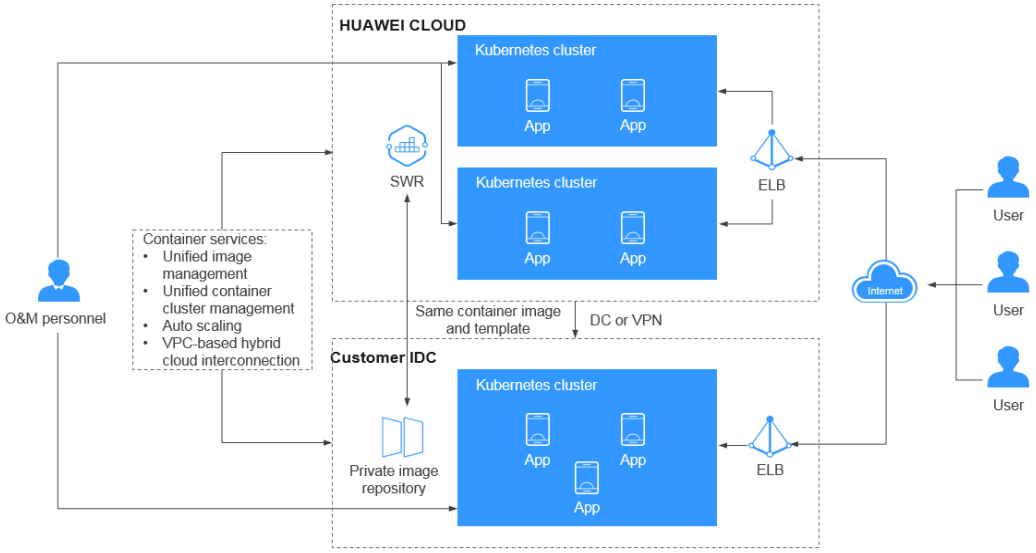
- Costos más bajos

  Los grupos de recursos de nube pública pueden responder rápidamente a los picos de carga mediante el aprovisionamiento automático de recursos. Las operaciones manuales y el mantenimiento ya no son necesarios y usted puede ahorrar mucho.

Servicios relacionados

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

Figura 5-6 Cómo funciona la nube híbrida



5.6 Programación de alto rendimiento

CCE integra Volcano para soportar cómputo de alto rendimiento.

Volcano es un sistema nativo de procesamiento por lotes de Kubernetes. Volcano proporciona una plataforma universal, escalable y estable para ejecutar trabajos de Big Data e IA. Es compatible con los marcos de cómputo generales para tareas de IA, big data, secuenciación de genes y renderizado. La excelencia de Volcano en la programación de tareas y la gestión de chips heterogéneos hace que el funcionamiento y la gestión de tareas sean más eficientes.

Escenario 1: Implementación híbrida de varios tipos de trabajos

Se desarrollan múltiples tipos de marcos de dominio para apoyar negocios en las diferentes industrias. Estos marcos, como Spark, TensorFlow y Flink, funcionan insustituiblemente en sus dominios de servicio. No trabajan solos, ya que los servicios y las empresas son cada vez más complejos. Sin embargo, la programación de recursos se convierte en un problema a medida que los clústeres en estos marcos crecen y un solo servicio puede tener cargas fluctuantes. Por lo tanto, un sistema de planificación unificado tiene una gran demanda.

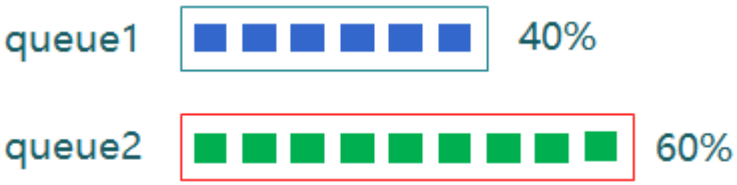
Volcano resume una capa básica común para el cómputo por lotes basado en Kubernetes. Complementa Kubernetes en la programación y proporciona abstracciones de trabajo flexibles y universales para marcos de cómputo. Estas abstracciones (trabajos de Volcano) se implementan a través de plantillas multitarea para describir varios tipos de trabajos (como TensorFlow, Spark, MPI, y PyTorch). Los diferentes tipos de trabajos se pueden ejecutar juntos, y Volcano utiliza su sistema de programación unificada para realizar el uso compartido de recursos del clúster.

Escenario de aplicación 2: Optimización de programación en los escenarios de varias colas

El aislamiento y el uso compartido de recursos a menudo se requieren cuando se utiliza un clúster de Kubernetes. Sin embargo, Kubernetes no admite colas. No puede compartir recursos cuando varios usuarios o departamentos comparten una máquina. Sin el uso compartido de recursos basado en colas, los trabajos de HPC y big data no se pueden ejecutar.

Volcano soporta múltiples mecanismos de intercambio de recursos con colas. Puede establecer **weight** de una cola. El clúster asigna recursos a la cola calculando la relación entre el peso de la cola y el peso total de todas las colas. También puede establecer **capability** de recursos de una cola para determinar el límite superior de recursos que puede utilizar la cola.

Por ejemplo, en la siguiente figura, la cola 1 se asigna el 40% de los recursos del clúster y el 60% para la cola 2. De esta manera, se pueden asignar dos colas a diferentes departamentos o proyectos para usar recursos en el mismo clúster. Si una cola tiene recursos inactivos, se pueden asignar a trabajos de otra cola.

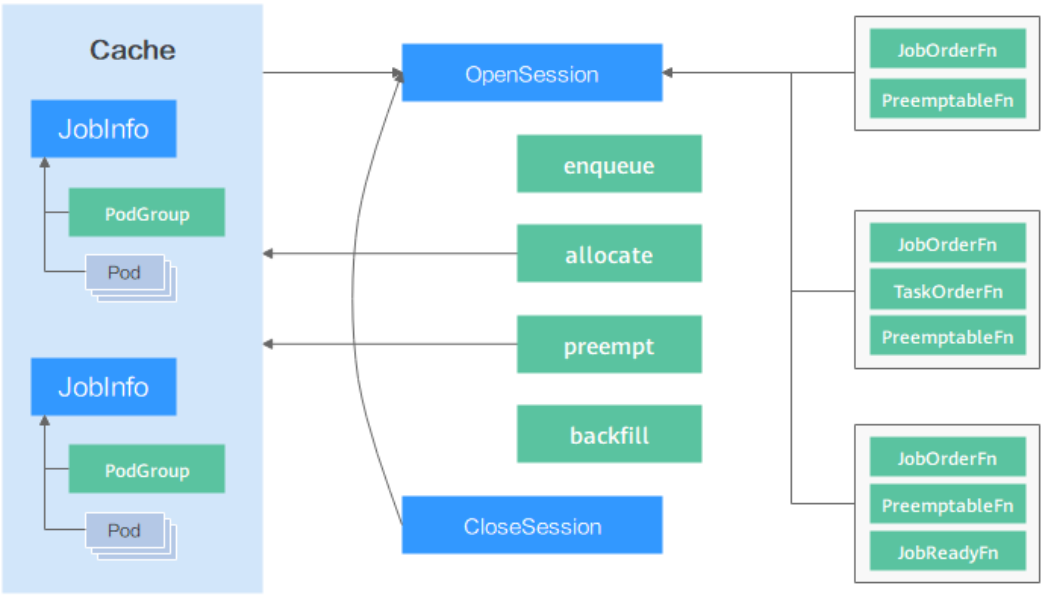


Escenario de la aplicación 3: Múltiples políticas avanzadas de programación

Los contenedores se programan en nodos que satisfacen sus requisitos en recursos de cómputo, como CPU, memoria y GPU. Normalmente, habrá más de un nodo calificado. Cada uno podría tener un volumen diferente de recursos disponibles para nuevas cargas de trabajo. Volcano analiza automáticamente la utilización de recursos de cada plan de programación y lo ayuda a lograr los resultados óptimos de implementación con gran facilidad.

La siguiente figura muestra cómo el planificador de Volcano programa los recursos. En primer lugar, el planificador carga la información de pod y PodGroup en el servidor API en la caché del planificador. En una sesión de programación, Volcano pasa por tres fases: "OpenSession", llamada de acción y "CloseSession". En OpenSession se carga la política de programación que configuraste en el complemento del planificador. Durante la llamada de

acción, las acciones configuradas se llaman una por una y se utiliza la política de programación cargada. En CloseSession se realizan las operaciones finales para completar la programación.



El planificador de Volcano proporciona plugins para soportar múltiples acciones de programación (como enqueue, allocate, preempt, reclaim y backfill) y políticas de programación (como gang, priority, drf, proportion y binpack). Puede configurarlos según sea necesario. Las API proporcionadas por el planificador también se pueden utilizar para el desarrollo personalizado.

Escenario de aplicación 4: Programación de recursos de alta precisión

Volcano proporciona políticas de programación de recursos de alta precisión para trabajos de inteligencia artificial y big data para mejorar la eficiencia informática. Tomemos a TensorFlow como ejemplo. Configure la afinidad entre ps y trabajador y la antiafinidad entre ps y ps, de modo que ps y trabajador al mismo nodo. Esto mejora el rendimiento de la red y la interacción de datos entre ps y el trabajador, mejorando así la eficiencia informática. Sin embargo, al programar pods, el planificador predeterminado de Kubernetes solo comprueba si las configuraciones de afinidad y antiafinidad de estos pods entran en conflicto con las de todos los pods en ejecución en el clúster y no tiene en cuenta los pods posteriores que también necesiten programación.

El algoritmo de topología de tareas proporcionado por Volcano calcula las prioridades de tareas y nodos en función de las configuraciones de afinidad y antiafinidad entre tareas en un trabajo. Las políticas de afinidad y antiafinidad de tareas en un trabajo y el algoritmo de topología de tareas garantizan que las tareas con configuraciones de afinidad se planifiquen preferentemente para el mismo nodo. y los pods con configuraciones antiafinidad están programados para los diferentes nodos. La diferencia entre el algoritmo de topología de tareas y el planificador predeterminado de Kubernetes es que el algoritmo de topología de tareas considera que los pods se programarán como un todo. Cuando los pods se programan por lotes, los ajustes de afinidad y antiafinidad entre los pods no programados se consideran y se aplican a los procesos de programación de pods en función de las prioridades.

Beneficios

La ejecución de contenedores en servidores en la nube acelerados por GPU de alto rendimiento mejora significativamente el rendimiento informático de IA de tres a cinco veces. Las GPU pueden costar mucho y compartir una GPU entre contenedores reduce en gran medida los costos de cómputo de IA. Además de las ventajas de rendimiento y costos, CCE también ofrece clústeres totalmente administrados que ocultarán toda la complejidad en la implementación y gestión de sus aplicaciones de IA para que pueda centrarse en el desarrollo de alto valor.

Ventajas

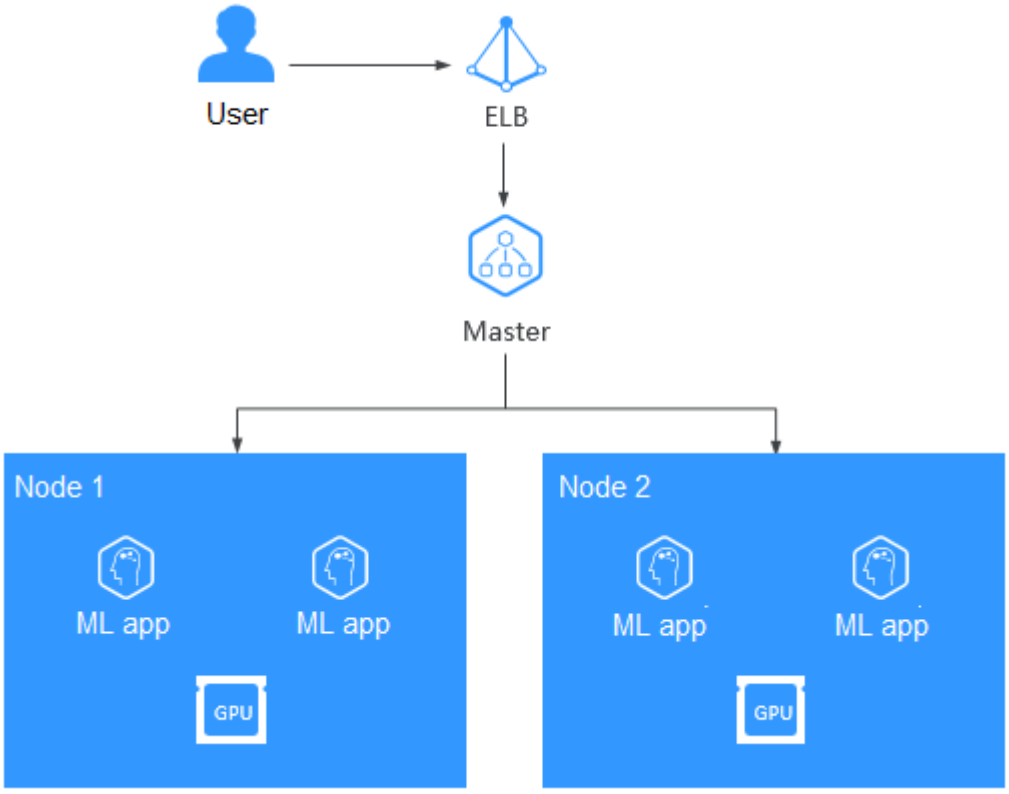
Al integrar Volcano, CCE tiene las siguientes ventajas en la ejecución de trabajos de cómputo de alto rendimiento, big data e inteligencia artificial:

- **Implementación híbrida** de trabajos de HPC, big data e IA
- **Programación optimizada de múltiples colas**: se pueden utilizar varias colas para compartir recursos con varios tenants y planificar grupos en función de prioridades y períodos de tiempo.
- **Políticas de programación avanzadas**: programación de grupos, programación justa, preferencia de recursos y topología de GPU
- **Plantilla multitarea**: Puede utilizar una plantilla para definir varias tareas en un solo trabajo de volcán, más allá del límite de los recursos nativos de Kubernetes. Volcano Jobs puede describir varios tipos de trabajo, como TensorFlow, MPI y PyTorch.
- **Plugins de extensión de trabajo**: El Volcano Controller le permite configurar plugins para personalizar la preparación y limpieza del entorno en etapas como el envío de trabajos y la creación de pods. Por ejemplo, antes de enviar un trabajo MPI común, puede configurar el complemento SSH para proporcionar la información SSH de los recursos de pod.

Servicios relacionados

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

Figura 5-7 Cómo funciona el cómputo de IA



6 Notas y restricciones

Esta sección describe las notas y restricciones sobre el uso de CCE.

Clústeres y nodos

- Después de crear un clúster, no se pueden cambiar los siguientes elementos:
  - Tipo de clúster. Por ejemplo, cambie un **Kunpeng cluster** a un **CCE cluster**.
  - Número de nodos principales en el clúster.
  - AZ de un nodo principal.
  - Configuración de red del clúster, como la VPC, la subred, el bloque CIDR de contenedor, el bloque CIDR de servicio, la configuración IPv6 y la configuración kube-proxy (reenvío).
  - Modelo de red. Por ejemplo, cambie la **tunnel network** a la **VPC network**.
- Las aplicaciones no se pueden migrar entre los diferentes espacios de nombres.
- Actualmente, las instancias (nodos) de ECS creadas admiten los modos de facturación de pago por uso y anual/mensual. Otros recursos (como los balanceadores de carga) admiten el modo de facturación de pago por uso. Puede cambiar el modo de facturación de pago por uso a anual/mensual en la consola de gestión para las instancias de ECS creadas.
- Los nodos creados durante la creación de clústeres admiten modos de facturación de pago por uso y anual/mensual, pero con las siguientes restricciones:
  - Si el clúster que se va a crear es de pago por uso, los nodos creados en el clúster también deben ser de pago por uso.
  - Si el clúster que se va a crear se factura anualmente/mensualmente, los nodos del clúster se pagan por uso o se facturan anualmente/mensualmente.
  - Si los nodos agregados después de la creación del clúster se facturan anualmente/mensualmente, deben renovarse por separado del clúster.

  Nota: Si adquiere un nodo después de crear un clúster, el modo de facturación del nodo no está restringido por el del clúster.
- Los recursos subyacentes, como los ECS (nodos), están limitados por las cuotas y su inventario. Por lo tanto, solo algunos nodos pueden crearse correctamente durante la creación del clúster, el ajuste del clúster o el ajuste automático.
- Las especificaciones ECS (nodo) deben ser superiores a 2 núcleos y 4 GB de memoria.

- Para acceder a un clúster de CCE a través de una VPN, asegúrese de que el bloque de CIDR de VPN no entre en conflicto con el bloque de CIDR de VPC donde reside el clúster y el bloque de CIDR de contenedor.

Redes

- De forma predeterminada, se accede a un NodePort Service dentro de una VPC. Si necesita usar un EIP para acceder a un NodePort Service a través de redes públicas, vincule un EIP al nodo del clúster de antemano.
- Los servicios de LoadBalancer permiten acceder a cargas de trabajo desde redes públicas a través de **ELB**. Este modo de acceso tiene las siguientes restricciones:
  - Se recomienda que los balanceadores de carga creados automáticamente no sean utilizados por otros recursos. De lo contrario, estos balanceadores de carga no se pueden eliminar por completo, lo que provoca los recursos residuales.
  - No cambie el nombre de oyente para el balanceador de carga en clústeres de v1.15 y anteriores. De lo contrario, no se puede acceder al balanceador de carga.
- Restricciones en las políticas de red:
  - Solo los clústeres que utilizan el modelo de red de túnel admiten políticas de red.
  - Las salidas no son compatibles con las políticas de red.
  - El aislamiento de red no es compatible con las direcciones IPv6.

Volúmenes

- Restricciones en los volúmenes de EVS:
  - De forma predeterminada, CCE crea discos de EVS facturados en modo **pago por uso**. Para usar discos de EVS facturados en modo **anual/mensual**, consulte la sección **Discos de EVS de facturación anual/mensual**.
  - Los discos de EVS no se pueden conectar a través de AZ y no pueden ser utilizados por múltiples cargas de trabajo, múltiples pods de la misma carga de trabajo o múltiples trabajos.
  - Los datos de un disco compartido no se pueden compartir entre los nodos de un clúster de CCE. Si el mismo disco de EVS está conectado a varios nodos, pueden producirse conflictos de lectura y escritura y conflictos de caché de datos. Al crear una implementación, se recomienda crear solo un pod si desea usar discos de EVS.
  - Para los clústeres anteriores a v1.19.10, si se utiliza una política de HPA para escalar una carga de trabajo con volúmenes de EVS montados, los pods existentes no se pueden leer ni escribir cuando se programa un nuevo pod en otro nodo.

    Para los clústeres de v1.19.10 y las versiones posteriores, si se utiliza una política de HPA para escalar una carga de trabajo con un volumen de EVS montado, no se puede iniciar un nuevo pod porque no se pueden conectar los discos deEVS.
  - Cuando crea un StatefulSet y agrega un volumen de almacenamiento en la nube, no se pueden utilizar los volúmenes de EVS existentes.
  - Los discos de EVS que tienen particiones o que tienen sistemas de archivos no ext4 no se pueden importar.
  - El almacenamiento de contenedores en clústeres de CCE de Kubernetes 1.13 o posterior admite la encriptación. Actualmente, la encriptación E2E solo se admite en ciertas regiones.
  - Los volúmenes de EVS no se pueden crear en proyectos de empresa especificados. Sólo se admite el proyecto de empresa predeterminado.

- Restricciones en los volúmenes de SFS:
  - El almacenamiento de contenedores en clústeres de CCE de Kubernetes 1.13 o posterior admite la encriptación. Actualmente, la encriptación E2E solo se admite en ciertas regiones.
  - No se pueden crear los volúmenes en los proyectos de empresa especificados. Sólo se admite el proyecto de empresa predeterminado.
- Restricciones en los volúmenes de OBS:
  - Los clústeres de CCE de v1.7.3-r8 y de anterior no admiten los volúmenes de OBS. Debe actualizar estos clústeres o crear los clústeres de una versión posterior que admita OBS.
  - Los clústeres de Kunpeng no soportan obsfs. Por lo tanto, no se pueden montar los sistemas de archivos paralelos.
  - No se pueden crear los volúmenes en los proyectos de empresa especificados. Sólo se admite el proyecto de empresa predeterminado.
- Restricciones en las instantáneas y copias de seguridad:
  - La función de instantánea está disponible **solo para los clústeres de v1.15 o posterior** y requiere el complemento everest basado en CSI.
  - El subtipo (E/S común, E/S alta o E/S ultra alta), modo de disco (SCSI o VBD), encriptación de datos, estado de uso compartido, y la capacidad de un disco de EVS creado a partir de una instantánea debe ser la misma que la del disco asociado a la instantánea. Estos atributos no se pueden modificar después de ser consultados o establecidos.

Servicios

Un servicio es un objeto de recurso de Kubernetes que define un conjunto lógico de pods y una política para acceder a ellos.

Se puede crear un máximo de servicios de 6,000 en cada espacio de nombres.

Recursos del clúster de CCE

Hay cuotas de recursos para sus clústeres de CCE en cada región.

Concepto	Restricciones en los usuarios comunes
Número total de clústeres en una región	50
Número de nodos en un clúster (escala de gestión de clústeres)	Puede seleccionar 50, 200, 1,000 o 2,000 nodos. Se admite un máximo de 5,000 de nodos.
Número máximo de pods de contenedores creados en cada nodo de trabajo	Este número se puede establecer en la consola al crear un clúster. En el modelo de red de VPC, se puede crear un máximo de 256 pods.

Recursos de nube subyacentes dependientes

Categoría	Concepto	Restricciones en los usuarios comunes
Cómputo	Pods	1,000
	Núcleos	8,000
	Capacidad de RAM (MB)	16384000
Redes	VPC por cuenta	5
	Subredes por cuenta	100
	Grupos de seguridad por cuenta	100
	Reglas de grupo de seguridad por cuenta	5000
	Rutas por tabla de ruta	100
	Rutas por VPC	100
	Interconexión de VPC por región	50
	ACL de red por cuenta	200
	Gateway de conexión de nivel 2 por cuenta	5
Balanceo de carga	Balanceadores de carga elásticos	50
	Oyentes del balanceador de carga	100
	Certificados del balanceador de carga	120
	Políticas de reenvío del balanceador de carga	500
	Grupo de host de backend del balanceador de carga	500
	Servidor de backend del balanceador de carga	500

7 Detalles de precios

Conceptos de facturación:

Cloud Container Engine (CCE) es gratuito. Solo paga por los recursos (como los nodos) creados cuando utiliza CCE. Hay dos tipos de artículos de facturación:

1. **Clusters**: La tarifa de clúster es el costo de los recursos utilizados por los nodos principales. La tarifa varía según el tipo de clúster y el tamaño del clúster. Los tipos de clúster incluyen clúster de VM y clúster de BMS (el número de nodos principales determina si un clúster está altamente disponible). El tamaño del clúster (también llamado escala de gestión) indica el número máximo de nodos permitidos en un clúster.

   La escala de gestión indica el número de ECS o de BMS en un clúster.

   Para obtener más información, consulte **Detalles de precios de CCE**.
2. **IaaS resources**: se factura el costo de los recursos de IaaS creados para ejecutar nodos de trabajo en el clúster. Los recursos de IaaS, que se crean manualmente o automáticamente, incluyen ECS, discos de EVS, EIP, ancho de banda y balanceadores de carga.

   Para obtener más información sobre los precios, consulte **Detalles de precios del producto**.

Modos de facturación

CCE se factura sobre una base de pago por uso o anual/mensual.

- Pago por uso: Es un modo de pago después de uso. La facturación se inicia cuando se aprovisiona un recurso y se detiene cuando se elimina el recurso. Puede utilizar los recursos en la nube según sea necesario y dejar de pagarlos cuando ya no los necesite. No hay pago por adelantado por exceso de capacidad.

Los siguientes son principios de precios en el caso de la hibernación del clúster de CCE o el apagado del nodo. Tenga en cuenta que hay muchos tipos de nodos de clúster y ECS se utiliza como ejemplo.

- **Cluster hibernation**: después de hibernar un clúster, se detendrá la facturación de los recursos utilizados por los nodos principales.
- **Node shutdown**: la facturación del nodo del trabajador se detiene cuando se detiene el nodo. Tenga en cuenta que la hibernación de un clúster no detendrá los nodos de trabajo en el clúster. Para detener un ECS, inicie sesión en la consola de ECS. Para obtener más información, consulte **Detener un nodo**.

  Los ECS detenidos no se facturan. Para obtener más información, consulte **Facturación de ECS**.

- Anual/mensual: Es un modo de pago antes de usar. La facturación anual/mensual proporciona un descuento más significativo que el pago por uso y se recomienda para el uso a largo plazo de los servicios en la nube. Cuando usted compra un paquete anual/mensual, el sistema deducirá el costo del paquete de su cuenta en la nube según las especificaciones elegidas.
- Cambio del modo de facturación: el modo de facturación no se puede cambiar dentro del ciclo de facturación.

- Los clústeres siguen un plan de precios por niveles. Los precios de cada nivel varían según el tamaño y el tipo del clúster.
- Una vez que una suscripción mensual/anual ha caducado o un recurso de pago por uso está en mora, Huawei Cloud proporciona un período de tiempo durante el cual puede renovar el recurso o recargar su cuenta. Dentro del período de gracia, todavía puede acceder y utilizar su servicio en la nube. Para obtener más información, consulte **¿Qué es un período de gracia? Cuánto dura el período de gracia de Huawei Cloud. ¿Qué es un período de retención? Cuánto dura el período de retención de Huawei Cloud.**

Cambios de configuración

De pago por uso a facturación anual/mensual: puede cambiar el modo de facturación de clúster de pago por uso a facturación anual/mensual. Después del cambio, todos los nodos principales, nodos de trabajo y recursos en la nube (como discos de EVS y EIP) utilizados por su clúster se facturarán anualmente/mensualmente y se generará un nuevo pedido. Los nodos y los recursos en la nube estarán listos para su uso inmediatamente después de pagar el nuevo pedido.

De la facturación anual/mensual al pago por uso: los clústeres que se facturan anualmente/mensualmente no pueden cambiar a pago por uso dentro del ciclo de facturación. Tenga en cuenta que los clústeres de pago por uso se pueden eliminar directamente, pero los clústeres que se facturan anualmente/mensualmente no se pueden eliminar. Para dejar de usar los clústeres que se facturan anualmente/mensualmente, vaya al Centro de facturación y **cancele la suscripción a los mismos**.

Notas

- Los cupones en efectivo no se devolverán después de degradar las especificaciones de los servidores en la nube que se compran con cupones en efectivo.

- Tendrá que pagar la diferencia de precio entre las especificaciones originales y nuevas después de actualizar las especificaciones del servidor en la nube.
- La reducción de las especificaciones del servidor en la nube (la cantidad de recursos de CPU o memoria) perjudicará el rendimiento del servidor en la nube.
- Si degrada las especificaciones del servidor en la nube y luego las actualiza a las especificaciones originales, todavía tendrá que pagar la diferencia de precio incurrida por la actualización.

8 Gestión de permisos

CCE le permite asignar permisos a los usuarios y grupos de usuarios de IAM en sus cuentas de tenant. CCE combina las ventajas de Identity and Access Management (IAM) y Kubernetes Role-based Access Control (RBAC) para proporcionar una variedad de métodos de autorización, incluida la autorización de grano fino/token de IAM y la autorización de ámbito de clúster/espacio de nombres.

Los permisos de CCE se describen de las siguientes maneras:

- **Cluster-level permissions**: la gestión de permisos a nivel de clúster evoluciona de la función de autorización de políticas del sistema de IAM. Los usuarios de IAM del mismo grupo de usuarios tienen los mismos permisos. En IAM, puede configurar las políticas del sistema para describir qué grupos de usuarios de IAM pueden realizar las operaciones en los recursos del clúster. Por ejemplo, puede conceder al grupo de usuarios A que cree y elimine el clúster X, agregue un nodo o instale un complemento, mientras que concede al grupo de usuarios B que vea información sobre el clúster X.

  Los permisos a nivel de clúster implican API de CCE que no son de Kubernetes y admiten las políticas de IAM detalladas y las capacidades de gestión de proyectos empresariales.

- **Namespace-level permissions**: puede regular el acceso de los usuarios o grupos de usuarios a los **Kubernetes resources**, como cargas de trabajo, trabajos y servicios, en un único espacio de nombres basado en sus roles RBAC de Kubernetes. El CCE también se ha mejorado sobre la base de las capacidades de código abierto. Admite la autorización de RBAC basada en el usuario o grupo de usuarios de IAM, y la autenticación de RBAC en el acceso a las API mediante tokens de IAM.

  Los permisos a nivel de espacio de nombres implican API de CCE Kubernetes y se mejoran en función de las capacidades RBAC de Kubernetes. Los permisos de nivel de espacio de nombres se pueden conceder a los usuarios o grupos de usuarios de IAM para la autenticación y la autorización, pero son independientes de las políticas de IAM detalladas. Para obtener más información, consulte **Uso de la autorización RBAC**.

- **Cluster-level permissions**solo se configuran para recursos relacionados con clústeres (como clústeres y nodos). También debe configurar los **permisos de espacio de nombres** para operar los recursos de Kubernetes (como cargas de trabajo, trabajos y servicios).
- Después de crear un clúster de v1.11.7-r2 o posterior, CCE le asigna automáticamente los permisos cluster-admin de todos los espacios de nombres del clúster, lo que significa que tiene control total sobre el clúster y todos los recursos de todos los espacios de nombres.

Permisos al nivel de clúster (asignados mediante políticas de sistema de IAM)

De forma predeterminada, los nuevos usuarios de IAM no tienen permisos asignados. Debe agregar un usuario a uno o más grupos y adjuntar políticas o roles de permisos a estos grupos. Los usuarios heredan permisos de los grupos a los que se agregan y pueden realizar operaciones específicas a servicios en la nube según los permisos.

CCE es un servicio a nivel de proyecto implementado y accedido en regiones físicas específicas. Para asignar los permisos de CCE a un grupo de usuarios, especifique el ámbito como proyectos específicos de la región y seleccione proyectos para que los permisos surtan efecto. Si se selecciona **All projects**, los permisos surtirán efecto para el grupo de usuarios en todos los proyectos específicos de la región. Al acceder a CCE, los usuarios deben cambiar a una región en la que se les haya autorizado a usar el servicio de CCE.

Puede conceder permisos a los usuarios mediante roles y políticas.

- Roles: Tipo de mecanismo de autorización de grano grueso que define permisos relacionados con las responsabilidades del usuario. Este mecanismo proporciona solo un número limitado de roles de nivel de servicio para la autorización. Al usar roles para asignar permisos, también debe asignar otros roles de los que dependen los permisos para que surtan efecto. Sin embargo, los roles no son una opción ideal para la autorización detallada y el control de acceso seguro.
- Políticas: Un tipo de mecanismo de autorización detallado que define los permisos necesarios para realizar operaciones en recursos de nube específicos bajo ciertas condiciones. Este mecanismo permite una autorización más flexible basada en políticas, cumpliendo los requisitos para un control de acceso seguro. Por ejemplo, puede asignar a los usuarios únicamente los permisos para gestionar un determinado tipo de clústeres y nodos.

Tabla 8-1 enumera todos los permisos del sistema admitidos por CCE.

Tabla 8-1 Permisos del sistema admitidos por CCE

Nombre de rol/ política	Descripción	Tipo	Dependencias
CCE Administrator	Permisos de lectura y escritura para clústeres de CCE y todos los recursos (incluidas cargas de trabajo, nodos, trabajos y servicios) en los clústeres	Rol	Los usuarios a los que se han concedido permisos de esta política también deben tener permisos de las siguientes políticas: **Global service project**: Visor de bucket de OBS y Administrador de OBS **Region-specific projects**: tenant invitado, administrador de servidor, administrador de ELB, administrador de SFS, administrador de SWR y FullAccess de APM **NOTA** Los usuarios con políticas de **CCE Administrator** y **NAT Gateway Administrator** pueden usar funciones de NAT Gateway para clústeres.
CCE FullAccess	Permisos de operación comunes en los recursos de clúster de CCE, excluyendo los permisos a nivel de espacio de nombres para los clústeres (con Kubernetes RBAC habilitado) y las operaciones de administrador con privilegios, como la configuración de delegación y la generación de certificados de clúster	Política	No hay.
CCE ReadOnly Access	Permisos para ver los recursos del clúster de CCE, excluyendo los permisos a nivel de espacio de nombres de los clústeres (con Kubernetes RBAC habilitado)	Política	No hay.

Tabla 8-2 Operaciones comunes soportadas por las políticas del sistema CCE

Operación	CCE ReadOnlyAccess	CCE FullAccess	CCE Administrator
Creación de un clúster	x	√	√
Eliminación de un clúster	x	√	√
Actualización de un clúster, por ejemplo, actualizar los parámetros de programación de nodos de clúster y proporcionar soporte RBAC a clústeres	x	√	√
Actualización de un clúster	x	√	√
Despierta de un clúster	x	√	√
Hibernación de un clúster	x	√	√
Listado de todos los clústeres	√	√	√
Consulta de los detalles del clúster	√	√	√
Adición de un nodo	x	√	√
Eliminación de uno o más nodos	x	√	√
Actualización de un nodo de clúster, por ejemplo, actualizar el nombre del nodo	x	√	√
Consulta de detalles de nodo	√	√	√
Listado de todos los nodos	√	√	√
Listado de todos los trabajos	√	√	√
Supresión de uno o más trabajos de clúster	x	√	√
Consulta de detalles del trabajo	√	√	√
Creación de un volumen de almacenamiento	x	√	√
Eliminación de un volumen de almacenamiento	x	√	√

Operación	CCE ReadOnlyAccess	CCE FullAccess	CCE Administrator
Realización de operaciones en todos los recursos de Kubernetes	√	√	√
Realización de todas las operaciones en un Elastic Cloud Server (ECS)	x	√	√
Realización de todas las operaciones en discos de Elastic Volume Service (EVS) Los discos de EVS se pueden conectar a servidores en la nube y escalar a una mayor capacidad cuando sea necesario.	x	√	√
Realización de todas las operaciones en VPC Un clúster debe ejecutarse en una VPC. Al crear un espacio de nombres, debe crear o asociar una VPC para el espacio de nombres de modo que todos los contenedores del espacio de nombres se ejecuten en la VPC.	x	√	√
Consulta de detalles de todos los recursos en un ECS En CCE, un nodo es un ECS con múltiples discos de EVS.	√	√	√
Listado de todos los recursos en un ECS	√	√	√
Consulta de detalles sobre todos los recursos de discos de EVS. Los discos de EVS se pueden conectar a los servidores en la nube y escalar a una mayor capacidad siempre que sea necesario.	√	√	√

Operación	CCE ReadOnlyAccess	CCE FullAccess	CCE Administrator
Listado de todos los recursos de EVS	√	√	√
Consulta de detalles sobre todos los recursos de VPC Un clúster debe ejecutarse en una VPC. Al crear un espacio de nombres, debe crear o asociar una VPC para el espacio de nombres de modo que todos los contenedores del espacio de nombres se ejecuten en la VPC.	√	√	√
Listado de todos los recursos de VPC	√	√	√
Consulta de detalles sobre todos los recursos de Elastic Load Balance (ELB)	x	x	√
Listado de todos los recursos de ELB	x	x	√
Consulta de los detalles de recursos del Scalable File Service (SFS)	√	√	√
Listado de todos los recursos de SFS	√	√	√
Consulta de los detalles de recursos de Application Operations Management (AOM)	√	√	√
Listado de recursos de AOM	√	√	√
Realización de todas las operaciones en reglas de ajuste automático de AOM	√	√	√

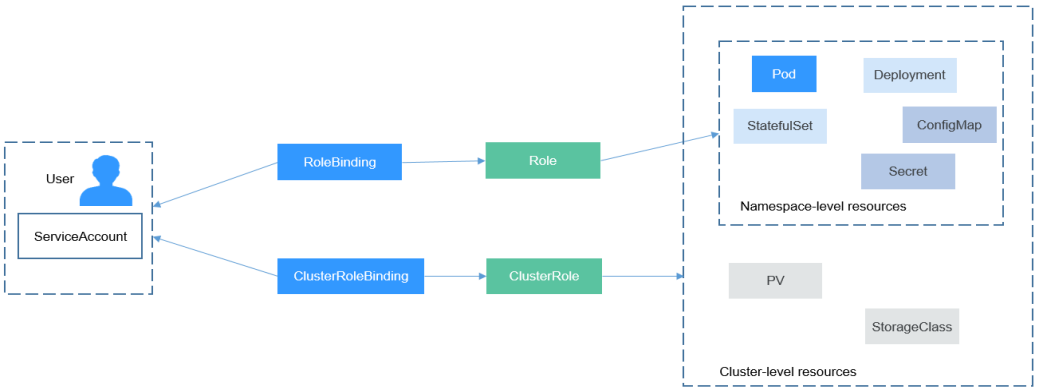
Permisos a nivel de espacio de nombres (asignados mediante Kubernetes RBAC)

Puede regular el acceso de los usuarios o los grupos de usuarios a los recursos de Kubernetes en un único espacio de nombres basado en sus roles de Kubernetes RBAC. La API de RBAC declara cuatro tipos de objetos de Kubernetes: Role, ClusterRole, RoleBinding y ClusterRoleBinding, que se describen a continuación:

- Role: define un conjunto de reglas para acceder a los recursos de Kubernetes en un espacio de nombres.
- RoleBinding: define la relación entre usuarios y roles.
- ClusterRole define un conjunto de reglas para acceder a los recursos de Kubernetes en un clúster (incluidos todos los espacios de nombres).
- ClusterRoleBinding: define la relación entre los usuarios y los roles de clúster.

Role y ClusterRole especifican acciones que se pueden realizar en recursos específicos. RoleBinding y ClusterRoleBinding vinculan roles a usuarios específicos, grupos de usuarios o cuentas de servicio. Consulta la siguiente figura.

Figura 8-1 Vinculación de roles



En la consola de CCE, puede asignar permisos a un usuario o grupo de usuarios para tener acceso a recursos en uno o varios espacios de nombres. De forma predeterminada, la consola de CCE proporciona las siguientes ClusterRoles:

- view: tiene permiso para ver recursos de espacio de nombres.
- edit: tiene permiso para modificar los recursos del espacio de nombres.
- admin: tiene todos los permisos en el espacio de nombres.
- cluster-admin: tiene todos los permisos en el clúster.
- psp-global: controla aspectos de seguridad sensibles de la especificación del pod.

Además de cluster-admin, admin, edit, y view, puede definir Roles y RoleBindings para configurar los permisos para agregar, eliminar, modificar y consultar recursos, como pods, implementaciones y servicios, en el espacio de nombres.

Enlaces útiles

- **Descripción general del servicio IAM**
- **Concesión de permisos a nivel de clúster**
- **Políticas de permisos y acciones admitidas**

9 Conceptos básicos

9.1 Conceptos básicos

CCE proporciona clústeres Kubernetes con gran capacidad de escalamiento, de alto rendimiento y de clase empresarial; además es compatible con contenedores Docker. Con CCE, puede implementar, gestionar y escalar fácilmente aplicaciones en contenedores en la nube.

La consola de gráfica de CCE permite las experiencias de usuario E2E. Además, CCE es compatible con las API nativas de Kubernetes y kubectl. Antes de usar CCE, se recomienda que comprenda los conceptos básicos relacionados.

Clúster

Un clúster es un grupo de uno o más servidores en la nube (también conocidos como nodos) en la misma subred. Tiene todos los recursos en la nube (incluyendo VPC y recursos informáticos) necesarios para ejecutar contenedores.

Nodo

Un nodo es un servidor en la nube (máquina virtual o física) que ejecuta una instancia de Docker Engine. Los contenedores se implementan, ejecutan y administran en nodos. El agente de nodo (kubelet) se ejecuta en cada nodo para gestionar instancias de contenedor en el nodo. Se puede escalar el número de nodos de un clúster.

Grupo de nodos

Un grupo de nodos contiene un nodo o un grupo de nodos con una configuración idéntica en un clúster.

Virtual Private Cloud (VPC)

Una VPC es una red virtual aislada lógicamente que facilita la gestión y la configuración de redes internas de forma segura. Los recursos de la misma VPC pueden comunicarse entre sí, pero aquellos de diferentes VPC no pueden comunicarse entre sí de forma predeterminada. Las VPC proporcionan las mismas funciones de red que las redes físicas y también servicios de red avanzados, como direcciones IP elásticas y grupos de seguridad.

Grupo de seguridad

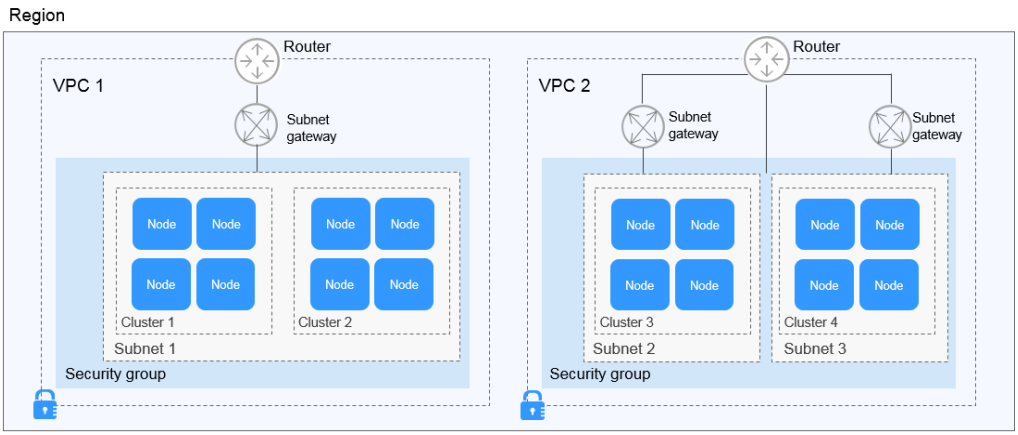
Un grupo de seguridad es un conjunto de reglas de control de acceso para los ECS que tengan los mismos requisitos de protección de seguridad y que sean de confianza recíproca en una VPC. Después de crear un grupo de seguridad, se pueden crear diferentes reglas de acceso para que el grupo de seguridad proteja los ECS que se agreguen a este grupo de seguridad.

Relación entre clústeres, VPC, grupos de seguridad y nodos

Como se muestra en **Figura 9-1**, una región puede comprender múltiples VPC. Una VPC consta de una o más subredes. Las subredes se comunican entre sí a través de un gateway de subred. Se crea un clúster en una subred. Hay tres escenarios:

- Se crean los diferentes clústeres en las diferentes VPC.
- Se crean los diferentes clústeres en la misma subred.
- Se crean los diferentes clústeres en las diferentes subredes.

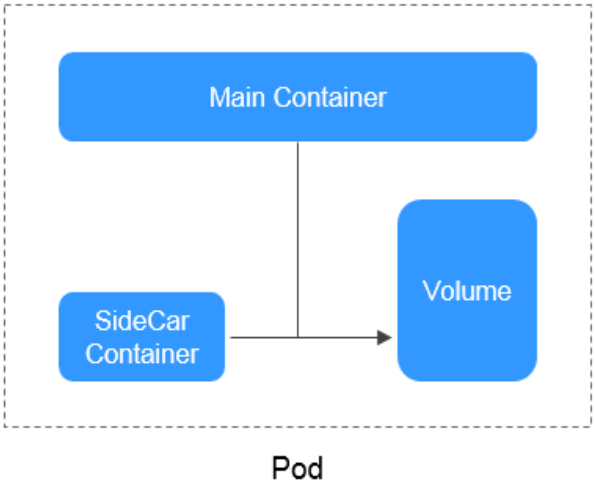
Figura 9-1 Relación entre clústeres, VPC, grupos de seguridad y nodos



Pod

Un pod es la unidad más pequeña y sencilla del modelo de objetos de Kubernetes que cree o implemente. Un pod encapsula un contenedor de aplicación (o, en algunos casos, varios contenedores), recursos de almacenamiento, una dirección IP de red única y opciones que rigen la ejecución de los contenedores.

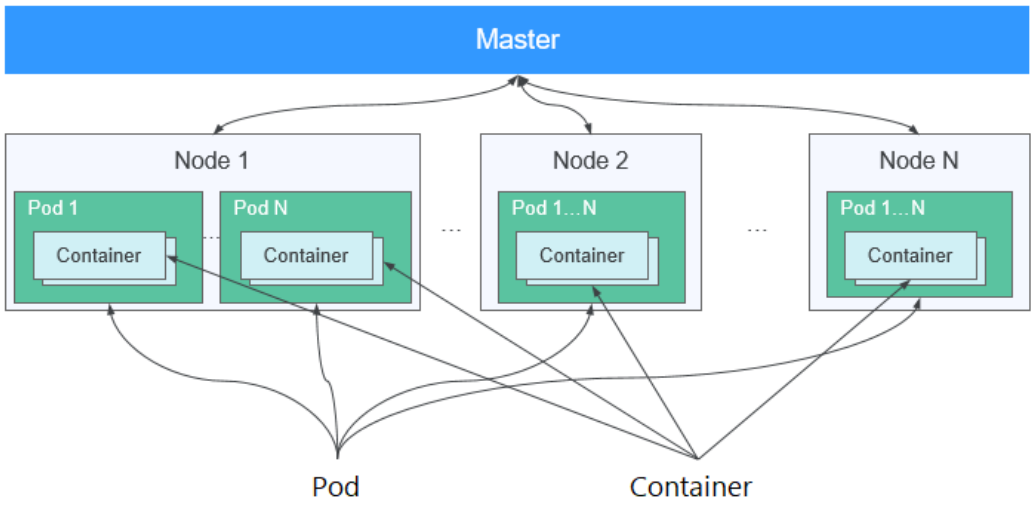
Figura 9-2 Pod



Contenedor

Un contenedor es una instancia en ejecución de una imagen Docker. Se pueden ejecutar varios contenedores en un nodo. Los contenedores son en realidad procesos de software. A diferencia de los procesos de software tradicionales, los contenedores tienen un espacio de nombres separado y no se ejecutan directamente en un host.

Figura 9-3 Relaciones entre pods, contenedores y nodos

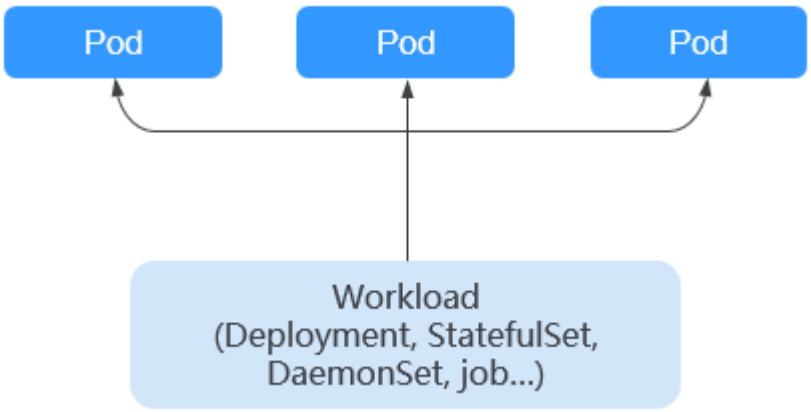


Carga de trabajo

Una carga de trabajo es una aplicación que se ejecuta en Kubernetes. No importa cuántos componentes haya en su carga de trabajo, puede ejecutarlo en un grupo de pods de Kubernetes. Una carga de trabajo es un modelo abstracto de un grupo de pods en Kubernetes. Las cargas de trabajo clasificadas en Kubernetes incluyen Deployments, StatefulSets, DaemonSets, trabajos y trabajos cron.

- **Deployment**: Los pods son completamente independientes entre sí y funcionalmente idénticos. Cuentan con ajuste automático y actualización de balanceo. Los ejemplos típicos incluyen Nginx y WordPress.
- **StatefulSet**: Los pods no son completamente independientes entre sí. Tienen un almacenamiento estable y persistente y cuentan con una implementación y eliminación ordenadas. Ejemplos típicos incluyen MySQL-HA y etcd.
- **DaemonSet**: Un DaemonSet asegura que todos o algunos nodos ejecuten un pod. Es aplicable a los pods que se ejecutan en cada nodo. Los ejemplos típicos incluyen Ceph, Fluentd y Prometheus Node Exporter.
- **Job**: es una tarea de una sola vez que se ejecuta hasta su finalización. Se puede ejecutar inmediatamente después de ser creado. Antes de crear una carga de trabajo, puede ejecutar un trabajo para cargar una imagen en el repositorio de imágenes.
- **Cron job**: Ejecuta un trabajo periódicamente en un horario determinado. Puede realizar la sincronización de tiempo para todos los nodos activos en un punto de tiempo fijo.

Figura 9-4 Relación entre cargas de trabajo y pods

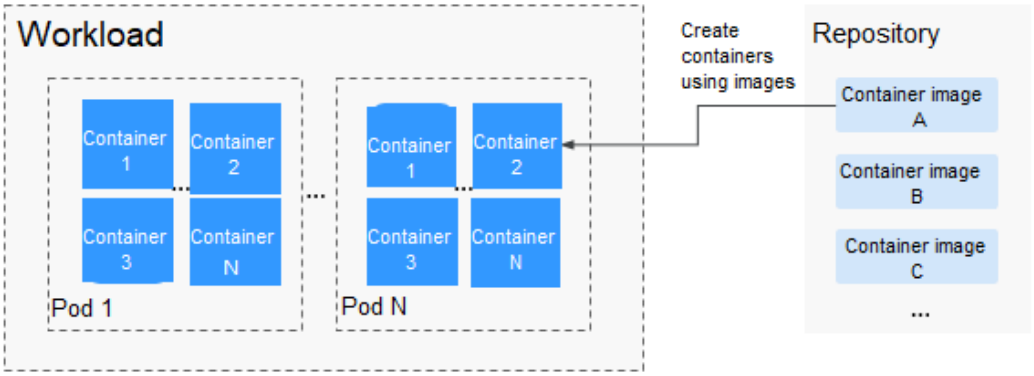


Imagen

Docker crea un estándar de la industria para aplicaciones de envasado en contenedores. Las imágenes de Docker son como plantillas que incluyen todo lo necesario para ejecutar contenedores y se utilizan para crear contenedores de Docker. En otras palabras, la imagen de Docker es un sistema de archivos especial que incluye todo lo necesario para ejecutar contenedores: programas, bibliotecas, recursos y archivos de configuración. También contiene parámetros de configuración (como volúmenes anónimos, variables de entorno y usuarios) requeridos dentro del tiempo de ejecución de un contenedor. Una imagen no contiene ningún dato dinámico. Su contenido permanece sin cambios después de ser construido. Al implementar aplicaciones en contenedores, puede usar imágenes de Docker Hub, SoftWare Repository for Container (SWR) y sus registros de imágenes privados. Por ejemplo, una imagen de Docker puede contener un sistema operativo Ubuntu completo, en el que solo se instalan los programas y dependencias necesarios.

Las imágenes se convierten en contenedores en tiempo de ejecución, es decir, los contenedores se crean a partir de imágenes. Los contenedores pueden crearse, iniciarse, detenerse, eliminarse y suspenderse.

Figura 9-5 Relación entre imágenes, contenedores y cargas de trabajo



Espacio de nombres

Un espacio de nombres es una colección abstracta de recursos y objetos. Permite que los recursos se organicen en grupos que no se superpongan. Se pueden crear varios espacios de nombres dentro de un clúster y aislarlos unos de otros. Esto permite que los espacios de nombres compartan los mismos servicios de clúster sin afectarse entre sí. Ejemplos:

- Puede implementar cargas de trabajo en un entorno de desarrollo en un espacio de nombres e implementar cargas de trabajo en un entorno de prueba en otro espacio de nombres.
- Pods, servicios, ReplicationControllers e implementaciones pertenecen a un espacio de nombres (se llama **default** de forma predeterminada), mientras que los nodos y PersistentVolumes no pertenecen a ningún espacio de nombres.

Servicio

Un servicio es un método abstracto que expone un grupo de aplicaciones que se ejecutan en un pod como servicios de red.

Kubernetes le proporciona un mecanismo de detección de servicios sin modificar las aplicaciones. En este mecanismo, Kubernetes proporciona a los pods sus propias direcciones IP y un único DNS para un grupo de pods, y equilibra la carga entre ellos.

Kubernetes le permite especificar un servicio de un tipo requerido. Los valores y acciones de los diferentes tipos de Servicios son los siguientes:

- **ClusterIP**: el servicio ClusterIP, como tipo de servicio predeterminado, se expone a través de la dirección IP interna del clúster. Si se selecciona este modo, solo se puede acceder a Servicios dentro del clúster.
- **NodePort**: Los servicios de NodePort están expuestos a través de la dirección IP y el puerto estático de cada nodo. Se crea automáticamente un servicio de ClusterIP, al que se enrutará un servicio de NodePort. Al enviar una solicitud a <NodeIP>:<NodePort>, puede acceder a un servicio de NodePort desde fuera de un clúster.
- **LoadBalancer (ELB)**: LoadBalancer (ELB) Services se exponen mediante el uso de balanceadores de carga del proveedor de nube. Los balanceadores de carga externos pueden enrutar a los servicios de NodePort y ClusterIP.
- **DNAT**: Un gateway de DNAT traduce direcciones para nodos de clúster y permite que varios nodos de clúster compartan un EIP. Los servicios de DNAT proporcionan una mayor fiabilidad que los servicios de NodePort basados en EIP, en los que el EIP está

unido a un solo nodo y una vez que el nodo está inactivo, todas las solicitudes entrantes a la carga de trabajo se distribuirán.

Equilibrio de carga de capa 7 (entrada)

Un ingreso es un conjunto de reglas de enrutamiento para las solicitudes que ingresan a un clúster. Proporciona servicios con direcciones URL, equilibrio de carga, terminación SSL y enrutamiento HTTP para el acceso externo al clúster.

Política de red

Las políticas de red proporcionan un control de red basado en políticas para aislar las aplicaciones y reducir la superficie de ataque. Una política de red utiliza selectores de etiquetas para simular redes segmentadas tradicionales y controla el tráfico entre ellas y el tráfico desde el exterior.

ConfigMap

Un ConfigMap se utiliza para almacenar datos de configuración o archivos de configuración como pares clave-valor. Las ConfigMaps son similares a los secretos, pero proporcionan un medio para trabajar con cadenas que no contienen información confidencial.

Secreto

Los secretos resuelven el problema de configuración de datos confidenciales como contraseñas, tokens y claves, y no expondrán los datos confidenciales en imágenes o especificaciones de pod. Un secreto se puede usar como un volumen o una variable de entorno.

Etiqueta

Una etiqueta es un par clave-valor y está asociado con un objeto, por ejemplo, un pod. Las etiquetas se utilizan para identificar características especiales de los objetos y son significativas para los usuarios. Sin embargo, las etiquetas no tienen un significado directo para el sistema del núcleo.

Selector de etiquetas

El selector de etiquetas es el mecanismo de agrupación principal de Kubernetes. Identifica un grupo de objetos de recurso con las mismas características o atributos a través del selector de etiquetas cliente o usuario.

Anotación

Las anotaciones se definen en pares clave-valor como lo son las etiquetas.

Las etiquetas tienen reglas de nomenclatura estrictas. Definen los metadatos de los objetos de Kubernetes y los utilizan los selectores de etiquetas.

Las anotaciones son información adicional definida por el usuario para que las herramientas externas busquen un objeto de recurso.

PersistentVolume

Un PersistentVolume (PV) es un almacenamiento de red en un clúster. Similar a un nodo, también es un recurso de clúster.

PersistentVolumeClaim

Un PV es un recurso de almacenamiento, y un PersistentVolumeClaim (PVC) es una solicitud de un PV. El PVC es similar al pod. Los pods consumen recursos de nodo y los PVC consumen recursos PV. Los pods solicitan recursos de CPU y memoria, y los PVC solicitan volúmenes de datos de un tamaño y modo de acceso específicos.

Auto Scaling - HPA

Horizontal Pod Autoscaling (HPA) es una función que implementa el ajuste horizontal de pods en Kubernetes. El mecanismo de ajuste de ReplicationController se puede utilizar para escalar sus clústeres de Kubernetes.

Afinidad y antiafinidad

Si una aplicación no está en contenedores, varios componentes de la aplicación pueden ejecutarse en la misma máquina virtual y los procesos se comunican entre sí. Sin embargo, en el caso de la contenedorización, los procesos de software se empaquetan en diferentes contenedores y cada contenedor tiene su propio ciclo de vida. Por ejemplo, el proceso de transacción se empaqueta en un contenedor mientras que el proceso de monitorización/registro y el proceso de almacenamiento local se empaquetan en otros contenedores. Si los procesos de contenedores estrechamente relacionados se ejecutan en nodos distantes, el enrutamiento entre ellos será costoso y lento.

- Afinidad: los contenedores se programan en el nodo más cercano. Por ejemplo, si la aplicación A y la aplicación B interactúan frecuentemente entre sí, es necesario usar la característica de afinidad para mantener las dos aplicaciones lo más cerca posible o incluso permitir que se ejecuten en el mismo nodo. De esta manera, no se producirá ninguna pérdida de rendimiento debido al enrutamiento lento.
- Antiafinidad: las instancias de la misma aplicación se extienden a través de diferentes nodos para lograr una mayor disponibilidad. Una vez que un nodo está inactivo, las instancias de otros nodos no se ven afectadas. Por ejemplo, si una aplicación tiene varias réplicas, es necesario utilizar la función antiafinidad para desplegar las réplicas en diferentes nodos. De esta manera, no se producirá un único punto de fallo.

Afinidad del nodo

Al seleccionar etiquetas, puede programar pods para nodos específicos.

Antiafinidad de nodos

Al seleccionar etiquetas, puede evitar que los pods se programen en nodos específicos.

Afinidad de pod

Puede implementar pods en el mismo nodo para reducir el consumo de recursos de red.

Antiafinidad de pod

Puede implementar pods en diferentes nodos para reducir el impacto de las averías del sistema. También se recomienda la implementación antiafinidad para cargas de trabajo que pueden interferir entre sí.

Cuota de recursos

Las cuotas de recursos se utilizan para limitar el uso de recursos de los usuarios.

Límite de recursos (LimitRange)

De forma predeterminada, todos los contenedores de Kubernetes no tienen límite de CPU ni de memoria. LimitRange (**limits** para abreviar) se utiliza para agregar un límite de recursos a un espacio de nombres, incluidas las cantidades mínimas, máximas y predeterminadas de recursos. Cuando se crea un pod, los recursos se asignan de acuerdo con los parámetros **limits**.

Variable de entorno

Una variable de entorno es una variable cuyo valor puede afectar a la forma en que se comportará un contenedor en ejecución. Se puede definir un máximo de 30 variables de entorno en el momento de la creación del contenedor. Puede modificar las variables de entorno incluso después de implementar las cargas de trabajo, lo que aumenta la flexibilidad en la configuración de la carga de trabajo.

La función de establecer variables de entorno en CCE es la misma que la de especificar ENV en un Dockerfile.

9.2 Cloud Native 2.0 y Huawei Cloud

Como uno de los primeros en adoptar la tecnología de contenedores, Huawei ha implementado la tecnología en múltiples productos internos desde 2013 y comenzó a usar ampliamente Kubernetes en 2014. En este curso, Huawei ha acumulado una rica experiencia práctica y proporciona servicios de contenedores completamente examinados y de pila completa para que los usuarios empresariales migren aplicaciones a la nube y tengan éxito en la era de Cloud Native 2.0.

Ahora en Huawei Cloud, puede utilizar servicios de infraestructura nativa en la nube estandarizados y fáciles de implementar para ejecutar sus aplicaciones.

Cloud Native 2.0

Desarrollo de Cloud Native

Las tecnologías nativas de la nube, como contenedores, microservicios y orquestación dinámica, están en auge y se han convertido en una fuerza impulsora importante para la innovación de servicios. Muchas empresas de industrias como las finanzas, la manufactura y la Internet han aplicado estas tecnologías a sus servicios principales. Los casos de uso en más escenarios de servicio están en movimiento, y el ecosistema de la industria se está expandiendo.

De "On Cloud" a "In Cloud"

Las nuevas aplicaciones empresariales se basan en la tecnologías nativas de la nube. Las aplicaciones, los datos y la IA se administran en la nube a lo largo de su ciclo de vida. Las aplicaciones existentes se coordinan orgánicamente con otras nuevas.

Nuevas empresas nativas en la nube

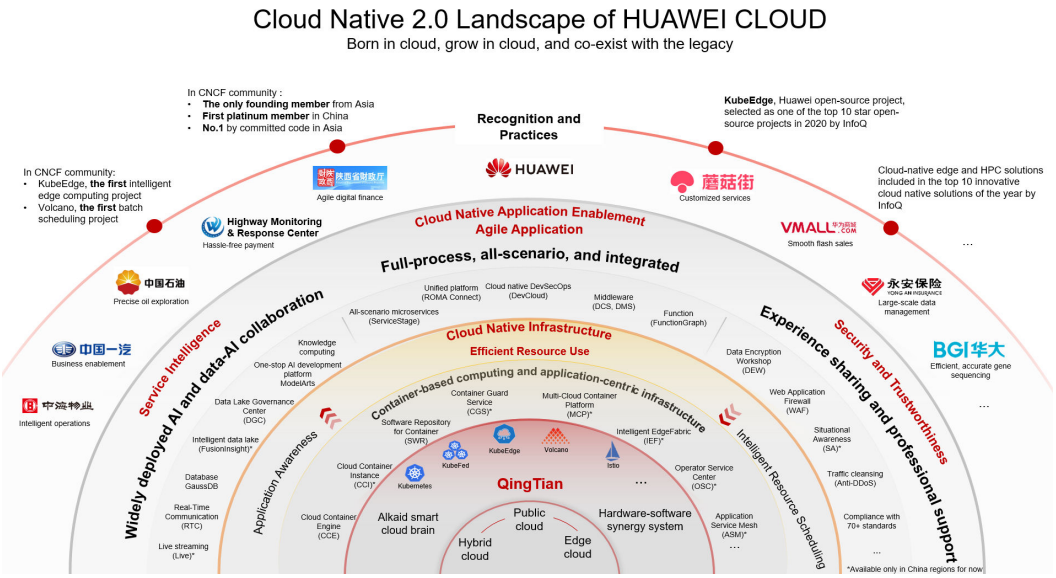
Cloud Native 2.0 es una nueva fase para la actualización inteligente de las empresas. Huawei Cloud está listo para proporcionar los recursos que necesitará para esta actualización que incluye un uso eficiente de los recursos, aplicaciones ágiles, inteligencia empresarial y servicios seguros y confiables.

Paisaje nativo de la nube de Huawei Cloud

Huawei Cloud está implementando la nube nativa en los servicios de infraestructura, lo que los hace centrados en las aplicaciones.

Estos servicios de infraestructura incluyen Cloud Container Engine (CCE), Software Repository for Container (SWR), Intelligent EdgeFabric (IEF), y Application Orchestration Service (AOS). Sobre la base de estos servicios, Huawei Cloud desarrolla cuatro soluciones nativas en la nube (cómputo de metal puro, cómputo de alto rendimiento, nube híbrida y cómputo perimetral) para construir la infraestructura de alto rendimiento, la arquitectura de servicios distribuidos y el ecosistema de aplicaciones nativas en la nube integral.

Figura 9-6 Ofertas de Huawei Cloud para Cloud Native 2.0



Infraestructura nativa de la nube

Huawei Cloud ofrece a los clientes los servicios de infraestructura nativa en la nube para ayudar a los clientes a redefinir su infraestructura, habilitar las aplicaciones ubicuas y refactorizar la arquitectura de aplicaciones. Con estos servicios, las aplicaciones pueden ejecutarse en una base compartida y colaborar entre sí a través de nubes y entre nubes y bordes para acelerar la innovación de servicios.

- **Cloud Container Engine (CCE)** le permite crear clústeres de Kubernetes de clase empresarial, altamente escalables y de alto rendimiento para ejecutar contenedores. CCE proporciona servicios de contenedores de pila completa, incluida la gestión del ciclo de

vida de clústeres y aplicaciones, servicio de mesh, gráficos Helm, complementos y programación. Con CCE, se pueden desplegar, gestionar y escalar fácilmente las aplicaciones en contenedores en Huawei Cloud.

- **Software Repository for Container (SWR)** aloja imágenes de contenedores que se pueden utilizar para implementar rápidamente aplicaciones en contenedores. Proporciona una gestión fácil, segura y confiable de las imágenes de contenedores a lo largo de su ciclo de vida.
- **Container Guard Service (CGS)** escanea vulnerabilidades y configuraciones en imágenes, ayudando a las empresas a detectar el entorno de contenedores, que no puede ser encontrado por el software de seguridad tradicional. CGS también ofrece funciones como la configuración de listas blancas de procesos, la protección de archivos de solo lectura y la detección de escape de contenedores para minimizar los riesgos de seguridad de un contenedor en ejecución.
- **Intelligent EdgeFabric (IEF)** extiende las aplicaciones en la nube a nodos perimetrales y asocia datos perimetrales y en la nube, cumpliendo con los requisitos del cliente para el control remoto, el procesamiento de datos, el análisis, la toma de decisiones y la inteligencia de los recursos informáticos perimetrales.IEF also provides unified on-cloud O&M capabilities, such as device/application monitoring and log collection, to achieve edge-cloud synergy.
- **Multi-Cloud Container Platform (MCP)** se desarrolla sobre tecnologías de contenedores de Huawei Cloud y la federación de clústeres avanzada para la comunidad. Puede gestionar de forma centralizada múltiples clústeres en las nubes y permite la implementación uniforme y la distribución del tráfico de la aplicación de múltiples clústeres. Además de la recuperación ante desastres multinube, MCP también puede separar servicios y datos, desarrollo y producción, e informática y servicios.

Habilitación de aplicaciones nativas en la nube

Huawei Cloud permite a los clientes con capacidades nativas en la nube de pila completa para admitir aplicaciones ágiles, inteligencia empresarial, servicios seguros y confiables y una evolución continua.

Aplicación ágil

- **DevCloud** es una plataforma integral de DevOps basada en la nube construida con las prácticas de Huawei de casi tres décadas en R&D, junto con sus ideas de R&D de vanguardia y herramientas de R&D de última generación. Estos servicios en la nube listos para usar le permiten gestionar proyectos, host de código, ejecutar canalizaciones, comprobar código y crear, implementar, probar y lanzar sus aplicaciones en la nube en cualquier momento y en cualquier lugar.
- **ServiceStage** es una plataforma de gestión de aplicaciones y microservicios que facilita la implementación, el monitoreo, la operación y el gobierno de las aplicaciones. ServiceStage ofrece una solución de pila completa para que las empresas desarrollen aplicaciones web, móviles y de microservicios. Esta solución ayuda a las empresas a migrar fácilmente varias aplicaciones a la nube, lo que permite a las empresas centrarse en la innovación de servicios para la transformación digital.
- **ROMA Connect** es una plataforma de integración de datos y aplicaciones de pila completa diseñada para diversos escenarios de servicio. ROMA Connect proporciona integración ligera de mensajes, datos, API, dispositivos y modelos para aplicaciones en la nube y en las instalaciones en todas las regiones para simplificar la cloudificación empresarial, ayudando a las empresas a lograr la transformación digital.

- **Distributed Message Service (DMS) for Kafka** es un servicio de espera de mensajes basado en Apache Kafka de código abierto. Proporciona instancias premium de Kafka con recursos aislados de cómputo, almacenamiento y ancho de banda. DMS for Kafka le permite aplicar recursos y configurar temas, particiones y réplicas en función de los requisitos de servicio. Se puede utilizar de inmediato y le libera de la implementación y O&M para que pueda centrarse en el desarrollo ágil de sus aplicaciones.
- **FunctionGraph** aloja y calcula funciones impulsadas por eventos. Todo lo que debe hacer es escribir su código y configurar las condiciones.

Inteligencia de servicio

- **ModelArts** es una plataforma de desarrollo de IA integral. Para el aprendizaje automático y el aprendizaje profundo, es compatible con el preprocesamiento de datos, el etiquetado de datos semiautomatizado, la formación distribuida, la creación automatizada de modelos y la implementación bajo demanda de modelos de nube de borde de dispositivo. ModelArts puede ayudar a los desarrolladores de IA a crear modelos rápidamente y a gestionar los ciclos de vida de los flujos de trabajo de IA.
- **GaussDB(for MySQL)** es un servicio de base de datos distribuida de clase empresarial de próxima generación que es totalmente compatible con MySQL. Utiliza una arquitectura de almacenamiento de cómputo desacoplada y data functions virtualization (DFV) que escala automáticamente hasta 128 TB por la instancia de base de datos. No hay que preocuparse por el sharding y prácticamente no hay riesgo de pérdida de datos. Combina la alta disponibilidad y el rendimiento de las bases de datos comerciales con la rentabilidad de las bases de datos de código abierto.

Seguridad y Confiabilidad

- **Data Security Center (DSC)** es una plataforma de protección de datos en la nube de próxima generación que protege sus activos con funciones como clasificación de riesgos, identificación de datos confidenciales, seguimiento de fuentes de marcas de agua y enmascaramiento de datos estáticos. DSC supervisa la seguridad de los datos y le brinda una visión completa de la seguridad de sus datos en la nube.
- **Host Security Service (HSS)** le ayuda a identificar y gestionar los activos de sus servidores, eliminar riesgos y defenderse de intrusiones y manipulación de páginas web. También hay funciones avanzadas de protección y operaciones de seguridad disponibles para ayudarle a detectar y manejar fácilmente las amenazas.
- **Anti-DDoS Service (ADS)** provides multiple security solutions to defend against DDoS attacks. ADS includes CNAD Basic (Anti-DDoS), CNAD Pro, and AAD.

9.3 Asignaciones entre los términos de CCE y de Kubernetes

Kubernetes (K8s) es un sistema de código abierto para automatizar la implementación, el ajuste y la gestión de clústeres de contenedores. Es una herramienta de orquestación de contenedores y una solución líder basada en la arquitectura distribuida de la tecnología de contenedores. Kubernetes se basa en la tecnología de Docker de código abierto que automatiza la implementación, la programación de recursos, el descubrimiento de servicios y el ajuste dinámico de aplicaciones en contenedores.

En este tema se describen las asignaciones entre los términos de CCE y de Kubernetes.

Tabla 9-1 Asignaciones entre los términos de CCE y de Kubernetes

CCE	Kubernetes
Clúster	Cluster
Nodo	Node
Grupo de nodos	NodePool
Contenedor	Container
Imagen	Image
Espacio de nombres	Namespace
Despliegue	Deployment
StatefulSet	StatefulSet
DaemonSet	DaemonSet
Trabajo	Job
Trabajo Cron	CronJob
Pod	Pod
Servicio	Service
ClusterIP	Cluster IP
NodePort	NodePort
LoadBalancer	LoadBalancer
Balanceo de carga de capa 7	Ingress
Política de red	NetworkPolicy
Gráfico	Template
ConfigMap	ConfigMap
Secreto	Secret
Etiqueta	Label
Selector de etiquetas	LabelSelector
Anotación	Annotation
Volumen	PersistentVolume
PersistentVolumeClaim	PersistentVolumeClaim
Ajuste automático	HPA
Afinidad del nodo	NodeAffinity
Antiafinidad de nodos	NodeAntiAffinity

CCE	Kubernetes
Afinidad de pod	PodAffinity
Antiafinidad de pod	PodAntiAffinity
Webhook	Webhook
Punto de conexión	Endpoint
Cuota	Resource Quota
Límite de recursos	Limit Range

9.4 Clúster de Turbo de CCE

Para implementar contenedores a gran escala, necesita un tipo de clúster de contenedores más potente que ofrezca un mayor rendimiento, un ajuste más rápido y una programación más eficiente.

Con los clústeres de Turbo de CCE, puede disfrutar de la informática acelerada, las redes y la programación para impulsar las innovaciones en las aplicaciones.

Ventajas del clúster

- **Cómputo acelerado**

  En una infraestructura con sinergia software-hardware, la informática consume menos recursos pero ofrece un mejor rendimiento.
- **Redes aceleradas**

  El modelo de Cloud Native Network 2.0 elimina la pérdida de rendimiento al aplanar la red y permite una comunicación más fluida entre las aplicaciones.
- **Programación acelerada**

  La programación híbrida inteligente simplifica la gestión de aplicaciones y recursos.
- **Aislamiento de seguridad**

  El motor de contenedor seguro proporciona aislamiento de seguridad a nivel de VM para aplicaciones.

Comparación entre clústeres de Turbo CCE y de CCE

En la siguiente tabla se enumeran las diferencias entre los clústeres de Turbo CCE y los clústeres de CCE:

Tabla 9-2 Tipos de clúster

Dimensión	Subdimensión	Clúster de Turbo de CCE	Clúster de CCE
Clúster	Posicionamiento	Clúster de contenedores de próxima generación para Cloud Native 2.0 con cómputo, redes y programación acelerados	Clúster estándar para el uso comercial común
	Tipo de nodo	Implementación híbrida de máquinas virtuales y servidores bare-metal	Implementación híbrida de máquinas virtuales y servidores bare-metal
Redes	Modelo	**Cloud Native Network 2.0**: se aplica a escenarios de gran escala y alto rendimiento. Escala de red: 2000 nodos	**Cloud-native network 1.0** para escenarios que no requieren un alto rendimiento o implican una implementación a gran escala. ● Modelo de red de túneles ● Modelo de red de VPC
	Rendimiento	La red de la VPC y la red de contenedores se aplanan en una sola, logrando una pérdida de rendimiento nula.	La red de VPC se superpone con la red de contenedores, causando cierta pérdida de rendimiento.
	Aislamiento de la red de contenedores	Los pods se pueden asociar directamente con grupos de seguridad para configurar políticas de aislamiento para recursos dentro y fuera de un clúster.	● Modelo de red de túnel: las políticas de aislamiento de red se admiten para la comunicación entre clústeres (mediante la configuración de políticas de red). ● Modelo de red de VPC: no se admite el aislamiento.
Seguridad	Aislamiento	● Máquina física: contenedores Kata, proporcionando aislamiento a nivel de VM. ● VM: Se implementan los contenedores comunes.	Los contenedores comunes son desplegados y aislados por Cgroups.

Cómo comprar

Obtenga información sobre cómo **comprar y configurar un clúster de Turbo CCE**.

9.5 Las regiones y las AZ

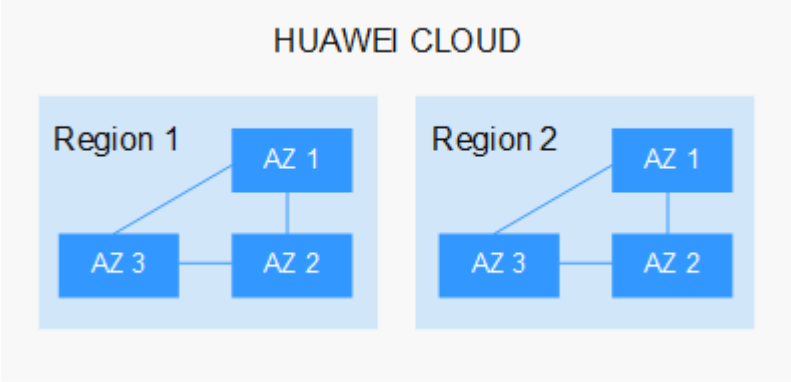
Definición

Una región y una zona de disponibilidad (AZ) identifican la ubicación de un centro de datos. Puede crear recursos en una región específica y AZ.

- Las regiones se dividen en función de la ubicación geográfica y la latencia de la red. Los servicios públicos, como Elastic Cloud Server (ECS), Elastic Volume Service (EVS), Object Storage Service (OBS), Virtual Private Cloud (VPC), Elastic IP (EIP) y Image Management Service (IMS), se comparten dentro de la misma región. Las regiones se clasifican como regiones universales y regiones dedicadas. Una región universal proporciona servicios en la nube universales para los dominios comunes. Una región dedicada proporciona servicios del mismo tipo solamente o para los dominios específicos.
- Una AZ contiene uno o más centros de datos físicos. Cada AZ cuenta con instalaciones independientes de electricidad, de refrigeración, de extinción de incendios y a prueba de humedad. Dentro de una AZ, los recursos de computación, red, almacenamiento y otros se dividen de forma lógica en múltiples clústeres. Las AZ en una región están interconectadas a través de fibras ópticas de alta velocidad. Esto es útil si va a implementar sistemas a través de AZ para lograr una mayor disponibilidad.

Figura 9-7 muestra la relación entre la región y AZ.

Figura 9-7 Las regiones y las AZ



Huawei Cloud ofrece servicios en muchas regiones de todo el mundo. Puede seleccionar una región y una AZ según sea necesario. Para obtener más información, consulte **Productos globales y servicios**.

¿Cómo seleccionar una región?

Al seleccionar una región, tenga en cuenta los siguientes factores:

- Localización

  Seleccione una región cercana a usted o a sus usuarios de destino para reducir la latencia de la red y mejorar la velocidad de acceso. Las regiones continentales de China proporcionan básicamente la misma infraestructura, calidad de red BGP, así como las operaciones y configuraciones de recursos. Si usted o sus usuarios objetivo están en el

continente de China, no es necesario tener en cuenta las diferencias de latencia de la red al seleccionar una región.

- Si usted o sus usuarios objetivo se encuentran en la región de Asia Pacífico, excepto China continental, seleccione la región **CN-Hong Kong**, **AP-Bangkok**, o **AP-Singapore**.
- Si usted o sus usuarios objetivo se encuentran en Sudáfrica, seleccione la región **AF-Johannesburg**.
- Si usted o sus usuarios objetivo están en Europa, seleccione la región **EU-Paris**.
- Si usted o sus usuarios objetivo están en América Latina, seleccione la región **LA-Santiago**.

La región **LA-Santiago** se encuentra en Chile.

- Precio del recurso

  Los precios de los recursos pueden variar en diferentes regiones. Para obtener más información, consulte **Detalles de precios del producto**.

Selección de una AZ

Al implementar recursos, tenga en cuenta los requisitos de las aplicaciones en cuanto a la recuperación ante desastres (DR) y la latencia de la red.

- Para una alta capacidad de DR, implemente recursos en diferentes AZ dentro de la misma región.
- Para una menor latencia de red, implemente recursos en la misma AZ.

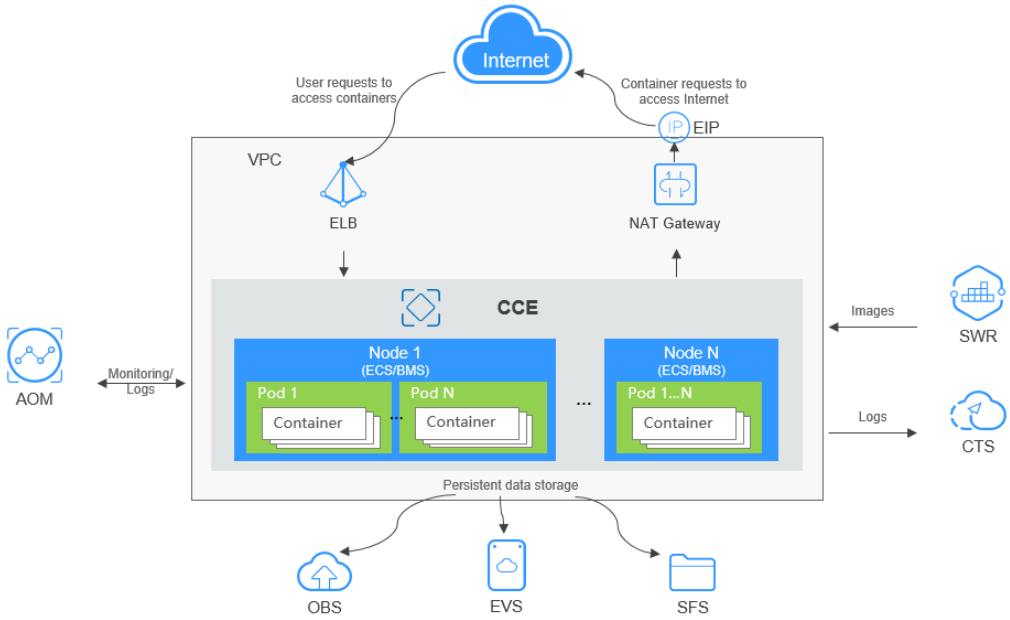
Regiones y puntos de conexión

Al utilizar una API para acceder a recursos, debe especificar una región y un punto de conexión. Para obtener más información, consulte **Regiones y puntos de conexión**.

10 Servicios relacionados

CCE trabaja con los siguientes servicios en la nube y requiere los permisos para acceder a ellos.

Figura 10-1 Relaciones entre CCE y otros servicios



Relaciones entre CCE y otros servicios

Tabla 10-1 Relaciones entre CCE y otros servicios

Servicio	Relación	Características Relacionadas
Elastic Cloud Server (ECS)	Un ECS con varios discos de EVS es un nodo en CCE. Puede elegir las especificaciones de ECS durante la creación de nodos.	● **Compra de un nodo** ● **Aceptación de nodos existentes en un clúster**

Servicio	Relación	Características Relacionadas
Virtual Private Cloud (VPC)	Por razones de seguridad, todos los clústeres creados por CCE deben ejecutarse en **las VPC**. Al crear un espacio de nombres, debe crear una VPC o vincular una VPC existente al espacio de nombres para que todos los contenedores del espacio de nombres se ejecuten en esta VPC.	**Compra de un clúster de CCE**
Elastic Load Balance (ELB)	CCE trabaja con ELB para equilibrar la carga de las solicitudes de acceso de una carga de trabajo a través de múltiples pods. Cuando se utiliza **ELB**, la dirección del balanceador de carga, en lugar de la dirección de la carga de trabajo, se expone a los usuarios. Las solicitudes de usuario llegan primero a ELB a través de una red pública y luego son enrutadas por ELB a diferentes pods de la carga de trabajo.	● **Creación de un Deployment** ● **Creación de un StatefulSet** ● **LoadBalancer**
NAT Gateway	El servicio NAT Gateway proporciona la traducción de direcciones de red de origen (SNAT) para instancias de contenedor en una VPC. La función SNAT traduce las direcciones IP privadas de estas instancias de contenedor a la misma EIP, que es una dirección IP pública accesible en Internet. Puede definir las reglas de SNAT en el **gateway de NAT** para permitir que los contenedores accedan a Internet.	● **Creación de un Deployment** ● **Creación de un StatefulSet** ● **DNAT**
Software Repository for Container (SWR)	Se utiliza un repositorio de imágenes para almacenar y gestionar las imágenes de Docker. Puede crear cargas de trabajo a partir de imágenes en **SWR**.	● **Creación de un Deployment** ● **Creación de un StatefulSet**

Servicio	Relación	Características Relacionadas
Elastic Volume Service (EVS)	Los discos de EVS se pueden conectar a servidores en la nube y escalar a una mayor capacidad cuando sea necesario. Un ECS con varios discos de EVS es un nodo en CCE. Puede elegir las especificaciones de ECS durante la creación de nodos.	**Uso de los volúmenes de EVS**
Object Storage Service (OBS)	OBS proporciona un almacenamiento en la nube para datos de cualquier tamaño que es estable, seguro, rentable y se basa en objetos. Con OBS, puede crear, modificar y eliminar bucket, así como cargar, descargar y eliminar objetos. CCE permite crear un volumen OBS y adjuntarlo a una ruta dentro de un contenedor.	**Uso de volúmenes de OBS**
Scalable File Service (SFS)	SFS es un servicio de almacenamiento de archivos compartido y totalmente administrado. Compatible con el protocolo del sistema de archivos de red, los sistemas de archivos SFS pueden escalar elásticamente hasta petabytes, lo que garantiza el máximo rendimiento de las aplicaciones con un uso intensivo de datos y ancho de banda. Puede utilizar sistemas de archivos de SFS como almacenamiento persistente para contenedores y adjuntar los sistemas de archivos a contenedores al crear una carga de trabajo.	**Uso de volúmenes de SFS**
Application Operations Management (AOM)	AOM recopila los archivos de log de contenedores en formatos como .log de CCE y los volca a AOM. En la consola de AOM, puede consultar y ver fácilmente los archivos de log. Además, AOM supervisa el uso de recursos de CCE. Puede definir umbrales de métricas para que el uso de recursos de CCE active el ajuste automático.	**Recopilación de logs de salida estándar de contenedores**

Servicio	Relación	Características Relacionadas
Cloud Trace Service (CTS)	CTS registra las operaciones en sus recursos en la nube, lo que le permite consultar, auditar y realizar un seguimiento de las solicitudes de operación de recursos iniciadas desde la consola de gestión o las API abiertas, así como las respuestas a estas solicitudes.	**Operaciones de CCE soportadas por CTS**