Documento técnico de AWS

Información general de las instancias de spot de Amazon EC2

Información general de las instancias de spot de Amazon EC2: Documento técnico de AWS

Table of Contents

Información general de las instancias de spot de Amazon EC2

Fecha de publicación: 5 de marzo de 2021 (Historial de revisión y colaboradores)

Resumen

En este documento se busca capacitarlo para maximizar el valor de sus inversiones, mejorar la precisión de las previsiones y la previsibilidad de los costes, crear una cultura de propiedad y transparencia de costes y medir continuamente su estado de optimización.

En este documento se suministra información general sobre las instancias de Spot de Amazon EC2 así como también sobre las prácticas recomendadas para usarlas de manera eficiente.

Introducción

Además de las instancias bajo demanda, instancias reservadas y Savings Plans, el cuarto modelo de precios de Amazon Elastic Compute Cloud (Amazon EC2) son las instancias de spot.

Con las instancias de spot puede utilizar la capacidad sobrante de computación Amazon EC2 con descuentos de hasta el 90 % en comparación con los precios bajo demanda. Esto significa que puede reducir significativamente el coste de ejecutar las aplicaciones o aumentar la capacidad de computación y el rendimiento de las aplicaciones con el mismo presupuesto. La única diferencia entre las instancias bajo demanda y las instancias de spot es que EC2 puede interrumpir las últimas con dos minutos de anticipación cuando EC2 necesite recuperar la capacidad.

A diferencia de las instancias reservadas o los Savings Plans, las instancias de spot no requieren un compromiso para lograr ahorros de costes en comparación con los precios bajo demanda. Sin embargo, dado que EC2 puede terminar las instancias de spot si no hay capacidad disponible en el grupo de capacidad (una combinación de un tipo de instancia y una zona de disponibilidad) en la que se ejecutan, son las más adecuadas para cargas de trabajo flexibles.

Cuándo usar instancias de spot

Puede utilizar las instancias de spot para distintas aplicaciones flexibles y tolerantes a errores. Los ejemplos incluyen servidores web sin estado, puntos de conexión de API, aplicaciones de análisis y macrodatos, cargas de trabajo en contenedores, computación de alto rendimiento y alto rendimiento de CI/CD (HPC/HTC), cargas de trabajo de renderización y otras cargas de trabajo flexibles.

Las instancias de spot no son adecuadas para cargas de trabajo que no sean flexibles, con estado, sin tolerancia a errores o estrechamente acopladas entre nodos de instancia. Las instancias de spot tampoco se recomiendan para cargas de trabajo que no sean tolerantes a períodos ocasionales en los que la capacidad de destino no esté completamente disponible. Advertimos encarecidamente contra el uso de instancias de spot para estas cargas de trabajo o para intentar conmutar por error a las instancias bajo demanda para gestionar las interrupciones.

Cómo lanzar instancias de spot

El servicio más recomendado para lanzar instancias de spot es Amazon EC2 Auto Scaling, ya que le permite lanzar y mantener la capacidad deseada y solicitar recursos automáticamente para reemplazar los que se interrumpan o se terminen manualmente. Al configurar un grupo de Auto Scaling, solo se necesita especificar los tipos de instancia y la capacidad deseada en función de las necesidades de la aplicación. Para obtener más información, consulte Grupos de Auto Scaling en la Guía del usuario de Amazon EC2 Auto Scaling.

Si necesita más flexibilidad, tener creados sus propios flujos de trabajo de lanzamiento de instancias o desea controlar aspectos individuales de los lanzamientos de instancias o los mecanismos de escalado, le recomendamos que evalúe el uso de la flota de EC2 en modo instantáneo como alternativa a EC2 Auto Scaling. Esta API sincrónica le permite especificar una lista de tipos de instancias y requisitos de lanzamiento, y proporciona una capacidad más flexible que la llamada a la API RunInstances de EC2 para lanzar instancias de spot o instancias bajo demanda.

Cuando utiliza los servicios de AWS para ejecutar sus cargas de trabajo en la nube, también puede usarlos para lanzar instancias de spot. Entre los ejemplos se incluyen Amazon EMR, Amazon EKS, Amazon ECS, AWS Batch y AWS Elastic Beanstalk. También puede lanzar instancias de spot utilizando herramientas de terceros que se integran con la nube de AWS.

Puede automatizar los lanzamientos de instancias de spot utilizando herramientas de infraestructura como código (AWS CloudFormation, AWS CDK) o la API, la CLI o los SDK de AWS. Spot Blueprints proporciona un asistente guiado que le permite generar plantillas de infraestructura como código para AWS Cloudformation y Hashicorp terraform que se adhieren a las prácticas recomendadas de spot.

Cómo funcionan las instancias de spot

Las instancias de spot se ejecutan exactamente igual que otras instancias de EC2 mientras están en funcionamiento. Sin embargo, Amazon EC2 puede interrumpirlas cuando EC2 necesite recuperar la capacidad.

Cuando EC2 interrumpe la instancia de spot, termina, detiene o hiberna la instancia, según el comportamiento de interrupción que elija.

Si EC2 interrumpe la instancia de spot durante la primera hora, antes de que transcurra una hora completa de tiempo de ejecución, no se le cobrará por la hora parcial utilizada. Sin embargo, si detiene o cancela su instancia de spot, pagará por cualquier hora parcial utilizada (como sucede en las instancias bajo demanda o reservadas). Para obtener información sobre cómo se facturan las instancias de spot interrumpidas que se ejecuten en diferentes sistemas operativos, consulte Facturación de instancias de spot interrumpidas en l a Guía del usuario de EC2.

El precio de spot de cada tipo de instancia en cada zona de disponibilidad viene determinado por las tendencias a largo plazo en la oferta y la demanda de capacidad sobrante de EC2. El usuario paga el precio de spot que está en vigor, facturado hasta el segundo más cercano.

Puede opcionalmente especificar un precio máximo para las instancias de spot. Si no se especifica un precio máximo, el precio máximo es por defecto el precio bajo demanda. Tenga en cuenta que nunca se paga más del precio de spot que está en vigor cuando la instancia de spot está en ejecución. Le recomendamos que no especifique un precio máximo, sino que deje que el precio máximo sea el precio bajo demanda. Un precio máximo elevado no aumenta las posibilidades de lanzar una instancia de spot ni reduce las posibilidades de que se interrumpa la instancia de spot (ya que EC2 aún puede interrumpir su instancia de spot cuando necesite recuperar la capacidad).

El precio de spot de un tipo de instancia en una zona de disponibilidad puede cambiar en cualquier momento, pero en general, no cambia con frecuencia. AWS publica el precio de spot actual y los precios históricos de las instancias de spot a través de la API DescribeSpotPriceHistory, así como en la consola de administración de AWS, que refleja los datos de la API. Esto puede ayudarle a evaluar los niveles y el calendario de las fluctuaciones de los precios de spot a lo largo del tiempo.

Administración de interrupciones de instancias de spot

La mejor manera de gestionar correctamente las interrupciones de instancias de spot y minimizar el impacto en el rendimiento o la disponibilidad es diseñar su aplicación para que sea tolerante a fallos. Para lograrlo, puede aprovechar las recomendaciones de reequilibrio de instancia de EC2 y los avisos de interrupción de instancia de spot.

Una recomendación de reequilibrio de instancia de EC2 es una señal que le notifica en caso de que una instancia de spot tenga un riesgo de interrupción elevado. La señal brinda la oportunidad de administrar de forma proactiva la instancia de spot antes del aviso de interrupción de dos minutos de la instancia de spot. Puede decidir reequilibrar su carga de trabajo con instancias de spot nuevas o existentes que no tengan un riesgo elevado de interrupción. Hemos facilitado el uso de esta señal proporcionando la característica de reequilibrio de capacidad en los grupos de EC2 Auto Scaling. Para obtener más información, consulte Reequilibrio de la capacidad de Amazon EC2 Auto Scaling.

Un aviso de interrupción de instancia de spot es una advertencia que se emite dos minutos antes de que Amazon EC2 interrumpa una instancia de spot. Si su carga de trabajo es “flexible en cuanto al tiempo”, puede configurar sus instancias de spot para que se detengan o hibernen, en lugar de terminarlas, cuando se interrumpan. Amazon EC2 detiene o hiberna automáticamente las instancias de spot en caso de interrupción y las reanuda automáticamente cuando se tiene capacidad disponible.

Puede utilizar la recomendación de reequilibrio de instancias de EC2 o el aviso de interrupción de instancias de spot para diseñar su carga de trabajo teniendo en cuenta la tolerancia a errores, de modo que pueda capturar notificaciones y guardar el estado de un trabajo en el almacenamiento (por ejemplo, Amazon S3, Amazon EFS o Amazon FSx), conservar los archivos de registro de la instancia (o transmitirlos continuamente para un enfoque más tolerante a fallos), drenar las conexiones de un equilibrador de carga, etc.

Algunos servicios de AWS y de terceros ya gestionan las interrupciones de spot para que disminuya el impacto en su aplicación. Por ejemplo, Amazon EKS que ejecuta grupos de nodos administrados con instancias de spot lanza automáticamente nodos de Kubernetes de sustitución cuando se entregan recomendaciones de reequilibrio o avisos de interrupción para un nodo existente.

Límites de instancias de Spot

Existe un límite en el número de instancias de spot en ejecución y solicitadas por cuenta de AWS por región. Los límites de las instancias de spot se administran en función del número de unidades de procesamiento centrales virtuales (vCPU) que las instancias de spot en ejecución utilizan o utilizarán en espera del cumplimiento de las solicitudes de instancias de spot abiertas. Si termina sus instancias de spot pero no cancela las solicitudes de instancias de spot, las solicitudes se descontarán de su límite de vCPU de instancias spot hasta que Amazon EC2 detecte las terminaciones de las instancias de spot y cierre las solicitudes.

Hay seis límites de instancias de spot:

- Todas las solicitudes de instancias de spot estándar (A, C, D, H, I, M, R, T, Z)
- Todas las solicitudes de instancias de spot F
- Todas las solicitudes de instancias de spot G
- Todas las solicitudes de instancias de spot Inf
- Todas las solicitudes de instancias de spot P
- Todas las solicitudes de instancias de spot X

Cada límite especifica el límite de vCPU virtual de una o más familias de instancias. Para obtener información acerca de las diferentes familias, generaciones y tamaños de instancias, consulte Tipos de instancia de Amazon EC2.

Con los límites de vCPU, puede usar sus límites en términos del número de vCPU necesarias para lanzar cualquier combinación de tipo de instancias que cumplan las necesidades cambiantes de su aplicación. Por ejemplo, supongamos que el límite de todas las solicitudes de instancias de spot estándar es de 256 vCPU, puede solicitar 32 instancias de spot `m5.2xlarge` (32 x 8 vCPU) o 16 instancias de spot `c5.4xlarge` (16 x 16 vCPU), o una combinación de cualquier tipo y tamaño de instancia de spot estándar con un total de 256 vCPU.

Para obtener más información, consulte Monitorear los límites y el uso de instancias de spot puntuales y Solicitar un aumento del límite de instancias puntuales en la Guía del usuario de instancias de Linux de Amazon EC2.

Prácticas recomendadas para instancias de spot

El uso de las siguientes prácticas recomendadas en su aplicación dependerá del diseño de esta y de sus requisitos de tipo de instancia y de presupuesto.

- Sea flexible con respecto a los tipos de instancia. Un grupo de instancias de spot es un conjunto de instancias EC2 no utilizadas con el mismo tipo de instancia (por ejemplo, m 5.large) y zona de disponibilidad (por ejemplo: us-east-1a). Debe ser flexible en cuanto a los tipos de instancia que solicita y las zonas de disponibilidad en las que puede implementar la carga de trabajo. Esto le da a las instancias de spot una mejor oportunidad de encontrar y asignar la cantidad necesaria de capacidad de cómputo. Por ejemplo, no pida solo c5.large si estaría dispuesto a usar larges de las familias c5, c4 y m5.
- Utilice la estrategia de asignación optimizada de la capacidad. Las estrategias de asignación en grupos de EC2 Auto Scaling ayudan a aprovisionar la capacidad de destino sin necesidad de buscar manualmente los grupos de instancias de spot con capacidad sobrante. Recomendamos utilizar la estrategia de capacidad optimizada, ya que aprovisiona automáticamente a instancias desde los grupos de instancias de spot que se encuentran con mayor disponibilidad. Dado que la capacidad de las instancias de spot proviene de grupos con una capacidad óptima, esto reduce la posibilidad de que sus instancias de spot se interrumpan. Para obtener más información acerca de las estrategias de asignación, consulte Instancias de spot en la Guía del usuario de Amazon EC2 Auto Scaling.
- Usar el reequilibrio de capacidad proactivo. El reequilibrio de capacidad lo ayuda a mantener la disponibilidad de la carga de trabajo mediante el aumento proactivo de su grupo de Auto Scaling con una nueva instancia de spot antes de que una instancia de spot en ejecución reciba el aviso de interrupción de dos minutos. Cuando se habilita el reequilibrio de capacidad, el Auto Scaling intenta reemplazar de forma proactiva las instancias de spot que han recibido una recomendación de reequilibrio, lo que brinda la oportunidad de reequilibrar la carga de trabajo en nuevas instancias spot que no tienen un riesgo elevado de interrupción.
- Utilice los servicios de AWS integrados para administrar sus instancias de spot. Otros servicios de AWS se integran con instancias de spot para reducir los costes informáticos generales sin necesidad de administrar las instancias o flotas individuales. Le recomendamos que considere las siguientes soluciones para sus cargas de trabajo aplicables: Amazon EMR, Amazon ECS, AWS Batch, Amazon EKS, SageMaker, AWS Elastic Beanstalk y Amazon GameLift. Para obtener más información sobre las prácticas recomendadas de Spot con estos servicios, consulte el sitio web de Amazon EC2 Spot Instances Workshops.

- Elija la herramienta de lanzamiento moderna y correcta para instancias de spot. Si uno de los servicios integrados de AWS no se ajusta bien a su carga de trabajo y aún necesita crear su aplicación con control sobre el lanzamiento de instancias de spot, utilice la herramienta adecuada. Para la mayoría de las cargas de trabajo, debe utilizar EC2 Auto Scaling porque proporciona un conjunto de características más completo para una amplia variedad de cargas de trabajo, como aplicaciones respaldadas por ELB, cargas de trabajo en contenedores y trabajos de procesamiento de colas. Si necesita más control sobre las solicitudes individuales y busca una herramienta de “solo lanzamiento”, utilice la flota de EC2 en modo instantáneo como sustituto directo de RunInstances, pero con un conjunto más amplio de capacidades, como estrategias de asignación y diversificación de tipos de instancias.

Integración de spot con otros servicios de AWS

Las instancias de spot de Amazon EC2 se integran con varios servicios de AWS.

Integración de Amazon EMR

Puede ejecutar clústeres de Amazon EMR en instancias de spot y reducir significativamente el coste de procesar grandes cantidades de datos para sus cargas de trabajo de análisis. Puede ejecutar sus clústeres de EMR mezclando fácilmente instancias de spot con instancias bajo demanda y reservadas mediante la característica Flotas de instancias de EMR. Puede utilizar estrategias de asignación de EMR para lanzar instancias de spot desde los grupos de capacidad más disponibles.

Integración de EC2 Auto Scaling

Puede utilizar los grupos de Amazon EC2 Auto Scaling para lanzar y administrar instancias de spot, mantener la disponibilidad de las aplicaciones, diversificar el tipo de instancia y la selección de opciones de compra (bajo demanda/spot) y escalar su capacidad de Amazon EC2 mediante políticas de escalado dinámicas, programadas y predictivas. Para obtener más información, consulte Solicitud de instancias de spot para aplicaciones flexibles y tolerantes a fallos en la Guía del usuario de Amazon EC2 Auto Scaling.

Integración de Amazon EKS

Puede optimizar los costes de sus cargas de trabajo basadas en Kubernetes con Amazon EKS, lanzando instancias de spot en grupos de nodos administrados de EKS. Los grupos de nodos administrados de EKS administran todo el ciclo de vida de las instancias de spot, sustituyendo las instancias de spot que pronto se interrumpirán por instancias lanzadas recientemente, para reducir las posibilidades de impacto en el rendimiento o la disponibilidad de las aplicaciones cuando se interrumpen las instancias de spot (cuando EC2 necesita recuperar la capacidad). Para obtener más información, consulte Grupos de nodos administrados en la Guía del usuario de Amazon EKS.

Integración de Amazon ECS

Puede ejecutar clústeres de Amazon ECS en instancias de spot para reducir el coste operativo de ejecutar aplicaciones en contenedores. Amazon ECS admite el vaciado automático de instancias de

spot que pronto se interrumpirán. Para obtener más información, consulte Using Spot Instances en la Amazon Elastic Container Service Developer Guide.

Integración de Amazon ECS con AWS Fargate Spot

Si sus tareas en contenedores son interrumpibles y flexibles, puede optar por ejecutar sus tareas de ECS con el proveedor de capacidad de instancias de spot de AWS Fargate, lo que significa que sus tareas se ejecutarán en AWS Fargate, una plataforma de contenedores sin servidor, y se beneficiará del ahorro de costes impulsado por Fargate Spot. Para obtener más información, consulte AWS Fargate capacity providers en la Amazon Elastic Container Service Developer Guide.

Integración de Amazon Batch

AWS Batch planifica, programa y ejecuta cargas de trabajo de computación en lotes en AWS. AWS Batch solicita de manera dinámica instancias de spot en su nombre, lo que reduce aún más el coste de la ejecución de sus trabajos por lotes.

Integración de Amazon SageMaker

Amazon SageMaker facilita la capacitación de modelos de aprendizaje automático mediante instancias de spot administradas. La formación de spot administrado puede optimizar el coste de los modelos de formación hasta en un 90 % en comparación con las instancias bajo demanda. SageMaker administra las interrupciones de Spot en su nombre. Para obtener más información, consulteManaged Spot Training en Amazon SageMaker en la Guía para desarrolladores de Amazon SageMaker.

Integración de Amazon Gamelift

Amazon GameLift es una solución de alojamiento de servidores para juegos que implementa, opera y escala servidores en la nube para juegos multijugador. La compatibilidad con instancias de spot en Amazon Gamelift le brinda la oportunidad de reducir significativamente sus costes de alojamiento. Al crear flotas de recursos de alojamiento, puede elegir entre instancias bajo demanda o instancias de spot. Si bien las instancias de spot pueden interrumpirse con dos minutos de notificación, FleetIQ de Amazon GameLift minimiza la posibilidad de interrupciones. Para obtener más información, consulte Uso de instancias de spot con GameLift en la Guía para desarrolladores de Amazon GameLift.

Integración de AWS Elastic Beanstalk

AWS Elastic Beanstalk es un servicio fácil de utilizar para implementar y escalar servicios y aplicaciones web desarrollados con Java, .NET, PHP, Node.js, Python, Ruby, Go y Docker en servidores familiares, como Apache, Nginx, Passenger e IIS. Puede cargar simplemente su código y Elastic Beanstalk se encarga automáticamente de la implementación, desde el aprovisionamiento de capacidad y balanceador de carga hasta el escalado automático y el monitoreo del estado de las aplicaciones. Puede usar instancias de spot en sus entornos de Elastic Beanstalk para optimizar los costes de la infraestructura subyacente de sus aplicaciones web. Para obtener información sobre el uso de instancias de spot con Elastic Beanstalk, consulte Compatibilidad con instancias de spot en la Guía para desarrolladores de AWS Elastic Beanstalk.

Conclusión

Tanto si tiene necesidades de computación como si desea aumentar la capacidad sin aumentar su presupuesto, las instancias de spot pueden ser una excelente manera de optimizar sus costes de AWS y/o crear teniendo en cuenta la escalabilidad. Al diseñar adecuadamente sus cargas de trabajo, puede aprovechar las instancias de spot para una amplia gama de necesidades. Para obtener más información, consulte Amazon EC2 Spot Instances.

Recursos

- Centro de arquitectura de AWS
- Documentos técnicos de AWS
- Arquitectura mensual de AWS
- Blog de arquitectura de AWS
- Vídeos de This Is My Architecture
- Documentación de AWS

Historial de revisión y colaboradores

Historial de revisión

Para recibir notificaciones sobre las actualizaciones de este documento técnico, suscríbase a la fuente RSS.

update-history-change	update-history-description	update-history-date
Actualización menor	Diseño de página ajustado.	30 de abril de 2021
Actualización menor	Se ha actualizado el contenido para reflejar las prácticas recomendadas actuales. El nombre del documento técnico cambió de "Uso de las instancias de spot de EC2 a escala" a "Información general de instancias de spot de Amazon EC2" para que refleje mejor el contenido.	5 de marzo de 2021
Actualización menor	Se han actualizado los límites de instancias de spot.	3 de febrero de 2021
Publicación inicial	Uso de instancias de Spot de Amazon EC2 a escala publicado	1 de marzo de 2018

> **Note**
>
> Para suscribirse a las actualizaciones de RSS, debe disponer de un complemento de RSS habilitado para el navegador que utilice.

Colaboradores

En este documento han participado las siguientes personas y organizaciones:

- Amilcar Alfaro, Sr. Director de marketing de productos, AWS
- Erin Carlson, directora de marketing de AWS
- Keith Jarrett, director de BD de WW - Optimización de costes, desarrollo empresarial de AWS
- Ran Sheinberg, arquitecto principal de soluciones, AWS