Whitepaper da AWS

Visão geral das instâncias spot do Amazon EC2

Visão geral das instâncias spot do Amazon EC2: Whitepaper da AWS

Table of Contents

Visão geral das instâncias spot do Amazon EC2

Data de publicação: 5 de março de 2021 (Histórico do documento e colaboradores)

Resumo

Este documento visa capacitar você para maximizar o valor de seus investimentos, melhorar a exatidão das previsões e a previsibilidade de custos, criar uma cultura de senso de propriedade e transparência de custos e medir continuamente seu status de otimização.

Este documento oferece uma visão geral das instâncias spot do Amazon EC2 e das práticas recomendadas para usá-las com eficácia.

Introdução

Além de sob demanda, instâncias reservadas e Savings Plans, o quarto modelo de preço do Amazon Elastic Compute Cloud (Amazon EC2) é o de instâncias spot.

Com instâncias spot, você pode usar a capacidade de computação extra do Amazon EC2 com descontos de até 90% em comparação com o preço sob demanda. Isso significa que você pode reduzir significativamente o custo de execução das aplicações ou aumentar a capacidade computacional e a taxa de transferência da aplicação para o mesmo orçamento. A única diferença entre as instâncias sob demanda e as instâncias spot é que as spot poderão ser interrompidas pelo EC2 com notificação prévia de dois minutos quando o EC2 precisar da capacidade.

Diferentemente das instâncias reservadas ou dos Savings Plans, as instâncias spot não exigem um compromisso para obter economia em relação ao preço sob demanda. No entanto, como as instâncias spot podem ser terminadas pelo EC2 se não houver capacidade disponível no grupo de capacidade (uma combinação de um tipo de instância e uma zona de disponibilidade) em que estão sendo executadas, elas são mais adequadas para workloads flexíveis.

Quando usar instâncias spot

Você pode usar instâncias spot para várias aplicações tolerantes a falhas e flexíveis. Os exemplos incluem servidores Web sem estado, endpoints de API, aplicações de big data e análise, workloads em contêineres, computação de alta performance e alta taxa de transferência de CI/CD (HPC/HTC), workloads de renderização e outras workloads flexíveis.

As instâncias spot não são adequadas para workloads que são inflexíveis, com estado, intolerantes a falhas ou fortemente acopladas entre nós de instância. As instâncias spot também não são recomendadas para workloads intolerantes a períodos ocasionais quando a capacidade de destino não está completamente disponível. Não recomendamos o uso de instâncias spot para essas workloads ou a tentativa de failover para instâncias sob demanda a fim de lidar com interrupções.

Como executar instâncias spot

O serviço mais recomendado para a execução de instâncias spot é o Amazon EC2 Auto Scaling, pois ele permite executar e manter a capacidade desejada e solicitar automaticamente recursos para substituir qualquer instância que seja interrompida ou terminada manualmente. Ao configurar um grupo do Auto Scaling, você só precisa especificar os tipos de instância e a capacidade desejada com base nas necessidades da aplicação. Para obter mais informações, consulte Grupos do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Se você precisar de mais flexibilidade, tiver criado seus próprios fluxos de trabalho de execução de instâncias ou quiser controlar aspectos individuais das execuções de instâncias ou dos mecanismos de escalabilidade, recomendamos que avalie o uso da frota do EC2 no modo instantâneo como uma alternativa ao EC2 Auto Scaling. Essa API síncrona permite especificar uma lista de tipos de instância e requisitos de execução, além de fornecer recursos mais flexíveis do que a chamada de API RunInstances do EC2 para executar instâncias spot ou instâncias sob demanda.

Ao usar os serviços da AWS para executar suas workloads na nuvem, você também pode usá-los para executar instâncias spot. Os exemplos incluem Amazon EMR, Amazon EKS, Amazon ECS, AWS Batch e AWS Elastic Beanstalk. Você também pode executar instâncias spot usando ferramentas de terceiros que se integram à Nuvem AWS.

Você pode automatizar as execuções de instâncias spot usando as ferramentas de infraestrutura como código (AWS CloudFormation, AWS CDK), a API, a CLI ou os SDKs da AWS. O Spot Blueprints oferece um assistente guiado que permite gerar modelos de infraestrutura como código para o AWS CloudFormation e o Hashicorp Terraform que aderem às práticas recomendadas de instâncias spot.

Como as instâncias spot funcionam

As instâncias spot se comportam exatamente como outras instâncias do EC2 durante a execução. No entanto, elas podem ser interrompidas pelo Amazon EC2 quando este precisar da capacidade de volta.

Quando o EC2 interrompe a instância spot, ele termina, interrompe ou hiberna a instância, dependendo do comportamento de interrupção que você escolher.

Se o EC2 interromper a instância spot na primeira hora, antes de uma hora inteira de tempo de execução, você não será cobrado pelo tempo parcial usado. No entanto, se você interromper ou encerrar a instância spot, pagará por qualquer tempo parcial usado (assim como paga para instâncias reservadas ou sob demanda). Para obter informações sobre como você é cobrado por instâncias spot interrompidas em execução em diferentes sistemas operacionais, consulte Faturamento de instâncias spot interrompidas no Guia do usuário do EC2.

O preço spot para cada tipo de instância em cada zona de disponibilidade é determinado pelas tendências de longo prazo na oferta e na demanda de capacidade extra do EC2. Você paga o preço spot que está em vigor, cobrado até o segundo mais próximo.

Se preferir, você poderá especificar um preço máximo para as instâncias spot. Se você não especificar um preço máximo, o padrão será o preço sob demanda. Observe que você nunca paga mais do que o preço spot que está em vigor quando a instância spot está em execução. Recomendamos que você não especifique um preço máximo, mas deixe o preço sob demanda como o preço máximo padrão. Um preço máximo alto não aumenta suas chances de executar uma instância spot e não diminui suas chances de que a instância spot seja interrompida (porque o EC2 ainda pode interromper a instância spot quando precisar da capacidade de volta).

O preço spot para um tipo de instância em uma zona de disponibilidade pode ser alterado a qualquer momento, mas, em geral, não muda com frequência. A AWS publica o preço spot atual e os preços históricos das instâncias spot por meio da API DescribeSpotPriceHistory, bem como no Console de Gerenciamento da AWS, que reflete os dados da API. Isso pode ajudar a avaliar os níveis e as flutuações no preço spot ao longo do tempo.

Gerenciar interrupções de instâncias spot

A melhor maneira de lidar com interrupções de instâncias spot e minimizar o impacto na performance ou disponibilidade é arquitetar a aplicação para ser tolerante a falhas. Para fazer isso, você pode aproveitar as recomendações de rebalanceamento de instâncias do EC2 e avisos de interrupção de instâncias spot.

Uma recomendação de rebalanceamento de instâncias do EC2 é um sinal que notifica você quando uma instância spot está em alto risco de interrupção. O sinal oferece a oportunidade de gerenciar proativamente a instância spot antes do aviso de interrupção de dois minutos. Você pode decidir rebalancear sua workload para instâncias spot novas ou existentes que não tenham alto risco de interrupção. Facilitamos o uso desse sinal fornecendo o recurso de rebalanceamento de capacidade em grupos do EC2 Auto Scaling. Para obter mais informações, consulte Rebalanceamento de capacidade do Amazon EC2 Auto Scaling.

Um aviso de interrupção de instância spot é um aviso emitido dois minutos antes de o Amazon EC2 interromper uma instância spot. Se a workload tiver “flexibilidade de tempo”, você poderá configurar as instâncias spot para serem interrompidas ou hibernadas, em vez de serem terminadas, quando forem interrompidas. O Amazon EC2 interrompe ou hiberna automaticamente as instâncias spot em caso de interrupção e as retoma automaticamente quando tivermos capacidade disponível.

Você pode usar a recomendação de rebalanceamento de instância do EC2 e/ou o aviso de interrupção de instância spot para arquitetar a workload tendo em mente a tolerância a falhas. Dessa forma, você poderá capturar notificações e salvar o estado de um trabalho no armazenamento (por exemplo, Amazon S3, Amazon EFS ou Amazon FSx), manter arquivos de log da instância (ou transmiti-las continuamente para uma abordagem mais tolerante a falhas), drenar conexões de um balanceador de carga etc.

Alguns serviços da AWS e de terceiros já lidam com interrupções spot para que você diminua o impacto na aplicação. Por exemplo, o Amazon EKS executando grupos de nós gerenciados com instâncias spot inicia automaticamente nós substitutos do Kubernetes quando uma recomendação de rebalanceamento ou avisos de interrupção são entregues a um nó existente.

Limites de instâncias spot

Há um limite para o número de instâncias spot em execução e solicitadas por conta da AWS por região. Os limites da instância spot são gerenciados em termos do número de unidades de processamento central virtual (vCPUs) que as instâncias spot em execução estão usando ou usarão enquanto aguardam o atendimento de solicitações de instância spot abertas. Se você terminar as instâncias spot, mas não cancelar as solicitações de instâncias spot, as solicitações serão contabilizadas em relação ao limite de vCPUs da instância spot até que o Amazon EC2 detecte as terminações da instância spot e feche as solicitações.

Existem seis limites da instâncias spot:

- Todas as solicitações de instância spot padrão (A, C, D, H, I, M, R, T, Z)
- Todas as solicitações de instâncias spot F
- Todas as solicitações de instâncias spot G
- Todas as solicitações de instâncias spot Inf
- Todas as solicitações de instâncias spot P
- Todas as solicitações de instâncias spot X

Cada limite especifica o limite de vCPU para uma ou mais famílias de instâncias. Para obter informações sobre as diferentes famílias, gerações e tamanhos de instâncias, consulte Tipos de instância do Amazon EC2.

Com limites de vCPU, é possível usar seu limite em termos do número de vCPUs necessárias para executar qualquer combinação de tipos de instância que atenda às necessidades em constante mudança da sua aplicação. Por exemplo, digamos que o limite de todas as solicitações de instância spot padrão seja 256 vCPUs, você poderia solicitar 32 instâncias spot `m5.2xlarge` (32 x 8 vCPUs) ou 16 instâncias spot `c5.4xlarge` (16 x 16 vCPUs) ou uma combinação de qualquer tipo e tamanho de instância spot padrão que totalize 256 vCPUs.

Para obter mais informações, consulte Monitorar limites e uso de instâncias spot e Solicitar um aumento de limite de instância spot no Guia do usuário do Amazon EC2 para instâncias Linux.

Melhores práticas para instâncias spot

Os seus requisitos de tipo de instância e orçamento, e o projeto da aplicação determinarão como aplicar as práticas recomendadas a seguir para a aplicação:

- Seja flexível sobre os tipos de instância. Um grupo de instâncias spot é um conjunto de instâncias do EC2 não utilizadas com o mesmo tipo de instância (por exemplo, m5.large) e zona de disponibilidade (por exemplo, us-east-1a). Você deve ser flexível sobre quais tipos de instância solicita e em quais zonas de disponibilidade pode implantar a carga de trabalho. Isso dá ao Spot uma chance melhor de encontrar e alocar a quantidade necessária de capacidade computacional. Por exemplo, não peça apenas c5.large se você está disposto a usar grandes das famílias c4, m5 e m4.
- Use a estratégia de alocação otimizada por capacidade. As estratégias de alocação nos grupos do EC2 Auto Scaling ajudam a provisionar a capacidade prevista sem a necessidade de procurar manualmente os grupos de instância spot com capacidade adicional. Recomendamos o uso da estratégia de otimização por capacidade, pois ela provisiona automaticamente as instâncias dos grupos de instância spot mais disponíveis. Como a capacidade da instância spot é originada de grupos com capacidade ideal, isso diminui a possibilidade de que as instâncias spot sejam interrompidas. Para obter mais informações sobre estratégias de alocação, consulte Instâncias spot no Guia do usuário do Amazon EC2 Auto Scaling.
- Use o rebalanceamento proativo de capacidade. O rebalanceamento de capacidade ajuda a manter a disponibilidade da workload aumentando proativamente o grupo do Auto Scaling com uma nova instância spot antes que uma instância spot em execução receba o aviso de interrupção de dois minutos. Quando o rebalanceamento de capacidade está habilitado, o Auto Scaling tenta substituir proativamente as instâncias spot que receberam uma recomendação de rebalanceamento, oferecendo a oportunidade de rebalancear a workload para novas instâncias spot que não apresentam alto risco de interrupção.
- Use serviços integrados da AWS para gerenciar as instâncias spot. Outros serviços da AWS integram-se ao Spot para reduzir os custos gerais de computação sem a necessidade de gerenciar instâncias ou frotas individuais. Recomendamos que você considere as seguintes soluções para as workloads aplicáveis: Amazon EMR, Amazon ECS, AWS Batch, Amazon EKS, SageMaker, AWS Elastic Beanstalk e Amazon GameLift. Para saber mais sobre as práticas recomendadas de spot com esses serviços, consulte o Amazon EC2 Spot Instances Workshops Website (site de workshops sobre instâncias spot do Amazon EC2).

- Escolha a ferramenta de execução moderna e correta para instâncias spot. Se um dos serviços integrados da AWS não for adequado para sua workload e você ainda precisar construir a aplicação com controle sobre a execução de instâncias spot, use a ferramenta certa. Para a maioria das workloads, você deve usar o EC2 Auto Scaling, pois ele fornece um conjunto de recursos mais abrangente a uma ampla variedade de workloads, como aplicações com base no ELB, workloads em contêineres e trabalhos de processamento de filas. Se você precisar de mais controle sobre solicitações individuais e estiver procurando uma ferramenta “somente iniciar”, use a frota do EC2 no modo instantâneo como um substituto imediato para RunInstances, mas com um conjunto mais amplo de recursos, como diversificação de tipos de instância e estratégias de alocação.

Integração de spot a outros serviços da AWS

As instâncias spot do Amazon EC2 se integram a vários serviços da AWS.

Integração ao Amazon EMR

Você pode executar clusters do Amazon EMR em instâncias spot e reduzir significativamente o custo de processamento de grandes quantidades de dados para as workloads de análise. Você pode executar os clusters do EMR misturando facilmente instâncias spot com instâncias reservadas e sob demanda usando o recurso frotas de instâncias do EMR. Você pode usar estratégias de alocação do EMR para executar instâncias spot por meio dos grupos de capacidade mais disponíveis.

Integração ao EC2 Auto Scaling

Você pode usar grupos do Amazon EC2 Auto Scaling para executar e gerenciar instâncias spot, manter a disponibilidade da aplicação, diversificar o tipo de instância e a seleção de opções de compra (sob demanda/spot) e escalar sua capacidade do Amazon EC2 usando políticas de escalabilidade dinâmicas, programadas e preditivas. Para obter mais informações, consulte Solicitar instâncias spot para aplicações flexíveis e tolerantes a falhas no Guia do usuário do Amazon EC2 Auto Scaling.

Integração ao Amazon EKS

Você pode otimizar os custos das workloads baseadas no Kubernetes usando o Amazon EKS, executando instâncias spot em grupos de nós gerenciados do EKS. Os grupos de nós gerenciados do EKS gerenciam todo o ciclo de vida da instância spot, substituindo aquelas que serão interrompidas em breve por instâncias recém-executadas, visando reduzir as chances de impacto na performance ou na disponibilidade da aplicação quando as instâncias spot forem interrompidas (quando o EC2 precisar da capacidade de volta). Para saber mais, consulte Grupos de nós gerenciados no Guia do usuário do Amazon EKS.

Integração ao Amazon ECS

Você pode executar clusters do Amazon ECS em instâncias spot para reduzir o custo operacional da execução de aplicações em contêineres. O Amazon ECS é compatível com a drenagem automática

de instâncias spot que serão interrompidas em breve. Para obter mais informações, consulte Usar instâncias spot no Guia do desenvolvedor do Amazon Elastic Container Service.

Integração do Amazon ECS ao AWS Fargate Spot

Se suas tarefas em contêineres forem flexíveis e puderem ser interrompidas, será possível optar por executar tarefas do ECS com o provedor de capacidade spot do AWS Fargate, o que significa que as tarefas serão executadas no AWS Fargate, uma plataforma de contêineres sem servidor, e você se beneficiará da economia de custos impulsionada pelo Fargate Spot. Para obter mais informações, consulte Provedores de capacidade do AWS Fargate no Guia do desenvolvedor do Amazon Elastic Container Service.

Integração ao Amazon Batch

O AWS Batch planeja, programa e executa workloads de computação em lote dos clientes na AWS. O AWS Batch solicita dinamicamente instâncias spot em seu nome, reduzindo ainda mais o custo de execução dos trabalhos em lotes.

Integração ao Amazon SageMaker

O Amazon SageMaker facilita o treinamento de modelos de machine learning usando instâncias spot gerenciadas. O treinamento de spots gerenciadas pode otimizar o custo dos modelos de treinamento em até 90% em relação às instâncias sob demanda. O SageMaker gerencia as interrupções de spot em seu nome. Para obter mais informações, consulte Treinamento de spots gerenciadas no Amazon SageMaker no Guia do desenvolvedor do Amazon SageMaker.

Integração ao Amazon Gamelift

O Amazon GameLift é uma solução de hospedagem de servidores de jogos que implanta, opera e escala servidores de nuvem para jogos multijogador. A compatibilidade com instâncias spot no Amazon Gamelift oferece a oportunidade de reduzir significativamente os custos de hospedagem. Ao criar frotas de recursos de hospedagem, você pode escolher entre instâncias sob demanda ou instâncias spot. Embora as instâncias spot possam ser interrompidas com dois minutos de notificação, o FleetIQ do Amazon GameLift minimiza a chance de interrupções. Para obter mais informações, consulte Usar instâncias spot com o GameLift no Guia do desenvolvedor do Amazon GameLift.

Integração ao AWS Elastic Beanstalk

O AWS Elastic Beanstalk é um serviço de fácil utilização para implantação e escalabilidade de serviços e aplicações Web desenvolvidos com Java, .NET, PHP, Node.js, Python, Ruby, Go e Docker em servidores familiares como Apache, Nginx, Passenger e IIS. Basta carregar seu código e o Elastic Beanstalk se encarrega automaticamente da implantação, desde o provisionamento de capacidade, o balanceamento de carga e a autoescalabilidade até o monitoramento da integridade da aplicação. Você pode usar instâncias spot em seus ambientes do Elastic Beanstalk para otimizar os custos da infraestrutura subjacente das aplicações Web. Para obter informações sobre como usar instâncias spot com o Elastic Beanstalk, consulte Compatibilidade com instâncias spot no Guia do desenvolvedor do AWS Elastic Beanstalk.

Conclusão

Se você tem necessidades de computação flexíveis ou deseja aumentar a capacidade sem aumentar seu orçamento, as instâncias spot podem ser uma ótima maneira de otimizar os custos da AWS e/ou construir levando em consideração a escalabilidade. Ao arquitetar adequadamente suas workloads, você pode aproveitar as instâncias spot para uma ampla variedade de necessidades. Para obter mais informações, consulte Instâncias spot do Amazon EC2.

Recursos

- Central de arquitetura da AWS
- Whitepapers da AWS
- Arquitetura mensal da AWS
- Blog de arquitetura da AWS
- Vídeos “This is my Architecture”
- Documentação da AWS

Histórico do documento e colaboradores

Histórico do documento

Para ser notificado sobre atualizações deste whitepaper, inscreva-se no RSS feed.

update-history-change	update-history-description	update-history-date
Atualização secundária	Layout de página ajustado.	30 de abril de 2021
Atualização secundária	Conteúdo atualizado para refletir as práticas recomendadas atuais. O nome do whitepaper mudou de “Como usar instâncias spot do Amazon EC2 em grande escala” para “Visão geral das instâncias spot do Amazon EC2” a fim de refletir melhor o conteúdo.	5 de março de 2021
Atualização secundária	Os limites da instância spot foram atualizados.	3 de fevereiro de 2021
Publicação inicial	Publicação de “Como usar instâncias spot do Amazon EC2 em grande escala”.	1 de março de 2018

> **Note**
>
> Para assinar atualizações RSS, você deve ter um plugin RSS habilitado para o navegador que está usando.

Contribuidores

As seguintes organizações e pessoas contribuíram para este documento:

- Amilcar Alfaro, gerente sênior de marketing de produtos da AWS
- Erin Carlson, gerente de marketing da AWS
- Keith Jarrett, líder de BD WW, Otimização de custos e desenvolvimento de negócios da AWS
- Ran Sheinberg, arquiteto-chefe de soluções da AWS