Livre blanc AWS

Présentation des instances Spot Amazon EC2

Présentation des instances Spot Amazon EC2: Livre blanc AWS

Table of Contents

Présentation des instances Spot Amazon EC2

Date de publication : 5 mars 2021 (Historique du document et contributeurs)

Résumé

Ce document vise à vous donner les moyens d'optimiser la valeur de vos investissements, d'améliorer la précision des prévisions et la prévisibilité des coûts, de créer une culture de propriété et de transparence des coûts, et de mesurer en permanence votre statut d'optimisation.

Ce document présente les instances Spot Amazon EC2 et les bonnes pratiques à adopter pour les utiliser efficacement.

Introduction

Outre les instances à la demande, les instances réservées et les Savings Plans, le quatrième modèle de tarification Amazon Elastic Compute Cloud (Amazon EC2) est celui des instances Spot.

Avec les instances Spot, vous pouvez utiliser la capacité de calcul disponible d'Amazon EC2 avec des remises allant jusqu'à 90 % par rapport à la tarification à la demande. Cela signifie que vous pouvez réduire considérablement le coût de fonctionnement de vos applications ou augmenter la capacité de calcul de votre application pour le même budget. La seule différence entre les instances à la demande et les instances Spot est que ces dernières peuvent être interrompues par EC2 après un préavis de deux minutes lorsqu'EC2 a besoin de récupérer la capacité.

Contrairement aux instances réservées ou aux Savings Plans, les instances Spot ne nécessitent pas d'engagement afin de réaliser des économies par rapport à la tarification à la demande. Toutefois, étant donné que les instances Spot peuvent être arrêtées par EC2 s'il n'y a pas de capacité disponible dans le groupe de capacité (combinaison d'un type d'instance et d'une zone de disponibilité) dans lequel elles s'exécutent, elles conviennent mieux aux charges de travail flexibles.

Quand utiliser les instances Spot

Vous pouvez utiliser les instances Spot pour une variété d'applications flexibles et tolérantes aux pannes. Les exemples incluent les serveurs web sans état, les points de terminaison d'API, les applications Big Data et d'analytique, les charges de travail conteneurisées, le calcul CI/CD haute performance et haut débit (HPC/HTC), les charges de travail de rendu et d'autres charges de travail flexibles.

Les instances Spot ne conviennent pas aux charges de travail inflexibles, avec état, intolérantes aux pannes ou étroitement couplées entre des nœuds d'instance. Les instances Spot ne sont pas recommandées non plus pour les charges de travail qui ne tolèrent pas les périodes occasionnelles où la capacité cible n'est pas complètement disponible. Nous vous déconseillons vivement d'utiliser des instances Spot pour ces charges de travail ou de tenter de basculer vers des instances à la demande pour gérer les interruptions.

Comment lancer des instances Spot

Le service le plus recommandé pour le lancement d'instances Spot est Amazon EC2 Auto Scaling, car il vous permet de lancer et de maintenir la capacité souhaitée, et de demander automatiquement des ressources pour remplacer celles qui sont perturbées ou interrompues manuellement. Lorsque vous configurez un groupe Auto Scaling, il vous suffit de spécifier les types d'instances et la capacité souhaitée en fonction des besoins de votre application. Pour plus d'informations, consultez Groupes Auto Scaling dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.

Si vous avez besoin de plus de flexibilité, si vous avez créé vos propres flux de lancement d'instances ou si vous souhaitez contrôler les aspects individuels des lancements d'instances ou les mécanismes de mise à l'échelle, nous vous recommandons d'évaluer l'utilisation de la flotte EC2 en mode instantané comme alternative à EC2 Auto Scaling. Cette API synchrone vous permet de spécifier une liste de types d'instances et de conditions de lancement, et offre des fonctionnalités plus flexibles que l'appel d'API EC2 RunInstances pour le lancement d'instances Spot ou d'instances à la demande.

Lorsque vous utilisez les services AWS pour exécuter vos charges de travail dans le cloud, vous pouvez également les utiliser pour lancer des instances Spot. Les exemples incluent Amazon EMR, Amazon EKS, Amazon ECS, AWS Batch et AWS Elastic Beanstalk. Vous pouvez également lancer des instances Spot à l'aide d'outils tiers qui s'intègrent au Cloud AWS.

Vous pouvez automatiser les lancements d'instances Spot en utilisant l'infrastructure en tant qu'outils de code (AWS CloudFormation, AWS CDK) ou l'API, l'interface de ligne de commande ou les kits SDK AWS. Les plans Spot fournissent un assistant guidé qui vous permet de générer une infrastructure en tant que modèles de code pour AWS Cloudformation et Hashicorp Terraform, conformément aux bonnes pratiques en matière d'instances Spot.

Fonctionnement des instances Spot

Les instances Spot ont des performances d'exécution parfaitement identiques à celles des autres instances Amazon EC2. Toutefois, elles peuvent être interrompues par Amazon EC2 lorsqu'EC2 a besoin de récupérer cette capacité.

Lorsque EC2 interrompt votre instance Spot, il résilie, arrête ou met l'instance en veille prolongée, selon le comportement d'interruption que vous choisissez.

Si EC2 interrompt votre instance Spot au cours de la première heure, avant une heure complète d'exécution, vous n'êtes pas facturé pour l'heure partielle utilisée. Toutefois, si vous arrêtez ou résiliez votre instance Spot, vous payez pour toute heure partielle utilisée (comme vous le faites pour les instances réservées ou à la demande). Pour plus d'informations sur la façon dont vous êtes facturé pour les instances Spot interrompues exécutées sur différents systèmes d'exploitation, consultez Facturation des instances Spot interrompues dans le Guide de l'utilisateur EC2.

Le prix Spot de chaque type d'instance dans chaque zone de disponibilité est déterminé par les tendances à long terme de l'offre et de la demande de capacité de réserve d'EC2. Vous payez le prix Spot en vigueur, facturé à la seconde près.

Le cas échéant, vous pouvez spécifier un prix maximum pour les instances Spot. Si vous ne spécifiez pas de prix maximum, la valeur par défaut est le prix à la demande. Notez que vous ne payez jamais plus que le prix Spot en vigueur lorsque votre instance Spot est en cours d'exécution. Nous vous recommandons de ne pas spécifier de prix maximum, mais plutôt d'utiliser le prix à la demande comme prix maximum par défaut. Un prix maximum élevé n'augmente pas vos chances de lancer une instance Spot et ne diminue pas vos chances de voir votre instance Spot interrompue (car EC2 peut toujours interrompre votre instance Spot lorsqu'il a besoin de récupérer sa capacité).

Le prix Spot d'un type d'instance dans une zone de disponibilité peut changer à tout moment, mais en général, il ne change pas fréquemment. AWS publie le prix Spot actuel et les prix historiques des instances Spot via l'API DescribeSpotPriceHistory, ainsi que dans AWS Management Console, ce qui reflète les données de l'API. Vous pouvez ainsi évaluer le rythme et l'ampleur des variations du prix Spot au fil du temps.

Gérer les interruptions d'instances Spot

La meilleure façon de gérer convenablement les interruptions d'instances Spot et de limiter l'impact sur vos performances ou votre disponibilité consiste à concevoir votre application de manière à ce qu'elle soit tolérante aux pannes. Pour ce faire, vous pouvez tirer parti des recommandations de rééquilibrage d'instance EC2 et des avis d'interruption des instances Spot.

La recommandation de rééquilibrage d'instance EC2 est une nouvelle fonction qui vous avertit lorsqu'une instance Spot présente un risque élevé d'interruption. Le signal vous donne la possibilité de gérer de manière proactive l'instance Spot avant l'avis d'interruption de deux minutes. Vous pouvez décider de rééquilibrer votre charge de travail en instance Spot nouvelle ou existante qui ne présente pas de risque élevé d'interruption. Nous avons facilité l'utilisation de ce signal en utilisant la fonction Rééquilibrage de capacité dans les groupes EC2 Auto Scaling. Pour plus d'informations, consultez Rééquilibrage de capacité Amazon EC2 Auto Scaling.

Un avis d'interruption d'instance Spot est un avertissement émis deux minutes avant qu'Amazon EC2 n'interrompe une instance Spot. Si votre charge de travail est « flexible dans le temps », vous pouvez configurer vos instances Spot pour qu'elles soient arrêtées ou mises en veille prolongée, au lieu d'être arrêtées, lorsqu'elles sont interrompues. Amazon EC2 arrête ou met automatiquement en veille prolongée vos instances Spot en cas d'interruption, et les reprend automatiquement lorsque la capacité est disponible.

Vous pouvez utiliser la recommandation de rééquilibrage de l'instance EC2 et/ou l'avis d'interruption d'instance Spot pour structurer votre charge de travail en tenant compte de la tolérance aux pannes, afin de pouvoir capturer des notifications et d'enregistrer l'état d'un travail sur le stockage (par exemple, Amazon S3, Amazon EFS ou Amazon FSx), conserver les fichiers journaux de l'instance (ou les diffuser en continu pour une approche plus tolérante aux pannes), drainer les connexions à partir d'un équilibreur de charge, etc.

Certains services AWS et tiers gèrent déjà les interruptions d'instances Spot afin de réduire l'impact sur votre application. Par exemple, Amazon EKS exécutant des groupes de nœuds gérés avec des instances Spot lance automatiquement des nœuds Kubernetes de remplacement lorsqu'une recommandation de rééquilibrage ou des avis d'interruption sont envoyés pour un nœud existant.

Limites associées aux instances Spot

Le nombre d'instances Spot en cours d'exécution et demandées par compte AWS et par région est limité. Les limites associées aux instances Spot sont gérées en termes de nombre d'unités de traitement centralisées virtuelles (vCPU) que vos instances Spot en cours d'exécution utilisent ou utiliseront en attendant le traitement des demandes d'instance Spot ouvertes. Si vous résiliez vos instances Spot, mais que vous n'annulez pas les demandes d'instances Spot, les demandes sont comptabilisées dans votre limite de vCPU d'instances Spot jusqu'à ce qu'Amazon EC2 détecte les résiliations d'instance Spot et ferme les demandes.

Il existe six limites associées aux instances Spot :

- Toutes les demandes d'instances Spot standard (A, C, D, H, I, M, R, T, Z)
- Toutes les demandes d'instances Spot F
- Toutes les demandes d'instances Spot G
- Toutes les demandes d'instances Spot Inf
- Toutes les demandes d'instances Spot P
- Toutes les demandes d'instances Spot X

Chaque limite spécifie la limite de vCPU pour une ou plusieurs familles d'instances. Pour de plus amples informations sur les différentes familles, générations et tailles d'instances, veuillez consulter Types d'instances Amazon EC2.

Avec les limites de vCPU, vous pouvez utiliser votre limite en fonction du nombre de vCPU nécessaires pour lancer toute combinaison de types d'instance qui répond à l'évolution de vos besoins en termes d'applications. Par exemple, supposons que votre limite de demandes d'instances Spot standard soit de 256 vCPU, vous pouvez demander 32 `m5.2xlarge`instances Spot (32 x 8 vCPU) ou 16 `c5.4xlarge`instances Spot (16 x 16 vCPU), ou une combinaison de tous types et tailles d'instance Spot standard totalisant 256 vCPU.

Pour de plus amples informations, veuillez consultez Surveiller les limites et l'utilisation des instances Spot et Demander une augmentation de la limite des instances Spot dans le Guide de l'utilisateur Amazon EC2 pour les instances Linux.

Bonnes pratiques relatives aux instances Spot

Vos exigences en matière de type d'instance et de budget, ainsi que la manière dont votre application est conçue, détermineront de quelle manière appliquer les bonnes pratiques pour votre application.

- Faites preuve de souplesse dans votre choix de type d'instance. Un groupe d'instances Spot est un ensemble d'instances EC2 inutilisées avec le même type d'instance (par exemple, m5.large) et la même zone de disponibilité (par exemple, us-east-1a). Vous devez être flexible quant aux types d'instance que vous demandez et aux zones de disponibilité dans lesquelles vous pouvez déployer votre charge de travail. Cela donne à Spot une meilleure chance de trouver et d'allouer la quantité requise de capacité de calcul. Par exemple, ne demandez pas simplement c5.large si vous seriez prêt à utiliser des instances larges des familles c4, m5 et m4.
- Utilisez la stratégie d'allocation optimisée pour la capacité. Les stratégies d'allocation dans les groupes EC2 Auto Scaling vous aident à provisionner votre capacité cible sans avoir à rechercher manuellement des groupes d'instances Spot avec une capacité de réserve. Nous vous recommandons d'utiliser la stratégie optimisée par la capacité, car elle alloue automatiquement les instances des groupes d'instances Spot les plus disponibles. Étant donné que la capacité de votre instance Spot provient de groupes ayant une capacité optimale, cela réduit la possibilité que vos instances Spot soient interrompues. Pour plus d'informations sur les stratégies d'allocation, consultez Instances spot dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.
- Utilisez un rééquilibrage de capacité proactif. Le rééquilibrage de capacité vous permet de maintenir la disponibilité de la charge de travail en augmentant de manière proactive votre groupe Auto Scaling avec une nouvelle instance Spot avant qu'une instance Spot en cours ne reçoive l'avis d'interruption d'instance Spot deux minutes avant l'arrêt. Lorsque le rééquilibrage de capacité est activé, Auto Scaling tente de remplacer de manière proactive les instances Spot qui ont reçu une recommandation de rééquilibrage, ce qui permet de rééquilibrer votre charge de travail vers de nouvelles instances Spot qui ne présentent pas un risque élevé d'interruption.
- Utilisez les services AWS intégrés pour gérer vos instances Spot. D'autres services AWS s'intègrent à Spot pour réduire les coûts de calcul globaux sans avoir à gérer les instances ou les flottes individuels. Nous vous recommandons d'envisager les solutions suivantes pour vos charges de travail applicables : Amazon EMR, Amazon ECS, AWS Batch, Amazon EKS, SageMaker, AWS Elastic Beanstalk et Amazon GameLift. Pour en savoir plus sur les bonnes pratiques relatives à ces services, consultez le site web des ateliers sur les instances Spot Amazon EC2.
- Choisissez l'outil de lancement moderne et approprié pour les instances Spot. Si l'un des services intégrés AWS n'est pas adapté à votre charge de travail et que vous devez toujours créer

votre application en contrôlant le lancement des instances Spot, utilisez le bon outil. Pour la plupart des charges de travail, vous devez utiliser EC2 Auto Scaling, car il fournit un ensemble de fonctionnalités plus complet pour une grande variété de charges de travail, telles que les applications basées sur ELB, les charges de travail conteneurisées et les tâches de traitement de file d'attente. Si vous avez besoin de plus de contrôle sur les demandes individuelles et que vous recherchez un outil de « lancement uniquement », utilisez EC2 Fleet en mode instantané en remplacement de RunInstances, mais avec un ensemble plus large de fonctionnalités, telles que la diversification des types d'instances et les stratégies d'allocation.

Intégration des instances Spot à d'autres services AWS

Les instances Spot Amazon EC2 s'intègrent à plusieurs services AWS.

Intégration à Amazon EMR

Vous pouvez exécuter des clusters Amazon EMR sur des instances Spot et réduire de manière significative le coût de traitement de grandes quantités de données pour vos charges de travail analytiques. Vous pouvez exécuter vos clusters EMR en mélangeant facilement des instances Spot avec des instances réservées et à la demande à l'aide de la fonctionnalité Flottes d'instances EMR. Vous pouvez utiliser des stratégies d'allocation EMR pour lancer des instances Spot à partir des groupes de capacité les plus disponibles.

Intégration à EC2 Auto Scaling

Vous pouvez utiliser des groupes Amazon EC2 Auto Scaling pour lancer et gérer des instances Spot, maintenir la disponibilité des applications, diversifier le type d'instance et la sélection d'options d'achat (à la demande/Spot), et dimensionner votre capacité Amazon EC2 à l'aide de politiques de mise à l'échelle dynamiques, planifiées et prédictives. Pour plus d'informations, consultez Demander des instances Spot pour des applications flexibles et tolérantes aux pannes dans le Guide de l'utilisateur Amazon EC2 Auto Scaling.

Intégration à Amazon EKS

Vous pouvez optimiser vos charges de travail basées sur Kubernetes à l'aide d'Amazon EKS, en lançant des instances Spot dans des groupes de nœuds gérés par EKS. Les groupes de nœuds gérés par EKS gèrent l'ensemble du cycle de vie des instances Spot, en remplaçant les instances Spot qui seront bientôt interrompues par des instances nouvellement lancées, afin de réduire les risques d'impact sur les performances ou la disponibilité de vos applications lorsque les instances Spot sont interrompues (lorsqu'EC2 a besoin de récupérer la capacité). Pour en savoir plus, consultez Groupes de nœuds gérés dans le Guide de l'utilisateur Amazon EKS.

Intégration à Amazon ECS

Vous pouvez exécuter des clusters Amazon ECS sur des instances Spots afin de réduire le coût opérationnel lié à l'exécution d'applications conteneurisées. Amazon ECS prend en charge le

drainage automatique des instances Spot qui seront bientôt interrompues. Pour de plus amples informations, veuillez consulter Utilisation d'instances Spot dans le Guide du développeur Amazon Elastic Container Service.

Intégration d'Amazon ECS à AWS Fargate Spot

Si vos tâches conteneurisées sont interruptibles et flexibles, vous pouvez choisir d'exécuter vos tâches ECS avec le fournisseur de capacité Spot AWS Fargate, ce qui signifie que vos tâches seront exécutées sur AWS Fargate, une plateforme de conteneurs sans serveur, et vous bénéficierez des économies liées à Fargate Spot. Pour de plus amples informations, veuillez consulter Fournisseurs de capacité AWS Fargate dans le Guide du développeur Amazon Elastic Container Service.

Intégration à Amazon Batch

AWS Batch planifie, programme et exécute les charges de travail de calcul par lot des clients sur AWS. AWS Batch formule également des demandes d'instances Spot à votre place de manière dynamique, ce qui permet de réduire les coûts d'exécution de vos tâches par lot.

Intégration à Amazon SageMaker

Amazon SageMaker facilite l'entraînement de modèles de machine learning à l'aide d'instances Spot gérées. L'entraînement Spot géré peut optimiser le coût des modèles d'entraînement jusqu'à 90 % par rapport aux instances à la demande. SageMaker gère les interruptions Spot en votre nom. Pour de plus amples informations, veuillez consulter Entraînement Spot géré dans Amazon SageMaker dans le Guide du développeur Amazon SageMaker.

Intégration à Amazon Gamelift

Amazon GameLift est une solution d'hébergement de serveurs de jeu qui permet de déployer, d'exploiter et de mettre à l'échelle des serveurs cloud destinés à des jeux multijoueur. La prise en charge des instances Spot dans Amazon Gamelift vous permet de réduire considérablement vos coûts d'hébergement. Lors de la création de flottes de ressources d'hébergement, vous pouvez choisir entre des instances à la demande ou des instances Spot. Alors que les instances Spot peuvent être interrompues avec deux minutes de notification, FleetIQ d'Amazon GameLift limite les risques d'interruption. Pour de plus amples informations, veuillez consulter Utilisation des instances Spot avec GameLift dans le Guide du développeur Amazon GameLift.

Intégration à Elastic Beanstalk

AWS Elastic Beanstalk est un service simple à utiliser pour déployer et dimensionner des applications et services web développés avec Java, .NET, PHP, Node.js, Python, Ruby, Go et Docker sur des serveurs connus, tels qu'Apache, Nginx, Passenger et IIS. Il vous suffit de charger votre code pour qu'Elastic Beanstalk gère automatiquement les étapes du déploiement, de l'allocation des capacités à l'équilibrage de charge, en passant par la scalabilité automatique et la surveillance de l'état de l'application. Vous pouvez utiliser des instances Spot dans vos environnements Elastic Beanstalk pour optimiser les coûts de l'infrastructure sous-jacente de vos applications web. Pour de plus amples informations sur l'utilisation des instances Spot avec Elastic Beanstalk, veuillez consulter Prise en charge des instances Spot dans le Guide du développeur AWS Elastic Beanstalk.

Conclusion

Que vous ayez des besoins en calcul flexibles ou que vous souhaitiez augmenter votre capacité sans accroître votre budget, les instances Spot peuvent être un excellent moyen d'optimiser vos coûts AWS et/ou de créer en tenant compte de l'évolutivité. En concevant correctement l'architecture de vos charges de travail, vous pouvez tirer parti des instances Spot pour répondre à un large éventail de besoins. Pour plus d'informations, consultez Amazon EC2 Spot instances.

Ressources

- Centre d'architecture AWS
- Livres blancs AWS
- AWS Architecture Monthly
- Blog sur l'architecture AWS
- Vidéos This Is My Architecture
- Documentation AWS

Historique du document et contributeurs

Historique du document

Pour être informé des mises à jour de ce livre blanc, abonnez-vous au flux RSS.

update-history-change	update-history-description	update-history-date
Mise à jour mineure	Mise en page ajustée.	30 avril 2021
Mise à jour mineure	Contenu mis à jour pour refléter les bonnes pratiques actuelles. Le titre du livre blanc est passé de « Exploiter les instances Spot Amazon EC2 à l'échelle » à « Présentation des instances Spot Amazon EC2 » afin de mieux refléter le contenu.	5 mars 2021
Mise à jour mineure	La section Limites associées aux instances Spot a été mise à jour.	3 février 2021
Publication initiale	Livre blanc Exploiter les instances Spot Amazon EC2 à grande échelle publié.	1er mars 2018

> **Note**
>
> Pour vous abonner aux mises à jour RSS, vous devez activer un plug-in RSS pour le navigateur que vous utilisez.

Collaborateurs

Les personnes et organisations suivantes ont participé à la préparation du présent document :

- Amilcar Alfaro, Sr. Responsable du marketing produit, AWS
- Erin Carlson, Responsable marketing, AWS
- Keith Jarrett, Responsable mondial du développement commercial, optimisation des coûts, développement commercial AWS
- Ran Sheinberg, Architecte de solutions principal, AWS