Whitepaper AWS

Panoramica delle istanze Spot di Amazon EC2

Panoramica delle istanze Spot di Amazon EC2: Whitepaper AWS

Table of Contents

Panoramica delle istanze Spot di Amazon EC2

Data di pubblicazione: 5 marzo 2021 (Cronologia dei documenti e collaboratori)

Riassunto

Questo documento mira a permetterti di massimizzare il valore dei tuoi investimenti, migliorare l'accuratezza delle previsioni e la prevedibilità dei costi, creare una cultura della proprietà e della trasparenza dei costi e misurare continuamente il tuo stato di ottimizzazione.

Questo documento offre una panoramica delle istanze Spot di Amazon EC2 e illustra le best practice per utilizzarle con la massima efficienza.

Introduzione

Oltre alle istanze on demand, alle istanze riservate e ai Savings Plan, il quarto modello di prezzo di Amazon Elastic Compute Cloud (Amazon EC2) è costituito dalle istanze Spot.

Con le istanze Spot, puoi utilizzare la capacità di calcolo di riserva di Amazon EC2 con sconti fino al 90% rispetto ai prezzi on demand. Ciò significa che puoi ridurre in modo significativo il costo di esecuzione delle applicazioni o far crescere la capacità di calcolo e la velocità effettiva dell'applicazione mantenendo lo stesso budget. L'unica differenza tra istanze on demand e istanze Spot è che queste ultime possono essere interrotte da EC2 con una notifica di due minuti, nel momento in cui EC2 necessità di capacità.

A differenza delle istanze riservate o dei Savings Plan, le istanze Spot non richiedono un impegno per ottenere risparmi sui costi rispetto ai prezzi on demand. Tuttavia, poiché le istanze Spot possono essere terminate da EC2 se non c'è capacità disponibile nel pool di capacità (una combinazione di un tipo di istanza e una zona di disponibilità) in cui sono in esecuzione, sono più adatte per carichi di lavoro flessibili.

Quando usare le istanze Spot

Puoi utilizzare le istanze Spot per diverse applicazioni flessibili e con tolleranza ai guasti. Gli esempi comprendono server Web senza stato, endpoint API, applicazioni di analisi dei dati e Big Data, carichi di lavoro containerizzati, elaborazione CI/CD ad alte prestazioni e ad elevata velocità effettiva (HPC/HTC), carichi di lavoro di rendering e altri carichi di lavoro flessibili.

Le istanze Spot non sono adatte per carichi di lavoro non flessibili, con stato, senza tolleranza ai guasti o strettamente accoppiati tra nodi di istanze. Le istanze Spot sono inoltre sconsigliate per carichi di lavoro che non tollerano periodi occasionali in cui la capacità target non è completamente disponibile. Avvertiamo vivamente di non utilizzare le istanze Spot per questi carichi di lavoro o per tentare un failover su istanze on demand per gestire le interruzioni.

Come avviare le istanze Spot

Il servizio più consigliato per l'avvio delle istanze Spot è Amazon EC2 Auto Scaling perché consente di avviare e mantenere la capacità desiderata e di richiedere automaticamente risorse per sostituire quelle che sono state interrotte o terminate manualmente. Quando configuri un gruppo Auto Scaling, devi specificare solo i tipi di istanza e la capacità desiderata in base alle esigenze dell'applicazione. Per ulteriori informazioni, consulta gruppi Auto Scaling nella Guida per l'utente di Amazon EC2 Auto Scaling.

Se hai bisogno di maggiore flessibilità, hai creato i tuoi flussi di lavoro di avvio delle istanze o desideri controllare singoli aspetti dell'avvio delle istanze o dei meccanismi di dimensionamento, ti consigliamo di valutare l'uso del parco istanze EC2 in modalità istantanea come alternativa a EC2 Auto Scaling. Questa API sincrona consente di specificare un elenco di tipi di istanze e requisiti di avvio e fornisce funzionalità più flessibili rispetto alla chiamata API RunInstances EC2 per l'avvio di istanze Spot o istanze on demand.

Quando utilizzi i servizi AWS per l'esecuzione dei carichi di lavoro nel cloud, puoi utilizzarli anche per avviare istanze Spot. Gli esempi comprendono Amazon EMR, Amazon EKS, Amazon ECS, AWS Batch e AWS Elastic Beanstalk. Puoi anche avviare istanze Spot utilizzando strumenti di terze parti che si integrano con AWS Cloud.

Puoi automatizzare l'avvio di istanze Spot utilizzando strumenti Infrastructure as Code (AWS CloudFormation, AWS CDK) o l'API AWS, CLI o SDK. Spot Blueprints fornisce una procedura guidata che consente di generare modelli Infrastructure as Code per AWS Cloudformation e Hashicorp Terraform che aderiscono alle best practice di Spot.

Come funzionano le istanze Spot

Le istanze Spot offrono esattamente le stesse prestazioni di tutte le altre istanze EC2. Tuttavia, possono essere interrotte da Amazon EC2 quando EC2 ha bisogno di recuperare la capacità.

Quando EC2 interrompe l'istanza Spot, termina, arresta o iberna l'istanza, a seconda del comportamento di interruzione scelto.

Se EC2 interrompe l'istanza Spot nella prima ora, prima di un'ora di esecuzione completa, non verrà addebitato alcun costo per l'ora parziale utilizzata. Tuttavia, se interrompi o termini l'istanza Spot, paghi per qualsiasi ora parziale utilizzata (come faresti per le istanze on demand o le istanze riservate). Per informazioni sulla fatturazione delle istanze Spot interrotte in esecuzione su diversi sistemi operativi, consulta Fatturazione delle istanze Spot interrotte nella Guida per l'utente di EC2.

Il prezzo istanza Spot per ciascun tipo di istanza in ciascuna zona di disponibilità è determinato da trend a lungo termine di offerta e domanda di capacità EC2 inutilizzata. Paghi il prezzo istanza Spot in vigore, fatturato al secondo più vicino.

Eventualmente, è possibile specificare un prezzo massimo per le istanze Spot. Se non si specifica un prezzo massimo, il prezzo massimo di default è il prezzo on demand. Tieni presente che non paghi mai più del prezzo istanza Spot in vigore quando la tua istanza Spot è in esecuzione. Ti consigliamo di non specificare un prezzo massimo, ma piuttosto di impostare il prezzo massimo di default sul prezzo on demand. Un prezzo massimo elevato non aumenta le tue possibilità di avviare un'istanza Spot e non riduce le tue possibilità di interruzione dell'istanza Spot (perché EC2 può comunque interrompere la tua istanza Spot quando ha bisogno di recuperare la capacità).

Il prezzo istanza Spot per un tipo di istanza in una zona di disponibilità può cambiare in qualsiasi momento, ma in generale non lo fa frequentemente. AWS pubblica il prezzo istanza Spot corrente e i prezzi storici delle istanze Spot tramite l'API DescribeSpotPriceHistory e nella Console di gestione AWS, che rispecchia i dati dell'API. In questo modo è più semplice valutare i livelli e le tempistiche delle fluttuazioni del prezzo istanza Spot nel corso del tempo.

Gestione delle interruzioni dell'istanza Spot

Il modo migliore per gestire correttamente le interruzioni delle istanze Spot e ridurre al minimo l'impatto sulle prestazioni o sulla disponibilità è progettare l'applicazione con tolleranza ai guasti. A tale scopo, è possibile sfruttare i suggerimenti per il ribilanciamento delle istanze EC2 e gli avvisi di interruzione dell'istanza Spot.

Un suggerimento di ribilanciamento dell'istanza EC2 è un segnale di notifica di un rischio elevato di interruzione per un'istanza Spot. Il segnale ti dà la possibilità di gestire l'istanza Spot in modo proattivo rispetto all'avviso di interruzione dell'istanza Spot di due minuti. È possibile decidere di ribilanciare il carico di lavoro su istanze Spot nuove o esistenti che non presentano un rischio elevato di interruzione. Abbiamo semplificato l'utilizzo di questo segnale utilizzando la funzione di ribilanciamento della capacità nei gruppi Auto Scaling di EC2. Per ulteriori informazioni, consulta Ribilanciamento della capacità di Amazon EC2 Auto Scaling.

Una notifica di interruzione di un'istanza Spot è un avviso emesso due minuti prima che Amazon EC2 interrompa un'istanza Spot. Se il carico di lavoro è "flessibile in termini di tempo", è possibile configurare le istanze Spot in modo che vengano arrestate o ibernate, invece di essere terminate, quando vengono interrotte. Amazon EC2 arresta o iberna automaticamente le istanze Spot in caso di interruzione e riprende automaticamente le istanze quando abbiamo capacità disponibile.

Puoi utilizzare il suggerimento per il ribilanciamento delle istanze EC2 e/o la notifica di interruzione dell'istanza Spot per progettare il carico di lavoro tenendo presente la tolleranza agli errori, in modo da poter acquisire le notifiche e salvare lo stato di un processo nell'archiviazione (ad esempio, Amazon S3, Amazon EFS o Amazon FSx), mantenere i file di log dall'istanza (o trasmetterli in streaming continuamente per un approccio con maggiore tolleranza ai guasti), drenare le connessioni da un bilanciatore del carico, ecc.

Alcuni servizi AWS e di terze parti gestiscono già le interruzioni Spot per ridurre l'impatto sulla tua applicazione. Amazon EKS ad esempio, che esegue gruppi di nodi gestiti con istanze Spot avvia automaticamente i nodi Kubernetes sostitutivi quando vengono forniti un suggerimento di ribilanciamento o avvisi di interruzione per un nodo esistente.

Limiti di istanze Spot

Esiste un limite al numero di istanze Spot in esecuzione e richieste di account AWS per regione. I limiti vengono gestiti in termini di numero di unità di elaborazione centrali virtuali (vCPU) che vengono utilizzate o verranno utilizzate dalle istanze Spot in esecuzione in attesa dell'evasione delle richieste di istanze Spot aperte. Se termini ma non annulli le istanze Spot, le richieste vengono conteggiate in base al limite di vCPU dell'istanza Spot finché Amazon EC2 non rileva la terminazione delle istanze Spot e chiude le richieste.

Esistono sei limiti di istanze Spot:

- Tutte le richieste di istanze Spot standard (A, C, D, H, I, M, R, T, Z)
- Tutte le richieste di istanze Spot F
- Tutte le richieste di istanze Spot G
- Tutte le richieste di istanze Spot Inf
- Tutte le richieste di istanze Spot P
- Tutte le richieste di istanze Spot X

Ogni limite specifica il limite vCPU per una o più famiglie di istanze. Per informazioni sulle diverse famiglie di istanze, sulle generazioni e le dimensioni, consulta Tipi di istanza Amazon EC2.

Con i limiti vCPU, puoi utilizzare il tuo limite in termini di numero di vCPU richiesti per avviare una qualsiasi combinazione di tipi di istanza che soddisfano le mutevoli esigenze dell'applicazione. Ad esempio, supponiamo che il limite di tutte le richieste di istanze Spot standard sia di 256 vCPU, che sia possibile richiedere 32 `m5.2xlarge` istanze Spot (32 x 8 vCPU) o 16 `c5.4xlarge` istanze Spot (16 x 16 vCPU) o una combinazione di qualsiasi tipo e dimensione di istanza Spot standard per un totale di 256 vCPU.

Per maggiori informazioni, consulta Monitoraggio di limiti e utilizzo delle istanze Spot e Richiedere un aumento del limite di istanze Spot nella Guida per l'utente di Amazon EC2 per le istanze Linux.

Best practice per le istanze Spot

I requisiti relativi al tipo di istanza e al budget, nonché la progettazione dell'applicazione, determineranno come applicare le seguenti best practice per la tua applicazione.

- Scegli i tipi di istanza in modo flessibile. Un pool di istanze Spot è un insieme di istanze EC2 inutilizzate con lo stesso tipo di istanza (ad esempio m5.large) e zona di disponibilità (ad esempio, us-east-1a). È necessario essere flessibili sui tipi di istanza richiesti e sulle zone di disponibilità in cui è possibile distribuire il carico di lavoro. Questo offre a Spot una migliore possibilità di trovare e allocare la quantità di capacità di elaborazione richiesta. Ad esempio, non richiedere solo c5.large se sei disposto a usare grandi quantità delle famiglie c5, c4 e m5.
- Utilizza una strategia di allocazione ottimizzata della capacità. Le strategie di allocazione nei gruppi EC2 Auto Scaling consentono di effettuare il provisioning della capacità target senza la necessità di cercare manualmente i pool di istanze Spot con capacità inutilizzata. È consigliabile utilizzare la strategia ottimizzata della capacità perché questa effettua automaticamente il provisioning delle istanze dai pool di istanze Spot più disponibili. Poiché la capacità delle istanze Spot proviene da pool con capacità ottimale, ciò riduce la possibilità che le istanze Spot vengano interrotte. Per ulteriori informazioni sulle strategie di allocazione, consulta Istanze Spot nella Guida per l'utente di Amazon EC2 Auto Scaling.
- Utilizza il ribilanciamento proattivo della capacità. Il ribilanciamento della capacità consente di mantenere la disponibilità del carico di lavoro aumentando in modo proattivo il gruppo Auto Scaling con una nuova istanza Spot prima che un'istanza Spot in esecuzione riceva l'avviso di interruzione di due minuti. Quando il ribilanciamento della capacità è abilitato, l'Auto Scaling tenta di sostituire in modo proattivo le istanze Spot che hanno ricevuto un suggerimento di ribilanciamento, offrendo l'opportunità di ribilanciare il carico di lavoro con nuove istanze Spot che non presentano un elevato rischio di interruzione.
- Utilizza i servizi AWS integrati per gestire le istanze Spot. Altri servizi AWS si integrano con Spot per ridurre i costi di calcolo complessivi senza la necessità di gestire singole istanze o parchi istanze. Ti consigliamo di considerare le seguenti soluzioni per i carichi di lavoro applicabili: Amazon EMR, Amazon ECS, AWS Batch, Amazon EKS, SageMaker, AWS Elastic Beanstalk e Amazon GameLift. Per ulteriori informazioni sulle best practice Spot con questi servizi, consulta il sito Web dei workshop sulle istanze Spot di Amazon EC2.
- Scegli lo strumento di avvio moderno e corretto per le istanze Spot. Se uno dei servizi integrati AWS non è adatto al tuo carico di lavoro e devi comunque creare la tua applicazione con il controllo sull'avvio delle istanze Spot, usa lo strumento giusto. Per la maggior parte dei carichi

di lavoro, è consigliabile utilizzare EC2 Auto Scaling perché fornisce un set di caratteristiche più completo per un'ampia varietà di carichi di lavoro, come le applicazioni supportate da ELB, i carichi di lavoro containerizzati e i processi di elaborazione delle code. Se hai bisogno di un maggiore controllo sulle singole richieste e stai cercando uno strumento di "solo avvio", usa il parco istanze EC2 in modalità istantanea come sostituzione drop-in di RunInstances, ma con un set più ampio di funzionalità, come la diversificazione del tipo di istanza e delle strategie di allocazione.

Integrazione di Spot con altri servizi AWS

Le istanze Spot di Amazon EC2 si integrano con diversi servizi AWS.

Integrazione con Amazon EMR

È possibile eseguire i cluster Amazon EMR su istanze Spot e ridurre in modo significativo i costi di elaborazione di grandi quantità di dati per i carichi di lavoro di analisi dei dati. È possibile eseguire i cluster EMR combinando facilmente istanze Spot con istanze on demand e istanze riservate utilizzando la caratteristica Parchi istanze EMR. È possibile utilizzare le strategie di allocazione EMR per avviare istanze Spot dai pool di capacità più disponibili.

Integrazione con EC2 Auto Scaling

Puoi utilizzare i gruppi Amazon EC2 Auto Scaling per avviare e gestire istanze Spot, mantenere la disponibilità delle applicazioni, diversificare il tipo di istanza e la selezione delle opzioni di acquisto (on demand/Spot) e dimensionare la capacità di Amazon EC2 utilizzando policy di dimensionamento dinamiche, pianificate e predittive. Per maggiori informazioni, consulta la sezione Richiesta di istanze spot per applicazioni flessibili e con tolleranza ai guasti nella Guida per l'utente di Amazon EC2 Auto Scaling.

Integrazione con Amazon EKS

Puoi ottimizzare i costi dei carichi di lavoro basati su Kubernetes utilizzando Amazon EKS, avviando le istanze Spot nei gruppi di nodi gestiti EKS. I gruppi di nodi gestiti EKS gestiscono l'intero ciclo di vita delle istanze Spot, sostituendo quelle che saranno presto interrotte con istanze appena avviate, per ridurre le possibilità di impatto sulle prestazioni o sulla disponibilità delle applicazioni quando le istanze Spot vengono interrotte (quando EC2 ha bisogno di recuperare la capacità). Per ulteriori informazioni, consulta Gruppi di nodi gestiti nella Guida per l'utente di Amazon EKS.

Integrazione con Amazon ECS

Puoi eseguire i cluster Amazon ECS su istanze Spot per ridurre i costi operativi dell'esecuzione di applicazioni containerizzate. Amazon ECS supporta lo svuotamento automatico delle istanze Spot che saranno presto interrotte. Per ulteriori informazioni, consulta Utilizzo delle istanze Spot in Guida per gli sviluppatori di Amazon Elastic Container Service.

Integrazione di Amazon ECS con AWS Fargate Spot

Se le tue attività containerizzate sono interrompibili e flessibili, puoi scegliere di eseguire le tue attività ECS con il provider di capacità AWS Fargate Spot. Ciò significa che le tue attività verranno eseguite su AWS Fargate, una piattaforma di container serverless, e potrai beneficiare dei risparmi sui costi con Fargate Spot. Per ulteriori informazioni, consulta AWS Fargate capacity providers nella Amazon Elastic Container Service Developer Guide.

Integrazione con Amazon Batch

AWS Batch pianifica, programma ed esegue carichi di lavoro in batch in AWS. AWS Batch può inoltrare dinamicamente le richieste di istanze Spot per tuo conto, riducendo il costo di esecuzione dei processi in batch.

Integrazione con Amazon SageMaker

Amazon SageMaker semplifica l'addestramento di modelli di machine learning utilizzando istanze Spot gestite. L'addestramento di Spot gestito può ottimizzare il costo dei modelli di addestramento fino al 90% rispetto alle istanze on demand. SageMaker gestisce le interruzioni Spot per tuo conto. Per ulteriori informazioni, consulta Managed Spot Training in Amazon SageMaker nella Amazon Elastic Container Service Developer Guide.

Integrazione con Amazon Gamelift

Amazon GameLift è una soluzione di hosting di server di giochi che implementa, gestisce e dimensiona i server cloud per giochi multigiocatore. Il supporto per le istanze Spot in Amazon Gamelift ti offre l'opportunità di ridurre notevolmente i costi di hosting. Quando si creano parchi istanze di risorse di hosting, è possibile scegliere tra istanze on demand o istanze Spot. Mentre le istanze Spot potrebbero essere interrotte con due minuti di notifica, FleetIQ di Amazon GameLift riduce al minimo la possibilità di interruzioni. Per maggiori informazioni, consulta Using Spot Instances with GameLift nella Guida per sviluppatori di Amazon GameLift.

Integrazione con Elastic Beanstalk

AWS Elastic Beanstalk è un servizio di semplice utilizzo per implementare e dimensionare applicazioni Web e servizi sviluppati con Java, .NET, PHP, Node.js, Python, Ruby, Go e Docker

su server comuni come Apache, Nginx, Passenger e IIS. È sufficiente caricare il codice e AWS Elastic Beanstalk gestisce automaticamente l'implementazione, dal provisioning della capacità, dal bilanciamento del carico e dalla scalabilità automatica al monitoraggio dell'integrità delle applicazioni. Puoi utilizzare le istanze Spot negli ambienti Elastic Beanstalk per ottimizzare i costi dell'infrastruttura sottostante delle tue applicazioni Web. Per informazioni sull'utilizzo delle istanze Spot con Elastic Beanstalk, consulta il Supporto istanze Spot nella AWS Elastic Beanstalk Guida per lo sviluppatore.

Conclusione

Sia che tu abbia esigenze di calcolo flessibili o desideri accrescere la capacità senza aumentare il budget, le istanze Spot possono essere un ottimo modo per ottimizzare i costi AWS e/o costruire pensando alla scalabilità. Grazie alla corretta progettazione dei carichi di lavoro, è possibile sfruttare le istanze Spot per un'ampia gamma di esigenze. Per ulteriori informazioni, consulta Istanze Spot di Amazon EC2.

Risorse

- Centro di progettazione AWS
- Whitepaper AWS
- Mensile di architettura AWS
- Blog sull'architettura di AWS
- Video La mia architettura
- Documentazione AWS

Cronologia dei documenti e collaboratori

Cronologia dei documenti

Per ricevere una notifica sugli aggiornamenti di questo whitepaper, iscriviti al feed RSS.

update-history-change	update-history-description	update-history-date
Aggiornamento di minore entità	Layout di pagina modificato.	30 aprile 2021
Aggiornamento di minore entità	Contenuti aggiornati per riflettere le best practice attuali. Il nome del whitepaper è cambiato da "Come utilizzare le istanze Spot di Amazon EC2 in modo scalabile" a "Panoramica sulle istanze Spot di Amazon EC2" per riflettere meglio i contenuti.	5 marzo 2021
Aggiornamento di minore entità	I limiti delle istanze Spot sono stati aggiornati.	3 febbraio 2021
Pubblicazione iniziale	Come utilizzare le istanze Spot di Amazon EC2 in modo scalabile pubblicato.	1 marzo 2018

> **Note**
>
> Per iscriversi e ricevere gli aggiornamenti RSS, è necessario disporre di un plug-in RSS abilitato per il browser in uso.

Collaboratori

Hanno contribuito alla stesura di questo documento:

- Amilcar Alfaro, Sr. Product Marketing Manager, AWS
- Erin Carlson, Marketing Manager, AWS
- Keith Jarrett, WW BD Lead - Cost Optimization, AWS Business Development
- Ran Sheinberg, Principal Solutions Architect, AWS