AWS 白皮書

Amazon EC2 Spot 執行個體概觀

Amazon EC2 Spot 執行個體概觀: AWS 白皮書

Table of Contents

Amazon EC2 Spot 執行個體概觀

發佈日期：2021 年 3 月 5 日 (文件歷史記錄與貢獻者)

摘要

本白皮書旨在協助您盡可能提高投資回報的價值、提高預測準確度和成本可預測性，建立所有權和成本透明度的文化，並持續衡量您的最佳化狀態。

本白皮書提供 Amazon EC2 Spot 執行個體的概觀，以及有效使用這些執行個體的最佳實務。

簡介

除了隨需、預留執行個體及 Savings Plans，第四個 Amazon Elastic Compute Cloud (Amazon EC2) 定價方式是 Spot 執行個體。

您可以透過 Spot 執行個體使用閒置的 Amazon EC2 運算容量，且相較於隨需定價，其折扣最高可達 90%。這表示您可以大幅降低執行應用程式的成本，或是以相同的預算增加應用程式的運算容量和輸送量。隨需執行個體與 Spot 執行個體的唯一差異是，當 EC2 需要收回容量時，EC2 只需在兩分鐘前通知即可中斷 Spot 執行個體。

與預留執行個體或 Savings Plans 不同，Spot 執行個體不需要承諾即可針對隨需定價節省成本。但是，因為 Spot 執行個體可能會在執行的容量集區 (執行個體類型和可用區域的組合) 中沒有可用容量時遭 EC2 終止，所以最適合靈活的工作負載。

何時使用 Spot 執行個體

您可以將 Spot 執行個體用於各種具備容錯能力和靈活性的應用程式。範例包括無狀態的 Web 伺服器、API 端點、大數據和分析應用程式、容器化工作負載、CI/CD 高效能及高輸送量運算 (HPC/HTC)、渲染工作負載和其他靈活的工作負載。

Spot 執行個體不適合執行個體節點之間不靈活、具備狀態、不具備容錯能力或緊密結合的工作負載。針對無法容忍目標容量偶爾會無法完全提供使用的工作負載，也不建議使用 Spot 執行個體。我們強烈建議不要將 Spot 執行個體用於這些工作負載，或試圖容錯移轉到隨需執行個體以處理中斷。

如何啟動 Spot 執行個體

最建議啟動 Spot 執行個體的服務是 Amazon EC2 Auto Scaling，因為其可以讓您啟動和維持所需的容量，並自動請求資源替換任何遭到中斷或手動終止的容量。當您設定 Auto Scaling 群組時，您只需要根據應用程式需求指定執行個體類型和所需的容量。如需詳細資訊，請參閱《Amazon EC2 Auto Scaling 使用者指南》中的〈Auto Scaling 群組〉。

當您需要更多靈活性，已建置您自己的執行個體啟動工作流程，或是需要控制執行個體啟動或擴展機制的各個方面時，建議您評估透過立即模式使用 EC2 機群，做為 EC2 Auto Scaling 的替代項目。此同步 API 可讓您指定執行個體類型清單及啟動需求，並針對啟動 Spot 執行個體或隨需執行個體，提供比 EC2 RunInstances API 呼叫更大的靈活性。

當您使用 AWS 服務來執行您的雲端工作負載時，您也可以使用這些服務來啟動 Spot 執行個體。範例包括 Amazon EMR、Amazon EKS、Amazon ECS、AWS Batch 及 AWS Elastic Beanstalk。您也可以使用與 AWS 雲端整合的第三方工具啟動 Spot 執行個體。

您可以透過使用基礎設施即程式碼工具 (AWS CloudFormation、AWS CDK) 或 AWS API、CLI 或 SDK 自動化 Spot 執行個體的啟動程序。Spot 藍圖提供了指引精靈，可讓您產生符合 Spot 最佳實務的 AWS CloudFormation 和 Hashicorp Terraform 基礎設施即程式碼範本。

Spot 執行個體的運作方式

Spot 執行個體在執行時的運作方式與其他 EC2 執行個體完全相同。但是，當 Amazon EC2 需要回收容量時，EC2 可以中斷這些執行個體。

當 EC2 中斷您的 Spot 執行個體時，取決於您選擇的中斷行為，其會終止、停止或使執行個體休眠。

若 EC2 在您的 Spot 執行個體執行滿一個小時前中斷該執行個體，您無需為所使用的時間支付費用。但是，如果您停止或終止 Spot 執行個體，不足一小時的部分將按一小時收費 (與您為隨需或預留執行個體支付費用的方式相同)。如需在不同作業系統上執行 Spot 執行個體而後中斷的費用資訊，請參閱《EC2 使用者指南》中的〈針對已中斷的 Spot 執行個體收費〉。

每個可用區域中每個執行個體類型的 Spot 價格取決於 EC2 閒置容量的長期供需趨勢。您支付的費用將是當時有效的 Spot 價格，並會按最接近的秒數計費。

或者，您可以選擇為您的 Spot 執行個體指定最高價格。如果您沒有指定最高價格，預設的最高價格是隨需價格。請注意，您支付的費用永遠不會超過在執行 Spot 執行個體時有效的 Spot 價格。我們建議您不要指定最高價格，而是讓最高價格維持預設的隨需價格。高昂的最高價格不會增加啟動 Spot 執行個體的機會，也不會減少 Spot 執行個體遭到中斷的機會 (因為 EC2 仍然會在需要回收容量時中斷您的 Spot 執行個體)。

可用區域中執行個體類型的 Spot 價格可能會隨時變更，但一般不會經常變更。AWS 會透過 DescribeSpotPriceHistory API 及 AWS 管理主控台 (其反映了前述 API 的資料) 發佈目前的 Spot 價格和 Spot 執行個體的歷史價格。這些資訊有助於您評定 Spot 價格隨時間浮動的程度和時間。

管理 Spot 執行個體中斷

要優雅地處理 Spot 執行個體中斷，並盡可能降低對效能或可用性的影響，最佳方式便是將您的應用程式架構設計為具備容錯能力。若要達到此目的，您可以利用 EC2 執行個體重新平衡建議和 Spot 執行個體中斷通知。

EC2 執行個體重新平衡建議是一種新的訊號，可在 Spot 執行個體的中斷風險提高時通知您。該訊號使您有機會在兩分鐘的 Spot 執行個體中斷通知前主動管理 Spot 執行個體。您可以決定將工作負載重新平衡至中斷風險較低的新增或現有 Spot 執行個體。我們透過在 EC2 Auto Scaling 群組中提供容量重新平衡功能，方便您使用這個訊號。如需詳細資訊，請參閱 Amazon EC2 Auto Scaling 容量重新平衡。

Spot 執行個體中斷通知是在 Amazon EC2 中斷 Spot 執行個體前兩分鐘所發出的警告。如果您的工作負載「時間很彈性」，您可以將 Spot 執行個體設為在中斷時停止或休眠，而不是終止。Amazon EC2 會在中斷時自動停止或使您的 Spot 執行個體休眠，並在我們有可用容量時自動繼續執行個體。

您可以使用 EC2 執行個體重新平衡建議和 (或) Spot 執行個體中斷通知，在設計工作負載架構時將容錯能力納入考量，讓您可以擷取通知，並將任務的狀態儲存至儲存體 (例如 Amazon S3、Amazon EFS 或 Amazon FSx)、保存執行個體的記錄日誌 (或持續串流以提高容錯能力)、用盡負載平衡器的連線等。

某些 AWS 和第三方服務已會替您處理 Spot 中斷，以減少對您應用程式所造成的影響。例如，使用 Spot 執行個體執行受管節點群組的 Amazon EKS 會在現有節點收到重新平衡建議或中斷通知時，自動啟動替代的 Kubernetes 節點。

Spot 執行個體限制

每個區域每個 AWS 帳戶執行和請求的 Spot 執行個體數量都有限制。Spot 執行個體限制的管理依據，是您執行中 Spot 執行個體正在或即將使用，且待處理 Spot 執行個體請求的「虛擬中央處理單元 (vCPU)」數量。如果您終止 Spot 執行個體但未取消 Spot 執行個體請求，請求會計入您的 Spot 執行個體 vCPU 限制，直到 Amazon EC2 偵測到 Spot 執行個體終止並關閉請求為止。

Spot 執行個體有六個限制：

- 所有標準 (A、C、D、H、I、M、R、T、Z) Spot 執行個體請求
- 所有 F Spot 執行個體請求
- 所有 G Spot 執行個體請求
- 所有 Inf Spot 執行個體請求
- 所有 P Spot 執行個體請求
- 所有 X Spot 執行個體請求

每個限制各指定一或多個執行個體系列的 vCPU 限制。如需不同執行個體系列、世代和大小的相關資訊，請參閱 Amazon EC2 執行個體類型。

透過 vCPU 限制，您可以根據為了滿足不斷變化的應用程式需求，而啟動任意執行個體類型組合所需的 vCPU 數量來運用限制。例如，假設您所有的標準 Spot 執行個體請求限制是 256 個 vCPU，您可以請求 32 個 `m5.2xlarge` Spot 執行個體 (32 x 8 個 vCPU)，或是 16 個 `c5.4xlarge` Spot 執行個體 (16 x 16 個 vCPU)，或是在合計 256 個 vCPU 的情況下組合任何標準 Spot 執行個體類型。

如需詳細資訊，請參閱《適用於 Linux 的 Amazon EC2 使用者指南》中的〈監控 Spot 執行個體限制和用量〉和〈請求增加 Spot 執行個體限制〉。

Spot 執行個體最佳實務

您的執行個體類型需求、預算需求以及應用程式設計將決定如何為應用程式應用下列最佳實務：

- 靈活使用各種執行個體類型。 Spot 執行個體集區是一組未使用的 EC2 執行個體，具有相同執行個體類型 (例如 m5.large) 和可用區域 (例如 us-east-1a)。您應該對於請求的執行個體類型，以及可在其中部署工作負載的可用區域具有彈性。這讓 Spot 有更好的機會找到並配置您所需的運算容量。例如，如果您願意使用 c4、m5 和 m4 系列的 large，則不要只要求使用 c5.large。
- 使用容量最佳化配置策略。 EC2 Auto Scaling 群組中的配置策略可協助您佈建目標容量，而無需手動尋找具有閒置容量的 Spot 執行個體集區。建議您使用容量最佳化策略，因為這種策略會自動從可用性最高的 Spot 執行個體集區佈建執行個體。由於您的 Spot 執行個體容量來自具有最佳容量的集區，因此可降低您 Spot 執行個體遭到中斷的機率。如需配置策略的詳細資訊，請參閱《Amazon EC2 Auto Scaling 使用者指南》中的〈Spot 執行個體〉。
- 使用主動容量重新平衡。 容量重新平衡可協助您維持工作負載可用性，方法是在執行中的 Spot 執行個體收到兩分鐘的 Spot 執行個體中斷通知之前，使用新的 Spot 執行個體主動擴增您的 Auto Scaling 群組。啟用容量重新平衡時，Auto Scaling 會嘗試主動替換收到重新平衡建議的 Spot 執行個體，讓您有機會將工作負載重新平衡到中斷風險較低的新 Spot 執行個體。
- 使用整合的 AWS 服務管理您的 Spot 執行個體。 其他 AWS 服務與 Spot 整合，可降低整體運算成本，且無需管理個別執行個體或機群。建議您針對適用的工作負載，考慮下列解決方案：Amazon EMR、Amazon ECS、AWS Batch、Amazon EKS、SageMaker、AWS Elastic Beanstalk 和 Amazon GameLift。若要進一步了解這些服務的 Spot 最佳實務，請參閱 Amazon EC2 Spot 執行個體研討會網站。
- 為 Spot 執行個體選擇現代化且正確的啟動工具。 若其中一個 AWS 整合服務不適合您的工作負載，但您仍然需要建置應用程式並控制 Spot 執行個體的啟動，請使用適當的工具。針對大多數工作負載，都建議您使用 EC2 Auto Scaling，因為其為各種工作負載 (例如採用 ELB 的應用程式、容器化工作負載和佇列處理任務) 提供了更全面的功能集。如果您需要對個別請求具備更多控制，並且正在尋找「僅啟動」的工具，請透過立即模式使用 EC2 機群做為 RunInstances 的直接替代方案。其具備了更廣泛的功能集，例如執行個體類型多樣化和配置策略。

Spot 與其他 AWS 服務的整合

Amazon EC2 Spot 執行個體與多種 AWS 服務整合。

Amazon EMR 整合

您可以在 Spot 執行個體上執行 Amazon EMR 叢集，並大幅降低為分析工作負載處理大量資料的成本。客戶可以透過使用 EMR 執行個體機群功能，輕鬆混合 Spot 執行個體與隨需執行個體和預留執行個體，以執行您的 EMR 叢集。您可以使用 EMR 配置策略，從可用性最高的容量集區啟動 Spot 執行個體。

EC2 Auto Scaling 整合

您可以使用 Amazon EC2 Auto Scaling 群組啟動和管理 Spot 執行個體、維護應用程式可用性、多樣化執行個體類型和購買選項 (隨需/Spot)，並使用動態、排程和預測性擴展策略擴展 Amazon EC2 容量。如需詳細資訊，請參閱《Amazon EC2 Auto Scaling 使用者指南》中的〈為具備容錯能力和靈活性的應用程式請求 Spot 執行個體〉。

Amazon EKS 整合

您可以使用 Amazon EKS，在 EKS 受管節點群組中啟動 Spot 執行個體，以對 Kubernetes 式工作負載進行成本最佳化。EKS 受管節點群組會管理整個 Spot 執行個體的生命週期，將即將遭到中斷的 Spot 執行個體替換成剛啟動的執行個體，以在 Spot 執行個體遭到中斷時 (EC2 需要回收容量時) 降低對您應用程式效能或可用性造成影響的機率。若要進一步了解，請參閱《Amazon EKS 使用者指南》中的〈受管節點群組〉。

Amazon ECS 整合

您可以在 Spot 執行個體上執行 Amazon ECS 叢集，以降低執行容器化應用程式的營運成本。Amazon ECS 支援自動用盡即將中斷的 Spot 執行個體。如需詳細資訊，請參閱《Amazon Elastic Container Service 開發人員指南》中的〈使用 Spot 執行個體〉。

具備 AWS Fargate Spot 整合的 Amazon ECS

如果您的容器化任務可中斷且彈性高，您可以選擇使用 AWS Fargate Spot 容量提供者執行您的 ECS 任務，這表示您的任務將會在 AWS Fargate (一種無伺服器容器平台) 上執行，並且您將可以透過

Fargate Spot 節省成本。如需詳細資訊，請參閱《Amazon Elastic Container Service 開發人員指南》中的〈AWS Fargate 容量提供者〉。

Amazon Batch 整合

AWS Batch 可在 AWS 上規劃、排程及執行您的批次運算工作負載。AWS Batch 還能代您動態請求 Spot 執行個體，進一步降低執行批次任務的成本。

Amazon SageMaker 整合

Amazon SageMaker 可讓您使用受管 Spot 執行個體，輕鬆訓練機器學習模型。與隨需執行個體相較，受管 Spot 訓練將模型訓練成本最佳化的幅度可達 90%。SageMaker 會代您管理 Spot 中斷。如需詳細資訊，請參閱《Amazon SageMaker 開發人員指南》中的〈Amazon SageMaker 中的受管 Spot 訓練〉。

Amazon GameLift 整合

Amazon GameLift 是可為多人遊戲部署、操作並擴展雲端伺服器的遊戲伺服器託管解決方案。Amazon Gamelift 中的 Spot 執行個體支援可讓您有機會大幅降低託管成本。建立託管資源機群時，您可以在隨需執行個體或 Spot 執行個體間進行選擇。雖然 Spot 執行個體可能遭到中斷 (有兩分鐘的通知)，但 Amazon GameLift 的 FleetIQ 可將中斷的機會降至最低。如需詳細資訊，請參閱《Amazon GameLift 開發人員指南》中的〈搭配 GameLift 使用 Spot 執行個體〉。

AWS Elastic Beanstalk 整合

AWS Elastic Beanstalk 是一項易用的服務，用於在熟悉的伺服器 (例如 Apache、Nginx、Passenger 和 IIS) 上部署和擴展以 Java、.NET、PHP、Node.js、Python、Ruby、Go 和 Docker 開發的 Web 應用程式和服務。您只需上傳您的程式碼，Elastic Beanstalk 就會自動進行部署，從容量佈建、負載平衡、自動擴展到應用程式運作狀態監控等。您可以在 Elastic Beanstalk 環境中使用 Spot 執行個體來最佳化 Web 應用程式底層基礎設施的成本。如需搭配 Elastic Beanstalk 使用 Spot 執行個體的資訊，請參閱《AWS Elastic Beanstalk 開發人員指南》中的〈Spot 執行個體支援〉。

結論

無論您是有靈活的運算需求，還是希望在不增加預算的情況下調整容量，Spot 執行個體都是一種最佳化 AWS 成本和 (或) 進行大規模建置的絕佳方式。透過適當地設計工作負載架構，您可以將 Spot 執行個體應用於各種需求。如需詳細資訊，請參閱 Amazon EC2 Spot 執行個體。

資源

- AWS 架構中心
- AWS 白皮書
- AWS 每月架構
- AWS 架構部落格
- This Is My Architecture 影片
- AWS 文件

文件歷史記錄與貢獻者

文件歷史記錄

若要收到此白皮書更新的通知，請訂閱 RSS 摘要。

update-history-change	update-history-description	update-history-date
小幅度更新	調整頁面配置。	2021 年 4 月 30 日
小幅度更新	更新內容，以反映目前的最佳實務。將白皮書的名稱從《大規模利用 Amazon EC2 Spot 執行個體》變更為《Amazon EC2 Spot 執行個體概觀》以更忠實地反映內容。	2021 年 3 月 5 日
小幅度更新	更新 Spot 執行個體限制。	2021 年 2 月 3 日
初次出版	大規模利用 Amazon EC2 Spot 執行個體	2018 年 3 月 1 日

> **Note**
>
> 若要訂閱 RSS 更新，您必須為正在使用的瀏覽器啟用 RSS 外掛程式。

貢獻者

協力完成本文件的個人與組織如下：

- Amilcar AlfaroAWS 資深產品行銷經理
- AWS 行銷經理 Erin Carlson
- AWS 業務開發 WW BD 負責人 – 成本最佳化 Keith Jarrett
- AWS 首席解決方案架構師 Ran Sheinberg