AWS 白皮書

AWS 上的即時通訊

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS 上的即時通訊: AWS 白皮書

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon 的商標和商業外觀不得用於任何非 Amazon 的產品或服務，也不能以任何可能造成客戶混淆、任何貶低或使 Amazon 名譽受損的方式使用 Amazon 的商標和商業外觀。所有其他非 Amazon 擁有的商標均為其各自擁有者的財產，這些擁有者可能隸屬於 Amazon，或與 Amazon 有合作關係，或由 Amazon 贊助。

Table of Contents

摘要 1
摘要 1
您是 Well-Architected 嗎？ 1
簡介 2
RTC 架構的基本元件 3
軟開關/PBX 4
工作階段邊界控制器 (SBC) 4
PSTN 連線 4
PSTN 閘道 4
SIP 中繼線 4
媒體閘道 （轉換器） 4
WebRTC 中的推送通知 5
WebRTC 和 WebRTC 閘道 5
上的高可用性和可擴展性 AWS 8
作用中待命狀態伺服器之間 HA 的浮動 IP 模式 8
RTC 解決方案的適用性 8
RTC 架構的適用性 10
使用 Application Load Balancer 和 Auto Scaling AWS 在 上針對 WebRTC 進行負載平衡 Application Load Balancer 10
使用 Network Load Balancer 或 AWS Marketplace 產品的 SIP 實作 11
跨區域 DNS 型負載平衡和容錯移轉 12
資料耐久性和具有持久性儲存體的 HA 13
使用 AWS Lambda、Amazon Route 53 和 Amazon EC2 Auto Scaling 進行動態擴展 14
具有 Amazon Kinesis Video Streams 的高可用性 WebRTC 15
搭配 Amazon Chime Voice Connector 的高可用性 SIP 中繼 15
欄位的最佳實務 16
建立 SIP 浮水印 16
執行詳細監控 17
使用 DNS 進行負載平衡，並將 IPs 浮動進行容錯移轉 17
使用多個可用區域 19
將流量保持在一個可用區域內，並使用 EC2 置放群組 19
使用增強型聯網 EC2 執行個體類型 20
安全考量 21
結論 22

縮寫 23
貢獻者 25
文件修訂 26
注意 27
AWS 詞彙表 28
..... xxix

上的即時通訊 AWS

在 上設計高可用性和可擴展即時通訊 (RTC) 工作負載的最佳實務 AWS

發佈日期：2022 年 5 月 5 日 (文件修訂)

摘要

如今，許多組織都希望降低成本，並實現即時語音、訊息和多媒體工作負載的可擴展性。本白皮書概述在 Amazon Web Services () 上管理即時通訊 (RTC AWS) 工作負載的最佳實務，並包含符合這些要求的參考架構。此白皮書為熟悉如何實現這些工作負載高可用性和可擴展性之即時通訊的個人提供指南。

本白皮書包含顯示如何設定 RTC 工作負載的參考架構 AWS，以及最佳化解決方案以符合最終使用者需求的最佳實務，同時最佳化雲端。Evolved Packet Core (EPC) 超出此白皮書的範圍，但此處詳述的最佳實務可以套用至虛擬網路函數 (VNFs)。

您是 Well-Architected 嗎？

AWS Well-Architected Framework 可協助您了解在雲端建置系統時所做決策的優缺點。架構的六個支柱可讓您了解架構最佳實務，以設計和操作可靠、安全、高效、經濟實惠且永續的系統。使用 AWS Well-Architected Tool（需要AWS 管理主控台登入），您可以針對每個支柱回答一組問題，根據這些最佳實務來檢閱工作負載。

如需雲端架構的更多專家指引和最佳實務，請參閱AWS 架構中心，參考架構部署、圖表和白皮書。

簡介

電信應用程式使用語音、視訊和簡訊做為頻道，是許多組織及其最終使用者的關鍵需求。這些即時通訊 (RTC) 工作負載具有特定的延遲和可用性要求，可透過遵循相關的設計最佳實務來滿足這些要求。過去，RTC 工作負載已部署在具有專用資源的傳統內部部署資料中心。

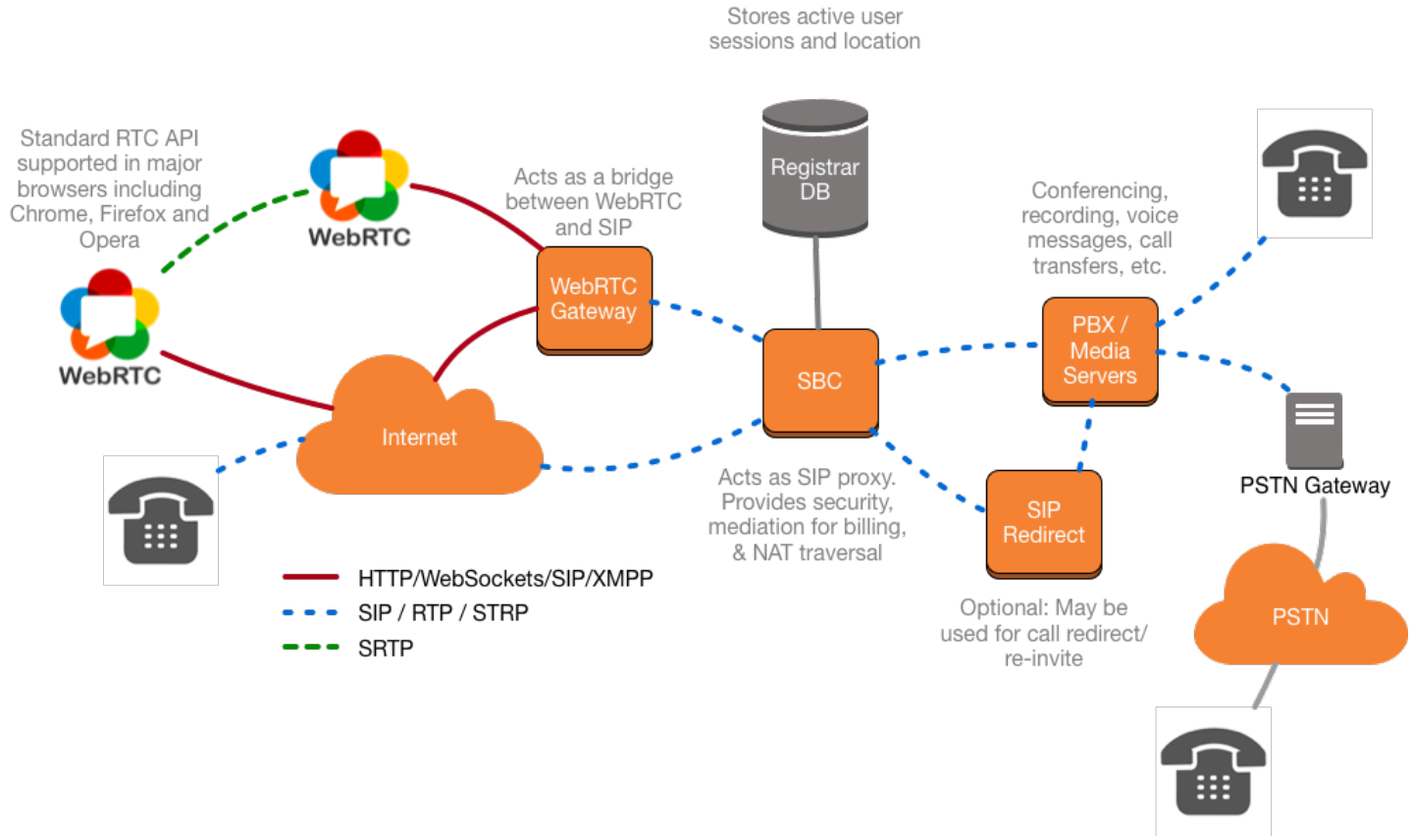
RTC 工作負載需要高度可擴展、彈性和可用的環境。如今，客戶使用 AWS 以降低成本、提高敏捷性、彈性和上市時間來執行 RTC 工作負載。

RTC 架構的基本元件

在電信產業中，RTC 通常是指兩個端點之間具有最低延遲的即時媒體工作階段。這些工作階段可能與下列相關：

- 雙方之間的語音工作階段（例如電話系統、行動或 IP 語音 (VoIP)）
- 即時通訊（例如聊天和即時中繼聊天 (IRC)）
- 即時視訊工作階段（例如視訊會議和遠端臨場）

上述每個解決方案都有一些共同的元件（例如提供身分驗證、授權和存取控制、轉碼、緩衝和轉送等的元件），以及一些傳輸媒體類型（例如廣播服務、訊息伺服器和佇列等）特有的元件。本節著重於定義以語音和視訊為基礎的 RTC 系統和所有相關元件，如下圖所示。



RTC 的必要架構元件

軟開關/PBX

軟體切換或 PBX 是語音電話系統的大腦，並提供智慧，以使用不同的元件在企業內外建立、維護和轉接語音通話。企業的所有訂閱者都必須向 softswitch 註冊，才能接收或撥打電話。軟體切換的重要功能是追蹤每個訂閱者，以及如何使用語音網路中的其他元件來聯絡他們。

工作階段邊界控制器 (SBC)

工作階段邊界控制器 (SBC) 位於語音網路的邊緣，並追蹤所有傳入和傳出流量 （ 控制和資料平面 ）。SBC 的其中一個關鍵責任是保護語音系統免受惡意使用。SBC 可用來與工作階段啟動通訊協定 (SIP) 中繼線互連，以進行外部連線。有些 SBCs也提供轉碼功能，可將 CODECs 從一種格式轉換為另一種格式。大多數 SBCs也提供網路地址轉譯 (NAT) 周遊功能，即使在防火牆網路之間也有助於確保通話的建立。

PSTN 連線

IP 語音 (VoIP) 解決方案使用公有切換電話網路 (PSTN) 閘道和 SIP 幹線來連接舊版 PSTN 網路。

PSTN 閘道

PSTN 閘道會使用 CODEC 轉碼，在 SIP 和 SS7 之間轉換訊號，並在即時傳輸通訊協定 (RTP) 和分時多工 (TDM) 之間轉換媒體。PSTN 閘道一律位於靠近 PSTN 網路的邊緣。

SIP 中繼線

在 SIP 中繼線中，企業不會結束對 TDM (SS7 型 ） 網路的呼叫，而是透過 IP 保留企業和電信之間的流程。大多數 SIP 中繼線都是使用 SBCs。企業必須同意電信業預先定義的安全規則，例如允許特定範圍的 IP 地址、連接埠等。

媒體閘道 （ 轉換器 ）

使用者使用音訊和/或視訊以及選用的資料和其他資訊進行即時通訊。若要進行通訊，這兩個裝置必須能夠針對每個媒體軌道達成共識，以便成功通訊和呈現共用媒體。所有與 WebRTC 相容的瀏覽器都必須支援線上定位使用者支援 (OPUS) 和 G711，適用於音訊、VP8 和 H.264 影片的受限基準描述檔。

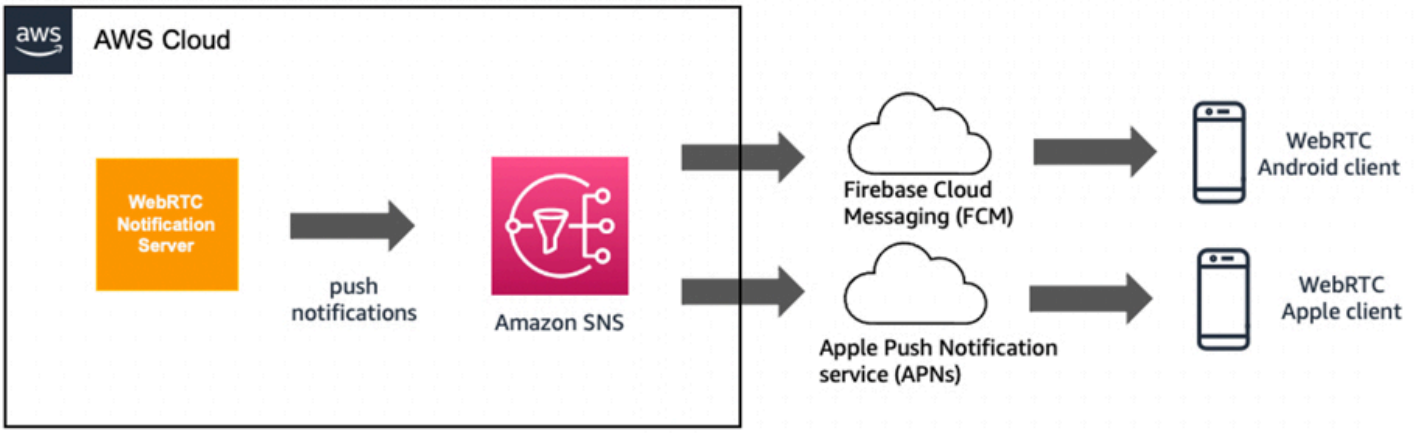
WebRTC 生態系統外的典型語音解決方案允許各種類型的 CODECs。一些常見的 CODECs 是 G.711 μ-law for North America、G.711 A-law、G.729 和 G.722。當使用兩個不同 CODECs 的兩個裝置彼此

通訊時，媒體閘道會轉換裝置之間的 CODEC 流程。換句話說，媒體閘道會處理媒體，並確保終端裝置能夠彼此通訊。

WebRTC 中的推送通知

WebRTC 實作在行動裝置上非常常見。與 Web 瀏覽器不同，行動裝置無法長時間保持 Websocket 連線開啟。因此，它需要依賴來自 WebRTC 伺服器的推送通知來處理所有結束請求，例如呼叫和訊息。

Amazon Simple Notification Service (Amazon SNS) 可讓您將推送通知傳送至行動裝置上的應用程式。這些應用程式可能會在 Apple iOS 或 Android 等各種作業系統上執行。下圖顯示推送通知流程的高階概觀，從 WebRTC 通知伺服器到 WebRTC 行動端點。

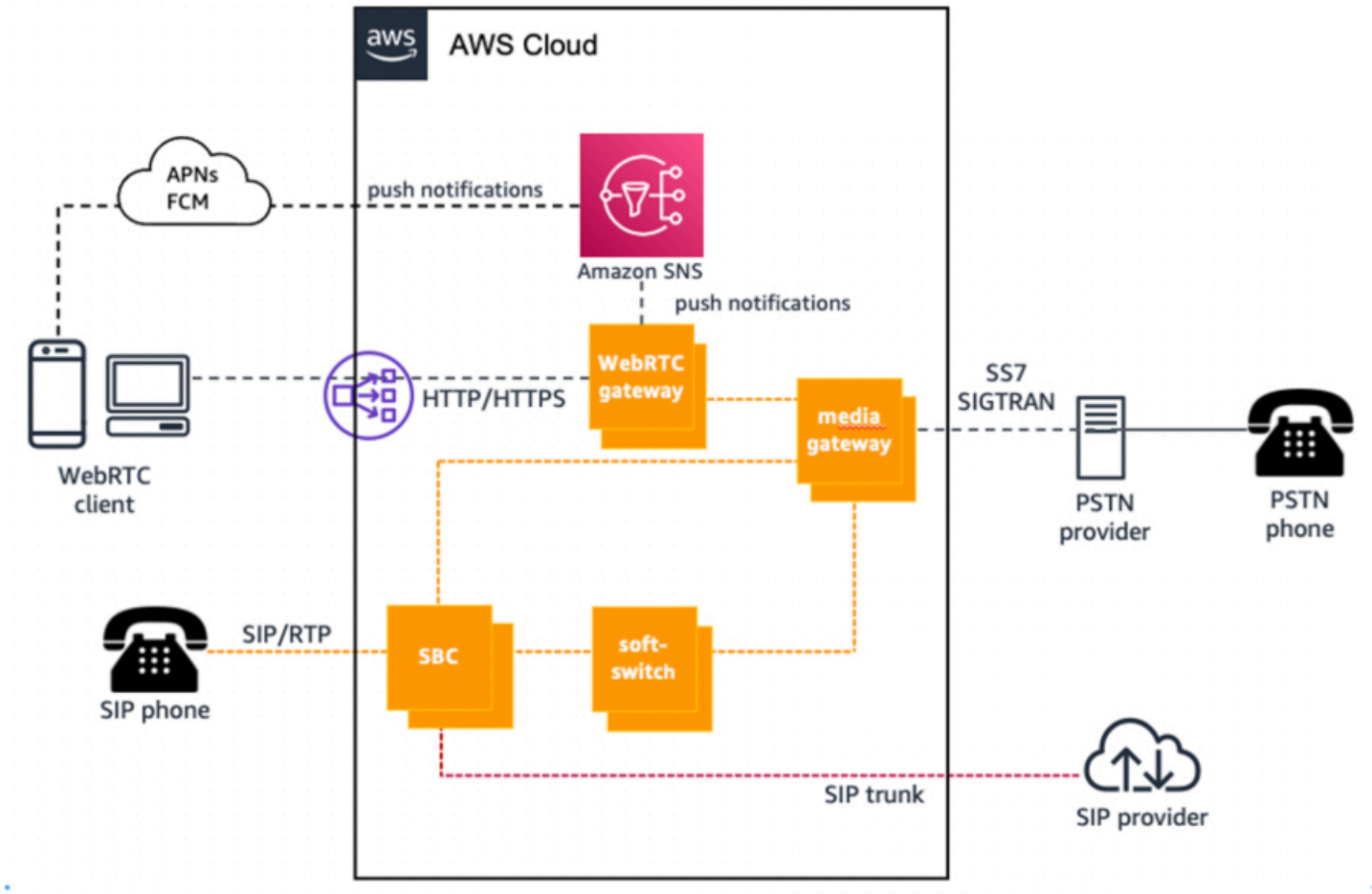


適用於推送通知的 Amazon SNS

WebRTC 和 WebRTC 閘道

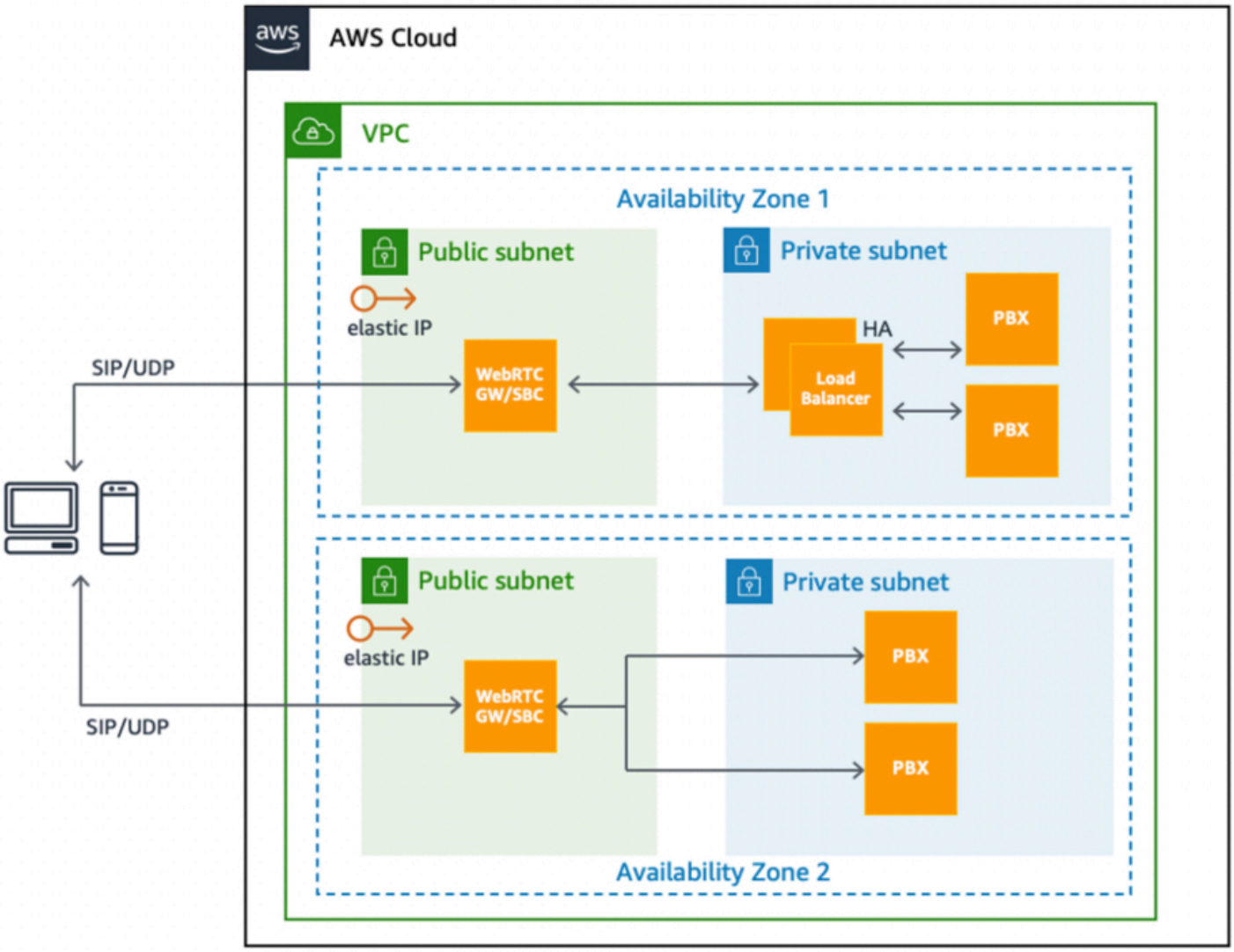
Web 即時通訊 (WebRTC) 可讓您從 Web 瀏覽器建立呼叫，或使用 API 從後端伺服器請求資源。該技術以雲端技術為設計考量，因此提供各種 APIs可用於建立呼叫。由於並非所有語音解決方案 （ 包括 SIP) 都支援這些 APIs，因此需要 WebRTC 閘道才能將 API 呼叫轉譯為 SIP 訊息，反之亦然。

下圖顯示高可用性 WebRTC 架構的設計模式。來自 WebRTC 用戶端的傳入流量由 Application Load Balancer (ALB) 平衡，而 WebRTC 會在屬於 Amazon EC2 Auto Scaling 群組的 Amazon Elastic Compute Cloud (Amazon EC2) 執行個體上執行。 Amazon EC2 Auto Scaling



語音 RTC 系統的基本拓撲

SIP 和 RTP 流量的另一個設計模式是跨可用區域以主動被動模式使用 Amazon EC2 上的 SBCs 對，如下圖所示。在這裡，彈性 IP 地址可在失敗時動態在執行個體之間移動，其中無法使用網域名稱服務 (DNS)。



在虛擬私有雲端 (VPC) 中使用 Amazon EC2 的 RTC 架構

上的高可用性和可擴展性 AWS

大多數即時通訊供應商都符合服務水準，可提供從 99.9% 到 99.999% 的可用性。根據您想要的高可用性 (HA) 程度，您必須在應用程式的完整生命週期中採取越來越複雜的措施。AWS 建議遵循這些準則，以實現強大程度的高可用性：

- 將系統設計為沒有單一故障點。使用無狀態和具狀態元件的自動化監控、故障偵測和容錯移轉機制
  - 單一故障點 (SPOF) 通常使用 N+1 或 2N 備援組態來消除，其中 N+1 是透過作用中節點之間的負載平衡來達成，而 2N 則是透過作用中待命組態中的一對節點來達成。
  - AWS 有數種方法可透過這兩種方法達到 HA，例如透過可擴展、負載平衡的叢集或假設作用中待命對。
- 正確檢測和測試系統的可用性。
- 準備手動機制的操作程序，以回應、緩解和復原失敗。

本節重點介紹如何使用 提供的功能來達成單一故障點 AWS。具體而言，本節描述了核心 AWS 功能和設計模式的子集，可讓您建置高可用性的即時通訊應用程式。

作用中待命狀態伺服器之間 HA 的浮動 IP 模式

浮動 IP 設計模式是一種眾所周知的機制，可在作用中和待命硬體節點對（媒體伺服器）之間實現自動容錯移轉。靜態次要虛擬 IP 地址會指派給作用中節點。作用中節點和待命節點之間的持續監控會偵測失敗。如果作用中節點失敗，監控指令碼會將虛擬 IP 指派給就緒待命節點，而待命節點會接管主要作用中函數。如此一來，虛擬 IP 就會在作用中節點和待命節點之間浮動。

RTC 解決方案的適用性

服務中不一定會有相同元件的多個作用中執行個體，例如 N 節點的作用中叢集。作用中待命組態為 HA 提供最佳機制。例如，RTC 解決方案中的具狀態元件，例如媒體伺服器或會議伺服器，甚至是 SBC 或資料庫伺服器，都非常適合進行作用中待命設定。SBC 或媒體伺服器在指定時間有數個長時間執行的工作階段或通道處於作用中狀態，而且在 SBC 作用中執行個體失敗的情況下，端點可以重新連線至待命節點，而不會因為浮動 IP 而進行任何用戶端組態。

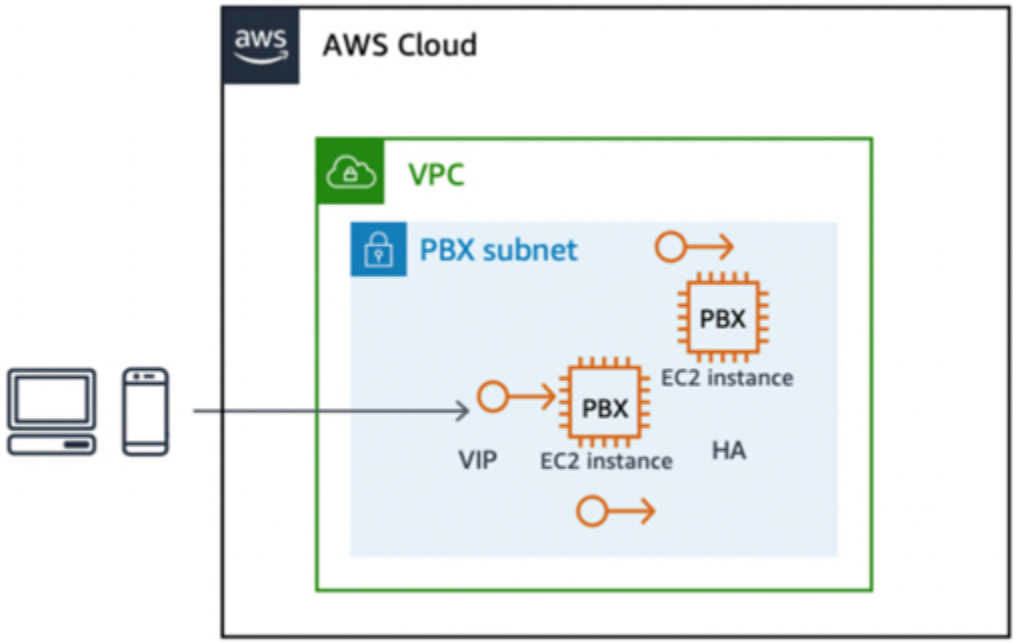
實作於 AWS

您可以使用 Amazon Elastic Compute Cloud (Amazon EC2)、Amazon EC2 API、彈性 IP 地址和 Amazon EC2 上次要私有 IP 地址支援的核心功能，在 AWS 上實作此模式。

若要在 上實作浮動 IP 模式 AWS：

1. 啟動兩個 EC2 執行個體以擔任主要節點和次要節點的角色，其中主要節點預設為作用中狀態。
2. 將額外的次要私有 IP 地址指派給主要 EC2 執行個體。
3. 與虛擬 IP (VIP) 類似的彈性 IP 地址會與次要私有地址相關聯。此次要私有地址是外部端點用來存取應用程式的地址。
4. 需要一些作業系統 (OS) 組態，才能將次要 IP 地址新增為主要網路介面的別名。
5. 應用程式必須繫結至此彈性 IP 地址。如果是星號軟體，您可以透過進階星號 SIP 設定來設定繫結。
6. 在每個節點上執行監控指令碼：自訂、Linux 上的 KeepAlive、Corosync 等，以監控對等節點的狀態。如果目前作用中節點失敗，對等會偵測到此失敗，並叫用 Amazon EC2 API 將次要私有 IP 地址重新指派給自己。

   因此，正在接聽與次要私有 IP 地址相關聯的 VIP 的應用程式可透過待命節點供端點使用。



使用彈性 IP 地址在具狀態 EC2 執行個體之間進行容錯移轉

優勢

此方法是一種可靠的低預算解決方案，可防止 EC2 執行個體、基礎設施或應用程式層級的故障。

限制和可擴展性

此設計模式通常限制在單一可用區域內。它可以跨兩個可用區域實作，但具有變化。在此情況下，浮動彈性 IP 地址會透過可用的重新關聯彈性 IP 地址 API，在不同可用區域中的作用中和待命節點之間重新關聯。在上圖所示的容錯移轉實作中，會捨棄進行中的呼叫，且端點必須重新連線。也可以使用基礎工作階段資料的複寫來擴展此實作，以提供工作階段或媒體連續性的無縫容錯移轉。

使用 WebRTC 和 SIP 進行可擴展性和 HA 的負載平衡

根據預先定義的規則平衡作用中執行個體叢集的負載，例如循環配置、親和性或延遲等，是廣受 HTTP 請求無狀態本質歡迎的設計模式。事實上，如果有許多 RTC 應用程式元件，負載平衡是可行的選項。

負載平衡器可做為對所需應用程式請求的反向代理或進入點，其本身設定為同時在多個作用中節點中執行。在任何指定時間點，負載平衡器會將使用者請求導向至已定義叢集中的其中一個作用中節點。負載平衡器會針對其目標叢集中的節點執行運作狀態檢查，而且不會將傳入請求傳送至未通過運作狀態檢查的節點。因此，負載平衡可實現高可用性的基本程度。此外，由於負載平衡器會以每秒的間隔對所有叢集節點執行主動和被動運作狀態檢查，因此容錯移轉的時間幾乎是瞬間的。

根據負載平衡器中定義的系統規則，決定要引導哪個節點，包括：

- 循環配置
- 工作階段或 IP 親和性，可確保工作階段內或來自相同 IP 的多個請求傳送至叢集中的相同節點
- 以延遲為基礎的
- 負載型

RTC 架構的適用性

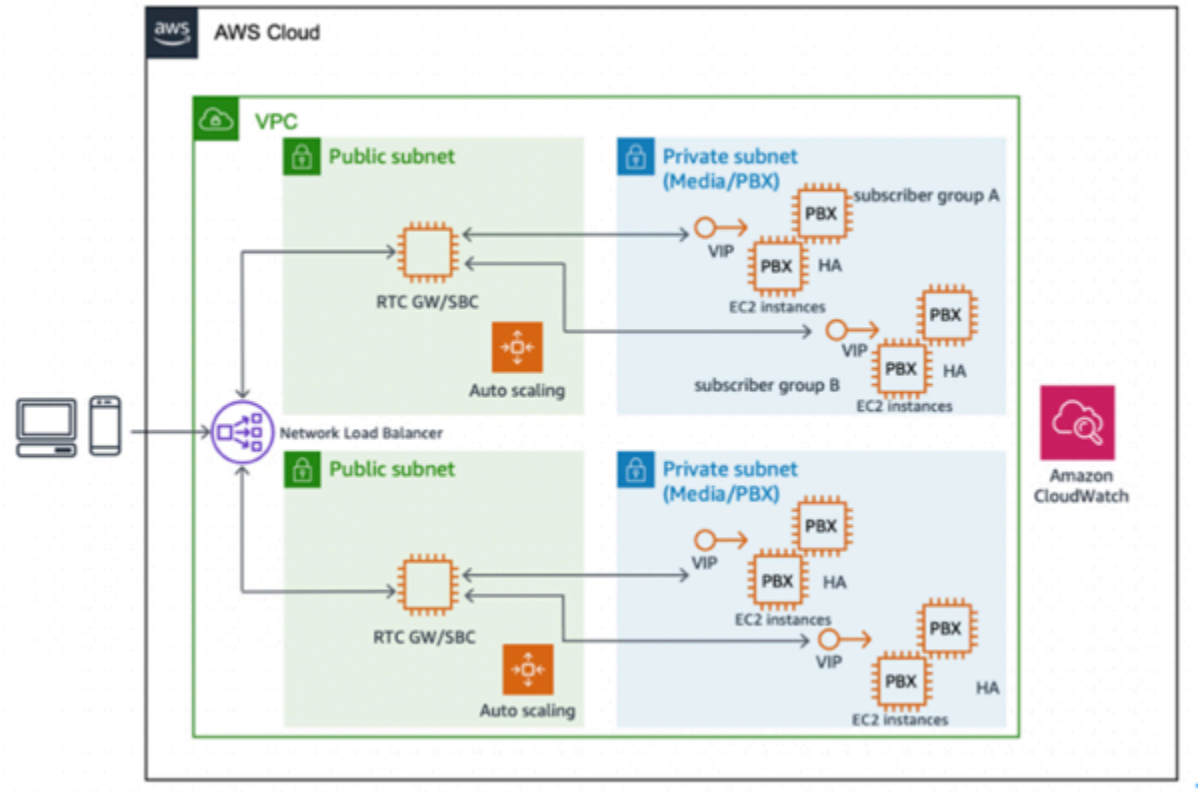
WebRTC 通訊協定可讓 WebRTC Gateways 透過以 HTTP 為基礎的負載平衡器輕鬆平衡負載，例如 Elastic Load Balancing (ELB)、Application Load Balancer (ALB) 或 Network Load Balancer (NLB)。由於大多數 SIP 實作都依賴傳輸控制通訊協定 (TCP) 和使用者資料包通訊協定 (UDP) 的傳輸，因此您需要網路或連線層級負載平衡，同時支援 TCP 和 UDP 型流量。

使用 Application Load Balancer 和 Auto Scaling 在 上 AWS 針對 WebRTC 進行負載平衡 Application Load Balancer

在以 WebRTC 為基礎的通訊中，Elastic Load Balancing 提供全受管、高可用性和可擴展的負載平衡器，做為請求的進入點，然後導向至與 Elastic Load Balancing 相關聯的 EC2 執行個體目標叢集。由於 WebRTC 請求是無狀態的，因此您可以使用 Amazon EC2 Auto Scaling 來提供全自動化且可控制的可擴展性、彈性和高可用性。

Application Load Balancer 提供全受管負載平衡服務，可使用多個可用區域進行高可用性和可擴展性。這支援 WebSocket 請求的負載平衡，這些請求會處理 WebRTC 應用程式的訊號，以及使用長時間執行的 TCP 連線在用戶端和伺服器之間進行雙向通訊。Application Load Balancer 也支援內容型路由和黏性工作階段，使用負載平衡器產生的 Cookie，將請求從相同用戶端路由至相同的目標。如果您啟用黏性工作階段，相同的目標會收到請求，並且可以使用 Cookie 來復原工作階段內容。

下圖顯示目標拓撲。



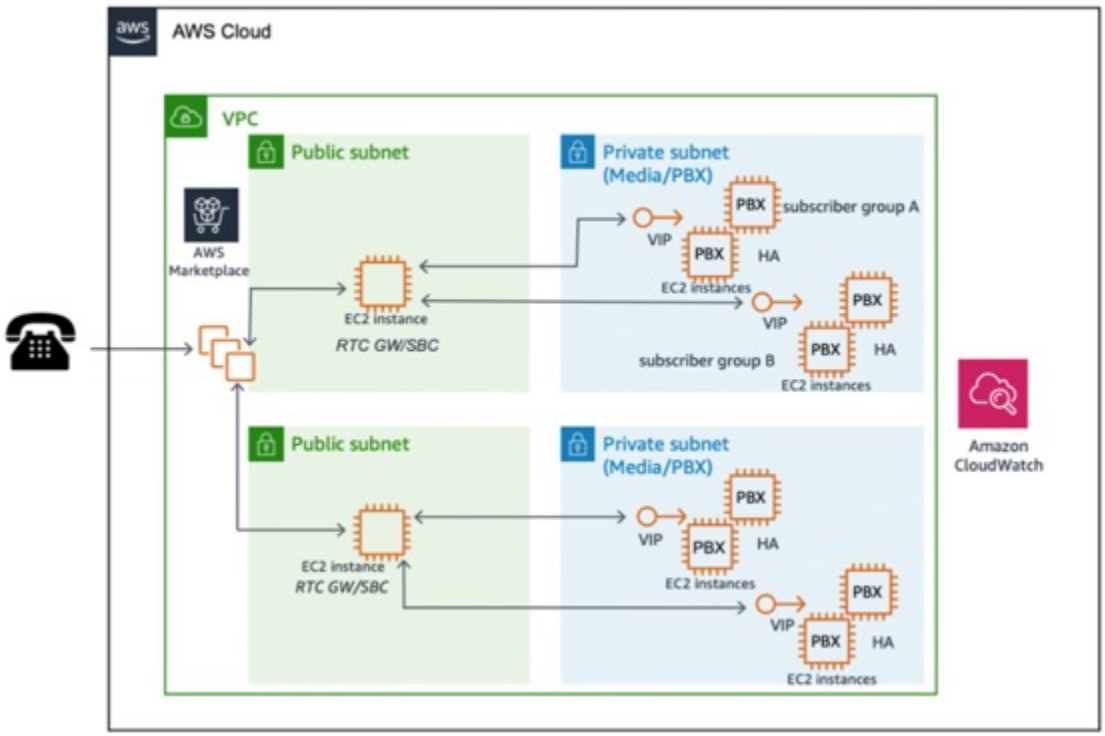
WebRTC 可擴展性和高可用性架構

使用 Network Load Balancer 或 AWS Marketplace 產品的 SIP 實作

在以 SIP 為基礎的通訊中，連線是透過 TCP 或 UDP 進行，大多數 RTC 應用程式使用 UDP。如果 SIP/TCP 是選擇的訊號通訊協定，則可以使用 Network Load Balancer 進行完全受管、高可用性、可擴展性和效能負載平衡。

Network Load Balancer 會在連線層級 （第四層） 運作，根據 IP 通訊協定資料將連線路由至目標，例如 Amazon EC2 執行個體、容器和 IP 地址。網路負載平衡非常適合 TCP 或 UDP 流量負載平衡，能夠每秒處理數百萬個請求，同時保持超低延遲。它與其他熱門的 AWS 服務整合，例如 Amazon EC2 Auto Scaling、Amazon Elastic Container Service (Amazon ECS)、Amazon Elastic Kubernetes Service (Amazon EKS) 和 AWS CloudFormation。

如果啟動 SIP 連線，另一個選項是使用AWS Marketplace商用off-the-shelf(COTS)。 AWS Marketplace 提供多種產品，可處理 UDP 和其他類型的第四層連線負載平衡。COTS 通常包括對高可用性的支援，並通常與 Amazon EC2 Auto Scaling 等功能整合，以進一步增強可用性和可擴展性。下圖顯示目標拓撲：



使用 AWS Marketplace 產品的 SIP 型 RTC 可擴展性

跨區域 DNS 型負載平衡和容錯移轉

Amazon Route 53 提供全域 DNS 服務，可做為公有或私有端點，供 RTC 用戶端註冊和與媒體應用程式連線。使用 Amazon Route 53，DNS 運作狀態檢查可設定為將流量路由至運作狀態良好的端點，或獨立監控應用程式的運作狀態。

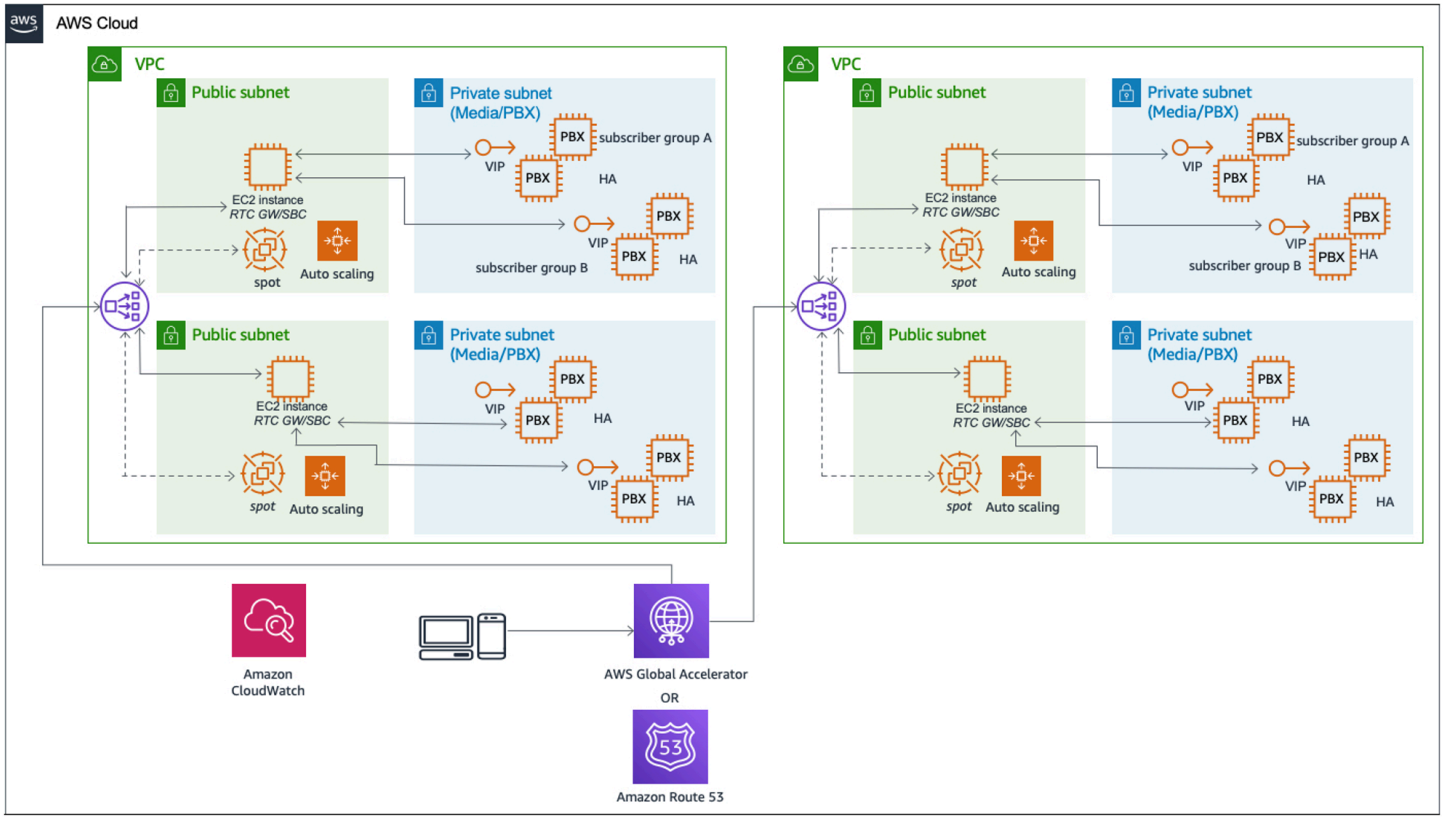
Amazon Route 53 流量流程功能可讓您透過各種路由類型輕鬆管理全域流量，包括以延遲為基礎的路由、地理 DNS、地理鄰近性和加權圓形配置，所有這些都可以與 DNS 容錯移轉結合，以啟用各種低延遲、容錯的架構。Amazon Route 53 Traffic Flow 簡易視覺化編輯器可讓您管理最終使用者路由到應用程式端點的方式，無論是在單一 AWS 區域中還是分佈於全球。

在全域部署的情況下，Route 53 中以延遲為基礎的路由政策特別有用，可將客戶引導至媒體伺服器最接近的存在點，以改善與即時媒體交換相關聯的服務品質。

請注意，若要強制執行容錯移轉至新的 DNS 地址，必須排清用戶端快取。此外，DNS 變更在散佈到全域 DNS 伺服器時，可能會有延遲。您可以使用存留時間屬性來管理 DNS 查詢的重新整理間隔。此屬性可在設定 DNS 政策時設定。

為了快速聯絡全球使用者或滿足使用單一公有 IP 的需求， AWS Global Accelerator 也可以用於跨區域容錯移轉。 AWS Global Accelerator 是一種網路服務，可改善具有本機和全域 reach 的應用程式可用性和效能。 AWS Global Accelerator 提供靜態 IP 地址，可做為應用程式端點的固定進入點，例如在單一或多個 AWS 區域中的 Application Load Balancer、Network Load Balancer 或 Amazon EC2 執行個體。它使用 AWS 全域網路來最佳化從使用者到應用程式的路徑，進而改善效能，例如 TCP 和 UDP 流量的延遲。

AWS Global Accelerator 會持續監控應用程式端點的運作狀態，並在目前端點運作狀態不佳時，自動將流量重新導向至最近的運作狀態良好端點。如需其他安全需求，加速Site-to-Site VPN 會使用 AWS Global Accelerator 透過 AWS Global Network 和 AWS 節點智慧路由流量，來改善 VPN 連線的效能。



使用 AWS Global Accelerator 或 Amazon Route 53 的跨區域高可用性設計

資料耐久性和具有持久性儲存體的 HA

大多數 RTC 應用程式依賴持久性儲存來存放和存取資料，以進行身分驗證、授權、會計（工作階段資料、通話詳細資訊記錄等）、操作監控和記錄。在傳統資料中心，確保持久性儲存元件（資料庫、檔

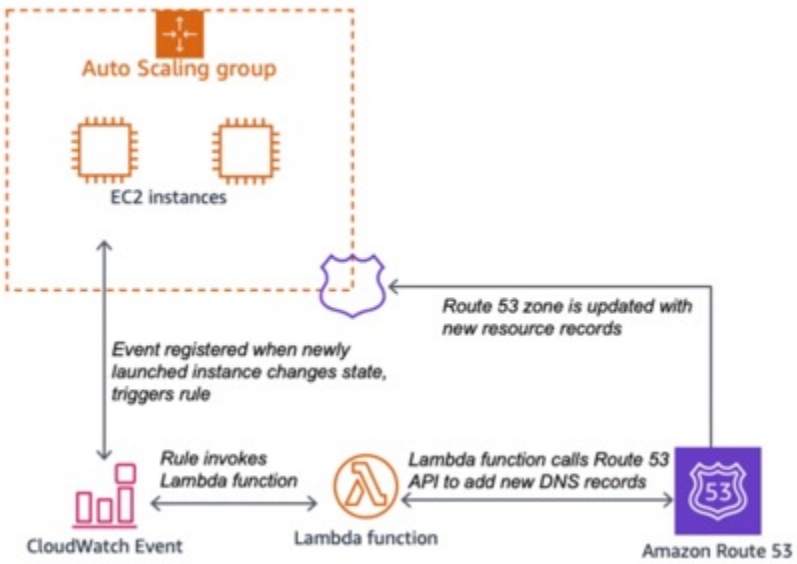
案系統等）的高可用性和耐用性通常需要透過設定儲存區域網路 (SAN)、獨立磁碟冗餘陣列 (RAID) 設計和備份、還原和容錯移轉處理程序進行繁重的提升。可 AWS 雲端 大幅簡化和增強資料耐久性和可用性的傳統資料中心實務。

對於物件儲存和檔案儲存，Amazon Simple Storage Service (Amazon S3) 和 Amazon Elastic File System (Amazon EFS) 等 AWS 服務可提供受管高可用性和可擴展性。Amazon S3 的資料耐久性為 99.999999999% (11 個九）。

對於交易資料儲存，客戶可以選擇利用全受管 Amazon Relational Database Service (Amazon RDS)，該服務支援具有高可用性部署的 Amazon Aurora、PostgreSQL、MySQL、MariaDB、Oracle 和 Microsoft SQL Server。對於註冊商函數、訂閱者設定檔或會計記錄儲存（例如 CDRs)，Amazon RDS 提供容錯、高度可用和可擴展的選項。

使用 AWS Lambda、Amazon Route 53 和 Amazon EC2 Auto Scaling 進行動態擴展

AWS 允許功能的鏈結，以及能夠根據基礎設施事件將自訂無伺服器函數整合為服務。在 RTC 應用程式中具有多種用途的這類設計模式之一，是自動擴展生命週期勾點與 Amazon CloudWatch Events、Amazon Route 53 和 AWS Lambda function 的組合。 AWS Lambda 函數可以嵌入任何動作或邏輯。下圖示範這些功能如何透過自動化來強化系統可靠性和可擴展性。



透過 Amazon Route 53 的動態更新自動擴展

搭配 Amazon Kinesis Video Streams 的高可用性 WebRTC

Amazon Kinesis Video Streams 透過 WebRTC 提供即時媒體串流，允許使用者擷取、處理和儲存媒體串流，以進行播放、分析和機器學習。這些串流高度可用、可擴展且符合 WebRTC 標準。Amazon Kinesis Video Streams 包含 WebRTC 訊號端點，可用於快速對等探索和建立安全連線。它包括 NAT 的受管工作階段周遊公用程式 (STUN) 和使用 NAT 周圍的轉送 (TURN) 端點的周遊，以在對等之間即時交換媒體。它也包含免費的開放原始碼 SDK，可直接與攝影機韌體整合，以啟用與 Amazon Kinesis Video Streams 端點的安全通訊，進而允許對等探索和媒體串流。最後，它提供適用於 Android、iOS 和 JavaScript 的用戶端程式庫，允許符合 WebRTC 規範的行動和 Web 播放器安全地探索並與攝影機裝置連線，以進行媒體串流和雙向通訊。

搭配 Amazon Chime Voice Connector 的高可用性 SIP 中繼

Amazon Chime Voice Connector pay-as-you-go的 SIP 轉接服務，讓公司能夠使用其電話系統撥打和/或接聽安全且便宜的電話。Amazon Chime Voice Connector 是服務供應商 SIP 中繼線或整合式服務數位網路 (ISDN) 主要費率界面 (PRIs的低成本替代方案。客戶可以選擇啟用傳入呼叫、傳出呼叫或兩者。

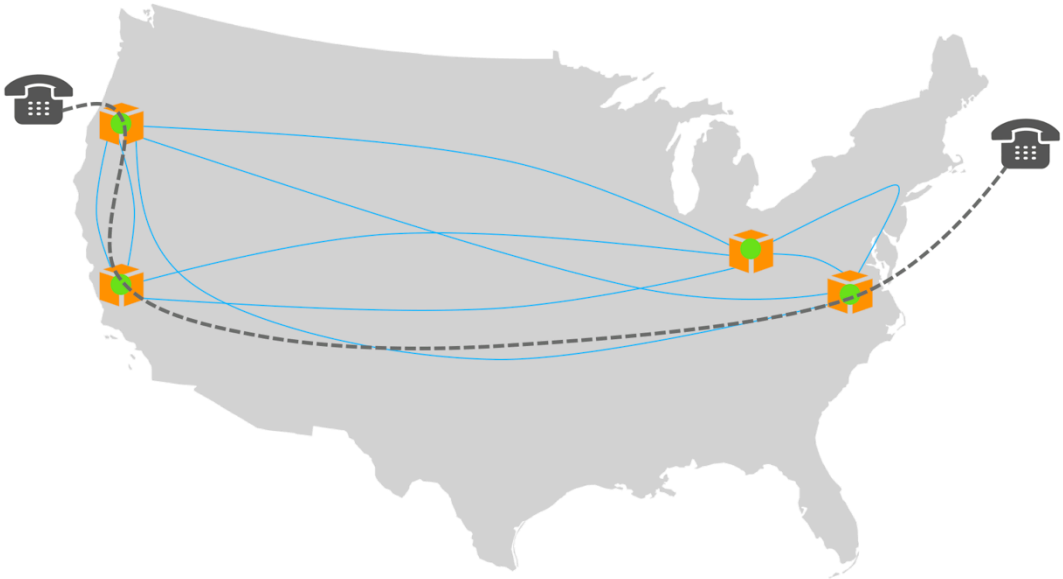
服務使用 AWS 網路跨多個 提供高可用性的通話體驗 AWS 區域。您可以從 SIP 轉接電話通話串流音訊，或將 SIP 型媒體錄製 (SIPREC) 饋送至 Amazon Kinesis Video Streams，以即時從業務通話中獲得洞見。您可以透過與 Amazon Transcribe 和其他常見機器學習程式庫整合，快速建置音訊分析的應用程式。

欄位的最佳實務

本節摘要說明一些執行大型即時工作階段啟動協定 (SIP) 工作負載的最大和最成功的 AWS 客戶所實作的最佳實務。想要在公有雲端中執行自己的 SIP 基礎設施 AWS 的客戶會發現這些最佳實務很有價值，因為它們有助於在發生不同類型的故障時提高系統的可靠性和彈性。雖然其中一些最佳實務是 SIP 特有的，但其中大多數適用於在 上執行的任何即時通訊應用程式 AWS。

建立 SIP 浮水印

AWS 具有強大、可擴展且備援的網路骨幹，可提供不同網路之間的連線 AWS 區域。當光纖切割等網路事件降級 AWS 骨幹連結時，流量會使用網路層級路由通訊協定，例如邊界閘道通訊協定 (BGP)，快速容錯移轉到備援路徑。此網路層級流量工程對客戶而言是黑色方塊 AWS ，大多數甚至不會注意到這些容錯移轉事件。不過，執行語音、高品質影片和低延遲訊息等即時工作負載的客戶，有時會注意到這些事件。那麼， AWS 客戶如何實作自己的流量工程，而不是網路層級 AWS 所提供的流量工程？ 解決方案正在部署許多不同的 SIP 基礎設施 AWS 區域。做為呼叫控制功能的一部分，SIP 也提供透過特定 SIP 代理路由呼叫的功能。



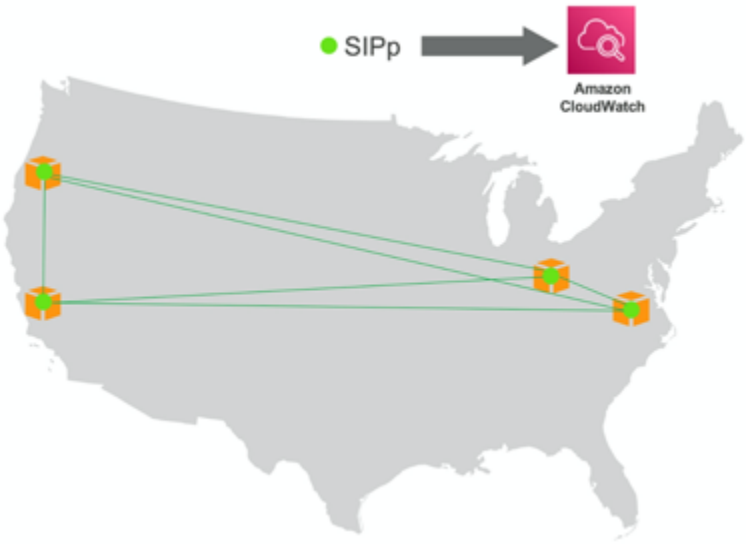
使用 SIP 路由覆寫網路路由

在上圖中，SIP 基礎設施 （ 由立方體內的綠點表示 ） 會在所有四個美國區域執行。實心藍線代表 AWS 骨幹的虛構描述。如果未實作 SIP 路由，則源自美國西部海岸且目的地為美國東部海岸的呼叫會經過直接連接奧勒岡和維吉尼亞區域的骨架連結。圖表顯示客戶如何覆寫網路層級路由，並在奧勒岡和維吉

尼亞之間使用 SIP 路由透過加州路由進行相同的呼叫。這種類型的 SIP 流量工程可以使用 SIP 代理和媒體閘道，根據 SIP 重新傳輸和客戶特定業務偏好設定等網路指標來實作。

執行詳細監控

即時語音和視訊應用程式的最終使用者預期效能會與傳統電話服務達到相同水準。因此，當他們遇到應用程式的問題時，最終會損害供應商的聲譽。若要主動而非被動，必須在為最終使用者提供服務的系統的每個部分部署詳細的監控。



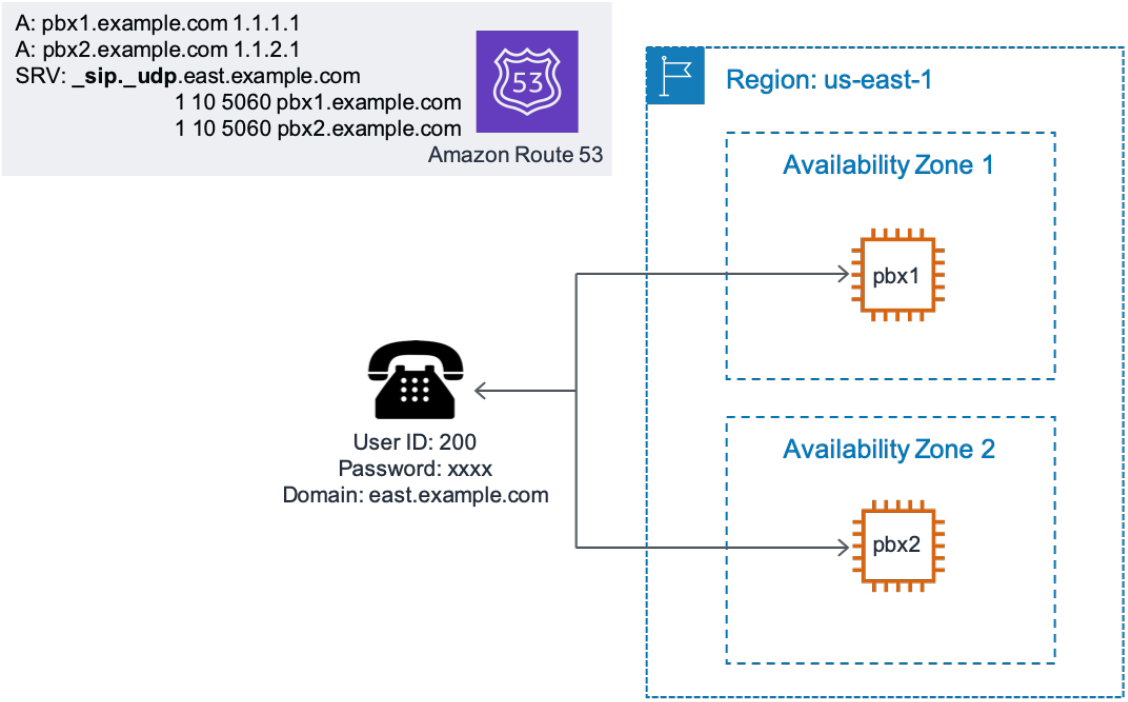
使用 SIPp 監控 VoIP 基礎設施

許多開放原始碼工具，例如 iPerf 或 SIPp，以及 VOIPMonitor，可用於監控 SIP/RTP 流量。在上述範例中，在用戶端和伺服器模式中執行 SIP 的節點正在測量 SIP 指標，例如成功通話和全部四個美國之間的 SIP 重新傳輸 AWS 區域。然後，您可以使用自訂指令碼將這些指標匯出至 Amazon CloudWatch。使用 CloudWatch，客戶可以根據特定閾值在這些自訂指標上建立警示。然後，您可以根據這些 CloudWatch 警示的狀態採取自動或手動修補動作。

對於不想配置開發和維護自訂監控系統所需的工程資源的客戶，市場上有許多良好的 VoIP 監控解決方案可用，例如 ThousandEyes。修復動作的範例是根據增加的 SIP 重新傳輸來變更 SIP 路由。

使用 DNS 進行負載平衡，並將 IPs 浮動進行容錯移轉

支援 DNS SRV 功能的 IP 電話用戶端可以透過將用戶端負載平衡到不同的 SBCs/PBXs，有效率地使用基礎設施中內建的備援。



使用 DNS SRV 記錄來載入平衡 SIP 用戶端

上圖顯示客戶如何使用 SRV 記錄來平衡 SIP 流量。支援 SRV 標準的任何 IP 電話用戶端都會在 SRV 類型 DNS 記錄中尋找 sip. <transport protocol>字首。在此範例中，來自 DNS 的答案區段包含兩個 PBXs兩者都在不同的 AWS 可用區域中執行。不過，除了端點 URIs之外，SRV 記錄還包含三個額外的資訊：

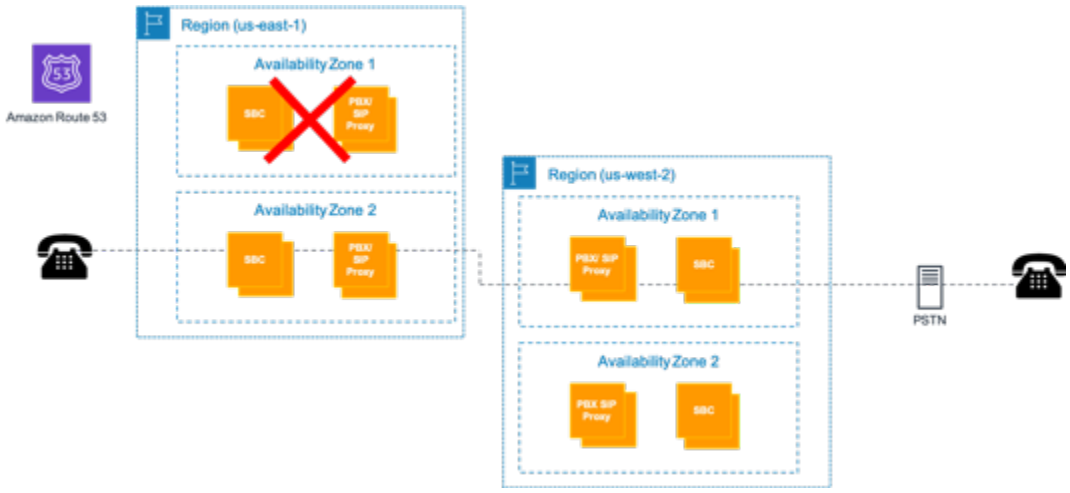
- 第一個數字是優先順序 （ 上述範例中為 1)。優先順序較低優先於較高。
- 第二個數字是 權重 （ 上述範例中為 10)。
- 第三個數字是要使用的連接埠 (5060)。

由於兩個 PBXs伺服器的優先順序相同 (1)，因此用戶端會使用權重來平衡兩個 PBXs之間的負載。在這種情況下，由於權重相同，因此 SIP 流量應該在兩個 PBXs 之間平均地平衡負載。

DNS 可以是用戶端負載平衡的好解決方案，但透過變更/更新 DNS 'A' 記錄來實作容錯移轉呢？ 由於在用戶端和中繼節點內的 DNS 快取行為中發現不一致，因此不建議使用此方法。SIP 節點叢集之間可用區域內容錯移轉的更好方法是使用 EC2 IP 重新指派，其中受損主機的 IP 地址會立即使用 EC2 API 重新指派給運作狀態良好的主機。搭配詳細的監控和運作狀態檢查解決方案，失敗節點的 IP 重新指派可確保流量及時移至運作狀態良好的主機，將最終使用者中斷降至最低。

使用多個可用區域

每個 AWS 區域 都會細分為不同的可用區域。每個可用區域都有自己的電源、冷卻和網路連線，因此形成隔離的故障網域。在 的建構中 AWS，我們鼓勵客戶在多個可用區域中執行工作負載。這可確保客戶應用程式甚至可以承受完整的可用區域故障 - 這是非常罕見的事件。此建議也適用於即時 SIP 基礎設施。



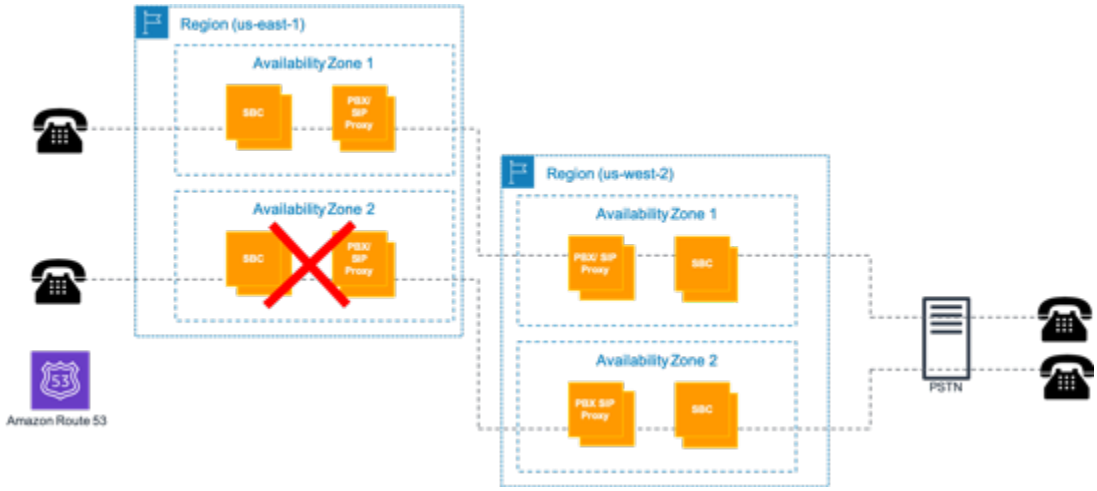
處理可用區域失敗

假設災難事件（例如類別 5 颶風）會導致 us-east-1 區域中的完整可用區域中斷。基礎設施如圖表所示執行時，所有最初向故障可用區域中的節點註冊的 SIP 用戶端，都應該向在可用區域 #2 中執行的 SIP 節點重新註冊。（使用 SIP 用戶端/電話測試此行為，以確保支援此行為。）雖然在可用區域中斷時作用中的 SIP 呼叫會遺失，但任何新的呼叫都會透過可用區域 2 路由。

總而言之，DNS SRV 記錄應將用戶端指向多個「A」記錄，每個可用區域中各有一個記錄。每個「A」記錄應反過來指向該可用區域中 SBCs/PBXs的多個 IP 地址，同時提供可用區域內和跨可用區域彈性。如果 IP 是公有的，則可以使用 IPs 重新指派來實作內部和跨可用區域容錯移轉。不過，私有 IPs 無法跨可用區域重新指派。如果客戶使用私有 IP 定址，則必須依賴使用備份 SBC/PBX 重新註冊的 SIP 用戶端進行可用區域間容錯移轉。

將流量保持在一個可用區域內，並使用 EC2 置放群組

也稱為可用區域親和性，此最佳實務也適用於完全可用區域故障的罕見事件。建議您消除任何跨可用區域流量，讓任何進入一個可用區域的 SIP 或 RTP 流量都應該保留在該可用區域中，直到它離開該區域為止。



可用區域親和性（最多 50% 的作用中通話會遺失）

上圖顯示使用可用區域親和性的簡化架構。此方法的比較優勢會變得明確，因為有一個方法考量到完整的可用區域中斷的影響。如圖表所述，如果失去可用區域 2，則最多 50% 的作用中呼叫會受到影響（假設可用區域之間的負載平衡相等）。如果未實作可用區域親和性，則某些呼叫會在一個區域中的可用區域之間流動，而失敗很可能會影響超過 50% 的作用中呼叫。

為了將流量延遲降至最低，AWS 也建議您考慮在每個可用區域中使用 EC2 置放群組。在相同 EC2 置放群組中啟動的執行個體具有更高的頻寬和更低的延遲，因為 EC2 可確保這些執行個體的網路接近度彼此相對。

使用增強型聯網 EC2 執行個體類型

在 Amazon EC2 上選擇正確的執行個體類型可確保系統可靠性以及基礎設施的高效使用。EC2 提供廣泛的執行個體類型選擇，已針對不同的使用案例進行最佳化。執行個體類型包含 CPU、記憶體、儲存體和聯網功能的各種組合，供您靈活選擇適用於應用程式的適當資源組合。這些增強型聯網執行個體類型可確保在其上執行的 SIP 工作負載能夠存取一致的頻寬，並相對降低彙總延遲。Amazon EC2 最近新增了 Elastic Network Adapter (ENA)，可提供高達 100 Gbps 的頻寬。您可以在 EC2 執行個體類型頁面上找到 EC2 執行個體類型的最新目錄和相關功能。

對於大多數客戶而言，最新一代的 Compute Optimized 執行個體應該提供成本的最佳值。例如，C5N 支援頻寬高達 100 Gbps 的新彈性網路轉接器，每秒數百萬個封包 (PPS)。大多數即時應用程式也會受益於使用 Intel Data Plane 開發人員套件 (DPDK)，這可以大幅提高網路封包處理。

不過，最佳實務是根據您的需求對各種 EC2 執行個體類型進行基準測試，以查看哪種執行個體類型最適合您。基準測試也可讓您尋找其他組態參數，例如特定執行個體類型一次可處理的呼叫數量上限。

安全考量

RTC 應用程式元件通常直接在面向網際網路的 Amazon EC2 執行個體上執行。除了 TCP 之外，流程使用 UDP 和 SIP 等通訊協定。在這些情況下， 會 AWS Shield Standard 保護 Amazon EC2 執行個體免於常見的基礎設施層 （ 第 3 層和第 4 層 ） DDoS 攻擊，例如 UDP 反射攻擊、DNS 反射、NTP 反射、SSDP 反射等。 AWS Shield Standard 會使用各種技術，例如優先順序型流量調整，這些技術會在偵測到定義良好的 DDoS 攻擊簽章時自動接合。

AWS 也透過啟用彈性 IP 地址 AWS Shield Advanced ，為這些應用程式提供大型且複雜的 DDoS 攻擊進階保護。 AWS Shield Advanced 提供增強的 DDoS 偵測，可自動偵測 EC2 執行個體 AWS 的資源類型和大小，並套用適當的預先定義緩解措施，以防範 SYN 或 UDP 洪水。使用 AWS Shield Advanced，客戶也可以透過與 24 小時全年無休的 AWS DDoS 回應團隊 (DRT) 互動來建立自己的自訂緩解設定檔。 AWS Shield Advanced 也可確保在 DDoS 攻擊期間，您的所有 Amazon VPC 網路存取控制清單 (ACLs) 都會在 AWS 網路邊界自動強制執行，讓您存取額外的頻寬和清理容量，以緩解大型容量 DDoS 攻擊。

結論

即時通訊 (RTC) 工作負載可以部署在 上， AWS 以達到可擴展性、彈性和高可用性，同時符合金鑰需求。今天，有幾位客戶使用 AWS、其合作夥伴和開放原始碼解決方案，以更低的成本和更快的敏捷性執行 RTC 工作負載，並減少全球足跡。

本白皮書提供的參考架構和最佳實務，可協助客戶成功設定 上的 RTC 工作負載， AWS 並最佳化解決方案，以滿足最終使用者的需求，同時最佳化雲端。

縮寫

本文件中使用的縮寫包括：

ACL — 存取控制清單

ALB — Application Load Balancer

APNs — Apple 推送通知服務

BGP — 邊界閘道通訊協定

CDR — 通話詳細資訊記錄

COTS — off-the-shelf軟體

DDoS — denial-of-service

DNS — 網域名稱系統

DPDK — Intel Data Plane 開發人員套件

DRT — DDoS 回應團隊

ENA — 彈性網路轉接器

EPC – 不斷發展的封包核心

FCM — Firebase 雲端傳訊

HA — 高可用性

IRC — 網路中繼聊天

ISDN — 整合式服務數位網路

NAT — 網路地址轉譯

OPUS — 線上定位使用者支援

PBX — 私有分支交換

PRI — 主要速率介面

PSTN — 公有切換電話網路

RAID — 獨立磁碟的備援陣列

RTC — 即時通訊

RTP — 即時傳輸通訊協定

SAN — 儲存區域網路

SBC — 工作階段邊界控制器

SIP — 工作階段起始通訊協定

SPOF — 單點故障

SRV — 服務

SS7 — 訊號系統 n.7

STUN — NAT 的工作階段周遊公用程式

SYN — 同步

TCP — 傳輸控制通訊協定

TDM — 分時多工

TURN — 在 NAT 周圍使用轉送進行周遊

UDP — 使用者資料包通訊協定

URI — 統一資源識別符

VIP — 虛擬 IP

VNF — 虛擬網路函數

VoIP — IP 語音

VPC — 虛擬私有雲端

WebRTC — Web 即時通訊

貢獻者

下列個人和組織為本文件作出了貢獻：

- Mounir Chennana | Amazon Web Services 資深解決方案架構師
- Mohammed Al-Mehdar | Amazon Web Services 資深解決方案架構師
- Ejaz Sial | Amazon Web Services 資深解決方案架構師
- Ahmad Khan，Amazon Web Services 資深解決方案架構師
- Tipu Qureshi | Amazon Web Services AWS 支援首席工程師
- Hasan Khan，Amazon Web Services 資深技術客戶經理
- Shoma Chakravarty，WW 技術主管，電信，Amazon Web Services

文件修訂

若要收到此白皮書更新的通知，請訂閱 RSS 摘要。

變更	描述	日期
白皮書已更新	針對最新的服務和功能進行更新。	2022 年 5 月 5 日
白皮書已更新	針對最新的服務和功能進行更新。	2020 年 2 月 13 日
初次出版	白皮書已首次發佈。	2018 年 10 月 1 日

注意

客戶有責任對本文件中的資訊進行自己的獨立評定。本文件：(a) 僅供參考，(b) 代表目前的 AWS 產品產品和實務，這些產品和實務可能隨時變更，恕不另行通知，且 (c) 不會從 AWS 及其附屬公司、供應商或授權方建立任何承諾或保證。AWS 產品或服務以「原樣」方式提供，不提供任何類型的保證、陳述或條件，無論明示或暗示。AWS 對其客戶的責任與義務應由 AWS 協議管轄，本文並非 AWS 與其客戶之間的任何協議的一部分，也並非上述協議的修改。

© 2022 Amazon Web Services, Inc. 或其附屬公司。保留所有權利。

AWS 詞彙表

如需最新的 AWS 術語，請參閱 AWS 詞彙表 參考中的AWS 詞彙表。

本文為英文版的機器翻譯版本，如內容有任何歧義或不一致之處，概以英文版為準。