使用隨機分區進行工作負載隔離

Colm MacCárthaigh

使用隨機分區進行工作負載隔離

如今，Amazon Route 53 託管著眾多全球最大型企業和最受歡迎的網站，但其起步非常低調。

進行 DNS 託管

AWS 開始提供服務後不久，AWS 客戶明確表示，他們希望能夠在其網域的「根」上使用我們的 Amazon Simple Storage Service (S3)、Amazon CloudFront 和 Elastic Load Balancing 服務，即用於 "amazon.com" 之類的名稱，而不僅僅是 "www.amazon.com" 之類的名稱。

這似乎很簡單。然而，由於 DNS 通訊協定的設計決定要追溯到 20 世紀 80 年代，因此它比看上去要難。DNS 具有一項稱為 CNAME 的功能，可讓網域擁有者將其網域的一部分卸載至另一個託管服務供應商，但這在根級或頂層網域不起作用。為滿足客戶的需求，我們必須實際託管客戶的網域。我們在託管客戶的網域時，可以返回 Amazon S3、Amazon CloudFront 或 Elastic Load Balancing 任何目前的 IP 地址集。這些服務正在不斷擴展並新增 IP 地址，因此，客戶無法輕鬆地在其網域組態中進行硬編碼。

託管 DNS 並非易事。若 DNS 出現問題，整個企業都可能會離線。但在確定需求之後，我們便著手按照 Amazon 慣用的方式來解決——亟不可待。我們組建了一個小型工程師團隊，然後開始工作。

處理 DDOS 攻擊

若要詢問任何 DNS 供應商，其面臨的最大挑戰是什麼，他們會告訴您最大的挑戰是處理分散式拒絕服務 (DDOS) 攻擊。DNS 建基於 UDP 通訊協定，這意味著 DNS 請求在許多荒涼的網際網路上具有欺騙性。由於 DNS 也是關鍵的基礎架構，因此這種組合讓其成為以下人員的誘人目標：試圖勒索企業的不法分子、出於各種原因而設法觸發停機的「啟動人員」、以及偶爾誤入歧途的令人討厭的製造商，但他們似乎沒有意識到自己犯下了嚴重的罪行，以及個人實際造成的後果。無論是什麼原因，每天都有數千種針對網域的 DDOS 攻擊。

緩解這些攻擊的一種方法是使用大量的伺服器容量。雖然擁有良好的容量基準很重要，但這種方法並不能切實擴展。供應商新增的每台伺服器都需要花費數千美元，但如果攻擊者使用受感染的殭屍網絡，他們可極其廉價地增加更多假用戶端。對於供應商而言，新增大量伺服器容量是一種失敗的策略。

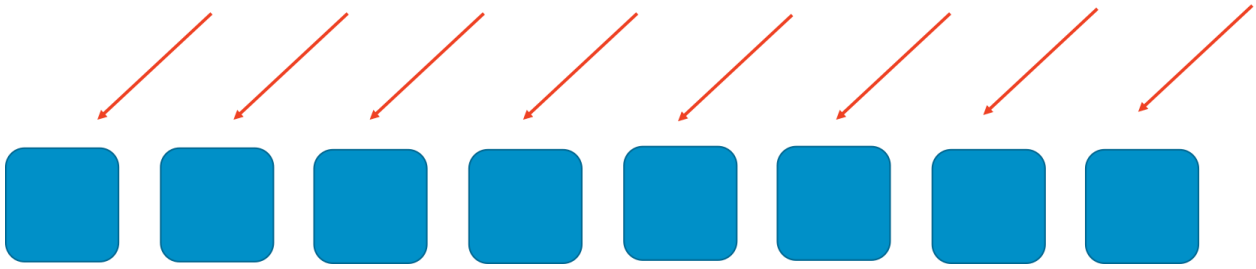
我們在建置 Amazon Route 53 時，DNS 防禦的最新技術是專用網路設備，可以利用各種技巧以極高的速率「清理」流量。我們在 Amazon 上有許多此類設備用於我們現有的內部 DNS 服務，我

們還與硬體供應商討論了其他可用工具。我們發現，購買足夠的設備來完全覆蓋每個 Route 53 網域將花費數千萬美元，並且讓我們的排程增加數月時間才能交付、安裝和執行。這與我們計劃的緊迫性或節儉的作風都不相符，因此我們從未認真考慮過它們。我們需要找到一種方法，僅花費資源來保護實際遭受攻擊的網域。我們轉向了古老的原則，即必要性是發明之母。我們的必要性是使用最少量的資源，快速打造世界一流的 100% 正常運作的 DNS 服務。我們的發明是隨機分區。

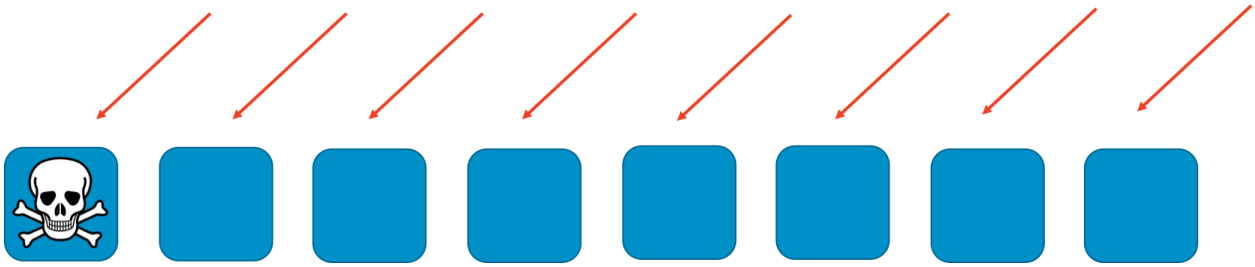
什麼是隨機分區？

隨機分區操作簡單，但功能很強大。它甚至比我們最初設想的還要強大。我們不斷使用，將其打造為一種核心模式，讓 AWS 能夠提供經濟高效的多租用戶服務，從而為每個客戶提供單一租用戶體驗。

若要了解隨機分區的工作原理，首先要考慮如何透過一般分區讓系統更具可擴展性和彈性。想像由八個工作程式組成的可水平擴展的系統或服務。下圖說明了工作程式及其要求。工作程式可以是伺服器、佇列或資料庫等，構成系統的任何「物件」。



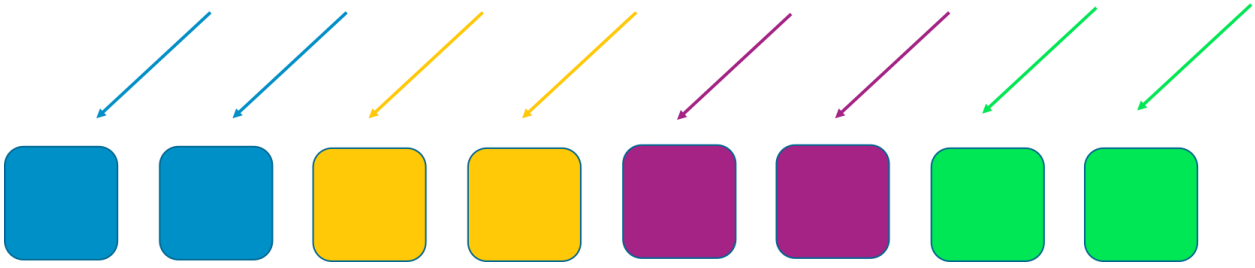
若沒有任何分區，工作程式機群將處理所有工作。每個工作程式都必須能夠處理任何請求。這對於提高效率和冗餘性非常有用。若一個工作程式失敗，其他七個能夠承擔工作，因此系統中需要的寬限容量相對較小。然而，若特定類型的請求或大量請求 (如 DDOS 攻擊) 觸發失敗，則會出現大問題。以下兩個圖顯示了這種攻擊的進程。



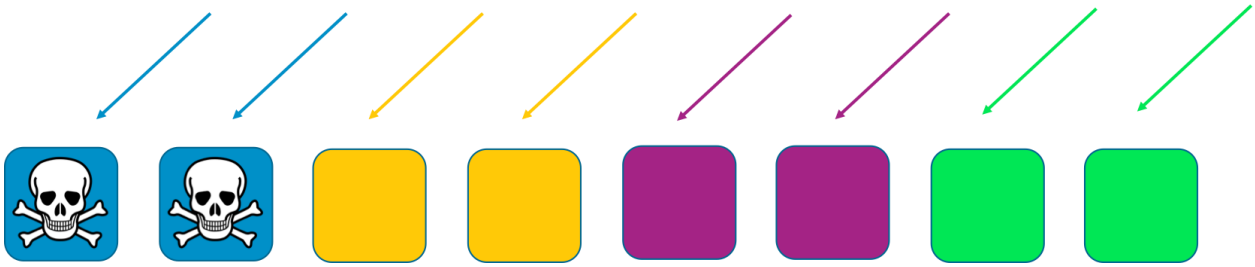
該問題將受影響的第一個工作程式扼殺，但隨著其餘工作程式的接管，繼續逐步影響其他工作程式。該問題很快會讓所有工作程式及整個服務癱瘓。

這種失敗的影響範圍是「所有事物和所有人」。 整個服務中斷。每個客戶都會受到影響。正如我們在可用性工程中所說的那樣：這不是最佳選擇。

透過常規分區，我們可以很好地改善。若將機群分為 4 個工作程式分區，就能以效率來換取影響範圍。以下兩個圖顯示了分區如何限制 DDOS 攻擊的影響。



在此示例中，每個分區都有兩個工作程式。我們在各個分區之間分割資源，例如客戶網域。我們仍有冗餘，但因為每個分區只有兩個工作程式，所以我們必須在系統中保留更多的備用容量來處理任何失敗。作為回報，影響範圍大大減小。

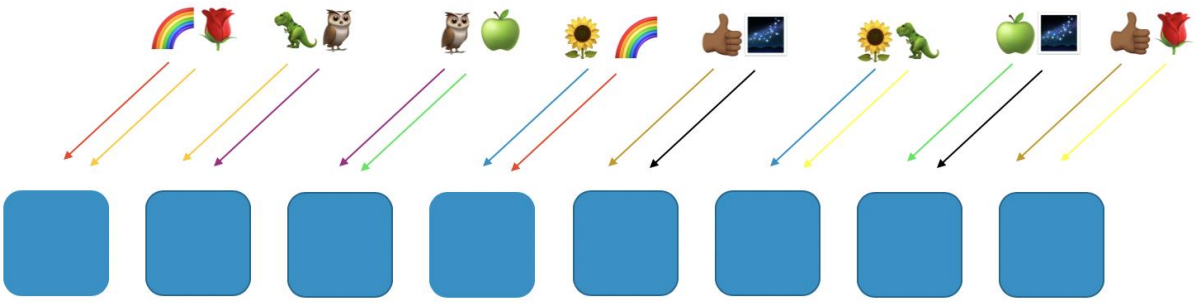


在這個分區世界中，影響範圍隨分區數量增加而減小。這裡有四個分區，若客戶遇到問題，則託管服務的分區，以及該分區的所有其他客戶都可能受到影響。但該分區僅佔整體服務的四分之一。25% 的影響比 100% 影響要好得多。使用隨機分區，則效用可成倍增加。

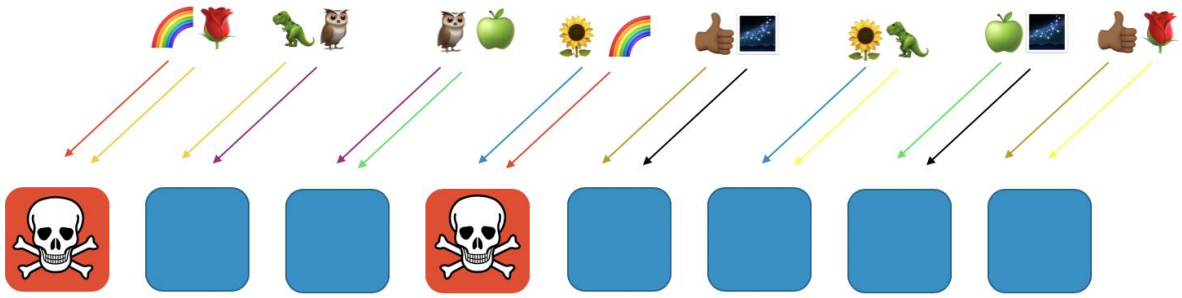
透過隨機分區，我們分別建立了兩個工作程式的虛擬分區，然後將客戶或資源或任何我們想隔離的事物，指派給其中一個虛擬分區。

下圖顯示了隨即分區佈局範例，其中包含八個工作程式和八位客戶，每位客戶被指派給兩個工作程式。通常，我們的客戶要比工作程式多得多，但我們若縮小規模，會更容易跟進。我們將重點關注兩位客戶——彩虹客戶和玫瑰客戶。

在範例中，我們將彩虹客戶支配給第一個工作程式和第四個工作程式。這兩個工作程式的組合構成了客戶的隨機分區。其他客戶及其自己的兩個混合工作程式構成不同的虛擬分區。例如，玫瑰客戶也被指派給第一個工作程式，但另一個是第八個工作程式。



若彩虹客戶所在的第一個和第四個工作程式遇到問題 (例如有害請求或大量請求)，則問題將影響該虛擬分區，但完全不會影響其他任何隨即分區。實際上，最多是另一個隨即分區會受到影響。若請求者可容錯並能解決此問題 (例如重試)，則可繼續為客戶或其餘分區資源提供不間斷服務，如下圖所示。



換句話說，雖然服務彩虹客戶的所有工作程式都可能遇到問題或襲擊，但其他工作程式完全不受影響。對客戶而言，這意味著即使玫瑰客戶和向日葵客戶各自與彩虹共用一個工作程式，也不會受到影響。玫瑰客戶可以從第八個工作程式獲得服務，向日葵客戶可以從第六個工作程式獲得服務，如下圖所示。

當發生問題時，我們仍會損失全部服務的四分之一，但這種指派客戶或資源的方式意味著，隨即分區的影響範圍得到極大改善。若有八個工作程式，則可設置 28 個由兩個工作程式組成的獨特組合，這意味著有 28 個可能的隨機分區。若我們有成千上萬或更多的客戶，並且將每個客戶指派給一個隨即分區，那麼問題造成的影響範圍僅為 $1/28^{th}$。這比常規分區的效用提升了 7 倍。

令人欣喜的是，擁有的工作程式和客戶越多，該數據會成倍增加。在這些方面，大多數擴展挑戰變得越來越艱難，但隨即分區會變得更加有效。實際上，若有足夠的工作程式，則隨即分區可能會比客戶多，並且每位客戶都可以被隔離。

Amazon Route 53 與隨即分區

所有這些如何協助 Amazon Route 53？ 藉助 Route 53，我們決定將容量擴展至共 2048 個虛擬名稱伺服器中。這些是虛擬伺服器，因為其與託管 Route 53 的實體伺服器不對應。我們可以四處

移動以協助管理容量。然後將每個客戶網域指派給四個虛擬名稱伺服器的隨機分區。數據顯示可能會有 7300 億個可能的隨機分區。我們有如此多可能的隨機分區，因此可以為每個網域指派唯一的隨機分區。實際上，我們可以進一步改善，並確保沒有客戶網域與任何其他客戶網域共用兩個以上的虛擬名稱伺服器。

結果令人驚歎。若某個客戶網域是 DDOS 攻擊的目標，則指派該網域所在的四個虛擬名稱伺服器流量將激增，但其他客戶的網域不會受影響。我們不會因目標客戶遭受攻擊而損失慘重。隨機分區意味著我們可以識別目標客戶，並將其隔離至特別建立的專用攻擊容量區。此外，我們還開發了自己的 AWS Shield 流量清理器專有層。而即使發生這些事件，隨即分區也能讓局面迥然不同，從而確保 Route 53 客戶擁有無縫的整體體驗。

結論

我們已經在眾多其他系統中內嵌隨即分區。此外，我們還提出了一些改進措施，例如遞迴隨即分區，其中我們對多層項目進行分區，從而隔離客戶的客戶。隨機分區具有很大的適應性。這是安排現有資源的明智方法。此外，通常也無須支付額外費用，因此在開源節流方面可帶來極大的改善。

若您有意自行使用隨機分區，請查閱我們的開放原始碼 Route 53 Infima 庫。這個庫包含可用於指派或安排資源的幾種不同的隨即分區實作方式。