Whitepaper zu AWS

Echtzeit-Kommunikation in AWS

Echtzeit-Kommunikation in AWS: Whitepaper zu AWS

Table of Contents

Echtzeit-Kommunikation in AWS

Bewährte Methoden für die Entwicklung hochverfügbarer und skalierbarer RTC-Workloads (Echtzeit-Kommunikation) in AWS.

Erscheinungsdatum: 13. Februar 2020 (Dokumentversionen)

Überblick

Heutzutage möchten viele Unternehmen ihre Kosten senken und skalierbare Sprach-, Messaging- und Multimedia-Workloads in Echtzeit nutzen. In diesem Dokument werden bewährte Methoden für die Verwaltung von Echtzeit-Kommunikationsworkloads in AWS beschrieben. Außerdem enthält es Referenzarchitekturen, um diese Anforderungen zu erfüllen. Dieses Dokument ist ein Leitfaden für Personen, die mit der Echtzeitkommunikation vertraut sind, um Hochverfügbarkeit und Skalierbarkeit für diese Workloads zu erreichen.

Einführung

Telekommunikationsanwendungen, die Sprache, Video und Messaging als Kanäle verwenden, sind für viele Unternehmen und ihre Endbenutzer eine wichtige Anforderung. Diese Workloads für die Echtzeitkommunikation (RTC) haben spezifische Latenz- und Verfügbarkeitsanforderungen, die mit bewährten Designmethoden erfüllt werden können. In der Vergangenheit wurden RTC-Workloads in traditionellen On-Premises-Rechenzentren mit dedizierten Ressourcen bereitgestellt.

Dank eines ausgereiften und wachsenden Funktionsumfangs können RTC-Workloads jedoch trotz strenger Service-Level-Anforderungen in Amazon Web Services (AWS) bereitgestellt werden und gleichzeitig von Skalierbarkeit, Elastizität und hoher Verfügbarkeit profitieren. Derzeit nutzen mehrere Kunden Lösungen von AWS, seinen Partnern sowie Open-Source-Lösungen, um RTC-Workloads mit einer potenziellen Verteilung in Minutenschnelle sowie umfangreichen Funktionen von AWS-Services günstiger und agiler auszuführen.

Kunden nutzen AWS-Funktionen wie Enhanced Networking mit einem Elastic Network Adapter (ENA) und EC2-Instances (Amazon Elastic Compute Cloud) der neuesten Generation, um vom Data Plane Development Kit (DPDK), Single-Root-I/O-Virtualisierung (SR-IOV), Huge Pages, NVM Express (NVMe), Unterstützung für Non-Uniform Memory Access (NUMA) sowie Bare-Metal-Instances zur Erfüllung der RTC-Workload-Anforderungen zu profitieren. Diese Instances bieten eine Netzwerkbandbreite von bis zu 100 Gbit/s und eine entsprechende Anzahl an Paketen pro Sekunde und somit eine höhere Leistung für netzwerkintensive Anwendungen. Für die Skalierung bietet Elastic Load Balancing den Application Load Balancer mit WebSocket-Unterstützung sowie den Network Load Balancer, der Millionen von Anfragen pro Sekunde verarbeiten kann. AWS Global Accelerator stellt für die Netzwerkbeschleunigung statische IP-Adressen bereit, die als fester Einstiegspunkt zu Ihren Anwendungsendpunkten in AWS dienen. Unterstützung statischer IP-Adressen für die Lastenverteilung. AWS Direct Connect stellt für reduzierte Latenz, niedrigere Kosten und einen erhöhten Bandbreitendurchsatz eine dedizierte Netzwerkverbindung von On-Premises zu AWS her. Hochverfügbares verwaltetes SIP-Trunking wird vom Amazon Chime Voice Connector bereitgestellt. Amazon Kinesis Video Streams mit WebRTC streamen problemlos Zwei-Wege-Medien in Echtzeit mit hoher Verfügbarkeit.

Dieses Dokument enthält Referenzarchitekturen, die die Einrichtung von RTC-Workloads in AWS veranschaulichen, sowie bewährte Methoden zur Optimierung von Lösungen, um sie auf die Cloud vorzubereiten und gleichzeitig die Anforderungen der Endbenutzer zu erfüllen. Der Evolved Packet Core (EPC) ist für dieses Whitepaper nicht verfügbar, aber die bewährten Methoden können auf virtuelle Netzwerkfunktionen (VNFs) angewendet werden.

Grundlegende Komponenten der RTC-Architektur

In der Telekommunikationsbranche bezieht sich die Echtzeitkommunikation (RTC) üblicherweise auf Live-Mediensitzungen zwischen zwei Endpunkten mit minimaler Latenz. Beispiele für solche Sitzungen sind:

- Eine Sprachsitzung zwischen zwei Parteien (z. B. Telefonanlage, Mobiltelefon, VoIP)
- Instant Messaging (z. B. Chatten, IRC)
- Live-Videositzung (z. B. Videokonferenzen, TelePresence)

Alle oben erwähnten Lösungen haben gewisse Komponenten gemein (z. B. Komponenten, die Authentifizierung, Autorisierung und Zugriffssteuerung, Transcodierung, Pufferung und Weiterleitung usw. ermöglichen) und manche dieser Komponenten sind je nach Art der übertragenen Medien einzigartig (z. B. Rundfunkdienst, Messaging-Server und Warteschlangen usw.). In diesem Abschnitt wird die Definition eines sprach- und videobasierten RTC-Systems und aller zugehörigen Komponenten erörtert, die in Abbildung 1 dargestellt sind.

Abbildung 1: Wesentliche Komponenten der Architektur für RTC

Themen

Softswitch/Nebenstellenanlage

Ein Softswitch bzw. eine Nebenstellenanlage ist das Gehirn eines Sprachtelefonsystems und bietet mithilfe verschiedener Komponenten Informationen zum Einrichten, Verwalten und Weiterleiten eines Sprachanrufs innerhalb oder außerhalb des Unternehmens. Alle Abonnenten des Unternehmens müssen sich beim Softswitch registrieren, um einen Anruf entgegenzunehmen oder zu tätigen. Eine wichtige Funktion des Softswitches besteht darin, jeden Teilnehmer sowie seine Kontaktdaten

nachzuverfolgen, damit er mithilfe der anderen Komponenten innerhalb des Sprachnetzwerks erreicht werden kann.

Session Border Controller (SBC)

Ein Session Border Controller (SBC) befindet sich am Rand eines Sprachnetzwerks und verfolgt den gesamten eingehenden und ausgehenden Datenverkehr (sowohl auf Kontroll- als auch auf Datenebene). Eine der Hauptaufgaben eines SBC besteht darin, das Sprachsystem vor Missbrauch zu schützen. Der SBC kann verwendet werden, um eine Verbindung mit SIP-Trunks (Session Initiation Protocol) für externe Konnektivität herzustellen. Einige SBCs bieten auch Transcodierungsfunktionen zum Konvertieren von CODECS von einem Format in ein anderes. Außerdem bieten die meisten SBCs auch NAT-Traversal-Funktionen, mit denen sichergestellt werden kann, dass Anrufe auch über Firewall-Netzwerke hergestellt werden.

PSTN-Konnektivität

Voice-over-IP-Lösungen (VoIP) verwenden PSTN-Gateways und SIP-Trunks, um Verbindungen mit Legacy-PSTN-Netzwerken herzustellen.

PSTN-Gateway

Das PSTN-Gateway (Public Switched Telephone Network) wandelt die Signalisierung (zwischen SIP und SS7) und Medien (zwischen RTP und Zeitmultiplex [TDM] mithilfe der CODEC-Transcodierung) um. PSTN-Gateways befinden sich immer am Rand in der Nähe des PSTN-Netzwerks.

SIP-Trunk

In einem SIP-Trunk beendet das Unternehmen seine Anrufe in ein TDM-Netzwerk (SS7-basiert) nicht, sondern wickelt den Datenfluss zwischen Unternehmen und Telekommunikation weiter über IP ab. Die meisten SIP-Trunks werden mithilfe von SBCs eingerichtet. Das Unternehmen muss die vordefinierten Sicherheitsregeln der Telekommunikation festlegen, z. B. das Zulassen eines bestimmten Bereichs von IP-Adressen, Ports usw.

Mediengateway (Transcoder)

Eine typische Sprachlösung unterstützt verschiedene Arten von CODECs. Einige der gängigen CODECs sind das G.711 μ-law in Nordamerika, das G.711 A-law außerhalb Nordamerikas sowie G.729 und G.722. Wenn zwei Geräte, die zwei verschiedene CODECs verwenden, miteinander

kommunizieren, übersetzt ein Medienserver den CODEC-Datenfluss zwischen den Geräten. Mit anderen Worten verarbeitet ein Mediengateway Medien und stellt sicher, dass die Endgeräte miteinander kommunizieren können.

WebRTC und WebRTC-Gateway

Mit der Web-Echtzeitkommunikation (WebRTC) können Sie mithilfe der API einen Anruf von einem Webbrowser aus einrichten oder Ressourcen vom Backend-Server abfragen. Die Technologie wurde mit Blick auf die Cloud-Technologie entwickelt und bietet daher verschiedene APIs, die zum Einrichten eines Anrufs verwendet werden könnten. Da nicht alle Sprachlösungen (einschließlich SIP) diese APIs unterstützen, ist das WebRTC-Gateway erforderlich, um API-Aufrufe in SIP-Nachrichten zu übersetzen und umgekehrt.

Abbildung 2 zeigt ein Entwurfsmuster für eine hochverfügbare WebRTC-Architektur. Der eingehende Datenverkehr von WebRTC-Clients wird durch einen Amazon Application Load Balancer ausgeglichen, wobei WebRTC auf EC2-Instances ausgeführt wird, die Teil einer Auto-Scaling-Gruppe sind.

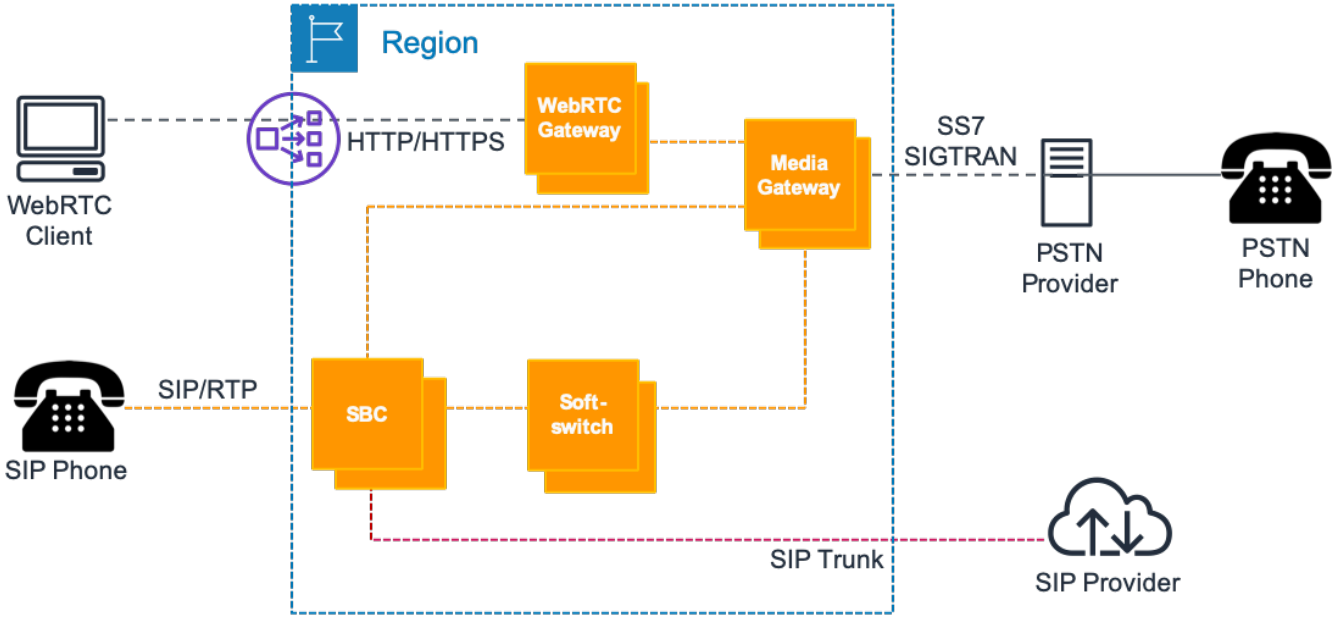

Abbildung 2: Grundlegende Topologie eines RTC-Systems für Sprache

Ein weiteres Entwurfsmuster für SIP- und RTP-Datenverkehr ist die Verwendung von SBC-Paaren in Amazon EC2 im aktiven passiven Modus in mehreren Availability Zones (Abbildung 3). Hier kann bei einem Ausfall eine elastische IP-Adresse dynamisch zwischen Instances verschoben werden, für die DNS nicht verwendet werden kann.

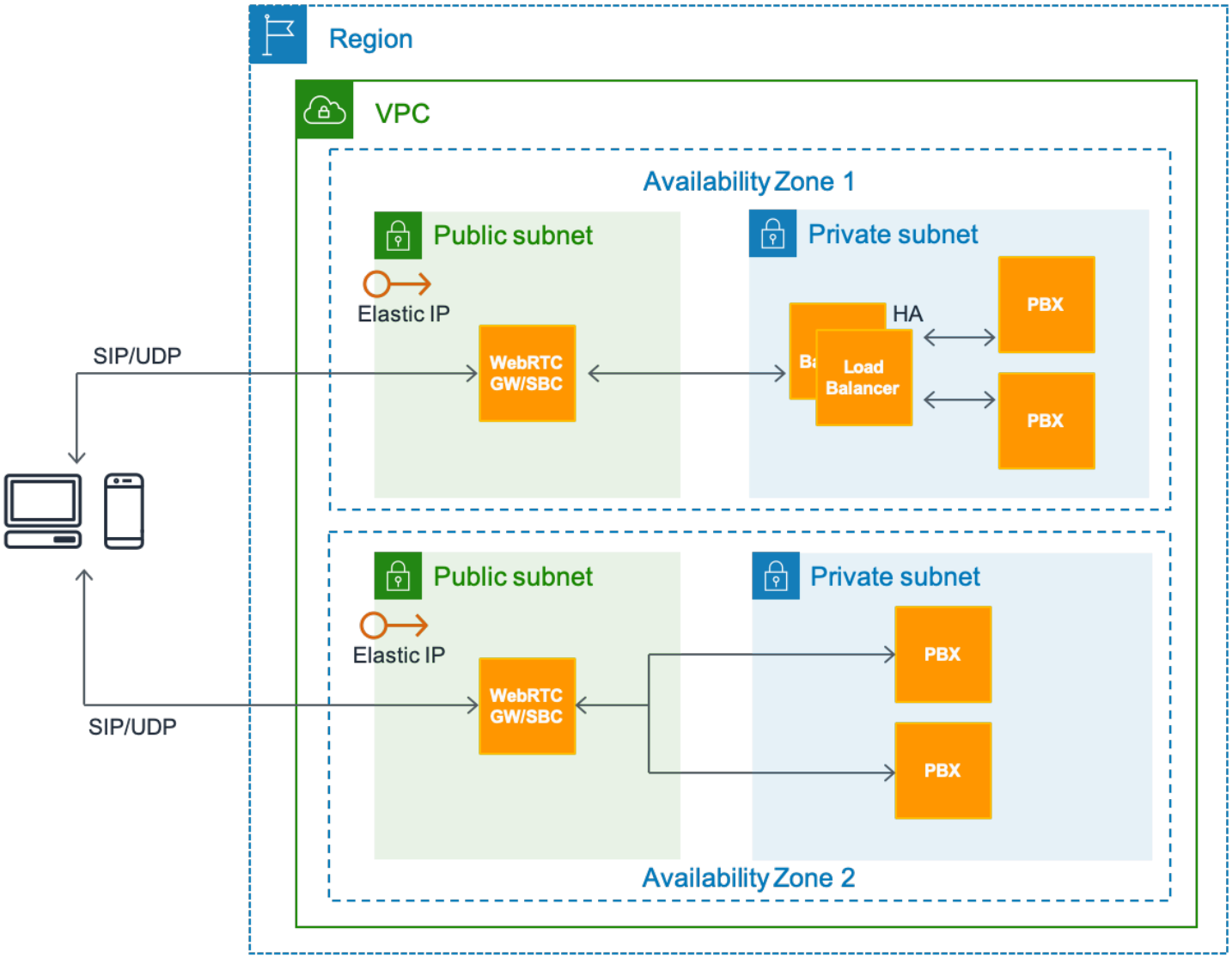

Abbildung 3: RTC-Architektur mit Amazon EC2 in einer VPC

Hohe Verfügbarkeit und Skalierbarkeit in AWS

Die meisten Anbieter von Echtzeitkommunikation bieten Service-Levels mit einer Verfügbarkeit von 99,9 % bis 99,999 %. Je nach Umfang der gewünschtem Hochverfügbarkeit (HA) müssen Sie während des gesamten Lebenszyklus der Anwendung immer komplexere Maßnahmen ergreifen. Wir empfehlen, folgende Richtlinien zu befolgen, um eine robuste Hochverfügbarkeit zu erreichen:

- Entwerfen Sie ein System, dass keinen Single-Point-of-Failure hat. Verwenden Sie automatisierte Überwachungs-, Fehlererkennungs- und Failover-Mechanismen für zustandslose und zustandsbehaftete Komponenten.

- Single Points of Failure (SPOF) werden üblicherweise mit einer N+1- oder 2N-Redundanzkonfiguration beseitigt. N+1 wird über eine Lastenverteilung zwischen Aktiv-Aktiv-Knoten und 2N durch ein Knotenpaar in einer Aktiv-Standby-Konfiguration erreicht.
- AWS verfügt über mehrere Methoden, um mit beiden Ansätzen HA zu erreichen, z. B. durch einen skalierbaren Cluster mit Lastenverteilung oder die Annahme eines Aktiv-Standby-Paars.

- Angemessene Verfügbarkeit von Instrumenten und Testsystemen.
- Bereiten Sie Betriebsverfahren für manuelle Mechanismen vor, um auf einen Ausfall zu reagieren, ihn abzuwehren und eine Wiederherstellung durchzuführen.

In diesem Abschnitt wird erläutert, wie Sie mit den Funktionen in AWS ein System ohne SPOF erreichen. In diesem Abschnitt werden eine Reihe wichtiger AWS-Funktionen und -Designmuster beschrieben, mit denen Sie hochverfügbare Echtzeit-Kommunikationsanwendungen auf der Plattform erstellen können.

Themen

- Hochverfügbares SIP-Trunking mit Amazon Chime Voice Connector

Floating-IP-Muster für HA zwischen zustandsbehafteten Aktiv-Standby-Servern

Das Floating-IP-Designmuster ist ein bekannter Mechanismus, um ein automatisches Failover zwischen einem aktiven und einem Standby-Paar von Hardwareknoten (Medienserver) durchzuführen. Dem aktiven Knoten wird eine statische sekundäre virtuelle IP-Adresse zugewiesen. Ausfälle werden durch die kontinuierliche Überwachung der aktiven Knoten und Standby-Knoten erkannt. Wenn der aktive Knoten ausfällt, weist das Überwachungsskript die virtuelle IP dem betriebsbereiten Standby-Knoten zu und der Standby-Knoten übernimmt die wichtigste aktive Funktion. Auf diese Weise schwebt die virtuelle IP zwischen dem aktiven Knoten und dem Standby-Knoten.

Themen

Anwendbarkeit in RTC-Lösungen

Es ist nicht immer möglich, mehrere aktive Instances derselben Komponente auszuführen, z. B. einen Aktiv-Aktiv-Cluster mit N Knoten. Für HA ist eine Aktiv-Standby-Konfiguration die beste Wahl. Beispielsweise eignen sich die zustandsbehafteten Komponenten in einer RTC-Lösung (z. B. Medienserver oder Konferenzserver oder sogar ein SBC- oder Datenbankserver) gut für eine Aktiv-Standby-Einrichtung. Auf einem SBC- oder Medienserver sind zu einem bestimmten Zeitpunkt mehrere Sitzungen oder Kanäle mit langer Laufzeit aktiv, und falls die aktive SBC-Instance ausfällt, können sich die Endpunkte dank der Floating-IP ohne clientseitige Konfiguration wieder mit dem Standby-Knoten verbinden.

Implementierung in AWS

Sie können dieses Muster in AWS mithilfe der Kernfunktionen von Amazon Elastic Compute Cloud (Amazon EC2), Amazon EC2 API, elastischen IP-Adressen und Unterstützung von Amazon EC2 für sekundäre private IP-Adressen implementieren.

1. Starten Sie zwei EC2-Instances, um die Rollen von Primär- und Sekundärknoten zu übernehmen. Voraussetzung ist, dass sich der Primärknoten standardmäßig im aktiven Zustand befindet.
2. Weisen Sie der primären EC2-Instance eine zusätzliche sekundäre private IP-Adresse zu.
3. Eine elastische IP-Adresse, die einer virtuellen IP (VIP) ähnelt, ist mit der sekundären privaten Adresse verknüpft. Diese sekundäre private Adresse ist die Adresse, die von externen Endpunkten für den Zugriff auf die Anwendung verwendet wird.
4. Es müssen gewisse Konfigurationen am Betriebssystem vorgenommen werden, damit die sekundäre IP-Adresse als Alias zur primären Netzwerkschnittstelle hinzugefügt wird.
5. Die Anwendung muss an diese elastische IP-Adresse binden. Im Fall der Asterisk-Software können Sie die Bindung über erweiterte Asterisk-SIP-Einstellungen konfigurieren.
6. Führen Sie auf jedem Knoten ein Überwachungsskript wie Custom, KeepAlive auf Linux, Corosync usw. aus, um den Status des Peer-Knotens zu überwachen. Falls der aktuelle aktive Knoten ausfällt, erkennt der Peer-Knoten diesen Fehler und ruft die Amazon EC2-API auf, um die sekundäre private IP-Adresse sich selbst neu zuzuweisen.
7. Dadurch wird die Anwendung, die auf der mit der sekundären privaten IP-Adresse verknüpften VIP lauschte, über den Standby-Knoten für Endpunkte verfügbar.

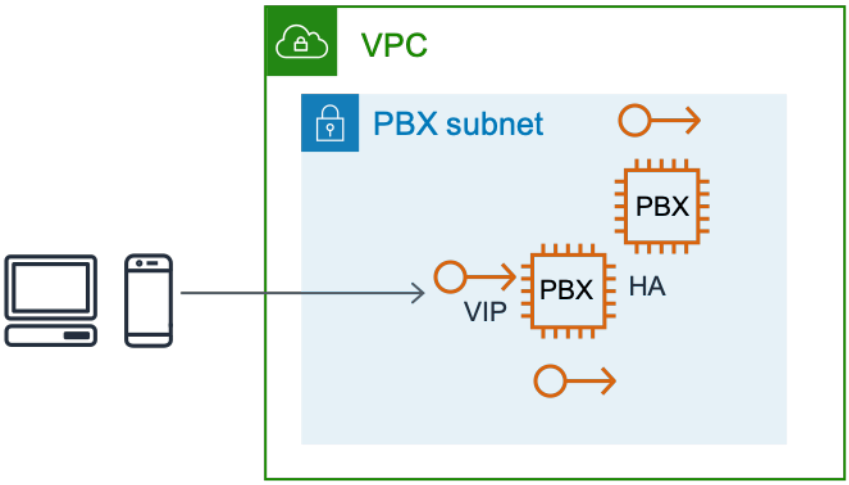

Abbildung 4: Failover zwischen zustandsbehafteten EC2-Instances unter Verwendung der elastischen IP-Adresse

Vorteile

Dieser Ansatz ist eine zuverlässige preiswerte Lösung, die vor Ausfällen auf EC2-Instance-, Infrastruktur- oder Anwendungsebene schützt.

Einschränkungen und Erweiterbarkeit

Dieses Designmuster ist normalerweise auf eine einzige Availability Zone beschränkt. Es kann über zwei Availability Zones hinweg implementiert werden, jedoch mit einer Variante. In diesem Fall wird die elastische Floating-IP-Adresse über die verfügbare API für elastische IP-Adressen erneut zwischen dem aktiven Knoten und dem Standby-Knoten in verschiedenen Availability Zones zugeordnet. Bei der in Abbildung 4 gezeigten Failover-Implementierung werden laufende Aufrufe verworfen und die Endpunkte müssen sich erneut verbinden. Es ist möglich, diese Implementierung um die Replikation der zugrunde liegenden Sitzungsdaten zu erweitern, um ein nahtloses Failover von Sitzungen oder die Medienkontinuität zu gewährleisten.

Lastenverteilung für Skalierbarkeit und HA mit WebRTC und SIP

Die Lastenverteilung eines Clusters aktiver Instances auf der Grundlage vordefinierter Regeln wie Round-Robin, Affinität oder Latenz usw. ist ein Designmuster, das aufgrund der zustandslosen HTTP-Anforderungen weit verbreitet ist. Tatsächlich ist die Lastenverteilung bei vielen RTC-Anwendungskomponenten eine praktikable Option.

Die Lastenverteilung fungiert als Reverse-Proxy oder Einstiegspunkt für Anforderungen an die gewünschte Anwendung, die selbst so konfiguriert ist, dass sie in mehreren aktiven Knoten gleichzeitig ausgeführt wird. Zu einem bestimmten Zeitpunkt leitet die Lastenverteilung eine Benutzeranforderung an einen der aktiven Knoten im definierten Cluster weiter. Die Lastenverteilungen führen eine Zustandsprüfung der Knoten in ihrem Zielcluster durch und senden keine eingehende Anforderung an einen Knoten, der die Zustandsprüfung nicht besteht. Dadurch wird dank der Lastenverteilung eine grundlegende Hochverfügbarkeit erreicht. Da eine Lastenverteilung in Intervallen von weniger als einer Sekunde aktive und passive Zustandsprüfungen aller Clusterknoten durchführt, findet das Failover fast augenblicklich statt.

Die Entscheidung, an welche Knoten Anforderungen geleitet werden sollen, basiert auf Systemregeln, die in der Lastenverteilung definiert sind, einschließlich:

- Round Robin
- Sitzungs- oder IP-Affinität, die sicherstellt, dass mehrere Anforderungen innerhalb einer Sitzung oder von derselben IP an denselben Knoten im Cluster gesendet werden
- Latenzbasierte Regeln
- Lastbasierte Regeln

Themen

Anwendbarkeit in RTC-Architekturen

Mit dem WebRTC-Protokoll können WebRTC Gateways über eine HTTP-basierte Lastenverteilung wie Elastic Load Balancing, Application Load Balancer oder Network Load Balancer eine einfache Lastenverteilung vornehmen. Da die meisten SIP-Implementierungen auf den Transport über TCP und UDP angewiesen sind, ist eine Lastenverteilung auf Netzwerk- oder Verbindungsebene mit Unterstützung für TCP- und UDP-basierten Datenverkehr erforderlich.

Lastenverteilung in AWS für WebRTC mit Application Load Balancer und Auto Scaling

Im Fall von WebRTC-basierter Kommunikation bietet Elastic Load Balancing eine vollständig verwaltete, hochverfügbare und skalierbare Lastenverteilung, die als Einstiegspunkt für Anforderungen dient, die dann an einen Zielcluster von EC2-Instances weitergeleitet werden, die mit Elastic Load Balancing verknüpft sind. Da WebRTC-Anforderungen zustandslos sind, können Sie Amazon EC2 Auto Scaling verwenden, um eine vollautomatische und kontrollierbare Skalierbarkeit, Elastizität und Hochverfügbarkeit zu gewährleisten.

Der Application Load Balancer bietet einen vollständig verwalteten Lastenverteilungsdienst, der mit mehreren Availability Zones hochverfügbar und skalierbar ist. Dieser Dienst unterstützt die Lastenverteilung von WebSocket-Anforderungen, die die Signalisierung für WebRTC-Anwendungen und die bidirektionale Kommunikation zwischen Client und Server über eine TCP-Verbindung mit langer Laufzeit verarbeiten. Der Application Load Balancer unterstützt auch inhaltsbasiertes Routing sowie Sticky Sessions und leitet Anforderungen vom selben Client mit von der Lastenverteilung generierten Cookies an dasselbe Ziel weiter. Wenn Sie Sticky Sessions aktivieren, erhält dasselbe Ziel die Anforderung und kann den Sitzungsinhalt über das Cookie wiederherstellen.

Abbildung 5 zeigt die Zieltopologie.

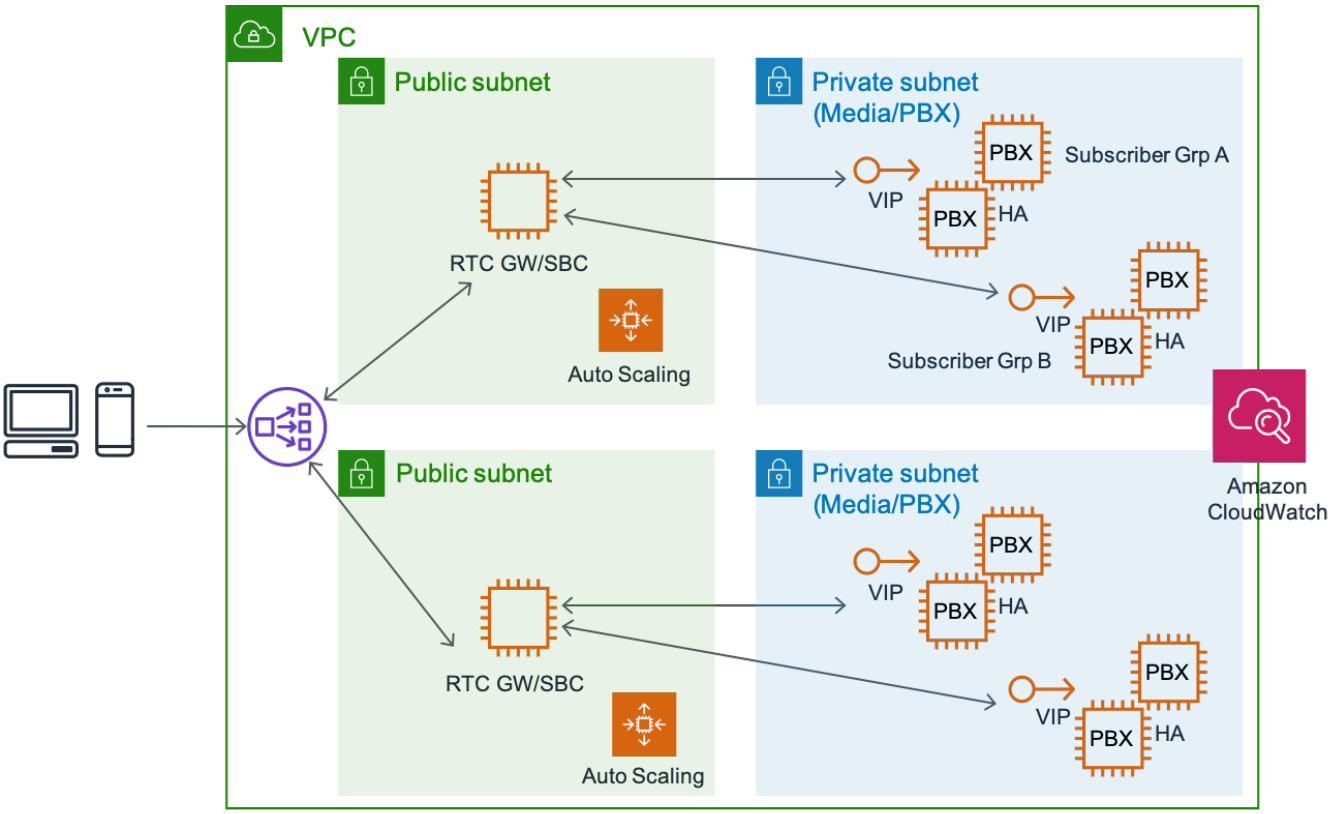

Abbildung 5: WebRTC-Skalierbarkeit und Hochverfügbarkeitsarchitektur

Implementierung für SIP mit Network Load Balancer oder dem AWS Marketplace-Produkt

Bei SIP-basierter Kommunikation werden die Verbindungen über TCP oder UDP hergestellt, wobei die meisten RTC-Anwendungen UDP verwenden. Wenn SIP/TCP das bevorzugte Signalprotokoll ist, kann der Network Load Balancer für eine vollständig verwaltete, hochverfügbare, skalierbare und leistungsfähige Lastenverteilung verwendet werden.

Ein Network Load Balancer arbeitet auf Verbindungsebene (Ebene 4) und leitet Verbindungen zu Zielen wie Amazon-EC2-Instances, Containern und IP-Adressen basierend auf IP-Protokolldaten weiter. Der Network Load Balancer ist ideal für die Lastenverteilung von TCP- oder UDP-Datenverkehr und kann Millionen von Anforderungen pro Sekunde bei extrem niedrigen Latenzen verarbeiten. Er ist in andere beliebte AWS-Services wie AWS Auto ScalingAmazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) und AWS CloudFormation integriert.

Wenn SIP-Verbindungen initiiert werden, kann auch AWS Marketplace kommerzielle Standardsoftware (COTS) verwendet werden. AWS Marketplace bietet viele Produkte, die Lastenverteilungen von UDP-Verbindungen sowie anderer Arten von Verbindungen der Ebene 4 vornehmen können. Diese COTS bieten in der Regel Hochverfügbarkeit und sind üblicherweise in Funktionen integriert, z. B. AWS Auto Scaling zur weiteren Verbesserung der Verfügbarkeit und Skalierbarkeit. Abbildung 6 zeigt die Zieltopologie:

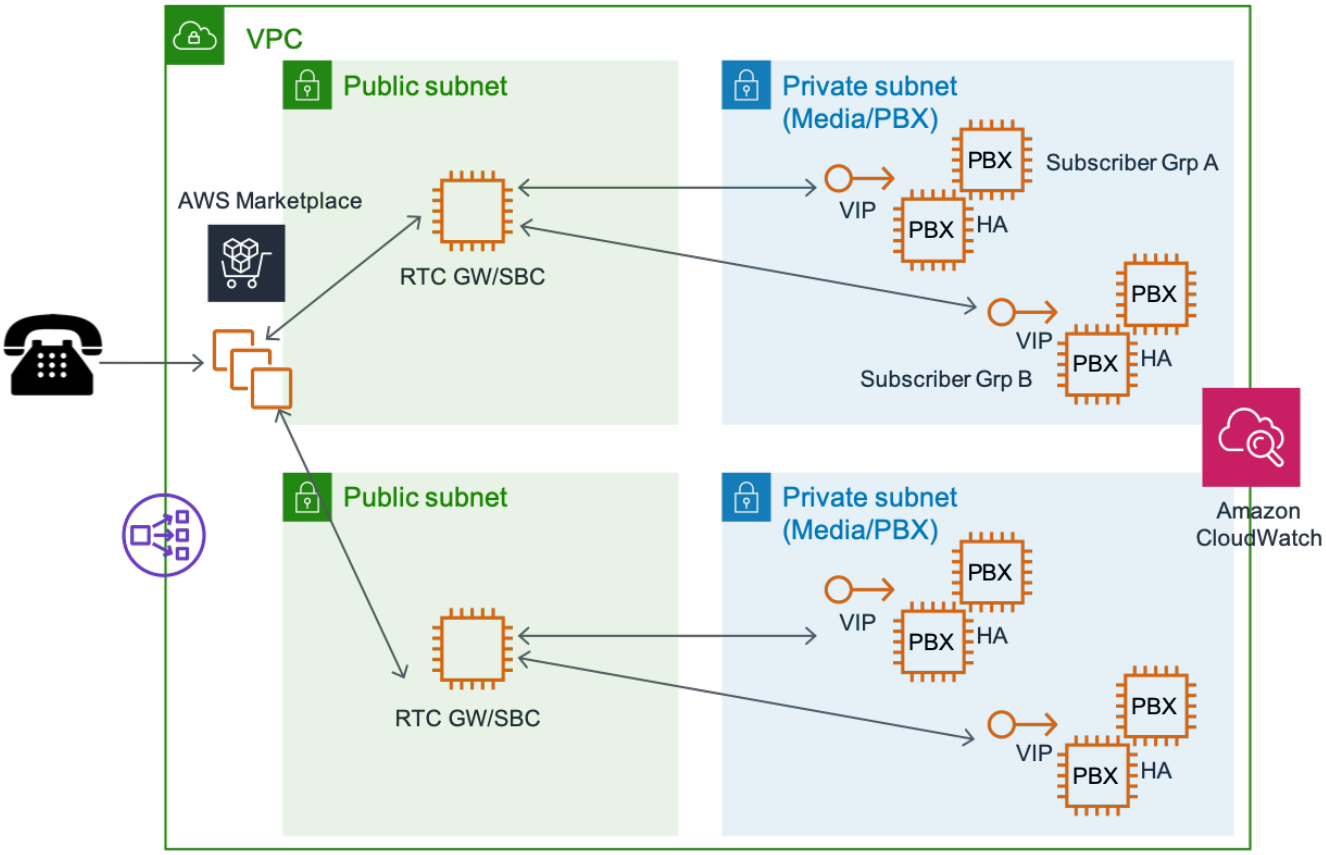

Abbildung 6: SIP-basierte RTC-Skalierbarkeit mit dem AWS Marketplace-Produkt

Regionsübergreifende DNS-basierte Lastenverteilung und Failover

Amazon Route 53 bietet einen globalen DNS-Service, der als öffentlicher oder privater Endpunkt für RTC-Clients zur Registrierung und Verbindung mit Medienanwendungen verwendet werden kann. Mit Amazon Route 53 können DNS-Zustandsprüfungen konfiguriert werden, um den Datenverkehr an fehlerfreie Endpunkte weiterzuleiten oder den Zustand Ihrer Anwendung unabhängig zu überwachen. Mit der Funktion Amazon Route 53 Traffic Flow können Sie Datenverkehr auf einfache Weise durch verschiedene Routingtypen, wie latenzbasiertes Routing, GEO-DNS, Geoproximität und Weighted Round Robin, global verwalten. Diese Typen lassen sich alle für eine fehlertolerante Architektur mit kurzer Latenz mit DNS Failover kombinieren. Mit dem einfachen visuellen Editor für den Amazon

Route 53-Datenverkehrsfluss können Sie das Routing Ihrer Endbenutzer zu den Endpunkten Ihrer Anwendung ganz einfach verwalten – und zwar unabhängig davon, ob es um eine einzelne oder weltweit verteilte AWS-Region geht.

Bei globalen Bereitstellungen ist die latenzbasierte Routing-Richtlinie in Route 53 besonders nützlich, um Kunden zum nächstgelegenen Point of Presence für einen Medienserver zu leiten und so die Servicequalität im Zusammenhang mit dem Medienaustausch in Echtzeit zu verbessern.

Beachten Sie, dass Client-Caches bereinigt werden müssen, um ein Failover auf eine neue DNS-Adresse zu erzwingen. Bei DNS-Änderungen können Verzögerungen auftreten, da sie über globale DNS-Server verteilt werden. Sie können das Aktualisierungsintervall für DNS-Lookups mit dem Attribut „Time to Live“ verwalten. Dieses Attribut lässt sich bei der Einrichtung von DNS-Richtlinien konfigurieren.

Um globale Benutzer schnell zu erreichen oder die Anforderungen einer einzelnen öffentlichen IP zu erfüllen, kann AWS Global Accelerator auch für regionsübergreifende Failover verwendet werden. AWS Global Accelerator ist ein Netzwerkservice, der die Verfügbarkeit und Leistung von Anwendungen mit lokaler und globaler Reichweite verbessert. AWS Global Accelerator bietet statische IP-Adressen, die als fester Einstiegspunkt für Ihre Anwendungsendpunkte dienen, z. B. Ihre Application Load Balancers, Network Load Balancers oder Amazon-EC2-Instances in einer oder mehreren AWS-Regionen. Es verwendet das globale AWS-Netzwerk, damit Ihre Benutzer leichter auf Ihre Anwendungen zugreifen können, indem die Leistung verbessert wird, z. B. die Latenz Ihres TCP- und UDP-Datenverkehrs. AWS Global Accelerator überwacht kontinuierlich den Zustand Ihrer Anwendungsendpunkte und leitet den Datenverkehr automatisch an die nächsten fehlerfreien Endpunkte um, falls auf den aktuellen Endpunkten Fehler auftreten. Für zusätzliche Sicherheitsanforderungen verwendet Accelerated Site-to-Site-VPN AWS Global Accelerator, um die Leistung von VPN-Verbindungen zu verbessern, indem der Datenverkehr intelligent durch das AWS Global Network und die AWS-Edge-Standorte geleitet wird.

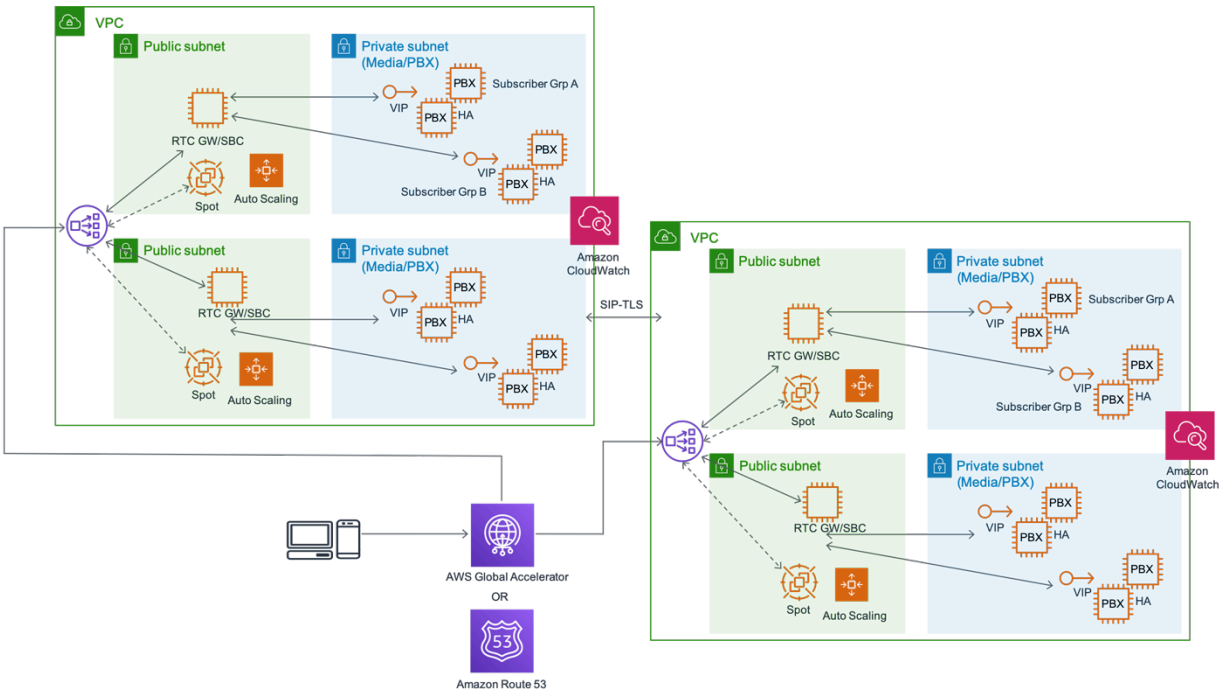

Abbildung 7: Regionsübergreifendes Design mit hoher Verfügbarkeit unter Verwendung von AWS Global Accelerator oder Amazon Route 53

Datenbeständigkeit und HA mit persistentem Speicher

Die meisten RTC-Anwendungen sind auf persistenten Speicher angewiesen, um Daten zur Authentifizierung, Autorisierung, Buchhaltung (Sitzungsdaten, Anrufdetaildatensätze usw.), Betriebsüberwachung und Protokollierung zu speichern und darauf zuzugreifen. In einem herkömmlichen Rechenzentrum ist in der Regel ein großer Arbeitsaufwand erforderlich, um eine hohe Verfügbarkeit und Beständigkeit der persistenten Speicherkomponenten (Datenbanken, Dateisysteme usw.) zu gewährleisten, da ein SAN, RAID-Design sowie Prozesse für Backup, Wiederherstellung und die Verarbeitung von Failovers eingerichtet werden müssen. Die AWS Cloud vereinfacht und verbessert die traditionellen Praktiken von Rechenzentren in Bezug auf Datenbeständigkeit und Verfügbarkeit erheblich.

Für Objektspeicher und Dateispeicher bieten AWS-Services wie Amazon Simple Storage Service (Amazon S3) und Amazon Elastic File System (Amazon EFS) verwaltete Hochverfügbarkeit und Skalierbarkeit. Amazon S3 hat eine Datenlebensdauer von 11 Neunen.

Für die Speicherung von Transaktionsdaten können Kunden den vollständig verwalteten Amazon Relational Database Service (Amazon RDS) nutzen, der Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle und Microsoft SQL Server mit Hochverfügbarkeitsbereitstellungen

unterstützt. Amazon RDS bietet eine fehlertolerante, hochverfügbare und skalierbare Option für die Registratorfunktion, das Abonnentenprofil oder die Speicherung von Buchhaltungsunterlagen (z. B. Anrufdetailsdatensätzen).

Dynamische Skalierung mit AWS Lambda, Amazon Route 53 und AWS Auto Scaling

Mit AWS können Funktionen verkettet und benutzerdefinierte Serverless-Funktionen als Service auf der Grundlage von Infrastrukturereignissen integriert werden. Ein solches Designmuster ist in RTC-Anwendungen vielseitig einsetzbar. Es ist eine Kombination von Lebenszyklus-Hooks mit automatischer Skalierung mit Amazon CloudWatch Events, Amazon Route 53 und AWS Lambda-Funktionen. In die AWS Lambda-Funktionen kann jede Aktion oder Logik eingebettet werden. Abbildung 8 zeigt, wie diese verketteten Funktionen die Systemzuverlässigkeit und Skalierbarkeit durch Automatisierung verbessern können.

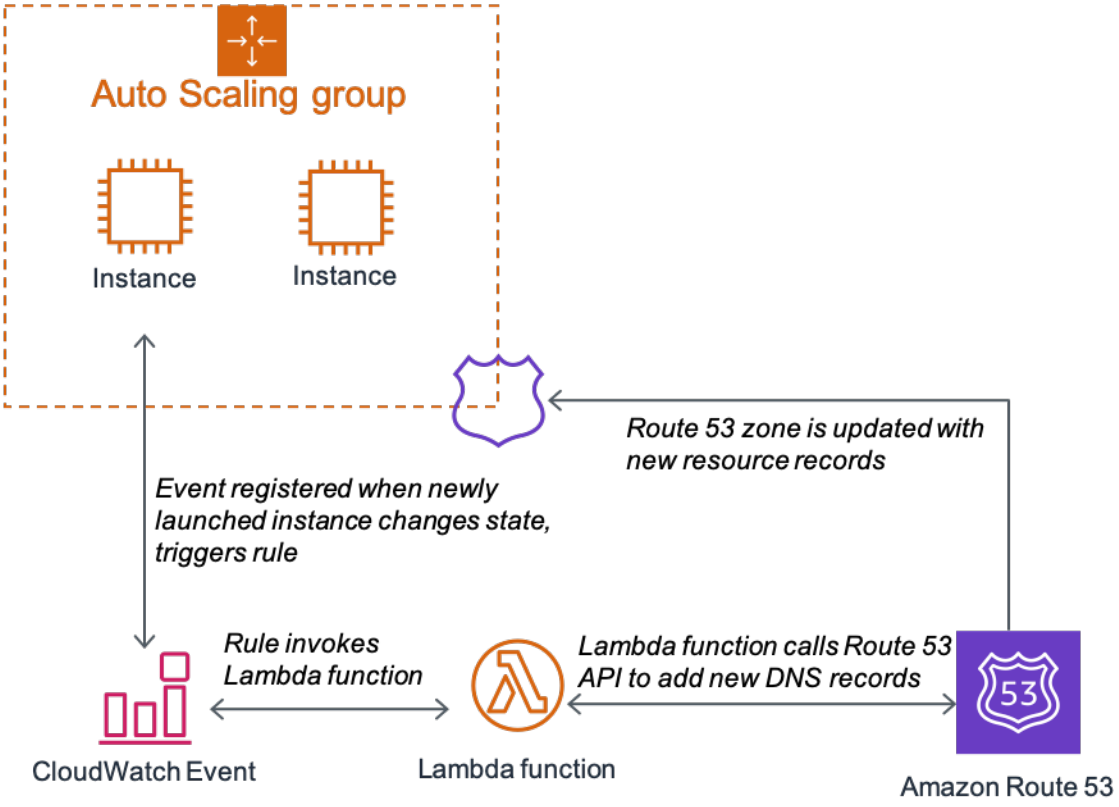

Abbildung 8: Automatische Skalierung mit dynamischen Aktualisierungen von Amazon Route 53

Hochverfügbares WebRTC mit Kinesis Video Streams

Amazon Kinesis Video Streams bietet Medienstreaming in Echtzeit über WebRTC, damit Benutzer Medienstreams zur Wiedergabe, Analyse und für maschinelles Lernen erfassen, verarbeiten und speichern können. Diese Streams sind hochverfügbar, skalierbar und entsprechen den WebRTC-Standards. Amazon Kinesis Video Streams enthält einen WebRTC-Signalendpunkt für eine schnelle Peer-Erkennung und die Einrichtung einer sicheren Verbindung. Es beinhaltet verwaltete Session Traversal Utilities für NAT (STUN) und Traversal Using-Relais um NAT-Endpunkte (TURN) herum für den Echtzeitaustausch von Medien zwischen Peers. Es enthält auch ein kostenloses Open-Source-SDK, das direkt in die Kamera-Firmware integriert ist, um eine sichere Kommunikation mit Kinesis Video Streams-Endpunkten für Peer-Discovery und Medienstreaming zu ermöglichen. Schließlich bietet es Client-Bibliotheken für Android, iOS und JavaScript, mit denen WebRTC-konforme Mobile- und Webplayer sicher ein Kameragerät für Medienstreaming und zweiseitige Kommunikation erkennen und sich mit ihnen verbinden können.

Hochverfügbares SIP-Trunking mit Amazon Chime Voice Connector

Amazon Chime Voice Connector bietet einen SIP-Trunking-Service mit nutzungsbasierter Zahlung, mit dem Firmen sichere und preiswerte Anrufe mit ihren Telefonsystemen tätigen und/oder empfangen können. Amazon Chime Voice Connector ist eine kostengünstige Alternative zu Dienstanbieter-SIP-Trunks oder den Primary Rate Interfaces (PRIs) des Integrated Services Digital Network (ISDN). Kunden können wahlweise ausgehende oder eingehende Anrufe oder beides aktivieren. Dieser Service nutzt das AWS-Netzwerk, um eine hochverfügbare Anruferfahrung über mehrere AWS-Regionen zu ermöglichen. Sie können Audio von SIP-Trunking-Telefonanrufen oder weitergeleiteten SIP-basierten Medienaufzeichnungsfeeds (SIPREC) an Amazon Kinesis Video Streams streamen, um in Echtzeit Erkenntnisse aus Geschäftsgesprächen zu gewinnen. Durch die Integration in Amazon Transcribe und andere gängige Bibliotheken für maschinelles Lernen können Sie schnell Anwendungen für die Audioanalyse erstellen.

Bewährte Methoden aus der Praxis

In diesem Abschnitt werden die bewährten Methoden zusammengefasst, die von einigen der größten und erfolgreichsten AWS-Kunden implementiert wurden, die umfassende Echtzeit-SIP-Workloads (Session Initiation Protocol) ausführen. AWS-Kunden, die ihre eigene SIP-Infrastruktur in der öffentlichen Cloud betreiben möchten, finden diese bewährten Methoden wertvoll, da sie die Zuverlässigkeit und Ausfallsicherheit des Systems bei verschiedenartigen Ausfällen erhöhen können. Obwohl einige dieser bewährten Methoden SIP-spezifisch sind, gelten die meisten von ihnen für jede auf AWS ausgeführte Echtzeit-Kommunikationsanwendung.

Themen

Ein SIP-Overlay erstellen

AWS verfügt über ein robustes, skalierbares und redundantes Netzwerk-Backbone, das Konnektivität zwischen verschiedenen Regionen bietet. Wenn ein Netzwerkereignis, z. B. die Unterbrechung einer Glasfaserverbindung, eine AWS-Backbone-Verbindung verschlechtert, wird der Datenverkehr mithilfe von Routing-Protokollen auf Netzwerkebene wie BGP schnell auf redundante Pfade umgeleitet. Dieses Traffic-Engineering auf Netzwerkebene ist eine Blackbox für AWS-Kunden. Die meisten werden diese Failover-Ereignisse nie bemerken. Kunden, die Echtzeit-Workloads wie Sprache, hochwertige Videos und Messaging mit niedriger Latenz ausführen, fallen diese Ereignisse jedoch manchmal auf. Wie kann ein AWS-Kunde sein eigenes Traffic-Engineering zusätzlich zu den von AWS auf Netzwerkebene bereitgestellten Optionen implementieren? Die Lösung stellt die SIP-Infrastruktur in vielen verschiedenen AWS-Regionen bereit. Mit den Anrufsteuerungsfunktionen bietet SIP auch die Möglichkeit, Anrufe über bestimmte SIP-Proxys weiterzuleiten.

Abbildung 9: Verwenden von SIP-Routing zum Überschreiben des Netzwerkrouting

In Abbildung 9 wird die SIP-Infrastruktur (durch grüne Punkte dargestellt) in allen vier US-Regionen ausgeführt. Die blauen Linien sind eine fiktive Darstellung des AWS-Backbones. Wenn kein SIP-Routing implementiert ist, erfolgt ein Anruf von der Westküste der USA an die US-Ostküste über die Backbone-Verbindung, die die Regionen Oregon und Virginia direkt miteinander verbindet. Das Diagramm zeigt, wie ein Kunde das Routing auf Netzwerkebene überschreiben und denselben Anruf zwischen Oregon und Virginia mithilfe von SIP-Routing über Kalifornien tätigen kann. Diese Art von SIP-Traffic-Engineering kann mithilfe von SIP-Proxys und Mediengateways basierend auf Netzwerkmetriken wie SIP-Neuübertragungen und kundenspezifischen Geschäftspräferenzen implementiert werden.

Durchführen einer detaillierten Überwachung

Endbenutzer von Sprach- und Videoanwendungen in Echtzeit erwarten dasselbe Leistungsniveau wie mit herkömmlichen Telefondiensten. Wenn also Probleme mit einer Anwendung auftreten, schädigt dies dem Ruf des Anbieters. Für ein proaktives anstatt eines reaktiven Managements muss dringend eine detaillierte Überwachung in allen Systembereichen bereitgestellt werden, die Endbenutzer bedienen.

Abbildung 10: Verwenden von SIPp zur Überwachung der VoIP-Infrastruktur

Es gibt viele Open-Source-Tools wie iPerf oder SIPp und VOIPMonitor, um den SIP/RTP-Datenverkehr zu überwachen. Im vorherigen Beispiel messen Knoten, auf denen SIPp im Client- und Servermodus ausgeführt wird, SIP-Metriken wie erfolgreiche Anrufe und SIP-Neuübertragungen zwischen allen vier AWS-Regionen in der USA. Diese Metriken können anschließend mit einem benutzerdefinierten Skript in Amazon CloudWatch exportiert werden. Mit CloudWatch können Kunden Alarme für diese benutzerdefinierten Metriken basierend auf einem bestimmten Schwellenwert erstellen. Abhängig vom Status dieser CloudWatch-Alarme können dann automatische oder manuelle Abhilfemaßnahmen ergriffen werden.

Für Kunden, die keine technischen Ressourcen zuweisen möchten, die für die Entwicklung und Wartung eines benutzerdefinierten Überwachungssystems erforderlich sind, gibt es auf dem Markt viele gute VoIP-Überwachungslösungen wie ThousandEyes. Ein Beispiel für eine Abhilfemaßnahme ist ein geändertes SIP-Routing basierend auf vermehrten SIP-Neuübertragungen.

Verwendung von DNS zur Lastenverteilung und Floating-IPs für Failover

IP-Telefonie-Clients, die DNS-SRV-Funktionen unterstützen, können die integrierte Redundanz der Infrastruktur effizient nutzen, indem sie eine Lastenverteilung der Clients auf verschiedene SBC/PBX-Anlagen vornehmen.

Abbildung 11: Verwenden von DNS-SRV-Datensätzen für eine Lastenverteilung von SIP-Clients

Abbildung 11 zeigt, wie Kunden die SRV-Datensätze für eine Lastenverteilung des SIP-Datenverkehrs verwenden können. Jeder IP-Telefonie-Client, der den SRV-Standard unterstützt, sucht in einem DNS-Datensatz vom Typ SRV nach dem Präfix sip._<transport protocol>. Im Beispiel enthält der Antwortabschnitt von DNS beide Nebenstellenanlagen, die in verschiedenen AWS Availability Zones ausgeführt werden. Neben den Endpunkt-URIs enthält der SRV-Datensatz jedoch drei zusätzliche Informationen:

- Die erste Zahl ist die Priorität (1 im obigen Beispiel). Es wird eine niedrigere Priorität gegenüber einer höheren Priorität vorgezogen.
- Die zweite Zahl ist das Gewicht (10 im obigen Beispiel).
- Und die dritte Zahl ist der zu verwendende Port (5060).

Da beide PBX-Server die gleiche Priorität haben (1), verwenden die Clients das Gewicht für die Lastenverteilung zwischen den beiden Nebenstellenanlagen. Da die Gewichtungen in diesem Fall gleich sind, sollte eine gleichmäßige Lastenverteilung des SIP-Verkehrs zwischen den beiden Nebenstellenanlagen vorgenommen werden.

DNS kann eine gute Lösung für die Lastenverteilung der Clients sein, aber was ist mit der Implementierung von Failovers durch die Änderung/Aktualisierung von DNS-A-Datensätzen? Von dieser Methode wird aufgrund von Inkonsistenzen im DNS-Caching-Verhalten innerhalb des Clients und der Zwischenknoten abgeraten. Ein besserer Ansatz für das Intra-AZ-Failover zwischen einem Cluster von SIP-Knoten ist die Verwendung der EC2-IP-Neuzuweisung, bei der die IP-Adresse eines beeinträchtigten Hosts mithilfe der EC2-API sofort einem fehlerfreien Host zugewiesen wird. In Kombination mit einer Lösung für detaillierte Überwachung und Zustandsprüfungen stellt die IP-Neuzuweisung eines ausgefallenen Knotens sicher, dass der Datenverkehr rechtzeitig auf einen fehlerfreien Host umgeleitet wird, wodurch Unterbrechungen für die Endbenutzer minimiert werden.

Verwendung mehrerer Availability Zones

Jede AWS-Region ist in separate Availability Zones unterteilt. Die einzelnen Availability Zones verfügen über ihre eigene Stromversorgung, Kühlung und Netzwerkkonnektivität und bilden somit eine isolierte Ausfalldomäne. Innerhalb der Konstrukte von AWS wird immer empfohlen, dass Kunden ihre Workloads in mehr als einer Availability Zone ausführen. Dadurch halten Kundenanwendungen auch einem vollständigen Ausfall der Availability Zone stand, obwohl dies nur selten vorkommt. Diese Empfehlung gilt auch für Echtzeit-SIP-Infrastrukturen.

Abbildung 12: Behebung von Ausfällen der Availability Zone

Angenommen, ein katastrophales Ereignis (wie ein Orkan der Kategorie 5) verursacht einen vollständigen Ausfall der Availability Zone in der Region US-Ost-1. Wenn die Infrastruktur wie im Diagramm dargestellt ausgeführt wird, sollten sich alle SIP-Clients, die ursprünglich bei den Knoten in der ausgefallenen Availability Zone registriert waren, erneut bei den SIP-Knoten registrieren, die in Availability Zone 2 ausgeführt werden. (Testen Sie dieses Verhalten mit Ihren SIP-Clients/Telefonen, um sicherzustellen, dass es unterstützt wird.). Obwohl die aktiven SIP-Aufrufe zum Zeitpunkt des Ausfalls der Availability Zone verloren gehen, werden alle neuen Anrufe über Availability Zone 2 geroutet.

Zusammenfassend sollten DNS-SRV-Datensätze den Client auf mehrere A-Datensätze verweisen, einen in jeder Availability Zone. Jeder dieser „A“ -Datensätze sollte wiederum auf mehrere IP-Adressen von SBC/Nebenstellenanlagen in dieser Availability Zone verweisen, die sowohl Intra- als auch Inter-AZ-Ausfallsicherheit bieten. Sowohl Intra- als auch Inter-AZ-Failover können mithilfe der IP-Neuzuweisung implementiert werden, wenn die IP-Adressen öffentlich sind. Private IP-Adressen können jedoch nicht über Availability Zones hinweg neu zugewiesen werden. Wenn ein Kunde private IP-Adressen verwendet, muss er sich darauf verlassen, dass sich die SIP-Clients für das Inter-AZ-Failover erneut bei der Backup-SBC/Nebenstellenanlage registrieren.

Beschränken Sie den Datenverkehr auf eine Availability Zone und verwenden Sie EC2-Placement-Gruppen.

Diese bewährte Methode wird auch als Availability-Zone-Affinität bezeichnet und kann ebenso in dem seltenen Fall eines vollständigen Ausfalls der Availability Zone angewendet werden. Es wird empfohlen, sämtlichen AZ-übergreifenden Datenverkehr zu beseitigen, sodass der SIP- oder RTP-Datenverkehr, der in eine Availability Zone gelangt, in dieser Availability Zone bleibt, bis er die Region verlässt.

Abbildung 13: Affinität der Availability Zone (höchstens 50 % der aktiven Anrufe gehen verloren)

Abbildung 13 zeigt eine vereinfachte Architektur, die die Affinität der Availability Zone verwendet. Der komparative Vorteil dieses Ansatzes wird deutlich, wenn man die Auswirkungen eines vollständigen Ausfalls der Availability Zone bedenkt. Wie im Diagramm dargestellt, sind bei Verlust der Availability Zone 2 höchstens 50 % der aktiven Aufrufe betroffen (vorausgesetzt, es besteht eine identische Lastenverteilung zwischen den Availability Zones). Wäre die Affinität für die Availability Zone nicht implementiert worden, würden sich manche Aufrufe zwischen verschiedenen Availability Zones in einer Region bewegen und bei einem Ausfall wären höchstwahrscheinlich mehr als 50 % der aktiven Anrufe betroffen.

Um die Latenz für den Datenverkehr zu minimieren, empfehlen wir außerdem, EC2-Placement-Gruppen innerhalb jeder Availability Zone zu verwenden. Instances, die innerhalb derselben EC2-Placement-Gruppe gestartet werden, haben eine höhere Bandbreite und eine geringere Latenz, da EC2 die Netzwerknähe dieser Instances zueinander gewährleistet.

Verwenden von EC2-Instance-Typen für Enhanced Networking

Die Systemzuverlässigkeit sowie die effiziente Nutzung der Infrastruktur erfordert die Wahl des richtigen Instance-Typs in Amazon EC2. EC2 bietet eine große Auswahl von Instance-Typen, die für unterschiedliche Anwendungsfälle optimiert sind. Instance-Typen unterstützen verschiedene Kombinationen von CPU, Arbeitsspeicher, Speicher und Netzwerkkapazität. So können Sie flexibel die ideale Ressourcenzusammenstellung für Ihre Anwendungen auswählen. Mit diesen erweiterten Networking-Instance-Typen wird sichergestellt, dass die auf ihnen ausgeführten SIP-Workloads Zugriff auf einheitliche Bandbreite und eine vergleichsweise geringere Gesamtlatenz haben. Eine neue Erweiterung von Amazon EC2 ist der Elastic Network Adapter (ENA), der bis zu 100 Gbit/s Bandbreite bietet. Den neuesten Katalog mit EC2-Instance-Typen und den zugehörigen Funktionen finden Sie auf der Seite EC2-Instance-Typen.

Für die meisten Kunden bietet die neueste Generation von für die Datenverarbeitung optimierten Instances das beste Preis-Leistungs-Verhältnis. Zum Beispiel unterstützt der C5N den neuen Elastic-Network Adapter mit einer Bandbreite von bis zu 100 Gbit/s mit Millionen von Paketen pro Sekunde (PPS). Die meisten Echtzeitanwendungen würden auch von dem Intel Data Plane Developer Kit (DPDK) profitieren, das die Netzwerkpaketverarbeitung erheblich verbessern kann.

Es empfiehlt sich jedoch immer, die verschiedenen EC2-Instance-Typen gemäß Ihren Anforderungen zu bewerten, um den für Sie besten Instance-Typ zu bestimmen. Mit dem Benchmarking können

Sie auch andere Konfigurationsparameter finden, z. B. die maximale Anzahl von Aufrufen, die ein bestimmter Instance-Typ gleichzeitig verarbeiten kann.

Sicherheitsüberlegungen

RTC-Anwendungskomponenten werden normalerweise direkt in mit dem Internet verbundenen Amazon-EC2-Instances ausgeführt. Neben TCP verwenden Abläufe Protokolle wie UDP und SIP. In diesen Fällen schützt AWS Shield Standard Amazon-EC2-Instances vor DDoS-Angriffen auf allgemeiner Infrastrukturebene (Ebene 3 und 4) wie UDP-Reflexionsangriffen, DNS-Reflexion, NTP-Reflexion, SSDP-Reflexion usw. AWS Shield Standard verwendet verschiedene Techniken wie prioritätsbasiertes Traffic Shaping, die automatisch aktiviert werden, wenn eine eindeutig definierte DDoS-Angriffssignatur erkannt wird.

AWS bietet auch erweiterten Schutz vor großen und ausgereiften DDoS-Angriffen für diese Anwendungen, indem AWS Shield Advanced auf elastischen IP-Adressen aktiviert wird. AWS Shield Advanced bietet eine erweiterte DDoS-Erkennung, die automatisch den Typ der AWS-Ressource und die Größe der EC2-Instance erkennt und geeignete vordefinierte Abwehrmaßnahmen zum Schutz vor SYN- oder UDP-Floods anwendet. Mit AWS Shield Advanced können Kunden auch ihre eigenen benutzerdefinierten Abwehrprofile erstellen. Sie können dazu rund um die Uhr das AWS DDoS Response Team (DRT) kontaktieren. AWS Shield Advanced stellt außerdem sicher, dass bei einem DDoS-Angriff alle Ihre Amazon VPC Network Zugriffskontrolllisten (ACLs) automatisch am Rand des AWS-Netzwerks durchgesetzt werden. Dadurch erhalten Sie zusätzliche Bandbreite und Scrubbing-Kapazität, um große volumetrische DDoS-Angriffe abzuwehren.

Fazit

Workloads für die Echtzeitkommunikation (RTC) können in Amazon Web Services (AWS) bereitgestellt werden, um Skalierbarkeit, Elastizität und Hochverfügbarkeit zu erreichen und gleichzeitig die wichtigsten Anforderungen zu erfüllen. Heutzutage nutzen viele Kunden Lösungen von AWS, seinen Partnern sowie Open Source-Lösungen, um RTC-Workloads mit eingeschränkter globaler Präsenz günstiger und agiler auszuführen.

Mit den in diesem Whitepaper enthaltenen Referenzarchitekturen und bewährten Methoden können Kunden RTC-Workloads in AWS erfolgreich einrichten und die Lösungen optimieren, um sie auf die Cloud vorzubereiten und gleichzeitig die Anforderungen der Endbenutzer zu erfüllen.

Mitwirkende

Dieses Dokument ist unter der Mitarbeit folgender Personen und Unternehmen entstanden:

- Ahmad Khan, Leitender Lösungsarchitekt, Amazon Web Services
- Tipu Qureshi, Chefingenieur, AWS Support, Amazon Web Services
- Paul Moran, Leitender Technical Account Manager, Amazon Web Services
- Shoma Chakravarty, Technischer Leiter für WW, Telekommunikation, Amazon Web Services

Dokumentversionen

Abonnieren Sie den RSS-Feed, um über Aktualisierungen des Whitepapers benachrichtigt zu werden.

Update-Historie-Änderung	Update-Historie-Beschreibung	Update-Historie-Datum
Whitepaper aktualisiert	Für die neuesten Services und Funktionen aktualisiert.	13. Februar 2020
Erste Veröffentlichung	Erstveröffentlichung des Whitepapers.	1. Oktober 2018

Hinweise

Kunden sind eigenverantwortlich für die unabhängige Bewertung der Informationen in diesem Dokument zuständig. Dieses Dokument: (a) dient rein zu Informationszwecken, (b) spiegelt die aktuellen Produktangebote und Verfahren von AWS wider, die sich ohne vorherige Mitteilung ändern können, und (c) impliziert keinerlei Verpflichtungen oder Zusicherungen seitens AWS und dessen Tochtergesellschaften, Lieferanten oder Lizenzgebern. AWS-Produkte oder -Services werden im vorliegenden Zustand und ohne ausdrückliche oder stillschweigende Gewährleistungen, Zusicherungen oder Bedingungen bereitgestellt. Die Verantwortung und Haftung von AWS gegenüber seinen Kunden wird durch AWS-Vereinbarungen geregelt. Dieses Dokument ist weder ganz noch teilweise Teil der Vereinbarungen zwischen AWS und seinen Kunden und ändert diese Vereinbarungen auch nicht.