AWS-Whitepaper

Kommunikation in Echtzeit auf AWS

Kommunikation in Echtzeit auf AWS: AWS-Whitepaper

Table of Contents

Kommunikation in Echtzeit auf AWS

Bewährte Methoden für den Entwurf hochverfügbarer und skalierbarer Workloads für Echtzeitkommunikation (RTC) auf AWS

Datum der Veröffentlichung: 5. Mai 2022 () Dokumentversionen

Überblick

Heutzutage sind viele Unternehmen bestrebt, die Kosten zu senken und Skalierbarkeit für Sprach-, Messaging- und Multimedia-Workloads in Echtzeit zu erreichen. In diesem paper werden die bewährten Methoden für die Verwaltung von Workloads für Echtzeitkommunikation (RTC) auf Amazon Web Services (AWS) beschrieben und Referenzarchitekturen zur Erfüllung dieser Anforderungen vorgestellt. Dieses paper dient Personen, die mit Echtzeitkommunikation vertraut sind, als Leitfaden, um eine hohe Verfügbarkeit und Skalierbarkeit für diese Workloads zu erreichen.

Dieses paper enthält Referenzarchitekturen, die zeigen, wie RTC-Workloads eingerichtet werden AWS, sowie bewährte Methoden zur Optimierung der Lösungen, um die Anforderungen der Endbenutzer zu erfüllen und gleichzeitig für die Cloud zu optimieren. Der Evolved Packet Core (EPC) ist in diesem Whitepaper nicht enthalten, aber die hier beschriebenen Best Practices können auf virtuelle Netzwerkfunktionen () angewendet werden. VNFs

Sind Sie Well-Architected?

Das AWS Well-Architected Framework hilft Ihnen dabei, die Vor- und Nachteile der Entscheidungen zu verstehen, die Sie beim Aufbau von Systemen in der Cloud treffen. Die sechs Säulen des Frameworks ermöglichen es Ihnen, bewährte Architekturpraktiken für den Entwurf und Betrieb zuverlässiger, sicherer, effizienter, kostengünstiger und nachhaltiger Systeme kennenzulernen. Mithilfe des AWS Well-Architected Tool, das kostenlos im Abschnitt verfügbar ist AWS Management Console(Anmeldung erforderlich), können Sie Ihre Workloads anhand dieser bewährten Methoden überprüfen, indem Sie für jede Säule eine Reihe von Fragen beantworten.

Weitere Expertentipps und bewährte Methoden für Ihre Cloud-Architektur — Referenzarchitekturbereitstellungen, Diagramme und Whitepapers — finden Sie im Architecture Center.AWS

Einführung

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

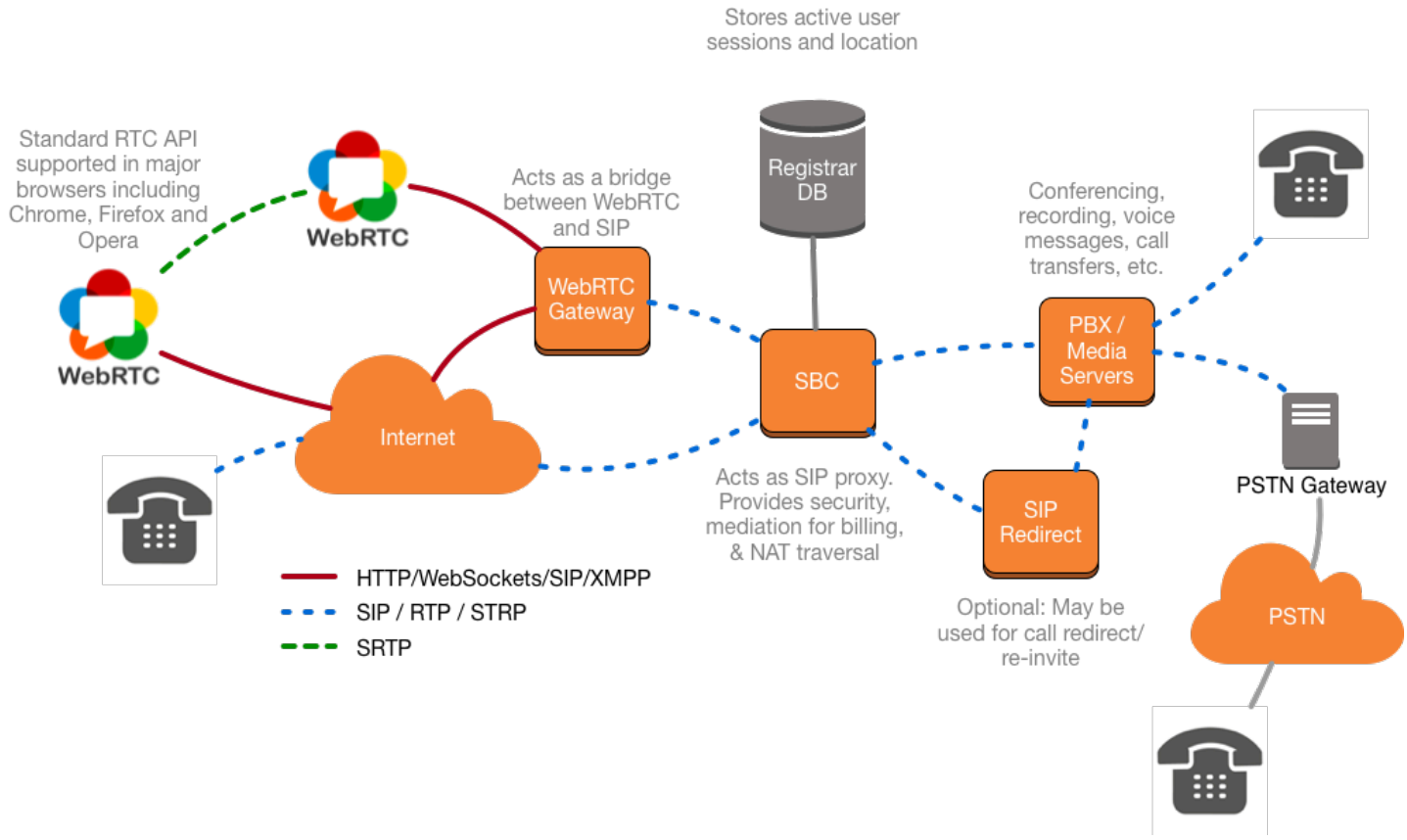
RTC-Workloads erfordern eine hochgradig skalierbare, belastbare und verfügbare Umgebung. Heute nutzen AWS Kunden RTC-Workloads mit geringeren Kosten, verbesserter Agilität, Flexibilität und Markteinführungszeit.

Grundlegende Komponenten der RTC-Architektur

In der Telekommunikationsbranche bezieht sich RTC üblicherweise auf Live-Mediensitzungen zwischen zwei Endpunkten mit minimaler Latenz. Diese Sitzungen könnten sich auf Folgendes beziehen:

- Eine Sprachsitzung zwischen zwei Parteien (z. B. eine Telefonanlage, ein Mobiltelefon oder Voice over IP (VoIP))
- Sofortnachrichten (wie Chatten und Instant Relay Chat (IRC))
- Live-Videositzung (wie Videokonferenzen und Telepräsenz)

Jede der oben genannten Lösungen hat einige Komponenten gemeinsam (z. B. Komponenten für Authentifizierung, Autorisierung und Zugriffskontrolle, Transcodierung, Pufferung und Relay usw.) und einige Komponenten, die für den Typ der übertragenen Medien spezifisch sind (z. B. Broadcast-Service, Messaging-Server und Warteschlangen usw.). Dieser Abschnitt konzentriert sich auf die Definition eines sprach- und videobasierten RTC-Systems und aller zugehörigen Komponenten, wie in der folgenden Abbildung dargestellt.



Wesentliche Architekturkomponenten für RTC

SoftSwitch/PBX

Ein Softswitch oder PBX ist das Gehirn eines Sprachtelefonsystems und stellt mithilfe verschiedener Komponenten Informationen für die Einrichtung, Aufrechterhaltung und Weiterleitung eines Sprachanrufs innerhalb oder außerhalb des Unternehmens bereit. Alle Teilnehmer des Unternehmens müssen sich beim Softswitch registrieren, um einen Anruf entgegennehmen oder tätigen zu können. Eine wichtige Funktion des Softswitches besteht darin, den Überblick über jeden Teilnehmer zu behalten und zu erfahren, wie 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 Verkehr (sowohl Kontroll- als auch Datenebene). Eine der Hauptaufgaben eines SBC besteht darin, das Sprachsystem vor böswilliger Nutzung zu schützen. Der SBC kann zur Verbindung mit SIP-Trunks (Session Initiation Protocol) für externe Konnektivität verwendet werden. Einige bieten SBCs auch Transcodierungsfunktionen für die Konvertierung CODECsvon einem Format in ein anderes. Die meisten bieten SBCs auch NAT-Traversal-Funktionen (Network Address Translation), mit deren Hilfe sichergestellt werden kann, dass Anrufe auch in Netzwerken mit Firewalls hergestellt werden.

PSTN-Konnektivität

Voice over IP (VoIP) -Lösungen verwenden Public Switched Telephone Network (PSTN) -Gateways und SIP-Trunks, um eine Verbindung mit älteren PSTN-Netzwerken herzustellen.

PSTN-Gateway

Das PSTN-Gateway konvertiert die Signalisierung zwischen SIP SS7 und Medien zwischen Real Time Transport Protocol (RTP) und Time Division Multiplexing (TDM) mithilfe von CODEC-Transcodierung. PSTN-Gateways befinden sich immer am Rand in der Nähe des PSTN-Netzwerks.

SIP-Trunk

Bei einem SIP-Trunk leitet das Unternehmen seine Anrufe nicht an ein (auf TDM SS7 basierendes) Netzwerk weiter, sondern der Datenfluss zwischen dem Unternehmen und dem

Telekommunikationsunternehmen erfolgt weiterhin über IP. Die meisten SIP-Trunks werden mithilfe von eingerichtet. SBCs Das Unternehmen muss sich auf die vordefinierten Sicherheitsregeln der Telekommunikationsunternehmen einigen, z. B. die Zulassung eines bestimmten Bereichs von IP-Adressen, Ports usw.

Media Gateway (Transcoder)

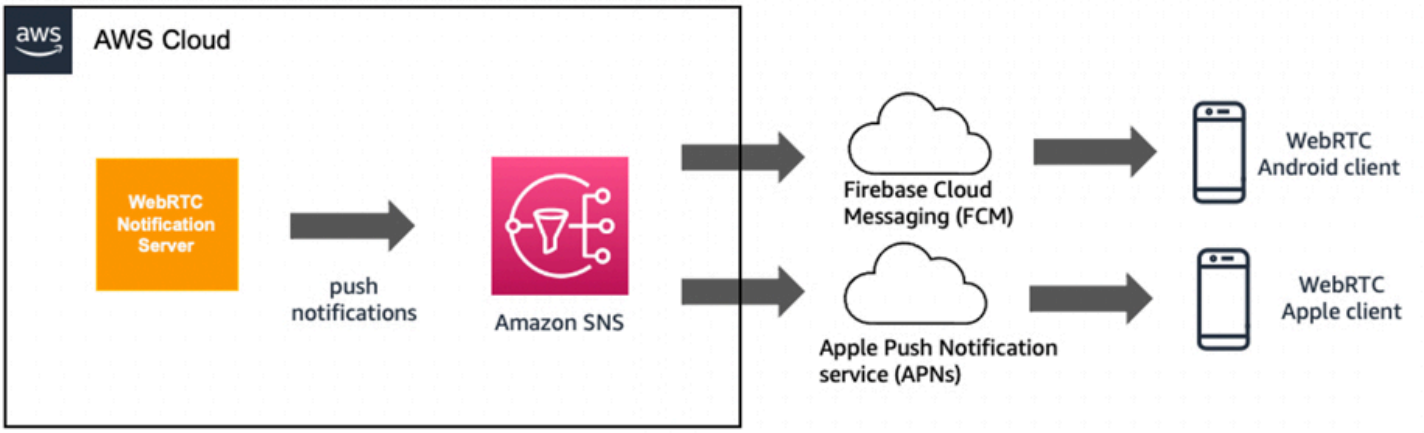
Benutzer kommunizieren in Echtzeit mithilfe von Audio und/oder Video sowie optionalen Daten und anderen Informationen. Für die Kommunikation müssen sich die beiden Geräte auf einen für beide Seiten verständlichen Codec für jede Medienspur einigen, damit sie erfolgreich kommunizieren und die gemeinsam genutzten Medien präsentieren können. Alle WebRTC-kompatiblen Browser müssen Online Positioning User Support (OPUS) und G711 für Audio und das H.264 Constrained Baseline-Profil für VP8Video unterstützen.

Eine typische Sprachlösung außerhalb des WebRTC-Ökosystems ermöglicht verschiedene Arten von. CODECs Einige der gebräuchlichsten CODECs sind G.711 μ-Law für Nordamerika, G.711 A-law, G.729 und G.722. Wenn zwei Geräte, die zwei unterschiedliche Geräte verwenden, CODECs miteinander kommunizieren, übersetzt das Media Gateway den CODEC-Fluss zwischen den Geräten. Mit anderen Worten, ein Media Gateway verarbeitet Medien und stellt sicher, dass die Endgeräte miteinander kommunizieren können.

Push-Benachrichtigungen in WebRTC

WebRTC-Implementierungen sind auf Mobilgeräten sehr verbreitet. Im Gegensatz zu Webbrowsern kann ein Mobilgerät eine Websocket-Konnektivität nicht lange aufrechterhalten. Daher muss es sich für alle endenden Anfragen, wie Anrufe und Nachrichten, auf Push-Benachrichtigungen vom WebRTC-Server verlassen.

Mit Amazon Simple Notification Service (Amazon SNS) können Sie Push-Benachrichtigungen an Apps auf Mobilgeräten senden. Diese Apps könnten auf verschiedenen Betriebssystemen wie Apple iOS oder Android laufen. Die folgende Abbildung zeigt einen allgemeinen Überblick über den Fluss von Push-Benachrichtigungen, von einem WebRTC-Benachrichtigungsserver zu mobilen WebRTC-Endpunkten.

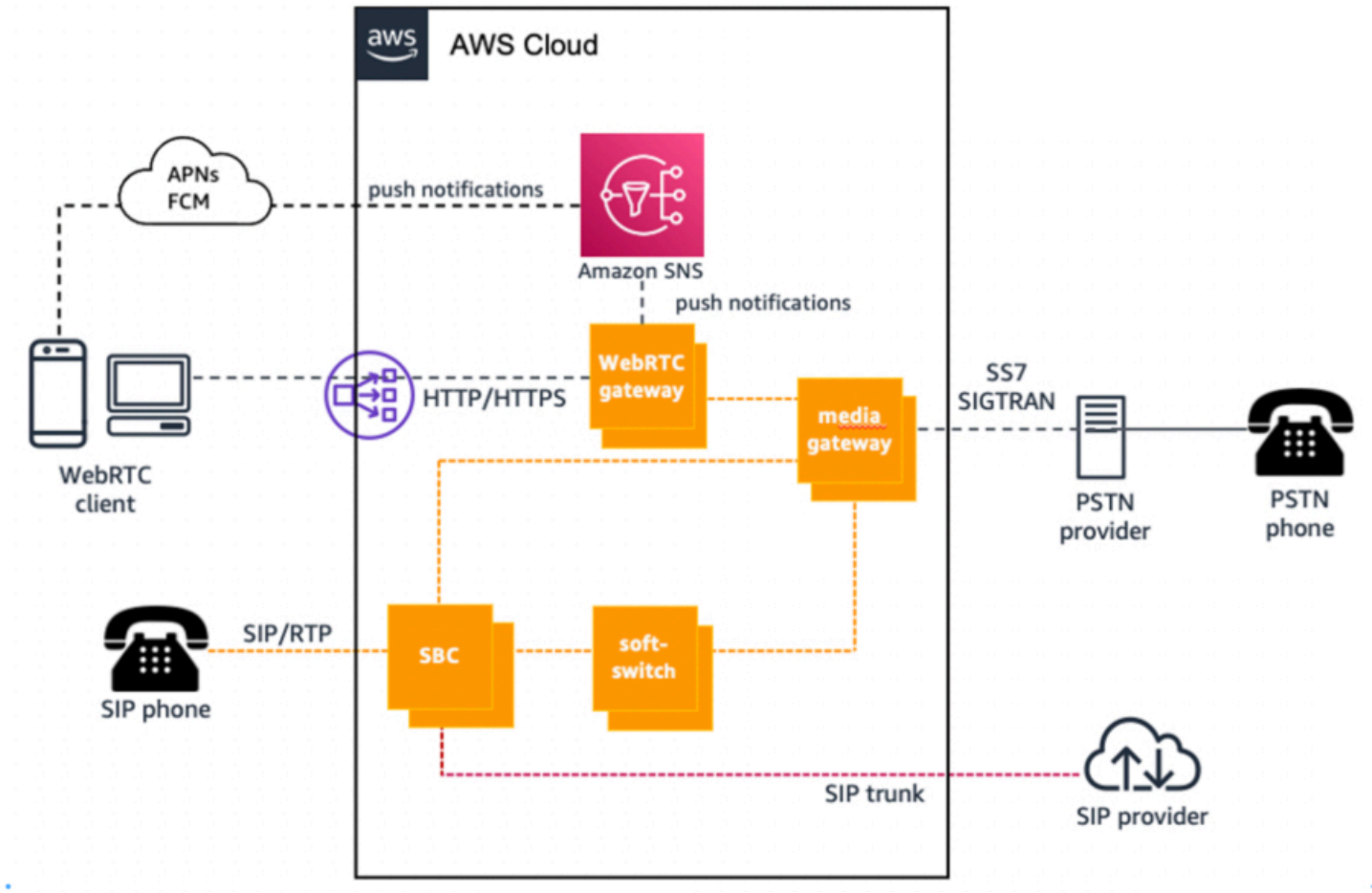


Amazon SNS für Push-Benachrichtigungen

WebRTC und WebRTC-Gateway

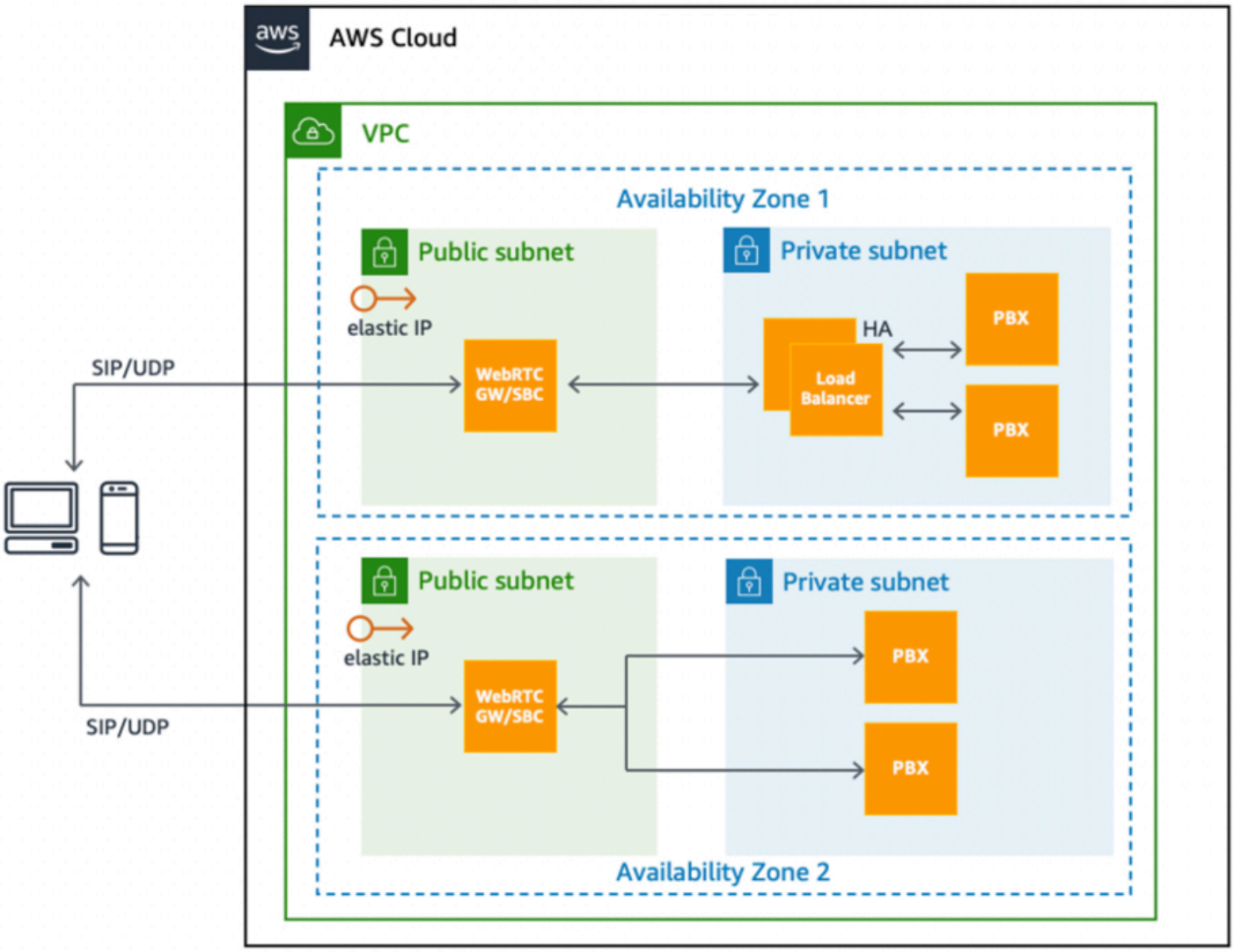
Mit Web Real-Time Communication (WebRTC) können Sie mithilfe der API einen Anruf von einem Webbrowser aus einrichten oder Ressourcen vom Backend-Server anfordern. Die Technologie wurde unter Berücksichtigung der Cloud-Technologie entwickelt und bietet daher verschiedene Funktionen, mit APIs denen ein Anruf hergestellt werden kann. Da nicht alle Sprachlösungen (einschließlich SIP) diese unterstützen APIs, muss das WebRTC-Gateway API-Aufrufe in SIP-Nachrichten übersetzen und umgekehrt.

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



Eine grundlegende Topologie eines RTC-Systems für Sprache

Ein weiteres Entwurfsmuster für SIP- und RTP-Verkehr besteht darin, Paare von SBCs auf Amazon EC2 im Aktiv-Passiv-Modus über Availability Zones hinweg zu verwenden, wie in der folgenden Abbildung dargestellt. Hier kann eine Elastic IP-Adresse bei einem Ausfall dynamisch zwischen Instances verschoben werden, wobei der Domain Name Service (DNS) nicht verwendet werden kann.



RTC-Architektur mit Amazon EC2 in einer Virtual Private Cloud (VPC)

Hohe Verfügbarkeit und Skalierbarkeit auf AWS

Die meisten Anbieter von Echtzeitkommunikation orientieren sich an Service Levels, die eine Verfügbarkeit von 99,9% bis 99,999% bieten. Je nachdem, welchen Grad an Hochverfügbarkeit (HA) Sie wünschen, müssen Sie während des gesamten Lebenszyklus der Anwendung immer ausgefeiltere Maßnahmen ergreifen. AWS empfiehlt, die folgenden Richtlinien zu befolgen, um ein stabiles Maß an Hochverfügbarkeit zu erreichen:

- Entwerfen Sie das System so, dass es keine einzige Fehlerquelle gibt. Verwenden Sie automatische Überwachungs-, Fehlererkennungs- und Failover-Mechanismen sowohl für statusfreie als auch für zustandsbehaftete Komponenten
  - Single Points of Failure (SPOF) werden in der Regel durch eine N+1- oder 2N-Redundanzkonfiguration vermieden, wobei N+1 durch Lastenausgleich zwischen aktiv-aktiven Knoten und 2N durch ein Knotenpaar in einer Active-Standby-Konfiguration erreicht wird.
  - AWS bietet mehrere Methoden, um HA mit beiden Ansätzen zu erreichen, z. B. durch einen skalierbaren Cluster mit Lastenausgleich oder die Annahme eines Aktiv-Standby-Paars.
- Richtiges Instrumentieren und Testen der Systemverfügbarkeit.
- Bereiten Sie Betriebsverfahren für manuelle Mechanismen vor, um auf den Ausfall zu reagieren, ihn zu mindern und ihn zu beheben.

Dieser Abschnitt konzentriert sich darauf, wie mithilfe der verfügbaren Funktionen erreicht werden kann, dass kein einziger Ausfallpunkt erreicht wird. AWS In diesem Abschnitt wird insbesondere ein Teil der AWS Kernfunktionen und Entwurfsmuster beschrieben, mit denen Sie hochverfügbare Echtzeitkommunikationsanwendungen erstellen können.

Floating-IP-Muster für HA zwischen aktiven und statusbehafteten Standby-Servern

Das Floating-IP-Entwurfsmuster ist ein bekannter Mechanismus, um einen automatischen Failover zwischen einem aktiven und einem Standby-Paar von Hardwareknoten (Medienservern) zu erreichen. Dem aktiven Knoten wird eine statische sekundäre virtuelle IP-Adresse zugewiesen. Durch die kontinuierliche Überwachung zwischen dem aktiven Knoten und dem Standby-Knoten wird ein Fehler erkannt. Wenn der aktive Knoten ausfällt, weist das Überwachungsskript die virtuelle IP dem bereiten Standby-Knoten zu und der Standby-Knoten übernimmt die primäre aktive Funktion. Auf diese Weise schwebt die virtuelle IP zwischen dem aktiven und dem Standby-Knoten.

Anwendbarkeit in RTC-Lösungen

Es ist nicht immer möglich, mehrere aktive Instanzen derselben Komponente in Betrieb zu haben, z. B. ein aktiv-aktives Cluster mit N Knoten. Eine Aktiv-Standby-Konfiguration bietet den besten Mechanismus für HA. Beispielsweise eignen sich die statusbehafteten Komponenten in einer RTC-Lösung, wie der Medienserver oder Konferenzserver oder sogar ein SBC- oder Datenbankserver, gut für eine Active-Standby-Konfiguration. Auf einem SBC- oder Medienserver sind zu einem bestimmten Zeitpunkt mehrere lang laufende Sitzungen oder Kanäle aktiv. Falls die aktive SBC-Instance ausfällt, können sich die Endpunkte aufgrund der Floating-IP ohne clientseitige Konfiguration wieder mit dem Standby-Knoten verbinden.

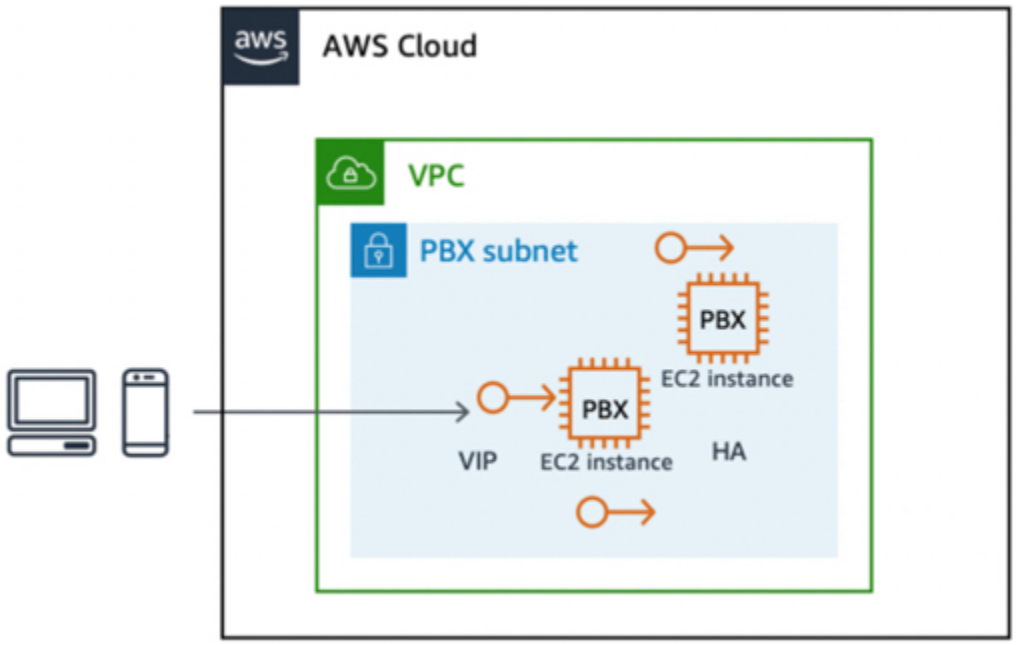
Implementierung am AWS

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

Um das Floating-IP-Muster zu implementieren auf AWS:

1. Starten Sie zwei EC2 Instances, um die Rollen des primären und des sekundären Knotens zu übernehmen, wobei davon ausgegangen wird, dass sich der primäre Knoten standardmäßig im aktiven Zustand befindet.
2. Weisen Sie der primären EC2 Instanz eine zusätzliche sekundäre private IP-Adresse zu.
3. Eine elastische IP-Adresse, die einer virtuellen IP (VIP) ähnelt, ist der sekundären privaten Adresse zugeordnet. Diese sekundäre private Adresse ist die Adresse, die von externen Endpunkten für den Zugriff auf die Anwendung verwendet wird.
4. Eine Konfiguration des Betriebssystems (OS) ist erforderlich, damit die sekundäre IP-Adresse als Alias zur primären Netzwerkschnittstelle hinzugefügt wird.
5. Die Anwendung muss an diese elastische IP-Adresse gebunden werden. Bei der Asterisk-Software können Sie die Bindung über erweiterte Asterisk-SIP-Einstellungen konfigurieren.
6. Führen Sie auf jedem Knoten ein Überwachungsskript (benutzerdefiniert, KeepAlive unter Linux, Corosync usw.) aus, um den Status des Peer-Knotens zu überwachen. Falls der aktuell aktive Knoten ausfällt, erkennt der Peer diesen Fehler und ruft die EC2 Amazon-API auf, um sich selbst die sekundäre private IP-Adresse neu zuzuweisen.

   Daher wird die Anwendung, die den mit der sekundären privaten IP-Adresse verknüpften VIP abgehört hat, für Endgeräte über den Standby-Knoten verfügbar.



Failover zwischen EC2 Stateful-Instances, die eine elastische IP-Adresse verwenden

Vorteile

Dieser Ansatz ist eine zuverlässige Low-Budget-Lösung, die vor Ausfällen auf EC2 Instanz-, Infrastruktur- oder Anwendungsebene schützt.

Einschränkungen und Erweiterbarkeit

Dieses Entwurfsmuster ist in der Regel auf eine einzelne Availability Zone beschränkt. Es kann in zwei Availability Zones implementiert werden, jedoch mit einer Variation. In diesem Fall wird die Floating Elastic IP-Adresse über die verfügbare Elastic IP Address API zwischen aktivem Knoten und Standby-Knoten in verschiedenen Availability Zones neu zugeordnet. Bei der in der vorherigen Abbildung gezeigten Failover-Implementierung werden laufende Anrufe gelöscht und die Endgeräte müssen erneut eine Verbindung herstellen. Es ist möglich, diese Implementierung um die Replikation der zugrunde liegenden Sitzungsdaten zu erweitern, um ein nahtloses Failover der Sitzungen oder auch die Medienkontinuität zu gewährleisten.

Lastenausgleich für Skalierbarkeit und HA mit WebRTC und SIP

Der Lastenausgleich eines Clusters aktiver Instances auf der Grundlage vordefinierter Regeln wie Round-Robin, Affinität oder Latenz usw. ist ein Entwurfsmuster, das aufgrund der statusfreien Natur von HTTP-Anfragen weit verbreitet ist. Tatsächlich ist der Lastenausgleich bei vielen RTC-Anwendungskomponenten eine praktikable Option.

Der Load Balancer fungiert als Reverse-Proxy oder als Einstiegspunkt für Anfragen an die gewünschte Anwendung, die selbst so konfiguriert ist, dass sie auf mehreren aktiven Knoten gleichzeitig ausgeführt wird. Zu einem beliebigen Zeitpunkt leitet der Load Balancer eine Benutzeranfrage an einen der aktiven Knoten im definierten Cluster weiter. Load Balancer führen eine Integritätsprüfung für die Knoten in ihrem Zielcluster durch und senden keine eingehende Anfrage an einen Knoten, der die Zustandsprüfung nicht besteht. Daher wird durch den Lastenausgleich ein grundlegender Grad an Hochverfügbarkeit erreicht. Da ein Load Balancer aktive und passive Integritätsprüfungen für alle Clusterknoten in Intervallen von weniger als einer Sekunde durchführt, erfolgt der Failover zudem fast augenblicklich.

Die Entscheidung, welcher Knoten geleitet werden soll, basiert auf den im Load Balancer definierten Systemregeln, darunter:

- Rundenturnier
- Sitzungs- oder IP-Affinität, wodurch sichergestellt wird, dass mehrere Anfragen innerhalb einer Sitzung oder von derselben IP an denselben Knoten im Cluster gesendet werden
- Basierend auf Latenz
- Lastbasiert

Anwendbarkeit in RTC-Architekturen

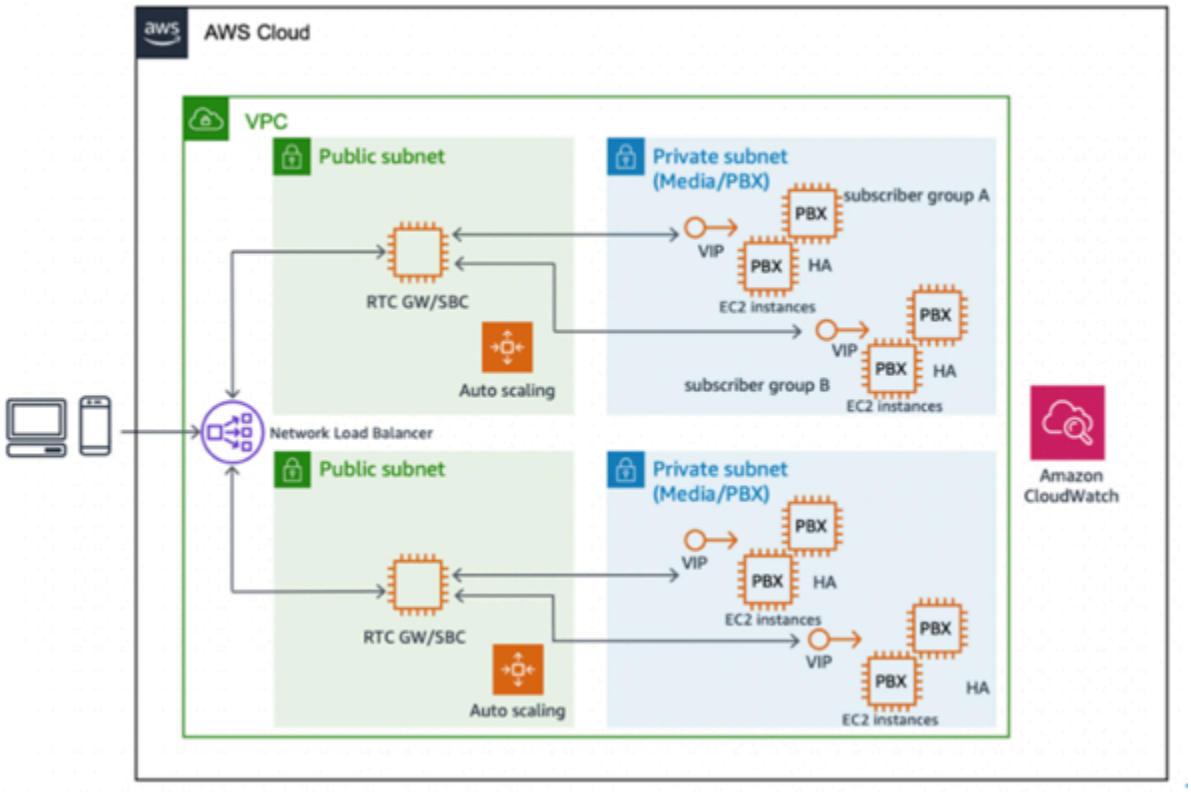
Das WebRTC-Protokoll ermöglicht den einfachen Lastenausgleich von WebRTC-Gateways über einen HTTP-basierten Load Balancer wie Elastic Load Balancing (ELB), Application Load Balancer (ALB) oder Network Load Balancer (NLB). Da die meisten SIP-Implementierungen auf den Transport sowohl über das Transmission Control Protocol (TCP) als auch über das User Datagram Protocol (UDP) angewiesen sind, benötigen Sie einen Lastenausgleich auf Netzwerk- oder Verbindungsebene mit Unterstützung für TCP- und UDP-basierten Datenverkehr.

Load Balancing AWS für WebRTC mit Application Load Balancer und Auto Scaling aktiviert

Im Fall von WebRTC-basierter Kommunikation bietet Elastic Load Balancing einen vollständig verwalteten, hochverfügbaren und skalierbaren Load Balancer, der als Einstiegspunkt für Anfragen dient, die dann an einen Zielcluster von EC2 Instances weitergeleitet werden, die mit Elastic Load Balancing verknüpft sind. Da WebRTC-Anfragen zustandslos sind, können Sie Amazon EC2 Auto Scaling verwenden, um vollautomatische und kontrollierbare Skalierbarkeit, Elastizität und Hochverfügbarkeit bereitzustellen.

Der Application Load Balancer bietet einen vollständig verwalteten Lastausgleichsdienst, der über mehrere Availability Zones hochverfügbar und skalierbar ist. Dies unterstützt den Lastenausgleich von WebSocket Anfragen, die die Signalisierung für WebRTC-Anwendungen verarbeiten, und die bidirektionale Kommunikation zwischen dem Client und dem Server über eine lang andauernde TCP-Verbindung. Der Application Load Balancer unterstützt auch inhaltsbasiertes Routing und Sticky-Sessions, bei denen Anfragen von demselben Client mithilfe von Load Balancer-generierten Cookies an dasselbe Ziel weitergeleitet werden. Wenn Sie Sticky Sessions aktivieren, empfängt dasselbe Ziel die Anfrage und kann das Cookie verwenden, um den Sitzungskontext wiederherzustellen.

Die folgende Abbildung zeigt die Zieltopologie.



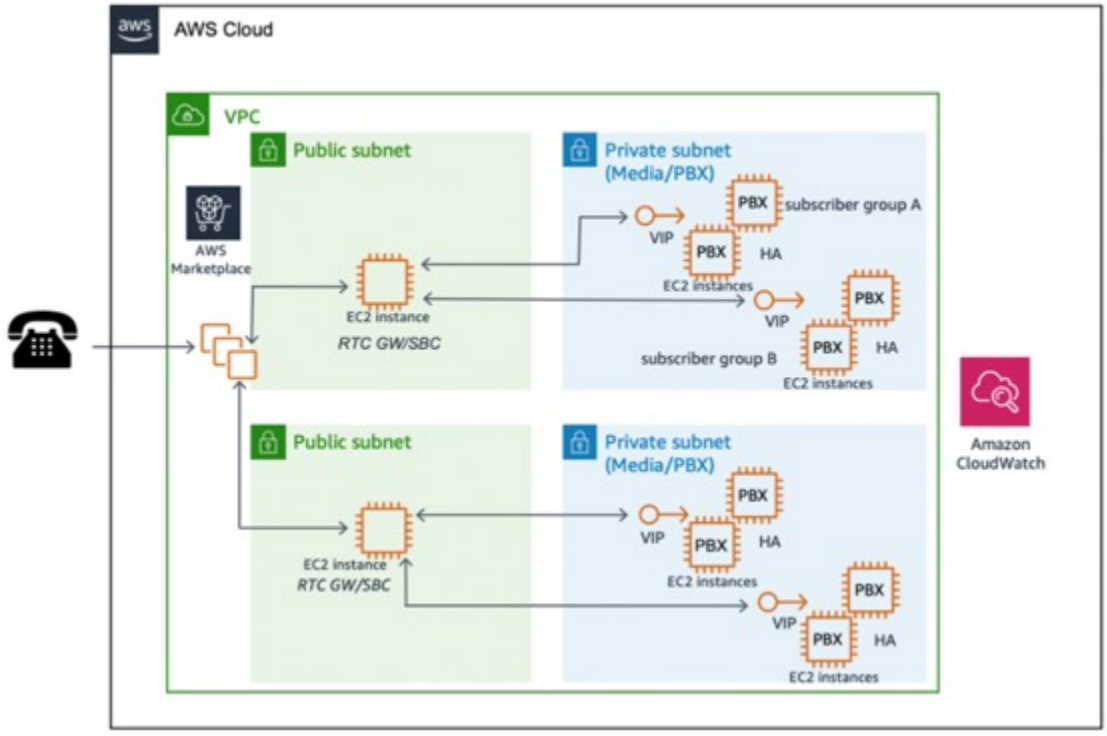
WebRTC-Skalierbarkeit und Hochverfügbarkeitsarchitektur

Implementierung für SIP mit Network Load Balancer oder einem Produkt AWS Marketplace

Bei SIP-basierter Kommunikation werden die Verbindungen über TCP oder UDP hergestellt, wobei die meisten RTC-Anwendungen UDP verwenden. Wenn SIP/TCP das Signalprotokoll der Wahl ist, ist es möglich, den Network Load Balancer für einen vollständig verwalteten, hochverfügbaren, skalierbaren und leistungsfähigen Lastenausgleich zu verwenden.

Ein Network Load Balancer arbeitet auf der Verbindungsebene (Layer 4) und leitet Verbindungen zu Zielen wie EC2 Amazon-Instances, Containern und IP-Adressen auf der Grundlage von IP-Protokolldaten weiter. Der Netzwerklastenausgleich eignet sich ideal für den Lastenausgleich des TCP- oder UDP-Datenverkehrs und ist in der Lage, Millionen von Anfragen pro Sekunde zu verarbeiten und gleichzeitig extrem niedrige Latenzen aufrechtzuerhalten. Es ist in andere beliebte AWS-Services wie Amazon EC2 Auto Scaling, Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) und integriert. AWS CloudFormation

Wenn SIP-Verbindungen initiiert werden, besteht eine weitere Option darin, AWS Marketplacekommerzielle off-the-shelf Software (COTS) zu verwenden. The AWS Marketplace bietet viele Produkte, die mit UDP und anderen Arten des Lastenausgleichs von Layer-4-Verbindungen umgehen können. COTS bieten in der Regel Unterstützung für Hochverfügbarkeit und lassen sich häufig in Funktionen wie Amazon EC2 Auto Scaling integrieren, um die Verfügbarkeit und Skalierbarkeit weiter zu verbessern. Die folgende Abbildung zeigt die Zieltopologie:



SIP-basierte RTC-Skalierbarkeit mit dem Produkt AWS Marketplace

Regionsübergreifender DNS-basierter Lastenausgleich 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 so konfiguriert werden, dass sie den Datenverkehr an fehlerfreie Endpunkte weiterleiten oder den Zustand Ihrer Anwendung unabhängig überwachen.

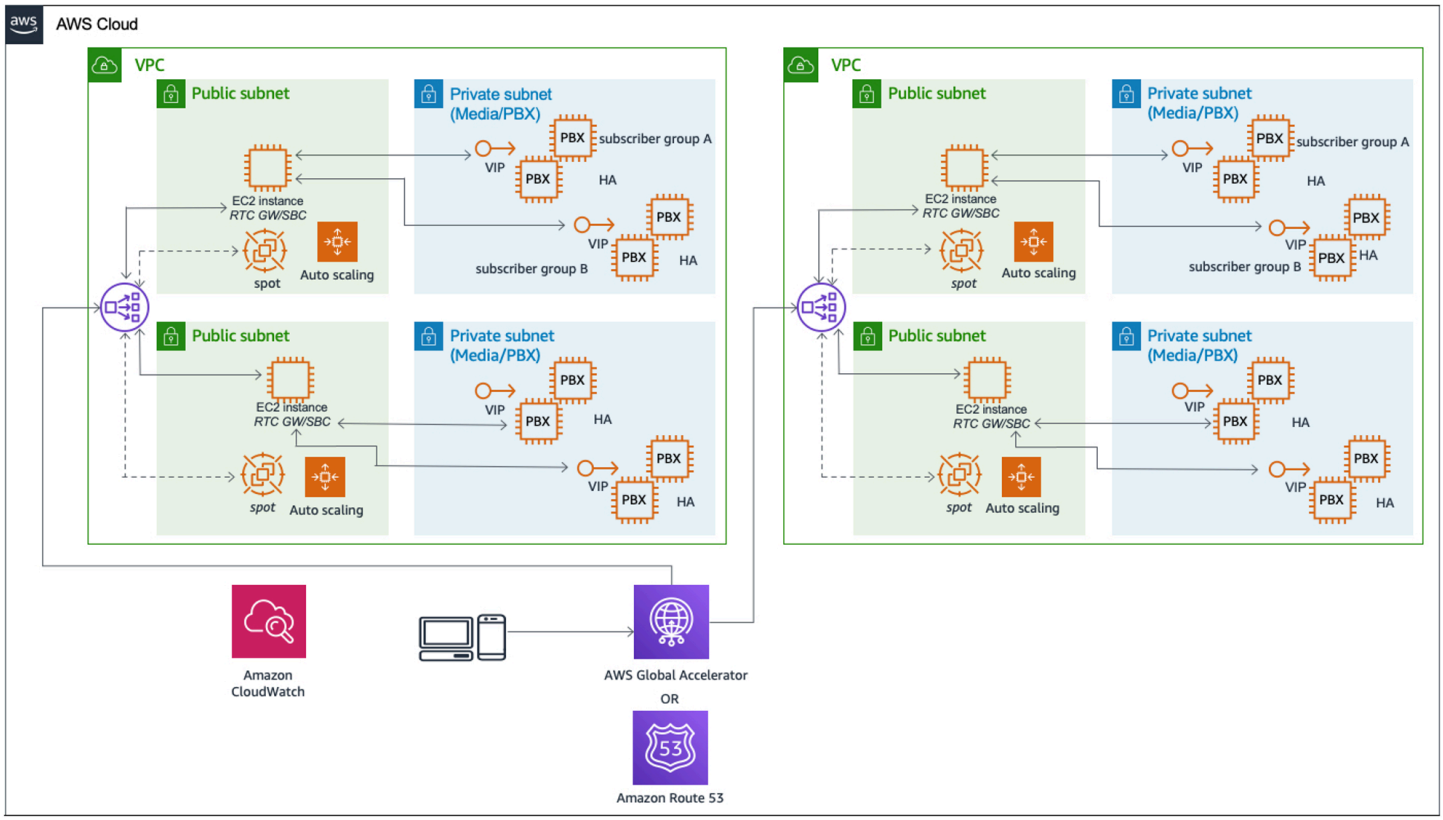
Die Amazon Route 53 Traffic Flow-Funktion erleichtert Ihnen die globale Verwaltung des Datenverkehrs mithilfe einer Vielzahl von Routingtypen, darunter latenzbasiertes Routing, Geo-DNS, Geoproximity und Weighted Round Robin. All diese Optionen können mit DNS-Failover kombiniert werden, um eine Vielzahl von fehlertoleranten Architekturen mit niedriger Latenz zu ermöglichen. Mit dem einfachen visuellen Editor von Amazon Route 53 Traffic Flow können Sie verwalten, wie Ihre Endbenutzer zu den Endpunkten Ihrer Anwendung weitergeleitet werden — ob in einer einzelnen AWS-Region oder auf der ganzen Welt verteilt.

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 die Client-Caches geleert werden müssen, um einen Failover auf eine neue DNS-Adresse zu erzwingen. Außerdem kann es bei DNS-Änderungen zu Verzögerungen kommen, wenn sie über globale DNS-Server verteilt werden. Sie können das Aktualisierungsintervall für DNS-Lookups mit dem Time to Live-Attribut verwalten. Dieses Attribut kann zum Zeitpunkt der Einrichtung von DNS-Richtlinien konfiguriert werden.

AWS Global Accelerator Kann auch für regionsübergreifendes Failover verwendet werden, um globale Benutzer schnell zu erreichen oder um die Anforderungen der Verwendung einer einzigen öffentlichen IP zu erfüllen. AWS Global Acceleratorist ein Netzwerkdienst, der die Verfügbarkeit und Leistung von Anwendungen mit lokaler und globaler Reichweite verbessert. AWS Global Accelerator stellt statische IP-Adressen bereit, die als fester Einstiegspunkt zu Ihren Anwendungsendpunkten dienen, wie z. B. Ihren Application Load Balancers, Network Load Balancers oder EC2 Amazon-Instances in einer oder mehreren AWS-Regionen. Es nutzt das globale AWS-Netzwerk, um den Pfad von Ihren Benutzern zu Ihren Anwendungen zu optimieren und so die Leistung zu verbessern, 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 nächstgelegenen fehlerfreien Endpunkte weiter, falls die aktuellen Endgeräte nicht mehr richtig funktionieren. Für zusätzliche Sicherheitsanforderungen verwendet Accelerated Site-to-Site VPN, AWS Global Accelerator um die Leistung von VPN-Verbindungen zu verbessern, indem der Datenverkehr intelligent über das globale AWS-Netzwerk und die AWS-Edge-Standorte geleitet wird.



Regionsübergreifendes Hochverfügbarkeitsdesign mit AWS Global Accelerator oder Amazon Route 53

Datenbeständigkeit und HA mit persistentem Speicher

Die meisten RTC-Anwendungen verlassen sich auf persistenten Speicher, um Daten für Authentifizierung, Autorisierung, Abrechnung (Sitzungsdaten, Anrufdetails usw.), Betriebsüberwachung und Protokollierung zu speichern und darauf zuzugreifen. In einem herkömmlichen Rechenzentrum erfordert die Sicherstellung einer hohen Verfügbarkeit und Beständigkeit der persistenten Speicherkomponenten (Datenbanken, Dateisysteme usw.) in der Regel viel Arbeit. Dazu gehören die Einrichtung eines Storage Area Network (SAN), das RAID-Design (Redundant Array of Independent Disks) und Prozesse für Backup, Wiederherstellung und

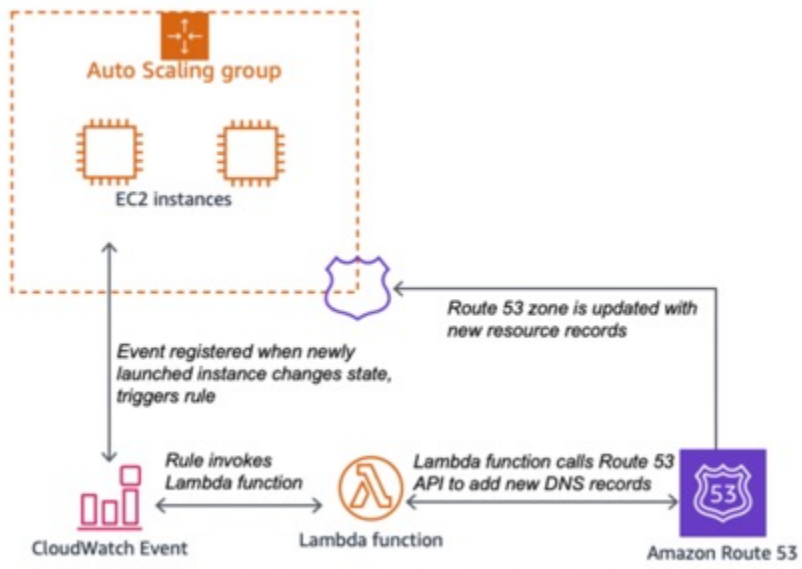
Failover-Verarbeitung. Dies AWS Cloud vereinfacht und verbessert die traditionellen Abläufe in Rechenzentren in Bezug auf Datenbeständigkeit und Verfügbarkeit erheblich.

Für Objekt- und Dateispeicherung bieten AWS Dienste wie Amazon Simple Storage Service (Amazon S3) und Amazon Elastic File System (Amazon EFS) verwaltete Hochverfügbarkeit und Skalierbarkeit. Amazon S3 hat eine Datenbeständigkeit von 99,999999999% (11 Neun).

Für die Speicherung von Transaktionsdaten haben Kunden die Möglichkeit, den vollständig verwalteten Amazon Relational Database Service (Amazon RDS) zu nutzen, der Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle und Microsoft SQL Server mit Hochverfügbarkeitsbereitstellungen unterstützt. Für die Registrar-Funktion, das Abonnentenprofil oder die Speicherung von Buchhaltungsunterlagen (z. B. CDRs) bietet Amazon RDS eine fehlertolerante, hochverfügbare und skalierbare Option.

Dynamische Skalierung mit AWS Lambda Amazon Route 53 und Amazon EC2 Auto Scaling

AWS ermöglicht die Verkettung von Funktionen und die Möglichkeit, benutzerdefinierte serverlose Funktionen als Service auf der Grundlage von Infrastrukturereignissen zu integrieren. Ein solches Entwurfsmuster, das in RTC-Anwendungen vielseitig einsetzbar ist, ist die Kombination von automatischen Skalierungslebenszyklus-Hooks mit Amazon CloudWatch Events, Amazon Route 53 und AWS LambdaFunktionen. AWS Lambda Funktionen können jede Aktion oder Logik einbetten. Die folgende Abbildung zeigt, wie diese miteinander verketteten Funktionen die Zuverlässigkeit und Skalierbarkeit des Systems durch Automatisierung verbessern können.



Automatische Skalierung mit dynamischen Updates für Amazon Route 53

Hochverfügbares WebRTC mit Amazon Kinesis Video Streams

Amazon Kinesis Video Streams bietet Medienstreaming in Echtzeit über WebRTC, sodass Benutzer Medienstreams für Wiedergabe, Analyse und maschinelles Lernen erfassen, verarbeiten und speichern können. Diese Streams sind hochverfügbar, skalierbar und entsprechen den WebRTC-Standards. Amazon Kinesis Video Streams enthalten einen WebRTC-Signalendpunkt für schnelle Peer-Erkennung und sicheren Verbindungsaufbau. Es umfasst Managed Session Traversal Utilities for NAT (STUN) und Traversal Using Relays around NAT (TURN) -Endpunkte für den Medienaustausch zwischen Peers in Echtzeit. Es enthält auch ein kostenloses Open-Source-SDK, das direkt in die Kamera-Firmware integriert ist, um eine sichere Kommunikation mit Amazon Kinesis Video Streams Streams-Endpunkten zu ermöglichen und Peer-Discovery und Medienstreaming zu ermöglichen. Schließlich bietet es Clientbibliotheken für Android und iOS, JavaScript die es WebRTC-kompatiblen Mobil- und Webplayern ermöglichen, ein Kameragerät für Medienstreaming und bidirektionale Kommunikation sicher zu erkennen und eine Verbindung mit diesem herzustellen.

Hochverfügbares SIP-Trunking mit Amazon Chime Voice Connector

Amazon Chime Voice Connector bietet einen pay-as-you-go SIP-Trunking-Service, der es Unternehmen ermöglicht, sichere und kostengünstige Telefonanrufe mit ihren Telefonsystemen zu tätigen und/oder zu empfangen. Amazon Chime Voice Connector ist eine kostengünstige Alternative zu SIP-Trunks von Dienstanbietern oder ISDN (Integrated Services Digital Network) -Primary Rate Interfaces (). PRIs Kunden haben die Möglichkeit, eingehende Anrufe, ausgehende Anrufe oder beides zu aktivieren.

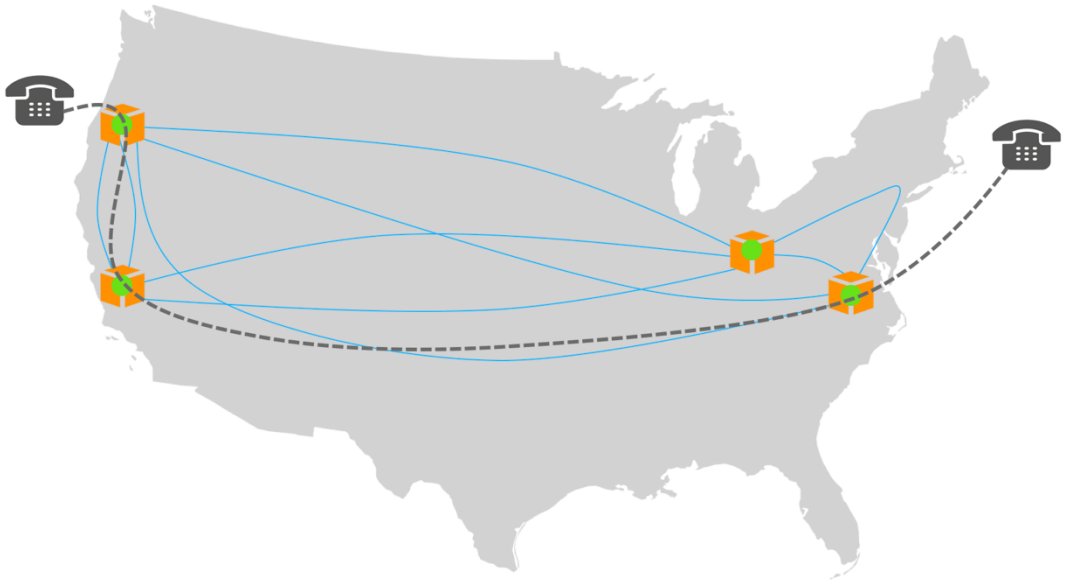
Der Dienst nutzt das AWS Netzwerk, um ein hochverfügbares Anruferlebnis über mehrere Kanäle hinweg bereitzustellen. AWS-Regionen Sie können Audio von SIP-Trunking-Telefonanrufen oder SIPREC-Feeds (Forward SIP-Based Media Recording) an Amazon Kinesis Video Streams streamen, um in Echtzeit Erkenntnisse aus Geschäftsanrufen zu gewinnen. Durch die Integration mit Amazon Transcribe und anderen gängigen Bibliotheken für maschinelles Lernen können Sie schnell Anwendungen für Audioanalysen erstellen.

Bewährte Verfahren aus der Praxis

In diesem Abschnitt werden die bewährten Methoden zusammengefasst, die von einigen der größten und erfolgreichsten AWS Kunden implementiert wurden, die große Echtzeit-Workloads mit Session Initiation Protocol (SIP) ausführen. AWS Kunden, die ihre eigene SIP-Infrastruktur in der Public Cloud betreiben möchten, würden diese Best Practices als wertvoll erachten, da sie dazu beitragen können, die Zuverlässigkeit und Widerstandsfähigkeit des Systems bei verschiedenen Arten von Ausfällen zu erhöhen. Obwohl einige dieser Best Practices SIP-spezifisch sind, sind die meisten von ihnen auf jede Echtzeitkommunikationsanwendung anwendbar, auf der AWS ausgeführt wird.

Erstellen Sie ein SIP-Overlay

AWS verfügt über ein robustes, skalierbares und redundantes Netzwerk-Backbone, das Konnektivität zwischen verschiedenen AWS-Regionen Geräten ermöglicht. Wenn ein Netzwerkereignis, z. B. ein Glasfaserausfall, eine AWS Backbone-Verbindung beeinträchtigt, wird der Datenverkehr mithilfe von Routing-Protokollen auf Netzwerkebene wie dem Border Gateway Protocol (BGP) schnell auf redundante Pfade umgeleitet. Diese Verkehrstechnik auf Netzwerkebene ist für AWS Kunden eine Blackbox, und die meisten bemerken diese Failover-Ereignisse nicht einmal. Kunden, die Echtzeit-Workloads wie Sprach-, Video- und Nachrichtendienste mit geringer Latenz ausführen, bemerken diese Ereignisse jedoch manchmal. Wie kann ein AWS Kunde also zusätzlich zu dem, was auf AWS Netzwerkebene bereitgestellt wird, seine eigene Verkehrstechnik implementieren? Die Lösung besteht darin, die SIP-Infrastruktur auf vielen verschiedenen Ebenen bereitzustellen AWS-Regionen. Als Teil der Funktionen zur Anrufsteuerung bietet SIP auch die Möglichkeit, Anrufe über bestimmte SIP-Proxys weiterzuleiten.

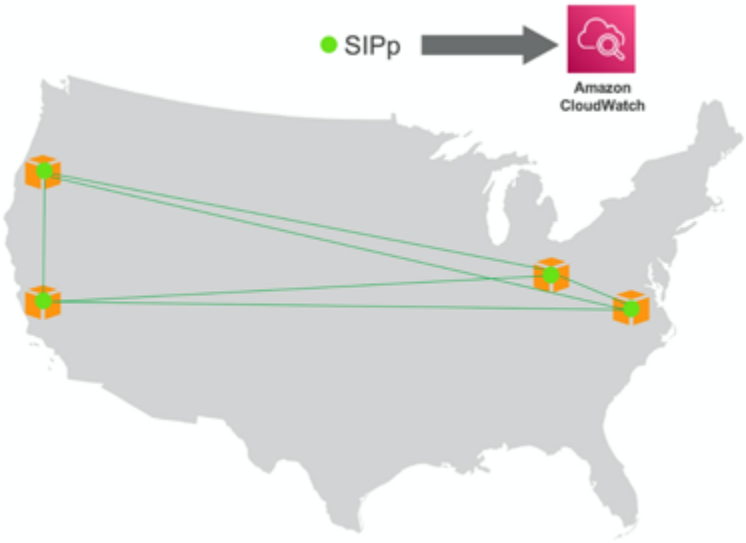


Verwenden von SIP-Routing zum Überschreiben des Netzwerkroutings

In der vorherigen Abbildung läuft die SIP-Infrastruktur (dargestellt durch grüne Punkte in den Würfeln) in allen vier US-Regionen. Die durchgezogenen blauen Linien stellen eine fiktive Darstellung des AWS Backbones dar. Wenn kein SIP-Routing implementiert ist, wird ein Anruf, der von der Westküste der USA ausgeht und für die Ostküste der USA bestimmt ist, über die Backbone-Verbindung weitergeleitet, die die Regionen Oregon und Virginia direkt verbindet. Das Diagramm zeigt, wie ein Kunde das Routing auf Netzwerkebene außer Kraft setzen und denselben Anruf zwischen Oregon und Virginia mithilfe von SIP-Routing über Kalifornien tätigen kann. Diese Art von SIP-Verkehrstechnik kann mithilfe von SIP-Proxys und Media Gateways auf der Grundlage von Netzwerkmetriken wie SIP-Weiterübertragungen und kundenspezifischen Geschäftspräferenzen implementiert werden.

Führen Sie eine detaillierte Überwachung durch

Endbenutzer von Sprach- und Videoanwendungen in Echtzeit erwarten dasselbe Leistungsniveau wie bei herkömmlichen Telefoniediensten. Wenn sie also Probleme mit einer Anwendung haben, schadet das letztlich dem Ruf des Anbieters. Um proaktiv und nicht reaktiv zu sein, ist es unerlässlich, dass an jedem Teil des Systems, der Endbenutzer bedient, eine detaillierte Überwachung eingerichtet wird.



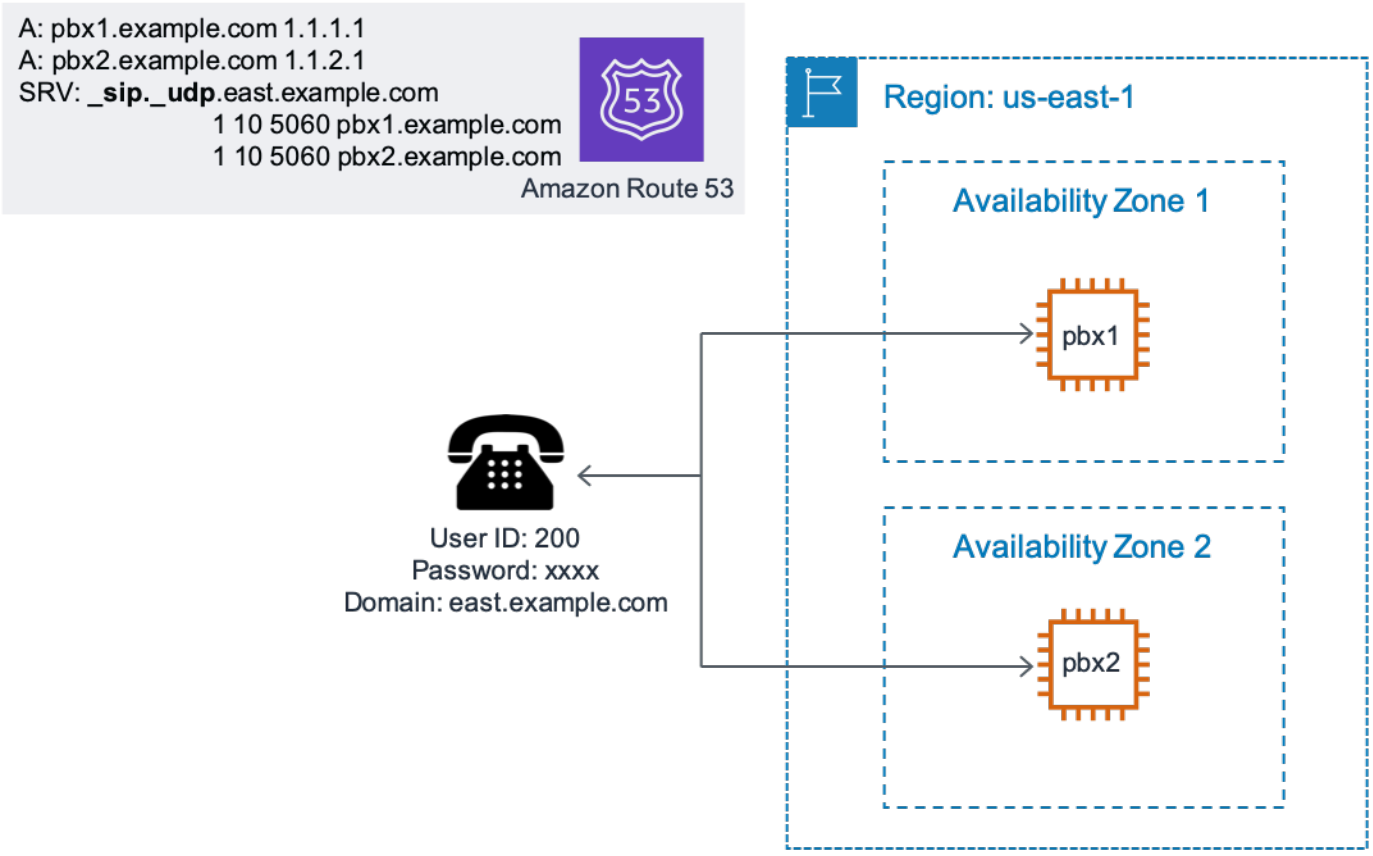
Verwendung SIPp zur Überwachung der VoIP-Infrastruktur

Viele Open-Source-Tools wie iPerf oder SIPpund stehen zur Überwachung des VOIPMonitorSIP/RTP-Verkehrs zur Verfügung. Im vorherigen Beispiel messen Knoten, auf denen SIP im Client- und Servermodus ausgeführt wird, SIP-Metriken wie erfolgreiche Anrufe und SIP-Rückübertragungen zwischen allen vier USA. AWS-Regionen Diese Metriken können dann CloudWatch mit einem benutzerdefinierten Skript nach Amazon exportiert werden. Mithilfe dieser CloudWatch Option können Kunden Alarme für diese benutzerdefinierten Messwerte erstellen, die auf einem bestimmten Schwellenwert basieren. Abhängig vom Status dieser CloudWatch Alarme können dann automatische oder manuelle Abhilfemaßnahmen ergriffen werden.

Für Kunden, die keine technischen Ressourcen bereitstellen möchten, die für die Entwicklung und Wartung eines maßgeschneiderten Überwachungssystems erforderlich sind, sind auf dem Markt viele gute VoIP-Überwachungslösungen erhältlich, wie ThousandEyesz. Ein Beispiel für eine Abhilfemaßnahme ist die Änderung des SIP-Routings aufgrund von vermehrten SIP-Neuübertragungen.

Verwenden Sie DNS für den Lastenausgleich und Floating IPs für den Failover

IP-Telefonieclients, die die DNS-SRV-Funktion unterstützen, können die in die Infrastruktur integrierte Redundanz effizient nutzen, indem sie die Clients auf unterschiedliche/verteilen. SBCs PBXs



Verwendung von DNS-SRV-Einträgen für den Lastenausgleich von SIP-Clients

Die vorherige Abbildung zeigt, wie Kunden die SRV-Einträge für den Lastenausgleich des SIP-Verkehrs verwenden können. Jeder IP-Telefonieclient, der den SRV-Standard unterstützt, sucht nach dem SIP. `<transport protocol>`Präfix in einem DNS-Eintrag vom Typ SRV. Im Beispiel enthält der Antwortbereich von DNS beide, die in unterschiedlichen AWS Availability Zones PBXs ausgeführt werden. Zusätzlich zum Endpunkt URIs enthält der SRV-Eintrag jedoch drei zusätzliche Informationen:

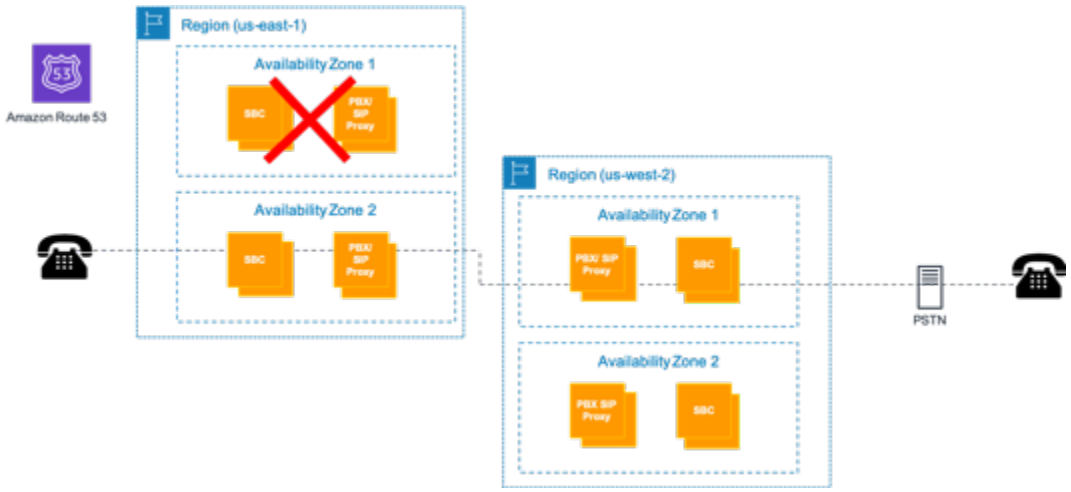
- Die erste Zahl ist die Priorität (1 im obigen Beispiel). Eine niedrigere Priorität wird einer höheren Priorität vorgezogen.
- Die zweite Zahl ist das Gewicht (10 im obigen Beispiel).
- Und die dritte Zahl ist der zu verwendende Port (5060).

Da die Priorität für beide PBXs Server dieselbe ist (1), verwenden die Clients die Gewichtung für den Lastenausgleich zwischen den beiden PBXs. Da die Gewichtungen in diesem Fall identisch sind, sollte der SIP-Verkehr gleichmäßig auf die beiden verteilt werden PBXs.

DNS kann eine gute Lösung für den Client-Lastenausgleich sein, aber wie sieht es mit der Implementierung eines Failovers durch Ändern/Aktualisieren von DNS-A-Einträgen aus? Von dieser Methode wird abgeraten, da das DNS-Caching-Verhalten innerhalb des Client und der dazwischenliegenden Knoten inkonsistent ist. Ein besserer Ansatz für Intra-AZ-Failover zwischen einem Cluster von SIP-Knoten ist die EC2 IP-Neuzuweisung, bei der die IP-Adresse eines beeinträchtigten Hosts mithilfe der API sofort einem fehlerfreien Host zugewiesen wird. EC2 In Kombination mit einer detaillierten Überwachungs- und Integritätsprüfungslösung stellt die IP-Neuzuweisung eines ausgefallenen Knotens sicher, dass der Datenverkehr rechtzeitig auf einen funktionsfähigen Host umgeleitet wird, wodurch Störungen für den Endbenutzer minimiert werden.

Verwenden Sie mehrere Availability Zones

Jede AWS-Region ist in separate Availability Zones unterteilt. Jede Availability Zone verfügt über ihre eigene Stromversorgung, Kühlung und Netzwerkkonnektivität und bildet somit eine isolierte Ausfalldomäne. Im Rahmen von werden Kunden ermutigt AWS, ihre Workloads in mehr als einer Availability Zone auszuführen. Dadurch wird sichergestellt, dass Kundenanwendungen selbst einem kompletten Ausfall der Availability Zone standhalten können — ein an sich schon sehr seltenes Ereignis. Diese Empfehlung steht auch für eine Echtzeit-SIP-Infrastruktur.



Behandlung von Fehlern in der Availability Zone

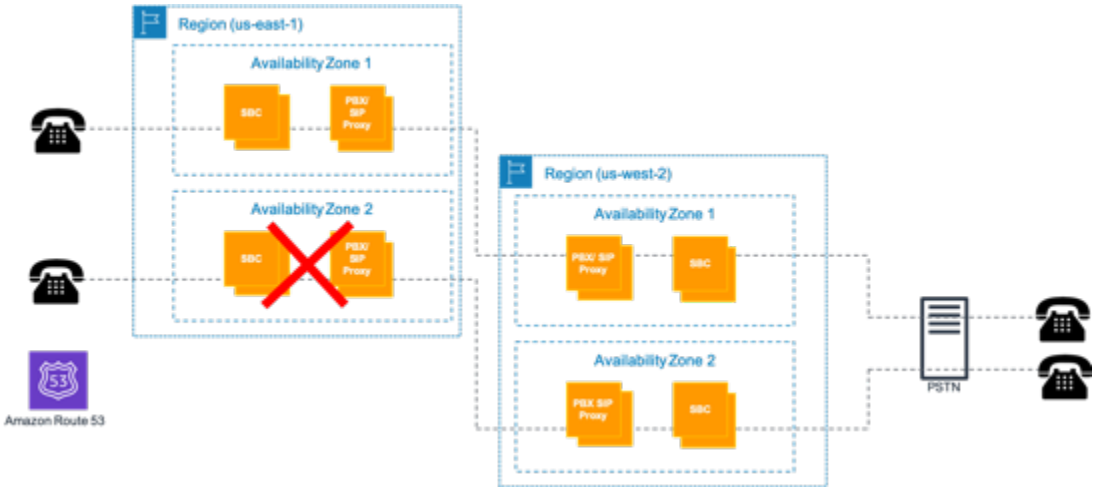
Angenommen, ein katastrophales Ereignis (z. B. ein Hurrikan der Kategorie 5) führt zu einem vollständigen Ausfall der Availability Zone in der Region us-east-1. Wenn die Infrastruktur wie im Diagramm dargestellt läuft, sollten sich alle SIP-Clients, die ursprünglich bei den Knoten in der ausgefallenen Availability Zone registriert waren, erneut bei den SIP-Knoten registrieren, die in Availability Zone #2 laufen. (Testen Sie dieses Verhalten mit Ihren SIP-Clients/-Telefonen, um

sicherzustellen, dass es unterstützt wird.) Obwohl die aktiven SIP-Anrufe zum Zeitpunkt des Ausfalls der Availability Zone verloren gehen, werden alle neuen Anrufe über Availability Zone 2 weitergeleitet.

Zusammenfassend lässt sich sagen, dass DNS-SRV-Einträge den Client auf mehrere A-Einträge verweisen sollten, einen in jeder Availability Zone. Jeder dieser A-Einträge sollte wiederum auf mehrere IP-Adressen von SBCs/PBXs in dieser Availability Zone verweisen, sodass sowohl innerhalb der Availability Zone als auch zwischen Availability Zones Resilienz gewährleistet ist. Sowohl Failover innerhalb als auch zwischen Availability Zones können mithilfe von IP-Neuzuweisung implementiert werden, wenn sie öffentlich sind. IPs Private Daten können IPs jedoch nicht für mehrere Availability Zones neu zugewiesen werden. Wenn ein Kunde private IP-Adressierung verwendet, müsste er sich darauf verlassen, dass sich die SIP-Clients für den Failover zwischen den Availability Zones erneut bei der Ersatz-SBC/PBX registrieren.

Halten Sie den Datenverkehr innerhalb einer Availability Zone und verwenden Sie Platzierungsgruppen EC2

Diese bewährte Methode, auch Availability Zone Affinity genannt, gilt auch für den seltenen Fall eines vollständigen Ausfalls der Availability Zone. Es wird empfohlen, jeglichen AZ-übergreifenden Verkehr zu eliminieren, sodass jeder SIP- oder RTP-Verkehr, der in eine Availability Zone eintritt, in dieser Availability Zone verbleibt, bis er die Region verlässt.



Affinität zur Availability Zone (höchstens 50% der aktiven Anrufe gehen verloren)

Die vorherige Abbildung zeigt eine vereinfachte Architektur, die Availability Zone-Affinität verwendet. Der komparative Vorteil dieses Ansatzes wird deutlich, wenn man die Auswirkungen eines vollständigen Ausfalls der Availability Zone berücksichtigt. Wie im Diagramm dargestellt, sind bei einem Ausfall der Availability Zone 2 höchstens 50% der aktiven Anrufe betroffen (unter der Annahme

eines gleichen Lastenausgleichs zwischen den Availability Zones). Wäre Availability Zone Affinity nicht implementiert worden, würden einige Anrufe zwischen Availability Zones in einer Region fließen, und ein Ausfall würde höchstwahrscheinlich mehr als 50% der aktiven Anrufe betreffen.

Um die Latenz für den Datenverkehr zu minimieren, empfiehlt AWS außerdem, die Verwendung von EC2 Platzierungsgruppen innerhalb jeder Availability Zone in Betracht zu ziehen. Instances, die innerhalb derselben EC2 Platzierungsgruppe gestartet werden, haben eine höhere Bandbreite und eine geringere Latenz, wodurch die Netzwerknähe dieser Instances zueinander EC2 gewährleistet ist.

Verwenden Sie erweiterte EC2 Netzwerkinstanztypen

Die Wahl des richtigen Instance-Typs bei Amazon EC2 gewährleistet die Zuverlässigkeit des Systems sowie die effiziente Nutzung der Infrastruktur. EC2 bietet eine große Auswahl an Instance-Typen, die für unterschiedliche Anwendungsfälle optimiert sind. Die Instance-Typen umfassen unterschiedliche Kombinationen von CPU-, Arbeitsspeicher-, Speicher- und Netzwerkkapazitäten und geben Ihnen die nötige Flexibilität, um die richtige Mischung von Ressourcen für Ihre Anwendungen zu wählen. Diese erweiterten Netzwerk-Instance-Typen stellen sicher, dass die auf ihnen ausgeführten SIP-Workloads Zugriff auf eine konsistente Bandbreite und eine vergleichsweise geringere Gesamtlatenz haben. Neu bei Amazon EC2 ist der Elastic Network Adapter (ENA), der bis zu 100 Gbit/s Bandbreite bietet. Den neuesten Katalog der EC2 Instance-Typen und der zugehörigen Funktionen finden Sie auf der Seite mit den EC2 Instance-Typen.

Für die meisten Kunden sollte die neueste Generation von Compute Optimized Instances das beste Preis-Leistungs-Verhältnis bieten. Der C5N unterstützt beispielsweise den neuen Elastic Network Adapter mit einer Bandbreite von bis zu 100 Gbit/s und Millionen von Paketen pro Sekunde (PPS). Die meisten Echtzeitanwendungen würden auch von der Verwendung des Intel Data Plane Developer Kit (DPDK) profitieren, das die Verarbeitung von Netzwerkpaketen erheblich beschleunigen kann.

Es ist jedoch immer eine bewährte Methode, die verschiedenen EC2 Instance-Typen entsprechend Ihren Anforderungen zu vergleichen, um herauszufinden, welcher Instance-Typ für Sie am besten geeignet ist. Mithilfe von Benchmarking können Sie auch andere Konfigurationsparameter ermitteln, z. B. die maximale Anzahl von Aufrufen, die ein bestimmter Instance-Typ gleichzeitig verarbeiten kann.

Sicherheitsüberlegungen

RTC-Anwendungskomponenten werden in der Regel direkt auf mit dem Internet verbundenen EC2 Amazon-Instances ausgeführt. Zusätzlich zu TCP verwenden Flows Protokolle wie UDP und SIP. AWS Shield Standard Schützt in diesen Fällen EC2 Amazon-Instances vor gängigen Angriffen auf Infrastrukturebene (Layer 3 und 4) DDo S, wie z. B. UDP-Reflection-Angriffen, DNS-Reflection, NTP-Reflection, SSDP-Reflection usw. AWS Shield Standard verwendet verschiedene Techniken wie prioritätsbasiertes Traffic Shaping, die automatisch aktiviert werden, wenn eine genau definierte DDo S-Angriffssignatur erkannt wird.

AWS bietet außerdem erweiterten Schutz vor großen und ausgeklügelten DDo S-Angriffen für diese Anwendungen durch die Aktivierung von Elastic AWS Shield Advanced IP-Adressen. AWS Shield Advanced bietet eine verbesserte DDo S-Erkennung, die automatisch die Art der AWS Ressource und die Größe der EC2 Instanz erkennt und entsprechende vordefinierte Abhilfemaßnahmen mit Schutz vor SYN- oder UDP-Floods anwendet. Mit können Kunden auch ihre eigenen benutzerdefinierten Schadensbegrenzungsprofile erstellen AWS Shield Advanced, indem sie das rund um die Uhr verfügbare AWS DDo S Response Team (DRT) hinzuziehen. AWS Shield Advanced stellt außerdem sicher, dass während eines DDo S-Angriffs alle Ihre Amazon VPC Network Access Control Lists (ACLs) automatisch an der AWS Netzwerkgrenze durchgesetzt werden, sodass Sie Zugriff auf zusätzliche Bandbreite und Scrubbing-Kapazität haben, um große volumetrische S-Angriffe abzuwehren. DDo

Schlussfolgerung

Workloads für Echtzeitkommunikation (RTC) können eingesetzt werden, AWS um Skalierbarkeit, Elastizität und Hochverfügbarkeit zu erreichen und gleichzeitig die wichtigsten Anforderungen zu erfüllen. Heute nutzen mehrere Kunden AWS, seine Partner und Open-Source-Lösungen, um RTC-Workloads mit geringeren Kosten und schnellerer Agilität sowie einer geringeren globalen Präsenz auszuführen.

Die in diesem Whitepaper vorgestellten Referenzarchitekturen und Best Practices können Kunden dabei helfen, RTC-Workloads erfolgreich einzurichten AWS und die Lösungen so zu optimieren, dass sie die Anforderungen der Endbenutzer erfüllen und gleichzeitig für die Cloud optimieren.

Akronyme

In diesem Dokument werden unter anderem folgende Akronyme verwendet:

ACL — Zugriffskontrollliste

ALB — Application Load Balancer

APNs — Apple-Push-Benachrichtigungsdienst

BGP — Border Gateway-Protokoll

CDR — Aufzeichnungen mit Anrufdetails

COTS — kommerzielle Software off-the-shelf

DDoS — verteilt denial-of-service

DNS — Domainnamensystem

DPDK — Entwicklerkit für Intel Data Plane

DRT — DDo Das Reaktionsteam

ENA — Elastischer Netzwerkadapter

EPC — Weiterentwickelter Paketkern

FCM — Firebase Cloud-Nachrichten

HA — Hochverfügbarkeit

IRC — Internet-Relay-Chat

ISDN — Digitales Netzwerk für integrierte Dienste

NAT — Netzwerkadressübersetzung

OPUS — Benutzerunterstützung für Online-Positionierung

PBX — Private Zweigstellenbörse

PRI — Schnittstelle zum Primärtarif

PSTN — Öffentliches Telefonnetz

RAID — Redundantes Array unabhängiger Festplatten

RTC — Kommunikation in Echtzeit

RTP — Echtzeit-Transportprotokoll

SAN — Speichernetzwerk

SBC — Grenzcontroller für Sitzungen

SIP — Protokoll zur Sitzungsinitiierung

SPOF — einzelne Fehlerquellen

SRV — Dienst

SS7 — Signalsystem n.7

STUN — Dienstprogramme zur Sitzungsdurchquerung für NAT

SYN — Synchronisieren

TCP — Übertragungssteuerungsprotokoll

TDM — Zeitmultiplexing

TURN — Durchquerung mithilfe von Relais rund um NAT

UDP — Benutzer-Datagramm-Protokoll

URI — Einheitliche Ressourcen-Identifikatoren

VIP — virtuelle IP

VNF — Virtuelle Netzwerkfunktion

VoIP — Telefonie über IP

VPC — Virtuelle private Cloud

WebRTC — Web-Kommunikation in Echtzeit

Mitwirkende

Folgende Personen und Organisationen haben zu diesem Dokument beigetragen:

- Mounir Chennana, leitender Lösungsarchitekt, Amazon Web Services
- Mohammed Al-Mehdar, leitender Lösungsarchitekt, Amazon Web Services
- Ejaz Sial, leitender Lösungsarchitekt, Amazon Web Services
- Ahmad Khan, leitender Lösungsarchitekt, Amazon Web Services
- Tipu Qureshi, Chefingenieur AWS -Support, Amazon Web Services
- Hasan Khan, leitender technischer Kundenbetreuer, Amazon Web Services
- Shoma Chakravarty, Technischer Leiter bei WW, Telekommunikation, Amazon Web Services

Dokumentversionen

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

Änderung	Beschreibung	Datum
Whitepaper aktualisiert	Für die neuesten Dienste und Funktionen aktualisiert.	5. Mai 2022
Whitepaper aktualisiert	Für die neuesten Dienste und Funktionen aktualisiert.	13. Februar 2020
Erste Veröffentlichung	Das Whitepaper wurde zuerst veröffentlicht.	1. Oktober 2018

Hinweise

Kunden sind dafür verantwortlich, Ihre eigene unabhängige Bewertung der Informationen in diesem Dokument vorzunehmen. Dieses Dokument: (a) dient nur zu Informationszwecken, (b) stellt aktuelle AWS-Produktangebote und -praktiken dar, die ohne vorherige Ankündigung geändert werden können, und (c) stellt keine Verpflichtungen oder Zusicherungen von AWS und seinen verbundenen Unternehmen, Lieferanten oder Lizenzgebern dar. AWS-Produkte oder -Services werden „wie sie sind“ ohne ausdrückliche oder stillschweigende Garantien, Zusicherungen oder Bedingungen jeglicher Art bereitgestellt. Die Verantwortung und Haftung von AWS gegenüber seinen Kunden werden durch AWS-Vereinbarungen geregelt. Dieses Dokument gehört, weder ganz noch teilweise, nicht zu den Vereinbarung von AWS mit seinen Kunden und ändert diese Vereinbarungen auch nicht.

AWS Glossar

Die neueste AWS Terminologie finden Sie im AWS Glossar in der AWS-Glossar Referenz.

Die vorliegende Übersetzung wurde maschinell erstellt. Im Falle eines Konflikts oder eines Widerspruchs zwischen dieser übersetzten Fassung und der englischen Fassung (einschließlich infolge von Verzögerungen bei der Übersetzung) ist die englische Fassung maßgeblich.