Livre blanc AWS

Communication en temps réel sur AWS

Communication en temps réel sur AWS: Livre blanc AWS

Table of Contents

Communication en temps réel sur AWS

Meilleures pratiques pour concevoir des charges de travail de communication en temps réel (RTC) hautement disponibles et évolutives sur AWS

Date de publication : 5 mai 2022 (Révisions du document)

Résumé

Aujourd'hui, de nombreuses entreprises cherchent à réduire leurs coûts et à atteindre l'évolutivité pour les charges de travail vocales, de messagerie et multimédia en temps réel. Ce paper décrit les meilleures pratiques pour gérer les charges de travail de communication en temps réel (RTC) sur Amazon Web Services (AWS) et inclut des architectures de référence répondant à ces exigences. Ce paper sert de guide aux personnes habituées à la communication en temps réel sur la manière d'atteindre une disponibilité et une évolutivité élevées pour ces charges de travail.

Ce paper inclut des architectures de référence qui montrent comment configurer les charges de travail RTC AWS, ainsi que les meilleures pratiques pour optimiser les solutions afin de répondre aux besoins des utilisateurs finaux tout en les optimisant pour le cloud. L'Evolved Packet Core (EPC) n'est pas abordé dans ce livre blanc, mais les meilleures pratiques détaillées ici peuvent être appliquées aux fonctions de réseau virtuel (). VNFs

Êtes-vous Well-Architected ?

Le AWS Well-Architected Framework vous aide à comprendre les avantages et les inconvénients des décisions que vous prenez lors de la création de systèmes dans le cloud. Les six piliers du cadre vous permettent d'apprendre les meilleures pratiques architecturales pour concevoir et exploiter des systèmes fiables, sécurisés, efficaces, rentables et durables. À l'aide du AWS Well-Architected Tool, disponible gratuitement dans le AWS Management Console(connexion requise), vous pouvez évaluer votre charge de travail par rapport à ces meilleures pratiques en répondant à une série de questions pour chaque pilier.

Pour obtenir des conseils d'experts supplémentaires et les meilleures pratiques relatives à votre architecture cloud (déploiements d'architecture de référence, diagrammes et livres blancs), consultez le Centre d'architecture.AWS

Introduction

Les applications de télécommunication utilisant la voix, la vidéo et la messagerie comme canaux constituent une exigence essentielle pour de nombreuses organisations et leurs utilisateurs finaux. Ces charges de travail de communication en temps réel (RTC) ont des exigences de latence et de disponibilité spécifiques qui peuvent être satisfaites en suivant les meilleures pratiques de conception pertinentes. Dans le passé, les charges de travail RTC étaient déployées dans des centres de données locaux traditionnels dotés de ressources dédiées.

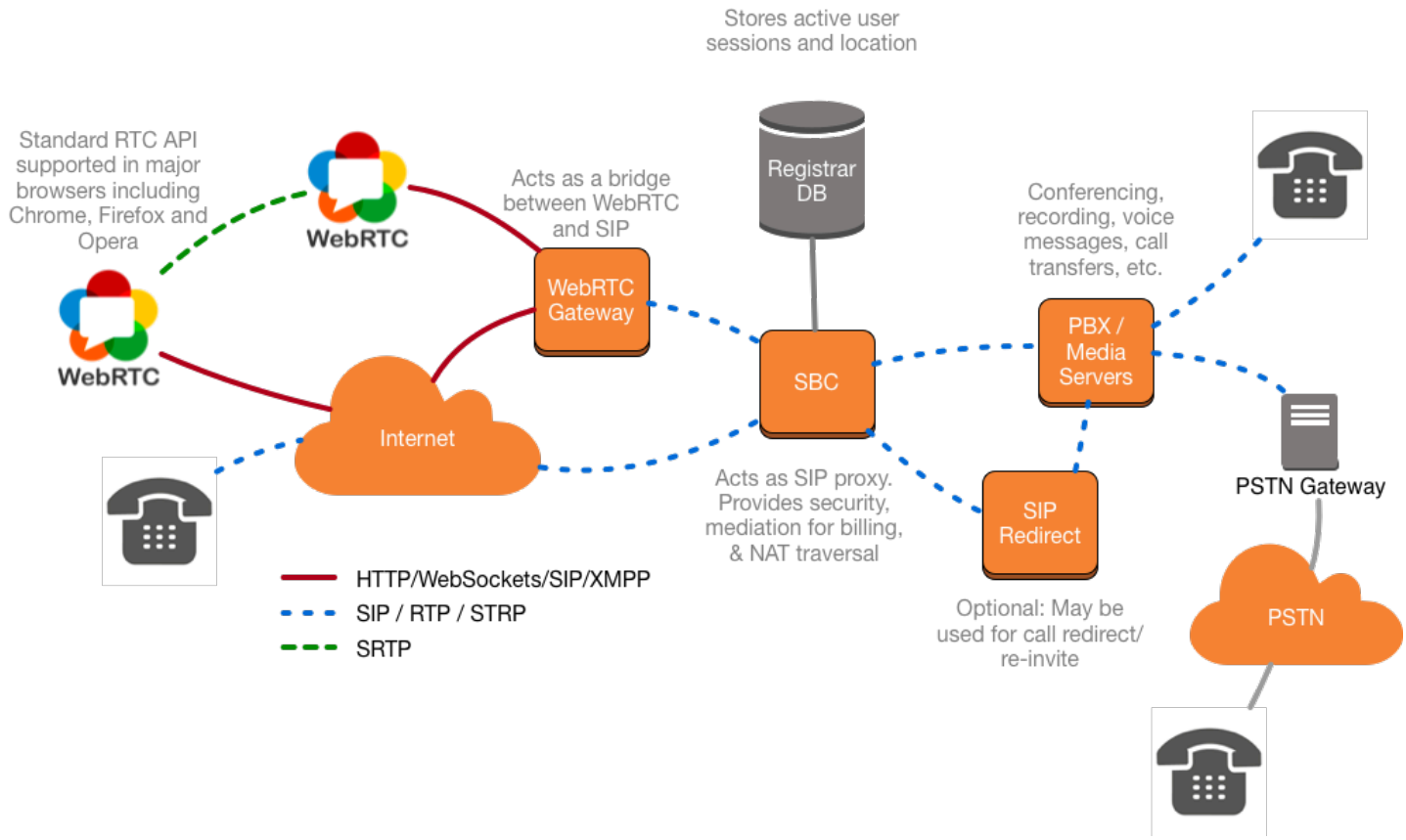
Les charges de travail RTC nécessitent un environnement hautement évolutif, résilient et disponible. Aujourd'hui, les clients ont l'habitude AWS d'exécuter des charges de travail RTC à moindre coût, en améliorant l'agilité, l'élasticité et les délais de mise sur le marché.

Composants fondamentaux de l'architecture RTC

Dans le secteur des télécommunications, le RTC fait généralement référence à des sessions multimédia en direct entre deux terminaux avec une latence minimale. Ces sessions peuvent être liées à :

- Une session vocale entre deux parties (système téléphonique, mobile ou voix sur IP (VoIP))
- Messagerie instantanée (telle que le chat et le chat par relais instantané (IRC))
- Séance vidéo en direct (comme la visioconférence et la téléprésence)

Chacune des solutions précédentes possède certains composants en commun (tels que des composants qui fournissent l'authentification, l'autorisation et le contrôle d'accès, le transcodage, la mise en mémoire tampon et le relais, etc.) et des composants spécifiques au type de média transmis (tels que le service de diffusion, le serveur de messagerie et les files d'attente, etc.). Cette section se concentre sur la définition d'un système RTC basé sur la voix et la vidéo et de tous les composants associés, comme illustré dans la figure suivante.



Composants architecturaux essentiels pour RTC

Softswitch/PBX

Un commutateur logiciel ou PBX est le cerveau d'un système de téléphonie vocale et fournit des informations pour établir, maintenir et acheminer un appel vocal au sein ou en dehors de l'entreprise en utilisant différents composants. Tous les abonnés de l'entreprise doivent s'inscrire auprès du softswitch pour recevoir ou passer un appel. L'une des fonctionnalités importantes du softswitch est de suivre chaque abonné et de savoir comment le joindre en utilisant les autres composants du réseau vocal.

Contrôleur de session en bordure (SBC)

Un contrôleur de session frontalier (SBC) se trouve à la périphérie d'un réseau vocal et assure le suivi de tout le trafic entrant et sortant (plans de contrôle et de données). L'une des principales responsabilités d'un SBC est de protéger le système vocal contre toute utilisation malveillante. Le SBC peut être utilisé pour s'interconnecter avec des troncs du protocole d'initiation de session (SIP) pour une connectivité externe. Certains offrent SBCs également des fonctionnalités de transcodage pour la conversion CODECsd'un format à un autre. La plupart offrent SBCs également des fonctionnalités de traversée de la traduction d'adresses réseau (NAT), ce qui permet de garantir que les appels sont établis, même sur des réseaux protégés par un pare-feu.

Connectivité PSTN

Les solutions de voix sur IP (VoIP) utilisent des passerelles du réseau téléphonique public commuté (PSTN) et des troncs SIP pour se connecter aux réseaux PSTN existants.

Passerelle PSTN

La passerelle PSTN convertit le signal entre le protocole SIP SS7 et le média entre le protocole de transport en temps réel (RTP) et le multiplexage par répartition dans le temps (TDM) à l'aide du transcodage CODEC. Les passerelles PSTN sont toujours situées à la périphérie, à proximité du réseau PSTN.

Tronc SIP

Dans une liaison SIP, l'entreprise n'envoie pas ses appels sur un réseau TDM (SS7 basé sur le TDM), mais les flux entre l'entreprise et l'opérateur téléphonique restent sur IP. La plupart des troncs SIP sont établis en utilisant SBCs. L'entreprise doit se mettre d'accord sur les règles de sécurité

prédéfinies par les opérateurs télécoms, telles que l'autorisation d'un certain nombre d'adresses IP, de ports, etc.

Passerelle multimédia (transcodeur)

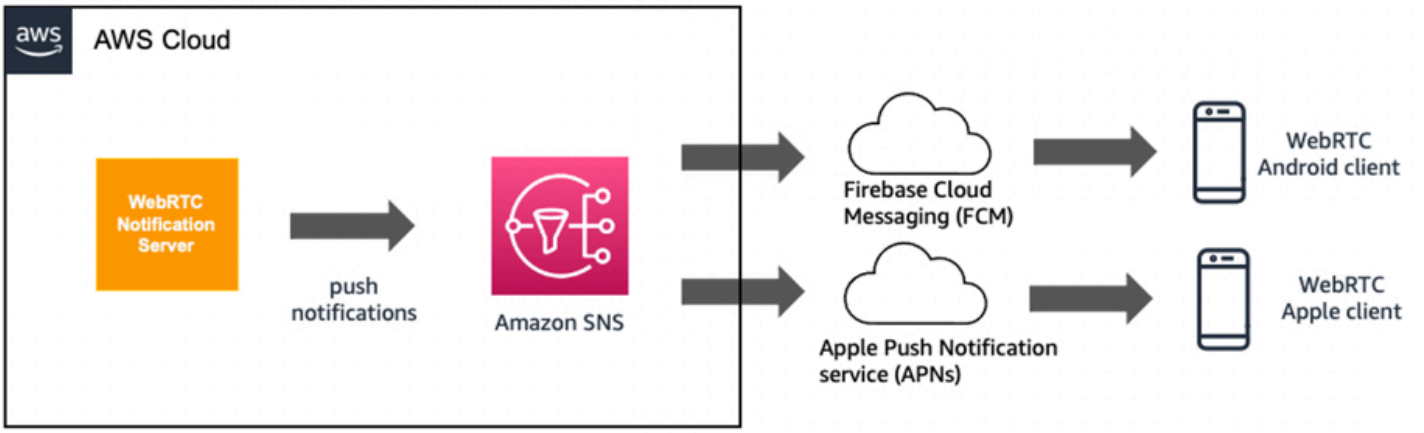
Les utilisateurs communiquent en temps réel par audio et/ou vidéo, ainsi que par le biais de données facultatives et d'autres informations. Pour communiquer, les deux appareils doivent être en mesure de s'entendre sur un codec mutuellement compris pour chaque piste multimédia, afin de pouvoir communiquer et présenter avec succès le contenu multimédia partagé. Tous les navigateurs compatibles avec WebRTC doivent prendre en charge le support utilisateur de positionnement en ligne (OPUS) et le G711 pour le son, ainsi que le profil de ligne de base contrainte VP8H.264 pour la vidéo.

Une solution vocale typique en dehors de l'écosystème WebRTC permet différents types de. CODECs Parmi les plus courantes, citons la loi CODECs G.711 μ-law pour l'Amérique du Nord, la loi G.711 A, la G.729 et la G.722. Lorsque deux appareils utilisant deux appareils différents CODECs communiquent entre eux, la passerelle multimédia traduit le flux de CODEC entre les appareils. En d'autres termes, une passerelle multimédia traite le contenu multimédia et garantit que les appareils finaux sont en mesure de communiquer entre eux.

Notifications push dans WebRTC

Les implémentations du WebRTC sont très courantes sur les appareils mobiles. Contrairement aux navigateurs Web, un appareil mobile ne peut pas maintenir une connectivité Websocket ouverte pendant longtemps. Par conséquent, il doit s'appuyer sur les notifications push du serveur WebRTC pour toutes les demandes de fin, telles que les appels et les messages.

Amazon Simple Notification Service (Amazon SNS) vous permet d'envoyer des notifications push à des applications sur des appareils mobiles. Ces applications peuvent fonctionner sur différents systèmes d'exploitation tels qu'Apple iOS ou Android. La figure suivante présente un aperçu général du flux de notifications push, d'un serveur de notifications WebRTC aux points de terminaison mobiles WebRTC.

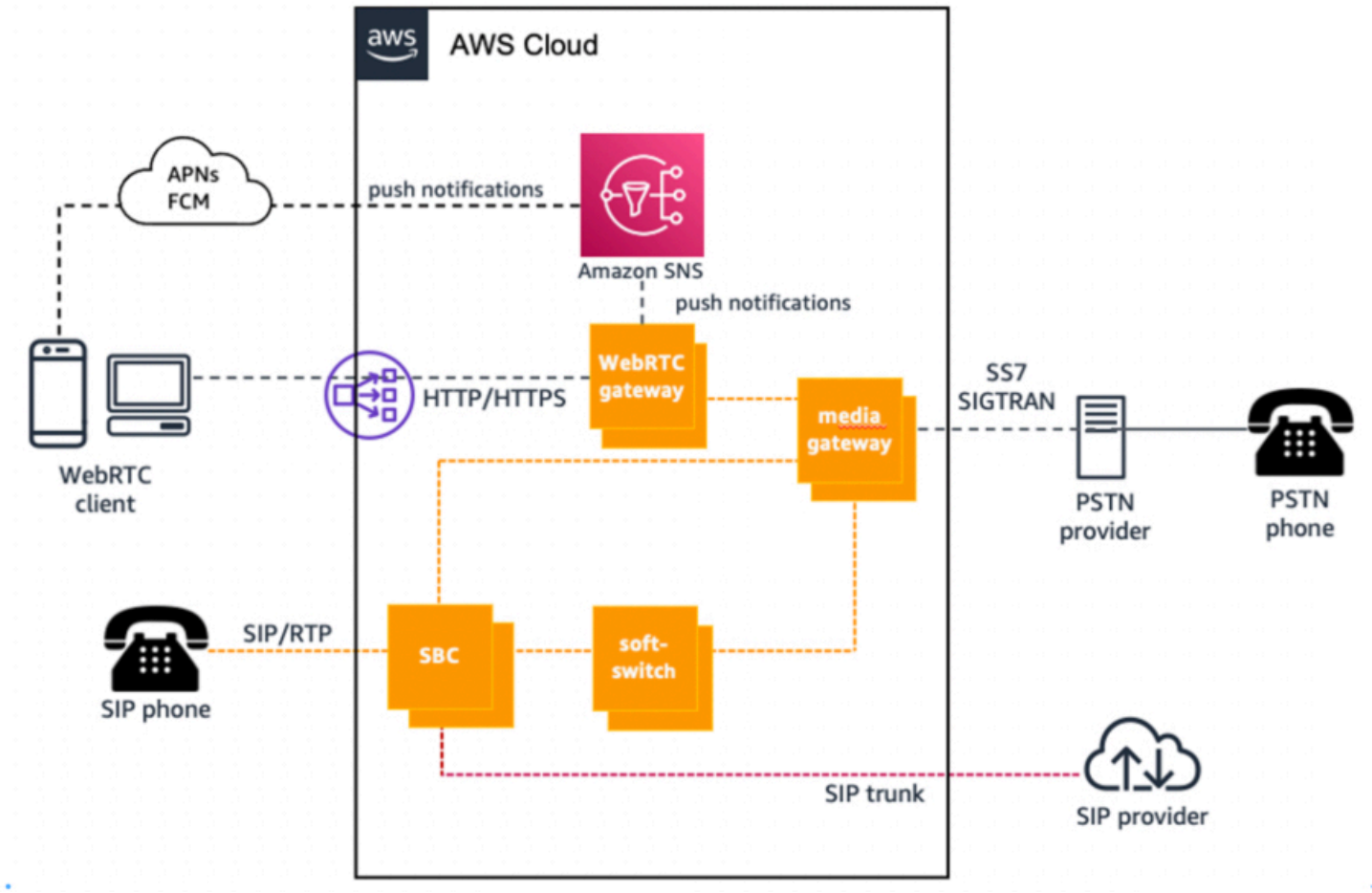


Amazon SNS pour les notifications push

Passerelle WebRTC et WebRTC

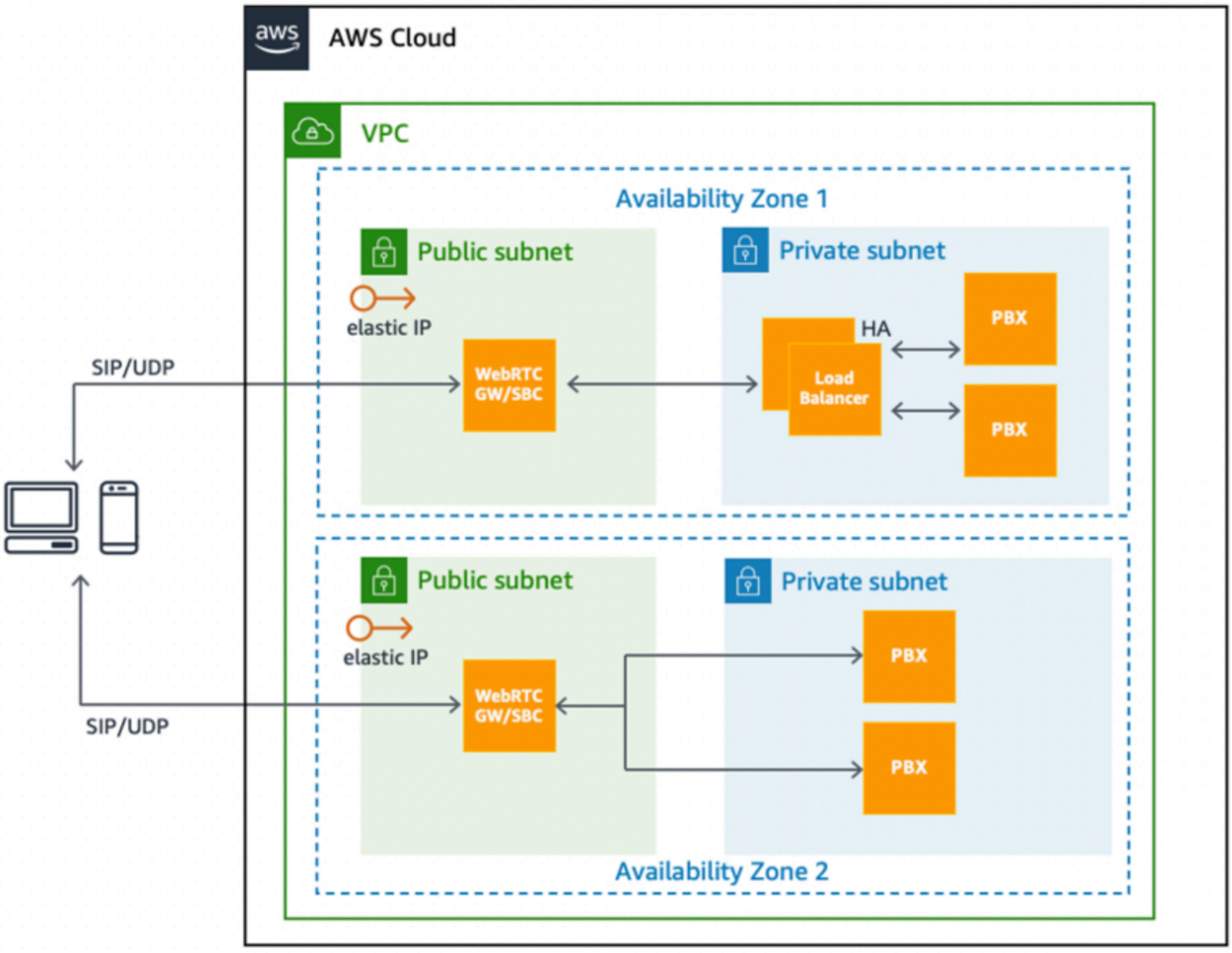
La communication Web en temps réel (WebRTC) vous permet d'établir un appel depuis un navigateur Web ou de demander des ressources au serveur principal à l'aide de l'API. La technologie est conçue en tenant compte de la technologie cloud et fournit donc divers APIs éléments qui pourraient être utilisés pour établir un appel. Comme toutes les solutions vocales (y compris SIP) ne les prennent pas en charge APIs, la passerelle WebRTC est requise pour traduire les appels d'API en messages SIP et vice versa.

La figure suivante montre un modèle de conception pour une architecture WebRTC à haute disponibilité. Le trafic entrant provenant des clients WebRTC est équilibré par un Application Load Balancer (ALB), WebRTC s'exécutant sur des instances Amazon Elastic Compute Cloud (Amazon) faisant partie d'un groupe EC2 Amazon Auto Scaling. EC2



Topologie de base d'un système RTC pour la voix

Un autre modèle de conception pour le trafic SIP et RTP consiste à utiliser des paires SBCs sur Amazon EC2 en mode actif-passif entre les zones de disponibilité, comme le montre la figure suivante. Ici, une adresse IP élastique peut être déplacée dynamiquement entre les instances en cas de défaillance, lorsque le service de nom de domaine (DNS) ne peut pas être utilisé.



Architecture RTC utilisant Amazon EC2 dans un cloud privé virtuel (VPC)

Haute disponibilité et évolutivité sur AWS

La plupart des fournisseurs de communications en temps réel s'alignent sur des niveaux de service garantissant une disponibilité comprise entre 99,9 % et 99,999 %. En fonction du degré de haute disponibilité (HA) que vous souhaitez, vous devez prendre des mesures de plus en plus sophistiquées tout au long du cycle de vie de l'application. AWS recommande de suivre ces directives pour atteindre un niveau élevé de haute disponibilité :

- Concevez le système de manière à ce qu'il n'y ait aucun point de défaillance unique. Utilisez des mécanismes automatisés de surveillance, de détection des défaillances et de basculement pour les composants statiques et dynamiques
  - Les points de défaillance uniques (SPOF) sont généralement éliminés avec une configuration de redondance N+1 ou 2N, où N+1 est obtenu via l'équilibrage de charge entre les nœuds actifs-actifs, et 2N est obtenu par une paire de nœuds en configuration actif-veille.
  - AWS dispose de plusieurs méthodes pour atteindre la haute disponibilité grâce aux deux approches, par exemple via un cluster évolutif à charge équilibrée ou en supposant une paire actif-veille.
- Disponibilité correcte de l'instrument et du système de test.
- Préparez des procédures opérationnelles pour les mécanismes manuels afin de réagir, d'atténuer et de récupérer après une panne.

Cette section explique comment éliminer tout point de défaillance unique à l'aide des fonctionnalités disponibles sur AWS. Plus précisément, cette section décrit un sous-ensemble de AWS fonctionnalités de base et de modèles de conception qui vous permettent de créer des applications de communication en temps réel hautement disponibles.

Modèle IP flottant pour la haute disponibilité entre des serveurs dynamiques actifs et en veille

Le modèle de conception IP flottante est un mécanisme bien connu pour réaliser un basculement automatique entre une paire de nœuds matériels actifs et en veille (serveurs multimédias). Une adresse IP virtuelle secondaire statique est attribuée au nœud actif. La surveillance continue entre les nœuds actifs et de secours permet de détecter les défaillances. En cas de défaillance du nœud actif, le script de surveillance attribue l'adresse IP virtuelle au nœud de secours prêt et le nœud de secours

prend en charge la fonction active principale. De cette façon, l'adresse IP virtuelle flotte entre le nœud actif et le nœud de secours.

Applicabilité dans les solutions RTC

Il n'est pas toujours possible d'avoir plusieurs instances actives du même composant en service, par exemple un cluster actif-actif de N nœuds. Une configuration active en veille constitue le meilleur mécanisme pour la haute disponibilité. Par exemple, les composants dynamiques d'une solution RTC, tels que le serveur multimédia ou le serveur de conférence, ou même un serveur SBC ou un serveur de base de données, conviennent parfaitement à une configuration active en veille. Un serveur SBC ou multimédia possède plusieurs sessions ou canaux actifs de longue durée à un moment donné, et en cas de défaillance de l'instance active SBC, les points de terminaison peuvent se reconnecter au nœud de secours sans aucune configuration côté client en raison de l'adresse IP flottante.

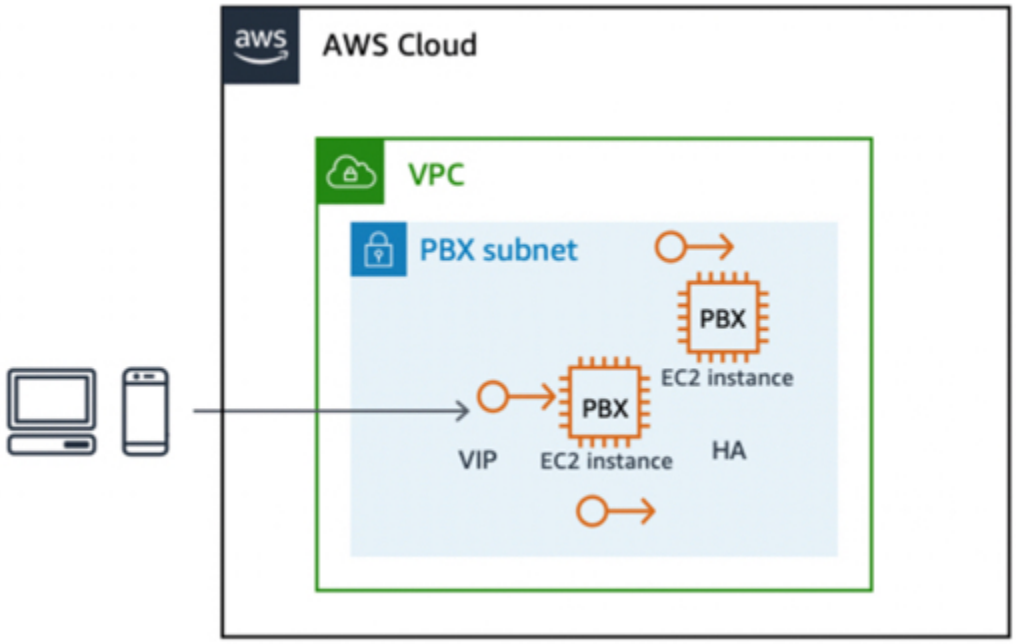
Mise en œuvre le AWS

Vous pouvez implémenter ce modèle sur AWS à l'aide des fonctionnalités de base d'Amazon Elastic Compute Cloud (Amazon EC2), de EC2 l'API Amazon, des adresses IP élastiques et du support d'Amazon EC2 pour les adresses IP privées secondaires.

Pour implémenter le modèle IP flottant sur AWS :

1. Lancez deux EC2 instances pour assumer les rôles de nœuds principal et secondaire, le nœud principal étant supposé être actif par défaut.
2. Attribuez une adresse IP privée secondaire supplémentaire à l' EC2 instance principale.
3. Une adresse IP élastique, similaire à une adresse IP virtuelle (VIP), est associée à l'adresse privée secondaire. Cette adresse privée secondaire est l'adresse utilisée par les points de terminaison externes pour accéder à l'application.
4. Une certaine configuration du système d'exploitation (OS) est requise pour que l'adresse IP secondaire soit ajoutée en tant qu'alias à l'interface réseau principale.
5. L'application doit être liée à cette adresse IP élastique. Dans le cas du logiciel Asterisk, vous pouvez configurer la liaison via les paramètres avancés du SIP Asterisk.
6. Exécutez un script de surveillance (personnalisé, KeepAlive sous Linux, Corosync, etc.) sur chaque nœud pour surveiller l'état du nœud homologue. En cas de défaillance du nœud actif actuel, l'homologue détecte cette défaillance et invoque l' EC2 API Amazon pour se réattribuer l'adresse IP privée secondaire.

Par conséquent, l'application qui écoutait sur le VIP associé à l'adresse IP privée secondaire devient accessible aux terminaux via le nœud de secours.



Basculement entre des EC2 instances dynamiques à l'aide d'une adresse IP élastique

Avantages

Cette approche est une solution fiable à petit budget qui protège contre les défaillances au niveau de l' EC2 instance, de l'infrastructure ou de l'application.

Limites et extensibilité

Ce modèle de conception est généralement limité à une seule zone de disponibilité. Il peut être mis en œuvre dans deux zones de disponibilité, mais avec des variantes. Dans ce cas, l'adresse IP élastique flottante est réassociée entre le nœud actif et le nœud de secours dans différentes zones de disponibilité via l'API d'adresse IP élastique de réassociation disponible. Dans l'implémentation du basculement illustrée dans la figure précédente, les appels en cours sont abandonnés et les terminaux doivent se reconnecter. Il est possible d'étendre cette implémentation avec la réplication des données de session sous-jacentes afin de permettre un basculement fluide des sessions ou une continuité multimédia.

Équilibrage de charge pour l'évolutivité et la haute disponibilité avec WebRTC et SIP

L'équilibrage de charge d'un cluster d'instances actives basé sur des règles prédéfinies, telles que le round robin, l'affinité ou la latence, etc., est un modèle de conception largement popularisé par la nature apatride des requêtes HTTP. En fait, l'équilibrage de charge est une option viable dans le cas de nombreux composants d'application RTC.

L'équilibreur de charge fait office de proxy inverse ou de point d'entrée pour les demandes adressées à l'application souhaitée, elle-même configurée pour s'exécuter simultanément sur plusieurs nœuds actifs. À tout moment, l'équilibreur de charge dirige une demande utilisateur vers l'un des nœuds actifs du cluster défini. Les équilibreurs de charge vérifient l'état des nœuds de leur cluster cible et n'envoient aucune demande entrante à un nœud qui échoue au contrôle de santé. Par conséquent, un degré fondamental de haute disponibilité est atteint grâce à l'équilibrage de charge. De plus, étant donné qu'un équilibreur de charge effectue des contrôles de santé actifs et passifs sur tous les nœuds du cluster à des intervalles inférieurs à une seconde, le temps de basculement est quasi instantané.

La décision quant au nœud à diriger est basée sur les règles du système définies dans l'équilibreur de charge, notamment :

- tournoi à la ronde
- Affinité de session ou d'adresse IP, qui garantit que plusieurs demandes au sein d'une session ou provenant de la même adresse IP sont envoyées au même nœud du cluster
- Basé sur la latence
- Basé sur la charge

Applicabilité dans les architectures RTC

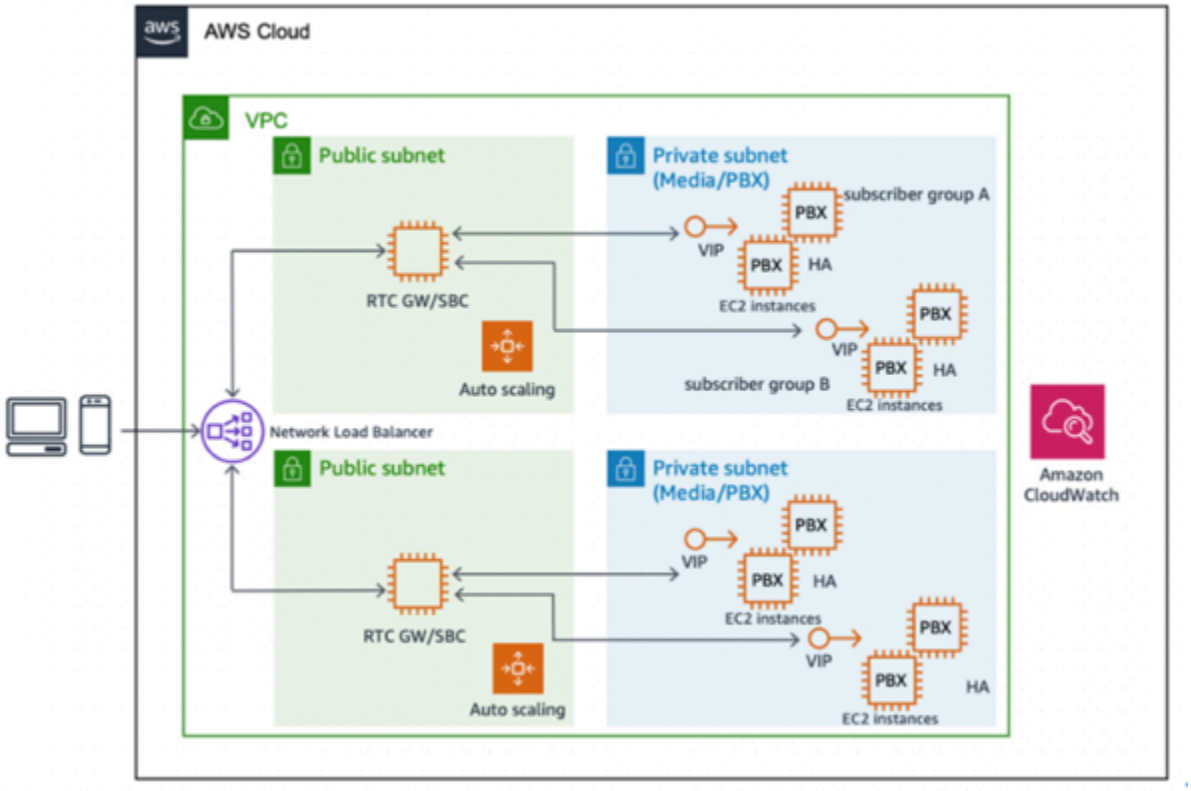
Le protocole WebRTC permet d'équilibrer facilement la charge des passerelles WebRTC via un équilibreur de charge basé sur le protocole HTTP, tel que Elastic Load Balancing (ELB), Application LoadBalancer (ALB) ou Network Load Balancer (NLB). La plupart des implémentations SIP reposant sur le transport via le protocole TCP (Transmission Control Protocol) et le protocole UDP (User Datagram Protocol), vous avez besoin d'un équilibrage de charge au niveau du réseau ou de la connexion, avec prise en charge du trafic TCP et UDP.

L'équilibrage de charge est activé AWS pour WebRTC à l'aide d'Application Load Balancer et d'Auto Scaling

Dans le cas des communications basées sur le WebRTC, Elastic Load Balancing fournit un équilibreur de charge entièrement géré, hautement disponible et évolutif qui sert de point d'entrée pour les demandes, qui sont ensuite dirigées vers un cluster cible EC2 d'instances associé à Elastic Load Balancing. Les demandes WebRTC étant apatrides, vous pouvez utiliser Amazon Auto EC2 Scaling pour fournir une évolutivité, une élasticité et une haute disponibilité entièrement automatisées et contrôlables.

L'Application Load Balancer fournit un service d'équilibrage de charge entièrement géré, hautement disponible à l'aide de plusieurs zones de disponibilité et évolutif. Cela prend en charge l'équilibrage de charge des WebSocket demandes qui gèrent la signalisation pour les applications WebRTC et la communication bidirectionnelle entre le client et le serveur à l'aide d'une connexion TCP de longue durée. L'Application Load Balancer prend également en charge le routage basé sur le contenu et les sessions persistantes, en acheminant les demandes du même client vers la même cible à l'aide de cookies générés par l'équilibreur de charge. Si vous activez les sessions persistantes, la même cible reçoit la demande et peut utiliser le cookie pour récupérer le contexte de session.

La figure suivante montre la topologie cible.



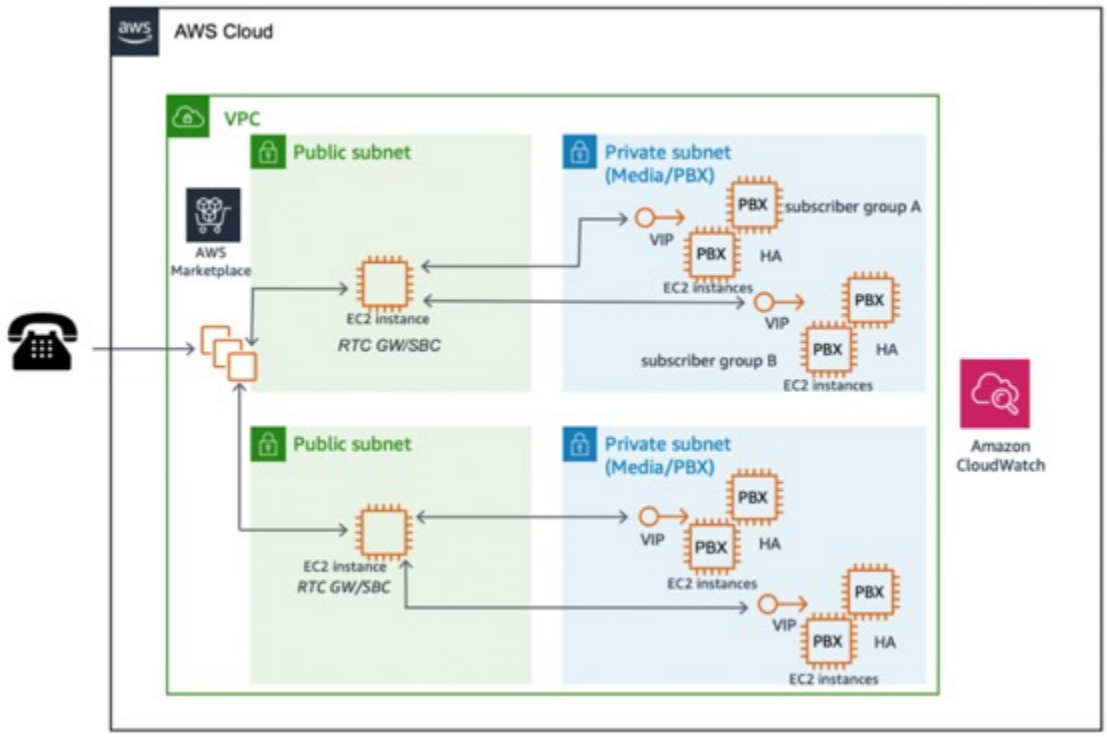
Évolutivité du WebRTC et architecture de haute disponibilité

Implémentation pour le protocole SIP à l'aide de Network Load Balancer ou d'un produit AWS Marketplace

Dans le cas des communications basées sur SIP, les connexions sont établies via TCP ou UDP, la majorité des applications RTC utilisant UDP. Si le protocole de signal de choix est le SIP/TCP, il est possible d'utiliser le Network Load Balancer pour un équilibrage de charge entièrement géré, hautement disponible, évolutif et performant.

Un Network Load Balancer fonctionne au niveau de la connexion (couche 4) et achemine les connexions vers des cibles telles que les EC2 instances Amazon, les conteneurs et les adresses IP en fonction des données du protocole IP. Idéal pour l'équilibrage de la charge du trafic TCP ou UDP, l'équilibrage de charge réseau est capable de traiter des millions de demandes par seconde tout en maintenant des latences extrêmement faibles. Il est intégré à d'autres services AWS populaires, tels qu'Amazon EC2 Auto Scaling, Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) et. AWS CloudFormation

Si des connexions SIP sont initiées, une autre option consiste à utiliser un off-the-shelf logiciel AWS Marketplacecommercial (COTS). Elle AWS Marketplace propose de nombreux produits capables de gérer l'UDP et d'autres types d'équilibrage de charge de connexion de couche 4. Les COTS incluent généralement la prise en charge de la haute disponibilité et s'intègrent généralement à des fonctionnalités, telles qu'Amazon EC2 Auto Scaling, pour améliorer encore la disponibilité et l'évolutivité. La figure suivante montre la topologie cible :



Évolutivité RTC basée sur le protocole SIP avec le produit AWS Marketplace

Équilibrage de charge et basculement entre régions basés sur le DNS

Amazon Route 53 fournit un service DNS mondial qui peut être utilisé comme point de terminaison public ou privé pour permettre aux clients RTC de s'inscrire et de se connecter à des applications multimédia. Avec Amazon Route 53, les contrôles de santé du DNS peuvent être configurés pour acheminer le trafic vers des points de terminaison sains ou pour surveiller indépendamment l'état de santé de votre application.

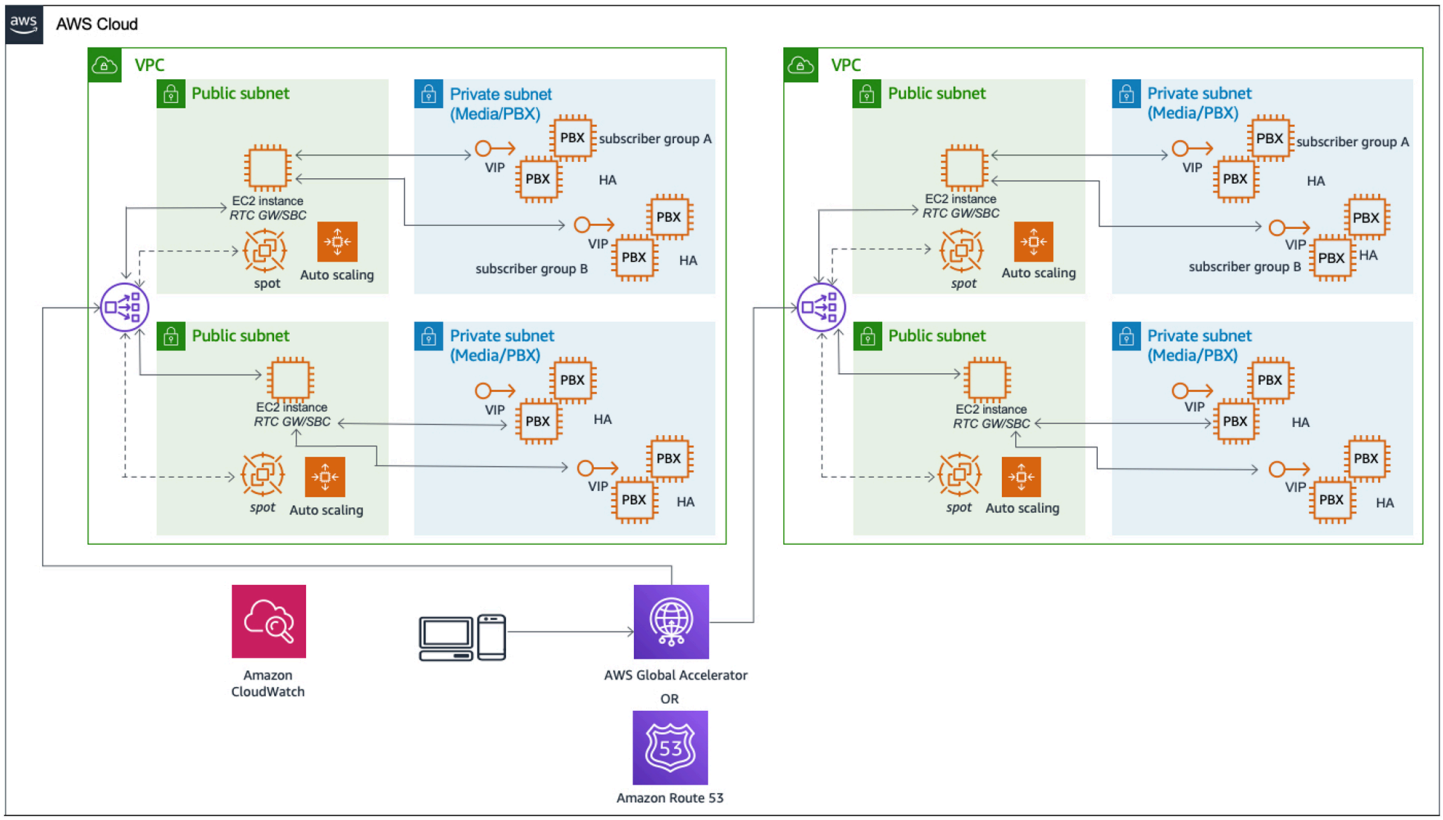
La fonctionnalité Amazon Route 53 Traffic Flow vous permet de gérer facilement le trafic mondial par le biais de différents types de routage, notamment le routage basé sur la latence, le géo-DNS, la géoproximité et le round robin pondéré. Tous ces éléments peuvent être combinés au DNS Failover pour permettre une variété d'architectures à faible latence et tolérantes aux pannes. L'éditeur visuel simple Amazon Route 53 Traffic Flow vous permet de gérer la manière dont vos utilisateurs finaux sont acheminés vers les points de terminaison de votre application, que ce soit dans une seule région AWS ou répartis dans le monde entier.

Dans le cas des déploiements mondiaux, la politique de routage basée sur la latence de Route 53 est particulièrement utile pour diriger les clients vers le point de présence le plus proche d'un serveur multimédia afin d'améliorer la qualité de service associée aux échanges multimédias en temps réel.

Notez que pour appliquer un basculement vers une nouvelle adresse DNS, les caches des clients doivent être vidés. En outre, les modifications du DNS peuvent être retardées lorsqu'elles sont propagées sur les serveurs DNS mondiaux. Vous pouvez gérer l'intervalle d'actualisation pour les recherches DNS à l'aide de l'attribut Time to Live. Cet attribut est configurable au moment de configurer les politiques DNS.

Pour atteindre rapidement les utilisateurs du monde entier ou pour répondre aux exigences liées à l'utilisation d'une adresse IP publique unique, il AWS Global Accelerator peut également être utilisé pour le basculement entre régions. AWS Global Acceleratorest un service réseau qui améliore la disponibilité et les performances des applications à portée locale et mondiale. AWS Global Accelerator fournit des adresses IP statiques qui agissent comme un point d'entrée fixe vers les points de terminaison de vos applications, tels que vos équilibreurs de charge d'application, vos équilibreurs de charge réseau ou les EC2 instances Amazon dans une ou plusieurs régions AWS. Il utilise le réseau mondial AWS pour optimiser le chemin entre vos utilisateurs et vos applications, en améliorant les performances, telles que la latence de votre trafic TCP et UDP.

AWS Global Accelerator surveille en permanence l'état des points de terminaison de votre application et redirige automatiquement le trafic vers les points de terminaison sains les plus proches en cas de défaillance des terminaux actuels. Pour répondre AWS Global Accelerator à des exigences de sécurité supplémentaires, le Site-to-Site VPN accéléré améliore les performances des connexions VPN en acheminant intelligemment le trafic via le réseau mondial AWS et les emplacements périphériques d'AWS.



Conception de haute disponibilité interrégionale à l'aide d'AWS Global Accelerator ou d'Amazon Route 53

Durabilité des données et haute disponibilité avec stockage persistant

La plupart des applications RTC s'appuient sur le stockage persistant pour stocker et accéder aux données à des fins d'authentification, d'autorisation, de comptabilité (données de session, enregistrements détaillés des appels, etc.), de surveillance opérationnelle et de journalisation. Dans un centre de données traditionnel, garantir la haute disponibilité et la durabilité des composants de stockage persistants (bases de données, systèmes de fichiers, etc.) nécessite généralement de lourdes tâches via la configuration d'un réseau de stockage (SAN), la conception d'un réseau redondant de disques indépendants (RAID) et des processus de sauvegarde, de restauration et de basculement. Cela simplifie et améliore AWS Cloud considérablement les pratiques traditionnelles des centres de données en matière de durabilité et de disponibilité des données.

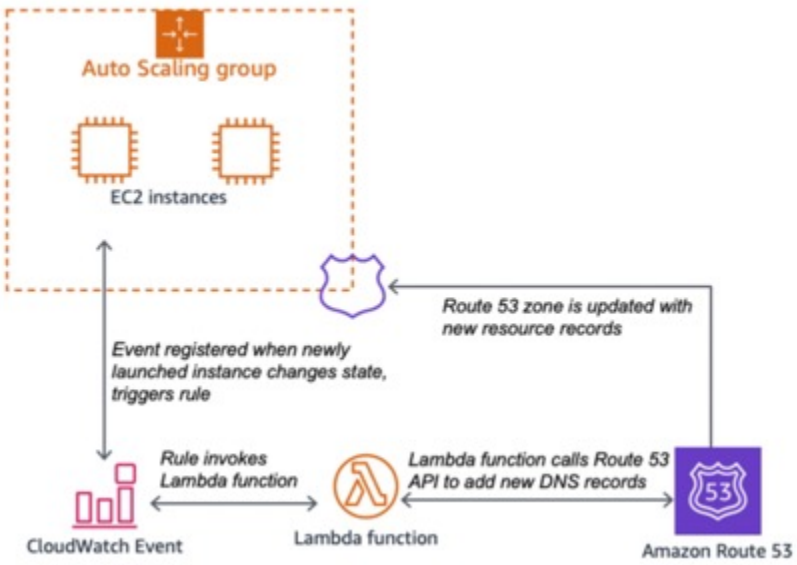
Pour le stockage d'objets et le stockage de fichiers, AWS des services tels qu'Amazon Simple Storage Service (Amazon S3) et Amazon Elastic File System (Amazon EFS) fournissent une haute

disponibilité et une évolutivité gérées. Amazon S3 a une durabilité des données de 99,999999999 % (11 neuf).

Pour le stockage des données transactionnelles, les clients ont la possibilité de tirer parti de l'Amazon Relational Database Service (Amazon RDS) entièrement géré qui prend en charge Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle et Microsoft SQL Server avec des déploiements à haute disponibilité. Pour la fonction d'enregistrement, le profil des abonnés ou le stockage des dossiers comptables (tels que CDRs), Amazon RDS fournit une option tolérante aux pannes, hautement disponible et évolutive.

Dimensionnement dynamique avec AWS Lambda Amazon Route 53 et Amazon EC2 Auto Scaling

AWS permet d'enchaîner les fonctionnalités et d'intégrer des fonctions sans serveur personnalisées en tant que service en fonction des événements de l'infrastructure. L'un de ces modèles de conception qui a de nombreuses utilisations polyvalentes dans les applications RTC est la combinaison du dimensionnement automatique des hooks du cycle de vie avec Amazon CloudWatch Events, Amazon Route 53 et AWS Lambdales fonctions. AWS Lambda les fonctions peuvent intégrer n'importe quelle action ou logique. La figure suivante montre comment ces fonctionnalités associées peuvent améliorer la fiabilité et l'évolutivité du système grâce à l'automatisation.



Dimensionnement automatique avec mises à jour dynamiques d'Amazon Route 53

WebRTC à haute disponibilité avec Amazon Kinesis Video Streams

Amazon Kinesis Video Streams propose un streaming multimédia en temps réel via WebRTC, permettant aux utilisateurs de capturer, de traiter et de stocker des flux multimédias à des fins de lecture, d'analyse et d'apprentissage automatique. Ces flux sont hautement disponibles, évolutifs et conformes aux normes WebRTC. Amazon Kinesis Video Streams inclut un point de terminaison de signalisation WebRTC pour une détection rapide des pairs et l'établissement de connexions sécurisées. Il inclut des utilitaires gérés de traversée de session pour NAT (STUN) et des utilitaires de traversée utilisant des relais autour des points de terminaison NAT (TURN) pour l'échange de contenu multimédia en temps réel entre pairs. Il inclut également un SDK open source gratuit qui s'intègre directement au microprogramme de la caméra pour permettre une communication sécurisée avec les points de terminaison Amazon Kinesis Video Streams, permettant ainsi la découverte par les pairs et le streaming multimédia. Enfin, il fournit des bibliothèques clientes pour Android et iOS JavaScript qui permettent aux lecteurs mobiles et Web compatibles WebRTC de découvrir et de se connecter en toute sécurité à un appareil photo pour le streaming multimédia et la communication bidirectionnelle.

Trunking SIP hautement disponible avec Amazon Chime Voice Connector

Amazon Chime Voice Connector fournit un service de jonction pay-as-you-go SIP qui permet aux entreprises de passer et/ou de recevoir des appels téléphoniques sécurisés et peu coûteux avec leurs systèmes téléphoniques. Amazon Chime Voice Connector est une alternative peu coûteuse aux liaisons SIP des fournisseurs de services ou aux interfaces à débit primaire (ISDN) du réseau numérique à intégration de services (). PRIs Les clients ont la possibilité d'activer les appels entrants, sortants ou les deux.

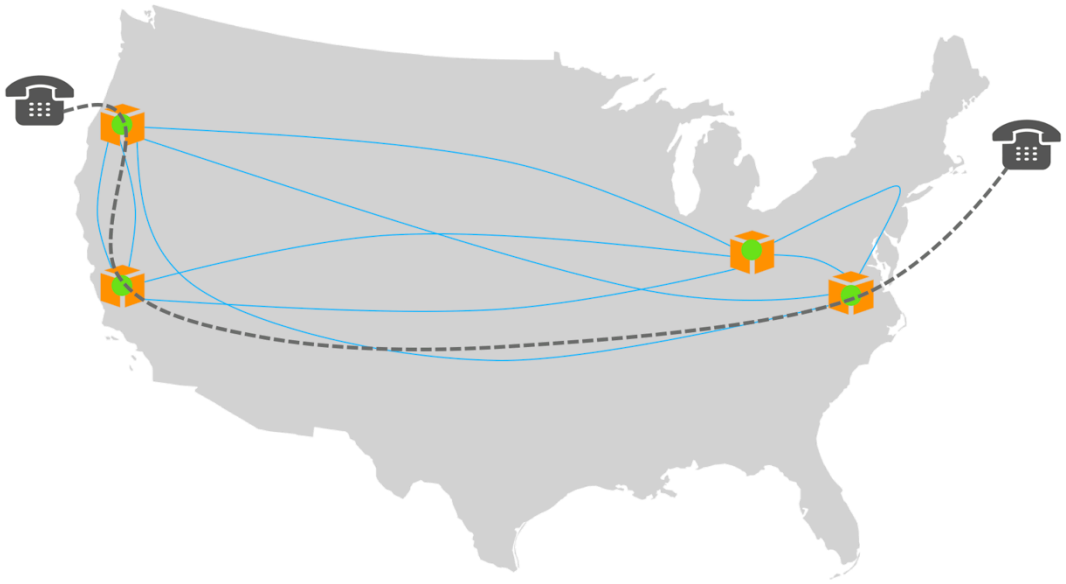
Le service utilise le AWS réseau pour offrir une expérience d'appel hautement disponible sur plusieurs réseaux Régions AWS. Vous pouvez diffuser du son à partir d'appels téléphoniques à jonction SIP ou transférer des flux d'enregistrement multimédia basés sur le protocole SIP (SIPREC) vers Amazon Kinesis Video Streams pour obtenir des informations sur les appels professionnels en temps réel. Vous pouvez créer rapidement des applications d'analyse audio grâce à l'intégration à Amazon Transcribe et à d'autres bibliothèques d'apprentissage automatique courantes.

Les meilleures pratiques du terrain

Cette section résume les meilleures pratiques mises en œuvre par certains des AWS clients les plus importants et les plus prospères qui exécutent d'importantes charges de travail liées au protocole SIP (Session Initiation Protocol) en temps réel. AWS les clients qui souhaitent gérer leur propre infrastructure SIP dans le cloud public trouveront ces meilleures pratiques utiles, car elles peuvent contribuer à accroître la fiabilité et la résilience du système en cas de différents types de défaillances. Bien que certaines de ces meilleures pratiques soient spécifiques au protocole SIP, la plupart d'entre elles sont applicables à toute application de communication en temps réel exécutée sur AWS.

Création d'une superposition SIP

AWS dispose d'un backbone réseau robuste, évolutif et redondant qui assure la connectivité entre différents Régions AWS. Lorsqu'un événement réseau, tel qu'une coupure de fibre, dégrade une liaison AWS dorsale, le trafic est rapidement transféré vers des chemins redondants à l'aide de protocoles de routage au niveau du réseau, tels que le Border Gateway Protocol (BGP). Cette ingénierie du trafic au niveau du réseau est une boîte noire pour les AWS clients et la plupart d'entre eux ne remarquent même pas ces événements de basculement. Cependant, les clients qui exécutent des charges de travail en temps réel, telles que la voix, la vidéo de haute qualité et la messagerie à faible latence, remarquent parfois ces événements. Alors, comment un AWS client peut-il mettre en œuvre sa propre ingénierie du trafic en plus de ce qui est fourni AWS au niveau du réseau ? La solution consiste à déployer une infrastructure SIP sur de nombreux sites différents Régions AWS. Dans le cadre des fonctionnalités de contrôle des appels, le protocole SIP permet également d'acheminer les appels via des proxys SIP spécifiques.

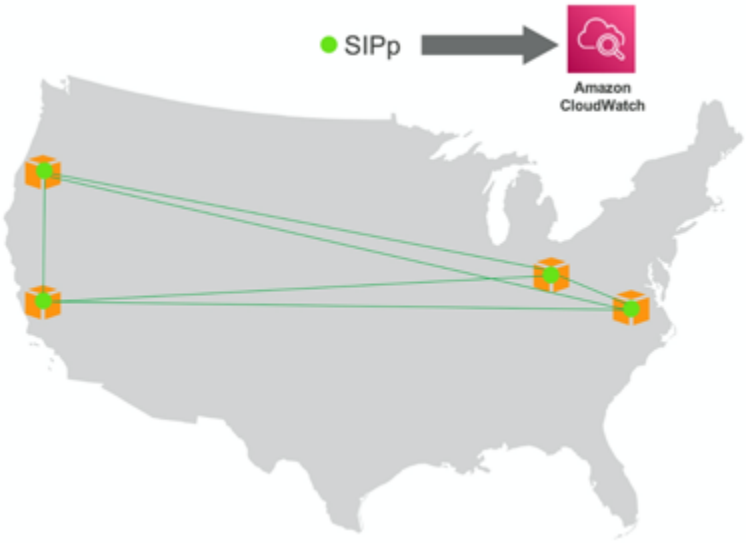


Utilisation du routage SIP pour remplacer le routage réseau

Dans la figure précédente, l'infrastructure SIP (représentée par des points verts à l'intérieur des cubes) fonctionne dans les quatre régions des États-Unis. Les lignes bleues continues représentent une représentation fictive de l' AWS épine dorsale. Si aucun routage SIP n'est mis en œuvre, un appel provenant de la côte ouest des États-Unis et à destination de la côte est des États-Unis passe par le lien principal qui relie directement les régions de l'Oregon et de la Virginie. Le schéma montre comment un client peut annuler le routage au niveau du réseau et passer le même appel entre l'Oregon et la Virginie en passant par la Californie à l'aide du routage SIP. Ce type d'ingénierie du trafic SIP peut être mis en œuvre à l'aide de proxys SIP et de passerelles multimédia en fonction de paramètres réseau tels que les retransmissions SIP et les préférences commerciales spécifiques des clients.

Effectuez une surveillance détaillée

Les utilisateurs finaux des applications vocales et vidéo en temps réel attendent le même niveau de performance qu'avec les services de téléphonie traditionnels. Ainsi, lorsqu'ils rencontrent des problèmes avec une application, cela finit par nuire à la réputation du fournisseur. Pour être proactif plutôt que réactif, il est impératif de déployer une surveillance détaillée à chaque partie du système qui dessert les utilisateurs finaux.



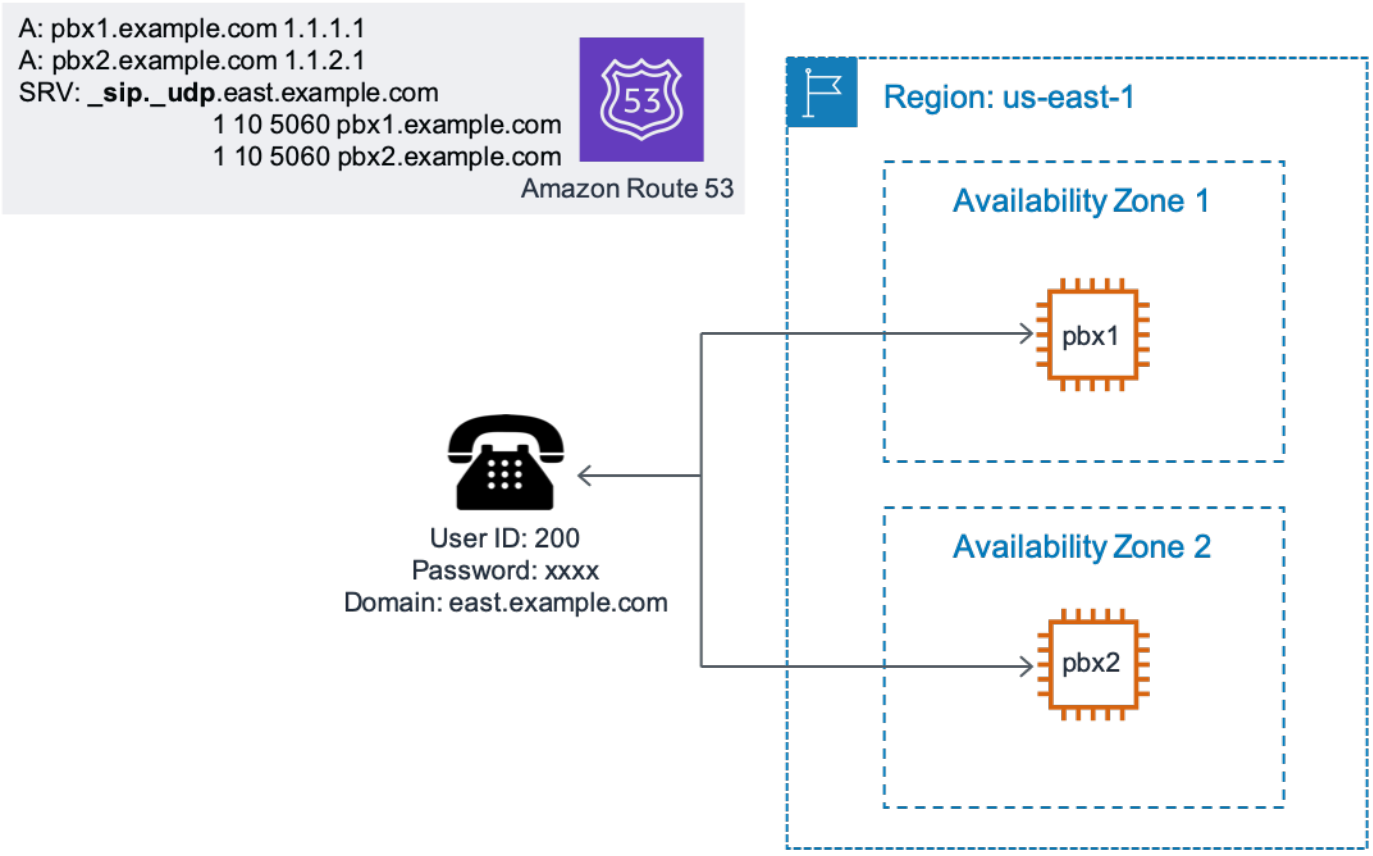
Utilisation SIPp pour surveiller l'infrastructure VoIP

De nombreux outils open source, tels que iPerf ou SIPp, et VOIPMonitor, sont disponibles pour surveiller le trafic SIP/RTP. Dans l'exemple précédent, les nœuds exécutant le protocole SIP en mode client et en mode serveur mesurent des métriques SIP telles que les appels réussis et les retransmissions SIP entre les quatre États-Unis Régions AWS. Ces métriques peuvent ensuite être exportées vers Amazon CloudWatch à l'aide d'un script personnalisé. À l'aide de ces indicateurs personnalisés CloudWatch, les clients peuvent créer des alarmes sur la base d'une certaine valeur de seuil. Des mesures correctives automatiques ou manuelles peuvent ensuite être prises en fonction de l'état de ces CloudWatch alarmes.

Pour les clients qui ne souhaitent pas allouer les ressources d'ingénierie nécessaires au développement et à la maintenance d'un système de surveillance personnalisé, de nombreuses bonnes solutions de surveillance VoIP sont disponibles sur le marché, telles que. ThousandEyes Un exemple d'action corrective consiste à modifier le routage SIP en fonction de l'augmentation du nombre de retransmissions SIP.

Utiliser le DNS pour l'équilibrage de charge et le système flottant IPs pour le basculement

Les clients de téléphonie IP qui prennent en charge la fonctionnalité DNS SRV peuvent utiliser efficacement la redondance intégrée à l'infrastructure en équilibrant la charge des clients entre différents/. SBCs PBXs



Utilisation des enregistrements DNS SRV pour équilibrer la charge des clients SIP

La figure précédente montre comment les clients peuvent utiliser les enregistrements SRV pour équilibrer la charge du trafic SIP. Tout client de téléphonie IP compatible avec la norme SRV recherchera le SIP. `<transport protocol>`préfixe dans un enregistrement DNS de type SRV. Dans l'exemple, la section de réponse du DNS contient les deux versions PBXs exécutées dans différentes zones de AWS disponibilité. Cependant, outre le point de terminaison URIs, l'enregistrement SRV contient trois informations supplémentaires :

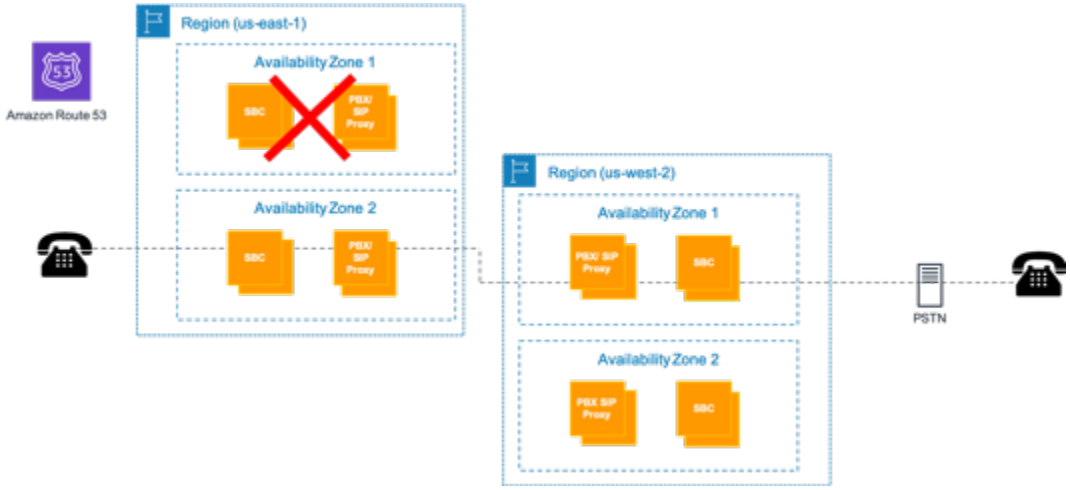
- Le premier chiffre est la priorité (1 dans l'exemple ci-dessus). Une priorité inférieure est préférable à une priorité plus élevée.
- Le deuxième chiffre est le poids (10 dans l'exemple ci-dessus).
- Et le troisième chiffre est le port à utiliser (5060).

La priorité étant la même (1) pour les deux PBXs serveurs, les clients utilisent le poids pour équilibrer la charge entre les deux PBXs. Dans ce cas, étant donné que les pondérations sont les mêmes, le trafic SIP doit être équilibré de manière égale entre les deux PBXs.

Le DNS peut être une bonne solution pour équilibrer la charge des clients, mais qu'en est-il de la mise en œuvre du basculement en modifiant ou en mettant à jour les enregistrements DNS « A » ? Cette méthode est déconseillée en raison de l'incohérence constatée dans le comportement de mise en cache du DNS au sein du client et des nœuds intermédiaires. Une meilleure approche pour le basculement intra-AZ entre un cluster de nœuds SIP consiste à utiliser la réattribution EC2 IP, qui permet de réattribuer instantanément l'adresse IP d'un hôte altéré à un hôte sain à l'aide de l'API. EC2 Associée à une solution de surveillance détaillée et de vérification de l'état de santé, la réattribution IP d'un nœud défaillant garantit le transfert du trafic vers un hôte sain en temps opportun, minimisant ainsi les perturbations pour l'utilisateur final.

Utiliser plusieurs zones de disponibilité

Chacune Région AWS est subdivisée en zones de disponibilité distinctes. Chaque zone de disponibilité possède ses propres réseaux d'alimentation, de refroidissement et de connectivité réseau et constitue ainsi un domaine de défaillance isolé. Dans le cadre des concepts de AWS, les clients sont invités à exécuter leurs charges de travail dans plusieurs zones de disponibilité. Cela garantit que les applications des clients peuvent résister même à une défaillance complète de la zone de disponibilité, un événement très rare en soi. Cette recommandation concerne également l'infrastructure SIP en temps réel.



Gestion des défaillances de zone de disponibilité

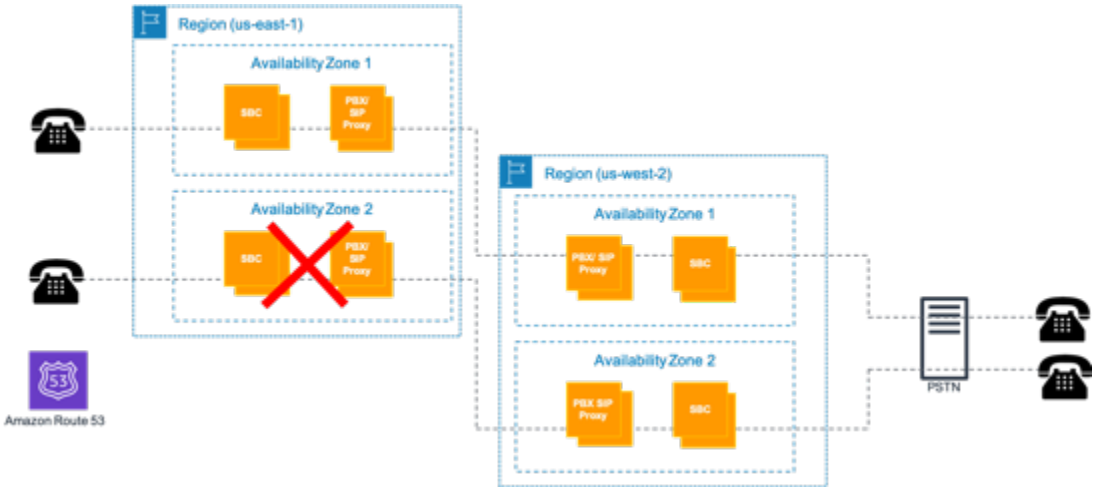
Supposons qu'un événement catastrophique (tel qu'un ouragan de catégorie 5) entraîne une interruption complète de la zone de disponibilité dans la région us-east-1. Lorsque l'infrastructure fonctionne comme indiqué dans le schéma, tous les clients SIP initialement enregistrés auprès des nœuds de la zone de disponibilité défaillante doivent se réenregistrer auprès des nœuds SIP exécutés dans la zone de disponibilité #2. (Testez ce comportement avec vos clients/téléphones SIP

pour vous assurer qu'il est pris en charge.) Bien que les appels SIP actifs au moment de la panne de la zone de disponibilité soient perdus, tous les nouveaux appels sont acheminés via la zone de disponibilité 2.

En résumé, les enregistrements DNS SRV doivent pointer le client vers plusieurs enregistrements « A », un dans chaque zone de disponibilité. Chacun de ces enregistrements « A » doit, à son tour, pointer vers plusieurs adresses IP SBCs de/ PBXs dans cette zone de disponibilité, ce qui garantit la résilience entre les zones de disponibilité et d'interdisponibilité. Le basculement entre zones intra et interdisponibles peut être mis en œuvre en utilisant la réattribution d'adresses IP si elles sont publiques. IPs Le mode privé ne peut IPs toutefois pas être réattribué entre les zones de disponibilité. Si un client utilise un adressage IP privé, il devra compter sur le réenregistrement des clients SIP auprès du SBC/PBX de sauvegarde pour le basculement dans la zone d'interdisponibilité.

Maintenez le trafic dans une seule zone de disponibilité et utilisez des groupes EC2 de placement

Également connue sous le nom d'affinité de zone de disponibilité, cette bonne pratique s'applique également aux rares cas de défaillance complète d'une zone de disponibilité. Il est recommandé d'éliminer tout trafic inter-AZ afin que tout trafic SIP ou RTP entrant dans une zone de disponibilité reste dans cette zone de disponibilité jusqu'à ce qu'il quitte la région.



Affinité de zone de disponibilité (au maximum, 50 % des appels actifs sont perdus)

La figure précédente montre une architecture simplifiée qui utilise l'affinité de zone de disponibilité. L'avantage comparatif de cette approche apparaît clairement si l'on tient compte des effets d'une interruption complète de la zone de disponibilité. Comme le montre le schéma, si la zone de disponibilité 2 est perdue, 50 % des appels actifs sont affectés au maximum (en supposant un

équilibrage de charge égal entre les zones de disponibilité). Si l'affinité de zone de disponibilité n'avait pas été mise en œuvre, certains appels circuleraient entre les zones de disponibilité d'une région et une panne affecterait très probablement plus de 50 % des appels actifs.

Pour minimiser la latence du trafic, AWS vous recommande également d'utiliser des groupes de EC2 placement au sein de chaque zone de disponibilité. Les instances lancées au sein du même groupe de EC2 placement ont une bande passante plus élevée et une latence réduite, EC2 ce qui garantit la proximité réseau de ces instances les unes par rapport aux autres.

Utiliser des types d' EC2 instances réseau améliorés

Choisir le bon type d'instance sur Amazon EC2 garantit la fiabilité du système ainsi qu'une utilisation efficace de l'infrastructure. EC2 propose une large sélection de types d'instances optimisés pour s'adapter à différents cas d'utilisation. Les types d'instance incluent diverses combinaisons de capacité de processeur, de mémoire, de stockage et de mise en réseau et vous offrent la flexibilité nécessaire pour choisir les combinaisons de ressources les plus adaptées à vos applications. Ces types d'instances réseau améliorés garantissent que les charges de travail SIP qui y sont exécutées ont accès à une bande passante constante et à une latence globale comparativement plus faible. Un ajout récent à Amazon EC2 est la disponibilité de l'adaptateur réseau Elastic (ENA) qui fournit jusqu'à 100 Gbit/s de bande passante. Le dernier catalogue des types d' EC2 instances et des fonctionnalités associées se trouve sur la page des types d'EC2 instances.

Pour la plupart des clients, la dernière génération d'instances optimisées pour le calcul devrait offrir le meilleur rapport qualité-prix. Par exemple, le C5N prend en charge le nouvel adaptateur réseau Elastic avec une bande passante allant jusqu'à 100 Gbit/s avec des millions de paquets par seconde (PPS). La plupart des applications en temps réel bénéficieraient également de l'utilisation de l'Intel Data Plane Developer Kit (DPDK), qui peut considérablement améliorer le traitement des paquets réseau.

Cependant, il est toujours recommandé de comparer les différents types d' EC2instances en fonction de vos besoins afin de déterminer le type d'instance qui vous convient le mieux. L'analyse comparative vous permet également de trouver d'autres paramètres de configuration, tels que le nombre maximum d'appels qu'un certain type d'instance peut traiter à la fois.

Considérations sur la sécurité

Les composants de l'application RTC s'exécutent généralement directement sur des EC2 instances Amazon connectées à Internet. Outre le protocole TCP, les flux utilisent des protocoles tels que UDP et SIP. Dans ces cas, AWS Shield Standard protège les EC2 instances Amazon contre les attaques des couches d'infrastructure communes (couches 3 et 4) DDo S, telles que les attaques par réflexion UDP, la réflexion DNS, la réflexion NTP, la réflexion SSDP, etc. AWS Shield Standard utilise diverses techniques, telles que la mise en forme du trafic basée sur les priorités, qui sont automatiquement activées lorsqu'une signature d'attaque DDo S bien définie est détectée.

AWS fournit également une protection avancée contre les attaques DDo S de grande envergure et sophistiquées pour ces applications en activant AWS Shield Advanced les adresses IP élastiques. AWS Shield Advanced fournit une détection DDo S améliorée qui détecte automatiquement le type de AWS ressource et la taille de l' EC2instance et applique des mesures d'atténuation prédéfinies appropriées avec des protections contre les inondations SYN ou UDP. Les clients peuvent également créer leurs propres profils d'atténuation personnalisés en faisant appel à l'équipe AWS DDo S Response Team (DRT) 24 heures sur 24, 7 jours sur 7. AWS Shield Advanced AWS Shield Advanced garantit également que lors d'une attaque DDo S, toutes vos listes de contrôle d'accès réseau Amazon VPC (ACLs) sont automatiquement appliquées à la frontière du AWS réseau, vous donnant ainsi accès à une bande passante et à une capacité de nettoyage supplémentaires pour atténuer les attaques S volumétriques DDo de grande envergure.

Conclusion

Les charges de travail de communication en temps réel (RTC) peuvent être déployées AWS pour atteindre l'évolutivité, l'élasticité et la haute disponibilité tout en répondant aux exigences clés. Aujourd'hui, plusieurs clients utilisent AWS, ses partenaires et des solutions open source pour exécuter les charges de travail RTC à moindre coût, avec une agilité accrue et une empreinte mondiale réduite.

Les architectures de référence et les meilleures pratiques présentées dans ce white paper peuvent aider les clients à configurer avec succès les charges de travail RTC AWS et à optimiser les solutions afin de répondre aux besoins des utilisateurs finaux tout en optimisant leur utilisation pour le cloud.

Acronymes

Les acronymes utilisés dans ce document incluent :

ACL — Liste de contrôle d'accès

ALB — Application Load Balancer

APNs — Service de notification push Apple

BGP — Protocole de passerelle frontalière

CDR — Enregistrements détaillés des appels

COTS — logiciel commercial off-the-shelf

DDoS — distribué denial-of-service

DNS — Système de noms de domaine

DPDK — Kit de développement Intel Data Plane

Équipe d'intervention DRT — DDo S

ENA — Adaptateur réseau élastique

EPC — Evolved Packet Core

FCM — Messagerie cloud Firebase

HA — Haute disponibilité

IRC — Chat par relais Internet

ISDN — Réseau numérique à services intégrés

NAT — traduction d'adresses réseau

OPUS — support utilisateur pour le positionnement en ligne

PBX — Succursale privée

PRI — Interface à débit primaire

PSTN — Réseau téléphonique public commuté

RAID : ensemble redondant de disques indépendants

RTC — communication en temps réel

RTP —Protocole de transport en temps réel

SAN : réseau de stockage

SBC — contrôleur de session en bordure

SIP —Protocole d'initiation de session

SPOF : points de défaillance uniques

SRV — Service

SS7 — Système de signalisation n.7

STUN — Utilitaires de traversée de session pour NAT

SYN —Synchroniser

TCP — Protocole de contrôle de transmission

TDM — multiplexage par répartition dans le temps

TURN — Traversée à l'aide de relais autour du NAT

UDP — Protocole de datagramme utilisateur

URI — Identifiants de ressources uniformes

VIP — IP virtuelle

VNF — Fonction de réseau virtuel

VoIP — VoIP

VPC — Cloud privé virtuel

WebRTC — communication Web en temps réel

Collaborateurs

Les personnes et organisations suivantes ont contribué à l'élaboration du présent document :

- Mounir Chennana, architecte de solutions senior, Amazon Web Services
- Mohammed Al-Mehdar, architecte de solutions senior, Amazon Web Services
- Ejaz Sial, architecte de solutions senior, Amazon Web Services
- Ahmad Khan, architecte de solutions senior, Amazon Web Services
- Tipu Qureshi, ingénieur principal, Amazon Web AWS Support Services
- Hasan Khan, responsable technique senior des comptes, Amazon Web Services
- Shoma Chakravarty, responsable technique des télécommunications chez Amazon Web Services chez WW

Révisions du document

Pour être informé des mises à jour de ce livre blanc, abonnez-vous au flux RSS.

Modification	Description	Date
Livre blanc mis à jour	Mise à jour pour les derniers services et fonctionnalités.	5 mai 2022
Livre blanc mis à jour	Mise à jour pour les derniers services et fonctionnalités.	13 février 2020
Publication initiale	Livre blanc publié pour la première fois.	1 octobre 2018

Avis

Il incombe aux clients de procéder à une évaluation indépendante des informations contenues dans le présent document. Ce document : (a) est fourni à titre informatif uniquement, (b) représente les offres de produits et les pratiques actuelles d'AWS, qui sont susceptibles d'être modifiées sans préavis, et (c) ne crée aucun engagement ni aucune garantie de la part d'AWS et de ses filiales, fournisseurs ou concédants de licence. Les produits ou services AWS sont fournis « tels quels » sans garanties, déclarations ou conditions d'aucune sorte, qu'elles soient explicites ou implicites. Les responsabilités et obligations d'AWS vis-à-vis de ses clients sont régies par les contrats AWS. Le présent document ne fait partie d'aucun, et ne modifie aucun, contrat entre AWS et ses clients.

AWS Glossaire

Pour la AWS terminologie la plus récente, consultez le AWS glossaire dans la Glossaire AWS référence.

Les traductions sont fournies par des outils de traduction automatique. En cas de conflit entre le contenu d'une traduction et celui de la version originale en anglais, la version anglaise prévaudra.