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

Bonnes pratiques pour la conception de charges de travail de communication en temps réel (RTC) à hautes disponibilité et évolutives sur AWS

Date de publication : 13 février 2020 (Révisions du document)

Résumé

Aujourd'hui, de nombreuses organisations cherchent à réduire leurs coûts et à obtenir la capacité de mise à l'échelle des charges de travail vocales, de messagerie et multimédias en temps réel. Ce livre blanc présente les bonnes pratiques pour gérer les charges de travail de communication en temps réel sur AWS et inclut des architectures de référence pour répondre à ces exigences. Ce document sert de guide aux personnes familiarisées avec la communication en temps réel sur la manière d'atteindre une haute disponibilité et une capacité de mise à l'échelle pour ces charges de travail.

Introduction

Les applications de télécommunication utilisant la voix, la vidéo et la messagerie comme canaux constituent une exigence clé 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 bonnes pratiques de conception pertinentes. Par le passé, les charges de travail RTC ont été déployées dans des centres de données sur site traditionnels avec des ressources dédiées.

Cependant, en raison d'un ensemble de fonctions matures et en plein essor, les charges de travail RTC peuvent être déployées sur Amazon Web Services (AWS) malgré des exigences de niveau de service strictes tout en alliant capacité de mise à l'échelle, élasticité et haute disponibilité. Aujourd'hui, plusieurs clients utilisent AWS, ses partenaires et des solutions open source pour exécuter des charges de travail RTC pour moins de frais, une agilité plus rapide, la capacité de se mondialiser en quelques minutes et les fonctions riches des services AWS.

Les clients tirent parti des fonctions AWS telles que les réseaux améliorés avec un Elastic Network Adapter (ENA) et la dernière génération d'instances Amazon Elastic Compute Cloud (EC2) pour bénéficier du kit de développement de plan de données (DPDK), de la virtualisation des I/O à racine unique (SR- IOV), de grandes pages, NVM Express (NVMe), la prise en charge de l'accès mémoire non uniforme (NUMA) ainsi que des instances à matériel nu pour répondre aux exigences de charge de travail RTC. Ces instances offrent une bande passante réseau allant jusqu'à 100 Gbit/s et des paquets proportionnels par seconde, offrant de meilleures performances pour les applications gourmandes en réseau. Pour la mise à l'échelle, Elastic Load Balancing propose Application Load Balancer, qui prend en charge WebSocket et Network Load Balancer pouvant traiter des millions de demandes par seconde. Pour l'accélération du réseau, AWS Global Accelerator fournit des adresses IP statiques qui agissent comme un point d'entrée fixe vers les points de terminaison de vos applications dans AWS. Il prend en charge les adresses IP statiques pour l'équilibreur de charge. Pour réduire la latence, les coûts et augmenter le débit de bande passante, AWS Direct Connect établit une connexion réseau dédiée depuis le site vers AWS. La jonction SIP gérée haute disponibilité est fournie par Amazon Chime Voice Connector. Amazon Kinesis Video Streams avec WebRTC permet de diffuser facilement du contenu multimédia bidirectionnel en temps réel avec une haute disponibilité.

Ce livre blanc présente des architectures de référence qui expliquent comment configurer des charges de travail RTC sur AWS et les bonnes pratiques pour optimiser les solutions afin de répondre aux exigences des utilisateurs finaux tout en optimisant pour le cloud. Le cœur de paquet évolué

(EPC) n'est pas inclus dans ce livre blanc, mais les bonnes pratiques détaillées peuvent être appliquées aux fonctions de réseau virtuel (VNF).

Composants de base de l'architecture RTC

Dans l'industrie des télécommunications, la communication en temps réel (RTC) fait généralement référence à des sessions multimédias en direct entre deux points de terminaison avec une latence minimale. Ces sessions peuvent être liées aux sujets suivants :

- Session vocale entre deux parties (p. ex. système téléphonique, mobile, VoIP)
- Messagerie instantanée (p. ex. chat, IRC)
- Session vidéo en direct (p. ex. vidéoconférence, téléprésence)

Chacune des solutions précédentes a certains composants en commun (p. ex. 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 certains composants uniques au type de média transmis (p. ex. service de diffusion, serveur de messagerie et 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 connexes illustrés dans la figure 1.

Figure 1 : Composants architecturaux essentiels pour RTC

Rubriques

Commutateur logiciel/PBX

Un commutateur logiciel ou PBX est le cerveau d'un système téléphonique vocal et fournit des informations pour établir, maintenir et acheminer un appel vocal à l'intérieur ou à l'extérieur de l'entreprise en utilisant différents composants. Tous les abonnés de l'entreprise doivent s'inscrire auprès du commutateur logiciel pour recevoir ou passer un appel. Une fonctionnalité importante du commutateur logiciel est de suivre chaque abonné et de savoir comment le joindre en utilisant les autres composants du réseau vocal.

Session Border Controller (SBC)

Un Session Border Controller (SBC) se trouve à la périphérie d'un réseau vocal et assure le suivi de tout le trafic entrant et sortant (contrôle et plans de données). L'une des responsabilités clé d'un SBC est de protéger le système vocal contre toute utilisation malveillante. Le SBC peut être utilisé pour s'interconnecter avec des jonctions SIP (Session Initiation Protocol) pour une connectivité externe. Certains SBC offrent également des fonctionnalités de transcodage pour convertir les CODECS d'un format à un autre. Enfin, la plupart des SBC fournissent également des fonctionnalités NAT Traversal qui aident à 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 RTPC et des jonctions SIP pour se connecter aux réseaux RTPC hérités.

Passerelle PSTN

La passerelle du réseau téléphonique commuté public (RTCP) convertit la signalisation (entre SIP et SS7) et le média (entre le RTP et le multiplexage temporel [TDM] à l'aide du transcodage CODEC). Les passerelles PSTN sont toujours situées à la périphérie, à proximité du réseau RTPC.

Jonction SIP

Dans une jonction SIP, l'entreprise ne met pas fin à ses appels sur un réseau TDM (SS7), mais les flux entre l'entreprise et les opérateurs de télécommunications restent sur IP. La plupart des jonctions SIP sont établies à l'aide de SBC. L'entreprise doit se mettre d'accord sur les règles de sécurité prédéfinies de la compagnie de télécommunications, telles que l'autorisation d'une certaine plage d'adresses IP, de ports, etc.

Passerelle multimédia (transcodeur)

Une solution vocale classique permet différents types de codecs. Certains des CODEC courants sont G.711 μ-law pour l'Amérique du Nord, G.711 A-law pour l'extérieur de l'Amérique du Nord, G.729 et G.722. Lorsque deux périphériques utilisant deux CODEC différents communiquent entre eux, un serveur multimédia traduit le flux de CODEC entre les appareils. En d'autres termes, une passerelle multimédia traite les médias et garantit que les terminaux sont en mesure de communiquer entre eux.

WebRTC et passerelles WebRTC

La communication en temps réel pour le web (WebRTC) vous permet d'établir un appel à partir d'un navigateur web ou de demander des ressources au serveur backend à l'aide d'une API. La technologie est conçue en tenant compte de la technologie cloud et fournit donc diverses API qui pourraient être utilisées pour établir un appel. Étant donné que toutes les solutions vocales (y compris SIP) ne prennent pas en charge ces API, la passerelle WebRTC est requise pour traduire les appels d'API en messages SIP et vice versa.

La figure 2 montre un modèle de conception pour une architecture WebRTC à haute disponibilité. Le trafic entrant des clients WebRTC est équilibré par un Application Load Balancer Amazon avec WebRTC exécuté sur des instances EC2 qui font partie d'un groupe Auto Scaling.

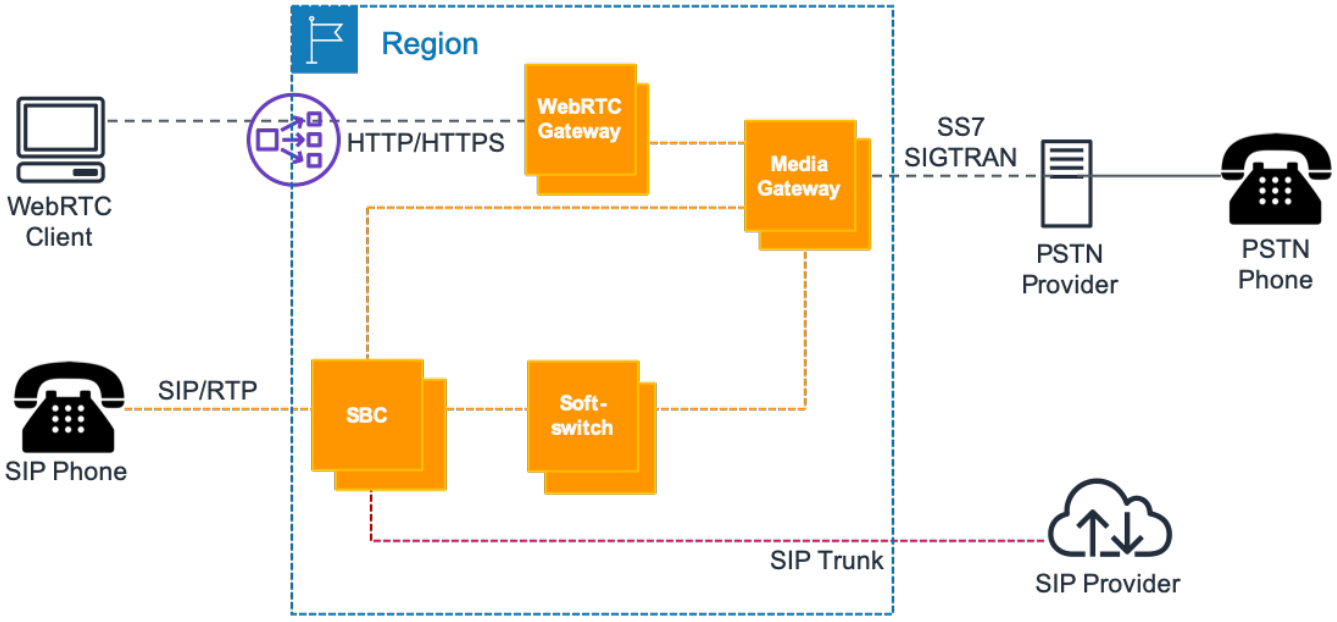

Figure 2 : Topologie de base d'un système RTC pour la voix

Un autre modèle de conception pour le trafic SIP et RTP consiste à utiliser des paires de SBC sur Amazon EC2 en mode passif actif dans les zones de disponibilité (figure 3). Ici, une adresse IP élastique peut être déplacée dynamiquement entre les instances en cas d'échec lorsque le DNS ne peut pas être utilisé.

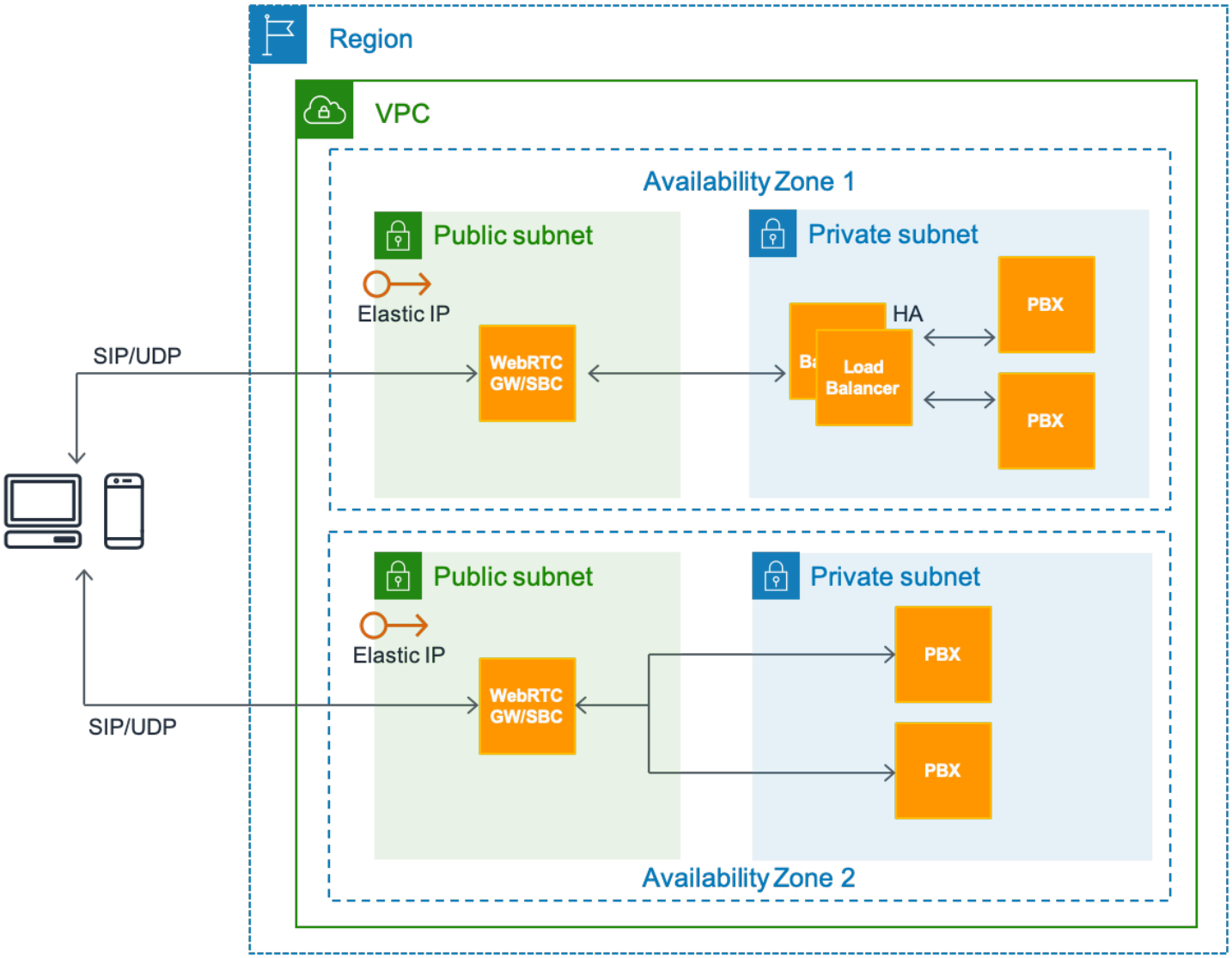

Figure 3 : Architecture RTC utilisant Amazon EC2 dans un VPC

Haute disponibilité et capacité de mise à l'échelle sur AWS

La plupart des fournisseurs de communications en temps réel s'alignent sur des niveaux de service qui offrent une disponibilité de 99,9 % à 99,999 %. Selon le degré de haute disponibilité que vous souhaitez, vous devez prendre des mesures de plus en plus sophistiquées tout au long du cycle de vie complet de l'application. Nous vous recommandons de suivre ces directives pour atteindre un degré élevé de haute disponibilité :

- Concevez le système de manière à ne pas avoir de point d'échec unique. Utilisez des mécanismes automatisés de surveillance, de détection des échecs et de basculement pour les composants avec et sans état.

- Les points uniques de défaillance (SPOF) sont généralement éliminés avec une configuration de redondance N+1 ou 2N, où N+1 est obtenu via la répartition de charge entre les nœuds actifs-actifs, et 2N est atteint par une paire de nœuds avec une configuration active-secours.
- AWS propose plusieurs méthodes pour atteindre la haute disponibilité par le biais des deux approches, comme par le biais d'un cluster évolutif à charge équilibrée ou en endossant une paire active-secours.

- Équipez et testez correctement la disponibilité du système.
- Préparez les procédures d'exploitation des mécanismes manuels destinés à répondre à l'échec, à l'atténuer et à s'en remettre.

Cette section se concentre sur comment n'avoir aucun point unique de défaillance à l'aide des fonctionnalités disponibles sur AWS. Plus précisément, cette section décrit un sous-ensemble des principales fonctionnalités et modèles de conception d'AWS qui vous permettent de créer des applications de communication en temps réel hautement disponibles sur la plateforme.

Rubriques

- Mise à l'échelle dynamique avec AWS Lambda, Amazon Route 53 et AWS Auto Scaling
- WebRTC haute disponibilité avec Kinesis Video Streams
- Jonction SIP à haute disponibilité avec Amazon Chime Voice Connector

Modèle IP flottant pour la haute disponibilité entre des serveurs avec état actif-secours

Le modèle de conception IP flottante est un mécanisme bien connu pour réaliser un basculement automatique entre une paire de nœuds matériels actifs et de secours (serveurs multimédias). Une adresse IP virtuelle secondaire statique est attribuée au nœud actif. La surveillance continue entre les nœuds actifs et de secours détecte les échecs. Si le nœud actif échoue, 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.

Rubriques

- Applicabilité dans les solutions RTC
- Implémentation sur AWS
- Avantages
- Limitations et extensibilité

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 actif-secours fournit le meilleur mécanisme pour la haute disponibilité. Par exemple, les composants avec état d'une solution RTC, tels que le serveur multimédia ou le serveur de conférence, ou même un SBC ou un serveur de base de données, conviennent parfaitement à une configuration actif-secours. Un SBC ou un serveur multimédia possède plusieurs sessions ou canaux actifs de longue durée à un moment donné, et en cas d'échec 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.

Implémentation sur AWS

Vous pouvez mettre en œuvre ce modèle sur AWS à l'aide des fonctionnalités principales d'Amazon Elastic Compute Cloud (Amazon EC2), de l'API Amazon EC2, des adresses IP élastiques et de la prise en charge sur Amazon EC2 pour les adresses IP privées secondaires.

1. Lancez deux instances EC2 pour assumer les rôles de nœuds primaires et secondaires, le nœud primaire étant supposé être actif par défaut.
2. Attribuez une adresse IP privée secondaire supplémentaire à l'instance EC2 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 est requise pour que l'adresse IP secondaire soit ajoutée en tant qu'alias à l'interface réseau principale.
5. L'application doit se lier à cette adresse IP élastique. Dans le cas du logiciel Asterisk, vous pouvez configurer la liaison via les paramètres SIP avancés d'Asterisk.
6. Exécutez un script de surveillance (personnalisé, KeepAlive sur Linux, Corosync, etc.) sur chaque nœud pour contrôler l'état du nœud homologue. Dans le cas où le nœud actif actuel échoue, l'homologue détecte cet échec et appelle l'API Amazon EC2 pour se réattribuer l'adresse IP privée secondaire.
7. Par conséquent, l'application qui écoutait l'adresse IP virtuelle associée à l'adresse IP privée secondaire devient disponible pour les points de terminaison via le nœud de secours.

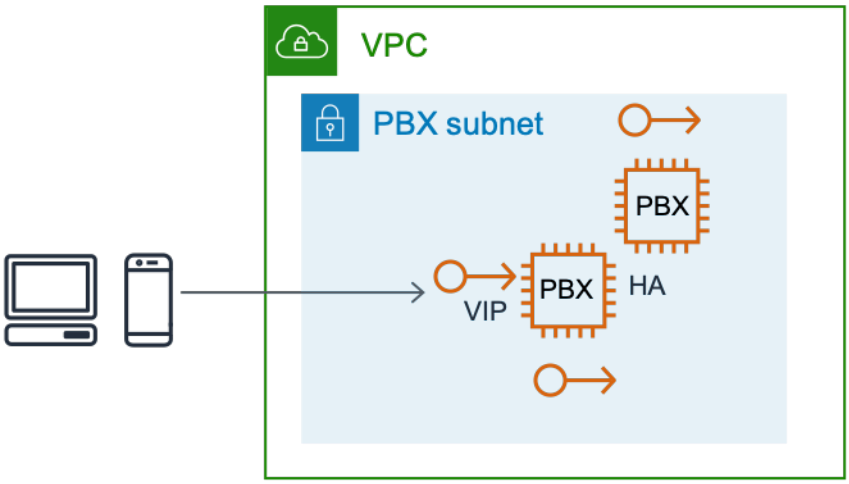

Figure 4 : Basculement entre des instances EC2 avec état à l'aide d'une adresse IP élastique

Avantages

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

Limitations 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 une variante. 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 de réassociation d'adresse IP élastique disponible. Dans l'implémentation du basculement illustrée à la figure 4, les appels en cours sont abandonnés et les points de terminaison doivent se reconnecter. Il est possible d'étendre cette implémentation à la réplication des données de sessions sous-jacentes afin d'assurer un basculement transparent des sessions ou la continuité des médias.

Répartition de charge pour la capacité de mise à l'échelle et la haute disponibilité avec WebRTC et SIP

La répartition de charge d'un cluster d'instances actives basé sur des règles prédéfinies, telles que le tourniquet, l'affinité ou la latence, etc., est un modèle de conception largement popularisé par la nature sans état des requêtes HTTP. La répartition de charge est une option viable dans le cas de nombreux composants d'application RTC.

L'équilibreur de charge agit en tant que proxy inverse ou 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 effectuent une surveillance de l'état sur les nœuds de leur cluster cible et n'envoient pas de demande entrante à un nœud qui échoue à la surveillance de l'état. Par conséquent, la répartition de charge permet d'atteindre un degré fondamental de haute disponibilité. En outre, puisqu'un équilibreur de charge effectue des surveillances de l'état actives et passives sur tous les nœuds de cluster à des intervalles inférieurs à la seconde, le temps de basculement est presque instantané.

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

- Tourniquet

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

Rubriques

Applicabilité dans les architectures RTC

Le protocole WebRTC permet aux passerelles WebRTC d'être facilement équilibrées par l'intermédiaire d'un équilibreur de charge basé sur HTTP, tel qu'Elastic Load Balancing, Application Load Balancer ou Network Load Balancer. La plupart des implémentations SIP reposant sur le transport sur TCP et UDP, une répartition de charge au niveau du réseau ou de la connexion avec prise en charge du trafic basé sur TCP et UDP est nécessaire.

Répartition de charge sur AWS pour WebRTC à l'aide d'Application Load Balancer et d'Auto Scaling

Dans le cas des communications basées sur 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 d'instances EC2 associées à Elastic Load Balancing. En outre, puisque les demandes WebRTC sont sans état, vous pouvez utiliser Amazon EC2 Auto Scaling pour fournir une capacité de mise à l'échelle, une élasticité et une haute disponibilité entièrement automatisées et contrôlables.

L'Application Load Balancer fournit un service de répartition de charge entièrement géré, évolutif et hautement disponible à l'aide de plusieurs zones de disponibilité. Cela prend en charge la répartition de charge des requêtes WebSocket 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 permanentes, en acheminant les demandes du même client vers la même cible à l'aide de

cookies générés par l'équilibreur de charge. Si vous autorisez les sessions permanentes, la même cible reçoit la requête et peut utiliser le cookie pour retrouver le contexte de la session.

La figure 5 montre la topologie cible.

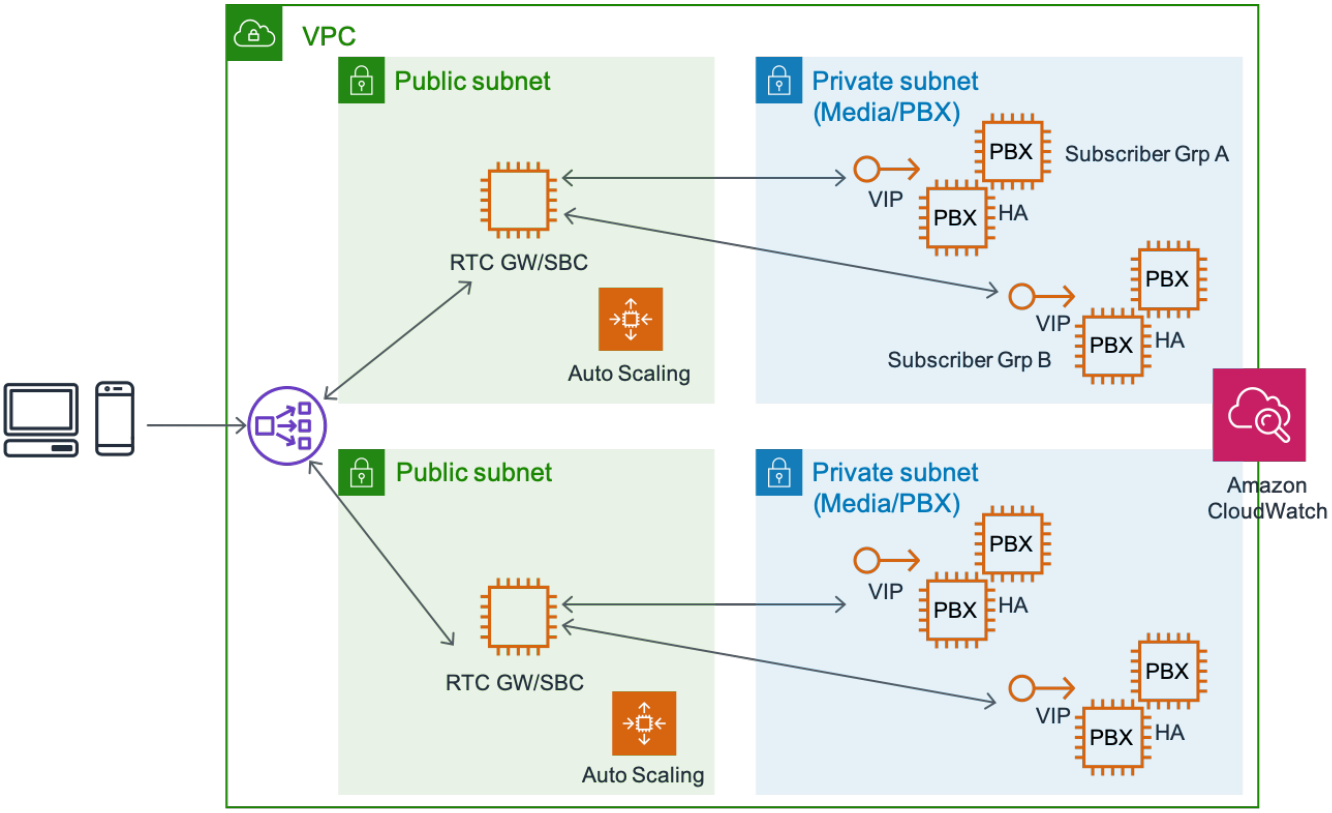

Figure 5 : Capacité de mise à l'échelle WebRTC et architecture haute disponibilité

Implémentation pour SIP avec Network Load Balancer ou le produit AWS Marketplace

Dans le cas des communications basées sur SIP, les connexions se font via TCP ou UDP, la majorité des applications RTC utilisant UDP. Si SIP/TCP est le protocole de signal de choix, il est possible d'utiliser le Network Load Balancer pour une répartition de charge entièrement gérée, hautement disponible, évolutive et performante.

Un Network Load Balancer fonctionne au niveau de la connexion (couche 4), acheminant les connexions vers des cibles telles que les instances Amazon EC2, les conteneurs et les adresses IP en fonction des données du protocole IP. L'équilibrage de charge du réseau est parfaitement adapté pour la répartition de charge du trafic TCP ou UDP, et peut traiter des millions de requêtes par seconde, tout en maintenant des temps de latence extrêmement faibles. Il est intégré à d'autres

services AWS populaires, tels qu'AWS 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 logiciel commercial prêt à l'emploi (COTS) AWS Marketplace. AWS Marketplace propose de nombreux produits capables de gérer l'UDP et d'autres types de répartition de charge de connexion de couche 4. Ces COTS incluent généralement la prise en charge de la haute disponibilité et sont habituellement intégrés à des fonctions, telles qu'AWS Auto Scaling, pour améliorer davantage la disponibilité et la capacité de mise à l'échelle. La figure 6 montre la topologie cible :

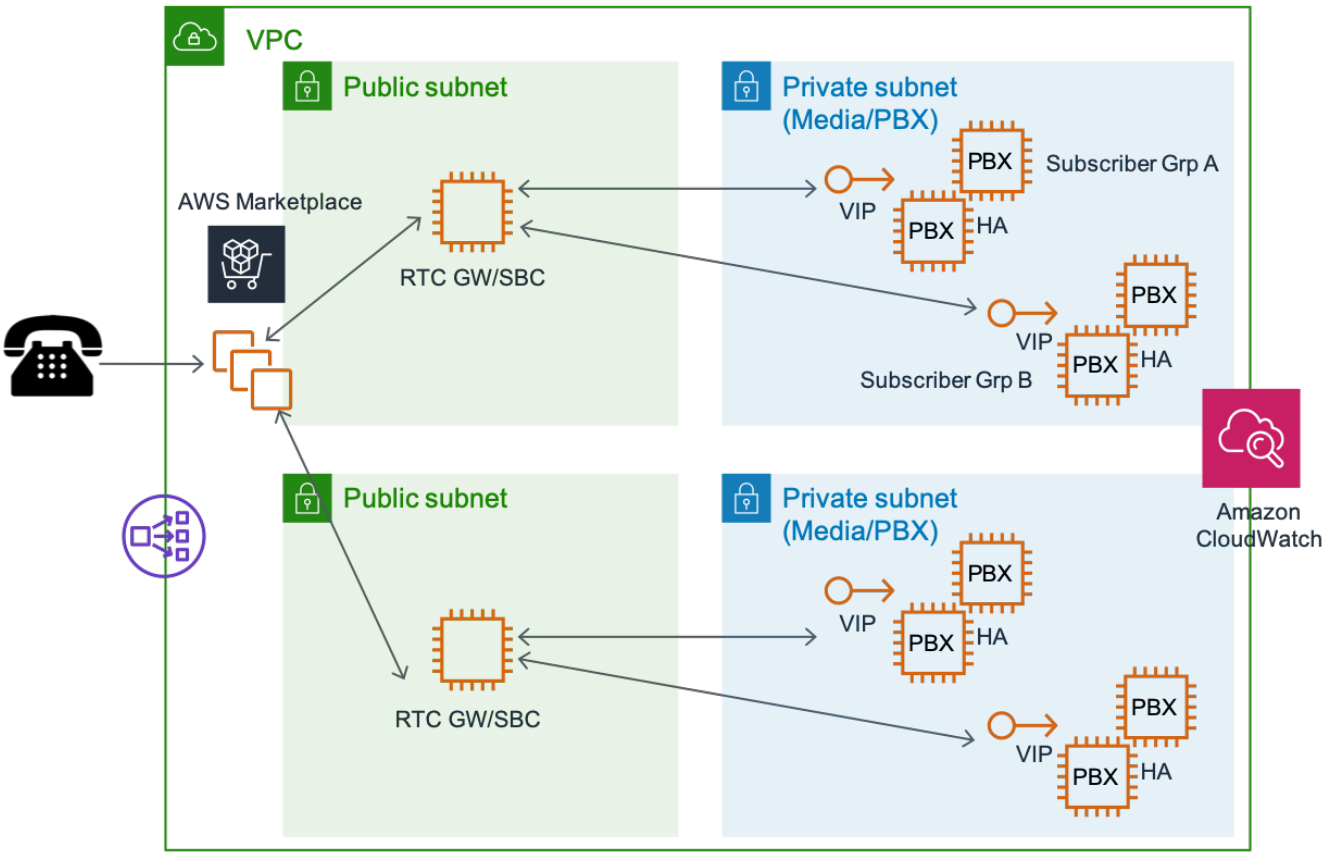

Figure 6 : Capacité de mise à l'échelle RTC basée sur SIP avec le produit AWS Marketplace

Répartition de charge basée sur DNS entre régions et basculement

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'enregistrer et de se connecter à des applications multimédias. Avec Amazon Route 53, les surveillances de l'état DNS peuvent être configurées pour acheminer le trafic vers des points de terminaison sains ou pour surveiller indépendamment l'état de votre application. La fonction Amazon Route 53 Traffic Flow vous permet de gérer facilement

et globalement le trafic via divers types de routage, notamment le routage basé sur la latence, le Geo DNS, la géoproximité et le WRR (technique du tourniquet pondéré), c'est-à-dire ceux pouvant être combinés au basculement DNS pour vous permettre de bâtir différentes architectures à faible latence, tolérantes aux pannes. L'éditeur visuel simple Amazon Route 53 Traffic Flow vous permet de facilement gérer la façon dont les utilisateurs finaux sont acheminés vers les points de terminaison de votre application, dans une région AWS unique ou distribuée aux quatre coins du monde.

Dans le cas de déploiements globaux, 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 clients doivent être vidés. En outre, les modifications DNS peuvent avoir un certain retard car elles sont propagées sur les serveurs DNS mondiaux. Vous pouvez gérer l'intervalle d'actualisation des recherches DNS avec l'attribut Durée de vie. Cet attribut est configurable au moment de la configuration des politiques DNS.

Pour atteindre rapidement des utilisateurs internationaux ou pour répondre aux exigences liées à l'utilisation d'une adresse IP publique unique, AWS Global Accelerator peut également être utilisé pour le basculement entre régions. AWS Global Accelerator est un service de réseaux qui améliore la disponibilité et les performances des applications ayant une portée locale et mondiale. AWS Global Accelerator fournit des adresses IP statiques qui agissent en tant que point d'entrée fixe vers vos points de terminaison d'application, tels que vos Application Load Balancers, vos Network Load Balancers ou vos instances Amazon EC2 dans une ou plusieurs régions AWS. Il utilise le réseau mondial AWS pour optimiser le chemin entre vos utilisateurs et vos applications, améliorant ainsi les performances, telles que la latence de votre trafic TCP et UDP. AWS Global Accelerator contrôle en permanence l'état de vos points de terminaison d'application et redirige automatiquement le trafic vers les points de terminaison sains les plus proches en cas de défaillance des points de terminaison actuels. Pour des exigences de sécurité supplémentaires, le Site-to-Site VPN accéléré utilise AWS Global Accelerator pour améliorer les performances des connexions VPN en acheminant intelligemment le trafic via le réseau mondial AWS et les emplacements périphériques AWS.

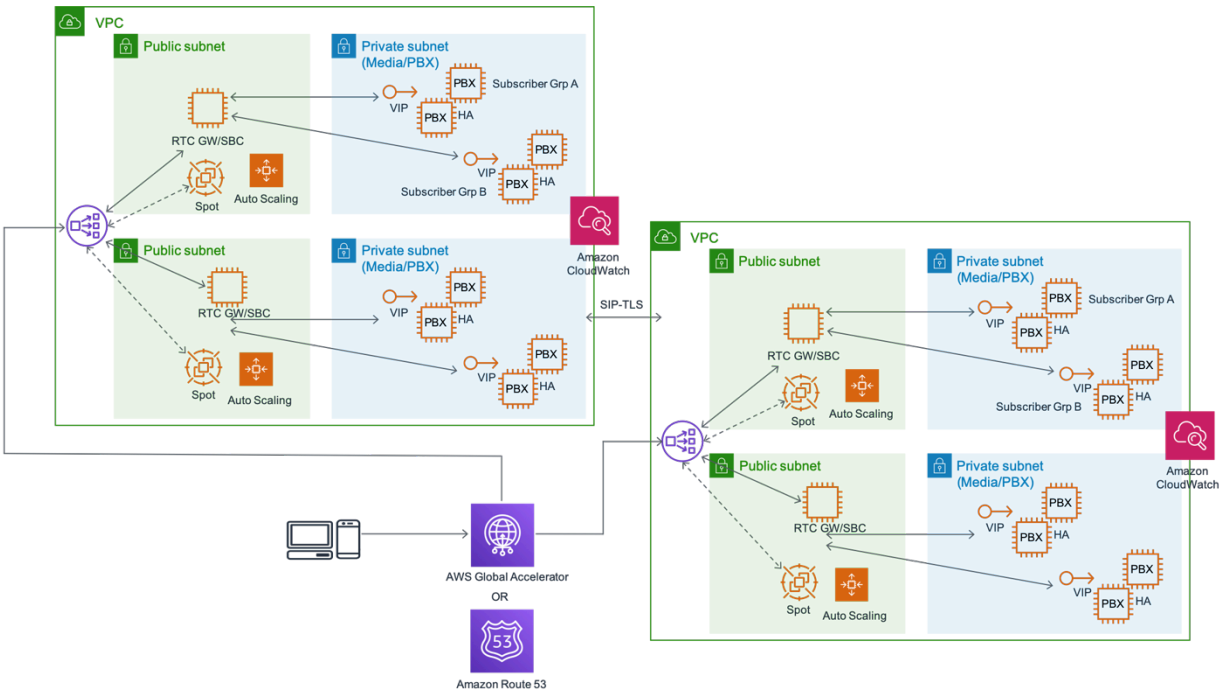

Figure 7 : Conception à haute disponibilité interrégion avec AWS Global Accelerator ou Amazon Route 53

Durabilité des données et haute disponibilité avec stockage permanent

La plupart des applications RTC s'appuient sur le stockage permanent pour stocker et accéder aux données à des fins d'authentification, d'autorisation, de comptabilisation (données de session, registres détaillés des appels, etc.), de surveillance opérationnelle et de journalisation. Dans un centre de données traditionnel, la garantie d'une haute disponibilité et d'une durabilité pour les composants de stockage permanent (bases de données, systèmes de fichiers, etc.) nécessite généralement de lourdes tâches via la configuration d'un SAN, la conception RAID et les processus de sauvegarde, de restauration et de traitement du basculement. Le cloud AWS simplifie et améliore 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, les services AWS tels qu'Amazon Simple Storage Service (Amazon S3) et Amazon Elastic File System (Amazon EFS) fournissent une haute disponibilité et une capacité de mise à l'échelle gérées. Amazon S3 a une durabilité des données de 99,999999999 %.

Pour le stockage des données transactionnelles, les clients ont la possibilité de tirer parti d'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 de registre, le profil d'abonné ou le stockage des registres comptables (p. ex. les CDR), Amazon RDS fournit une option tolérante aux pannes, hautement disponible et évolutive.

Mise à l'échelle dynamique avec AWS Lambda, Amazon Route 53 et AWS Auto Scaling

AWS permet de chaîner des fonctions et d'intégrer des capacités sans serveur personnalisées en tant que service basé sur les événements d'infrastructure. L'un de ces modèles de conception qui présente de nombreuses utilisations polyvalentes dans les applications RTC est la combinaison de hooks de cycle de vie de scalabilité automatique avec Amazon CloudWatch Events, Amazon Route 53 et des fonctions AWS Lambda. AWS Lambda peut intégrer n'importe quelle action ou logique. La figure 8 montre comment ces fonctions chaînées peuvent améliorer la fiabilité et la capacité de mise à l'échelle du système grâce à l'automatisation.

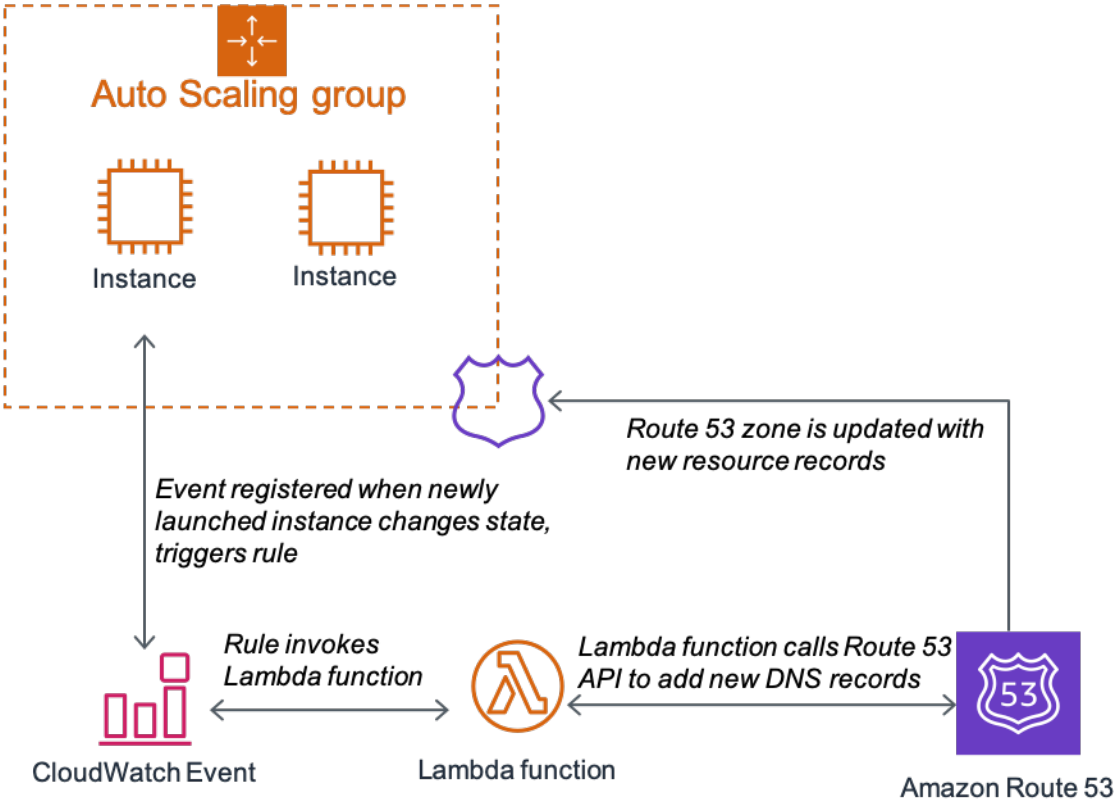

Figure 8 : Scalabilité automatique avec mises à jour dynamiques d'Amazon Route 53

WebRTC haute disponibilité avec Kinesis Video Streams

Amazon Kinesis Video Streams propose une diffusion multimédia en temps réel via WebRTC, ce qui permet aux utilisateurs de capturer, de traiter et de stocker des flux multimédias à des fins de lecture, d'analytique et de machine learning. Ces flux sont hautement disponibles, évolutifs et conformes aux normes WebRTC. Amazon Kinesis Video Streams inclut un point de terminaison de signalisation WebRTC pour permettre une découverte rapide des pairs et établir une connexion sécurisée. Cela inclut des points de terminaison gérés appelés Session Traversal Utilities for NAT (STUN) et Traversal Using Relays around NAT (TURN) dédiés aux échanges multimédias en temps réel entre pairs. Cela inclut également un kit SDK open source gratuit qui s'intègre directement au micrologiciel de la caméra afin de permettre une communication sécurisée avec les points de terminaison de Kinesis Video Streams et de découvrir les pairs et le streaming multimédia. Enfin, cela fournit des bibliothèques client pour Android, iOS et JavaScript qui autorisent les lecteurs mobiles et web conformes à WebRTC de découvrir en toute sécurité et de se connecter à une caméra pour un streaming multimédia et une communication bidirectionnelle.

Jonction SIP à haute disponibilité avec Amazon Chime Voice Connector

Amazon Chime Voice Connector offre un service de jonctions SIP avec paiement à l'utilisation qui permet aux entreprises de passer et/ou recevoir des appels téléphoniques sécurisés et peu coûteux avec leurs systèmes téléphoniques. Amazon Chime Voice Connector est une alternative peu coûteuse aux jonctions SIP des fournisseurs de services ou à l'interface à débit primaire du réseau numérique à intégration de services (RNIS). Il permet aux clients d'activer les appels entrants, les appels sortants ou les deux. Ce service tire parti du réseau AWS pour offrir une expérience d'appel hautement disponible dans plusieurs régions AWS. Vous pouvez diffuser de l'audio à partir d'appels téléphoniques à jonction SIP ou de flux de registre multimédia basé sur SIP (SIPREC) transférés vers Amazon Kinesis Video Streams pour obtenir des informations sur les appels professionnels en temps réel. Vous pouvez rapidement créer des applications pour l'analytique audio grâce à l'intégration à Amazon Transcribe et à d'autres bibliothèques courantes de machine learning.

Bonnes pratiques sur le terrain

Cette section vise à résumer les bonnes pratiques mises en œuvre par certains des clients AWS les plus importants et les plus performants qui exécutent de grandes charges de travail SIP (Session Initiation Protocol) en temps réel. Les clients AWS qui souhaitent exécuter leur propre infrastructure SIP dans le cloud public trouveront ces bonnes pratiques utiles car elles peuvent contribuer à augmenter la fiabilité et la résilience du système en cas de différents types d'échecs. Bien que certaines de ces bonnes pratiques soient spécifiques au protocole SIP, la plupart d'entre elles s'appliquent à toute application de communication en temps réel exécutée sur AWS.

Rubriques

Créer une superposition SIP

AWS dispose d'un backbone réseau robuste, évolutif et redondant qui assure une connectivité entre différentes régions. Lorsqu'un événement réseau, tel qu'une coupure de fibre, dégrade une liaison de backbone AWS, le trafic est rapidement basculé vers des chemins redondants à l'aide de protocoles de routage au niveau du réseau, tels que BGP. Cette ingénierie du trafic au niveau du réseau est une boîte noire pour les clients AWS et la plupart ne remarquent même pas ces événements de basculement. Cependant, les clients qui exécutent des charges de travail en temps réel, des échanges vocaux, des vidéos de haute qualité et une messagerie à faible latence, remarquent parfois ces événements. Alors, comment un client AWS peut-il mettre en œuvre sa propre ingénierie du trafic en plus de ce qui est fourni par AWS au niveau du réseau ? La solution consiste à déployer une infrastructure SIP dans de nombreuses régions AWS différentes. Dans le cadre des fonctions de contrôle des appels, SIP permet également d'acheminer les appels via des proxys SIP spécifiques.

Figure 9 : Utilisation du routage SIP pour remplacer le routage réseau

Dans la figure 9, l'infrastructure SIP (représentée par des points verts) fonctionne dans les quatre régions américaines. Les lignes bleues sont une représentation fictive du backbone AWS. Si aucun routage SIP n'est mis en œuvre, un appel provenant de la côte ouest des États-Unis et destiné à la côte est des États-Unis passe par la liaison de backbone qui relie directement les régions de l'Oregon et de la Virginie. Le diagramme montre comment un client peut remplacer le routage au niveau du réseau et passer le même appel entre l'Oregon et la Virginie acheminé via 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 multimédias en fonction de métriques réseau telles que les retransmissions SIP et les préférences commerciales spécifiques du client.

Effectuer une surveillance détaillée

Les utilisateurs finaux d'applications vocales et vidéo en temps réel s'attendent au 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.

Figure 10 : Utilisation de SIPp pour contrôler l'infrastructure VoIP

De nombreux outils open source, tels que iPerf ou SIPp , et VoIPMonitor, sont disponibles et peuvent être utilisés pour contrôler le trafic SIP/RTP. Dans l'exemple précédent, les nœuds exécutant SIPp en modes client et serveur mesurent des métriques SIP telles que les appels réussis et les retransmissions SIP entre les quatre régions AWS américaines. Ces métriques peuvent ensuite être exportées vers Amazon CloudWatch à l'aide d'un script personnalisé. À l'aide de CloudWatch, les clients peuvent créer des alarmes sur ces métriques personnalisées en fonction d'une certaine valeur seuil. Des mesures correctives automatiques ou manuelles peuvent ensuite être prises en fonction de l'état de ces alarmes CloudWatch.

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 solutions de surveillance VoIP de qualité sont disponibles sur le marché, telles que ThousandEyes. Un exemple d'action corrective consiste à modifier le routage SIP en fonction de l'augmentation des retransmissions SIP.

Utiliser DNS pour la répartition de charge et les adresses IP flottantes pour le basculement

Les clients de téléphonie IP qui prennent en charge la capacité SRV DNS peuvent utiliser efficacement la redondance intégrée à l'infrastructure en équilibrant la charge des clients vers différents SBC ou PBX.

Figure 11 : Utilisation des registres SRV DNS pour équilibrer la charge des clients SIP

La figure 11 montre comment les clients peuvent utiliser les registres SRV pour équilibrer la charge du trafic SIP. Tout client de téléphonie IP prenant en charge la norme SRV recherchera le préfixe sip._<transport protocol> dans un registre DNS de type SRV. Dans l'exemple, la section de réponse du DNS contient les deux PBX exécutés dans différentes zones de disponibilité AWS. Cependant, en plus des URI de point de terminaison, le registre SRV contient trois informations supplémentaires :

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

Puisque la priorité est la même (1) pour les deux serveurs PBX, les clients utilisent le poids pour équilibrer la charge entre les deux PBX. Dans ce cas, puisque les poids sont identiques, la charge du trafic SIP doit être équilibrée de manière égale entre les deux PBX.

DNS peut être une bonne solution pour la répartition de charge du client, mais qu'en est-il de la mise en œuvre du basculement en changeant ou en mettant à jour les registres DNS « A » ? Cette méthode est déconseillée en raison d'une incohérence constatée dans le comportement de mise en cache de 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 réaffectation IP EC2 où l'adresse IP d'un hôte altéré est instantanément réattribuée à un hôte sain à l'aide de l'API EC2. Associée à une solution détaillée de contrôle et de surveillance de l'état, la réattribution IP d'un nœud défaillant garantit que le trafic est transféré vers un hôte sain en temps opportun, ce qui minimise les perturbations pour l'utilisateur final.

Utiliser plusieurs zones de disponibilité

Chaque région AWS est subdivisée en zones de disponibilité distinctes. Chaque zone de disponibilité possède sa propre alimentation, son propre refroidissement et sa propre connectivité réseau et forme ainsi un domaine d'échec isolé. Dans les constructions d'AWS, il est toujours recommandé que les clients exécutent leurs charges de travail dans plusieurs zones de disponibilité. Cela garantit que les applications des clients peuvent résister à un échec complet de la zone de disponibilité, un événement très rare en soi. Cette recommandation concerne également l'infrastructure SIP en temps réel.

Figure 12 : Gestion des échecs de la zone de disponibilité

Supposons qu'un événement catastrophique (tel qu'un ouragan de catégorie 5) provoque une panne complète de la zone de disponibilité dans la région us-east-1. L'infrastructure étant exécutée comme indiqué dans le diagramme, 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 n°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 n°2.

En résumé, les registres SRV DNS doivent pointer le client vers plusieurs registres « A », un dans chaque zone de disponibilité. Chacun de ces registres « A » doit, à son tour, pointer vers plusieurs adresses IP de SBC ou PBX dans cette zone de disponibilité, offrant ainsi une résilience intra et inter-AZ. Le basculement intra-AZ et inter-AZ peut être mis en œuvre en utilisant la réattribution IP si les adresses IP sont publiques. Les adresses IP privées ne peuvent toutefois pas être réaffectées entre les zones de disponibilité. Si un client utilise un adressage IP privé, il doit alors compter sur le fait que les clients SIP se réenregistrent auprès du SBC/PBX de sauvegarde pour le basculement inter-AZ.

Conserver le trafic dans une zone de disponibilité et utiliser des groupes de placement EC2

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

Figure 13 : Affinité de la zone de disponibilité (50 % des appels actifs maximums sont perdus)

La figure 13 illustre une architecture simplifiée qui utilise l'affinité de la zone de disponibilité. L'avantage comparatif de cette approche apparaît clairement si l'on tient compte des effets d'une panne complète de la zone de disponibilité. Comme le montre le diagramme, si la zone de disponibilité n°2 est perdue, 50 % des appels actifs sont affectés au maximum (en supposant une répartition de charge égale entre les zones de disponibilité). Si l'affinité de la zone de disponibilité n'avait pas été mise en œuvre, certains appels circuleraient entre les zones de disponibilité d'une région et un échec affecterait probablement plus de 50 % des appels actifs.

En outre, pour réduire la latence du trafic, nous vous recommandons également d'envisager d'utiliser des groupes de placement EC2 au sein de chaque zone de disponibilité. Les instances lancées au sein du même groupe de placement EC2 ont une bande passante plus élevée et une latence réduite, car EC2 garantit la proximité réseau de ces instances les unes par rapport aux autres.

Utiliser les types d'instances EC2 de réseaux améliorés

Le choix du bon type d'instance sur Amazon EC2 garantit la fiabilité du système ainsi qu'une utilisation efficace de l'infrastructure. EC2 fournit un vaste éventail de types d'instances optimisés pour différents cas d'utilisation. Ces types d'instances correspondent à différentes combinaisons en termes de capacités de CPU, de mémoire, de stockage et de réseaux. Vous pouvez ainsi choisir un ensemble de ressources parfaitement adapté à vos applications. Ces types d'instances de réseaux améliorés garantissent que les charges de travail SIP qui s'exécutent sur eux ont accès à une bande passante constante et à une latence agrégée comparativement plus faible. Un ajout récent à Amazon EC2 est la disponibilité de l'Elastic Network Adapter (ENA) qui fournit jusqu'à 100 Gbit/s de bande passante. Le dernier catalogue des types d'instances EC2 et des fonctions associées se trouve sur la page des types d'instance EC2.

Pour la plupart des clients, la dernière génération d'instances optimisées pour le calcul doit fournir le meilleur rapport qualité-prix. Par exemple, le C5N prend en charge le nouvel Elastic Network Adapter avec une bande passante maximale de 100 Gbit/s avec des millions de paquets par seconde (PPS). La plupart des applications en temps réel bénéficieraient également de l'utilisation du kit de développement de plan de données (DPDK) Intel qui peut considérablement améliorer le traitement des paquets réseau.

Toutefois, il est toujours recommandé de comparer les différents types d'instances EC2 en fonction de vos besoins afin de déterminer quel type d'instance 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 de sécurité

Les composants de l'application RTC s'exécutent généralement directement sur des instances Amazon EC2 orientées Internet. Outre le protocole TCP, les flux utilisent des protocoles tels que UDP et SIP. Dans ces cas, AWS Shield Standard protège les instances Amazon EC2 contre les attaques DDoS de la couche d'infrastructure commune (couches 3 et 4), 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 DDoS bien définie est détectée.

AWS fournit également une protection avancée contre les attaques DDoS importantes et sophistiquées pour ces applications en activant AWS Shield Advanced sur les adresses IP élastiques. AWS Shield Advanced fournit une détection DDoS améliorée qui détecte automatiquement le type de ressource AWS et la taille de l'instance EC2 et applique des mesures d'atténuation prédéfinies appropriées avec des protections contre les flux SYN ou UDP. Avec AWS Shield Advanced, les clients peuvent également créer leurs propres profils d'atténuation personnalisés en faisant appel à l'équipe d'intervention DDoS (DRT) d'AWS 24 h/24 et 7 j/7. AWS Shield Advanced garantit également que lors d'une attaque DDoS, toutes vos listes de contrôle d'accès au réseau (ACL) Amazon VPC sont automatiquement appliquées à la frontière du réseau AWS, ce qui vous donne accès à une bande passante supplémentaire et à une capacité de nettoyage pour atténuer les attaques DDoS volumétriques importantes.

Conclusion

Les charges de travail de communication en temps réel (RTC) peuvent être déployées sur Amazon Web Services (AWS) pour atteindre capacité de mise à l'échelle, élasticité et haute disponibilité tout en répondant aux exigences clé. Aujourd'hui, plusieurs clients utilisent AWS, ses partenaires et des solutions open source pour exécuter des charges de travail RTC avec un coût réduit, une agilité plus rapide et une empreinte mondiale réduite.

Les architectures de référence et les bonnes pratiques fournies dans ce livre blanc peuvent aider les clients à configurer avec succès des charges de travail RTC sur AWS et à optimiser les solutions concernant les besoins des utilisateurs finaux et le cloud.

Participants

Les personnes et organisations suivantes ont participé à la préparation du présent document :

- Ahmad Khan, architecte de solutions principal, Amazon Web Services
- Tipu Qureshi, ingénieur principal AWS Support, Amazon Web Services
- Hasan Khan, gestionnaire de compte technique principal, Amazon Web Services
- Shoma Chakravarty, responsable technique mondial, télécommunications, Amazon Web Services

Révisions du document

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

update-history-change	update-history-description	update-history-date
Livre blanc mis à jour	Mis à jour avec les derniers services et fonctions.	13 février 2020
Publication initiale	Première publication du livre blanc.	1er octobre 2018

Mentions légales

Les clients sont responsables de leur propre évaluation indépendante des informations contenues dans ce document. Le présent document : (a) est fourni à titre informatif uniquement, (b) représente les offres et pratiques actuelles de produits AWS, qui sont susceptibles d'être modifiées sans préavis, et (c) ne crée aucun engagement ou assurance de la part d'AWS et de ses affiliés, fournisseurs ou concédants de licences. Les produits ou services AWS sont fournis « en l'état » sans garantie, représentation ou condition, de quelque nature que ce soit, explicite ou implicite. Les responsabilités et obligations d'AWS envers ses clients sont déterminées par les contrats AWS, et le présent document ne fait pas partie d'un contrat entre AWS et ses clients, ni le modifie.