Whitepaper da AWS

Comunicação em tempo real na AWS

Comunicação em tempo real na AWS: Whitepaper da AWS

Table of Contents

Comunicação em tempo real na AWS

Práticas recomendadas para criar workloads de comunicação em tempo real (RTC) de alta disponibilidade e escalabilidade na AWS

Data de publicação: 13 de fevereiro de 2020 (Revisões do documento)

Resumo

Atualmente, muitas organizações buscam reduzir custos e obter escalabilidade para workloads de voz, mensagens e multimídia em tempo real. Este documento descreve as práticas recomendadas para gerenciar workloads de comunicação em tempo real na AWS e inclui arquiteturas de referência para atender a esses requisitos. Este documento serve como um guia para indivíduos familiarizados com a comunicação em tempo real. Ele mostra como obter alta disponibilidade e escalabilidade para essas workloads.

Introdução

Aplicações de telecomunicações que usam voz, vídeo e mensagens como canais são um requisito fundamental para muitas organizações e seus usuários finais. Essas workloads de comunicação em tempo real (RTC) têm requisitos específicos de latência e disponibilidade que podem ser atendidos seguindo as práticas recomendadas de design relevantes. No passado, as workloads de RTC eram implantadas em datacenters on-premises tradicionais com recursos dedicados.

No entanto, devido a um conjunto maduro e crescente de recursos, as workloads de RTC podem ser implantadas na Amazon Web Services (AWS), apesar dos rigorosos requisitos de nível de serviço, beneficiando-se da escalabilidade, elasticidade e alta disponibilidade. Hoje, vários clientes estão usando a AWS, seus parceiros e soluções de código aberto para executar workloads de RTC com custo reduzido, mais agilidade, capacidade de se tornar global em poucos minutos e recursos avançados dos serviços da AWS.

Os clientes utilizam recursos da AWS, como redes avançadas com um Elastic Network Adapter (ENA) e a última geração de instâncias do Amazon Elastic Compute Cloud (EC2) para se beneficiar de um kit de desenvolvimento de plano de dados (DPDK), virtualização de E/S de raiz única (SR-IOV), páginas enormes, NVM Express (NVMe), suporte a acesso não uniforme à memória (NUMA) e instâncias bare metal para atender aos requisitos de workloads de RTC. Essas instâncias oferecem largura de banda de rede de até 100 Gbps e pacotes proporcionais por segundo, melhorando a performance de aplicações com uso intenso de rede. Para escalabilidade, o Elastic Load Balancing oferece o Application Load Balancer, que oferece suporte ao WebSocket e o Network Load Balancer, capaz de gerenciar milhões de solicitações por segundo. Para aceleração de rede, o AWS Global Accelerator fornece endereços IP estáticos que atuam como um ponto de entrada fixo para endpoints de aplicações na AWS. Ele tem suporte para endereços IP estáticos para o balanceador de carga. Para reduzir a latência, o custo e aumentar a taxa de transferência de largura de banda, o AWS Direct Connect estabelece uma conexão de rede dedicada do ambiente on-premises para a AWS. O entroncamento SIP gerenciado altamente disponível é fornecido pelo Amazon Chime Voice Connector. O Amazon Kinesis Video Streams com WebRTC transmite facilmente mídias bidirecionais em tempo real com alta disponibilidade.

Este documento inclui arquiteturas de referência que mostram como configurar workloads de RTC na AWS e práticas recomendadas para otimizar as soluções para atender aos requisitos do usuário final, otimizando para a nuvem. O núcleo de pacote evoluído (EPC) está fora do escopo deste whitepaper, mas as práticas recomendadas detalhadas podem ser aplicadas a funções de rede virtual (VNFs).

Componentes fundamentais da arquitetura RTC

No setor de telecomunicações, a comunicação em tempo real (RTC) geralmente se refere a sessões de conteúdo ao vivo entre dois endpoints com latência mínima. Essas sessões podem ser relacionadas a:

- Uma sessão de voz entre duas partes (por exemplo, sistema telefônico, celular, VoIP)
- Mensagens instantâneas (por exemplo, bate-papo, IRC)
- Sessão de vídeo ao vivo (por exemplo, videoconferência, telepresença)

Cada uma das soluções anteriores tem alguns componentes em comum (por exemplo, componentes que fornecem autenticação, autorização e controle de acesso, transcodificação, buffer e retransmissão etc.) e alguns componentes exclusivos do tipo de mídia transmitida (por exemplo, serviço de transmissão, servidor de mensagens e filas e assim por diante). Esta seção se concentra na definição de um sistema de RTC baseado em voz e vídeo e todos os componentes relacionados ilustrados na Figura 1.

Figura 1: componentes arquitetônicos essenciais da RTC

Tópicos

Softswitch/PBX

Um softswitch ou PBX é o cérebro de um sistema de telefonia de voz. Ele fornece inteligência para estabelecer, manter e rotear uma chamada de voz dentro ou fora da empresa usando diversos componentes. Todos os assinantes da empresa devem se inscrever no softswitch para receber ou fazer uma chamada. Uma funcionalidade importante do softswitch é acompanhar cada assinante e como alcançá-los usando os outros componentes da rede de voz.

Controlador de borda de sessão (SBC)

Um controlador de borda de sessão (SBC) fica na borda de uma rede de voz e mantém o controle de todo o tráfego de entrada e saída (planos de controle e dados). Uma das principais responsabilidades de um SBC é proteger o sistema de voz contra o uso malicioso. O SBC pode ser usado para interconexão com troncos SIP (Session Initiation Protocol) para conectividade externa. Alguns SBCs também fornecem recursos de transcodificação para converter CODECS de um formato em outro. Por fim, a maioria dos SBCs também fornece recursos de passagem de NAT, o que ajuda a garantir que as chamadas sejam estabelecidas, mesmo em redes com firewall.

Conectividade PSTN

As soluções de voz sobre IP (VoIP) usam gateways PSTN e troncos SIP para se conectar a redes PSTN herdadas.

Gateway PSTN

O gateway da rede telefônica pública comutada (PSTN) converte a sinalização (entre SIP e SS7) e a mídia (entre RTP e multiplexação por divisão de tempo [TDM] usando a transcodificação CODEC). Os gateways PSTN sempre ficam na borda próxima à rede PSTN.

Tronco SIP

Em um tronco SIP, a empresa não encerra suas chamadas em uma rede TDM (baseada em SS7), mas os fluxos entre a empresa e as telecomunicações permanecem por IP. A maioria dos troncos SIP é estabelecida com SBCs. A empresa deve concordar com as regras de segurança predefinidas da empresa de telecomunicações, como permitir um determinado intervalo de endereços IP, portas e assim por diante.

Gateway de mídia (transcodificador)

Uma solução de voz típica permite vários tipos de CODECs. Alguns dos CODECs comuns são G.711 μ-law para a América do Norte, G.711 A-law para fora da América do Norte, G.729 e G.722. Quando dois dispositivos que estão usando dois CODECs diferentes se comunicam entre si, um servidor de mídia traduz o fluxo de CODEC entre os dispositivos. Em outras palavras, um gateway de mídia processa a mídia e garante que os dispositivos finais possam se comunicar uns com os outros.

Gateway WebRTC e WebRTC

A comunicação em tempo real pela Web (WebRTC) permite estabelecer uma chamada de um navegador da Web ou solicitar recursos do servidor de back-end usando a API. A tecnologia foi criada tendo em mente a tecnologia de nuvem e, portanto, fornece várias APIs que podem ser usadas para estabelecer uma chamada. Como nem todas as soluções de voz (incluindo SIP) oferecem suporte a essas APIs, o gateway WebRTC precisa converter chamadas de API em mensagens SIP e vice-versa.

A Figura 2 mostra um padrão de design para uma arquitetura WebRTC altamente disponível. O tráfego de entrada de clientes WebRTC é balanceado por um Application Load Balancer da Amazon com WebRTC em execução em instâncias do EC2 que fazem parte de um grupo do Auto Scaling.

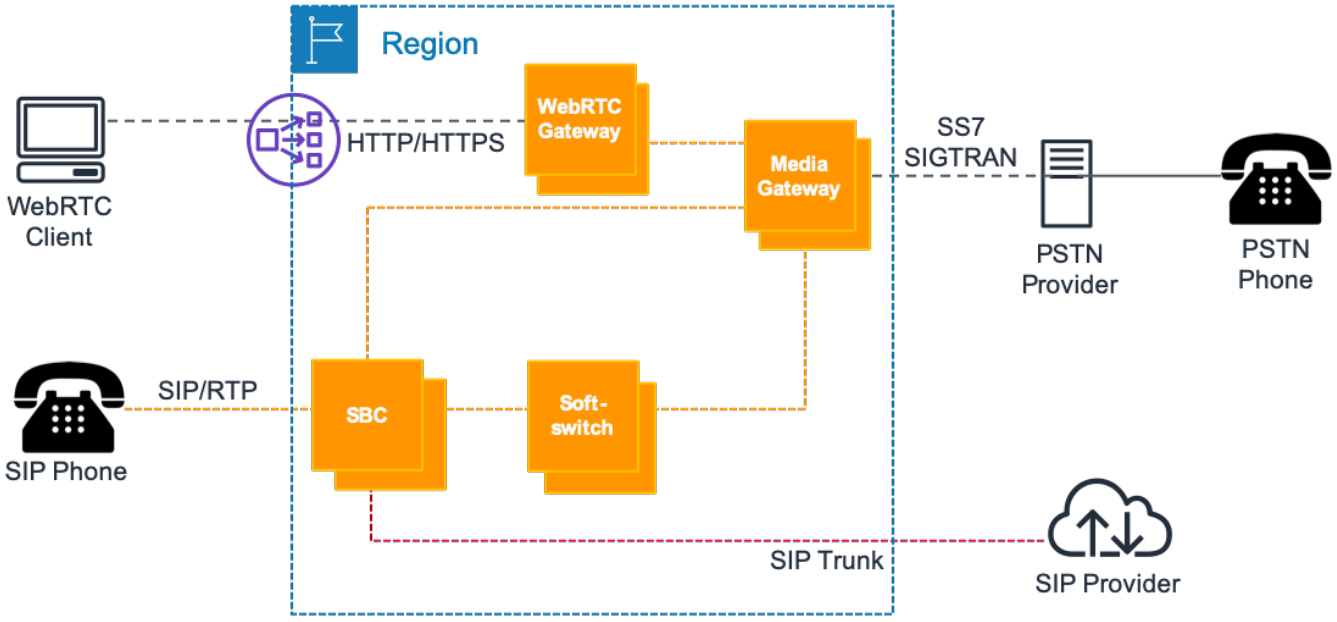

Figura 2: uma topologia básica de um sistema de RTC para voz

Outro padrão de design para o tráfego SIP e RTP é usar pares de SBCs no Amazon EC2 no modo passivo ativo nas zonas de disponibilidade (Figura 3). Aqui, um endereço IP elástico pode ser movido dinamicamente entre instâncias em caso de falha, onde o DNS não pode ser usado.

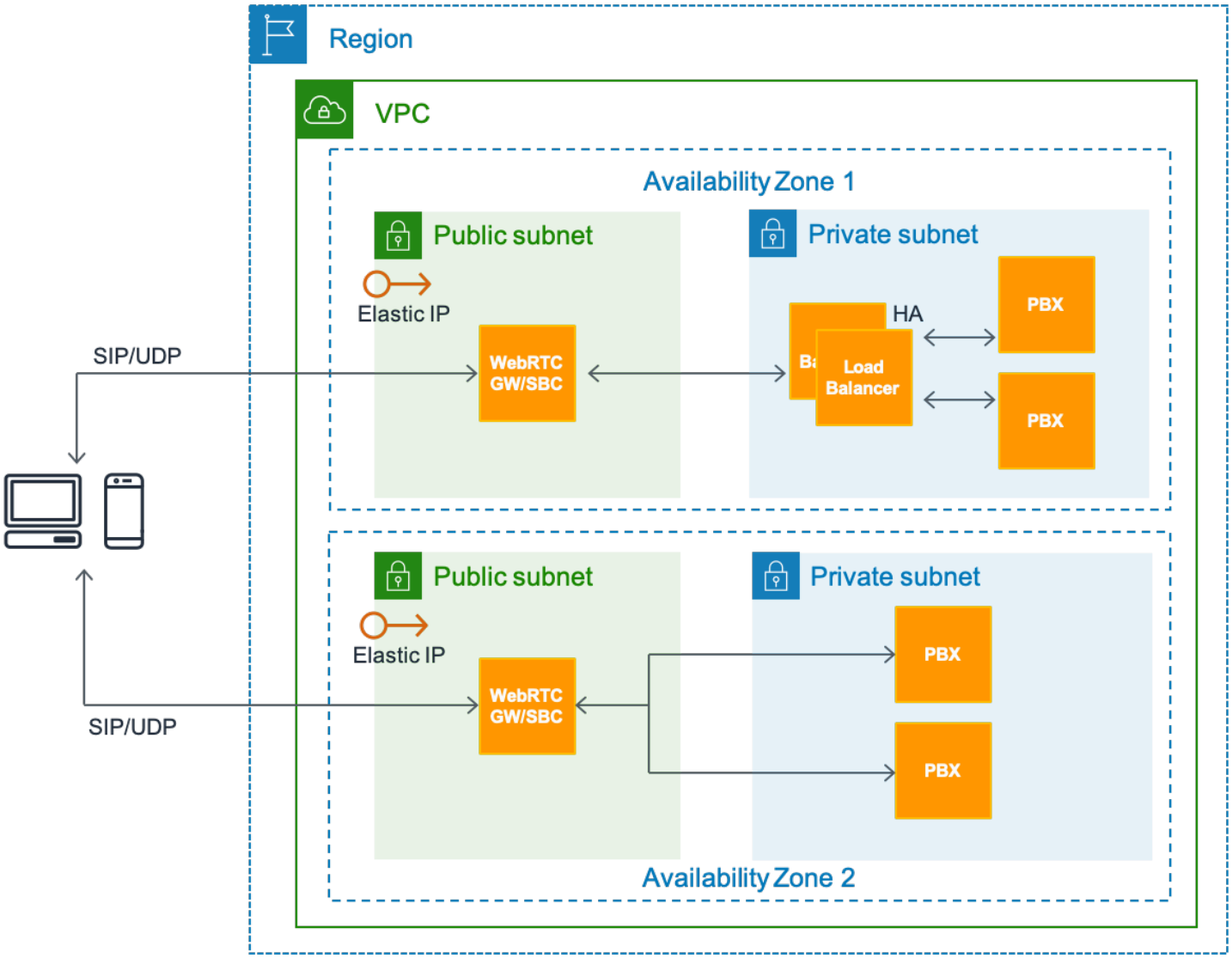

Figura 3: arquitetura RTC usando o Amazon EC2 em uma VPC

Alta disponibilidade e escalabilidade na AWS

A maioria dos provedores de comunicações em tempo real se alinha aos níveis de serviço que fornecem disponibilidade de 99,9% a 99,999%. Dependendo do grau de alta disponibilidade (HA) desejado, você deve tomar medidas cada vez mais sofisticadas ao longo de todo o ciclo de vida da aplicação. Recomendamos seguir estas diretrizes para obter um grau robusto de alta disponibilidade:

- Crie o sistema para não ter nenhum ponto de falha. Use mecanismos automatizados de monitoramento, detecção de falhas e failover para componentes sem estado e com estado

- Pontos únicos de falha (SPOF) geralmente são eliminados com uma configuração de redundância N+1 ou 2N, em que N+1 é obtido por meio do balanceamento de carga entre nós ativos-ativos, e 2N é obtido por um par de nós na configuração ativos-em espera.
- A AWS tem vários métodos para obter HA por meio de ambas as abordagens, como por meio de um cluster escalável com balanceamento de carga ou assumindo um par ativos-em espera.

- Instrumente e teste corretamente a disponibilidade do sistema.
- Prepare procedimentos operacionais para mecanismos manuais para responder, mitigar e se recuperar da falha.

Esta seção se concentra em como não ter nenhum ponto de falha usando os recursos disponíveis na AWS. Especificamente, esta seção descreve um subconjunto dos principais recursos e padrões de design da AWS que permitem criar aplicações de comunicação em tempo real altamente disponíveis na plataforma.

Tópicos

Padrão de IP flutuante para HA entre servidores com estado ativo e em espera

O padrão de design de IP flutuante é um mecanismo bem conhecido para obter failover automático entre um par ativo e em espera de nós de hardware (servidores de mídia). Um endereço IP virtual secundário estático é atribuído ao nó ativo. O monitoramento contínuo entre os nós ativo e em espera detecta falhas. Se o nó ativo falhar, o script de monitoramento atribuirá o IP virtual ao nó pronto em espera, que assumirá a função ativa primária. Desse modo, o IP virtual flutua entre o nó ativo e o nó em espera.

Tópicos

Aplicabilidade em soluções RTC

Nem sempre é possível ter várias instâncias ativas do mesmo componente em serviço, como um cluster ativo-ativo de N nós. Uma configuração ativo-em espera fornece o melhor mecanismo para HA. Por exemplo, os componentes com estado em uma solução RTC, como o servidor de mídia ou o servidor de conferência, ou até mesmo um SBC ou servidor de banco de dados, são adequados para uma configuração ativo-em espera. Um SBC ou servidor de mídia tem várias sessões ou canais de longa duração ativos em um determinado momento e, no caso de falha da instância ativa do SBC, os endpoints podem se reconectar ao nó em espera sem qualquer configuração do lado do cliente devido ao IP flutuante.

Implementação na AWS

Você pode implementar esse padrão na AWS usando os principais recursos do Amazon Elastic Compute Cloud (Amazon EC2), API do Amazon EC2, endereços IP elásticos e suporte no Amazon EC2 para endereços IP privados secundários.

1. Execute duas instâncias do EC2 para assumir as funções de nós primários e secundários, em que o primário está no estado ativo por padrão.
2. Atribua um endereço IP privado secundário adicional à instância primária do EC2.

3. Um endereço IP elástico, que é semelhante a um IP virtual (VIP), está associado ao endereço privado secundário. Esse é o endereço usado por endpoints externos para acessar a aplicação.
4. Algumas configurações do sistema operacional são necessárias para que o endereço IP secundário seja adicionado como um alias à interface de rede primária.
5. A aplicação deve ser vinculada a esse endereço IP elástico. No caso do software Asterisk, você pode configurar a vinculação por meio de configurações avançadas do Asterisk SIP.
6. Execute um script de monitoramento — personalizado, KeepAlive no Linux, Corosync e assim por diante — em cada nó para monitorar o estado do nó par. Caso o nó ativo atual falhe, o par detecta essa falha e invoca a API do Amazon EC2 para reatribuir o endereço IP privado secundário a si mesmo.
7. Portanto, a aplicação que estava escutando no VIP associado ao endereço IP privado secundário fica disponível para endpoints por meio do nó em espera.

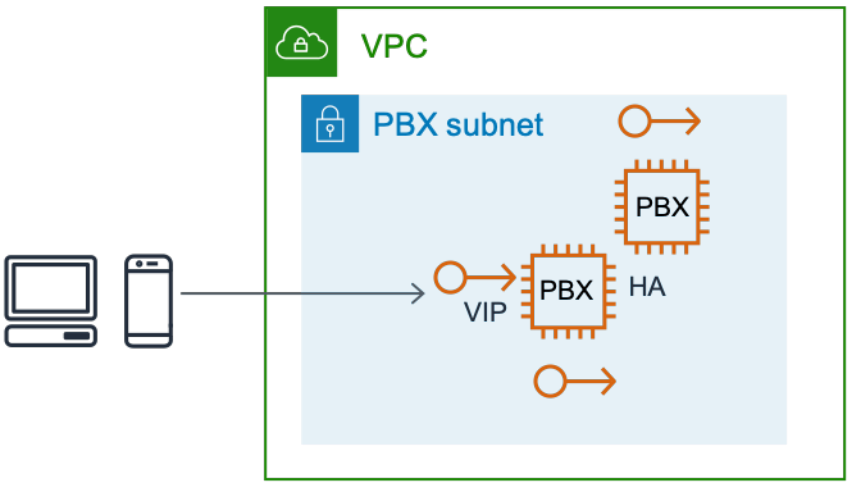

Figura 4: failover entre instâncias do EC2 com estado usando um endereço IP elástico

Benefícios

Essa abordagem é uma solução confiável de baixo orçamento que protege contra falhas no nível da infraestrutura, da aplicação ou da instância do EC2.

Limitações e extensibilidade

Esse padrão de design é normalmente limitado a uma única zona de disponibilidade. Ele pode ser implementado em duas zonas de disponibilidade, mas com uma variação. Nesse caso, o endereço IP elástico flutuante é reassociado entre o nó ativo e o nó em espera em diferentes zonas de

disponibilidade por meio da API de endereço IP elástico reassociada disponível. Na implementação de failover mostrada na Figura 4, as chamadas em andamento são descartadas, e os endpoints devem se reconectar. É possível estender essa implementação com a replicação dos dados da sessão subjacente para fornecer failover contínuo de sessões ou continuidade de mídia também.

Balanceamento de carga para escalabilidade e HA com WebRTC e SIP

O balanceamento de carga de um cluster de instâncias ativas com base em regras predefinidas, como Round Robin, afinidade ou latência, e assim por diante, é um padrão de design amplamente popularizado pela natureza sem estado das solicitações HTTP. Na verdade, o balanceamento de carga é uma opção viável no caso de muitos componentes de aplicações RTC.

O balanceador de carga atua como proxy reverso ou ponto de entrada para solicitações à aplicação desejada, que é configurada para ser executada em vários nós ativos simultaneamente. Em um determinado momento, o balanceador de carga direciona uma solicitação do usuário para um dos nós ativos no cluster definido. Os balanceadores de carga executam uma verificação de integridade em relação aos nós em seu cluster de destino e não enviam uma solicitação de entrada para um nó que falha na verificação de integridade. Portanto, um grau fundamental de alta disponibilidade é alcançado por meio do balanceamento de carga. Além disso, como um balanceador de carga executa verificações de integridade ativas e passivas em todos os nós do cluster em intervalos de menos de um segundo, o tempo de failover é quase instantâneo.

A decisão sobre qual nó direcionar se baseia nas regras do sistema definidas no balanceador de carga, incluindo:

- Round Robin
- Afinidade de IP ou sessão, que garante que várias solicitações dentro de uma sessão ou do mesmo IP sejam enviadas ao mesmo nó no cluster
- Com base em latência
- Com base em carga

Tópicos

- Implementação para SIP usando o Network Load Balancer ou um produto do AWS Marketplace

Aplicabilidade em arquiteturas RTC

O protocolo WebRTC possibilita que os gateways WebRTC sejam facilmente balanceados por meio de um balanceador de carga baseado em HTTP, como Elastic Load Balancing, Application Load Balancer ou Network Load Balancer. Com a maioria das implementações SIP contando com transporte por TCP e UDP, é necessário balanceamento de carga no nível de rede ou conexão com suporte para tráfego baseado em TCP e UDP.

Balanceamento de carga na AWS para WebRTC usando o Application Load Balancer e o Auto Scaling

No caso de comunicações baseadas em WebRTC, o Elastic Load Balancing fornece um balanceador de carga totalmente gerenciado, altamente disponível e escalável para servir como ponto de entrada para solicitações, que são direcionadas para um cluster de destino de instâncias do EC2 associadas ao Elastic Load Balancing. Além disso, como as solicitações WebRTC não têm estado, você pode usar o Amazon EC2 Auto Scaling para fornecer escalabilidade, elasticidade e alta disponibilidade totalmente automatizadas e controláveis.

O Application Load Balancer fornece um serviço de balanceamento de carga totalmente gerenciado altamente disponível que usa várias zonas de disponibilidade e é escalável. Ele oferece suporte ao balanceamento de carga de solicitações WebSocket que gerenciam a sinalização para aplicações WebRTC e comunicação bidirecional entre o cliente e o servidor usando uma conexão TCP de longa duração. O Application Load Balancer também oferece suporte a roteamento baseado em conteúdo e sessões persistentes, roteando solicitações do mesmo cliente para o mesmo destino usando cookies gerados pelo balanceador de carga. Se você habilitar as sessões persistentes, o mesmo destino receberá a solicitação, e você poderá usar o cookie para recuperar o contexto da sessão.

A Figura 5 mostra a topologia de destino.

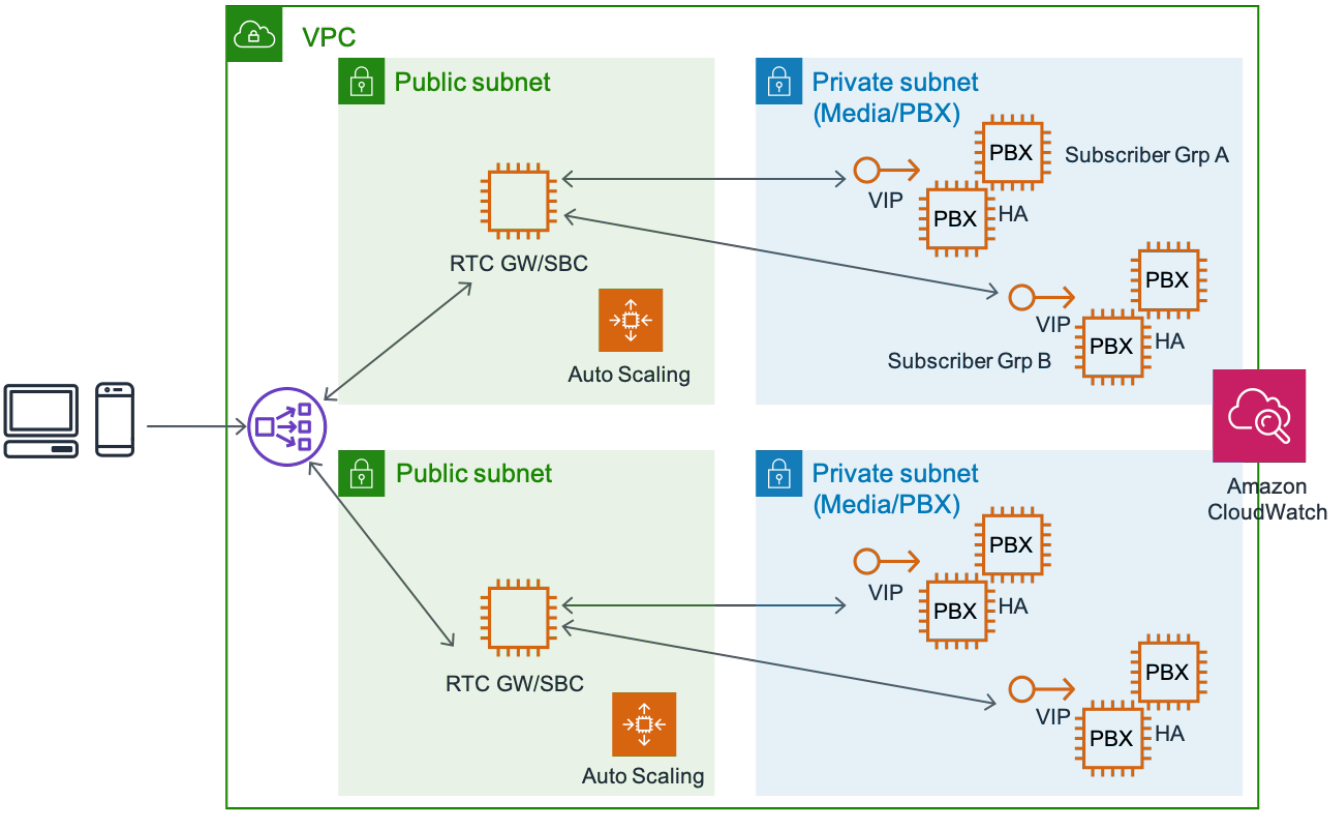

Figura 5: arquitetura de escalabilidade e alta disponibilidade do WebRTC

Implementação para SIP usando o Network Load Balancer ou um produto do AWS Marketplace

No caso de comunicações baseadas em SIP, as conexões são feitas por TCP ou UDP, com a maioria das aplicações RTC usando UDP. Se o SIP/TCP for o protocolo de sinal preferido, é possível usar o Network Load Balancer para balanceamento de carga totalmente gerenciado, altamente disponível, escalável e de alta performance.

Um Network Load Balancer opera no nível da conexão (camada 4), roteando conexões para destinos como instâncias do Amazon EC2, contêineres e endereços IP com base nos dados do protocolo IP. Ideal para balanceamento de carga de tráfego TCP ou UDP, o Network Load Balancer é capaz de processar milhões de solicitações por segundo, mantendo latências ultrabaixas. Ele é integrado a outros serviços populares da AWS, como o AWS Auto Scaling, o Amazon Elastic Container Service (Amazon ECS), o Amazon Elastic Kubernetes Service (Amazon EKS) e o AWS CloudFormation.

Se as conexões SIP forem iniciadas, outra opção é usar o software comercial pronto para uso do AWS Marketplace (COTS). O AWS Marketplace oferece muitos produtos que podem gerenciar

UDP e outros tipos de balanceamento de carga de conexão da camada 4. Esses COTS geralmente incluem suporte para alta disponibilidade e geralmente são integrados a recursos, como o AWS Auto Scaling, para melhorar ainda mais a disponibilidade e a escalabilidade. A Figura 6 mostra a topologia de destino:

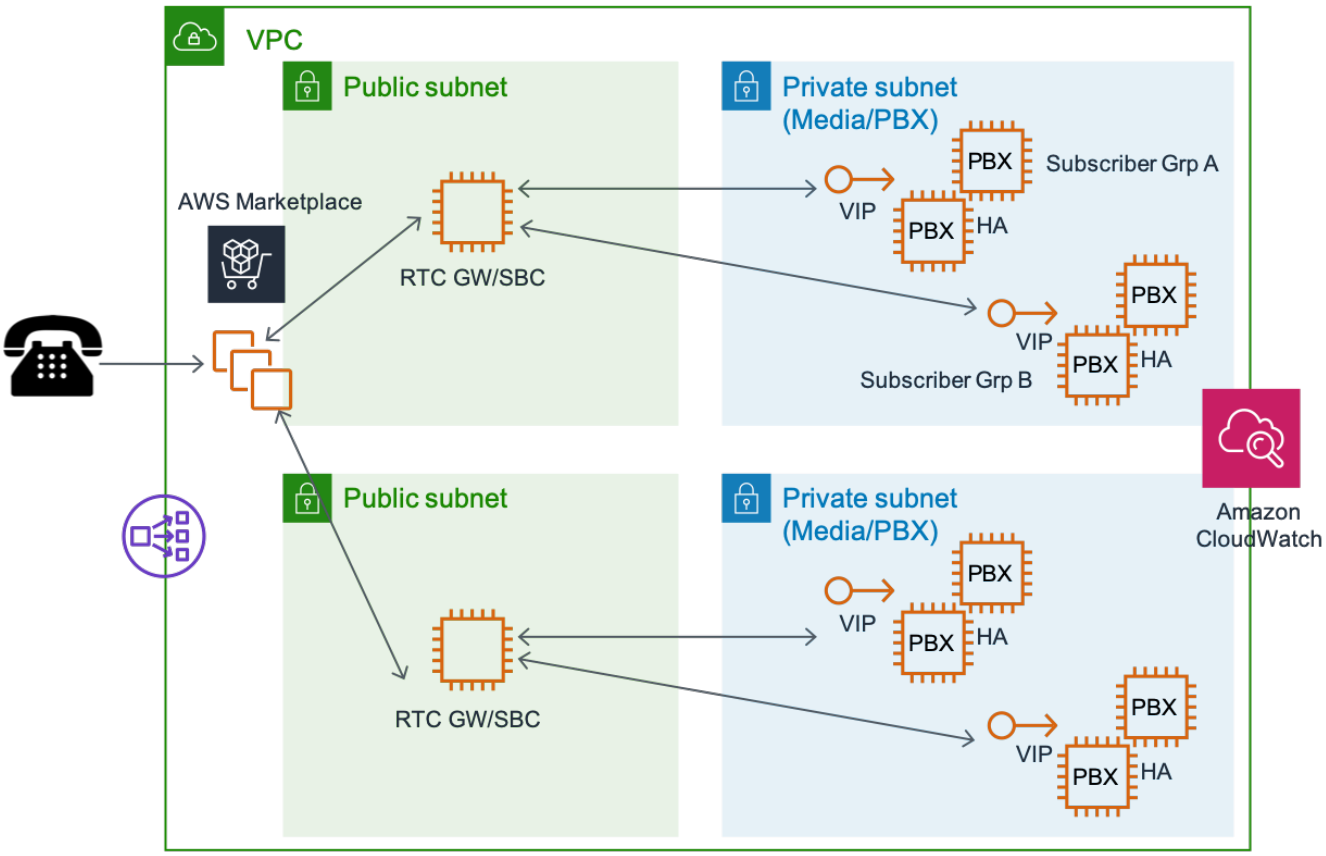

Figura 6: escalabilidade RTC baseada em SIP com um produto do AWS Marketplace

Balanceamento de carga e failover baseados em DNS entre regiões

O Amazon Route 53 fornece um serviço DNS global que pode ser usado como um endpoint público ou privado para clientes RTC registrarem e se conectarem a aplicações de mídia. Com o Amazon Route 53, as verificações de integridade do DNS podem ser configuradas para rotear o tráfego para endpoints íntegros ou monitorar de forma independente a integridade da sua aplicação. O recurso de fluxo de tráfego do Amazon Route 53 facilita o gerenciamento do tráfego globalmente por meio de diversos tipos de roteamento, incluindo roteamento baseado em latência, Geo DNS, geoproximidade e Weighted Round Robin, e todos podem ser combinados com DNS Failover para habilitar várias arquiteturas de baixa latência e tolerantes a falhas. O editor visual simples do Amazon Route 53

Traffic Flow permite gerenciar como os usuários finais são roteados para os endpoints da aplicação, que podem estar em uma única região da AWS ou distribuídos pelo mundo.

No caso de implantações globais, a política de roteamento baseada em latência no Route 53 é especialmente útil para direcionar os clientes para o ponto de presença mais próximo de um servidor de mídia para melhorar a qualidade do serviço associado a trocas de mídia em tempo real.

Para aplicar um failover a um novo endereço DNS, os caches do cliente devem ser liberados. Além disso, as alterações de DNS podem ter um atraso à medida que são propagadas pelos servidores DNS globais. Você pode gerenciar o intervalo de atualização para pesquisas de DNS com o atributo Vida útil. Esse atributo é configurável no momento da configuração das políticas de DNS.

Para alcançar usuários globais rapidamente ou atender aos requisitos de uso de um único IP público, o AWS Global Accelerator também pode ser usado para failover entre regiões. O AWS Global Accelerator é um serviço de rede que melhora a disponibilidade e a performance de aplicações com alcance local e global. O AWS Global Accelerator fornece endereços IP estáticos que atuam como um ponto de entrada fixo para os endpoints de aplicações, como Application Load Balancers, Network Load Balancers ou instâncias do Amazon EC2 em uma única ou em várias regiões da AWS. Ele usa a rede global da AWS para otimizar o caminho de seus usuários para suas aplicações, melhorando a performance, como a latência do tráfego TCP e UDP. O AWS Global Accelerator monitora continuamente a integridade de seus endpoints de aplicações e redireciona automaticamente o tráfego para os endpoints íntegros mais próximos caso os endpoints atuais sejam corrompidos. Para requisitos de segurança adicionais, o Accelerated Site-to-Site VPN usa o AWS Global Accelerator para melhorar a performance das conexões VPN roteando o tráfego de forma inteligente por meio da Rede Global da AWS e dos locais de borda da AWS.

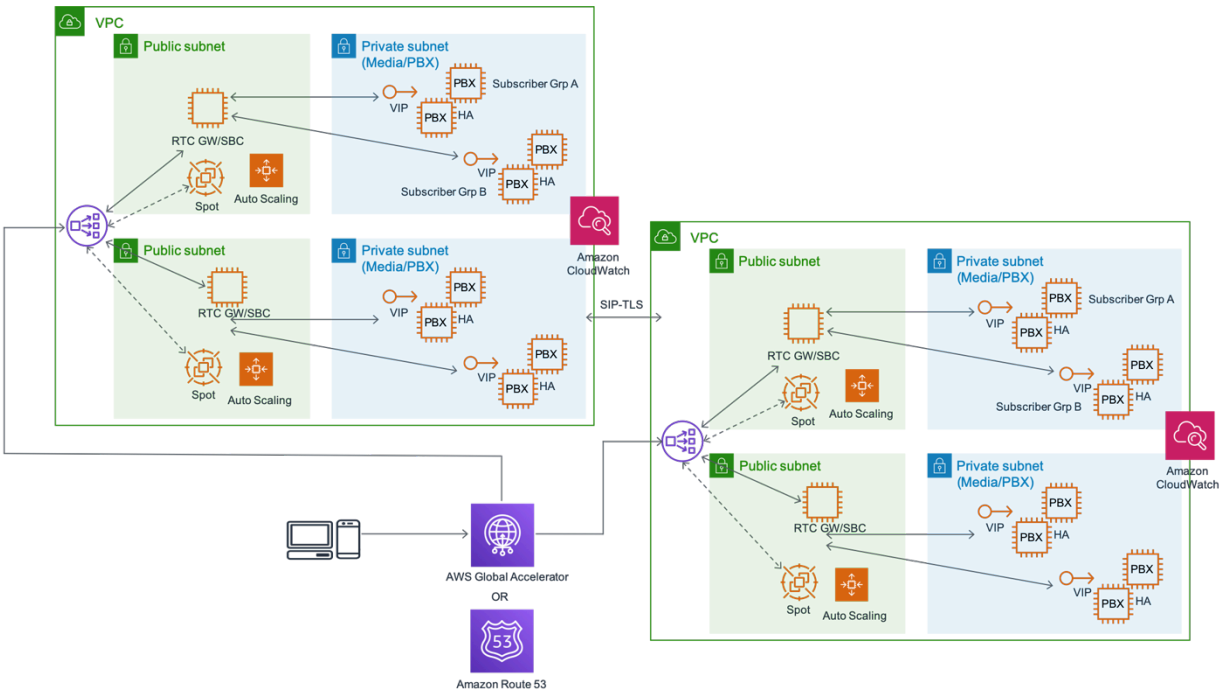

Figura 7: design de alta disponibilidade entre regiões usando o AWS Global Accelerator ou o Amazon Route 53

Durabilidade de dados e HA com armazenamento persistente

A maioria das aplicações RTC depende de armazenamento persistente para armazenar e acessar dados para autenticação, autorização, contabilidade (dados de sessão, registros de detalhes de chamadas etc.), monitoramento operacional e registro em log. Em um datacenter tradicional, garantir alta disponibilidade e durabilidade para os componentes de armazenamento persistente (bancos de dados, sistemas de arquivos etc.) normalmente requer trabalho pesado por meio da configuração de uma SAN, design RAID e processos de backup, restauração e processamento de failover. A Nuvem AWS simplifica e aprimora muito as práticas tradicionais de datacenter em relação à durabilidade e à disponibilidade de dados.

Para armazenamento de objetos e de arquivos, os serviços da AWS, como o Amazon Simple Storage Service (Amazon S3) e o Amazon Elastic File System (Amazon EFS), oferecem alta disponibilidade e escalabilidade gerenciadas. O Amazon S3 tem durabilidade de dados de 11 noves.

Para armazenamento de dados transacionais, os clientes têm a opção de aproveitar o Amazon Relational Database Service (Amazon RDS) totalmente gerenciado que oferece suporte ao Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle e Microsoft SQL Server com implantações de alta disponibilidade. Para a função de registrador, perfil do assinante ou armazenamento de registros

contábeis (por exemplo, CDRs), o Amazon RDS oferece uma opção tolerante a falhas, altamente disponível e escalável.

Escalabilidade dinâmica com o AWS Lambda, o Amazon Route 53 e o AWS Auto Scaling

A AWS permite o encadeamento de recursos e a capacidade de incorporar funções personalizadas sem servidor como um serviço com base em eventos de infraestrutura. Um desses padrões de design que tem muitos usos versáteis em aplicações RTC é a combinação de ganchos de ciclo de vida de escalabilidade automática com o Amazon CloudWatch Events, o Amazon Route 53 e funções do AWS Lambda. As funções do AWS Lambda podem incorporar qualquer ação ou lógica. A Figura 8 demonstra como esses recursos encadeados podem melhorar a confiabilidade e a escalabilidade do sistema com automação.

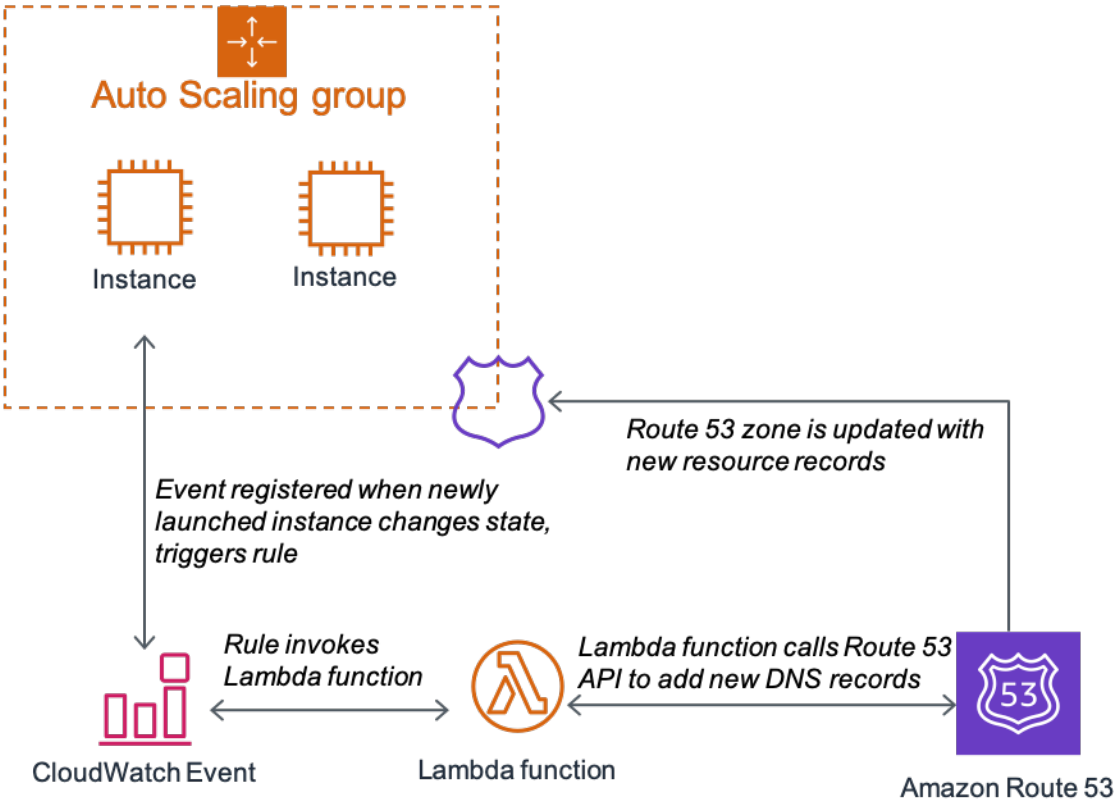

Figura 8: escalabilidade automática com atualizações dinâmicas para o Amazon Route 53

WebRTC altamente disponível com Kinesis Video Streams

O Amazon Kinesis Video Streams oferece transmissão de mídia em tempo real via WebRTC, permitindo que os usuários capturem, processem e armazenem transmissões de mídia para reprodução, análise e Machine Learning. Essas transmissões são altamente disponíveis, escaláveis e compatíveis com os padrões WebRTC. O Amazon Kinesis Video Streams inclui um endpoint de sinalização do WebRTC para agilizar a descoberta de pares e estabelecer uma conexão segura. Ele inclui endpoints STUN (Session Traversal Utilities for NAT) e TURN (Traversal Using Relays around NAT) gerenciados para garantir a troca de mídia entre os pares em tempo real. Inclui também um SDK de código aberto gratuito que se integra diretamente com firmware de câmera para proteger a comunicação com endpoints do Kinesis Video Streams, oferecendo descoberta de pares e streaming de mídia. Por fim, ele fornece bibliotecas de clientes para Android, iOS e JavaScript que permite que webplayers e dispositivos móveis compatíveis com WebRTC para descobrir e se conectar em segurança com um dispositivo de câmera para streaming de mídia e comunicação bidirecional.

Entroncamento SIP altamente disponível com o Amazon Chime Voice Connector

O Amazon Chime Voice Connector é um serviço de entroncamento SIP que você paga conforme o uso. Ele permite que as empresas façam e/ou recebam chamadas telefônicas seguras e baratas com os próprios sistemas de telefonia. O Amazon Chime Voice Connector é uma alternativa de baixo custo para troncos SIP do provedor de serviços ou de interfaces de taxa primária (PRIs) da rede digital de serviços integrados (ISDN). Os clientes têm a opção de habilitar chamadas recebidas, feitas ou ambas. O serviço utiliza a rede da AWS para gerar uma experiência de chamadas altamente disponível em várias regiões da AWS. Você pode transmitir áudio de chamadas telefônicas de entroncamento SIP ou feeds encaminhados de gravação de mídia baseada em SIP (SIPREC) para o Amazon Kinesis Video Streams para obter insights de chamadas empresariais em tempo real. Você pode criar rapidamente aplicações para análise de áudio por meio da integração com o Amazon Transcribe e outras bibliotecas comuns de Machine Learning.

Práticas recomendadas do campo

Esta seção tem como objetivo resumir as práticas recomendadas implementadas por alguns dos maiores e mais bem-sucedidos clientes da AWS que executam grandes workloads com o Session Initiation Protocol (SIP – Protocolo de início de sessão) em tempo real. Essas práticas recomendadas são valiosas para os clientes da AWS que desejam executar sua própria infraestrutura SIP na nuvem pública, pois elas podem ajudar a aumentar a confiabilidade e a resiliência do sistema em caso de diferentes tipos de falhas. Embora algumas dessas práticas recomendadas sejam específicas para SIP, a maioria delas é aplicável a qualquer aplicação de comunicação em tempo real executado na AWS.

Tópicos

Crie uma sobreposição de SIP

A AWS tem uma estrutura de rede robusta, escalável e redundante que fornece conectividade entre diferentes regiões. Quando um evento de rede, como um corte de fibra, degrada um link de estrutura da AWS, o tráfego é rapidamente transferido para caminhos redundantes usando protocolos de roteamento no nível da rede, como o BGP. Essa engenharia de tráfego no nível da rede é uma caixa preta para os clientes da AWS, e a maioria nem percebe esses eventos de failover. No entanto, os clientes que executam workloads em tempo real, como voz, vídeo de alta qualidade e mensagens de baixa latência, às vezes percebem esses eventos. Então, como um cliente da AWS pode implementar sua própria engenharia de tráfego além do que é fornecido pela AWS no nível da rede? A solução é implantar a infraestrutura SIP em muitas regiões diferentes da AWS. Como parte dos recursos de controle de chamadas, o SIP também permite rotear chamadas por meio de proxies SIP específicos.

Figura 9: usar o roteamento SIP para substituir o roteamento de rede

Na Figura 9, a infraestrutura SIP (representada por pontos verdes) está sendo executada em todas as quatro regiões dos EUA. As linhas azuis são uma representação fictícia da estrutura da AWS. Se nenhum roteamento SIP for implementado, uma chamada originada na costa oeste e destinada à costa leste dos EUA passará pelo link da estrutura que conecta diretamente as regiões do Oregon e da Virgínia. O diagrama mostra como um cliente pode substituir o roteamento no nível da rede e fazer a mesma chamada entre o Oregon e a Virgínia roteada pela Califórnia usando o roteamento SIP. Esse tipo de engenharia de tráfego SIP pode ser implementado usando proxies SIP e gateways de mídia com base em métricas de rede, como retransmissões SIP e preferências empresariais específicas do cliente.

Realize o monitoramento detalhado

Os usuários finais de aplicações de voz e vídeo em tempo real esperam o mesmo nível de performance obtido com os serviços de telefonia tradicionais. Assim sendo, quando eles enfrentam problemas com uma aplicação, isso acaba prejudicando a reputação do provedor. Para ser proativo em vez de reativo, é fundamental que o monitoramento detalhado seja implantado em todas as partes do sistema que atende aos usuários finais.

Figura 10: usar o SIPp para monitorar a infraestrutura de VoIP

Muitas ferramentas de código aberto, como iPerf ouSIPp e VOIPMonitor, estão disponíveis e podem ser usadas para monitorar o tráfego SIP/RTP. No exemplo anterior, os nós que executam o SIPp nos modos cliente e servidor estão avaliando métricas SIP, como chamadas bem-sucedidas e retransmissões SIP entre as quatro regiões da AWS dos EUA. Essas métricas podem ser exportadas para o Amazon CloudWatch usando um script personalizado. Usando o CloudWatch, os clientes podem criar alarmes nessas métricas personalizadas com base em um determinado valor limite. Ações de correção automáticas ou manuais podem ser realizadas com base no estado desses alarmes do CloudWatch.

Para os clientes que não desejam alocar os recursos de engenharia necessários para desenvolver e manter um sistema de monitoramento personalizado, há muitas soluções de monitoramento de VoIP disponíveis no mercado, como a ThousandEyes. Um exemplo de ação de correção é alterar o roteamento SIP com base no aumento das retransmissões SIP.

Use DNS para balanceamento de carga e IPs flutuantes para failover

Os clientes de telefonia IP compatíveis com o recurso DNS SRV podem usar com eficiência a redundância incorporada à infraestrutura, balanceando a carga de clientes para diferentes SBCs/PBXs.

Figura 11: usar registros SRV DNS para balancear a carga de clientes SIP

A Figura 11 mostra como os clientes podem usar os registros SRV para balancear a carga do tráfego SIP. Qualquer cliente de telefonia IP compatível com o padrão SRV procurará o prefixo sip._<transport protocol> em um registro DNS do tipo SRV. No exemplo, a seção de resposta do DNS contém os dois PBXs em execução em diferentes zonas de disponibilidade da AWS. No entanto, além dos URIs de endpoint, o registro SRV contém três informações adicionais:

- O primeiro número é a prioridade (1 no exemplo acima). Uma prioridade mais baixa é preferida em relação a outra mais alta.
- O segundo número é o peso (10 no exemplo acima).
- E o terceiro número é a porta a ser usada (5060).

Como a prioridade é a mesma (1) para ambos os servidores PBXs, os clientes usam o peso para balancear a carga entre os dois PBXs. Nesse caso, como os pesos são os mesmos, o tráfego SIP deve ter sua carga igualmente balanceada entre os dois PBXs.

O DNS pode ser uma boa solução para o balanceamento de carga do cliente, mas e a implementação de failover alterando/atualizando os registros “A” do DNS? Esse método é desencorajado devido à inconsistência encontrada no comportamento de cache do DNS no cliente e nos nós intermediários. Uma abordagem melhor para o failover intra-AZ entre um cluster de nós SIP é usar a reatribuição de IP do EC2 em que o endereço IP de um host corrompido é instantaneamente reatribuído a um host íntegro usando a API do EC2. Junto com uma solução de monitoramento detalhado e verificação de integridade, a reatribuição de IP de um nó com falha garante que o tráfego seja transferido para um host íntegro em tempo hábil, minimizando as interrupções para o usuário final.

Use várias zonas de disponibilidade

Cada região da AWS é subdividida em zonas de disponibilidade separadas. Cada zona de disponibilidade tem sua própria energia, resfriamento e conectividade de rede e, portanto, forma um domínio de falha isolado. Dentro das instalações da AWS, é sempre recomendável que os clientes executem suas workloads em mais de uma zona de disponibilidade. Isso garante que as aplicações do cliente possam suportar até mesmo uma falha completa da zona de disponibilidade, um evento muito raro por si só. Essa recomendação também se aplica à infraestrutura SIP em tempo real.

Figura 12: tratamento de falha da zona de disponibilidade

Vamos supor que um evento catastrófico (como o furacão de categoria 5) cause uma interrupção completa da zona de disponibilidade na região us-east-1. Com a infra-estrutura em execução conforme mostrado no diagrama, todos os clientes SIP que foram originalmente registrados com os nós na zona de disponibilidade que apresenta falha devem se registrar novamente com os nós SIP em execução na Zona de disponibilidade 2. (Teste esse comportamento com seus clientes/telefones SIP para confirmar a compatibilidade.). Embora as chamadas SIP ativas no momento da interrupção da zona de disponibilidade sejam perdidas, todas as novas chamadas são roteadas por meio da Zona de disponibilidade 2.

Para resumir, os registros SRV DNS devem direcionar o cliente para vários registros “A”, um em cada zona de disponibilidade. Cada um desses registros “A” deve, por sua vez, direcionar para vários endereços IP de SBCs/PBXs nessa zona de disponibilidade, fornecendo resiliência intra e inter-AZ. Os failovers intra e inter-AZ podem ser implementados com a reatribuição de IP, se os IPs forem públicos. No entanto, os IPs privados não podem ser reatribuídos entre as zonas de disponibilidade. Se um cliente estiver usando endereçamento IP privado, ele terá que depender de um novo registro dos clientes SIP com o SBC/PBX de backup para failover inter-AZ.

Mantenha o tráfego dentro de uma zona de disponibilidade e use os grupos de posicionamento do EC2

Também conhecida como afinidade da zona de disponibilidade, essa prática recomendada também se aplica ao raro evento de uma falha completa da zona de disponibilidade. É recomendável eliminar todo o tráfego entre as zonas de disponibilidade de modo que todo o tráfego SIP ou RTP que entre em uma zona de disponibilidade permaneça nela até sair da região.

Figura 13: afinidade da zona de disponibilidade (no máximo, 50% das chamadas ativas são perdidas)

A Figura 13 mostra uma arquitetura simplificada que usa a afinidade da zona de disponibilidade. A vantagem comparativa dessa abordagem fica clara se considerarmos os efeitos de uma interrupção completa da zona de disponibilidade. Conforme mostrado no diagrama, se a Zona de disponibilidade 2 for perdida, no máximo 50% das chamadas ativas serão afetadas (supondo um balanceamento de carga igual entre as zonas de disponibilidade). Se a afinidade da zona de disponibilidade não tivesse sido implementada, algumas chamadas fluiriam entre zonas de disponibilidade em uma região e, provavelmente, uma falha afetaria mais de 50% das chamadas ativas.

Além disso, para minimizar a latência do tráfego, também recomendamos que você considere o uso de grupos de posicionamento do EC2 em cada zona de disponibilidade. As instâncias iniciadas no mesmo grupo de posicionamento do EC2 têm maior largura de banda e latência reduzida, pois o EC2 garante a proximidade da rede dessas instâncias em relação umas às outras.

Use tipos de instância do EC2 de rede avançada

A escolha do tipo de instância certo no Amazon EC2 garante a confiabilidade do sistema, bem como o uso eficiente da infraestrutura. O EC2 oferece uma ampla seleção de tipos de instâncias otimizadas para atender a diferentes casos de uso. Os tipos de instância consistem em várias combinações de CPU, memória, armazenamento e capacidade de rede e oferecem flexibilidade de escolha da composição adequada de recursos para as suas aplicações. Esses tipos de instâncias de rede aprimoradas garantem que as workloads SIP executadas nelas tenham acesso a uma largura de banda consistente e latência agregada comparativamente menor. Uma adição recente ao Amazon EC2 é a disponibilidade do Elastic Network Adapter (ENA), que fornece até 100 Gbps de largura de banda. O catálogo mais recente de tipos de instância do EC2 e recursos associados pode ser encontrado na página de tipos de instância do EC2.

Para a maioria dos clientes, a última geração de instâncias otimizadas para computação deve oferecer o melhor valor pelo custo. Por exemplo, o C5N oferece suporte ao novo Elastic Network Adapter com largura de banda de até 100 Gbps com milhões de pacotes por segundo (PPS). A maioria das aplicações em tempo real também se beneficiaria do uso do Intel Data Plane Developer Kit (DPDK), que pode aumentar consideravelmente o processamento de pacotes de rede.

No entanto, é sempre uma prática recomendada comparar os vários tipos de instância do EC2 de acordo com seus requisitos para ver qual tipo de instância funciona melhor para você. A comparação também permite que você encontre outros parâmetros de configuração, como o número máximo de chamadas que um determinado tipo de instância pode processar por vez.

Considerações sobre segurança

Os componentes de aplicações RTC normalmente são executados diretamente na Internet, voltados para instâncias do Amazon EC2. Além de TCP, os fluxos usam protocolos como UDP e SIP. Nesses casos, o AWS Shield Standard protege as instâncias do Amazon EC2 contra ataques DDoS de camada de infraestrutura comum (Camadas 3 e 4), como ataques de reflexão UDP, reflexão de DNS, reflexão de NTP, reflexão de SSDP e assim por diante. O AWS Shield Standard usa várias técnicas, como modelagem de tráfego baseada em prioridade, que são acionadas automaticamente quando uma assinatura de ataque DDoS bem definida é detectada.

A AWS também oferece proteção avançada contra ataques DDoS grandes e sofisticados para essas aplicações ativando o AWS Shield Advanced em endereços IP elásticos. O AWS Shield Advanced fornece detecção avançada de DDoS que detecta automaticamente o tipo de recurso da AWS e o tamanho da instância do EC2 e aplica atenuações predefinidas apropriadas com proteções contra SYN ou UDP floods. Com o AWS Shield Advanced, os clientes também podem criar seus próprios perfis de atenuação personalizados usando o AWS DDoS Response Team (DRT) 24 horas por dia, 7 dias por semana. O AWS Shield Advanced também garante que, durante um ataque de DDoS, todas as listas de controle de acesso (ACLs) de rede da Amazon VPC sejam aplicadas automaticamente na borda da rede da AWS, permitindo acessar capacidade adicional de largura de banda e análise para atenuar grandes ataques DDoS volumétricos.

Conclusão

As workloads de comunicação em tempo real (RTC) podem ser implantadas na Amazon Web Services (AWS) para obter escalabilidade, elasticidade e alta disponibilidade, atendendo aos principais requisitos. Hoje, vários clientes estão usando a AWS, seus parceiros e soluções de código aberto para executar workloads de RTC com custo reduzido e mais agilidade, bem como uma pegada global reduzida.

As arquiteturas de referência e as práticas recomendadas apresentadas neste whitepaper podem ajudar os clientes a configurar com sucesso workloads de RTC na AWS e otimizar as soluções para atender aos requisitos do usuário final, otimizando para a nuvem.

Colaboradores

Os indivíduos e empresas a seguir contribuíram para este documento:

- Ahmad Khan, arquiteto de soluções sênior, Amazon Web Services
- Tipu Qureshi, engenheiro principal, AWS Support, Amazon Web Services
- Hasan Khan, gerente técnico de contas sênior da Amazon Web Services
- Shoma Chakravarty, líder técnico WW de telecomunicações da Amazon Web Services

Revisões do documento

Para ser notificado sobre atualizações deste whitepaper, inscreva-se no RSS feed.

update-history-change	update-history-description	update-history-date
Whitepaper atualizado	Atualizado para os serviços e recursos mais recentes.	13 de fevereiro de 2020
Publicação inicial	Primeira publicação do whitepaper.	1 de outubro de 2018

Avisos

Os clientes são responsáveis por fazer sua própria avaliação independente das informações neste documento. Este documento é: (a) fornecido apenas para fins informativos, (b) representa as ofertas e práticas de produtos atuais da AWS, que estão sujeitas a alterações sem aviso prévio e (c) não cria nenhum compromisso ou garantia da AWS e suas afiliadas, fornecedores ou licenciadores. Os produtos ou serviços da AWS são fornecidos no “estado em que se encontram”, sem garantias, declarações ou condições de qualquer tipo, explícitas ou implícitas. As responsabilidades e obrigações da AWS com seus clientes são regidas por contratos da AWS, e este documento não modifica nem faz parte de nenhum contrato entre a AWS e seus clientes.