Whitepaper da AWS

# Comunicação em tempo real na AWS

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

# Comunicação em tempo real na AWS: Whitepaper da AWS

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

As marcas comerciais e imagens comerciais da Amazon não podem ser usadas no contexto de nenhum produto ou serviço que não seja da Amazon, nem de qualquer maneira que possa gerar confusão entre os clientes ou que deprecie ou desprestigie a Amazon. Todas as outras marcas comerciais que não são propriedade da Amazon pertencem aos respectivos proprietários, os quais podem ou não ser afiliados, estar conectados ou ser patrocinados pela Amazon.

# Table of Contents

Resumo ..... 1
Resumo ..... 1
Você é Well-Architected? ..... 1
Introdução ..... 2
Componentes fundamentais da arquitetura RTC ..... 3
Softswitch/PBX ..... 4
Controlador de borda de sessão (SBC) ..... 4
Conectividade PSTN ..... 4
Gateway PSTN ..... 4
Tronco SIP ..... 4
Gateway de mídia (transcodificador) ..... 5
Notificações push no WebRTC ..... 5
WebRTC e gateway WebRTC ..... 6
Alta disponibilidade e escalabilidade em AWS ..... 8
Padrão de IP flutuante para HA entre servidores ativos e em espera ..... 8
Aplicabilidade em soluções RTC ..... 9
Aplicabilidade em arquiteturas RTC ..... 11
Balanceamento de carga ativado AWS para WebRTC usando Application Load Balancer e Auto Scaling ..... 11
Implementação para SIP usando o Network Load Balancer ou um produto AWS Marketplace ..... 12
Balanceamento de carga e failover baseados em DNS entre regiões ..... 14
Durabilidade de dados e HA com armazenamento persistente ..... 15
Escalabilidade dinâmica com AWS Lambda Amazon Route 53 e Amazon EC2 Auto Scaling ..... 16
WebRTC altamente disponível com Amazon Kinesis Video Streams ..... 17
Entroncamento SIP de alta disponibilidade com o Amazon Chime Voice Connector ..... 17
Melhores práticas do campo ..... 18
Crie uma sobreposição SIP ..... 18
Execute monitoramento detalhado ..... 19
Use DNS para balanceamento de carga e IPs flutuação para failover ..... 20
Use várias zonas de disponibilidade ..... 22
Mantenha o tráfego dentro de uma zona de disponibilidade e use grupos de EC2 posicionamento ..... 23
Use tipos de EC2 instância de rede aprimorados ..... 24

Considerações sobre segurança ..... 25
Conclusão ..... 26
Acrônimos ..... 27
Colaboradores ..... 29
Revisões do documento ..... 30
Avisos ..... 31
AWS Glossário ..... 32
..... xxxiii

# Comunicação em tempo real em AWS

Melhores práticas para projetar cargas de trabalho de comunicação em tempo real (RTC) altamente disponíveis e escaláveis em AWS

Data de publicação: 5 de maio de 2022 (Revisões do documento)

## Resumo

Atualmente, muitas organizações buscam reduzir custos e obter escalabilidade para cargas de trabalho de voz, mensagens e multimídia em tempo real. Este paper descreve as melhores práticas para gerenciar cargas de trabalho de comunicação em tempo real (RTC) na Amazon Web Services (AWS) e inclui arquiteturas de referência para atender a esses requisitos. Este paper serve como um guia para pessoas familiarizadas com a comunicação em tempo real sobre como obter alta disponibilidade e escalabilidade para essas cargas de trabalho.

Este paper inclui arquiteturas de referência que mostram como configurar cargas de trabalho RTC e as melhores práticas para otimizar as soluções para atender aos requisitos do usuário final e AWS, ao mesmo tempo, otimizar para a nuvem. O Evolved Packet Core (EPC) está fora do escopo deste whitepaper, mas as melhores práticas detalhadas aqui podem ser aplicadas às funções de rede virtual (). VNFs

## Sua arquitetura está bem planejada?

A AWS Well-Architected Framework ajuda você a entender os prós e os contras das decisões que você toma ao criar sistemas na nuvem. Os seis pilares do framework permitem a você conhecer as melhores práticas de arquitetura para criar e operar sistemas confiáveis, seguros, econômicos e sustentáveis na nuvem. Usando o AWS Well-Architected Tool, disponível gratuitamente no Console de gerenciamento da AWS(é necessário fazer login), você pode analisar suas cargas de trabalho em relação a essas melhores práticas respondendo a um conjunto de perguntas para cada pilar.

Para obter orientações especializadas e melhores práticas adicionais para a arquitetura de sua nuvem (implantações de arquitetura de referência, diagramas e whitepapers), consulte o Centro de Arquitetura da AWS.

# Introdução

Aplicativos de telecomunicações que usam voz, vídeo e mensagens como canais são um requisito fundamental para muitas organizações e seus usuários finais. Essas cargas de trabalho de comunicação em tempo real (RTC) têm requisitos específicos de latência e disponibilidade que podem ser atendidos seguindo as melhores práticas de design relevantes. No passado, as cargas de trabalho RTC eram implantadas em data centers locais tradicionais com recursos dedicados.

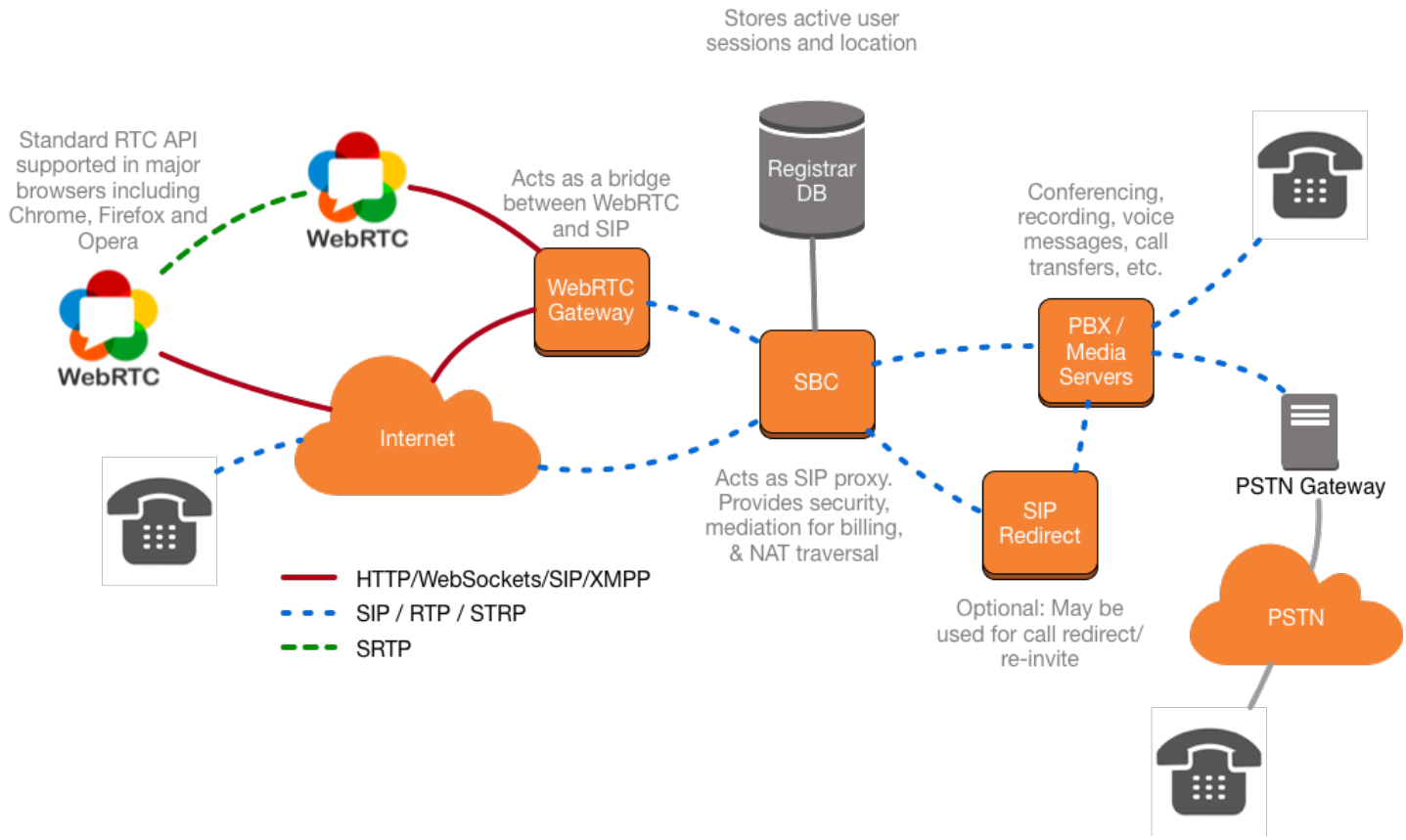
As cargas de trabalho RTC exigem um ambiente altamente escalável, resiliente e disponível. Atualmente, os clientes costumam AWS executar cargas de trabalho RTC com custo reduzido, maior agilidade, elasticidade e tempo de lançamento no mercado.

# Componentes fundamentais da arquitetura RTC

No setor de telecomunicações, o RTC geralmente se refere a sessões de mídia ao vivo entre dois endpoints com latência mínima. Essas sessões podem estar relacionadas a:

- Uma sessão de voz entre duas partes (como um sistema telefônico, celular ou Voz sobre IP (VoIP))
- Mensagens instantâneas (como bate-papo e Instant Relay Chat (IRC))
- Sessão de vídeo ao vivo (como videoconferência e telepresença)

Cada uma das soluções anteriores tem alguns componentes em comum (como componentes que fornecem autenticação, autorização e controle de acesso, transcodificação, armazenamento em buffer e retransmissão, etc.) e alguns componentes exclusivos do tipo de mídia transmitida (como serviço de transmissão, servidor de mensagens e filas, etc.). Esta seção se concentra na definição de um sistema RTC baseado em voz e vídeo e todos os componentes relacionados, conforme ilustrado na figura a seguir.




Componentes arquitetônicos essenciais para RTC

# Softswitch/PBX

Um softswitch ou PBX é o cérebro de um sistema de telefonia por voz e fornece inteligência para estabelecer, manter e rotear uma chamada de voz dentro ou fora da empresa usando componentes diferentes. Todos os assinantes da empresa precisam se registrar 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 acompanha 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 do protocolo de iniciação de sessão (SIP) para conectividade externa. Alguns SBCs também oferecem recursos de transcodificação para conversão CODECsde um formato para outro. A maioria SBCs também fornece recursos de travessia de tradução de endereços de rede (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 de rede telefônica pública comutada (PSTN) e troncos SIP para se conectar a redes PSTN antigas.

## Gateway PSTN

O gateway PSTN converte a sinalização entre SIP e mídia entre o Real Time Transport Protocol (RTP) SS7 e a 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 (SS7 baseada), mas os fluxos entre a empresa e a empresa de telecomunicações permanecem por IP. A maioria dos troncos SIP é estabelecida usando. 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)

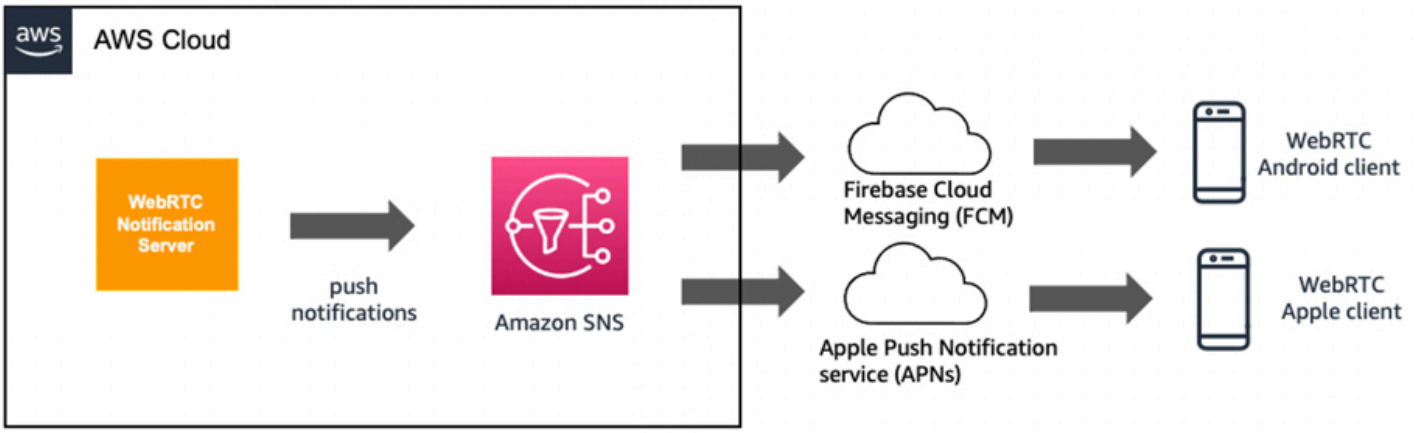
Os usuários se comunicam em tempo real usando áudio e/ou vídeo, além de dados opcionais e outras informações. Para se comunicar, os dois dispositivos precisam ser capazes de concordar com um codec mutuamente compreendido para cada faixa de mídia, para que possam se comunicar e apresentar com sucesso a mídia compartilhada. Todos os navegadores compatíveis com WebRTC devem oferecer suporte ao usuário de posicionamento on-line (OPUS) e G711 para áudio e perfil de linha de base restrita VP8H.264 para vídeo.

Uma solução de voz típica fora do ecossistema WebRTC permite vários tipos de. CODECs Algumas das mais comuns CODECs são G.711 μ-law para a América do Norte, G.711 A-law, G.729 e G.722. Quando dois dispositivos que usam dois dispositivos diferentes CODECs se comunicam entre si, o gateway 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 entre si.

# Notificações push no WebRTC

As implementações do WebRTC são muito comuns em dispositivos móveis. Ao contrário dos navegadores da web, um dispositivo móvel não consegue manter a conectividade de um websocket aberta por muito tempo. Portanto, ele precisa contar com notificações push do servidor WebRTC para todas as solicitações finais, como chamadas e mensagens.

O Amazon Simple Notification Service (Amazon SNS) permite que você envie notificações push para aplicativos em dispositivos móveis. Esses aplicativos podem ser executados em vários sistemas operacionais, como Apple iOS ou Android. A figura a seguir mostra uma visão geral de alto nível do fluxo de notificações push, de um servidor de notificação WebRTC para endpoints móveis WebRTC.

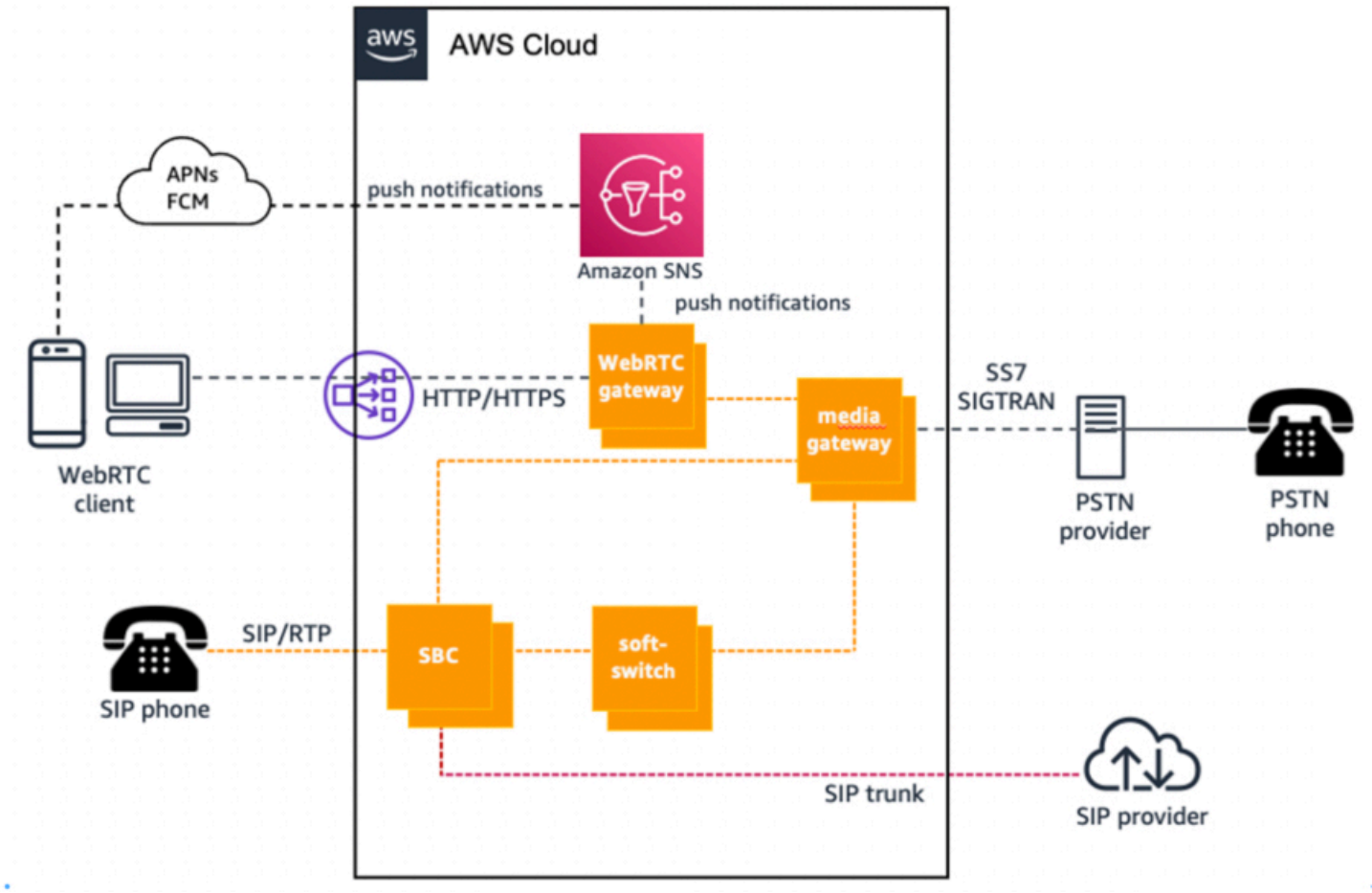



Amazon SNS para notificações push

# WebRTC e gateway WebRTC

A comunicação em tempo real na Web (WebRTC) permite que você estabeleça uma chamada a partir de um navegador da Web ou solicite recursos do servidor de back-end usando a API. A tecnologia foi projetada com a tecnologia de nuvem em mente 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 elas APIs, o gateway WebRTC é necessário para traduzir chamadas de API em mensagens SIP e vice-versa.

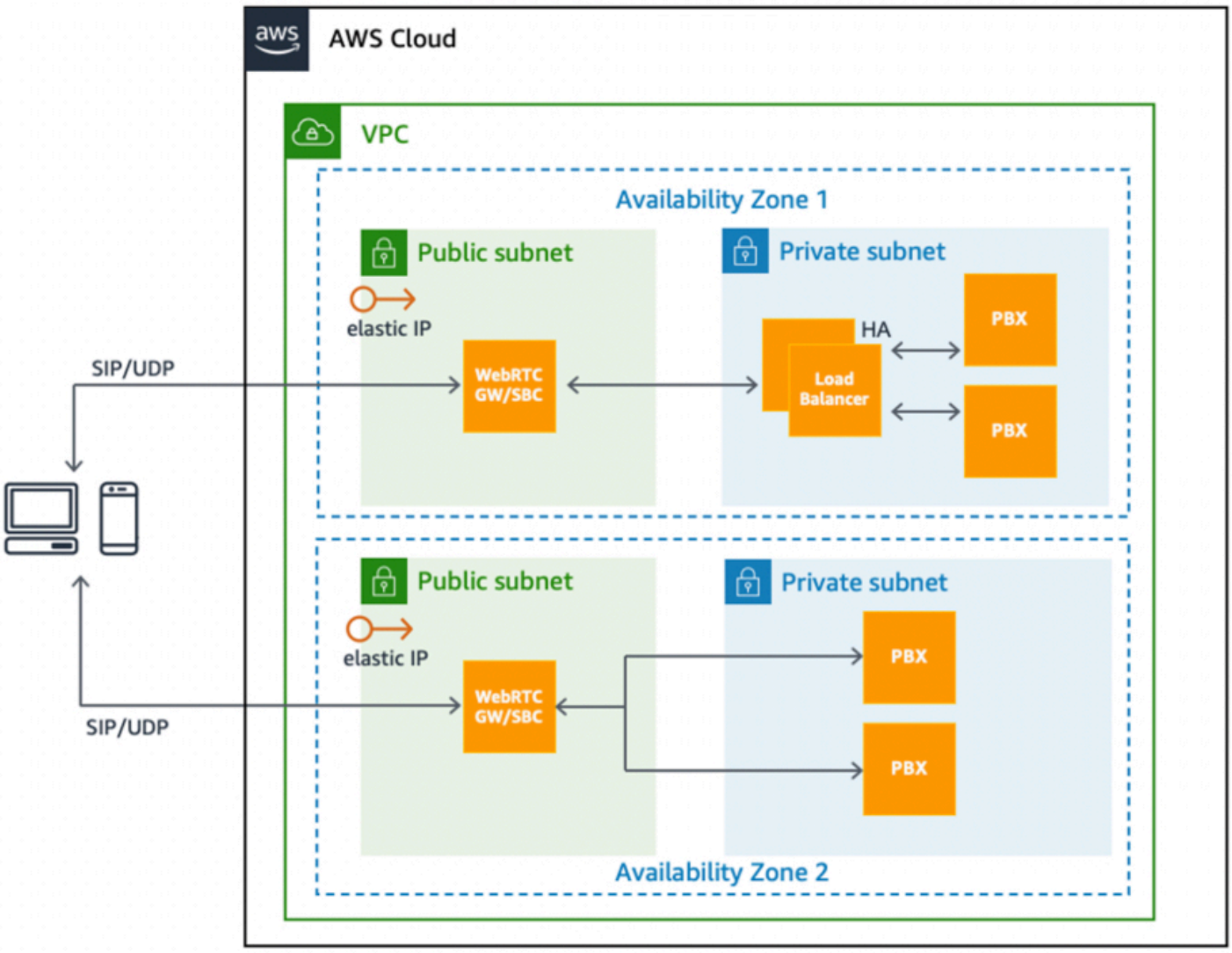
A figura a seguir mostra um padrão de design para uma arquitetura WebRTC altamente disponível. O tráfego de entrada dos clientes WebRTC é balanceado por um Application Load Balancer (ALB) com o WebRTC executado em instâncias do Amazon Elastic Compute Cloud (Amazon) que fazem parte de um grupo do Amazon Auto Scaling. EC2 EC2




Uma topologia básica de um sistema RTC para voz

Outro padrão de design para tráfego SIP e RTP é usar pares de EC2 na SBCs Amazon no modo ativo-passivo em todas as zonas de disponibilidade, conforme mostrado na figura a seguir. Aqui,

um endereço IP elástico pode ser movido dinamicamente entre instâncias em caso de falha, onde o Domain Name Service (DNS) não pode ser usado.




Arquitetura RTC usando a Amazon EC2 em uma nuvem privada virtual (VPC)

# Alta disponibilidade e escalabilidade em 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 do aplicativo. A AWS recomenda seguir essas diretrizes para alcançar um grau robusto de alta disponibilidade:

- Projete o sistema para que não tenha um único ponto de falha. Use mecanismos automatizados de monitoramento, detecção de falhas e failover para componentes sem estado e com estado
  - Os 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 e ativos e 2N é obtido por um par de nós na configuração ativa em espera.
  - A AWS tem vários métodos para obter HA por meio de ambas as abordagens, como por meio de um cluster escalável e com balanceamento de carga ou assumindo um par ativo e em espera.
- Instrumentar corretamente e testar 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 atingir um único ponto de falha usando os recursos disponíveis em AWS. Especificamente, esta seção descreve um subconjunto dos principais AWS recursos e padrões de design que permitem criar aplicativos de comunicação em tempo real altamente disponíveis.

## Padrão de IP flutuante para HA entre servidores ativos e em espera

O padrão de design de IP flutuante é um mecanismo bem conhecido para obter failover automático entre um par de nós de hardware ativos e em espera (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 em espera pronto e o nó em espera assumirá a função ativa primária. Dessa forma, o IP virtual flutua entre o nó ativo e o em espera.

# 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 ativa em espera fornece o melhor mecanismo para HA. Por exemplo, os componentes com estado em uma solução RTC, como o servidor de mídia ou de conferência, ou até mesmo um servidor SBC ou de banco de dados, são adequados para uma configuração ativa e em espera. Um servidor SBC ou de mídia tem várias sessões ou canais de longa execução 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 nenhuma configuração do lado do cliente devido ao IP flutuante.

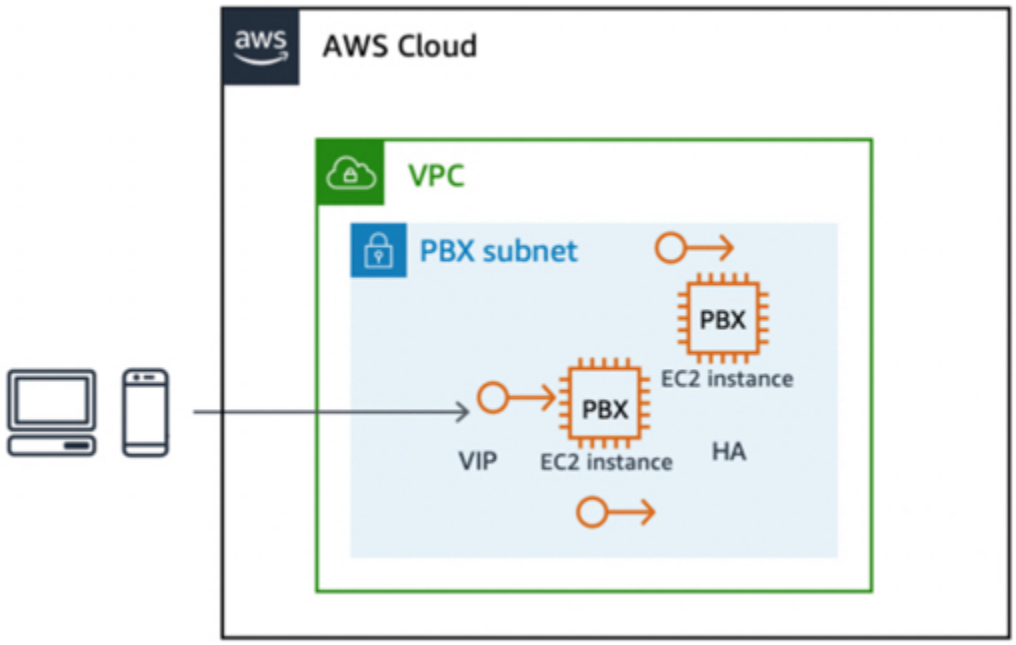
## Implementação em AWS

Você pode implementar esse padrão na AWS usando os principais recursos da Amazon Elastic Compute Cloud (Amazon EC2), Amazon EC2 API, endereços IP elásticos e suporte na Amazon EC2 para endereços IP privados secundários.

Para implementar o padrão de IP flutuante em AWS:

1. Inicie duas EC2 instâncias para assumir as funções dos nós primário e secundário, em que se presume que o primário esteja no estado ativo por padrão.
2. Atribua um endereço IP privado secundário adicional à EC2 instância primária.
3. Um endereço IP elástico, que é semelhante a um IP virtual (VIP), está associado ao endereço privado secundário. Esse endereço privado secundário é o endereço usado pelos endpoints externos para acessar o aplicativo.
4. Algumas configurações do sistema operacional (SO) são necessárias para que o endereço IP secundário seja adicionado como um alias à interface de rede primária.
5. O aplicativo deve se vincular 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 em Linux, Corosync e assim por diante — em cada nó para monitorar o estado do nó de mesmo nível. Caso o nó ativo atual falhe, o par detecta essa falha e invoca a API da EC2 Amazon para reatribuir o endereço IP privado secundário a si mesmo.

   Portanto, o aplicativo que estava escutando o VIP associado ao endereço IP privado secundário fica disponível para os endpoints por meio do nó em espera.




Failover entre EC2 instâncias com estado usando um endereço IP elástico

Benefícios

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

Limitações e extensibilidade

Esse padrão de design geralmente é 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 reassociação de endereços IP elásticos disponível. Na implementação de failover mostrada na figura anterior, as chamadas em andamento são descartadas e os endpoints devem se reconectar. É possível estender essa implementação com a replicação dos dados subjacentes da sessão para fornecer um failover contínuo das sessões ou também a continuidade da mídia.

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, etc., 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 do aplicativo RTC.

O balanceador de carga atua como proxy reverso ou ponto de entrada para solicitações ao aplicativo desejado, que por sua vez está configurado para ser executado em vários nós ativos simultaneamente. Em qualquer momento, o balanceador de carga direciona uma solicitação do usuário para um dos nós ativos no cluster definido. Os balanceadores de carga realizam uma verificação de integridade nos nós em seu cluster de destino e não enviam uma solicitação de entrada para um nó que falhe na verificação de integridade. Portanto, um grau fundamental de alta disponibilidade é alcançado pelo balanceamento de carga. Além disso, como um balanceador de carga realiza 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 é baseada nas regras do sistema definidas no balanceador de carga, incluindo:

- Ida e volta
- Afinidade de sessão ou IP, que garante que várias solicitações em uma sessão ou do mesmo IP sejam enviadas para o mesmo nó no cluster
- Baseado em latência
- Baseado na carga

## Aplicabilidade em arquiteturas RTC

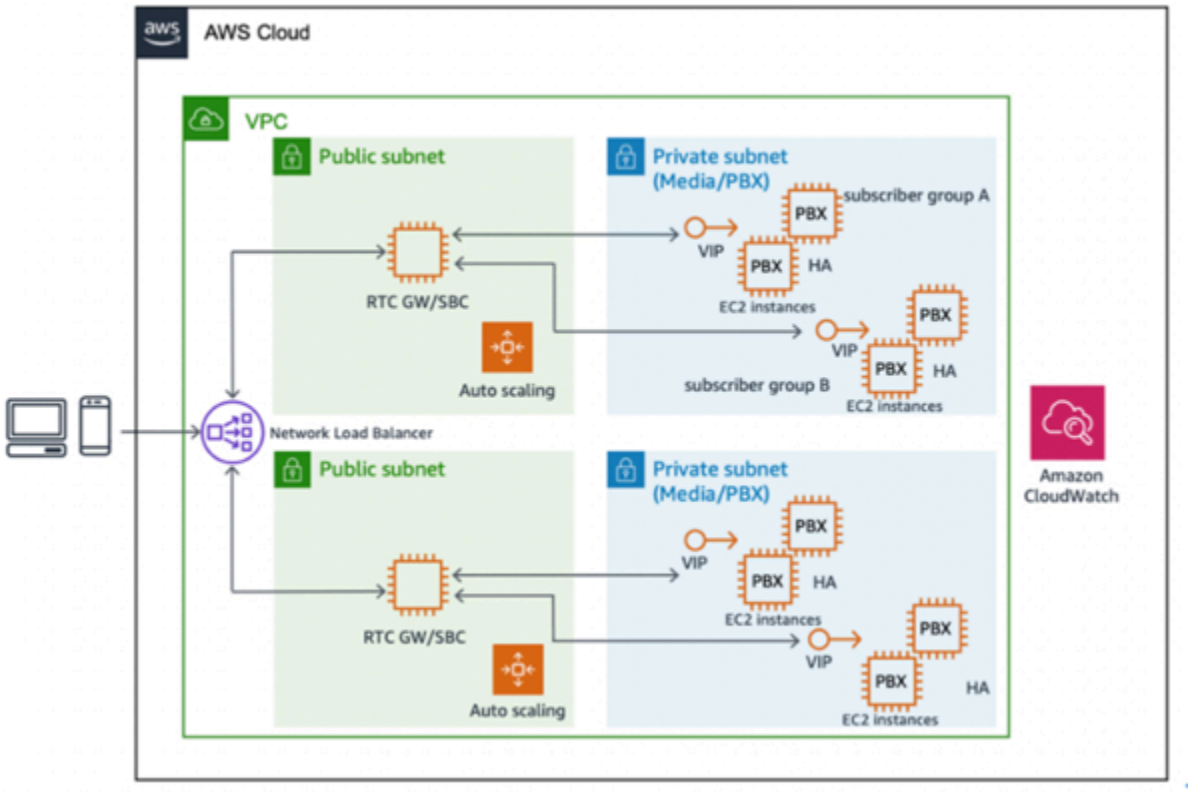
O protocolo WebRTC possibilita que os WebRTC Gateways sejam facilmente balanceados por meio de um balanceador de carga baseado em HTTP, como Elastic Load Balancing (ELB), Application Load Balancer (ALB) ou Network Load Balancer (NLB). Com a maioria das implementações de SIP dependendo do transporte pelo Protocolo de Controle de Transmissão (TCP) e pelo Protocolo de Datagrama de Usuário (UDP), você precisa de balanceamento de carga em nível de rede ou conexão, com suporte para tráfego baseado em TCP e UDP.

## Balanceamento de carga ativado AWS para WebRTC usando Application Load Balancer e 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 então direcionadas para um cluster de instâncias de destino associado ao Elastic Load Balancing. EC2 Como as solicitações do 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 usando várias zonas de disponibilidade e escalável. Isso suporta o balanceamento de carga de WebSocket solicitações que manipulam a sinalização para aplicativos WebRTC e a comunicação bidirecional entre o cliente e o servidor usando uma conexão TCP de longa duração. O Application Load Balancer também suporta roteamento baseado em conteúdo e sessões fixas, roteando solicitações do mesmo cliente para o mesmo destino usando cookies gerados pelo balanceador de carga. Se você ativar sessões fixas, o mesmo destino receberá a solicitação e poderá usar o cookie para recuperar o contexto da sessão.

A figura a seguir mostra a topologia de destino.



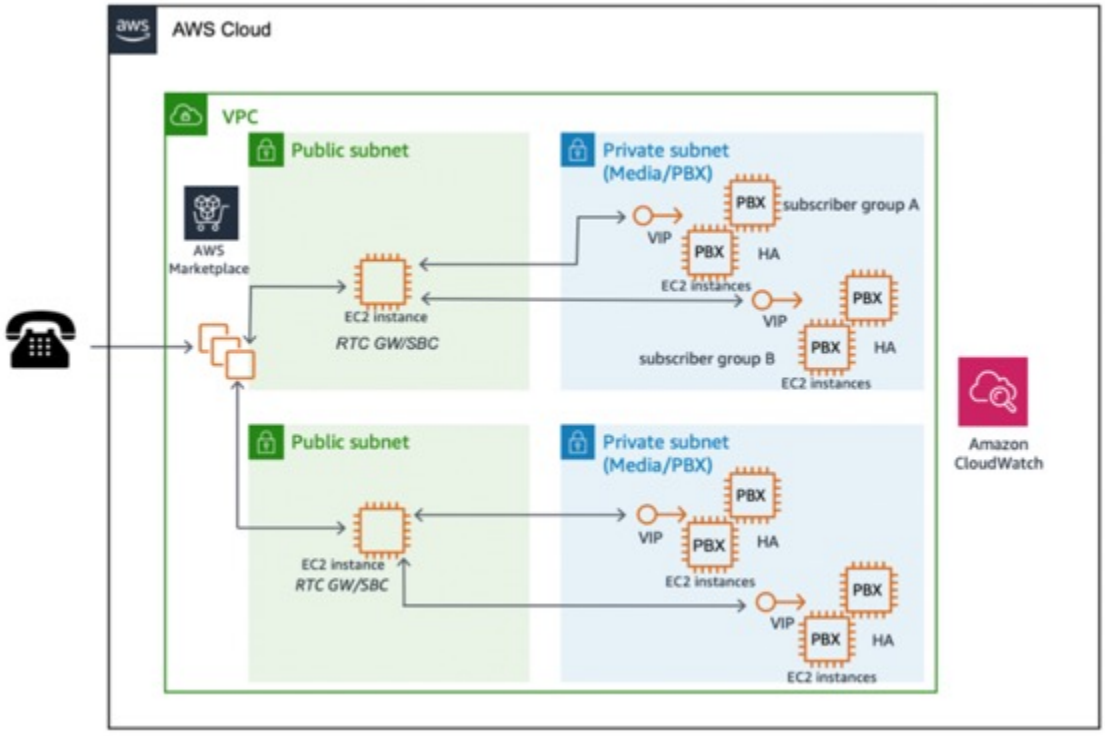

Escalabilidade e arquitetura de alta disponibilidade do WebRTC

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

No caso de comunicações baseadas em SIP, as conexões são feitas por TCP ou UDP, com a maioria dos aplicativos 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 desempenho.

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

Se as conexões SIP forem iniciadas, outra opção é usar off-the-shelf software AWS Marketplacecomercial (COTS). Ele AWS Marketplace oferece muitos produtos que podem lidar com UDP e outros tipos de balanceamento de carga de conexão de camada quatro. O COTS normalmente inclui suporte para alta disponibilidade e geralmente se integra a recursos, como o Amazon EC2 Auto Scaling, para aumentar ainda mais a disponibilidade e a escalabilidade. A figura a seguir mostra a topologia de destino:




Escalabilidade de RTC baseada em SIP com o produto AWS Marketplace

# Balanceamento de carga e failover baseados em DNS entre regiões

O Amazon Route 53 fornece um serviço de DNS global que pode ser usado como um endpoint público ou privado para clientes RTC se registrarem e se conectarem com aplicativos de mídia. Com o Amazon Route 53, as verificações de saúde do DNS podem ser configuradas para rotear o tráfego para endpoints íntegros ou para monitorar de forma independente a integridade do seu aplicativo.

O recurso de fluxo de tráfego do Amazon Route 53 facilita o gerenciamento global do tráfego por meio de vários tipos de roteamento, incluindo roteamento baseado em latência, geoDNS, geoproximidade e round robin ponderado, tudo isso pode ser combinado com o DNS Failover para permitir uma variedade de arquiteturas de baixa latência e tolerantes a falhas. O editor visual simples do Amazon Route 53 Traffic Flow permite que você gerencie como seus usuários finais são encaminhados para os endpoints do seu aplicativo, seja em uma única região da AWS ou distribuídos em todo o mundo.

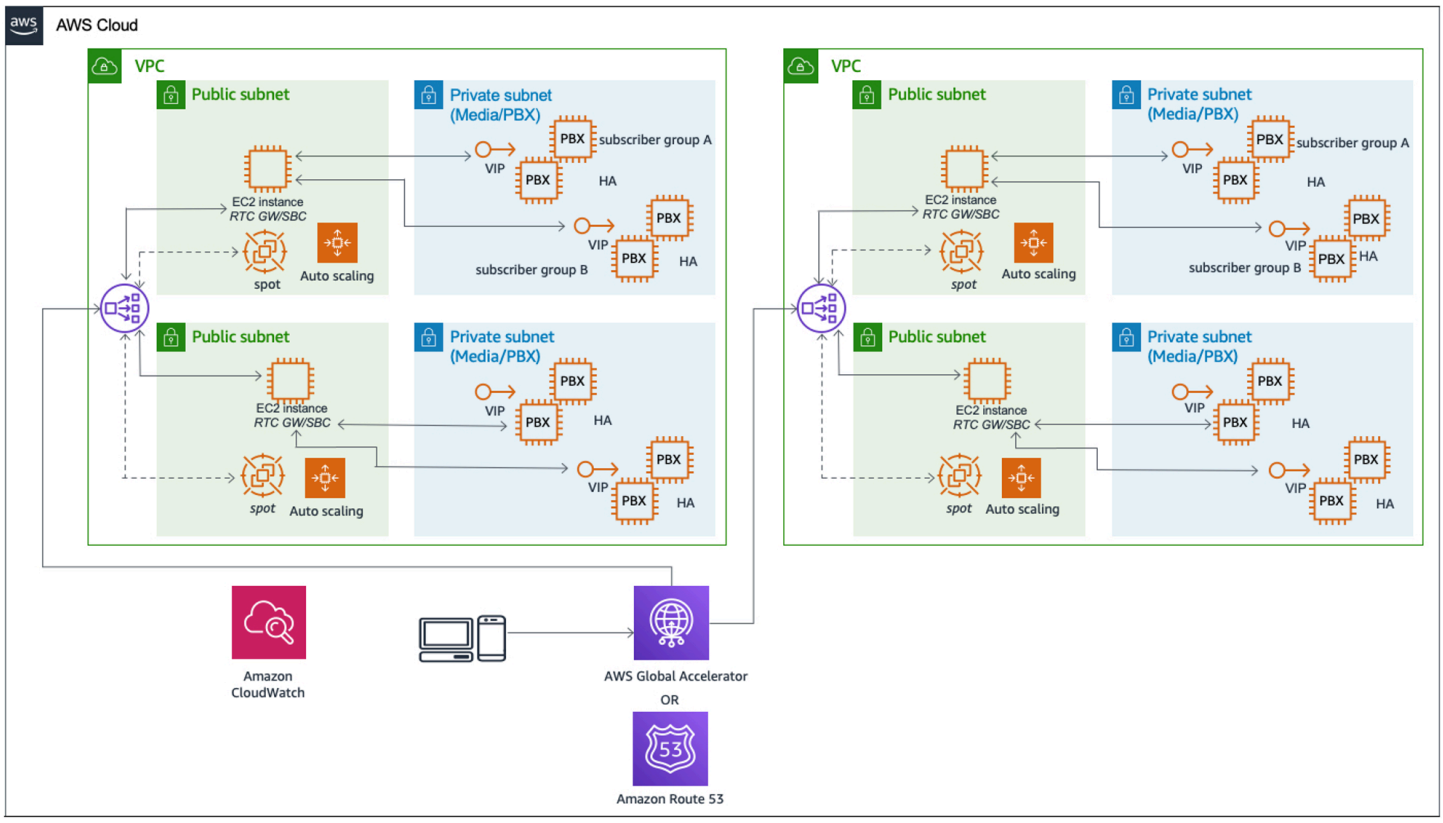
No caso de implantações globais, a política de roteamento baseada em latência no Route 53 é especialmente útil para direcionar os clientes ao ponto de presença mais próximo de um servidor de mídia para melhorar a qualidade do serviço associado às trocas de mídia em tempo real.

Observe que, para impor um failover para um novo endereço DNS, os caches do cliente devem ser esvaziados. 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 Time to Live. 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, também AWS Global Accelerator pode ser usado para failover entre regiões. AWS Global Acceleratoré um serviço de rede que melhora a disponibilidade e o desempenho de aplicativos com alcance local e global. AWS Global Accelerator fornece endereços IP estáticos que atuam como um ponto de entrada fixo para seus endpoints de aplicativos, como seus Application Load Balancers, Network Load Balancers ou EC2 instâncias da Amazon em uma ou várias regiões da AWS. Ele usa a rede global da AWS para otimizar o caminho dos usuários até os aplicativos, melhorando o desempenho, como a latência do tráfego TCP e UDP.

AWS Global Accelerator monitora continuamente a integridade dos endpoints do seu aplicativo e redireciona automaticamente o tráfego para os endpoints íntegros mais próximos no caso de os endpoints atuais ficarem insalubres. Para requisitos adicionais de segurança, o Accelerated Site-

to-Site VPN é usado AWS Global Accelerator para melhorar o desempenho das conexões VPN ao rotear o tráfego de forma inteligente pela Rede Global da AWS e pelos pontos de presença da AWS.




Projeto 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 dos aplicativos RTC depende do armazenamento persistente para armazenar e acessar dados para autenticação, autorização, contabilidade (dados da sessão, registros detalhados de chamadas etc.), monitoramento operacional e registro. Em um data center tradicional, garantir alta disponibilidade e durabilidade dos componentes de armazenamento persistente (bancos de dados, sistemas de arquivos etc.) normalmente exige trabalho pesado por meio da configuração de uma rede de área de armazenamento (SAN), design de matriz redundante de discos independentes (RAID) e processos para backup, restauração e processamento de failover. Nuvem AWS Isso simplifica e aprimora muito as práticas tradicionais de data center em relação à durabilidade e disponibilidade dos dados.

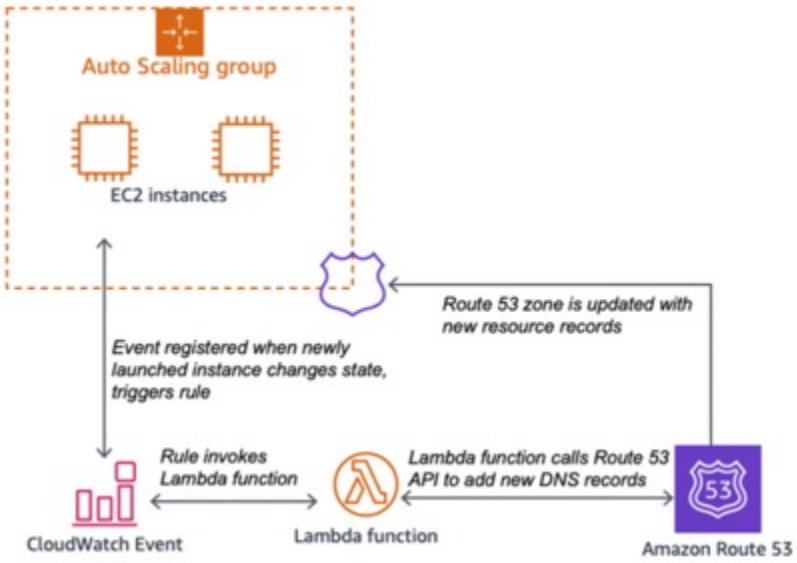
Para armazenamento de objetos e armazenamento de arquivos, AWS serviços como o Amazon Simple Storage Service (Amazon S3) e o Amazon Elastic File System (Amazon EFS) oferecem alta

disponibilidade e escalabilidade gerenciadas. O Amazon S3 tem uma durabilidade de dados de 99,999999999% (11 nove).

Para o 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 de assinante ou armazenamento de registros contábeis (como CDRs), o Amazon RDS fornece uma opção tolerante a falhas, altamente disponível e escalável.

## Escalabilidade dinâmica com AWS Lambda Amazon Route 53 e Amazon EC2 Auto Scaling

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 aplicativos RTC é a combinação de ganchos de ciclo de vida de escalabilidade automática com Amazon Events CloudWatch , Amazon Route 53 e funções. AWS Lambda AWS Lambda as funções podem incorporar qualquer ação ou lógica. A figura a seguir demonstra como esses recursos encadeados podem melhorar a confiabilidade e a escalabilidade do sistema com a automação.




Escalabilidade automática com atualizações dinâmicas para o Amazon Route 53

# WebRTC altamente disponível com Amazon Kinesis Video Streams

O Amazon Kinesis Video Streams oferece streaming de mídia em tempo real via WebRTC, permitindo que os usuários capturem, processem e armazenem streams de mídia para reprodução, análise e aprendizado de máquina. Esses fluxos são altamente disponíveis, escaláveis e compatíveis com os padrões WebRTC. O Amazon Kinesis Video Streams inclui um endpoint de sinalização WebRTC para rápida descoberta de pares e estabelecimento seguro de conexão. Inclui utilitários gerenciados de passagem de sessão para NAT (STUN) e travessia usando relés em torno de pontos finais NAT (TURN) para troca de mídia em tempo real entre pares. Ele também inclui um SDK gratuito de código aberto que se integra diretamente ao firmware da câmera para permitir a comunicação segura com os endpoints do Amazon Kinesis Video Streams, permitindo a descoberta entre pares e o streaming de mídia. Por fim, ele fornece bibliotecas de clientes para Android e iOS e JavaScript que permitem que players móveis e web compatíveis com WebRTC descubram e se conectem com segurança a um dispositivo de câmera para streaming de mídia e comunicação bidirecional.

# Entroncamento SIP de alta disponibilidade com o Amazon Chime Voice Connector

O Amazon Chime Voice Connector fornece um serviço de entroncamento pay-as-you-go SIP que permite que as empresas façam e/ou recebam chamadas telefônicas seguras e econômicas com seus sistemas telefônicos. O Amazon Chime Voice Connector é uma alternativa de baixo custo aos troncos SIP do provedor de serviços ou às interfaces de taxa primária () da Rede Digital de Serviços Integrados (ISDN). PRIs Os clientes têm a opção de ativar chamadas de entrada, chamadas de saída ou ambas.

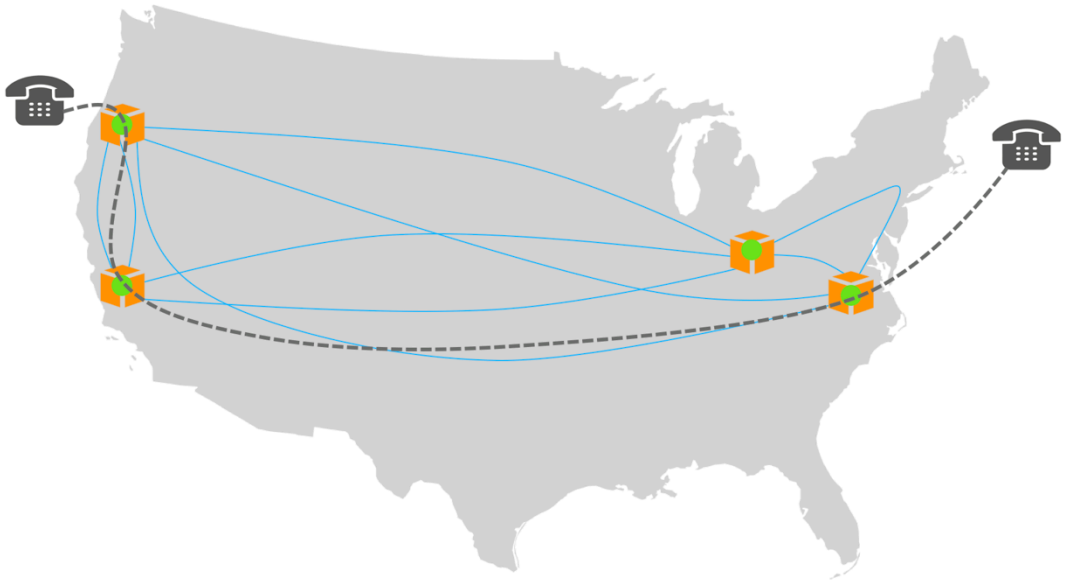
O serviço usa a AWS rede para oferecer uma experiência de chamada altamente disponível em vários Regiões da AWS. Você pode transmitir áudio de chamadas telefônicas de entroncamento SIP ou feeds encaminhados de gravação de mídia baseada em SIP (SIPREC) para o Amazon Kinesis Video Streams para obter informações sobre chamadas comerciais em tempo real. Você pode criar rapidamente aplicativos para análise de áudio por meio da integração com o Amazon Transcribe e outras bibliotecas comuns de aprendizado de máquina.

# Melhores práticas do campo

Esta seção resume as melhores práticas implementadas por alguns dos maiores e mais bem-sucedidos AWS clientes que executam grandes cargas de trabalho do Protocolo de Iniciação de Sessão (SIP) em tempo real. AWS os clientes que desejam executar sua própria infraestrutura SIP na nuvem pública achariam essas melhores práticas valiosas, pois podem ajudar a aumentar a confiabilidade e a resiliência do sistema em caso de diferentes tipos de falhas. Embora algumas dessas melhores práticas sejam específicas do SIP, a maioria delas é aplicável a qualquer aplicativo de comunicação em tempo real executado nele AWS.

## Crie uma sobreposição SIP

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

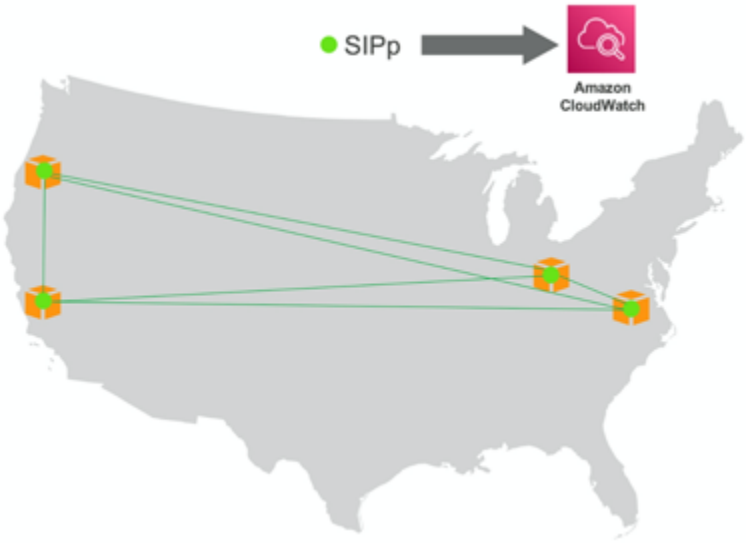


Usando o roteamento SIP para substituir o roteamento de rede

Na figura anterior, a infraestrutura SIP (representada por pontos verdes dentro dos cubos) está funcionando nas quatro regiões dos EUA. As linhas azuis sólidas representam uma representação fictícia da espinha dorsal. AWS Se nenhum roteamento SIP for implementado, uma chamada originada na costa oeste dos EUA e destinada à costa leste dos EUA passará pelo link de backbone que conecta diretamente as regiões de Oregon e Virgínia. O diagrama mostra como um cliente pode ignorar o roteamento no nível da rede e fazer a mesma chamada entre Oregon e 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 comerciais específicas do cliente.

## Execute monitoramento detalhado

Os usuários finais de aplicativos de voz e vídeo em tempo real esperam o mesmo nível de desempenho que obtêm com os serviços de telefonia tradicionais. Então, quando eles enfrentam problemas com um aplicativo, isso acaba prejudicando a reputação do provedor. Para ser proativo em vez de reativo, é imperativo que o monitoramento detalhado seja implantado em todas as partes do sistema que atendem aos usuários finais.



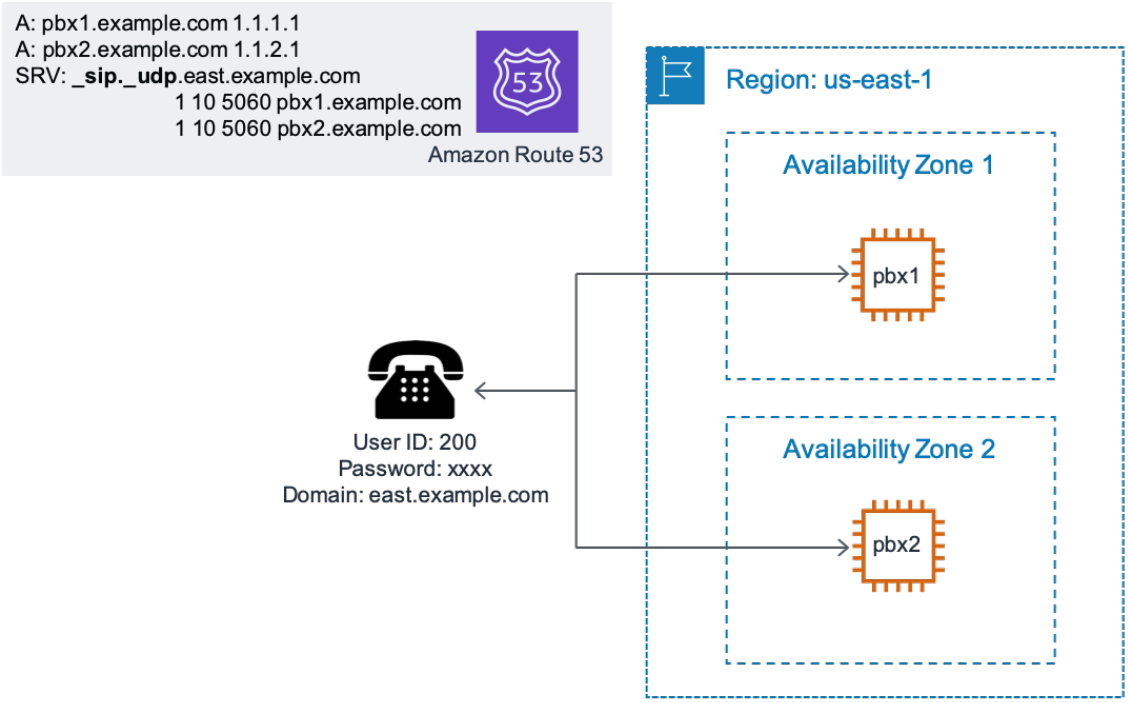

Usando SIPp para monitorar a infraestrutura de VoIP

Muitas ferramentas de código aberto, como iPerf ou e SIPpVOIPMonitor, estão disponíveis para uso no monitoramento do tráfego SIP/RTP. No exemplo anterior, os nós que executam SIP nos modos cliente e servidor estão medindo métricas de SIP, como chamadas bem-sucedidas e retransmissões SIP entre os quatro EUA. Regiões da AWS Essas métricas podem então ser exportadas para a Amazon CloudWatch usando um script personalizado. Usando CloudWatch, os clientes podem criar alarmes sobre essas métricas personalizadas com base em um determinado valor limite. Ações de remediação automática ou manual podem então ser tomadas com base no estado desses CloudWatch alarmes.

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

## Use DNS para balanceamento de carga e IPs flutuação para failover

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




Usando registros DNS SRV para balancear a carga de clientes SIP

A figura anterior mostra como os clientes podem usar os registros SRV para balancear a carga do tráfego SIP. Qualquer cliente de telefonia IP que suporte o padrão SRV procurará o sip. `<transport protocol>`prefixo em um registro DNS do tipo SRV. No exemplo, a seção de resposta do DNS contém as duas em PBXs execução em diferentes zonas de AWS disponibilidade. No entanto, além do endpoint URIs, o registro SRV contém três informações adicionais:

- O primeiro número é a Prioridade (1 no exemplo acima). Uma prioridade mais baixa é preferível à maior.
- O segundo número é o Peso (10 no exemplo acima).
- E o terceiro número é a porta a ser usada (5060).

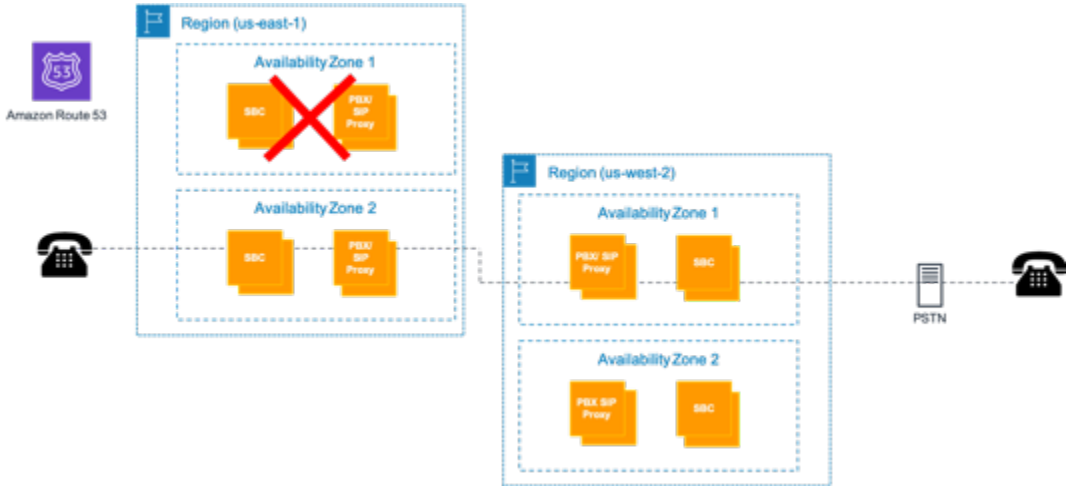
Como a prioridade é a mesma (1) para os dois PBXs servidores, os clientes usam o peso para balancear a carga entre os dois PBXs. Nesse caso, como os pesos são os mesmos, o tráfego SIP deve ter a carga balanceada igualmente entre os dois. PBXs

O DNS pode ser uma boa solução para o balanceamento de carga do cliente, mas que tal implementar o failover alterando/atualizando os registros 'A' do DNS? Esse método é desencorajado devido à inconsistência encontrada no comportamento de cache do DNS nos nós cliente e intermediário. Uma abordagem melhor para o failover intra-AZ entre um cluster de nós SIP é usar a

reatribuição de EC2 IP em que o endereço IP de um host danificado é instantaneamente reatribuído a um host saudável usando a API. EC2 Combinada com uma solução detalhada de monitoramento e verificação de integridade, a reatribuição de IP de um nó com falha garante que o tráfego seja transferido para um host saudável em tempo hábil, minimizando a interrupção do usuário final.

## Use várias zonas de disponibilidade

Cada uma Região da AWS é subdividida em zonas de disponibilidade separadas. Cada zona de disponibilidade tem sua própria conectividade de energia, resfriamento e rede e, portanto, forma um domínio de falha isolado. Dentro das construções do AWS, os clientes são incentivados a executar suas cargas de trabalho em mais de uma zona de disponibilidade. Isso garante que os aplicativos do cliente possam resistir até mesmo a uma falha completa na Zona de Disponibilidade — um evento muito raro por si só. Essa recomendação também significa infraestrutura SIP em tempo real.




Lidando com falhas na zona de disponibilidade

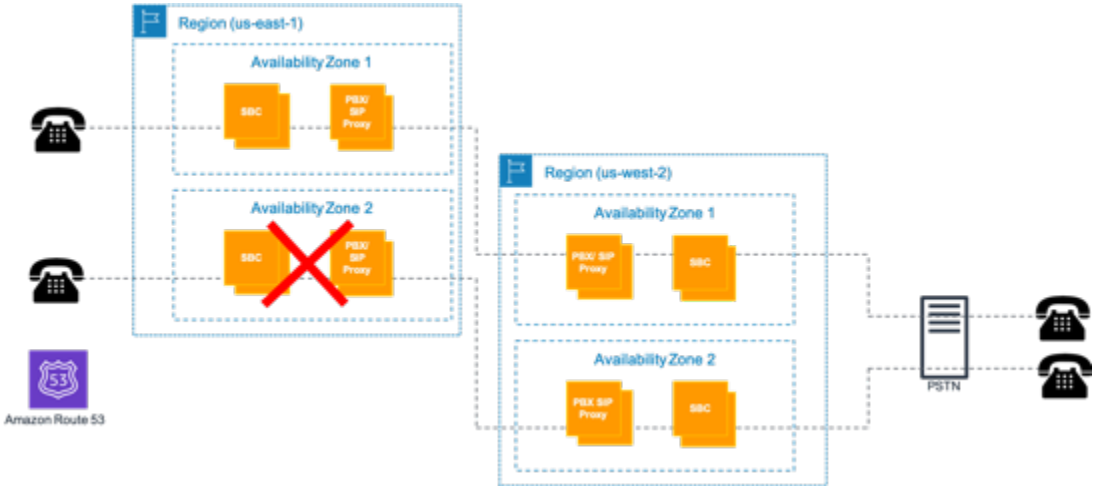
Suponha que um evento catastrófico (como um furacão de categoria cinco) cause uma interrupção completa da Zona de Disponibilidade na região us-east-1. Com a infraestrutura funcionando conforme mostrado no diagrama, todos os clientes SIP que foram originalmente registrados com os nós na Zona de Disponibilidade com 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 garantir que ele seja compatível.) Embora as chamadas SIP ativas no momento da interrupção da Zona de Disponibilidade sejam perdidas, todas as novas chamadas são roteadas pela Zona de Disponibilidade 2.

Resumindo, os registros SRV do DNS devem direcionar o cliente para vários registros 'A', um em cada zona de disponibilidade. Cada um desses registros “A” deve, por sua vez, apontar para

vários endereços IP de SBCs/PBXs nessa zona de disponibilidade, fornecendo resiliência intra e interdisponibilidade. O failover na zona de interdisponibilidade e na zona de interdisponibilidade pode ser implementado usando a reatribuição de IP, se forem públicos. IPs O privado IPs, no entanto, não pode ser reatribuído entre as zonas de disponibilidade. Se um cliente estiver usando endereçamento IP privado, ele precisará contar com o novo registro dos clientes SIP com o SBC/PBX de backup para o failover da zona de interdisponibilidade.

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

Também conhecida como Afinidade da Zona de Disponibilidade, essa prática recomendada também se aplica ao raro evento de uma falha completa na Zona de Disponibilidade. É recomendável eliminar qualquer tráfego entre AZ, de forma que qualquer tráfego SIP ou RTP que entre em uma zona de disponibilidade permaneça nessa zona de disponibilidade até sair da região.




Afinidade com a zona de disponibilidade (no máximo, 50% das chamadas ativas são perdidas)

A figura anterior mostra uma arquitetura simplificada que usa a afinidade da Zona de Disponibilidade. A vantagem comparativa dessa abordagem fica clara se levarmos em conta os efeitos de uma interrupção completa da zona de disponibilidade. Conforme ilustrado no diagrama, se a Zona de Disponibilidade 2 for perdida, 50% das chamadas ativas serão afetadas no máximo (supondo um balanceamento de carga igual entre as Zonas de Disponibilidade). Se o Availability Zone Affinity não tivesse sido implementado, algumas chamadas fluiriam entre as zonas de disponibilidade em uma região e uma falha provavelmente afetaria mais de 50% das chamadas ativas.

Para minimizar a latência do tráfego, a AWS também recomenda que você considere usar grupos de EC2 posicionamento em cada zona de disponibilidade. As instâncias lançadas dentro do mesmo

grupo de EC2 posicionamento têm maior largura de banda e latência reduzida, o EC2 que garante a proximidade entre essas instâncias na rede.

## Use tipos de EC2 instância de rede aprimorados

Escolher o tipo de instância certo na Amazon EC2 garante a confiabilidade do sistema, bem como o uso eficiente da infraestrutura. EC2 fornece uma ampla seleção de tipos de instância otimizados para atender a diferentes casos de uso. Os tipos de instância incluem combinações variadas de capacidade de CPU, memória, armazenamento e redes e oferecem a flexibilidade de escolher a combinação de recursos adequada para suas aplicações. Esses tipos de instância de rede aprimorados garantem que as cargas de trabalho SIP executadas nelas tenham acesso a uma largura de banda consistente e a uma latência agregada comparativamente menor. Uma adição recente à 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 EC2 instância e recursos associados pode ser encontrado na página de tipos de EC2 instância.

Para a maioria dos clientes, a última geração de instâncias otimizadas para computação deve oferecer a melhor relação custo-benefício. Por exemplo, o C5N suporta o novo adaptador de rede elástico com largura de banda de até 100 Gbps com milhões de pacotes por segundo (PPS). A maioria dos aplicativos 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 EC2 instância de acordo com seus requisitos para ver qual tipo de instância funciona melhor para você. O benchmarking 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 do aplicativo RTC geralmente são executados diretamente na Internet, voltados para EC2 instâncias da Amazon. Além do TCP, os fluxos usam protocolos como UDP e SIP. Nesses casos, AWS Shield Standard protege as EC2 instâncias da Amazon contra ataques comuns da camada de infraestrutura (camadas 3 e 4) DDo S, como ataques de reflexão UDP, reflexão de DNS, reflexão de NTP, reflexão de SSDP e assim por diante. AWS Shield Standard usa várias técnicas, como modelagem de tráfego com base em prioridades, que são ativadas automaticamente quando uma assinatura de ataque DDo S bem definida é detectada.

AWS também fornece proteção avançada contra ataques DDo S grandes e sofisticados para esses aplicativos, AWS Shield Advanced ativando endereços IP elásticos. AWS Shield Advanced fornece detecção aprimorada de DDo S que detecta automaticamente o tipo de AWS recurso e o tamanho da EC2 instância e aplica mitigações predefinidas apropriadas com proteções contra inundações de SYN ou UDP. Com AWS Shield Advanced, os clientes também podem criar seus próprios perfis de mitigação personalizados ao engajar a equipe de resposta (DRT) da DDo AWS S, 24 horas por dia, 7 dias por semana. AWS Shield Advanced também garante que, durante um ataque DDo S, todas as suas listas de controle de acesso à rede Amazon VPC (ACLs) sejam aplicadas automaticamente na borda da AWS rede, fornecendo acesso a largura de banda adicional e capacidade de depuração para mitigar ataques S volumétricos de grande porte. DDo

# Conclusão

As cargas de trabalho de comunicação em tempo real (RTC) podem ser implantadas AWS para obter escalabilidade, elasticidade e alta disponibilidade e, ao mesmo tempo, atender aos principais requisitos. Atualmente, vários clientes estão usando a AWS, seus parceiros e soluções de código aberto para executar cargas de trabalho RTC com custo reduzido e agilidade mais rápida, além de uma pegada global reduzida.

As arquiteturas de referência e as melhores práticas fornecidas neste white paper podem ajudar os clientes a configurar com sucesso as cargas de trabalho de RTC AWS e otimizar as soluções para atender aos requisitos do usuário final e, ao mesmo tempo, otimizar para a nuvem.

# Acrônimos

Os acrônimos usados neste documento incluem:

ACL — Lista de controle de acesso

ALB — Application Load Balancer

APNs — Serviço de notificação push da Apple

BGP — Protocolo de gateway de fronteira

CDR — Registros de detalhes de chamadas

COTS — software comercial off-the-shelf

DDoS — distribuído denial-of-service

DNS — Sistema de Nomes de Domínio

DPDK — Kit para desenvolvedores do Intel Data Plane

DRT — Equipe de Resposta DDo S

ENA — Adaptador de rede elástico

EPC — Evolved Packet Core

FCM — Firebase Cloud Messaging

HA — Alta disponibilidade

IRC — Internet Relay Chat

ISDN — Rede digital de serviços integrados

NAT — tradução de endereços de rede

OPUS — suporte ao usuário de posicionamento online

PBX — Central de Câmbio Privado

PRI — Interface de taxa primária

PSTN — Rede telefônica pública comutada

RAID — matriz redundante de discos independentes

RTC — comunicação em tempo real

RTP — Protocolo de transporte em tempo real

SAN — Rede de área de armazenamento

SBC — controlador de fronteira de sessão

SIP — Protocolo de iniciação de sessão

SPOF — pontos únicos de falha

SRV — Serviço

SS7 — Sistema de Sinalização n.7

STUN — Utilitários de passagem de sessão para NAT

SYN — Sincronizar

TCP — Protocolo de Controle de Transmissão

TDM — multiplexação por divisão de tempo

TURN — Travessia usando relés em torno do NAT

UDP — Protocolo de datagrama de usuário

URI — Identificadores uniformes de recursos

VIP — IP virtual

VNF — Função de rede virtual

VoIP — Voz sobre IP

VPC — Nuvem privada virtual

WebRTC — comunicação web em tempo real

# Colaboradores

Os seguintes indivíduos e organizações contribuíram para este documento:

- Mounir Chennana, arquiteto sênior de soluções, Amazon Web Services
- Mohammed Al-Mehdar, arquiteto sênior de soluções, Amazon Web Services
- Ejaz Sial, arquiteto sênior de soluções, Amazon Web Services
- Ahmad Khan, arquiteto sênior de soluções, Amazon Web Services
- Tipu Qureshi, engenheira principal da Amazon Web AWS Support Services
- Hasan Khan, gerente técnico sênior de contas, Amazon Web Services
- Shoma Chakravarty, líder técnica da WW, Telecom, Amazon Web Services

# Revisões do documento

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

| Alteração | Descrição | Data |
| --- | --- | --- |
| Whitepaper atualizado | Atualizado para os serviços e recursos mais recentes. | 05 de maio de 2022 |
| Whitepaper atualizado | Atualizado para os serviços e recursos mais recentes. | 13 de fevereiro de 2020 |
| Publicação inicial | Whitepaper publicado pela primeira vez. | 1º de outubro de 2018 |

# Avisos

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

© 2022 Amazon Web Services, Inc. ou suas afiliadas. Todos os direitos reservados.

# AWS Glossário

Para obter a AWS terminologia mais recente, consulte o AWS glossário na Glossário da AWS Referência.

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.