Documento técnico de AWS

Comunicación en tiempo real en AWS

Comunicación en tiempo real en AWS: Documento técnico de AWS

Table of Contents

Comunicación en tiempo real en AWS

Prácticas recomendadas para el diseño de cargas de trabajo de comunicación en tiempo real (RTC) de alta disponibilidad y escalabilidad en AWS

Fecha de publicación: 13 de febrero de 2020 (Revisiones del documento)

Resumen

En la actualidad, muchas organizaciones desean reducir los costes y conseguir escalabilidad para las cargas de trabajo de voz, mensajería y multimedia en tiempo real. En este documento, se describen las prácticas recomendadas para administrar cargas de trabajo de comunicación en tiempo real en AWS y se incluyen arquitecturas de referencia para cumplir estos requisitos. Este documento sirve de guía para aquellas personas que estén familiarizadas con la comunicación en tiempo real sobre la forma de lograr una alta disponibilidad y escalabilidad para estas cargas de trabajo.

Introducción

Las aplicaciones de telecomunicaciones que utilizan voz, vídeo y mensajería como canales son un requisito clave para muchas organizaciones y sus usuarios finales. Estas cargas de trabajo de comunicación en tiempo real (RTC) tienen requisitos específicos de latencia y disponibilidad que se pueden cumplir siguiendo las prácticas recomendadas de diseño relevantes. En el pasado, las cargas de trabajo de RTC se implementaban en centros de datos locales tradicionales con recursos dedicados.

Sin embargo, gracias a un conjunto de características maduro y cada vez mayor, las cargas de trabajo de RTC se pueden implementar en Amazon Web Services (AWS) a pesar de los estrictos requisitos de nivel de servicio, al mismo tiempo que se benefician de escalabilidad, elasticidad y alta disponibilidad. En la actualidad, varios clientes utilizan AWS, a sus socios y soluciones de código abierto para ejecutar cargas de trabajo de RTC a un coste menor, con más agilidad, con la capacidad de globalizarse en minutos y con las características completas de los servicios de AWS.

Los clientes utilizan características de AWS, como las redes mejoradas con un Elastic Network Adapter (ENA) y la última generación de instancias de Amazon Elastic Compute Cloud (EC2) para beneficiarse del kit de desarrollo de plano de datos (DPDK), la virtualización de E/S de raíz única (SR- IOV), páginas enormes, NVM Express (NVMe), la compatibilidad con el acceso a memoria no uniforme (NUMA), así como instancias bare metal para cumplir los requisitos de carga de trabajo de RTC. Estas instancias ofrecen un ancho de banda de red de hasta 100 Gbps y paquetes proporcionales por segundo, lo que ofrece un rendimiento mayor para las aplicaciones que utilizan la red de forma intensiva. Para escalar, Elastic Load Balancing ofrece Application Load Balancer, que dispone de soporte para WebSocket, y Network Load Balancer, que puede gestionar millones de solicitudes por segundo. Para la aceleración de la red, AWS Global Accelerator proporciona direcciones IP estáticas que actúan como punto de entrada fijo a los puntos de conexión de su aplicación en AWS. Admite direcciones IP estáticas para el equilibrador de carga. Para reducir la latencia, el coste y aumentar el rendimiento del ancho de banda, AWS Direct Connect establece una conexión de red dedicada desde las instalaciones locales a AWS. El conector de voz de Amazon Chime proporciona troncales SIP administrados de alta disponibilidad. Amazon Kinesis Video Streams con WebRTC transmite fácilmente contenido multimedia de manera bidireccional en tiempo real con alta disponibilidad.

Este documento incluye arquitecturas de referencia que muestran cómo configurar cargas de trabajo de RTC en AWS y prácticas recomendadas para optimizar las soluciones para cumplir los requisitos del usuario final y, al mismo tiempo, optimizarlas para la nube. El núcleo de paquetes evolucionado

(EPC) está fuera del alcance de este documento técnico, pero las prácticas recomendadas detalladas se pueden aplicar a las funciones de red virtual (VNF).

Componentes fundamentales de la arquitectura RTC

En el sector de las telecomunicaciones, la comunicación en tiempo real (RTC) suele referirse a sesiones de contenido multimedia en directo entre dos puntos de conexión con una latencia mínima. Estas sesiones podrían estar relacionadas con:

- Una sesión de voz entre dos partes (por ejemplo, sistema telefónico, móvil, VoIP)
- Mensajería instantánea (por ejemplo, chat, IRC)
- Sesión de vídeo en directo (por ejemplo, videoconferencia, telepresencia)

Cada una de las soluciones anteriores tiene algunos componentes en común (por ejemplo, componentes que proporcionan autenticación, autorización y control de acceso, transcodificación, almacenamiento en búfer y retransmisión, etc.) y algunos componentes únicos para el tipo de contenido multimedia transmitido (por ejemplo, servicio de difusión, servidor de mensajería y colas, etc.). Esta sección se centra en la definición de un sistema RTC basado en voz y vídeo y todos los componentes relacionados que se ilustran en la figura 1.

Figura 1: Componentes arquitectónicos esenciales para RTC

Temas

SoftSwitch/PBX

Un softswitch o PBX es el cerebro de un sistema telefónico de voz y proporciona la inteligencia para establecer, mantener y enrutar una llamada de voz dentro o fuera de la empresa mediante el uso de diferentes componentes. Todos los suscriptores de la empresa deben registrarse en el softswitch para recibir o realizar una llamada. Una funcionalidad importante del softswitch es realizar un seguimiento de cada suscriptor y descubrir la forma de llegar a ellos mediante el uso de los otros componentes de la red de voz.

Controlador de borde de sesión (SBC)

Un controlador de borde de sesión (SBC) se encuentra en el borde de una red de voz y realiza un seguimiento de todo el tráfico entrante y saliente (tanto en los planos de control como de datos). Una de las principales responsabilidades de un SBC es proteger el sistema de voz de cualquier uso malintencionado. El SBC se puede utilizar para interconectarse con los troncales del protocolo de inicio de sesión (SIP) para conseguir conectividad externa. Algunos SBC también poseen capacidades de transcodificación para convertir CODECS de un formato a otro. Por último, la mayoría de los SBC también disponen de capacidades de recorrido de NAT que ayudan a garantizar que las llamadas se establezcan, incluso a través de redes con firewall.

Conectividad PSTN

Las soluciones de voz sobre IP (VoIP) utilizan puertas de enlace PSTN y troncales SIP para conectarse con redes PSTN heredadas.

Puerta de enlace PSTN

La puerta de enlace de la red telefónica pública conmutada (PSTN) convierte la señalización (entre SIP y SS7) y el contenido multimedia (entre RTP y la multiplexación por división de tiempo [TDM] mediante transcodificación CODEC). Las puertas de enlace PSTN siempre se encuentran en el borde, cerca de la red PSTN.

Troncal SIP

En un troncal SIP, la empresa no termina sus llamadas en una red TDM (basada en SS7), sino que los flujos entre la empresa y la compañía de telecomunicaciones se siguen realizando a través de IP. La mayoría de los troncales SIP se establecen mediante SBC. La empresa debe acordar las reglas de seguridad predefinidas de las telecomunicaciones, como permitir un cierto rango de direcciones IP, puertos, etc.

Puerta de enlace multimedia (transcodificador)

Una solución de voz típica permite varios tipos de códecs. Entre los códecs más comunes se encuentran G.711 μ-law para Norteamérica, G.711 A-law para fuera de Norteamérica, G.729 y G.722. Cuando dos dispositivos que utilizan dos códecs diferentes se comunican entre sí, un servidor multimedia traduce el flujo de códecs entre los dispositivos. En otras palabras, una puerta de

enlace multimedia procesa el contenido multimedia y garantiza que los dispositivos finales puedan comunicarse entre sí.

WebRTC y puerta de enlace WebRTC

La comunicación web en tiempo real (WebRTC) permite establecer una llamada desde un navegador web o solicitar recursos del servidor backend mediante la API. La tecnología está diseñada teniendo en cuenta la tecnología de la nube y, por lo tanto, dispone de varias API que podrían usarse para establecer una llamada. Dado que no todas las soluciones de voz (incluido SIP) admiten estas API, la puerta de enlace WebRTC es necesaria para traducir las llamadas a la API en mensajes SIP y viceversa.

La figura 2 muestra un patrón de diseño para una arquitectura WebRTC de alta disponibilidad. El tráfico entrante desde los clientes de WebRTC se equilibra mediante un Application Load Balancer de Amazon con WebRTC ejecutándose en instancias de EC2 que forman parte de un grupo de Auto Scaling.

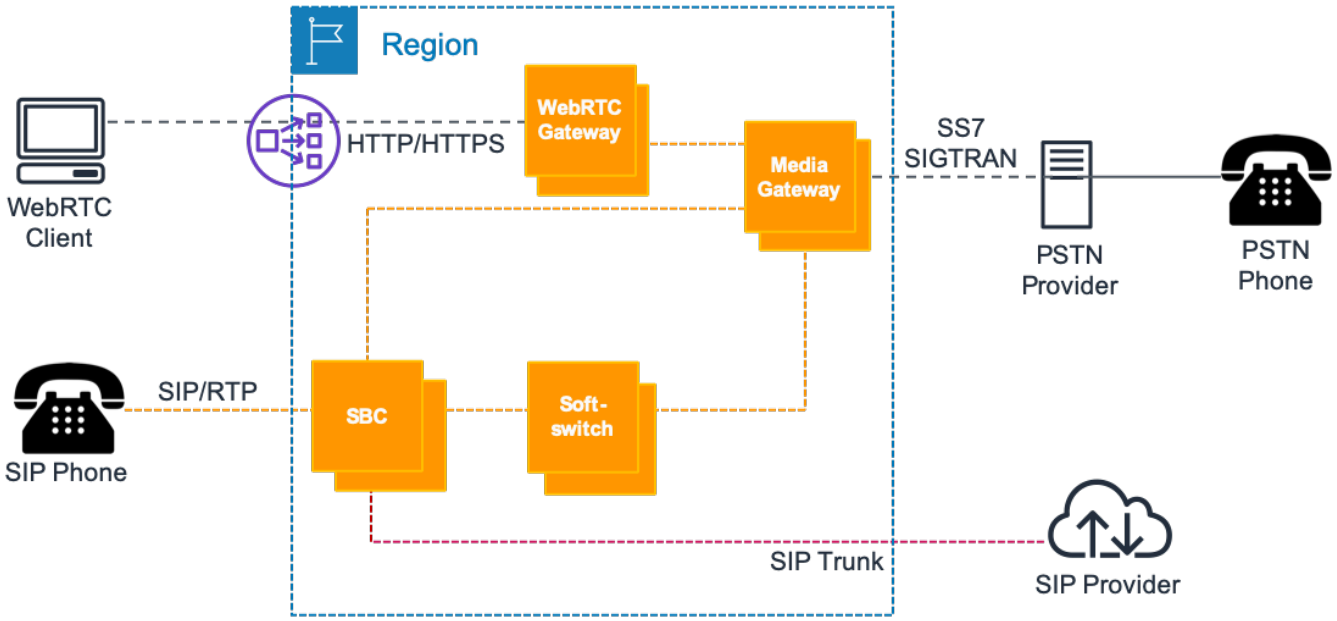

Figura 2: Topología básica de un sistema RTC para voz

Otro patrón de diseño para el tráfico SIP y RTP consiste en utilizar pares de SBC en Amazon EC2 en modo activo pasivo en todas las zonas de disponibilidad (Figura 3). En este caso, una dirección IP elástica se puede mover dinámicamente entre instancias si hay un error que impide usar DNS.

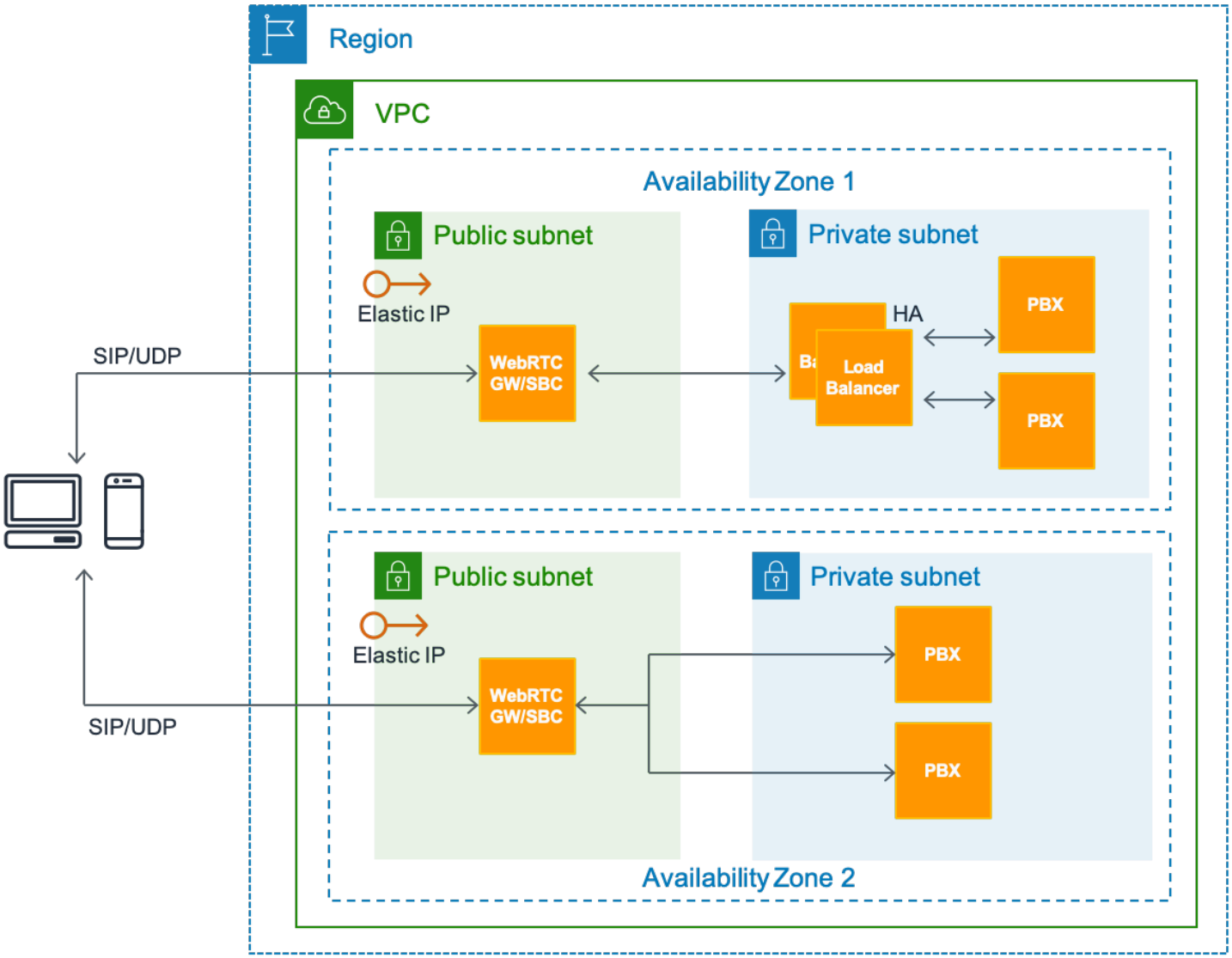

Figura 3: Arquitectura RTC con Amazon EC2 en una VPC

Alta disponibilidad y escalabilidad en AWS

La mayoría de los proveedores de comunicaciones en tiempo real tienen unos niveles de servicio que ofrecen una disponibilidad de entre el 99,9 % y el 99,999 %. Según el grado de alta disponibilidad (HA) que desee, debe tomar medidas cada vez más sofisticadas a lo largo del ciclo de vida completo de la aplicación. Recomendamos seguir estas directrices para conseguir un buen nivel de alta disponibilidad:

- Diseñe el sistema para que no tenga un único punto de error. Utilice mecanismos automatizados de supervisión, detección de errores y conmutación por error para componentes sin estado y con estado

- Los únicos puntos de error (SPOF) suelen eliminarse con una configuración de redundancia N+1 o 2N, en la que N+1 se logra con un equilibrio de carga entre los nodos activo-activo, y 2N se logra mediante un par de nodos en una configuración activa-en espera.
- AWS tiene varios métodos para lograr la alta disponibilidad a través de ambos enfoques, por ejemplo, mediante un clúster escalable y con equilibrio de carga o asumiendo un par activo-en espera.

- Instrumente y pruebe la disponibilidad del sistema correctamente.
- Prepare los procedimientos operativos para que los mecanismos manuales respondan al error, lo mitiguen y se recuperen de él.

Esta sección se centra en cómo conseguir que no haya un único punto de error con las capacidades disponibles en AWS. Específicamente, en esta sección se describe un subconjunto de capacidades y patrones de diseño principales de AWS que permiten crear aplicaciones de comunicación en tiempo real de alta disponibilidad en la plataforma.

Temas

- WebRTC de alta disponibilidad con Kinesis Video Streams
- Troncal SIP de alta disponibilidad con conector de voz Amazon Chime

Patrón de IP flotante para alta disponibilidad entre servidores con estado activos-en espera

El patrón de diseño de IP flotante es un mecanismo bien conocido para lograr una conmutación por error automática entre un par de nodos de hardware activos y en espera (servidores multimedia). Se asigna una dirección IP virtual secundaria estática al nodo activo. La supervisión continua entre los nodos activos y en espera detecta errores. Si el nodo activo falla, el script de supervisión asigna la IP virtual al nodo en espera preparado y el nodo en espera asume la función activa principal. De esta manera, la IP virtual flota entre el nodo activo y en espera.

Temas

- Aplicabilidad en soluciones RTC
- Implementación en AWS
- Beneficios
- Limitaciones y extensibilidad

Aplicabilidad en soluciones RTC

No siempre es posible tener en servicio varias instancias activas del mismo componente, como un clúster activo-activo de nodos N. Una configuración activa-en espera proporciona el mejor mecanismo para la alta disponibilidad. Por ejemplo, los componentes con estado de una solución RTC, como el servidor multimedia o el servidor de conferencias, o incluso un servidor SBC o de base de datos, son adecuados para una configuración activa-en espera. Un servidor multimedia o SBC tiene varias sesiones o canales de larga ejecución activos en un momento determinado y, en el caso de que la instancia activa de SBC falle, los puntos de conexión pueden volver a conectarse al nodo en espera sin ninguna configuración en el lado del cliente debido a la IP flotante.

Implementación en AWS

Puede implementar este patrón en AWS utilizando las capacidades principales de Amazon Elastic Compute Cloud (Amazon EC2), la API de Amazon EC2, las direcciones de IP elástica y la compatibilidad con Amazon EC2 para direcciones IP privadas secundarias.

1. Lance dos instancias de EC2 para que asuman los roles de nodos principales y secundarios, donde se supone que el principal está en estado activo de forma predeterminada.
2. Asigne una dirección IP privada secundaria adicional a la instancia de EC2 principal.
3. Hay asociada una dirección IP elástica, que es similar a una IP virtual (VIP), a la dirección privada secundaria. Esta dirección privada secundaria es la dirección que utilizan los puntos de conexión externos para acceder a la aplicación.
4. Se necesita una configuración del sistema operativo determinada para que la dirección IP secundaria se añada como un alias a la interfaz de red principal.
5. La aplicación debe vincularse a esta dirección IP elástica. En el caso del software Asterisk, puede configurar el enlace a través de la configuración avanzada de SIP de Asterisk.
6. Ejecute un script de supervisión (personalizado, KeepAlive en Linux, Corosync, etc.) en cada nodo para supervisar el estado del nodo del mismo nivel. En el caso de que el nodo activo actual falle, el nodo del mismo nivel detecta este error e invoca a la API de Amazon EC2 para reasignar la dirección IP privada secundaria a sí mismo.
7. Por lo tanto, la aplicación que estaba escuchando en el VIP asociado a la dirección IP privada secundaria está disponible para los puntos de conexión a través del nodo en espera.

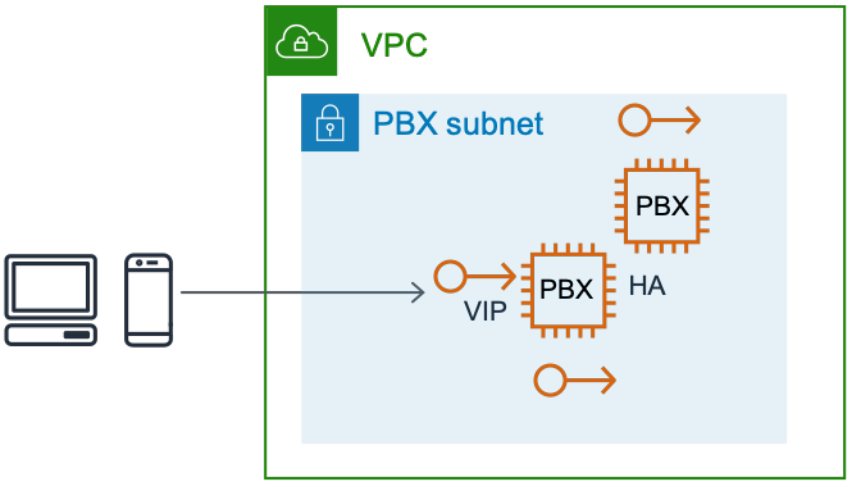

Figura 4: Conmutación por error entre instancias de EC2 con estado mediante la dirección IP elástica

Beneficios

Este enfoque es una solución fiable de bajo presupuesto que protege contra errores en el nivel de instancia de EC2, infraestructura o aplicación.

Limitaciones y extensibilidad

Este patrón de diseño suele limitarse a una sola zona de disponibilidad. Se puede implementar en dos zonas de disponibilidad, pero con una variación. En este caso, la dirección IP elástica flotante se vuelve a asociar entre el nodo activo y en espera en diferentes zonas de disponibilidad a través de la API de dirección IP elástica de reasociación disponible. En la implementación de la conmutación por error que se muestra en la figura 4, las llamadas en curso se interrumpen y los puntos de conexión deben volver a conectarse. Es posible ampliar esta implementación con la replicación de los datos de sesión subyacentes para proporcionar también una conmutación por error perfecta de continuidad de las sesiones o multimedia.

Equilibrio de carga para escalabilidad y alta disponibilidad con WebRTC y SIP

Equilibrar la carga de un clúster de instancias activas en función de reglas predefinidas (como los turnos rotativos, la afinidad o la latencia, etc.) es un patrón de diseño muy popularizado por la naturaleza sin estado de las solicitudes HTTP. De hecho, el equilibrio de carga es una opción viable si hay muchos componentes de aplicaciones RTC.

El equilibrador de carga actúa como un proxy inverso o punto de entrada para realizar solicitudes a la aplicación deseada, que a su vez está configurada para ejecutarse en varios nodos activos simultáneamente. En un momento dado, el equilibrador de carga dirige la solicitud de un usuario a uno de los nodos activos del clúster definido. Los equilibradores de carga realizan una comprobación de estado en los nodos de su clúster de destino y no envían una solicitud entrante a un nodo que no supere esta comprobación. Por lo tanto, gracias al equilibrio de carga, se logra un nivel fundamental de alta disponibilidad. Además, dado que un equilibrador de carga realiza comprobaciones de estado activas y pasivas en todos los nodos del clúster a intervalos de menos de un segundo, el tiempo de conmutación por error es casi instantáneo.

La decisión sobre a qué nodo debe dirigirse se basa en las reglas del sistema que están definidas en el equilibrador de carga, que incluyen:

- Turno rotativo
- Afinidad de sesión o IP, que garantiza que varias solicitudes dentro de una sesión o desde la misma IP se envíen al mismo nodo del clúster
- Basada en latencia

- Basada en carga

Temas

Aplicabilidad en arquitecturas RTC

El protocolo WebRTC hace posible que las puertas de enlace de WebRTC se equilibren fácilmente mediante un equilibrador de carga basado en HTTP, como Elastic Load Balancing, Application Load Balancer o Network Load Balancer. Dado que la mayoría de las implementaciones de SIP se basan en el transporte tanto a través de TCP como de UDP, se necesita un equilibrio de carga a nivel de red o conexión que admita tráfico basado en TCP y UDP.

Equilibrio de carga en AWS para WebRTC mediante Application Load Balancer y Auto Scaling

En el caso de las comunicaciones basadas en WebRTC, Elastic Load Balancing proporciona un equilibrador de carga escalable, de alta disponibilidad y completamente administrado que sirve de punto de entrada para las solicitudes, que luego se dirigen a un clúster de destino de instancias de EC2 asociadas con Elastic Load Balancing. Además, dado que las solicitudes de WebRTC no tienen estado, puede utilizar Amazon EC2 Auto Scaling para proporcionar escalabilidad y elasticidad totalmente automatizadas y controlables, además de alta disponibilidad.

Application Load Balancer proporciona un servicio de equilibrio de carga totalmente administrado que tiene alta disponibilidad mediante múltiples zonas de disponibilidad y es escalable. Esto admite el equilibrio de carga de las solicitudes de WebSocket que gestionan la señalización de las aplicaciones WebRTC y la comunicación bidireccional entre el cliente y el servidor mediante una conexión TCP de larga duración. Application Load Balancer también admite el enrutamiento basado en contenido y sesiones permanentes, además de enrutar las solicitudes del mismo cliente al mismo destino mediante cookies generadas por el equilibrador de carga. Si habilita las sesiones permanentes, el mismo destino recibe la solicitud y puede utilizar la cookie para recuperar el contexto de la sesión.

La figura 5 muestra la topología del destino.

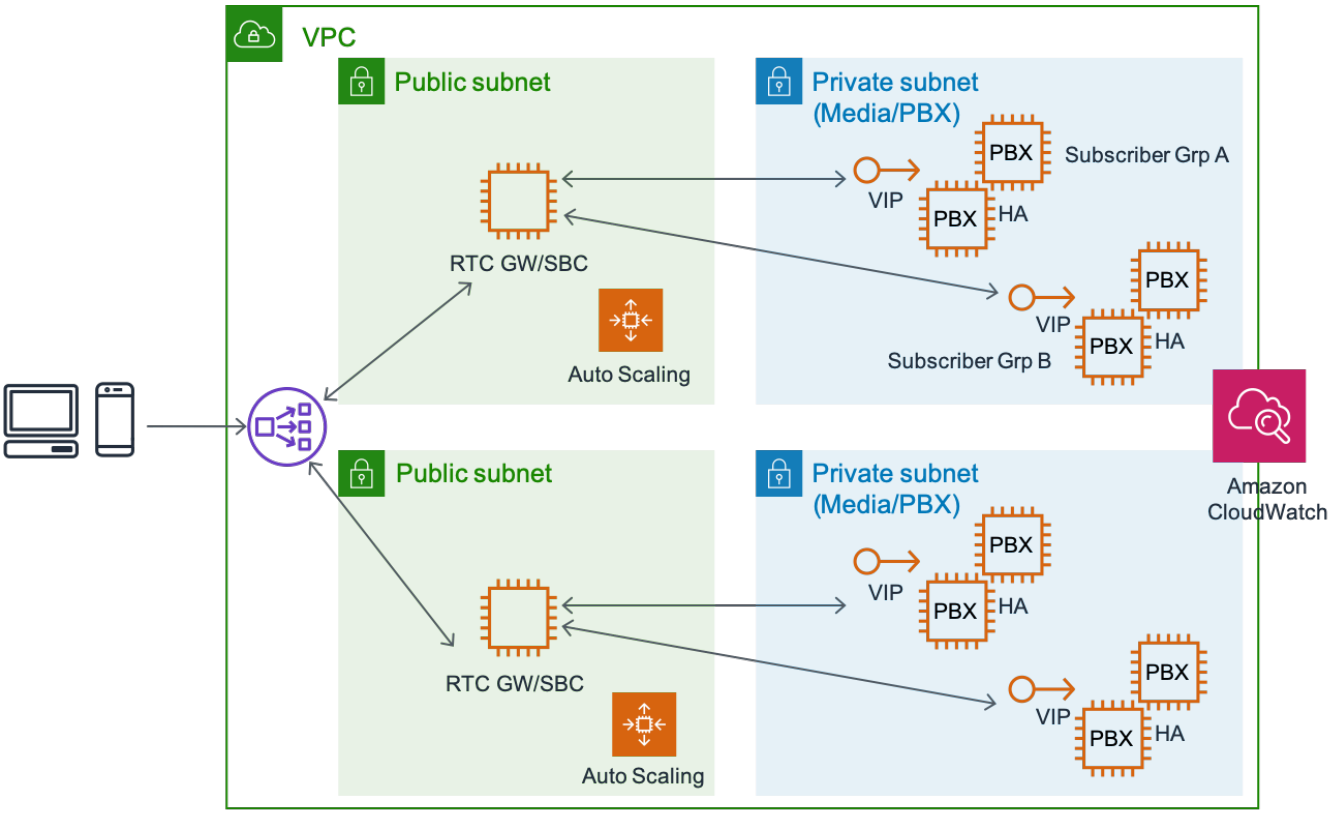

Figura 5: Escalabilidad de WebRTC y arquitectura de alta disponibilidad

Implementación para SIP mediante Network Load Balancer o el producto AWS Marketplace

En el caso de comunicaciones basadas en SIP, las conexiones se realizan a través de TCP o UDP, y la mayoría de las aplicaciones RTC utilizan UDP. Si SIP/TCP es el protocolo de señal preferido, entonces es factible utilizar Network Load Balancer para conseguir un equilibrio de carga de rendimiento, escalable, de alta disponibilidad y totalmente administrado.

Network Load Balancer funciona en el nivel de conexión (capa 4) y enruta las conexiones a destinos como instancias de Amazon EC2, contenedores y direcciones IP en función de los datos del protocolo IP. El equilibrador de carga de red, que es ideal para equilibrar la carga de tráfico TCP o UDP, puede ocuparse de millones de solicitudes por segundo y mantener latencias ultrabajas. Se integra con otros servicios populares de AWS, como AWS Auto Scaling Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) y AWS CloudFormation.

Si se inician conexiones SIP, otra opción es usar el software comercial listo para usar (COTS) AWS Marketplace. AWS Marketplace dispone de muchos productos que pueden gestionar UDP y otros

tipos de equilibrios de carga de conexión de capa 4. Estos COTS suelen incluir soporte para alta disponibilidad y, por lo general, se integran con características, tales como AWS Auto Scaling, para mejorar aún más la disponibilidad y la escalabilidad. La figura 6 muestra la topología del destino:

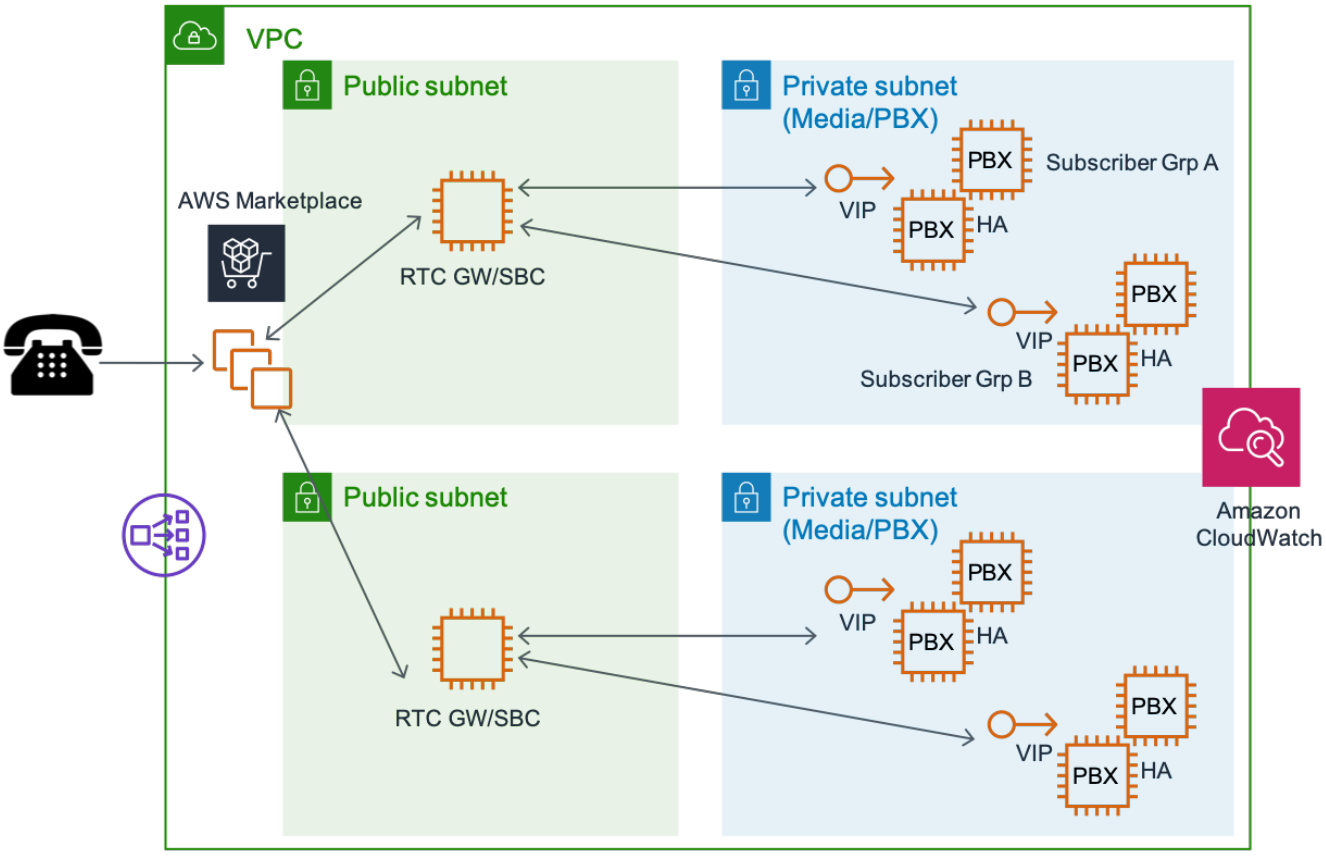

Figura 6: Escalabilidad de RTC basada en SIP con el producto AWS Marketplace

Equilibrio de carga y conmutación por error basados en DNS entre regiones

Amazon Route 53 ofrece un servicio de DNS global que se puede utilizar como punto de conexión público o privado para que los clientes de RTC se registren y se conecten con aplicaciones multimedia. Con Amazon Route 53, se pueden configurar comprobaciones de estado de DNS para enrutar el tráfico a puntos de conexión en buen estado o para supervisar de forma independiente el estado de su aplicación. La característica Amazon Route 53 Traffic Flow simplifica la administración del tráfico de manera global a través de varios tipos de enrutamiento, incluido el enrutamiento basado en la latencia, el DNS geográfico, la geoproximidad y el turno rotativo ponderado. Todos ellos se pueden combinar con la conmutación por error de DNS para permitir diversas arquitecturas de baja latencia y tolerantes a errores. Mediante el sencillo editor visual de Amazon Route 53 Traffic

Flow, puede administrar fácilmente el modo en que se redirige a los usuarios finales a los puntos de conexión de la aplicación, tanto si están en una sola región de AWS como si se encuentran distribuidos por todo el mundo.

En el caso de implementaciones globales, la política de enrutamiento basada en la latencia de Route 53 es especialmente útil para dirigir a los clientes al punto de presencia más cercano para que un servidor multimedia mejore la calidad del servicio asociado con los intercambios de contenido multimedia en tiempo real.

Tenga en cuenta que, para aplicar una conmutación por error a una nueva dirección DNS, las cachés de los clientes deben vaciarse. Además, los cambios de DNS podrían tener un retardo cuando se propagan a través de los servidores DNS globales. Puede administrar el intervalo de actualización para las búsquedas de DNS con el atributo de período de vida. Este atributo se puede configurar en el momento de configurar las políticas de DNS.

Para llegar rápidamente a los usuarios globales o para cumplir los requisitos de usar una única IP pública, también se puede utilizar AWS Global Accelerator para la conmutación por error entre regiones. AWS Global Accelerator es un servicio de red que mejora la disponibilidad y el rendimiento de las aplicaciones con alcance local y global. AWS Global Accelerator proporciona direcciones IP estáticas que actúan como punto de entrada fijo a los puntos de conexión de la aplicación, como los Application Load Balancer, los Network Load Balancer o las instancias de Amazon EC2 en una o varias regiones de AWS. Utiliza la red global de AWS para optimizar la ruta de los usuarios a las aplicaciones, lo que mejora el rendimiento, como la latencia del tráfico TCP y UDP. AWS Global Accelerator supervisa continuamente el estado de los puntos de conexión de su aplicación y redirige automáticamente el tráfico a los puntos de conexión en buen estado más cercanos en caso de que los puntos de conexión actuales no funcionen correctamente. Para los requisitos de seguridad adicionales, Site-to-Site VPN acelerada utiliza AWS Global Accelerator para mejorar el rendimiento de las conexiones de VPN mediante el enrutamiento inteligente del tráfico a través de la red global de AWS y las ubicaciones de borde de AWS.

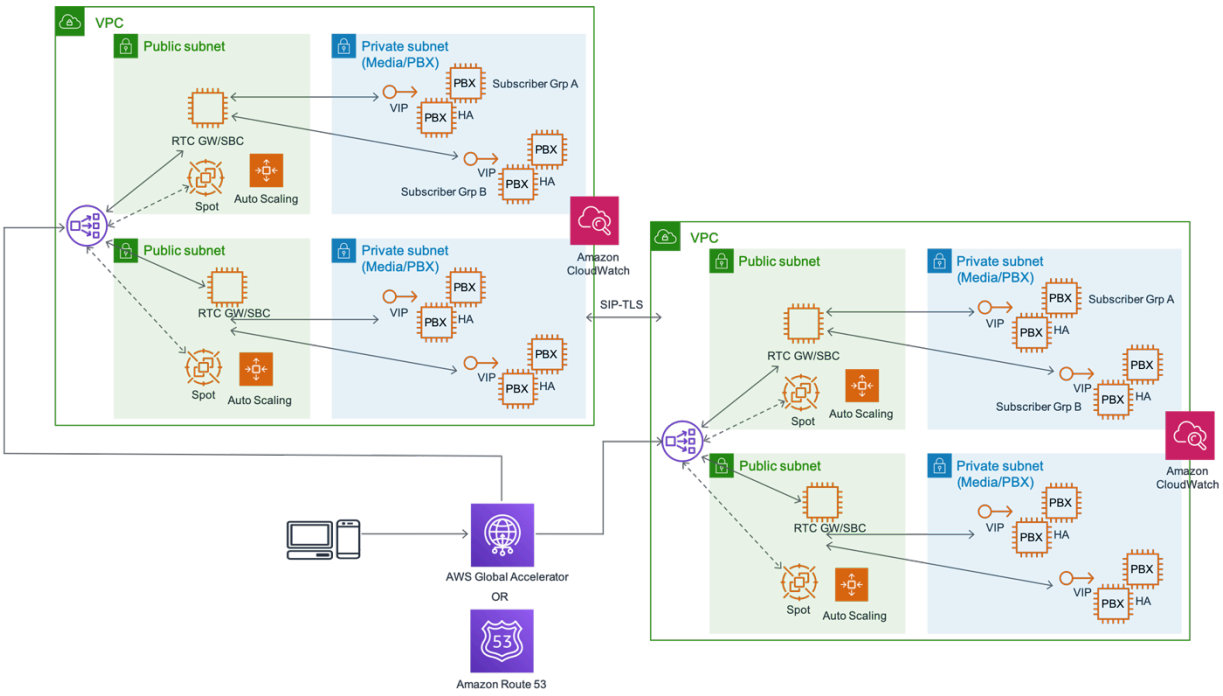

Figura 7: Diseño de alta disponibilidad interregional con AWS Global Accelerator o Amazon Route 53

Durabilidad de los datos y alta disponibilidad con almacenamiento persistente

La mayoría de las aplicaciones RTC utilizan el almacenamiento persistente para almacenar los datos y acceder a ellos para la autenticación, la autorización, la contabilidad (datos de sesión, registros de detalles de llamadas, etc.), la supervisión operativa y el registro. En un centro de datos tradicional, para garantizar una alta disponibilidad y durabilidad de los componentes de almacenamiento persistente (bases de datos, sistemas de archivos, etc.) normalmente se requiere un transporte pesado mediante la configuración de un diseño RAID de SAN y procesos para el procesamiento de copia de seguridad, restauración y conmutación por error. La nube de AWS simplifica y mejora considerablemente las prácticas tradicionales de los centros de datos relacionados con la durabilidad y disponibilidad de los datos.

Para el almacenamiento de objetos y archivos, los servicios de AWS como Amazon Simple Storage Service (Amazon S3) y Amazon Elastic File System (Amazon EFS) disponen de alta disponibilidad y escalabilidad administradas. Amazon S3 tiene una durabilidad de datos de 11 nueves.

Para el almacenamiento de datos transaccionales, los clientes tienen la opción de utilizar Amazon Relational Database Service (Amazon RDS) completamente administrado, que admite Amazon

Aurora, PostgreSQL, MySQL, MariaDB, Oracle y Microsoft SQL Server con implementaciones de alta disponibilidad. Para la función de registrador, el perfil de suscriptor o el almacenamiento de registros contables (por ejemplo, CDR), Amazon RDS dispone de una opción tolerante a errores, de alta disponibilidad y escalable.

Escalado dinámico con AWS LambdaAmazon Route 53 y AWS Auto Scaling

AWS permite el encadenamiento de características y ofrece la capacidad de incorporar funciones como servicio sin servidor personalizadas basadas en eventos de infraestructura. Uno de esos patrones de diseño que tiene numerosos usos versátiles en las aplicaciones RTC es la combinación de enlaces de ciclo de vida de escalado automático con Amazon CloudWatch Events, Amazon Route 53 y funciones AWS Lambda. Las funciones AWS Lambda pueden integrar cualquier acción o lógica. La figura 8 demuestra cómo estas características encadenadas pueden mejorar la fiabilidad y la escalabilidad del sistema con automatización.

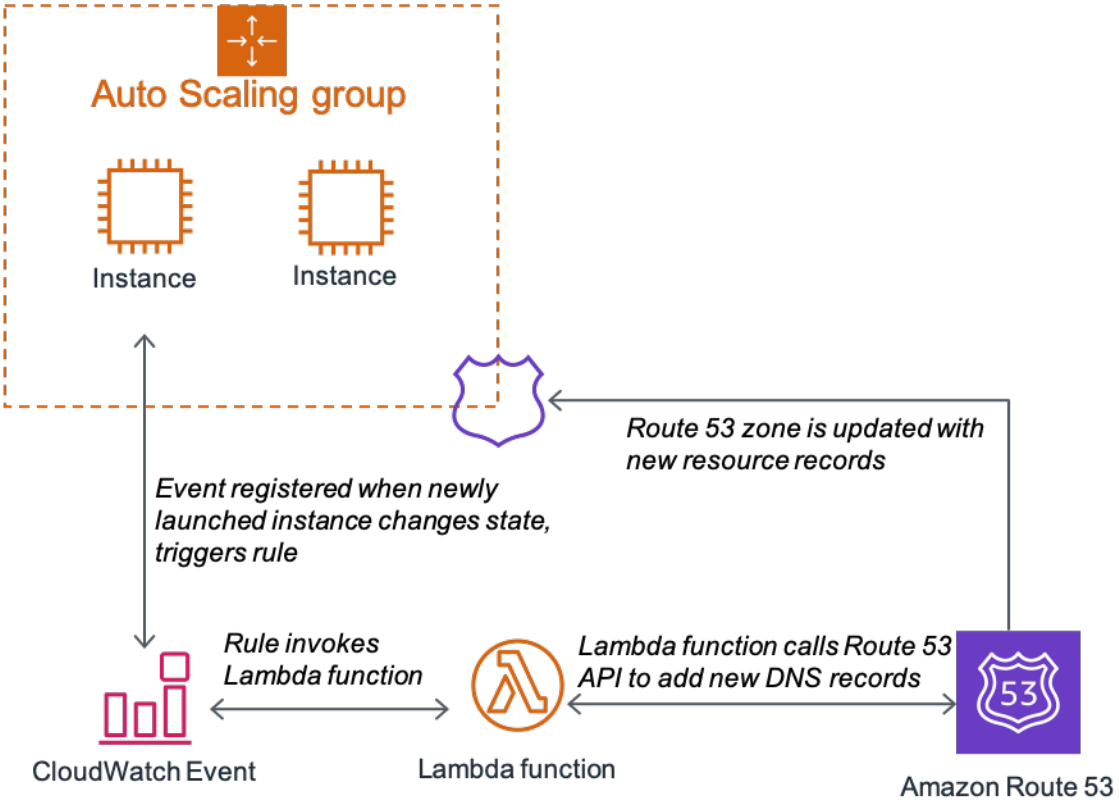

Figura 8: Escalado automático con actualizaciones dinámicas de Amazon Route 53

WebRTC de alta disponibilidad con Kinesis Video Streams

Amazon Kinesis Video Streams ofrece streaming de contenido multimedia en tiempo real a través de WebRTC, lo que permite a los usuarios obtener, procesar y almacenar transmisiones multimedia para tareas de reproducción, análisis y machine learning. Estas transmisiones son de alta disponibilidad, escalables y cumplen los estándares de WebRTC. Amazon Kinesis Video Streams incluye un punto de conexión de señalización de WebRTC para realizar la detección entre pares rápidamente y establecer conexiones seguras. Incluye puntos de conexión administrados de Session Traversal Utilities for NAT (STUN) y Traversal Using Relays around NAT (TURN) para un intercambio en tiempo real de contenido multimedia entre pares. También incluye un SDK de código abierto gratuito que se integra directamente con el firmware de la cámara para permitir una comunicación segura con los puntos de conexión de Kinesis Video Streams para la detección entre pares y el streaming multimedia. Por último, proporciona al cliente bibliotecas para Android, iOS y JavaScript que permite a los reproductores web y móviles compatibles con WebRTC descubrir y conectarse de forma segura con un dispositivo de cámara para streaming multimedia y la comunicación bidireccional.

Troncal SIP de alta disponibilidad con conector de voz Amazon Chime

El conector de voz de Amazon Chime es un servicio de troncal SIP que se paga por uso y permite a las compañías hacer o recibir llamadas telefónicas económicas y seguras con sus sistemas de teléfono. El conector de voz de Amazon Chime es una alternativa de bajo coste a los troncales SIP de proveedores de servicios o a las interfaces de velocidad primaria (PRI) de la red digital de servicios integrados (ISDN). Los clientes tienen la opción de habilitar las llamadas entrantes, las salientes o ambas. El servicio aprovecha la red de AWS para ofrecer una experiencia de llamadas de alta disponibilidad en varias regiones de AWS. Puede transmitir audio desde llamadas telefónicas de troncal SIP o fuentes de grabación multimedia basada en SIP (SIPREC) reenviadas a Amazon Kinesis Video Streams para obtener información de las llamadas empresariales en tiempo real. Puede crear rápidamente aplicaciones de análisis de audio mediante la integración con Amazon Transcribe y otras bibliotecas comunes de machine learning.

Prácticas recomendadas sobre el terreno

Esta sección tiene como objetivo resumir las prácticas recomendadas que han implementado algunos de los clientes de AWS más importantes y con más éxito que ejecutan cargas de trabajo de gran tamaño del protocolo de inicio de sesión (SIP) en tiempo real. Estas prácticas recomendadas les resultarán muy útiles a los clientes de AWS que deseen ejecutar su propia infraestructura de SIP en la nube pública, ya que pueden ayudar a aumentar la fiabilidad y la resiliencia del sistema en caso de que se produzcan diferentes tipos de errores. Aunque algunas de estas prácticas recomendadas son específicas de SIP, la mayoría sirven para cualquier aplicación de comunicación en tiempo real que se ejecute en AWS.

Temas

Crear una superposición de SIP

AWS posee una red troncal sólida, escalable y redundante que proporciona conectividad entre diferentes regiones. Cuando un evento de la red, como un corte de la fibra, degrada una red troncal de AWS, el tráfico se conmuta por error rápidamente a rutas redundantes mediante protocolos de enrutamiento a nivel de red, como BGP. Esta ingeniería del tráfico a nivel de la red es una caja negra para los clientes de AWS y la mayoría ni siquiera se da cuenta de estos eventos de conmutación por error. Sin embargo, los clientes que ejecutan cargas de trabajo en tiempo real, como voz, vídeo de alta calidad y mensajes de baja latencia, a veces sí que notan estos eventos. Por tanto, ¿cómo un cliente de AWS puede implementar su propia ingeniería del tráfico sobre lo que AWS ofrece a nivel de red? La solución consiste en implementar la infraestructura de SIP en muchas regiones de AWS diferentes. Como parte de las características de control de llamadas, SIP también posee la capacidad de enrutar llamadas a través de proxies de SIP específicos.

Figura 9: Uso del enrutamiento de SIP para anular el enrutamiento de la red

En la figura 9, la infraestructura de SIP (que está representada con puntos verdes) se ejecuta en las cuatro regiones de EE. UU. Las líneas azules son una representación ficticia de la red troncal de AWS. Si no se implementara el enrutamiento de SIP, una llamada que se originase en la costa oeste de EE. UU. con destino a la costa este de EE. UU. pasaría por la red troncal que conecta directamente las regiones de Oregón y Virginia. El diagrama muestra cómo un cliente puede anular el enrutamiento a nivel de red y hacer la misma llamada entre Oregón y Virginia enrutada a través de California mediante el enrutamiento de SIP. Este tipo de ingeniería de tráfico de SIP se puede implementar utilizando proxies de SIP y puertas de enlace multimedia en función de métricas de red tales como retransmisiones SIP y preferencias comerciales específicas del cliente.

Realizar una supervisión detallada

Los usuarios finales de las aplicaciones de voz y vídeo en tiempo real esperan el mismo nivel de rendimiento que el que obtienen con los servicios de telefonía tradicionales. Por lo tanto, los problemas que tienen con una aplicación terminan perjudicando la reputación del proveedor. Para ser proactivo en lugar de reactivo, es imprescindible implementar una supervisión detallada en cada parte del sistema que atienda a los usuarios finales.

Figura 10: Uso de SIPp para supervisar la infraestructura de VoIP

Hay disponibles muchas herramientas de código abierto, como iPerf o SIPp y VOIPMonitor, que se pueden usar para supervisar el tráfico SIP/RTP. En el ejemplo anterior, los nodos que ejecutan SIPp en los modos cliente y servidor miden métricas de SIP, como las llamadas correctas y las retransmisiones de SIP entre las cuatro regiones de AWS de EE. UU. Estas métricas se pueden exportar a Amazon CloudWatch mediante un script personalizado. Con CloudWatch, los clientes pueden crear alarmas sobre estas métricas personalizadas en función de un valor umbral determinado. A continuación, se pueden tomar medidas de corrección automáticas o manuales en función del estado de estas alarmas de CloudWatch.

Para los clientes que no desean asignar los recursos de ingeniería necesarios para desarrollar y mantener un sistema de supervisión personalizado, en el mercado hay disponibles muchas buenas soluciones de supervisión de VoIP, como ThousandEyes. Un ejemplo de una acción de corrección sería cambiar el enrutamiento de SIP en función del aumento de las retransmisiones de SIP.

Usar DNS para el balanceador de carga e IP flotantes para la conmutación por error

Los clientes de telefonía IP que admiten la capacidad de SRV de DNS pueden utilizar de manera eficiente la redundancia incorporada en la infraestructura mediante el balanceador de carga de los clientes en diferentes SBC/PBX.

Figura 11: Uso de registros SRV de DNS para el equilibrio de carga de clientes SIP

La figura 11 muestra cómo los clientes pueden usar los registros de SRV para equilibrar la carga del tráfico de SIP. Cualquier cliente de telefonía IP que admita el estándar SRV buscará el prefijo sip._<transport protocol> en un registro de DNS de tipo SRV. En el ejemplo, la sección de respuestas de DNS contiene los dos PBX que se ejecutan en diferentes zonas de disponibilidad de AWS. Sin embargo, además de los URI de punto de conexión, el registro SRV contiene tres datos adicionales:

- El primer número es la prioridad (1 en el ejemplo anterior). Es preferible que haya una prioridad baja que una alta.
- El segundo número es la ponderación (10 en el ejemplo anterior).
- Y el tercer número es el puerto que se utilizará (5060).

Como la prioridad es la misma (1) para ambos servidores de PBX, los clientes utilizan la ponderación para equilibrar la carga entre los dos PBX. En este caso, dado que las ponderaciones son las mismas, el equilibro de carga del tráfico de SIP debe ser igual entre los dos PBX.

El DNS puede ser una buena solución para el equilibrio de carga del cliente, pero ¿y si se implementara la conmutación por error al cambiar o actualizar los registros “A” de DNS? No se recomienda utilizar este método porque se ha encontrado una incoherencia en el comportamiento de almacenamiento en caché de DNS dentro de los nodos intermedios y del cliente. Un enfoque mejor para la conmutación por error dentro de AZ entre un clúster de nodos SIP es utilizar la reasignación de IP de EC2, en la que la dirección IP de un host dañado se reasigna instantáneamente a un host en buen estado mediante la API de EC2. En combinación con una solución detallada de supervisión y comprobación de estado, la reasignación de IP de un nodo con errores garantiza que el tráfico se traslade a un host en buen estado de la manera oportuna, lo que minimiza las interrupciones del usuario final.

Varias zonas de disponibilidad

Cada región de AWS se subdivide en distintas zonas de disponibilidad. Cada zona de disponibilidad tiene su propia fuente de alimentación, sistema de refrigeración y conectividad de red y, por lo tanto, forma un dominio de errores aislado. Dentro de las construcciones de AWS, siempre se recomienda que los clientes ejecuten sus cargas de trabajo en más de una zona de disponibilidad. Esto garantiza que las aplicaciones de los clientes soporten incluso un error completo de la zona de disponibilidad, lo que es un evento muy raro ya de por sí. Esta recomendación también se refiere a la infraestructura de SIP en tiempo real.

Figura 12: Gestión de los errores de la zona de disponibilidad

Supongamos que un evento catastrófico (como un huracán de categoría 5) provoca una interrupción completa de la zona de disponibilidad en la región us-east-1. Con la infraestructura ejecutándose como se muestra en el diagrama, todos los clientes SIP que se registraron originalmente con los nodos de la zona de disponibilidad con errores deben volver a registrarse en los nodos SIP que se ejecutan en la zona de disponibilidad 2. (Pruebe este comportamiento con sus teléfonos/clientes SIP para asegurarse de que sea compatible). Aunque las llamadas SIP activas en el momento de la interrupción de la zona de disponibilidad se pierden, las llamadas nuevas se enrutan a través de la zona de disponibilidad 2.

En resumen, los registros SRV de DNS deben dirigir al cliente a varios registros “A”, uno en cada zona de disponibilidad. Cada uno de esos registros “A” debe, a su vez, apuntar a varias direcciones IP de SBC/PBX en esa zona de disponibilidad, lo que proporciona resiliencia dentro y entre zonas de disponibilidad. La conmutación por error dentro y entre zonas de disponibilidad se puede implementar mediante la reasignación de IP si las IP son públicas. Sin embargo, las IP privadas no se pueden reasignar en las zonas de disponibilidad. Si un cliente utiliza direcciones IP privadas, tendrá que confiar en que los clientes SIP se registren de nuevo con la SBC/PBX de copia de seguridad para la conmutación por error entre zonas de disponibilidad.

Mantener el tráfico dentro de una zona de disponibilidad y usar grupos de ubicación de EC2

Esta práctica recomendada, también conocida como afinidad de zonas de disponibilidad, se aplica en el raro caso de que se produzca un error completo de la zona de disponibilidad. Se recomienda eliminar todo el tráfico entre zonas de disponibilidad, de modo que cualquier tráfico de SIP o RTP que entre en una zona de disponibilidad permanezca en ella hasta que salga de la región.

Figura 13: Afinidad de zonas de disponibilidad (como máximo, se pierden el 50 % de las llamadas activas)

La figura 13 muestra una arquitectura simplificada que utiliza la afinidad de zonas de disponibilidad. La ventaja comparativa de este enfoque es evidente si se tienen en cuenta los efectos de una interrupción completa de la zona de disponibilidad. Como se muestra en el diagrama, si se pierde la zona de disponibilidad 2, eso afectaría como máximo al 50 % de las llamadas activas (suponiendo que el equilibrio de carga es igual entre las zonas de disponibilidad). Si no se implementara la afinidad de zonas de disponibilidad, algunas llamadas fluirían entre las zonas de disponibilidad de una región y es probable que un error afecte a más del 50 % de las llamadas activas.

Además, para minimizar la latencia del tráfico, también recomendamos considerar el uso de grupos de ubicación de EC2 dentro de cada zona de disponibilidad. Las instancias lanzadas dentro del mismo grupo de ubicación de EC2 tienen un ancho de banda mayor y una latencia menor, ya que EC2 garantiza la proximidad de red de estas instancias entre sí.

Usar tipos de instancias de EC2 de redes mejoradas

La elección del tipo de instancia correcto en Amazon EC2 garantiza la fiabilidad del sistema y el uso eficiente de la infraestructura. EC2 proporciona una amplia variedad de tipos de instancias optimizados para diferentes casos de uso. Los tipos de instancias abarcan varias combinaciones de capacidad de CPU, memoria, almacenamiento y redes. Le proporcionan flexibilidad para elegir la combinación de recursos adecuada para sus aplicaciones. Estos tipos de instancias de redes mejoradas garantizan que las cargas de trabajo SIP que se ejecutan en ellas tengan acceso a un ancho de banda constante y a una latencia agregada comparativamente menor. Una reciente novedad de Amazon EC2 es la disponibilidad del Elastic Network Adapter (ENA) que proporciona hasta 100 Gbps de ancho de banda. El catálogo más reciente de tipos de instancias de EC2 y características asociadas se puede encontrar en la página de tipos de instancias de EC2.

Para la mayoría de los clientes, la última generación de instancias optimizadas para la computación debería ofrecer la mejor relación calidad-precio. Por ejemplo, el C5N admite el nuevo Elastic Network Adapter con un ancho de banda de hasta 100 Gbps con millones de paquetes por segundo (PPS). La mayoría de las aplicaciones en tiempo real también se beneficiarían del uso del Intel Data Plane Developer Kit (DPDK), que puede mejorar enormemente el procesamiento de paquetes.

Sin embargo, siempre se recomienda comparar los distintos tipos de instancias de EC2 de acuerdo con los requisitos para ver qué tipo de instancia funciona mejor en cada caso. Esta comparación

también permite encontrar otros parámetros de configuración, como el número máximo de llamadas que puede procesar a la vez un tipo de instancia determinado.

Consideraciones de seguridad

Los componentes de las aplicaciones RTC suelen ejecutarse directamente en instancias de Amazon EC2 orientadas a Internet. Además de TCP, los flujos utilizan protocolos como UDP y SIP. En estos casos, AWS Shield Standard protege las instancias de Amazon EC2 de los ataques de DDoS de la capa de infraestructura común (capas 3 y 4), como los ataques de reflexión de UDP, reflexión de DNS, reflexión de NTP, reflexión de SSDP, etc. AWS Shield Standard utiliza varias técnicas, como el modelado de tráfico basado en prioridades, que se activan automáticamente cuando se detecta una firma de ataque de DDoS bien definida.

AWS también proporciona protección avanzada contra ataques de DDoS grandes y sofisticados para estas aplicaciones al habilitar AWS Shield Advanced en direcciones IP elásticas. AWS Shield Advanced proporciona una detección de DDoS mejorada que detecta automáticamente el tipo de recurso de AWS y el tamaño de la instancia de EC2 y aplica las mitigaciones predefinidas apropiadas con protecciones contra inundaciones UDP o SYN. Con AWS Shield Advanced, los clientes también pueden crear sus propios perfiles de mitigación personalizados junto con el equipo de respuesta a DDoS (DRT) de AWS que trabaja las 24 horas del día y los 7 días de la semana. AWS Shield Advanced también garantiza que, durante un ataque de DDoS, todas sus listas de control de acceso (ACL) a redes de Amazon VPC se apliquen automáticamente en el borde de la red de AWS, lo que le proporciona acceso a ancho de banda adicional y capacidad de depuración para mitigar ataques de DDoS volumétricos de gran tamaño.

Conclusión

Las cargas de trabajo de la comunicación en tiempo real (RTC) se pueden implementar en Amazon Web Services (AWS) para lograr escalabilidad, elasticidad y alta disponibilidad y, al mismo tiempo, cumplir los requisitos clave. En la actualidad, varios clientes utilizan AWS, a sus socios y las soluciones de código abierto para ejecutar cargas de trabajo de RTC a un coste menor y con más agilidad, además de reducir el espacio global.

Las arquitecturas de referencia y las prácticas recomendadas que se proporcionan en este documento técnico pueden ayudar a los clientes a configurar correctamente las cargas de trabajo de RTC en AWS y a optimizar las soluciones para cumplir los requisitos de los usuarios finales mientras realizan la optimización para la nube.

Colaboradores

Las siguientes personas y organizaciones contribuyeron a redactar este documento:

- Ahmad Khan, arquitecto de soluciones sénior, Amazon Web Services
- Tipu Qureshi, ingeniero principal, AWS Support, Amazon Web Services
- Hasan Khan, gerente de cuentas técnicas sénior, Amazon Web Services
- Shoma Chakravarty, líder técnico de WW, telecomunicaciones, Amazon Web Services

Revisiones del documento

Para recibir notificaciones sobre las actualizaciones de este documento técnico, suscríbase a la fuente RSS.

update-history-change	update-history-description	update-history-date
Documento técnico actualizado	Actualizado para los servicios y características más recientes .	13 de febrero de 2020
Publicación inicial	Documento técnico publicado por primera vez.	1 de octubre de 2018

Avisos

Los clientes son responsables de realizar sus propias evaluaciones de la información contenida en este documento. Este documento: (a) solo tiene fines informativos, (b) representa las prácticas y las ofertas de productos vigentes de AWS, que están sujetas a cambios sin previo aviso, y (c) no crea ningún compromiso ni garantía de AWS y sus empresas afiliadas, proveedores o concesionarios de licencias. Los productos o servicios de AWS se proporcionan “tal cual”, sin garantías, representaciones ni condiciones de ningún tipo, ya sean explícitas o implícitas. Las responsabilidades y obligaciones de AWS en relación con sus clientes se rigen por los acuerdos de AWS, y este documento no modifica ni forma parte de ningún acuerdo entre AWS y sus clientes.