Whitepaper AWS

Comunicazione in tempo reale su AWS

Comunicazione in tempo reale su AWS: Whitepaper AWS

Table of Contents

Comunicazione in tempo reale su AWS

Best practice per la progettazione di carichi di lavoro scalabili, a elevata disponibilità e dotati di comunicazione in tempo reale su AWS

Data di pubblicazione: 13 febbraio 2020 (Revisioni del documento)

Riassunto

Oggi, molte organizzazioni stanno cercando di ridurre i costi e raggiungere la scalabilità per carichi di lavoro vocali, di messaggistica e multimediali in tempo reale. Questo documento illustra le best practice per la gestione dei carichi di lavoro di comunicazione in tempo reale su AWS e include architetture di riferimento per soddisfare questi requisiti. Questo documento serve come guida per le persone che hanno familiarità con la comunicazione in tempo reale su come ottenere disponibilità e scalabilità elevate per questi carichi di lavoro.

Introduzione

Le applicazioni di telecomunicazione che utilizzano voce, video e messaggistica come canali sono un requisito fondamentale per molte organizzazioni e per i loro utenti finali. Questi carichi di lavoro di comunicazione in tempo reale (RTC) hanno requisiti specifici di latenza e disponibilità che possono essere soddisfatti seguendo le best practice di progettazione pertinenti. In passato, i carichi di lavoro RTC sono stati implementati nei tradizionali data center On-Premise con risorse dedicate.

Tuttavia, a causa di un insieme maturo e in crescita di funzionalità, i carichi di lavoro RTC possono essere distribuiti su Amazon Web Services (AWS) nonostante i severi requisiti dei livelli di servizio, beneficiando allo stesso tempo di scalabilità, elasticità ed elevata disponibilità. Oggi, diversi clienti utilizzano AWS, i suoi partner e soluzioni open source per eseguire carichi di lavoro RTC con costi ridotti, agilità più rapida, capacità di diventare globali in pochi minuti e funzionalità avanzate dai servizi AWS.

I clienti sfruttano le funzionalità di AWS come il networking avanzato con un Elastic Network Adapter (ENA) e l'ultima generazione di istanze Amazon Elastic Compute Cloud (EC2) per trarre vantaggio dal kit di sviluppo del piano dati (DPDK), dalla virtualizzazione I/O a radice singola (SR- IOV), pagine enormi, supporto NVM Express (NVMe), accesso alla memoria non uniforme (NUMA) e istanze bare metal per soddisfare i requisiti del carico di lavoro RTC. Queste istanze offrono una larghezza di banda di rete fino a 100 Gbps e pacchetti proporzionati al secondo, offrendo prestazioni migliori per le applicazioni a uso intensivo di rete. Per la scalabilità, Elastic Load Balancing offre Application Load Balancer, che offre supporto WebSocket e Network Load Balancer in grado di gestire milioni di richieste al secondo. Per l'accelerazione della rete, AWS Global Accelerator fornisce indirizzi IP statici che fungono da punto di ingresso fisso agli endpoint delle applicazioni in AWS. Supporta gli indirizzi IP statici per il sistema di bilanciamento del carico. Per ridurre latenza, costi e una maggiore velocità di trasmissione della larghezza di banda,AWS Direct Connect stabilisce una connessione di rete dedicata da On-Premise ad AWS. Il trunking SIP gestito a elevata disponibilità è fornito da Amazon Chime Voice Connector. Amazon Kinesis Video Streams con WebRTC trasmette facilmente contenuti multimediali bidirezionali in tempo reale con elevata disponibilità.

Questo documento include architetture di riferimento che mostrano come impostare carichi di lavoro RTC su AWS e best practice per ottimizzare le soluzioni per soddisfare i requisiti degli utenti finali ottimizzando al contempo per il cloud. L'evolved packet core (EPC) non rientra nell'ambito di questo Whitepaper, ma le best practice dettagliate possono essere applicate alle funzioni di rete virtuale (VNF).

Componenti fondamentali dell'architettura RTC

Nel settore delle telecomunicazioni, la comunicazione in tempo reale (RTC) si riferisce comunemente a sessioni multimediali in tempo reale tra due endpoint con una latenza minima. Queste sessioni possono essere correlate a:

- Una sessione vocale tra due parti (ad esempio, sistema telefonico, cellulare, VoIP)
- Messaggistica istantanea (ad es. chat, IRC)
- Sessione video in diretta (ad es. videoconferenza, telepresenza)

Ognuna delle soluzioni precedenti ha alcuni componenti in comune (ad esempio, componenti che forniscono autenticazione, autorizzazione e controllo degli accessi, transcodifica, buffering e relay e così via) e alcuni componenti unici per il tipo di supporto trasmesso (ad esempio, servizio di trasmissione, server di messaggistica e code e così via). Questa sezione si concentra sulla definizione di un sistema RTC basato su voce e video e su tutti i componenti correlati illustrati nella Figura 1.

Figura 1: Componenti architettonici essenziali per RTC

Argomenti

Softswitch/PBX

Un softswitch o PBX è il cervello di un sistema di telefonia vocale e fornisce intelligenza per stabilire, mantenere e instradare una chiamata vocale all'interno o all'esterno dell'azienda utilizzando diversi componenti. Tutti gli abbonati dell'azienda sono tenuti a registrarsi con il softswitch per ricevere o effettuare una chiamata. Una funzionalità importante del softswitch è tenere traccia di ciascun abbonato e di come raggiungerlo utilizzando gli altri componenti all'interno della rete vocale.

Controller di frontiera di sessione (SBC)

Un controller di frontiera di sessione (SBC) si trova ai margini di una rete vocale e tiene traccia di tutto il traffico in entrata e in uscita (sia piani di controllo che piani dati). Una delle principali responsabilità di un SBC è proteggere il sistema vocale da un uso dannoso. L'SBC può essere utilizzato per interconnettersi con i trunk SIP (Session Initiation Protocol) per la connettività esterna. Alcuni SBC forniscono anche funzionalità di transcodifica per la conversione di CODEC da un formato all'altro. Infine, la maggior parte degli SBC fornisce anche funzionalità NAT Traversal che aiutano a garantire che le chiamate vengano stabilite anche attraverso reti con firewall.

Connettività PSTN

Le soluzioni Voice over IP (VoIP) utilizzano gateway PSTN e trunk SIP per connettersi alle reti PSTN legacy.

Gateway PSTN

Il gateway di rete telefonica pubblica commutata (PSTN) converte la segnalazione (tra SIP e SS7) e i media (tra RTP e multiplexing a divisione temporale [TDM] utilizzando la transcodifica CODEC). I gateway PSTN si trovano sempre ai margini della rete PSTN.

Trunk SIP

In un trunk SIP, l'azienda non termina le sue chiamate su una rete TDM (basata su SS7), ma piuttosto i flussi tra enterprise e telco rimangono su IP. La maggior parte dei trunk SIP viene stabilita utilizzando SBC. L'azienda deve concordare le regole di sicurezza predefinite delle telecomunicazioni, come consentire un certo intervallo di indirizzi IP, porte e così via.

Media Gateway (Transcoder)

Una tipica soluzione vocale consente vari tipi di CODEC. Alcuni dei CODEC comuni sono G.711 μ-law per il Nord America, G.711 A-law per i paesi al di fuori del Nord America, G.729 e G.722. Quando due dispositivi che utilizzano due CODEC diversi comunicano tra loro, un media server traduce il flusso CODEC tra i dispositivi. In altre parole, un gateway multimediale elabora i media e garantisce che i dispositivi finali siano in grado di comunicare tra loro.

Gateway WebRTC e WebRTC

La comunicazione Web in tempo reale (WebRTC) consente di stabilire una chiamata da un browser Web o richiedere risorse dal server di backend utilizzando l'API. La tecnologia è progettata pensando alla tecnologia cloud e quindi fornisce varie API che potrebbero essere utilizzate per stabilire una chiamata. Poiché non tutte le soluzioni vocali (incluso SIP) supportano queste API, il gateway WebRTC è necessario per tradurre le chiamate API in messaggi SIP e viceversa.

La Figura 2 mostra un modello di progettazione per un'architettura WebRTC a elevata disponibilità. Il traffico in entrata dai client WebRTC è bilanciato da un sistema di bilanciamento del carico delle applicazioni Amazon con WebRTC in esecuzione su istanze EC2 che fanno parte di un gruppo Auto Scaling.

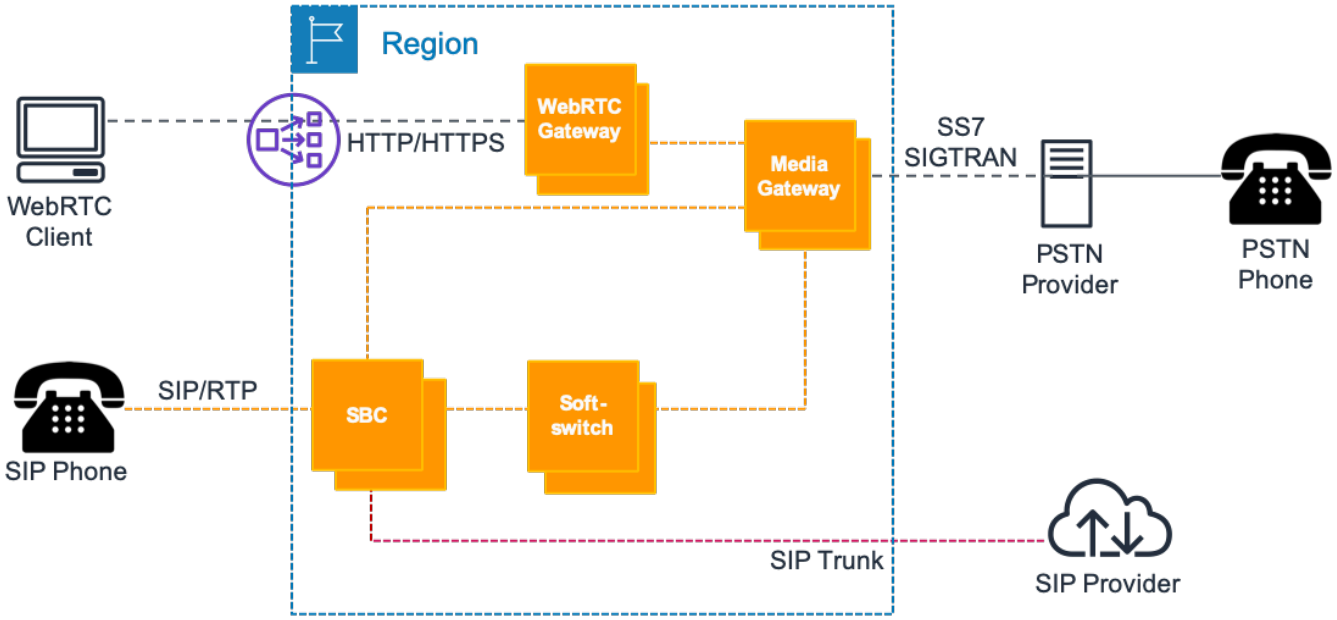

Figura 2: Una topologia di base di un sistema RTC per la voce

Un altro modello di progettazione per il traffico SIP e RTP consiste nell'utilizzare coppie di SBC su Amazon EC2 in modalità attiva/passiva nelle zone di disponibilità (Figura 3). In questo caso, un indirizzo IP elastico può essere spostato dinamicamente tra le istanze in caso di guasto, laddove non è possibile utilizzare il DNS.

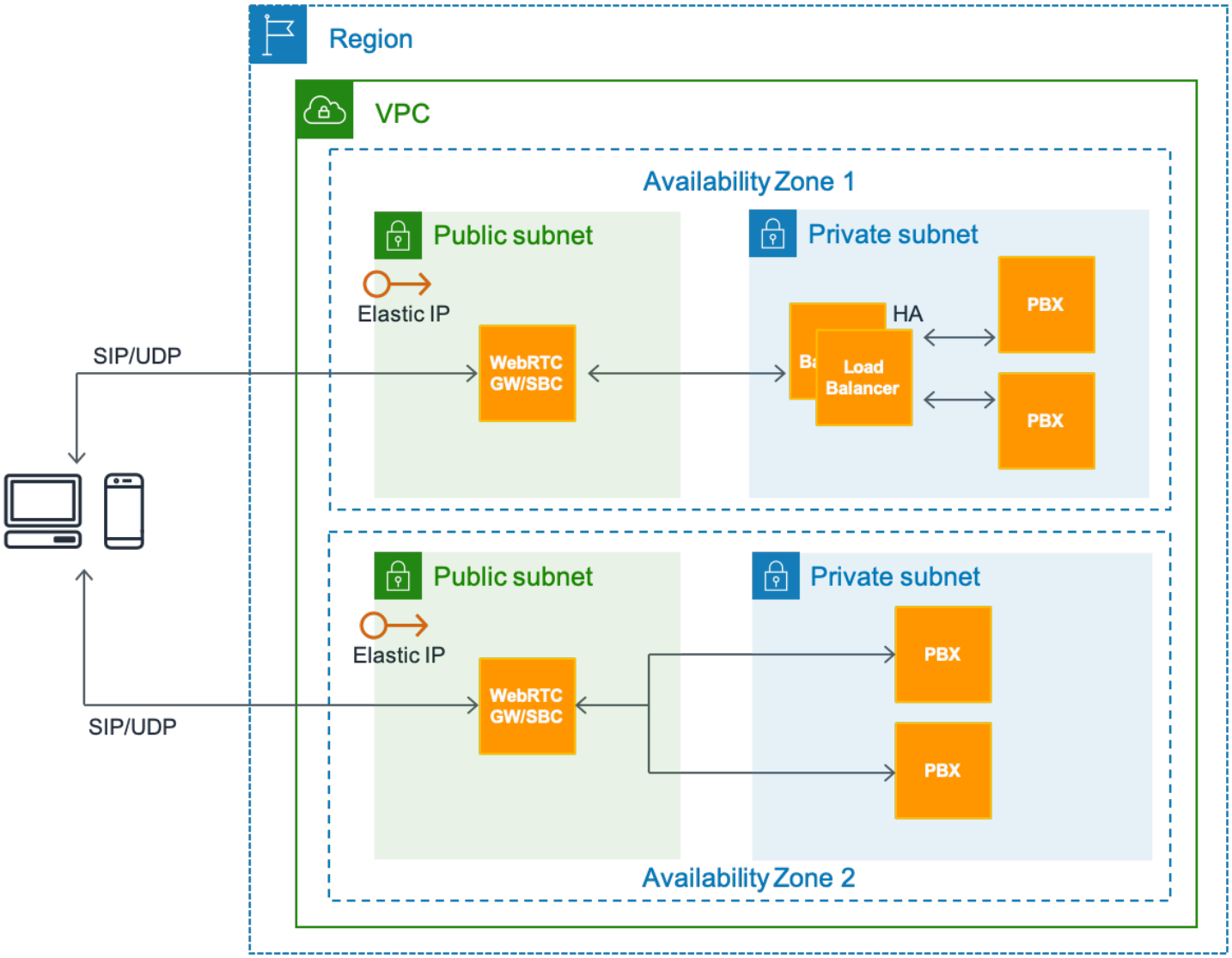

Figura 3: Architettura RTC che utilizza Amazon EC2 in un VPC

Scalabilità e disponibilità elevate su AWS

La maggior parte dei provider di comunicazioni in tempo reale si allinea ai livelli di servizio che forniscono disponibilità dal 99,9% al 99,999%. A seconda del grado di elevata disponibilità (HA) desiderato, è necessario adottare misure sempre più sofisticate lungo l'intero ciclo di vita dell'applicazione. Consigliamo di seguire queste linee guida per ottenere un grado di disponibilità elevata eccellente:

- Progettare il sistema in modo che non abbia un singolo punto di errore. Utilizzo di meccanismi automatici di monitoraggio, rilevamento dei guasti e failover per componenti senza stato e con stato

- I singoli punti di errore (SPOF) vengono generalmente eliminati con una configurazione di ridondanza N+1 o 2N, dove N+1 si raggiunge tramite il bilanciamento del carico tra nodi attivi-attivi e 2N viene raggiunto da una coppia di nodi in configurazione attivo-standby.
- AWS ha diversi metodi per raggiungere l'HA attraverso entrambi gli approcci, ad esempio attraverso un cluster scalabile e con bilanciamento del carico o assumendo una coppia attiva-standby.

- Disponibilità corretta dello strumento e del sistema di test.
- Preparare le procedure operative per i meccanismi manuali per rispondere, mitigare e ripristinare l'errore.

Questa sezione si concentra su come non raggiungere un singolo punto di errore utilizzando le funzionalità disponibili in AWS. In particolare, questa sezione descrive un sottoinsieme delle funzionalità principali di AWS e dei modelli di progettazione che consentono di creare applicazioni di comunicazione in tempo reale a elevata disponibilità sulla piattaforma.

Argomenti

- Trunking SIP a elevata disponibilità con Amazon Chime Voice Connector

Modello IP mobile per elevata disponibilità tra server stateful attivi e in standby

Il modello di progettazione IP mobile è un meccanismo noto per ottenere il failover automatico tra una coppia di nodi hardware attivi e in standby (server multimediali). Un indirizzo IP virtuale secondario statico viene assegnato al nodo attivo. Il monitoraggio continuo tra i nodi attivi e quelli in standby rileva i guasti. Se il nodo attivo non riesce, lo script di monitoraggio assegna l'IP virtuale al nodo di standby pronto e il nodo di standby assume la funzione attiva primaria. In questo modo, l'IP virtuale fluttua tra il nodo attivo e quello in standby.

Argomenti

Applicabilità nelle soluzioni RTC

Non è sempre possibile avere più istanze attive dello stesso componente in servizio, ad esempio un cluster attivo-attivo di N nodi. Una configurazione active-standby fornisce il meccanismo migliore per l'elevata disponibilità. Ad esempio, i componenti stateful di una soluzione RTC, come il media server o il server di conferenza, o anche un server SBC o database, sono adatti per una configurazione attiva-standby. Un SBC o un media server ha diverse sessioni o canali di lunga durata attivi in un dato momento e, in caso di errore dell'istanza attiva SBC, gli endpoint possono riconnettersi al nodo di standby senza alcuna configurazione lato client a causa dell'IP mobile.

Implementazione su AWS

Puoi implementare questo modello in AWS utilizzando le funzionalità di base di Amazon Elastic Compute Cloud (Amazon EC2), l'API di Amazon EC2, gli indirizzi IP elastici e il supporto su Amazon EC2 per gli indirizzi IP privati secondari.

1. Avviare due istanze EC2 per assumere il ruolo di nodi primari e secondari, in cui si presume che il primario sia attivo per impostazione predefinita.

2. Assegnare un indirizzo IP privato secondario aggiuntivo all'istanza EC2 primaria.
3. Un indirizzo IP elastico, simile a un IP virtuale (VIP), è associato all'indirizzo privato secondario. Questo indirizzo privato secondario è l'indirizzo utilizzato dagli endpoint esterni per accedere all'applicazione.
4. È necessaria una certa configurazione del sistema operativo per aggiungere l'indirizzo IP secondario come alias all'interfaccia di rete primaria.
5. L'applicazione deve collegarsi a questo indirizzo IP elastico. Nel caso del software Asterisk, è possibile configurare l'associazione tramite le impostazioni SIP Asterisk avanzate.
6. Eseguire uno script di monitoraggio (personalizzato, KeepAlive su Linux, Corosync e così via) su ogni nodo per monitorare lo stato del nodo peer. Nel caso in cui il nodo attivo corrente fallisca, il peer rileva questo errore e richiama l'API di Amazon EC2 per riassegnare l'indirizzo IP privato secondario a se stesso.
7. Pertanto, l'applicazione che era in ascolto sul VIP associato all'indirizzo IP privato secondario diventa disponibile per gli endpoint tramite il nodo di standby.

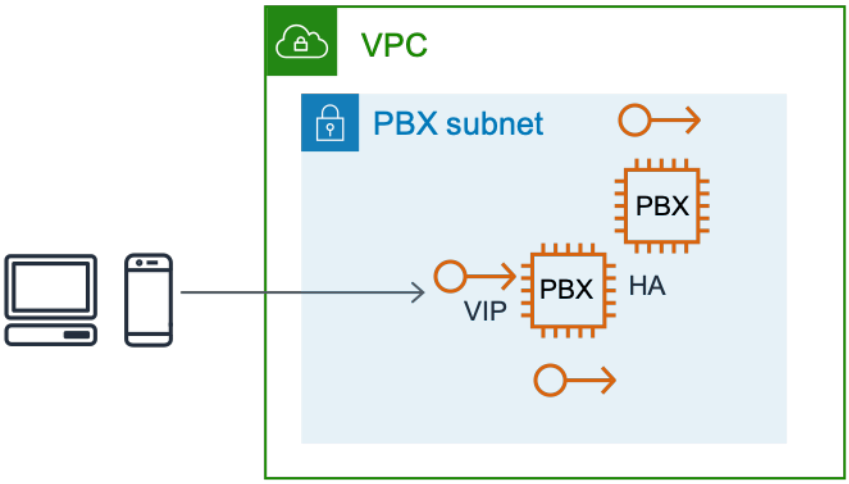

Figura 4: Failover tra istanze EC2 stateful che utilizzano indirizzi IP elastici

Vantaggi

Questo approccio è una soluzione affidabile a basso budget che protegge dai guasti a livello di istanza, infrastruttura o applicazione EC2.

Limitazioni ed estensibilità

Questo modello di progettazione è in genere limitato all'interno di una singola zona di disponibilità. Può essere implementato in due zone di disponibilità, ma con una variazione. In questo caso, l'indirizzo IP elastico flottante viene riassociato tra il nodo attivo e quello in standby in diverse zone di disponibilità tramite l'API dell'indirizzo IP elastico nuovamente associato disponibile. Nell'implementazione del failover illustrata nella Figura 4, le chiamate in corso vengono eliminate e gli endpoint devono riconnettersi. È possibile estendere questa implementazione con la replica dei dati di sessione sottostanti per fornire anche il failover continuo delle sessioni o la continuità dei supporti.

Bilanciamento del carico per scalabilità ed elevata disponibilità con WebRTC e SIP

Il bilanciamento del carico di un cluster di istanze attive basato su regole predefinite, come round robin, affinità o latenza e così via, è un modello di progettazione ampiamente diffuso dalla natura stateless delle richieste HTTP. In effetti, il bilanciamento del carico è un'opzione valida nel caso di molti componenti dell'applicazione RTC.

Il bilanciatore del carico funge da proxy inverso o punto di ingresso per le richieste all'applicazione desiderata, che a sua volta è configurata per l'esecuzione simultanea in più nodi attivi. In un dato momento, il bilanciatore del carico indirizza una richiesta dell'utente a uno dei nodi attivi nel cluster definito. I sistemi di bilanciamento del carico eseguono un controllo dello stato sui nodi del cluster di destinazione e non inviano una richiesta in arrivo a un nodo che non supera il controllo dello stato. Pertanto, un livello fondamentale di elevata disponibilità è raggiunto dal bilanciamento del carico. Inoltre, poiché un sistema di bilanciamento del carico esegue controlli di integrità attivi e passivi su tutti i nodi del cluster a intervalli inferiori al secondo, il tempo per il failover è quasi istantaneo.

La decisione su quale nodo indirizzare si basa sulle regole di sistema definite nel bilanciatore del carico, tra cui:

- Round robin
- Affinità di sessione o IP, che garantisce che più richieste all'interno di una sessione o dallo stesso IP vengano inviate allo stesso nodo del cluster
- Dipendente dalla latenza
- Dipendente dal carico

Argomenti

Applicabilità nelle architetture RTC

Il protocollo WebRTC consente ai gateway WebRTC di essere facilmente bilanciati nel carico tramite un bilanciatore del carico basato su HTTP, come Elastic Load Balancing, Application Load Balancer o Network Load Balancer. Con la maggior parte delle implementazioni SIP che si basano sul trasporto su TCP e UDP, è necessario il bilanciamento del carico a livello di rete o di connessione con supporto per il traffico basato su TCP e UDP.

Bilanciamento del carico su AWS per WebRTC utilizzando Application Load Balancer e Auto Scaling

Nel caso di comunicazioni basate su WebRTC, Elastic Load Balancing fornisce un bilanciatore del carico completamente gestito, altamente disponibile e scalabile che funge da punto di ingresso per le richieste, che vengono poi indirizzate a un cluster di destinazione di istanze EC2 associate a Elastic Load Balancing. Inoltre, poiché le richieste WebRTC sono stateless, puoi utilizzare Amazon EC2 Auto Scaling per fornire scalabilità, elasticità ed elevata disponibilità completamente automatizzate e controllabili.

Application Load Balancer fornisce un servizio di bilanciamento del carico completamente gestito che è altamente disponibile, poiché utilizza più zone di disponibilità, e scalabile. Questo supporta il bilanciamento del carico delle richieste WebSocket che gestiscono la segnalazione per le applicazioni WebRTC e la comunicazione bidirezionale tra client e server utilizzando una connessione TCP di lunga durata. Application Load Balancer supporta anche il routing basato sul contenuto e le sessioni permanenti, instradando le richieste dallo stesso client alla stessa destinazione utilizzando i cookie generati dal bilanciatore del carico. Se abiliti la sessione permanente, il medesimo obiettivo riceve la richiesta e può utilizzare il cookie per recuperare il contesto della sessione.

La Figura 5 mostra la topologia di destinazione.

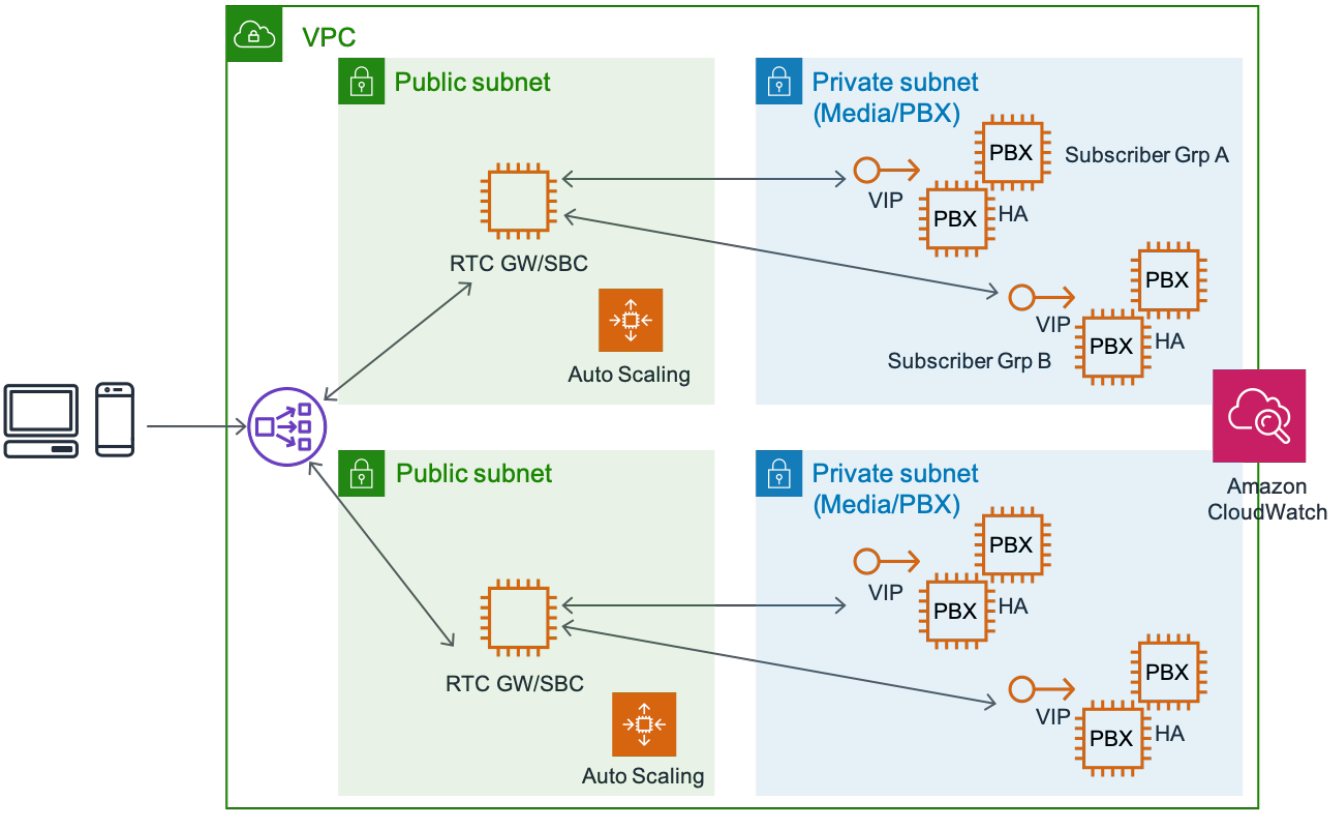

Figura 5: Scalabilità WebRTC e architettura a elevata disponibilità

Implementazione per SIP utilizzando Network Load Balancer o Marketplace AWS

Nel caso di comunicazioni basate su SIP, le connessioni vengono effettuate tramite TCP o UDP, con la maggior parte delle applicazioni RTC che utilizzano UDP. Se SIP/TCP è il protocollo di segnale preferito, allora è possibile utilizzare Network Load Balancer per un bilanciamento del carico completamente gestito, a elevata disponibilità, scalabile e delle prestazioni.

Un Network Load Balancer opera a livello di connessione, ovvero a livello 4, instradando le connessioni verso destinazioni designate, che possono essere istanze Amazon EC2, container e indirizzi IP in base ai dati del protocollo IP. Ideale per il bilanciamento del carico del traffico TCP o UDP, il bilanciamento del carico di rete è in grado di gestire milioni di richieste al secondo mantenendo latenze ultra basse. È integrato con altri servizi AWS popolari, come AWS Auto Scaling, Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) e AWS CloudFormation.

Se vengono avviate connessioni SIP, un'altra opzione è quella di utilizzare un software Marketplace AWS commerciale standard (COTS). Marketplace AWS offre molti prodotti in grado di gestire UDP e altri tipi di bilanciamento del carico di connessione Layer 4. Questi COTS in genere includono il supporto per l'elevata disponibilità e sono comunemente integrati con funzionalità, ad esempio AWS Auto Scaling, per migliorare ulteriormente la disponibilità e la scalabilità. La Figura 6 mostra la topologia di destinazione:

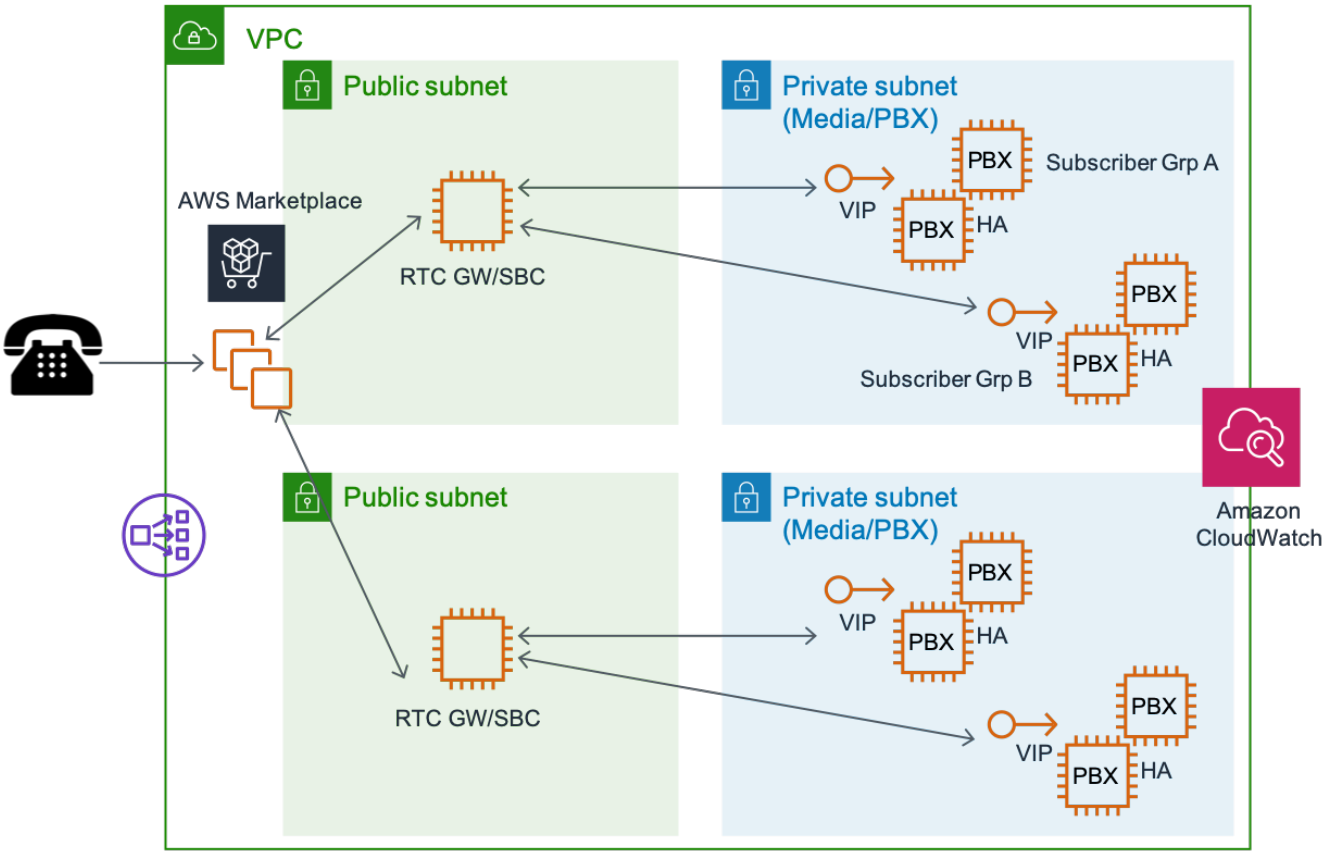

Figura 6: Scalabilità RTC basata su SIP con il prodotto Marketplace AWS

Bilanciamento del carico e failover basati su DNS su più regioni

Amazon Route 53 fornisce un servizio DNS globale che può essere utilizzato come endpoint pubblico o privato per i client RTC per la registrazione e la connessione con le applicazioni multimediali. Con Amazon Route 53, i controlli dello stato del DNS possono essere configurati per instradare il traffico verso endpoint sani o per monitorare in modo indipendente lo stato di un'applicazione. Amazon Route 53 Traffic Flow semplifica la gestione del traffico globale attraverso vari tipi di instradamento, inclusi instradamento basato sulla latenza, Geo DNS, geoprossimità e Weighted Round Robin, ognuno dei quali può essere combinato con il failover DNS per abilitare una serie di architetture a bassa latenza e con tolleranza di errore. L'intuitivo editor visuale di Amazon Route 53 Traffic Flow consente di

gestire l'instradamento degli utenti finali agli endpoint delle applicazioni, sia in una singola regione AWS sia in un ambiente distribuito in tutto il mondo.

Nel caso di implementazioni globali, la politica di instradamento basata sulla latenza in Route 53 è particolarmente utile per indirizzare i clienti al punto di presenza più vicino per un media server per migliorare la qualità del servizio associato agli scambi di media in tempo reale.

Si noti che per applicare un failover a un nuovo indirizzo DNS, le cache dei client devono essere svuotate. Inoltre, le modifiche DNS possono presentare un ritardo in quanto vengono propagate tra i server DNS globali. È possibile gestire l'intervallo di aggiornamento per le ricerche DNS con l'attributo Time to Live. Questo attributo è configurabile al momento dell'impostazione dei criteri DNS.

Per raggiungere rapidamente gli utenti globali o per soddisfare i requisiti di utilizzo di un singolo IP pubblico, AWS Global Accelerator può essere utilizzato anche per il failover interregionale. AWS Global Accelerator è un servizio di rete che migliora la disponibilità e le prestazioni per le applicazioni con portata locale e globale. AWS Global Accelerator fornisce indirizzi IP statici che fungono da punto di ingresso fisso per gli endpoint delle applicazioni, ad esempio Application Load Balancer, Network Load Balancer o istanze Amazon EC2 in una o più regioni AWS. Utilizza la rete globale AWS per ottimizzare il percorso dagli utenti alle applicazioni, migliorando le prestazioni, come la latenza del traffico TCP e UDP. AWS Global Accelerator monitora continuamente lo stato degli endpoint delle applicazioni e reindirizza automaticamente il traffico agli endpoint integri più vicini nel caso in cui gli endpoint attuali diventino non integri. Per ulteriori requisiti di sicurezza, la VPN accelerata da sito a sito utilizza AWS Global Accelerator per migliorare le prestazioni delle connessioni VPN instradando in modo intelligente il traffico attraverso la rete globale AWS e le edge location AWS.

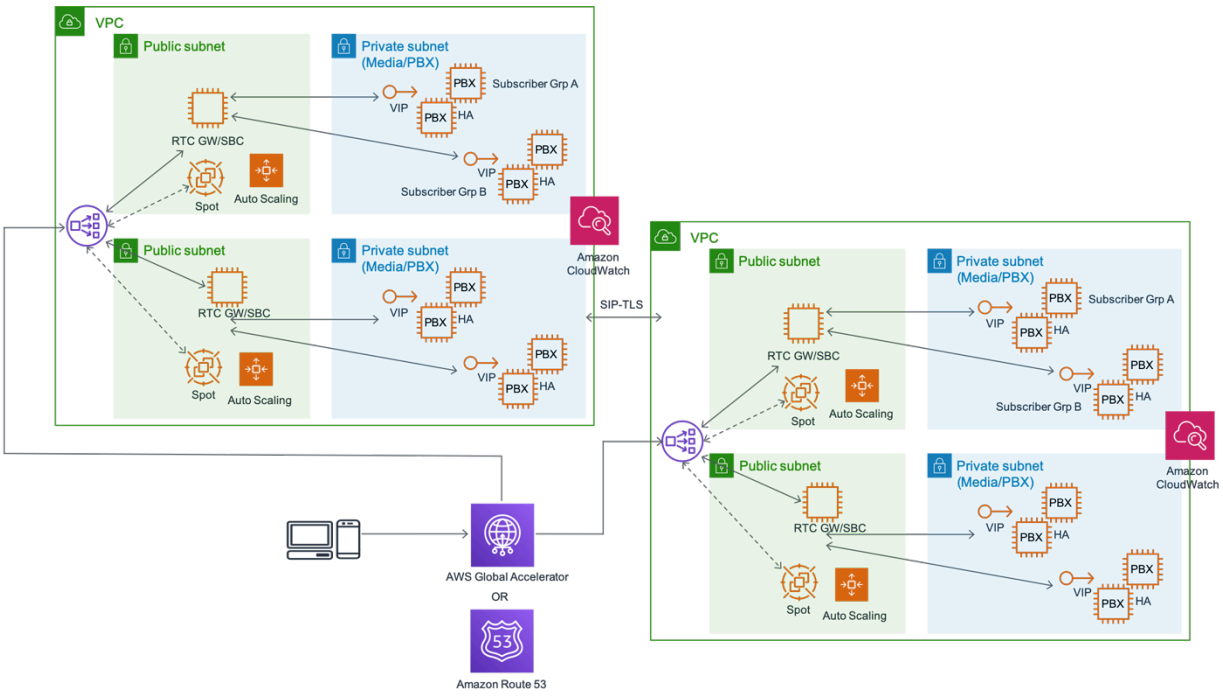

Figura 7: Progettazione a elevata disponibilità tra regioni utilizzando AWS Global Accelerator o Amazon Route 53

Durabilità dei dati ed elevata disponibilità con archiviazione persistente

La maggior parte delle applicazioni RTC si basa sull'archiviazione persistente per archiviare e accedere ai dati per l'autenticazione, l'autorizzazione, la contabilità (dati di sessione, record dei dettagli delle chiamate, ecc.), il monitoraggio operativo e la registrazione. In un data center tradizionale, la garanzia di elevata disponibilità e durata per i componenti di archiviazione persistenti (database, file system e così via) richiede in genere un carico pesante attraverso l'installazione di una SAN, la progettazione RAID e processi per il backup, il ripristino e l'elaborazione del failover. Il cloud AWS semplifica e migliora notevolmente le pratiche tradizionali dei data center in merito alla durabilità e alla disponibilità dei dati.

Per l'archiviazione di oggetti e di file, i servizi AWS come Amazon Simple Storage Service (Amazon S3) e Amazon Elastic File System (Amazon EFS) forniscono elevata disponibilità e scalabilità gestite. Amazon S3 ha una durabilità dei dati di 11 nove.

Per l'archiviazione dei dati transazionali, i clienti hanno la possibilità di sfruttare Amazon Relational Database Service (Amazon RDS) completamente gestito che supporta Amazon Aurora, PostgreSQL,

MySQL, MariaDB, Oracle e Microsoft SQL Server con distribuzioni a elevata disponibilità. Per la funzione di registrar, il profilo dell'abbonato o l'archiviazione dei record contabili (ad esempio, CDR), Amazon RDS offre un'opzione fault-tolerant, altamente disponibile e scalabile.

Scalabilità dinamica con AWS Lambda, Amazon Route 53 e AWS Auto Scaling

AWS consente il concatenamento di funzionalità e la possibilità di incorporare funzioni serverless personalizzate come servizio in base agli eventi dell'infrastruttura. Uno di questi modelli di progettazione che ha molti usi versatili nelle applicazioni RTC è la combinazione di hook del ciclo di vita con scalabilità automatica con Amazon CloudWatch Events, Amazon Route 53 e funzioni AWS Lambda. Le funzioni AWS Lambda possono incorporare qualsiasi azione o logica. La Figura 8 dimostra come queste funzionalità concatenate possano migliorare l'affidabilità e la scalabilità del sistema con l'automazione.

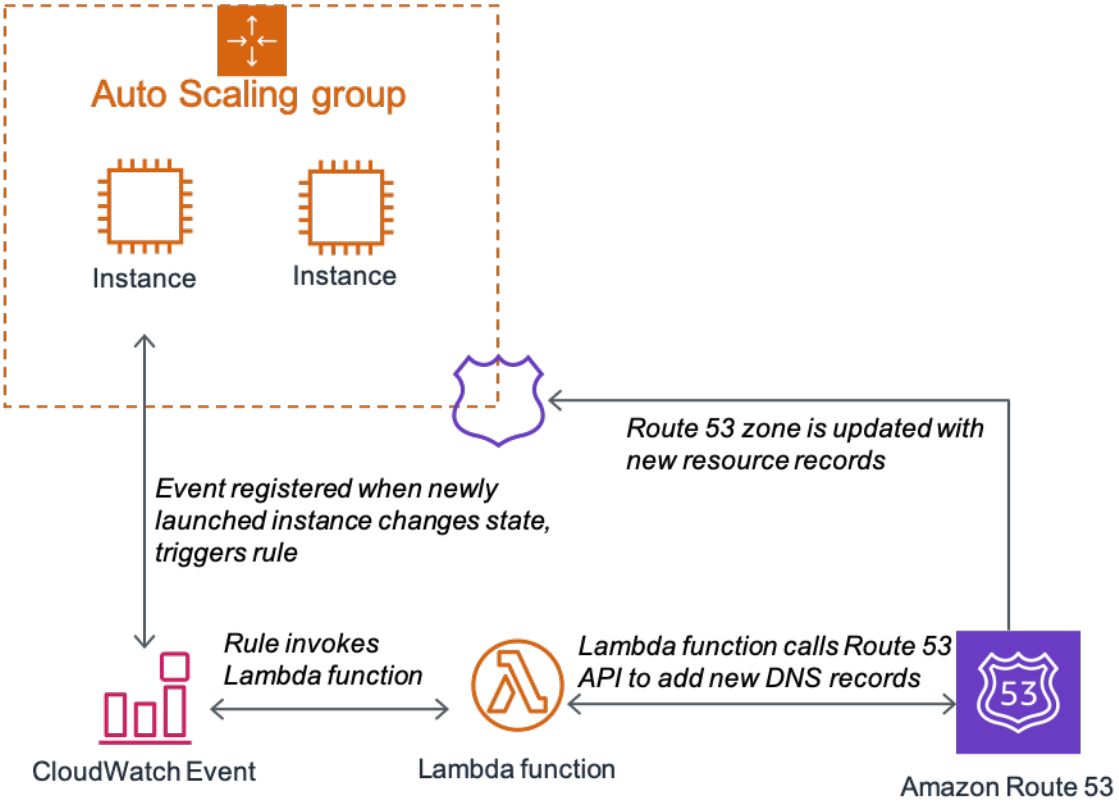

Figura 8: Scalabilità automatica con aggiornamenti dinamici di Amazon Route 53

WebRTC a elevata disponibilità con Kinesis Video Streams

Amazon Kinesis Video Streams offre streaming multimediale in tempo reale tramite WebRTC, consentendo agli utenti di acquisire, elaborare e archiviare flussi multimediali per la riproduzione, l'analisi e l'apprendimento automatico. Questi flussi sono altamente disponibili, scalabili e conformi agli standard WebRTC. Amazon Kinesis Video Streams include gli endpoint di segnalazione WebRTC per una rapida individuazione peer e la creazione di una connessione sicura. Include endpoint Session Traversal Utilities for NAT (STUN) e Traversal Using Relays around NAT (TURN) gestiti per uno scambio in tempo reale di contenuti multimediali tra peer. Inoltre, include un SDK open source gratuito che si integra direttamente con il firmware della videocamera per garantire una comunicazione sicura con gli endpoint Kinesis Video Streams per l'individuazione peer e streaming multimediale. Da ultimo, fornisce librerie client per Android, iOS e JavaScript che consentono ai giocatori Web e di dispositivi mobili di essere conformi con WebRTC per scoprire e connettersi in modo sicuro a una fotocamera per lo streaming multimediale e la comunicazione bidirezionale.

Trunking SIP a elevata disponibilità con Amazon Chime Voice Connector

Amazon Chime Voice Connector offre un servizio trunking SIP con pagamento in base al consumo che consente alle aziende di effettuare e/o ricevere telefonate economiche e sicure tramite i propri sistemi telefonici. Amazon Chime Voice Connector è un’alternativa low-cost ai fornitori di servizi trunking SIP o Integrated Services Digital Network (ISDN) Primary Rate Interfaces (PRIs). I clienti hanno la possibilità di abilitare le telefonate in entrata, in uscita o entrambe. Il servizio sfrutta la rete di AWS per fornire un’esperienza di chiamata ampiamente disponibile su più regioni AWS. Puoi trasmettere in streaming l'audio dalle chiamate telefoniche trunking SIP o dai feed SIPREC (SIP based media recording) inoltrati ad Amazon Kinesis Video Streams per ottenere informazioni dettagliate dalle chiamate di lavoro in tempo reale. Puoi creare rapidamente applicazioni per l'analisi audio tramite l'integrazione con Amazon Transcribe e altre librerie di machine learning comuni.

Best practice sul campo

Questa sezione ha lo scopo di riassumere le best practice implementate da alcuni dei clienti AWS più importanti e di maggior successo che eseguono carichi di lavoro SIP (Session Initiation Protocol) di grandi dimensioni in tempo reale. I clienti AWS che desiderano eseguire la propria infrastruttura SIP sul cloud pubblico troveranno queste best practice preziose in quanto possono contribuire ad aumentare l'affidabilità e la resilienza del sistema in caso di diversi tipi di guasti. Sebbene alcune di queste best practice siano specifiche di SIP, la maggior parte di esse è applicabile a qualsiasi applicazione di comunicazione in tempo reale in esecuzione su AWS.

Argomenti

Creare un overlay SIP

AWS dispone di una dorsale di rete (backbone) robusta, scalabile e ridondante che fornisce connettività tra diverse regioni. Quando un evento di rete, ad esempio un taglio di fibre, degrada un collegamento della dorsale di rete (backbone) AWS, il traffico viene rapidamente trasferito su percorsi ridondanti utilizzando protocolli di routing a livello di rete, come BGP. Questa ingegneria del traffico a livello di rete è una scatola nera per i clienti AWS e la maggior parte di essi non si accorge nemmeno degli eventi di failover. Tuttavia, i clienti che eseguono carichi di lavoro in tempo reale, come traffico voce, video di alta qualità e messaggi a bassa latenza, a volte notano questi eventi. Quindi, come può un cliente AWS implementare la propria ingegneria del traffico in aggiunta a ciò che viene fornito da AWS a livello di rete? La soluzione sta nel distribuire l'infrastruttura SIP in molte regioni AWS diverse. Come parte delle funzionalità di controllo delle chiamate, SIP offre anche la possibilità di instradare le chiamate attraverso proxy SIP specifici.

Figura 9: Utilizzo del routing SIP per ignorare il routing di rete

Nella Figura 9, l'infrastruttura SIP (rappresentata da punti verdi) è in esecuzione in tutte e quattro le regioni degli Stati Uniti. Le linee blu non sono altro che una rappresentazione immaginaria della dorsale di rete (backbone) AWS. Se non viene implementato alcun routing SIP, una chiamata che ha origine nella costa occidentale degli Stati Uniti, destinata alla costa orientale degli Stati Uniti, passa attraverso la dorsale di rete (backbone) che collega direttamente le regioni dell'Oregon e della Virginia. Il diagramma mostra come un cliente potrebbe ignorare il routing a livello di rete ed effettuare la stessa chiamata tra Oregon e Virginia instradata attraverso la California utilizzando il routing SIP. Questo tipo di ingegneria del traffico SIP può essere implementato utilizzando proxy SIP e gateway multimediali in base a metriche di rete come ritrasmissioni SIP e preferenze aziendali specifiche del cliente.

Eseguire un monitoraggio dettagliato

Gli utenti finali di applicazioni voce e video in tempo reale si aspettano lo stesso livello di prestazioni che raggiungono con i servizi di telefonia tradizionale. Quindi, quando si verificano problemi con un'applicazione, si finisce per danneggiare la reputazione del provider. Per essere proattivi piuttosto che reattivi, è fondamentale implementare un monitoraggio dettagliato in ogni parte del sistema che serve gli utenti finali.

Figura 10: Utilizzo di SIPp per monitorare l'infrastruttura VoIP

Sono disponibili molti strumenti open source come iPerf o SiPp e VOIPMonitor, che possono essere utilizzati per monitorare il traffico SIP/RTP. Nell'esempio precedente, i nodi che eseguono SIPp in modalità client e server misurano parametri SIP come chiamate riuscite e ritrasmissioni SIP tra tutte e quattro le regioni AWS statunitensi. Questi parametri possono quindi essere esportati in Amazon CloudWatch utilizzando uno script personalizzato. Utilizzando CloudWatch, i clienti possono creare allarmi su questi parametri personalizzati in base a un determinato valore di soglia. È quindi possibile intraprendere azioni di correzione automatiche o manuali in base allo stato di questi allarmi CloudWatch.

Per i clienti che non vogliono allocare le risorse ingegneristiche necessarie per sviluppare e mantenere un sistema di monitoraggio personalizzato, sono disponibili sul mercato molte buone soluzioni di monitoraggio VoIP, come ThousandEyes. Un esempio di azione correttiva è la modifica del routing SIP in base all'aumento delle ritrasmissioni SIP.

Usare il DNS per il bilanciamento del carico e gli IP mobili per il failover

I client di telefonia IP che supportano la funzionalità DNS SRV possono utilizzare in modo efficiente la ridondanza integrata nell'infrastruttura bilanciando il carico dei client su diversi SBC/PBX.

Figura 11: Utilizzo di record DNS SRV per bilanciare il carico dei client SIP

La Figura 11 mostra come i clienti possono utilizzare i record SRV per bilanciare il carico del traffico SIP. Qualsiasi client di telefonia IP che supporti lo standard SRV cercherà il prefisso sip._<transport protocol> in un record DNS di tipo SRV. Nell'esempio, la sezione delle risposte del DNS contiene entrambi i PBX in esecuzione in diverse zone di disponibilità AWS. Tuttavia, oltre agli URI degli endpoint, il record SRV contiene tre informazioni aggiuntive:

- Il primo numero è la Priorità (1 nell'esempio precedente). È preferibile una priorità più bassa rispetto a una più alta.
- Il secondo numero è il Peso (10 nell'esempio precedente).
- Il terzo numero è la Porta da utilizzare (5060).

Poiché la priorità è la stessa (1) per entrambi i server PBX, i client utilizzano il peso per bilanciare il carico tra i due PBX. In questo caso, poiché i pesi sono gli stessi, il traffico SIP dovrebbe essere bilanciato equamente tra i due PBX.

Il DNS può essere una buona soluzione per il bilanciamento del carico dei client, ma cosa accade per l'implementazione del failover modificando/aggiornando i record "A" DNS? Questo metodo è sconsigliato a causa dell'incoerenza riscontrata nel comportamento di memorizzazione nella cache DNS all'interno dei nodi client e intermedi. Un approccio migliore per il failover intra-AZ tra un cluster di nodi SIP consiste nell'utilizzare la riassegnazione IP EC2 in cui l'indirizzo IP di un host danneggiato viene immediatamente riassegnato a un host sano utilizzando l'API EC2. Abbinata a una soluzione di monitoraggio e controllo dello stato dettagliata, la riassegnazione IP di un nodo guasto garantisce che il traffico venga trasferito su un host sano in modo tempestivo, riducendo al minimo le interruzioni dell'utente finale.

Utilizzare zone di disponibilità multiple

Ogni regione AWS è suddivisa in zone di disponibilità separate. Ogni zona di disponibilità ha la propria alimentazione, raffreddamento e connettività di rete e costituisce quindi un dominio di guasto isolato. All'interno dei costrutti di AWS, è sempre preferibile che i clienti eseguano i loro carichi di lavoro in più di una zona di disponibilità. Ciò garantisce che le applicazioni dei clienti siano in grado di sopportare anche un guasto completo della zona di disponibilità, un evento molto raro di per sé. Questo suggerimento è anche sinonimo di infrastruttura SIP in tempo reale.

Figura 12: Gestione dell'errore della zona di disponibilità

Supponiamo che un evento catastrofico (come un uragano di categoria 5) causi un'interruzione completa della zona di disponibilità nella regione us-east-1. Con l'infrastruttura in esecuzione come mostrato nel diagramma, tutti i client SIP originariamente registrati con i nodi nella zona di disponibilità con guasto devono registrarsi nuovamente con i nodi SIP in esecuzione nella zona di disponibilità #2. Verifica questo comportamento con i tuoi client/telefoni SIP per assicurarti che sia supportato. Sebbene le chiamate SIP attive al momento dell'interruzione della zona di disponibilità vengano perse, tutte le nuove chiamate vengono instradate attraverso la zona di disponibilità #2.

Per riassumere, i record DNS SRV devono indirizzare il client a più record "A", uno per ogni zona di disponibilità. Ciascuno di questi record "A" deve, a sua volta, puntare a più indirizzi IP di SBC/PBX in quella zona di disponibilità fornendo resilienza sia intra- che inter-AZ. Il failover intra e inter-AZ può essere implementato utilizzando la riassegnazione IP se gli IP sono pubblici. Gli IP privati, tuttavia, non possono essere riassegnati tra le zone di disponibilità. Se un cliente utilizza un indirizzo IP privato, deve fare affidamento sui client SIP che si registrano nuovamente con il sistema SBC/PBX di backup per il failover intra-AZ.

Mantenere il traffico all'interno di una zona di disponibilità e utilizzare i gruppi di posizionamento EC2

Conosciuta anche come affinità della zona di disponibilità, questa best practice si applica anche al raro caso di guasto di una zona di disponibilità completa. Si consiglia di eliminare il traffico tra zone di disponibilità in modo tale che qualsiasi tipo di traffico SIP o RTP che entra in una zona di disponibilità rimanga in quella zona di disponibilità fino a quando non esce dalla regione.

Figura 13: Affinità della zona di disponibilità (al massimo, il 50% delle chiamate attive viene perso)

La Figura 13 mostra un'architettura semplificata che utilizza l'affinità della zona di disponibilità. Il vantaggio comparativo di questo approccio diventa chiaro se si tiene conto degli effetti di un'interruzione completa della zona di disponibilità. Come illustrato nel diagramma, se la zona di disponibilità #2 viene persa, il 50% delle chiamate attive ne risentirà al massimo (supponendo un bilanciamento del carico uguale tra le zone di disponibilità). Se l'affinità delle zone di disponibilità non fosse stata implementata, alcune chiamate fluirebbero tra le zone di disponibilità in una regione e il guasto interesserebbe con molte probabilità più del 50% delle chiamate attive.

Inoltre, per ridurre al minimo la latenza per il traffico, consigliamo di considerare l'utilizzo di gruppi di posizionamento EC2 all'interno di ciascuna zona di disponibilità. Le istanze lanciate all'interno dello stesso gruppo di posizionamento EC2 hanno una larghezza di banda più elevata e una latenza ridotta poiché EC2 garantisce la prossimità di rete di queste istanze l'una rispetto all'altra.

Utilizzare tipi di istanze EC2 per reti avanzate

La scelta del tipo di istanza giusto su Amazon EC2 garantisce l'affidabilità del sistema e un uso efficiente dell'infrastruttura. EC2 offre un'ampia gamma di tipi di istanze ottimizzati per soddisfare diversi casi d'uso. I tipi di istanze comprendono diverse combinazioni di capacità di CPU, memoria, archiviazione e rete, offrendo la flessibilità di poter scegliere la combinazione di risorse adeguata per le proprie applicazioni. Questi tipi di istanze di rete avanzate assicurano che i carichi di lavoro SIP in esecuzione su di esse abbiano accesso a una larghezza di banda costante e una latenza aggregata relativamente inferiore. Una recente aggiunta ad Amazon EC2 è la disponibilità dell'Elastic Network Adapter (ENA) che fornisce fino a 100 Gbps di larghezza di banda. Il catalogo più recente dei tipi di istanze EC2 e delle funzionalità associate è disponibile nella pagina dei tipi di istanze EC2.

Per la maggior parte dei clienti, l'ultima generazione di istanze ottimizzate per il calcolo dovrebbe fornire il miglior rapporto qualità-prezzo. Ad esempio, il C5N supporta il nuovo adattatore di rete elastico con larghezza di banda fino a 100 Gbps con milioni di pacchetti al secondo (PPS). La maggior parte delle applicazioni in tempo reale trarrebbe vantaggio dall'utilizzo dell'Intel Data Plane Developer Kit (DPDK) che può migliorare notevolmente l'elaborazione dei pacchetti di rete

Tuttavia, è sempre consigliabile confrontare i vari tipi di istanze EC2 in base alle proprie esigenze per vedere quale tipo di istanza è più adatta al caso specifico. Il benchmarking consente inoltre di trovare altri parametri di configurazione, come il numero massimo di chiamate che un determinato tipo di istanza può elaborare alla volta.

Considerazioni sulla sicurezza

I componenti dell'applicazione RTC vengono in genere eseguiti direttamente su istanze Amazon EC2 con accesso a Internet. Oltre al TCP, i flussi utilizzano protocolli come UDP e SIP. In questi casi, AWS Shield Standard protegge le istanze di Amazon EC2 dagli attacchi DDoS a livello di infrastruttura comune (Layer 3 e 4), come attacchi di riflessione UDP, riflessione DNS, riflessione NTP, riflessione SSDP e così via. AWS Shield Standard utilizza varie tecniche come il traffic shaping basato sulle priorità che vengono attivate automaticamente quando viene rilevata una firma di attacco DDoS ben definita.

AWS fornisce anche una protezione avanzata contro attacchi DDoS di grandi dimensioni e sofisticati per queste applicazioni abilitando AWS Shield Advanced sugli indirizzi IP elastici. AWS Shield Advanced fornisce un rilevamento DDoS avanzato che rileva automaticamente il tipo di risorsa AWS e le dimensioni dell'istanza EC2 e applica le opportune misure di mitigazione predefinite con protezioni contro i flood SYN o UDP. Con AWS Shield Advanced, i clienti possono anche creare profili di mitigazione personalizzati coinvolgendo il team di risposta DDoS (DRT) di AWS 24 ore su 24, 7 giorni su 7. AWS Shield Advanced garantisce inoltre che durante un attacco DDoS, tutte le liste di controllo accessi (ACL) di rete Amazon VPC vengano applicate automaticamente al confine della rete AWS, fornendo accesso a larghezza di banda aggiuntiva e capacità di scrubbing per mitigare gli attacchi DDoS volumetrici di grandi dimensioni.

Conclusione

I carichi di lavoro di comunicazione in tempo reale (RTC) possono essere distribuiti su Amazon Web Services (AWS) per ottenere scalabilità, elasticità ed elevata disponibilità soddisfacendo i requisiti chiave. Oggi, diversi clienti utilizzano AWS, i suoi partner e soluzioni open source per eseguire carichi di lavoro RTC con costi ridotti e agilità più rapida, oltre a un impatto globale ridotto.

Le architetture di riferimento e le best practice fornite in questo Whitepaper possono aiutare i clienti a configurare con successo carichi di lavoro RTC su AWS e ottimizzare le soluzioni per soddisfare i requisiti degli utenti finali ottimizzando al contempo per il cloud.

Collaboratori

Hanno contribuito alla stesura di questo documento:

- Ahmad Khan, Senior Solutions Architect, Amazon Web Services
- Tipu Qureshi, Principal Engineer, AWS Support, Amazon Web Services
- Hasan Khan, Senior Technical Account Manager, Amazon Web Services
- Shoma Chakravarty, WW Technical Leader, Telecom, Amazon Web Services

Revisioni del documento

Per ricevere una notifica sugli aggiornamenti di questo whitepaper, iscriviti al feed RSS.

update-history-change	update-history-description	update-history-date
Whitepaper aggiornato	Aggiornato per i servizi e le funzionalità più recenti.	13 febbraio 2020
Pubblicazione iniziale	Prima pubblicazione del Whitepaper	1 ottobre 2018

Avvisi

I clienti sono responsabili della propria valutazione autonoma delle informazioni contenute in questo documento. Questo documento: (a) è solo a scopo informativo, (b) mostra le offerte e le pratiche attuali dei prodotti AWS soggette a modifiche senza preavviso, e (c) non crea alcun impegno o garanzia da parte di AWS e dei suoi affiliati, fornitori o licenziatari. I prodotti o servizi AWS sono forniti "così come sono" senza garanzie, dichiarazioni o condizioni di alcun tipo, sia esplicite che implicite. Le responsabilità e gli obblighi di AWS verso i propri clienti sono disciplinati dagli accordi AWS e il presente documento non fa parte né modifica alcun accordo tra AWS e i suoi clienti.