Whitepaper AWS

Comunicazione in tempo reale su AWS

Comunicazione in tempo reale su AWS: Whitepaper AWS

Table of Contents

Sintesi 1
Sintesi 1
Sei Well-Architected? 1
Introduzione 2
Componenti fondamentali dell'architettura RTC 3
Softswitch/PBX 4
Controller di frontiera di sessione (SBC) 4
Connettività PSTN 4
Gateway PSTN 4
Tronco SIP 4
Gateway multimediale (transcodificatore) 5
Notifiche push in WebRTC 5
Gateway WebRTC e WebRTC 6
Disponibilità e scalabilità elevate su AWS 8
Pattern IP mobile per HA tra server con stato di attivazione e standby 8
Applicabilità nelle soluzioni RTC 9
Applicabilità nelle architetture RTC 11
Load Balancing attivo AWS per WebRTC utilizzando Application Load Balancer e Auto Scaling 11
Implementazione per SIP tramite Network Load Balancer o un prodotto Marketplace AWS ... 12
Failover e bilanciamento del carico basati su DNS tra regioni 13
Durabilità dei dati e HA con storage persistente 15
Scalabilità dinamica con AWS Lambda Amazon Route 53 e Amazon EC2 Auto Scaling 16
WebRTC ad alta disponibilità con Amazon Kinesis Video Streams 16
Trunking SIP ad alta disponibilità con Amazon Chime Voice Connector 17
Le migliori pratiche sul campo 18
Crea un overlay SIP 18
Eseguire un monitoraggio dettagliato 19
Usa DNS per il bilanciamento del carico e floating per il failover IPs 20
Utilizza più zone di disponibilità 22
Mantieni il traffico all'interno di una zona di disponibilità e utilizza i gruppi di EC2 collocamento 23
Usa tipi di EC2 istanze di rete avanzati 24
Considerazioni relative alla sicurezza 25

Conclusioni 26
Acronimi 27
Collaboratori 29
Revisioni del documento 30
Note 31
AWS Glossario 32
..... xxxiii

Comunicazione in tempo reale su AWS

Migliori pratiche per la progettazione di carichi di lavoro di comunicazione in tempo reale (RTC) altamente disponibili e scalabili su AWS

Data di pubblicazione: 5 maggio 2022 () Revisioni del documento

Sintesi

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 paper descrive le migliori pratiche per la gestione dei carichi di lavoro di comunicazione in tempo reale (RTC) su Amazon Web Services (AWS) e include architetture di riferimento per soddisfare questi requisiti. Questo paper serve da guida per le persone che hanno dimestichezza con la comunicazione in tempo reale su come ottenere disponibilità e scalabilità elevate per questi carichi di lavoro.

Questo paper include architetture di riferimento che mostrano come configurare i carichi di lavoro RTC e le migliori pratiche per ottimizzare le soluzioni per soddisfare i requisiti degli utenti finali ottimizzando al contempo per il cloud. AWS L'Evolved Packet Core (EPC) non rientra nell'ambito di questo white paper, ma le migliori pratiche qui descritte possono essere applicate alle funzioni di rete virtuali (). VNFs

Sei Well-Architected?

Il AWS Well-Architected Framework ti aiuta a comprendere i pro e i contro delle decisioni che prendi quando crei sistemi nel cloud. I sei pilastri del Framework consentono di apprendere le migliori pratiche architettoniche per progettare e gestire sistemi affidabili, sicuri, efficienti, convenienti e sostenibili. Utilizzando AWS Well-Architected Tool, disponibile gratuitamente in AWS Management Console(è richiesto il login), puoi esaminare i tuoi carichi di lavoro rispetto a queste best practice rispondendo a una serie di domande per ogni pilastro.

Per ulteriori indicazioni e best practice da parte degli esperti per la tua architettura cloud (implementazioni dell'architettura di riferimento, diagrammi e white paper), consulta l'Architecture Center.AWS

Introduzione

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

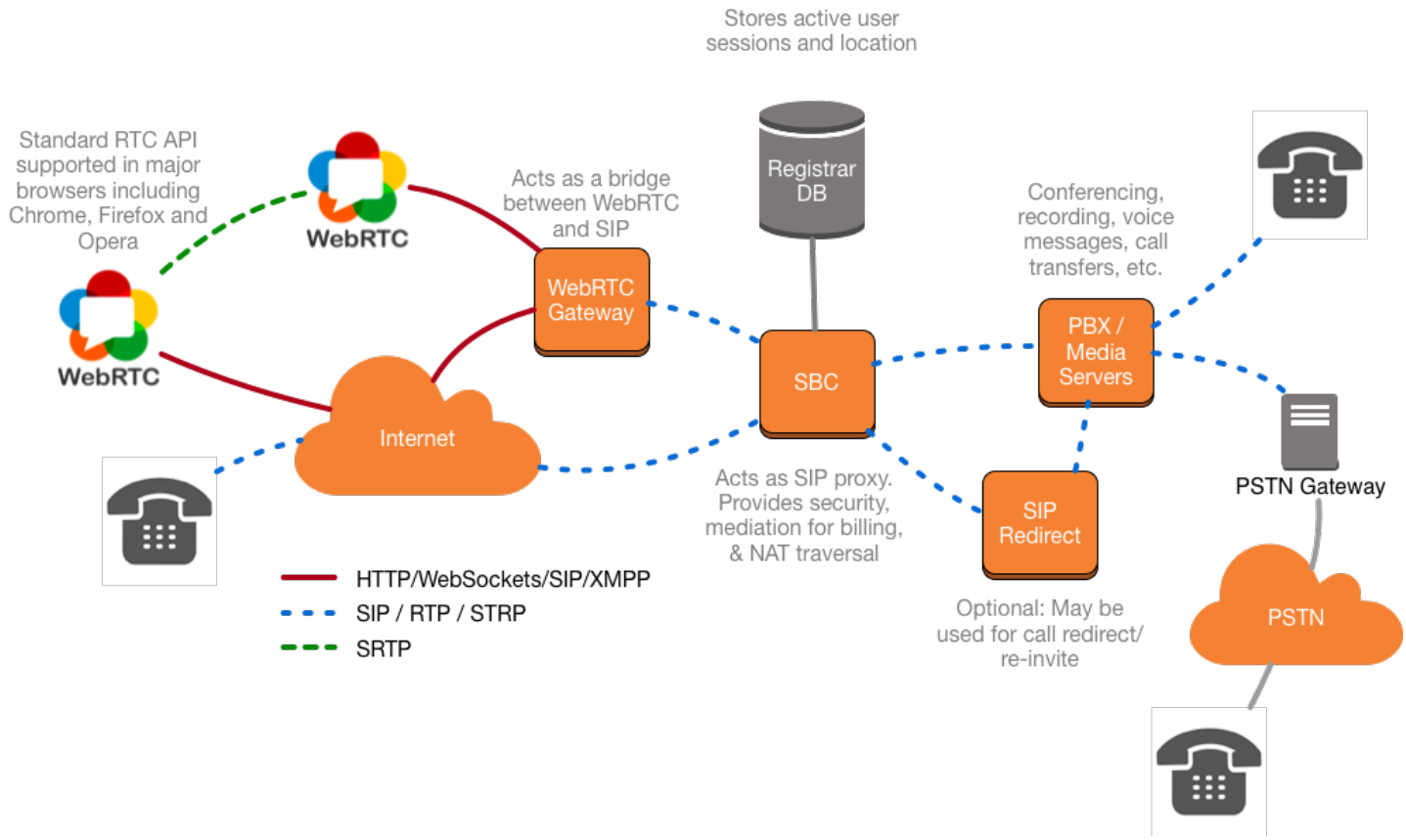
I carichi di lavoro RTC richiedono un ambiente altamente scalabile, resiliente e disponibile. Oggi, i clienti eseguono carichi AWS di lavoro RTC con costi ridotti, maggiore agilità, elasticità e time-to-market.

Componenti fondamentali dell'architettura RTC

Nel settore delle telecomunicazioni, RTC si riferisce comunemente a sessioni multimediali in diretta tra due endpoint con una latenza minima. Queste sessioni potrebbero essere correlate a:

- Una sessione vocale tra due parti (ad esempio un sistema telefonico, un cellulare o Voice over IP (VoIP))
- Messaggistica istantanea (ad esempio chat e Instant Relay Chat (IRC))
- Sessione video in diretta (ad esempio videoconferenze e telepresenza)

Ciascuna delle soluzioni precedenti presenta alcuni componenti in comune (ad esempio componenti che forniscono autenticazione, autorizzazione e controllo degli accessi, transcodifica, buffering e inoltro e così via) e alcuni componenti specifici del tipo di file multimediale trasmesso (come il servizio di trasmissione, il server di messaggistica e le code e così via). Questa sezione si concentra sulla definizione di un sistema RTC basato su voce e video e di tutti i relativi componenti, come illustrato nella figura seguente.



Componenti architettonici essenziali per RTC

SoftSwitch/PBX

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

Session border controller (SBC)

Un session border controller (SBC) si trova ai margini di una rete vocale e tiene traccia di tutto il traffico in entrata e in uscita (sia sul piano di controllo che su quello dati). Una delle responsabilità principali di un SBC è proteggere il sistema vocale da un uso malevolo. L'SBC può essere utilizzato per interconnettersi con i trunk SIP (Session Initiation Protocol) per la connettività esterna. Alcuni offrono SBCs anche funzionalità di transcodifica per la conversione da un formato all'altro. CODECs La maggior parte offre SBCs anche funzionalità di attraversamento degli indirizzi di rete (NAT), che aiutano a garantire che le chiamate vengano stabilite, anche su reti dotate di firewall.

Connettività PSTN

Le soluzioni Voice over IP (VoIP) utilizzano gateway PSTN (Public Switched Telephone Network) e trunk SIP per connettersi alle reti PSTN precedenti.

Gateway PSTN

Il gateway PSTN converte la segnalazione tra SIP e media tra Real Time Transport Protocol (RTP) SS7 e Time Division Multiplexing (TDM) utilizzando la transcodifica CODEC. I gateway PSTN si trovano sempre nella periferia, vicino alla rete PSTN.

Tronco SIP

In un trunk SIP, l'azienda non termina le chiamate su una rete TDM (SS7 basata), ma piuttosto i flussi tra l'azienda e le telecomunicazioni rimangono su IP. La maggior parte dei SIP Trunk viene creata utilizzando. SBCs L'azienda deve concordare le regole di sicurezza predefinite delle telecomunicazioni, ad esempio consentire un determinato intervallo di indirizzi IP, porte e così via.

Gateway multimediale (transcodificatore)

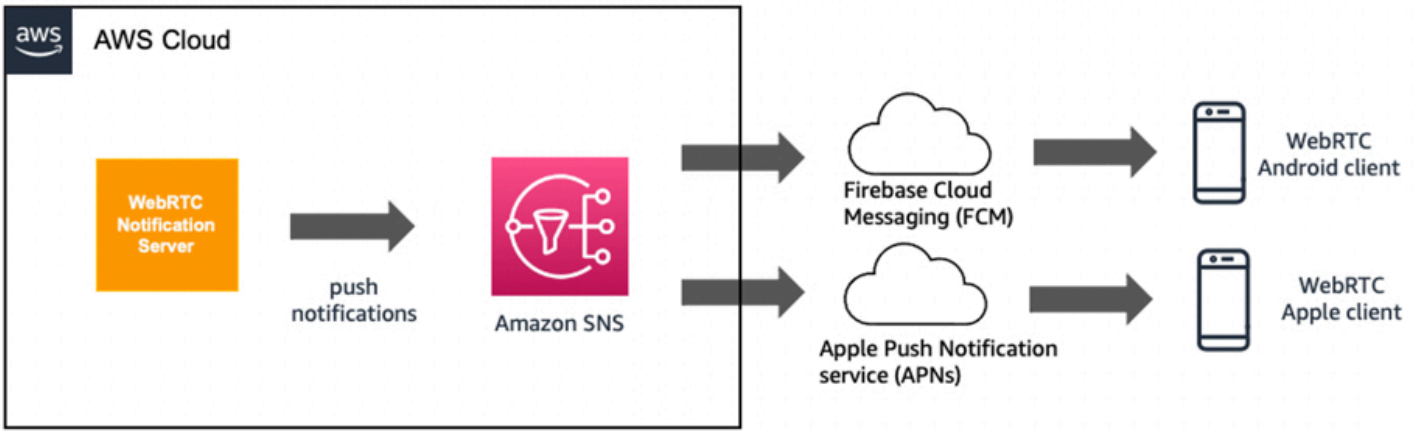
Gli utenti comunicano in tempo reale utilizzando audio e/o video, oltre a dati opzionali e altre informazioni. Per comunicare, i due dispositivi devono essere in grado di concordare un codec di reciproca comprensione per ogni traccia multimediale, in modo da poter comunicare e presentare correttamente i contenuti multimediali condivisi. Tutti i browser compatibili con WebRTC devono supportare il supporto utente per il posizionamento online (OPUS) e G711 per l'audio e il profilo H.264 Constrained Baseline per il video. VP8

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

Notifiche push in WebRTC

Le implementazioni WebRTC sono molto comuni sui dispositivi mobili. A differenza dei browser Web, un dispositivo mobile non può mantenere aperta una connettività websocket per molto tempo. Pertanto, deve fare affidamento sulle notifiche push del server WebRTC per tutte le richieste finali, come chiamate e messaggi.

Amazon Simple Notification Service (Amazon SNS) ti consente di inviare notifiche push alle app sui dispositivi mobili. Queste app potrebbero funzionare su vari sistemi operativi come Apple iOS o Android. La figura seguente mostra una panoramica di alto livello del flusso delle notifiche push, da un server di notifica WebRTC agli endpoint mobili WebRTC.

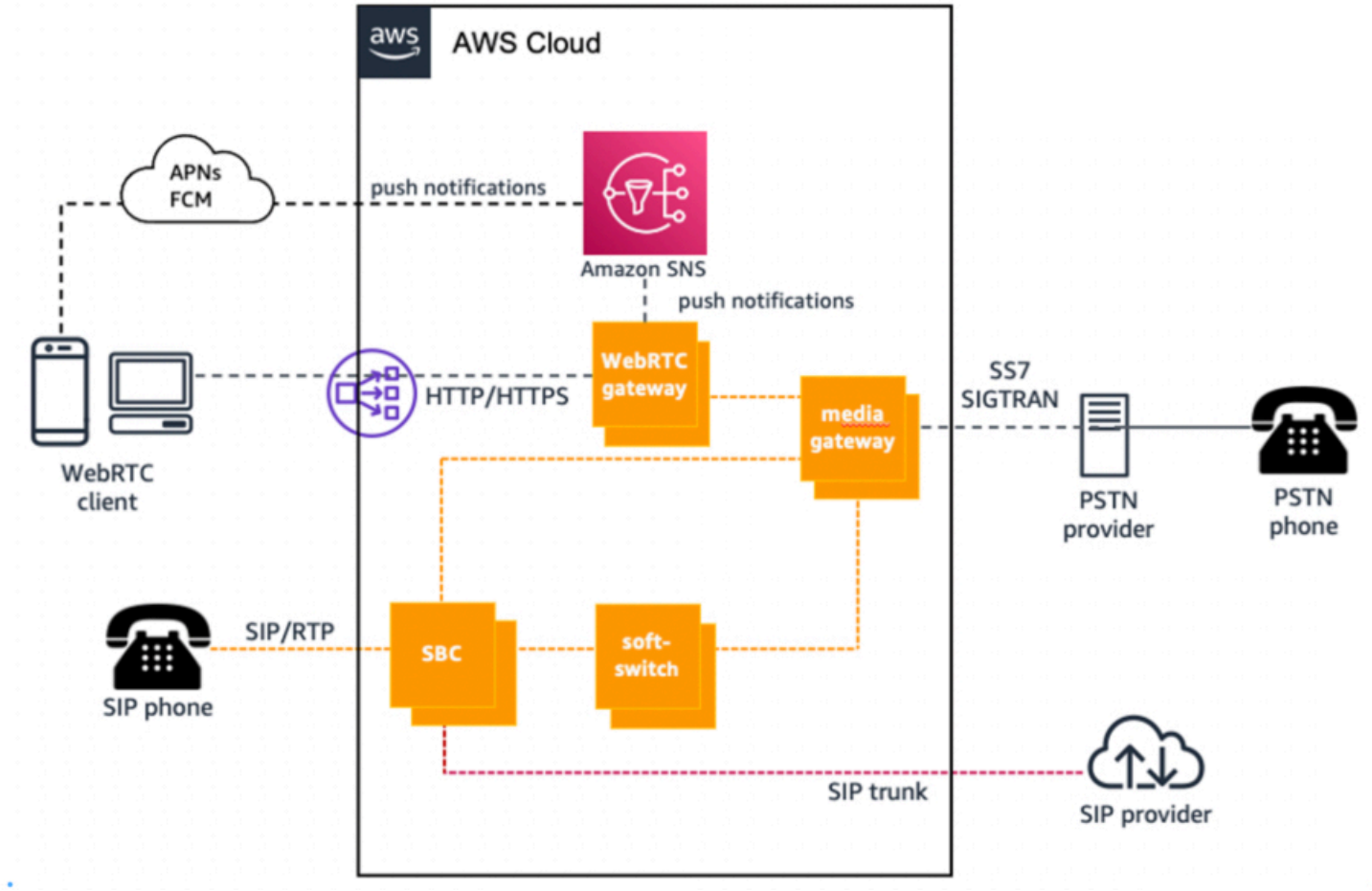


Amazon SNS per notifiche push

Gateway WebRTC e WebRTC

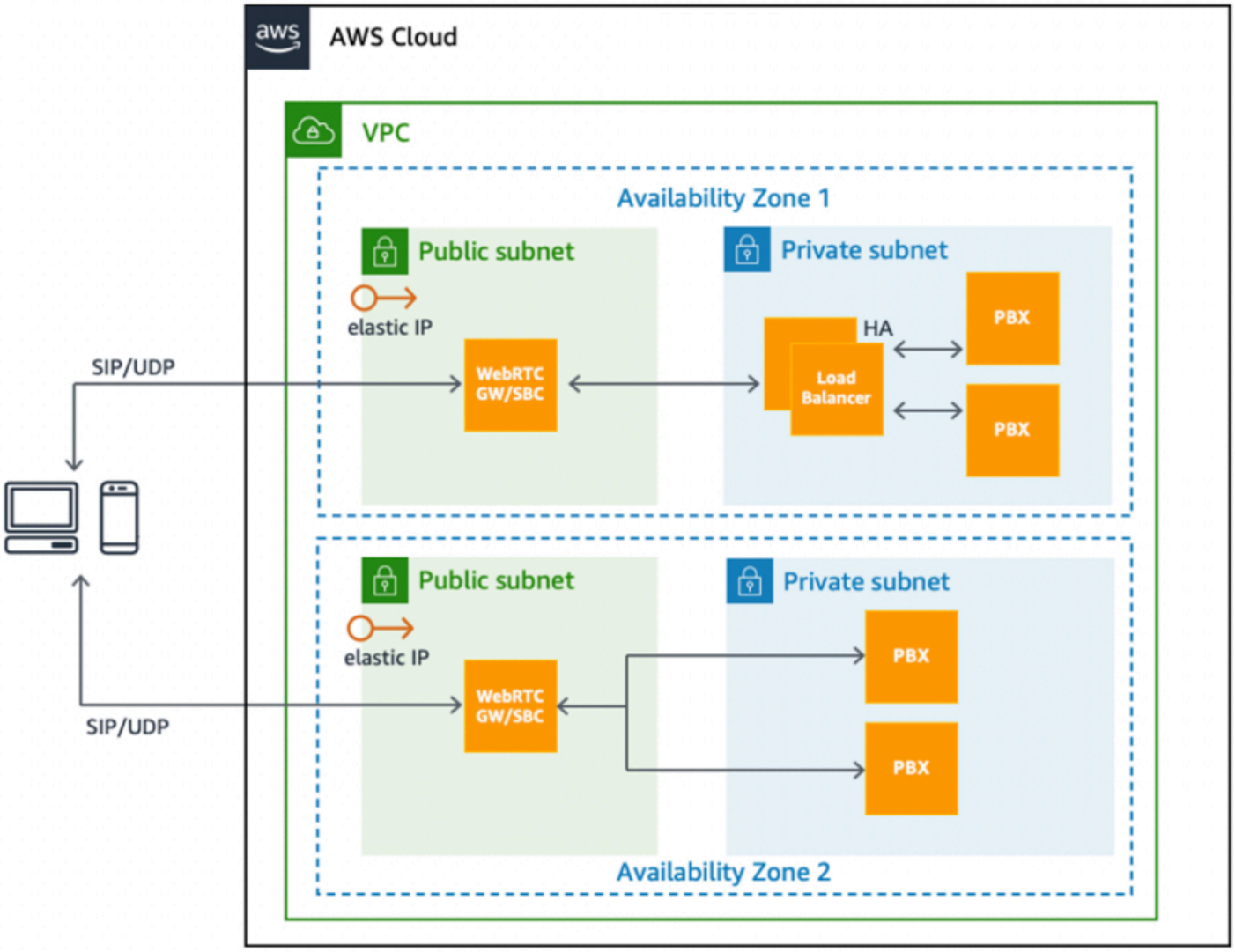
La comunicazione Web in tempo reale (WebRTC) consente di stabilire una chiamata da un browser Web o richiedere risorse dal server di backend utilizzando l'API. La tecnologia è progettata pensando alla tecnologia cloud e pertanto fornisce diverse opzioni APIs che possono essere utilizzate per stabilire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

La figura seguente mostra un modello di progettazione per un'architettura WebRTC ad alta disponibilità. Il traffico in entrata dai client WebRTC è bilanciato da un Application Load Balancer (ALB) con WebRTC in esecuzione su istanze Amazon Elastic Compute Cloud (Amazon) che fanno parte di un gruppo Amazon Auto Scaling. EC2 EC2



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 dati SBCs su Amazon EC2 in modalità attiva-passiva tra le zone di disponibilità, come illustrato nella figura seguente. Qui, un indirizzo IP elastico può essere spostato dinamicamente tra le istanze in caso di errore, laddove il Domain Name Service (DNS) non può essere utilizzato.



Architettura RTC che utilizza Amazon EC2 in un cloud privato virtuale (VPC)

Disponibilità e scalabilità elevate su AWS

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

- Progetta il sistema in modo che non abbia un singolo punto di errore. Utilizza meccanismi automatici di monitoraggio, rilevamento degli errori e failover per componenti stateless e stateful
  - I punti di errore singoli (SPOF) vengono generalmente eliminati con una configurazione di ridondanza N+1 o 2N, in cui N+1 viene ottenuto tramite il bilanciamento del carico tra nodi attivi-attivi e 2N viene ottenuto da una coppia di nodi in configurazione active-standby.
  - AWS offre diversi metodi per raggiungere l'HA attraverso entrambi gli approcci, ad esempio tramite un cluster scalabile con carico bilanciato o presupponendo una coppia attiva/standby.
- Disponibilità corretta dello strumento e del sistema di test.
- Prepara le procedure operative per i meccanismi manuali in grado di rispondere, mitigare e ripristinare i guasti.

Questa sezione si concentra su come evitare un singolo punto di errore utilizzando le funzionalità disponibili su AWS. In particolare, questa sezione descrive un sottoinsieme di AWS funzionalità e modelli di progettazione di base che consentono di creare applicazioni di comunicazione in tempo reale ad alta disponibilità.

Pattern IP mobile per HA tra server con stato attivo e in standby

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

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 offre il meccanismo migliore per l'HA. Ad esempio, i componenti stateful di una soluzione RTC, come il server multimediale o il server di conferenza, o anche un SBC o un server di database, sono adatti per una configurazione attiva e in standby. Un SBC o un server multimediale ha diverse sessioni o canali di lunga durata attivi in un determinato momento e, in caso di guasto dell'istanza SBC active, gli endpoint possono riconnettersi al nodo di standby senza alcuna configurazione lato client a causa dell'IP mobile.

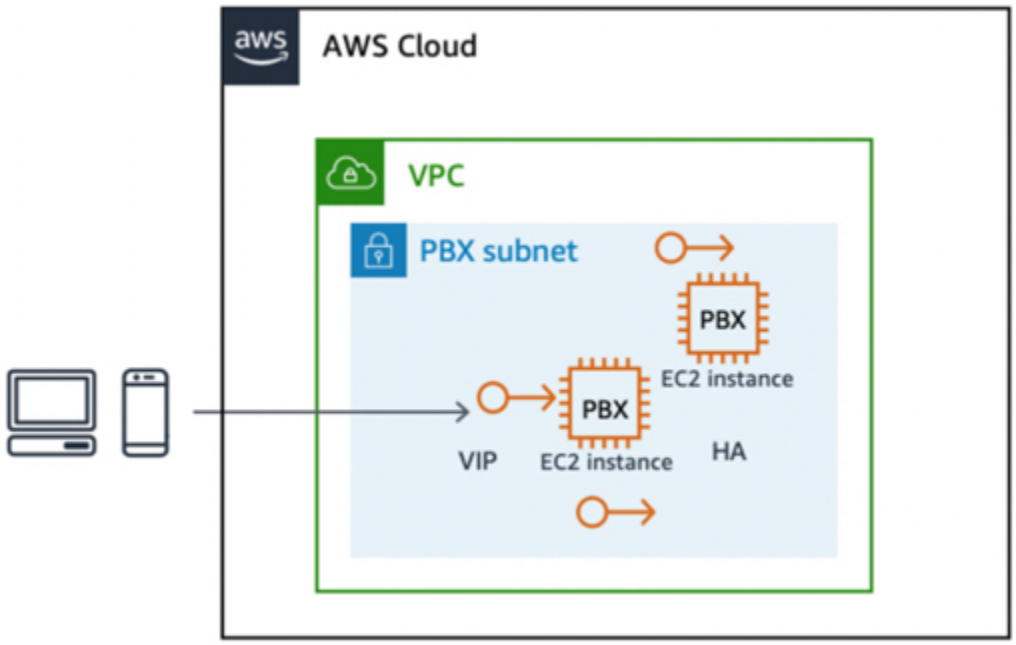
Implementazione su AWS

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

Per implementare il pattern IP mobile su: AWS

1. Avvia due EC2 istanze per assumere i ruoli di nodi primari e secondari, dove per impostazione predefinita si presume che il primario sia in stato attivo.
2. Assegna un indirizzo IP privato secondario aggiuntivo all'istanza principale EC2 .
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 (OS) per aggiungere l'indirizzo IP secondario come alias all'interfaccia di rete principale.
5. L'applicazione deve essere associata a questo indirizzo IP elastico. Nel caso del software Asterisk, è possibile configurare l'associazione tramite le impostazioni SIP avanzate di Asterisk.
6. Esegui uno script di monitoraggio, personalizzato, KeepAlive su Linux, Corosync e così via, su ciascun nodo per monitorare lo stato del nodo peer. In caso di guasto del nodo attivo corrente, il peer rileva l'errore e richiama l'API EC2 Amazon per riassegnare a se stesso l'indirizzo IP privato secondario.

   Pertanto, l'applicazione che era in ascolto sul VIP associata all'indirizzo IP privato secondario diventa disponibile per gli endpoint tramite il nodo di standby.



Failover tra istanze con stato utilizzando un indirizzo IP elastico EC2

Vantaggi

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

Limitazioni ed estensibilità

Questo modello di progettazione è in genere limitato a una singola zona di disponibilità. Può essere implementato in due zone di disponibilità, ma con una variante. In questo caso, l'indirizzo IP elastico mobile viene riassociato tra il nodo attivo e quello di standby in diverse zone di disponibilità tramite l'API di riassociazione degli indirizzi IP elastici disponibile. Nell'implementazione del failover mostrata nella figura precedente, le chiamate in corso vengono interrotte e gli endpoint devono riconnettersi. È possibile estendere questa implementazione con la replica dei dati di sessione sottostanti per garantire un failover senza interruzioni delle sessioni o anche la continuità dei supporti.

Bilanciamento del carico per scalabilità e HA 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 grazie alla natura stateless delle richieste HTTP. In effetti, il bilanciamento del carico è un'opzione valida nel caso di molti componenti dell'applicazione RTC.

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

La decisione su quale nodo dirigere si basa sulle regole di sistema definite nel load balancer, 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
- Basato sulla latenza
- Basato sul carico

Applicabilità nelle architetture RTC

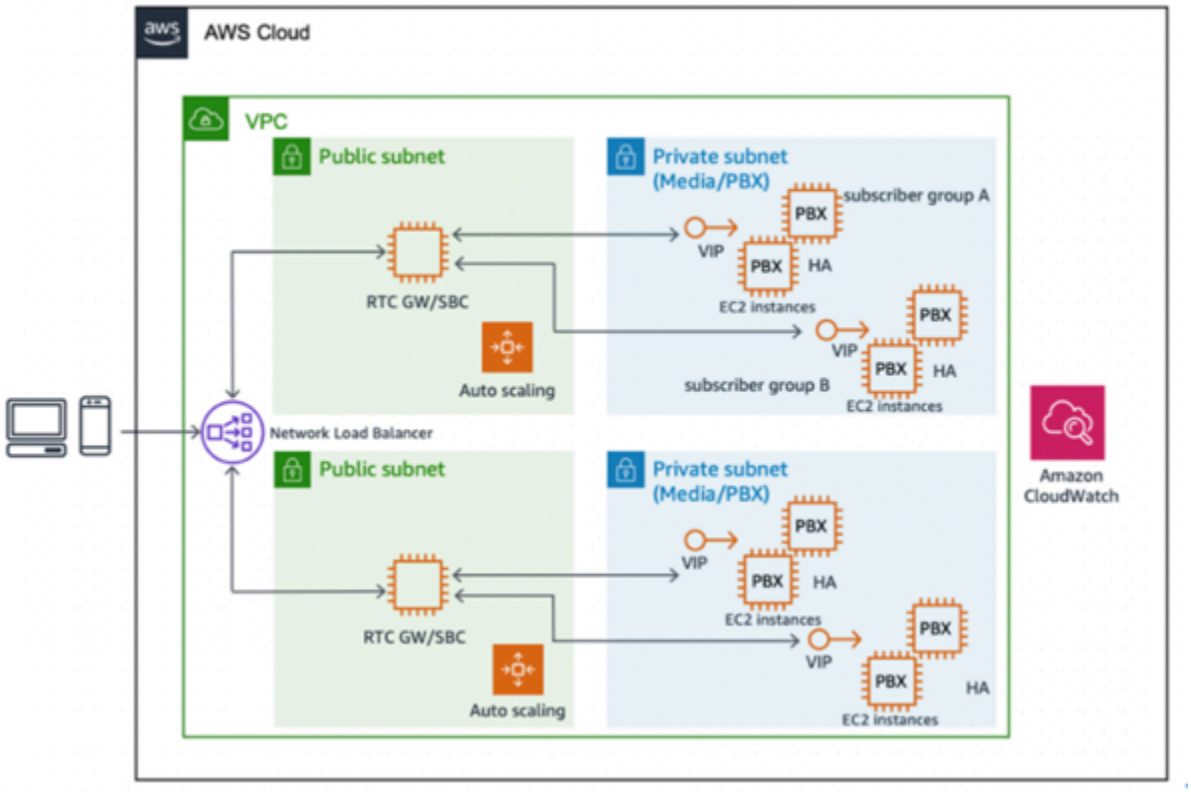
Il protocollo WebRTC consente di bilanciare facilmente il carico dei gateway WebRTC tramite un sistema di bilanciamento del carico basato su HTTP, come Elastic Load Balancing (ELB), Application Load Balancer (ALB) o Network Load Balancer (NLB). Poiché la maggior parte delle implementazioni SIP si basa sul trasporto tramite TCP (Transmission Control Protocol) e UDP (User Datagram Protocol), è necessario un bilanciamento del carico a livello di rete o di connessione con supporto per il traffico basato su TCP e UDP.

Attivazione del bilanciamento del carico AWS per WebRTC tramite Application Load Balancer e Auto Scaling

Nel caso delle comunicazioni basate su WebRTC, Elastic Load Balancing fornisce un sistema di bilanciamento 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 istanze di destinazione associato a Elastic Load Balancing. EC2 Poiché le richieste WebRTC sono stateless, puoi utilizzare Amazon EC2 Auto Scaling per fornire scalabilità, elasticità e alta disponibilità completamente automatizzate e controllabili.

L'Application Load Balancer fornisce un servizio di bilanciamento del carico completamente gestito che è altamente disponibile utilizzando più zone di disponibilità e scalabile. Ciò supporta il bilanciamento del carico delle WebSocket richieste che gestiscono la segnalazione per le applicazioni WebRTC e la comunicazione bidirezionale tra client e server utilizzando una connessione TCP a lunga durata. L'Application Load Balancer supporta anche il routing basato sui contenuti e le sessioni permanenti, instradando le richieste dallo stesso client allo stesso target utilizzando i cookie generati dal load balancer. Se abiliti le sessioni permanenti, lo stesso target riceve la richiesta e può utilizzare il cookie per ripristinare il contesto della sessione.

La figura seguente mostra la topologia di destinazione.



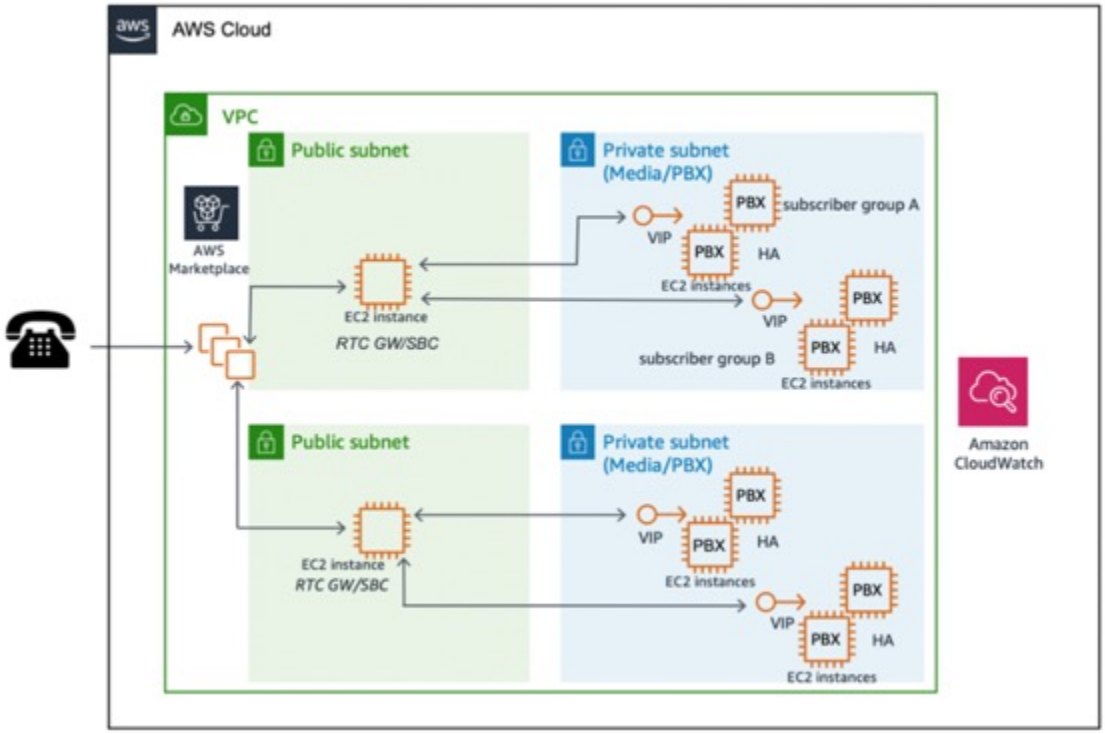
Scalabilità WebRTC e architettura ad alta disponibilità

Implementazione per SIP tramite Network Load Balancer o un prodotto Marketplace AWS

Nel caso delle 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, altamente disponibile, scalabile e prestazionale.

Un Network Load Balancer opera a livello di connessione (Layer four), instradando le connessioni verso destinazioni come EC2 istanze Amazon, 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 estremamente basse. È integrato con altri servizi AWS popolari, come Amazon EC2 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 utilizzare software commerciale (COTS). Marketplace AWS off-the-shelf Marketplace AWS Offre molti prodotti in grado di gestire UDP e altri tipi di bilanciamento del carico di connessione di livello quattro. I COTS in genere includono il supporto per l'alta disponibilità e si integrano comunemente con funzionalità, come Amazon EC2 Auto Scaling, per migliorare ulteriormente la disponibilità e la scalabilità. La figura seguente mostra la topologia di destinazione:



Scalabilità RTC basata su SIP con il prodotto AWS Marketplace

Failover e bilanciamento del carico basati su DNS tra regioni

Amazon Route 53 fornisce un servizio DNS globale che può essere utilizzato come endpoint pubblico o privato per consentire ai client RTC di registrarsi e connettersi con applicazioni multimediali. Con Amazon Route 53, i controlli dello stato del DNS possono essere configurati per indirizzare il traffico verso endpoint integri o per monitorare in modo indipendente lo stato dell'applicazione.

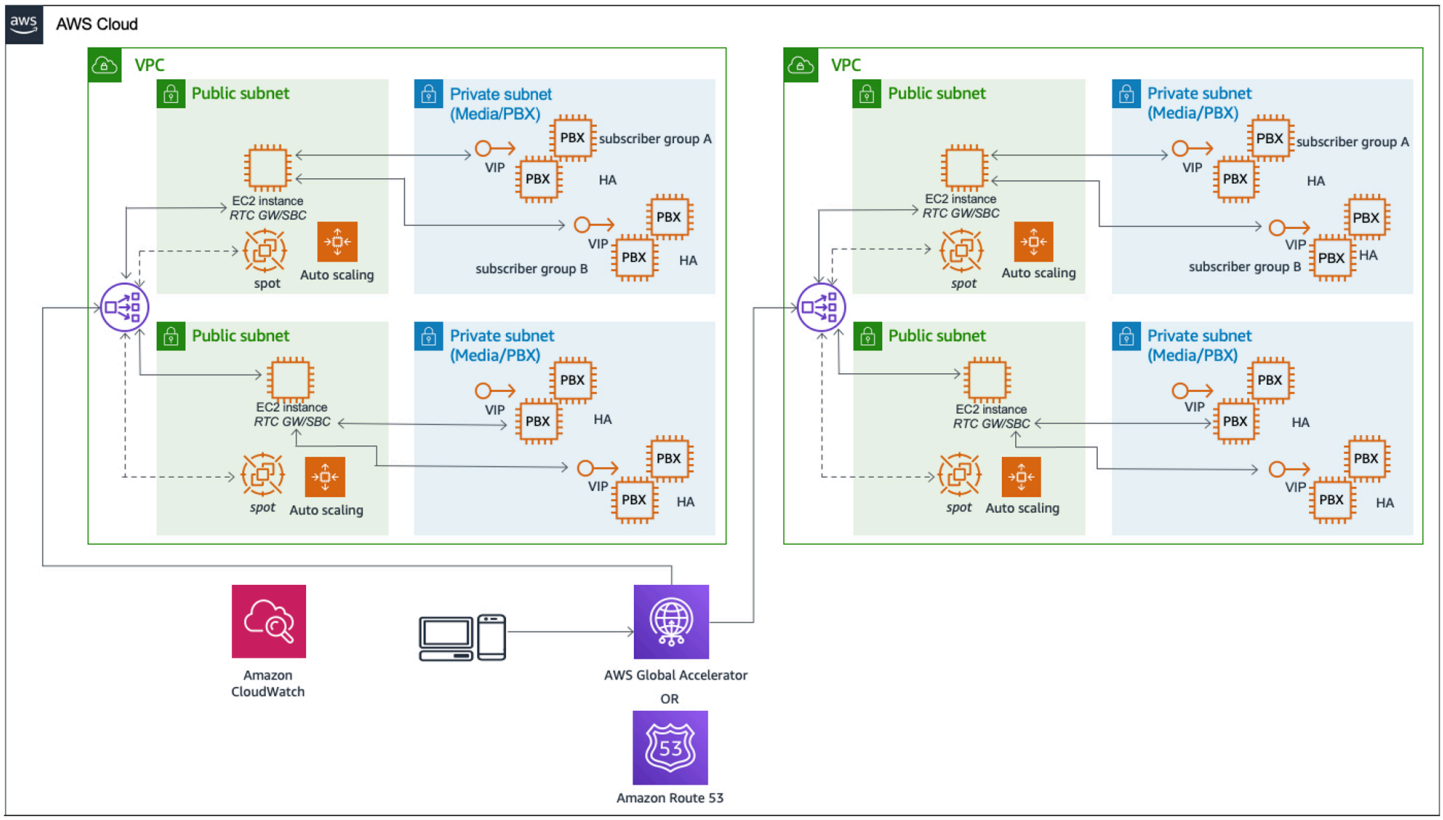
La funzionalità Amazon Route 53 Traffic Flow semplifica la gestione del traffico a livello globale attraverso una varietà di tipi di routing, tra cui routing basato sulla latenza, geodNS, geoproximity e weighted round robin, tutti combinabili con DNS Failover per abilitare una varietà di architetture a bassa latenza e tolleranti ai guasti. Il semplice editor visivo di Amazon Route 53 Traffic Flow ti consente di gestire il modo in cui gli utenti finali vengono indirizzati agli endpoint della tua applicazione, sia in una singola regione AWS che distribuiti in tutto il mondo.

Nel caso di implementazioni globali, la politica di routing basata sulla latenza di Route 53 è particolarmente utile per indirizzare i clienti verso il punto di presenza più vicino per un server multimediale e migliorare la qualità del servizio associato agli scambi di contenuti multimediali in tempo reale.

Tieni presente che per applicare un failover su un nuovo indirizzo DNS, le cache dei client devono essere svuotate. Inoltre, le modifiche DNS possono avere un ritardo in quanto vengono propagate su 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 della configurazione delle politiche 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 tra regioni. AWS Global Acceleratorè 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 accesso fisso agli endpoint delle applicazioni, come Application Load Balancer, Network Load Balancer o istanze EC2 Amazon 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 verso gli endpoint integri più vicini nel caso in cui gli endpoint correnti non funzionino correttamente. Per requisiti di sicurezza aggiuntivi, Accelerated Site-to-Site VPN migliora AWS Global Accelerator le prestazioni delle connessioni VPN instradando in modo intelligente il traffico attraverso la rete globale AWS e le edge location AWS.



Progettazione ad alta disponibilità interregionale con AWS Global Accelerator o Amazon Route 53

Durabilità dei dati e HA con storage persistente

La maggior parte delle applicazioni RTC si basa sullo storage persistente per archiviare e accedere ai dati per l'autenticazione, l'autorizzazione, la contabilità (dati della sessione, record dei dettagli delle chiamate, ecc.), il monitoraggio operativo e la registrazione. In un data center tradizionale, garantire disponibilità e durabilità elevate per i componenti di storage persistenti (database, file system e così via) richiede in genere un lavoro impegnativo tramite la configurazione di una rete SAN (Storage Area Network), la progettazione di Redundant Array of Independent Disks (RAID) e processi di backup, ripristino ed elaborazione di failover. Semplifica e migliora Cloud AWS notevolmente le pratiche tradizionali dei data center relative alla durabilità e alla disponibilità dei dati.

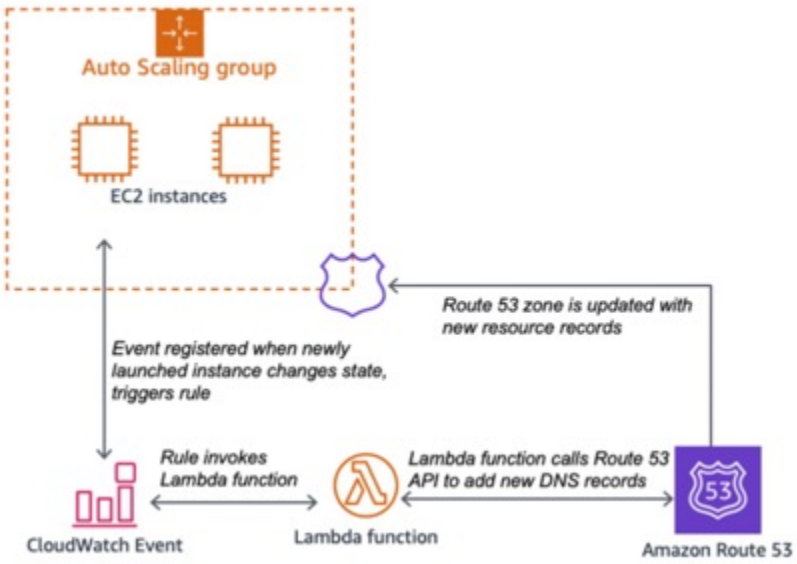
Per lo storage di oggetti e file, AWS servizi come Amazon Simple Storage Service (Amazon S3) e Amazon Elastic File System (Amazon EFS) forniscono disponibilità e scalabilità elevate gestite. Amazon S3 ha una durabilità dei dati del 99,99999% (11 nove).

Per lo storage dei dati transazionali, i clienti hanno la possibilità di sfruttare il servizio Amazon Relational Database Service (Amazon RDS) completamente gestito che supporta Amazon Aurora,

PostgreSQL, MySQL, MariaDB, Oracle e Microsoft SQL Server con implementazioni ad alta disponibilità. Per la funzione di registrazione, il profilo dell'abbonato o l'archiviazione dei registri contabili (ad esempio CDRs), Amazon RDS offre un'opzione con tolleranza ai guasti, altamente disponibile e scalabile.

Scalabilità dinamica con AWS Lambda Amazon Route 53 e Amazon EC2 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 per il ciclo di vita con scalabilità automatica con Amazon Events CloudWatch , Amazon Route 53 e funzioni. AWS Lambda AWS Lambda le funzioni possono incorporare qualsiasi azione o logica. La figura seguente mostra come queste funzionalità concatenate possono migliorare l'affidabilità e la scalabilità del sistema con l'automazione.



Scalabilità automatica con aggiornamenti dinamici ad Amazon Route 53

WebRTC ad alta disponibilità con Amazon Kinesis Video Streams

Amazon Kinesis Video Streams offre streaming multimediale in tempo reale tramite WebRTC, che consente agli utenti di acquisire, elaborare e archiviare flussi multimediali per la riproduzione, l'analisi e l'apprendimento automatico. Questi stream sono altamente disponibili, scalabili e conformi agli standard WebRTC. Amazon Kinesis Video Streams include un endpoint di segnalazione

WebRTC per una rapida scoperta tra pari e la creazione di connessioni sicure. Include gli endpoint Session Traversal Utilities for NAT (STUN) e Traversal Using Relays around NAT (TURN) gestiti per lo scambio di contenuti multimediali in tempo reale tra colleghi. Include anche un SDK open source gratuito che si integra direttamente con il firmware della fotocamera per consentire una comunicazione sicura con gli endpoint Amazon Kinesis Video Streams, consentendo il peer discovery e lo streaming multimediale. Infine, fornisce librerie client per Android, iOS e JavaScript che consentono ai lettori mobili e web compatibili con WebRTC di rilevare e connettersi in modo sicuro a un dispositivo con fotocamera per lo streaming multimediale e la comunicazione bidirezionale.

Trunking SIP ad alta disponibilità con Amazon Chime Voice Connector

Amazon Chime Voice Connector offre un servizio di trunking pay-as-you-go SIP che consente alle aziende di effettuare e/o ricevere chiamate telefoniche sicure ed economiche con i propri sistemi telefonici. Amazon Chime Voice Connector è un'alternativa a basso costo ai trunk SIP dei provider di servizi o alle interfacce ISDN (Integrated Services Digital Network) (). PRIs I clienti hanno la possibilità di abilitare le chiamate in entrata, le chiamate in uscita o entrambe.

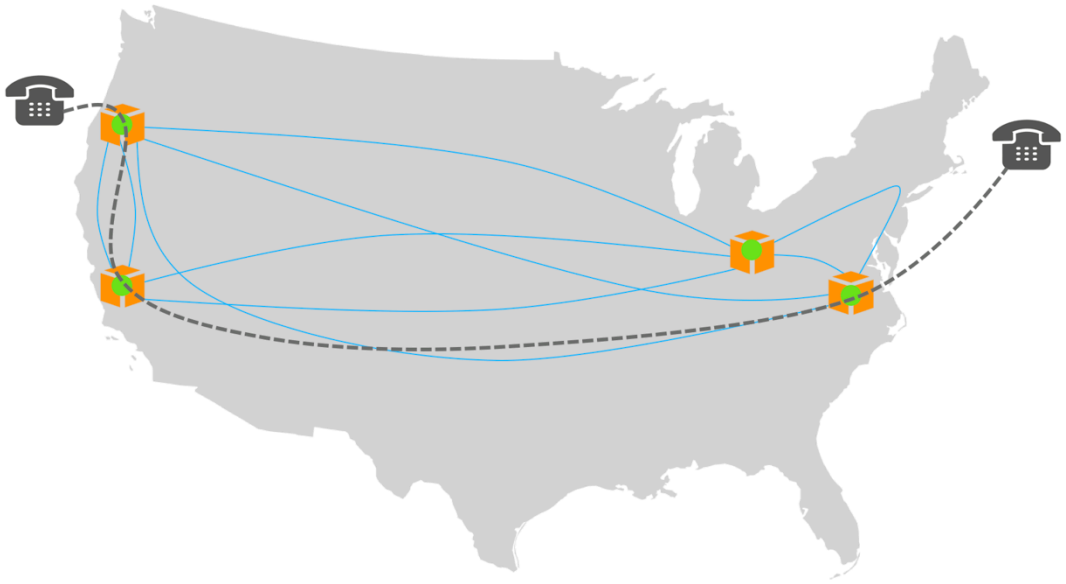
Il servizio utilizza la AWS rete per offrire un'esperienza di chiamata ad alta disponibilità su più canali. Regioni AWS Puoi trasmettere l'audio dalle chiamate telefoniche con trunking SIP o dai feed di registrazione multimediale basati su SIP (SIPREC) inoltrati ad Amazon Kinesis Video Streams per ottenere informazioni dettagliate dalle chiamate di lavoro in tempo reale. Puoi creare rapidamente applicazioni per l'analisi audio attraverso l'integrazione con Amazon Transcribe e altre librerie di machine learning comuni.

Le migliori pratiche sul campo

Questa sezione riassume le best practice implementate da alcuni dei AWS clienti più importanti e di maggior successo che eseguono grandi carichi di lavoro SIP (Session Initiation Protocol) in tempo reale. AWS i clienti che desiderano eseguire la propria infrastruttura SIP nel cloud pubblico troverebbero utili queste best practice 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 per SIP, la maggior parte di esse è applicabile a qualsiasi applicazione di comunicazione in tempo reale in esecuzione. AWS

Crea un overlay SIP

AWS dispone di una dorsale di rete robusta, scalabile e ridondante che fornisce connettività tra diversi. Regioni AWS Quando un evento di rete, ad esempio un'interruzione della fibra, danneggia un collegamento AWS backbone, il traffico viene rapidamente trasferito su percorsi ridondanti utilizzando protocolli di routing a livello di rete, come Border Gateway Protocol (BGP). Questa ingegneria del traffico a livello di rete è una scatola nera per i AWS clienti e la maggior parte non si accorge nemmeno di questi eventi di failover. Tuttavia, i clienti che eseguono carichi di lavoro in tempo reale, ad esempio voce, video di alta qualità e messaggistica a bassa latenza, a volte notano questi eventi. Quindi, come può un AWS cliente implementare la propria ingegneria del traffico oltre a ciò che viene fornito AWS a livello di rete? La soluzione prevede l'implementazione dell'infrastruttura SIP in molti modi diversi. Regioni AWS Come parte delle funzionalità di controllo delle chiamate, SIP offre anche la possibilità di instradare le chiamate attraverso proxy SIP specifici.

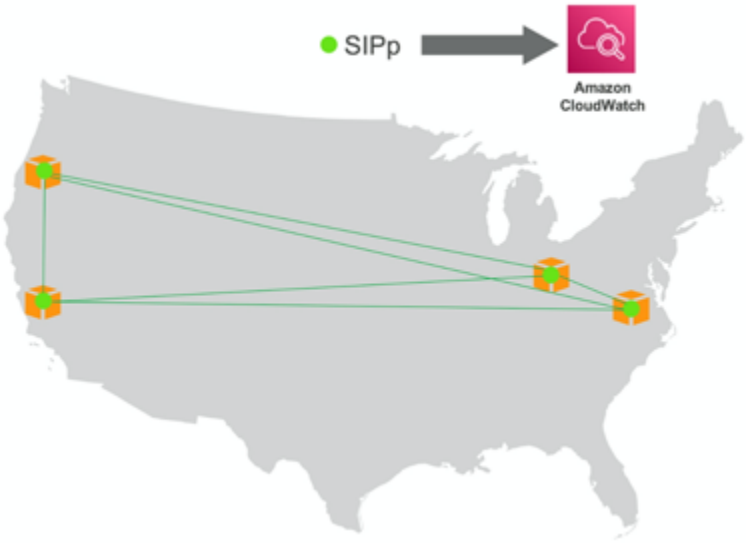


Utilizzo del routing SIP per sostituire il routing di rete

Nella figura precedente, l'infrastruttura SIP (rappresentata da punti verdi all'interno dei cubi) è in funzione in tutte e quattro le regioni degli Stati Uniti. Le linee blu continue rappresentano una rappresentazione fittizia della spina dorsale. AWS Se non viene implementato alcun routing SIP, una chiamata proveniente dalla costa occidentale degli Stati Uniti e destinata alla costa orientale degli Stati Uniti passa attraverso il collegamento dorsale 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 indirizzata attraverso la California utilizzando il routing SIP. Questo tipo di ingegneria del traffico SIP può essere implementata utilizzando proxy SIP e gateway multimediali in base a metriche di rete come le ritrasmissioni SIP e le preferenze aziendali specifiche del cliente.

Esegui un monitoraggio dettagliato

Gli utenti finali di applicazioni vocali e video in tempo reale si aspettano lo stesso livello di prestazioni dei servizi di telefonia tradizionali. Pertanto, quando riscontrano problemi con un'applicazione, finiscono per danneggiare la reputazione del provider. Per essere proattivi anziché reattivi, è fondamentale implementare un monitoraggio dettagliato in ogni parte del sistema che serve gli utenti finali.



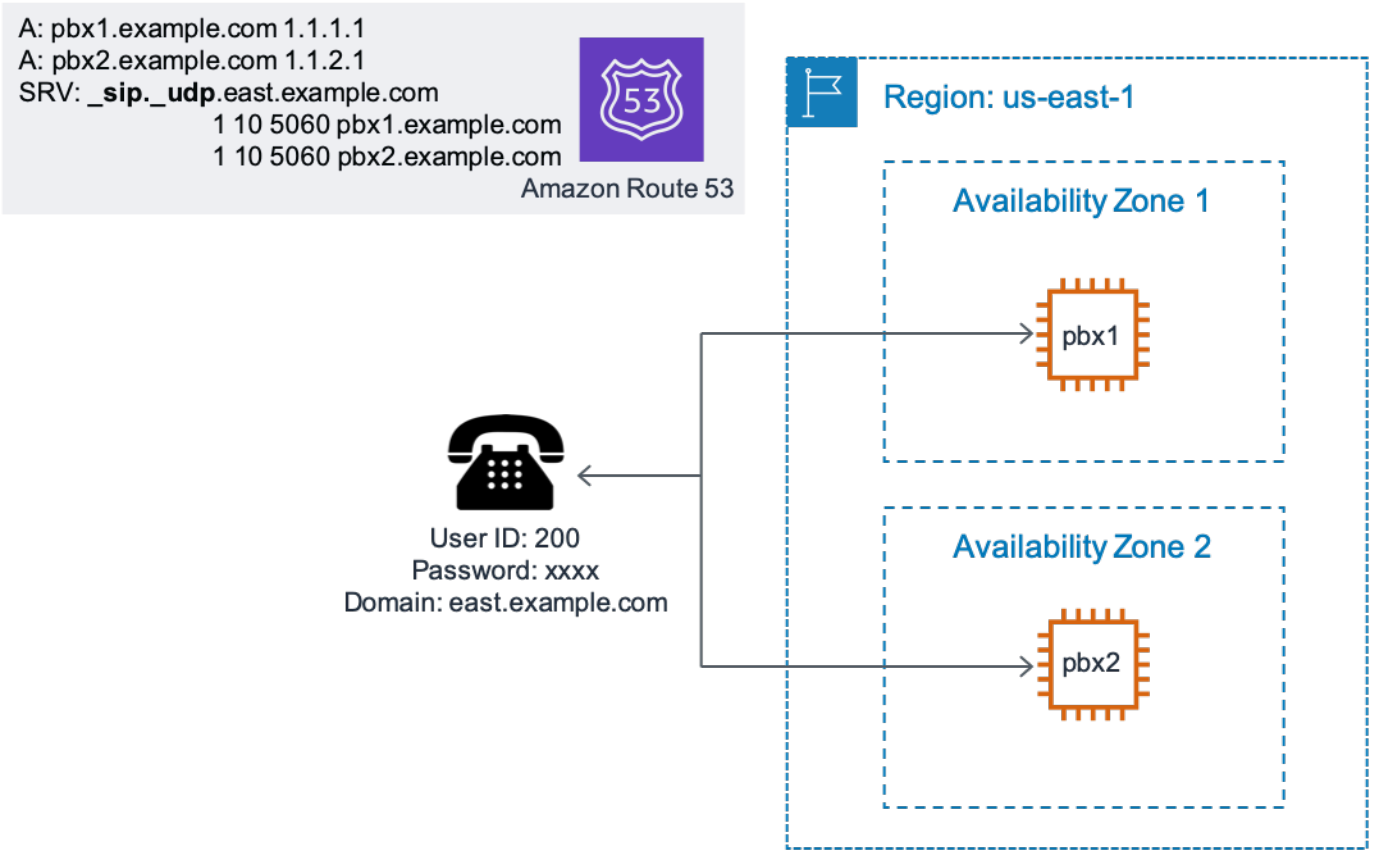
Utilizzo SIPp per monitorare l'infrastruttura VoIP

Molti strumenti open source, come iPerf o SIPp, and VOIPMonitor, sono disponibili per il monitoraggio del traffico SIP/RTP. Nell'esempio precedente, i nodi che eseguono SIP in modalità client e server misurano metriche SIP come Successful Calls e SIP Retrasmit tra tutti e quattro gli Stati Uniti. Regioni AWS Queste metriche possono quindi essere esportate in Amazon CloudWatch utilizzando uno script personalizzato. Utilizzando CloudWatch, i clienti possono creare allarmi sulla base di queste metriche personalizzate in base a un determinato valore di soglia. È quindi possibile intraprendere azioni di riparazione 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, sul mercato sono disponibili molte buone soluzioni di monitoraggio VoIP, come. ThousandEyes Un esempio di azione correttiva è la modifica del routing SIP in base a un aumento delle ritrasmissioni SIP.

Utilizzate il DNS per il bilanciamento del carico e il floating per il failover IPs

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/. SBCs PBXs



Utilizzo dei record DNS SRV per bilanciare il carico dei client SIP

La figura precedente 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 sip. `<transport protocol>`prefisso in un record DNS di tipo SRV. Nell'esempio, la sezione di risposta di DNS contiene entrambi i file in PBXs esecuzione in diverse zone di disponibilità. AWS Tuttavia, oltre all'endpoint URIs, il record SRV contiene tre informazioni aggiuntive:

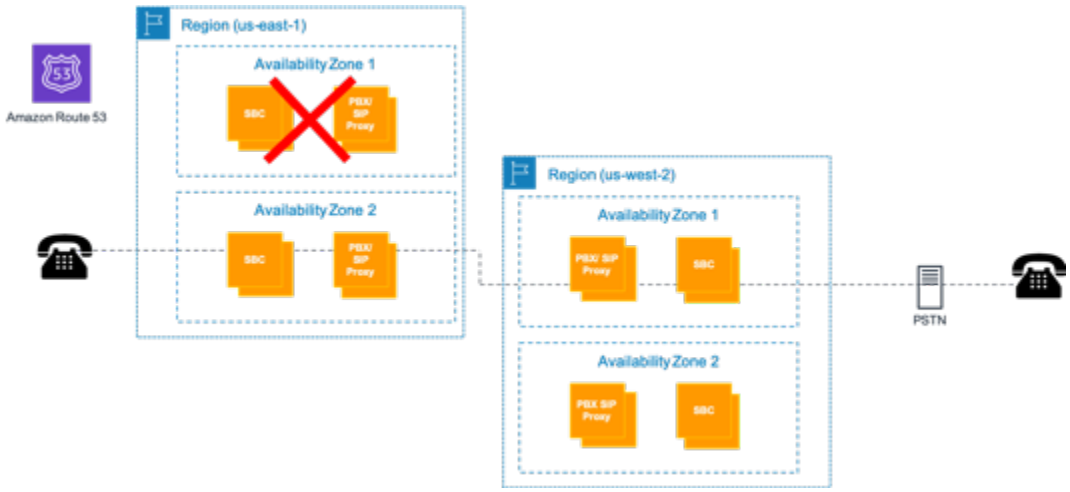
- Il primo numero è la priorità (1 nell'esempio precedente). È preferibile una priorità più bassa rispetto a una priorità più alta.
- Il secondo numero è il Peso (10 nell'esempio precedente).
- E il terzo numero è la porta da utilizzare (5060).

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

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

Utilizza più zone di disponibilità

Ciascuna Regione AWS è suddivisa in zone di disponibilità separate. Ogni zona di disponibilità è dotata di alimentazione, raffreddamento e connettività di rete proprie e forma quindi un dominio di errore isolato. Nell'ambito di AWS, i clienti sono incoraggiati a eseguire i propri carichi di lavoro in più di una zona di disponibilità. Ciò garantisce che le applicazioni dei clienti possano resistere anche a un guasto completo della zona di disponibilità, un evento di per sé molto raro. Questa raccomandazione vale anche per l'infrastruttura SIP in tempo reale.



Gestione dell'errore nella zona di disponibilità

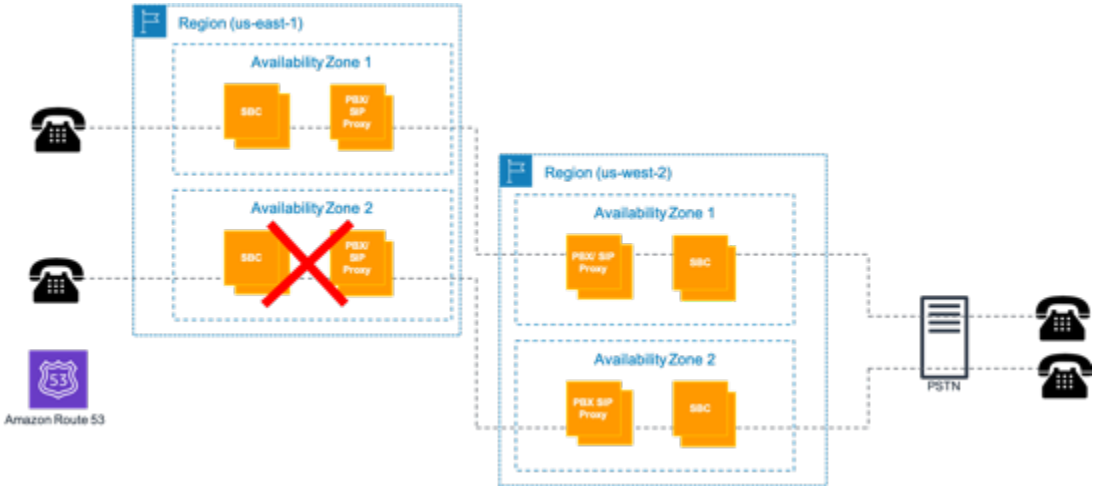
Supponiamo che un evento catastrofico (come un uragano di categoria cinque) provochi un'interruzione completa della zona di disponibilità nella regione us-east-1. Con l'infrastruttura in funzione come mostrato nel diagramma, tutti i client SIP originariamente registrati con i nodi della zona di disponibilità guasta 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 vadano perse, tutte le nuove chiamate vengono instradate attraverso la Zona di disponibilità 2.

Riassumendo, i record DNS SRV devono indirizzare il client verso più record 'A', uno in ogni zona di disponibilità. Ciascuno di questi record 'A' dovrebbe, a sua volta, puntare a più indirizzi IP di SBCs/PBXs in quella zona di disponibilità, garantendo la resilienza sia all'interno che all'interno della zona di disponibilità. Il failover tra zone di disponibilità e tra zone di disponibilità può essere implementato utilizzando la riassegnazione IP, se sono pubblici. IPs Il privato IPs, tuttavia, non può essere riassegnato tra le zone di disponibilità. Se un cliente utilizza un indirizzamento IP privato, dovrebbe fare affidamento sulla nuova registrazione dei client SIP con l'SBC/PBX di backup per il failover tra zone di disponibilità.

Mantieni il traffico all'interno di una zona di disponibilità e utilizza i gruppi di collocamento EC2

Nota anche come Availability Zone Affinity, questa best practice si applica anche al raro caso di un guasto completo della zona di disponibilità. Si consiglia di eliminare il traffico inter-AZ in modo che il traffico SIP o RTP che entra in una zona di disponibilità rimanga in quella zona di disponibilità fino a quando non esce dalla regione.



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

La figura precedente mostra un'architettura semplificata che utilizza l'affinità Availability Zone. Il vantaggio comparativo di questo approccio diventa evidente se si tiene conto degli effetti di un'interruzione completa della Availability Zone. Come illustrato nel diagramma, se si perde la Zona di disponibilità 2, ne risente al massimo il 50% delle chiamate attive (presupponendo che il bilanciamento del carico sia uguale tra le zone di disponibilità). Se non fosse stata implementata

l'affinità delle zone di disponibilità, alcune chiamate fluirebbero tra le zone di disponibilità di una regione e un errore molto probabilmente influirebbe su più del 50% delle chiamate attive.

Per ridurre al minimo la latenza per il traffico, AWS consiglia inoltre di prendere in considerazione l'utilizzo di gruppi di EC2 collocamento all'interno di ciascuna zona di disponibilità. Le istanze lanciate all'interno dello stesso gruppo di EC2 collocamento hanno una larghezza di banda maggiore e una latenza ridotta, in modo da EC2 garantire la vicinanza di rete di queste istanze l'una rispetto all'altra.

Utilizza tipi di istanze di rete avanzati EC2

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 selezione di tipi di istanze ottimizzate per adattarsi a diversi casi d'uso. I tipi di istanza comprendono diverse combinazioni di CPU, memoria, archiviazione e capacità di rete, inoltre offrono la flessibilità necessaria per scegliere la combinazione di risorse appropriata per le applicazioni. Questi tipi di istanze di rete avanzate garantiscono che i carichi di lavoro SIP in esecuzione su di esse abbiano accesso a una larghezza di banda costante e a una latenza aggregata relativamente inferiore. Un'aggiunta recente 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 EC2 istanze e delle funzionalità associate è disponibile nella pagina dei tipi di EC2 istanze.

Per la maggior parte dei clienti, l'ultima generazione di istanze Compute Optimized dovrebbe offrire il miglior rapporto qualità-prezzo. Ad esempio, il C5N supporta il nuovo Elastic Network Adapter 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 anche dall'utilizzo dell'Intel Data Plane Developer Kit (DPDK), che può potenziare notevolmente l'elaborazione dei pacchetti di rete.

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

Considerazioni relative alla sicurezza

I componenti delle applicazioni RTC in genere vengono eseguiti direttamente su EC2 istanze Amazon rivolte a Internet. Oltre al TCP, i flussi utilizzano protocolli come UDP e SIP. In questi casi, AWS Shield Standard protegge EC2 le istanze Amazon dagli attacchi DDo S del 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 la modellazione del traffico basata sulla priorità, che vengono attivate automaticamente quando viene rilevata una firma di attacco S ben definita. DDo

AWS fornisce inoltre una protezione avanzata contro attacchi DDo S di grandi dimensioni e sofisticati per queste applicazioni AWS Shield Advanced abilitando indirizzi IP elastici. AWS Shield Advanced fornisce un rilevamento DDo S avanzato che rileva automaticamente il tipo di AWS risorsa e la dimensione dell' EC2istanza e applica le mitigazioni predefinite appropriate con protezioni contro le inondazioni SYN o UDP. Con AWS Shield Advanced, i clienti possono anche creare i propri profili di mitigazione personalizzati coinvolgendo l' DDoAWS S Response Team (DRT) 24 ore su 24, 7 giorni su 7. AWS Shield Advanced assicura inoltre che durante un attacco DDo S, tutte le tue Amazon VPC Network Access Control List (ACLs) vengano applicate automaticamente ai confini della AWS rete, fornendoti l'accesso a larghezza di banda e capacità di pulizia aggiuntive per mitigare attacchi S volumetrici di grandi dimensioni. DDo

Conclusioni

È possibile implementare carichi di lavoro di comunicazione in tempo reale (RTC) AWS per ottenere scalabilità, elasticità e alta disponibilità soddisfacendo al contempo i requisiti chiave. Oggi, diversi clienti utilizzano AWS, i suoi partner e soluzioni open source per eseguire carichi di lavoro RTC con costi ridotti, maggiore agilità e un impatto globale ridotto.

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

Acronimi

Gli acronimi utilizzati in questo documento includono:

ACL — Elenco di controllo degli accessi

ALB — Application Load Balancer

APNs — Servizio Apple Push Notification

BGP — Protocollo Border Gateway

CDR — Record dei dettagli delle chiamate

COTS — software commerciale off-the-shelf

DDoS — distribuito denial-of-service

DNS: sistema di nomi di dominio

DPDK — Kit per sviluppatori Intel Data Plane

DRT — DDo S Response Team

ENA — Adattatore di rete elastico

EPC — Evolved Packet Core

FCM — Firebase Cloud Messaging

HA — Alta disponibilità

IRC — Internet Relay Chat

ISDN — Rete digitale di servizi integrati

NAT: traduzione degli indirizzi di rete

OPUS: supporto utente per il posizionamento online

PBX — Borsa di filiali private

PRI — Primary Rate Interface

PSTN — Rete telefonica pubblica commutata

RAID: array ridondante di dischi indipendenti

RTC: comunicazione in tempo reale

RTP: protocollo di trasporto in tempo reale

SAN — Storage Area Network

SBC — controller di frontiera di sessione

SIP — Protocollo di avvio della sessione

SPOF: singoli punti di errore

SRV — Servizio

SS7 — Sistema di segnalamento n.7

STUN — Session Traversal Utilities per NAT

SYN — Sincronizza

TCP — Transmission Control Protocol

TDM — multiplazione a divisione temporale

TURN — Attraversamento tramite relè attorno al NAT

UDP — User Datagram Protocol

URI — Identificatori di risorse uniformi

VIP — IP virtuale

VNF — Funzione di rete virtuale

VoIP — Voice over IP

VPC: cloud privato virtuale

WebRTC: comunicazione web in tempo reale

Collaboratori

Le seguenti persone e organizzazioni hanno contribuito a questo documento:

- Mounir Chennana, Architetto di soluzioni senior, Amazon Web Services
- Mohammed Al-Mehdar, Architetto di soluzioni senior, Amazon Web Services
- Ejaz Sial, architetto di soluzioni senior, Amazon Web Services
- Ahmad Khan, architetto di soluzioni senior, Amazon Web Services
- Tipu Qureshi, ingegnere principale Supporto AWS, Amazon Web Services
- Hasan Khan, responsabile tecnico senior, Amazon Web Services
- Shoma Chakravarty, responsabile tecnico mondiale, Telecomunicazioni, Amazon Web Services

Revisioni del documento

Per ricevere una notifica sugli aggiornamenti del presente whitepaper, iscriviti al feed RSS.

Modifica	Descrizione	Data
Aggiornamento del whitepaper	Aggiornato per i servizi e le funzionalità più recenti.	5 maggio 2022
Aggiornamento del whitepaper	Aggiornato per i servizi e le funzionalità più recenti.	13 febbraio 2020
Pubblicazione iniziale	Whitepaper pubblicato per la prima volta.	1 ottobre 2018

Note

I clienti sono responsabili della propria valutazione indipendente delle informazioni contenute nel presente documento. Questo documento: (a) è solo a scopo informativo, (b) rappresenta le attuali offerte e pratiche di prodotti AWS, che sono soggette a modifiche senza preavviso, e (c) non crea alcun impegno o garanzia da parte di AWS e delle sue affiliate, fornitori o licenzianti. I prodotti o i servizi AWS sono forniti «così come sono» senza garanzie, dichiarazioni o condizioni di alcun tipo, esplicite o implicite. Le responsabilità di AWS nei confronti dei propri clienti sono definite dai contratti AWS e il presente documento non costituisce parte né modifica qualsivoglia contratto tra AWS e i suoi clienti.

AWS Glossario

Per la AWS terminologia più recente, consultate il AWS glossario nella sezione Reference.Glossario AWS

Le traduzioni sono generate tramite traduzione automatica. In caso di conflitto tra il contenuto di una traduzione e la versione originale in Inglese, quest'ultima prevarrà.