AWS Whitepaper

Limiti di AWS Fault Isolation

Limiti di AWS Fault Isolation: AWS Whitepaper

Table of Contents

AWS Well-Architected Turb

Data di pubblicazione: 16 novembre 2022 (Revisioni del documento)

Sintesi

Amazon Web Services (AWS) fornisce diversi limiti di isolamento, come zone di disponibilità (AZ), regioni, piani di controllo e piani dati. Questo paper descrive in dettaglio come AWS utilizzare questi confini per creare servizi zonali, regionali e globali. Include anche indicazioni prescrittive su come considerare le dipendenze da questi diversi servizi e su come migliorare la resilienza dei carichi di lavoro creati utilizzandoli.

Well-Architected Trac

Il AWSWell-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 la progettazione e il funzionamento di sistemi affidabili, sicuri, efficienti, convenienti e sostenibili. Utilizzando il AWS Well-Architected Tool, disponibile gratuitamente in AWS Management Console, puoi esaminare i tuoi carichi di lavoro rispetto a queste best practice rispondendo a una serie di domande per ogni pilastro.

Per una guida più esperta e le best practice per la tua architettura cloud (implementazioni di architetture di riferimento, diagrammi e white paper), consulta l'Architecture Center. AWS

Introduzione

AWSgestisce un'infrastruttura globale per fornire servizi cloud che aiutano i clienti a distribuire carichi di lavoro in modo flessibile, sicuro, scalabile e altamente disponibile. L'AWSinfrastruttura utilizza diversi costrutti di isolamento dei guasti per aiutare i clienti a raggiungere i propri obiettivi di resilienza. Questi limiti di isolamento dei guasti consentono ai clienti di progettare i propri carichi di lavoro per sfruttare l'ambito prevedibile di contenimento degli impatti che forniscono. È anche importante capire in che modo AWS i servizi sono progettati utilizzando questi limiti, in modo da poter fare scelte intenzionali sulle dipendenze selezionate per il carico di lavoro.

Questo paper riassumerà innanzitutto l'infrastruttura AWS globale e i limiti di isolamento dei guasti che fornisce, nonché alcuni dei modelli utilizzati per progettare i nostri servizi. Utilizzando questa

base di comprensione, il paper illustrerà successivamente i diversi ambiti di servizi AWS forniti: zonale, regionale e globale. Presenterà inoltre le migliori pratiche per la creazione di architetture che utilizzano questi limiti di isolamento e diversi ambiti di servizio per migliorare la resilienza dei carichi di lavoro su cui vengono eseguiti. AWS In particolare, fornisce indicazioni prescrittive su come gestire la dipendenza dai servizi globali riducendo al minimo i singoli punti di errore. Questo ti aiuterà a fare scelte informate sulle tue AWS dipendenze e su come progettare il tuo carico di lavoro per l'alta disponibilità (HA) e il disaster recovery (DR).

infrastruttura AWS

Questa sezione presenta un riepilogo dell'infrastruttura AWS globale e dei limiti di isolamento dei guasti che fornisce. Inoltre, questa sezione fornirà una panoramica del concetto di piani di controllo e piani dati, che sono distinzioni fondamentali nella AWS progettazione dei servizi. Queste informazioni forniscono la base per comprendere in che modo i limiti di isolamento dei guasti, il piano di controllo e il piano dati di un servizio si applicano ai tipi di AWS servizio descritti nella sezione successiva.

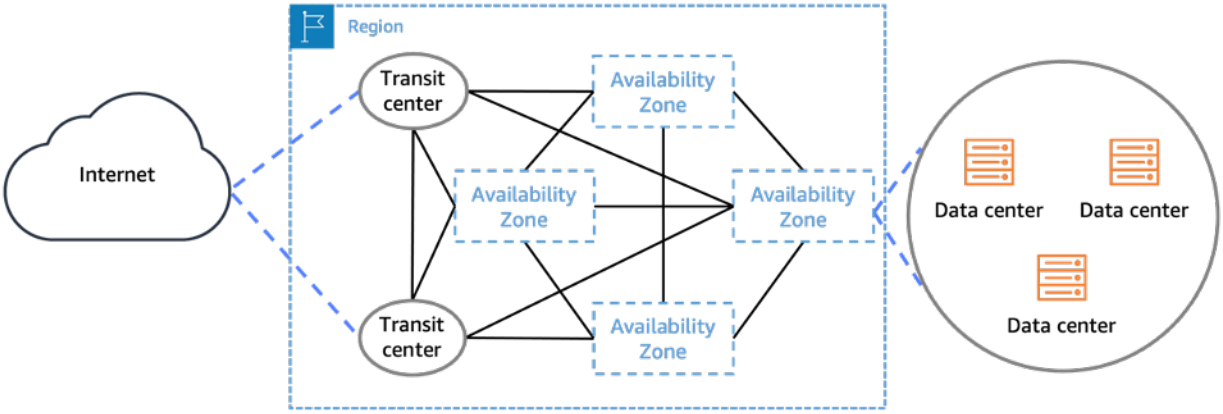
Argomenti

Zone di disponibilità

AWSgestisce oltre 100 zone di disponibilità in diverse regioni del mondo (i numeri attuali sono disponibili qui: AWSGlobal Infrastructure). Una zona di disponibilità è uno o più data center discreti con infrastruttura di alimentazione, rete e connettività indipendenti e ridondanti in un unico. Regione AWS Le zone di disponibilità in una regione sono significativamente distanti l'una dall'altra, fino a 60 miglia (~100 km) per prevenire guasti correlati, ma abbastanza vicine da utilizzare la replica sincrona con una latenza di un millisecondo. Sono progettate per non essere influenzate simultaneamente da uno scenario di destino condiviso, come l'alimentazione elettrica, l'interruzione delle risorse idriche, l'isolamento delle fibre, i terremoti, gli incendi, i tornado o le inondazioni. I punti di guasto più comuni, come i generatori e le apparecchiature di raffreddamento, non sono condivisi tra le zone di disponibilità e sono progettati per essere alimentati da sottostazioni elettriche indipendenti. Quando AWS distribuisce gli aggiornamenti ai propri servizi, le distribuzioni nelle zone di disponibilità della stessa regione vengono separate nel tempo per evitare guasti correlati.

Tutte le zone di disponibilità di una regione sono interconnesse con reti ad alta larghezza di banda e bassa latenza, tramite fibra metropolitana dedicata e completamente ridondante. Ogni zona di disponibilità in una regione si connette a Internet attraverso due centri di transito in cui AWS si collegano più provider Internet di primo livello (per ulteriori informazioni, consulta Overview of Amazon Web Services).

Queste funzionalità forniscono un forte isolamento delle zone di disponibilità l'una dall'altra, che chiamiamo Availability Zone Independence (AZI). Il costrutto logico delle zone di disponibilità e della loro connettività a Internet è illustrato nella figura seguente.



Le zone di disponibilità sono costituite da uno o più data center fisici collegati in modo ridondante tra loro e a Internet

Regioni

Ciascuna Regione AWS è composta da più zone di disponibilità indipendenti e fisicamente separate all'interno di un'area geografica. Tutte le regioni hanno attualmente tre o più zone di disponibilità. Le regioni stesse sono isolate e indipendenti dalle altre regioni con alcune eccezioni riportate più avanti in questo documento (fare riferimento alle operazioni globali in una singola regione). Questa separazione tra le regioni limita gli errori del servizio, quando si verificano, a una singola regione. In questo caso le normali operazioni delle altre regioni non sono influenzate. Inoltre, le risorse e i dati creati in una regione non esistono in nessun'altra regione, a meno che non si utilizzi esplicitamente una funzionalità di replica o copia offerta da un AWS servizio o si replichi personalmente la risorsa.

Regioni AWS attuali e pianificate a dicembre 2022

AWSLocal Zones

AWSLe Local Zones sono un tipo di implementazione dell'infrastruttura che colloca servizi di elaborazione, storage, database e altri AWSservizi selezionati vicino a grandi centri abitati e industriali. È possibile utilizzare AWS servizi, come i servizi di elaborazione e archiviazione, nella zona locale per eseguire applicazioni a bassa latenza all'edge o semplificare le migrazioni al cloud ibrido. Le Local Zones dispongono di ingressi e uscite Internet locali per ridurre la latenza, ma sono anche connesse alla regione madre tramite la rete privata ridondante e ad alta larghezza di banda di Amazon, che offre alle applicazioni in esecuzione nelle Local Zones AWS un accesso rapido, sicuro e senza interruzioni all'intera gamma di servizi.

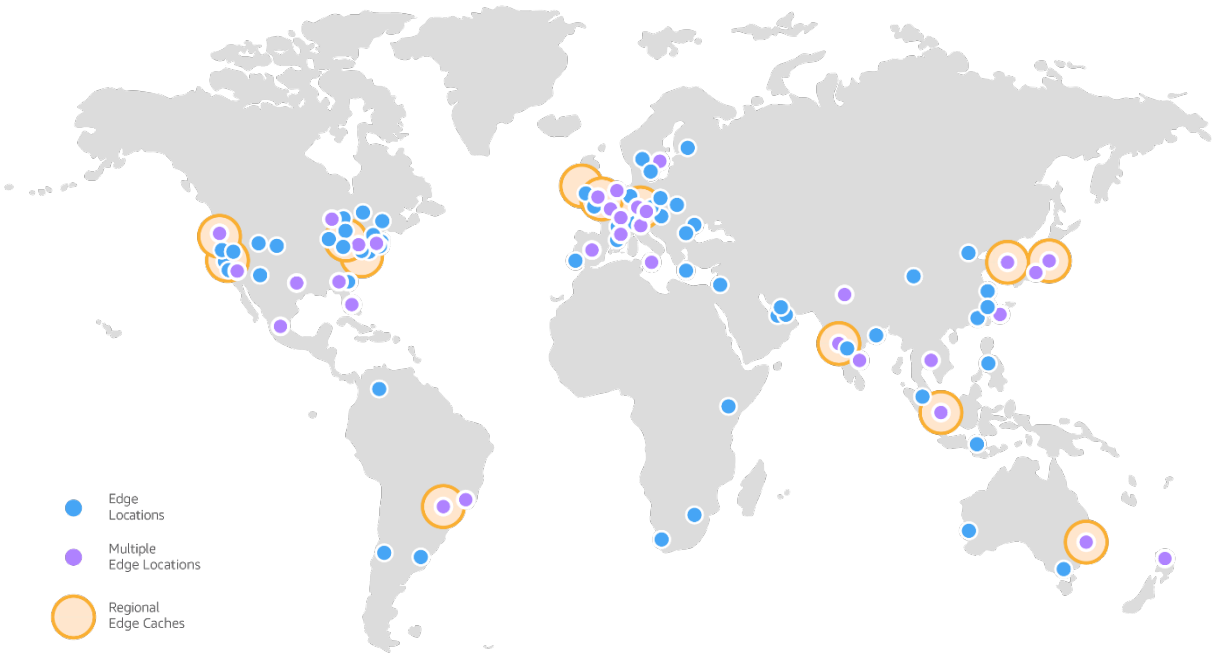
AWS Outposts

AWS Outpostsè una famiglia di soluzioni completamente gestite che forniscono AWS infrastrutture e servizi praticamente a qualsiasi location locale o periferica per un'esperienza ibrida davvero coerente. Le soluzioni Outposts consentono di estendere ed eseguire AWS servizi nativi in locale e sono disponibili in una varietà di fattori di forma, dai server Outposts 1U e 2U ai rack Outposts 42U e implementazioni su più rack.

ConAWS Outposts, puoi eseguire determinati AWS servizi localmente e connetterti a un'ampia gamma di servizi disponibili nella versione principale. Regione AWS AWS Outpostssono rack di elaborazione e archiviazione completamente gestiti e configurabili realizzati con hardware AWS progettato che consente ai clienti di eseguire elaborazione e archiviazione in locale, connettendosi senza problemi all'ampia gamma AWS di servizi nel cloud.

Punti di presenza

Oltre alle zone di disponibilità, gestisce AWS anche una rete di punti di presenza (PoP) distribuita a livello globale. Regioni AWS Questi PoPs ospitano Amazon CloudFront, una rete di distribuzione dei contenuti (CDN); Amazon Route 53, un servizio di risoluzione DNS (Domain Name System) pubblico; e AWS Global Accelerator (AGA), un servizio di ottimizzazione della rete perimetrale. La rete edge globale è attualmente composta da oltre 410 PoPs, tra cui più di 400 edge location, e 13 cache regionali di livello medio in oltre 90 città in 48 paesi (lo stato attuale è disponibile qui: Amazon CloudFront Key Features).



Rete CloudFront perimetrale globale Amazon

Ogni PoP è isolato dagli altri, il che significa che un guasto che interessa un singolo PoP o un'area metropolitana non ha alcun impatto sul resto della rete globale. La AWS rete collabora con migliaia di operatori di telecomunicazioni Tier 1/2/3 a livello globale, è ben collegata a tutte le principali reti di accesso per prestazioni ottimali e dispone di centinaia di terabit di capacità implementata. Le edge

location sono collegate alla Regioni AWS dorsale di AWS rete, una fibra parallela multipla da 100 GbE completamente ridondante che circola in tutto il mondo e si collega a decine di migliaia di reti per migliorare il recupero delle origini e l'accelerazione dinamica dei contenuti.

Partizioni

AWSraggruppa le regioni in partizioni. Ogni regione si trova esattamente in una partizione e ogni partizione ha una o più regioni. Le partizioni hanno istanze indipendenti di AWS Identity and Access Management (IAM) e forniscono un confine rigido tra le regioni in partizioni diverse. AWSLe regioni commerciali sono nella `aws` partizione, le regioni in Cina sono nella partizione e AWS GovCloud le regioni sono nella `aws-cn` partizione. `aws-us-gov` Alcuni AWS servizi sono progettati per fornire funzionalità interregionali, come Amazon S3 Cross-Region Replication o AWSTransit Gateway Inter-Region Peering. Questi tipi di funzionalità sono supportati solo tra regioni della stessa partizione. Non è possibile utilizzare le credenziali IAM di una partizione per interagire con le risorse in una partizione diversa.

Piani di controllo e piani dati

AWSsepara la maggior parte dei servizi nei concetti di piano di controllo e piano dati. Questi termini provengono dal mondo delle reti, in particolare dai router. Il piano dati del router, che è la sua funzionalità principale, sposta i pacchetti in base a regole. Ma le politiche di routing devono essere create e distribuite da qualche parte, ed è qui che entra in gioco il piano di controllo.

I piani di controllo forniscono le API amministrative utilizzate per creare, leggere/descrivere, aggiornare, eliminare ed elencare le risorse (CRUDL). Ad esempio, tutte le azioni del piano di controllo sono le seguenti: avvio di una nuova istanza Amazon Elastic Compute Cloud (Amazon EC2), creazione di un bucket Amazon Simple Storage Service (Amazon S3) e descrizione di una coda Amazon Simple Queue Service (Amazon SQS). Quando avvii un'istanza EC2, il piano di controllo deve eseguire diverse attività come trovare un host fisico con capacità, allocare le interfacce di rete, preparare un volume Amazon Elastic Block Store (Amazon EBS), generare credenziali IAM, aggiungere le regole del Security Group e altro ancora. I piani di controllo tendono ad essere sistemi di orchestrazione e aggregazione complicati.

Il piano dati è ciò che fornisce la funzione principale del servizio. Ad esempio, le seguenti sono tutte le parti del piano dati per ciascuno dei servizi coinvolti: l'istanza EC2 in esecuzione stessa, che legge e scrive su un volume EBS, riceve e inserisce oggetti in un bucket S3 e Route 53 che risponde alle query DNS ed esegue controlli di integrità.

I piani dati sono intenzionalmente meno complicati, con meno parti mobili rispetto ai piani di controllo, che di solito implementano un sistema complesso di flussi di lavoro, logica aziendale e database. Ciò rende statisticamente meno probabile che si verifichino eventi di errore nel piano dati rispetto al piano di controllo. Sebbene sia il piano dati che quello di controllo contribuiscano al funzionamento e al successo complessivi del servizio, li AWS considera componenti distinti. Questa separazione offre vantaggi sia in termini di prestazioni che di disponibilità.

Stabilità statica

Una delle caratteristiche di resilienza più importanti dei AWS servizi è la cosiddetta stabilità AWS statica. Ciò che significa questo termine è che i sistemi operano in uno stato statico e continuano a funzionare normalmente senza la necessità di apportare modifiche in caso di guasto o indisponibilità delle dipendenze. Un modo per farlo è prevenire le dipendenze circolari nei nostri servizi che potrebbero impedire il corretto ripristino di uno di tali servizi. Un altro modo per farlo è mantenere lo stato esistente. Riteniamo che i piani di controllo abbiano statisticamente maggiori probabilità di fallire rispetto ai piani dati. Sebbene il piano dati dipenda in genere dai dati che arrivano dal piano di controllo, il piano dati mantiene lo stato esistente e continua a funzionare anche in caso di compromissione del piano di controllo. L'accesso alle risorse dal piano dati, una volta effettuato il provisioning, non dipende dal piano di controllo e pertanto non è influenzato da alcuna compromissione del piano di controllo. In altre parole, anche se la capacità di creare, modificare o eliminare risorse è compromessa, le risorse esistenti rimangono disponibili. Ciò rende AWS i piani dati staticamente stabili rispetto a una compromissione del piano di controllo. È possibile implementare diversi modelli per essere staticamente stabili contro diversi tipi di errori di dipendenza.

Un esempio di stabilità statica è disponibile in Amazon EC2. Una volta lanciata, un'istanza EC2 è disponibile tanto quanto il server fisico in un data center. Non dipende da alcuna API del piano di controllo per rimanere in esecuzione o per riprendere a funzionare dopo un riavvio. La stessa proprietà vale per altre AWS risorse come VPC, bucket e oggetti Amazon S3 e volumi Amazon EBS.

La stabilità statica è un concetto profondamente radicato nella AWS progettazione dei suoi servizi, ma è anche un modello che può essere utilizzato dai clienti. In effetti, la maggior parte delle linee guida sulle migliori pratiche per utilizzare i diversi tipi di AWS servizi in modo resiliente consiste nell'implementare la stabilità statica per gli ambienti di produzione. I meccanismi di ripristino e mitigazione più affidabili sono quelli che richiedono il minor numero di modifiche per ottenere il ripristino. Invece di affidarsi al piano di controllo EC2 per lanciare nuove istanze EC2 da ripristinare in caso di guasto della zona di disponibilità, disporre di tale capacità aggiuntiva in anticipo aiuta a raggiungere la stabilità statica. Pertanto, l'eliminazione delle dipendenze dai piani di controllo (le

API che implementano le modifiche alle risorse) nel percorso di ripristino aiuta a produrre carichi di lavoro più resilienti. Per ulteriori dettagli sulla stabilità statica, i piani di controllo e i piani dati, consulta l'articolo della Amazon Builders' Library sulla stabilità statica con zone di disponibilità.

Riepilogo

AWSutilizza diversi contenitori di errore nella nostra infrastruttura per creare l'isolamento dei guasti. I contenitori di guasto dell'infrastruttura principale sono partizioni, regioni, zone di disponibilità, piani di controllo e piani dati. Successivamente, esamineremo diversi tipi di AWS servizi, come questi contenitori di guasti vengono utilizzati nella loro progettazione e come progettare i carichi di lavoro con essi per renderli resilienti.

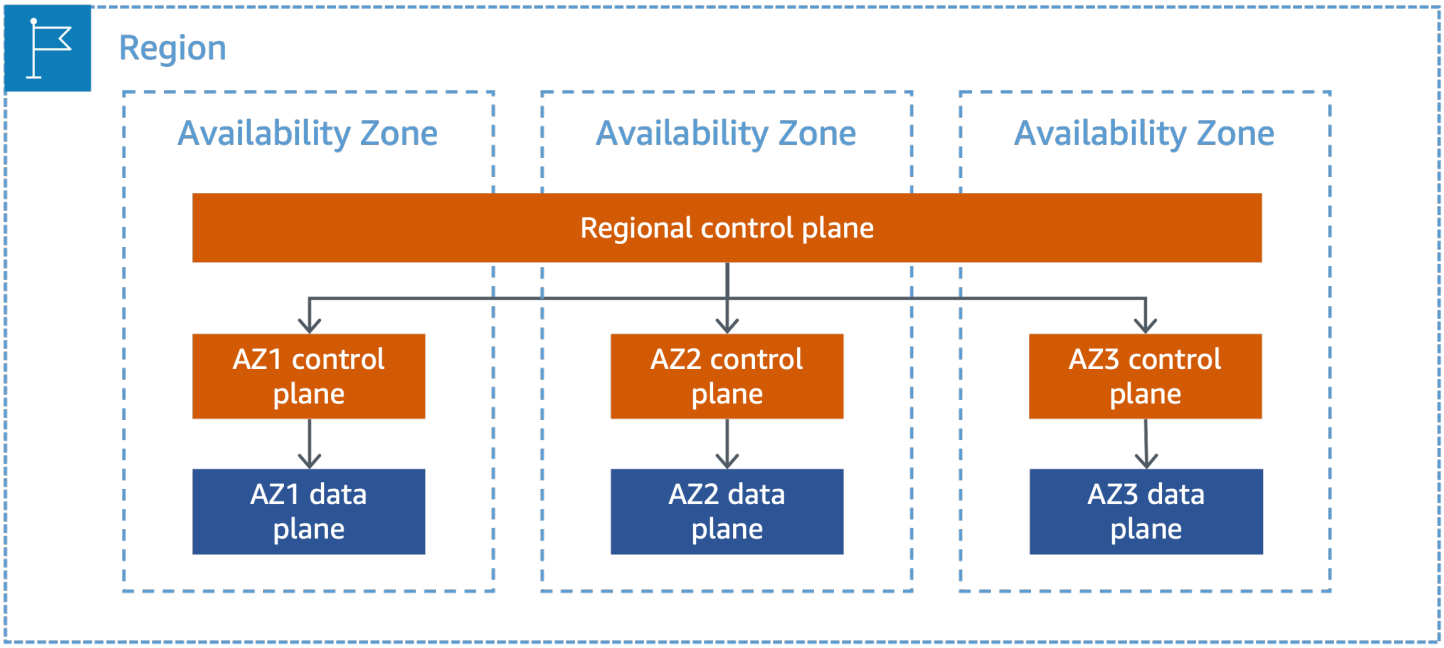
AWS tipi di servizio

AWS gestisce tre diverse categorie di servizi in base al limite di isolamento dei guasti: zonale, regionale e globale. Questa sezione descriverà più dettagliatamente come sono stati progettati questi diversi tipi di servizi in modo da poter determinare in che modo gli errori all'interno di un servizio di un determinato tipo di servizio influiranno sul carico di lavoro su cui è in esecuzione. AWS Fornisce inoltre indicazioni di alto livello su come progettare i carichi di lavoro per utilizzare questi servizi in modo resiliente. Per quanto riguarda i servizi globali, questo documento fornisce anche linee guida prescrittive Appendice B - Guida all'assistenza globale della rete Edge che possono aiutare a prevenire l'impatto sui carichi di lavoro derivanti da alterazioni del piano di controllo AWS dei servizi, aiutandovi a far fronte in modo sicuro alla dipendenza dai servizi globali e riducendo al minimo l'introduzione di singoli punti di errore. Appendice A - Guida al servizio partizionale

Argomenti

Servizi zonali

Availability Zone Independence (AZI) consente di AWS offrire servizi zonali, come Amazon EC2 e Amazon EBS. Un servizio zonale è un servizio che offre la possibilità di specificare in quale zona di disponibilità vengono distribuite le risorse. Questi servizi operano in modo indipendente in ogni zona di disponibilità all'interno di una regione e, cosa ancora più importante, falliscono indipendentemente anche in ogni zona di disponibilità. Ciò significa che i componenti di un servizio in una zona di disponibilità non dipendono dai componenti di altre zone di disponibilità. Possiamo farlo perché un servizio zonale ha piani dati zonali. In alcuni casi, come nel caso di EC2, il servizio include anche piani di controllo zonali per operazioni allineate a zone, come l'avvio di un'istanza EC2. Per questi servizi, fornisce AWS anche un endpoint del piano di controllo regionale per facilitare l'interazione con il servizio. Il piano di controllo regionale offre inoltre funzionalità con ambito regionale e funge da livello di aggregazione e routing in aggiunta ai piani di controllo zonali. Ciò è illustrato nella figura seguente.



Un servizio zonale con piani di controllo e piani dati isolati zonalmente

Le zone di disponibilità offrono ai clienti la possibilità di gestire carichi di lavoro di produzione con maggiore disponibilità, tolleranza ai guasti e scalabilità rispetto a quanto sarebbe possibile con un singolo data center. Quando un carico di lavoro utilizza più zone di disponibilità, i clienti sono meglio isolati e protetti dai problemi che influiscono sull'infrastruttura fisica di una singola zona di disponibilità. Questo aiuta i clienti a creare servizi ridondanti tra le zone di disponibilità e, se progettati correttamente, a rimanere operativi anche in caso di guasti in una zona di disponibilità. I clienti possono sfruttare AZI per creare carichi di lavoro resilienti e altamente disponibili. L'implementazione di AZI nell'architettura consente di ripristinare rapidamente un errore isolato in una zona di disponibilità, poiché le risorse in una zona di disponibilità riducono al minimo o eliminano l'interazione con le risorse in altre zone di disponibilità. Questo aiuta a rimuovere le dipendenze tra zone di disponibilità, semplificando così l'evacuazione delle zone di disponibilità. Consulta Advanced Multi-AZ Resilience Patterns per maggiori dettagli sulla creazione di meccanismi di evacuazione delle zone di disponibilità. Inoltre, è possibile sfruttare ulteriormente le zone di disponibilità seguendo alcune delle stesse best practice AWS utilizzate per i propri servizi, ad esempio implementando le modifiche a una sola zona di disponibilità alla volta o rimuovendo una zona di disponibilità dal servizio se una modifica in quella zona di disponibilità va male.

La stabilità statica è anche un concetto importante per le architetture Multi-Availability Zone. Una delle modalità di errore da pianificare con le architetture Multi-Availability Zone è la perdita di una zona di disponibilità, che può comportare la perdita della capacità di una zona di disponibilità. Se non è stata predisposta una capacità sufficiente per gestire la perdita di una zona di disponibilità, la

capacità residua potrebbe essere sovraccaricata dal carico corrente. Inoltre, sarà necessario fare affidamento sui piani di controllo dei servizi zonali utilizzati per sostituire la capacità persa, che può essere meno affidabile di un design staticamente stabile. In questo caso, predisporre in anticipo una capacità aggiuntiva sufficiente può aiutarvi a mantenere la stabilità statica in caso di perdita di un dominio di errore, come una zona di disponibilità, grazie alla possibilità di continuare le normali operazioni senza la necessità di modifiche dinamiche.

Puoi scegliere di utilizzare un gruppo di istanze EC2 con scalabilità automatica distribuite su più zone di disponibilità per scalare dinamicamente in entrata e in uscita, in base alle esigenze del tuo carico di lavoro. La scalabilità automatica è ideale per i cambiamenti graduali di utilizzo che si verificano nell'arco di minuti o decine di minuti. Tuttavia, il lancio di nuove istanze EC2 richiede tempo, soprattutto se le istanze richiedono il bootstrap (ad esempio l'installazione di agenti, file binari delle applicazioni o file di configurazione). Durante questo periodo, la capacità residua potrebbe essere sopraffatta dal carico attuale. Inoltre, l'implementazione di nuove istanze tramite la scalabilità automatica si basa sul piano di controllo EC2. Ciò presenta un compromesso: per essere staticamente stabili alla perdita di una singola zona di disponibilità, è necessario predisporre un numero sufficiente di istanze EC2 nelle altre zone di disponibilità per gestire il carico che è stato spostato dalla zona di disponibilità compromessa, invece di affidarsi alla scalabilità automatica per il provisioning di nuove istanze. Tuttavia, il pre-provisioning di capacità aggiuntiva può comportare costi aggiuntivi.

Ad esempio, durante il normale funzionamento, supponiamo che il carico di lavoro richieda sei istanze per servire il traffico dei clienti in tre zone di disponibilità. Per garantire la stabilità statica in caso di guasto di una singola zona di disponibilità, è necessario implementare tre istanze in ciascuna zona di disponibilità, per un totale di nove. Se in una singola zona di disponibilità si guastasse un numero di istanze pari a sei, ne resterebbero comunque sei e saresti in grado di continuare a servire il traffico dei clienti senza dover fornire e configurare nuove istanze in caso di guasto. Il raggiungimento della stabilità statica della capacità EC2 comporta costi aggiuntivi, poiché, in questo caso, si utilizzano istanze aggiuntive del 50%. Non tutti i servizi in cui è possibile effettuare il pre-provisioning delle risorse comportano costi aggiuntivi, come il pre-provisioning di un bucket S3 o di un utente. Dovrai valutare qualsiasi compromesso derivante dall'implementazione della stabilità statica rispetto al rischio di superare il tempo di ripristino desiderato per il tuo carico di lavoro.

AWS Local Zones and Outposts avvicinano il piano dati di AWS servizi selezionati agli utenti finali. I piani di controllo per questi servizi risiedono nella regione madre. L'istanza Local Zone o Outposts avrà dipendenze dal piano di controllo per i servizi zonali come EC2 ed EBS sulla zona di disponibilità in cui hai creato la tua sottorete Local Zone o Outposts. Inoltre dipenderanno dai piani di controllo regionali per i servizi regionali come Elastic Load Balancing (ELB), i gruppi di sicurezza e il piano di

controllo Kubernetes gestito da Amazon Elastic Kubernetes Service (Amazon EKS) (se usi EKS). Per ulteriori informazioni specifiche su Outposts, consulta la documentazione e le domande frequenti sull'assistenza e la manutenzione. Implementa la stabilità statica quando usi Local Zones o Outposts per migliorare la resilienza al fine di controllare i problemi o le interruzioni del piano nella connettività di rete verso la regione madre.

Servizi regionali

I servizi regionali sono servizi che AWS si basano su più zone di disponibilità in modo che i clienti non debbano capire come utilizzare al meglio i servizi zonali. Raggruppiamo logicamente il servizio distribuito in più zone di disponibilità per presentare ai clienti un unico endpoint regionale. Amazon SQS e Amazon DynamoDB sono esempi di servizi regionali. Sfruttano l'indipendenza e la ridondanza delle zone di disponibilità per ridurre al minimo i guasti dell'infrastruttura come categoria di rischio di disponibilità e durabilità. Amazon S3, ad esempio, distribuisce richieste e dati su più zone di disponibilità ed è progettato per il ripristino automatico in caso di guasto di una zona di disponibilità. Tuttavia, interagisci solo con l'endpoint regionale del servizio.

AWS ritiene che la maggior parte dei clienti possa raggiungere i propri obiettivi di resilienza in una singola regione utilizzando servizi regionali o architetture Multi-AZ che si basano su servizi zonali. Tuttavia, alcuni carichi di lavoro potrebbero richiedere una ridondanza aggiuntiva ed è possibile utilizzare l'isolamento di Regioni AWS per creare architetture multiregionali per scopi di elevata disponibilità o continuità aziendale. La separazione fisica e logica tra di esse evita guasti correlati tra Regioni AWS di loro. In altre parole, come se fossi un cliente EC2 e potessi trarre vantaggio dall'isolamento delle zone di disponibilità distribuendole su di esse, puoi ottenere lo stesso vantaggio per i servizi regionali distribuendoli in più regioni. Ciò richiede l'implementazione di un'architettura multiregionale per la tua applicazione, che può aiutarti a resistere ai danni di un servizio regionale.

Tuttavia, ottenere i vantaggi di un'architettura multiregionale può essere difficile; richiede un lavoro accurato per sfruttare l'isolamento regionale senza annullare nulla a livello di applicazione. Ad esempio, se si esegue il failover di un'applicazione tra regioni, è necessario mantenere una netta separazione tra gli stack di applicazioni in ciascuna regione, tenere conto di tutte le dipendenze delle applicazioni ed eseguire il failover di tutte le parti dell'applicazione contemporaneamente. Per raggiungere questo obiettivo con un'architettura complessa basata su microservizi che presenta molte dipendenze tra le applicazioni, è necessario pianificare e coordinare diversi team di progettazione e business. Consentire ai singoli carichi di lavoro di prendere le proprie decisioni di failover rende il coordinamento meno complesso, ma introduce un comportamento modale grazie alla significativa differenza di latenza che si verifica tra le regioni rispetto a quella all'interno di una singola regione.

AWS al momento non fornisce una funzionalità di replica sincrona tra regioni. Quando si utilizza un datastore replicato in modo asincrono (fornito da AWS) tra regioni, esiste la possibilità di perdita o incoerenza dei dati in caso di failover dell'applicazione tra regioni. Per mitigare possibili incoerenze, è necessario un processo di riconciliazione dei dati affidabile in cui avere fiducia e che potrebbe essere necessario operare su più archivi di dati in tutto il portafoglio di carichi di lavoro, oppure è necessario essere disposti ad accettare la perdita di dati. Infine, devi fare pratica con il failover per sapere che funzionerà quando ne avrai bisogno. La rotazione regolare dell'applicazione da una regione all'altra per fare pratica con il failover è un notevole investimento in termini di tempo e risorse. Se si decide di utilizzare un datastore replicato in modo sincrono in più regioni per supportare le applicazioni eseguite contemporaneamente da più di una regione, le caratteristiche prestazionali e la latenza di un database di questo tipo che si estende per centinaia o migliaia di miglia sono molto diverse da quelle di un database che opera in una singola regione. Ciò richiede di pianificare lo stack di applicazioni da zero per tenere conto di questo comportamento. Inoltre, rende la disponibilità di entrambe le regioni una forte dipendenza, il che potrebbe comportare una riduzione della resilienza del carico di lavoro.

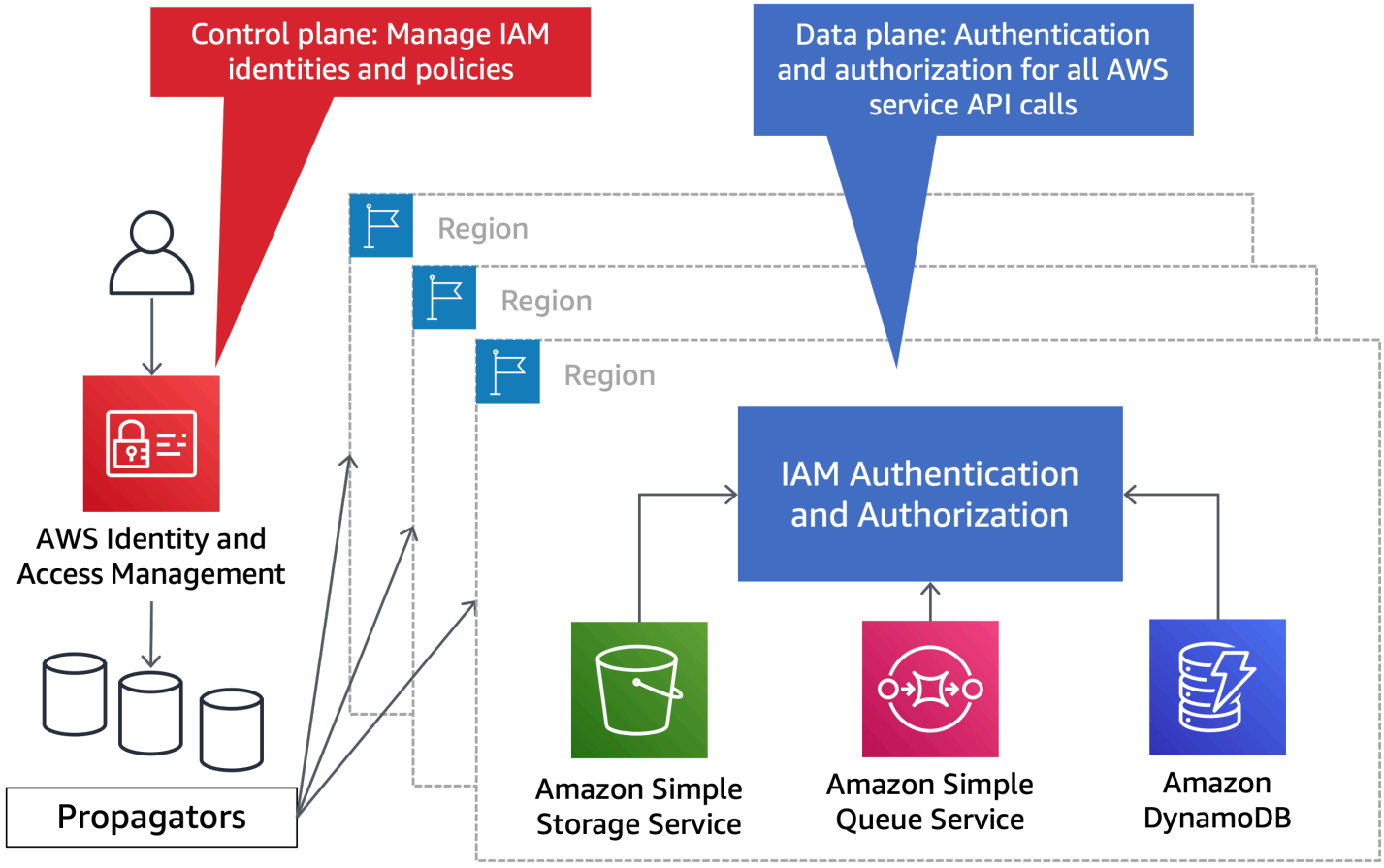
Servizi globali

Oltre ai AWS servizi regionali e zonali, esiste un piccolo insieme di AWS servizi i cui piani di controllo e piani dati non esistono indipendentemente in ciascuna regione. Poiché le loro risorse non sono specifiche per regione, vengono comunemente definite globali. AWS I servizi globali seguono ancora lo schema di AWS progettazione convenzionale che prevede la separazione del piano di controllo dal piano dati per raggiungere la stabilità statica. La differenza significativa per la maggior parte dei servizi globali è che il piano di controllo è ospitato in un unico piano Regione AWS, mentre il piano dati è distribuito a livello globale. Esistono tre diversi tipi di servizi globali e un set di servizi che possono apparire globali in base alla configurazione selezionata.

Le sezioni seguenti identificheranno ogni tipo di servizio globale e il modo in cui i relativi piani di controllo e piani dati sono separati. È possibile utilizzare queste informazioni per guidare la creazione di meccanismi affidabili di alta disponibilità (HA) e disaster recovery (DR) senza dover dipendere da un piano di controllo del servizio globale. Questo approccio aiuta a rimuovere singoli punti di errore nell'architettura ed evita potenziali impatti interregionali, anche quando si opera in una regione diversa da quella in cui è ospitato il piano di controllo del servizio globale. Inoltre, consente di implementare in modo sicuro meccanismi di failover che non si basano su piani di controllo dei servizi globali.

Servizi globali unici per partizione

In ogni partizione sono presenti alcuni AWS servizi globali (denominati in questo paper servizi partizionali). I servizi partizionali forniscono il proprio piano di controllo in un unico piano. Regione AWS Alcuni servizi partizionali, come AWS Network Manager, sono disponibili solo sul piano di controllo e orchestrano il piano dati di altri servizi. Altri servizi partizionali, come IAM, dispongono di un proprio piano dati isolato e distribuito su tutti gli elementi della partizione. Regioni AWS Gli errori in un servizio partizionale non influiscono sulle altre partizioni. Nella `aws` partizione, il piano di controllo del servizio IAM si trova nella `us-east-1` regione, con piani dati isolati in ciascuna regione della partizione. I servizi partizionali dispongono inoltre di piani di controllo e piani dati indipendenti nelle partizioni e. `aws-us-gov` `aws-cn` La separazione tra piano di controllo e piano dati per IAM è illustrata nel diagramma seguente.



IAM ha un unico piano di controllo e un piano dati regionalizzato

Di seguito sono riportati i servizi partizionali e la loro posizione del piano di controllo nella partizione: `aws`

- AWS IAM () `us-east-1`

- AWS Organizations (`us-east-1`)
- AWS Gestione dell'account (`us-east-1`)
- Route 53 Application Recovery Controller (ARC`us-west-2`) () - Questo servizio è presente solo nella `aws` partizione
- AWS Gestore di rete () `us-west-2`
- DNS privato Route 53 () `us-east-1`

Se uno di questi piani di controllo del servizio presenta un evento che influisce sulla disponibilità, potresti non essere in grado di utilizzare le operazioni di tipo CRUDL fornite da questi servizi. Pertanto, se la strategia di ripristino dipende da queste operazioni, un impatto sulla disponibilità sul piano di controllo o sulla regione che ospita il piano di controllo ridurrà le possibilità di successo del ripristino. Appendice A - Guida al servizio partizionalefornisce strategie per rimuovere le dipendenze dai piani di controllo dei servizi globali durante il ripristino.

> **Raccomandazione**
>
> Non fate affidamento sui piani di controllo dei servizi partizionali nel percorso di ripristino. Affidatevi invece alle operazioni del piano dati di questi servizi. Appendice A - Guida al servizio partizionalePer ulteriori dettagli su come progettare i servizi partizionali, consulta la sezione.

Servizi globali nella rete perimetrale

Il prossimo set di AWS servizi globali prevede un piano di controllo nella `aws` partizione e ospita i relativi piani dati nell'infrastruttura dei punti di presenza globali (PoP) (e potenzialmente Regioni AWS anche). È PoPs possibile accedere ai piani dati ospitati dalle risorse di qualsiasi partizione e da Internet. Ad esempio, Route 53 utilizza il proprio piano di controllo nella `us-east-1` regione, ma il piano dati è distribuito su centinaia di siti PoPs a livello globale, oltre a ciascuno di essi Regione AWS (per supportare i DNS pubblici e privati di Route 53 all'interno della regione). Anche i controlli dello stato di Route 53 fanno parte del piano dati e vengono eseguiti dalle otto unità Regioni AWS della `aws` partizione. I client possono risolvere il DNS utilizzando le zone ospitate pubbliche di Route 53 da qualsiasi punto di Internet, comprese altre partizioni come GovCloud, nonché da un AWS Virtual Private Cloud (VPC). Di seguito sono riportati i servizi di rete edge globali e la loro posizione del piano di controllo nella partizione: `aws`

- DNS pubblico Route 53 () `us-east-1`
- Amazon CloudFront (`us-east-1`)
- AWS WAF Classico per CloudFront (`us-east-1`)
- AWS WAF per CloudFront (`us-east-1`)
- Amazon Certificate Manager (ACM) per CloudFront (`us-east-1`)
- AWS Global Accelerator (AGA) () `us-west-2`
- AWS Shield Advanced (`us-east-1`)

Se utilizzi i controlli di integrità AGA per istanze EC2 o indirizzi IP elastici, questi utilizzano i controlli di integrità di Route 53. La creazione o l'aggiornamento dei controlli di integrità AGA dipenderebbe dal piano di controllo della Route 53 in. `us-east-1` L'esecuzione dei controlli di integrità AGA utilizza il piano dati di controllo dello stato di salute della Route 53.

In caso di guasto che influisce sulla regione che ospita i piani di controllo di questi servizi o di un guasto che influisce sul piano di controllo stesso, potrebbe non essere possibile utilizzare le operazioni di tipo CRUDL fornite da questi servizi. Se nella strategia di ripristino si è fatto affidamento su queste operazioni, è possibile che tale strategia abbia meno probabilità di successo rispetto a quando si fa affidamento solo sul piano dati di questi servizi.

> **Raccomandazione**
>
> Non fare affidamento sul piano di controllo dei servizi di rete perimetrale nel percorso di ripristino. Affidati invece alle operazioni del piano dati di questi servizi. Appendice B - Guida all'assistenza globale della rete EdgePer ulteriori dettagli su come progettare servizi globali nella rete perimetrale, vedi.

Operazioni globali in una singola regione

L'ultima categoria è composta da specifiche operazioni sul piano di controllo nell'ambito di un servizio con un ambito di impatto globale, non da interi servizi come nelle categorie precedenti. Sebbene si interagisca con i servizi zonali e regionali nella regione specificata, alcune operazioni dipendono da una singola regione diversa da quella in cui si trova la risorsa. Si tratta di servizi diversi dai servizi forniti in una sola regione; Appendice C - Servizi a regione singola per un elenco di tali servizi, consulta la sezione.

Durante un errore che influisce sulla dipendenza globale sottostante, potresti non essere in grado di utilizzare le azioni di tipo CRUDL delle operazioni dipendenti. Se nella strategia di ripristino si è fatto affidamento su queste operazioni, è possibile che tale strategia abbia meno probabilità di successo rispetto a quando si fa affidamento solo sul piano dati di questi servizi. È necessario evitare di dipendere da queste operazioni per la strategia di ripristino.

Di seguito è riportato un elenco di servizi da cui altri servizi possono dipendere e che hanno una portata globale:

- Itinerario 53

  Diversi AWS servizi creano risorse che forniscono uno o più nomi DNS specifici della risorsa. Ad esempio, quando si effettua il provisioning di un Elastic Load Balancer (ELB), il servizio crea record DNS pubblici e controlli di integrità in Route 53 per l'ELB. Ciò si basa sul piano di controllo della Route 53 in. `us-east-1` Gli altri servizi che utilizzi potrebbero inoltre richiedere il provisioning di un ELB, creare record DNS pubblici di Route 53 o creare controlli di integrità della Route 53 come parte dei flussi di lavoro del piano di controllo. Ad esempio, il provisioning di una risorsa API REST di Amazon API Gateway, di un database Amazon Relational Database Service (Amazon RDS) o di un dominio OpenSearch Amazon Service comporta la creazione di record DNS in Route 53. Di seguito è riportato un elenco di servizi il cui piano di controllo dipende dal piano di controllo di Route 53 `us-east-1` per creare, aggiornare o eliminare record DNS, zone ospitate e/o creare controlli di integrità di Route 53. Questo elenco non è esaustivo; ha lo scopo di evidenziare alcuni dei servizi più comunemente utilizzati le cui azioni del piano di controllo per la creazione, l'aggiornamento o l'eliminazione delle risorse dipendono dal piano di controllo della Route 53:

  - API REST e HTTP di Amazon API Gateway
  - Istanze Amazon RDS
  - Database Amazon Aurora
  - Bilanciatori di carico Amazon ELB
  - AWS PrivateLink Endpoint VPC
  - AWS Lambda URL
  - Amazon ElastiCache
  - OpenSearch Servizio Amazon
  - Amazon CloudFront
  - Amazon MemoryDB per Redis
  - Amazon Neptune

- Acceleratore Amazon DynamoDB (DAX)
- AGA
- Amazon Elastic Container Service (Amazon ECS) con Service Discovery basato su DNS (che utilizza l'API per gestire AWS Cloud Map il DNS di Route 53)
- Piano di controllo Amazon EKS Kubernetes

  È importante notare che il servizio DNS VPC per i nomi host delle istanze EC2 esiste indipendentemente in ciascuno di essi Regione AWS e non dipende dal piano di controllo di Route 53. I record AWS creati per le istanze EC2 nel servizio DNS VPC, ad esempio,, e `ip-10-0-10.ec2.internal` `ip-10-0-1-5.compute.us-west-2.compute.internal` `i-0123456789abcdef.ec2.internali-0123456789abcdef.us-west-2.compute.internal`, non si basano sul piano di controllo di Route 53 in. `us-east-1`

  > **Raccomandazione**
  >
  > Non fare affidamento sulla creazione, l'aggiornamento o l'eliminazione di risorse che richiedono la creazione, l'aggiornamento o l'eliminazione di record di risorse, zone ospitate o controlli dello stato di Route 53 nel percorso di ripristino. Effettua il pre-provisioning di queste risorse, come gli ELB, per evitare che nel percorso di ripristino si crei una dipendenza dal piano di controllo della Route 53.

- Amazon S3

  Le seguenti operazioni del piano di controllo di Amazon S3 dipendono da una partizione di base. `us-east-1` `aws` Un guasto che influisce su Amazon S3 o altri servizi potrebbe compromettere le azioni di questi piani di controllo `us-east-1` in altre regioni:

```
PutBucketCors
DeleteBucketCors
PutAccelerateConfiguration
PutBucketRequestPayment
PutBucketObjectLockConfiguration
PutBucketTagging
DeleteBucketTagging
PutBucketReplication
DeleteBucketReplication
```

```
PutBucketEncryption
DeleteBucketEncryption
PutBucketLifecycle
DeleteBucketLifecycle
PutBucketNotification
PutBucketLogging
DeleteBucketLogging
PutBucketVersioning
PutBucketPolicy
DeleteBucketPolicy
PutBucketOwnershipControls
DeleteBucketOwnershipControls
PutBucketAcl
PutBucketPublicAccessBlock
DeleteBucketPublicAccessBlock
```

Il piano di controllo per Amazon S3 Multi-Region Access Points (MRAP) è ospitato solo in `us-west-2` e le richieste di creazione, aggiornamento o eliminazione di MRAP si rivolgono direttamente a tale regione. Il piano di controllo per MRAP ha anche dipendenze sottostanti da AGA in`us-west-2`, Route 53 in e ACM in `us-east-1` ogni regione da cui MRAP è configurato per servire i contenuti. Non si dovrebbe dipendere dalla disponibilità del piano di controllo MRAP nel percorso di ripristino o nei piani dati dei propri sistemi. Questo è diverso dai controlli di failover MRAP che vengono utilizzati per specificare lo stato di routing attivo o passivo per ciascuno dei bucket nel MRAP. Queste API sono ospitate su cinque server Regioni AWS e possono essere utilizzate per spostare efficacemente il traffico utilizzando il piano dati del servizio.

Inoltre, i nomi dei bucket di Amazon S3 sono unici a livello globale e tutte le chiamate alle `DeleteBucket` API `CreateBucket` e dipendono `us-east-1` dalla garanzia dell'unicità dei nomi nella `aws` partizione, anche se la chiamata API è diretta alla regione specifica in cui si desidera creare il bucket. Infine, se hai flussi di lavoro critici per la creazione di bucket, non dovresti dipendere dalla disponibilità di alcuna ortografia specifica del nome di un bucket, in particolare quelli che seguono uno schema riconoscibile.

> **Raccomandazione**
>
> Non fare affidamento sull'eliminazione o sulla creazione di nuovi bucket S3 o sull'aggiornamento delle configurazioni dei bucket S3 come parte del percorso di ripristino. Predisponi tutti i bucket S3 richiesti con le configurazioni necessarie in modo da non dover

apportare modifiche per il ripristino dopo un errore. Questo approccio si applica anche agli MRAP.

- CloudFront

  Amazon API Gateway fornisce endpoint API ottimizzati per l'edge. La creazione di questi endpoint dipende dal piano di CloudFront controllo utilizzato `us-east-1` per creare la distribuzione davanti all'endpoint del gateway.

  > **Raccomandazione**
  >
  > Non fare affidamento sulla creazione di nuovi endpoint API Gateway ottimizzati per l'edge come parte del percorso di ripristino. Effettua il pre-provisioning di tutti gli endpoint API Gateway richiesti.

  Tutte le dipendenze discusse in questa sezione sono azioni del piano di controllo, non azioni del piano dati. Se i carichi di lavoro sono configurati per essere staticamente stabili, queste dipendenze non dovrebbero influire sul percorso di ripristino, tenendo presente che la stabilità statica richiede lavoro o servizi aggiuntivi per l'implementazione.

Servizi che utilizzano endpoint globali predefiniti

In alcuni casi, AWS i servizi forniscono un endpoint globale predefinito, come AWS Security Token Service (AWS STS). Altri servizi possono utilizzare questo endpoint globale predefinito nella loro configurazione predefinita. Ciò significa che un servizio regionale in uso potrebbe avere una dipendenza globale da un singolo servizio. Regione AWS I seguenti dettagli spiegano come rimuovere le dipendenze involontarie dagli endpoint globali predefiniti che ti aiuteranno a utilizzare il servizio in modo regionale.

AWS STS: STS è un servizio web che consente di richiedere credenziali temporanee con privilegi limitati per gli utenti IAM o per gli utenti autenticati (utenti federati). L'utilizzo di STS dal kit di sviluppo AWS software (SDK) e dall'interfaccia a riga di comando (CLI) è predefinito su. `us-east-1` Il servizio STS fornisce anche endpoint regionali. Questi endpoint sono abilitati per impostazione predefinita nelle regioni che sono anch'esse abilitate per impostazione predefinita. Puoi trarne vantaggio in qualsiasi momento configurando il tuo SDK o la CLI seguendo queste istruzioni:AWS Endpoint

regionalizzati STS. L'utilizzo di SIGv4A richiede anche credenziali temporanee richieste da un endpoint STS regionale. Non è possibile utilizzare l'endpoint STS globale per questa operazione.

> **Raccomandazione**
>
> Aggiorna la configurazione dell'SDK e della CLI per utilizzare gli endpoint STS regionali.

Accesso al Security Assertion Markup Language (SAML): i servizi SAML esistono ovunque. Regioni AWSPer utilizzare questo servizio, scegli l'endpoint SAML regionale appropriato, ad esempio https://us-west-2.signin.aws.amazon.com/saml. È necessario aggiornare le configurazioni nelle policy di fiducia e nell'Identity Provider (IdP) per utilizzare gli endpoint regionali. Per dettagli specifici, consulta la documentazione AWS SAML.

Se utilizzi un IdP che è anche ospitato su AWS, c'è il rischio che anche quest'ultimo venga compromesso durante un AWS evento di errore. Ciò potrebbe comportare l'impossibilità di aggiornare la configurazione IdP o la federazione completa. Dovresti predisporre gli utenti «break-glass» nel caso in cui il tuo IdP sia compromesso o non disponibile. Appendice A - Guida al servizio partizionalePer ulteriori informazioni su come creare utenti break-glass in modo staticamente stabile, consulta

> **Raccomandazione**
>
> Aggiorna le policy di trust dei ruoli IAM per accettare gli accessi SAML da più regioni. In caso di errore, aggiorna la configurazione IdP per utilizzare un endpoint SAML regionale diverso se l'endpoint preferito è danneggiato. Crea uno o più utenti break-glass nel caso in cui il tuo IdP sia compromesso o non disponibile.

AWS IAM Identity Center: Identity Center è un servizio basato sul cloud che semplifica la gestione centralizzata dell'accesso Single Sign-On alle applicazioni del cliente e al cloud. Account AWS Identity Center deve essere distribuito in un'unica regione di tua scelta. Tuttavia, il comportamento predefinito del servizio consiste nell'utilizzare l'endpoint SAML globale (https://signin.aws.amazon.com/saml), che è ospitato in. `us-east-1` Se hai distribuito Identity Center in un altro Regione AWS, devi aggiornare l'URL relaystate di ogni set di autorizzazioni in modo che abbia come target lo stesso endpoint della console regionale utilizzato per la distribuzione dell'Identity Center. Ad esempio, se hai distribuito Identity Center in`us-west-2`, devi aggiornare il relaystate dei

set di autorizzazioni per utilizzare https://us-west-2.console.aws.amazon.com. Ciò eliminerà qualsiasi dipendenza dalla distribuzione di Identity `us-east-1` Center.

Inoltre, poiché IAM Identity Center può essere distribuito solo in una singola regione, è necessario predisporre in anticipo gli utenti «break-glass» nel caso in cui l'implementazione sia compromessa. Appendice A - Guida al servizio partizionalePer ulteriori informazioni su come creare utenti break-glass in modo staticamente stabile, consulta

> **Raccomandazione**
>
> Imposta l'URL relaystate dei tuoi set di autorizzazioni in IAM Identity Center in modo che corrisponda alla regione in cui hai distribuito il servizio. Crea uno o più utenti eccezionali nel caso in cui l'implementazione di IAM Identity Center non sia disponibile.

Amazon S3 Storage Lens: Storage Lens fornisce una dashboard predefinita chiamata. default-account-dashboard La configurazione del dashboard e le metriche associate vengono archiviate in. `us-east-1` Puoi creare dashboard aggiuntivi in altre regioni specificando la regione principale per la configurazione della dashboard e i dati delle metriche.

> **Raccomandazione**
>
> Se hai bisogno di dati dalla dashboard predefinita di S3 Storage Lens durante un guasto che influisce sul servizious`-east-1`, crea una dashboard aggiuntiva in una regione di origine alternativa. Puoi anche duplicare qualsiasi altra dashboard personalizzata che hai creato in altre regioni.

Riepilogo dei servizi globali

I piani dati per i servizi globali applicano principi di isolamento e indipendenza simili a quelli AWS dei servizi regionali. Un guasto che influisce sul piano dati di IAM in una regione non influisce sul funzionamento del piano dati IAM in un'altra Regione AWS. Allo stesso modo, un guasto che influisce sul piano dati di Route 53 in un PoP non influisce sul funzionamento del piano dati Route 53 nel resto del. PoPs Pertanto, ciò che dobbiamo considerare sono gli eventi di disponibilità del servizio che influiscono sulla regione in cui opera il piano di controllo o influiscono sul piano di controllo stesso. Poiché esiste un solo piano di controllo per ogni servizio globale, un guasto che influisca su tale piano di controllo potrebbe avere effetti interregionali sulle operazioni di tipo CRUDL (che sono le

operazioni di configurazione che vengono in genere utilizzate per impostare o configurare un servizio anziché l'uso diretto del servizio).

Il modo più efficace per progettare carichi di lavoro in modo da utilizzare i servizi globali in modo resiliente consiste nell'utilizzare la stabilità statica. In uno scenario di errore, progetta il carico di lavoro in modo da non dover apportare modifiche utilizzando un piano di controllo per mitigare l'impatto o il failover su una posizione diversa. Per informazioni prescrittive su come utilizzare questi tipi di servizi globali Appendice B - Guida all'assistenza globale della rete Edge per rimuovere le dipendenze dal piano di controllo Appendice A - Guida al servizio partizionale ed eliminare i singoli punti di errore, fate riferimento a e per ottenere indicazioni prescrittive su come utilizzare questi tipi di servizi globali. Se hai bisogno dei dati di un'operazione del piano di controllo per il ripristino, memorizzali nella cache in un data store a cui è possibile accedere tramite il relativo piano dati, come un parametro AWS Systems Manager Parameter Store (SSM Parameter Store), una tabella DynamoDB o un bucket S3. Per motivi di ridondanza, puoi anche scegliere di archiviare i dati in una regione aggiuntiva. Ad esempio, seguendo le best practice per Route 53 Application Recovery Controller (ARC), è necessario codificare o aggiungere un segnalibro ai cinque endpoint del cluster regionale. Durante un evento di errore, potresti non essere in grado di accedere ad alcune operazioni API, incluse le operazioni dell'API Route 53 ARC che non sono ospitate nel cluster Data Plane estremamente affidabile. È possibile elencare gli endpoint per i cluster Route 53 ARC utilizzando l'operazione `DescribeCluster` API.

Di seguito è riportato un riepilogo di alcune delle configurazioni errate o degli anti-pattern più comuni che introducono dipendenze dai piani di controllo dei servizi globali:

- Apportare modifiche ai record della Route 53, ad esempio aggiornare il valore di un record A o modificare i pesi di un set di record ponderato, per eseguire il failover.
- Creazione o aggiornamento di risorse IAM, inclusi ruoli e policy IAM, durante un failover. In genere non è intenzionale, ma potrebbe essere il risultato di un piano di failover non testato.
- Affidarsi a IAM Identity Center per consentire agli operatori di accedere agli ambienti di produzione durante un evento di guasto.
- Affidarsi alla configurazione predefinita di IAM Identity Center per utilizzare la console `us-east-1` quando Identity Center è stato distribuito in un'altra regione.
- Apportare modifiche ai pesi del traffico AGA per eseguire manualmente un failover regionale.
- Aggiornamento della configurazione di origine di una CloudFront distribuzione per eliminare un'origine compromessa.
- Il provisioning di risorse di disaster recovery (DR), come le istanze ELB e RDS durante un evento di errore, dipende dalla creazione di record DNS in Route 53.

Di seguito è riportato un riepilogo dei consigli forniti in questa sezione per utilizzare i servizi globali in modo resiliente che consentirebbe di prevenire i precedenti anti-pattern comuni.

Riepilogo delle raccomandazioni

Non fare affidamento sui piani di controllo dei servizi partizionali nel percorso di ripristino. Affidatevi invece alle operazioni del piano dati di questi servizi. Appendice A - Guida al servizio partizionalePer ulteriori dettagli su come progettare i servizi partizionali, consulta la sezione.

Non fate affidamento sul piano di controllo dei servizi di rete perimetrale nel vostro percorso di ripristino. Affidati invece alle operazioni del piano dati di questi servizi. Appendice B - Guida all'assistenza globale della rete EdgePer ulteriori dettagli su come progettare servizi globali nella rete perimetrale, vedi.

Non fare affidamento sulla creazione, l'aggiornamento o l'eliminazione di risorse che richiedono la creazione, l'aggiornamento o l'eliminazione dei record di risorse, delle zone ospitate o dei controlli dello stato di Route 53 nel percorso di ripristino. Effettua il pre-provisioning di queste risorse, come gli ELB, per evitare che nel percorso di ripristino si crei una dipendenza dal piano di controllo della Route 53.

Non fare affidamento sull'eliminazione o sulla creazione di nuovi bucket S3 o sull'aggiornamento delle configurazioni dei bucket S3 come parte del percorso di ripristino. Predisponi tutti i bucket S3 richiesti con le configurazioni necessarie in modo da non dover apportare modifiche per il ripristino dopo un errore. Questo approccio si applica anche agli MRAP.

Non fare affidamento sulla creazione di nuovi endpoint API Gateway ottimizzati per l'edge come parte del percorso di ripristino. Effettua il pre-provisioning di tutti gli endpoint API Gateway richiesti.

Aggiorna la configurazione dell'SDK e della CLI per utilizzare gli endpoint STS regionali.

Aggiorna le policy di fiducia dei ruoli IAM per accettare gli accessi SAML da più regioni. In caso di errore, aggiorna la configurazione IdP per utilizzare un endpoint SAML regionale diverso se l'endpoint preferito è danneggiato. Crea utenti eccezionali nel caso in cui il tuo IdP sia compromesso o non disponibile.

Imposta l'URL relaystate dei tuoi set di autorizzazioni in IAM Identity Center in modo che corrisponda alla regione in cui hai distribuito il servizio. Crea uno o più utenti ineccepibili nel caso in cui la distribuzione dell'Identity Center non sia disponibile.

Se hai bisogno di dati dal pannello di controllo predefinito di S3 Storage Lens in caso di guasto che influisce sul servizio in`us-east-1`, crea un pannello di controllo aggiuntivo in una

regione alternativa. Puoi anche duplicare qualsiasi altra dashboard personalizzata che hai creato in altre regioni.

Conclusioni

AWSfornisce diversi costrutti per i limiti di isolamento dei guasti. È necessario considerare come progettare i servizi zonali, regionali e globali, nonché i potenziali impatti sul carico di lavoro e sulla capacità del carico di lavoro di ripristinarsi durante i problemi del piano di controllo. La stabilità statica è uno dei modi principali per evitare le dipendenze dal piano di controllo e creare meccanismi HA e DR affidabili e resilienti quando si utilizzano AWS i servizi.

Appendice A - Guida al servizio partizionale

Per i servizi partizionali, è necessario implementare la stabilità statica per mantenere la resilienza del carico di lavoro durante un danneggiamento del piano di controllo del AWS servizio. Di seguito vengono fornite indicazioni prescrittive su come considerare le dipendenze dai servizi partizionali e su ciò che può funzionare o meno durante un danneggiamento del piano di controllo.

AWS Identity and Access Management (IAM)

Il piano di controllo AWS Identity and Access Management (IAM) è composto da tutte le API IAM pubbliche (incluso Access Advisor ma non Access Analyzer o IAM Roles Anywhere). Ciò include azioni come `CreateRoleAttachRolePolicy`,`ChangePassword`,`UpdateSAMLProvider`, e`UpdateLoginProfile`. Il piano dati IAM fornisce l'autenticazione e l'autorizzazione per i principi IAM in ciascuno di essi. Regione AWS Durante un guasto al piano di controllo, le operazioni di tipo CRUDL per IAM potrebbero non funzionare, ma l'autenticazione e l'autorizzazione per i presidi esistenti continueranno a funzionare. STS è un servizio che riguarda solo il piano dati, separato da IAM e non dipende dal piano di controllo IAM.

Ciò significa che quando si pianificano le dipendenze da IAM, non è necessario fare affidamento sul piano di controllo IAM nel percorso di ripristino. Ad esempio, un progetto staticamente stabile per un utente amministratore «infrangibile» consiste nel creare un utente con le autorizzazioni appropriate allegate, impostare la password e fornire la chiave di accesso e la chiave di accesso segreta e quindi bloccare tali credenziali in un deposito fisico o virtuale. Se necessario durante un'emergenza, recupera le credenziali utente dal vault e utilizzale secondo necessità. Una non-statically-stable soluzione potrebbe consistere nel fornire i dati all'utente durante un errore, oppure nel predisporre l'utente, ma allegando la politica di amministrazione solo quando necessario. Questi approcci dipenderebbero dal piano di controllo IAM.

AWS Organizations

Il piano di AWS Organizations controllo è composto da tutte le API delle organizzazioni pubbliche come `AcceptHandshakeAttachPolicy`,, `CreateAccountCreatePolicy`, `eListAccounts`. Non esiste un piano dati perAWS Organizations. Orchestra il piano dati per altri servizi come IAM. Durante un guasto al piano di controllo, le operazioni di tipo CRUDL per Organizations potrebbero non funzionare, ma le politiche, come Service Control Policies (SCP) e Tag Policies, continueranno a funzionare e saranno valutate come parte del processo di autorizzazione IAM. Anche le funzionalità

di amministrazione delegata e le funzionalità multi-account in altri AWS servizi supportati dalle Organizations continueranno a funzionare.

Ciò significa che quando si pianificano le dipendenze daAWS Organizations, non è necessario fare affidamento sul piano di controllo dell'Organizations nel percorso di ripristino. Implementa invece la stabilità statica nel tuo piano di ripristino. Ad esempio, un non-statically-stable approccio potrebbe consistere nell'aggiornare gli SCP per rimuovere le restrizioni sui permessi consentiti Regioni AWS tramite la `aws:RequestedRegion` condizione o per abilitare le autorizzazioni di amministratore per ruoli IAM specifici. Ciò si basa sul piano di controllo dell'Organizations per effettuare questi aggiornamenti. Un approccio migliore sarebbe quello di utilizzare i tag di sessione per concedere l'uso delle autorizzazioni di amministratore. Il tuo Identity Provider (IdP) può includere tag di sessione che possono essere valutati in base alla `aws:PrincipalTag` condizione, il che ti aiuta a configurare dinamicamente le autorizzazioni per determinati principi, aiutando i tuoi SCP a rimanere statici. Ciò rimuove le dipendenze dai piani di controllo e utilizza solo le azioni del piano dati.

AWS Account Management

Il piano di controllo di AWS Account Management è ospitato in us-east-1 e comprende tutte le API pubbliche per la gestione di unAccount AWS, ad esempio e. `GetContactInformation` `PutContactInformation` Include anche la creazione o la chiusura di un nuovo Account AWS tramite la console di gestione. Le API per`CloseAccount`, `CreateAccountCreateGovCloudAccount`, e `DescribeAccount` fanno parte del piano di AWS Organizations controllo, anch'esso ospitato in us-east-1. Inoltre, la creazione di un GovCloud account all'esterno di AWS Organizations si basa sul piano di controllo di Account AWS gestione in us-east-1. Inoltre, GovCloud gli account devono essere collegati 1:1 a un elemento Account AWS nella `aws` partizione. La creazione di account nella `aws-cn` partizione non si basa su Il piano dati per Account AWS sono gli account stessi. Durante un guasto al piano di controllo, le operazioni di tipo CRUDL (come la creazione di un nuovo account o la raccolta e l'aggiornamento delle informazioni di contatto) potrebbero non funzionare. Account AWS I riferimenti all'account nelle policy IAM continueranno a funzionare.

Ciò significa che quando si pianificano dipendenze da AWS Account Management, non è necessario fare affidamento sul piano di controllo di Account Management nel percorso di ripristino. Sebbene il piano di controllo di Account Management non offra funzionalità dirette che normalmente utilizzeresti in una situazione di ripristino, in alcuni casi potrebbero esserlo. Ad esempio, una progettazione staticamente stabile consisterebbe nel predisporre tutto il necessario per il Account AWS failover. Un non-statically-stable progetto potrebbe essere quello di crearne uno nuovo Account AWS durante un evento di guasto per ospitare le risorse di ripristino di emergenza.

Route 53 Application Recovery Controller

Il piano di controllo per Route 53 ARC è costituito dalle API per il controllo e la preparazione al ripristino, come indicato in: Endpoint e quote di Amazon Route 53 Application Recovery Controller. È possibile gestire i controlli di idoneità, i controlli di routing e le operazioni dei cluster utilizzando il piano di controllo. Il piano dati di ARC è il cluster di ripristino, che gestisce i valori di controllo del routing interrogati dai controlli di integrità di Route 53 e implementa anche le regole di sicurezza. La funzionalità del piano dati di Route 53 ARC è accessibile tramite le API del cluster di ripristino, ad esempio. `https://aaaaaaaa.route53-recovery-cluster.eu-west-1.amazonaws.com`

Ciò significa che non dovresti fare affidamento sul piano di controllo Route 53 ARC nel tuo percorso di ripristino. Esistono due best practice che aiutano a implementare questa guida:

- Innanzitutto, aggiungi ai segnalibri o codifica i cinque endpoint del cluster regionale. Ciò elimina la necessità di utilizzare il funzionamento del piano DescribeCluster di controllo durante uno scenario di failover per scoprire i valori degli endpoint.
- In secondo luogo, utilizza le API del cluster Route 53 ARC utilizzando la CLI o l'SDK per eseguire aggiornamenti ai controlli di routing e non al. AWS Management Console In questo modo la console di gestione viene rimossa come dipendenza per il piano di failover e garantisce che dipenda solo dalle azioni del piano dati.

AWS Network Manager

Il servizio AWS Network Manager è principalmente un sistema che utilizza solo il piano di controllo ospitato in us-west-2. Il suo scopo è gestire centralmente la configurazione della tua Cloud AWS rete core di rete di AWS Transit Gateway ((((((((((((((((Account AWS Inoltre, aggrega le metriche Cloud WAN in us-west-2, a cui è possibile accedere anche tramite il piano dati. CloudWatch Se Network Manager è compromesso, il piano dati dei servizi che orchestra non ne risentirà. Le CloudWatch metriche per Cloud WAN (Se desideri che dati metrici storici, come i byte in entrata e in uscita per regione, capiscano quanto traffico potrebbe essere trasferito verso altre regioni durante un errore che ha un impatto su us-west-2 o per altri scopi operativi, puoi esportare tali metriche come dati CSV direttamente dalla CloudWatch console o utilizzando questo metodo: Pubblica le metriche Amazon in un file CSV. CloudWatch I dati si trovano nel `AWS/Network Manager` namespace ed è possibile eseguire questa operazione in base a una pianificazione scelta e archiviarli in S3 o in un altro archivio dati selezionato. Per implementare un piano di ripristino staticamente stabile, non utilizzare AWS Network Manager per aggiornare la rete né fare affidamento sui dati delle operazioni del piano di controllo per l'input del failover.

DNS privato Route 53

Le zone ospitate private Route 53 sono supportate in ogni partizione; tuttavia, le considerazioni per le zone ospitate private e le zone ospitate pubbliche in Route 53 sono le stesse. Fai riferimento ad Amazon Route 53 nell'Appendice B - Guida all'assistenza globale per la rete Edge.

Appendice B - Guida all'assistenza globale della rete Edge

Per i servizi globali della rete edge, è necessario implementare la stabilità statica per mantenere la resilienza del carico di lavoro durante un deterioramento del piano di controllo del AWS servizio.

Route 53

Il piano di controllo di Route 53 è composto da tutte le API pubbliche di Route 53 che coprono funzionalità per zone ospitate, record, controlli di integrità, registri di query DNS, set di deleghe riutilizzabili, politiche di traffico e tag di allocazione dei costi. È ospitazione in us-east-1. Il piano dei dati è il servizio DNS autoritativo, che funziona in oltre 200 punti PoP e che risponde alle query DNS basate sulle tue zone ospitate e sui dati di controllo dell'integrità. Regione AWS Inoltre, Route 53 dispone di un piano dati per i controlli sanitari, che è anche un servizio distribuito a livello globale su più dispositivi. Regioni AWS Questo piano dei dati esegue i controlli dell'integrità, aggrega i risultati e li consegna ai piani dati di Route 53. Durante un guasto al piano di controllo, le operazioni di tipo CRUDL per la Route 53 potrebbero non funzionare, ma la risoluzione e i controlli di integrità del DNS e gli aggiornamenti del routing derivanti da modifiche nei controlli sanitari continueranno a funzionare.

Ciò significa che quando si pianificano dipendenze dalla Route 53, non è necessario fare affidamento sul piano di controllo della Route 53 nel percorso di ripristino. Ad esempio, una progettazione staticamente stabile consiste nell'utilizzare lo stato dei controlli sanitari per eseguire failover tra regioni o per evacuare una zona di disponibilità. È possibile utilizzare i controlli di routing ARC (Application Recovery Controller) di Route 53 per modificare manualmente lo stato dei controlli di integrità e modificare le risposte alle domande DNS. Esistono modelli simili a quelli forniti da ARC che è possibile implementare in base alle proprie esigenze. Alcuni di questi modelli sono descritti nella sezione Creazione di meccanismi di ripristino di emergenza utilizzando Route 53 e nella sezione sugli interruttori automatici di controllo dello stato di salute di Advanced Multi-AZ Resilience Patterns. Se hai scelto di utilizzare un piano di ripristino di emergenza multiregionale, predisponi le risorse che richiedono la creazione di record DNS, come le istanze ELB e RDS. Un non-statically-stable progetto potrebbe consistere nell'aggiornare il valore di un record di risorse Route 53 tramite l'`ChangeResourceRecordSets`API, modificare il peso di un record ponderato o creare nuovi record per eseguire il failover. Questi approcci dipendono dal piano di controllo della Route 53.

Amazon CloudFront

Il piano di CloudFront controllo di Amazon è composto da tutte le CloudFront API pubbliche per la gestione delle distribuzioni ed è ospitato in us-east-1. Il piano dati è la distribuzione stessa fornita

dalla PoPs rete perimetrale. Esegue la gestione delle richieste, il routing e la memorizzazione nella cache dei contenuti di origine. Durante un guasto al piano di controllo, le operazioni di tipo CRUDL CloudFront (comprese le richieste di annullamento della validità) potrebbero non funzionare, ma i tuoi contenuti continueranno a essere memorizzati nella cache e serviti e i failover di origine continueranno a funzionare.

Ciò significa che quando si pianificano le dipendenze daCloudFront, non è necessario fare affidamento sul piano di CloudFront controllo nel percorso di ripristino. Ad esempio, una progettazione staticamente stabile prevede l'utilizzo di failover di origine automatizzati per mitigare l'impatto di una lesione su una delle origini. Puoi anche scegliere di creare il bilanciamento del carico di origine o il failover utilizzando Lamda @Edge, fare riferimento a Tre modelli di progettazione avanzati per applicazioni ad alta disponibilità utilizzando Amazon CloudFront e Utilizzo di Amazon CloudFront e Amazon S3 per creare applicazioni di geo-prossimità attive e multiregionali per maggiori dettagli su tale modello. Un non-statically-stable progetto potrebbe consistere nell'aggiornare manualmente la configurazione della distribuzione in risposta a un errore di origine. Questo approccio dipenderebbe dal piano CloudFront di controllo.

Amazon Certificate Manager

Se utilizzi certificati personalizzati con la tua CloudFront distribuzione, dipendi anche da ACM. L'utilizzo dello CloudFront scudo di controllo nella regione us-east-1. Durante un guasto al piano di controllo, i certificati esistenti configurati nella distribuzione continueranno a funzionare così come i rinnovi automatici dei certificati. Non fare affidamento sulla modifica della configurazione della distribuzione o sulla creazione di nuovi certificati come parte del percorso di ripristino.

AWSWeb Application Firewall (WAF) e WAF Classic

Se lo utilizzi AWS WAF con la tua CloudFront distribuzione, dipendi dal piano di controllo WAF, anch'esso ospitato nella regione us-east-1. Durante un guasto al piano di controllo, gli elenchi di controllo degli accessi Web (ACL) configurati e le regole associate continuano a funzionare. Non fare affidamento sull'aggiornamento degli ACL Web WAF come parte del percorso di ripristino.

AWS Global Accelerator

Il piano di controllo AGA è composto da tutte le API AGA pubbliche ed è ospitato in us-west-2. Il piano dati è il routing di rete degli indirizzi IP anycast forniti da AGA agli endpoint registrati. AGA utilizza anche i controlli di integrità della Route 53 per determinare lo stato degli endpoint AGA, che fa parte del piano dati Route 53. Durante un guasto al piano di controllo, le operazioni di tipo CRUDL

per AGA potrebbero non funzionare. L'instradamento verso gli endpoint esistenti, così come i controlli di integrità, le chiamate stradali e le configurazioni di peso degli endpoint esistenti utilizzate per indirizzare o spostare il traffico verso altri endpoint e gruppi di endpoint, continueranno a funzionare.

Ciò significa che quando si pianificano dipendenze da AGA, non è necessario fare affidamento sul piano di controllo AGA nel percorso di ripristino. Ad esempio, una progettazione staticamente stabile consiste nell'utilizzare lo stato dei controlli di integrità configurati per eliminare gli endpoint non integri. Per esempi di questa configurazione, consulta Implementazione di applicazioni multiregionali AWS utilizzando AWS Global Accelerator. Un non-statically-stable progetto potrebbe consistere nel modificare le percentuali di chiamata del traffico AGA, modificare i gruppi di endpoint o rimuovere un endpoint da un gruppo di endpoint durante un problema. Questi approcci dipenderebbero dal piano di controllo AGA.

Shield

Il piano di controllo Amazon Shield Advanced è composto da tutte le API Shield Advanced pubbliche ed è ospitato in us-east-1. Ciò include funzionalità come `CreateProtectionCreateProtectionGroup`,`AssociateHealthCheck`,`DesribeDRTAccess`, `eListProtections`. Il piano dati è la protezione DDoS fornita da Shield Advanced e la creazione di metriche Shield Advanced. Shield Advanced utilizza anche i controlli di integrità della Route 53 (che fanno parte del piano dati Route 53), se li hai configurati. Durante un guasto al piano di controllo, le operazioni di tipo CRUDL per Shield Advanced potrebbero non funzionare, ma la protezione DDoS configurata per le tue risorse, così come le risposte ai cambiamenti nei controlli sanitari, continueranno a funzionare.

Ciò significa che non è necessario fare affidamento sul piano di controllo Shield Advanced nel percorso di ripristino. Sebbene il piano di controllo Shield Advanced non fornisca funzionalità dirette che normalmente si utilizzano in una situazione di ripristino, in alcuni casi è possibile che ciò avvenga. Ad esempio, una progettazione staticamente stabile prevede che le risorse di ripristino di emergenza siano già configurate per far parte di un gruppo di protezione e che vengano associati controlli di integrità anziché configurare tale protezione dopo il verificarsi dell'errore. Ciò impedisce di dipendere dal piano di controllo Shield Advanced per il ripristino.

Appendice C - Servizi a regione singola

Di seguito è riportato un elenco di servizi o funzionalità specifiche di quel servizio (che sono elencate tra parentesi dopo il nome del servizio), disponibili solo in una singola regione. Le stesse linee guida per l'implementazione della stabilità statica fornite per altri servizi globali si applicano a questi servizi quando è necessario pianificare le dipendenze dai relativi piani di controllo e piani dati.

- Alexa for Business
- Marketplace AWS(API Marketplace AWS del catalogo, Marketplace AWS Commerce Analytics, Marketplace AWS Entitlement Service)
- Billing and Cost Management (AWS Cost Explorerreport su AWS costi e utilizzo, AWS budget, Savings Plans)
- AWS BugBust
- Amazon Mechanical Turk
- Amazon Chime
- Amazon Chime SDK (audio PSTN, messaggistica, identità)
- AWSChatbot
- AWS DeepRacer
- AWSDevice Farm
- Amazon GameSparks
- Amazon Honeycode

Fattori determinanti

I contributori a questo documento includono:

- Michael Haken, Principal Solutions Architect, Amazon Web Services

Revisioni del documento

Per ricevere una notifica sugli aggiornamenti del white paper, è possibile iscriversi al feed RSS.

Modifica	Descrizione	Data
Revisione minore	Guida aggiornata per allinearsi alle best practice IAM. Per ulteriori informazioni, consulta la sezione Best practice per la sicurezza in IAM	9 febbraio 2023
Pubblicazione iniziale	Whitepaper pubblicato.	16 novembre 2022

Glossario AWS

Per la terminologia AWS più recente, consultare il glossario AWS nella documentazione di riferimento per Glossario AWS.

Note

I clienti sono responsabili della propria valutazione indipendente delle informazioni contenute in questo documento. Questo documento: (a) è solo a scopo informativo, (b) rappresenta le attuali offerte e pratiche di AWS prodotti, che sono soggette a modifiche senza preavviso e (c) non crea alcun impegno o garanzia da parte AWS e dei suoi affiliati, fornitori o licenzianti. AWSi prodotti o i servizi sono forniti «così come sono» senza garanzie, dichiarazioni o condizioni di alcun tipo, espresse o implicite. Le responsabilità e le responsabilità nei AWS confronti dei propri clienti sono controllate da AWS accordi e il presente documento non fa parte, né modifica, alcun accordo tra AWS e i suoi clienti.

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.