AWS Documento técnico

Límites de aislamiento de errores de AWS

Límites de aislamiento de errores de AWS: AWS Documento técnico

Table of Contents

AWS Fault Aislation límites

Fecha de publicación: 16 de noviembre del 2022 (Revisiones del documento)

Resumen

Amazon Web Services (AWS) proporciona diferentes límites de aislamiento, como zonas de disponibilidad (AZ), regiones, planos de control y planos de datos. Este paper detalla cómo se AWS utilizan estos límites para crear servicios zonales, regionales y globales. También incluye una guía prescriptiva sobre cómo considerar las dependencias de estos diferentes servicios y cómo mejorar la resiliencia de las cargas de trabajo que se generan al utilizarlos.

¿El enfoque Well-Architected de Well-Architected?

El marco AWS Well-Architected le ayuda a comprender las ventajas y desventajas de las decisiones que toma al crear sistemas en la nube. Los seis pilares del Marco le permiten aprender las mejores prácticas arquitectónicas para diseñar y operar sistemas confiables, seguros, eficientes, rentables y sostenibles. Con el AWS Well-Architected Tool, disponible sin coste alguno en el AWS Management Console, puede comparar sus cargas de trabajo con estas mejores prácticas respondiendo a una serie de preguntas para cada pilar.

Para obtener más orientación de expertos y mejores prácticas para su arquitectura de nube (implementaciones de arquitectura de referencia, diagramas y documentos técnicos), consulte el Centro de arquitectura. AWS

Introducción

AWSopera una infraestructura global para proporcionar servicios en la nube que ayudan a los clientes a implementar cargas de trabajo de manera flexible, segura, escalable y con alta disponibilidad. La AWS infraestructura utiliza múltiples construcciones de aislamiento de fallas para ayudar a los clientes a alcanzar sus objetivos de resiliencia. Estos límites de aislamiento de fallas permiten a los clientes diseñar sus cargas de trabajo para aprovechar el alcance predecible de la contención de impactos que ofrecen. También es importante entender cómo se diseñan AWS los servicios utilizando estos límites, de modo que pueda tomar decisiones intencionales sobre las dependencias que seleccione para su carga de trabajo.

Este paper resumirá primero la infraestructura AWS global y los límites de aislamiento de fallas que proporciona, así como algunos de los patrones utilizados para diseñar nuestros servicios. Utilizando esta base de comprensión, el paper describirá a continuación los diferentes alcances de los servicios AWS prestados: zonales, regionales y globales. También presentará las mejores prácticas para crear arquitecturas que utilicen estos límites de aislamiento y diferentes ámbitos de servicio a fin de mejorar la resiliencia de las cargas de trabajo en las que se ejecuta. AWS En particular, proporciona una guía prescriptiva sobre cómo adoptar las dependencias de los servicios globales y, al mismo tiempo, minimizar los puntos únicos de falla. Esto le ayudará a tomar decisiones informadas sobre sus AWS dependencias y sobre cómo diseñar su carga de trabajo para la alta disponibilidad (HA) y la recuperación ante desastres (DR).

infraestructura de AWS

En esta sección se presenta un resumen de la infraestructura AWS global y de los límites de aislamiento de fallas que proporciona. Además, en esta sección se proporcionará una visión general del concepto de planos de control y planos de datos, que son distinciones fundamentales a la hora de AWS diseñar sus servicios. Esta información proporciona la base para comprender cómo se aplican los límites de aislamiento de fallas y el plano de control y el plano de datos de un servicio a los tipos de AWS servicio que analizamos en la siguiente sección.

Temas

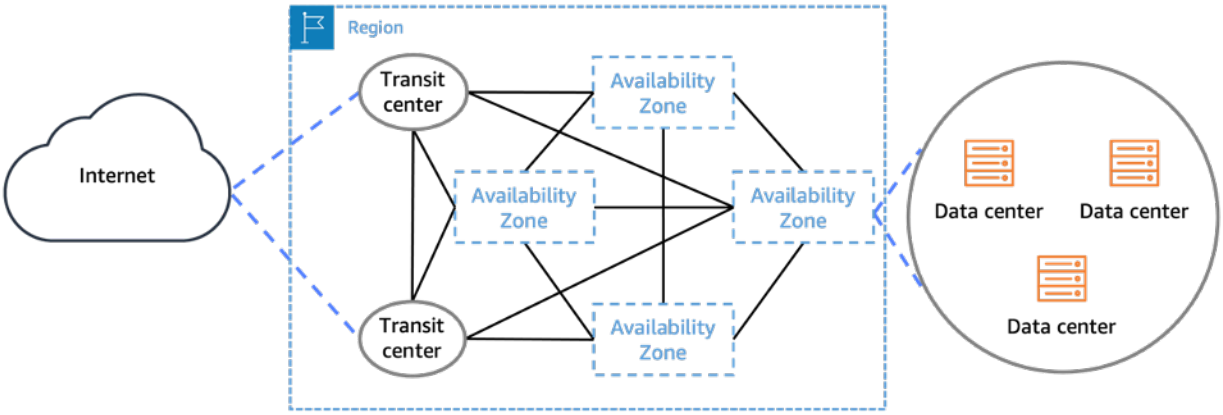
Zonas de disponibilidad

AWSopera más de 100 zonas de disponibilidad en varias regiones del mundo (las cifras actuales se encuentran aquí: Infraestructura AWS global). Una zona de disponibilidad es uno o más centros de datos discretos con una infraestructura de alimentación, redes y conectividad independientes y redundantes en unRegión de AWS. Las zonas de disponibilidad de una región están significativamente distantes entre sí, hasta 60 millas (~100 km) para evitar fallos correlacionados, pero lo suficientemente cerca como para utilizar la replicación sincrónica con una latencia de milisegundos de un solo dígito. Están diseñadas para no verse afectadas simultáneamente por un escenario de destino compartido, como el suministro eléctrico, las interrupciones del suministro de agua, el aislamiento de la fibra, los terremotos, los incendios, los tornados o las inundaciones. Los puntos de fallo más comunes, como los generadores y los equipos de refrigeración, no se comparten entre las distintas zonas de disponibilidad y están diseñados para ser alimentados por subestaciones eléctricas independientes. Cuando AWS implementa actualizaciones de sus servicios,

las implementaciones en las zonas de disponibilidad de la misma región se separan en el tiempo para evitar fallos correlacionados.

Todas las zonas de disponibilidad de una región están interconectadas con redes de gran ancho de banda y baja latencia, a través de fibra metropolitana dedicada y totalmente redundante. Cada zona de disponibilidad de una región se conecta a Internet a través de dos centros de tránsito que AWS coinciden con varios proveedores de Internet de nivel 1 (para obtener más información, consulte Descripción general de Amazon Web Services).

Estas características proporcionan un fuerte aislamiento entre las zonas de disponibilidad, lo que denominamos independencia de la zona de disponibilidad (AZI). La estructura lógica de las zonas de disponibilidad y su conectividad a Internet se muestra en la siguiente figura.



Las zonas de disponibilidad constan de uno o más centros de datos físicos que están conectados de forma redundante entre sí y a Internet

Regiones

Cada una Región de AWS consta de varias zonas de disponibilidad independientes y físicamente separadas dentro de un área geográfica. Actualmente, todas las regiones tienen tres o más zonas de disponibilidad. Las propias regiones están aisladas y son independientes de otras regiones, con algunas excepciones que se indican más adelante en este documento (consulte la sección Operaciones globales de una sola región). Esta separación entre regiones limita los fallos de servicio, cuando se producen, a una sola región. En este caso, las operaciones normales de otras regiones no se ven afectadas. Además, los recursos y los datos que cree en una región no existen en ninguna otra, a menos que utilice explícitamente una función de replicación o copia ofrecida por un AWS servicio o replique el recurso usted mismo.

Regiones de AWS actuales y planificadas a partir de diciembre de 2022

AWSZonas Locales

AWSLas Zonas Locales son un tipo de implementación de infraestructura que coloca la computación, el almacenamiento, las bases de datos y otros AWSservicios selectos cerca de grandes centros industriales y de población. Puede usar AWS servicios, como los servicios de cómputo y almacenamiento, en la zona local para ejecutar aplicaciones de baja latencia en la periferia o simplificar las migraciones a la nube híbrida. Las Zonas Locales tienen acceso y salida de Internet local para reducir la latencia, pero también están conectadas a su región principal a través de la red privada redundante y de gran ancho de banda de Amazon, lo que proporciona a las aplicaciones que se ejecutan en las zonas AWS locales un acceso rápido, seguro y sin problemas a toda la gama de servicios.

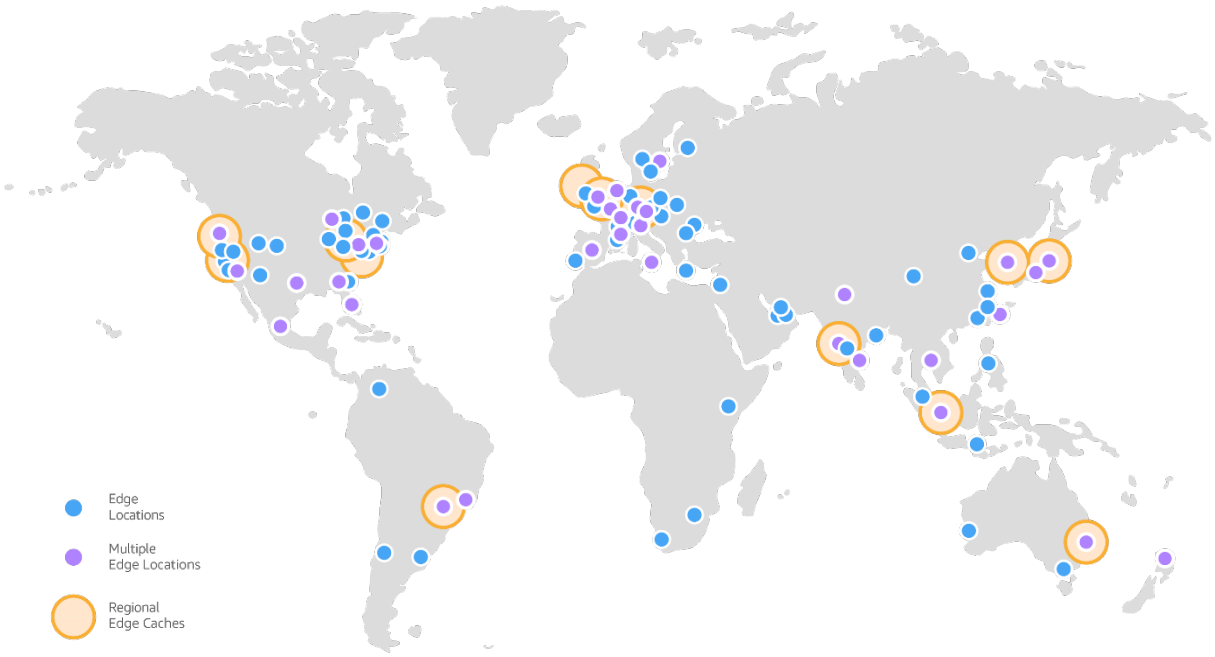
AWS Outposts

AWS Outpostses una familia de soluciones totalmente gestionadas que ofrecen AWS infraestructura y servicios a prácticamente cualquier ubicación local o perimetral para ofrecer una experiencia híbrida realmente coherente. Las soluciones de Outposts le permiten ampliar y ejecutar los AWS servicios nativos de forma local, y están disponibles en una variedad de formatos, desde servidores Outposts de 1U y 2U hasta racks Outposts de 42U y despliegues de múltiples racks.

Con élAWS Outposts, puede ejecutar determinados AWS servicios de forma local y conectarse a una amplia gama de servicios disponibles en el servidor principal. Región de AWS AWS Outpostsson racks de cómputo y almacenamiento totalmente gestionados y configurables, construidos con hardware AWS diseñado que permite a los clientes ejecutar el procesamiento y el almacenamiento en sus instalaciones, a la vez que se conectan sin problemas a AWS la amplia gama de servicios en la nube.

Puntos de presencia

Además de las zonas de disponibilidad Regiones de AWS y disponibilidad, AWS también opera una red de puntos de presencia (PoP) distribuidos a nivel mundial. Estos PoPs alojan Amazon CloudFront, una red de entrega de contenido (CDN); Amazon Route 53, un servicio de resolución del Sistema de Nombres de Dominio (DNS) público; y AWS Global Accelerator (AGA), un servicio de optimización de redes periféricas. La red perimetral global consta actualmente de más de 410 PoPs, incluidas más de 400 ubicaciones perimetrales y 13 cachés regionales de nivel medio en más de 90 ciudades de 48 países (el estado actual puede consultarse aquí: Amazon CloudFront Key Features).



Red perimetral CloudFront global de Amazon

Cada PoP está aislado de los demás, lo que significa que una falla que afecte a una sola PoP o área metropolitana no afectará al resto de la red global. La AWS red es similar a la de miles de operadores de telecomunicaciones de nivel 1/2/3 en todo el mundo, está bien conectada con las

principales redes de acceso para ofrecer un rendimiento óptimo y tiene una capacidad desplegada de cientos de terabits. Las ubicaciones periféricas se conectan a Regiones de AWS través de la AWS red troncal, una fibra paralela múltiple de 100 GbE totalmente redundante que rodea el mundo y se conecta con decenas de miles de redes para mejorar las búsquedas de origen y acelerar el contenido dinámico.

Particiones

AWSagrupa las regiones en particiones. Cada región está exactamente en una partición y cada partición tiene una o más regiones. Las particiones tienen instancias independientes de AWS Identity and Access Management (IAM) y proporcionan un límite estricto entre las regiones de las distintas particiones. AWSLas regiones comerciales están en la `aws` partición, las regiones de China están en la `aws-cn` partición y AWS GovCloud las regiones están en la `aws-us-gov` partición. Algunos AWS servicios están diseñados para ofrecer funcionalidad entre regiones, como la replicación entre regiones de Amazon S3 o la interconexión entre regiones de AWSTransit Gateway. Estos tipos de capacidades solo se admiten entre regiones de la misma partición. No puede usar las credenciales de IAM de una partición para interactuar con los recursos de otra partición.

Planos de control y planos de datos

AWSsepara la mayoría de los servicios en los conceptos de plano de control y plano de datos. Estos términos provienen del mundo de las redes, específicamente de los enrutadores. El plano de datos del router, que es su principal funcionalidad, consiste en mover los paquetes según reglas. Pero las políticas de enrutamiento deben crearse y distribuirse desde algún lugar, y ahí es donde entra en juego el plano de control.

Los planos de control proporcionan las API administrativas que se utilizan para crear, leer/describir, actualizar, eliminar y enumerar los recursos (CRUDL). Por ejemplo, todas las acciones del plano de control son las siguientes: lanzar una nueva instancia de Amazon Elastic Compute Cloud (Amazon EC2), crear un bucket de Amazon Simple Storage Service (Amazon S3) y describir una cola de Amazon Simple Queue Service (Amazon SQS). Al lanzar una instancia EC2, el plano de control debe realizar varias tareas, como encontrar un host físico con capacidad, asignar las interfaces de red, preparar un volumen de Amazon Elastic Block Store (Amazon EBS), generar credenciales de IAM, añadir las reglas del grupo de seguridad y mucho más. Los planos de control suelen ser sistemas complicados de organización y agregación.

El plano de datos es lo que proporciona la función principal del servicio. Por ejemplo, las siguientes son todas las partes del plano de datos de cada uno de los servicios involucrados: la propia instancia

de EC2 en ejecución, la lectura y escritura en un volumen de EBS, la obtención y colocación de objetos en un depósito de S3 y la respuesta de Route 53 a las consultas de DNS y a la realización de comprobaciones de estado.

Los planos de datos son intencionalmente menos complicados y tienen menos partes móviles en comparación con los planos de control, que suelen implementar un sistema complejo de flujos de trabajo, lógica empresarial y bases de datos. Esto hace que, desde el punto de vista estadístico, sea menos probable que se produzcan eventos de falla en el plano de datos que en el plano de control. Si bien tanto el plano de datos como el de control contribuyen al funcionamiento general y al éxito del servicio, los AWS considera componentes distintos. Esta separación tiene beneficios tanto en términos de rendimiento como de disponibilidad.

Estabilidad estática

Una de las características de resiliencia más importantes de los AWS servicios es lo que se AWS denomina estabilidad estática. Lo que este término significa es que los sistemas funcionan en un estado estático y siguen funcionando con normalidad sin necesidad de realizar cambios durante el fallo o la falta de disponibilidad de las dependencias. Una forma de hacerlo es evitar las dependencias circulares en nuestros servicios que podrían impedir que uno de esos servicios se recupere correctamente. Otra forma de hacerlo es manteniendo el estado actual. Consideramos el hecho de que los planos de control tienen estadísticamente más probabilidades de fallar que los planos de datos. Si bien el plano de datos suele depender de los datos que llegan desde el plano de control, el plano de datos mantiene su estado actual y sigue funcionando incluso cuando el plano de control está dañado. El acceso a los recursos del plano de datos, una vez aprovisionado, no depende del plano de control y, por lo tanto, no se ve afectado por ninguna alteración del plano de control. En otras palabras, incluso si la capacidad de crear, modificar o eliminar recursos se ve afectada, los recursos existentes permanecen disponibles. Esto hace que AWS los planos de datos se mantengan estáticamente estables ante una alteración en el plano de control. Puede implementar diferentes patrones para mantener una estabilidad estática frente a diferentes tipos de fallas de dependencia.

Puede encontrar un ejemplo de estabilidad estática en Amazon EC2. Una vez que se ha lanzado una instancia de EC2, está tan disponible como el servidor físico de un centro de datos. No depende de ninguna API del plano de control para seguir funcionando o volver a funcionar tras un reinicio. La misma propiedad se aplica a otros AWS recursos, como las VPC, los buckets y objetos de Amazon S3 y los volúmenes de Amazon EBS.

La estabilidad estática es un concepto que está profundamente arraigado en la forma en que AWS diseña sus servicios, pero también es un patrón que pueden utilizar los clientes. De hecho, la mayoría de las recomendaciones sobre mejores prácticas para utilizar los diferentes tipos de AWS servicios de forma resiliente consisten en implementar la estabilidad estática en los entornos de producción. Los mecanismos de recuperación y mitigación más confiables son los que requieren menos cambios para lograr la recuperación. En lugar de confiar en el plano de control de EC2 para lanzar nuevas instancias de EC2 para recuperarse de una zona de disponibilidad fallida, disponer de esa capacidad adicional previamente aprovisionada ayuda a lograr la estabilidad estática. Por lo tanto, eliminar las dependencias de los planos de control (las API que implementan los cambios en los recursos) de su ruta de recuperación ayuda a generar cargas de trabajo más resilientes. Para obtener más información sobre la estabilidad estática, los planos de control y los planos de datos, consulte el artículo Static Stability using Availability Zones de Amazon Builders Library.

Resumen

AWSutiliza diferentes contenedores de fallas en nuestra infraestructura para aislar las fallas. Los principales contenedores de fallas de la infraestructura son las particiones, las regiones, las zonas de disponibilidad, los planos de control y los planos de datos. A continuación, examinaremos los diferentes tipos de AWS servicios, cómo se utilizan estos contenedores de fallas en su diseño y cómo se deben diseñar las cargas de trabajo con ellos para que sean resilientes.

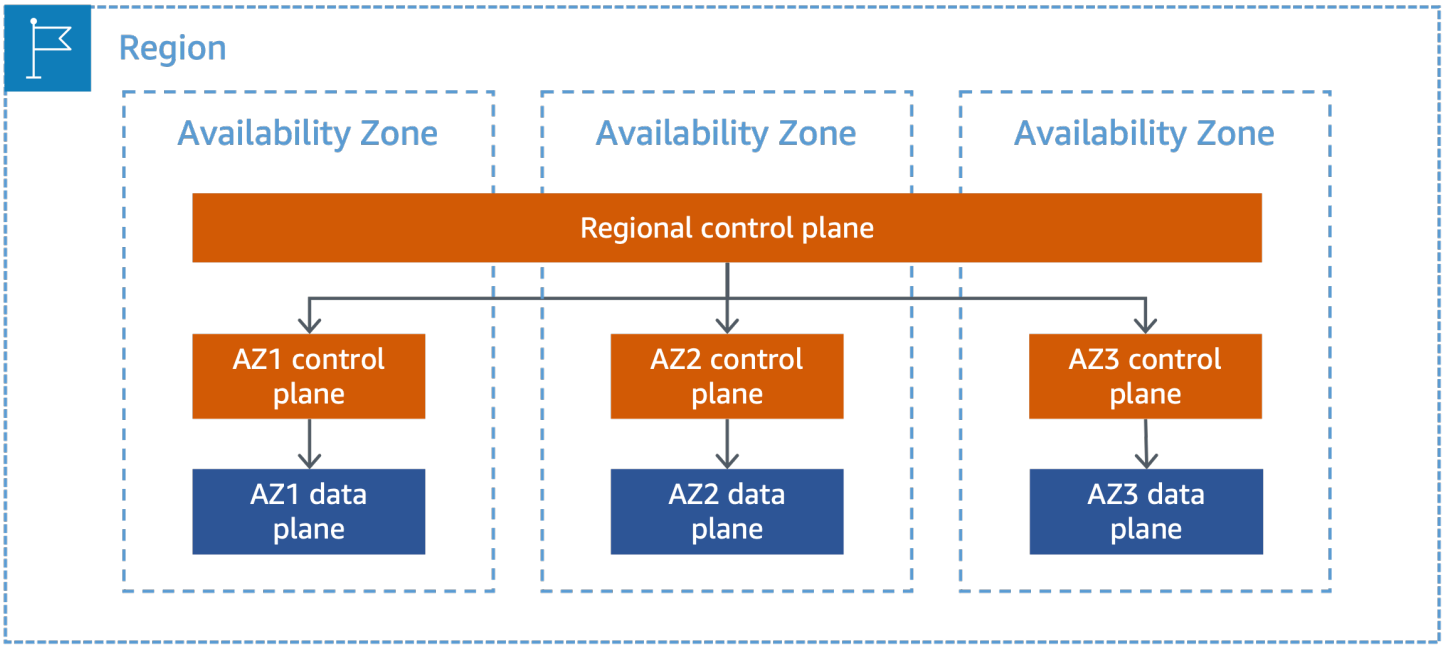
AWS tipos de servicio

AWS opera tres categorías diferentes de servicios en función de su límite de aislamiento de fallas: zonales, regionales y globales. En esta sección se describe con más detalle cómo se han diseñado estos distintos tipos de servicios para que pueda determinar cómo afectarán los fallos de un servicio de un determinado tipo de servicio a su carga de trabajo. AWS También proporciona orientación de alto nivel sobre cómo diseñar sus cargas de trabajo para utilizar estos servicios de forma resiliente. En el caso de los servicios globales, este documento también proporciona una guía prescriptiva Apéndice B: Guía de servicio global de redes periféricas que puede ayudarle a evitar que las deficiencias de los servicios en Apéndice A: Guía de servicios particionales el plano de control afecten a sus cargas de trabajo, lo que le permite depender de los AWS servicios globales de forma segura y, al mismo tiempo, minimizar la introducción de puntos únicos de fallo.

Temas

Servicios zonales

Availability Zone Independence (AZI) AWS permite ofrecer servicios zonales, como Amazon EC2 y Amazon EBS. Un servicio zonal es aquel que permite especificar en qué zona de disponibilidad se implementan los recursos. Estos servicios funcionan de forma independiente en cada zona de disponibilidad de una región y, lo que es más importante, también fallan de forma independiente en cada zona de disponibilidad. Esto significa que los componentes de un servicio de una zona de disponibilidad no dependen de los componentes de otras zonas de disponibilidad. Podemos hacerlo porque un servicio zonal tiene planos de datos zonales. En algunos casos, como en el caso de EC2, el servicio también incluye planos de control zonales para operaciones alineadas por zonas, como el lanzamiento de una instancia de EC2. Para esos servicios, AWS también proporciona un punto final en el plano de control regional para facilitar la interacción con el servicio. El plano de control regional también proporciona una funcionalidad de ámbito regional y sirve como capa de agregación y enrutamiento sobre los planos de control zonales. Esto se muestra en la siguiente figura.



Un servicio zonal con planos de control y planos de datos aislados por zonas

Las zonas de disponibilidad ofrecen a los clientes la posibilidad de operar cargas de trabajo de producción con mayor disponibilidad, tolerancia a errores y escalables de lo que sería posible en un solo centro de datos. Cuando una carga de trabajo utiliza varias zonas de disponibilidad, los clientes están mejor aislados y protegidos de los problemas que afectan a la infraestructura física de una única zona de disponibilidad. Esto ayuda a los clientes a crear servicios redundantes en todas las zonas de disponibilidad y, si se diseñan correctamente, permanecen operativos incluso si una zona de disponibilidad sufre errores. Los clientes pueden aprovechar AZI para crear cargas de trabajo resilientes y de alta disponibilidad. La implementación de AZI en su arquitectura le ayuda a recuperarse rápidamente de un error aislado en una zona de disponibilidad, ya que los recursos de una zona de disponibilidad minimizan o eliminan la interacción con los recursos de otras zonas de disponibilidad. Esto ayuda a eliminar las dependencias entre zonas de disponibilidad, lo que simplifica la evacuación de las zonas de disponibilidad. Consulte los patrones de resiliencia avanzados para zonas de disponibilidad múltiples (Multi-AZ) para obtener más información sobre la creación de mecanismos de evacuación en las zonas de disponibilidad. Además, puede aprovechar aún más las zonas de disponibilidad si sigue algunas de las mismas prácticas recomendadas que se AWS utilizan para sus propios servicios, como implementar solo los cambios en una única zona de disponibilidad a la vez o eliminar una zona de disponibilidad del servicio si un cambio en esa zona de disponibilidad no funciona correctamente.

La estabilidad estática también es un concepto importante para las arquitecturas de zonas de disponibilidad múltiple. Uno de los modos de error que debe tener en cuenta con las arquitecturas de

zonas de disponibilidad múltiple es la pérdida de una zona de disponibilidad, lo que puede provocar la pérdida de la capacidad de una zona de disponibilidad. Si no ha aprovisionado previamente suficiente capacidad para hacer frente a la pérdida de una zona de disponibilidad, la carga actual podría sobrecargar la capacidad restante. Además, tendrá que depender de los planos de control de los servicios zonales que utilice para reemplazar la capacidad perdida, lo que puede resultar menos fiable que un diseño estable desde el punto de vista estático. En este caso, aprovisionar previamente suficiente capacidad adicional puede ayudarle a mantener una estabilidad estática ante la pérdida de un dominio de errores, como una zona de disponibilidad, al poder continuar con sus operaciones normales sin necesidad de cambios dinámicos.

Puede optar por utilizar un grupo de instancias EC2 de escalado automático implementado en varias zonas de disponibilidad para escalar dinámicamente los niveles de entrada y salida, en función de las necesidades de su carga de trabajo. El escalado automático funciona bien para los cambios graduales en el uso que se producen entre minutos y decenas de minutos. Sin embargo, el lanzamiento de nuevas instancias de EC2 lleva tiempo, especialmente si las instancias requieren un arranque (por ejemplo, instalar agentes, binarios de aplicaciones o archivos de configuración). Durante este tiempo, la capacidad restante podría verse abrumada por la carga actual. Además, la implementación de nuevas instancias mediante el escalado automático depende del plano de control EC2. Esto supone una compensación: para mantener la estabilidad estática ante la pérdida de una única zona de disponibilidad, debe aprovisionar previamente suficientes instancias de EC2 en las demás zonas de disponibilidad para gestionar la carga que se ha desplazado fuera de la zona de disponibilidad afectada, en lugar de confiar en el escalado automático para aprovisionar nuevas instancias. Sin embargo, el aprovisionamiento previo de capacidad adicional puede suponer un coste adicional.

Por ejemplo, durante el funcionamiento normal, supongamos que su carga de trabajo requiere seis instancias para atender el tráfico de clientes en tres zonas de disponibilidad. Para mantener la estabilidad estática frente a un error en una sola zona de disponibilidad, debe implementar tres instancias en cada zona de disponibilidad, es decir, nueve en total. Si fallara una sola instancia equivalente a una zona de disponibilidad, aún le quedarían seis y podría seguir atendiendo al tráfico de clientes sin necesidad de aprovisionar ni configurar nuevas instancias durante el fallo. Lograr una estabilidad estática para la capacidad de EC2 conlleva un coste adicional, ya que, en este caso, se está ejecutando un 50% más de instancias. No todos los servicios en los que puede aprovisionar recursos previamente conllevarán costes adicionales, como el aprovisionamiento previo de un bucket de S3 o un usuario. Deberá sopesar las desventajas de implementar la estabilidad estática con el riesgo de superar el tiempo de recuperación deseado para su carga de trabajo.

AWS Las Zonas Locales y los Outposts acercan el plano de datos de determinados AWS servicios a los usuarios finales. Los planos de control de estos servicios residen en la región principal. Tu instancia de Local Zone o Outposts tendrá dependencias en el plano de control para servicios zonales como EC2 y EBS en la zona de disponibilidad en la que creaste la zona local o la subred de Outposts. También dependerán de los planos de control regionales para los servicios regionales, como Elastic Load Balancing (ELB), los grupos de seguridad y el plano de control de Kubernetes administrado por Amazon Elastic Kubernetes Service (Amazon EKS) (si usa EKS). Para obtener información adicional específica sobre Outposts, consulta la documentación y las preguntas frecuentes sobre soporte y mantenimiento. Implemente la estabilidad estática cuando utilice Zonas Locales o Outposts para ayudar a mejorar la resiliencia y controlar las deficiencias del avión o las interrupciones en la conectividad de la red con la Región principal.

Servicios regionales

Los servicios regionales son servicios que AWS se basan en varias zonas de disponibilidad para que los clientes no tengan que averiguar cómo aprovechar al máximo los servicios zonales. Agrupamos de forma lógica el servicio implementado en varias zonas de disponibilidad para ofrecer a los clientes un único punto final regional. Amazon SQS y Amazon DynamoDB son ejemplos de servicios regionales. Utilizan la independencia y la redundancia de las zonas de disponibilidad para minimizar los fallos de infraestructura como una categoría de riesgo de disponibilidad y durabilidad. Amazon S3, por ejemplo, distribuye las solicitudes y los datos entre varias zonas de disponibilidad y está diseñado para recuperarse automáticamente del fallo de una zona de disponibilidad. Sin embargo, solo interactúa con el punto final regional del servicio.

AWS cree que la mayoría de los clientes pueden alcanzar sus objetivos de resiliencia en una sola región mediante el uso de servicios regionales o arquitecturas Multi-AZ que se basan en servicios zonales. Sin embargo, algunas cargas de trabajo pueden requerir redundancia adicional, y puede utilizar el aislamiento Regiones de AWS para crear arquitecturas multirregionales con fines de alta disponibilidad o continuidad empresarial. La separación física y lógica entre ellas Regiones de AWS evita los fallos correlacionados entre sí. En otras palabras, al igual que si fuera un cliente de EC2 y pudiera beneficiarse del aislamiento de las zonas de disponibilidad mediante la implementación en todas ellas, puede obtener la misma ventaja con los servicios regionales si se despliegan en varias regiones. Esto requiere que implemente una arquitectura multirregional para su aplicación, lo que puede ayudarle a resistir las deficiencias de un servicio regional.

Sin embargo, lograr los beneficios de una arquitectura multirregional puede resultar difícil; requiere un trabajo cuidadoso para aprovechar el aislamiento regional y, al mismo tiempo, no deshacer nada a nivel de aplicación. Por ejemplo, si va a realizar la conmutación por error de una aplicación entre

regiones, debe mantener una separación estricta entre las pilas de aplicaciones de cada región, tener en cuenta todas las dependencias de la aplicación y realizar la conmutación por error de todas las partes de la aplicación a la vez. Lograrlo con una arquitectura compleja basada en microservicios que tiene muchas dependencias entre las aplicaciones requiere planificación y coordinación entre muchos equipos empresariales y de ingeniería. Permitir que las cargas de trabajo individuales tomen sus propias decisiones de conmutación por error hace que la coordinación sea menos compleja, pero introduce un comportamiento modal gracias a la importante diferencia de latencia que se produce entre las regiones y dentro de una sola región.

AWS por el momento, no proporciona una función de replicación sincrónica entre regiones. Cuando se utiliza un almacén de datos replicado de forma asíncrona (proporcionado por AWS) entre regiones, existe la posibilidad de que se pierdan datos o se produzcan incoherencias si se conmuta por error la aplicación entre regiones. Para mitigar las posibles incoherencias, necesita un proceso de reconciliación de datos fiable en el que pueda confiar y que pueda tener que funcionar en varios almacenes de datos de su cartera de cargas de trabajo, o bien estar dispuesto a aceptar la pérdida de datos. Por último, debe practicar la conmutación por error para saber que funcionará cuando la necesite. La rotación periódica de la solicitud entre regiones para practicar la conmutación por error supone una inversión considerable de tiempo y recursos. Si decide utilizar un almacén de datos replicado de forma sincrónica en todas las regiones para dar soporte a sus aplicaciones que se ejecutan desde más de una región al mismo tiempo, las características de rendimiento y la latencia de una base de datos de este tipo, que abarca cientos o miles de millas, son muy diferentes a las de una base de datos que opera en una sola región. Esto requiere que planifique su pila de aplicaciones desde cero para tener en cuenta este comportamiento. También hace que la disponibilidad de ambas regiones sea una fuerte dependencia, lo que podría reducir la resiliencia de la carga de trabajo.

Servicios globales

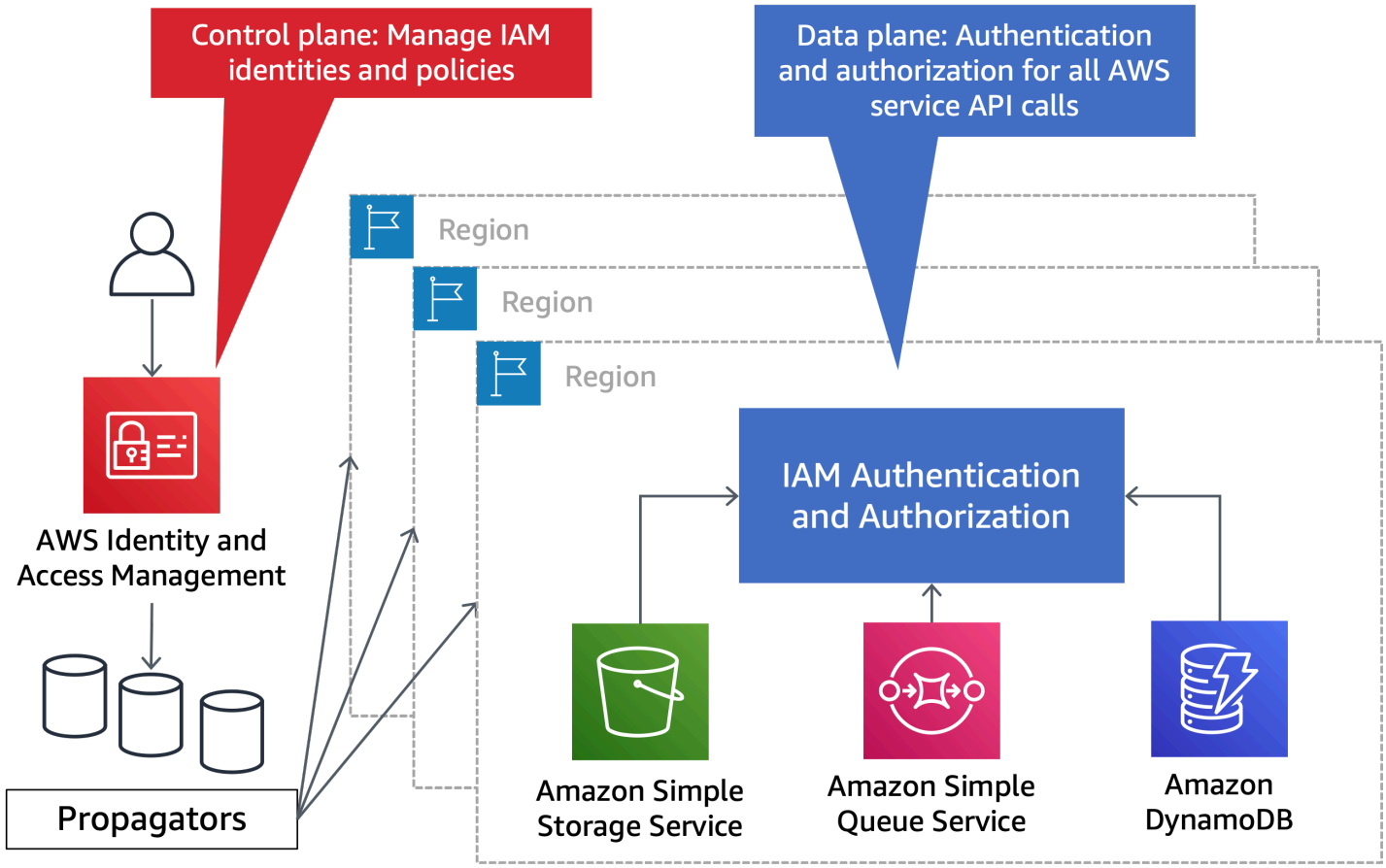
Además de los AWS servicios regionales y zonales, hay un pequeño conjunto de AWS servicios cuyos planos de control y de datos no existen de forma independiente en cada región. Como sus recursos no son específicos de una región, suelen denominarse globales. AWS Los servicios globales siguen el patrón de AWS diseño convencional de separar el plano de control y el plano de datos para lograr la estabilidad estática. La diferencia significativa para la mayoría de los servicios globales es que su plano de control está alojado en un único plano Región de AWS, mientras que su plano de datos está distribuido globalmente. Existen tres tipos diferentes de servicios globales y un conjunto de servicios que pueden parecer globales en función de la configuración seleccionada.

En las siguientes secciones se identificará cada tipo de servicio global y la forma en que se separan sus planos de control y de datos. Puede usar esta información como guía para crear mecanismos

confiables de alta disponibilidad (HA) y recuperación ante desastres (DR) sin necesidad de depender de un plano de control de servicios global. Este enfoque ayuda a eliminar los puntos únicos de falla en su arquitectura y evita posibles impactos entre regiones, incluso cuando opera en una región diferente de donde está alojado el plano de control de servicios global. También le ayuda a implementar de forma segura mecanismos de conmutación por error que no dependen de planos de control de servicios globales.

Servicios globales que son únicos por partición

Existen algunos AWS servicios globales en cada partición (denominados en este paper servicios particionales). Los servicios particionales proporcionan su plano de control en un solo plano. Región de AWS Algunos servicios particionales, como AWS Network Manager, son únicamente del plano de control y organizan el plano de datos de otros servicios. Otros servicios particionales, como IAM, tienen su propio plano de datos que está aislado y distribuido en toda la partición. Regiones de AWS Los errores en un servicio particional no afectan a las demás particiones. En la `aws` partición, el plano de control del servicio de IAM se encuentra en la `us-east-1` región, con planos de datos aislados en cada región de la partición. Los servicios particionales también tienen planos de control y de datos independientes en las particiones `aws-us-gov` y`aws-cn`. La separación del plano de control y el plano de datos para IAM se muestra en el siguiente diagrama.



IAM tiene un plano de control único y un plano de datos regionalizado

A continuación se muestran los servicios particionales y su ubicación en el plano de control en la partición: `aws`

- AWS IAM () `us-east-1`
- AWS Organizations (`us-east-1`)
- AWS Administración de cuentas () `us-east-1`
- Controlador de recuperación de aplicaciones (ARC) de Route 53 (`us-west-2`): este servicio solo está presente en la `aws` partición
- AWS Administrador de red (`us-west-2`)
- DNS privado de Route 53 (`us-east-1`)

Si alguno de estos aviones de control de servicio sufre un evento que afecte a la disponibilidad, es posible que no pueda utilizar las operaciones tipo CRUDL que ofrecen estos servicios. Por lo tanto, si su estrategia de recuperación depende de estas operaciones, si la disponibilidad se ve

afectada en el plano de control o en la región donde se encuentra el plano de control, se reducirán las probabilidades de que la recuperación se lleve a cabo satisfactoriamente. Apéndice A: Guía de servicios particionalesproporciona estrategias para eliminar las dependencias de los planes de control de servicios globales durante la recuperación.

> **Recomendación**
>
> No confíe en los planos de control de los servicios particionales en su ruta de recuperación. En su lugar, confíe en las operaciones del plano de datos de estos servicios. Consulte Apéndice A: Guía de servicios particionales para obtener detalles adicionales sobre cómo debe diseñar los servicios particionales.

Servicios globales en la red perimetral

El siguiente conjunto de AWS servicios globales tiene un plano de control en la `aws` partición y aloja sus planos de datos en la infraestructura de puntos de presencia (PoP) globales (y posiblemente Regiones de AWS también). Se PoPs puede acceder a los planos de datos alojados desde los recursos de cualquier partición, así como desde Internet. Por ejemplo, Route 53 opera su plano de control en la `us-east-1` región, pero su plano de datos está distribuido en cientos de sitios a PoPs nivel mundial, así como en cada uno de ellos Región de AWS (para admitir el DNS público y privado de Route 53 en la región). Las comprobaciones de estado de Route 53 también forman parte del plano de datos y se realizan desde ocho Regiones de AWS puntos de la `aws` partición. Los clientes pueden resolver el DNS mediante las zonas alojadas públicas de Route 53 desde cualquier lugar de Internet, incluidas otras particiones GovCloud, así como desde una Nube Privada AWS Virtual (VPC). A continuación se muestran los servicios de red perimetral globales y su ubicación en el plano de control en la `aws` partición:

- DNS público de Route 53 (`us-east-1`)
- Amazon CloudFront (`us-east-1`)
- AWS WAF Clásico para CloudFront (`us-east-1`)
- AWS WAF para CloudFront (`us-east-1`)
- Amazon Certificate Manager (ACM) para CloudFront (`us-east-1`)
- AWS Global Accelerator (AGA) () `us-west-2`
- AWS Shield Advanced (`us-east-1`)

Si utiliza las comprobaciones de estado de AGA para las instancias EC2 o las direcciones IP elásticas, estas utilizan las comprobaciones de estado de Route 53. La creación o actualización de las comprobaciones de estado de AGA dependería del plano de control de Route 53 en `us-east-1` el que se encuentre. La ejecución de los controles de estado de la AGA utiliza el plano de datos de los controles de estado de la Ruta 53.

Si se produce un fallo que afecte a la región donde se encuentran los aviones de control de estos servicios, o si se produce un fallo que afecte al propio plano de control, es posible que no pueda utilizar las operaciones tipo CRUDL que ofrecen estos servicios. Si ha utilizado estas operaciones en su estrategia de recuperación, es menos probable que esa estrategia tenga éxito que si se basa únicamente en el plano de datos de estos servicios.

> **Recomendación**
>
> No confíe en el plano de control de los servicios de red perimetrales de su ruta de recuperación. En su lugar, confíe en las operaciones del plano de datos de estos servicios. Consulte Apéndice B: Guía de servicio global de redes periféricas para obtener detalles adicionales sobre cómo diseñar servicios globales en la red perimetral.

Operaciones globales de una sola región

La última categoría se compone de operaciones específicas del plano de control dentro de un servicio que tienen un alcance de impacto global, y no de servicios completos como en las categorías anteriores. Al interactuar con los servicios zonales y regionales de la región que especifique, algunas operaciones tienen una dependencia subyacente de una sola región que es diferente de donde se encuentra el recurso. Son diferentes de los servicios que solo se proporcionan en una sola región; consulte Apéndice C: Servicios de una sola región para obtener una lista de esos servicios.

Durante un fallo que afecte a la dependencia global subyacente, es posible que no pueda utilizar las acciones de tipo CRUDL de las operaciones dependientes. Si ha utilizado estas operaciones en su estrategia de recuperación, es menos probable que esa estrategia tenga éxito que si se basa únicamente en el plano de datos de estos servicios. Debe evitar depender de estas operaciones en su estrategia de recuperación.

La siguiente es una lista de los servicios de los que pueden depender otros servicios y que tienen un alcance global:

- Ruta 53

Varios AWS servicios crean recursos que proporcionan un nombre DNS específico para cada recurso. Por ejemplo, cuando aprovisiona un Elastic Load Balancer (ELB), el servicio crea registros de DNS públicos y comprobaciones de estado en Route 53 para el ELB. Esto se basa en la entrada del plano de control de Route 53. `us-east-1` Es posible que otros servicios que utilice también necesiten aprovisionar un ELB, crear registros DNS públicos de Route 53 o crear comprobaciones de estado de Route 53 como parte de sus flujos de trabajo en el plano de control. Por ejemplo, el aprovisionamiento de un recurso de API REST de Amazon API Gateway, una base de datos del Amazon Relational Database Service (Amazon RDS) o un dominio de OpenSearch Amazon Service da como resultado la creación de registros DNS en Route 53. La siguiente es una lista de servicios cuyo plano de control depende del plano de control de Route 53 `us-east-1` para crear, actualizar o eliminar registros de DNS, zonas alojadas o crear comprobaciones de estado de Route 53. Esta lista no es exhaustiva; su objetivo es destacar algunos de los servicios más utilizados cuyas acciones del plano de control para crear, actualizar o eliminar recursos dependen del plano de control de Route 53:

- API REST y HTTP de Amazon API Gateway
- Instancias de Amazon RDS
- Bases de datos de Amazon Aurora
- Equilibradores de carga Amazon ELB
- AWS PrivateLink Puntos finales de VPC
- AWS Lambda URL
- Amazon ElastiCache
- OpenSearch Servicio Amazon
- Amazon CloudFront
- Amazon MemoryDB para Redis
- Amazon Neptune
- Amazon DynamoDB Accelerator (DAX) (Acelerador de Amazon DynamoDB (DAX)).
- MAGA
- Amazon Elastic Container Service (Amazon ECS) con detección de servicios basada en DNS (que utiliza AWS Cloud Map la API para gestionar el DNS de Route 53)
- Plano de control Amazon EKS Kubernetes

  Es importante tener en cuenta que el servicio DNS de VPC para los nombres de host de las instancias EC2 existe de forma independiente en cada una de

ellas Región de AWS y no depende del plano de control de Route 53. Los registros que se AWS crean para las instancias EC2 en el servicio DNS de la VPC, `ip-10-0-10.ec2.internal` como `ip-10-0-1-5.compute.us-west-2.compute.internali-0123456789abcdef.ec2.internal`, `i-0123456789abcdef.us-west-2.compute.internal` y, no dependen del plano de control de Route 53 en. `us-east-1`

> **Recomendación**
>
> No confíe en la creación, actualización o eliminación de recursos que requieran la creación, actualización o eliminación de los registros de recursos, las zonas alojadas o las comprobaciones de estado de Route 53 en su ruta de recuperación. Aprovisione previamente estos recursos, como los ELB, para evitar que su proceso de recuperación dependa del plano de control de Route 53.

- Amazon S3

Las siguientes operaciones del plano de control de Amazon S3 tienen una dependencia subyacente `us-east-1` en la `aws` partición. Un fallo que afecte a Amazon S3 o a otros servicios `us-east-1` podría provocar que las acciones de estos planos de control se vean afectadas en otras regiones:

```
PutBucketCors
DeleteBucketCors
PutAccelerateConfiguration
PutBucketRequestPayment
PutBucketObjectLockConfiguration
PutBucketTagging
DeleteBucketTagging
PutBucketReplication
DeleteBucketReplication
PutBucketEncryption
DeleteBucketEncryption
PutBucketLifecycle
DeleteBucketLifecycle
PutBucketNotification
PutBucketLogging
DeleteBucketLogging
```

```
PutBucketVersioning
PutBucketPolicy
DeleteBucketPolicy
PutBucketOwnershipControls
DeleteBucketOwnershipControls
PutBucketAcl
PutBucketPublicAccessBlock
DeleteBucketPublicAccessBlock
```

El plano de control de los puntos de acceso multirregionales (MRAP) de Amazon S3 está alojado únicamente en esa región `us-west-2` y las solicitudes para crear, actualizar o eliminar los MRAP se dirigen directamente a esa región. El plano de control del MRAP también tiene dependencias subyacentes de AGA in`us-west-2`, Route 53 in y ACM en cada región `us-east-1` desde la que el MRAP está configurado para ofrecer contenido. No debe depender de la disponibilidad del plano de control del MRAP en su ruta de recuperación ni en los planos de datos de sus propios sistemas. Esto es distinto de los controles de conmutación por error del MRAP, que se utilizan para especificar el estado de enrutamiento activo o pasivo de cada uno de los buckets del MRAP. Estas API están alojadas en cinco Regiones de AWS y se pueden utilizar para desplazar el tráfico de forma eficaz mediante el plano de datos del servicio.

Además, los nombres de los buckets de Amazon S3 son únicos a nivel mundial `CreateBucket` y todas las llamadas a `DeleteBucket` las API y dependen de ellas`us-east-1`, en la `aws` partición, para garantizar la exclusividad del nombre, aunque la llamada a la API esté dirigida a la región específica en la que desee crear el bucket. Por último, si tiene flujos de trabajo críticos para la creación de cubos, no debe depender de la disponibilidad de una ortografía específica para el nombre de un bucket, especialmente si sigue un patrón discernible.

> **Recomendación**
>
> No confíe en eliminar o crear nuevos buckets de S3 ni en actualizar las configuraciones de los buckets de S3 como parte de su proceso de recuperación. Aprovisione previamente todos los depósitos de S3 necesarios con las configuraciones necesarias para no tener que realizar cambios para recuperarse de un error. Este enfoque también se aplica a los MRAP.

- CloudFront

Amazon API Gateway proporciona puntos de enlace de API optimizados desde el punto de vista periférico. La creación de estos puntos de enlace depende del plano de CloudFront control utilizado `us-east-1` para crear la distribución frente al punto de enlace de la puerta de enlace.

> **Recomendación**
>
> No confíe en la creación de nuevos puntos finales de API Gateway optimizados desde el punto de vista periférico como parte de su ruta de recuperación. Aprovisione previamente todos los puntos finales de API Gateway necesarios.

Todas las dependencias analizadas en esta sección son acciones del plano de control, no acciones del plano de datos. Si sus cargas de trabajo están configuradas para ser estáticamente estables, estas dependencias no deberían afectar a su ruta de recuperación, teniendo en cuenta que la estabilidad estática requiere trabajo o servicios adicionales para implementarlas.

Servicios que utilizan puntos finales globales predeterminados

En algunos casos, los AWS servicios proporcionan un punto final global predeterminado, como el AWS Security Token Service (AWS STS). Otros servicios pueden usar este punto final global predeterminado en su configuración predeterminada. Esto significa que un servicio regional que esté utilizando podría tener una dependencia global de un solo servicio Región de AWS. En los siguientes detalles se explica cómo eliminar las dependencias no deseadas en los puntos finales globales predeterminados, lo que le ayudará a utilizar el servicio de forma regional.

AWS STS: STS es un servicio web que le permite solicitar credenciales temporales con privilegios limitados para los usuarios de IAM o para los usuarios que autentique (usuarios federados). El uso de STS desde el kit de desarrollo de AWS software (SDK) y la interfaz de línea de comandos (CLI) se establece de forma predeterminada en. `us-east-1` El servicio STS también proporciona puntos finales regionales. Estos puntos finales están habilitados de forma predeterminada en las regiones que también están habilitadas de forma predeterminada. Puede aprovecharlas en cualquier momento configurando su SDK o CLI siguiendo estas instrucciones: Puntos de conexión regionalizados de AWS STS. El uso de SiGV4a también requiere credenciales temporales solicitadas a un punto final de STS regional. No puede utilizar el punto final STS global para esta operación.

> **Recomendación**
>
> Actualice la configuración de su SDK y CLI para usar los puntos finales STS regionales.

Inicio de sesión en el lenguaje de marcado de aserciones de seguridad (SAML): todos los servicios de SAML están disponibles. Regiones de AWSPara usar este servicio, elige el punto de conexión SAML regional adecuado, como https://us-west-2.signin.aws.amazon.com/saml. Debe actualizar las configuraciones de sus políticas de confianza y del proveedor de identidad (IdP) para usar los puntos finales regionales. Consulte la documentación de AWS SAML para obtener detalles específicos.

Si utiliza un IdP que también está alojado en AWS, existe el riesgo de que también se vea afectado durante un evento de AWS error. Esto podría provocar que no pueda actualizar la configuración de su IdP o que no pueda federarse por completo. Debe aprovisionar previamente a los usuarios «break-glass» en caso de que su IdP esté averiado o no esté disponible. Consulte Apéndice A: Guía de servicios particionales para obtener más información sobre cómo crear usuarios de break-glass de una forma estable desde el punto de vista estático.

> **Recomendación**
>
> Actualice sus políticas de confianza en los roles de IAM para aceptar inicios de sesión con SAML desde varias regiones. En caso de que se produzca un error, actualice la configuración de su IdP para usar un punto final SAML regional diferente si su punto final preferido está dañado. Cree un usuario rompe-cristal en caso de que su IdP esté averiado o no esté disponible.

AWS IAM Identity Center: Identity Center es un servicio basado en la nube que facilita la gestión centralizada del acceso mediante inicio de sesión único a las aplicaciones de un cliente y en la nube. Cuentas de AWS Identity Center debe implementarse en la única región que elija. Sin embargo, el comportamiento predeterminado del servicio es utilizar el punto de conexión SAML global (https://signin.aws.amazon.com/saml), que está alojado en`us-east-1`. Si ha implementado Identity Center en otro Región de AWS, debe actualizar la URL del estado de retransmisión de cada conjunto de permisos para que se dirija al mismo punto final de la consola regional que utilizó en su implementación de Identity Center. Por ejemplo, si implementó Identity Center en`us-west-2`, debe actualizar el estado de retransmisión de sus conjuntos de permisos para usar https://us-west-2.console.aws.amazon.com. Esto eliminará cualquier dependencia de la implementación `us-east-1` de Identity Center.

Además, dado que IAM Identity Center solo se puede implementar en una sola región, debe aprovisionar previamente a los usuarios más avanzados en caso de que su implementación se vea afectada. Consulte Apéndice A: Guía de servicios particionales para obtener más información sobre cómo crear usuarios de breakglass de forma estática y estable.

> **Recomendación**
>
> Defina la URL del estado de retransmisión de sus conjuntos de permisos en el Centro de Identidad de IAM para que coincida con la región en la que está desplegado el servicio. Cree un usuario innovador en caso de que la implementación de su centro de identidad de IAM no esté disponible.

Amazon S3 Storage Lens: Storage Lens proporciona un panel predeterminado llamado default-account-dashboard. La configuración del panel y sus métricas asociadas se almacenan en `us-east-1`. Puede crear paneles adicionales en otras regiones especificando la región de origen para la configuración del panel y los datos de las métricas.

> **Recomendación**
>
> Si necesita datos del panel de control predeterminado de S3 Storage Lens durante un fallo que afecte al servicio `us-east-1`, cree un panel adicional en una región de origen alternativa. También puede duplicar cualquier otro panel personalizado que haya creado en otras regiones.

Resumen de los servicios globales

Los planos de datos para los servicios globales aplican principios de aislamiento e independencia similares a los de los AWS servicios regionales. Un error que afecte al plano de datos de IAM en una región no afectará al funcionamiento del plano de datos de IAM en otra. Región de AWS Del mismo modo, una falla que afecte al plano de datos de la Ruta 53 en un PoP no afectará al funcionamiento del plano de datos de la Ruta 53 en el resto de la región. PoPs Por lo tanto, lo que debemos tener en cuenta son los eventos de disponibilidad del servicio que afecten a la región en la que opera el plano de control o que afecten al propio plano de control. Como solo hay un plano de control para cada servicio global, una falla que afecte a ese plano de control podría tener efectos transregionales en las operaciones de tipo CRUDL (que son las operaciones de configuración que se utilizan normalmente para establecer o configurar un servicio, en lugar del uso directo del servicio).

La forma más eficaz de diseñar las cargas de trabajo para utilizar los servicios globales de forma resiliente es utilizar la estabilidad estática. En caso de fallo, diseñe su carga de trabajo de forma que no necesite realizar cambios con un plano de control para mitigar el impacto o la conmutación por error a una ubicación diferente. Consulte Apéndice A: Guía de servicios particionales y obtenga una guía prescriptiva sobre cómo utilizar este tipo de servicios globales Apéndice B: Guía de servicio global de redes periféricas para eliminar las dependencias del plano de control y eliminar los puntos únicos de fallo. Si necesita los datos de una operación del plano de control para la recuperación, almacene en caché estos datos en un almacén de datos al que se pueda acceder a través de su plano de datos, como un parámetro del Almacén de parámetros de AWS Systems Manager (SSM Parameter Store), una tabla de DynamoDB o un bucket de S3. Por motivos de redundancia, también puede optar por almacenar esos datos en una región adicional. Por ejemplo, siguiendo las prácticas recomendadas para el controlador de recuperación de aplicaciones (ARC) de Route 53, debe codificar o marcar los cinco puntos finales del clúster regional como favoritos. Durante un caso de error, es posible que no pueda acceder a algunas operaciones de la API, incluidas las operaciones de la API de Route 53 ARC que no están alojadas en el clúster del plano de datos extremadamente confiable. Puede enumerar los puntos finales de sus clústeres ARC de Route 53 mediante la operación de API. `DescribeCluster`

El siguiente es un resumen de algunos de los errores de configuración o antipatrones más comunes que introducen dependencias en los planos de control de los servicios globales:

- Realizar cambios en los registros de Route 53, como actualizar el valor de un registro A o cambiar las ponderaciones de un conjunto de registros ponderado, para realizar una conmutación por error.
- Crear o actualizar los recursos de IAM, incluidas las funciones y políticas de IAM, durante una conmutación por error. Por lo general, esto no es intencional, pero puede ser el resultado de un plan de conmutación por error no probado.
- Confiar en el IAM Identity Center para que los operadores accedan a los entornos de producción en caso de fallo.
- Confíe en la configuración predeterminada del IAM Identity Center para utilizar la consola `us-east-1` cuando haya implementado Identity Center en una región diferente.
- Al realizar cambios en el tráfico de AGA, hay que tener en cuenta las ponderaciones para realizar manualmente una conmutación por error regional.
- Actualizar la configuración de origen de una CloudFront distribución para evitar errores en un origen defectuoso.
- Aprovisionamiento de recursos de recuperación ante desastres (DR), como instancias de ELB y RDS durante una falla, que dependen de la creación de registros de DNS en Route 53.

A continuación, se presenta un resumen de las recomendaciones que se proporcionan en esta sección para utilizar los servicios globales de una forma flexible que ayude a prevenir los antipatrones habituales que solían producirse anteriormente.

Resumen de las recomendaciones

No confíe en los planos de control de los servicios particionales en su proceso de recuperación. En su lugar, confíe en las operaciones del plano de datos de estos servicios. Consulte Apéndice A: Guía de servicios particionales para obtener detalles adicionales sobre cómo debe diseñar los servicios particionales.
No confíe en el plano de control de los servicios de red perimetrales en su ruta de recuperación. En su lugar, confíe en las operaciones del plano de datos de estos servicios. Consulte Apéndice B: Guía de servicio global de redes periféricas para obtener detalles adicionales sobre cómo diseñar servicios globales en la red perimetral.
No confíe en la creación, actualización o eliminación de recursos que requieran la creación, actualización o eliminación de registros de recursos, zonas alojadas o comprobaciones de estado de Route 53 en su ruta de recuperación. Aprovisione previamente estos recursos, como los ELB, para evitar que su proceso de recuperación dependa del plano de control de Route 53.
No confíe en eliminar o crear nuevos depósitos de S3 ni en actualizar las configuraciones de los depósitos de S3 como parte de su ruta de recuperación. Aprovisione previamente todos los depósitos de S3 necesarios con las configuraciones necesarias para no tener que realizar cambios para recuperarse de un error. Este enfoque también se aplica a los MRAP.
No confíe en la creación de nuevos puntos finales de API Gateway optimizados desde el punto de vista periférico como parte de su ruta de recuperación. Aprovisione previamente todos los puntos finales de API Gateway necesarios.
Actualice la configuración de su SDK y CLI para usar los puntos finales STS regionales.
Actualice sus políticas de confianza en los roles de IAM para aceptar inicios de sesión con SAML desde varias regiones. En caso de que se produzca un error, actualice la configuración de su IdP para usar un punto final SAML regional diferente si su punto final preferido está dañado. Cree usuarios rompe-cristal en caso de que su IdP esté dañado o no esté disponible.
Defina la URL del estado de retransmisión de sus conjuntos de permisos en el Centro de Identidad de IAM para que coincida con la región en la que está desplegado el servicio. Cree un usuario único en caso de que la implementación de su Centro de Identidad no esté disponible.

Si necesita datos del panel de control predeterminado de S3 Storage Lens durante una falla que afecte al servicio`us-east-1`, cree un panel adicional en una región de origen alternativa. También puede duplicar cualquier otro panel personalizado que haya creado en otras regiones.

Conclusión

AWSproporciona varios constructos diferentes para los límites de aislamiento de fallas. Debe considerar la forma en que diseña los servicios zonales, regionales y globales, así como los posibles impactos en su carga de trabajo y en la capacidad de su carga de trabajo de recuperarse durante las deficiencias del plano de control. La estabilidad estática es una de las principales formas de evitar las dependencias del plano de control y crear mecanismos de HA y DR confiables y resilientes al utilizar los AWS servicios.

Apéndice A: Guía de servicios particionales

En el caso de los servicios particionales, debe implementar la estabilidad estática para mantener la resiliencia de la carga de trabajo durante un deterioro del plano de control de AWS servicios. A continuación se proporciona una guía prescriptiva sobre cómo considerar las dependencias de los servicios particionales, así como lo que funcionará y lo que no funcionará durante una discapacidad del plano de control.

AWS Identity and Access Management (IAM)

El plano de control AWS Identity and Access Management (IAM) consta de todas las API de IAM públicas (incluidas Access Advisor, pero no Access Analyzer ni IAM Roles Anywhere). Esto incluye acciones como `CreateRole`, `AttachRolePolicy`, `ChangePassword`, `UpdateSAMLProvider`, y`UpdateLoginProfile`. El plano de datos de IAM proporciona autenticación y autorización para los principales de IAM en cada uno de ellos. Región de AWS Durante una alteración del plano de control, es posible que las operaciones de tipo CRUDL para el IAM no funcionen, pero la autenticación y la autorización de los directores existentes seguirán funcionando. El STS es un servicio exclusivo del plano de datos que es independiente del IAM y no depende del plano de control del IAM.

Esto significa que, al planificar las dependencias de IAM, no debe confiar en el plano de control de IAM en su ruta de recuperación. Por ejemplo, un diseño estable desde el punto de vista estático para un usuario administrador «rompecristales» sería crear un usuario con los permisos adecuados adjuntos, establecer la contraseña y aprovisionar la clave de acceso y la clave de acceso secreta y, a continuación, bloquear esas credenciales en una bóveda física o virtual. Cuando sea necesario durante una emergencia, recupere las credenciales de usuario del almacén y utilícelas según sea necesario. Un non-statically-stable diseño consistiría en aprovisionar al usuario en caso de error o en hacer que el usuario se aprovisione previamente, pero solo adjuntar la política de administración cuando sea necesario. Estos enfoques dependerían del plano de control del IAM.

AWS Organizations

El plano de AWS Organizations control está formado por todas las API de las Organizations públicas`AcceptHandshake`, como `AttachPolicy``CreateAccount`,`CreatePolicy`, y`ListAccounts`. No hay un plano de datos paraAWS Organizations. Organiza el plano de datos para otros servicios, como IAM. Durante una alteración del plano de control, es posible que las

operaciones del tipo CRUDL para las Organizations no funcionen, pero las políticas, como las políticas de control de servicios (SCP) y las políticas de etiquetas, seguirán funcionando y se evaluarán como parte del proceso de autorización del IAM. Las capacidades de administración delegada y las funciones de cuentas múltiples en otros AWS servicios compatibles con las Organizations también seguirán funcionando.

Lo que esto significa es que, cuando planifique dependencias enAWS Organizations, no debe confiar en el plano de control de la Organizations en su ruta de recuperación. En su lugar, implemente la estabilidad estática en su plan de recuperación. Por ejemplo, un non-statically-stable enfoque podría consistir en actualizar los SCP para eliminar las restricciones permitidas Regiones de AWS a través de la `aws:RequestedRegion` condición o habilitar los permisos de administrador para funciones de IAM específicas. Esto depende del plano de control de la Organizations para realizar estas actualizaciones. Un mejor enfoque sería utilizar etiquetas de sesión para conceder el uso de permisos de administrador. Su proveedor de identidad (IdP) puede incluir etiquetas de sesión que se pueden evaluar en función de la `aws:PrincipalTag` condición, lo que le ayuda a configurar de forma dinámica los permisos para ciertos principios y, al mismo tiempo, a que sus SCP permanezcan estáticos. Esto elimina las dependencias de los planos de control y solo utiliza las acciones del plano de datos.

AWS Account Management

El plano de control de gestión de AWS cuentas está alojado en us-east-1 y consta de todas las API públicas para administrar unCuenta de AWS, como y. `GetContactInformation` `PutContactInformation` También incluye crear o cerrar uno nuevo a Cuenta de AWS través de la consola de administración. Las API de`CloseAccount`, `CreateAccount``CreateGovCloudAccount`, y `DescribeAccount` forman parte del plano de AWS Organizations control, que también está alojado en us-east-1. Además, la creación de una GovCloud cuenta fuera de AWS Organizations depende del plano de control Cuenta de AWS de administración de us-east-1. Además, GovCloud las cuentas deben estar vinculadas Cuenta de AWS en forma 1:1 a una `aws` partición. La creación de las cuentas en la `aws-cn` partición no depende de us-east-1. El plano de datos Cuentas de AWS son las propias cuentas. Durante una avería en el plano de control, es Cuentas de AWS posible que las operaciones del tipo CRUDL (como crear una cuenta nueva o obtener y actualizar la información de contacto) no funcionen. Las referencias a la cuenta en las políticas de IAM seguirán funcionando.

Esto significa que, cuando planifique dependencias en la administración de AWS cuentas, no debe confiar en el plano de control de la administración de cuentas en su ruta de recuperación. Si

bien el plano de control de administración de cuentas no proporciona la funcionalidad directa que normalmente utilizaría en una situación de recuperación, puede haber ocasiones en las que sí. Por ejemplo, un diseño estable desde el punto de vista estático consistiría en aprovisionar previamente todo lo que Cuentas de AWS necesita para la conmutación por error. Un non-statically-stable diseño consistiría en crear nuevos Cuentas de AWS durante un evento de falla para alojar sus recursos de DR.

Controlador de recuperación de aplicaciones Route 53

El plano de control de Route 53 ARC consta de las API para el control de la recuperación y la preparación para la recuperación, tal como se identifican en: puntos finales y cuotas del controlador de recuperación de aplicaciones de Amazon Route 53. Las comprobaciones de disponibilidad, los controles de enrutamiento y las operaciones del clúster se gestionan mediante el plano de control. El plano de datos de ARC es su clúster de recuperación, que administra los valores de control de enrutamiento consultados por las comprobaciones de estado de Route 53 y también implementa las reglas de seguridad. Se accede a la funcionalidad del plano de datos de Route 53 ARC a través de las API del clúster de recuperación, por ejemplo. `https://aaaaaaaa.route53-recovery-cluster.eu-west-1.amazonaws.com`

Esto significa que no debe confiar en el plano de control ARC de la Ruta 53 en su ruta de recuperación. Hay dos prácticas recomendadas que ayudan a implementar esta guía:

- Primero, marque como favoritos o codifique de forma rígida los cinco extremos del clúster regional. Esto elimina la necesidad de utilizar la operación del plano DescribeCluster de control durante un escenario de conmutación por error para descubrir los valores de los puntos finales.
- En segundo lugar, utilice las API del clúster ARC de Route 53 mediante la CLI o el SDK para realizar actualizaciones en los controles de enrutamiento y no en los. AWS Management Console Esto elimina la consola de administración como una dependencia del plan de conmutación por error y garantiza que dependa únicamente de las acciones del plano de datos.

AWS Network Manager

El servicio AWS Network Manager es principalmente un sistema exclusivo del plano de control alojado en us-west-2. Su propósito es administrar de forma centralizada la configuración de su Red central de red de área de Nube de AWS área de la red central de la Red central de la Red central de la red AWS Transit Gateway la red central de Cuentas de AWS la red de área de la red de área en las instalaciones. También agrega las métricas de Cloud WAN en us-west-2, a las que también

se puede acceder a través del CloudWatch plano de datos. Si Network Manager está dañado, el plano de datos de los servicios que organiza no se verá afectado. Las CloudWatch métricas de Cloud WAN también están disponibles en us-west-2. Si quieres datos métricos históricos, como los bytes de entrada y salida por región, para entender cuánto tráfico podría transferirse a otras regiones durante un error que afecte a us-west-2, o para otros fines operativos, puedes exportar esas métricas como datos CSV directamente desde la CloudWatch consola o utilizar este método: publica CloudWatch las métricas de Amazon en un archivo CSV. Los datos se encuentran en el espacio de `AWS/Network Manager` nombres y puede hacerlo según el horario que elija y almacenarlos en S3 o en otro banco de datos que seleccione. Para implementar un plan de recuperación estable desde el punto de vista estático, no utilice AWS Network Manager para realizar actualizaciones en la red ni confíe en los datos de las operaciones del plano de control para introducir la conmutación por error.

DNS privado de Route 53

Las zonas alojadas privadas de Route 53 se admiten en cada partición; sin embargo, las consideraciones para las zonas alojadas privadas y las zonas alojadas públicas en Route 53 son las mismas. Consulte Amazon Route 53 en el apéndice B: Guía de servicios globales de redes periféricas.

Apéndice B: Guía de servicio global de redes periféricas

En el caso de los servicios globales de redes periféricas, debe implementar la estabilidad estática para mantener la resiliencia de su carga de trabajo durante un deterioro del plano de control de AWS servicios.

Route 53

El plano de control de Route 53 consta de todas las API públicas de Route 53 que cubren la funcionalidad de las zonas alojadas, los registros, las comprobaciones de estado, los registros de consultas de DNS, los conjuntos de delegación reutilizables, las políticas de tráfico y las etiquetas de asignación de costos. Se encuentra alojado en la us-east-1. El plano de datos es el servicio DNS autorizado, que se ejecuta en más de 200 ubicaciones PoP y responde a las consultas de DNS Región de AWS en función de las zonas alojadas y los datos de comprobación de estado. Además, Route 53 tiene un plano de datos para las comprobaciones de estado, que también es un servicio distribuido a nivel mundial en varios. Regiones de AWS Este plano de datos realiza comprobaciones de estado, agrega los resultados y los datos de DNS público y privado de Route 53 y. Durante una alteración del plano de control, es posible que las operaciones de tipo CRUDL para la Ruta 53 no funcionen, pero la resolución del DNS y las comprobaciones de estado, y las actualizaciones del enrutamiento que resulten de los cambios en las comprobaciones de estado, seguirán funcionando.

Esto significa que cuando planifique dependencias en la Ruta 53, no debe confiar en el plano de control de la Ruta 53 en su ruta de recuperación. Por ejemplo, un diseño estable desde el punto de vista estático consistiría en utilizar el estado de las comprobaciones de estado para realizar conmutaciones por error entre regiones o para evacuar una zona de disponibilidad. Puede utilizar los controles de enrutamiento del controlador de recuperación de aplicaciones (ARC) de Route 53 para cambiar manualmente el estado de las comprobaciones de estado y modificar las respuestas a las consultas de DNS. Existen patrones similares a los que proporciona ARC que puede implementar en función de sus requisitos. Algunos de estos patrones se describen en Creación de mecanismos de recuperación ante desastres mediante la Ruta 53 y en la sección de disyuntores de verificación del estado avanzados de patrones de resiliencia Multi-AZ. Si ha optado por utilizar un plan de DR multirregional, aprovisione previamente los recursos que requieran la creación de registros DNS, como las instancias de ELB y RDS. Un non-statically-stable diseño consistiría en actualizar el valor de un registro de recursos de Route 53 mediante la `ChangeResourceRecordSets` API, cambiar el peso de un registro ponderado o crear nuevos registros para realizar la conmutación por error. Estos enfoques dependen del plano de control de la Ruta 53.

Amazon CloudFront

El plano de CloudFront control de Amazon consta de todas las CloudFront API públicas para administrar las distribuciones y está alojado en us-east-1. El plano de datos es la propia distribución servida desde PoPs la red perimetral. Realiza la gestión de solicitudes, el enrutamiento y el almacenamiento en caché del contenido de origen. Durante una alteración del plano de control, es posible que las operaciones de tipo CRUDL para CloudFront (incluidas las solicitudes de invalidación) no funcionen, pero el contenido seguirá almacenándose en caché y publicándose, y las conmutaciones por error de origen seguirán funcionando.

Esto significa que, cuando planifique dependencias enCloudFront, no debe confiar en el plano de CloudFront control en su ruta de recuperación. Por ejemplo, un diseño estable desde el punto de vista estático consistiría en utilizar conmutaciones por error de origen automatizadas para mitigar el impacto de una discapacidad en uno de sus orígenes. También puede optar por crear un equilibrio de carga de origen o una conmutación por error con Lamda @Edge. Consulte Tres patrones de diseño avanzados para aplicaciones de alta disponibilidad con Amazon CloudFront y Amazon S3 para crear aplicaciones de geolocalización activa CloudFront y activa multirregiones para obtener más información sobre ese patrón. Un non-statically-stable diseño consistiría en actualizar manualmente la configuración de la distribución en respuesta a un error de origen. Este enfoque dependería del plano CloudFront de control.

Certificate Manager

Si utiliza certificados personalizados en su CloudFront distribución, también depende de ACM. El uso de certificados personalizados en la CloudFront distribución se basa en el plano de control ACM en la región us-east-1. Durante una alteración del plano de control, los certificados existentes configurados en su distribución seguirán funcionando, al igual que las renovaciones automáticas de certificados. No confíe en cambiar la configuración de la distribución ni en crear nuevos certificados como parte de su ruta de recuperación.

AWSFirewall de aplicaciones web (WAF) y WAF Classic

Si lo utiliza AWS WAF con su CloudFront distribución, depende del plano de control de WAF, que también está alojado en la región us-east-1. Durante una alteración del plano de control, las listas de control de acceso web (ACL) configuradas y sus reglas asociadas siguen funcionando. No confíe en la actualización de las ACL web de WAF como parte de su ruta de recuperación.

AWS Global Accelerator

El plano de control AGA consta de todas las API AGA públicas y está alojado en us-west-2. El plano de datos es el enrutamiento de red de las direcciones IP anycast proporcionadas por AGA a los puntos finales registrados. AGA también utiliza las comprobaciones de estado de Route 53 para determinar el estado de sus puntos finales de AGA, que forman parte del plano de datos de Route 53. Durante una alteración del plano de control, es posible que las operaciones tipo CRUDL para AGA no funcionen. El enrutamiento a los puntos finales existentes, así como las configuraciones existentes de comprobaciones de estado, marcación de tráfico y peso de puntos finales que se utilizan para enrutar o desviar el tráfico a otros puntos finales y grupos de puntos finales, seguirán funcionando.

Esto significa que, al planificar las dependencias de AGA, no debe confiar en el plano de control de AGA en su ruta de recuperación. Por ejemplo, un diseño estable desde el punto de vista estático consistiría en utilizar el estado de las comprobaciones de estado configuradas para evitar errores en los puntos finales que no estén en buen estado. Consulte Implementación de aplicaciones multirregionales AWS con AWS Global Accelerator para ver ejemplos de esta configuración. Un non-statically-stable diseño consistiría en modificar los porcentajes de marcado de tráfico de la AGA, editar los grupos de puntos finales o eliminar un punto final de un grupo de puntos finales durante una discapacidad. Estos enfoques dependerían del plano de control de la AGA.

Amazon Shield

El plano de control de Amazon Shield Advanced consta de todas las API públicas de Shield Advanced y está alojado en us-east-1. Esto incluye funciones como `CreateProtectionCreateProtectionGroup`, `AssociateHealthCheckDesribeDRTAccess`, y`ListProtections`. El plano de datos es la protección contra ataques DDoS que proporciona Shield Advanced, así como la creación de las métricas de Shield Advanced. Shield Advanced también utiliza las comprobaciones de estado de Route 53 (que forman parte del plano de datos de Route 53), si las ha configurado. Durante una alteración del plano de control, es posible que las operaciones de tipo CRUDL para Shield Advanced no funcionen, pero la protección contra DDoS configurada para sus recursos, así como las respuestas a los cambios en las comprobaciones de estado, seguirán funcionando.

Esto significa que no debe confiar en el plano de control de Shield Advanced en su ruta de recuperación. Si bien el plano de control de Shield Advanced no proporciona la funcionalidad directa que normalmente utilizaría en una situación de recuperación, puede haber ocasiones en las que sí lo

haga. Por ejemplo, un diseño estable desde el punto de vista estático sería tener los recursos de DR ya configurados para formar parte de un grupo de protección y tener asociadas las comprobaciones de estado, en lugar de configurar esa protección después de que se produzca el error. Esto evita depender del plano de control de Shield Advanced para la recuperación.

Apéndice C: Servicios de una sola región

La siguiente es una lista de servicios o funciones específicas de ese servicio (que aparecen entre paréntesis después del nombre del servicio), que solo están disponibles en una región. La misma guía para implementar la estabilidad estática que se proporciona para otros servicios globales se aplica a estos servicios cuando necesita planificar las dependencias en sus planos de control y de datos.

- Alexa for Business
- AWS Marketplace(API de AWS Marketplace catálogo, análisis de AWS Marketplace comercio, AWS Marketplace servicio de derechos)
- Billing and Cost Management (AWS Cost Explorerinformes de AWS costos y uso, AWS presupuestos, Savings Plans)
- AWS BugBust
- Amazon Mechanical Turk
- Amazon Chime
- SDK de Amazon Chime (audio, mensajería e identidad de PSTN)
- AWSChatbot
- AWS DeepRacer
- AWSDevice Farm
- Amazon GameSparks
- Amazon Honeycode

Colaboradores

Los contribuyentes a este documento incluyen:

- Michael Haken, arquitecto principal de soluciones, Amazon Web Services

Revisiones del documento

Si desea recibir notificaciones sobre las actualizaciones de este documento, suscríbase a la fuente RSS.

Cambio	Descripción	Fecha
Revisión menor	Guía actualizada para implementar las prácticas recomendadas de IAM. Para obtener más información, consulte Prácticas recomendadas de seguridad en IAM.	9 de febrero de 2023
Publicación inicial	Se publicó un documento técnico.	16 de noviembre de 2022

Glosario de AWS

Para ver la terminología más reciente de AWS, consulte el Glosario de AWS en la Referencia de Glosario de AWS.

Avisos

Los clientes son responsables de realizar su propia evaluación independiente de la información de este documento. Este documento: (a) tiene únicamente fines informativos, (b) representa las ofertas y prácticas de AWS productos actuales, que están sujetas a cambios sin previo aviso, y (c) no crea ningún compromiso ni garantía por parte de AWS sus filiales, proveedores o licenciantes. AWSlos productos o servicios se proporcionan «tal cual» sin garantías, representaciones o condiciones de ningún tipo, ya sean expresas o implícitas. Las responsabilidades y responsabilidades de AWS sus clientes están reguladas por AWS acuerdos, y este documento no forma parte de ningún acuerdo entre AWS sus clientes ni lo modifica.

Las traducciones son generadas a través de traducción automática. En caso de conflicto entre la traducción y la version original de inglés, prevalecerá la version en inglés.