AWS Livro branco

Limites de isolamento de falhas da AWS

Limites de isolamento de falhas da AWS: AWS Livro branco

Table of Contents

AWS Fault

Data de publicação: 16 de novembro de 2022 (Revisões do documento)

Resumo

O Amazon Web Services (AWS) fornece diferentes limites de isolamento, como zonas de disponibilidade (AZ), regiões, planos de controle e planos de dados. Este paper detalha como AWS usa esses limites para criar serviços zonais, regionais e globais. Ele também inclui orientações prescritivas sobre como considerar as dependências desses diferentes serviços e como melhorar a resiliência das cargas de trabalho que você cria usando eles.

Você é Well-Architected para Dentes

O AWSWell-Architected Framework ajuda você a entender os prós e os contras das decisões que você toma ao criar sistemas na nuvem. Os seis pilares da Estrutura permitem que você aprenda as melhores práticas arquitetônicas para projetar e operar sistemas confiáveis, seguros, eficientes, econômicos e sustentáveis. Usando o AWS Well-Architected Tool, disponível gratuitamente no AWS Management Console, você pode analisar suas cargas de trabalho com base nessas melhores práticas respondendo a um conjunto de perguntas para cada pilar.

Para obter mais orientações de especialistas e melhores práticas para sua arquitetura de nuvem — implantações de arquitetura de referência, diagramas e whitepapers — consulte o Centro de Arquitetura. AWS

Introdução

AWSopera uma infraestrutura global para fornecer serviços em nuvem que ajudam os clientes a implantar cargas de trabalho de forma flexível, segura, escalável e altamente disponível. A AWS infraestrutura usa várias construções de isolamento de falhas para ajudar os clientes a atingirem seus objetivos de resiliência. Esses limites de isolamento de falhas permitem que os clientes projetem suas cargas de trabalho para aproveitar o escopo previsível de contenção de impactos que elas fornecem. Também é importante entender como os AWS serviços são projetados usando esses limites para que você possa fazer escolhas intencionais sobre as dependências selecionadas para sua carga de trabalho.

Este paper resumirá primeiro a infraestrutura AWS global e os limites de isolamento de falhas que ela fornece, bem como alguns dos padrões usados para projetar nossos serviços. Usando essa linha de base de entendimento, o paper descreverá a seguir os diferentes escopos de serviços AWS fornecidos: zonal, regional e global. Ele também apresentará as melhores práticas para criar arquiteturas que usam esses limites de isolamento e diferentes escopos de serviço para melhorar a resiliência das cargas de trabalho nas quais você executa. AWS Em particular, ele fornece orientação prescritiva sobre como assumir dependências de serviços globais e, ao mesmo tempo, minimizar pontos únicos de falha. Isso ajudará você a fazer escolhas informadas sobre suas AWS dependências e como você projeta sua carga de trabalho para alta disponibilidade (HA) e recuperação de desastres (DR).

AWSinfraestrutura

Esta seção apresenta um resumo da infraestrutura AWS global e dos limites de isolamento de falhas que ela fornece. Além disso, esta seção fornecerá uma visão geral do conceito de planos de controle e planos de dados, que são distinções críticas na forma como AWS projeta seus serviços. Essas informações fornecem a linha de base para entender como os limites de isolamento de falhas e o plano de controle e o plano de dados de um serviço se aplicam aos tipos de AWS serviço que discutiremos na próxima seção.

Tópicos

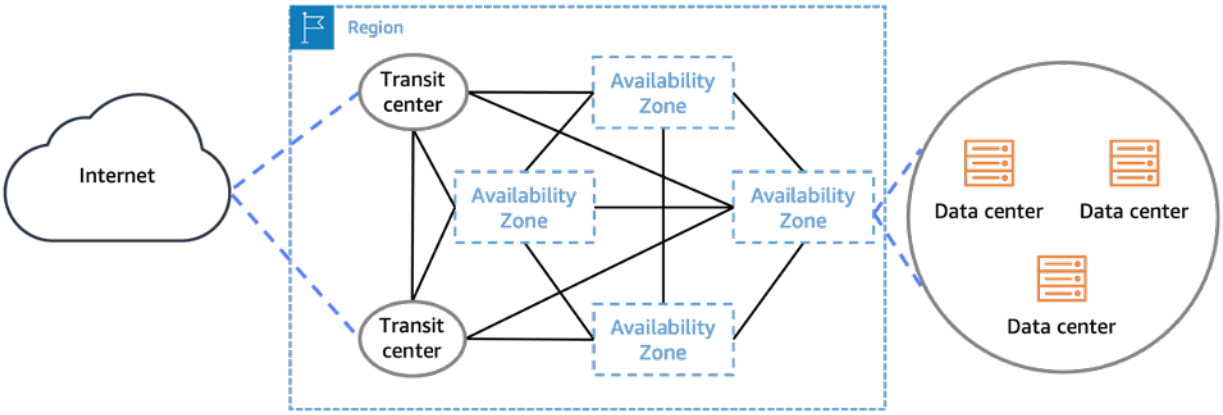
Zonas de disponibilidade

AWSopera mais de 100 zonas de disponibilidade em várias regiões ao redor do mundo (os números atuais podem ser encontrados aqui: Infraestrutura AWS global). Uma zona de disponibilidade é um ou mais data centers discretos com infraestrutura de energia, rede e conectividade independentes e redundantes em um. Região da AWS As zonas de disponibilidade em uma região estão significativamente distantes umas das outras, até 60 milhas (~ 100 km) para evitar falhas correlacionadas, mas próximas o suficiente para usar a replicação síncrona com latência de milissegundos de um dígito. Eles foram projetados para não serem afetados simultaneamente por um cenário de destino compartilhado, como energia elétrica, interrupção da água, isolamento de fibras, terremotos, incêndios, tornados ou inundações. Pontos comuns de falha, como geradores e equipamentos de resfriamento, não são compartilhados entre as zonas de disponibilidade e são projetados para serem fornecidos por subestações de energia independentes. Quando AWS

implanta atualizações em seus serviços, as implantações em zonas de disponibilidade na mesma região são separadas no tempo para evitar falhas correlacionadas.

Todas as zonas de disponibilidade em uma região são interconectadas com redes de alta largura de banda e baixa latência, por meio de fibra metropolitana dedicada e totalmente redundante. Cada zona de disponibilidade em uma região se conecta à Internet por meio de dois centros de trânsito onde há AWS pares com vários provedores de internet de nível 1 (para obter mais informações, consulte Visão geral da Amazon Web Services).

Esses recursos fornecem um forte isolamento das zonas de disponibilidade umas das outras, o que chamamos de Independência da Zona de Disponibilidade (AZI). A construção lógica das zonas de disponibilidade e sua conectividade com a Internet é mostrada na figura a seguir.



As zonas de disponibilidade consistem em um ou mais data centers físicos conectados de forma redundante entre si e à Internet

Regiões

Cada uma Região da AWS consiste em várias zonas de disponibilidade independentes e fisicamente separadas dentro de uma área geográfica. Atualmente, todas as regiões têm três ou mais zonas de disponibilidade. As próprias regiões são isoladas e independentes de outras regiões, com algumas exceções mencionadas posteriormente neste documento (consulte Operações globais de uma única região). Essa separação entre regiões limita as falhas de serviço, quando elas ocorrem, a uma única região. As operações normais de outras regiões não são afetadas neste caso. Além disso, os recursos e dados que você cria em uma região não existem em nenhuma outra região, a menos que você use explicitamente um recurso de replicação ou cópia oferecido por um AWS serviço ou replique o recurso você mesmo.

Regiões atuais e planejadas da AWS a partir de dezembro de 2022

AWSZonas Locais

AWSAs Zonas Locais são um tipo de implantação de infraestrutura que coloca computação, armazenamento, banco de dados e outros AWSserviços selecionados próximos a grandes centros populacionais e industriais. Você pode usar AWS serviços, como serviços de computação e armazenamento, na Zona Local para executar aplicativos de baixa latência na borda ou simplificar as migrações para a nuvem híbrida. As Zonas Locais têm entrada e saída locais de Internet para reduzir a latência, mas também estão conectadas à sua região principal por meio da rede privada redundante e de alta largura de banda da Amazon, oferecendo aos aplicativos executados em Zonas AWS Locais acesso rápido, seguro e contínuo a toda a gama de serviços.

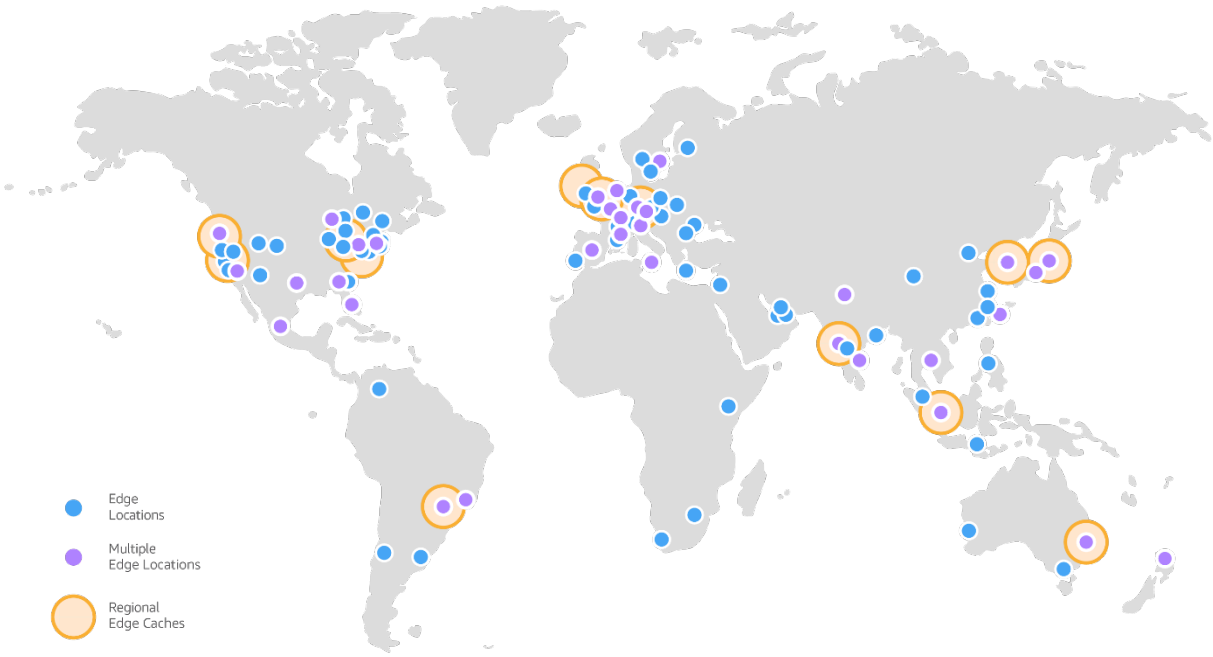
AWS Outposts

AWS Outpostsé uma família de soluções totalmente gerenciadas que fornece AWS infraestrutura e serviços para praticamente qualquer local local ou local periférico para uma experiência híbrida verdadeiramente consistente. As soluções Outposts permitem que você estenda e execute AWS serviços nativos no local e estão disponíveis em vários formatos, de servidores Outposts de 1U e 2U a racks de 42U Outposts e várias implantações de racks.

ComAWS Outposts, você pode executar AWSserviços selecionados localmente e se conectar a uma ampla variedade de serviços disponíveis na matrizRegião da AWS. AWS Outpostssão racks de computação e armazenamento totalmente gerenciados e configuráveis, construídos com hardware AWS projetado que permite que os clientes executem computação e armazenamento no local, enquanto se conectam perfeitamente à ampla variedade AWS de serviços na nuvem.

Pontos de presença

Além das zonas de disponibilidade Regiões da AWS e de disponibilidade, AWS também opera uma rede de ponto de presença (PoP) distribuída globalmente. Eles PoPs hospedam a Amazon CloudFront, uma rede de entrega de conteúdo (CDN); o Amazon Route 53, um serviço público de resolução de Sistema de Nomes de Domínio (DNS); e o AWS Global Accelerator (AGA), um serviço de otimização de rede de ponta. Atualmente, a rede de borda global consiste em mais de 410 PoPs, incluindo mais de 400 pontos de presença e 13 caches regionais de médio porte em mais de 90 cidades em 48 países (o status atual pode ser encontrado aqui: Amazon CloudFront Key Features).



Rede de ponta CloudFront global da Amazon

Cada PoP é isolado dos outros, o que significa que uma falha que afeta um único PoP ou área metropolitana não afeta o resto da rede global. A AWS rede é compatível com milhares de operadoras de telecomunicações de nível 1/2/3 em todo o mundo, está bem conectada a todas as principais redes de acesso para um desempenho ideal e tem centenas de terabits de capacidade

implantada. As localizações periféricas são conectadas ao backbone Regiões da AWS através da AWS rede, uma fibra paralela múltipla de 100 GbE totalmente redundante que circunda o globo e se conecta a dezenas de milhares de redes para melhorar as buscas de origem e acelerar o conteúdo dinâmico.

Partições

AWSagrupa regiões em partições. Cada região está em exatamente uma partição, e cada partição tem uma ou mais regiões. As partições têm instâncias independentes de AWS Identity and Access Management (IAM) e fornecem um limite rígido entre regiões em partições diferentes. AWSAs regiões comerciais estão na `aws` partição, as regiões na China estão na `aws-cn` partição e AWS GovCloud as regiões estão na `aws-us-gov` partição. Alguns AWS serviços são projetados para fornecer funcionalidade entre regiões, como a replicação entre regiões do Amazon S3 ou o emparelhamento entre regiões do AWSTransit Gateway. Esses tipos de recursos só são compatíveis entre regiões na mesma partição. Você não pode usar as credenciais do IAM de uma partição para interagir com recursos em uma partição diferente.

Ambientes de gerenciamento e planos de dados

AWSsepara a maioria dos serviços nos conceitos de plano de controle e plano de dados. Esses termos vêm do mundo das redes, especificamente dos roteadores. O plano de dados do roteador, que é sua principal funcionalidade, está movendo pacotes com base em regras. Mas as políticas de roteamento precisam ser criadas e distribuídas de algum lugar, e é aí que entra o plano de controle.

Os planos de controle fornecem as APIs administrativas usadas para criar, ler/descrever, atualizar, excluir e listar recursos (CRUDL). Por exemplo, todas as ações do plano de controle são: iniciar uma nova instância do Amazon Elastic Compute Cloud (Amazon EC2), criar um bucket do Amazon Simple Storage Service (Amazon S3) e descrever uma fila do Amazon Simple Queue Service (Amazon SQS). Quando você executa uma instância do EC2, o plano de controle precisa realizar várias tarefas, como encontrar um host físico com capacidade, alocar as interfaces de rede, preparar um volume do Amazon Elastic Block Store (Amazon EBS), gerar credenciais do IAM, adicionar as regras do Security Group e muito mais. Os planos de controle tendem a ser sistemas complicados de orquestração e agregação.

O plano de dados é o que fornece a função principal do serviço. Por exemplo, a seguir estão todas as partes do plano de dados para cada um dos serviços envolvidos: a própria instância do EC2 em execução, leitura e gravação em um volume do EBS, obtenção e colocação de objetos em um bucket do S3 e o Route 53 respondendo a consultas de DNS e realizando verificações de saúde.

Os planos de dados são intencionalmente menos complicados, com menos partes móveis em comparação com os planos de controle, que geralmente implementam um sistema complexo de fluxos de trabalho, lógica de negócios e bancos de dados. Isso torna os eventos de falha estatisticamente menos prováveis de ocorrerem no plano de dados em relação ao plano de controle. Embora os dados e o plano de controle contribuam para a operação geral e o sucesso do serviço, os AWS considera componentes distintos. Essa separação traz benefícios de desempenho e disponibilidade.

Estabilidade estática

Uma das características de resiliência mais importantes dos AWS serviços é o que AWS chama de estabilidade estática. O que esse termo significa é que os sistemas operam em um estado estático e continuam operando normalmente, sem a necessidade de fazer alterações durante a falha ou a indisponibilidade das dependências. Uma maneira de fazer isso é evitar dependências circulares em nossos serviços que possam impedir a recuperação bem-sucedida de um desses serviços. Outra forma de fazer isso é mantendo o estado existente. Consideramos o fato de que os planos de controle são estatisticamente mais propensos a falhar do que os planos de dados. Embora o plano de dados normalmente dependa dos dados que chegam do plano de controle, o plano de dados mantém seu estado existente e continua funcionando mesmo em face da deficiência do plano de controle. O acesso do plano de dados aos recursos, uma vez provisionado, não depende do plano de controle e, portanto, não é afetado por nenhuma deficiência no plano de controle. Em outras palavras, mesmo que a capacidade de criar, modificar ou excluir recursos seja prejudicada, os recursos existentes permanecerão disponíveis. Isso torna AWS os planos de dados estaticamente estáveis a uma deficiência no plano de controle. Você pode implementar padrões diferentes para ser estaticamente estável contra diferentes tipos de falhas de dependência.

Um exemplo de estabilidade estática pode ser encontrado no Amazon EC2. Depois que uma instância do EC2 é iniciada, ela fica tão disponível quanto o servidor físico em um data center. Ele não depende de nenhuma API do plano de controle para continuar em execução ou para começar a ser executado novamente após uma reinicialização. A mesma propriedade vale para outros AWS recursos, como VPCs, buckets e objetos do Amazon S3 e volumes do Amazon EBS.

A estabilidade estática é um conceito profundamente enraizado na forma como AWS projeta seus serviços, mas também é um padrão que pode ser usado pelos clientes. Na verdade, a maioria das diretrizes de melhores práticas para usar os diferentes tipos de AWS serviços de forma resiliente é implementar estabilidade estática em ambientes de produção. Os mecanismos de recuperação e mitigação mais confiáveis são aqueles que exigem menos mudanças para alcançar a recuperação.

Em vez de depender do plano de controle do EC2 para iniciar novas instâncias do EC2 para se recuperar de uma zona de disponibilidade com falha, ter essa capacidade extra pré-provisionada ajuda a obter estabilidade estática. Assim, eliminar as dependências nos planos de controle (as APIs que implementam mudanças nos recursos) em seu caminho de recuperação ajuda a produzir cargas de trabalho mais resilientes. Para obter mais detalhes sobre estabilidade estática, planos de controle e planos de dados, consulte o artigo Estabilidade estática usando zonas de disponibilidade da Amazon Builders' Library.

Resumo

AWSutiliza diferentes contêineres de falhas em nossa infraestrutura para criar isolamento de falhas. Os principais contêineres de falhas da infraestrutura são partições, regiões, zonas de disponibilidade, planos de controle e planos de dados. A seguir, examinaremos diferentes tipos de AWS serviços, como esses contêineres de falhas são utilizados em seu design e como você deve arquitetar cargas de trabalho com eles para serem resilientes.

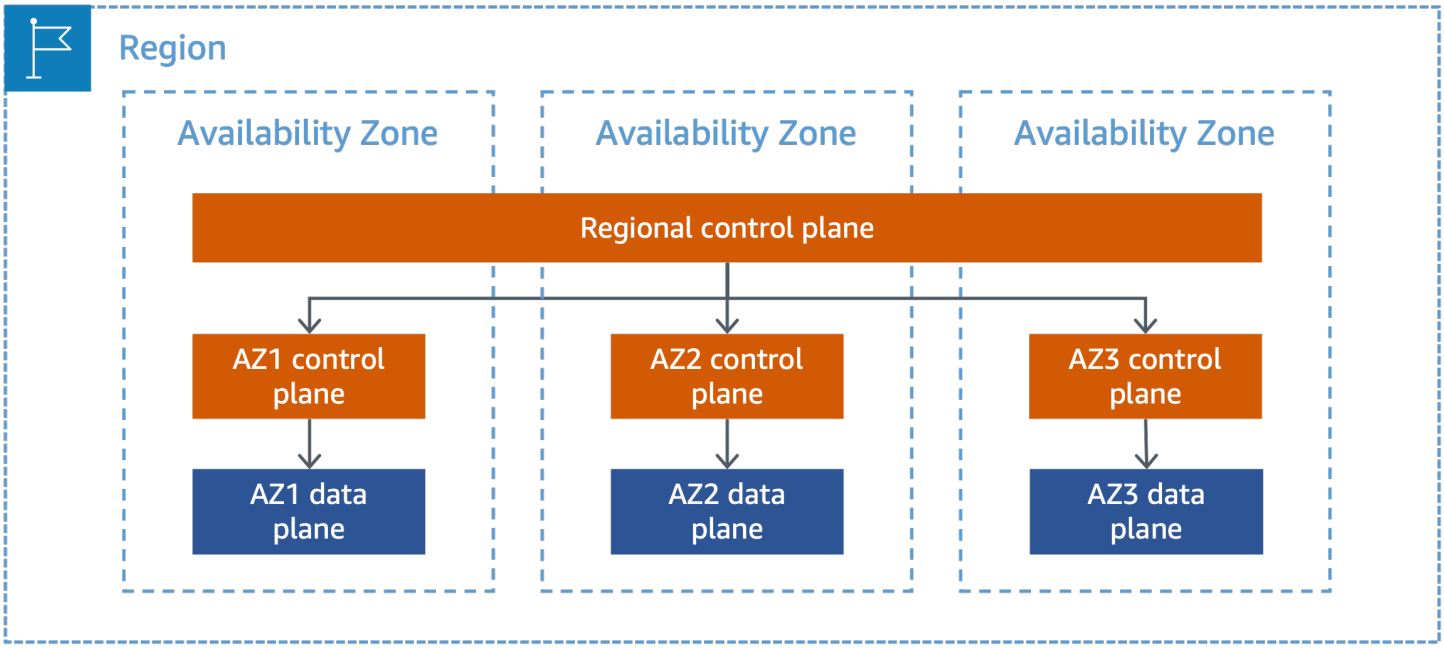
AWS tipos de serviço

AWS opera três categorias diferentes de serviços com base em seu limite de isolamento de falhas: zonal, regional e global. Esta seção descreverá com mais detalhes como esses diferentes tipos de serviços foram projetados para que você possa determinar como as falhas em um serviço de um determinado tipo de serviço afetarão sua carga de trabalho em AWS execução. Ele também fornece orientação de alto nível sobre como arquitetar suas cargas de trabalho para usar esses serviços de forma resiliente. Para serviços globais, este documento também fornece orientações prescritivas Apêndice B - Orientação de serviço global da rede Edge que podem ajudá-lo a evitar o impacto em Apêndice A - Orientação de serviço particional suas cargas de trabalho devido a deficiências nos AWS serviços do plano de controle, ajudando você a depender dos serviços globais com segurança e minimizando a introdução de pontos únicos de falha.

Tópicos

Serviços zonais

A Independência da Zona de Disponibilidade (AZI) permite AWS oferecer serviços zonais, como Amazon EC2 e Amazon EBS. Um serviço zonal é aquele que fornece a capacidade de especificar em qual zona de disponibilidade os recursos são implantados. Esses serviços operam de forma independente em cada zona de disponibilidade dentro de uma região e, o mais importante, também falham de forma independente em cada zona de disponibilidade. Isso significa que os componentes de um serviço em uma zona de disponibilidade não dependem de componentes em outras zonas de disponibilidade. Podemos fazer isso porque um serviço zonal tem planos de dados zonais. Em alguns casos, como no EC2, o serviço também inclui planos de controle zonal para operações alinhadas por zona, como iniciar uma instância do EC2. Para esses serviços, AWS também fornece um endpoint de plano de controle regional para facilitar a interação com o serviço. O plano de controle regional também fornece funcionalidade com escopo regional, além de servir como uma camada de agregação e roteamento sobre os planos de controle zonais. Isso é mostrado na figura a seguir.



Um serviço zonal com planos de controle e planos de dados isolados por zona

As zonas de disponibilidade oferecem aos clientes a capacidade de operar cargas de trabalho de produção mais altamente disponíveis, tolerantes a falhas e escaláveis do que seria possível em um único data center. Quando uma carga de trabalho usa várias zonas de disponibilidade, os clientes ficam melhor isolados e protegidos de problemas que afetam a infraestrutura física de uma única zona de disponibilidade. Isso ajuda os clientes a criar serviços redundantes em todas as zonas de disponibilidade e, se arquitetados corretamente, permanecerem operacionais mesmo se uma zona de disponibilidade apresentar falhas. Os clientes podem aproveitar o AZI para criar cargas de trabalho altamente disponíveis e resilientes. A implementação do AZI em sua arquitetura ajuda você a se recuperar rapidamente de uma falha isolada na zona de disponibilidade porque seus recursos em uma zona de disponibilidade minimizam ou eliminam a interação com recursos em outras zonas de disponibilidade. Isso ajuda a remover dependências entre zonas de disponibilidade, o que simplifica a evacuação da zona de disponibilidade. Consulte Padrões avançados de resiliência Multi-AZ para obter mais detalhes sobre a criação de mecanismos de evacuação da zona de disponibilidade. Além disso, você pode aproveitar ainda mais as zonas de disponibilidade seguindo algumas das mesmas práticas recomendadas AWS usadas em seus próprios serviços, como implantar apenas alterações em uma única zona de disponibilidade por vez ou remover uma zona de disponibilidade do serviço se uma alteração nessa zona de disponibilidade der errado.

A estabilidade estática também é um conceito importante para arquiteturas de zonas de multidisponibilidade. Um dos modos de falha que você deve planejar com arquiteturas de zona de disponibilidade múltipla é a perda de uma zona de disponibilidade, o que pode resultar na perda da

capacidade de uma zona de disponibilidade. Se você não pré-provisionou capacidade suficiente para lidar com a perda de uma zona de disponibilidade, isso pode fazer com que sua capacidade restante seja sobrecarregada pela carga atual. Além disso, você precisará depender dos planos de controle dos serviços zonais usados para substituir essa capacidade perdida, que pode ser menos confiável do que um projeto estaticamente estável. Nesse caso, pré-provisionar capacidade extra suficiente pode ajudá-lo a ficar estaticamente estável à perda de um domínio de falha, como uma zona de disponibilidade, ao ser capaz de continuar as operações normais sem a necessidade de alterações dinâmicas.

Você pode optar por usar um grupo de auto scaling de instâncias do EC2 implantadas em várias zonas de disponibilidade para aumentar e reduzir dinamicamente a escala, com base nas necessidades de sua carga de trabalho. O escalonamento automático funciona bem para mudanças graduais no uso que ocorrem de minutos a dezenas de minutos. No entanto, lançar novas instâncias do EC2 leva tempo, especialmente se suas instâncias exigirem inicialização (como instalação de agentes, binários de aplicativos ou arquivos de configuração). Durante esse tempo, sua capacidade restante pode ser sobrecarregada pela carga atual. Além disso, a implantação de novas instâncias por meio do auto scaling depende do plano de controle do EC2. Isso apresenta uma desvantagem: para ser estaticamente estável à perda de uma única zona de disponibilidade, você precisa pré-provisionar instâncias EC2 suficientes nas outras zonas de disponibilidade para lidar com a carga que foi transferida da zona de disponibilidade prejudicada, em vez de depender do escalonamento automático para provisionar novas instâncias. No entanto, o pré-provisionamento de capacidade extra pode gerar custos adicionais.

Por exemplo, durante a operação normal, vamos supor que sua carga de trabalho exija seis instâncias para atender ao tráfego de clientes em três zonas de disponibilidade. Para ser estaticamente estável contra uma única falha na zona de disponibilidade, você implantaria três instâncias em cada zona de disponibilidade, totalizando nove. Se uma única zona de disponibilidade de instâncias falhasse, você ainda teria seis e seria capaz de continuar atendendo ao tráfego de seus clientes sem a necessidade de provisionar e configurar novas instâncias durante a falha. Obter estabilidade estática para sua capacidade do EC2 tem um custo adicional, pois, nesse caso, você está executando 50% de instâncias adicionais. Nem todos os serviços em que você pode pré-provisionar recursos terão custos adicionais, como o pré-provisionamento de um bucket S3 ou de um usuário. Você precisará ponderar todas as desvantagens da implementação da estabilidade estática em relação ao risco de exceder o tempo de recuperação desejado para sua carga de trabalho.

AWS Locais Zones and Outposts aproximam o plano de dados de AWS serviços selecionados dos usuários finais. Os planos de controle desses serviços residem na região principal. Sua instância Local Zone ou Outposts terá dependências de plano de controle para serviços zonais como EC2

e EBS na Zona de Disponibilidade em que você criou sua Zona Local ou sub-rede Outposts. Eles também terão dependências de planos de controle regionais para serviços regionais, como o Elastic Load Balancing (ELB), grupos de segurança e o plano de controle Kubernetes gerenciado pelo Amazon Elastic Kubernetes Service (Amazon EKS) (se você usar o EKS). Para obter informações adicionais específicas sobre Outposts, consulte a documentação e as perguntas frequentes sobre suporte e manutenção. Implemente estabilidade estática ao usar Locais Zones ou Outposts para ajudar a melhorar a resiliência para controlar deficiências do plano ou interrupções na conectividade da rede com a região principal.

Serviços regionais

Os serviços regionais são serviços criados com base em várias zonas de disponibilidade para que os clientes não precisem descobrir como fazer o melhor uso dos serviços zonais. AWS Agrupamos logicamente o serviço implantado em várias zonas de disponibilidade para apresentar um único endpoint regional aos clientes. O Amazon SQS e o Amazon DynamoDB são exemplos de serviços regionais. Eles usam a independência e a redundância das zonas de disponibilidade para minimizar as falhas na infraestrutura como uma categoria de risco de disponibilidade e durabilidade. O Amazon S3, por exemplo, distribui solicitações e dados em várias zonas de disponibilidade e foi projetado para se recuperar automaticamente da falha de uma zona de disponibilidade. No entanto, você só interage com o endpoint regional do serviço.

AWS acredita que a maioria dos clientes pode atingir suas metas de resiliência em uma única região usando serviços regionais ou arquiteturas Multi-AZ que dependem de serviços zonais. No entanto, algumas cargas de trabalho podem exigir redundância adicional, e você pode usar o isolamento de Regiões da AWS para criar arquiteturas multirregionais para fins de HA ou continuidade de negócios. A separação física e lógica entre elas Regiões da AWS evita falhas correlacionadas entre elas. Em outras palavras, da mesma forma que se você fosse um cliente do EC2 e pudesse se beneficiar do isolamento das zonas de disponibilidade implantando em todas elas, você pode obter o mesmo benefício para serviços regionais implantando em várias regiões. Isso exige que você implemente uma arquitetura multirregional para seu aplicativo, o que pode ajudá-lo a ser resiliente às deficiências de um serviço regional.

No entanto, obter os benefícios de uma arquitetura multirregional pode ser difícil; é necessário um trabalho cuidadoso para tirar proveito do isolamento regional sem desfazer nada no nível do aplicativo. Por exemplo, se você estiver fazendo o failover de um aplicativo entre regiões, precisará manter uma separação estrita entre as pilhas de aplicativos em cada região, estar ciente de todas as dependências do aplicativo e realizar o failover de todas as partes do aplicativo em

conjunto. Conseguir isso com uma arquitetura complexa baseada em microsserviços que tem muitas dependências entre aplicativos requer planejamento e coordenação entre muitas equipes de engenharia e negócios. Permitir que cargas de trabalho individuais tomem suas próprias decisões de failover torna a coordenação menos complexa, mas introduz o comportamento modal por meio da diferença significativa na latência que ocorre entre regiões em comparação com dentro de uma única região.

AWS não fornece um recurso de replicação síncrona entre regiões no momento. Ao usar um armazenamento de dados replicado de forma assíncrona (fornecido por AWS) entre regiões, existe a possibilidade de perda ou inconsistência de dados quando você executa o failover do seu aplicativo entre regiões. Para mitigar possíveis inconsistências, você precisa de um processo confiável de reconciliação de dados no qual tenha confiança e talvez precise operar em vários armazenamentos de dados em seu portfólio de carga de trabalho, ou precisa estar disposto a aceitar a perda de dados. Por fim, você precisa praticar o failover para saber se ele funcionará quando você precisar. A rotação regular de seu aplicativo entre regiões para praticar o failover é um investimento substancial de tempo e recursos. Se você decidir usar um armazenamento de dados replicado de forma síncrona entre regiões para suportar seus aplicativos executados em mais de uma região simultaneamente, as características de desempenho e a latência desse banco de dados que se estende por centenas ou milhares de milhas são muito diferentes de um banco de dados operando em uma única região. Isso exige que você planeje sua pilha de aplicativos desde o início para considerar esse comportamento. Isso também torna a disponibilidade de ambas as regiões uma forte dependência, o que pode resultar na diminuição da resiliência de sua carga de trabalho.

Serviços globais

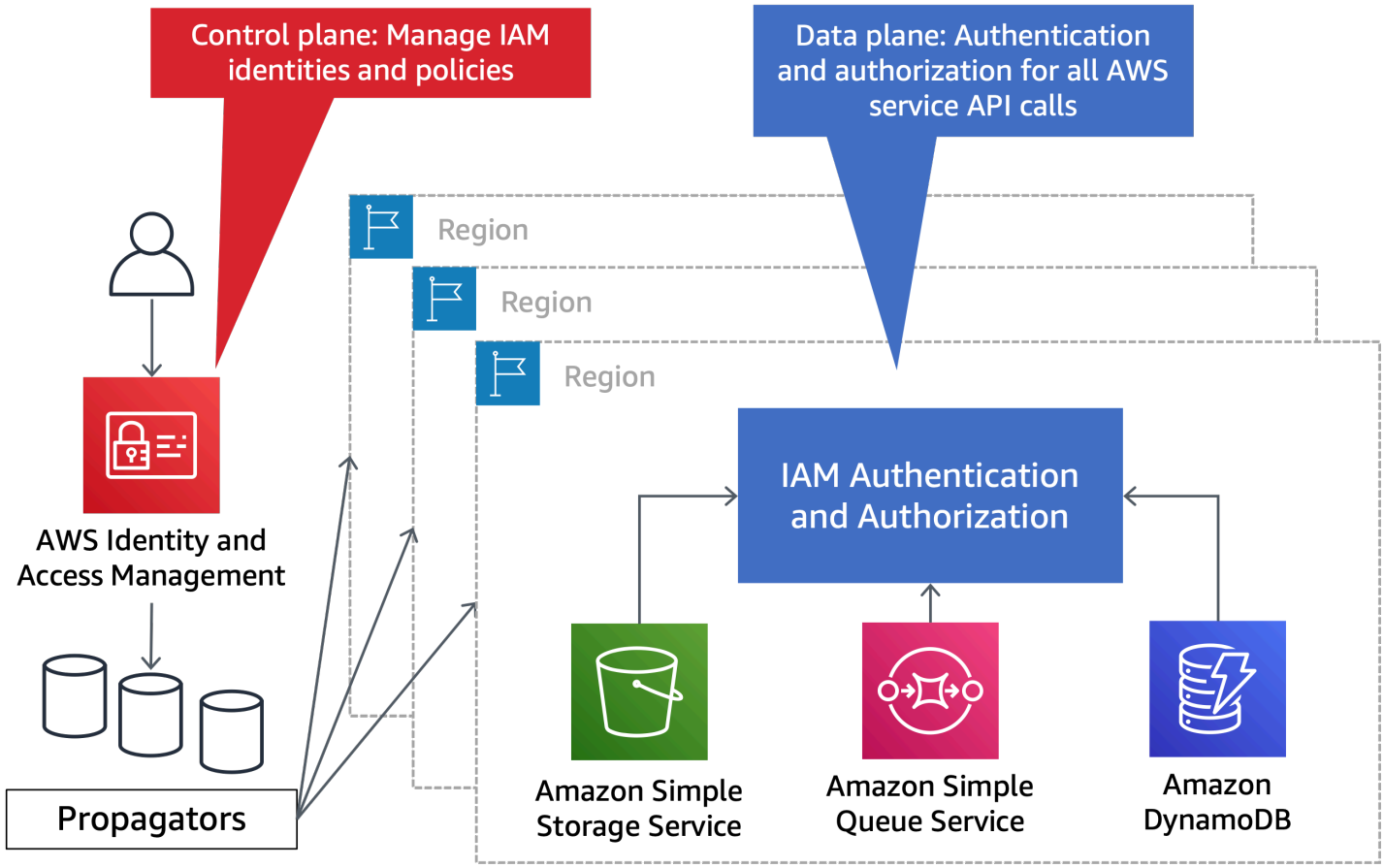
Além dos AWS serviços regionais e zonais, há um pequeno conjunto de AWS serviços cujos planos de controle e planos de dados não existem de forma independente em cada região. Como seus recursos não são específicos da região, eles são comumente chamados de globais. AWS Os serviços globais ainda seguem o padrão de AWS projeto convencional de separar o plano de controle e o plano de dados para obter estabilidade estática. A diferença significativa para a maioria dos serviços globais é que seu plano de controle é hospedado em um único plano Região da AWS, enquanto seu plano de dados é distribuído globalmente. Há três tipos diferentes de serviços globais e um conjunto de serviços que podem parecer globais com base na configuração selecionada.

As seções a seguir identificarão cada tipo de serviço global e como seus planos de controle e planos de dados são separados. Você pode usar essas informações para orientar como criar mecanismos confiáveis de alta disponibilidade (HA) e recuperação de desastres (DR) sem precisar depender de um plano de controle de serviço global. Essa abordagem ajuda a remover pontos únicos de falha

em sua arquitetura e evita possíveis impactos entre regiões, mesmo quando você está operando em uma região diferente de onde o plano de controle de serviço global está hospedado. Também ajuda você a implementar com segurança mecanismos de failover que não dependem de planos de controle de serviço global.

Serviços globais que são exclusivos por partição

Alguns AWS serviços globais existem em cada partição (referidos neste paper como serviços particionais). Os serviços particionais fornecem seu plano de controle em um único Região da AWS. Alguns serviços particionais, como o AWS Network Manager, são somente do plano de controle e orquestram o plano de dados de outros serviços. Outros serviços particionais, como o IAM, têm seu próprio plano de dados, isolado e distribuído em todos os Regiões da AWS da partição. Falhas em um serviço particional não afetam outras partições. Na `aws` partição, o plano de controle do serviço IAM está na `us-east-1` região, com planos de dados isolados em cada região da partição. Os serviços particionais também têm planos de controle e planos de dados independentes nas `aws-cn` partições `aws-us-gov` e. A separação do plano de controle e do plano de dados para o IAM é mostrada no diagrama a seguir.



O IAM tem um único plano de controle e um plano de dados regionalizado

A seguir estão os serviços particionais e sua localização no plano de controle na `aws` partição:

- AWS IAM (`us-east-1`)
- AWS Organizations (`us-east-1`)
- AWS Gerenciamento de contas (`us-east-1`)
- Route 53 Application Recovery Controller (ARC`us-west-2`) () - Este serviço está presente somente na `aws` partição
- AWS Gerente de rede (`us-west-2`)
- DNS privado do Route 53 () `us-east-1`

Se algum desses planos de controle de serviço tiver um evento que afete a disponibilidade, talvez você não consiga usar as operações do tipo CRUDL fornecidas por esses serviços. Portanto, se sua estratégia de recuperação depender dessas operações, um impacto na disponibilidade do plano de controle ou na região que hospeda o plano de controle reduzirá suas chances de recuperação bem-sucedida. Apêndice A - Orientação de serviço particionalfornece estratégias para remover dependências em planos de controle de serviços globais durante a recuperação.

> **Recomendação**
>
> Não confie nos planos de controle dos serviços particionais em seu caminho de recuperação. Em vez disso, confie nas operações do plano de dados desses serviços. Consulte Apêndice A - Orientação de serviço particional para obter detalhes adicionais sobre como você deve projetar serviços particionais.

Serviços globais na rede de ponta

O próximo conjunto de AWS serviços globais tem um plano de controle na `aws` partição e hospeda seus planos de dados na infraestrutura de pontos de presença globais (PoP) (e potencialmente Regiões da AWS também). Os planos de dados hospedados PoPs podem ser acessados a partir de recursos em qualquer partição, bem como na Internet. Por exemplo, o Route 53 opera seu plano de controle na `us-east-1` região, mas seu plano de dados é distribuído em centenas de países PoPs , bem como em cada um deles Região da AWS (para oferecer suporte ao DNS público e privado do Route 53 na região). As verificações de integridade do Route 53 também fazem parte do plano de dados e são executadas a partir de oito Regiões da AWS na `aws` partição. Os clientes podem resolver o DNS usando zonas hospedadas públicas do Route 53 de qualquer lugar na Internet,

incluindo outras partições GovCloud, como, por exemplo, de uma AWS Virtual Private Cloud (VPC). A seguir estão os serviços de rede de borda global e sua localização no plano de controle na `aws` partição:

- DNS público do Route 53 () `us-east-1`
- Amazon CloudFront (`us-east-1`)
- AWS WAF Clássico 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`)

Se você usa verificações de saúde AGA para instâncias EC2 ou endereços IP elásticos, elas usam verificações de saúde do Route 53. Criar ou atualizar as verificações de saúde do AGA dependeria do plano de controle do Route 53 em`us-east-1`. A execução das verificações de saúde da AGA utiliza o plano de dados de verificação de saúde do Route 53.

Durante uma falha afetando a região que hospeda os planos de controle desses serviços, ou uma falha afetando o próprio plano de controle, talvez você não consiga usar as operações do tipo CRUDL fornecidas por esses serviços. Se você depender dessas operações em sua estratégia de recuperação, essa estratégia pode ter menos probabilidade de sucesso do que se você confiasse apenas no plano de dados desses serviços.

> **Recomendação**
>
> Não confie no plano de controle dos serviços de rede de ponta em seu caminho de recuperação. Em vez disso, confie nas operações do plano de dados desses serviços. Consulte Apêndice B - Orientação de serviço global da rede Edge para obter detalhes adicionais sobre como projetar serviços globais na rede de borda.

Operações globais em uma única região

A categoria final é composta por operações específicas do plano de controle dentro de um serviço que tem um escopo de impacto global, não por serviços inteiros, como nas categorias anteriores. Enquanto você interage com serviços zonais e regionais na região especificada, determinadas

operações têm uma dependência subjacente em uma única região que é diferente de onde o recurso está localizado. Eles são diferentes dos serviços que são fornecidos somente em uma única região; consulte Apêndice C - Serviços de região única para obter uma lista desses serviços.

Durante uma falha que afeta a dependência global subjacente, talvez você não consiga usar as ações do tipo CRUDL das operações dependentes. Se você depender dessas operações em sua estratégia de recuperação, essa estratégia pode ter menos probabilidade de sucesso do que se você confiasse apenas no plano de dados desses serviços. Você deve evitar dependências dessas operações para sua estratégia de recuperação.

A seguir está uma lista de serviços dos quais outros serviços podem depender, que têm escopo global:

- Rota 53

  Vários AWS serviços criam recursos que fornecem nomes DNS específicos para cada recurso. Por exemplo, quando você provisiona um Elastic Load Balancer (ELB), o serviço cria registros DNS públicos e verificações de saúde no Route 53 para o ELB. Isso depende do plano de controle do Route 53 em`us-east-1`. Outros serviços que você usa também podem precisar provisionar um ELB, criar registros DNS públicos do Route 53 ou criar verificações de saúde do Route 53 como parte de seus fluxos de trabalho do plano de controle. Por exemplo, provisionar um recurso de API REST do Amazon API Gateway, um banco de dados do Amazon Relational Database Service (Amazon RDS) ou um domínio do Amazon OpenSearch Service resultam na criação de registros DNS no Route 53. A seguir está uma lista de serviços cujo plano de controle depende do plano de controle do Route 53 `us-east-1` para criar, atualizar ou excluir registros DNS, zonas hospedadas e/ou criar verificações de saúde do Route 53. Essa lista não é exaustiva; ela serve para destacar alguns dos serviços mais usados, cujas ações do plano de controle para criar, atualizar ou excluir recursos dependem do plano de controle do Route 53:

  - APIs REST e HTTP do Amazon API Gateway
  - Instâncias do Amazon RDS
  - Bancos de dados Amazon Aurora
  - Balanceadores de carga Amazon ELB
  - AWS PrivateLink Endpoints de VPC
  - AWS Lambda URLs
  - Amazon ElastiCache
  - OpenSearch Serviço Amazon

- Amazon CloudFront
- Amazon MemoryDB para Redis
- Amazon Neptune
- Amazon DynamoDB Accelerator (DAX)
- AGA
- Amazon Elastic Container Service (Amazon ECS) com Service Discovery baseado em DNS (que usa a API para gerenciar o DNS AWS Cloud Map do Route 53)
- Plano de controle Amazon EKS Kubernetes

  É importante observar que o serviço VPC DNS para nomes de host de instâncias do EC2 existe de forma independente em cada um Região da AWS e não depende do plano de controle do Route 53. Registros AWS criados para instâncias do EC2 no serviço VPC DNS, `ip-10-0-10.ec2.internal` como,,, `i-0123456789abcdef.us-west-2.compute.internal` e `ip-10-0-1-5.compute.us-west-2.compute.internali-0123456789abcdef.ec2.internal`, não dependem do plano de controle do Route 53 em. `us-east-1`

  > **Recomendação**
  >
  > Não confie na criação, atualização ou exclusão de recursos que exijam a criação, atualização ou exclusão de registros de recursos, zonas hospedadas ou verificações de saúde do Route 53 em seu caminho de recuperação. Pré-provisione esses recursos, como ELBs, para evitar a dependência do plano de controle do Route 53 em seu caminho de recuperação.

- Amazon S3

  As seguintes operações do plano de controle do Amazon S3 têm uma dependência subjacente `us-east-1` na partição. `aws` Uma falha afetando o Amazon S3 ou outros serviços `us-east-1` em pode fazer com que essas ações dos planos de controle sejam prejudicadas em outras regiões:

```
PutBucketCors
DeleteBucketCors
PutAccelerateConfiguration
PutBucketRequestPayment
```

```
PutBucketObjectLockConfiguration
PutBucketTagging
DeleteBucketTagging
PutBucketReplication
DeleteBucketReplication
PutBucketEncryption
DeleteBucketEncryption
PutBucketLifecycle
DeleteBucketLifecycle
PutBucketNotification
PutBucketLogging
DeleteBucketLogging
PutBucketVersioning
PutBucketPolicy
DeleteBucketPolicy
PutBucketOwnershipControls
DeleteBucketOwnershipControls
PutBucketAcl
PutBucketPublicAccessBlock
DeleteBucketPublicAccessBlock
```

O plano de controle dos pontos de acesso multirregionais (MRAP) do Amazon S3 está hospedado somente em `us-west-2` e as solicitações para criar, atualizar ou excluir MRAPs têm como alvo diretamente essa região. O plano de controle do MRAP também tem dependências subjacentes do AGA in`us-west-2`, Route 53 in `us-east-1` e ACM em cada região de onde o MRAP está configurado para fornecer conteúdo. Você não deve depender da disponibilidade do plano de controle do MRAP em seu caminho de recuperação ou nos planos de dados de seus próprios sistemas. Isso é diferente dos controles de failover do MRAP que são usados para especificar o status de roteamento ativo ou passivo para cada um dos seus buckets no MRAP. Essas APIs são hospedadas em cinco Regiões da AWS e podem ser usadas para mudar efetivamente o tráfego usando o plano de dados do serviço.

Além disso, os nomes de bucket do Amazon S3 são globalmente exclusivos e todas as chamadas para as `DeleteBucket` APIs `CreateBucket` e dependem`us-east-1`, na `aws` partição, para garantir a exclusividade do nome, mesmo que a chamada da API seja direcionada à região específica na qual você deseja criar o bucket. Por fim, se você tiver fluxos de trabalho críticos de criação de intervalos, não deverá depender da disponibilidade de nenhuma grafia específica do nome de um intervalo, especialmente aqueles que seguem um padrão perceptível.

> **Recomendação**
>
> Não confie na exclusão ou criação de novos buckets do S3 nem na atualização das configurações do bucket do S3 como parte do seu caminho de recuperação. Pré-provisione todos os buckets S3 necessários com as configurações necessárias para que você não precise fazer alterações para se recuperar de uma falha. Essa abordagem também se aplica aos MRAPs.

- CloudFront

  O Amazon API Gateway fornece endpoints de API otimizados para bordas. A criação desses endpoints depende do plano CloudFront de controle `us-east-1` para criar a distribuição na frente do endpoint do gateway.

  > **Recomendação**
  >
  > Não confie na criação de novos endpoints do API Gateway otimizados para borda como parte do seu caminho de recuperação. Pré-provisione todos os endpoints necessários do API Gateway.

  Todas as dependências discutidas nesta seção são ações do plano de controle, não ações do plano de dados. Se suas cargas de trabalho estiverem configuradas para serem estaticamente estáveis, essas dependências não devem afetar seu caminho de recuperação, lembrando que a estabilidade estática exige trabalho ou serviços adicionais para ser implementada.

Serviços que usam endpoints globais padrão

Em alguns casos, os AWS serviços fornecem um endpoint global padrão, como o AWS Security Token Service (AWS STS). Outros serviços podem usar esse endpoint global padrão em sua configuração padrão. Isso significa que um serviço regional que você está usando pode ter uma dependência global de um único Região da AWS. Os detalhes a seguir explicam como remover dependências não intencionais em endpoints globais padrão que ajudarão você a usar o serviço de forma regional.

AWS STS: O STS é um serviço web que permite solicitar credenciais temporárias com privilégios limitados para usuários do IAM ou para usuários autenticados (usuários federados). O uso de STS do kit de desenvolvimento de AWS software (SDK) e da interface de linha de comando (CLI) é padronizado como. `us-east-1` O serviço STS também fornece endpoints regionais. Esses endpoints são ativados por padrão em regiões que também são ativadas por padrão. Você pode tirar proveito deles a qualquer momento configurando seu SDK ou CLI seguindo estas instruções:AWS Endpoints regionalizados STS. O uso do SIGv4a também requer credenciais temporárias solicitadas de um endpoint regional do STS. Você não pode usar o endpoint global do STS para essa operação.

> Recomendação
>
> Atualize sua configuração de SDK e CLI para usar os endpoints regionais do STS.

Login da Security Assertion Markup Language (SAML): os serviços SAML existem em todos. Regiões da AWSPara usar esse serviço, escolha o endpoint SAML regional apropriado, como https://us-west-2.signin.aws.amazon.com/saml. Você deve fazer atualizações nas configurações em suas políticas de confiança e no Provedor de Identidade (IdP) para usar os endpoints regionais. Consulte a documentação do AWS SAML para obter detalhes específicos.

Se você estiver usando um IdP que também esteja hospedado AWS, existe o risco de que ele também seja afetado durante um AWS evento de falha. Isso pode fazer com que você não consiga atualizar sua configuração de IdP ou talvez não consiga federar totalmente. Você deve pré-provisionar usuários “quebra-vidro” caso seu IdP esteja comprometido ou indisponível. Consulte Apêndice A - Orientação de serviço particional para obter detalhes sobre como criar usuários de quebra-vidros de uma forma estaticamente estável.

> Recomendação
>
> Atualize suas políticas de confiança da função do IAM para aceitar logins SAML de várias regiões. Durante uma falha, atualize sua configuração de IdP para usar um endpoint SAML regional diferente se seu endpoint preferencial estiver comprometido. Crie um (s) usuário (s) inovador (s) caso seu IdP esteja comprometido ou indisponível.

AWS IAM Identity Center: o Identity Center é um serviço baseado em nuvem que facilita o gerenciamento centralizado do acesso de login único aos aplicativos do cliente e da nuvem. Contas da AWS O Identity Center deve ser implantado em uma única região de sua escolha. No entanto, o

comportamento padrão do serviço é usar o endpoint SAML global (https://signin.aws.amazon.com/saml), que está hospedado em. `us-east-1` Se você implantou o Identity Center em outro Região da AWS, você deve atualizar o URL relaystate de cada conjunto de permissões para atingir o mesmo endpoint de console regional da implantação do Identity Center. Por exemplo, se você implantou o Identity Center em`us-west-2`, você deve atualizar o estado de retransmissão dos seus conjuntos de permissões para usar https://us-west-2.console.aws.amazon.com. Isso removerá qualquer dependência `us-east-1` da sua implantação do Identity Center.

Além disso, como o IAM Identity Center só pode ser implantado em uma única região, você deve pré-provisionar usuários “inovadores” caso sua implantação seja prejudicada. Consulte Apêndice A - Orientação de serviço particional para obter detalhes sobre como criar usuários de quebra-vidros de uma forma estaticamente estável.

> Recomendação
>
> Defina a URL relaystate dos seus conjuntos de permissões no IAM Identity Center para corresponder à região em que você tem o serviço implantado. Crie um (s) usuário (s) inovador (s) caso sua implantação do IAM Identity Center não esteja disponível.

Lente de armazenamento Amazon S3: o Storage Lens fornece um painel padrão chamado. default-account-dashboard A configuração do painel e suas métricas associadas são armazenadas em`us-east-1`. Você pode criar painéis adicionais em outras regiões especificando a região de origem para a configuração do painel e os dados métricos.

> Recomendação
>
> Se você precisar de dados do painel padrão do S3 Storage Lens durante uma falha que afeta o serviço`us-east-1`, crie um painel adicional em uma região de origem alternativa. Você também pode duplicar qualquer outro painel personalizado que tenha criado em regiões adicionais.

Resumo dos serviços globais

Os planos de dados para serviços globais aplicam princípios de isolamento e independência semelhantes aos AWS serviços regionais. Uma falha que afeta o plano de dados do IAM em uma região não afeta a operação do plano de dados do IAM em outra Região da AWS. Da mesma forma,

uma falha afetando o plano de dados do Route 53 em um PoP não afeta a operação do plano de dados do Route 53 no resto do PoPs. Portanto, o que devemos considerar são os eventos de disponibilidade de serviços que afetam a região em que o plano de controle opera ou afetam o próprio plano de controle. Como há apenas um único plano de controle para cada serviço global, uma falha que afeta esse plano de controle pode ter efeitos entre regiões nas operações do tipo CRUDL (que são as operações de configuração normalmente usadas para instalar ou configurar um serviço, em oposição ao uso direto do serviço).

A maneira mais eficaz de arquitetar cargas de trabalho para usar serviços globais de forma resiliente é usar a estabilidade estática. Durante um cenário de falha, projete sua carga de trabalho para não precisar fazer alterações em um plano de controle para mitigar o impacto ou o failover em um local diferente. Consulte Apêndice A - Orientação de serviço particional e obtenha Apêndice B - Orientação de serviço global da rede Edge orientações prescritivas sobre como utilizar esses tipos de serviços globais para remover dependências do plano de controle e eliminar pontos únicos de falha. Se você precisar dos dados de uma operação do plano de controle para recuperação, armazene esses dados em um armazenamento de dados que possa ser acessado por meio de seu plano de dados, como um parâmetro do AWS Systems Manager Parameter Store (SSM Parameter Store), uma tabela do DynamoDB ou um bucket do S3. Para redundância, você também pode optar por armazenar esses dados em uma região adicional. Por exemplo, seguindo as melhores práticas do Route 53 Application Recovery Controller (ARC), você deve codificar ou marcar seus cinco endpoints de cluster regionais como favoritos. Durante um evento de falha, talvez você não consiga acessar algumas operações de API, incluindo operações de API do Route 53 ARC que não estão hospedadas no cluster de plano de dados extremamente confiável. Você pode listar os endpoints dos seus clusters ARC do Route 53 usando a operação da `DescribeCluster` API.

A seguir está um resumo de algumas das configurações incorretas ou antipadrões mais comuns que introduzem dependências nos planos de controle dos serviços globais:

- Fazer alterações nos registros do Route 53, como atualizar o valor de um registro A ou alterar os pesos de um conjunto de registros ponderados, para realizar o failover.
- Criar ou atualizar recursos do IAM, incluindo funções e políticas do IAM, durante um failover. Isso normalmente não é intencional, mas pode ser resultado de um plano de failover não testado.
- Contar com o IAM Identity Center para que os operadores tenham acesso aos ambientes de produção durante um evento de falha.
- Confiar na configuração padrão do IAM Identity Center para utilizar o console `us-east-1` quando você implantou o Identity Center em uma região diferente.

- Fazendo alterações nos pesos de discagem de tráfego do AGA para realizar manualmente um failover regional.
- Atualizar a configuração de origem de uma CloudFront distribuição para evitar uma origem danificada.
- Provisionamento de recursos de recuperação de desastres (DR), como instâncias de ELBs e RDS durante um evento de falha, que dependem da criação de registros DNS no Route 53.

A seguir está um resumo das recomendações fornecidas nesta seção para o uso de serviços globais de forma resiliente que ajudaria a evitar os antipadrões comuns anteriores.

Resumo da recomendação

Não confie nos planos de controle dos serviços particionais em seu caminho de recuperação. Em vez disso, confie nas operações do plano de dados desses serviços. Consulte Apêndice A - Orientação de serviço particional para obter detalhes adicionais sobre como você deve projetar serviços particionais.
Não confie no plano de controle dos serviços de rede de ponta em seu caminho de recuperação. Em vez disso, confie nas operações do plano de dados desses serviços. Consulte Apêndice B - Orientação de serviço global da rede Edge para obter detalhes adicionais sobre como projetar serviços globais na rede de borda.
Não confie na criação, atualização ou exclusão de recursos que exijam a criação, atualização ou exclusão de registros de recursos, zonas hospedadas ou verificações de saúde do Route 53 em seu caminho de recuperação. Pré-provisione esses recursos, como ELBs, para evitar a dependência do plano de controle do Route 53 em seu caminho de recuperação.
Não confie na exclusão ou criação de novos buckets do S3 nem na atualização das configurações do bucket do S3 como parte do seu caminho de recuperação. Pré-provisione todos os buckets S3 necessários com as configurações necessárias para que você não precise fazer alterações para se recuperar de uma falha. Essa abordagem também se aplica aos MRAPs.
Não confie na criação de novos endpoints do API Gateway otimizados para borda como parte do seu caminho de recuperação. Pré-provisione todos os endpoints necessários do API Gateway.
Atualize sua configuração de SDK e CLI para usar os endpoints regionais do STS.
Atualize suas políticas de confiança da função do IAM para aceitar logins SAML de várias regiões. Durante uma falha, atualize sua configuração de IdP para usar um endpoint

SAML regional diferente se seu endpoint preferencial estiver comprometido. Crie usuários incomparáveis caso seu IdP esteja comprometido ou indisponível.
Defina a URL relaystate dos seus conjuntos de permissões no IAM Identity Center para corresponder à região em que você tem o serviço implantado. Crie um (s) usuário (s) inovador (s) caso sua implantação do Identity Center não esteja disponível.
Se você precisar de dados do painel padrão do S3 Storage Lens durante uma falha que afeta o serviço`us-east-1`, crie um painel adicional em uma região de origem alternativa. Você também pode duplicar qualquer outro painel personalizado que tenha criado em regiões adicionais.

Conclusão

AWSfornece várias construções diferentes para limites de isolamento de falhas. Você deve considerar como arquitetar serviços zonais, regionais e globais, bem como os possíveis impactos em sua carga de trabalho e na capacidade de recuperação de sua carga de trabalho durante deficiências no plano de controle. A estabilidade estática é uma das principais formas de evitar dependências do plano de controle e criar mecanismos de HA e DR confiáveis e resilientes ao usar AWS serviços.

Apêndice A - Orientação de serviço particional

Para serviços particionais, você deve implementar a estabilidade estática para manter a resiliência de sua carga de trabalho durante uma falha no plano de controle AWS de serviço. O seguinte fornece orientação prescritiva sobre como considerar as dependências de serviços particionais, bem como o que funcionará ou não durante uma falha no plano de controle.

AWS Identity and Access Management (IAM)

O plano de controle AWS Identity and Access Management (IAM) consiste em todas as APIs públicas do IAM (incluindo o Access Advisor, mas não o Access Analyzer ou o IAM Roles Anywhere). Isso inclui ações como `CreateRole`, `AttachRolePolicy`, `ChangePassword`, `UpdateSAMLProvider`,, `UpdateLoginProfile` e. O plano de dados do IAM fornece autenticação e autorização para os diretores do IAM em cada umRegião da AWS. Durante uma interrupção do plano de controle, as operações do tipo CRUDL para IAM podem não funcionar, mas a autenticação e a autorização dos diretores existentes continuarão funcionando. O STS é um serviço exclusivo de plano de dados que é separado do IAM e não depende do plano de controle do IAM.

O que isso significa é que, ao planejar dependências no IAM, você não deve confiar no plano de controle do IAM em seu caminho de recuperação. Por exemplo, um design estaticamente estável para um usuário administrador “revolucionário” seria criar um usuário com as permissões apropriadas anexadas, ter a senha definida, provisionar a chave de acesso e a chave de acesso secreta e, em seguida, bloquear essas credenciais em um cofre físico ou virtual. Quando necessário durante uma emergência, recupere as credenciais do usuário do cofre e use-as conforme necessário. Um non-statically-stable design seria provisionar o usuário durante uma falha ou pré-provisionar o usuário, mas anexar a política administrativa somente quando necessário. Essas abordagens dependeriam do plano de controle do IAM.

AWS Organizations

O plano AWS Organizations de controle consiste em todas as APIs de Organizations públicas `AcceptHandshake`, `AttachPolicy`, como`CreateAccount`,`CreatePolicy`,, e. `ListAccounts` Não há um plano de dados paraAWS Organizations. Ele orquestra o plano de dados para outros serviços, como o IAM. Durante uma interrupção do plano de controle, as operações do tipo CRUDL para Organizations podem não funcionar, mas as políticas, como as Políticas de Controle de

Serviços (SCP) e as Políticas de Etiquetas, continuarão funcionando e sendo avaliadas como parte do processo de autorização do IAM. Os recursos administrativos delegados e os recursos de várias contas em outros AWS serviços que são suportados pelas Organizations também continuarão funcionando.

O que isso significa é que, ao planejar dependênciasAWS Organizations, você não deve confiar no plano de controle da Organizations em seu caminho de recuperação. Em vez disso, implemente estabilidade estática em seu plano de recuperação. Por exemplo, uma non-statically-stable abordagem pode ser atualizar os SCPs para remover as restrições permitidas Regiões da AWS por meio da `aws:RequestedRegion` condição ou habilitar permissões de administrador para funções específicas do IAM. Isso depende do plano de controle Organizations para fazer essas atualizações. Uma abordagem melhor seria usar tags de sessão para conceder o uso de permissões de administrador. Seu provedor de identidade (IdP) pode incluir tags de sessão que podem ser avaliadas em relação à `aws:PrincipalTag` condição, o que ajuda você a configurar dinamicamente as permissões para determinados diretores e, ao mesmo tempo, ajudar seus SCPs a permanecerem estáticos. Isso remove as dependências dos planos de controle e utiliza somente ações do plano de dados.

AWS Account Management

O plano de controle de gerenciamento de AWS contas está hospedado em us-east-1 e consiste em todas as APIs públicas para gerenciar umConta da AWS, como e. `GetContactInformation` `PutContactInformation` Também inclui a criação ou o fechamento de um novo Conta da AWS por meio do console de gerenciamento. As APIs de`CloseAccount`, `CreateAccountCreateGovCloudAccount`, e `DescribeAccount` fazem parte do plano de AWS Organizations controle, que também está hospedado em us-east-1. Além disso, a criação de uma GovCloud conta externa AWS Organizations depende do plano de controle Conta da AWS de gerenciamento em us-east-1. Além disso, GovCloud as contas devem ser vinculadas 1:1 a uma Conta da AWS na `aws` partição. A criação de contas na a `aws-cn` sua. A sua principal rede do U. O plano de dados Contas da AWS é para as contas em si. Durante uma falha no plano de controle, as operações do tipo CRUDL (como criar uma nova conta ou obter e atualizar informações de contato) podem não funcionar. Contas da AWS As referências à conta nas políticas do IAM continuarão funcionando.

O que isso significa é que, ao planejar dependências no Gerenciamento de AWS Contas, você não deve confiar no plano de controle do Gerenciamento de Contas em seu caminho de recuperação. Embora o plano de controle do Gerenciamento de Contas não forneça a funcionalidade direta que

você normalmente usaria em uma situação de recuperação, pode haver momentos em que você o faria. Por exemplo, um design estaticamente estável seria pré-provisionar tudo o que você precisa para o Contas da AWS failover. Um non-statically-stable projeto seria criar um novo Contas da AWS durante um evento de falha para hospedar seus recursos de DR.

Controlador de recuperação de aplicações do Route 53

O plano de controle do Route 53 ARC consiste em APIs para controle de recuperação e prontidão para recuperação, conforme identificado em: endpoints e cotas do Amazon Route 53 Application Recovery Controller. Você gerencia verificações de prontidão, controles de roteamento e operações de cluster usando o plano de controle. O plano de dados do ARC é seu cluster de recuperação, que gerencia os valores de controle de roteamento que são consultados pelas verificações de integridade do Route 53 e também implementa as regras de segurança. A funcionalidade do plano de dados do Route 53 ARC é acessada por meio de suas APIs de cluster de recuperação, como`https://aaaaaaaa.route53-recovery-cluster.eu-west-1.amazonaws.com`.

O que isso significa é que você não deve confiar no plano de controle ARC do Route 53 em seu caminho de recuperação. Há duas melhores práticas que ajudam a implementar essa orientação:

- Primeiro, marque ou codifique os cinco endpoints regionais do cluster. Isso elimina a necessidade de usar a operação do plano DescribeCluster de controle durante um cenário de failover para descobrir os valores do endpoint.
- Segundo, use as APIs de cluster do Route 53 ARC usando a CLI ou o SDK para realizar atualizações nos controles de roteamento e não no. AWS Management Console Isso remove o console de gerenciamento como uma dependência do seu plano de failover e garante que ele dependa somente das ações do plano de dados.

Gerenciador de rede AWS

O serviço AWS Network Manager é principalmente um sistema somente de plano de controle hospedado em us-west-2. A sua principal rede do Transit Gateway A sua principal rede do Transit Nuvem AWS Gateway entre Contas da AWS regiões e locais on-premises da. A sua principal rede do AWS Transit Gateway entre regiões e locais on-premises da. Ele também agrega suas métricas de Cloud WAN em us-west-2, que também podem ser acessadas por meio do CloudWatch plano de dados. Se o Network Manager estiver comprometido, o plano de dados dos serviços que ele orquestra não será afetado. As CloudWatch métricas para Cloud WAN também estão disponíveis em us-west-2. Se você quiser dados métricos históricos, como bytes de entrada e saída por região,

para entender quanto tráfego pode ser transferido para outras regiões durante uma falha que afeta us-west-2, ou para outros fins operacionais, você pode exportar essas métricas como dados CSV diretamente do CloudWatch console ou usando este método: publicar CloudWatch métricas da Amazon em um arquivo CSV. Os dados podem ser encontrados no `AWS/Network Manager` namespace e você pode fazer isso de acordo com um cronograma de sua escolha e armazená-los no S3 ou em outro armazenamento de dados selecionado. Para implementar um plano de recuperação estaticamente estável, não use o AWS Network Manager para fazer atualizações em sua rede nem confie nos dados das operações do plano de controle para entrada de failover.

DNS privado do Route 53

As zonas hospedadas privadas do Route 53 são suportadas em cada partição; no entanto, as considerações para zonas hospedadas privadas e zonas hospedadas públicas no Route 53 são as mesmas. Consulte o Amazon Route 53 no Apêndice B — Orientação de serviço global da rede Edge.

Apêndice B - Orientação de serviço global da rede Edge

Para serviços globais de rede de ponta, você deve implementar a estabilidade estática para manter a resiliência de sua carga de trabalho durante uma interrupção do plano AWS de controle de serviço.

Route 53

O plano de controle do Route 53 consiste em todas as APIs públicas do Route 53 que abrangem a funcionalidade de zonas hospedadas, registros, verificações de integridade, registros de consultas de DNS, conjuntos de delegações reutilizáveis, políticas de tráfego e tags de alocação de custos. Ele é hospedado em us-east-1. O plano de dados é o serviço DNS autoritativo, que é executado em mais de 200 locais PoP, bem como em cada umRegião da AWS, respondendo a consultas de DNS com base nas suas zonas hospedadas e dados de verificação de integridade. Além disso, o Route 53 tem um plano de dados para verificações de saúde, que também é um serviço distribuído globalmente em vários. Regiões da AWS Esse plano de dados realiza verificações de integridade, agrega os resultados e os entrega aos planos de dados do DNS público e privado do Route 53 e do AGA. Durante uma falha no plano de controle, as operações do tipo CRUDL para a Rota 53 podem não funcionar, mas as verificações de resolução e integridade do DNS e as atualizações no roteamento resultantes de alterações nas verificações de saúde continuarão funcionando.

O que isso significa é que, ao planejar dependências no Route 53, você não deve confiar no plano de controle do Route 53 em seu caminho de recuperação. Por exemplo, um design estaticamente estável seria usar o status das verificações de saúde para realizar failovers entre regiões ou evacuar uma zona de disponibilidade. Você pode usar os controles de roteamento do Route 53 Application Recovery Controller (ARC) para alterar manualmente o status das verificações de integridade e alterar as respostas às consultas de DNS. Existem padrões semelhantes aos fornecidos pelo ARC que você pode implementar com base em seus requisitos. Alguns desses padrões são descritos em Criando mecanismos de recuperação de desastres usando o Route 53 e na seção de disjuntores de verificação de integridade de padrões avançados de resiliência Multi-AZ. Se você optou por usar um plano de DR multirregional, pré-provisione recursos que exigem a criação de registros DNS, como ELBs e instâncias de RDS. Um non-statically-stable projeto seria atualizar o valor de um registro de recurso do Route 53 por meio da `ChangeResourceRecordSets` API, alterar o peso de um registro ponderado ou criar novos registros para realizar o failover. Essas abordagens dependem do plano de controle da Rota 53.

Amazon CloudFront

O plano CloudFront de controle da Amazon consiste em todas as CloudFront APIs públicas para gerenciar distribuições e está hospedado em us-east-1. O plano de dados é a distribuição em si servida pela PoPs rede periférica. Ele executa o tratamento da solicitação, o roteamento e o armazenamento em cache do seu conteúdo de origem. Durante uma falha no plano de controle, as operações do tipo CRUDL para CloudFront (incluindo solicitações de invalidação) podem não funcionar, mas seu conteúdo continuará sendo armazenado em cache e servido, e os failovers de origem continuarão funcionando.

O que isso significa é que, ao planejar dependências doCloudFront, você não deve confiar no plano de CloudFront controle em seu caminho de recuperação. Por exemplo, um projeto estaticamente estável seria usar failovers de origem automatizados para mitigar o impacto de uma deficiência em uma de suas origens. Você também pode optar por criar balanceamento de carga de origem ou failover usando o Lamda @Edge. Consulte Três padrões de design avançados para aplicativos de alta disponibilidade usando a Amazon CloudFront e o Amazon S3 para criar aplicativos de geoproximidade ativos CloudFront e ativos em várias regiões para obter mais detalhes sobre esse padrão. Um non-statically-stable projeto seria atualizar manualmente a configuração de sua distribuição em resposta a uma falha de origem. Essa abordagem dependeria do plano CloudFront de controle.

AWS Certificate Manager

Se você estiver usando certificados personalizados com sua CloudFront distribuição, também dependerá do ACM. Usar certificados personalizados com sua CloudFront distribuição depende do ambiente de gerenciamento da ACM na região us-east-1. Durante uma falha no plano de controle, seus certificados existentes configurados em sua distribuição continuarão funcionando, assim como as renovações automáticas de certificados. Não confie em alterar a configuração da distribuição ou criar novos certificados como parte do seu caminho de recuperação.

AWSFirewall de aplicativos Web (WAF) e WAF Classic

Se você estiver usando AWS WAF com sua CloudFront distribuição, você depende do plano de controle WAF, que também está hospedado na região us-east-1. Durante uma falha no plano de controle, as listas de controle de acesso à web (ACLs) configuradas e suas regras associadas continuam funcionando. Não confie na atualização de suas ACLs web do WAF como parte de seu caminho de recuperação.

AWS Global Accelerator

O plano de controle AGA consiste em todas as APIs públicas da AGA e está hospedado em us-west-2. O plano de dados é o roteamento de rede dos endereços IP anycast fornecidos pela AGA para seus endpoints registrados. O AGA também utiliza verificações de saúde do Route 53 para determinar a integridade de seus endpoints AGA, que fazem parte do plano de dados do Route 53. Durante uma falha no plano de controle, as operações do tipo CRUDL para AGA podem não funcionar. O roteamento para seus endpoints existentes, bem como as verificações de integridade, marcadores de tráfego e configurações de peso de terminais existentes usadas para rotear ou transferir o tráfego para outros endpoints e grupos de endpoints, continuarão funcionando.

O que isso significa é que, ao planejar dependências do AGA, você não deve confiar no plano de controle do AGA em seu caminho de recuperação. Por exemplo, um design estaticamente estável seria usar o status das verificações de saúde configuradas para evitar falhas em endpoints não íntegros. Consulte Implantação de aplicativos multirregionais AWS usando o AWS Global Accelerator para obter exemplos dessa configuração. Um non-statically-stable projeto seria modificar as porcentagens de discagem de tráfego AGA, editar grupos de endpoints ou remover um endpoint de um grupo de endpoints durante uma deficiência. Essas abordagens dependeriam do plano de controle da AGA.

Amazon Shield Advanced

O plano de controle Amazon Shield Advanced consiste em todas as APIs públicas do Shield Advanced e está hospedado em us-east-1. Isso inclui funcionalidades como `CreateProtectionCreateProtectionGroup`, `AssociateHealthCheckDesribeDRTAccess`,, `ListProtections` e. O plano de dados é a proteção contra DDoS fornecida pelo Shield Advanced, bem como a criação das métricas do Shield Advanced. O Shield Advanced também utiliza verificações de integridade do Route 53 (que fazem parte do plano de dados do Route 53), se você as tiver configurado. Durante uma falha no plano de controle, as operações do tipo CRUDL do Shield Advanced podem não funcionar, mas a proteção contra DDoS configurada para seus recursos, bem como as respostas às mudanças nas verificações de saúde, continuarão funcionando.

O que isso significa é que você não deve confiar no plano de controle Shield Advanced em seu caminho de recuperação. Embora o plano de controle Shield Advanced não forneça a funcionalidade direta que você normalmente usaria em uma situação de recuperação, pode haver momentos em que você o faria. Por exemplo, um design estaticamente estável seria ter seus recursos de DR já configurados para fazer parte de um grupo de proteção e ter verificações de integridade associadas

a eles, em vez de configurar essa proteção após a ocorrência da falha. Isso evita depender do plano de controle Shield Advanced para recuperação.

Apêndice C - Serviços de região única

A seguir está uma lista de serviços ou recursos específicos desse serviço (listados entre parênteses após o nome do serviço), que só estão disponíveis em uma única região. A mesma orientação para implementar a estabilidade estática fornecida para outros serviços globais se aplica a esses serviços quando você precisa planejar dependências em seus planos de controle e planos de dados.

- Alexa for Business
- AWS Marketplace(API AWS Marketplace de catálogo, análise de AWS Marketplace comércio, AWS Marketplace serviço de direitos)
- Billing and Cost Management (AWS Cost Explorerrelatórios de AWS custo e uso, AWS orçamentos, Savings Plans)
- AWS BugBust
- Amazon Mechanical Turk
- Amazon Chime
- Amazon Chime SDK (áudio, mensagens e identidade PSTN)
- AWSChatbot
- AWS DeepRacer
- AWSDevice Farm
- Amazon GameSparks
- Amazon Honeycode

Contribuidores

Os colaboradores deste documento incluem:

- Michael Haken, principal arquiteto de soluções, Amazon Web Services

Revisões do documento

Para ser notificado sobre atualizações dessa documentação técnica, inscreva-se no feed RSS.

Alteração	Descrição	Data
Revisão menor	Guia atualizado para alinhamento com as práticas recomendadas do IAM. Para obter mais informações, consulte Práticas recomendadas de segurança no IAM.	9 de fevereiro de 2023
Publicação inicial	Publicação no whitepaper.	16 de novembro de 2022

AWS Glossário

Para obter a terminologia mais recente da AWS, consulte o glossário da AWS na Referência do Glossário da AWS.

Avisos

Os clientes são responsáveis por fazer sua própria avaliação independente das informações contidas neste documento. Este documento: (a) é apenas para fins informativos, (b) representa as ofertas e práticas atuais de AWS produtos, que estão sujeitas a alterações sem aviso prévio, e (c) não cria nenhum compromisso ou garantia de suas afiliadas, fornecedores ou AWS licenciadores. AWSos produtos ou serviços são fornecidos “como estão” sem garantias, representações ou condições de qualquer tipo, expressas ou implícitas. As responsabilidades e obrigações de AWS seus clientes são controladas por AWS acordos, e este documento não faz parte, nem modifica, nenhum acordo AWS entre seus clientes.

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.