CLOUDFLARE

ARTIGO TÉCNICO

Resiliência da rede e dos serviços da Cloudflare

Conteúdo

Visão geral

A internet é construída sobre sistemas imperfeitos que são projetados para priorizar o tempo de atividade em detrimento de tudo o mais. Protocolos como TCP e BGP são baseados nos princípios de sistemas distribuídos, esperar falhas e lidar com elas, e a internet foi projetada de acordo com isso. No entanto, os pontos de falha ainda existem e podem causar impacto. Os provedores de rede precisam ser capazes de planejar para falhas, detectá-las quando ocorrerem e mitigar o impacto experimentado pelos usuários.

Devido à natureza em tempo real de muitos aplicativos na internet, as redes precisam não apenas cobrir a maior parte da internet possível, mas também responder e mitigar o impacto em tempo real. Tempos de inatividade da ordem de minutos não são aceitáveis: os clientes esperam uma correção em poucos segundos.

A Cloudflare oferece serviços de infraestrutura de rede crítica para apoiar a forma como as organizações protegem e se comunicam pela internet. Desenvolvemos nossa rede e os serviços que ela fornece para manter o mais alto nível de excelência operacional. A Cloudflare processa, em média, mais de 84 milhões de solicitações HTTP e 61 milhões de consultas de DNS por segundo, fornecendo serviços a milhões de ativos da internet e usuários.

Como a Cloudflare fornece serviços confiáveis nesta escala, dadas as características imprevisíveis da internet?

A resposta vem da arquitetura da rede da Cloudflare, que opera com recursos de resiliência projetados para atuar de forma independente e resistir a todo tipo de interrupção. Nossos recursos de computação, rede e armazenamento, juntamente com nossos processos operacionais, são projetados para tornar a Cloudflare tão confiável quanto o tom de discagem da antiga rede de serviço telefônico. A Cloudflare oferece o metafórico "tom de nuvem" para os serviços de rede e segurança dos quais os clientes dependem, independentemente das condições da internet em qualquer momento.

Este artigo técnico se aprofunda nos desafios de operar em tempo real na internet e como a Cloudflare está posicionada de forma única para resolver esses desafios.

Viver em um mundo imperfeito

A internet é um lugar imperfeito, ainda assim as organizações precisam dela para desenvolver seus negócios e administrar organizações que vinculam usuários, dados e dispositivos distribuídos a aplicativos em nuvem. Qualquer interrupção do serviço tem implicações graves. Ao longo dos anos, a Cloudflare relatou várias grandes interrupções de serviços de internet em todo o mundo, desde acidentes e ataques de DDoS a desastres naturais e muito mais.

A internet é uma rede diversificada, construída como um coletivo pouco interligado de milhares de redes participantes de vários tamanhos e capacidades. Essas redes têm diferentes níveis de serviço, algumas operando da melhor forma possível como uma extensão de boa vontade, e outras operando como parte de seus serviços comerciais. Por meio de acordos mútuos para troca de tráfego, que pode ou não ser destinado a hosts em suas próprias redes, a internet funciona como uma estrutura global através da participação benevolente de provedores de serviços e trocas de internet regionais e locais.

No entanto, com essa diversidade vem um grau de imprevisibilidade. O caminho que um determinado pacote percorre depende de uma sequência de melhores tentativas e de melhores esforços para a entrega. Na ausência de dados em tempo real ou modelagem preditiva sobre as condições reais da rede em um determinado momento, um pacote normalmente dá seu próximo salto usando rotas padrão ou o caminho mais curto presumido. Nenhuma dessas opções leva em conta o estado do serviço de rede em suas decisões de roteamento. Isso significa que os pacotes estão frequentemente sujeitos a uma série de condições debilitantes:

- No nível mais básico, as redes podem enfrentar falhas temporárias e degradações que reduzem a taxa de transferência e causam perda de pacotes.
- À medida que as redes ficam saturadas, o congestionamento prejudica o desempenho e a experiência dos usuários.
- Embora esses tipos de lentidão possam ocorrer em condições normais de operação, um número crescente de interrupções é causado intencionalmente por agentes hostis que consomem os recursos disponíveis de forma maliciosa.

Na Cloudflare, nossa missão é fazer a nossa parte para construir uma internet melhor. Para esse fim, criamos uma infraestrutura para tornar a internet mais rápida, mais confiável e mais segura. Isso é possível graças aos serviços da e através da rede da Cloudflare, uma das maiores redes do mundo que opera em nossos próprios servidores bare metal (sem virtualização), um backbone privado e uma enorme presença global que pode ser alcançada, em média, em 50 ms por 95% da população mundial.[1]

Os clientes geralmente querem entender como isso é feito. Como a Cloudflare projeta serviços nessa escala, sem comprometer a segurança e o desempenho e com a confiabilidade de que as organizações precisam? A resposta vem de um conjunto de objetivos norteadores, que definem como construímos nossa infraestrutura para que permaneça eficiente apesar das forças disruptivas, tanto externas quanto internas.

Como a Cloudflare projeta a resiliência

Muitas organizações se concentram em melhorar a disponibilidade tentando reagir de forma mais eficaz ou mais rápida às falhas. Isso requer investimento e testes constantes em seus recursos e processos de failover. Embora esses sejam objetivos válidos, a Cloudflare pensa de forma diferente, nossas equipes de engenharia de resiliência dedicam um esforço enorme para reduzir os cenários em que uma resposta de recuperação de desastres é necessária.

A engenharia de resiliência da Cloudflare começa com uma premissa simples: **Como você criaria uma infraestrutura crítica que permanecesse operacional presumindo que falhas vão acontecer?**

Quando as falhas são inevitáveis, os serviços resilientes da Cloudflare detectam e isolam as falhas para que não afetem a disponibilidade do serviço. As falhas são resolvidas fora de banda a partir da nossa entrega de serviços. Nosso objetivo é sermos resistentes a falhas em toda a nossa frota.

Para entender os princípios de resiliência, é útil definir primeiro os principais conceitos por trás dos caminhos de tráfego com a Cloudflare e os objetivos de design dos principais sistemas. No nível mais abstrato, a arquitetura da Cloudflare pode ser separada em dois segmentos: **o plano de controle e o plano de dados.** Cada um tem uma postura única em relação a resiliência e desastres.

Resiliência do plano de controle

O plano de controle fornece a interface de gerenciamento que estabelece a fonte da verdade para a configuração dos serviços de rede e segurança no ambiente do cliente. O plano de controle em si não processa tráfego (que é a função do plano de dados). Ele informa ao plano de dados quais políticas aplicar e gerencia as configurações em diferentes data centers.

Os serviços do plano de controle da Cloudflare geralmente são implantados em uma topografia mais tradicional e centralizada em três data centers logicamente relacionados, mas independentes, em uma região primária (por exemplo, EUA). Esses três data centers são replicados com capacidade equivalente em uma região secundária (por exemplo, UE). Os serviços do plano de controle são projetados para serem resilientes e manter uma entrega de serviços consistente no caso de falha de qualquer um dos data centers no local principal. A perda de data centers adicionais no local principal causaria um failover para os data centers de outra região (por exemplo, na Europa em vez de nos EUA).

De acordo com o foco da Cloudflare na resiliência antes da recuperação, continuamos a investir no aprimoramento de nossa postura de resiliência.

Por exemplo:

- Estamos cada vez mais utilizando as duas regiões do plano de controle em uma configuração ativo-ativo, o que aumenta simultaneamente nossa capacidade/responsividade, além de nossa tolerância a falhas. Consequentemente, podemos resistir a mais tipos de falhas sem interrupção do serviço ou necessidade de failover.
- Também estamos aumentando a granularidade da forma como transferimos serviços entre os vários sites do plano de controle, o que nos permite responder com mais precisão aos problemas de infraestrutura local.

Testes de caos

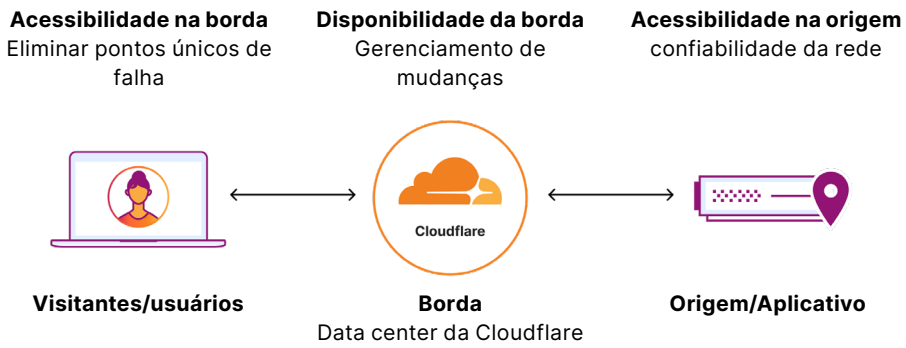
A resiliência não é uma atividade do tipo "configurar e esquecer". Mesmo os planos de resiliência mais robustos podem se revelar ineficazes devido ao "desvio" do sistema, o acúmulo lento e muitas vezes imperceptível de mudanças que pode degradar os comportamentos pretendidos e introduzir modos de falha imprevistos. Para abordar proativamente esse risco, são realizados testes de caos na Cloudflare regularmente, sondando sistematicamente possíveis vulnerabilidades relacionadas ao desvio.

Resiliência do plano de dados

O plano de dados processa o tráfego dos clientes da Cloudflare de acordo com as políticas estabelecidas no plano de controle. Embora os serviços do plano de dados recebam a orientação do plano de controle, eles não dependem do plano de controle para operar. Todas as políticas são mantidas através do Quicksilver, nosso armazenamento de chave-valor distribuído globalmente, para garantir que os serviços permaneçam operacionais com uma configuração válida conhecida em caso de qualquer interrupção de comunicação com o plano de controle.

O Anycast desempenha um papel importante na redundância dos data centers. Os data centers da Cloudflare, localizados em mais de 330 cidades, são localmente autônomos e, ainda assim, intercambiáveis entre si por meio de Anycast e BGP. Isso ocorre porque cada data center pode processar qualquer serviço localmente, sem ser codependente de serviços em outro data center. Com o Anycast, cada data center é efetivamente redundante com os demais. Como todos os data centers participam do Anycast, não há necessidade de instruir o cliente a mudar para um data center alternativo em outro endereço de IP.

Seja um consumidor visitando um site protegido pela Cloudflare, um funcionário acessando aplicativos conectados à internet ou um escritório conectando-se à sua WAN, todos esses cenários usam o BGP para encontrar o data center Anycast da Cloudflare mais próximo. Se o data center escolhido ficar indisponível, o BGP resolverá automaticamente para o próximo melhor data center da Cloudflare.



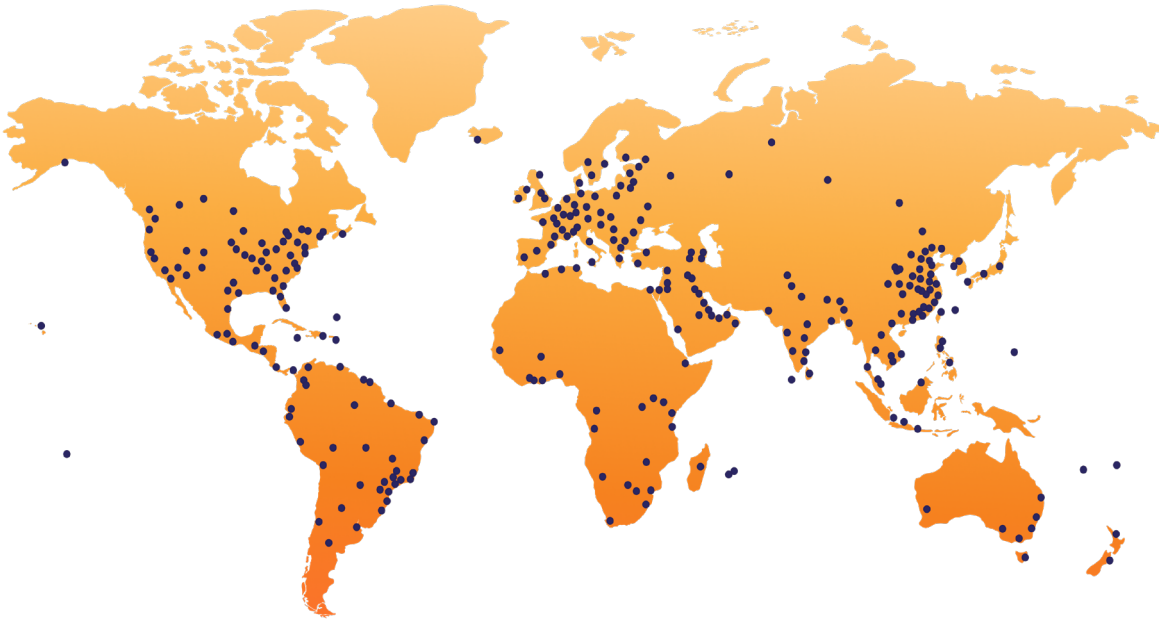
A Cloudflare aborda a resiliência do plano de dados resolvendo três problemas diferentes:

- Acessibilidade na borda: garantir que o tráfego do usuário final possa chegar aos data centers e eliminar pontos únicos de falha na entrada de tráfego
- Disponibilidade da borda: manter a qualidade do código e o tempo de atividade do software por meio de um gerenciamento de mudanças rigoroso
- Acessibilidade na origem: roteamento adaptável para aplicativos do cliente para garantir que não haja perda nos caminhos de saída

Acessibilidade na borda

A acessibilidade na borda é a capacidade dos usuários finais de alcançarem a rede da Cloudflare. É, sem dúvida, a parte mais importante no contexto do problema da resiliência. Se os provedores de internet (ISPs) ou os data centers forem desativados, a acessibilidade na borda será reduzida ou degradada, o que impede ou retarda os usuários de chegarem aonde precisam estar na internet. A Cloudflare aborda problemas de acessibilidade na borda de quatro maneiras principais:

1. Rede Anycast



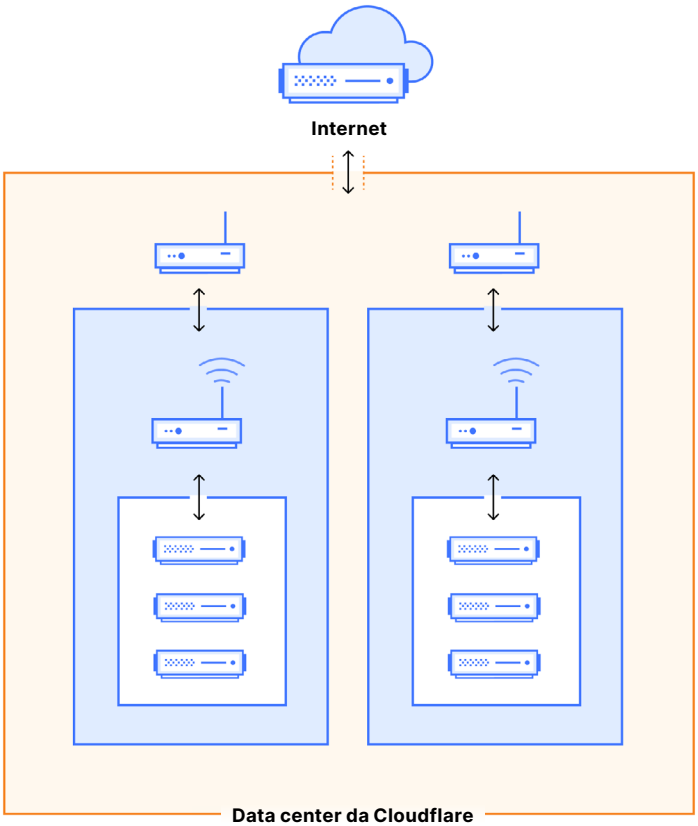
Um melhor desempenho de rede com resiliência está integrado à nossa arquitetura de rede, que depende fortemente da tecnologia Anycast. O Anycast, o superpoder de engenharia da Cloudflare, significa que o espaço de IP é anunciado em todos os lugares: se algum de nossos pontos de presença (PoPs) estiver off-line, o tráfego simplesmente será roteado para outros locais em vez de ser descartado. O fato de a Cloudflare estar presente em tantas cidades no mundo e fazer peering com redes locais significa que processamos o tráfego de clientes o mais próximo possível do usuário. Por exemplo, mesmo que um trânsito de provedor de internet seja desconectado ou que um data center perca energia, o tráfego permanece inalterado.

2. Todos os serviços da Cloudflare são executados em todos os locais.

Além do Anycast, os data centers da Cloudflare são projetados para processar o tráfego localmente, sem depender de cadeias de proxy para outros recursos de computação. Executamos quase todos os serviços em todas as máquinas. Isso significa que um data center pode ser colocado off-line sem impacto para o cliente. As máquinas dentro do data center podem ser substituídas umas pelas outras porque há centenas delas executando o mesmo serviço e capazes de entrar em ação caso uma delas apresente falha.

3. PoPs multi-colo

O design e os locais dos data centers da Cloudflare refletem as necessidades de nossos clientes. Uma rede Anycast nos permite adicionar/remover data centers à vontade, mas devemos monitorar constantemente o desempenho do cliente. Adaptamos as topologias de nossos data centers para permitir que seções de capacidade computacional (colos) falhem independentemente umas das outras. Esses data centers (com várias colos) são chamados de pontos de presença multi-colo, ou locais de MCP.



Esses locais separam a conectividade voltada para a internet da conectividade de computação interna para permitir que as colos sejam desativadas individualmente. Isso significa que, mesmo que haja um problema com uma colo, todo o PoP em uma região pode permanecer on-line, proporcionando maior tempo de atividade e desempenho ao cliente. Esse tipo de data center também remove pontos únicos de falha ao ter dispositivos redundantes na camada voltada para a internet: se um roteador voltado para a internet (de borda) falhar, o outro roteador de borda pode receber o tráfego e garantir que o local permaneça operacional.

Esse modelo operacional ajuda os locais de MCP a evitar o desvio do tráfego, a menos que seja absolutamente necessário, e, assim, aumenta ainda mais o tempo de atividade dos clientes da Cloudflare.

4. Cloudflare Traffic Manager

Os data centers MCP trabalham juntos para formar a rede da Cloudflare. Essa rede utiliza Anycast para ajudar a garantir que o tráfego do cliente seja atendido. O Anycast é aprimorado com gerenciamento de tráfego determinístico para garantir que as solicitações dos clientes sejam atendidas onde pudermos atendê-las, com o melhor desempenho possível. Esse sistema de gerenciamento de tráfego, o Traffic Manager, funciona examinando continuamente a rede da Cloudflare e desviando automaticamente o tráfego para longe de data centers que apresentam sobrecarga de CPU. Isso evita congestionamentos em data centers com alto tráfego; em vez disso, o tráfego é roteado de forma inteligente para outro data center capaz de processá-lo.

Disponibilidade da borda

A disponibilidade da borda refere-se à capacidade da Cloudflare de processar o tráfego assim que ele chega à nossa rede. Quando alterações nas ferramentas ou no software de rede resultam em alterações não intencionais, a disponibilidade pode diminuir e afetar a experiência dos usuários. Para evitar que incidentes resultantes de mudança de código aconteçam, a Cloudflare investiu fortemente nos seguintes controles de implantação:

1. Funil de implantação

Ao implantar software, garantir a qualidade do código começa com o monitoramento e a limitação da capacidade de desenvolvedores e clientes introduzirem mudanças no ecossistema. A Cloudflare limita o número de maneiras pelas quais qualquer pessoa pode introduzir mudanças em nossa infraestrutura para que possamos monitorar de perto cada alteração e garantir que elas passem por uma bateria de testes antes de serem implantadas em produção.

2. Gerenciamento de alterações do raio de alcance de impacto

Outra maneira pela qual a Cloudflare oferece suporte à disponibilidade da borda durante a implantação é limitando as implantações a data centers de teste, ou grupos de teste, antes de implementá-las amplamente. Nós nos referimos a isso como gerenciamento do raio de impacto.

Quando as alterações na rede são implementadas, elas podem ser ativadas globalmente em minutos. Ao conter as alterações em um ambiente de implantação e implementar outras alterações em cascata, podemos monitorar os efeitos da mudança identificando consequências intencionais ou não, antes de afetar áreas geográficas ou populações de usuários maiores.

Temos duas maneiras de limitar o impacto das alterações de código:

- Limitar o número de locais que recebem alterações; e
- Limitar o número de usuários que recebem alterações

Ao limitar o número de locais e máquinas que recebem alterações, podemos garantir que estamos realizando testes A/B adequados no código dentro de um único local para avaliar sua integridade antes de prosseguir. Ao limitar o número de usuários que recebem alterações, podemos testar as correções primeiro em grupos menores de usuários.

3. Implantação mediada pela integridade

A implantação mediada pela integridade é um sistema que avalia programaticamente a adequação de uma versão com base em métricas predefinidas que emitem um sinal de "aprovar" ou "reprovar" com base no possível impacto. Essa série de verificações automatizadas pode não apenas evitar a liberação de uma versão prejudicial, mas também pode reverter uma versão ao detectar o impacto.

Cada produto e serviço implantado via Cloudflare deve ter um objetivo de nível de serviço (SLO), que contém uma métrica que representa a integridade do produto e também uma meta abaixo da qual um produto seria considerado não íntegro.

Os SLOs têm taxas de consumo, ou limiares aceitáveis de falha. Qualquer serviço com monitoramento de integridade fornecerá SLOs para um sistema automatizado como parte da incorporação de uma alteração a ser implantada. Em cada escopo de implantação definido (planos gratuitos, um subconjunto de máquinas em Ashburn etc.), o sistema automatizado:

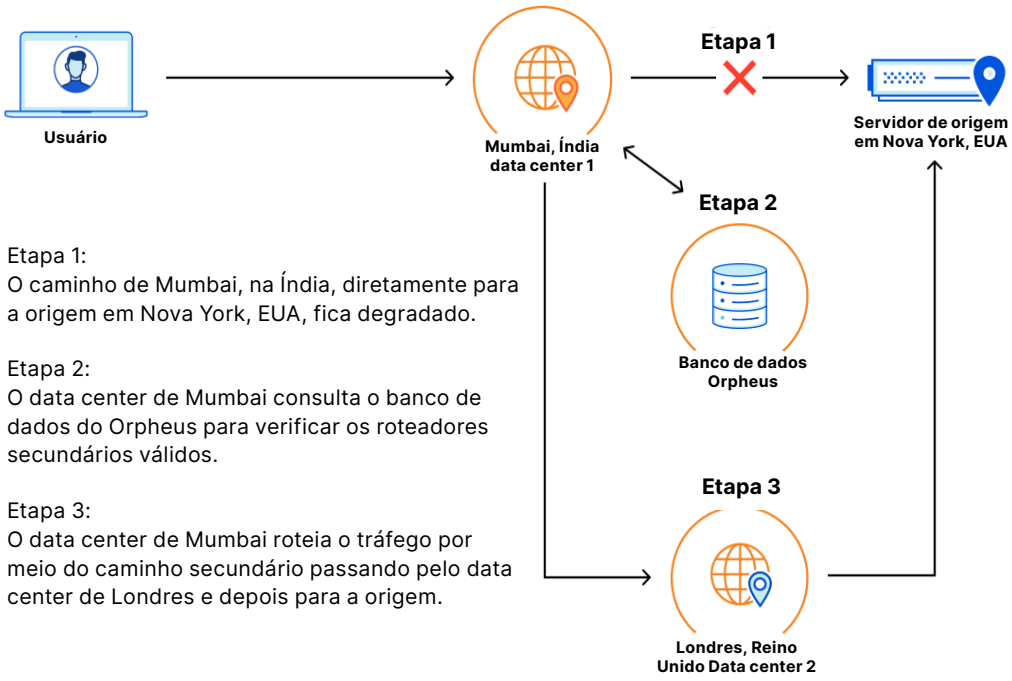
- Primeiro, monitora o SLO do serviço por um período de tempo definido para garantir que a integridade não fique abaixo do limite.
- Se o SLO permanecer dentro dos intervalos aceitáveis durante o período de estabilização definido, o sistema avançará automaticamente a implantação para um estágio maior.
- No entanto, se o SLO for violado, o sistema interrompe automaticamente a implantação e reverte para que o impacto seja mitigado automaticamente.

Essas etapas garantem que a integridade do cliente não seja afetada por implantações de código e que a duração seja a mais curta possível.

Acessibilidade na origem

A capacidade de acessibilidade na origem refere-se à habilidade da Cloudflare de chegar a destinos, seja a origem de um cliente, um site SaaS ou a internet pública por meio do Cloudflare Gateway. O roteamento de solicitações para onde elas precisam estar é crucial para os usuários que acessam a rede da Cloudflare. Por exemplo, o Roteamento inteligente Argo, ferramenta da Cloudflare que otimiza o desempenho (ou seja, tempo decorrido até o primeiro byte), examina constantemente a rede da Cloudflare para encontrar o caminho mais rápido para as origens. O Argo examina constantemente a rede da Cloudflare, encontrando o caminho mais rápido para as origens.

O Orpheus, complemento do Argo, tem uma filosofia semelhante, mas uma função diferente. O Orpheus existe para estabelecer conexões confiáveis com os servidores de origem. O Orpheus analisa especificamente as métricas que afetam a capacidade da Cloudflare de alcançar a origem (em oposição ao caminho mais rápido para a origem) e encontra caminhos que minimizam a perda de pacotes sem afetar o desempenho do estado estacionário. Isso significa que, quando surgem problemas, o tráfego é roteado automaticamente para contornar os erros detectados.



Etapa 1:
O caminho de Mumbai, na Índia, diretamente para a origem em Nova York, EUA, fica degradado.

Etapa 2:
O data center de Mumbai consulta o banco de dados do Orpheus para verificar os roteadores secundários válidos.

Etapa 3:
O data center de Mumbai roteia o tráfego por meio do caminho secundário passando pelo data center de Londres e depois para a origem.

Antes de a Cloudflare lançar o Orpheus em 2021, conseguíamos rotear com sucesso para as origens em 99,9% das vezes. Depois de implementar o Orpheus, nossa capacidade de rotear para as origens aumentou para 99,99%. No próximo ano, vamos expandir o Orpheus para proteger mais tipos de tráfego, agir em mais cenários de falha e trabalhar mais rápido para reduzir o tempo em que qualquer usuário é afetado.

Compromisso com a transparência operacional

Mesmo as redes mais resilientes e inovadoras irão sofrer interrupções. Quando incidentes surgem e os clientes são afetados, a Cloudflare segue uma resposta de comunicação de incidentes que inclui uma investigação minuciosa, um relatório de incidente interno, um relatório de incidente externo e, se necessário, atualizações de status em toda a janela de impacto.

Em certos casos em que os incidentes resultem em tal impacto, ou inovação, as análises pós-incidente serão publicadas no Blog da Cloudflare.

Conclusão

O esforço de engenharia por trás da rede da Cloudflare não é um trabalho fácil, mas é um trabalho que orgulhosamente fazemos pelos nossos clientes. A recompensa é construir uma plataforma de rede que beneficia nossos clientes e a comunidade mais ampla da internet como um todo.

Em última análise, ao priorizar a resiliência não apenas como uma preocupação técnica, mas como uma filosofia operacional essencial, estamos fazendo mais do que apenas reforçar defesas: estamos construindo ativamente uma empresa que está inerentemente pronta para o futuro. A abordagem proativa de testar e adaptar continuamente significa que podemos evoluir facilmente junto com as demandas em constante mudança de nossos clientes e do cenário dinâmico da própria internet.

Entendemos que muitos dos conceitos representados neste documento giram em torno de conceitos de rede aos quais muitas empresas não estão expostas, pois envolvem a arquitetura interna de operação de um ambiente de nuvem global de classe de operadora.

Se quiser um resumo aprofundado para saber mais sobre a engenharia de resiliência da Cloudflare, entre em contato com seu representante da Cloudflare.

+55 (11) 3230 4523 | enterprise@cloudflare.com | www.cloudflare.com/pt-br/

REV:BDES-8251.2025OCT28