Um desafio de capacidade iminente para ISPs de Nível 1 (Tier One)

Visão geral

Praticamente todos os ISPs de nível 1 (tier one) operam infraestruturas de rede IP amplas, de alta capacidade e altamente resilientes, capazes de entregar serviços de banda larga gigabit simétricos para dezenas de milhões de residências. Embora os níveis de tráfego tenham aumentado de forma constante na última década, os padrões de tráfego permaneceram (até recentemente) notavelmente consistentes. Como era de se esperar, esses padrões mapeiam fielmente as rotinas de estudo e trabalho de seus assinantes. A rede fica relativamente ociosa durante o dia (mesmo com o aumento do trabalho remoto e das consequentes reuniões via Zoom), seguida por picos no horário nobre e, para os jogadores de videogame, estendendo-se até tarde da noite. No entanto, o surgimento e o crescimento contínuo do streaming over-the-top (OTT) ao vivo — especialmente de esportes — interrompeu essa consistência e bagunçou completamente o cálculo de engenharia de tráfego de todos os grandes ISPs.

A transmissão ao vivo de grandes eventos esportivos, como jogos da NFL, pode gerar picos de tráfego de 10 vezes ou mais em comparação ao que os provedores de internet (ISPs) normalmente veriam em qualquer outro dia. Para jogos de playoffs, o Super Bowl ou qualquer partida com a presença de Taylor Swift, esses picos podem exceder 30 vezes o normal. A questão para os provedores é: “devemos investir em capacidade suficiente para lidar com picos que ocorrem com relativa raridade?” É uma faca de dois gumes: ou eles investem em quantidades massivas de capacidade (cara) que fica ociosa na maior parte do tempo, ou correm o risco de perder clientes (churn).

Como chegamos a este ponto?

Embora muita atenção tenha se concentrado nas dezenas de bilhões de dólares gastos pelas grandes plataformas de streaming em pacotes de direitos esportivos, existe um problema técnico latente no cerne de cada plataforma que está criando esse impasse. Sim, o dinheiro, somado ao contínuo movimento de cancelamento de TV a cabo (cord-cutting) e de pessoas que nunca assinaram esses serviços (cord-nevering), criou uma enorme gama de conteúdo da qual os assinantes de banda larga podem desfrutar. Como exemplo, a Amazon Prime Video e a NBCU/Peacock estiveram entre os vencedores em um recente acordo de direitos de 11 anos com a NBA, no valor de 77 bilhões de dólares. E isso é apenas a ponta do iceberg.

Aqui está o problema. Do ponto de vista técnico, o principal mecanismo de entrega de camada três no cerne de toda plataforma de streaming de vídeo é o mesmo: o IP unicast. É compreensível por que as plataformas de streaming optaram por esse mecanismo: elas precisam garantir que podem alcançar qualquer dispositivo usado por qualquer assinante em qualquer rede, e o menor denominador comum é o IP unicast ponto a ponto. Mas, do ponto de vista da rede, seria difícil encontrar uma maneira menos eficiente de entregar streaming de vídeo ao vivo para os assinantes.

Quase todas as tecnologias de entrega de vídeo usadas historicamente — transmissão aberta (broadcast), cabo, satélite, IPTV — utilizaram tecnologias ponto-multiponto (ou seja, multicast) infinitamente mais eficientes em termos de largura de banda. Possivelmente pela primeira vez na história, a indústria como um todo está se movendo intencionalmente em direção à ineficiência.

Para o vídeo sob demanda (VOD), o IP unicast não é um grande problema. Como a sessão de cada assinante é absolutamente única, o unicast é a única forma de viabilizar o streaming de VOD via OTT. No entanto, para conteúdos ao vivo — especialmente esportes — isso se torna um problema enorme. Cada dispositivo de visualização ativo em cada residência de assinantes exige uma sessão de IP

exclusiva. Multiplique isso por uma taxa de bits média de, digamos, 5 Mbps e você terá eventos de grande escala, como jogos da NFL, consumindo centenas de terabits por segundo da capacidade total da rede. Próximo ao núcleo (core), as CDNs ajudam a amenizar esse problema, mas, para provedores de internet (ISPs) de todos os tamanhos, a maior parte de sua infraestrutura de rede fica sobrecarregada em dias de jogos e é utilizada de forma ineficiente no restante do tempo.

O papel das CDNs tradicionais

As Redes de Entrega de Conteúdo (CDNs) são utilizadas há muito tempo para armazenar conteúdo em cache mais próximo do usuário final, a fim de melhorar tanto a qualidade de experiência (QoE) quanto a eficiência da rede. Mais recentemente, as CDNs desenvolvidas para a entrega de streaming de vídeo adicionaram funções de replicação de conteúdo que as permitem imitar topologias multicast e aumentar a eficiência da rede.

Historicamente, os POPs (Points of Presence) das CDNs têm sido posicionados em locais de núcleo (core) maiores, de alta conectividade e capacidade, como data centers de internet, provedores de interconexão (interexchange) e pontos de entrada de rede para grandes ISPs. Embora isso ajude a reduzir ou eliminar as taxas de peering (troca de tráfego), pouco faz para conservar a largura de banda nas redes dos provedores. Como mostrado na Figura 1, os POPs das CDNs tendem a ser implantados logo acima (upstream) ou logo dentro da infraestrutura de rede do ISP. Alguns ISPs de nível 1 (tier one) podem ter implantado POPs de CDN de forma mais ampla, mas estes ainda tendem a estar concentrados em grandes localizações centrais.

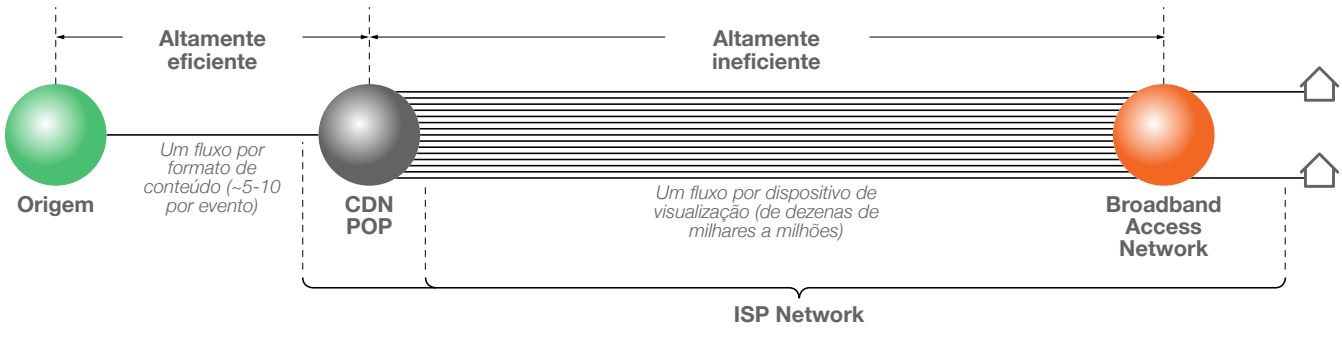
No que diz respeito aos ganhos de eficiência obtidos por meio da replicação de conteúdo, esses benefícios só são concretizados entre a origem do conteúdo e o ponto de presença (POP) ou cache da CDN. Nenhum ganho de eficiência de rede é percebido a jusante (downstream) do POP da CDN mais remoto. Isso representa um problema significativo para os ISPs de nível 1 (tier one), já que uma parcela muito grande de suas redes de núcleo e agregação está apenas transportando tráfego entre os caches da CDN e as redes de acesso de banda larga. Mesmo que os ISPs tier one sintam que possuem capacidade adequada para lidar com os picos associados ao streaming de esportes ao vivo, projetar algo visando a ineficiência seria, em algum momento, um bom princípio de design?

Quanto mais próximo da borda, melhor

De maneira geral, implantar caches de CDN mais próximos do assinante resulta em uma utilização mais eficiente da largura de banda da rede. Novamente, isso é especialmente válido para a transmissão ao vivo de grandes eventos, como esportes. No entanto, no passado, houve limites para o quão perto do assinante esses caches poderiam ser instalados. Existem dois fatores limitantes principais que os engenheiros de rede devem levar em consideração:

- **Eficiência do cache:** À medida que os caches se aproximam do perímetro absoluto da rede do ISP, por definição, eles atenderão a menos assinantes a jusante

1



POR QUE NÃO IP MULTICAST?

As plataformas de streaming de vídeo geralmente não utilizam o IP multicast por diversos motivos práticos, técnicos e comerciais. Embora o IP multicast seja um método teoricamente eficiente para entregar vídeo a muitos usuários simultaneamente, ele é, em grande parte, inadequado para a internet e os modelos de negócios atuais..

Falta de suporte na internet pública

- O IP multicast não possui suporte amplo na infraestrutura da internet pública (ou seja, em ISPs, roteadores e redes de backbone).
- A maioria dos roteadores na internet não está configurada para lidar com tráfego multicast, pois isso adiciona complexidade e riscos potenciais à segurança.

Desafios de complexidade e escalabilidade

- Gerenciar grupos multicast, adesão de membros e fluxo de tráfego é mais complexo do que no unicast.
- A infraestrutura de rede precisa coordenar a participação nos grupos (via protocolos IGMP e PIM), o que é difícil de realizar entre diferentes redes e provedores.

Segurança e controle de acesso

- Com o multicast, é difícil controlar quem recebe a transmissão — qualquer pessoa que se junte ao grupo pode, potencialmente, acessar os dados.
- O unicast permite que os serviços protejam as transmissões por usuário, apliquem a gestão de direitos digitais (DRM), gerenciem assinaturas e exibam anúncios direcionados.

Personalização e análise de dados (analytics)

- O streaming de vídeo moderno geralmente envolve o Adaptive Bitrate Streaming (ABR), conteúdo personalizado (ex: recomendações ou anúncios) e rastreamento de métricas por usuário. Isso requer sessões individuais, algo que o multicast não consegue suportar com eficiência..

A integração com CDNs funciona melhor com unicast

- As CDNs são otimizadas para tráfego unicast baseado em HTTP.
- Elas fazem o cache do conteúdo próximo aos usuários e gerenciam o balanceamento de carga, lógica de repetição (retry), entre outros.

(downstream). Especialmente para CDNs de propósito geral, a eficiência do cache é impulsionada pelas solicitações de conteúdo dos assinantes; quanto maior o número de solicitações, melhor será o desempenho dos algoritmos de cache da CDN ao coletar e armazenar o conteúdo correto. Com menos assinantes próximos à borda (edge), a eficiência do cache, ou “taxa de acerto” (hit rate), cai para níveis inaceitáveis para os caches convencionais.

- **Custo.** Outro fato inevitável é que, conforme os caches são implantados de forma mais profunda, o número de unidades dentro da rede do ISP aumenta. Em vez de poucos caches em grandes localidades da rede, os grandes ISPs podem ter dezenas de caches (embora menores) implantados na borda da rede. No total, isso elevará os custos gerais de hardware e conectividade da infraestrutura de cache. Eventualmente, esses custos podem exceder a economia gerada pela maior eficiência da largura de banda..

Por esses motivos, a maioria dos grandes ISPs que implantaram caches de CDNs de terceiros em suas redes os restringiu a um punhado de grandes localizações urbanas.

CACHES DE CDN DE PROPÓSITO GERAL VERSUS CACHES ESPECÍFICOS PARA VÍDEO

Os caches de propósito geral e os específicos para vídeo diferem principalmente em seus objetivos de design e nos tipos de dados que processam. Caches de propósito geral são projetados para armazenar temporariamente uma ampla gama de tipos de dados — como páginas da web, arquivos ou instruções de programas — para melhorar a velocidade de acesso e reduzir a latência em diversas tarefas computacionais. Esses caches focam na recuperação rápida e são otimizados para blocos de dados de pequeno a médio porte acessados com frequência.

Em contraste, os caches específicos para vídeo são especializados no manuseio de fluxos de dados grandes e contínuos associados à reprodução e ao streaming de vídeo. Esses caches são otimizados para gerenciar um alto rendimento (throughput) de dados e manter uma reprodução fluida, sem interrupções por buffering. Eles geralmente utilizam técnicas como pré-busca (prefetching) e armazenamento antecipado (buffering antes da reprodução), adaptadas aos padrões de acesso sequenciais e previsíveis do conteúdo de vídeo. Os caches específicos para vídeo também priorizam a minimização da latência e o gerenciamento eficiente de grandes arquivos de mídia, o que difere significativamente do foco em acesso aleatório dos caches de propósito geral. No geral, embora ambos os tipos de cache busquem aumentar o desempenho, os caches específicos para vídeo são ajustados com precisão para as demandas exclusivas da entrega de conteúdo multimídia.

Um caminho melhor: CDNs de múltiplas camadas com visibilidade na última milha

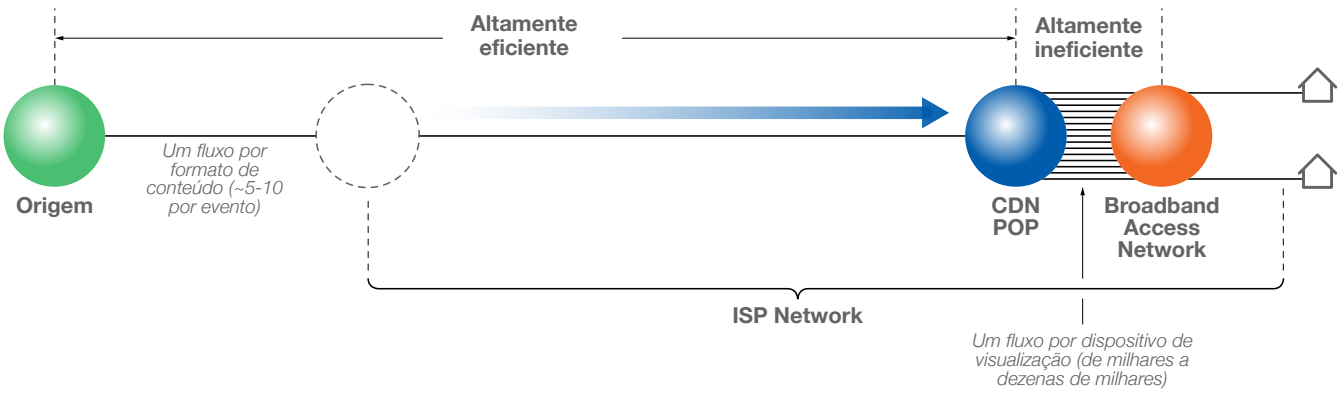
Claramente, a indústria se encontra em uma posição difícil de sustentar. Primeiro, não há como voltar atrás; o tráfego associado ao streaming OTT ao vivo continuará a crescer indefinidamente. Segundo, apesar dos esforços das plataformas de streaming para melhorar a tecnologia de compressão e entregar uma qualidade de conteúdo "apenas aceitável", seus espectadores investiram pesado em aparelhos 4K e 8K e esperam poder usá-los em todo o seu potencial. Isso impulsionará ainda mais o crescimento do tráfego. Os ISPs representam a rede através da qual 100% deste tráfego deve ser entregue. Isso significa que os grandes ISPs enfrentam a quase certeza de um crescimento contínuo na capacidade e nos custos de rede. Para piorar a situação, os ISPs são geralmente os primeiros a receber as reclamações dos clientes quando a qualidade do vídeo cai.

O que se exige é a ampliação do papel que os caches de CDN específicos para vídeo desempenham. Historicamente, eles têm sido o meio pelo qual as plataformas de streaming entregam seu conteúdo aos assinantes. O que é necessário agora são caches de CDN atuando como ferramentas de engenharia de capacidade para grandes operadoras de rede, especialmente ISPs.

Com isso em mente, a Netskrt desenvolveu um novo tipo de CDN, projetado especificamente para entregar streaming de vídeo de alta qualidade e, ao mesmo tempo, proporcionar ganhos de eficiência de rede para os ISPs. A CDN da Netskrt diferencia-se de qualquer outra CDN disponível atualmente em três aspectos importantes:

- **Consciência de conteúdo (Content aware):** Como uma CDN específica para vídeo, a CDN da Netskrt possui consciência de conteúdo. Dessa forma, conteúdos de vídeo (títulos, temporadas, anúncios) podem ser pré-posicionados nos dispositivos de cache, permitindo que eles alcancem altas taxas de acerto (cache hit ratios) em ambientes com populações de assinantes relativamente pequenas. Resultados do mundo real mostraram métricas de qualidade materialmente melhores para a Netskrt do que as alcançadas com CDNs convencionais.
- **Visibilidade na última milha:** A CDN da Netskrt é única por ser projetada para ser implantada primordialmente dentro das redes dos ISPs. Embora a CDN da Netskrt seja implantada em uma estrutura de múltiplas camadas, uma grande parte dessa capacidade é alocada na camada mais remota — a dos ISPs atendidos. Mesmo os locais de implantação a montante (upstream) têm visibilidade direta das redes dos ISPs e podem tomar decisões de entrega que otimizam tanto a utilização da capacidade quanto a qualidade da experiência do usuário final.
- **Capacidade de implantação profunda (Deep deployability):** Os caches da CDN da Netskrt estão disponíveis em formatos extremamente reduzidos e podem até ser implantados em infraestruturas existentes de bare metal ou baseadas em contêineres. Essa característica permite que eles sejam instalados mais próximos do assinante, maximizando os ganhos de eficiência de rede para os ISPs.

2



Como pode ser observado na Figura 2, a implantação de mais caches de CDN, porém menores, próximos ao perímetro da infraestrutura do ISP gera os maiores ganhos possíveis de eficiência de rede, ao mesmo tempo em que entrega uma QoE (Qualidade de Experiência) superior ao assinante. De fato, além da superdimensionamento da capacidade da rede, o cache altamente distribuído pode ser a única maneira de os grandes ISPs lidarem com o "tsunami" de tráfego de streaming ao vivo que inunda suas redes.

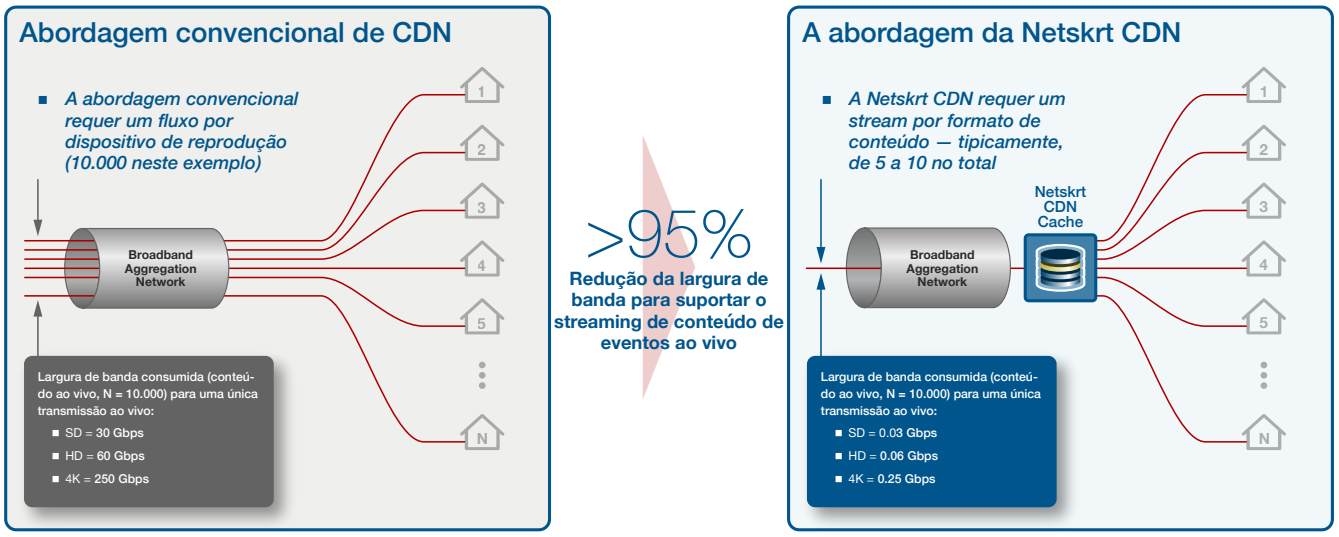
O caso de negócio para CDNs como ferramentas de engenharia de capacidade

Provedores de serviços de rede de todos os tipos são, basicamente, fornecedores de capacidade. Essa capacidade acarreta custos, tanto de capital (CAPEX) quanto operacionais (OPEX). O caso de uso básico descrito neste documento envolve a distribuição de dispositivos de cache para o perímetro de grandes redes de ISPs. Esse ponto geralmente se encontra no limite entre as instalações da rede de núcleo (core) e a rede de agregação de banda larga, embora diferentes ISPs possam adotar arquiteturas distintas. Na maioria dos casos, o objetivo do projeto é distribuir os caches de tal maneira que cada dispositivo atenda a populações de 100.000 a 250.000 assinantes. Este cenário reduzirá a quantidade de capacidade de rede necessária para atender aos requisitos de pico de tráfego. Essa redução na capacidade exigida diminuirá os custos; a única questão é em qual proporção.

A maioria das grandes plataformas de streaming utiliza estratégias de distribuição "multi-CDN", nas quais contratam múltiplas CDNs de terceiros e direcionam o tráfego com base na geografia e no desempenho em tempo real. Ao focar quase exclusivamente na capacidade de pico necessária para suportar o streaming de eventos ao vivo, a CDN da Netskrt demonstrou, em entregas reais de grandes eventos esportivos, reduzir a capacidade necessária em mais de 95% para o tráfego atribuído à Netskrt. Na maioria dos casos, para ASNs onde a Netskrt está implantada ou diretamente conectada, a porcentagem do tráfego total atribuído a ela varia entre 20% e 50%. É importante notar que a Netskrt recebe uma parcela desproporcional de tráfego das plataformas de streaming suportadas devido à sua "visibilidade na última milha" que, por sua vez, entrega métricas de qualidade consistentemente superiores.

Como exemplo, considere uma região de borda de um ISP tier one com 250.000 assinantes de banda larga a jusante. Suponha que, para um grande evento — por exemplo, um jogo de playoff da NFL — 40% das residências estejam assistindo ao jogo via streaming OTT e a taxa de bits média seja de 6 Mbps, o que equivale aproximadamente a

3



uma transmissão em HD. Sem a Netskrt implantada nesse ponto da rede do ISP, seriam necessários 600 Gbps de capacidade de pico nesse local. Com a Netskrt instalada, e assumindo que a plataforma de streaming direcione 50% do tráfego de assinantes aplicável para a Netskrt, o pico de capacidade cai para 300 Gbps.

A economia financeira resultante dessa redução na capacidade de pico só poderá ser calculada pelo próprio ISP. Em alguns casos, podem ser utilizados serviços de transporte baseados em uso, o que resultará em economias financeiras quase instantâneas. Em outros casos, o ISP pode ter projetado instalações de grande capacidade fixa, o que significa que as economias serão percebidas ao longo do tempo, conforme a necessidade de atualizar essas instalações seja reduzida. Em qualquer um dos casos, uma utilização mais eficiente da capacidade da rede sempre produzirá melhores resultados financeiros.

Resumo

O boom do streaming OTT, em grande parte arquitetado pelas plataformas de streaming, gerou uma infinidade de problemas imprevistos para as redes a jusante (downstream). Ao optarem, pelos motivos descritos acima, por utilizar conexões IP unicast para entregar conteúdo aos dispositivos de reprodução, as plataformas de streaming impuseram um fardo monumental aos ISPs. Basicamente, isso multiplicou a capacidade necessária para o streaming ao vivo pelo número de assinantes do ISP, em comparação com as tecnologias convencionais de transmissão de vídeo (broadcast). E embora o streaming ao vivo sempre tenha funcionado dessa maneira, foi apenas quando os grandes eventos esportivos começaram a ser transmitidos para dezenas de milhões de residências que a magnitude do problema tornou-se evidente.

Indiscutivelmente, o único mecanismo viável para que os grandes ISPs possam lidar com essa investida de tráfego é estender a infraestrutura de cache de vídeo não apenas para dentro de suas redes, mas distribuí-la até as extremidades mais remotas (edges).

info@netskrt.io

#2200 – 1050 West Pender St.
Vancouver, BC
Canada V6E 3S7

/company/netskrt-systems-inc/

/netskrt.io/