WHITE PAPER

The Adaptive Network: uma estrutura para compreender as implicações da Edge Cloud na rede

Resumo Executivo

Os serviços em nuvem são onipresentes. De usuários individuais a serviços de vídeo OTT (Over-The-Top) até empresas que implantam SaaS (Software as a Service), os serviços em nuvem são a forma como as pessoas e organizações consomem conteúdo e dados. Durante anos, grandes e centralizadas arquiteturas de data center e nuvem forneceram acesso a esses serviços.

Agora, uma nova geração de aplicativos cloud-native está surgindo em categorias como entretenimento, varejo, manufatura e o setor automotivo, que, em muitos casos, serão de computação mais intensiva e sensíveis à latência. As arquiteturas de nuvem centralizadas tradicionais não atenderão às expectativas de Qualidade da Experiência (QoE) para esses aplicativos e exigirão um modelo de nuvem mais dinâmico e distribuído. Como resultado, os recursos de computação e armazenamento em nuvem precisam se mover para mais perto da borda da rede, onde o conteúdo é gerado ou consumido, para atender à QoE esperada. Essa nova abordagem é conhecida como Edge Cloud.

Essa mudança para um modelo distribuído de Edge Cloud resultará em cerca de três vezes mais data centers na borda da rede do que existe atualmente e exigirá que todo o ecossistema de nuvem pense de maneira diferente sobre o papel da conectividade de rede[1] .

Este documento examina os motivadores e as implicações da edge computing e explora como a visão Adaptive Network™ da Ciena pode fornecer uma estrutura eficaz para a evolução rumo a uma arquitetura Edge Cloud distribuída.

Com os dados chegando mais perto da borda da rede, o mundo está mudando

Imagine alguém que está prestes a sair para uma viagem de negócios fazendo um pedido simples: *"Ei, Siri, qual o caminho para chegar ao aeroporto"*. Em segundos, o Apple Maps determina a rota mais curta e fornece instruções passo a passo. No caminho, uma notificação pop-up fornece novas direções para evitar congestionamentos à frente, adicionando apenas alguns minutos à viagem. Em cada caso, a solicitação do Apple Maps provavelmente foi atendida por um data center centralizado que poderia estar a milhares de quilômetros de distância, resultando em latência adicional (atraso) no processamento da solicitação. Para um aplicativo não crítico como o Apple Maps, esse tempo de resposta é aceitável e geralmente não afeta a capacidade do usuário de navegar corretamente até seu destino.

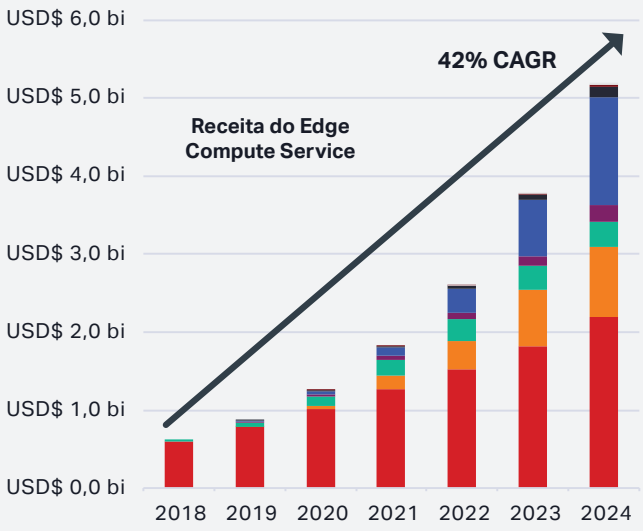
Em outro cenário, um cliente entra em sua mercearia favorita, fazendo check-in por meio de seu aplicativo de smartphone. Seus movimentos - escolher e colocar itens das prateleiras - são capturados por câmeras colocadas no teto da loja. A IA de visão computacional analisa essas imagens para determinar o que o cliente comprou e fatura o cartão de crédito diretamente, eliminando a necessidade de passar por um caixa no balcão de check-out. Recursos de computação significativos serão necessários nas lojas de varejo ou na borda para executar esse processamento de imagem quase em tempo real para oferecer uma experiência perfeita ao cliente.

Um motivador adicional para a computação de borda é o valor de processar as grandes quantidades de dados gerados localmente por esses aplicativos e reduzir o tráfego de backhaul de volta para a nuvem central. O objetivo é reduzir a latência e a quantidade de tráfego de backhaul para a nuvem central e atender com mais eficácia as análises em grande escala necessárias para enviar inferências e previsões para os dispositivos na borda, melhorando o desempenho do aplicativo.

Diversos aplicativos emergentes exigem latência menor do que pode ser atendida por meio de data centers centralizados, conforme mostrado na Figura 1. Espera-se que a receita para esses aplicativos tenha um CAGR de 42 por cento, passando de US$ 1,2 bilhão em 2020 para mais de US$ 5 bilhões em 2024, com os maiores impulsionadores de receita provenientes de redes de entrega de vídeo/conteúdo, cloud gaming e aplicativos automotivos[1].

1 Mobile Experts: "Edge Computing for Enterprises 2019", julho de 2019.

Aplicação	Descrição
Vídeo/CDN	É necessário armazenar o conteúdo em cache na borda local para melhorar o desempenho
Cloud Gaming	Substituição dos os consoles de jogos por streaming de jogos à partir de servidores na nuvem de baixa latência na Cloud Edge
IoT industrial/ automação	A edge computing reduz a latência para aplicativos remotos de missão crítica
Espaços inteligentes	Processamento necessário nos locais para reduzir os tempos de resposta
Automotivo	Infotainment, direção autônoma
AR/VR	Baixa latência para reduzir a cinetose (motion sickness)



Fonte: Mobile Experts, "Edge Computing for Enterprises 2019", Julho de 2019.

Figura 1. Principais aplicativos por receita para edge computing

Além da necessidade de oferecer suporte aos requisitos de latência da aplicação e de QoE, outros desafios vêm do lado móvel da rede. Aplicações móveis como cloud gaming e realidade aumentada/virtual são cada vez mais intensivos em computação, o que impacta negativamente o desempenho da bateria do dispositivo móvel. Nesse caso, quando a computação e o armazenamento do dispositivo móvel podem ser movidos para um data center na nuvem mais próximo do usuário, os usuários experimentam um desempenho melhor da bateria móvel.

Além disso, em resposta a questões de privacidade, vários governos estão exigindo que os provedores de serviços em nuvem armazenem os dados de seus clientes no país. Para um mercado como a Europa, muitas empresas de nuvem tradicionalmente hospedam seus recursos de nuvem em países como o Reino Unido e atendem clientes de forma centralizada em toda a Europa. Com o surgimento das fronteiras digitais estabelecidas pelos governos, os provedores de nuvem estão sendo obrigados a hospedar esses recursos de nuvem em data centers da borda dentro de cada país, mais perto de seus usuários.

A abordagem do setor para lidar com esses desafios é criar um modelo de nuvem mais distribuído e dinâmico, que envolve mover recursos de nuvem de data centers centralizados para mais perto do usuário, em data centers na borda.

Este documento aborda as implicações para a rede de uma abordagem de nuvem distribuída, conhecida como Edge Cloud. A Ciena define a Edge Cloud como um ecossistema de nuvem intercambiável que engloba componentes da edge computing (armazenamento e computação) de vários fornecedores, bem como uma rede escalável e com reconhecimento de aplicativos que interconecta data centers na borda que pode detectar e se adaptar às necessidades dos aplicativos de forma segura e em tempo real.

Então, onde está exatamente a Edge (borda)?

Embora muitos no setor estejam tentando estabelecer definições estáticas de onde a borda existe, a realidade é que a borda residirá em qualquer número de locais, dependendo das expectativas de QoE e dos requisitos/disponibilidade de recursos para um determinado aplicativo. A localização da Edge Cloud irá variar dependendo da perspectiva de um usuário, operadora de rede ou provedor de aplicativos.

Este documento se refere aos seguintes grupos de locais onde um aplicativo pode residir fisicamente, conforme mostrado na Figura 2:

1. **Metro Edge:** uma mistura de grandes data centers de vários clientes (Global Content Networks [GCNs] e Data Center Operators [DCOs]) e escritórios centrais de hub do Provedor de serviços de comunicações (CSP) adaptados como data centers, situados em locais regionais/ metropolitanos para atender a esse mercado

2. **Far Edge:** uma mistura de escritórios centrais do CSP, hed-end de Operadora de cabo multi-serviço (MSO) ou unidades distribuídas (DU) 5G móveis, situadas mais perto do usuário

3. **User/On-premises Edge:** uma mistura de localizações de grandes e pequenas empresas, incluindo data centers corporativos e filiais; pode se estender a hubs de transporte, locais de mineração e fábricas

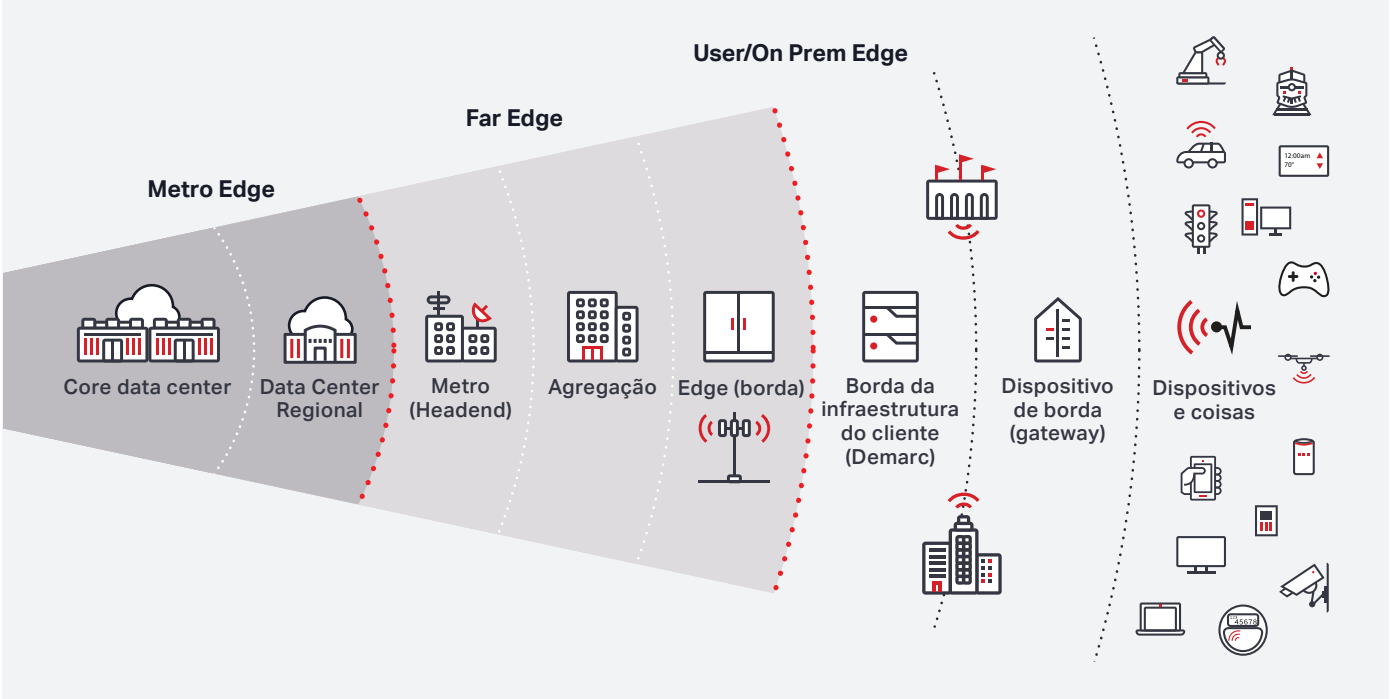
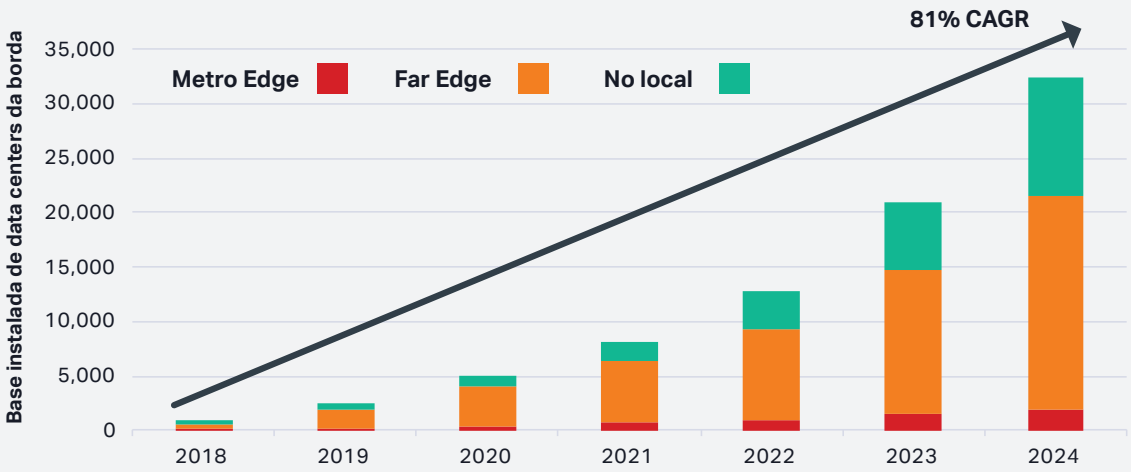


Figura 2. Localizações da borda

O surgimento da Edge Cloud está ofuscando essas linhas, à medida que as parcerias se formam entre uma ampla variedade de participantes da Edge Cloud com a capacidade de fornecer recursos na nuvem que abrangem várias nuvens, provedores de serviços ou soluções desenvolvidas pelo GCN ou pelos próprios clientes finais.

Hoje, existem aproximadamente 10.000 data centers no mundo todo. Com a mudança em expansão rumo à Edge Cloud, as previsões (conforme mostrado na Figura 3) mostram que haverá até três vezes mais locais de novos data centers nos agrupamentos Metro/Far Edge e User/ On-premises Edge nos próximos quatro anos[1].



Fonte: Mobile Experts, "Edge Computing for Enterprises 2019", Julho de 2019.

Figura 3. Crescimento de data centers edge computing

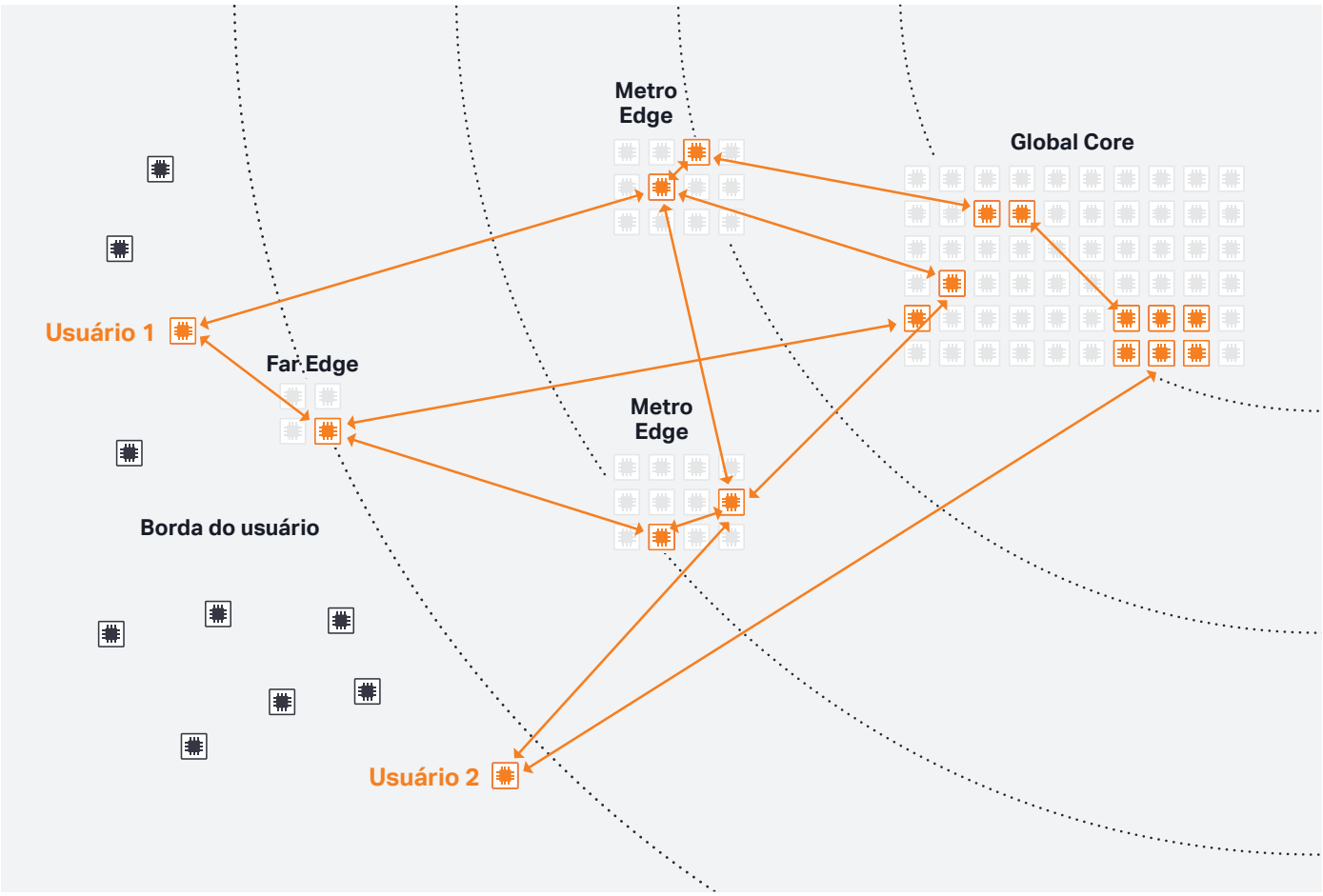


Figura 4. Natureza dinâmica de aplicativos habilitados por Edge Cloud

Não é apenas a definição da localização da borda que é fluida; é a natureza dos aplicativos na borda da rede que podem ser estáticos ou dinâmicos. Cada quadrado na Figura 4 representa um recurso de computação em que um aplicativo, microsserviço ou função de rede pode residir para oferecer suporte a um usuário final (empresa ou consumidor). Um aplicativo de usuário pode acessar recursos de computação em vários locais Edge Cloud durante o ciclo de vida de uso desse aplicativo. Diferentes usuários podem acessar os recursos da nuvem em qualquer um dos locais Far Edge, Metro Edge e/ou Global Core, dependendo da natureza do aplicativo e da disponibilidade dos recursos de nuvem necessários para atender aos requisitos de QoE durante a sessão do aplicativo. É a natureza dinâmica de como os aplicativos se movem em diferentes locais Edge Cloud o que exige novos requisitos de rede para suportar a Edge Cloud.

Quem são os provedores de Edge Cloud?

Além do desenvolvedor de aplicativos, vários provedores que aproveitarão a Edge Cloud - Hyperscalers, operadoras GCN, DCOs e CSPs - todos precisarão formar novos relacionamentos de negócios para habilitá-la. Conforme a corrida para construir a Edge Cloud se expande, esses provedores de ecossistema precisarão trabalhar juntos. Parcerias já estão surgindo e continuarão ocorrendo nos próximos anos.

A seção a seguir examina como os diferentes provedores de Edge Cloud devem evoluir suas estratégias de data center para habilitar a Edge Cloud (Figura 5).

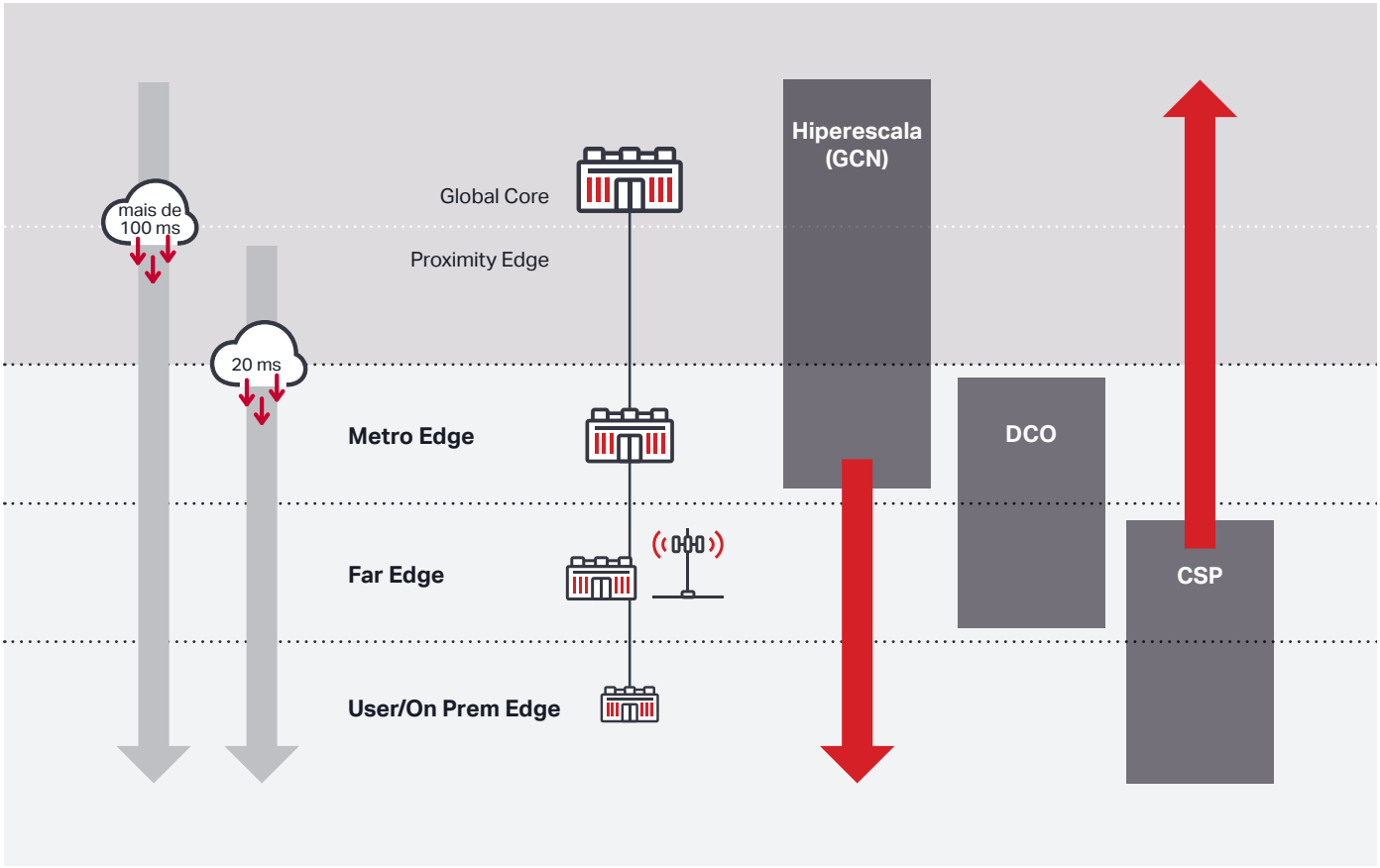


Figura 5. Provedores Edge Cloud versus site de data center de borda

GCNs: os GCNs criaram um amplo conjunto de data centers em hiperescala (global core) como parte de uma arquitetura de nuvem descentralizada. Eles também fizeram parceria com DCOs para expandir sua área metropolitana até a borda da rede. No entanto, para escalar para serviços em nuvem mais distribuídos e de baixa latência, eles precisarão expandir ainda mais sua presença na borda em locais de Far Edge e de User/On-premises Edge, seja criando seus próprios ou por meio de parcerias com CSPs que dominam esta área.

Parcerias estratégicas entre GCNs e CSPs já estão surgindo, como no caso da AT&T com o Google Cloud e a Azure, ou a Verizon com a AWS. Devido à natureza dinâmica de como os serviços da Edge Cloud serão consumidos, os GCNs esperam que as redes dos DCOs e dos CSPs forneçam maior conhecimento da Edge Cloud no contexto dos aplicativos alimentados pelo GCN que se executam sobre elas. Por exemplo, os GCNs que oferecem serviços de streaming de vídeo OTT em suas nuvens devem estar cientes de que há recursos de rede suficientes dos CSPs para fornecer uma QoE de streaming de vídeo consistente aos consumidores.

DCOs: as DCOs construíram um grande número de data centers em locais metropolitanos para permitir que seus clientes corporativos terceirizem sua infraestrutura de data center, ofereçam espaço e energia para provedores de nuvem e proporcionem uma troca para provedores de conteúdo e nuvem. Suas principais fontes de receita são derivadas de imóveis, energia e conectividade, mas elas reconhecem a necessidade de subir mais na pilha da nuvem para agregar valor e margens adicionais aos seus negócios. Elas continuarão desempenhando um papel fundamental na expansão da Edge Cloud.

CSPs: os CSPs atualmente dominam no fornecimento de conectividade e infraestrutura para o usuário final - seja corporativo ou consumidor - e criaram milhares de escritórios centrais/head-ends ao longo dos anos para fazer isso. Eles também estão virtualizando suas redes internas em uma arquitetura de borda com proximidade local ao usuário final. Como mencionado acima, os CSPs estão buscando novas parcerias com os GCNs, para conhecerem melhor os dos recursos de rede subjacentes que conectam seus locais de Edge Cloud aos aplicativos fornecidos pelos GCNs e vice-versa.

Requisitos para entregar serviços Edge Cloud

A natureza dinâmica da Edge Cloud exige que os vários participantes do ecossistema pensem de forma diferente sobre suas redes. Para ter sucesso com uma implantação de Edge Cloud, GCNs, DCOs e CSPs precisam entender os requisitos que suas redes enfrentarão e como eles devem responder. Os requisitos de rede chave da Edge Cloud são descritos a seguir:

1. **Conhecimento dos aplicativos:** as redes de aplicativos definirão a abordagem de rede de próxima geração para serviços e aplicativos em nuvem. Atualmente, os aplicativos operados ou hospedados principalmente por GCNs são executados em uma infraestrutura virtualizada que é abstraída da infraestrutura física. Para permitir a operação eficiente de aplicativos em rede nos recursos de edge computing geograficamente distribuídos, é necessário que as redes de aplicativos e infraestrutura física (sobreposição e subjacência, consulte a Figura 6) conheçam as características e requisitos umas das outras.
2. **Visibilidade e posicionamento da carga de trabalho da rede e do aplicativo:** atender às demandas dinâmicas na borda da rede requer que os aplicativos e provedores de rede tenham maior visibilidade nas camadas de infraestrutura e aplicativos, para observar onde o congestionamento está se formando e onde os problemas são antecipados. Esse nível de visibilidade deve ser suportado em todas as camadas e em um ambiente de vários fornecedores.
3. **Segurança:** conforme os aplicativos se tornam mais distribuídos e dinâmicos, manter uma postura de segurança consistente se torna cada vez mais complexo. Quando os recursos e aplicativos da nuvem são centralizados em um data center, as empresas podem padronizar a segurança técnica e física com mais facilidade. Uma abordagem Edge Cloud apresenta maior complexidade ao forçar o ecossistema a lidar com modelos emergentes de segurança zero-trust e parâmetros de segurança física, mas para três vezes mais locais de data center de borda em uma área amplamente distribuída.

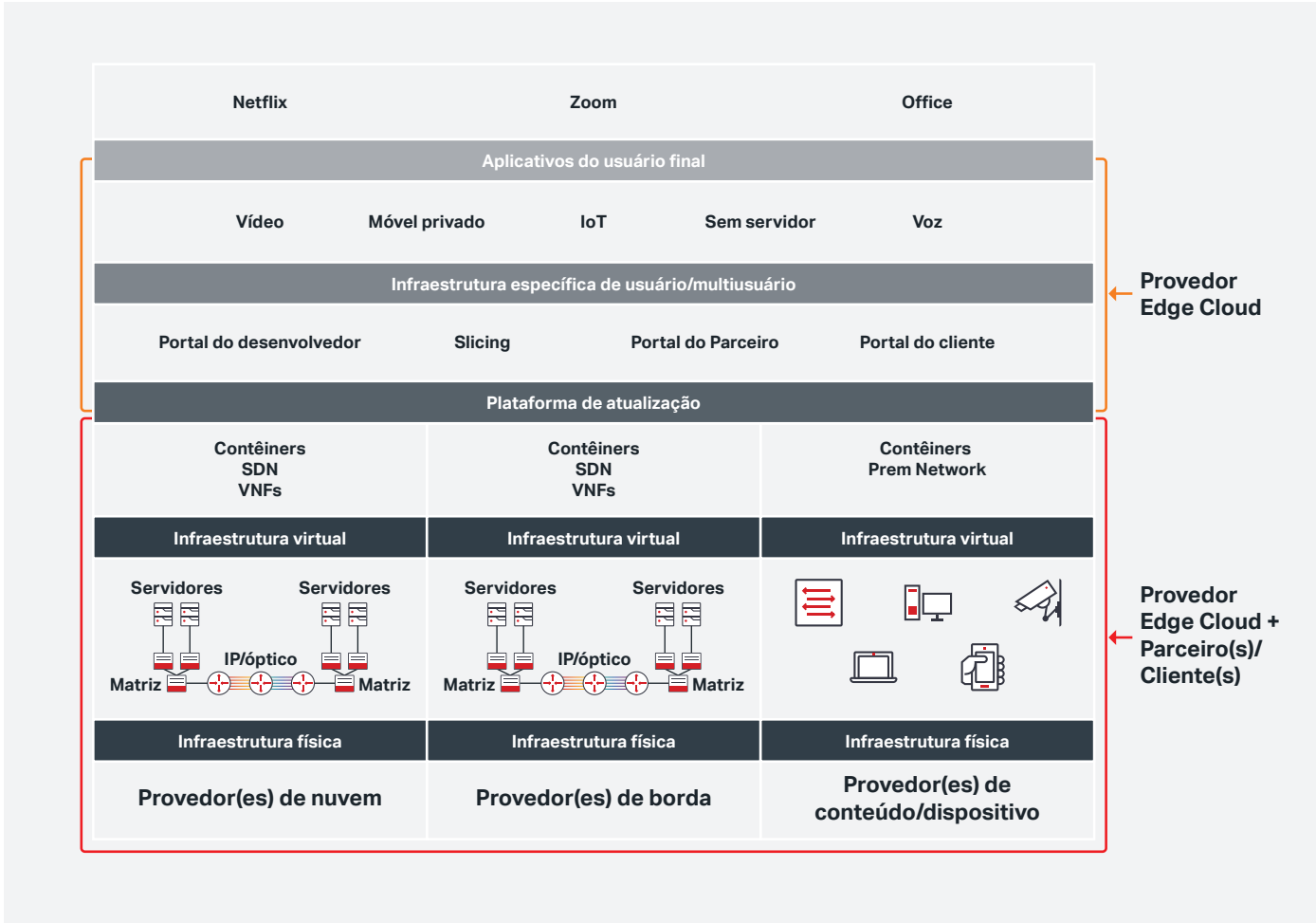


Figura 6. Componentes do ecossistema Edge Cloud

4. **Análise:** os dados em tempo real são um ativo comercial importante e o pool de dados continuará crescendo à medida que mais dados sejam coletados de terminais, especialmente com o crescimento de dispositivos IoT. A análise de dados precisa ser considerada de duas perspectivas:

   - Análise de dados em movimento - ingestão e análise de dados em tempo real/quase em tempo real; os principais metadados serão enviados para o norte (northbound) para armazenamento, o que ocorrerá perto da fonte, como na borda
   - Dados em repouso - os dados serão agregados/combinados com outros conjuntos de dados e analisados para business intelligence em um data center centralizado

   No entanto, os dados não são úteis, a menos que possam ser analisados para fornecer insights acionáveis e os resultados operacionalizados. Ferramentas de IA, Machine Learning (ML) e análises são essenciais para entender as demandas dos aplicativos de borda em constante mudança e melhorar o desempenho da rede e dos aplicativos.

5. **Capacidade escalável para redes de data center inter/intra-edge :** a adoção da edge computing cria novas expectativas para aplicativos de alto desempenho always on dentro e entre os data centers de borda e a nuvem central. Espera-se que a proporção de tráfego na parte metropolitana da rede cresça a uma taxa mais rápida do que o núcleo, e a adoção da computação para computação de alta largura de banda na borda apenas acelerará essa tendência. Essa inversão de capacidade core-to-metro exigirá uma infraestrutura óptica de pacote escalável no interior e entre esses data centers de borda, com conectividade contínua com a nuvem central.

6. **Orquestração e automação inteligente do Edge Cloud:** para otimizar a utilização dos recursos da Edge Cloud e, ao mesmo tempo, atender à demanda e aos requisitos de aplicativos dinâmicos, a automação inteligente, com uma visão panorâmica da rede e da Edge Cloud, é necessária. Instâncias separadas de orquestração funcionarão em virtualização de nuvem/borda, plataforma, infraestrutura e aplicativo para colocar e interconectar os componentes do aplicativo em hosts Edge Cloud adequados com base na localização do usuário final; recurso de aplicativo, QoS e especificações de serviço; capacidade, do host, custo e disponibilidade do host; capacidade e desempenho da rede; e restrições da operadora, regulamentações, usuário e outras restrições. Ao contrário da nuvem centralizada que atende a um grande número de clientes, cada aplicativo de borda é específico para um subconjunto muito menor de clientes e deve responder instantaneamente de forma dinâmica e automática aos requisitos de cada local do cliente em um ambiente com recursos limitados.

7. **Edge Cloud slicing para multilocação:** uma das oportunidades para provedores de rede da Edge Cloud é alocar dinamicamente diferentes recursos de nuvem e rede para cada cliente dentro e entre seus data centers de borda. Pode-se referir a isso como Edge Cloud slicing, que entrega recursos de computação, armazenamento e rede de ponta a ponta com base no aplicativo do cliente e nos requisitos de SLA.

O que é Edge Cloud?

Abordagem Adaptive Network™ para habilitar a Edge Cloud

O principal desafio para os provedores de borda é gerenciar de forma eficiente e inteligente a rede e os recursos dos aplicativos para os data centers Edge Cloud durante os períodos de pico de uso. A visão Adaptive Network da Ciena fornece uma estrutura para Edge Cloud que permite aos provedores alcançarem coletivamente uma rede ponta a ponta que fica mais inteligente e ágil a cada dia, com a escala necessária para responder dinamicamente às pressões colocadas sobre ela.

A Adaptive Network permite que os provedores de Edge Cloud otimizem suas infraestruturas existentes, incorporando novas tecnologias e formas de trabalhar para atender aos novos requisitos de Edge Cloud. A Adaptive Network é construída em quatro elementos básicos fundamentais - Infraestrutura Programável, Análise e Inteligência, Controle e Automação de Software e Serviços - que aprimoram a rede e os resultados de negócios de forma independente, mas são um multiplicador de força quando trabalham juntos.

Infraestrutura programável: um pacote de borda programável e infraestrutura óptica é aquele que pode ser acessado e configurado por meio de interfaces abertas comuns, é altamente escalável e equipado com a capacidade de exportar dados de desempenho da rede em tempo real para a camada de aplicativo da Edge Cloud e pode ajustar seus recursos , conforme necessário, para atender às demandas da camada do aplicativo. Isso será fundamental para habilitar uma rede com reconhecimento de aplicativo e fornecer escalabilidade para interconectar as matrizes Edge Cloud entre e dentro dos data centers da Edge Cloud. Além disso, o network slicing na camada de infraestrutura será essencial para os provedores habilitarem os serviços de multilocação da Edge Cloud para diferentes provedores de nuvem e sobreposições de aplicativos.

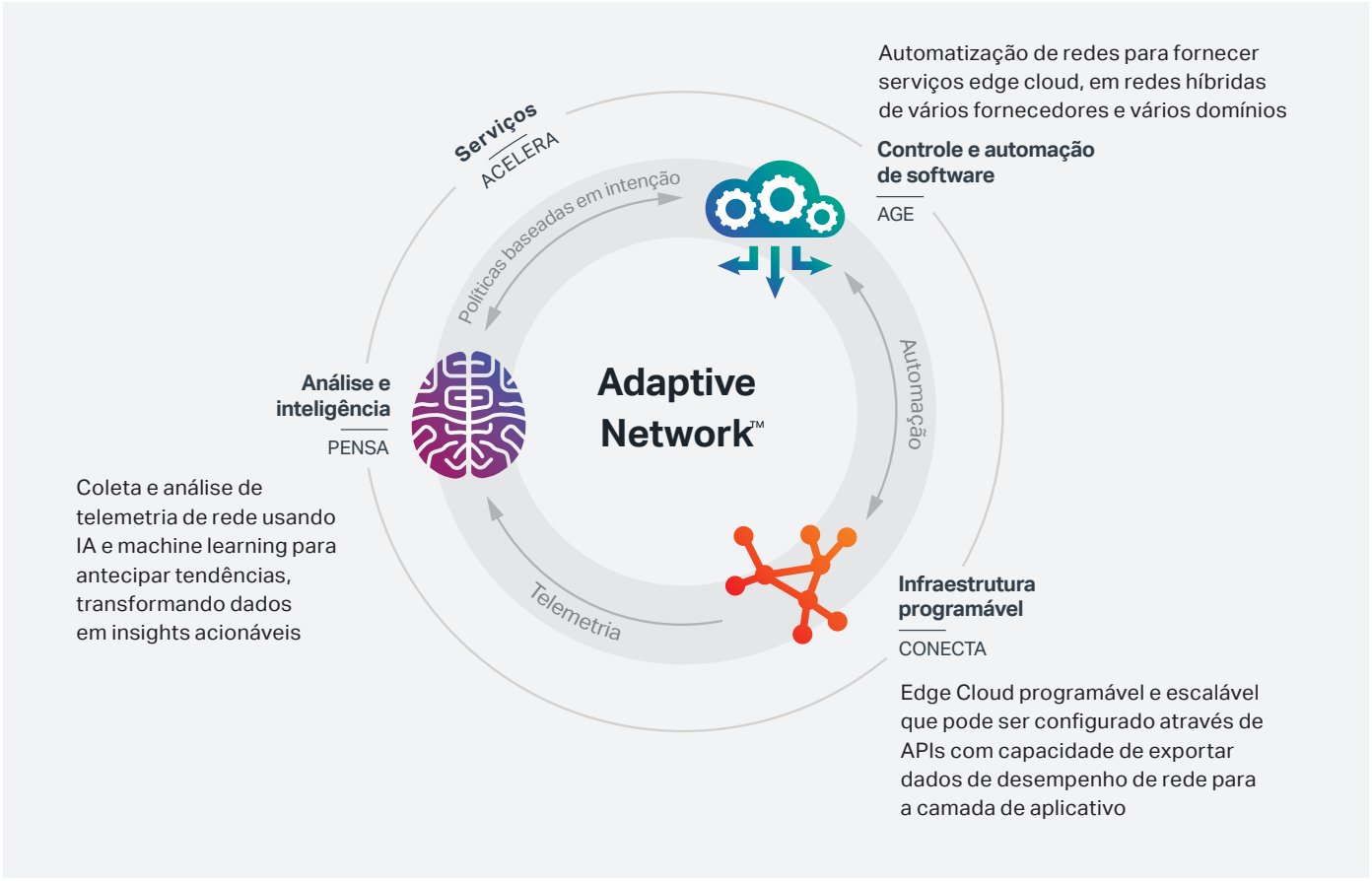


Figura 7. The Adaptive Network: uma estrutura para habilitar a Edge Cloud

Análise e inteligência: a Edge Cloud é uma extensão das práticas de computação e operação em nuvem, que dependem fortemente da automação informada por meio da interpretação de fluxos massivos de telemetria/indicadores de desempenho (KPI) chave dos recursos subjacentes. O reconhecimento e a automação do aplicativo dependerão significativamente da coleta e análise usando IA de rede, servidor e recursos virtualizados (contêiner), e da capacidade de antecipar tendências transformando montanhas de dados em insights acionáveis. Aproveitar esses insights fornecerá uma rede com reconhecimento de aplicativos que pode detectar e se adaptar às necessidades dos aplicativos de borda com segurança e em tempo real.

Controle e automação de software: automatizar a colocação de cargas de trabalho do Edge Cloud para atender às demandas de aplicativos em tempo quase real será fundamental para atender às principais restrições e objetivos da Edge Cloud. Por meio da implementação de SDN, NFV e APIs abertas, os provedores podem simplificar o ato de gerenciar, proteger e automatizar suas redes de ponta a ponta para fornecer serviços Edge Cloud em redes híbridas de vários fornecedores e vários domínios.

Serviços: serviços técnicos e profissionais são necessários para ajudar os provedores a determinar sua melhor estratégia e arquitetura para Edge Cloud e para construir, operar e melhorar continuamente suas redes, acelerando sua jornada para a Adaptive Network.

Visão da Adaptive Network
Saiba mais

Resumo

Essas ainda são as fases iniciais da implantação da Edge Cloud e da evolução para uma arquitetura de nuvem distribuída. A borda não deve ser considerada como um local de data center específico, mas residirá em qualquer número de locais, dependendo das expectativas de QoE e dos requisitos/disponibilidade de recursos de um determinado aplicativo. A localização de um aplicativo pode mudar para diferentes data centers de borda durante seu ciclo de vida, levando à necessidade de dimensionar a infraestrutura de maneira inteligente tanto dentro como entre os data centers de borda e para a nuvem central, enquanto automatiza as cargas de trabalho entre os locais na borda da rede.

Embora os GCNs tenham demonstrado com sucesso sua capacidade de escalar dentro de um modelo de nuvem pública e híbrida centralizado, a mudança para um modelo Edge Cloud distribuído exigirá parcerias com DCOs e CSPs para aproveitar a sua ampla infraestrutura e espaço físico mais próximos dos usuários finais. Para que um modelo Edge Cloud distribuído atinja todo o seu potencial, novos requisitos de rede deverão ser atendidos. Mais notavelmente, as camadas de aplicativo da pilha da nuvem devem estar dinamicamente cientes dos recursos nas camadas de rede, enquanto as camadas de rede devem manter reconhecimento da mudança de dinâmica na camada do aplicativo.

As soluções Adaptive Network da Ciena desempenham um papel crítico em algumas das maiores arquiteturas de data center e nuvem do mundo atualmente. Tendo o maior market share tanto na DCI Global quanto na DCI Metropolitana, a Ciena está bem posicionada para levar sua profunda experiência e liderança nos mercados de nuvem e DCI até o limite. A Adaptive Network fornece uma estrutura adicional para todos os provedores de ecossistema de borda seguirem, abordando os desafios do modelo dinâmico de Edge Cloud e aproveitando uma infraestrutura altamente programável e escalável, análise e automação para dimensionar dinamicamente os recursos de nuvem tanto da rede como do aplicativo conforme necessário para atender às expectativas do usuário final. Seguir a estrutura da Adaptive Network pode ajudar a garantir que o desempenho de um modelo Edge Cloud possa ser dimensionado e adaptado para atender às demandas em constante mudança da borda da rede.

ciena