DEFININDO OS QUATROS ARQUÉTIPOS DO "EDGE" E SUAS EXIGÊNCIAS TECNOLÓGICAS

Introdução

Ao longo dos últimos anos, "edge computing" se tornou uma das tendências mais faladas em TI, e por bom motivo. A Grand Valley Research projeta um **crescimento anual composto de 41 por cento da edge computing** entre 2018 e 2025. Quase todas as indústrias estão reconhecendo as limitações de suportar usuários e tecnologias emergentes por meio de infraestruturas de TI centralizadas; elas estão avançando o armazenamento e a computação para mais perto dos usuários e dispositivos.

Essa mudança está se tornando necessária devido à maior conectividade de dispositivos e pessoas, e aos enormes volumes de dados que eles geram e consomem. Segundo o **Cisco Visual Networking Index**, o tráfego IP global atingiu 1,2 zetabytes em 2016. Até 2021, ele quase triplicará, atingindo 3,3 zetabytes. Também até 2021, a Cisco projeta que o número de dispositivos conectados a redes IP será o triplo da população global. Isso representa mais de 23 bilhões de dispositivos conectados em apenas três anos. **Outras empresas estão fazendo projeções semelhantes:** até 2020, o Gartner projeta 20,8 bilhões de dispositivos conectados; o IDC, 28,1 bilhões; e a IHS Markit, 30,7 bilhões.

Grande parte desses dados da IoT serão dados de sensores móveis que precisarão ser transmitidos em redes sem fio ou móveis, em vez de conexões Internet cabeadas, pressionando a infraestrutura de rede móvel. **A projeção para o tráfego móvel é aumentar sete vezes até 2021**, o dobro da velocidade de crescimento do tráfego em IP fixo.

As mudanças de infraestrutura de computação e armazenamento necessárias para suportar o futuro inteligente e conectado, particularmente no nível local, serão profundas.

Entretanto, ao explorarmos as informações atualmente disponíveis acerca de edge computing, descobrimos a pouca ou total ausência de recursos que forneçam uma visão abrangente do ecossistema de perímetro. Uma análise atenta do Mercado revela uma grande variedade de casos de uso atuais e emergentes; embora compartilhem algumas semelhanças com base na definição ampla de edge computing, eles são também distintos de algumas maneiras significantes.

Juntamente com uma empresa de consultoria independente terceirizada, os especialistas em perímetro da Vertiv analisaram os casos de uso que compreendem o ecossistema de perímetro para desenvolver uma melhor compreensão dessas diferenças e suas implicações para a infraestrutura de apoio. Como resultado dessa análise, identificamos quatro arquétipos principais para aplicações de perímetro:

- Intenso em dados
- Sensível à latência humana
- Sensível à latência máquina-a-máquina
- Crítico para a vida

Este documento apresenta uma descrição de cada arquétipo com exemplos dos casos de uso de maior impacto, juntamente com uma visão geral de suas exigências de conectividade com concentradores locais, metropolitanos e regionais, que representam a camada de transmissão e o núcleo do perímetro e são, às vezes, diferenciados como edge computing, fog computing e cloud computing.

Entendendo os casos de uso de perímetro

Para identificar os quatro arquétipos, primeiramente foi necessário compreender os casos de uso para a tecnologia de perímetro. A equipe de pesquisa da Vertiv identificou e revisou mais de 100 casos de uso para tecnologia de perímetro e refinaram essa lista inicial até os 24 que terão o maior impacto sobre a infraestrutura de TI, para uma análise mais detalhada.

Essa análise observou as exigências de desempenho de cada caso de uso em termos de latência, disponibilidade e crescimento projetado, bem como exigências de segurança como a necessidade de criptografia, autenticação e conformidade regulatória. Foi também avaliada a necessidade de integração com aplicações existentes ou legadas e outras fontes de dados, e o número de potenciais locais necessários para suportar o caso de uso.

Mais importante, a equipe estudou as características dos dados de cada caso de uso e descobriu que as aplicações que sustentam cada um deles têm um conjunto datacêntrico de requisitos de carga de trabalho, além de seus requisitos de disponibilidade e segurança. Eles incluem volume de dados, como os dados são acessados, requisitos de transmissão de dados, integridade de dados e analytics de dados. Essa abordagem datacêntrica, filtrada por requisitos de disponibilidade e segurança, é fundamental para a compreensão e categorização dos requisitos de diversos casos de uso.

Uma lista dos 24 casos de uso, organizada por arquétipo, pode ser encontrada na Figura 1.

O ecossistema do perímetro

INTENSO EM DADOS	SENSÍVEL À LATÊNCIA MÁQUINA-A-MÁQUINA	CRÍTICO PARA A VIDA	SENSÍVEL À LATÊNCIA HUMANA
• Conectividade restrita • Cidades inteligentes • Fábricas inteligentes • Casas/edifícios inteligentes • Distribuição de conteúdo em HD • Computação de alto desempenho • Realidade virtual • Digitalização de petróleo e gás • Custo elevado de infraestrutura de rede	• Segurança inteligente • Rede de energia inteligente • Distribuição de conteúdo com baixa latência. • Mercado de arbitragem • Analytics em tempo real • Simulação de força de defesa	• Saúde digital • Carrros conectados/ autônomos • Drones • Transporte inteligente • Robôs autônomos	• Otimização de websites • Realidade aumentada • Varejo inteligente • Processamento de linguagem natural

Figura 1: *Arquétipos*

Arquétipo um: Intenso em dados

Largura de banda	Latência	Disponibilidade	Segurança
Grande	Média	Alta	Média

O arquétipo intenso em dados representa casos de uso nos quais a quantidade de dados inviabiliza a transferência via rede diretamente para a nuvem, ou da nuvem para o ponto de uso, devido a problemas de volume de dados, custo ou largura de banda.

Provavelmente, o exemplo mais amplamente discutido de aplicação de perímetro intensa em dados é a distribuição de conteúdo em alta definição. **Em 2016, o vídeo foi responsável por 73 por cento de todo o tráfego IP e espera-se que isso cresça para 82 por cento até 2021**, à medida que streaming de vídeo e realidade virtual continuem a crescer. Os principais provedores de conteúdo, como Amazon e Netflix, estão se associando ativamente a provedores de colocation para expandir suas redes de entrega e levar para perto dos usuários streaming de vídeo intenso em dados, para reduzir custos e latência.

No momento, **35 por cento do conteúdo acessado por um usuário de Internet na América do Norte é envido pela área municipal em que o usuário está localizado**. Projeta-se que isso aumente para 51 por cento até 2021, à medida que os provedores de conteúdo continuem a estender suas redes para o perímetro. Contudo, isso representa apenas a primeira onda de computação "core-to-edge" (do centro para o perímetro). À medida que a demanda por vídeo de alta definição continuar a crescer, os concentradores de dados locais auxiliarão cada vez mais os atuais concentradores metropolitanos para reduzir ainda mais os custos de largura de banda e os problemas de latência.

Outro ótimo exemplo do arquétipo intenso em dados é o uso de redes IoT para criar casas, edifícios, fábricas e cidades inteligentes. Uma enquete feita em 2018 pela 451 Research e a Vertiv encontrou que, embora apenas 33 por cento das 700 organizações pesquisadas houvessem implementado amplamente IoT, 56 por cento indicaram que, atualmente, pelo menos 25 por cento de sua capacidade de TI suportam IoT. A despeito de a IoT ainda estar em seus estágios iniciais, as organizações já estão se esforçando para lidar com o volume de dados que está sendo gerado.

Nesse caso, o desafio é o oposto daquele apresentado pela entrega de conteúdo em alta definição. Em vez de mover os dados para mais perto dos usuários, essas aplicações precisam mover as enormes quantidades de dados geradas por dispositivos e sistemas na origem para um local central, para processamento. Isso exigirá a evolução de uma arquitetura de rede "edge-to-core" (do perímetro para o centro).

A IoT e a Internet Industrial das Coisas (IoT) representam uma malha de sensores que geram enormes volumes de dados a cada hora. Esses dados suportam uma realimentação "percebe-infere-reage" que permite a visibilidade e o controle de tudo, desde eletrodomésticos até equipamentos industriais. Somente um subconjunto desses dados é transmitido para um data center local, regional ou na nuvem para processamento adicional, o que significa que uma computação massiva será necessária na extremidade do perímetro para permitir que dispositivos e sistemas tomem decisões e atuem sobre os dados fornecidos pelos sensores.

A mais simples dessas aplicações, a casa inteligente, precisa suportar múltiplos dispositivos e sistemas intensos em dados, incluindo entretenimento, sistemas de aquecimento, ventilação e ar-condicionado, e segurança.

Intenso em dados

Segundo a IHS Markit, **o mercado mundial de dispositivos residenciais conectados crescerá de mais de 100 milhões de unidades em 2017 para cerca de 600 milhões de unidades em 2021**.

Cidades e fábricas inteligentes ampliam os desafios de dados inerentes às casas inteligentes. Muitas cidades já estão pilotando ou avaliando tecnologia de cidade inteligente para melhorar fluxos de tráfego, apoiar serviços de emergência e reduzir custos.

As fábricas inteligentes, que alavancam a convergência de IoT, sistemas ciberfísicos e computação na nuvem para permitir que os fabricantes usem dados em tempo real para aumentar a eficiência, reduzir custos e adaptar-se a mudanças de demanda, estão sendo promovidas como a próxima revolução industrial. Segundo a McKinsey, fábricas e outros ambientes de produção têm o potencial de perceber o maior impacto financeiro da aplicação de IoT. Eles preveem que a IoT gerará um **valor econômico na casa de USD 1,2 trilhões a 3,7 trilhões** até 2025. Esse valor virá de novas eficiências energéticas, produtividade no trabalho, otimização de estoques e maior segurança para o trabalhador. Realizá-lo, porém, exigirá uma robusta infraestrutura local.

Na indústria de petróleo e gás, a digitalização já criou uma ampla melhoria na eficiência dos processos de exploração e extração, mas também introduziu enormes desafios de

gerenciamento de dados. Um único equipamento de perfuração pode gerar terabytes de dados a cada dia.

Outros casos de uso que se enquadram no arquétipo intenso em dados incluem realidade virtual, altos custos de infraestrutura de rede, computação de alto desempenho e ambientes com conectividade restrita, como áreas nas quais estão ocorrendo operações de recuperação após um desastre natural ou ciberataque.

O que todos esses casos de uso têm em comum é a necessidade de mover grandes volumes de dados para os usuários onde eles possam ser consumidos, ou de dispositivos e sistemas onde eles são gerados para um repositório central.

Arquétipo dois: Sensível à latência humana

Largura de banda	Latência	Disponibilidade	Segurança
Média	Alta	Média	Média

O arquétipo sensível à latência humana cobre casos de uso nos quais serviços são otimizados para consumo humano. Como o nome sugere, a velocidade é a característica que define este arquétipo.

O desafio da latência humana pode ser visto no caso de uso da otimização da experiência do cliente. Em aplicativos como e-commerce, a velocidade tem impacto direto sobre a experiência do usuário; sites otimizados para velocidade usando infraestrutura local se traduzem diretamente em aumento das visualizações de páginas e das vendas.

Sensível à latência humana
A Google descobriu que acrescentar um atraso de 500 milissegundos no tempo de resposta das páginas resultava em diminuição de 20 por cento no tráfego; a Yahoo observou que um atraso de 400 milissegundos causava redução de 5 a 9 por cento no tráfego.

Esse efeito também se estende ao processamento de pagamentos. A Amazon descobriu que um atraso de 10 milissegundos no processamento de pagamentos causava uma redução de 1% na receita auferida. A aprovação centralizada por meio de senha levava, em média, 7 segundos. Uma mudança para processamento local reduziu o tempo para 600 milissegundos, uma melhoria de 6.400 milissegundos, com cada 100 milissegundos resultando potencialmente em 1% extra de receita auferida.

Outro exemplo emergente de aplicação sensível à latência humana é o processamento de linguagem natural. Provavelmente, no futuro a voz será a forma primária de interação com as aplicações de TI cotidianas. Atualmente, o processamento de linguagem natural para Alexa e Siri é realizado na nuvem. Entretanto, à medida que aumentar o volume de usuários, aplicações e idiomas suportados, será necessário migrar esses recursos para mais perto dos usuários.

Outros casos de uso de latência humana identificados incluem o varejo inteligente, como as lojas Amazon Go, que não têm caixas registradoras; e tecnologias imersivas, como realidade aumentada, onde pequenos atrasos na latência podem significar a diferença entre diversão e náusea.

Em cada caso, atrasos na entrega de dados impactam diretamente a experiência do usuário com a tecnologia, como no processamento de idiomas e na realidade aumentada, ou nas vendas e lucratividade de um varejista com otimização de sites e varejo inteligente. À medida que esses casos de uso crescerem, aumentará também a necessidade de concentradores locais de processamento de dados.

Arquétipo três: Sensível à latência máquina-a-máquina

Largura de banda	Latência	Disponibilidade	Segurança
Média	Alta	Alta	Alta

O arquétipo sensível à latência máquina-a-máquina abrange casos de uso em que os serviços são otimizados para consumo máquina-a-máquina. Devido às máquinas conseguirem processar dados muito mais rapidamente do que os seres humanos, a velocidade é a característica que define este arquétipo. As consequências de não entregar dados nas velocidades necessárias podem ser ainda maiores neste caso do que no arquétipo sensível à latência humana.

Por exemplo, os sistemas usados em transações financeiras automatizadas, como negociação de commodities e ações, são sensíveis à latência. Nesses casos, os preços podem mudar em milissegundos, e sistemas que não têm os dados mais recentes quando necessário não podem otimizar transações, transformando ganhos potenciais em perdas.

Sensível à latência máquina-a-máquina
Segundo um estudo do Tabb Group, um corretor pode perder **até USD 4 milhões em receita por milissegundo** se sua plataforma eletrônica de negociação estiver 5 milissegundos atrás dos concorrentes.

A tecnologia de redes elétricas inteligentes também recai neste arquétipo. Essa tecnologia está sendo implementada na rede de distribuição elétrica para balancear automaticamente a oferta e a demanda, e gerenciar o uso de eletricidade de maneira sustentável, confiável e econômica. Ela permite que as redes de distribuição se restabeleçam automaticamente, otimizem seu custo e gerenciem fontes de energia intermitentes, supondo-se que os dados corretos estejam disponíveis no momento certo.

Outras aplicações sensíveis à latência máquina-a-máquina incluem sistemas de segurança inteligentes que dependem de reconhecimento de imagem, simulações de guerra militar e análises em tempo real.

Arquétipo quatro: Crítico para a vida

Largura de banda	Latência	Disponibilidade	Segurança
Média	Alta	Alta	Alta

O arquétipo crítico para a vida engloba casos de uso que impactam diretamente a saúde e segurança humana. Nesses casos de uso, velocidade e confiabilidade são primordiais.

Provavelmente, os melhores exemplos do arquétipo crítico para a vida são os veículos autônomos e drones, que proporcionam grandes benefícios quando operam conforme projetado; entretanto, se tomarem decisões erradas, podem pôr em perigo a saúde humana.

Os veículos autônomos progrediram mais rapidamente do que muitos esperavam, com várias empresas automotivas e de tecnologia já testando veículos ativamente. A maioria desses veículos tem um ser humano no banco do motorista, pronto para suplantar os controles automáticos se ocorrerem problemas, para minimizar o risco à saúde humana. Porém, no futuro próximo, veículos de entrega e sistemas de transporte sem motoristas estarão nas ruas. Se esses sistemas não tiverem os dados necessários quando precisarem deles, as consequências poderão ser desastrosas.

O mesmo se aplica aos drones. Podemos facilmente estar olhando para um futuro em que centenas de drones de entrega estarão sobrevoando uma cidade a qualquer dado momento.

Crítico para a vida
Grandes empresas de e-commerce e entrega de pacotes, como a Amazon e a DHL, já estão experimentando drones para entrega de pacotes.

O aumento do uso de tecnologia nos cuidados de saúde também representa um arquétipo crítico para a vida. Registros eletrônicos de saúde, cibermedicina, medicina personalizada (mapeamento de genoma) e dispositivos automonitorados estão remodelando os cuidados de saúde e gerando enormes volumes de dados.

Outros exemplos incluem transporte inteligente e robôs autônomos. As indústrias de transporte e logística estão buscando soluções datacêntricas para aprimorar a segurança de motoristas e passageiros, a eficiência de combustível e o gerenciamento de ativos. Nesse espaço, a tecnologia incluirá sistemas inteligentes de transporte, gerenciamento de frotas e telemática; sistemas de orientação e controle; aplicações de entretenimento de passageiros e comércio; sistemas de reservas, pedágio e bilhetagem; e sistemas de segurança e vigilância.

Exigências tecnológicas para concentradores locais e regionais

A infraestrutura necessária para suportar esses casos de uso atuais e estabelecidos consiste em quatro camadas de armazenamento e computação, além da infraestrutura de comunicações necessária para mover dados entre as camadas.

Na origem, tipicamente há o dispositivo que gera ou consome dados e uma extremidade de processamento. O dispositivo pode ser um sensor que monitore qualquer coisa, desde o status de energização de uma lâmpada, o acesso a uma porta, a temperatura de uma sala ou outras informações desejadas. A extremidade de processamento pode ser tão simples quanto o PC ou o tablet para o qual um consumidor está transmitindo vídeo, ou podem ser os microprocessadores incorporados em automóveis, robôs ou dispositivos vestíveis. Esses componentes são dependentes da aplicação e, tipicamente, projetados pelo fabricante do equipamento ou adaptados a dispositivos existentes.

Todo arquétipo também exigirá um concentrador de dados local, que fornece armazenamento e processamento próximos à origem. Em alguns casos, o concentrador local pode ser um data center independente. Mais comumente, será um sistema baseado em rack ou fileira, que forneça 30 a 300 kW de capacidade em um gabinete integrado que pode ser instalado em qualquer ambiente.

Esses sistemas de gabinete baseados em racks e fileiras integram comunicação, computação e armazenamento com proteção energética, controles ambientais e segurança física adequados. Para arquétipos que exigem alto grau de

disponibilidade, tais como o sensível à latência máquina-a-máquina e o crítico para a vida, o concentrador local deve incluir sistemas redundantes de nobreak e estar equipado para permitir gerenciamento e monitoramento remotos. Muitos casos de uso também exigirão criptografia de dados e outros recursos de segurança no concentrador local.

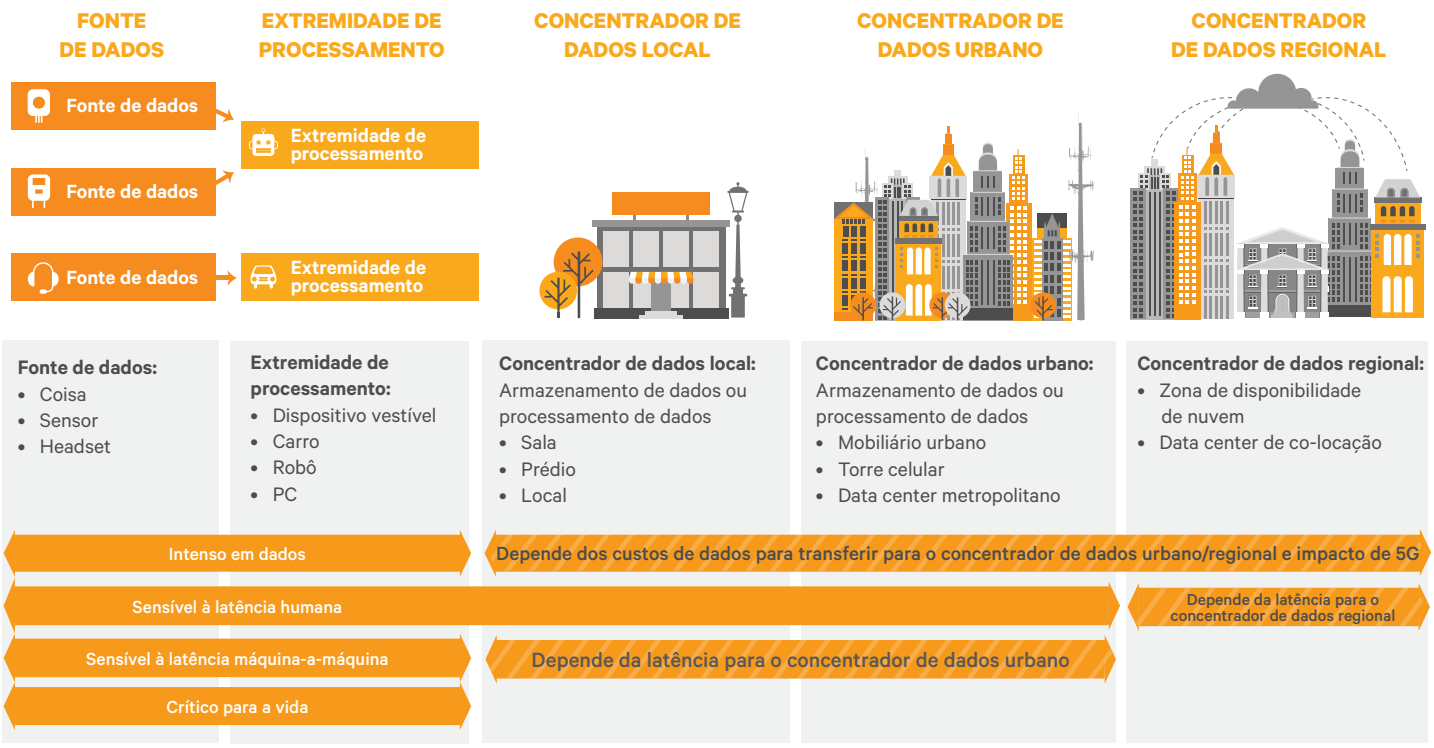
Para todos os arquétipos, exceto o crítico para a vida, o concentrador local exigirá a capacidade de conectar-se a um concentrador metropolitano e/ou regional, que fornecerá armazenamento de dados a longo prazo e recursos de suporte, como aprendizado de máquina.

O concentrador metropolitano aproveita a infraestrutura de telecomunicações existente para apoiar o concentrador local com armazenamento de dados de longo prazo e capacidades de processamento mais robustas. Provavelmente, o concentrador regional será um data center de nuvem operando na mesma região do concentrador local.

Tanto para o concentrador metropolitano quanto para o regional, devem ser considerados projetos modulares capazes de crescer facilmente além das especificações iniciais de projeto, de modo a responder a picos de demanda inesperados. Essas instalações devem também ser projetadas para crescer em termos de densidade. Aplicações com uso intenso de imagens, como realidade virtual, e aplicações com processamento intenso, como analytics e aprendizado de máquina, provavelmente exigirão densidades de rack que excedam a típica especificação de projeto de 10 kW. Em virtualmente todos os casos, esses concentradores devem fornecer um nível igual ou maior de redundância e segurança que o concentrador local.

Avançando

Na identificação das necessidades de carga de trabalho para os vinte e quatro casos de uso discutidos emergiram quatro arquétipos principais que podem orientar decisões referentes às exigências de infraestrutura e configuração para os casos de uso analisados, bem como para os que surgirão nos próximos anos. A Vertiv tomará este trabalho inicial sobre arquétipos como base para definir adicionalmente exigências e configurações tecnológicas específicas para cada arquétipo.



VertivCo.com | **Vertiv Brasil,** Av. das Nações Unidas, 11.633 - 4º andar. Novo Brooklin - São Paulo/SP - Brasil

SL-11490 (R08/18) PT