RESUMO DO ANALISTA

Patrocinado por: Dell

O investimento em infraestrutura de processamento de IA está crescendo em um ritmo acelerado. A boa notícia é que, para a infraestrutura de processamento de IA, mais de 50% dos sistemas não serão acelerados em 2024 e poderão ser executados em servidores e sistemas de rede Ethernet padrão.

A IA generativa está evoluindo a uma velocidade recorde à medida que as empresas iniciam a jornada rumo a essa tecnologia

Fevereiro de 2024

Escrito por: Brandon Hoff, diretor de pesquisa, Enabling Technologies: Networking and Comm, e Vijay Bhagavath, vice-presidente de pesquisa, Cloud and Datacenter Networks

Introdução

Uma pesquisa da IDC apresenta previsões e motivadores subjacentes que deverão afetar os investimentos em TI em 2024 e além. Líderes de tecnologia e seus colegas nas linhas de negócios (LOBs) podem usar este documento para orientar os respectivos esforços de planejamento estratégico.

As equipes de operações têm capturado dados, criado data lakes e utilizado a nuvem para armazenar seus dados. Agora, com a popularidade do ChatGPT, o momento do iPhone para a IA generativa (GenAI), as equipes de operações sabem o que podem fazer com seus conjuntos de dados. Como todos conhecem os benefícios da IA generativa, as equipes de operações também estão enfrentando uma pressão adicional dos investidores, dos executivos e do mercado para implementar uma estratégia eficaz de IA generativa. Há várias tecnologias e uma ampla variedade de opções que podem ser utilizadas para melhorar as operações de negócios e a produtividade dos funcionários, desde IA generativa até ML, gêmeos digitais e muito mais. A implementação bem-sucedida da tecnologia certa se tornará um KPI essencial para as equipes de operações e a empresa em geral.

DESCRIÇÃO RÁPIDA

IDEIAS PRINCIPAIS

Comece a planejar a infraestrutura de IA generativa:

- Integre a IA com mais rapidez do que os outros, primeiro movendo pelo menos uma cópia dos dados para o local por meio de uma iniciativa para extrair dados da nuvem.
- Invista na compreensão do valor dos algoritmos que sustentam a IA generativa, a IA, o ML e os gêmeos digitais nos negócios e priorize as decisões com base no retorno comercial.
- Três etapas: implementar servidores e sistemas de rede Ethernet padrão para a avaliação da IA generativa. Faça o scale-out da IA generativa para as cargas de trabalho de porte empresarial, conforme necessário, utilizando Ethernet padrão. Reequilibre as cargas de trabalho entre a infraestrutura local e externa para otimizar o CAPEX e o OpEx nos próximos três a cinco anos.

Compreensão da IA generativa atual

Com a explosão da demanda por IA, os provedores de serviços (SPs) em nuvem e as empresas estão construindo a respectiva infraestrutura em um ritmo acelerado. Os provedores de serviços em nuvem estão consumindo a grande maioria dos aceleradores de IA e criando a própria infraestrutura de IA. No entanto, esses aceleradores são caros, o que eleva o custo dos serviços de IA desses provedores. Os aceleradores de IA consistem em GPUs, TPUs, FPGAs, ASSPs e ASICs. Essas provedores de serviços em nuvem estão criando fábricas de IA para cargas de trabalho massivas que abrangem as necessidades de uma ampla variedade de empresas e são implementadas principalmente nos maiores ambientes de TI do mundo, abrangendo cerca de nove empresas.

Por outro lado, a infraestrutura de IA generativa para as cargas de trabalho de porte empresarial pode ser criada com sistemas padrão sem necessidade de aceleração. A IDC prevê que mais de 50% dos sistemas de IA generativa não serão acelerados em 2024. Portanto, qualquer pessoa pode começar a implementar uma infraestrutura de IA generativa com servidores e sistemas de rede padrão. GPUs também estão disponíveis para aqueles que precisam delas. Há várias opções para implementar a infraestrutura de IA, bem como vários tipos de IA generativa, IA, ML e gêmeos digitais que beneficiarão diferentes empresas de maneiras distintas. Há benefícios na execução da IA generativa em servidores padrão, já que as pilhas de software de IA generativa geralmente são compatíveis. As empresas que investem em infraestrutura padrão no local poderão promover as iniciativas de IA generativa mais rápido do que outras. É essencial que as equipes de TI comecem a avaliar os vários algoritmos de IA generativa, IA, ML e gêmeos digitais para identificar quais causam o maior impacto sobre os negócios.

Benefícios

A IA generativa e outros modelos fundamentais estão virando o jogo, elevando as tecnologias assistivas a um nível inédito e trazendo recursos avançados para os usuários não técnicos. A IA generativa tem o potencial de aumentar a eficiência e a produtividade, abrir novas oportunidades de crescimento, reduzir custos e oferecer uma vantagem competitiva para as empresas que a utilizam.

A criação de sua própria infraestrutura de IA generativa inicia a integração dessa tecnologia inovadora às operações de negócios e aprimora o conhecimento especializado no local nos conjuntos de tecnologia de IA generativa. A priorização dos investimentos em tecnologia que criam a infraestrutura inicial de IA generativa no local, com base em servidores empresariais e sistemas de rede Ethernet padrão, criará uma vantagem de time-to-market para as empresas que aproveitarem essa tecnologia transformadora.

Considerações

Aja agora para aproveitar a IA generativa

Há empolgação em torno da IA generativa, dados os resultados impressionantes que o ChatGPT e outros modelos oferecem. A IA generativa agrega valor, mas esse valor varia de acordo com a fonte de dados exclusivos e os algoritmos implementados. Salas de diretoria, investidores e executivos farão perguntas e buscarão entender como a IA generativa pode ajudar nos negócios.

Para as empresas que capturaram grandes quantidades de dados exclusivos não estruturados, a IA generativa promete criar um conteúdo original a partir dos dados existentes, o que deve ajudar a reestruturar a organização para garantir a inovação contínua. Uma abordagem de engatinhar, caminhar e correr faz sentido para entender o que a IA generativa pode trazer para os negócios e como seguir em frente.

Como criar sua infraestrutura inicial de IA generativa em Ethernet

Para as cargas de trabalho de porte empresarial, os sistemas padrão oferecem o desempenho necessário para iniciar a jornada rumo à IA generativa. Além disso, basear a infraestrutura de IA generativa em servidores e sistemas de rede Ethernet padrão permite usar sistemas operacionais, ferramentas de gerenciamento e ferramentas de gerenciamento de rede empresariais. Depois que os requisitos de

computação para LLMs que beneficiam os negócios forem compreendidos, o desempenho de computação poderá ser aprimorado com a seleção certa de aceleradores de IA generativa e IA. A questão fundamental é que a infraestrutura de IA precisa ser bem arquitetada. Um fabric bem arquitetado pode comportar de dezenas a milhares de nós de computação de IA.

Embora possa haver diferentes opções de sistema de rede para as cargas de trabalho de IA generativa, a opção universal, aberta e preferida por vários fornecedores é o sistema de rede Ethernet. As implementações iniciais de IA generativa podem ser compatíveis com o sistema de rede Ethernet padrão disponível atualmente para os clusters de IA generativa.

Como criar uma infraestrutura de IA generativa com Ultra Ethernet

À medida que cada empresa inicia as fases de caminhar e correr no desenvolvimento da IA generativa, pode fazer sentido criar a própria infraestrutura de IA generativa de scale-out. A criação de uma infraestrutura de IA generativa de scale-out exige duas adições importantes: aceleradores de IA de data center e sistema de rede de IA. Na infraestrutura típica de IA generativa de scale-out, oito GPUs de data center são implementadas em cada servidor e, para cada GPU, há uma NIC ou DPU de alta velocidade implementada para oferecer um sistema de rede de alto desempenho.

"Prevê-se que o mercado de switches Ethernet de datacenter para IA generativa no segmento empresarial cresça a um CAGR de 158,2%, de US$ 41,9 milhões em 2023 para US$ 1,0 bilhão em 2027", Vijay Bhagavath, IDC.

Um dos principais requisitos da infraestrutura de IA de scale-out é o sistema de rede de alto desempenho. Para LLMs de IA generativa, o gargalo no processamento é o tempo que os dados passam na rede. Para algumas cargas de trabalho, o tempo na rede pode ser até 60% do tempo de processamento de um LLM, o que deixa a infraestrutura de computação ociosa conforme os dados se movem entre clusters de computação. Para a rede de IA, há um sistema de rede aprimorado que está disponível agora e é entregue por meio do Ultra Ethernet Consortium, que promete uma interconexão com o mesmo desempenho das redes de supercomputação, escalável para o data center em nuvem e tão econômico e universal quanto a Ethernet. O sistema de rede de IA é essencial para lidar com o crescimento em grande escala das demandas de rede de IA generativa e HPC. A boa notícia é que o Ultra Ethernet Consortium é compatível com a maioria dos fornecedores de switch Ethernet.

Para o desempenho, são necessárias três tecnologias centrais: SerDes de alta velocidade, PHYs e Optics. Essas três tecnologias são usadas em Ethernet e outras tecnologias de sistema de rede. Portanto, essencialmente, não há vantagem de desempenho para nenhuma tecnologia específica de sistema de rede. Para obter o mais alto desempenho de Ethernet, a InfiniBand Trade Association lançou a iniciativa RDMA over Converged Ethernet (RoCE) e definiu o protocolo RoCE. O RoCE é compatível com switches de data center padrão, e há melhorias adicionais chegando ao mercado para impulsionar o desempenho, como os switches Ethernet de alto radix, os switches de corte, o balanceamento de carga e os links com maior largura de banda de até 800 GbE (4 de 200 GbE).

Os testes iniciais de LLMs de IA generativa podem fornecer percepções antecipadas sobre os benefícios que a IA generativa pode trazer aos negócios e auxiliar na criação de uma estratégia para LLM de IA generativa, bem como quais tipos de infraestrutura seriam necessários. Essencialmente, a pilha de software impulsiona os requisitos de semicondutores para a próxima etapa da evolução da IA generativa empresarial. Entender a pilha de software ajudará na implementação de uma infraestrutura de hardware otimizada.

Rebalanceamento da infraestrutura local versus externa à medida que os custos dos semicondutores se estabilizam

À medida que aumenta a oferta de GPUs de data center, mais fornecedores oferecerão essas GPUs, mais aceleradores de IA estarão disponíveis e mais capacidade de processamento de IA generativa estará disponível para as implementações no local. Paralelamente, os gargalos nos provedores de serviços em nuvem desaparecerão e espera-se que os custos se estabilizem. Quando isso acontecer, em três a cinco anos, o rebalanceamento das cargas de trabalho de IA generativa entre a infraestrutura no local e a infraestrutura em nuvem otimizará o CAPEX e o OpEx.

Conclusão

A IA generativa é a tecnologia inovadora de IA. As empresas precisarão ter uma estratégia/plano de IA generativa, que deve começar agora para as cargas de trabalho de porte empresarial a fim de continuar a jornada de integração dessa tecnologia inovadora às operações empresariais.

A demanda é alta, o que eleva os preços dos componentes e dos provedores de serviços em nuvem. Ao mesmo tempo, a IDC prevê que mais de 50% dos sistemas de IA generativa não serão acelerados em 2024. Portanto, qualquer pessoa pode começar a implementar sua infraestrutura de IA generativa com servidores e sistemas de rede padrão. GPUs também estão disponíveis para aqueles que precisam delas. Há várias opções para implementar a infraestrutura de IA, bem como vários tipos de IA generativa, IA, ML e gêmeos digitais que beneficiarão diferentes empresas de maneiras distintas.

> "O mercado continuou a subestimar o crescimento da IA generativa, e a IDC espera observar um crescimento robusto na infraestrutura e nos semicondutores de IA generativa", Brandon Hoff, IDC.

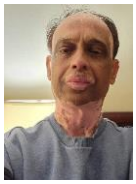
A previsão da IDC é que as empresas trarão de volta dados da nuvem para o processamento da IA generativa a fim de reduzir os custos de OpEx. As empresas começarão o desenvolvimento e os testes com IA generativa em hardware de sistema de rede Ethernet e computação padrão e investirão à medida que aprenderem quais LLMs funcionam para os negócios e o valor que podem extrair de seus dados exclusivos.

A criação da infraestrutura para testes de LLM de IA generativa em servidores prontos para uso e redes Ethernet empresariais revelará o valor da IA generativa para a empresa.

Sobre os analistas

***Brandon Hoff,** diretor de pesquisa, Enabling Technologies: Networking and Comm*

Brandon Hoff lidera a infraestrutura de sistema de rede e comunicações da IDC na equipe Enabling Technologies da empresa. O Sr. Hoff cobre tendências tecnológicas, cargas de trabalho, produtos, fornecedores, cadeia de suprimentos e estratégias de adoção por usuários finais em TI empresarial e data centers de provedores de serviços da Web, em nuvem e de telecomunicações.



***Vijay Bhagavath,** vice-presidente de pesquisa, Cloud and Datacenter Networks*

Vijay Bhagavath fornece liderança inspiradora e acionável e percepções pragmáticas sobre os mercados e as tecnologias de nuvem e sistema de rede de data center. Vijay tem uma compreensão profunda do mercado geral de sistema de rede, tecnologias, roteiros de produtos, diferenciação competitiva e estratégias de implantação, o que lhe permite fornecer comentários e orientações perspicazes para fornecedores, provedores de serviços em nuvem, compradores e profissionais de TI empresarial.

MENSAGEM DO PATROCINADOR

IDC Custom Solutions

O conteúdo deste artigo foi adaptado a partir de uma pesquisa existente da IDC, publicada em www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com