**Infraestrutura de IA**

**Introdução**

Como profissional de TI, você é responsável por implantar e manter a espinha dorsal dos recursos de IA da sua empresa. As configurações tradicionais de data center não são mais suficientes para atender às demandas dos modelos modernos de IA, que exigem imenso poder de processamento, redes de baixa latência e soluções de armazenamento de alto rendimento. A complexidade da integração desses componentes pode ser avassaladora. Uma infraestrutura mal projetada pode levar a problemas que ameaçam a capacidade de atingir os objetivos de negócios da sua organização.

Considere um cenário em que sua empresa está desenvolvendo uma nova aplicação baseada em IA. O sucesso dessa aplicação depende da sua capacidade de implantar uma infraestrutura de IA capaz de lidar com dados em larga escala e modelos complexos sem gargalos. O desafio é claro: como projetar e gerenciar uma infraestrutura de IA confiável, escalável e econômica?

Este treinamento foi desenvolvido especificamente para profissionais como você, que estão na vanguarda da implantação e do gerenciamento de infraestrutura de IA. Você aprenderá sobre os principais componentes das infraestruturas de IA tradicionais e modernas, incluindo soluções de computação, rede e armazenamento. Você também explorará os clusters de IA Nexus HyperFabric da Cisco, que oferecem uma abordagem simplificada e escalável para o gerenciamento de cargas de trabalho de IA.

Ao final deste treinamento, você será capaz de compreender e analisar os desafios de design de uma infraestrutura de IA que maximize a eficiência operacional, reduza custos e aprimore a vantagem competitiva da sua empresa. O treinamento o equipará com as habilidades necessárias para lidar com as complexidades da infraestrutura de IA permitindo que você apoie as iniciativas de IA da sua organização de forma eficaz.

**Infraestrutura de IA tradicional**

A infraestrutura de IA tradicional difere das infraestruturas de IA usadas hoje, principalmente em termos de escalabilidade, poder de processamento e capacidade de lidar com dados em larga escala e modelos complexos. A tabela a seguir ilustra diversas características da infraestrutura de IA tradicional:

| Infraestrutura de IA tradicional | Descrição |
| --- | --- |
| Hardware | O uso de CPUs foi transferido para GPUs em modelos de IA tradicionais mais recentes. |
| Os servidores geralmente são encontrados em implantações locais e geralmente exigem altos investimentos de capital e custos de manutenção. |
| Software | A IA tradicional geralmente usa software proprietário projetado apenas para uma tarefa específica. |
| Os algoritmos são muito mais simples do que as soluções de aprendizado profundo nos ambientes de IA modernos de hoje. |
| Todas as otimizações são tradicionalmente feitas manualmente. |
| Gestão de Dados | A quantidade de dados armazenados e processados ​​é menor em comparação aos requisitos atuais devido às limitações de armazenamento e processamento. |
| Muitas vezes, o tratamento de dados é um processo manual. |

A infraestrutura tradicional de IA pode ser dividida em três categorias: computação, rede e armazenamento. Para alcançar o sucesso de uma aplicação habilitada para IA, os administradores precisam implantar os recursos de hardware e software corretos, redes sem perdas de baixa latência e uma configuração de armazenamento que suporte as necessidades da aplicação de IA.

Texto, Carta

O conteúdo gerado por IA pode estar incorreto.

**Compute**

Servidores, ou nós, especialmente projetados são a base da infraestrutura tradicional de IA. Eles são construídos em servidores poderosos com processadores especializados chamados GPUs (Unidades de Processamento Gráfico). Descobriu-se nos últimos anos que as GPUs não são apenas ideais para acelerar placas gráficas e espaços renderizados em 3D, mas também para calcular somas algorítmicas (vetoriais) resultantes de processos de aprendizado profundo.

Esses nós de computação normalmente abrigam várias GPUs (4, 8 ou 16), cada uma com uma placa de rede de alta velocidade de 400 ou 800G e um switch interno para transferência de dados entre GPUs por meio de um backplane compartilhado. Um nó de computação com oito GPUs normalmente suporta oito placas de rede de 400G ou 800G, dependendo do silício usado no hardware de comutação.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Cada GPU se conectará a oito switches de folha, forçando um padrão de design de malha completa entre GPUs e folhas, formando um "cluster" ou "trilho" dentro da malha. Clusters de clusters também são empregados para criar malhas em hiperescala, às vezes compostas por dezenas de milhares de GPUs. Essa quantidade de GPUs é encontrada em modelos de parâmetros multibilionários, como o modelo de linguagem de grande porte da OpenAI, ChatGPT (175 bilhões), ou nos carros autônomos da Tesla.

**Rede**

A largura de banda óptica é um fator crítico para cargas de trabalho de IA/ML, pois influencia diretamente a velocidade, a eficiência e a escalabilidade dos processos de transferência de dados. Os modelos de IA têm padrões de tráfego variados que afetarão o design da infraestrutura. Por exemplo, alguns ciclos de aprendizado da IA ​​tradicional podem levar dias ou até semanas para serem concluídos com grandes conjuntos de dados. Quando a comunicação entre os clusters de servidores envolvidos nos ciclos de aprendizado apresenta alta latência ou perdas de pacotes, o processo pode levar muito mais tempo para ser concluído ou, em alguns casos, falhar.

Embora data centers de uso geral possam ter um bom desempenho com uma taxa de sobreassinatura de 20:1 (onde mais dispositivos compartilham a largura de banda disponível do que a rede pode suportar simultaneamente), o treinamento em IA normalmente requer uma taxa de assinatura de 1:1. No entanto, uma taxa de 4:1 é comum em alguns tipos de cargas de trabalho de IA. Maior largura de banda permite comunicação e processamento mais rápidos, resultando em melhor desempenho em aplicações de IA/ML. Ópticas 400G e 800G aumentam a eficiência da interconectividade entre servidores, switches e armazenamento, garantindo o desempenho ideal de toda a infraestrutura.

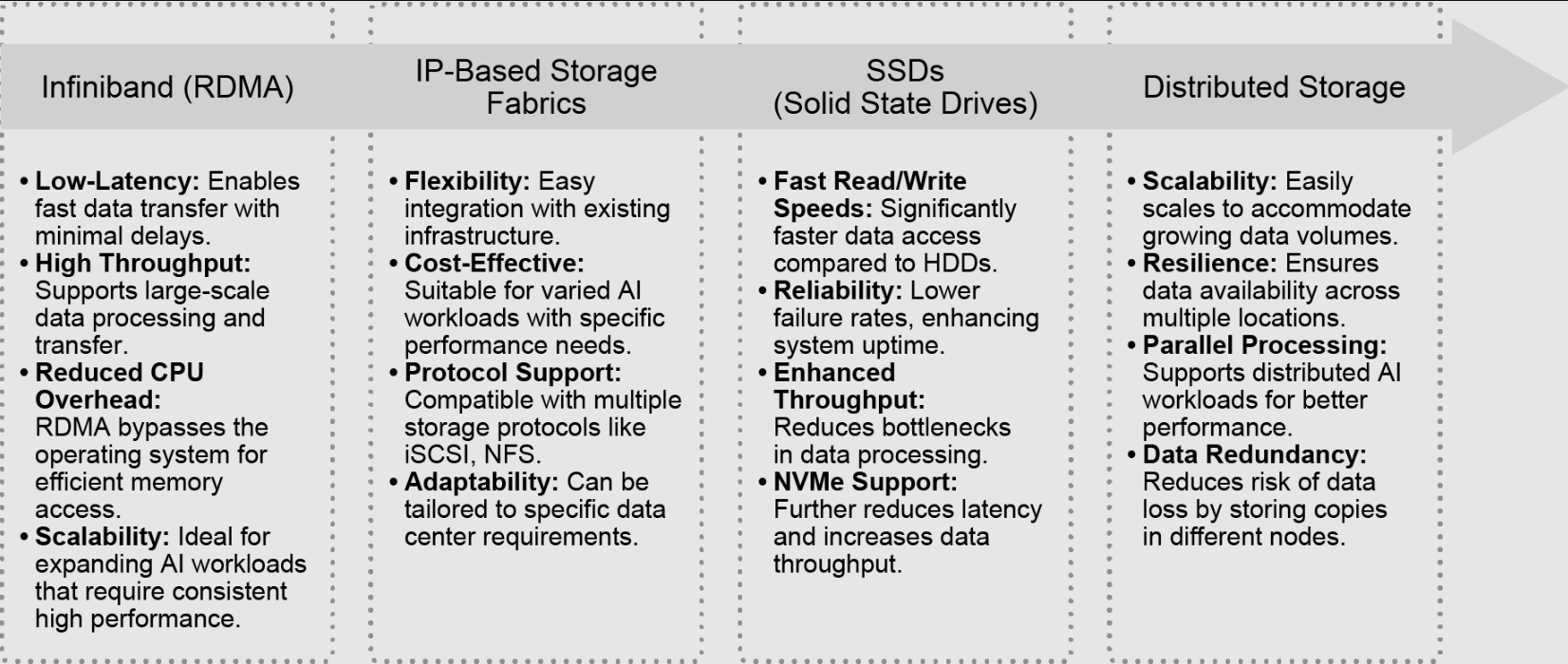
**Armazenar**

A eficiência do treinamento e da inferência de modelos de IA depende fortemente da infraestrutura de armazenamento. As soluções tradicionais de armazenamento de IA precisam não apenas lidar com grandes quantidades de dados, mas também fornecer alta taxa de transferência e baixa latência para atender às demandas de recursos de computação e rede. O sistema de armazenamento impacta tanto a velocidade de acesso aos dados quanto a escalabilidade geral da infraestrutura de IA.

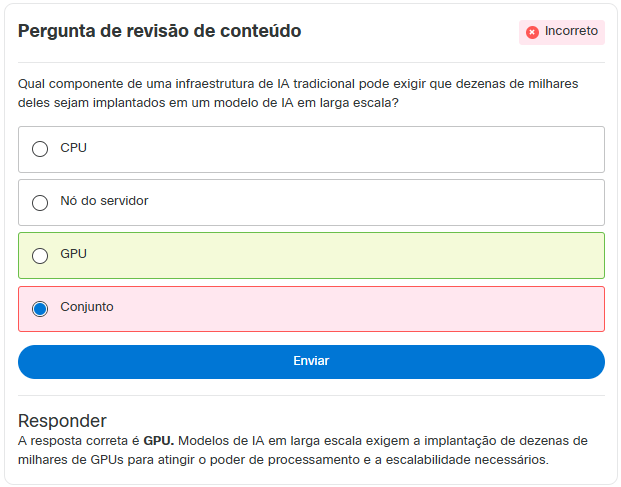
O InfiniBand com Acesso Direto à Memória Remoto (RDMA) é uma escolha comum para armazenamento de IA de alto desempenho devido à sua capacidade de reduzir a sobrecarga da CPU e fornecer transferência de dados de baixa latência e alta taxa de transferência. O RDMA permite o acesso direto à memória de um computador para a de outro sem envolver o sistema operacional de nenhum deles. O RDMA facilita o processamento de dados mais rápido e eficiente, essencial para cargas de trabalho de IA em larga escala.

No entanto, dependendo dos requisitos específicos dos modelos de IA e da natureza dos dados, alguns data centers tradicionais de IA podem optar por estruturas de armazenamento baseadas em IP. Esses sistemas podem oferecer maior flexibilidade e, muitas vezes, são mais fáceis de integrar à infraestrutura existente do data center, embora possam apresentar desvantagens em termos de desempenho em comparação com o InfiniBand.

Além da escolha da malha, o tipo de mídia de armazenamento utilizada também é crucial. Para aplicações de IA que exigem acesso rápido a grandes conjuntos de dados, SSDs (unidades de estado sólido) são frequentemente preferidos aos HDDs (unidades de disco rígido) tradicionais devido às suas velocidades de leitura/gravação significativamente mais rápidas. Além disso, unidades NVMe (memória não volátil expressa) de alto desempenho estão sendo cada vez mais utilizadas para aprimorar ainda mais a taxa de transferência de dados e reduzir a latência.



A escolha da arquitetura de armazenamento — seja ela centralizada, distribuída ou híbrida — também desempenha um papel crucial. Arquiteturas de armazenamento distribuído, por exemplo, costumam ser preferidas em infraestruturas de IA tradicionais para garantir escalabilidade e resiliência, permitindo que os dados sejam armazenados e acessados ​​em vários locais.



**Infraestrutura de IA moderna.**

Os modelos modernos de IA utilizam aprendizado profundo e abordagens orientadas por dados, com grandes conjuntos de dados e algoritmos avançados. Portanto, a IA moderna requer uma infraestrutura escalável para se adaptar a aplicações complexas do mundo real, que são sofisticadas demais para os métodos tradicionais de IA.

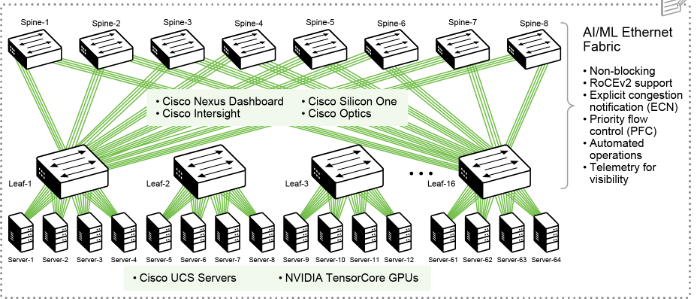
**Características da infraestrutura de IA moderna**

A tabela a seguir ilustra diversas características da infraestrutura de IA moderna:

| Infraestrutura de IA moderna | Descrição |
| --- | --- |
| Hardware | As GPUs são usadas com frequência e em grandes quantidades. |
| TPUs projetadas para treinamento e inferência de modelos de aprendizado profundo são frequentemente usadas para atender às demandas de desempenho do processamento intenso de IA. |
| ASICs projetados para serem o hardware mais rápido e otimizado para uma tarefa específica são comumente usados. |
| Software | Os algoritmos usados ​​em infraestruturas de IA modernas são focados em aprendizado profundo, além do aprendizado de máquina. |
| Os aprimoramentos de aprendizado de máquina permitem que as ferramentas selecionem automaticamente o melhor modelo para uma tarefa específica. |
| Estruturas de código aberto são usadas para permitir desenvolvimento rápido e simplificar implantações. |
| Gestão de Dados | O consumo e o armazenamento de dados são exponencialmente maiores em uma infraestrutura de IA moderna que usa aprendizado profundo do que em um modelo de IA tradicional. |
| A infraestrutura moderna de IA exige automação para o armazenamento e a recuperação de dados. Data lakes e data warehouses são comuns. |
| Implantação e integrações | O uso de serviços de contêiner permite a automação da implantação e a escalabilidade dos serviços conforme a demanda flutua. |
| As APIs são usadas para integração fácil com outros microsserviços |
| A automação é usada para cargas de trabalho que exigem alocação dinâmica de recursos, dimensionamento rápido e configuração de trilhos. |

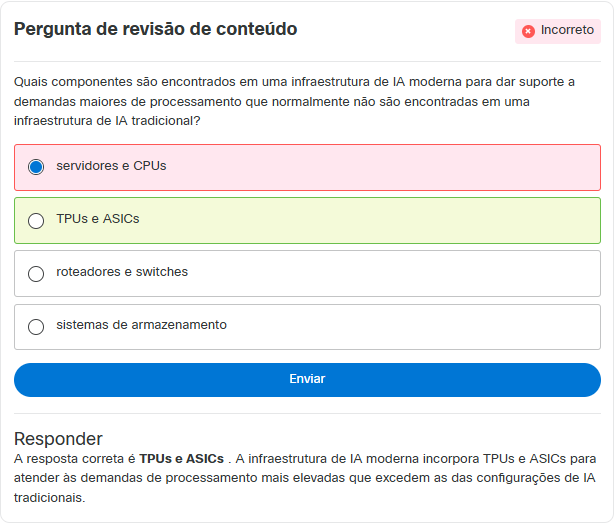
**Preparação do data center para IA/ML**

Para processar cargas de trabalho de IA/ML que envolvem grandes conjuntos de dados, é necessário distribuí-los entre várias GPUs em um cluster de IA/ML. Essa abordagem ajuda a equilibrar a carga por meio do processamento paralelo e a entregar resultados de alta qualidade rapidamente. Para atingir esse equilíbrio, é essencial ter uma rede de alto desempenho, sem bloqueios e com suporte para transporte sem perdas. Além disso, a infraestrutura deve incluir mecanismos de gerenciamento de congestionamento, como notificação explícita de congestionamento (ECN) e controle de fluxo de prioridade (PFC), para suportar o transporte RoCEv2 para transferência de informações de memória para memória da GPU.



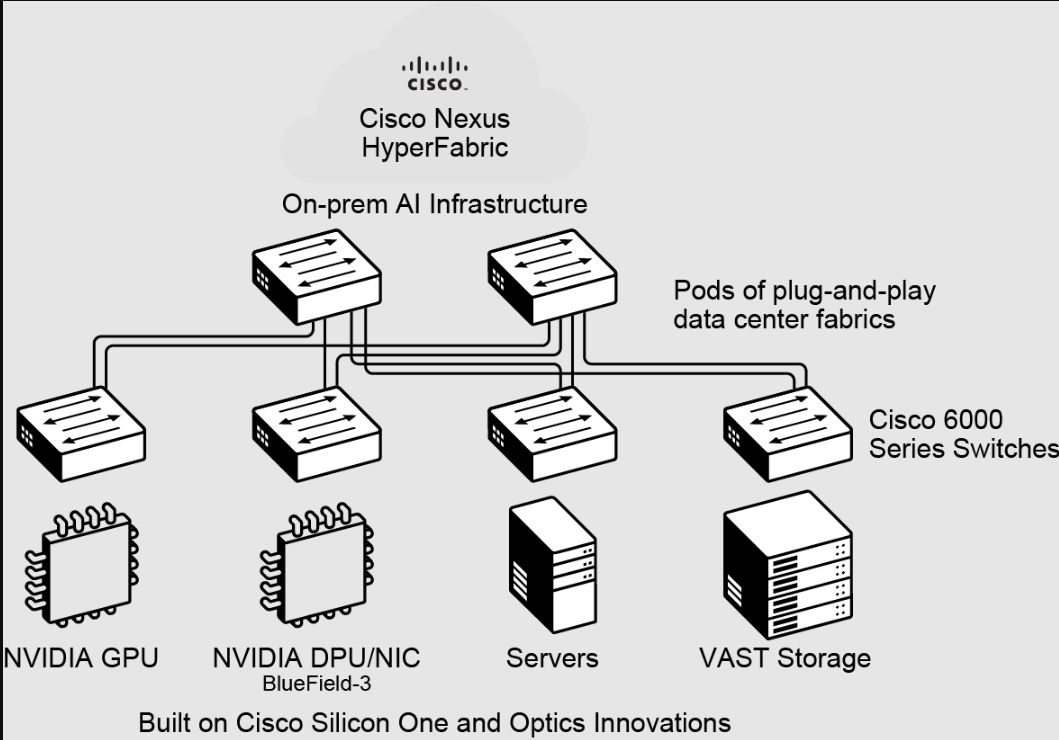
Ao combinar tecnologias líderes do setor da Cisco e da NVIDIA, os componentes de uma infraestrutura de IA moderna incluem o seguinte:

* Infraestrutura Ethernet do data center da Cisco que inclui switches Cisco Nexus série 9000 e roteadores Cisco série 8000
* Cisco Optics, que fornece conexões ópticas de alta velocidade para transmissão eficiente de dados.
* Cisco Silicon One, uma arquitetura de silício de rede flexível e com baixo consumo de energia para redes de data center de IA/ML de alto desempenho que controlam latência e perdas
* Geração Cisco M7 de servidores rack e blade Cisco UCS que permitem desempenho de computação ideal em uma ampla gama de cargas de trabalho com uso intensivo de IA e dados
* Ferramentas de gerenciamento e operações de infraestrutura, como Cisco Networking Cloud com Cisco Nexus Dashboard e Cisco Intersight, monitoramento de experiência digital com Cisco ThousandEyes e análise de telemetria entre domínios com a Cisco Observability Platform
* GPUs NVIDIA Tensor Core otimizadas para cargas de trabalho de IA/ML, usadas em servidores blade e rack Cisco UCS
* NVIDIA BlueField-3 SuperNICs, que são aceleradores de rede especialmente desenvolvidos para fornecer conectividade de rede de alto desempenho entre servidores GPU
* Unidades de processamento de dados (DPUs) NVIDIA BlueField-3 para descarregar, acelerar e isolar funções de rede, armazenamento, segurança e gerenciamento definidas por software em uma infraestrutura de nuvem
* NVIDIA AI Enterprise, um pacote de software de estruturas, modelos pré-treinados e ferramentas de desenvolvimento, bem como novos microsserviços NVIDIA NIM, para uma produção de IA mais segura, estável e com suporte.



Clusters de IA Cisco Nexus HyperFabric

Muitas organizações buscam simplificar a implantação e as operações em toda a infraestrutura de IA por meio de dados, computação, software, armazenamento e redes de alto desempenho integrados. A solução de cluster de IA Cisco Nexus HyperFabric foi projetada para simplificar as implantações corporativas com gerenciamento de nuvem plug-and-play para infraestrutura nativa de IA.



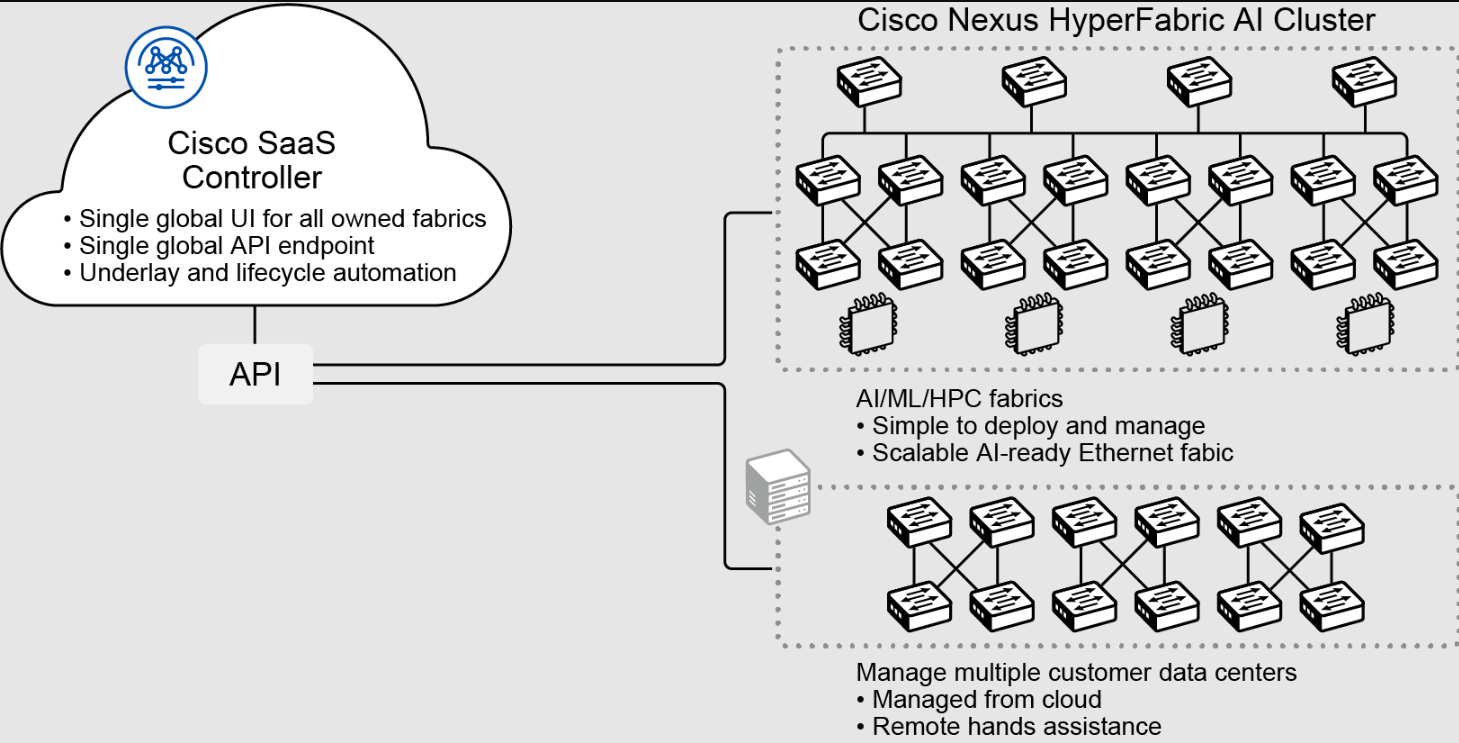
A solução de cluster de IA Cisco Nexus HyperFabric combina a rede nativa de IA da Cisco com a computação acelerada e o software de IA da NVIDIA, além de um armazenamento de dados robusto. A solução orienta as organizações em todo o processo do ciclo de vida da infraestrutura, desde o design até o pedido, implantação, configuração, validação e operações contínuas para implantações de IA.

O cluster Cisco Nexus HyperFabric AI compreende os seguintes componentes:

* Switches Cisco série 6000 para spine e leaf que oferecem desempenho de estrutura Ethernet de 400G e 800G
* Família Cisco Optics de módulos QSFP-DD para oferecer escolha ao cliente e entregar densidades super altas
* Software NVIDIA AI Enterprise para otimizar o desenvolvimento de cargas de trabalho de IA generativa de nível de produção
* Microsserviços de inferência NVIDIA NIM que aceleram a implantação de modelos de base, garantindo ao mesmo tempo a segurança dos dados
* GPUs NVIDIA Tensor Core projetadas para potencializar cargas de trabalho de IA generativa com desempenho e recursos de memória revolucionários
* Unidade de processamento de dados NVIDIA BlueField-3, processador DPU e BlueField-3 SuperNIC para acelerar redes de computação de IA, acesso a dados e cargas de trabalho de segurança
* A plataforma de dados VAST, que oferece armazenamento unificado, um banco de dados e um mecanismo de função orientado a dados desenvolvido para IA

**Recursos do Cisco Nexus HyperFabric**

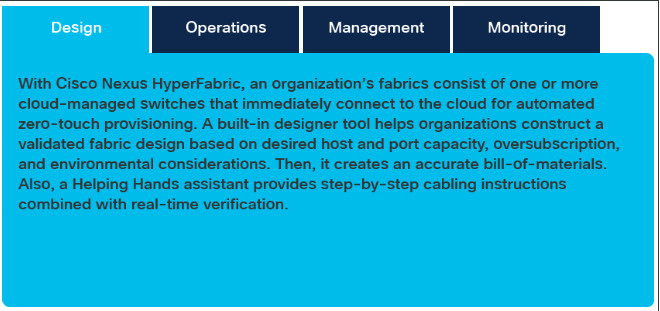
O Cisco Nexus HyperFabric é uma oferta de data center de malha de rede como serviço, desenvolvida especificamente para esse fim. Utilizando um controlador de nuvem gerenciado pela Cisco, a solução permite que os clientes projetem, implantem e gerenciem facilmente qualquer número de malhas, habilitadas ou não para IA, localizadas em qualquer lugar, abrangendo data centers primários, instalações de colocation e sites de ponta de centro de distribuição.



O Cisco Nexus HyperFabric reinventa o ciclo de vida das operações de TI do data center simplificando cada etapa do processo.

**Explore o ciclo de vida simplificado das operações de TI**

Clique nos estágios do ciclo de vida para saber como o Cisco Nexus HyperFabric simplifica cada estágio.



Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

**Integração de SDN com clusters de IA Cisco Nexus HyperFabric**

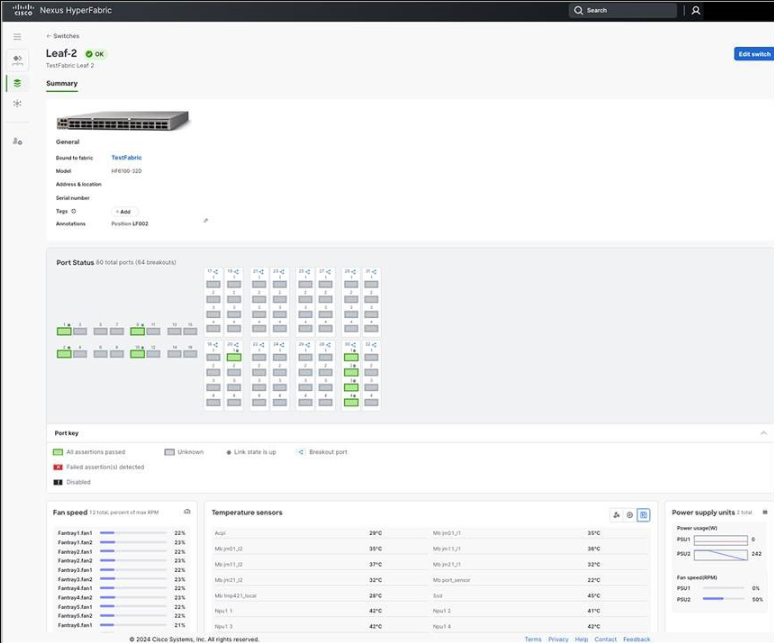
Os clusters de IA do Cisco Nexus HyperFabric integram-se perfeitamente à Rede Definida por Software (SDN) para aumentar a agilidade, a escalabilidade e a eficiência das cargas de trabalho de IA.

A SDN fornece gerenciamento e orquestração de rede centralizados, permitindo ajustes nas configurações de rede em tempo real. O Cisco Nexus HyperFabric é um controlador baseado em nuvem que oferece fabric-as-a-service, um ponto único de administração para todas as fabrics, independentemente da localização. Esse recurso ajuda a manter as demandas de alto desempenho de aplicações de IA nas quais o fluxo ininterrupto de dados entre nós e data centers é vital.

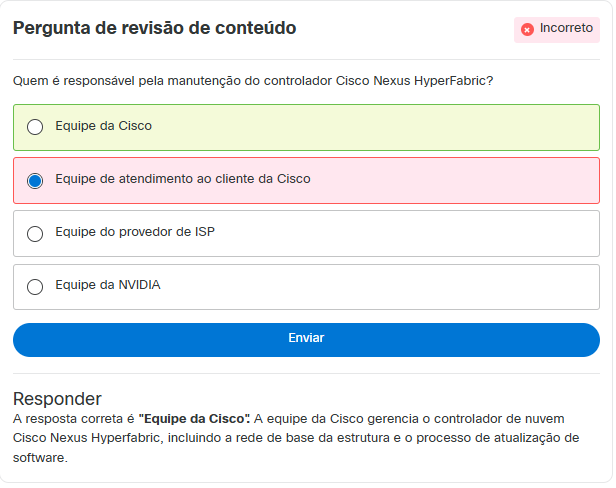
Interface gráfica do usuário

O conteúdo gerado por IA pode estar incorreto.

A SDN aprimora a escalabilidade do Cisco Nexus HyperFabric, permitindo ajustes dinâmicos nos recursos de rede, garantindo que a infraestrutura atenda às crescentes demandas sem comprometer o desempenho. Além disso, os recursos avançados de gerenciamento de tráfego com SDN permitem roteamento inteligente que reduz a latência e melhora o desempenho geral dos processos de treinamento e inferência de IA.



Além das melhorias de desempenho, a SDN fortalece a segurança do Cisco Nexus HyperFabric, permitindo políticas de segurança consistentes e mitigação de ameaças em tempo real em toda a rede. A integração da SDN também estende os recursos do Cisco Nexus HyperFabric para ambientes de nuvem e edge, oferecendo suporte a operações contínuas em aplicações de IA distribuídas.



Resumo

Considere as seguintes questões ao aplicar os conceitos que aprendeu em seu próprio ambiente:

* Como a integração de GPUs, TPUs e ASICs em sua infraestrutura atual pode melhorar o desempenho e a escalabilidade de seus modelos de IA?

Considere como esses dispositivos são usados ​​para lidar com grandes conjuntos de dados e algoritmos complexos de aprendizado profundo, utilizando estruturas de código aberto e sistemas automatizados de gerenciamento de dados.

* Quais benefícios sua organização poderia obter ao adotar os Cisco Nexus HyperFabric AI Clusters, especialmente em termos de implantação, gerenciamento e monitoramento de cargas de trabalho de IA?

Veja como os Cisco Nexus HyperFabric AI Clusters simplificam a implantação e as operações por meio de serviços gerenciados em nuvem, oferecem provisionamento zero-touch, monitoramento em tempo real e integração perfeita de SDN para aumentar a agilidade, a escalabilidade e a segurança das cargas de trabalho de IA.

* Como soluções modernas de infraestrutura de IA, como data lakes e recuperação automatizada de dados, podem melhorar seus processos de tratamento de dados?

Considere como as infraestruturas tradicionais de data center dependem principalmente de CPUs e servidores locais, com otimizações manuais e recursos limitados de gerenciamento de dados.