**Soluções de Recursos de Computação**

Introdução

A Cisco fornece soluções de recursos de computação, incluindo soluções de infraestrutura convergente e hiperconvergente e soluções específicas para inteligência artificial (IA), como FlashStack Data Center, Nutanix GPT-in-a-Box e Run:ai.

**Visão geral das soluções de infraestrutura hiperconvergente da Cisco**

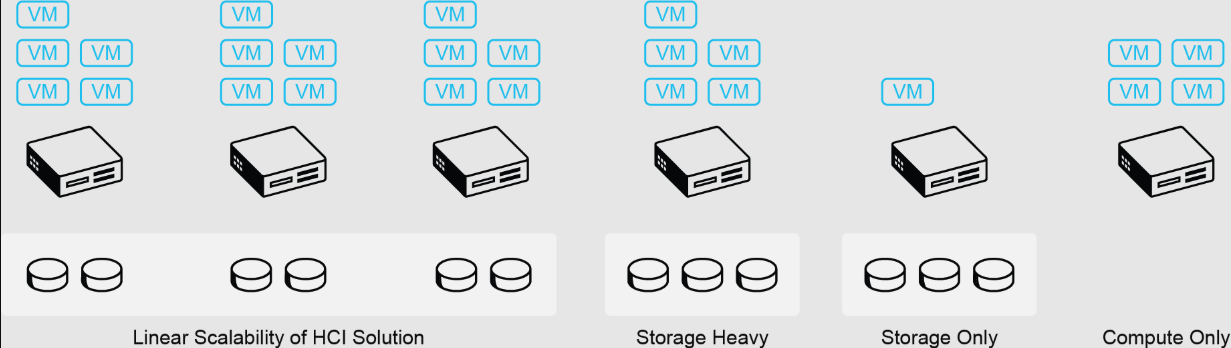
Soluções hiperconvergentes integradas oferecem uma infraestrutura de TI robusta, escalável e eficiente que simplifica a gestão do data center. As abordagens conjuntas da Cisco e da Nutanix permitem suporte abrangente, escalabilidade perfeita e proteção avançada de dados, o que as torna ideais para empresas modernas que buscam otimizar suas operações de TI e oferecer suporte a uma ampla gama de aplicações.

**Introdução às Soluções Hiperconvergentes**

Um dos aspectos mais desafiadores e caros do projeto de uma solução tradicional de três camadas é saber quantos recursos serão necessários ao longo do ciclo de vida da solução.

A infraestrutura hiperconvergente, por outro lado, é dimensionada de forma linear e previsível devido à sua arquitetura central, que redistribui dados automaticamente à medida que novos nós são adicionados, usando a ferramenta de expansão de nós integrada.

Como você pode ver na figura a seguir, você pode adicionar máquinas virtuais (VMs) com alto nível de armazenamento, somente armazenamento ou somente computação, de acordo com suas necessidades.



Você deve considerar o dimensionamento quando encontrar as seguintes situações:

* Se os recursos de armazenamento acabarem mais rápido que a computação, adicione nós com uso intenso de armazenamento.
* Se for necessário mais armazenamento, adicione nós somente de armazenamento (sem precisar licenciar um hipervisor).
  + Misture e combine nós flash e híbridos dentro do mesmo cluster para que seu design de cluster original não restrinja seus esforços.
* Se os recursos de computação acabarem mais rápido que o armazenamento, você poderá adicionar nós somente de computação para dar suporte a mais VMs sem adicionar armazenamento ao cluster.

A capacidade de comprar o que você precisa no curto prazo e comprar mais recursos conforme necessário elimina muitas suposições e atritos ao comprar infraestrutura.

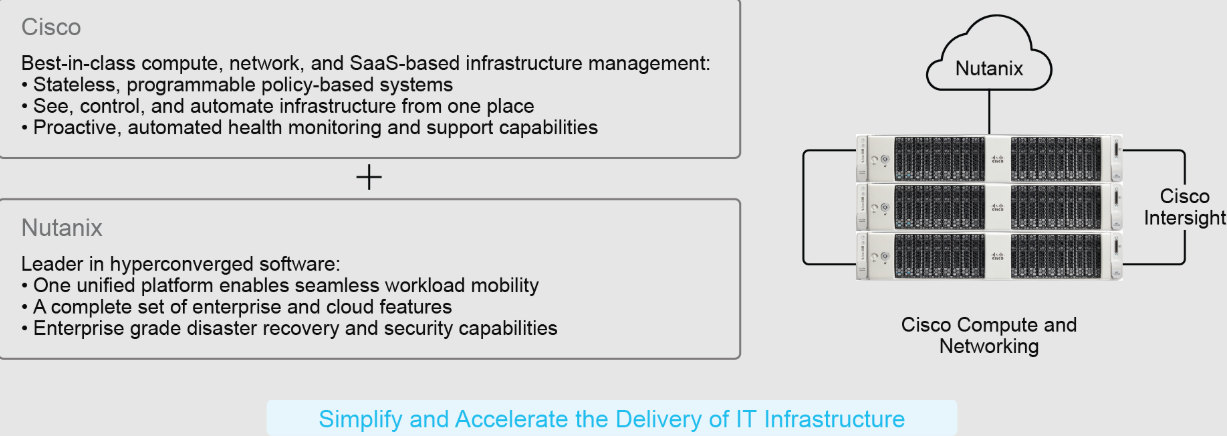
**Infraestrutura hiperconvergente simplifica a multinuvem híbrida**

A Cisco e a Nutanix combinam excelência na unificação de gerenciamento, infraestrutura e plataformas para simplificar as operações e estabelecer uma base ágil para dar suporte a qualquer aplicativo, independentemente de sua localização.

A Cisco oferece o melhor gerenciamento de infraestrutura baseado em computação, rede e software como serviço (SaaS). As soluções da Cisco contam com sistemas programáveis, sem estado e baseados em políticas, que permitem aos usuários visualizar, controlar e automatizar sua infraestrutura a partir de um único local. A Cisco também oferece recursos proativos e automatizados de monitoramento e suporte à saúde.

A Nutanix, por outro lado, é líder em software hiperconvergente. A plataforma unificada da Nutanix permite mobilidade perfeita da carga de trabalho e inclui um conjunto completo de recursos empresariais e de nuvem. O software da Nutanix também oferece recursos de segurança e recuperação de desastres de nível empresarial.

Conforme descrito na figura a seguir, uma combinação de ambas as soluções visa simplificar e acelerar a entrega de infraestrutura de TI.

****

**Abordagem unificada para benefícios da multinuvem híbrida**

A Cisco firmou uma parceria com a Nutanix para fornecer uma solução multicloud híbrida baseada na liderança combinada em gerenciamento de aplicativos, dados e infraestrutura. A Plataforma de Nuvem Nutanix, baseada em sua infraestrutura hiperconvergente líder do setor, agora está validada, certificada e integrada à infraestrutura de servidores da Cisco para ajudar as empresas a acelerar sua jornada multicloud híbrida.

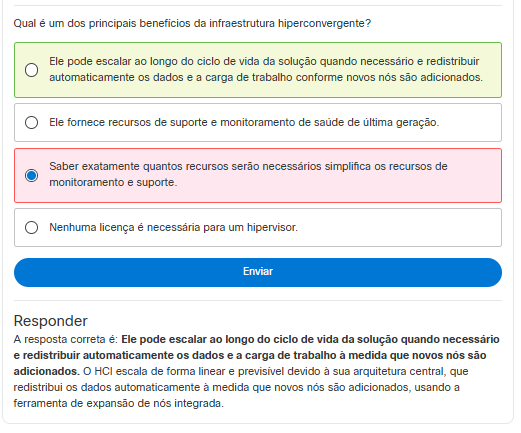
O Cisco Compute Hyperconverged com Nutanix é criado, gerenciado e suportado de forma holística para oferecer uma experiência mais integrada, promover a inovação e acelerar a jornada de multinuvem híbrida para os clientes, conforme ilustrado na figura a seguir.

Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

O Cisco Compute Hyperconverged com Nutanix simplifica e acelera a entrega e a operação de infraestrutura e aplicativos em escala global com estas vantagens:

* Esta solução completa é criada, gerenciada e suportada de forma holística e apresenta o melhor modelo operacional em nuvem da categoria.
* Esta solução se adapta aos requisitos dinâmicos de negócios e aplicativos com flexibilidade e opções em servidores Cisco; o que há de mais moderno em tecnologias de acelerador, rede e armazenamento; inovações em SaaS; e a liberdade de se conectar a várias nuvens.
* A inovação pode ocorrer com confiança, sabendo que recursos conjuntos de suporte aumentado e resiliência automatizada previnem e resolvem problemas mais rapidamente.

****

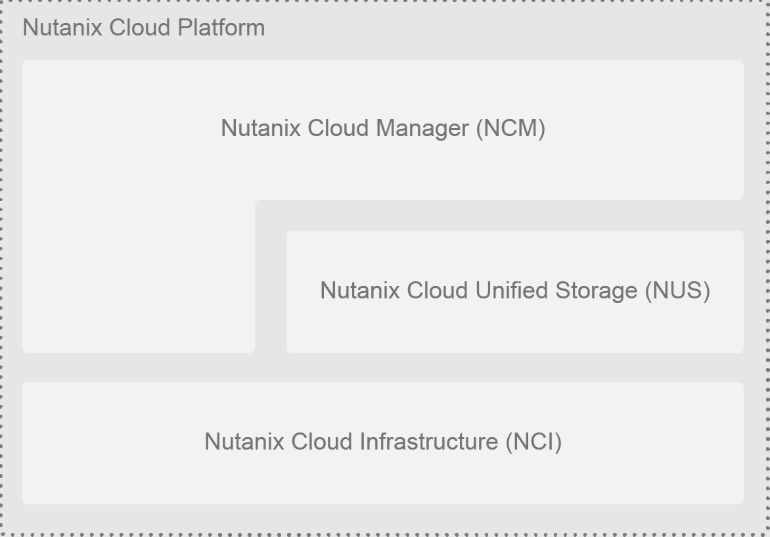
**Componentes da solução hiperconvergente da Cisco**

O Cisco Compute Hyperconverged com Nutanix é uma plataforma de software segura, resiliente e autocorretiva que permite que você crie sua infraestrutura multicloud híbrida para dar suporte a todos os tipos de cargas de trabalho e casos de uso em nuvens públicas e privadas.

O Cisco Compute Hyperconverged com Nutanix tem os seguintes componentes:

* Nutanix Cloud Platform (NCP), que inclui Nutanix Cloud Infrastructure (NCI), Nutanix Unified Storage (NUS) e Nutanix Cloud Manager (NCM)
* Suporte a hipervisor: Nutanix Acropolis Hypervisor (AHV) e VMware vSphere
* Um ou mais nós Cisco UCS suportados
* Gerenciamento de sistemas: Cisco Intersight Infrastructure Service

A hierarquia e a dependência dos componentes do NCP são ilustradas na figura a seguir.



**O NCM** é um plano de controle de software para provisionamento, operação, automação e governança de cargas de trabalho em nuvens.

**O NUS** é uma solução de armazenamento distribuída e definida por software que oferece a escala necessária para que as organizações atendam a qualquer carga de trabalho em qualquer lugar. Ele possibilita uma plataforma de armazenamento unificada, capaz de fornecer serviços de armazenamento em blocos, arquivos e objetos.

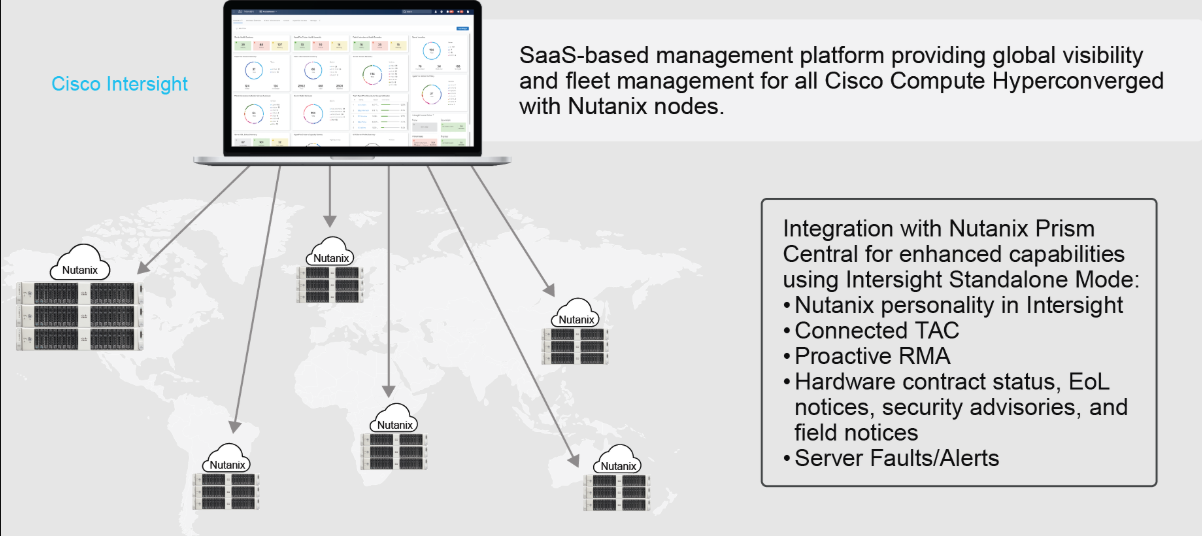
**A NCI** fornece um conjunto completo de software para unificar sua infraestrutura de nuvem híbrida e habilita as seguintes funcionalidades:

* Armazenamento escalável
* Hipervisor Nutanix Acropolis
* Infraestrutura hiperconvergente avançada
* Rede virtual
* Recuperação de desastres
* Serviços de contêineres
* Segurança de dados e rede

**A computação do Cisco Unified Computing System (UCS)** oferece uma arquitetura de servidor sem estado que combina computação e rede em uma única plataforma para potencializar seus aplicativos. A computação do Cisco UCS oferece as seguintes vantagens:

* Gestão simplificada
* Programabilidade completa da interface de programação de aplicativos (API)
* Configuração sem estado por meio de políticas e perfis lógicos
* Vários formatos e periféricos para atender aos requisitos do cliente

**O Cisco Intersight** é uma plataforma de gerenciamento baseada em SaaS que fornece visibilidade global e gerenciamento de frota para todos os nós Cisco Compute Hyperconverged com Nutanix.



A integração do Intersight com o Nutanix Prism Central permite estes recursos aprimorados usando o modo autônomo do Intersight:

* Personalidade da Nutanix na Intersight
* Cisco TAC conectado
* Autorização Proativa de Devolução de Material (RMA)
* Status do contrato de hardware, avisos de fim de vida útil (EOL), avisos de segurança e avisos de campo
* Falhas e alertas do servidor

Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

* Centro de Dados FlashStack

A infraestrutura convergente combina recursos de computação, rede, armazenamento, software de gerenciamento e automação em uma solução unificada. Com soluções validadas, a infraestrutura convergente pode simplificar e acelerar a implantação, reduzindo riscos. A infraestrutura convergente consiste em componentes agrupados, pré-testados e pré-validados. Os sistemas de infraestrutura convergente utilizam matrizes de armazenamento e o servidor não possui armazenamento local.

A infraestrutura convergente normalmente inclui os seguintes elementos:

* Calcular
* Armazenar
* Rede
* Hipervisor
* Software de gestão
* Capacidades de automação e orquestração

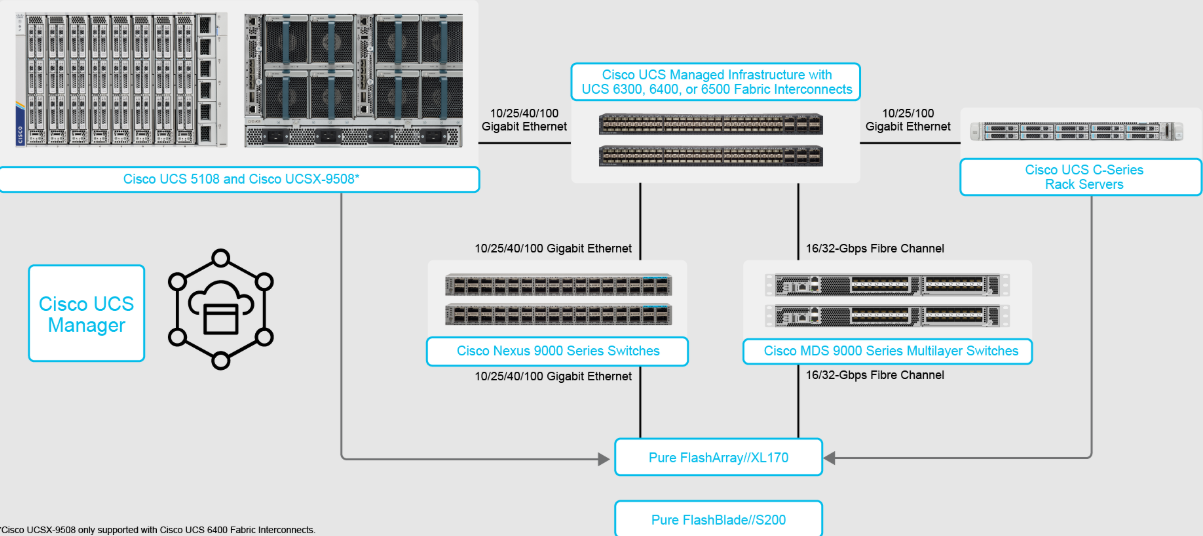
Fornecedores de infraestrutura convergente podem fornecer equipamentos testados, pré-montados e prontos para uso. Alternativamente, o equipamento pode ser montado no local do cliente, juntamente com um documento de arquitetura de referência, que é um guia detalhado sobre a implantação e configuração da solução.

Um exemplo de infraestrutura convergente é um data center FlashStack com tecnologia NVIDIA. Ele incorpora computação acelerada, software essencial de IA e modelos pré-treinados. Essa pilha simplifica a implantação de modelos de IA em diversas aplicações e oferece uma solução abrangente para uma ampla gama de casos de uso.

A arquitetura do FlashStack é construída usando os seguintes componentes de infraestrutura para computação, rede e armazenamento:

* Plataforma modular Cisco UCS X-Series usando nós de computação Cisco UCS X210c M7 com unidades de processamento gráfico (GPUs) NVIDIA
* Switches Cisco Nexus
* Switches Cisco MDS série 9000
* FlashArray de armazenamento puro

A implantação consiste em clusters do Red Hat OpenShift Container Platform implantados no VMware vSphere, instalados em nós de computação Cisco com GPUs NVIDIA. O Cisco Intersight gerencia os nós de computação. A camada de software da plataforma NVIDIA AI, NVIDIA AI Enterprise, alimenta o fluxo de trabalho de inferência.



Todos os componentes do FlashStack são integrados, permitindo que você implante a solução de forma rápida e econômica, eliminando muitos riscos associados à pesquisa, ao design, à construção e à implantação de soluções semelhantes desde o início. Um dos principais benefícios do FlashStack é sua capacidade de manter a consistência em escala.

A solução FlashStack usa o Cisco UCS C-Series ou o Cisco UCS X-Series com os seguintes componentes de hardware:

* Uma combinação das seguintes famílias de servidores:
  1. Chassi Cisco UCS X9508 com qualquer número de nós de computação Cisco UCS X210c M7.
  2. Servidores de rack Cisco UCS C-Series (C220 M7, C240 ​​M7, C245 M8, por exemplo).
* Interconexões de malha Cisco UCS 6454 de 4ª geração para oferecer suporte à conectividade Ethernet de 25 e 100 Gigabits de vários componentes.
* Design de comutação Cisco Nexus 93180YC-FX3 de alta velocidade baseado no sistema operacional Cisco Nexus (NX-OS) para oferecer suporte a conectividade Ethernet de até 100 Gigabits.
* Armazenamento de objetos e arquivos escaláveis ​​Pure Storage FlashBlade//S500 com conectividade Ethernet de 100 Gigabits para a malha de comutação Cisco Nexus.
* Armazenamento Pure FlashArray//XL170 com conectividade Ethernet de 25 Gigabits com a estrutura de comutação Cisco Nexus e conectividade Fibre Channel de 32 Gb com a estrutura de comutação Cisco MDS.

Os componentes do software consistem nos seguintes recursos:

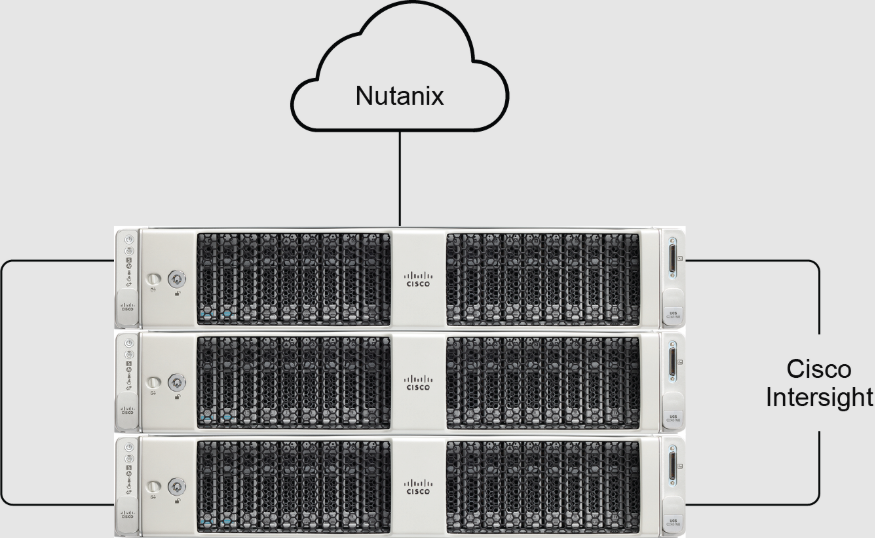
* Plataforma Cisco Intersight para implantar, manter e dar suporte aos componentes do FlashStack.
* Dispositivo virtual Cisco Intersight Assist para ajudar a conectar o Pure Storage FlashArray e o VMware vCenter com a plataforma Cisco Intersight.
* Para clusters virtualizados, o VMware vCenter 8.0 é necessário para configurar e gerenciar a infraestrutura virtual e a integração do ambiente virtual com o software Cisco Intersight.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

**Nutanix GPT-in-a-Box**

A solução Cisco Compute Hyperconverged com Nutanix GPT-in-a-Box elimina a complexidade da adoção de IA generativa. Ela fornece as etapas para a implantação da infraestrutura subjacente para esta solução em um único pacote. Esta solução combina servidores Cisco e operações SaaS com software Nutanix, utilizando os modelos de linguagem de grande porte (LLMs) mais populares para produzir uma plataforma pronta para IA totalmente validada que pode simplificar e impulsionar suas iniciativas de IA do data center até a borda.



A solução oferece três benefícios principais:

* **Simplicidade** : a solução oferece opções de gerenciamento SaaS e local e inclui operações do dia 0 ao dia N, incluindo perfis de serviço para computação, armazenamento e rede personalizados para a Nutanix para ajudar a simplificar e acelerar a implantação de cluster e fornecer melhor desempenho e resiliência.
* **Flexibilidade** : A solução Cisco Compute Hyperconverged com Nutanix aborda aplicações e casos de uso modernos para oferecer múltiplas opções de implantação de servidores Cisco UCS, as mais recentes tecnologias de aceleradores e drives e inovações SaaS de duas potências do setor, incluindo integrações com os principais provedores de nuvem pública. Além disso, a solução incorpora a melhor tecnologia de rede da Cisco, incluindo integrações com a Cisco Application Centric Infrastructure (ACI), para aprimorar o desempenho e a resiliência de cargas de trabalho com uso intensivo de dados em ambientes de nuvem híbrida.
* **Resiliência** : A solução conjunta utiliza apenas componentes de nível empresarial. Sua proteção de sistema aprimorada inclui um modelo de suporte colaborativo, recursos proativos de resiliência e segurança automatizados, além de sistemas de suporte e notas de caso para uma triagem mais rápida. Quando arquivos de log são carregados ou notas de caso são geradas, essas informações são compartilhadas, permitindo uma colaboração aprimorada entre as equipes de suporte para resolver problemas mais rapidamente e proporcionar uma experiência aprimorada ao cliente. A abordagem baseada em políticas minimiza erros humanos e desvios de configuração, o que resulta em implantações de cluster consistentes e confiáveis.

Ele também impõe uma postura geral de segurança por meio de autorizações centralizadas para evitar adulteração de configurações.

Esta arquitetura de referência combina estes elementos:

* Solução definida por software Nutanix GPT-in-a-Box
* Servidores Cisco Compute Hyperconverged C-Series
* GPU NVIDIA L40S
* Modo autônomo Cisco Intersight (sem interconexão de malha externa)
* Gerenciamento de sistemas: Nutanix Prism
* Uma gama dos LLMs mais populares

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

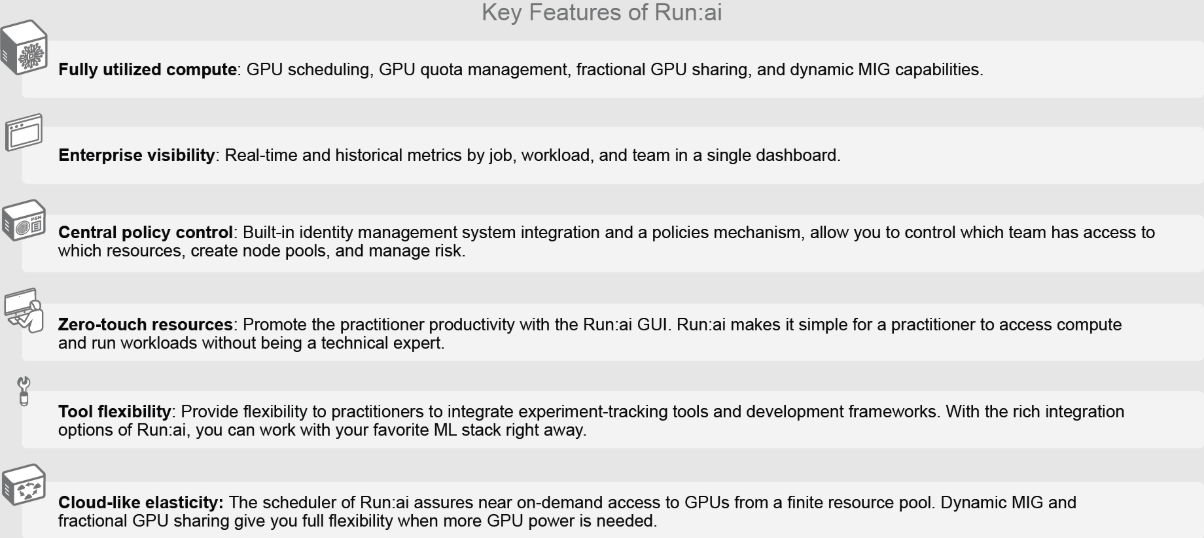
**Nutanix GPT-in-a-Box**

O Run:ai é uma plataforma de orquestração de IA que oferece soluções eficazes para gerenciar e otimizar fluxos de trabalho de IA. Quando integrado ao OpenShift no Cisco UCS X-Series, o Run:ai pode ajudar a otimizar cargas de trabalho de IA e aprendizado de máquina. O OpenShift, uma plataforma baseada em Kubernetes, fornece o ambiente perfeito para implantar e gerenciar o Run:ai e permite a conteinerização e a automação de cargas de trabalho de IA. O Cisco UCS X-Series, uma plataforma de computação modular altamente escalável e flexível, fornece o poder computacional e a capacidade necessários para lidar com tarefas de IA que exigem muitos recursos.

A integração do Run:ai com o OpenShift no Cisco UCS X-Series oferece uma solução para gerenciamento de carga de trabalho de IA. Ela permite que as organizações aloquem recursos dinamicamente, simplifiquem o gerenciamento

de carga de trabalho e acelerem a pesquisa em IA. Com o Run:ai, as empresas podem priorizar tarefas com eficiência, garantir a utilização ideal de recursos e reduzir custos operacionais.

A figura a seguir descreve os principais recursos do Run:ai.



**Capacidade de compartilhamento fracionário de GPU**

O Run:ai pode alocar um contêiner com uma quantidade específica de RAM de GPU. Se o seu código precisar de 4 GB de RAM, você pode enviar uma tarefa especificando a quantidade exata de memória de GPU necessária. Exceder a quantidade de RAM especificada resultará em uma exceção de falta de memória.

Com o recurso de fração da GPU, todas as cargas de trabalho em execução que usam a GPU compartilham a computação em paralelo e, em média, recebem uma parte solicitada da computação.

Por exemplo, considerando dois contêineres, um com uma carga de trabalho de GPU de 0,25 e outro com uma carga de trabalho de GPU de 0,75, ambos receberão, em média, uma proporção igual do poder computacional. Se uma das cargas de trabalho não usar a GPU, a outra receberá o restante do poder computacional da GPU.

**MIG dinâmico**

O NVIDIA MIG permite que GPUs baseadas na arquitetura NVIDIA Ampere (como o NVIDIA A100) sejam particionadas em instâncias de GPU separadas. O MIG pode particionar os recursos de computação da GPU disponíveis para fornecer uma qualidade de serviço (QoS) definida com isolamento de falhas para clientes como VMs, contêineres ou processos. O MIG permite que várias instâncias de GPU sejam executadas em paralelo em uma única GPU física da NVIDIA. A divisão é tanto de computação quanto de memória e tem tamanhos fixos. Até sete unidades de computação e memória em tamanhos fixos são suportadas com base nos vários perfis MIG que podem ser encontrados na documentação da NVIDIA. Um perfil típico pode ser o MIG 2g.10gb, que fornece 2/7 do poder de computação e 10 GB de RAM. A divisão é estática no sentido de que você deve chamar a API da NVIDIA ou o comando **nvidia-smi** para criar ou remover a partição MIG.

Para evitar a atribuição estática, o Run:ai oferece uma maneira de criar uma partição MIG dinamicamente. O Run.ai possui as seguintes características:

* Semelhante ao recurso de GPU fracionada, você pode solicitar a parte da memória da GPU necessária. O Run:ai chamará a API MIG da NVIDIA para gerar o menor perfil MIG possível para sua solicitação e alocá-lo ao seu contêiner.
* O MIG é configurado de acordo com a demanda da carga de trabalho, sem drenar as cargas de trabalho ou envolver um administrador de TI.
* O Run:ai desalocará a partição automaticamente quando a carga de trabalho terminar. A partição não será removida até que o agendador decida que ela é necessária em outro lugar.
* Em um único cluster de GPU, você pode ter alguns nós MIG que são alocados dinamicamente e alguns que não são.

**NVIDIA GPUDirect**

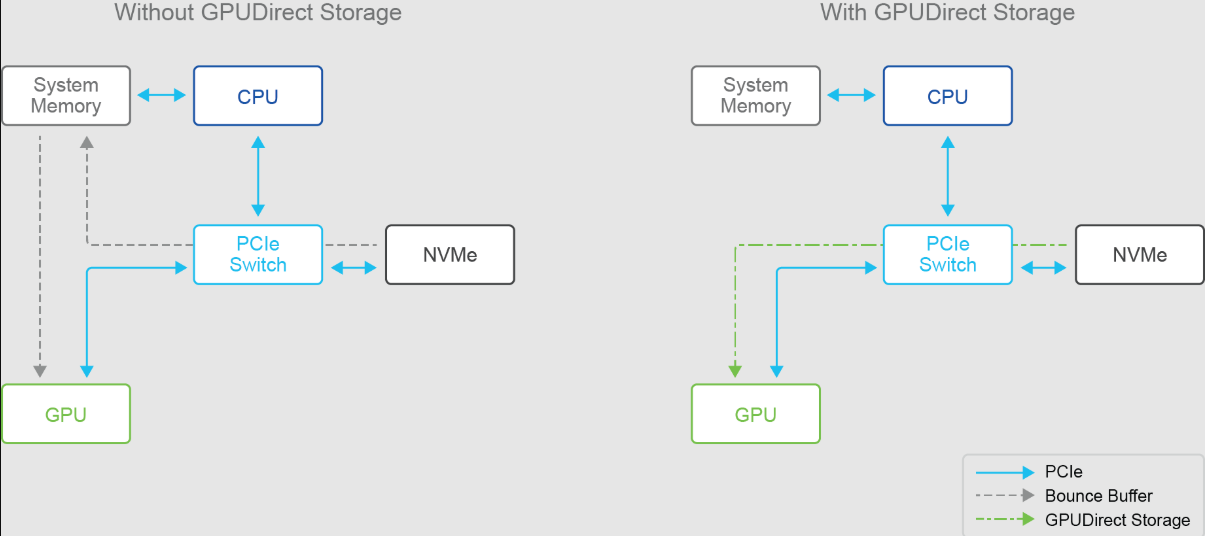
NVIDIA GPUDirect é uma família de tecnologias que aprimora a movimentação e o acesso de dados para GPUs de data center da NVIDIA. Usando o GPUDirect, adaptadores de rede e unidades de armazenamento podem ler e gravar diretamente na memória da GPU. Essa capacidade elimina cópias de memória desnecessárias, diminui a sobrecarga da CPU, reduz a latência e resulta em melhorias significativas de desempenho.

O NVIDIA GPUDirect inclui estas tecnologias:

* Armazenamento GPUDirect
* GPUDirect Remote Direct Memory Access (RDMA)

Da perspectiva do desenvolvedor, essas tecnologias são apresentadas por meio de um conjunto abrangente de APIs.

A figura a seguir compara um sistema com e sem GPUDirect Storage. O GPUDirect Storage permite um caminho de dados direto entre o armazenamento local ou remoto, como Non-Volatile Memory Express (NVMe) ou NVMe over Fabric (NVMe-oF), e a memória da GPU. O buffer de rejeição refere-se a uma área de memória intermediária usada para facilitar a transferência de dados entre a GPU e os dispositivos de armazenamento, como NVMe, quando a transferência direta de dados não é possível.



O GPUDirect Storage evita a utilização da CPU, permitindo que um mecanismo de acesso direto à memória (DMA) próximo à placa de interface de rede (NIC) ou ao armazenamento mova dados por um caminho direto para dentro ou para fora da memória da GPU. Além dos benefícios de acelerar a computação com GPUs em vez de CPUs, o GPUDirect Storage atua como um multiplicador de força quando pipelines inteiros de processamento de dados são transferidos para a execução na GPU. Essa função se torna especialmente importante quando o tamanho dos conjuntos de dados não cabe mais na memória do sistema e a E/S de dados para as GPUs se torna o gargalo determinante no tempo de processamento.

GPUDirect RDMA é outra tecnologia em computação de alto desempenho (HPC) e clusters de IA/ML, onde as GPUs nos nós remotos podem acessar diretamente a memória umas das outras, conforme ilustrado na figura a seguir.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

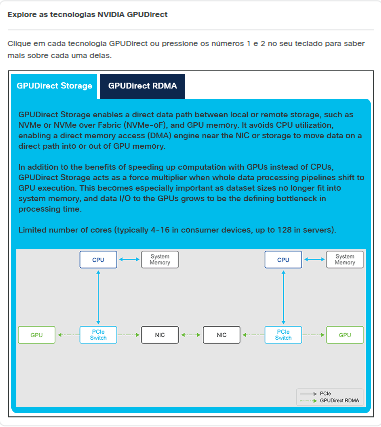
Projetado para aceleração de GPU, o GPUDirect RDMA fornece comunicação direta entre GPUs NVIDIA em sistemas remotos. Essa comunicação elimina as CPUs do sistema e as cópias de buffer de dados necessárias através da memória do sistema, resultando em um desempenho aproximadamente 10 vezes melhor.

O protocolo RDMA permite o acesso remoto direto à memória de um computador para a memória de outro, sem envolver o sistema operacional de nenhum deles. Exemplos de operações incluem gravação e leitura RDMA. O protocolo RDMA não deve ser confundido com o GPUDirect RDMA, que não está relacionado ao protocolo RDMA. O GPUDirect RDMA é uma das tecnologias habilitadas pela NVIDIA na família de tecnologias GPUDirect. Ele permite que a placa de rede envie ou receba dados diretamente, acessando a memória da GPU e ignorando as cópias de memória da CPU e as rotinas do sistema operacional. O GPUDirect RDMA funciona com InfiniBand ou RDMA sobre Ethernet Convergente (RoCE).

O GPUDirect RDMA está disponível no kit de ferramentas Compute Unified Device Architecture (CUDA).

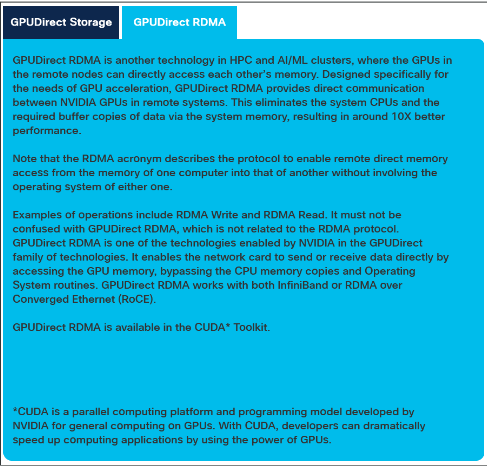
**Observação**

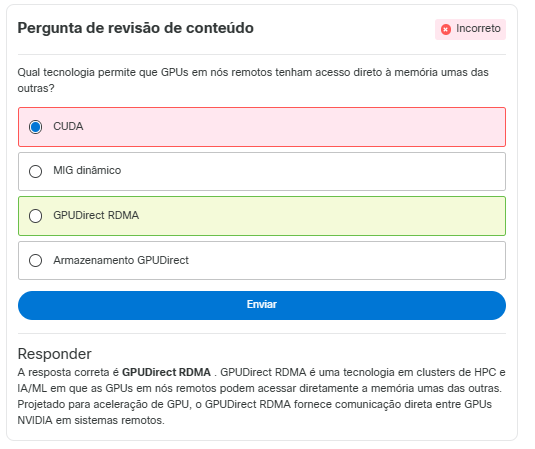
*CUDA é uma plataforma de computação paralela e um modelo de programação desenvolvido pela NVIDIA para computação geral em GPUs. Com CUDA, os desenvolvedores podem acelerar drasticamente os aplicativos de computação usando o poder das GPUs.*



Interface gráfica do usuário

O conteúdo gerado por IA pode estar incorreto.





Resumo

Você cobriu com sucesso soluções convergentes e hiperconvergentes, bem como soluções específicas para IA, como FlashStack Data Center, Nutanix GPT-in-a-Box e Run:ai.