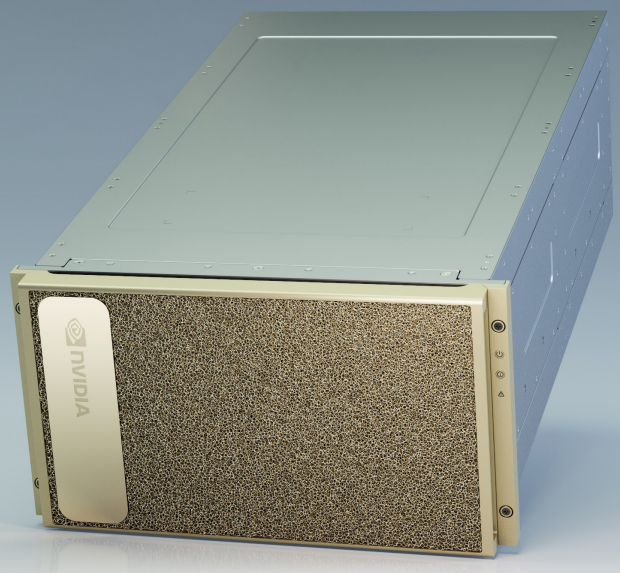
NVIDIA DGX A100
O SISTEMA UNIVERSAL PARA INFRAESTRUTURAS DE AI



O Desafio de Dimensionar a AI Empresarial

Toda empresa precisa se transformar usando a inteligência artificial (AI - Artificial Intelligence). Não apenas para sobreviver, mas para prosperar em tempos desafiadores. No entanto, as empresas precisam de uma plataforma para infraestrutura de AI que melhora as abordagens tradicionais, as quais historicamente envolveram arquiteturas de computação lentas que foram separadas por cargas de trabalho de análise, treinamento e inferência.

A abordagem antiga criava complexidade, gerava custos, restringia a velocidade de escala e não estava pronta para a AI moderna. Empresas, desenvolvedores, cientistas de dados e pesquisadores precisam de uma nova plataforma que unifique todas as cargas de trabalho de AI, simplificando a infraestrutura e acelerando o retorno sobre o investimento.

O Sistema Universal para Todas as Cargas de Trabalho de AI

A NVIDIA DGX™ A100 é o sistema universal para todas as cargas de trabalho de AI, desde a análise até treinamento e inferência. O DGX A100 define um novo nível para a densidade computacional, com 5 petaFLOPS de desempenho de AI em um tamanho 6U, substituindo a infraestrutura de computação legada por um sistema único e unificado. A DGX A100 também oferece a capacidade sem precedentes de fornecer alocação refinada de potência computacional, usando o recurso de GPU de várias instâncias na GPU NVIDIA A100 Tensor Core, que permite que os administradores atribuam recursos que são dimensionados corretamente para cargas de trabalho específicas. Isso garante que os maiores e mais complexos trabalhos sejam compatíveis, sem deixar de lado os mais simples e menores. Executando o conjunto de software DGX com software otimizado do NGC, a combinação de potência computacional densa e a flexibilidade de carga de trabalho completa torna a DGX A100 uma opção ideal para implantações de um único nó e clusters Slurm e Kubernetes de grande escala implantados com o NVIDIA DeepOps.

Acesso Direto ao NVIDIA DGXperts

A NVIDIA DGX A100 é mais do que um servidor, é uma plataforma completa de hardware e software baseada no conhecimento adquirido na maior pista de testes de DGX do mundo, NVIDIA DGX SATURNV, e auxiliada por milhares de DGXperts na NVIDIA. DGXperts são profissionais especialistas em AI que oferecem orientação prescritiva e experiência em design para ajudar a acelerar a transformação de AI. Eles desenvolveram uma ampla variedade de conhecimentos e experiências na última década para ajudar a maximizar o valor do seu investimento em DGX. Os DGXperts ajudam a garantir que aplicações críticas sejam ativadas rapidamente e permaneçam funcionando sem problemas, para um tempo de insights significativamente aprimorado.

ESPECIFICAÇÕES DO SISTEMA

GPUs	**8 GPUs NVIDIA A100 Tensor Core**
Memória da GPU	**320 GB total**
Desempenho	**5 petaFLOPS AI** **10 petaOPS INT8**
NVIDIA NVSwitches	**6**
Uso de Energia do Sistema	**6,5 kW Max**
CPU	**Dual AMD Rome 7742, total de 128 núcleos, 2,25 GHz (base), 3,4 GHz (máximo de boost)**
Memória do sistema	**1 TB**
Rede	**8 Mellanox ConnectX-6 VPI de porta única** **HDR InfiniBand de 200 GB/s** **1 Mellanox ConnectX-6 VPI de porta dupla** **Ethernet de 10/25/50/100/200 GB/s**
Armazenamento	**Sistema operacional: 2 unidades NVME M.2 de 1,92 TB** **Armazenamento interno: unidades NVME U.2 de 15 TB (4x 3,84 TB)**
Software	**Ubuntu Linux OS**
Peso do sistema	**271 lbs (123 kg)**
Peso do sistema empacotado	**315 lbs (143 kg)**
Dimensões do sistema	**Altura: 10,4 pol. (264 mm)** **Largura: 19 pol. (482,3 mm) máx.** **Comprimento: 35,3 pol. (897,1 mm) máx.**
Faixa de temperatura operacional	**5ºC a 30ºC (41ºF a 86ºF)**

Tempo Mais Rápido para a Solução

A NVIDIA DGX A100 conta com oito GPUs NVIDIA A100 Tensor Core, oferecendo aos usuários aceleração inigualável, além de ser totalmente otimizada para o software NVIDIA CUDA-X™ e o conjunto completo de soluções de data center da NVIDIA. As GPUs NVIDIA A100 trazem uma nova precisão, TF32, que funciona como FP32 ao mesmo tempo que oferece FLOPS 20 vezes mais altos para AI em relação à geração anterior. E o melhor de tudo, nenhuma alteração de código é necessária para obter essa velocidade. E, ao usar a precisão mista automática da NVIDIA, a A100 oferece 2 vezes mais desempenho com apenas uma linha de código adicional, usando a precisão FP16. A GPU A100 também tem um nível de 1,6 terabyte por segundo (TB/s) de largura de banda, que é mais do que 70% de aumento em relação à última geração. Além disso, a GPU A100 tem uma memória em chip significativamente maior, incluindo um cache nível 2 de 40 MB que é quase 7 vezes maior do que a geração anterior, maximizando o desempenho da computação. O DGX A100 também apresenta a última geração do NVIDIA NVLink™, que dobra a largura de banda de GPU para GPU para 600 gigabytes por segundo (GB/s), quase 10 vezes mais do que a PCIe da 4ª geração, e um novo NVIDIA NVSwitch que é 2 vezes mais rápido do que a geração anterior. Essa potência sem precedentes oferece o mais rápido tempo de solução, permitindo aos usuários enfrentar desafios que não eram possíveis nem práticos antes.

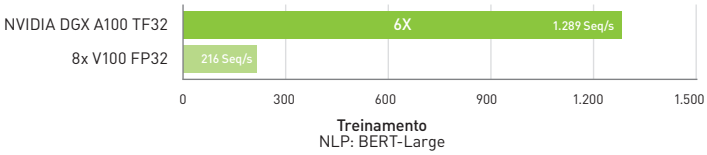
O Sistema de AI Mais Seguro do Mundo para Empresas

A NVIDIA DGX A100 oferece a melhor postura de segurança para sua empresa de AI, com uma abordagem em várias camadas que protege todos os principais componentes de hardware e software. Ampliando o Baseboard Management Controller (BMC), a placa da CPU, a GPU, as unidades com criptografia automática e a inicialização segura, a DGX A100 tem segurança incorporada, permitindo que a IT se concentre em operações de AI, em vez de gastar tempo na avaliação e mitigação de ameaças.

Escalabilidade de Data Center Inigualável com Mellanox

Com a arquitetura de I/O mais rápida do que qualquer sistema DGX, a NVIDIA DGX A100 é a base para clusters de AI maiores, como o NVIDIA DGX SuperPOD™, o esquema empresarial para infraestrutura dimensionável de AI. A DGX A100 tem oito adaptadores Mellanox ConnectX-6 VPI HDR InfiniBand de porta única para clusters e um adaptador de Ethernet ConnectX-6 VPI com porta dupla para armazenamento e rede, todos capazes de 200 GB/s.

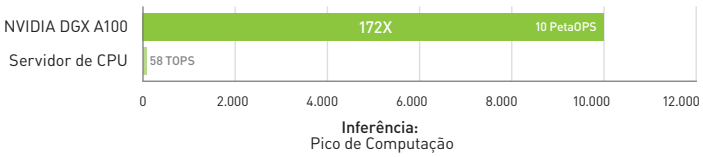
DGX A100 Oferece 6 Vezes Mais Desempenho de Treinamento



Treinamento
NLP: BERT-Large

Taxa de transferência de pré-treinamento BERT usando o PyTorch, incluindo (2/3)Fase 1 e (1/3)Fase 2 | Fase 1 Seq Len = 128, Fase 2 Seq Len = 512 | V100: DGX-1 com 8 V100 usando precisão FP32 | DGX A100: DGX A100 com 8 A100 usando precisão TF32

DGX A100 Oferece 172 Vezes Mais Desempenho de Inferência



Inferência:
Pico de Computação

Servidor de CPU: 2 Intel Platinum 8280 usando INT8 | DGX A100: DGX A100 com 8 A100 usando INT8 com dispersão estrutural

DGX A100 Oferece 13 Vezes Mais Desempenho de Análise de Dados

Análise:
PageRank

3.000 Servidores de CPU vs. 4 DGX A100 | Conjunto de Dados de Rastreamento Comum Publicado: 128B Edges, 2,6 TB Gráficos

A combinação de computação maciça acelerada por GPU com hardware de última geração e otimizações de software faz com que o DGX A100 possa ser dimensionado até centenas ou milhares de nós para enfrentar os maiores desafios, como a AI de conversação e a classificação de imagem em larga escala.

Soluções de Infraestrutura Comprovadas Criadas com Líderes de Data Center Confiáveis

Em combinação com os principais fornecedores de tecnologia de armazenamento e rede, oferecemos um portfólio de soluções de infraestrutura que incorporam o melhor da arquitetura de referência NVIDIA DGX POD™. Entregue como ofertas totalmente integradas e prontas para implantação com nossa rede de parceiros NVIDIA, essas soluções tornam as implantações de AI de data centers mais simples e mais rápidas para a IT.

Para saber mais sobre o NVIDIA DGX A100, acesse **www.nvidia.com/DGXA100**

NVIDIA