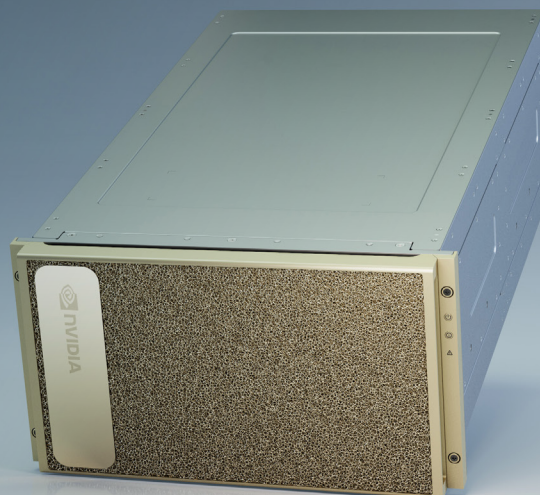




# NVIDIA DGX A100

## O Sistema Universal para a Infraestrutura 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, é necessária uma plataforma para a infraestrutura de AI que melhore as abordagens tradicionais, que historicamente envolveram arquiteturas de computação lentas 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

O NVIDIA DGX™ A100 é o sistema universal para todas as infraestruturas de AI, de análise a treinamento e inferência. O DGX A100 define um novo nível para a densidade computacional, com 5 petaFLOPS de desempenho de AI em um formato 6U, substituindo a infraestrutura de computação legada por um sistema único e unificado. O DGX A100 também oferece a capacidade sem precedentes de fornecer alocação refinada de potência computacional, usando o recurso de GPU Multi-Instância (MIG - Multi-Instance GPU) na GPU NVIDIA A100 Tensor Core. Isso permite que os administradores atribuam recursos do tamanho certo para cargas de trabalho específicas.

Disponível com até 640 gigabytes (GB) de memória total de GPU, o que aumenta o desempenho em tarefas de treinamento de grande escala em até 3 vezes e dobra o tamanho das instâncias de MIG, o DGX A100 pode lidar com as maiores e mais complexas tarefas, além das mais simples e menores. Executando o pacote de softwares DGX com software otimizado do NVIDIA NGC™, a combinação de potência computacional densa e flexibilidade de carga de trabalho completa torna o DGX A100 uma opção ideal para implantações de um único nó e clusters Slurm e Kubernetes de grande escala implantados com o NVIDIA DeepOps.

### Nível Inigualável de Suporte e Experiência

O NVIDIA DGX A100 é mais do que um servidor. É uma plataforma completa de hardware e software desenvolvida com base no conhecimento adquirido do maior campo de testes DGX do mundo, o NVIDIA DGX SATURNV, e apoiado por milhares de DGXperts na NVIDIA. Os DGXperts são profissionais de AI que desenvolveram uma grande variedade de conhecimentos e experiências na última década para ajudar a maximizar o valor do seu investimento em DGX. Eles ajudam a garantir que aplicações críticas sejam ativadas rapidamente e permaneçam funcionando sem problemas, para um tempo de insights significativamente aprimorado.

### ESPECIFICAÇÕES DO SISTEMA

| NVIDIA DGX A100 640GB            |                                                                                                                                           |                                                                                                                                               |
|----------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| GPUs                             | 8 GPUs NVIDIA A100 80GB Tensor Core                                                                                                       |                                                                                                                                               |
| Memória da GPU                   | Total de 640GB                                                                                                                            |                                                                                                                                               |
| Desempenho                       | AI de 5 petaFLOPS<br>10 petaOPS, INT8                                                                                                     |                                                                                                                                               |
| NVIDIA NVSwitches                | 6                                                                                                                                         |                                                                                                                                               |
| Uso de Energia do Sistema        | 6,5kW máx.                                                                                                                                |                                                                                                                                               |
| CPU                              | Dual AMD Rome 7742,<br>total de 128 núcleos, 2,25GHz (base),<br>3,4GHz (máximo de boost)                                                  |                                                                                                                                               |
| Memória do Sistema               | 2TB                                                                                                                                       |                                                                                                                                               |
| Redes                            | 8 NVIDIA ConnectX-7 de Porta Única<br>InfiniBand de 200GB/s<br>2 NVIDIA ConnectX-7 VPI de Porta Dupla<br>Ethernet de 10/25/50/100/200GB/s | 8 NVIDIA ConnectX-6 VPI de Porta Única<br>InfiniBand de 200GB/s<br>2 NVIDIA ConnectX-6 VPI de Porta Dupla<br>Ethernet de 10/25/50/100/200GB/s |
| Armazenamento                    | Sistema Operacional: 2 unidades NVME M.2 de 1,92TB<br>Armazenamento Interno: unidades NVMe U.2 de 30TB (8 de 3,84TB)                      |                                                                                                                                               |
| Software                         | Sistema Operacional Ubuntu Linux<br>Também é compatível com:<br>Red Hat Enterprise Linux<br>CentOS                                        |                                                                                                                                               |
| Peso do Sistema                  | 271,5lb (123,16kg) máx.                                                                                                                   |                                                                                                                                               |
| Peso do Sistema Empacotado       | 359,7lb (163,16kgs) máx.                                                                                                                  |                                                                                                                                               |
| Dimensões do Sistema             | Altura: 10,4pol. (264mm)<br>Largura: 19pol. (482,3mm) máx.<br>Comprimento: 35,3pol. (897,1mm) máx.                                        |                                                                                                                                               |
| Faixa de Temperatura Operacional | 5°C a 30°C (41°F a 86°F)                                                                                                                  |                                                                                                                                               |

## Menos Tempo para Solução

O NVIDIA DGX A100 tem oito GPUs NVIDIA A100 Tensor Core, o que oferece uma aceleração única, e é totalmente otimizado para o software NVIDIA CUDA-X™ e o pacote de soluções de data center de ponta a ponta da NVIDIA. As GPUs NVIDIA A100 trazem precisão Tensor Float 32 (TF32), o formato de precisão padrão para frameworks de AI TensorFlow e PyTorch. Isso funciona como o FP32, mas oferece operações flutuantes por segundo (FLOPS) 20 vezes maiores para AI em comparação com a geração anterior. O melhor de tudo é que não são necessárias alterações de código para atingir essa aceleração.

A GPU A100 de 80GB aumenta a largura de banda da memória da GPU em 30% em relação à GPU A100 de 40GB, tornando-a a primeira do mundo com 2 terabytes por segundo (TB/s). Ela também tem uma memória no chip significativamente maior do que a GPU NVIDIA da geração anterior, incluindo um cache de nível 2 de 40 megabytes (MB) que é quase 7 vezes maior, maximizando o desempenho de computação. O DGX A100 também estreia a terceira geração do NVIDIA®NVLink®, que dobra a largura de banda direta entre GPUs para 600 gigabytes por segundo (GB/s), quase 10 vezes maior que o PCIe 4.0, e um novo NVIDIA NVSwitch™ que é 2 vezes mais rápido 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

O NVIDIA DGX A100 oferece uma sólida postura de segurança para a 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, o DGX A100 tem segurança incorporada, permitindo que a TI se concentre em operações de AI, em vez de gastar tempo na avaliação e mitigação de ameaças.

## Escalabilidade de Data Center Inigualável com NVIDIA Networking

Com a arquitetura de I/O mais rápida de um sistema DGX, o NVIDIA DGX A100 é o componente fundamental para grandes clusters de AI, como o NVIDIA DGX SuperPOD™, o esquema empresarial para infraestrutura de AI dimensionável. O DGX A100 tem oito adaptadores NVIDIA ConnectX®-7 InfiniBand de porta única para clusters e até dois adaptadores ConnectX-7 VPI (InfiniBand ou Ethernet) com porta dupla para armazenamento e rede, todos capazes de 200GB/s. Com a conectividade ConnectX-7 com os switches NVIDIA Quantum-2 InfiniBand, é possível criar o DGX SuperPOD com menos switches e cabos, economizando nos custos operacionais e de capital na infraestrutura do data center. A combinação de computação maciça acelerada por GPU com hardware de última geração e otimizações de

### Treinamento de AI com Rendimento até 3 Vezes maior nos Modelos Maiores



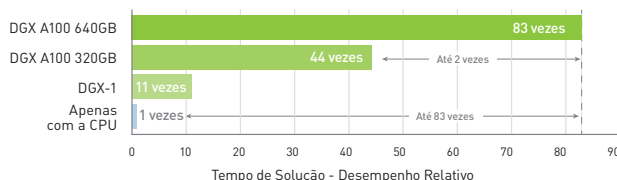
DLRM no framework HugeCTR, precisão = FP16 | 1 DGX A100 de 640GB de tamanho do lote = 48 | 2 DGX A100 de 320GB de tamanho do lote = 32 | 1 DGX-2 (16 V100 de 32GB) de tamanho do lote = 32. Acelerações Normalizadas em Número de GPUs.

### Rendimento Até 1,25 Vezes Maior para Inferência de AI



MLPerf 0.7 RNN-T medida com (1/7) fatias MIG. Framework: TensorRT 7.2, conjunto de dados = LibriSpeech, precisão = FP16.

### Rendimento até 83 Vezes Maior que o da CPU, Rendimento 2 vezes Maior que o da A100 de 320GB no Benchmark de Análise de Big Data



Benchmark de análise de big data | 30 consultas de varejo analíticas, ETL, ML, NLP em um conjunto de dados de 10TB | CPU: 19 vezes o Intel Xeon Gold 6252 de 2,10GHz, Hadoop | 16 vezes o DGX-1 (8 V100 32GB cada), RAPIDS/Dask | 12 vezes a DGX A100 de 320GB e 4 vezes a DGX A100 de 640GB, RAPIDS/Dask/BlazingSQL. Acelerações Normalizadas em Número de GPUs.

software faz com que o DGX A100 possa ser dimensionado até centenas ou milhares de nós para enfrentar os maiores desafios, como a AI conversacional e a classificação de imagem em larga escala.

## Soluções de Infraestrutura Comprovadas Criadas com Líderes de Data Center Confiáveis

Em conjunto com os principais provedores de tecnologia de armazenamento e rede, há um portfólio de soluções de infraestrutura que incorpora o melhor da arquitetura de referência NVIDIA DGX POD™. Entregue como ofertas totalmente integradas e prontas para implantação com nossa Rede de Parceiros da NVIDIA (NPN - Partner Network), essas soluções tornam as implantações de AI de data centers mais simples e mais rápidas.

Saiba mais

Para saber mais sobre o NVIDIA DGX A100, acesse [www.nvidia.com/pt-br/data-center/dgx-a100/](http://www.nvidia.com/pt-br/data-center/dgx-a100/)