NVIDIA DGX-2
THE WORLD'S MOST POWERFUL DEEP LEARNING SYSTEM FOR THE MOST COMPLEX AI CHALLENGES

The Challenge of Scaling to Meet the Demands of Modern AI and Deep Learning

Deep neural networks are rapidly growing in size and complexity, in response to the most pressing challenges in business and research. The computational capacity needed to support today's modern AI workloads has outpaced traditional data center architectures. Modern techniques that exploit increasing use of model parallelism are colliding with the limits of inter-GPU bandwidth, as developers build increasingly large accelerated computing clusters, pushing the limits of data center scale. A new approach is needed - one that delivers almost limitless AI computing scale in order to break through the barriers to achieving faster insights that can transform the world.

Performance to Train the Previously Impossible

Increasingly complex AI demands unprecedented levels of compute. NVIDIA® DGX-2™ is the world's first 2 petaFLOPS system, packing the power of 16 of the world's most advanced GPUs, accelerating the newest deep learning model types that were previously untrainable. With groundbreaking GPU scale, you can train models 4X bigger on a single node. In comparison with legacy x86 architectures, DGX-2's ability to train ResNet-50 would require the equivalent of 300 servers with dual Intel Xeon Gold CPUs costing over $2.7 million dollars.

NVIDIA NVSwitch—A Revolutionary AI Network Fabric

Leading edge research demands the freedom to leverage model parallelism and requires never-before-seen levels of inter-GPU bandwidth. NVIDIA has created NVSwitch to address this need. Like the evolution from dial-up to ultra-high speed broadband, NVSwitch delivers a networking fabric for the future, today. With NVIDIA DGX-2, model complexity and size are no longer constrained by the limits of traditional architectures. Embrace model-parallel training with a networking fabric in DGX-2 that delivers 2.4TB/s of bisection bandwidth for a 24X increase over prior generations. This new interconnect "superhighway" enables limitless possibilities for model types that can reap the power of distributed training across 16 GPUs at once.

SYSTEM SPECIFICATIONS

GPUs	**16X NVIDIA® Tesla V100**
GPU Memory	**512GB total**
Performance	**2 petaFLOPS**
NVIDIA CUDA® Cores	**81920**
NVIDIA Tensor Cores	**10240**
NVSwitches	**12**
Maximum Power Usage	**10kW**
CPU	**Dual Intel Xeon Platinum 8168, 2.7 GHz, 24-cores**
System Memory	**1.5TB**
Network	**8X 100Gb/sec Infiniband/100GigE Dual 10/25/40/50/100GbE**
Storage	**OS: 2X 960GB NVME SSDs Internal Storage: 30TB (8X 3.84TB) NVME SSDs**
Software	**Ubuntu Linux OS See Software stack for details**
System Weight	**360 lbs (163.29 kgs)**
Packaged System Weight	**400lbs (181.44kgs)**
System Dimensions	**Height: 17.3 in (440.0 mm) Width: 19.0 in (482.3 mm) Length: 31.3 in (795.4 mm) - No Front Bezel 32.8 in (834.0 mm) - With Front Bezel**
Operating Temperature Range	**5°C to 35°C (41°F to 95°F)**

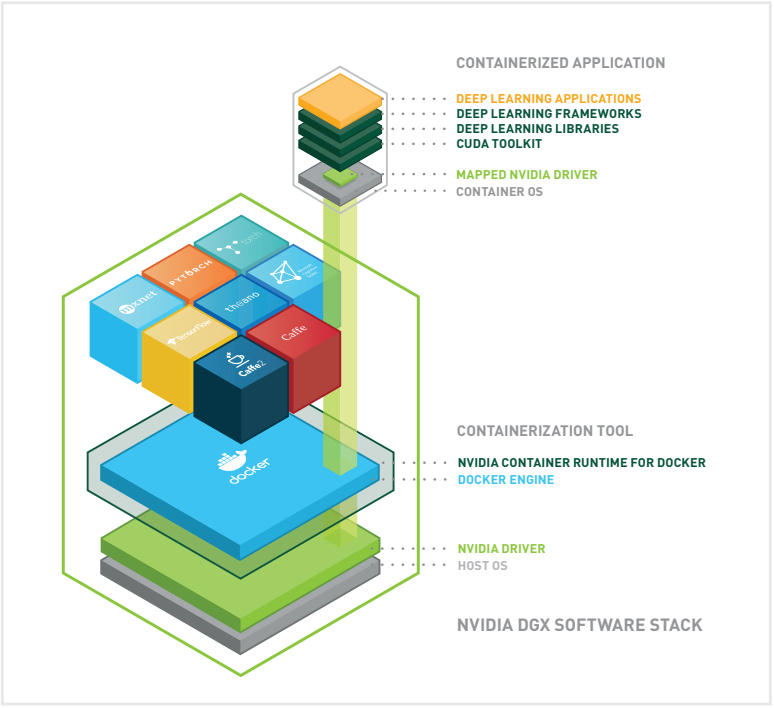
AI Scale on a Whole New Level

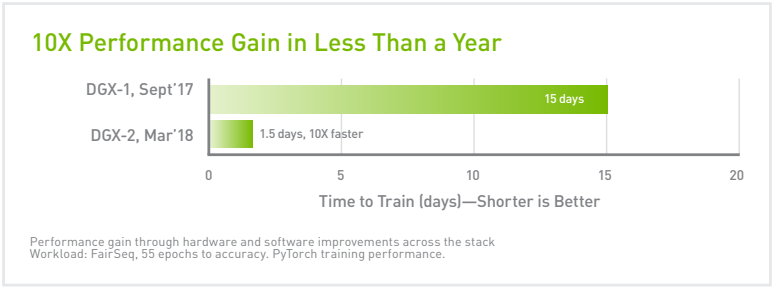
Modern enterprises need to rapidly deploy AI power in response to business imperatives, and need to scale-out AI, without scaling-up cost or complexity. We've built DGX-2 and powered it with DGX software that enables accelerated deployment, simplified operations – at scale. DGX-2 delivers a ready-to-go solution that offers the fastest path to scaling-up AI, along with virtualization support, to enable you to build your own private enterprise grade AI cloud. Now businesses can harness unrestricted AI power in a solution that scales effortlessly with a fraction of the networking infrastructure needed to bind accelerated computing resources together. With an accelerated deployment model, and an architecture purpose-built for ease of scale, your team can spend more time driving insights and less time building infrastructure.

Enterprise Grade AI Infrastructure

If your AI platform is critical to your business, you need one designed with reliability, availability and serviceability (RAS) in mind. DGX-2 is enterprise-grade, built for rigorous round-the-clock AI operations, and is purpose-built for RAS to reduce unplanned downtime, streamline serviceability, and maintain operation continuity.

Spend less time tuning and optimizing and more time focused on discovery. NVIDIA's enterprise-grade support saves you from the time-consuming job of troubleshooting hardware and open source software. With every DGX system, get started fast, train faster, and remain faster with an integrated solution that includes software, tools and NVIDIA expertise.





For more information, visit **www.nvidia.com/DGX-2**