
NVIDIA DGX SYSTEMS
PURPOSE-BUILT FOR AI

AI researchers depend on computing performance to gain insights and innovate faster, using the power of deep learning and analytics. **GPU technology offers a faster path to AI, but building a platform goes well beyond deploying a server and GPU's.**

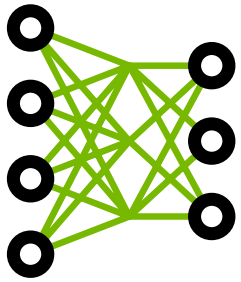
AI and deep learning can require a substantial commitment of your resources to deploy, manage, and scale your platform. An investment that could delay your project by months as you integrate a complex stack of components and software including frameworks, libraries, and drivers. You need a solution that's quickly deployed, effortless to manage, delivering breakthrough performance to power your AI initiative.


UNPARALLELED
VALUE

NVIDIA DGX SYSTEMS

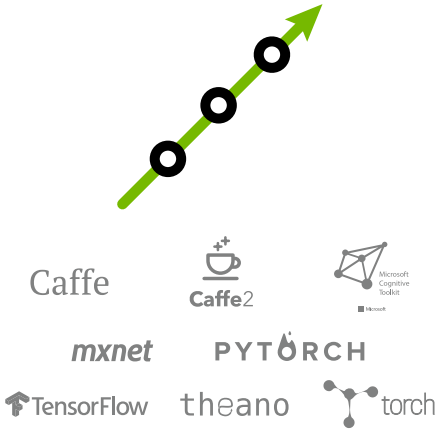
The Essential Instruments of AI Research

Introducing **NVIDIA® DGX™ Systems,** purpose-built and inspired by the unique demands of deep learning. More than GPU-powered supercomputers, NVIDIA designed each DGX solution with three key objectives in mind:



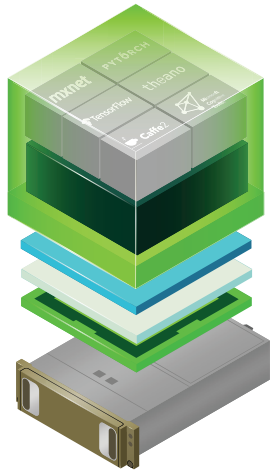
1. FASTEST PATH TO DEEP LEARNING

Get started quickly in deep learning, using a fully integrated hardware and software solution that's simple to deploy, and even simpler to configure.



2. EFFORTLESS PRODUCTIVITY

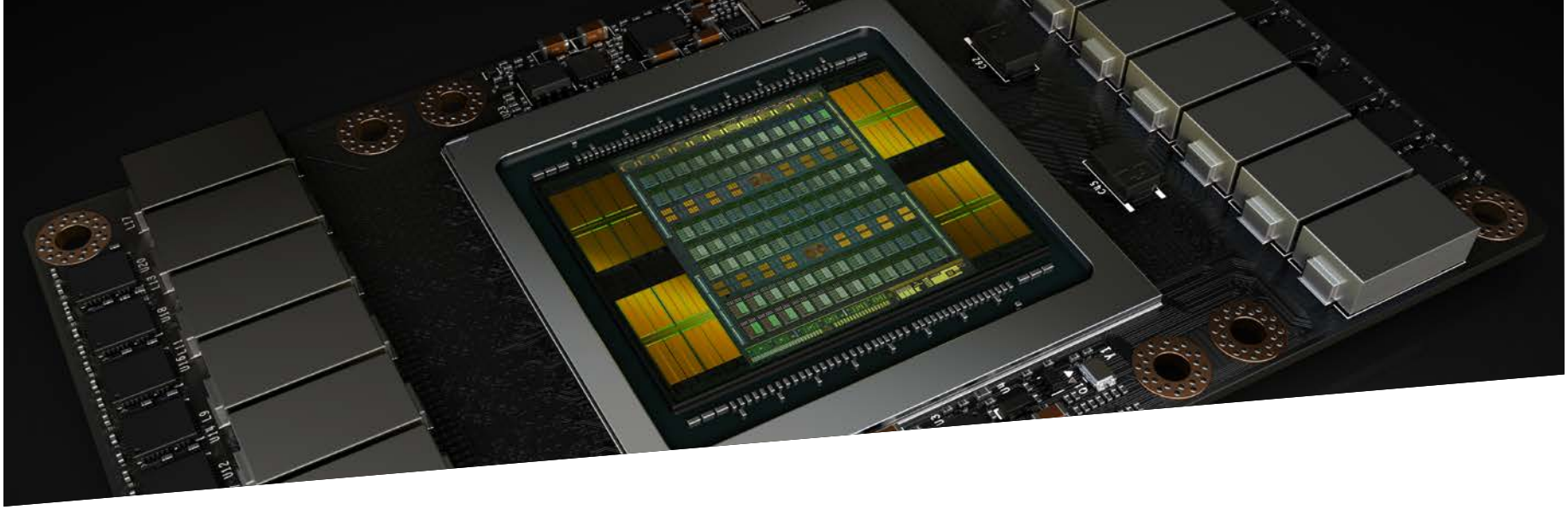
Spend more time focused on experimentation and insight with NVIDIA optimized deep learning (DL) frameworks, an intuitive, streamlined user interface (UI), and cloud management.



3. REVOLUTIONARY AI PERFORMANCE

Delivering the fastest performance through the power of our DGX software stack available only on DGX, built on the latest GPU technology including NVIDIA Volta™.

MASSACHUSETTS GENERAL HOSPITAL USES DGX-1 TO ACCELERATE ADOPTION OF AI IN RADIOLOGY **LEARN MORE**



UNPRECEDENTED PERFORMANCE

Foundational to this portfolio of systems is the groundbreaking AI performance it delivers. NVIDIA research and development has delivered an integrated software stack that powers each DGX System, offering better performance than do-it-yourself (DIY) approaches. Data scientists can reap the real-world benefits of patented NVIDIA software found nowhere else.

Built on the latest NVIDIA GPU technology including the NVIDIA Tesla® V100, DGX Systems are architected to deliver the fastest deep learning training possible, up to 3X faster than prior generations. Additionally, multi-system training and strong scaling with the next generation of NVIDIA NVLink™, enables a dramatic speedup in training time, allowing the highest possible computational density with excellent linearity of performance at scale.

This groundbreaking architecture, enables DGX Systems to break through the bottleneck inherent to commodity hardware solutions, with data rates that are up to 10X faster than traditional PCIe Gen3 interconnects, and 2X the throughput of prior generations. Also new with our Tesla V100 based DGX Systems are Tensor Cores–which deliver an exponential leap in A.I. performance—with 640 Tensor Cores in every Tesla V100. DGX Systems leverage this optimized architecture for the massively parallelized 4x4x4 matrix calculations found in today's deep neural networks.

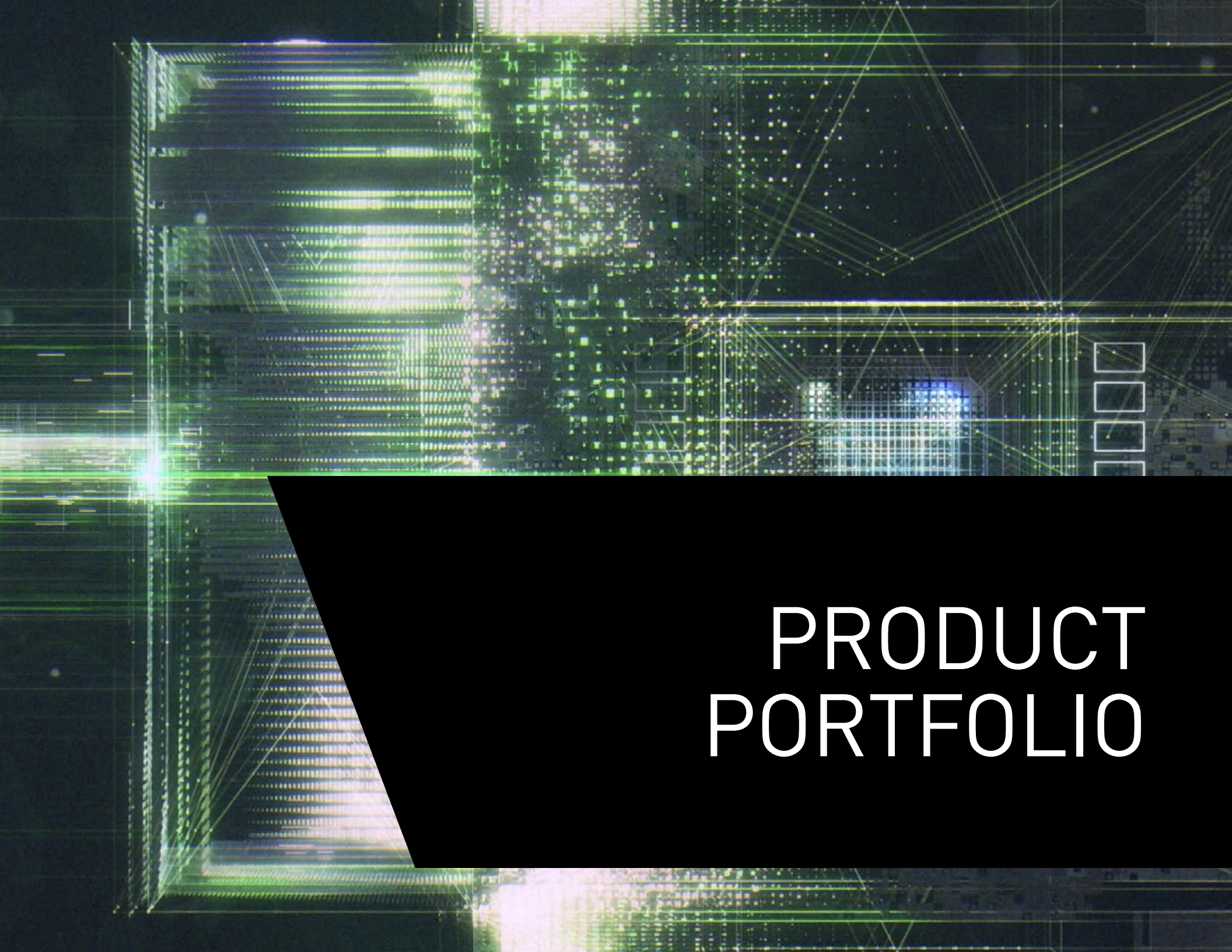
NVIDIA DELIVERED EXASCALE PERFORMANCE WITH DGX SATURNV **LEARN MORE**

EFFORTLESS PRODUCTIVITY

Optimizing and fine-tuning a deep learning environment requires millions of dollars in departmental OpEx and lost weeks or months of productivity. Unlike off-the-shelf commodity hardware, only DGX Systems incorporate the groundbreaking innovations delivered in NVIDIA's software, purpose-built for deep learning. Now you get the full scope of NVIDIA's years of work in every DGX System, with the world's only A.I. supercomputers that integrate these capabilities.

DGX Systems are the only supercomputers that include NVIDIA optimized deep learning frameworks. NVIDIA's team of engineers take today's frameworks further, with innovations including an optimized I/O pipeline matched to each DGX configuration, enhancements that maximize NVLink performance, and monthly updates to each deep learning framework delivered over the network. Also, while today's most popular deep learning frameworks are open source, every enterprise data scientist needs the peace of mind that comes with enterprise-grade support. NVIDIA DGX Systems are the only solutions in the marketplace that integrate the most popular open source deep learning frameworks, optimized for NVIDIA GPU and DGX configurations, backed by the deep learning expertise of NVIDIA engineers, solution architects, and community. Finally, DGX Systems provide a cloud-hosted management interface for your on-premises system and data.

Now you can enjoy an accelerated, streamlined experience for managing DGX nodes or clusters, and containers. Use this intuitive UI to take advantage of point and click job scheduling, with granular visibility of resource consumption, hardware health, as well as automated DGX software and security updates.



PRODUCT PORTFOLIO

NVIDIA DGX STATION

NVIDIA DGX-1

NVIDIA DGX-1

This is the first portfolio of its kind, designed to address the end-to-end lifecycle of deep learning development, exploration, and scale across an organization's teams. We've started with the already proven NVIDIA DGX-1™ which is now offered with the NVIDIA Tesla V100. DGX-1 delivers groundbreaking AI performance in the data center for production scale training with the computing capacity of 800 CPUs, and integrated scale-out cluster interconnect using Infiniband EDR. And with Tesla V100 based systems you get the added versatility of high-performance inference in the same system when you need it.

NVIDIA DGX-1 **LEARN MORE**

NVIDIA DGX STATION

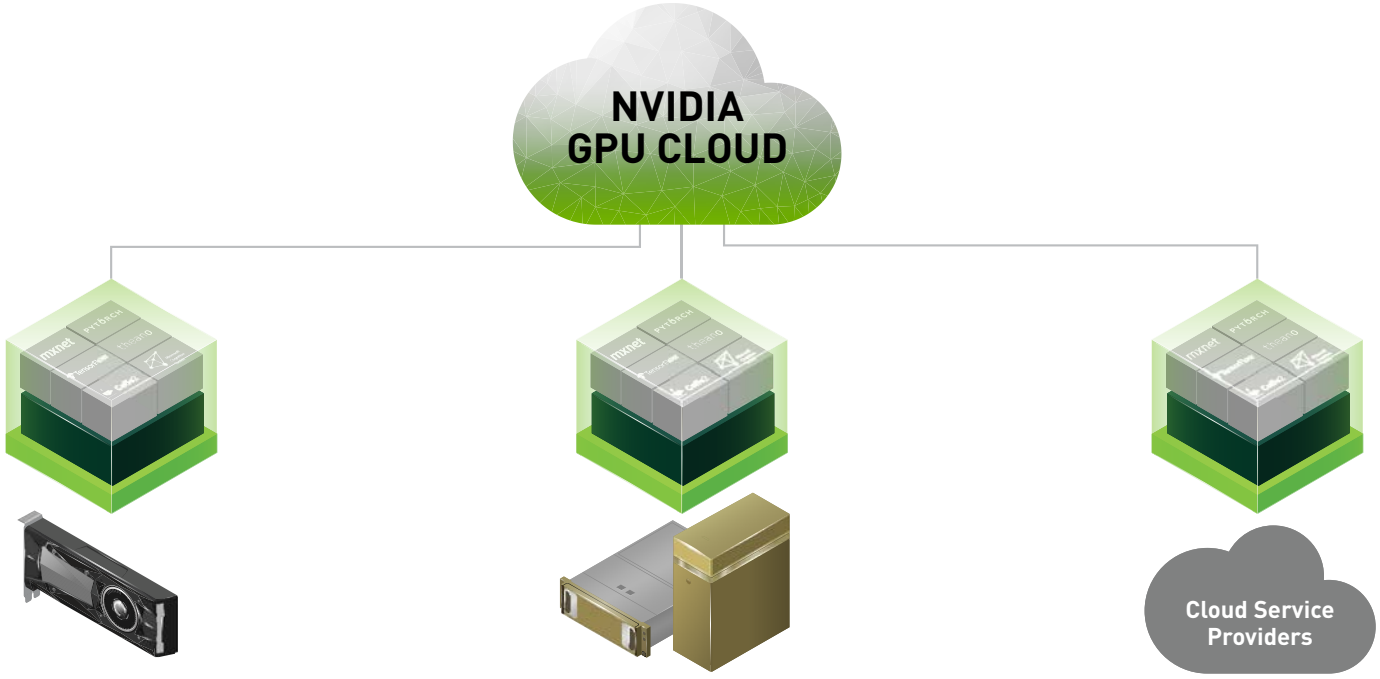
NVIDIA DGX Station™ is the newest addition to the DGX portfolio, and it solves the challenge of how to bring supercomputing power closer to those who need it so that they can be more productive from the convenience of their workspace. This is the industry's fastest personal AI supercomputer, taking the value and performance of NVIDIA DGX-1 and extending it to the data scientist's desk. DGX Station delivers the industry's only 4-way NVLink workstation, and have the computing capacity of 400 CPUs. All that is contained in a whisper quiet form factor that's optimized for the desk, powered by the same deep learning stack found on DGX-1.

DEEP DIVE INTO NVIDIA DGX STATION

LEARN MORE

NVIDIA GPU CLOUD

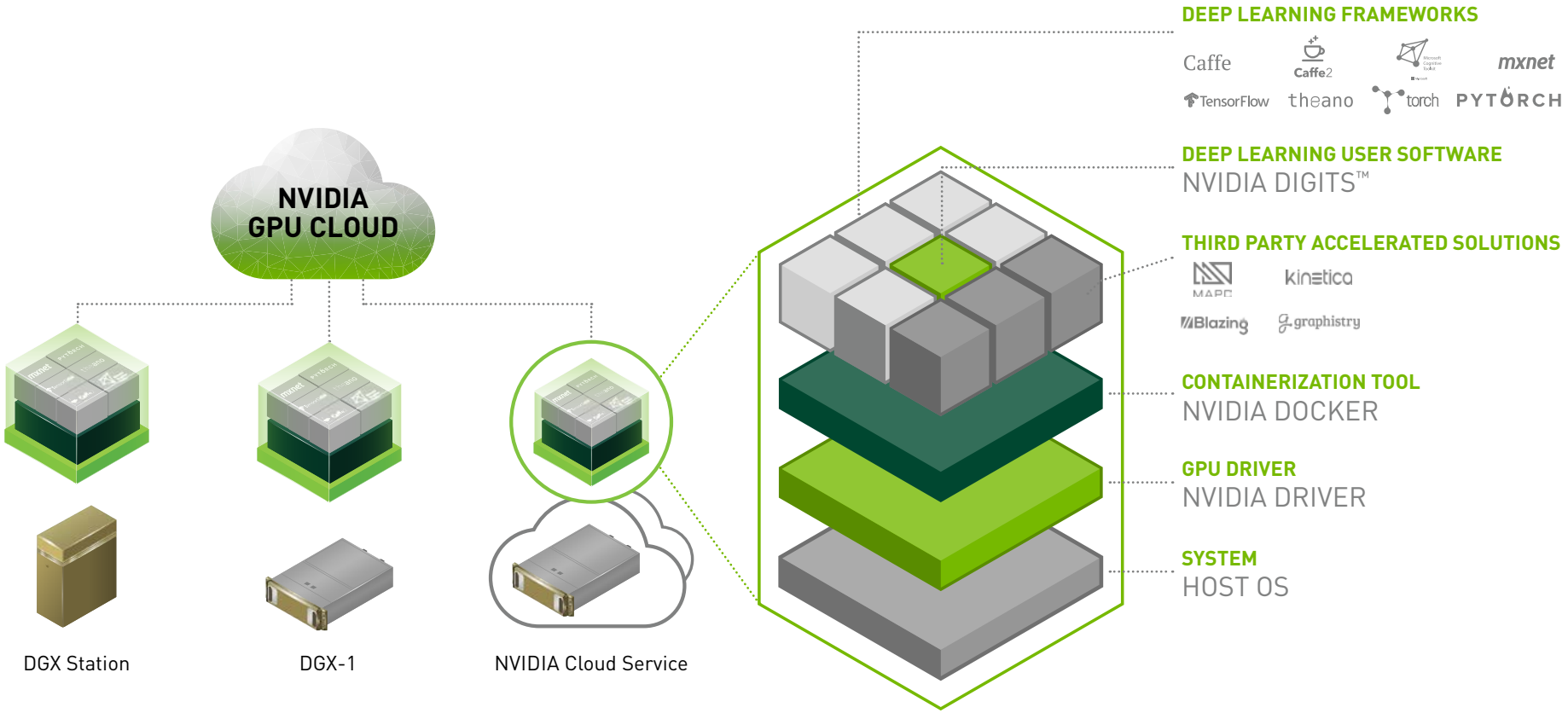
The 3rd pillar of this portfolio is the NVIDIA GPU Cloud, providing simplified, streamlined manageability of DGX Systems, and a brand new consumption model for GPU-accelerated deep learning–delivered from the cloud, built on NVIDIA GPU technology, and powered by a unified deep learning stack.



Cloud Service with the Highest Deep Learning Efficiency

SOFTWARE PLATFORM

COMMON SOFTWARE STACK ACROSS DGX FAMILY



DGX software running on NVIDIA GPUs delivers 30% faster training than a DIY system using comparable hardware without DGX optimized software.

NVIDIA DGX Systems are powered by a unified deep learning stack and cloud management, providing an intuitive interface and a powerful suite of core services that make it easy to manage your deep learning environment.

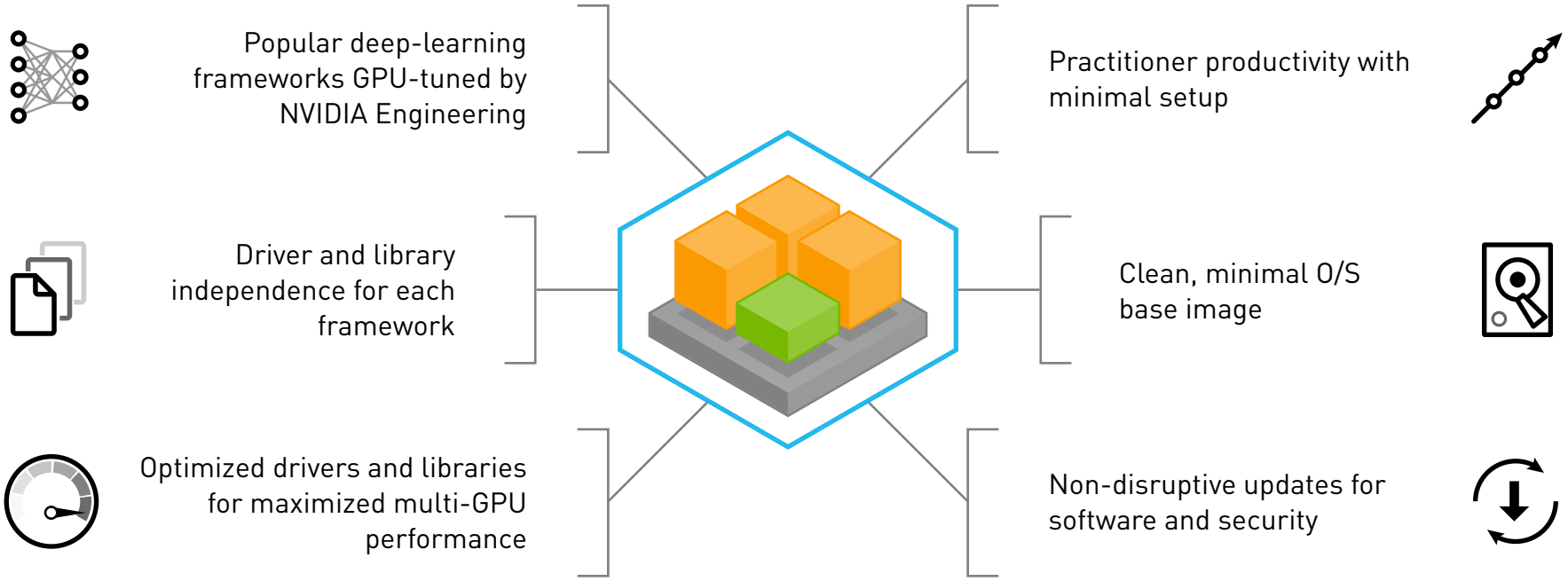
The DGX deep learning software stack is accessed from the NVIDIA GPU Cloud and integrates our optimized deep learning framework containers, libraries, drivers, and OS. It is tuned for maximum performance on the world's fastest GPU's. In fact, DGX software running on NVIDIA GPUs delivers 30% faster training than a DIY system using comparable hardware without DGX optimized software. If you had to replicate this level of software engineering investment, it would cost hundreds of thousands of dollars in OpEx.

Now organizations can unlock the full potential of their deep learning initiative with technology that "just works," right out of the box. Practitioners can get started faster, eliminate weeks or months spent cobbling together hardware and software componentry, simplify management of deep learning frameworks and underlying infrastructure, and realize faster time-to-insights.

WATCH WEBINAR TO DEEP-DIVE INTO THE DGX SOFTWARE STACK **LEARN MORE**

ENTERPRISE BENEFITS OF DGX SOFTWARE

NVIDIA DGX Systems Deliver Deep Learning Performance and Manageability



With DGX cloud management services, the bring-up experience is as simple as plug-in and power-up... start training in hours instead of weeks or months.

Let's explore the benefits of this software platform. Unlike DIY solutions we built DGX software with the goals of enabling a faster start, effortless productivity, and breakthrough performance as mentioned earlier.

DGX software takes the most popular deep learning frameworks, and through the painstaking work of NVIDIA's software engineering teams and DL experts, we ensure those frameworks are tuned for maximized performance on our GPUs. The containerized platform allows us to support multiple frameworks co-resident with each other, and each of them can use their versions of dependent libraries and drivers, for maximum flexibility.

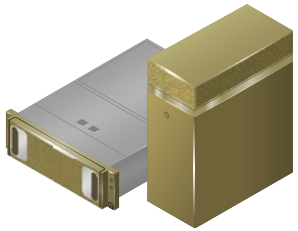
All of this keeps the base image to a clean, small footprint that's easily maintained along with streamlined over-the-network updates. And combined with our cloud management services, the bring-up experience is as simple as plug-in and power-up, allowing teams to start training in hours instead of weeks or months.

COMPARE DGX SYSTEMS WITH "DO IT YOURSELF" **LEARN MORE**

THE POWER TO RUN MULTIPLE FRAMEWORKS AT ONCE

NVIDIA DOCKER	NVIDIA DOCKER	NVIDIA DOCKER	NVIDIA DOCKER		NVIDIA DOCKER
TensorFlow	Microsoft Cognitive Toolkit	Caffe2	PYTORCH	• • •	Other Frameworks and Apps
TensorFlow Tuned Software	CNTK Tuned Software	Caffe2 Tuned Software	PyTorch Tuned Software		Tuned Software
CUDA RUNTIME	CUDA RUNTIME	CUDA RUNTIME	CUDA RUNTIME		CUDA RUNTIME

LINUX KERNEL + CUDA DRIVER



NVIDIA® DGX™ Systems

DGX Systems easily support the ability to run multiple versions of the same (or different) frameworks in parallel.

Many organizations support multiple projects and teams that require the ability to experiment across a wider array of deep learning platform configurations, often in parallel. This may involve running experiments on several different frameworks simultaneously, or even running multiple versions of the same framework, each supported by its own optimized software stack. This can often result in version, driver, or library conflicts due to this co-residency of software on the same base OS. It's for these reasons that DGX Systems are built on a containerized platform leveraging NVIDIA Docker.

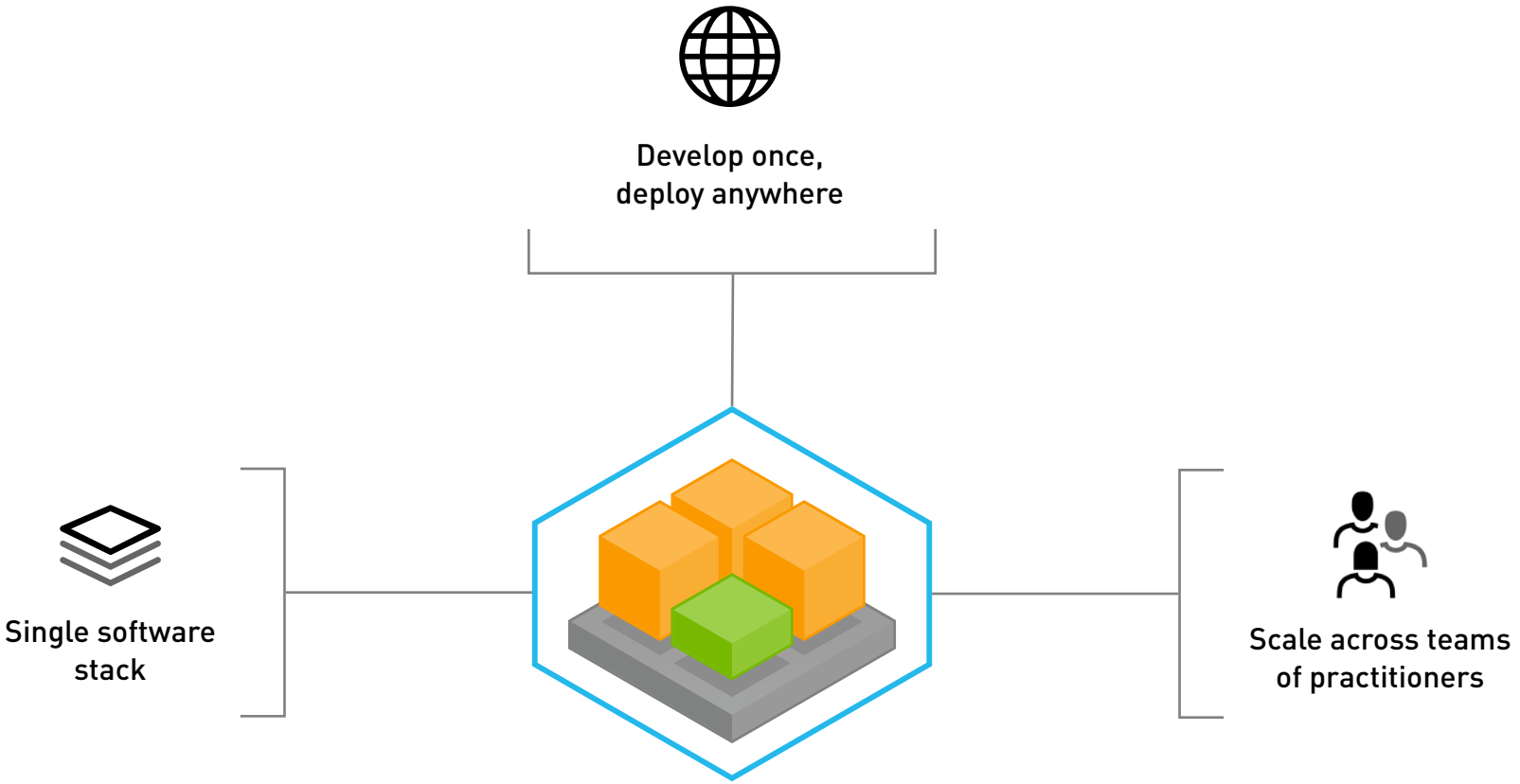
DGX Systems easily support the ability to run multiple versions of the same (or different) frameworks in parallel, on the same system. Each encapsulated within their own container, running with their own dependent libraries and drivers, including CUDA, cuDNN, and the NVIDIA Deep Learning SDK. All of which can be tuned independently based on the framework supported.

Teams no longer have to take a lowest common denominator approach in supporting a single OS image. Now you can get version independence and the flexibility to tune, tweak, and experiment within your own containerized stack. All of this with seamless portability and customization of your work without conflict, while maintaining a clean, minimized base OS image for the DGX System that's easy to update and maintain.

NVIDIA DOCKER ENCAPSULATES THE DGX SYSTEM SOFTWARE STACK **LEARN MORE**

DGX WITH NVIDIA GPU CLOUD

Benefits for Deep Learning Workflow



When you add it all up, probably the most important thing to note about DGX software is how it transforms DL workflow for an organization.

DGX cloud management via the NVIDIA GPU Cloud (NGC) is core to this transformation and includes essential capabilities including the container registry for centralized management of optimized frameworks made available by NVIDIA. This provides a simplified way to share frameworks across an extended organization of collaborators. Also included are job monitoring, scheduling, system health monitoring, and user administration. It supports both web and command line interface, allowing practitioners to use the tool that best fits their needs.

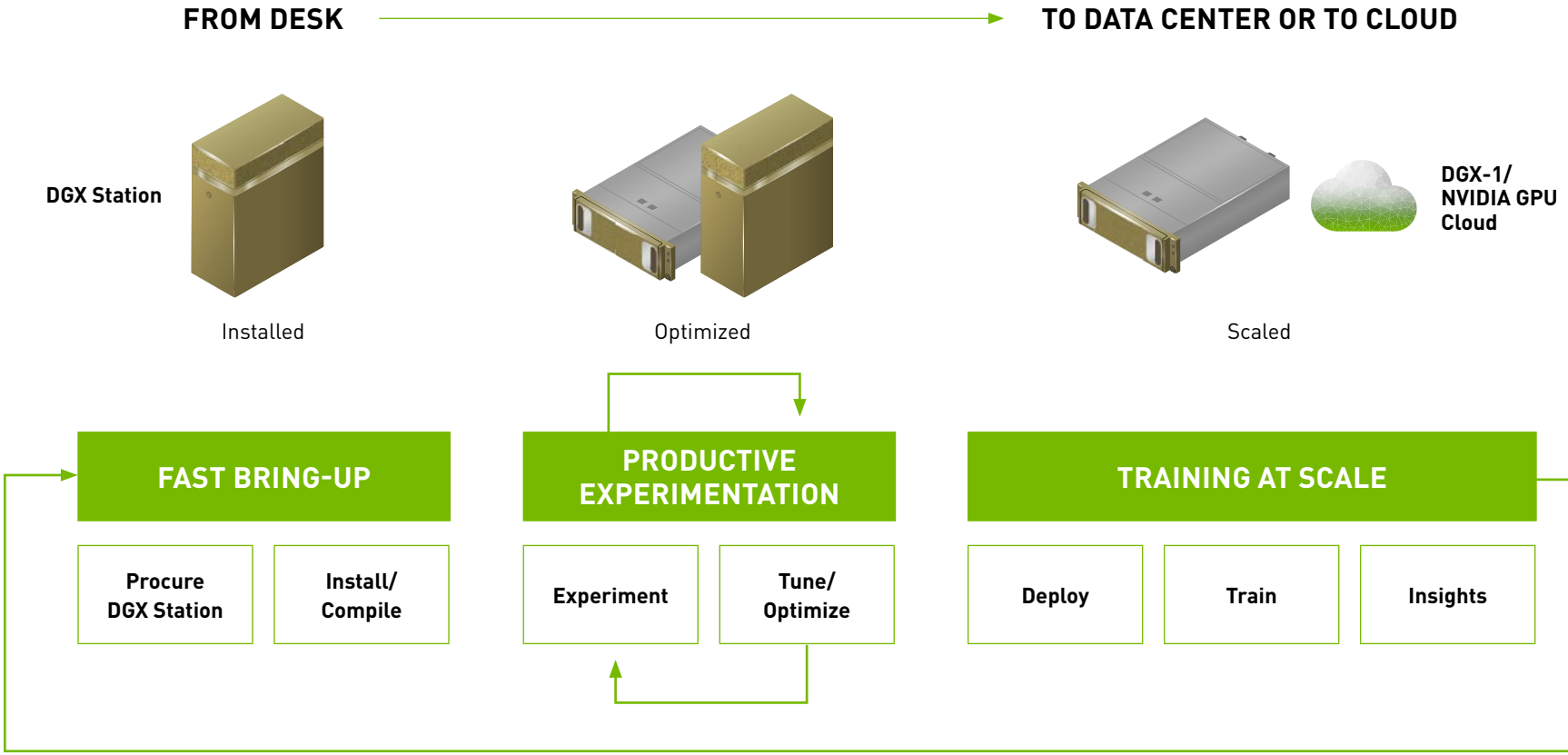
The combination of cloud management along with our optimized stack that's common to all our systems and the power of containers enables your organization to:

- **Leverage a software stack that insulates them from the dependent libraries and drivers, so they can focus on developing their frameworks**
- **Develop optimized frameworks once and deploy anywhere**
- **Scale your work with ease and improve collaboration across teams of practitioners**
- **Enjoy ubiquitous access to your optimized frameworks**

FROM DESK TO DATA CENTER TO CLOUD

DL FROM DEVELOPMENT TO PRODUCTION

DGX Systems Deliver an End-To-End Lifecycle Approach in Enabling Any Organization's Deep Learning Environment



This end-to-end approach enables the fastest, most streamlined cycle between deep learning experimentation and deriving insights at scale in production.

Starting with the researcher and developer, DGX Station extends the reach of AI supercomputing performance to the desk, enabling the fastest start possible in deep learning, with a rapid deployment experience that's as simple as plug-in, power-up. Data scientists can start training deep neural networks in hours instead of spending weeks configuring hardware and software. This fast bring-up, allows researchers to jump into productive experimentation almost immediately, enjoying the ability to run, tune, and optimize models with ease and speed.

Once they're ready for production training, DGX Station delivers the groundbreaking AI performance to support not only training but also inference, thanks to the 4 Tesla V100s with NVLink configuration. If greater scale in the data center is required, that same neural net, containerized on DGX Station, including supporting libraries and drivers, can be ported via the NVIDIA GPU Cloud and the DGX container registry, to any DGX System, including DGX-1 or the cloud.

This end-to-end approach enables the fastest, most streamlined cycle between deep learning experimentation and deriving insights at scale in production.

REAPING THE BENEFITS OF AI

Your AI initiative is critical to your organization's success, and dependent on a frequently optimized software stack and integrated hardware infrastructure. NVIDIA DGX Systems enable you to get the fastest start in AI and deep learning, delivering a rapid deployment experience, combined with effortless manageability, and revolutionary performance that lets you spend less time on IT, and more time gaining insights.

Combined with NVIDIA's enterprise support, your investment is protected with software upgrades and priority resolution of critical issues; giving you the peace of mind that your environment is tuned for maximized performance and uptime, enabling you to reap the benefits of AI and deep learning, faster than ever.

www.nvidia.com/dgx