



BREAKING THE BARRIERS WITH DL - SCALE ON AI SYSTEMS

Renee Yao | NVIDIA Senior Product Marketing Manager, Deep Learning and Analytics
September 2018 | Australia



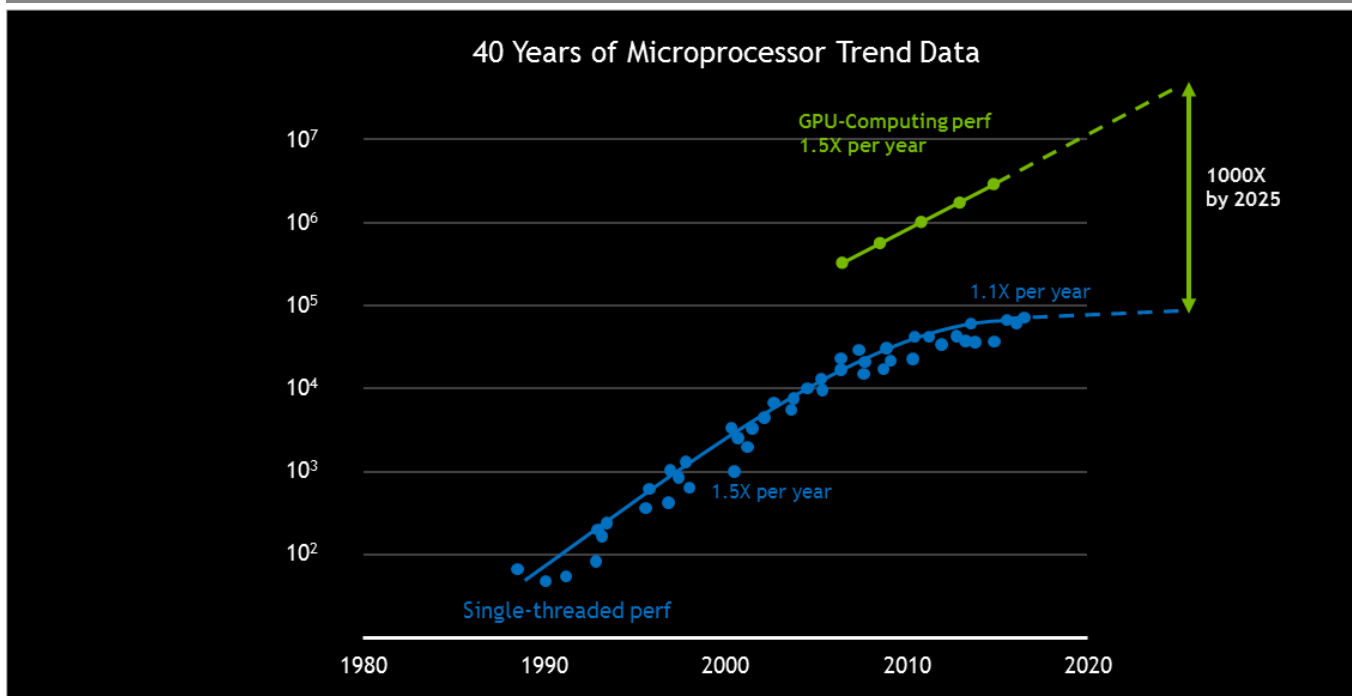
AGENDA

- The Rise of GPU Computing
- Why Infrastructure Matters to AI
- NVIDIA GPU Cloud
- DGX Systems
- Scaling AI Workloads
- Wrap-up

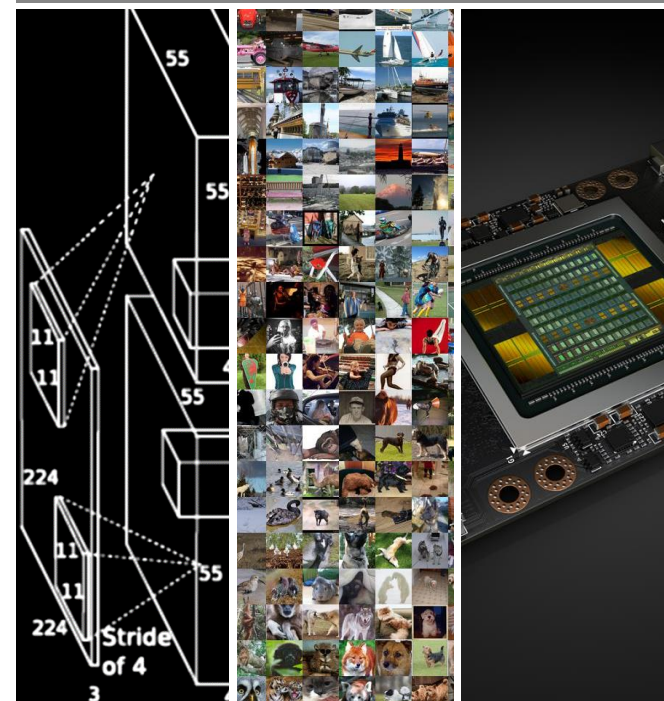
GPU COMPUTING AT THE HEART OF AI

New Advancements Leapfrog Moore's Law

Performance Beyond Moore's Law



Big Bang of Modern AI

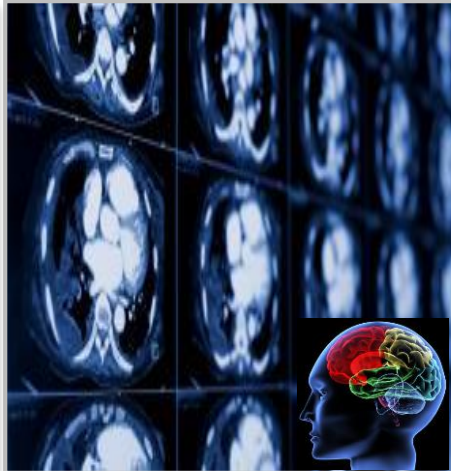


DEEP LEARNING IS SWEEPING ACROSS INDUSTRIES

Internet Services



Medicine



Media & Entertainment



Security & Defense



Autonomous Machines



- Image/Video classification
- Speech recognition
- Natural language processing

- Cancer cell detection
- Diabetic grading
- Drug discovery

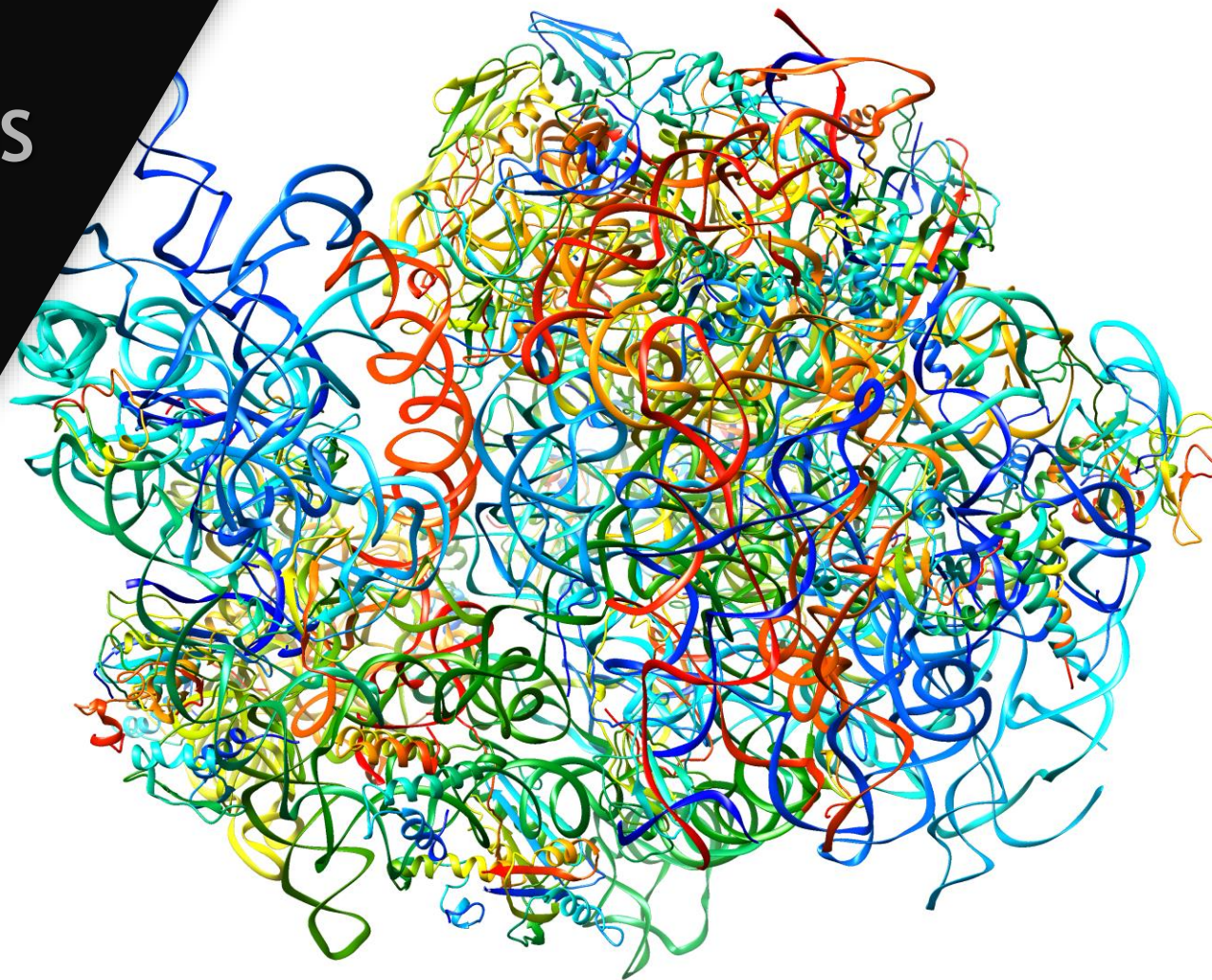
- Video captioning
- Content based search
- Real time translation

- Face recognition
- Video surveillance
- Cyber security

- Pedestrian detection
- Lane tracking
- Recognize traffic signs

SUPER DRUGS TO COMBAT SUPER BUGS

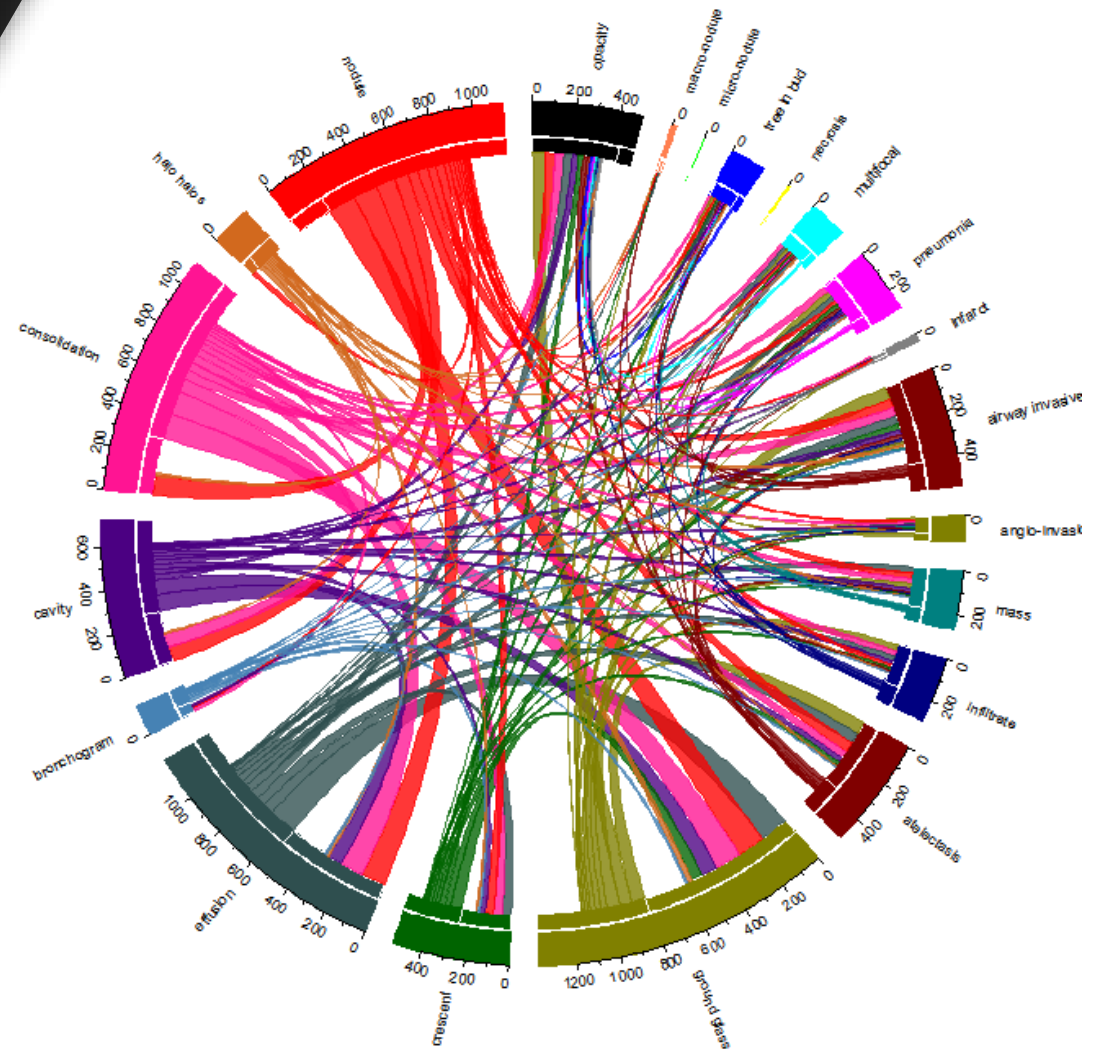
In the race of design more effective drugs and new treatments for infectious diseases, Australian scientists are using HPC and advanced imaging to visualize changes in ribosomes that occur in response to antibiotics. With MASSIVE's M3 supercomputer powered by >160 NVIDIA GPUs and 2 DGX-1V's to accelerate data processing, the team strives to identify new drugs that are lethal to bacteria despite structural changes in ribosomes.



One of Dr Matt Belousoff and Professor Trevor Lithgow ribosome structures illustrating the complicated details that can be determined using the Titan Krios at the Ramaciotti Centre for Cryo Electron Microscopy

AI PLATFORM TO ACCELERATE MEDICAL DISCOVERIES

Researchers at Monash University are using GPU-powered deep learning to detect invasive fungal disease from chest computed tomography imaging in hematology-oncology patients. Trained on data acquisition augmented by natural language processing of chest CT reports, the team is designing a platform to accelerate knowledge discovery of rare diseases.

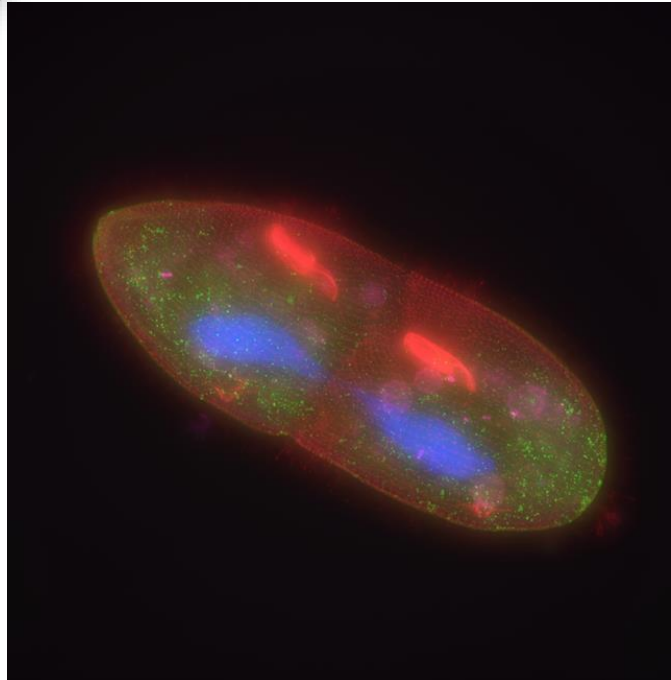


Co-occurrence analysis of key symptoms for fungal disease

Dr. Reza Haffari has built a deep learning based CT record classifier based on these features to help doctors in Alfred Hospital make medical diagnosis, achieving 86.96% accuracy.

FASTER INTERPRETATION OF MICROSCOPY DATA

The Research Computing Centre (RCC) at UQ uses the Wiener supercomputer to expedite research in a diverse range of imaging-intensive sciences. Using deconvolution algorithms, machine learning and pattern recognition techniques, Wiener provides near real-time outputs of deconvolved, tagged and appropriately characterized data, providing researchers with immediate feedback on data quality and allowing for faster interpretation of microscopy data.

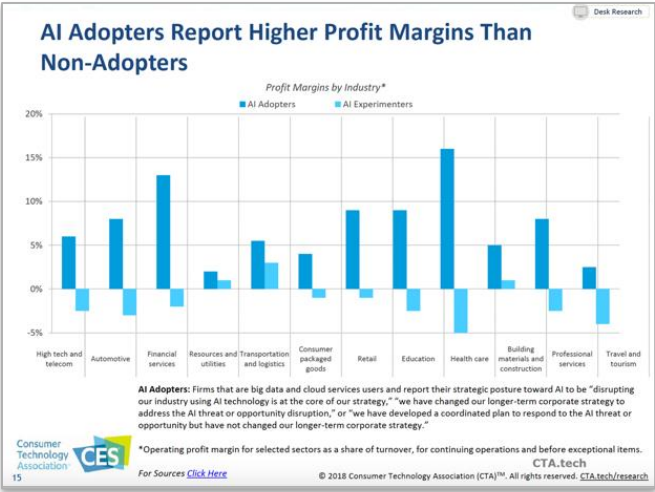


UQ's new Wiener supercomputer will add sharp detail to microscopic imagery

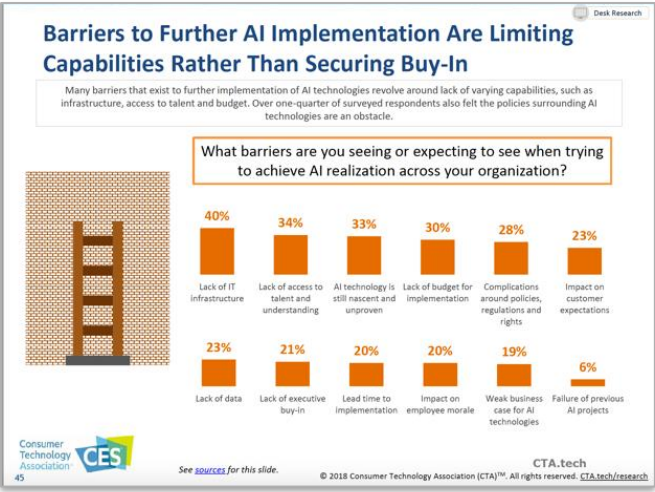


**WHY INFRASTRUCTURE
MATTERS TO AI**

AI ADOPTERS IMPEDED BY INFRASTRUCTURE



AI Boosts Profit Margins up to 15%



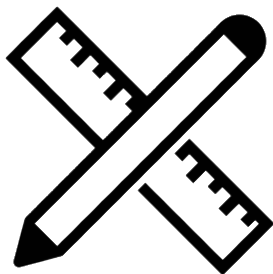
40% see infrastructure as impeding AI

source: 2018 CTA Market Research

THE CHALLENGE OF AI INFRASTRUCTURE

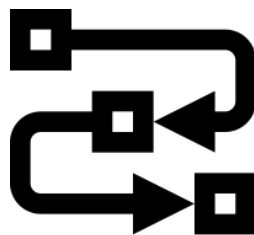
Short term thinking leads to longer term problems

DESIGN
GUESSWORK



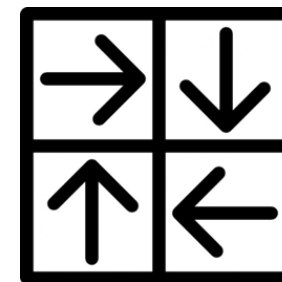
Ensuring the architecture delivers predictable performance that scales

DEPLOYMENT
COMPLEXITY

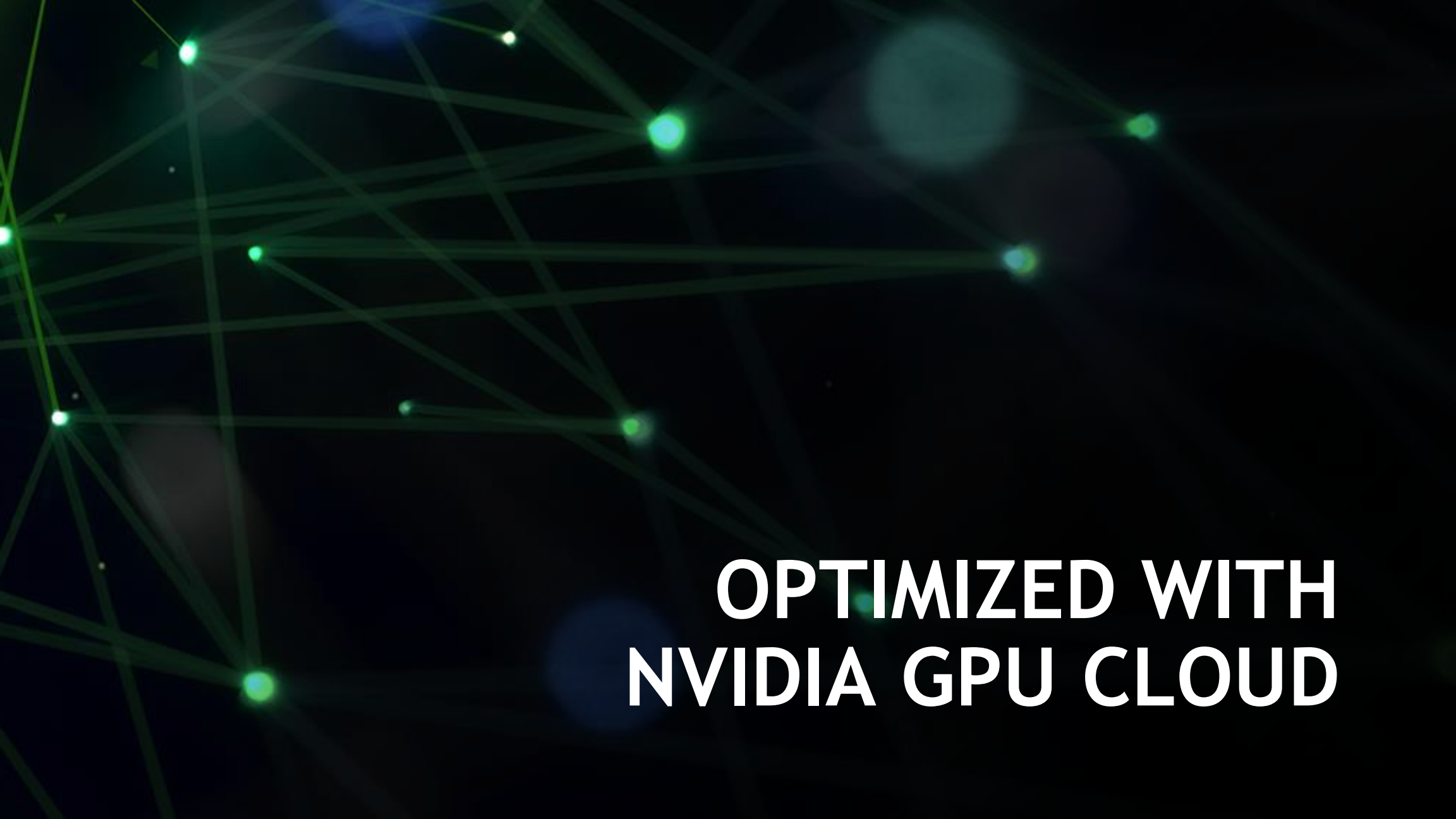


Procuring, installing and troubleshooting compute, storage, networking and software

MULTIPLE POINTS
OF SUPPORT



Contending with multiple vendors across multiple layers in the stack

An abstract network visualization featuring a complex web of thin, light green lines connecting various nodes. The nodes are represented by small, glowing green circles of varying sizes and brightness, scattered across a dark, almost black background. The overall effect is that of a dynamic, interconnected system, possibly representing a data network or a computational graph.

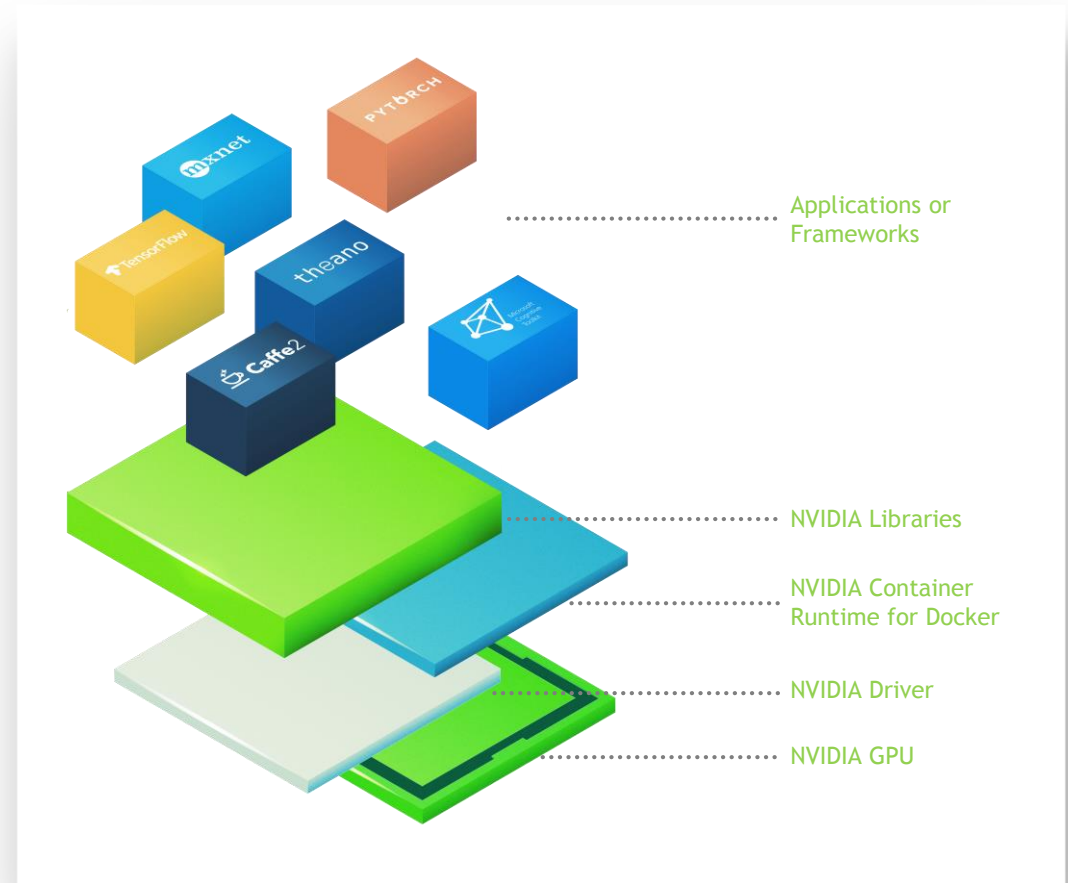
**OPTIMIZED WITH
NVIDIA GPU CLOUD**

CHALLENGES WITH COMPLEX SOFTWARE

Current DIY GPU-accelerated AI and HPC deployments are **complex** and **time consuming** to build, test and maintain

Development of software frameworks by the community is moving **very fast**

Requires high level of **expertise** to manage driver, library, framework dependencies



FROM DEVELOPMENT TO DEPLOYMENT

With NVIDIA GPU Cloud Containers

Deep Learning, HPC, HPC Visualization



NGC CONTAINER REGISTRY

10 Containers at Launch, 35 Containers Today

Deep Learning	HPC	HPC Visualization	NVIDIA/K8s	Partners
caffe	bigdft	index	Kubernetes on NVIDIA GPUs	chainer
caffe2	candle	paraview-holodeck		h20ai-driverless
cntk	chroma	paraview-index		kinetica
cuda	gamess	paraview-optix		mapd
digits	gromacs			paddlepaddle
inferenceserver	lammps			
mxnet	lattice-microbes			
pytorch	milc			
tensorflow	namd			
tensorrt	pgi			
theano	picongpu			
torch	relion			
	vmd			



NVIDIA
GPU CLOUD

KUBERNETES ON NVIDIA GPUS

GPU-Aware Automation and Orchestration

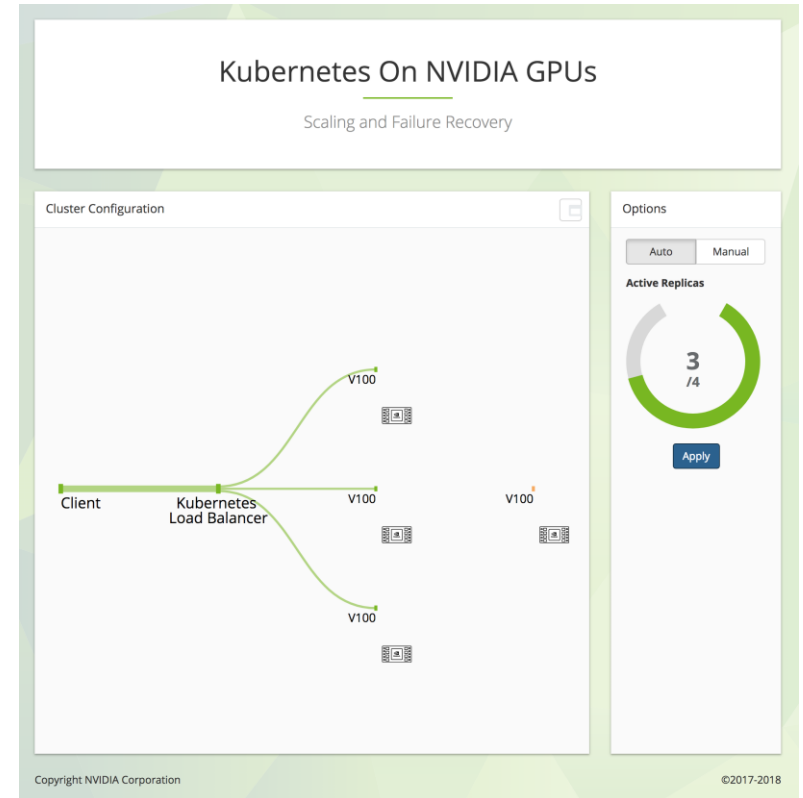
Deploy and manage multi node, hybrid cloud infrastructure for GPU-accelerated applications

Orchestrate resources on heterogeneous GPU clusters, including DGX Systems and NVIDIA GPU-enabled cloud service providers

Optimize GPU cluster utilization with active health monitoring

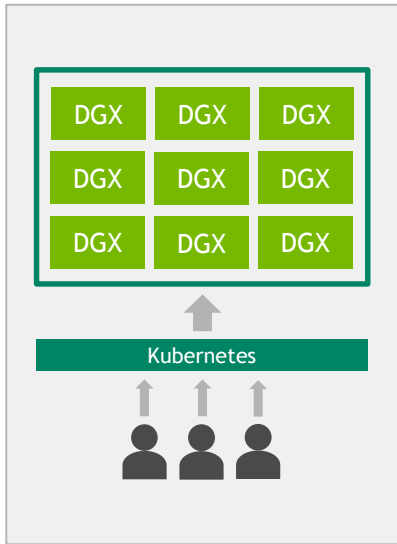
Deploy wide range of applications using multiple container technologies such as Docker and CRI-O

Installer containers are hosted on NGC

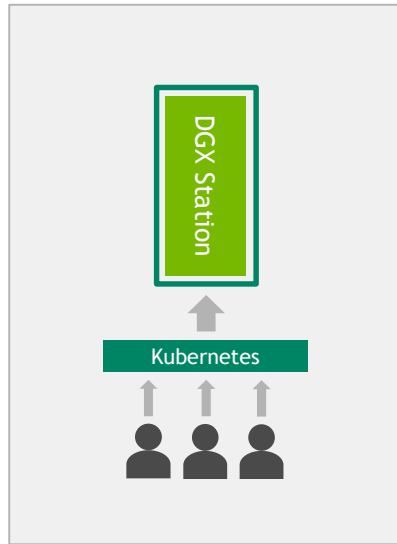


USE CASES

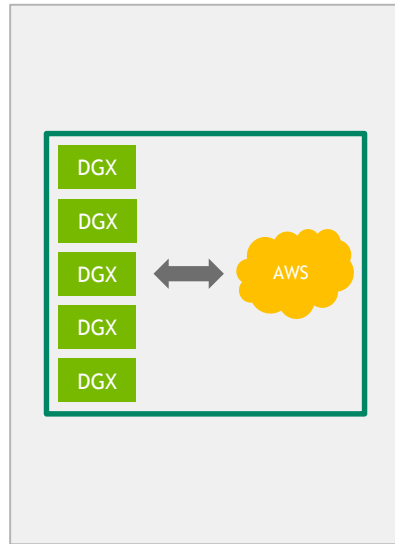
Where can it be leveraged?



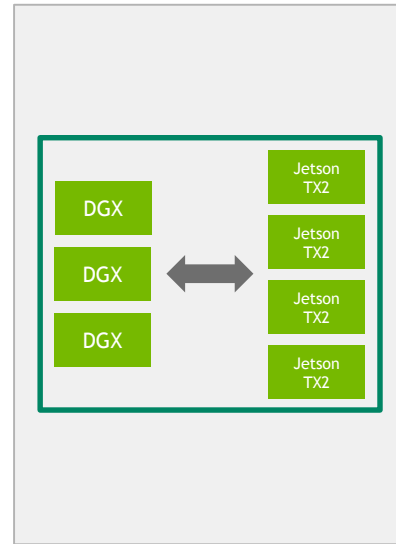
Many users, many nodes
On-prem



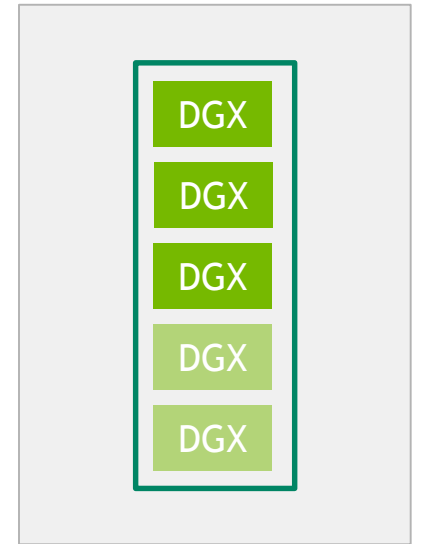
Many users, single node
On-prem



Cloud bursting
Hybrid



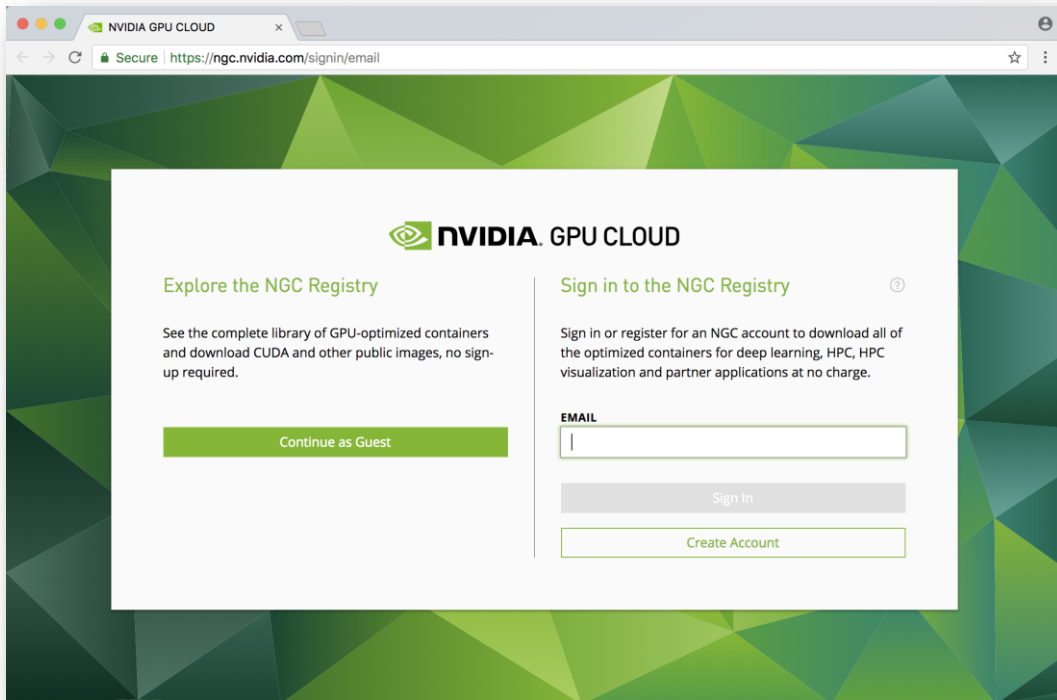
Edge/IoT
Multi-region



Production Inferencing
*

GET STARTED TODAY WITH NGC

Sign Up and Access Containers for Free



To learn more about all of the GPU-accelerated software from NGC, visit:

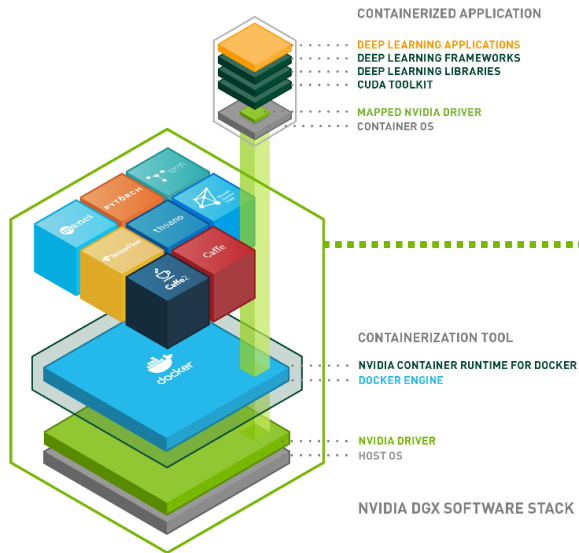
nvidia.com/cloud

To sign up or explore NGC, visit:

ngc.nvidia.com

PURPOSE-BUILT AI SUPERCOMPUTERS

NGC DL SOFTWARE STACK



- Universal SW for Deep Learning
- Predictable execution across platforms
- Pervasive reach

AI WORKSTATION

DGX Station



The Personal
AI Supercomputer

AI DATA CENTER

DGX-1



The Essential
Instrument for AI
Research

DGX-2



The World's Most Powerful
AI System for the Most
Complex AI Challenges

The background features a complex network of thin, glowing green lines connecting various nodes. Some nodes are bright green, while others are a soft blue. The overall effect is a sense of digital connectivity and data flow against a dark, almost black background.

**GET STARTED WITH
DGX STATION**

INTRODUCING NVIDIA DGX STATION



The Fastest Personal Supercomputer for Researchers and Data Scientists



Revolutionary form factor -
designed for the desk, whisper-quiet



Start experimenting in hours,
not weeks, powered by DGX Stack



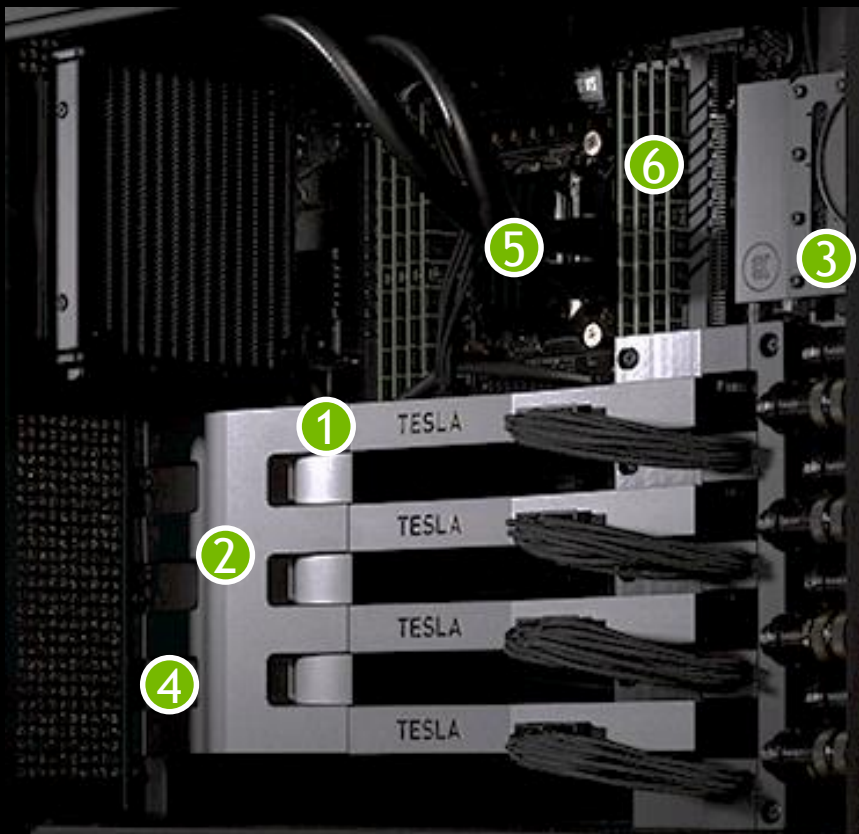
Productivity that goes from desk
to data center to cloud



Breakthrough performance and
precision - powered by Volta

INTRODUCING NVIDIA DGX STATION

Groundbreaking AI - at your desk



The Personal AI Supercomputer for Researchers and Data Scientists

Key Features

1. 4 x NVIDIA Tesla V100 GPU
2. 2nd-gen NVLink (4-way)
3. Water-cooled design
4. 3 x DisplayPort (4K resolution)
5. Intel Xeon E5-2698 20-core
6. 256GB DDR4 RAM

NEXT GEN AI-POWERED IMAGE RECOGNITION SOLUTIONS

Fast and easy to implement functionalities are vital in helping organizations enhance services. Imagga—a global provider of platform-as-a-service and customized on-premise AI-based image recognition solutions—helps developers and enterprises optimize image-based projects and quickly build scalable, image-intensive cloud apps.

With an API and solutions built on the NVIDIA DGX Station to speed training and inference, Imagga has sped service delivery by up to 88%, while maintaining high levels of accuracy.



Training History

Co-located server:

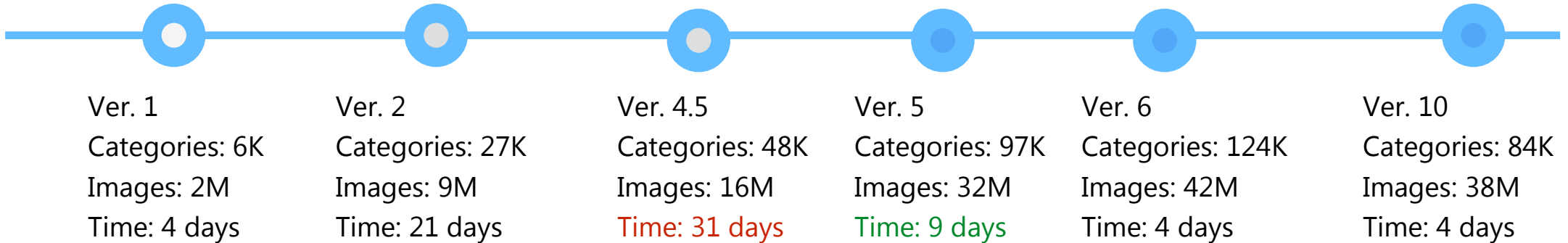
2x NVIDIA GTX Titan 1st gen
 2x 6GB VRAM
Intel Xeon E5-2620 @ 2.10GHz
 6x CPU Cores
64GB RAM

Azure instance:

4x NVIDIA Tesla K80
 4x 12GB VRAM
Intel Xeon E5-2690 v3 @ 2.60GHz
 12x CPU Cores
220GB RAM

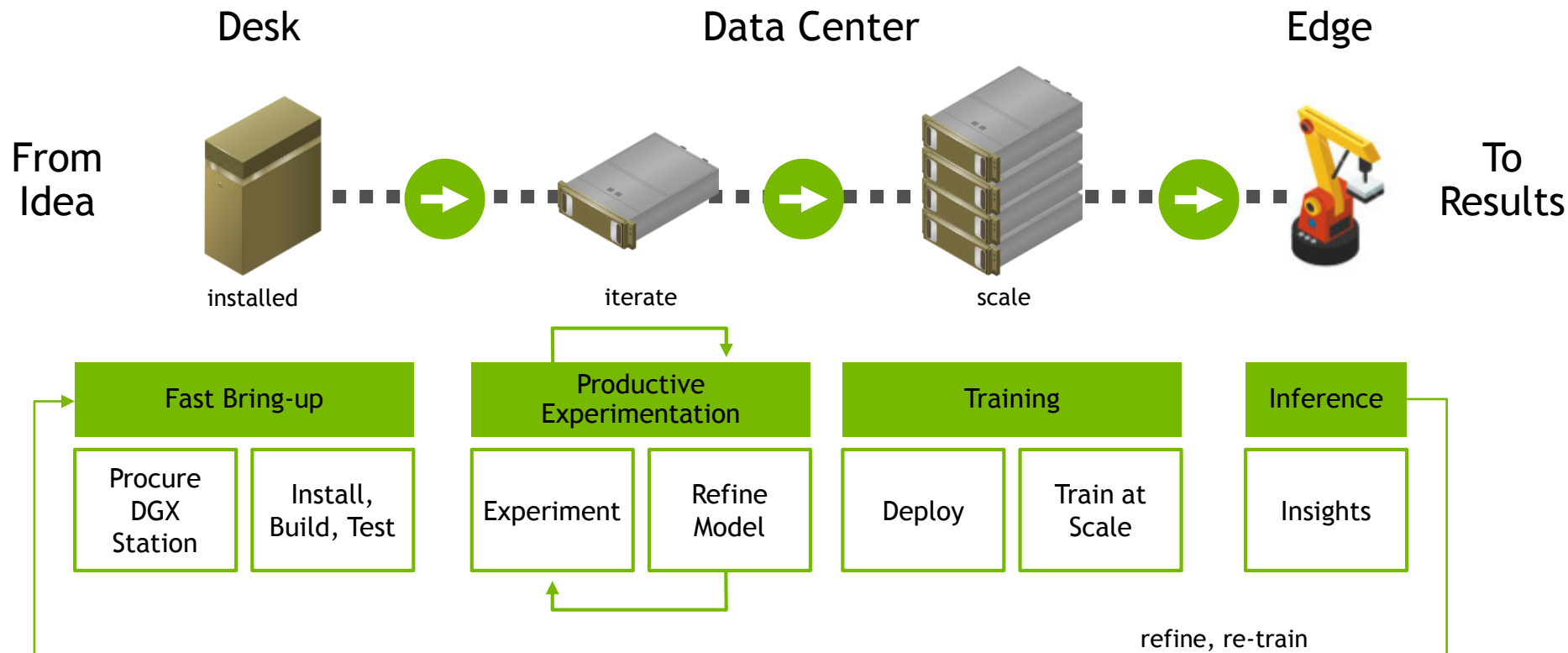
Co-located DGX Station:

4x NVIDIA Tesla V100
 4x 16GB VRAM
Intel Xeon E5-2698 v4 @ 2.20GHz
 20x CPU Cores
256GB RAM



DL FROM DEVELOPMENT TO PRODUCTION

Accelerate Deep Learning Value



AI-ENABLED DATA DELIVERS GREATER BUSINESS INSIGHTS

Insights from call center audio scoring help companies improve employee training, enhance lead qualification, increase sales, improve customer satisfaction, and reduce employee turnover. But because call scoring is manual and time-consuming, most call centers only listen to ~2% of their recorded calls.

Powered by the NVIDIA DGX Station and DGX-1 AI supercomputers, Deepgram delivers a game-changing technology that recognizes audio more completely and precisely, enabling companies to utilize 100% of their recorded calls for greater insights to fuel business decisions.



Call ID: 323689

Agent: Xu Bei | Contact: +16074157890 | July 12, 2017 | 25:43 PST



TAGS IDENTIFIED:

COMPLIANCE ISSUE ✕

COMPETITOR MENTION ✕

BILLING ISSUE ✕

00:00:56

I am calling because I received a wrong bill. I just paid two days ago and my payment is not reflected in the bill. This is the third time this is happened. **I'm very close to switching providers.**

00:01:00

sorry for the inconvenience madam may i have your account number please

00:01:02

five three four zero zero three six five four eight



00:56

18:08



DGX STATION WEBINAR SERIES

Title: Gaining Insights from Audio: AI-Enabled Voice Data Analytics for Enterprise

Date: Wednesday, September 12, 2018

Time: 09:00 AM Pacific Daylight Time



Scott Stephenson
CEO
Deepgram



Tony Paikeday
Director of Product Marketing, NVIDIA DGX portfolio of
A.I. supercomputers
NVIDIA

Title: Using AI to Look Deeper: Transforming Business with Next-Gen Image Recognition

Date: Wednesday, September 19, 2018

Time: 09:00 AM Pacific Daylight Time



Georgi Kadrev
CEO
Imagga



Renee Yao
Senior Product Marketing Manager, DGX Station
NVIDIA

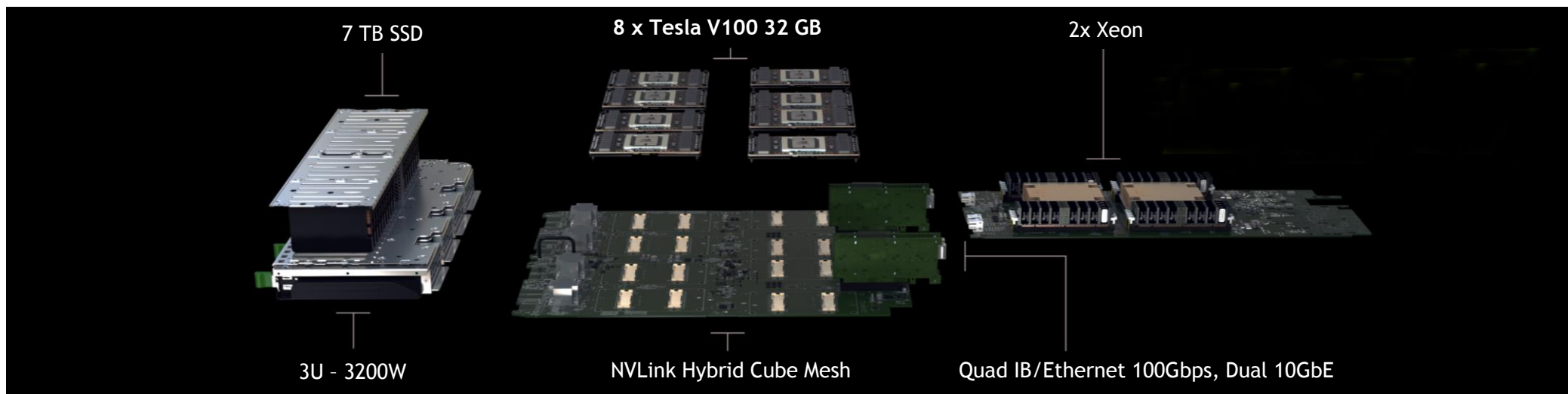
<https://webinars.on24.com/NVIDIA/stationwebinarseries>

A network diagram with green nodes and lines on a dark background. The nodes are represented by small, glowing green circles of varying sizes, connected by thin, light green lines. The lines form a complex web of connections, with some nodes having multiple links. The overall aesthetic is futuristic and technical, typical of a data center or network infrastructure visualization.

DEPLOY ON DGX-1

NVIDIA DGX-1 WITH VOLTA

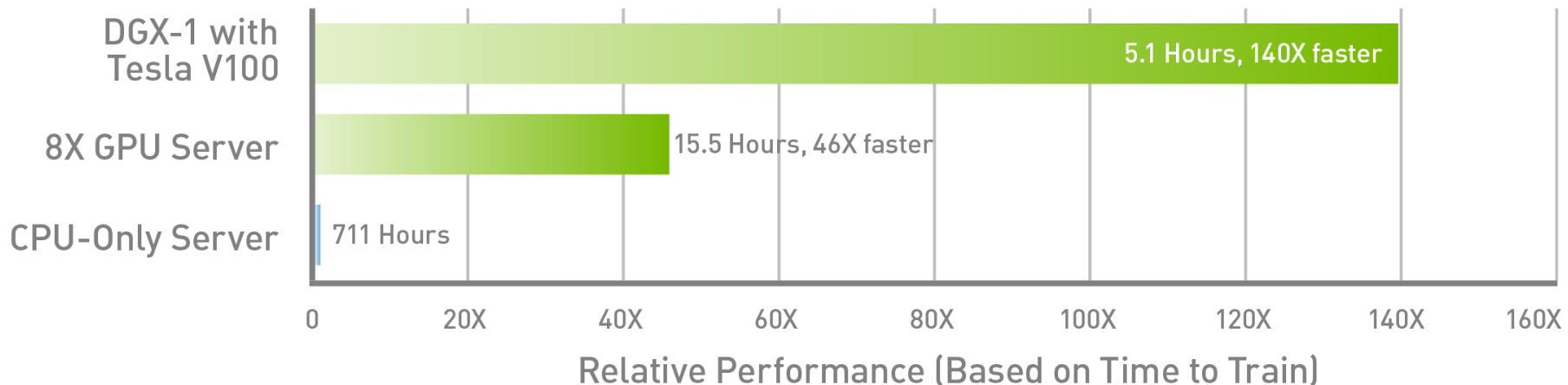
Highest Performance, Fully Integrated HW System



1 PetaFLOPS | **8x Tesla V100 32GB** | **300 Gb/s** NVLink Hybrid Cube Mesh
2x Xeon | 7 TB RAID 0 | Quad IB/Ethernet 100Gbps, Dual 10GbE | 3U – 3500W

DGX-1: 140X FASTER THAN CPU

NVIDIA DGX-1 Delivers 140X Faster Deep Learning Training



Workload: ResNet-50, 90 epochs to solution | CPU Server: Dual Xeon E5-2699v4, 2.6GHz



AI DETECTS GROWTH PROBLEMS IN CHILDREN

Detecting growth-related problems in children requires calculating their bone age. But it's an antiquated process that requires radiologists to match X-rays with images in a 1950s textbook. Massachusetts General Hospital, which conducts the largest hospital-based research program in the United States, developed an automated bone-age analyzer built on the NVIDIA DGX-1 with CUDA. The system is 99% accurate and delivers test results in seconds versus days.



MASSACHUSETTS
GENERAL HOSPITAL

AI HELPS CREATE BEAUTIFUL MUSIC

Music plays an important role in the emotional and communicative experience of media, but it's difficult for composers to produce music for the dynamic, user driven situations of interactive experiences. Monash researchers train deep recurrent neural networks on datasets of tens of thousands of music compositions as part of a new adaptive music system for interactive media.

The NVIDIA DGX-1 could train these networks much faster, allowing for more exploration of datasets and model architectures, and significantly improving the quality of results.



DL Driven Risk Models

One data-scientist, one DGX-1, six weeks, 48 months of data and an internally trained loss prediction model that met industry benchmarks.

The DGX-1 provided an instant uplift in analyst productivity--a treasured commodity in this market--allowing him to reduce software (TensorFlow) setup as well as network training time to iterate through new and evolving models.

The DL methods explored not only demonstrated acceleration in terms of training speed, but that sufficiently sophisticated loss prediction could be made to a reasonable level of accuracy against current industry benchmarks.



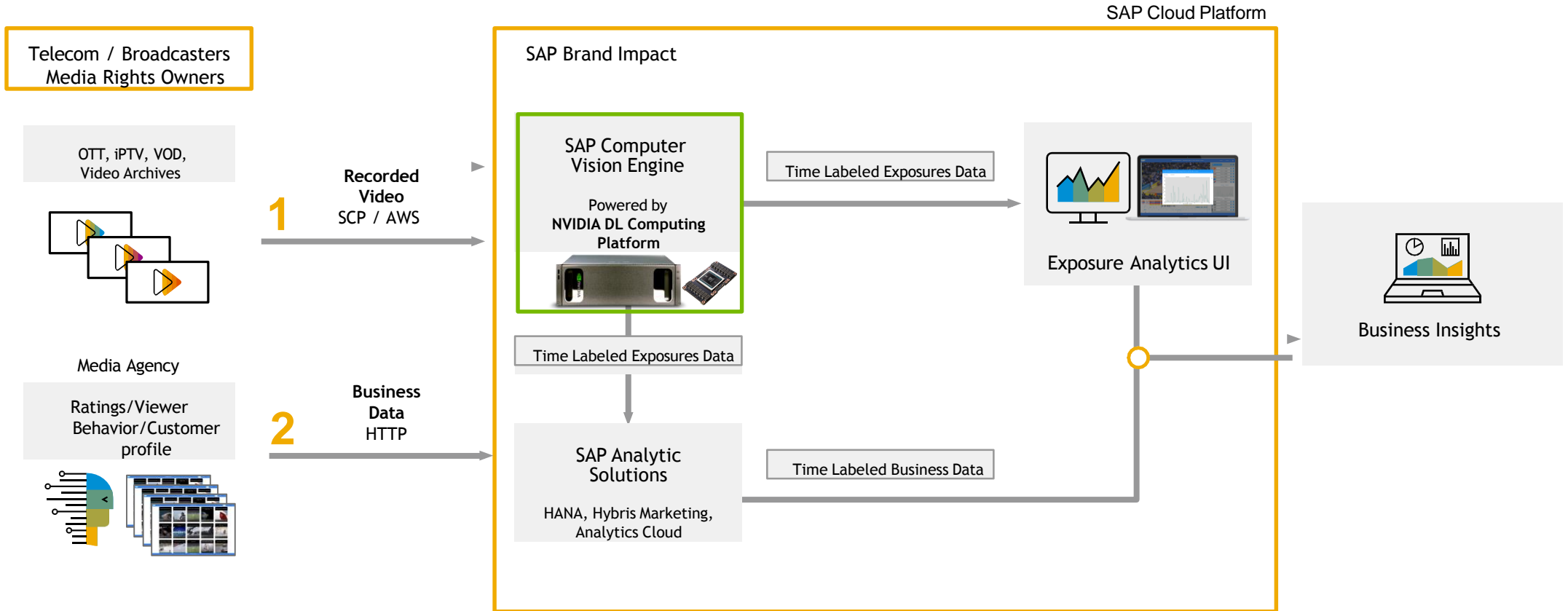
A NEW WAVE OF AI BUSINESS APPLICATIONS

Many companies rely on sponsoring televised events to promote their brands, yet impact is difficult to track. Manual tracking takes up to six weeks to measure ROI and even longer to adjust expenditures. SAP Brand Impact, powered by NVIDIA deep learning, measures brand attributes in near real-time with superhuman accuracy. With deep neural networks trained on the NVIDIA DGX-1 and the TensorRT inference engine, the SAP team achieves a 40x performance improvement and reduces hourly costs by 32x. Results are immediate, accurate and auditable.



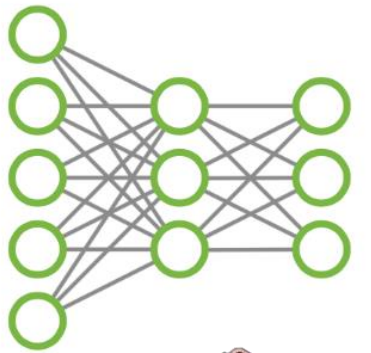
SAP Brand Impact

The Business Impact of Marketing and Sponsorship in Video Content



DRAMATIC BOOST IN ACCURACY WITH LARGER, MORE COMPLEX MODELS

SAP Brand Impact on DGX-1 (32 GB) for Object Detection



V100 (16 GB)

VGG-16
(16 Layers)



V100 (32 GB)

ResNet-152
(152 Layers)

40% Reduced Error Rate

More Complex
Models Now Possible



Dramatic Boost
in Accuracy

Dataset: Winter Sports 2018 Campaign; high definition resolution images (1920 x1080)

A network diagram with glowing green nodes and lines on a dark background. The nodes are connected by a complex web of lines, creating a dense network structure. The nodes are scattered across the frame, with some appearing brighter than others. The lines are thin and green, connecting the nodes in various directions. The overall aesthetic is futuristic and technological.

SCALE UP WITH DGX-2

NVIDIA DGX-2

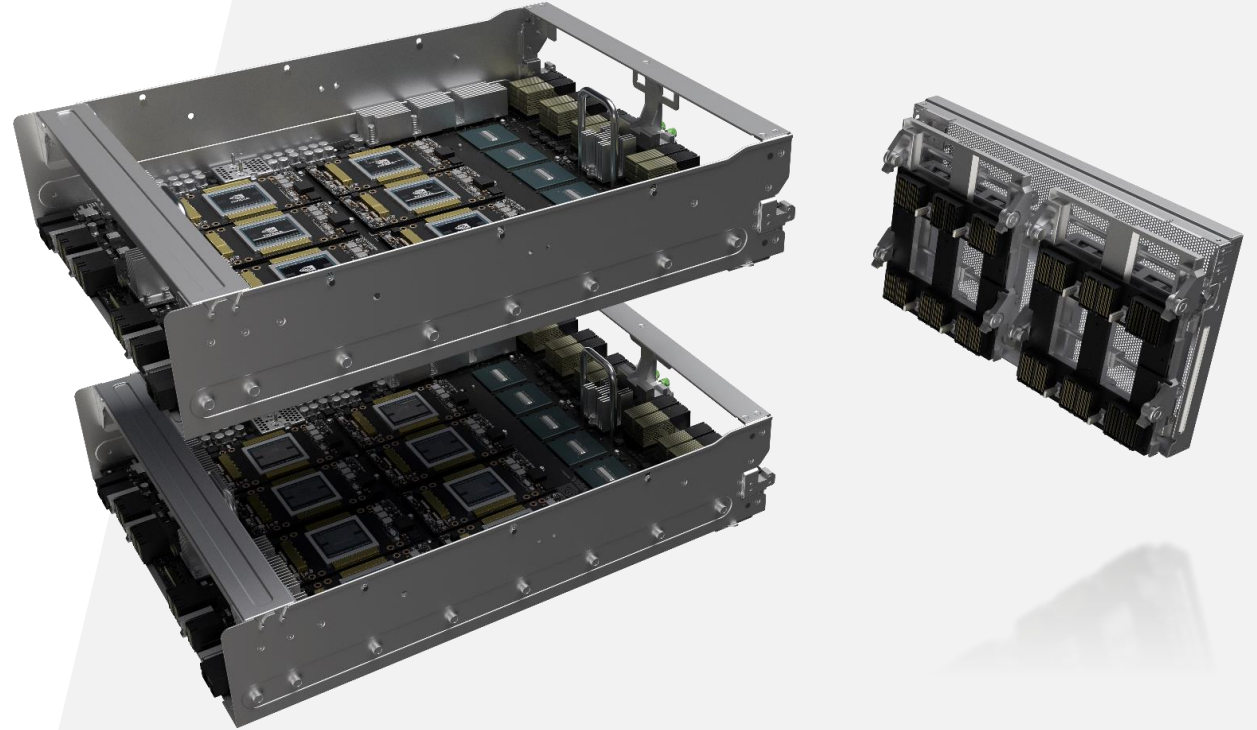
THE WORLD'S MOST POWERFUL DEEP LEARNING SYSTEM
FOR THE MOST COMPLEX DEEP LEARNING CHALLENGES

- First 2 PFLOPS System
- 16 V100 32GB GPUs Fully Interconnected
- NVSwitch: 2.4 TB/s bisection bandwidth
- 24X GPU-GPU Bandwidth
- 0.5 TB of Unified GPU Memory
- 10X Deep Learning Performance



THE WORLD'S FIRST 16 GPU AI PLATFORM

- Revolutionary SXM3 GPU package design
- Innovative 2 GPU board interconnect
- 32GB HBM2 stacked memory per GPU

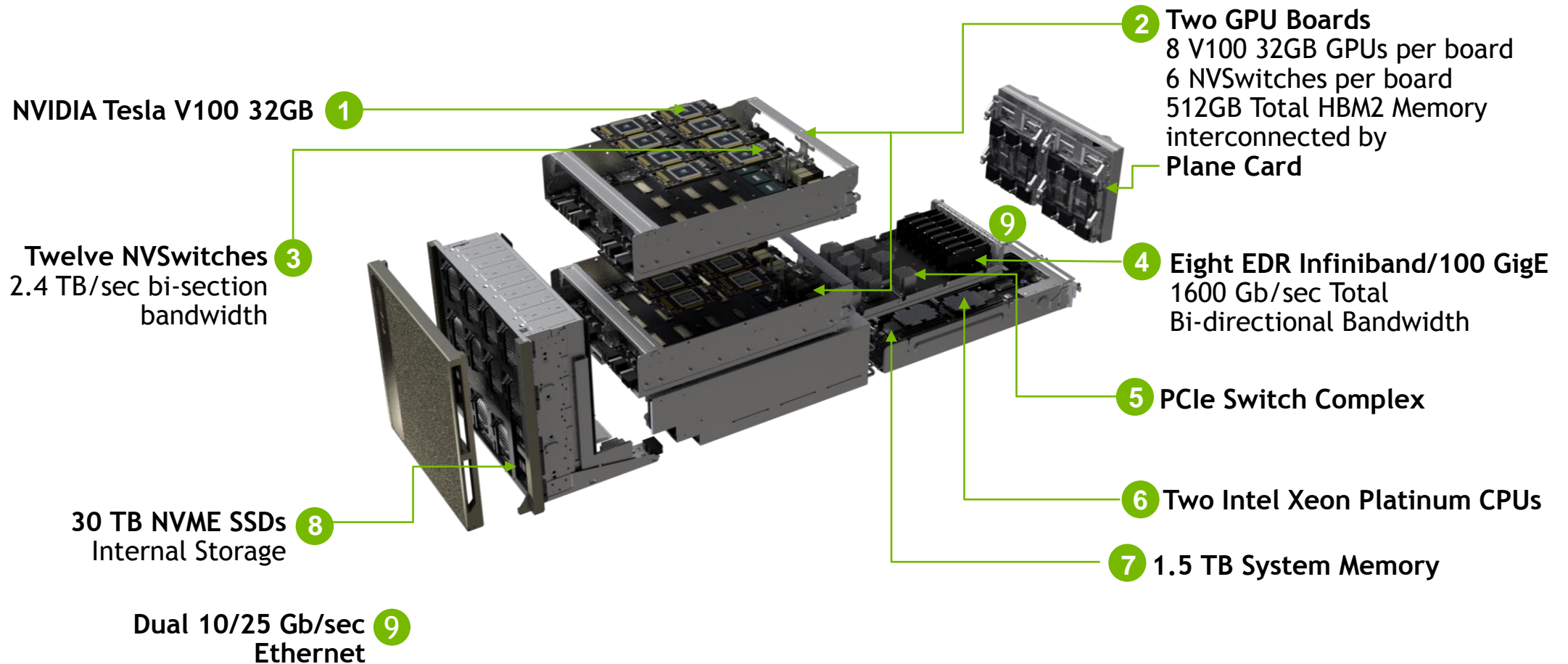


NVSWITCH: THE REVOLUTIONARY AI NETWORK FABRIC

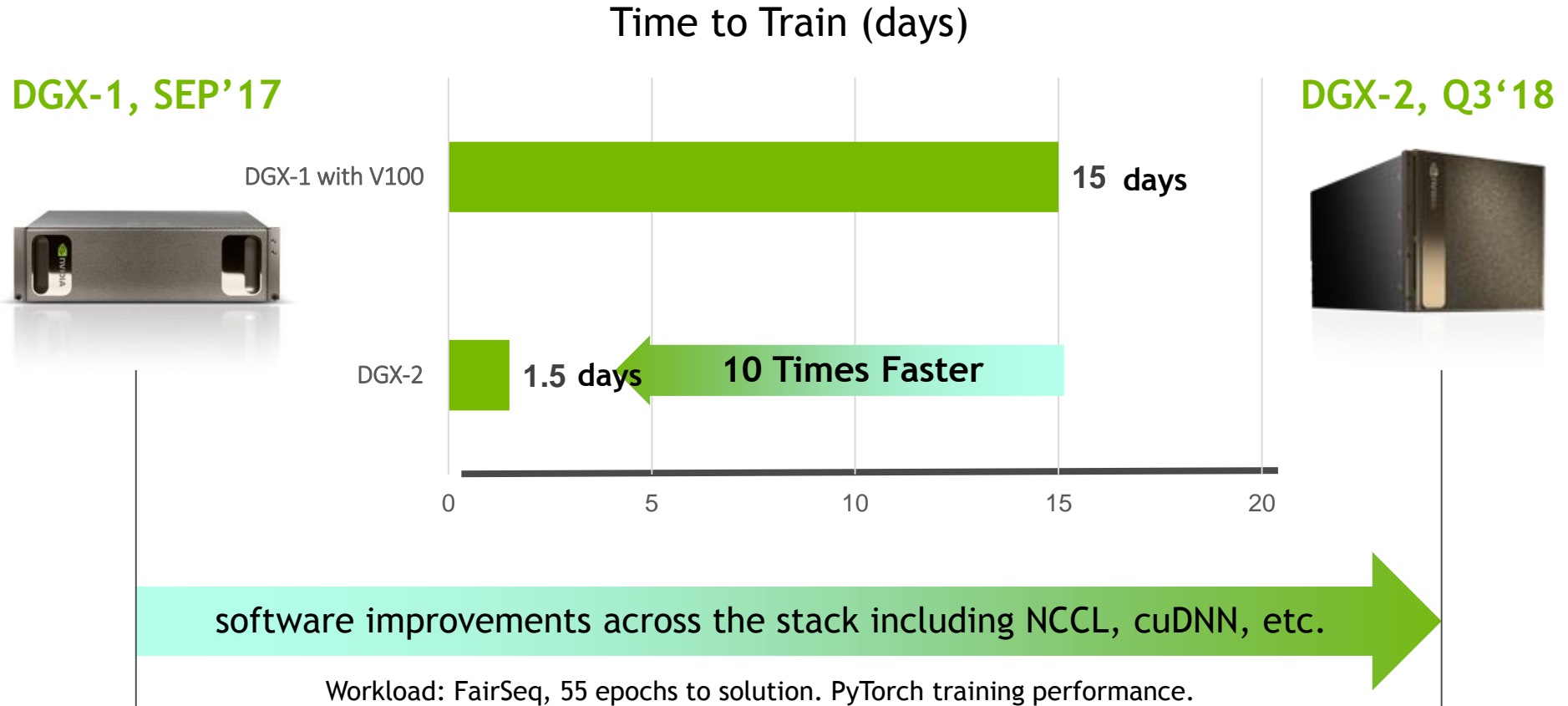


- Inspired by leading edge research that demands unrestricted model parallelism
- Like the evolution from dial-up to broadband, NVSwitch delivers a networking fabric for the future, today
- Delivering 2.4 TB/s bisection bandwidth, equivalent to a PCIe bus with 1,200 lanes
- NVSwitches on DGX-2 = all of Netflix HD <45s

DESIGNED TO TRAIN THE PREVIOUSLY IMPOSSIBLE



10X PERFORMANCE GAIN IN LESS THAN A YEAR





GETTING STARTED WITH DGX REFERENCE ARCHITECTURES

THE CHALLENGE OF AI INFRASTRUCTURE

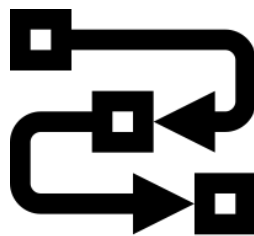
Short term thinking leads to longer term problems

DESIGN
GUESSWORK



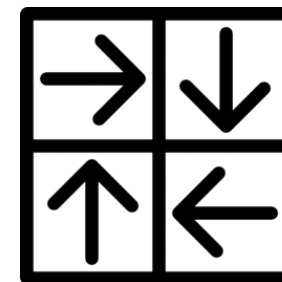
Ensuring the architecture delivers predictable performance that scales

DEPLOYMENT
COMPLEXITY



Procuring, installing and troubleshooting compute, storage, networking and software

MULTIPLE POINTS
OF SUPPORT



Contending with multiple vendors across multiple layers in the stack

THE VALUE OF AI INFRASTRUCTURE WITH DGX REFERENCE ARCHITECTURES

SCALABLE PERFORMANCE



Reference architectures from NVIDIA and leading storage partners

FASTER, SIMPLIFIED DEPLOYMENT



Simplified, validated, converged infrastructure offers

TRUSTED EXPERTISE AND SUPPORT

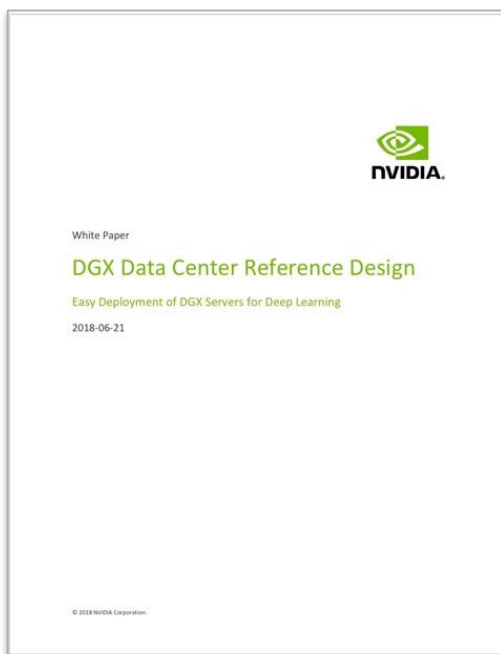


Available through select NPN partners as a turnkey solution

NVIDIA DGX REFERENCE ARCHITECTURES

Converged Infrastructure Solutions using DGX POD best practices

DGX POD Ref. Arch.



Storage Partner DGX-1 RA Solutions

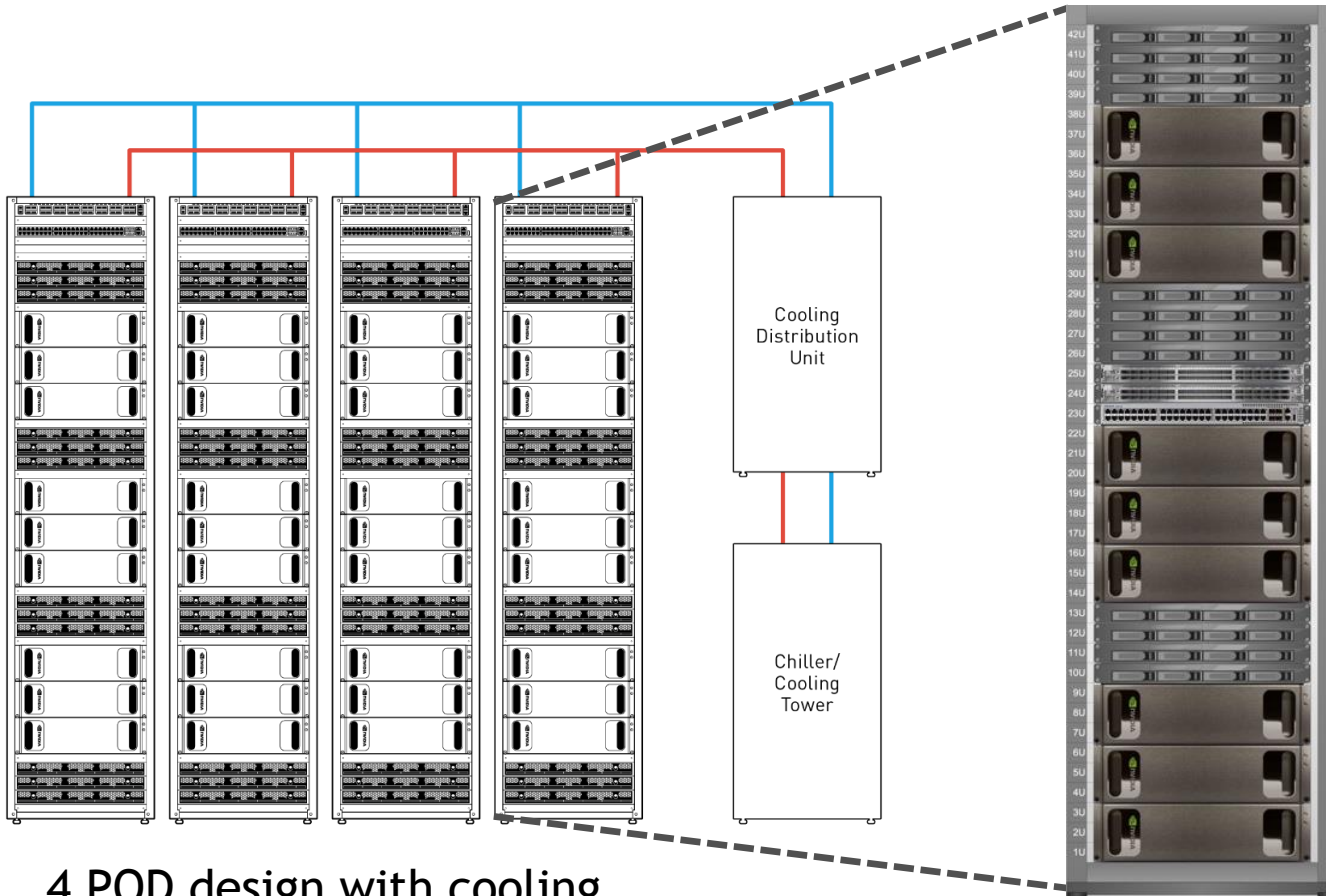


Common benefits:

- Eliminate design guesswork
- Faster, simpler deployment
- Predictable performance at scale
- Simplified, single-point of support

Backed by prioritized NPN partners

NVIDIA DGX POD™: HIGH-DENSITY COMPUTE REFERENCE ARCH.



4 POD design with cooling

DGX-1 POD

- NVIDIA DGX POD
- Support scalability to hundreds of nodes
- Based on proven SATURNV architecture

Nine DGX-1 Servers

- Eight Tesla V100 GPUs
- NVIDIA GPU Direct™ over RDMA support
- Run at MaxQ
- 100 GbE networking (up to 4 x 100 GbE)

Twelve Storage Nodes

- 192 GB RAM
- 3.8 TB SSD
- 100 TB HDD (1.2 PB Total HDD)
- 50 GbE networking

Network

- In-rack: 100 GbE to DGX-1 servers
- In-rack: 50 GbE to storage nodes
- Out-of-rack: 4 x 100 GbE (up to 8)

Rack

- 35 kW Power
- 42U x 1200 mm x 700 mm (minimum)
- Rear Door Cooler

DEVELOPING THE VEHICLES OF THE FUTURE

Zenuity, a joint venture of Volvo and Veoneer, aims to build autonomous driving software for production vehicles by 2021. They chose to build their deep learning infrastructure with NVIDIA DGX-1 servers and Pure FlashBlade systems to accelerate their AI initiative.



128 NODE DGX-1 CLUSTER SCALES DEEP LEARNING INFRASTRUCTURE

The field of AI holds tremendous promise to improve lives. Facebook A.I. Researchers (FAIR) are advancing the field of machine intelligence by creating new technologies that give people better ways to communicate. To manage the huge variety of projects, datasets, and ever-changing workloads, FAIR needed to update its research cluster. 128 NVIDIA DGXs are the main component of the new cluster and deliver the extreme performance and flexibility FAIR needs to advance AI.



NVIDIA DEEP LEARNING INSTITUTE

Hands-on training for data scientists and software engineers



Training organizations and individuals to solve challenging problems using Deep Learning

On-site workshops and online courses presented by certified experts

Covering complete workflows for proven application use cases. Image classification, object detection, natural language processing, recommendation systems, and more

www.nvidia.com/dli

AI RESOURCES

Special
promo!



DGX Systems

The world's fastest AI Supercomputers

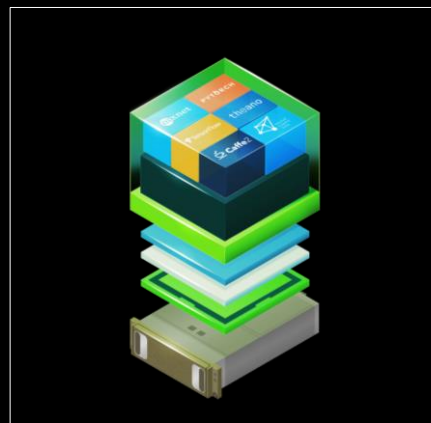
nvidia.com/DGX



DGX Station Webinars

To learn more about how SBB, Imagga, and Deepgram are using DGX Stations for their solutions

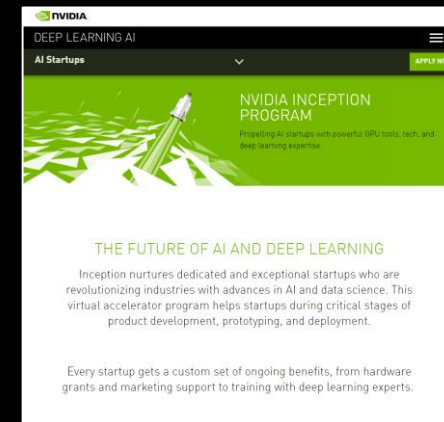
<https://webinars.on24.com/NVIDIA/stationwebinarseries>



NVIDIA GPU Cloud

To learn more about all of the GPU-accelerated software from NGC, visit: nvidia.com/cloud

To sign up or explore NGC: ngc.nvidia.com



Inception

Access to NVIDIA tech GPU and AI experts
Global marketing/sales
GPU venture introduction
nvidia.com/inception

