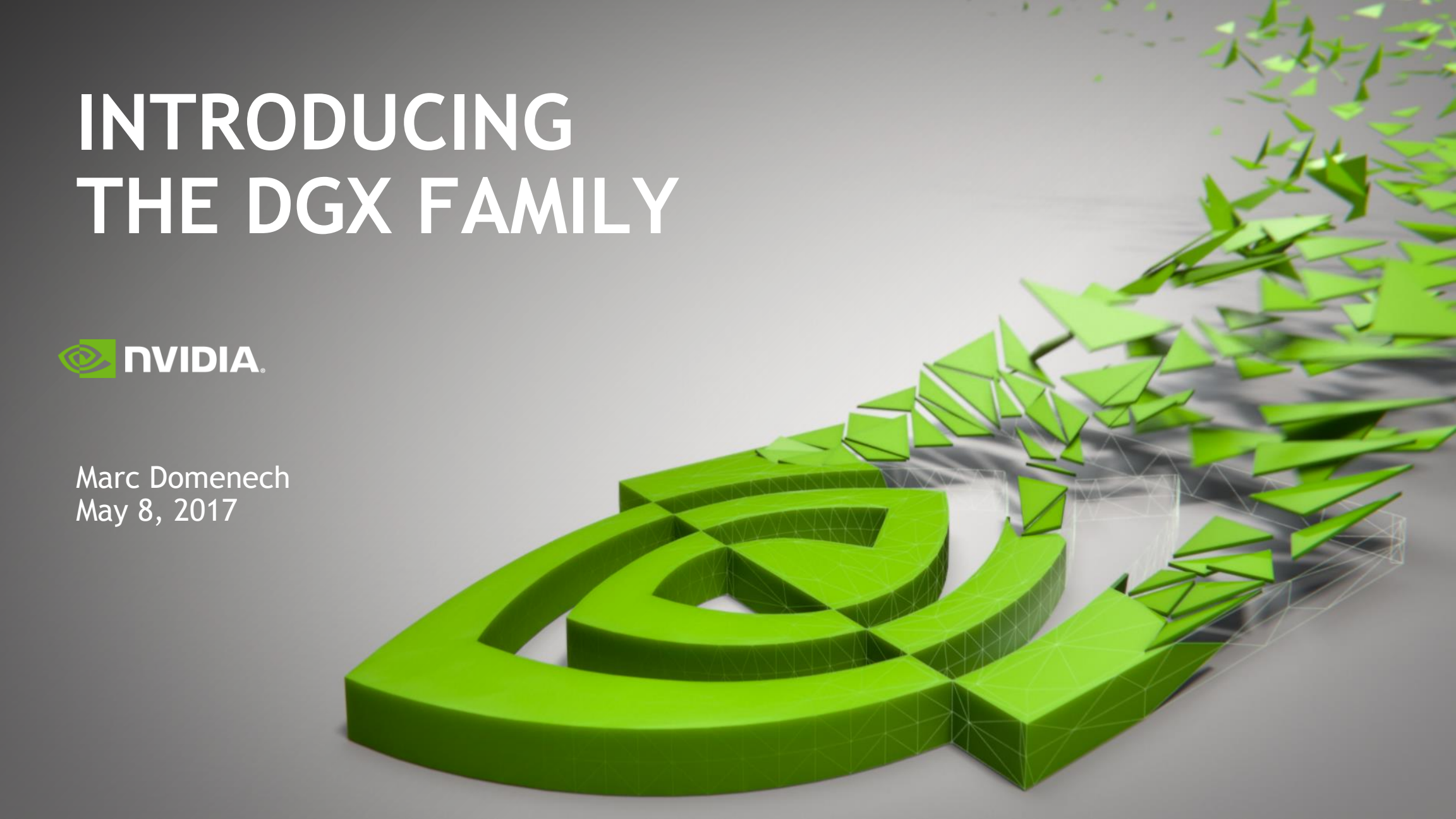
INTRODUCING THE DGX FAMILY



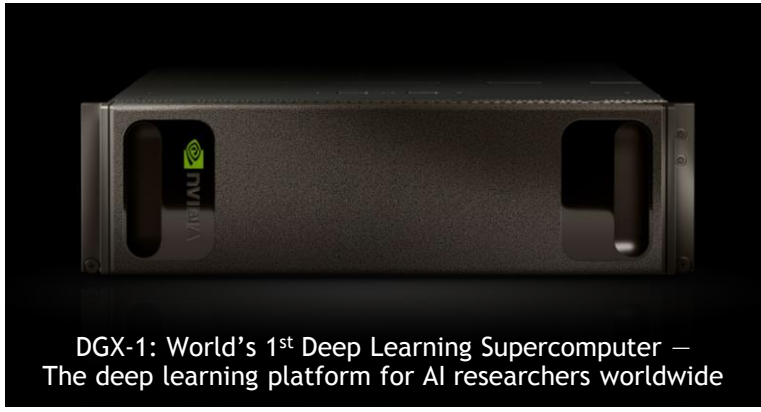
Marc Domenech
May 8, 2017

NVIDIA

Pioneered GPU Computing | Founded 1993 | $7B | 9,500 Employees

100M NVIDIA GeForce Gamers —
The world's largest gaming platform


Pioneering AI computing for
self-driving cars


DGX-1: World's 1st Deep Learning Supercomputer —
The deep learning platform for AI researchers worldwide


GE Revolution —
The GPU choice when it really matters


The visualization platform of every car company
and movie studio


The processor of #1 U.S. supercomputer and
9 of 10 of world's most energy-efficient supercomputers

DATA & ANALYTICS USE CASES

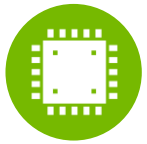
AUTOMOTIVE
Auto sensors reporting location, problems

COMMUNICATIONS
Location-based advertising

CONSUMER PACKAGED GOODS
Sentiment analysis of what's hot, problems

FINANCIAL SERVICES
Risk & portfolio analysis
New products

EDUCATION & RESEARCH
Experiment sensor analysis



HIGH TECHNOLOGY / INDUSTRIAL MFG.
Mfg. quality
Warranty analysis

LIFE SCIENCES
Clinical trials

MEDIA/ENTERTAINMENT
Viewers / advertising effectiveness

ON-LINE SERVICES / SOCIAL MEDIA
People & career matching

HEALTH CARE
Patient sensors, monitoring, EHRs

OIL & GAS
Drilling exploration sensor analysis

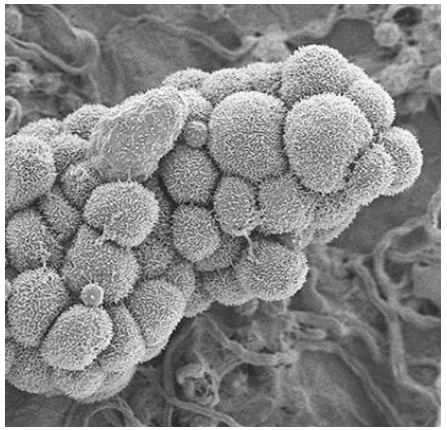
RETAIL
Consumer sentiment

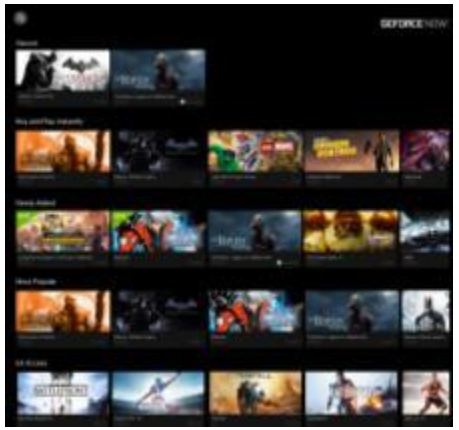
TRAVEL & TRANSPORTATION
Sensor analysis for optimal traffic flows

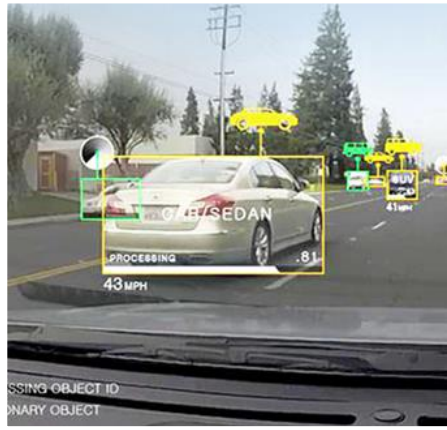
UTILITIES
Smart Meter analysis for network capacity,

LAW ENFORCEMENT & DEFENSE
Threat analysis - social media monitoring, photo analysis

DEEP LEARNING EVERYWHERE







INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation

MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery

MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

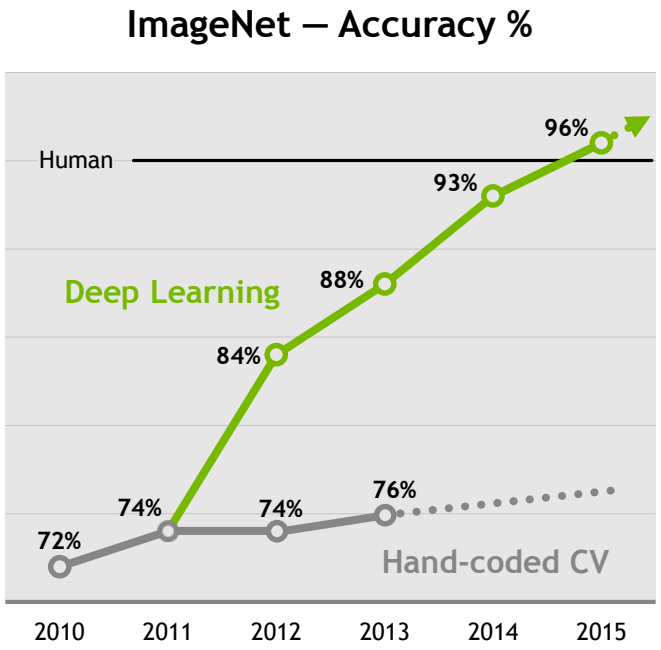
SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

AUTONOMOUS MACHINES

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

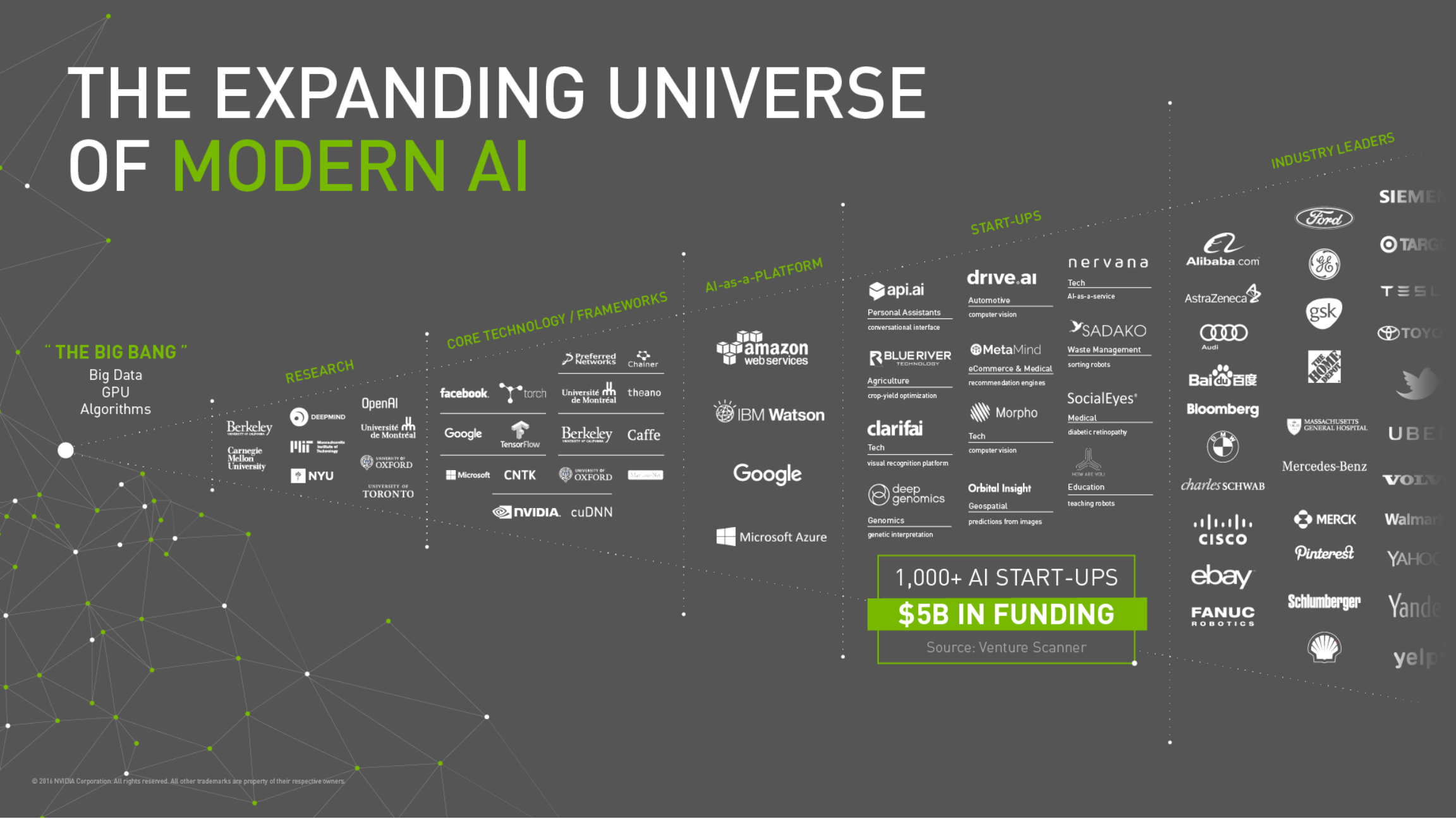
“SUPERHUMAN” RESULTS SPARK HYPERSCALE ADOPTION



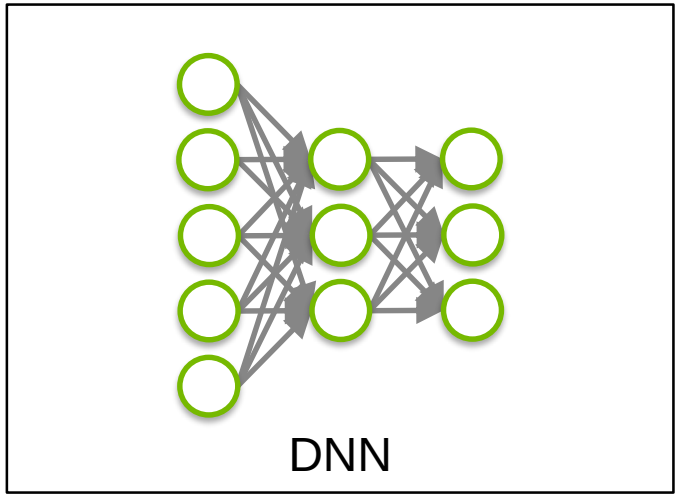
Alibaba/Aliyun
Amazon
Baidu
eBay
Facebook
Flickr
Google
iFLYTEK
iQIYI
JD.com
Orange
Periscope
Pinterest
Qihoo 360
Shazam
Skype
Sogou
Twitter
Yahoo Supermarket
Yandex
Yelp

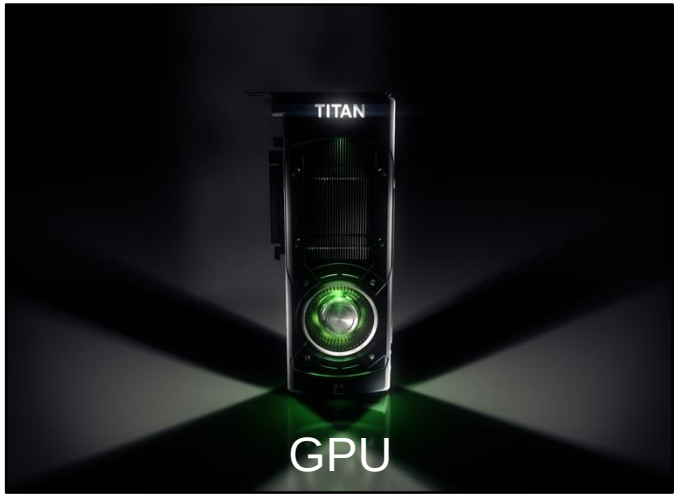
Cloud Services with AI Powered by NVIDIA

THE EXPANDING UNIVERSE OF MODERN AI



THE BIG BANG IN MACHINE LEARNING

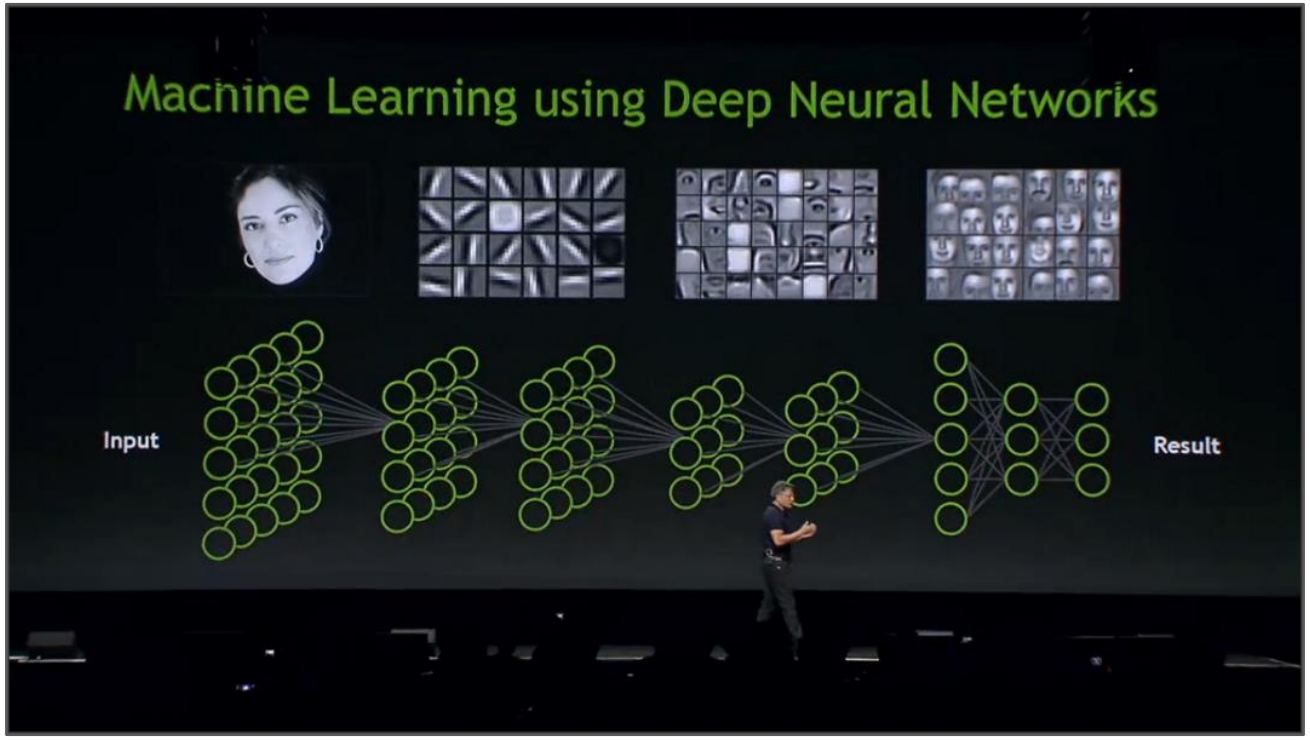




“ Google’s AI engine also reflects how the world of computer hardware is changing. (It) depends on machines equipped with GPUs… And it depends on these chips more than the larger tech universe realizes.”

WIRED

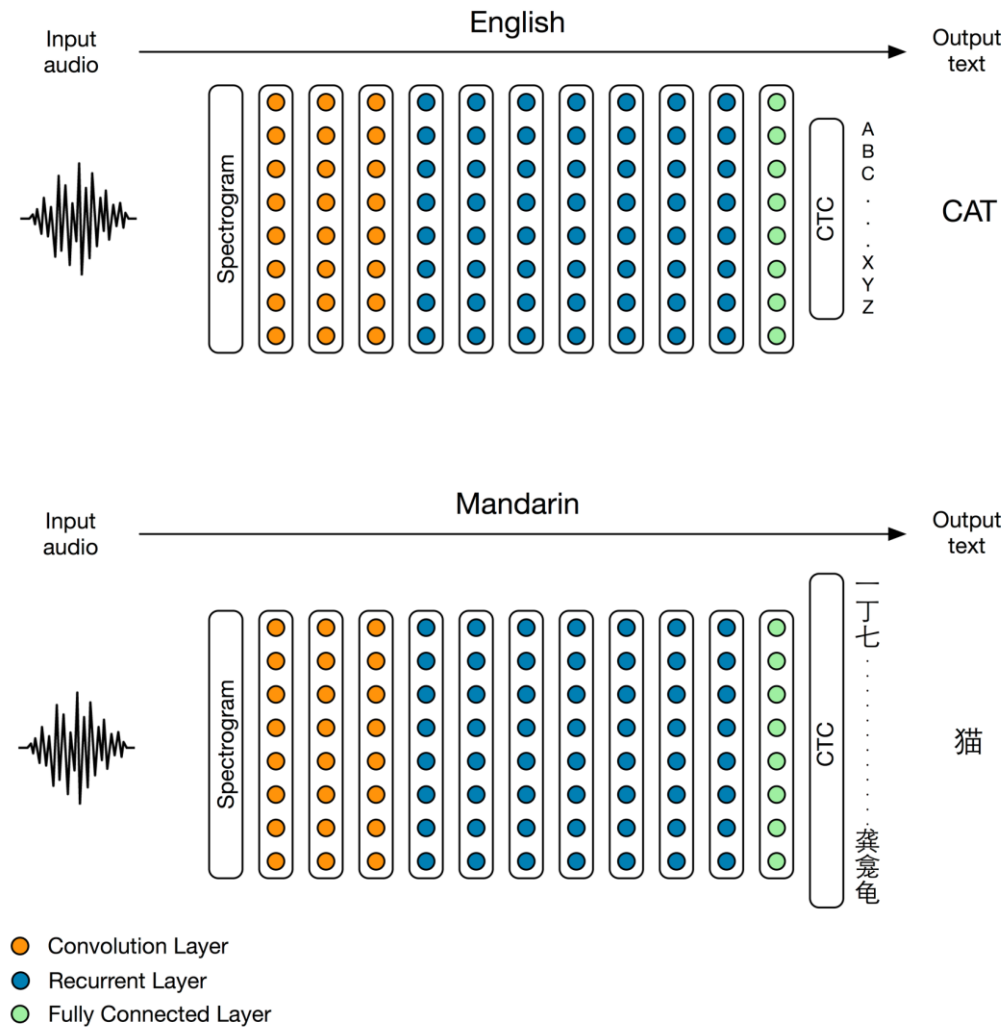
GPU DEEP LEARNING IGNITES AI



For China's 1.25 billion mobile users, web experience can be slow and frustrating with a keyboard because there are thousands of Chinese characters.

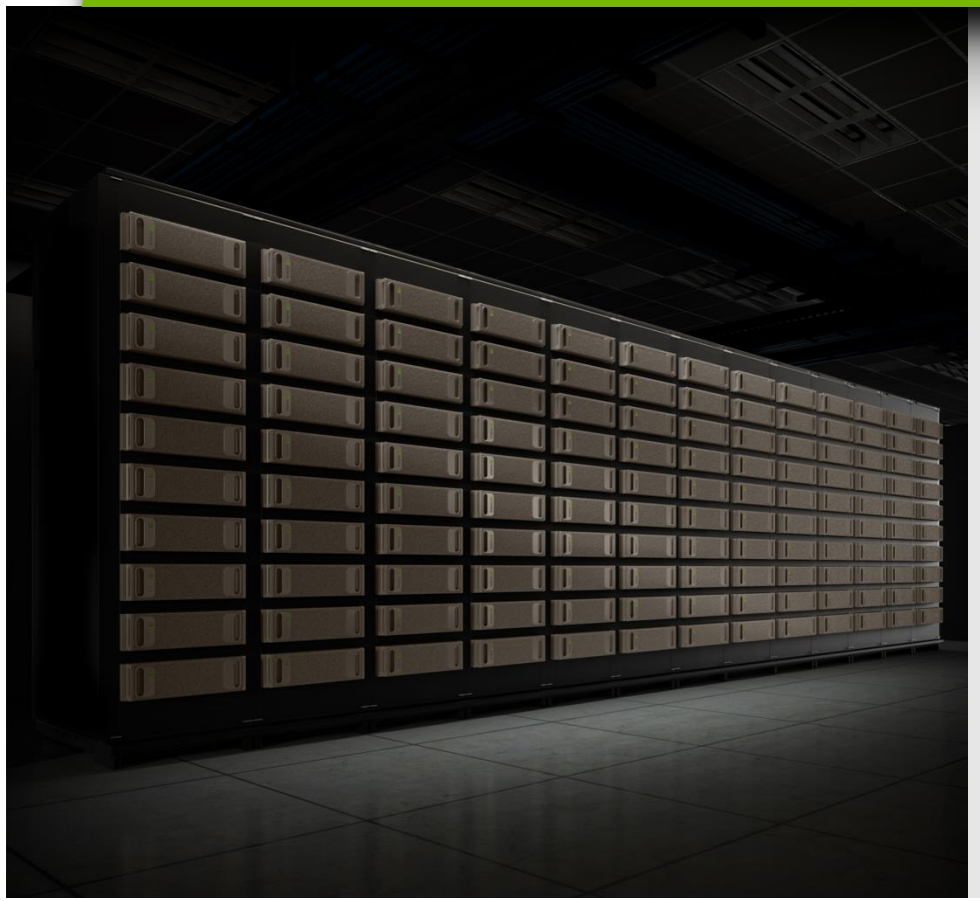
Baidu, China's largest search engine company, developed the world's most advanced speech recognition system, powered by deep learning. Deep Speech 2 is the world's first model to recognize both English and Mandarin while delivering super-human accuracy.

Baidu has deployed NVIDIA GPUs in production to power AI services like Deep Speech 2. GPUs deliver responsiveness that would not be possible on CPU servers.



NVIDIA DGX SATURNV

Giant Leap Towards Exascale AI



Fastest AI Supercomputer in TOP500

4.9 Petaflops Peak FP64
19.6 Petaflops Peak FP16
13 DGX-1 to get into Top500

Most Energy Efficient Supercomputer

#1 Green500
9.5 GFLOPS per Watt

Rocket for Cancer Moonshot

CANDLE Development Platform
Common platform with DOE labs – ANL, LLNL, ORNL, LANL

INTRODUCING THE DGX FAMILY

AI WORKSTATION

DGX Station

The Personal
AI Supercomputer

AI DATA CENTER

DGX-1

with

Tesla P100

The World's First
AI Supercomputer
in a Box

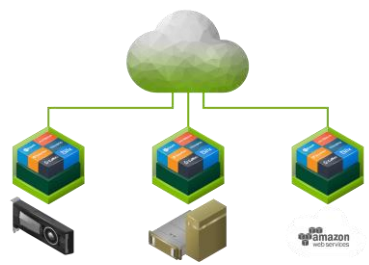
with

Tesla V100

The Essential
Instrument for AI
Research

CLOUD-SCALE AI

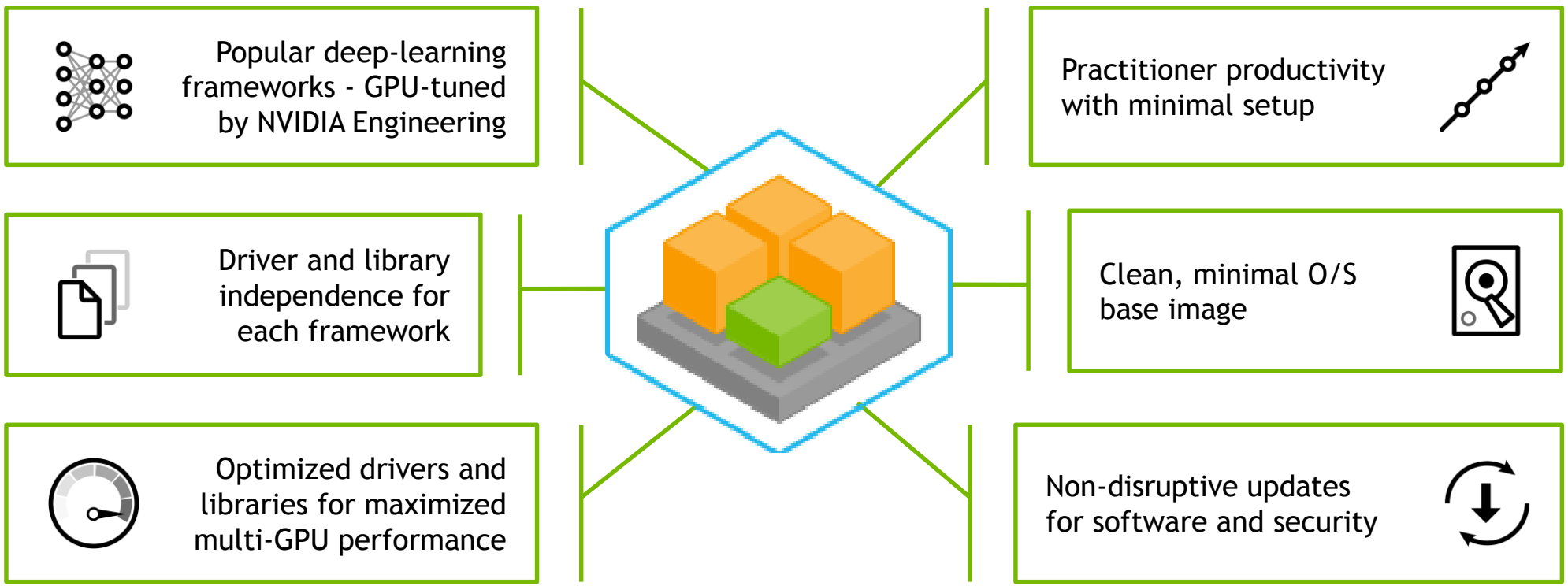
NVIDIA GPU Cloud



Cloud service with the highest
deep learning efficiency

ENTERPRISE BENEFITS OF DGX SOFTWARE

NVIDIA Investments in Deep Learning Performance and Manageability



10 STEPS TO SETUP A DIY SYSTEM

380 PAGES OF DOCS TO READ

Step 1. Install Ubuntu linux (10 pg)
Step 2. Install CUDA (41 pg)
Step 3. Install CUDNN (154 pg)
Step 4. Install and Upgrade PIP (20 pg)
Step 5. Install BAZEL (build TF source) (50 pg)
Step 6. Install TensorFlow (15 pg)
Step 7. Upgrade Protobuf (15 pg)
Step 8. Install Docker (75 pg)
Step 9. Test the installation
Step 10. Debug and fix install

NVIDIA DGX SYSTEMS

Deep Learning is a massive opportunity

Data Scientist's productivity is vital

NVIDIA is the choice of the deep learning world

DGX-1 is the fastest system for deep learning

For More Information: **nvidia.com/dgx-1**

NVIDIA DGX-1

AI Supercomputer-in-a-Box

170 TFLOPS | 8x Tesla P100 16GB | NVLink Hybrid Cube Mesh

2x Xeon | 8 TB RAID 0 | Quad IB 100Gbps, Dual 10GbE | 3U — 3200W

NVIDIA DGX-1

250 NODE HPC SUPERCOMPUTER-IN-A-BOX	
# Servers	250
Cost per server	$9,000
IB cost per node	$1,000
Total value	$2.5M
and more…	100X less power, plug-and-play…

NVIDIA DGX unlocks the full potential of NVIDIA GPU's – powered by software innovation

REVOLUTIONARY AI PERFORMANCE

3X system performance over prior generation

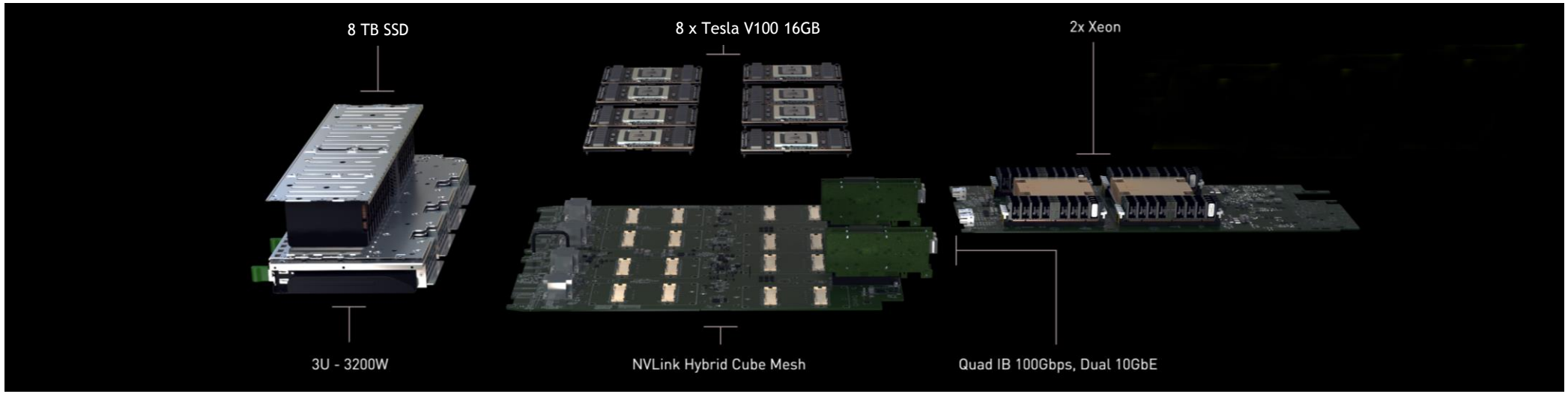
Software stack delivers additional 30% faster training performance vs other GPU systems

10X I/O performance with 2nd generation NVLink vs PCIe-connected GPU's

New Tensor Core architecture inspired by the demands of deep learning

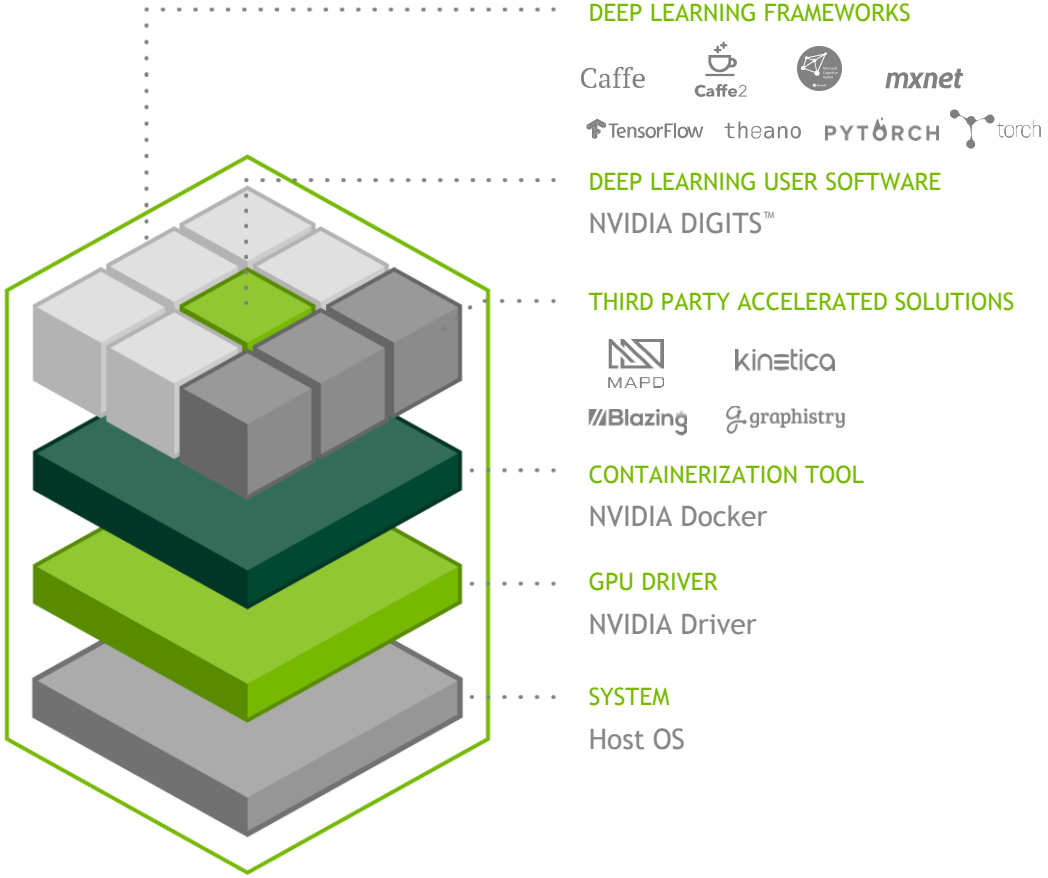
OUR STRATEGY IN THE DATACENTER: NVIDIA DGX-1

Highest Performance, Fully Integrated HW System



960 TFLOPS | 8x Tesla V100 16GB | 300 GB/s NVLink Hybrid Cube Mesh

2x Xeon | 8 TB RAID 0 | Quad IB 100Gbps, Dual 10GbE | 3U — 3200W



NVIDIA DGX-1 SOFTWARE STACK

Fully Integrated Software for Instant Productivity

Advantages:

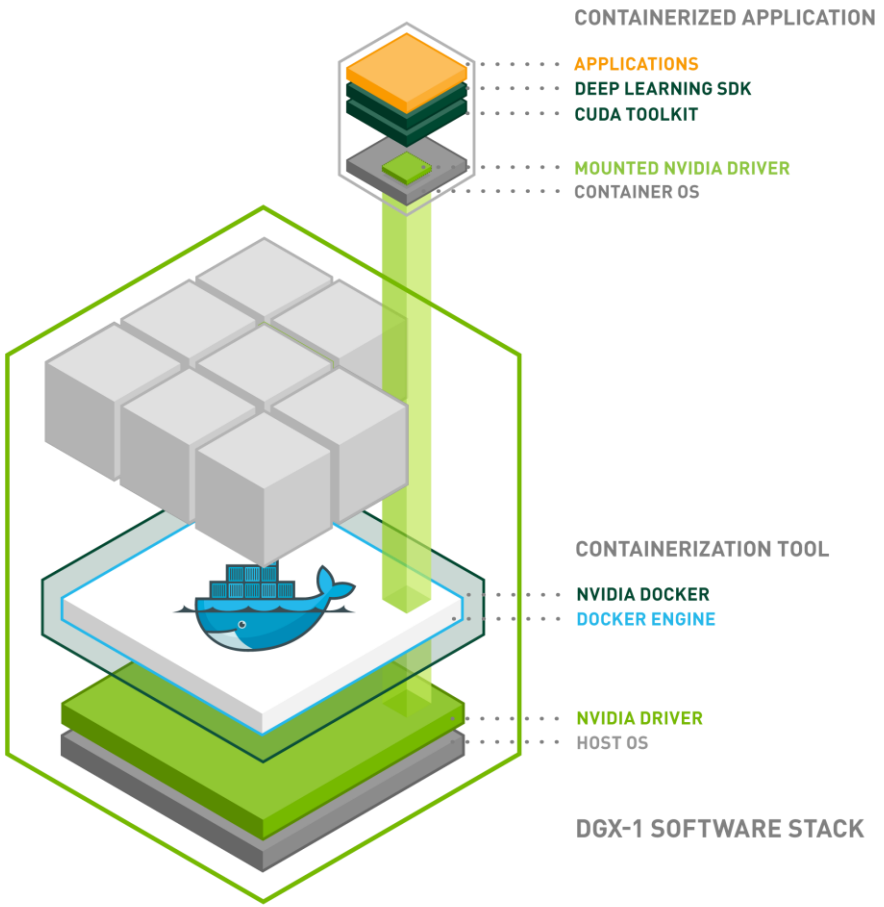
Instant productivity with NVIDIA optimized deep learning frameworks

- Caffe, CNTK, MXNet, PyTorch, TensorFlow, Theano, and Torch

Performance optimized across the entire stack

Faster Time-to-Insight with pre-built, tested, and ready to run framework containers

Flexibility to use different versions of libraries like libc, cuDNN in each framework container



SIMPLIFY PORTABILITY WITH NVIDIA DOCKER CONTAINERS

Benefits of Containers:

Simplify deployment of GPU-accelerated applications

Isolate individual frameworks or applications

Share, collaborate, and test applications across different environments

THE POWER TO RUN MULTIPLE FRAMEWORKS AT ONCE

Container Images portable across new driver versions

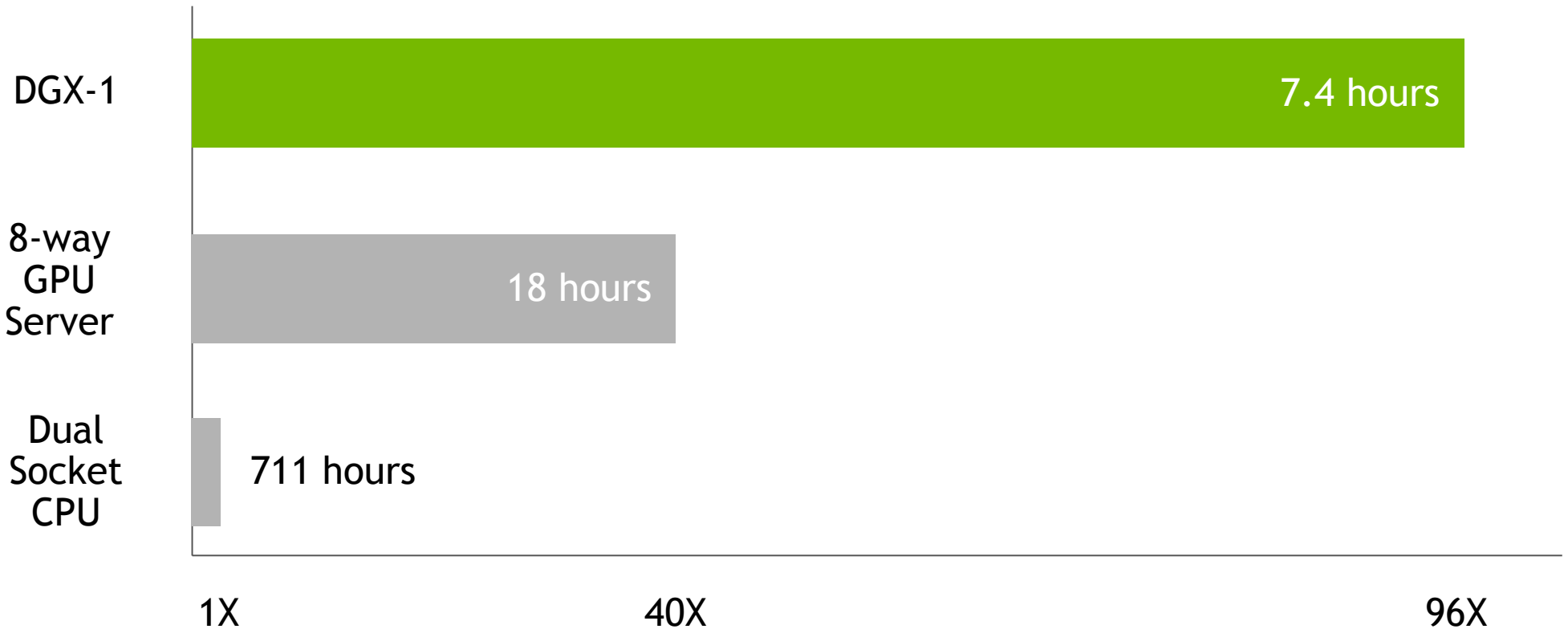
Containerized Applications

NVIDIA Docker	NVIDIA Docker	NVIDIA Docker	NVIDIA Docker		NVIDIA Docker
TensorFlow	Microsoft Cognitive Toolkit	Caffe2	PYTORCH	• • •	Other Frameworks and Apps
TF Tuned SW	CNTK Tuned SW	Caffe2 Tuned SW	Pytorch Tuned SW		Tuned SW
CUDA RT	CUDA RT	CUDA RT	CUDA RT		CUDA RT

Linux Kernel + CUDA Driver

NVIDIA ® DGX-1™

DGX-1: 96X FASTER THAN CPU



Workload: ResNet50, 90 epochs to solution | CPU Server: Dual Xeon E5-2699 v4, 2.6GHz

NVIDIA DGX-1 CUSTOMER MOMENTUM

Major Worldwide Branded Wins

RIKEN SUCCESS STORY

Fujitsu and NVIDIA Build AI Supercomputer With 24 DGX-1s

CHALLENGE

Enterprises and research organizations embracing AI/DL

Needed to accelerated research in areas including medicine, manufacturing and healthcare

Conventional HPC architectures too costly and inefficient

SOLUTION

Partnered with Fujitsu for scale-out AI architecture built on DGX-1

24 DGX-1's deliver 4 petaflops powering the RIKEN supercomputer

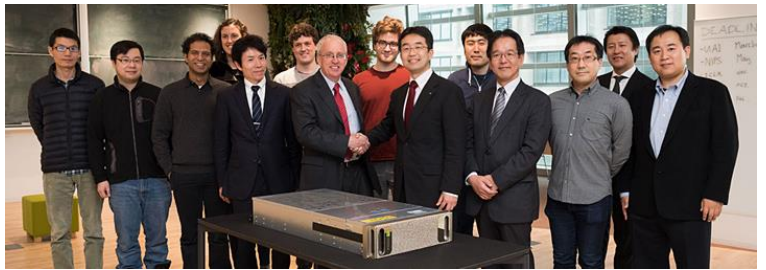
NVIDIA COSMOS streamlines AI researcher workflow, helping accelerate RIKEN productivity

IMPACT

Accelerated real-world implementation of scale-out AI

Enables RIKEN team to take advantage of next-gen DL algorithms

Helping create future in which AI finds solutions to societal issues



MASS GENERAL SUCCESS STORY

Man, Machine & Medicine: AI-Powered Research at MGH

CHALLENGE

Clinical data science center needed to apply ML to medicine

Data volume requires immense computational capacity to process

Immediate applications include radiology to improve accuracy, reduce variation

SOLUTION

1st medical institute in the world to leverage the DGX-1

Center for Clinical Data Sciences expands to partner hosp. (3X data)

Deployment has grown to scale-out architecture with 4 DGX-1's

IMPACT

New prostrate cancer pathology developed on DGX in 6 months

AI/DL becomes critical tool in physician's toolkit in 5-10 years

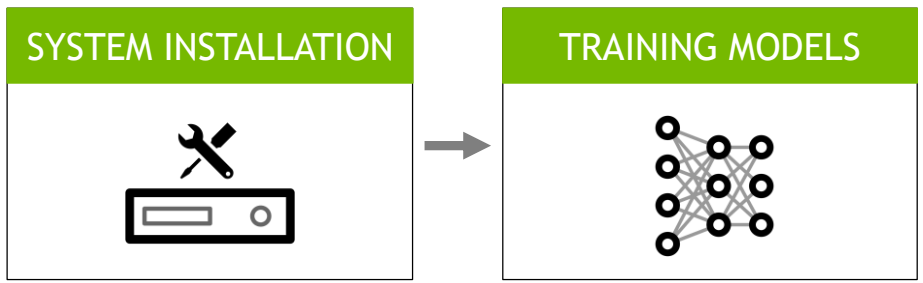
Advancements in diagnostics, genomics, genetics, imaging

BENEVOLENTAI: TRAINING REDUCED TO DAYS

Technology Review Article on DGX-1:

The Pint-Sized Supercomputer That Companies Are Scrambling to Get

https://www.technologyreview.com/s/603075/the-pint-sized-supercomputer-that-companies-are-scrambling-to-get/



Same Day

$ “The cost of renting enough servers on Amazon Web Services would surpass the system’s $129,000 price tag within a year.”

-Jackie Hunter, CEO, BenevolentAI

NVIDIA DGX-1

Days

Vs.

Other GPU System

Weeks of Training

DGX-1

3x-4x

FASTER TRAINING

NVIDIA DGX-1

The Essential Instrument of AI Research

Deep Learning is a massive opportunity

Data Scientist's productivity is vital

NVIDIA is the choice of the deep learning world

DGX-1 is the fastest system for deep learning

For More Information: **nvidia.com/dgx-1**

INTRODUCING NVIDIA DGX STATION

The Personal AI 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

DESIGNED FOR THE DESK

The Only Supercomputer Designed for Your Office

The power of 400 CPU's – at your fingertips

Consuming only 1500W, it draws only 1/20th the power

Emitting only 1/10th the noise of other workstations

EFFORTLESS PRODUCTIVITY

Productivity That Follows You From Desk to Data Center to Cloud

Caffe Caffe2 TensorFlow mxnet theano torch PYTORCH

Access popular deep learning frameworks, NVIDIA-optimized for maximum performance

DGX containers enable easier experimentation and keep base OS clean

Develop on DGX Station, scale on DGX-1 or the NVIDIA Cloud

3X FASTER THAN THE FASTEST WORKSTATIONS

Supercomputing performance at your desk

480 TFLOPS

Water-cooled performance – the only workstation built on 4 Tesla V100's

3X

3X the performance of today's fastest GPU workstations

30%

with 30% faster training over non-DGX stack solutions

5X

5X increase in I/O performance with 4-way next generation NVLink vs. PCIe-connected GPU's

DGX STATION ARCHITECTURE

The world's fastest GPU workstation with the equivalent compute capacity of 400 CPU's, consuming only 1/20th the power

NVIDIA Tesla V100

Next generation NVIDIA NVLink™ high-speed interconnect

Water-cooling system for whisper-quiet operation, and maximized performance

Intel Xeon CPU

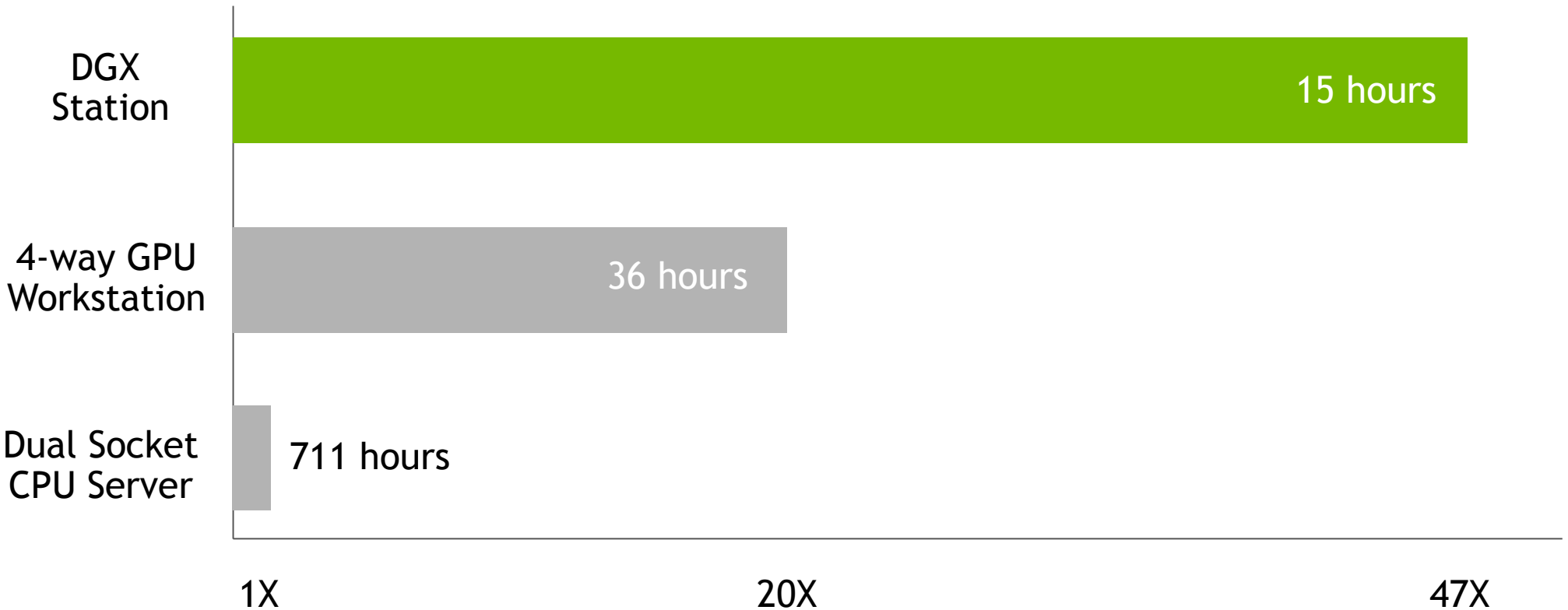
3 DisplayPort x 4K resolution

NVIDIA DGX STATION SPECIFICATIONS

At a Glance

GPUs	4x NVIDIA® Tesla® V100
TFLOPS (GPU FP16)	480
GPU Memory	16 GB per GPU
NVIDIA Tensor Cores	2,560 (total)
NVIDIA CUDA Cores	20,480 (total)
CPU	Intel Xeon E5-2698 v4 2.2 GHz (20-core)
System Memory	256 GB LRDIMM DDR4
Storage	Data: 3 x 1.92 TB SSD RAID 0 OS: 1 x 1.92 TB SSD
Network	Dual 10 Gb LAN
Display	3x DisplayPort, 4K Resolution
Acoustics	< 35 dB
Maximum Power Requirements	1500 W
Operating Temperature Range	10 - 30 °C
Software	Ubuntu Desktop Linux OS DGX Recommended GPU Driver CUDA Toolkit

DGX STATION: 47X FASTER THAN CPU



WHAT HAVE USERS BEEN SAYING?

NVIDIA Internal Researchers are impressed

"**I felt I won the software stack lottery** as nvidia-docker was already installed. I immediately pulled a container and started work on a CNTK NCCL project, the next day pulled another container to work on a TF biomedical project. I haven't looked back at how to reimage because felt too productive."

"**DGX Station runs extremely quiet.** I can barely hear it running from under the desk. This is a plus point for a workstation that's meant to be positioned in an office environment."

"For the numbers, it's taking about **1-2 hrs to train a 152 layer ResNet on a ~20GB dataset**, which is pretty good and keeping me active with experiments rolling, just on the workstation. It feels right for this work to allow fast iteration. The last time I did some serious model architecture/tuning work it took halfdays to days on Kepler GPUs."

DGX STATION:

The Personal AI Supercomputer

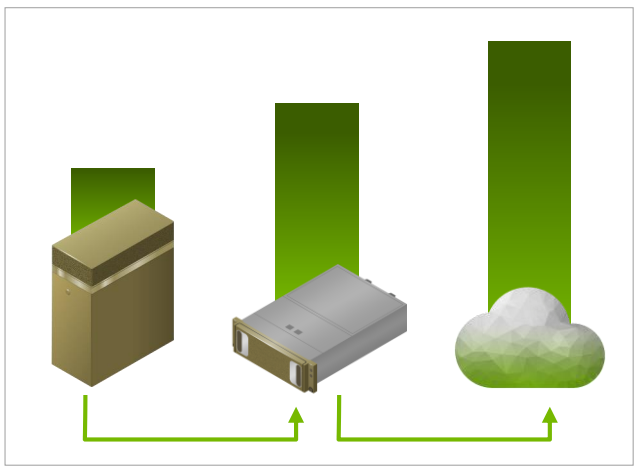
VOLTA-POWERED PERFORMANCE

4 racks of x86 servers –
in a workstation

DESIGNED FOR THE OFFICE

Desk-friendly
Whisper-quiet

EFFORTLESS PRODUCTIVITY



Experiment on Station
Scale on DGX-1 / Cloud

NVIDIA DGX SYSTEMS

The Personal AI Supercomputer

Introducing NVIDIA DGX Station

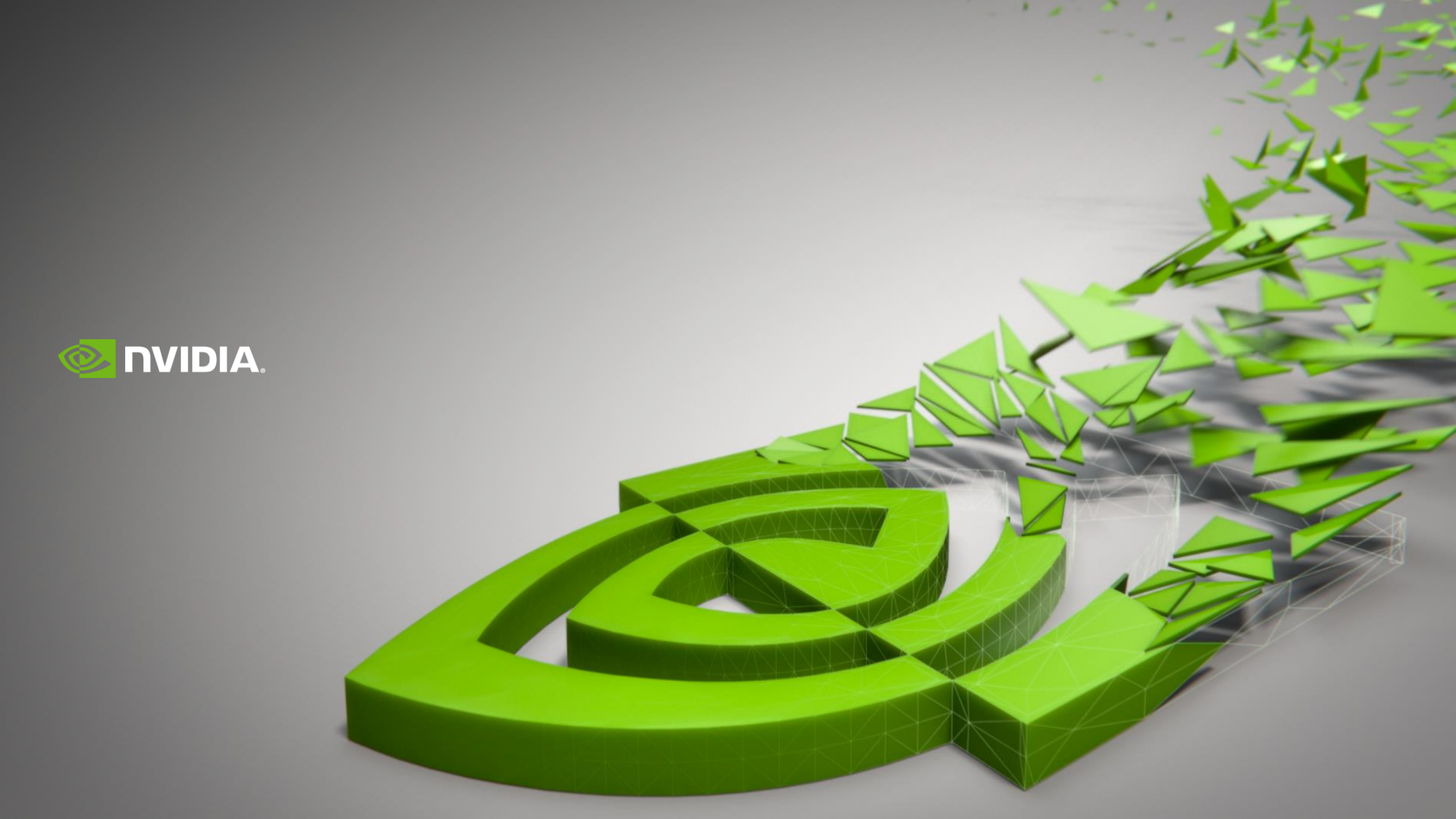
The Only Supercomputer Designed for Your Office

Get the Fastest Start in Deep Learning

Productivity That Follows You from Desk to Data Center

3X Faster than the Fastest Workstations

For More Information: nvidia.com/dgx-station


NVIDIA