THE AI COMPUTING COMPANY

Yaniv Benami
Sr. Manager Enterprise Sales

NVIDIA

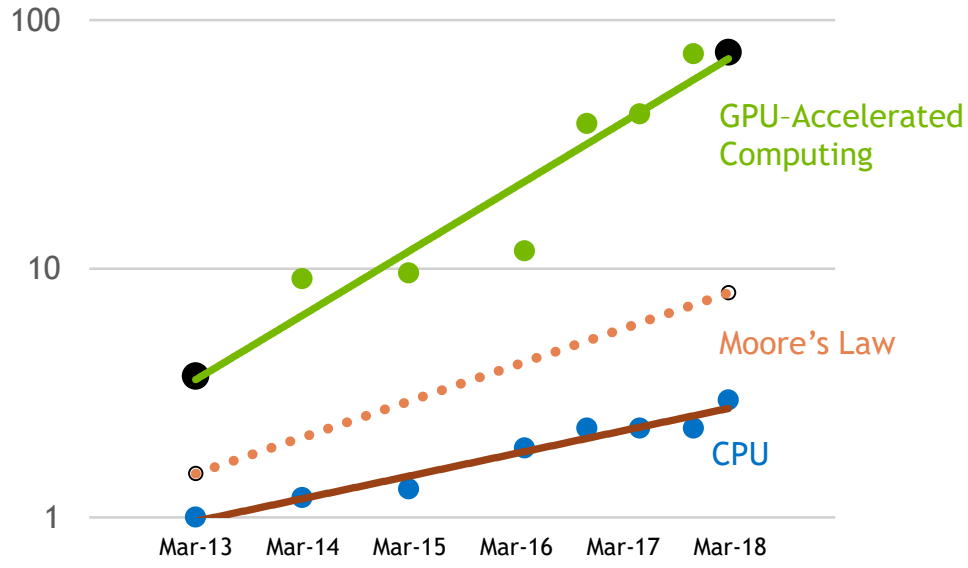
April 2018

BEYOND MOORE'S LAW

2013

- cuBLAS: 5.0
- cuFFT: 5.0
- cuRAND: 5.0
- cuSPARSE: 5.0
- NPP: 5.0
- Thrust: 1.5.3
- CUDA: 5.0
- Resource Mgr: r304
- Base OS: CentOS 6.2

Accelerated Server with M2090



Measured performance of Amber, CHROMA, GTC, LAMMPS, MILC, NAMD, Quantum Espresso, SPECFEM3D

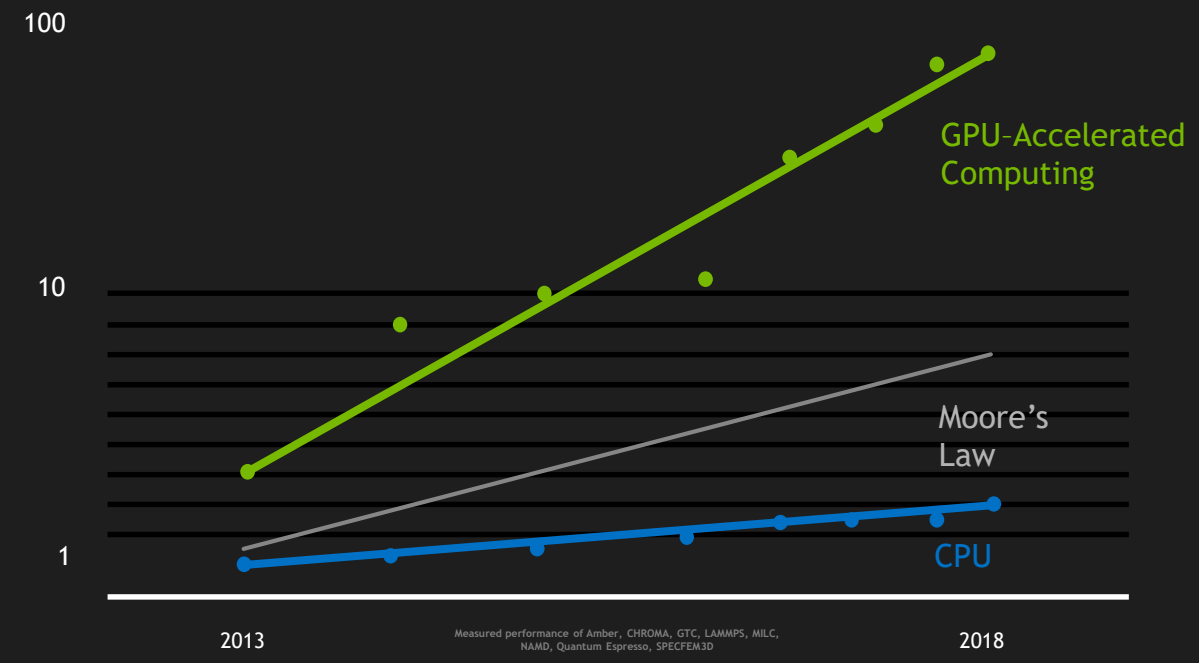
2018

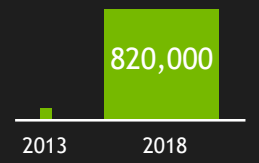
- cuBLAS: 9.0
- cuFFT: 9.0
- cuRAND: 9.0
- cuSOLVER: 9.0
- cuSPARSE: 9.0
- NPP: 9.0
- CUDA: 9.0
- Resource Mgr: r384
- Base OS: Ubuntu 16.04

Accelerated Server with V100

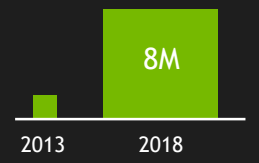
"NVIDIA Is So Far Ahead of the Curve"

—The Inquirer

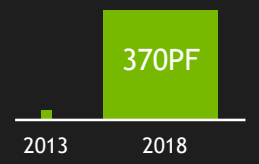




GPU Developers
10X in 5 Yrs



CUDA Downloads
5X in 5 Yrs



Total GPU FLOPS
of Top 50 Systems
15X in 5 Yrs

For 30 years, the dynamics of Moore's law held true. But CPU performance scaling has slowed. GPU computing is defining a new, supercharged law. It starts with a highly specialized parallel processor called the GPU and continues through system design, system software, algorithms, and all the way through optimized applications. The world is jumping on board.

NVIDIA DEVELOPER

Log in

NVIDIA SDK

The Essential Resource for GPU Developers

NVIDIA SDK

https://developer.nvidia.com/

DEEP LEARNING
Deep Learning SDK
High-performance tools and libraries for deep learning

SELF-DRIVING CARS
NVIDIA DriveWorks™
Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

VIRTUAL REALITY
NVIDIA VRWorks™
A comprehensive SDK for VR headsets, games and professional applications

GAME DEVELOPMENT
NVIDIA GameWorks™
Advanced simulation and rendering technology for game development

ACCELERATED COMPUTING
NVIDIA ComputeWorks™
Everything scientists and engineers need to build GPU-accelerated applications

DESIGN & VISUALIZATION
NVIDIA DesignWorks™
Tools and technologies to create professional graphics and advanced rendering applications

AUTONOMOUS MACHINES
NVIDIA JetPack™
Powering breakthroughs in autonomous machines, robotics and embedded computing

ADDITIONAL RESOURCES
More resources for GPU Developers

CUDA TOOLKIT - DOWNLOAD TODAY!

Everything You Need to Accelerate Applications

CUDA DOCUMENTATION

- Installation Guide
- Best Practices Guide
- Programming Guide
- CUDA Tools Guide
- API Reference
- Samples

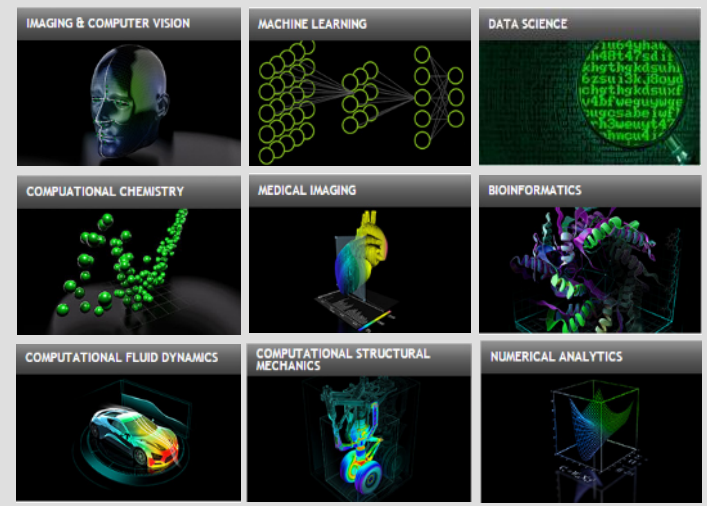
GETTING STARTED RESOURCES

GTC WEBINARS

UDACITY

qwik LABS

INDUSTRY APPLICATIONS



developer.nvidia.com/cuda-toolkit

CUDA EVERYWHERE

All Top 15 HPC Apps Accelerated

GROMACS
ANSYS Fluent
Gaussian
VASP
NAMD
Simula Abaqus
WRF
OpenFOAM
ANSYS
LS-DYNA
BLAST
LAMMPS
AMBER
Quantum Espresso
GAMESS

Caffe2

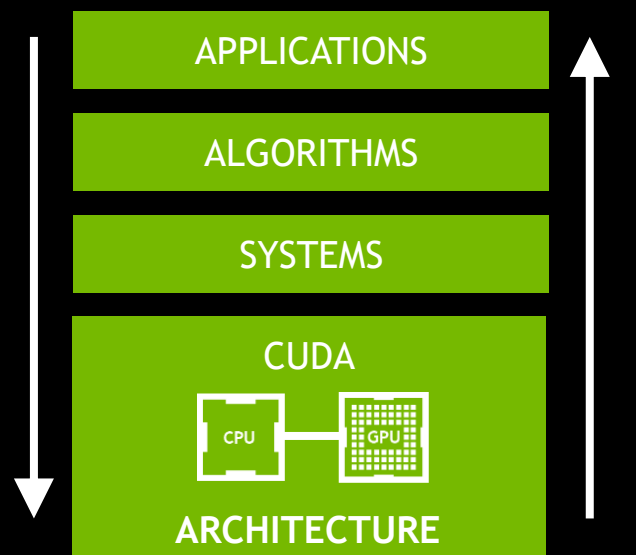
Microsoft Cognitive Toolkit

mxnet

PYTORCH

TensorFlow

theano



500+ GPU-ACCELERATED Apps & Databases

EVERY DEEP LEARNING FRAMEWORK ACCELERATED

INVESTING FROM TOP TO BOTTOM

GPU-ACCELERATED HPC APPLICATIONS

500+ APPLICATIONS

LIFE SCIENCES

50+ app

Including:
- Gaussian
- VASP
- AMBER
- HOOMD-Blue
- GAMESS

MFG, CAD, & CAE

111 apps

Including:
- Ansys Fluent
- Abaqus SIMULIA
- AutoCAD
- CST Studio Suite

PHYSICS

20 apps

Including:
- QUDA
- MILC
- GTC-P

OIL & GAS

17 apps

Including:
- RTM
- SPECFEM 3D

CLIMATE & WEATHER

4 apps

Including:
- Cosmos
- Gales
- WRF

DEEP LEARNING

32 apps

Including:
- Caffe2
- MXNet
- Tensorflow

MEDIA & ENT.

142 apps

Including:
- DaVinci Resolve
- Premiere Pro CC
- Redshift Renderer

FEDERAL & DEFENSE

13 apps

Including:
- ArcGIS Pro
- EVNI
- SocetGXP

DATA SCI. & ANALYTICS

23 apps

Including:
- MapD
- Kinetica
- Graphistry

SAFETY & SECURITY

15 apps

Including:
- Cyllance
- FaceControl
- Syndex Pro

COMP. FINANCE

16 apps

Including:
- O-Quant Options Pricing
- MUREX
- MISYS

TOOLS & MGMT.

15 apps

Including:
- Bright Cluster Manager
- HPCtoolkit
- Vampir

https://www.nvidia.com/en-us/data-center/gpu-accelerated-applications/catalog/

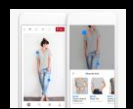
TESLA STACK

World's Leading Data Center Platform for Accelerating HPC and AI

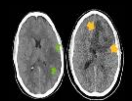
CUSTOMER USECASES


Speech

Translate


Recommender

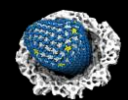
CONSUMER INTERNET

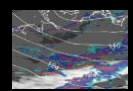

Healthcare

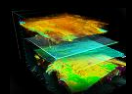

Manufacturing


Engineering

ENTERPRISE APPLICATIONS


Molecular Simulations


Weather Forecasting


Seismic Mapping

SUPERCOMPUTING

INDUSTRY FRAMEWORKS & APPLICATIONS

Caffe2 · Chainer · Microsoft Cognitive Toolkit · mxnet · PaddlePaddle · PYTORCH · TensorFlow

Amber · ANSYS · CHROMA · GROMACS FAST. FLEXIBLE. FREE. · LAMMPS · NAMD · SIMULIA · VASP · +550 Applications

NVIDIA SDK & LIBRARIES

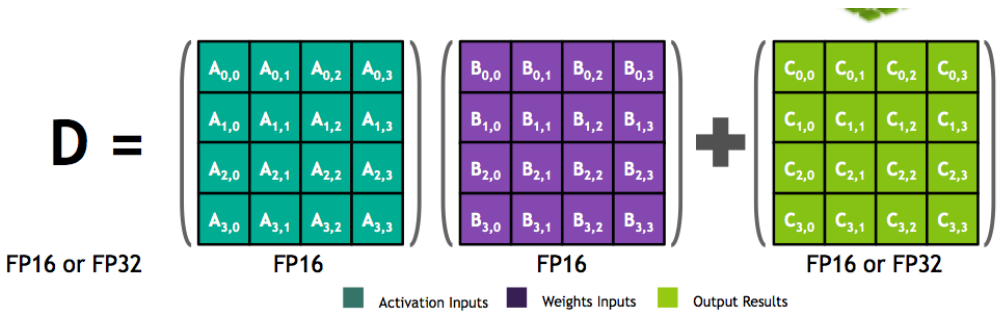
cuBLAS · cuDNN · cuFFT · cuRAND · cuSPARSE · DeepStream · NCCL · TensorRT · PGI OpenACC Directives for Accelerators

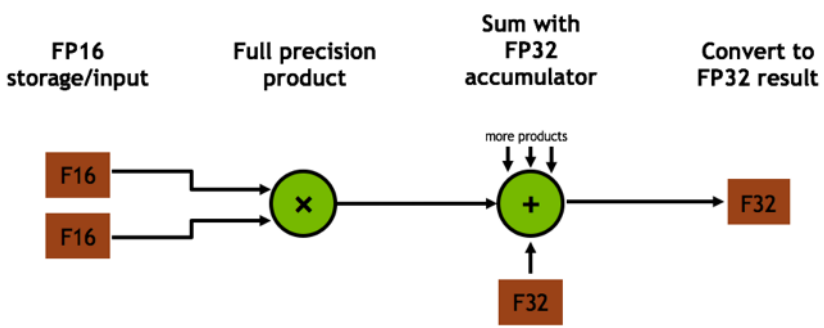
CUDA

TESLA GPUs & SYSTEMS

TENSOR CORE

Mixed Precision Matrix Math - 4x4 matrices



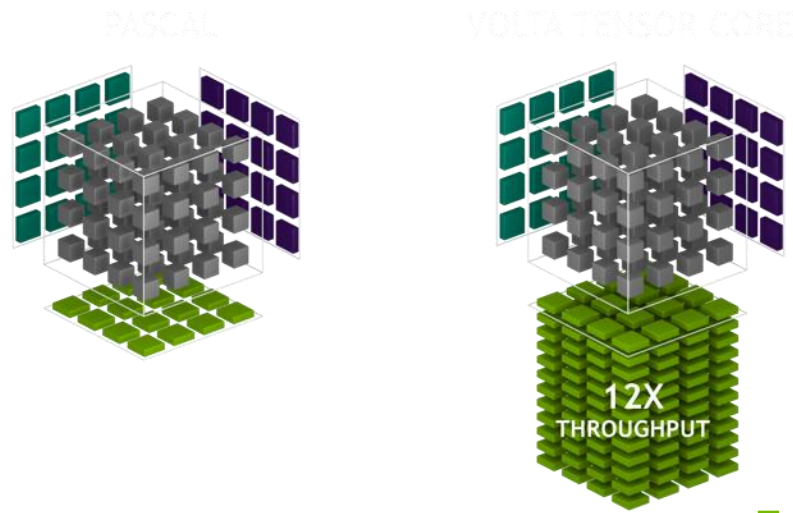


Also supports FP16 accumulator mode for inferencing

New CUDA TensorOp instructions & data formats

4x4x4 matrix processing array

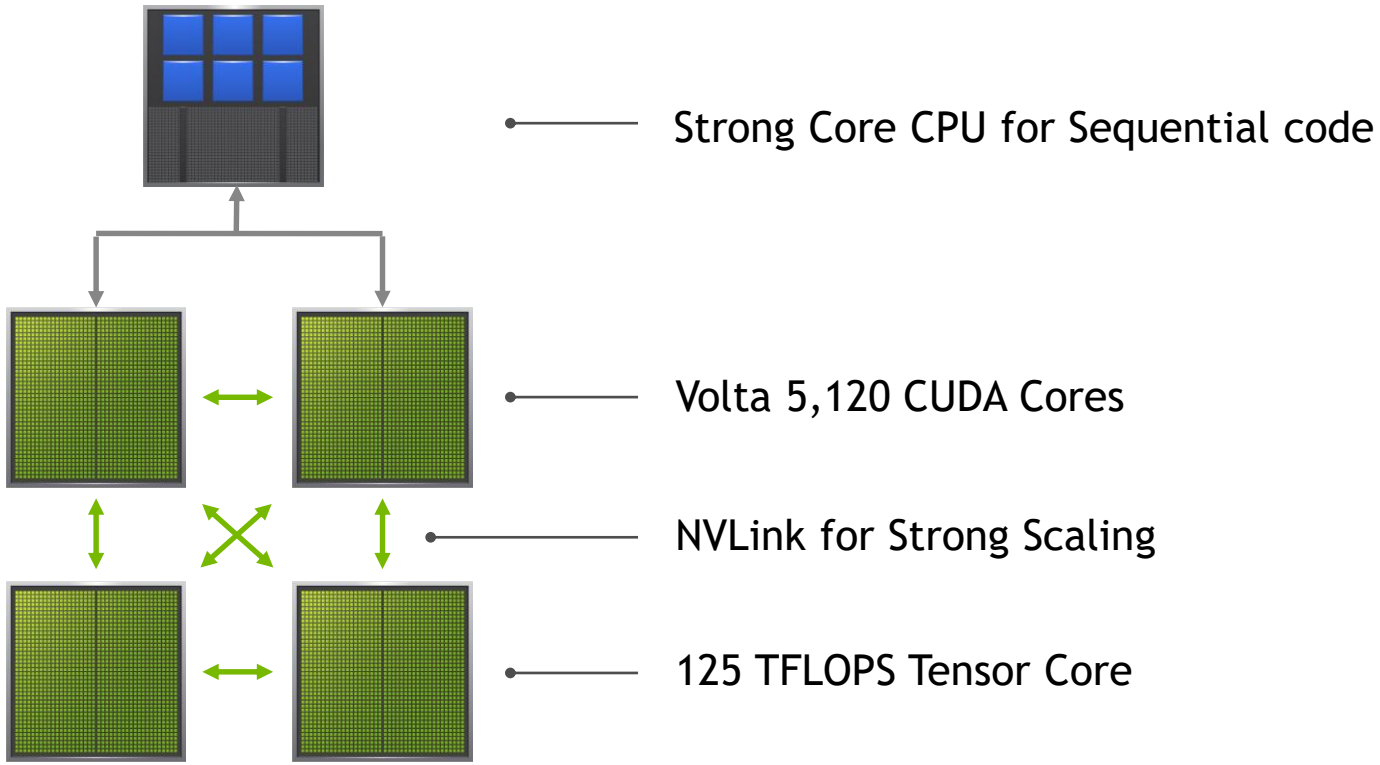
D[FP32] = A[FP16] * B[FP16] + C[FP32]



GPU PERFORMANCE COMPARISON

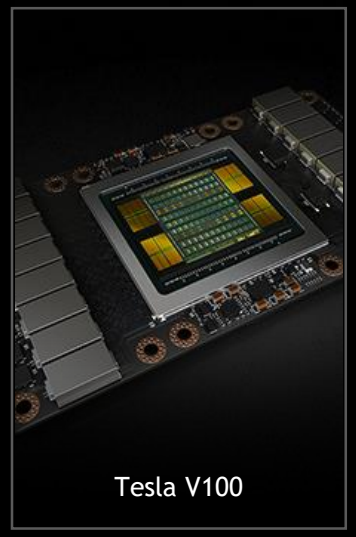
	P100	V100	Ratio
Training acceleration	10 TOPS	125 TOPS	12.5x
Inference acceleration	21 TFLOPS	125 TOPS	6x
FP64/FP32	5/10 TFLOPS	7.8/15.7 TFLOPS	1.5x
HBM2 Bandwidth	720 GB/s	900 GB/s	1.2x
NVLink Bandwidth	160 GB/s	300 GB/s	1.9x
L2 Cache	4 MB	6 MB	1.5x
L1 Caches	1.3 MB	10 MB	7.7x

ARCHITECTING MODERN DATACENTERS



"Creating Powerful System-level Solutions Will Give It an Edge Against Rivals Who Have Merely Developed a Good Chip"

—TheStreet



Tesla V100

NEW 32GB

DGX Systems

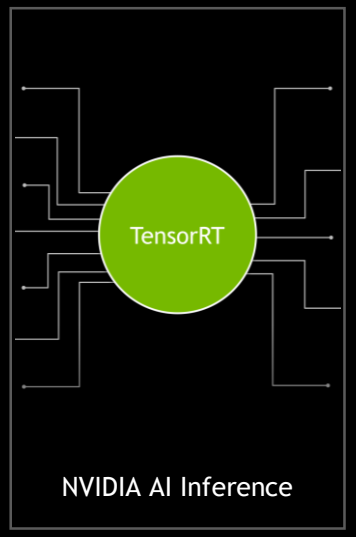
NEW with V100 32GB
NEW DGX-2

Every Cloud

NGC Now on AWS, GCP, AliCloud, Oracle

NVIDIA GPU Cloud

30 GPU-Optimized Containers



NVIDIA AI Inference

NEW TensorRT 4, TensorFlow, Kaldi, ONNX, WinML

TITAN V

Out of stock!

We are advancing GPU computing for deep learning and AI at the speed of light. We create the entire stack, and make it easily available in every computer, datacenter, and cloud. We supercharged NVIDIA AI with a new "double-sized" 32GB Volta GPU; announced the NVIDIA DGX-2, the power of 300 servers in a box; expanded our inference platform with TensorRT 4 and Kubernetes on NVIDIA GPU; and we built out the NVIDIA GPU Cloud registry with 30 GPU-optimized containers and made it available from more cloud service providers.

Quadro GV100

Reinventing the Workstation with Real-Time Ray Tracing and AI

ACCELERATING THE DELIVERY OF AI SOLUTIONS

AI-enabled transformations such as autonomous vehicles, personal assistants, and medical breakthroughs can greatly benefit society, but demand for applied AI is growing faster than the talent pool.

UnternehmerTUM is on a mission through its Applied.AI Initiative to accelerate the delivery of AI solutions by educating and connecting talent with state-of-the-art technology & industry companies. The government-backed initiative—which expects 3,000 participants and >30 new AI startups its first year—has selected the NVIDIA DGX-1V, DGX Station, and Deep Learning Institute to realize its vision for the Applied.AI Initiative as the leading innovation hub for AI in Germany and one of the top three centers in the world.

unternehmertum
Center for Innovation and Business Creation at TUM

Allianz

Bayerisches Staatsministerium für Wirtschaft und Medien, Energie und Technologie

Bundesministerium für Wirtschaft und Energie

Fraunhofer IIS

Giesecke & Devrient

Infineon

Landeshauptstadt München

THE LINDE GROUP

Munich RE

Porsche Consulting

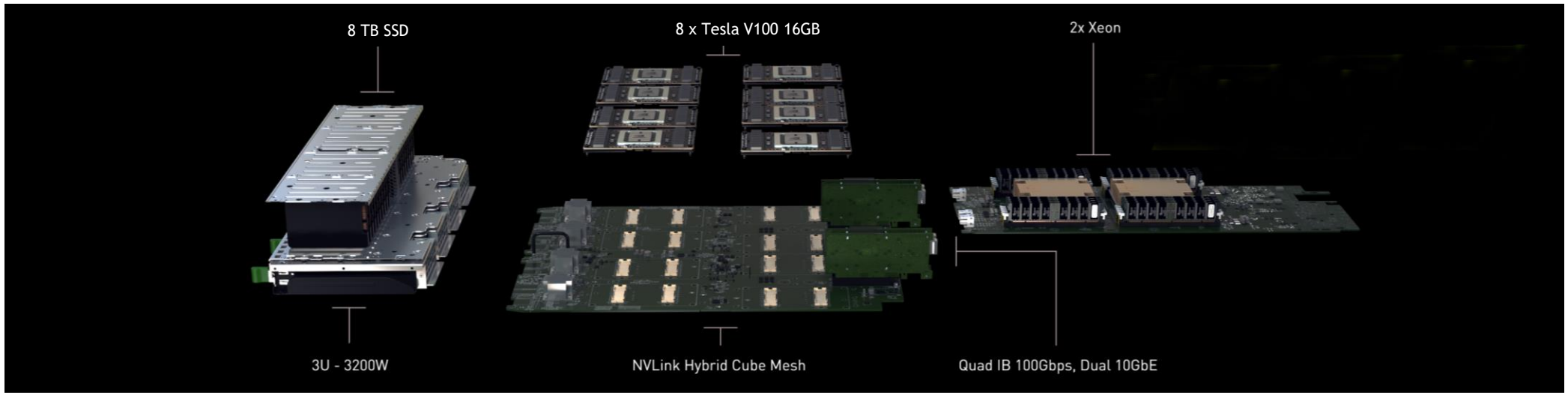
SAP

SIEMENS

WACKER

NVIDIA DGX-1 ARCHITECTURE

Highest Performance, Fully Integrated HW System

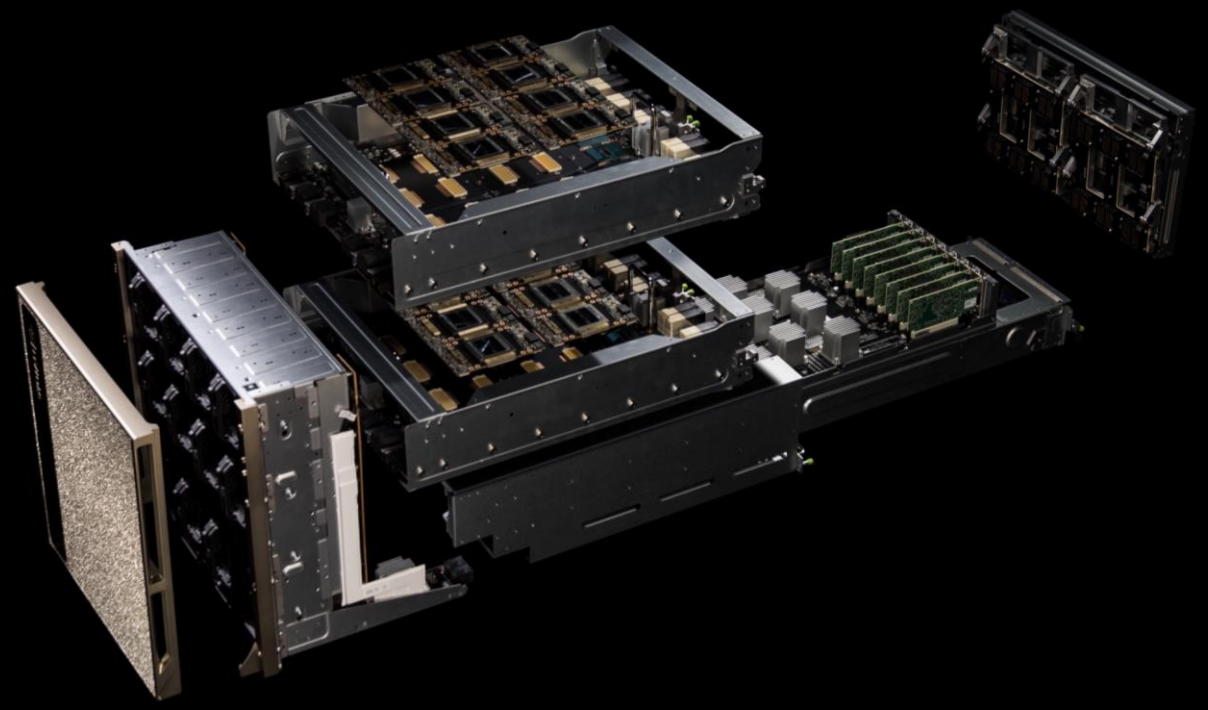


1 PFLOPS | 8x Tesla V100 32/16GB | 300 GB/s NVLink Hybrid Cube Mesh

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

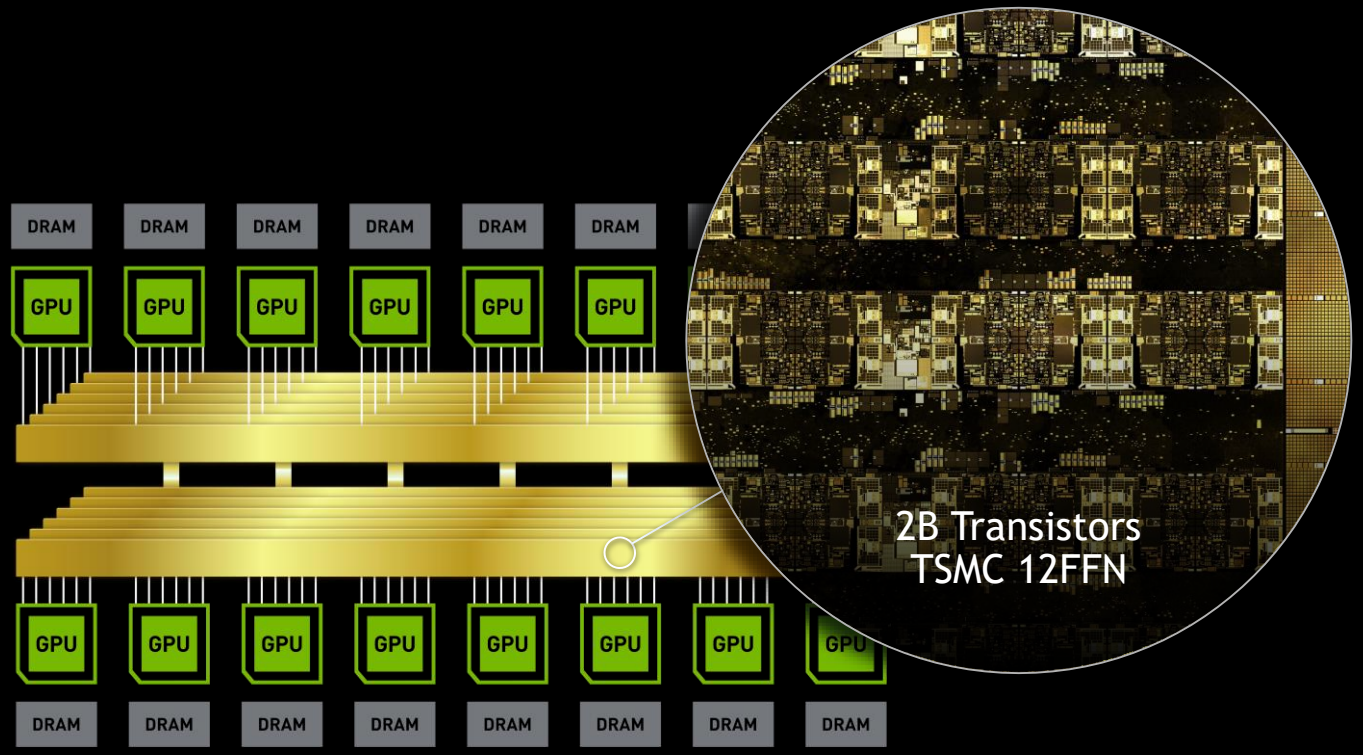
“NVIDIA Gave a Look Inside Its DGX-2, the Star of This Year’s GTC”

—EE Times



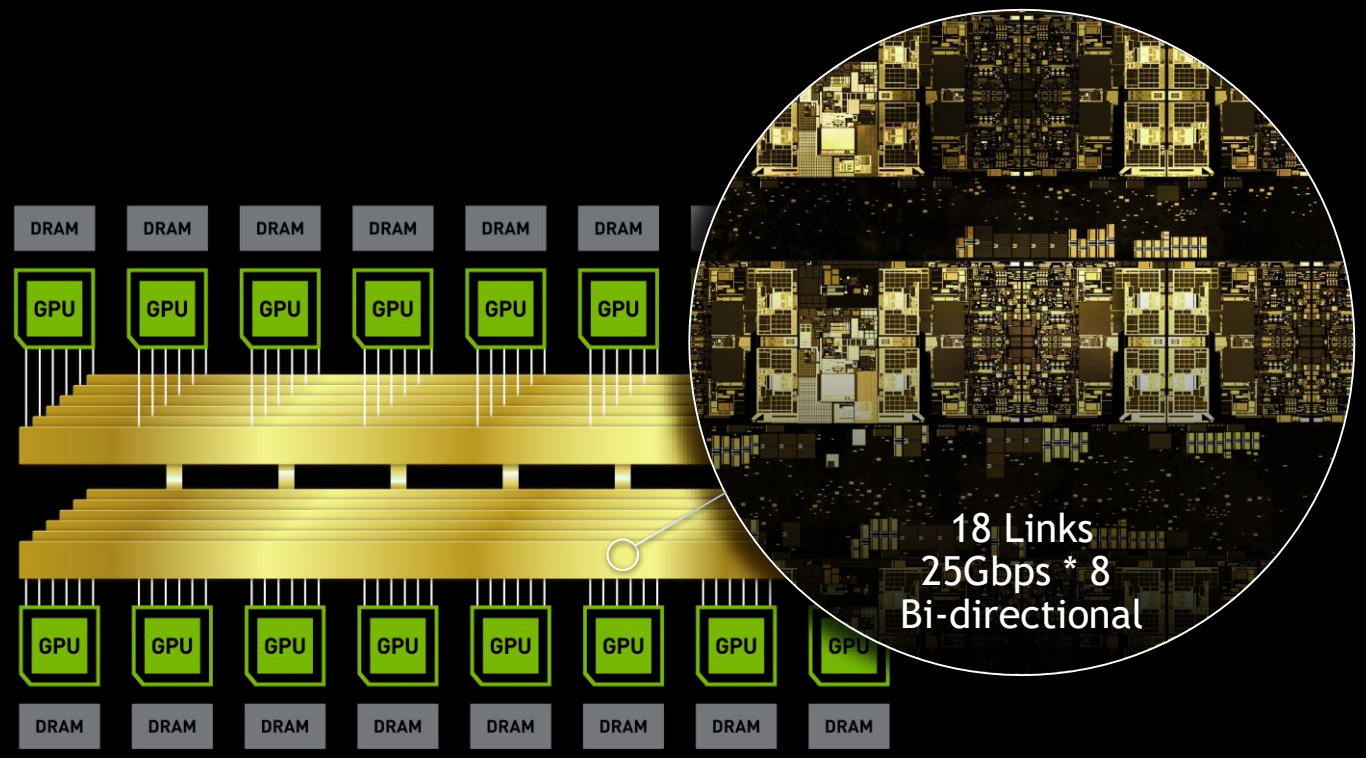
AI researchers want gigantic GPUs. We launched a breakthrough in deep learning computing with the introduction of NVIDIA DGX-2, the first single server capable of delivering two petaflops of computational power. DGX-2 features NVSwitch, a revolutionary GPU interconnect fabric which enables its 16 Tesla V100 GPUs to simultaneously communicate at a record speed of 2.4 terabytes per second. Programming DGX-2 is like programming “the largest GPU in the world.”

“THE WORLD’S LARGEST GPU”



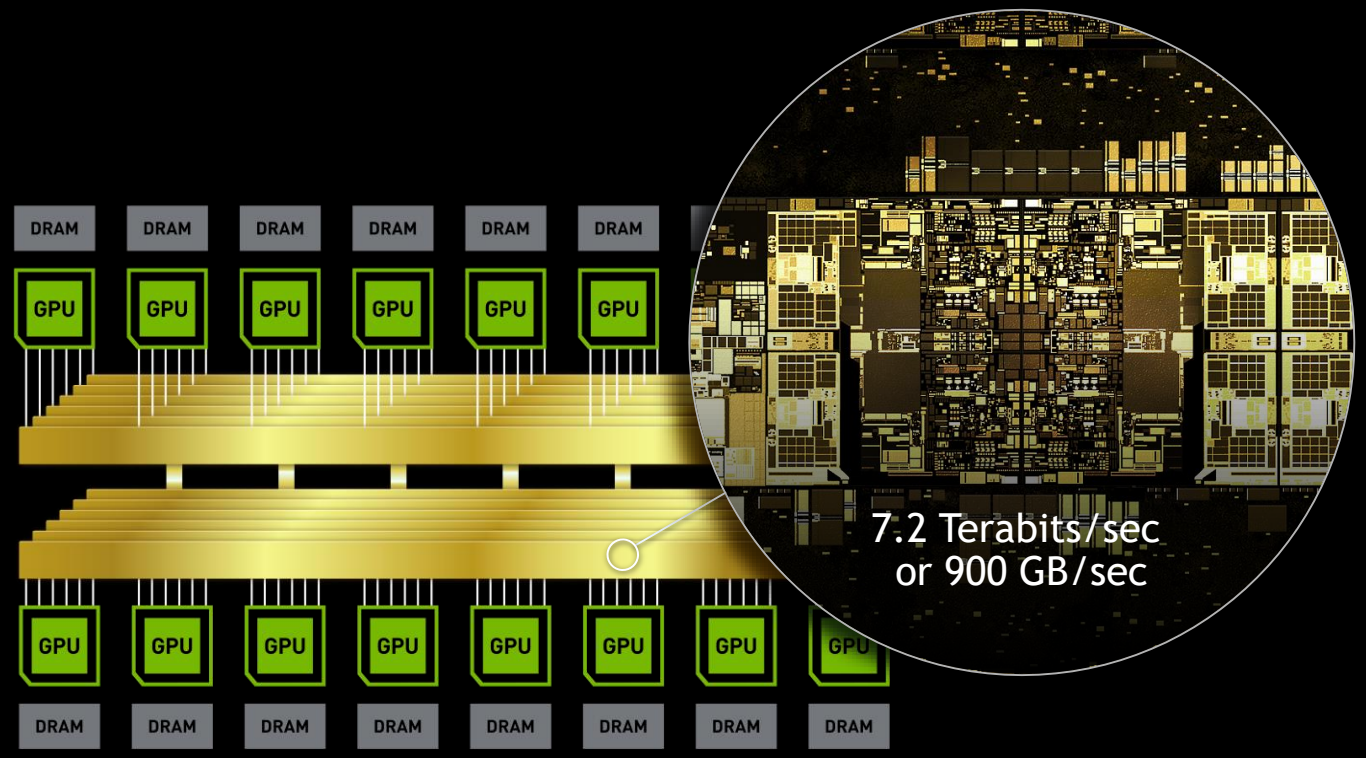
16 Tesla V100 32GB Connected by NVSwitch | On-chip Memory Fabric Semantic Extended Across All GPUs
512GB HBM2 and 14.4TB/sec Aggregate | 81,920 CUDA Cores | 2,000 TFLOPS Tensor Cores

“THE WORLD’S LARGEST GPU”



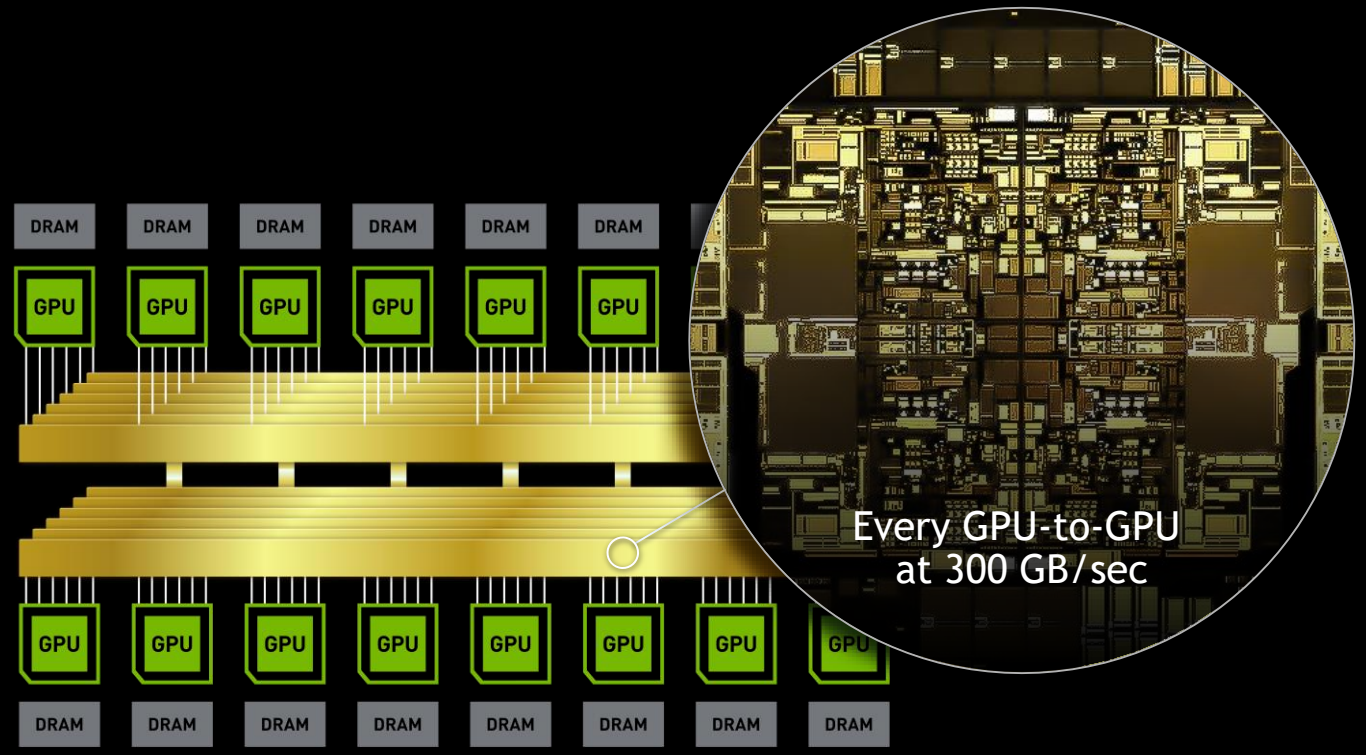
16 Tesla V100 32GB Connected by NVSwitch | On-chip Memory Fabric Semantic Extended Across All GPUs
512GB HBM2 and 14.4TB/sec Aggregate | 81,920 CUDA Cores | 2,000 TFLOPS Tensor Cores

“THE WORLD’S LARGEST GPU”



16 Tesla V100 32GB Connected by NVSwitch | On-chip Memory Fabric Semantic Extended Across All GPUs
512GB HBM2 and 14.4TB/sec Aggregate | 81,920 CUDA Cores | 2,000 TFLOPS Tensor Cores

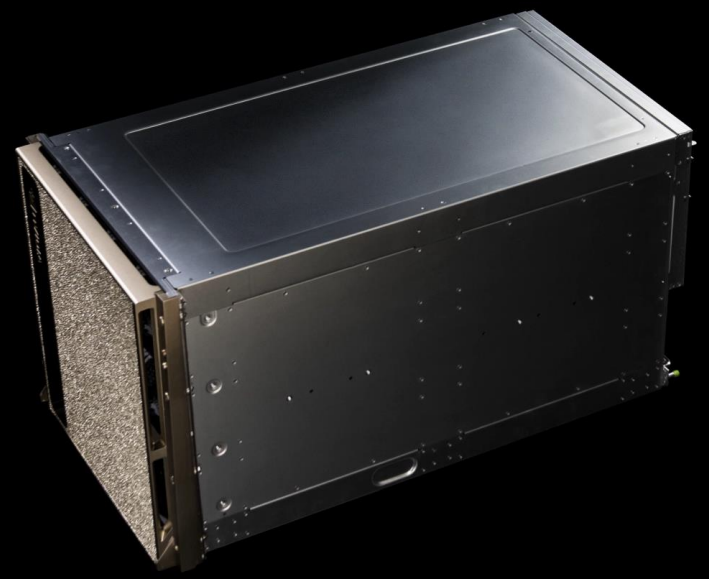
“THE WORLD’S LARGEST GPU”



16 Tesla V100 32GB Connected by NVSwitch | On-chip Memory Fabric Semantic Extended Across All GPUs
512GB HBM2 and 14.4TB/sec Aggregate | 81,920 CUDA Cores | 2,000 TFLOPS Tensor Cores

ANNOUNCING NVIDIA DGX-2

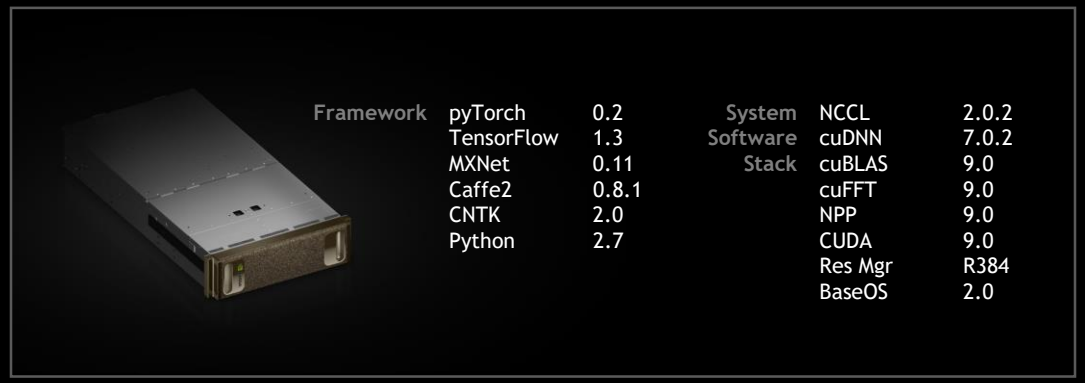
THE LARGEST GPU EVER CREATED



2 PFLOPS | 512GB HBM2 | 10 kW | 350 lbs

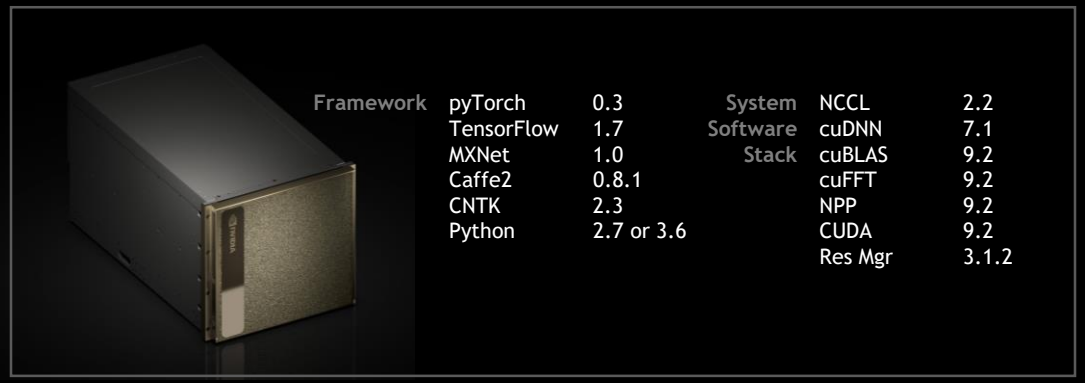
10X IN 6 MONTHS

DGX-1 V100 16GB — SEPT '17

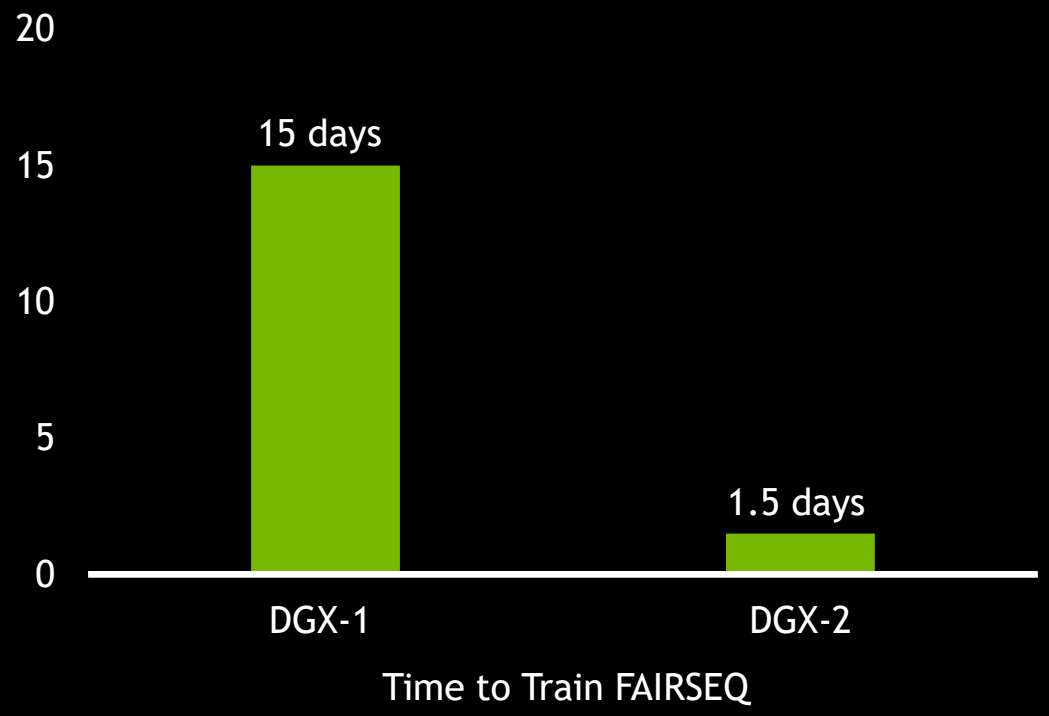


Framework	pyTorch	0.2	System Software Stack	NCCL	2.0.2
	TensorFlow	1.3		cuDNN	7.0.2
	MXNet	0.11		cuBLAS	9.0
	Caffe2	0.8.1		cuFFT	9.0
	CNTK	2.0		NPP	9.0
	Python	2.7		CUDA	9.0
				Res Mgr	R384
				BaseOS	2.0

DGX-2 V100 32GB — MAR '18

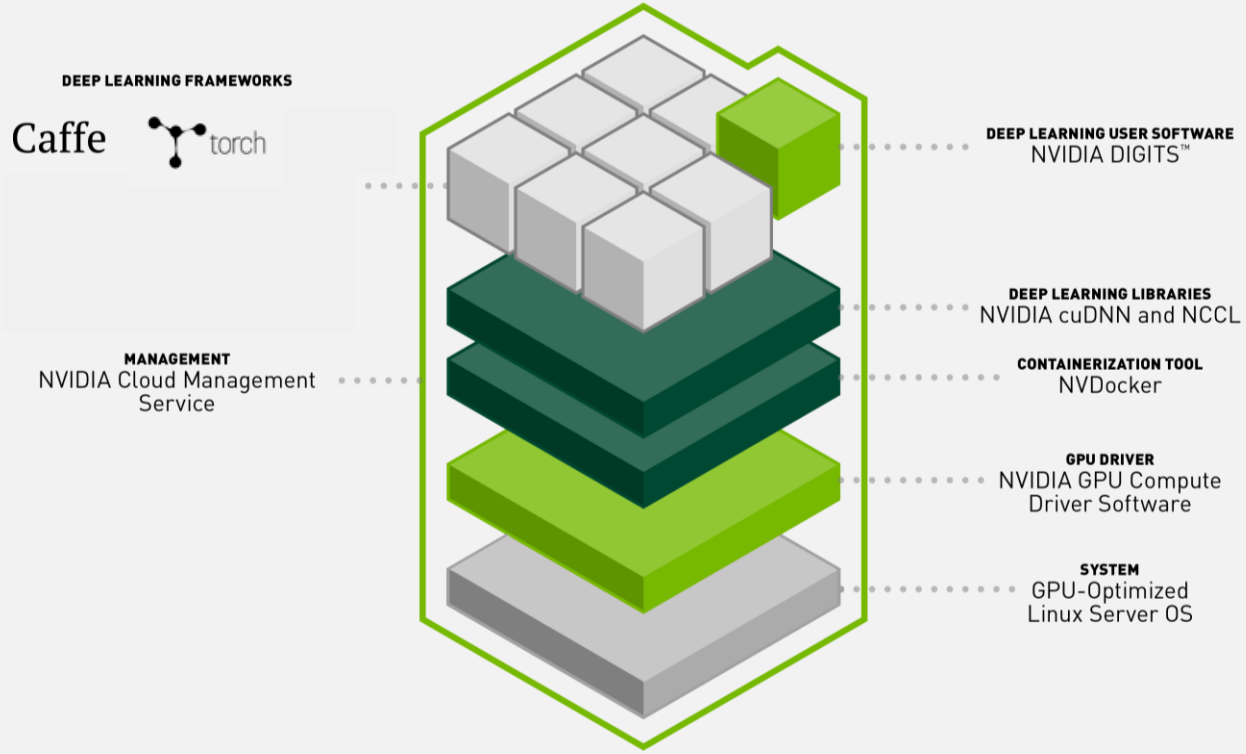


Framework	pyTorch	0.3	System Software Stack	NCCL	2.2
	TensorFlow	1.7		cuDNN	7.1
	MXNet	1.0		cuBLAS	9.2
	Caffe2	0.8.1		cuFFT	9.2
	CNTK	2.3		NPP	9.2
	Python	2.7 or 3.6		CUDA	9.2
				Res Mgr	3.1.2



DGX INTEGRATED STACK

Fully integrated Deep Learning platform



Instant productivity — plug-and-play, supporting every AI framework

Performance optimized across the entire stack

Always up-to-date via the cloud

Mixed framework environments — virtualized and containerized

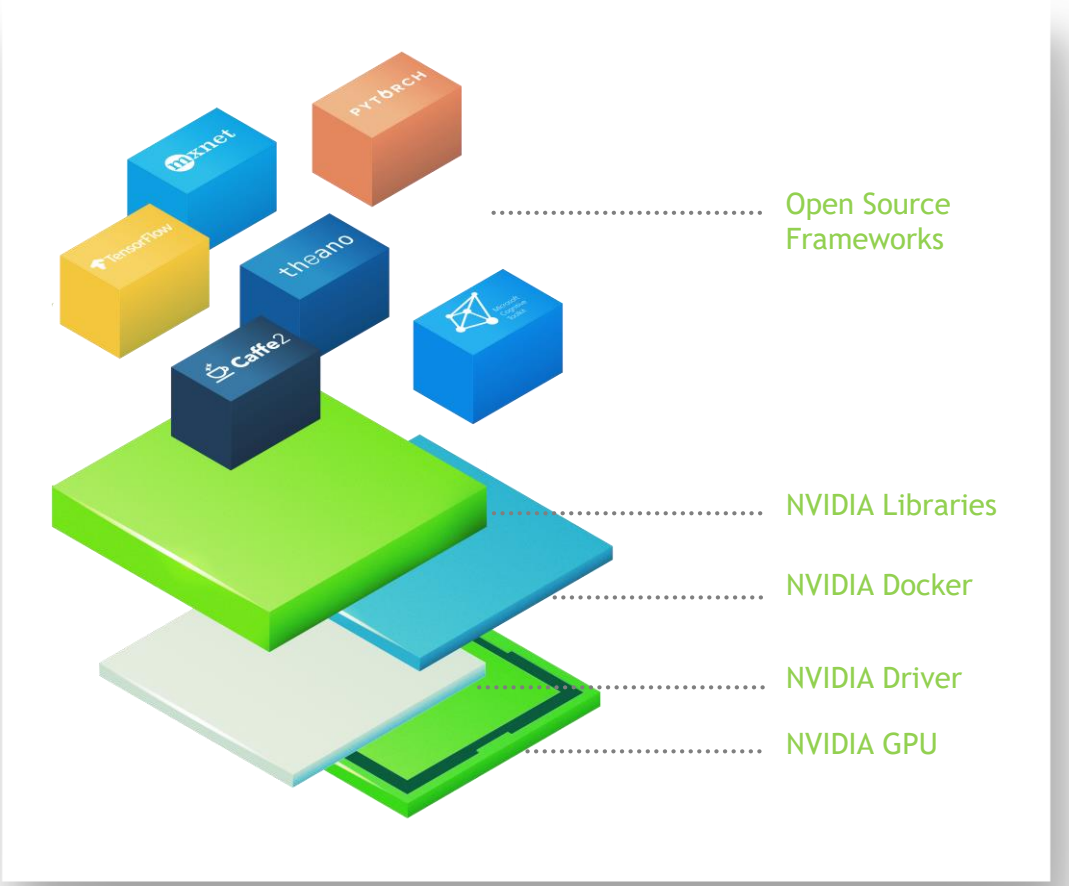
Direct access to NVIDIA experts

CHALLENGES WITH DEEP LEARNING

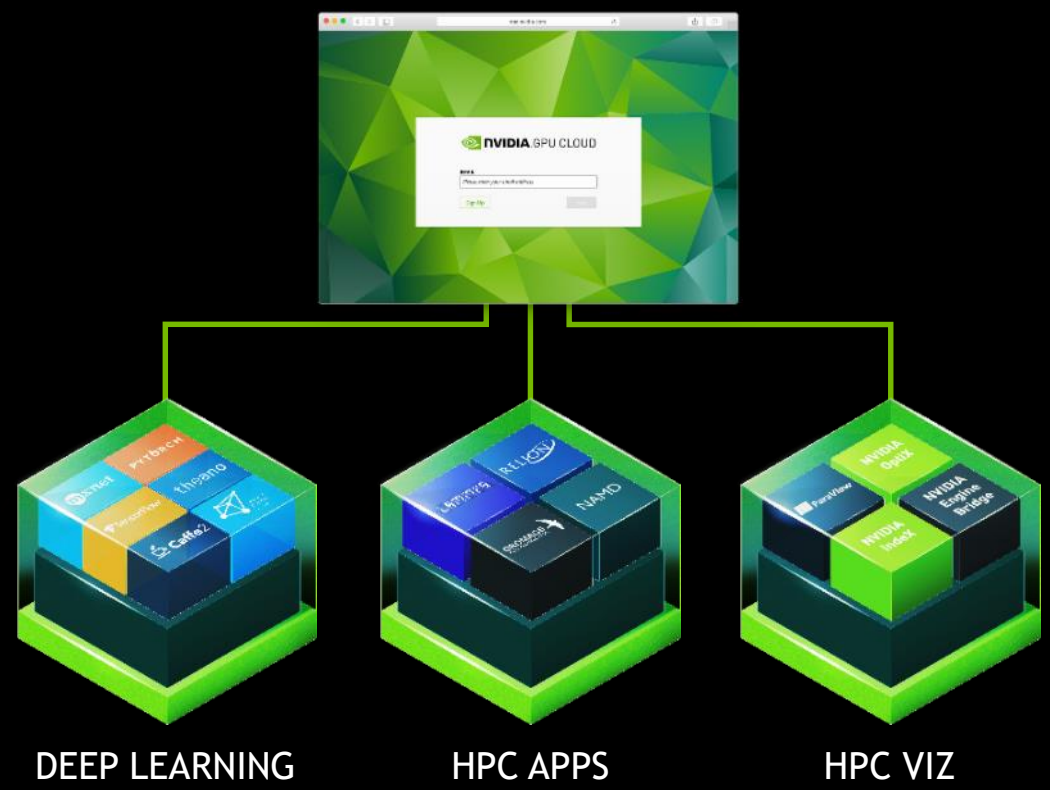
Current DIY deep learning environments are complex and time consuming to build, test and maintain

Development of frameworks by the community is moving very quickly

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



NVIDIA GPU CLOUD
SIMPLIFYING AI & HPC



NGC CONTAINER REGISTRY

10 @ LAUNCH

DEEP LEARNING

caffe
caffe2
cntk
cuda
digits
mxnet
pytorch
tensorflow
theano
torch

32 @ GTC 2018

DEEP LEARNING

caffe
caffe2
cntk
cuda
digits
mxnet
pytorch
tensorflow
tensorRT
theano
torch

HPC VIZ

paraview-holodeck
paraview-index
paraview-optix
IndeX*
VMD*

HPC

gamess
gromacs
lammps
namd
relion
PIConGPU*
CHROMA*
MILC*
CANDLE*
Lattice Microbes*

PARTNER

h2o
mapd
chainer
paddlepaddle
Kinetica*

GET STARTED TODAY WITH NGC

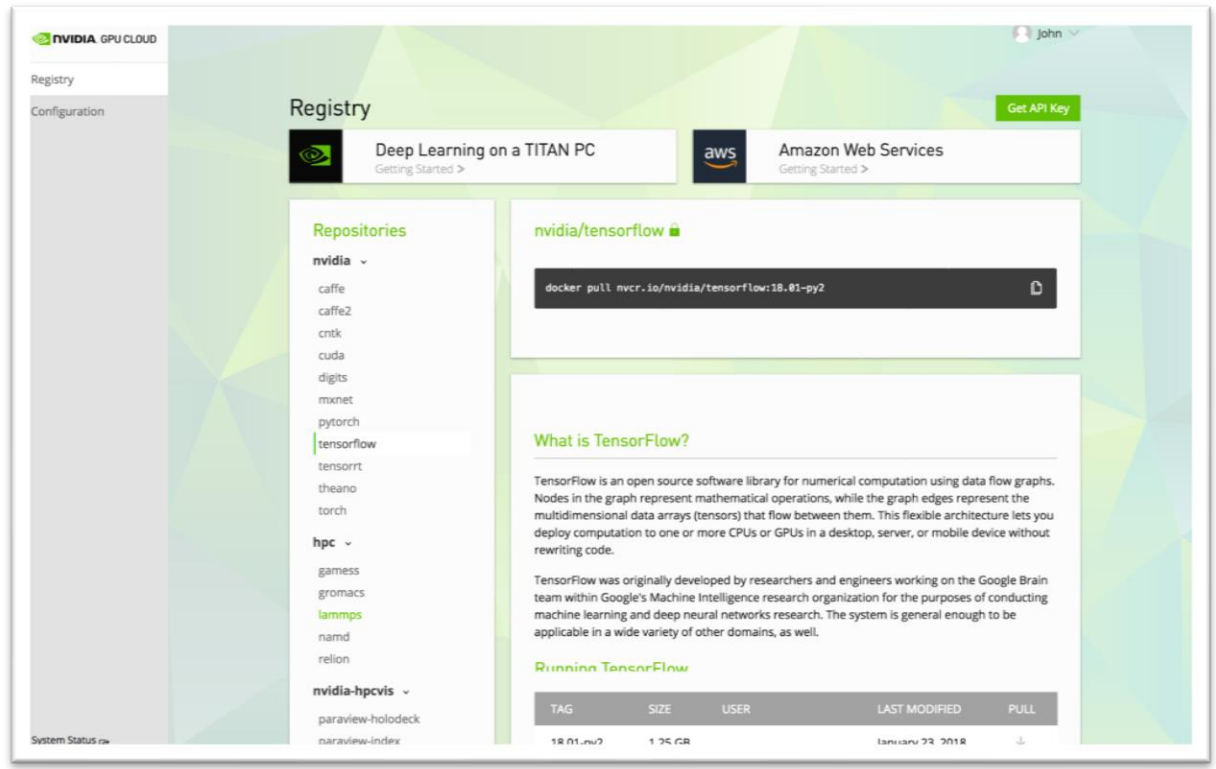
Sign up for no cost access

To learn more about all of the GPU-accelerated software on NVIDIA GPU Cloud, visit:

nvidia.com/cloud

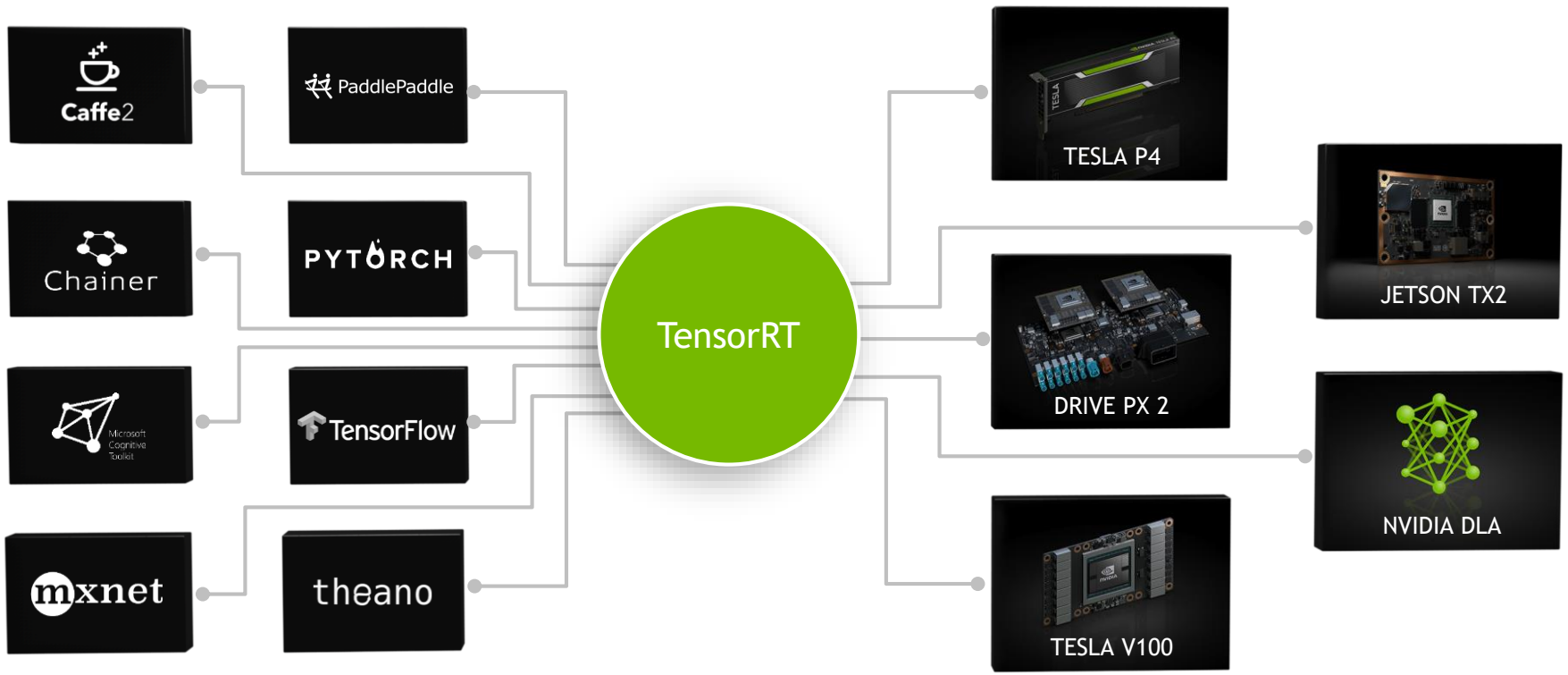
To sign up, go to:

nvidia.com/ngcsignup



NVIDIA TENSORRT

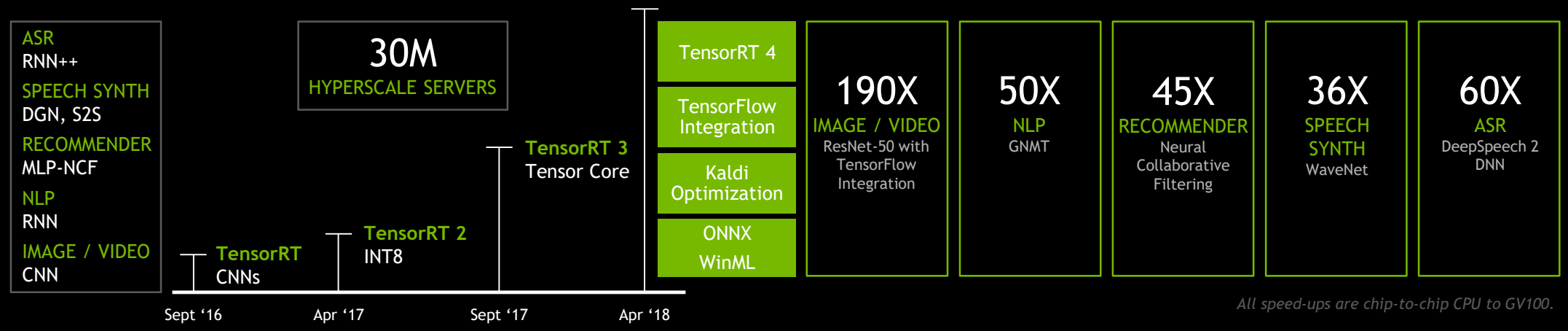
Programmable Inference Accelerator



Compile and Optimize Neural Networks | Support for Every Framework
Optimize for Each Target Platform

"NVIDIA Strengthened Its Inference Push by Unveiling TensorRT 4"

—TheStreet

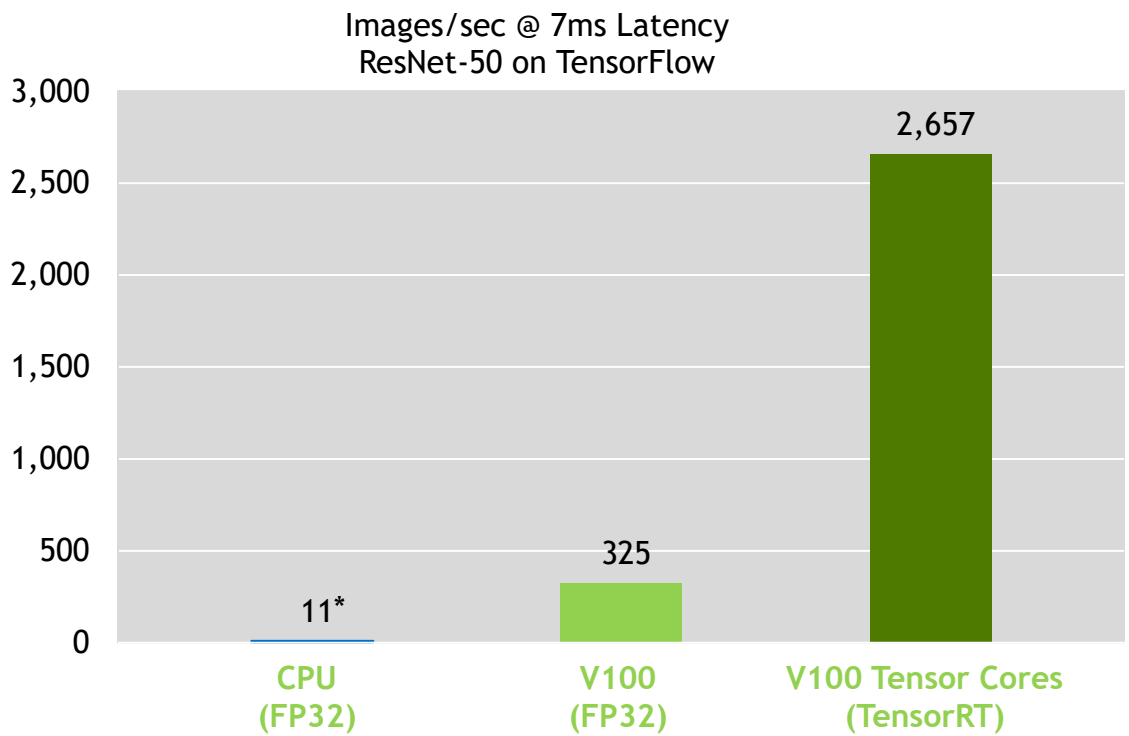


All speed-ups are chip-to-chip CPU to GV100.

Every hyperscale server — millions — will be accelerated for AI someday. The workload is complex — remember PLASTER — and the optimizing compiler technologies are still being invented. We announced TensorRT 4, the latest version of our inference software, and its integration into Google's popular TensorFlow framework. We announced that Kaldi, the most popular framework for speech recognition, is now optimized for GPUs. NVIDIA's close collaboration with partners such as Amazon, Facebook, and Microsoft makes it easier for developers to take advantage of GPU acceleration using ONNX and WinML. Hyperscale datacenters can save big money with NVIDIA Inference Acceleration.

TensorRT INTEGRATED WITH TensorFlow

Delivers 8x Faster Inference



* Min CPU latency measured was 83 ms. It is not < 7 ms.

CPU: Skylake Gold 6140, 2.5GHz, Ubuntu 16.04; 18 CPU threads.
Volta V100 SXM; CUDA (384.111; v9.0.176);
Batch size: CPU=1, TF_GPU=2, TF-TRT=16 w/ latency=6ms

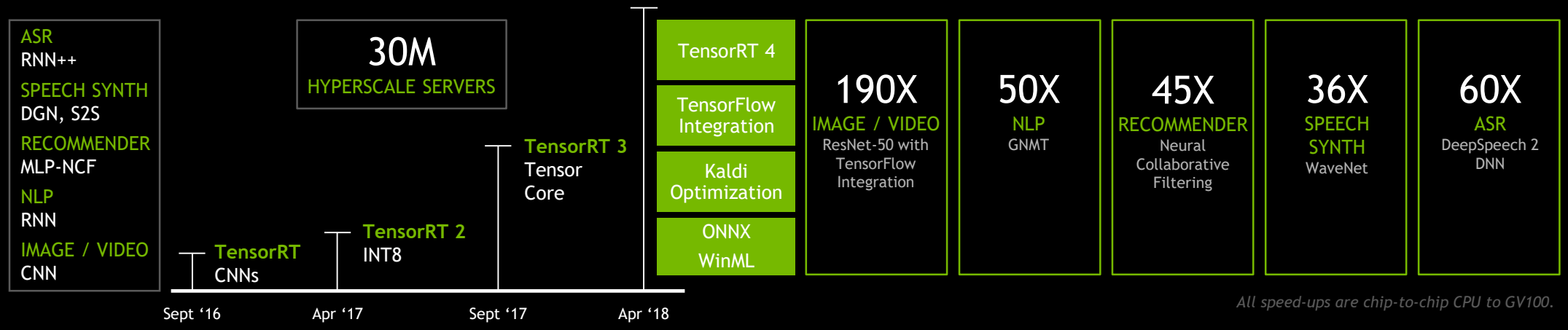
- AI Researchers
- Data Scientists

TensorFlow
TensorRT

Available in TensorFlow 1.7

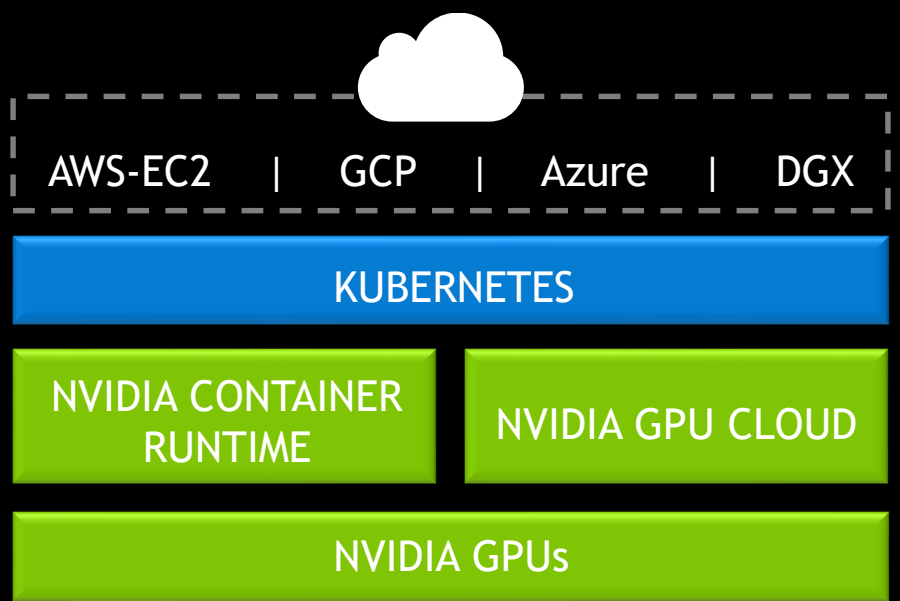
https://github.com/tensorflow/tensorflow

NVIDIA AI INFERENCE



All speed-ups are chip-to-chip CPU to GV100.

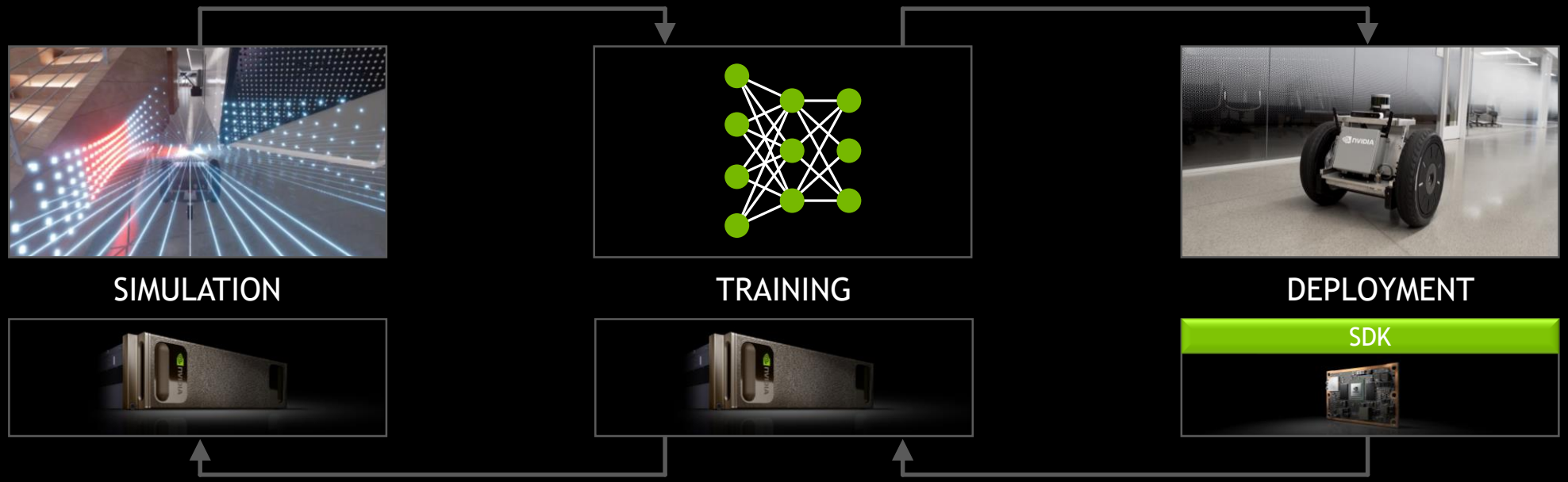
Container Orchestration for DL Training & Inference



KUBERNETES on NVIDIA GPUs

- Scale-up Thousands of GPUs Instantly
- Self-healing Cluster Orchestration
- GPU Optimized Out-of-the-Box
- Powered by NVIDIA Container Runtime
- Included with Enterprise Support on DGX
- Available end of April 2018

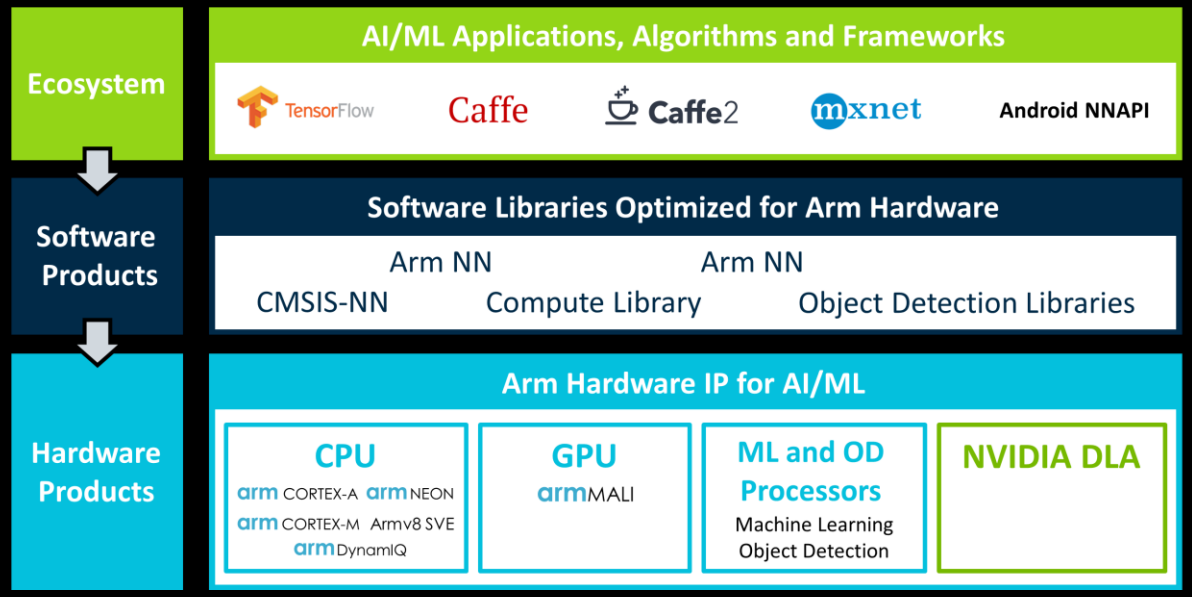
NVIDIA ISAAC ROBOTICS PLATFORM



https://developer.nvidia.com/isaac-sdk

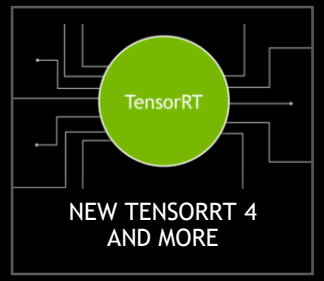
"A New A.I. Era Dawns for Chip Makers"

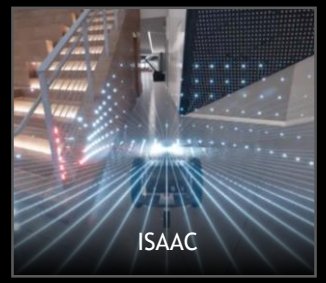
—Barron's



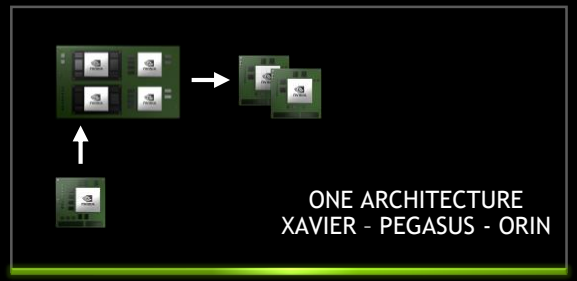
Billions of smart sensing devices will connect to the internet someday. NVIDIA and Arm announced a partnership to bring deep learning inferencing to the wide array of mobile, consumer electronics, and Internet of Things devices. Arm has integrated the NVDLA inference accelerator into its Project Trillium platform for machine learning. The collaboration will make it simple for IoT chip companies to integrate AI into their designs and help put intelligent, affordable products into the hands of billions of consumers.

THE GPU COMPUTING REVOLUTION CONTINUES





Kubernetes
On
NVIDIA
GPUs



GRAPHICS

AI

AUTO

NEW PLATFORMS

END-TO-END PRODUCT FAMILY

TRAINING

DESKTOP

GPU-Accelerated Container Registry

TITAN V

DGX Station

DATA CENTER

GPU-Accelerated Container Registry

DGX-1 Server

TESLA V100

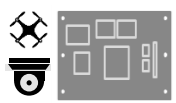
INFERENCE

DATA CENTER

TESLA P4

TESLA V100

EMBEDDED



JETPACK SDK

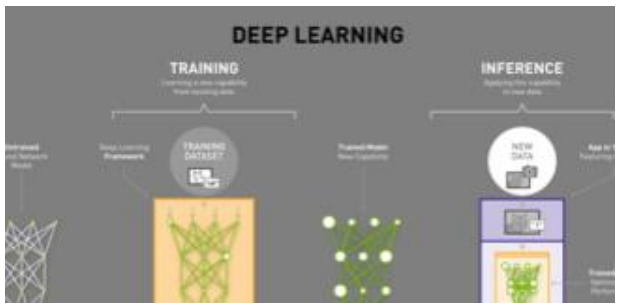
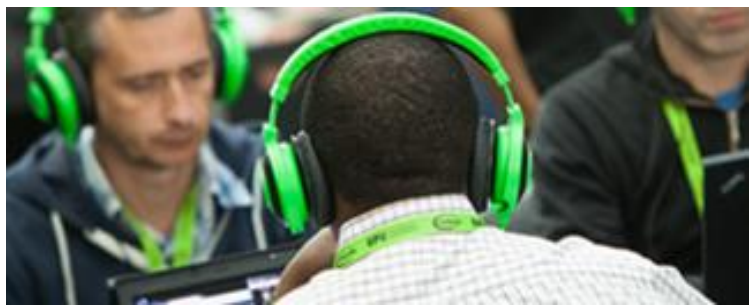
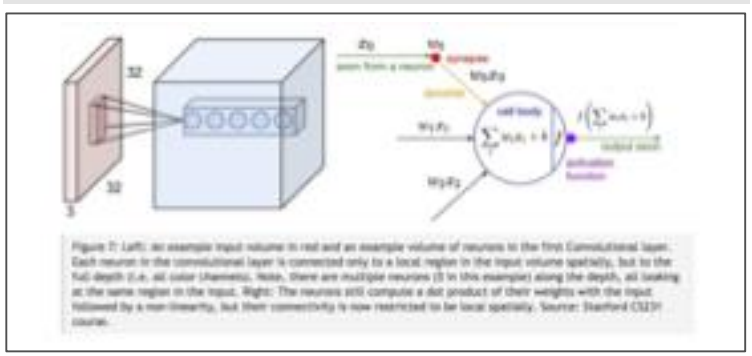
Jetson

AUTOMOTIVE

DriveWorks SDK

Drive PX

DLI RESOURCES

FOR EVERYONE	FOR DEVELOPERS, DATA SCIENTISTS, RESEARCHERS	
		
Intro Materials	Self-paced Labs	Onsite Workshops
		
Case Studies	Courses	Technical Blogs

http://www.nvidia.com/dlilabs

NVIDIA®