

An abstract network visualization on a black background. It features a dense web of thin green lines connecting numerous small, glowing green and yellow nodes. A prominent, highly interconnected cluster of yellow nodes is located on the right side of the image, while the left side shows a more sparse network of green nodes. The overall effect is one of complex connectivity and data flow.

GTC 2018



“GTC was the introduction to the future of AI, a protector, a healer, a helper, a guardian, a visionary, and just a little slice of amazing.”

—IT Business Edge



“Clearly the adoption of GPU computing is growing.”

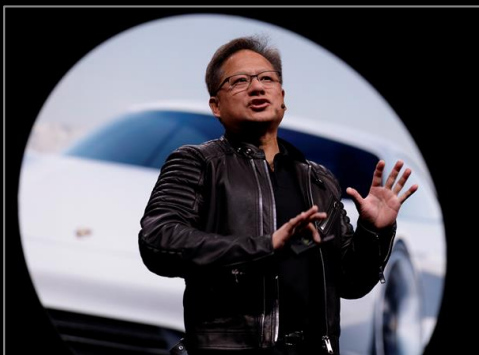
—ZDNet

“If the excitement at GTC this week is any indication, the market will need lots and lots of GPUs in the years to come.”

—Datanami



Deep Learning Institute
10,000 training instances



Keynote
550K views



Show Floor
150 exhibitors



Inception Awards
\$1M to 3 AI startups

GTC is ground zero of the GPU computing movement. NVIDIA pioneered this computing model for those doing groundbreaking work that cannot be done without a supercharged computer. GTC 2018 was a four-day event with 8,300 registered attendees who were offered more than 600 technical sessions. More than 200 reporters and 80 industry analysts experienced first-hand NVIDIA's lineup of announcements.

“NVIDIA’s CEO, Jensen Huang, is a consummate showman and he didn’t disappoint this year.”

—IT Business Edge

“Ray Tracing and Giant GPUs Electrify GTC 2018”

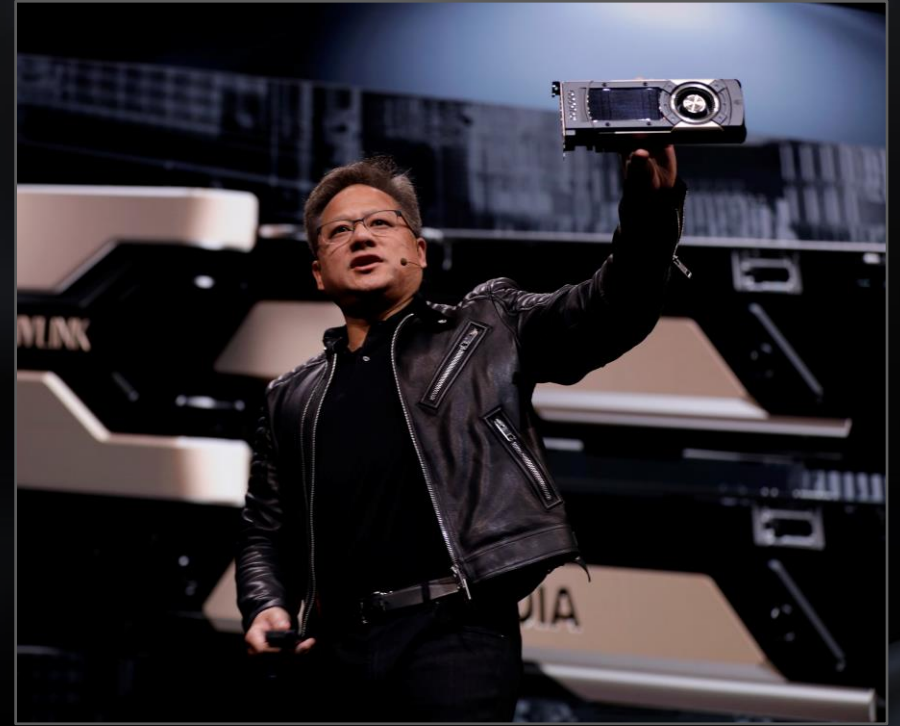
—Electronic Design



Real-time ray tracing has been the dream of computer scientists since it was first described nearly 40 years ago. NVIDIA stunned the audience with its new NVIDIA RTX technology — a platform for real-time ray tracing.

*“NVIDIA Is Really Throwing
Down the Hammer with
the Quadro GV100”*

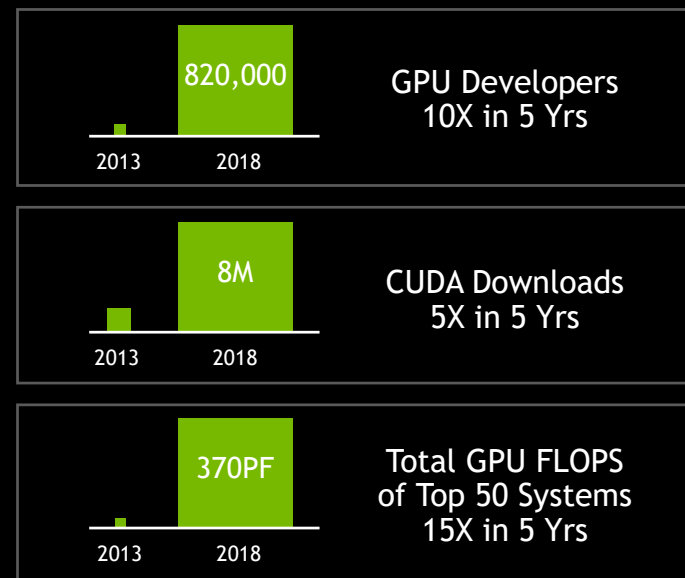
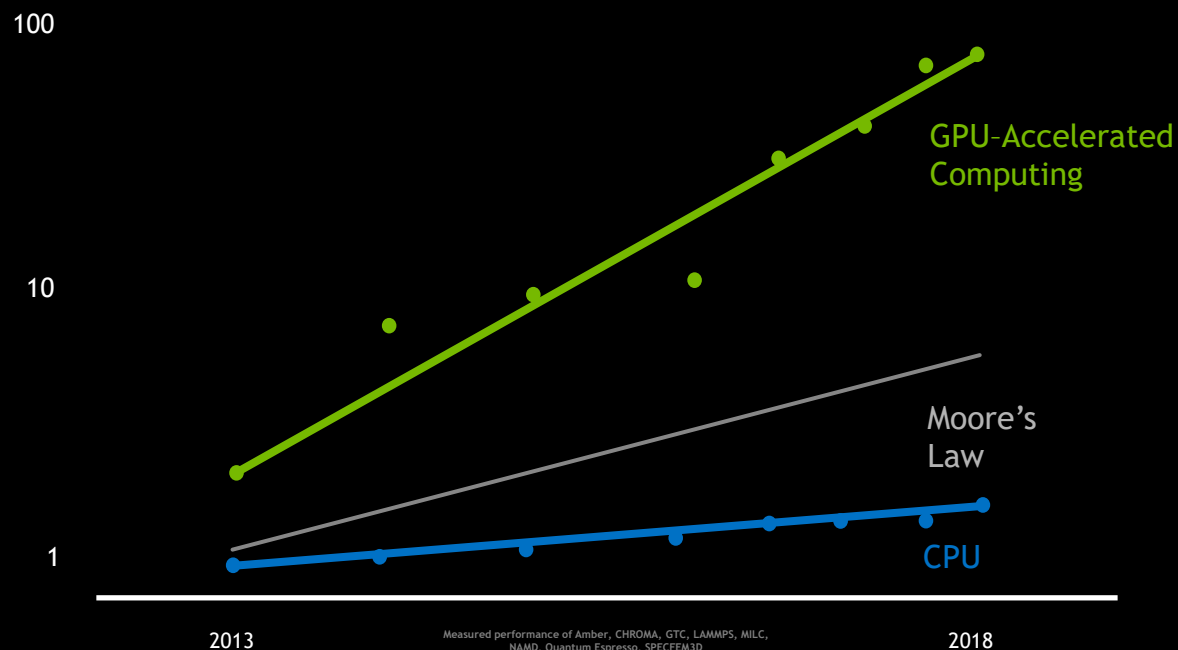
—Hot Hardware



The Quadro GV100 with NVIDIA RTX technology is the greatest advance in computer graphics of the past 15 years, since our introduction of the programmable shader. NVIDIA RTX is the culmination of 10 years of research, combining a new GPU architecture, algorithms, and deep learning in a way that no one else can. NVIDIA has reinvented computer graphics, again.

“NVIDIA Is So Far Ahead of the Curve”

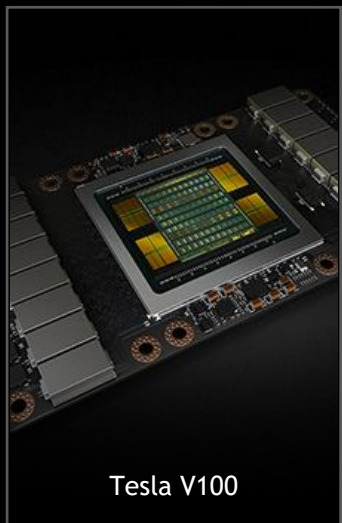
—The Inquirer



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.

“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



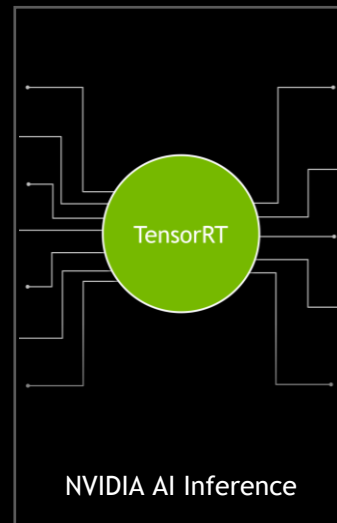
Every Cloud

NGC Now on AWS, GCP,
AliCloud, Oracle



NVIDIA GPU Cloud

30 GPU-Optimized
Containers



NVIDIA AI Inference

NEW TensorRT 4, TensorFlow,
Kaldi, ONNX, WinML



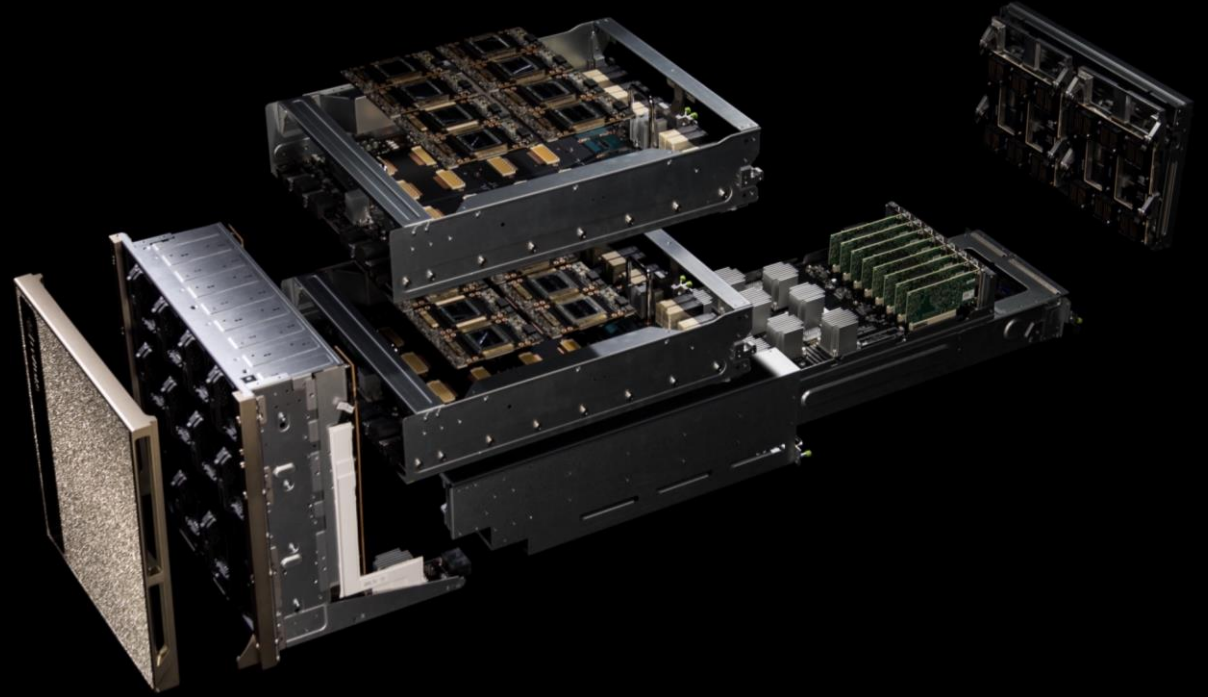
TITAN V

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.

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

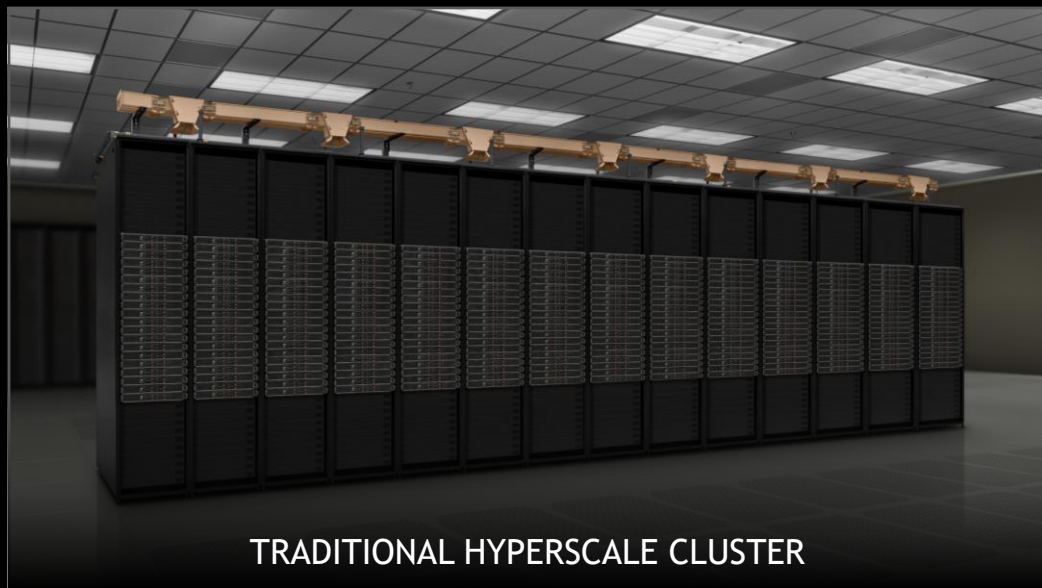
—EE Times



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.”

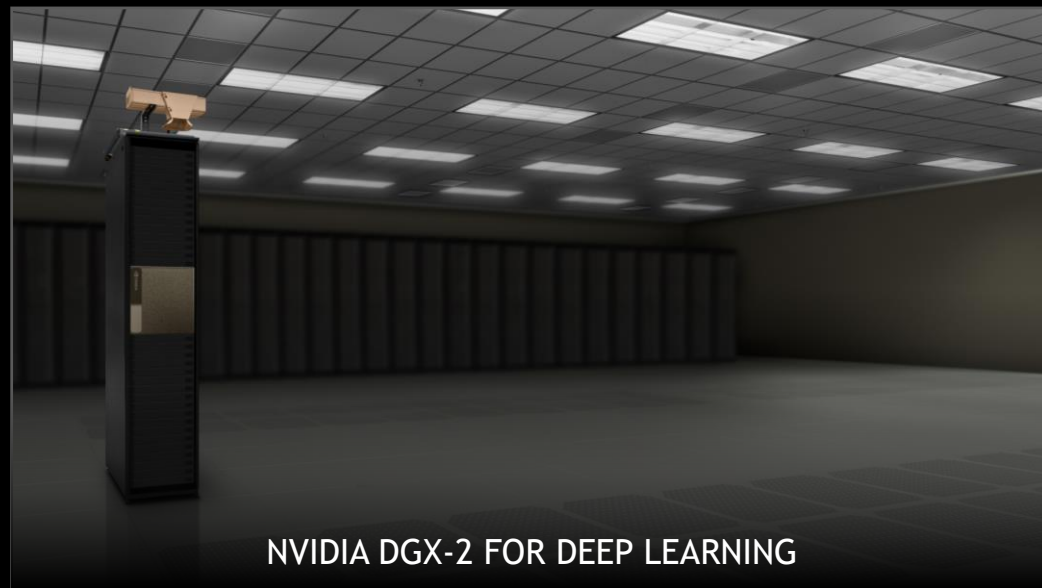
“The More GPUs You Buy, The More You Save”

—Jensen Huang



TRADITIONAL HYPERSCALE CLUSTER

300 Dual-CPU Servers | \$3M | 180 kW



NVIDIA DGX-2 FOR DEEP LEARNING

1 DGX-2 | \$399K | 10kW
1/8 the Cost 1/60 the Space 1/18 the Power

Convolutional Networks



Encoder/Decoder



ReLU



BatchNorm



Concat

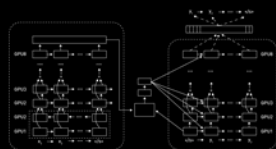


Dropout



Pooling

Recurrent Networks



LSTM



GRU



Beam Search



WaveNet



CTC



Attention

Generative Adversarial Networks



3D-GAN



MedGAN



Conditional GAN

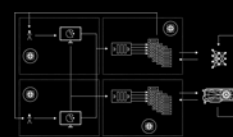


Coupled GAN

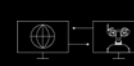


Speech Enhancement GAN

Reinforcement Learning



DQN



Simulation



DDPG

New Species



Capsule Nets



Mixture of Experts




Neural Collaborative Filtering



Block Sparse LSTM

There is a Cambrian explosion of neural networks. Since AlexNet, thousands of new models have emerged. With hundreds of layers and billions of parameters, their complexity has soared by 500X in just 5 years. The hyperscale datacenters that host them serve billions of people, cost billions to operate, and are among the most complex computers the world has ever made. Maintaining great quality of service while minimizing cost is incredibly difficult. Jensen helps us remember with PLASTER.



PROGRAMMABILITY

LATENCY

ACCURACY

SIZE

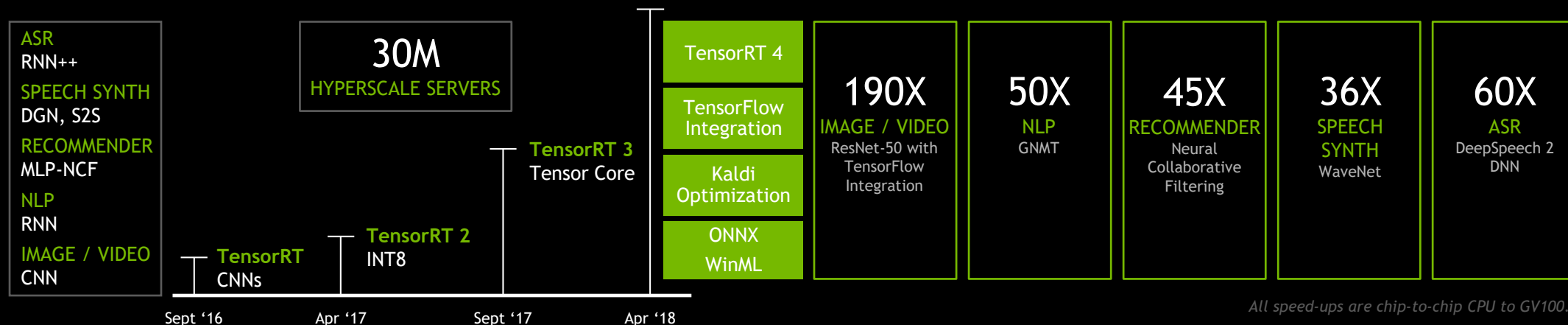
THROUGHPUT

ENERGY EFFICIENCY

RATE OF LEARNING

“NVIDIA Strengthened Its Inference Push by Unveiling TensorRT 4”

—TheStreet

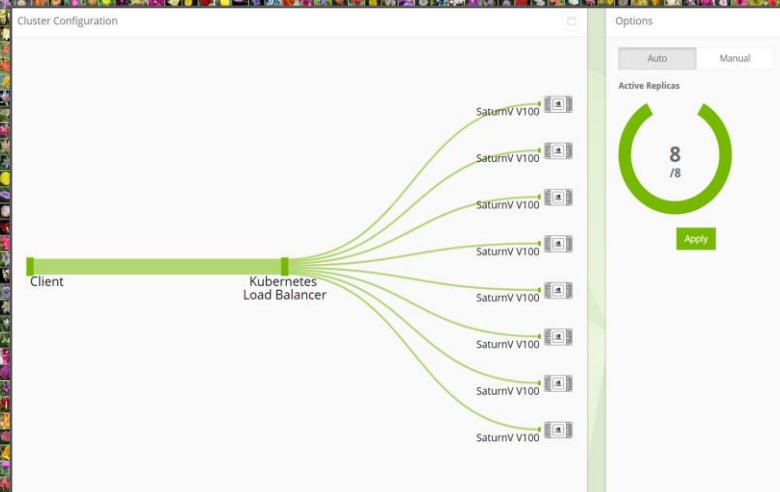
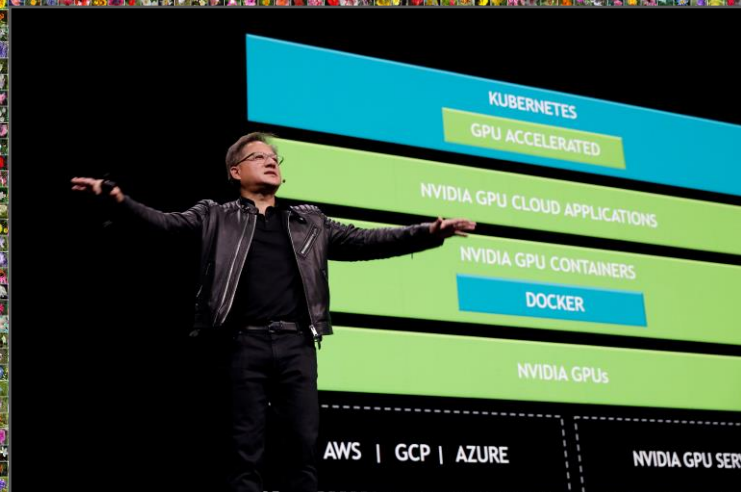


Every hyperscale server — millions — will be accelerated for AI someday. The workload is complex — remember PLASTER — and the optimizing compiler technologies are still being invented. We announced TensorRT 4, the latest version of our inference software, and its integration into Google’s popular TensorFlow 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.

“NVIDIA Brings Joy by Bringing GPU Acceleration to Kubernetes”

—TechCrunch

Kubernetes on NVIDIA GPUs
Scale Out Infrastructure for the Accelerated Datacenter

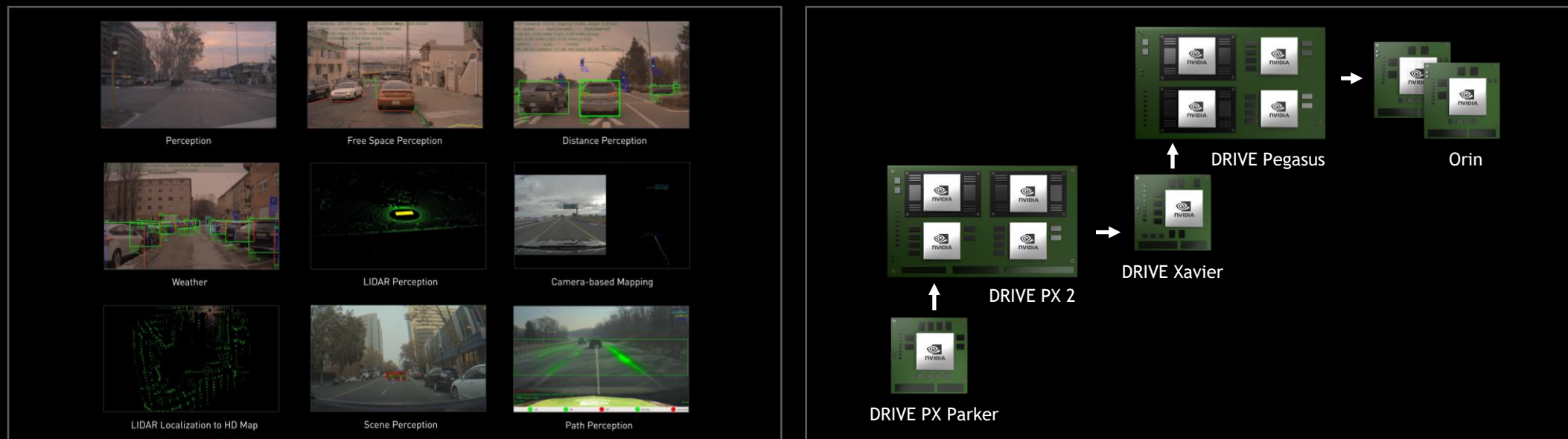


Images Per Sec: 5269

Kubernetes on NVIDIA GPUs is here! Hyperscale datacenters are made up of thousands of CPUs, memory modules, networking and storage devices, and system software. Orchestration is key. We announced Kubernetes-ONG, software that allows orchestration of resources across multi-cloud GPU clusters. Provision on-prem and cloud GPU resources with a single command — even in the event of a system failure.

“NVIDIA Already Has Their Eye on What’s Next ... And That Is Orin”

—AnandTech



Autonomous vehicles will modernize the \$10 trillion transportation industry — making our roads safer and our cities more efficient. NVIDIA DRIVE is a scalable AI car platform that spans the entire range of autonomous driving, from traffic-jam pilots to robotaxis. More than 370 companies contributing to the AV industry have adopted DRIVE. At GTC, we demonstrated our latest DRIVE software stack and gave a sneak peek at our next DRIVE AI car computer, “Orin.”

“There Are Too Many Situations to Train For and Not Enough Testing Time on the Highways”

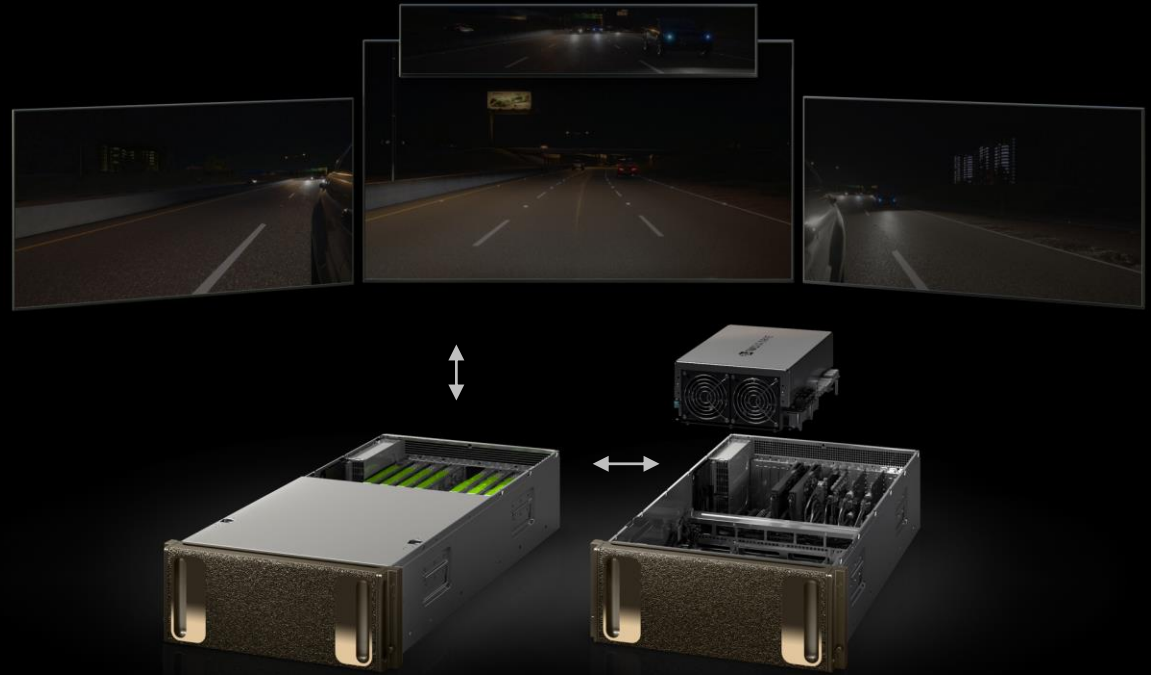
—VentureBeat



Each year, 10 trillion miles are driven around the world. Test cars can eventually cover millions of miles, an insignificant fraction of all the scenarios. We need to cover billions of miles to create a safe and reliable system. We have to invent something new.

“NVIDIA Redefines Autonomous Vehicle Testing with VR Simulation System”

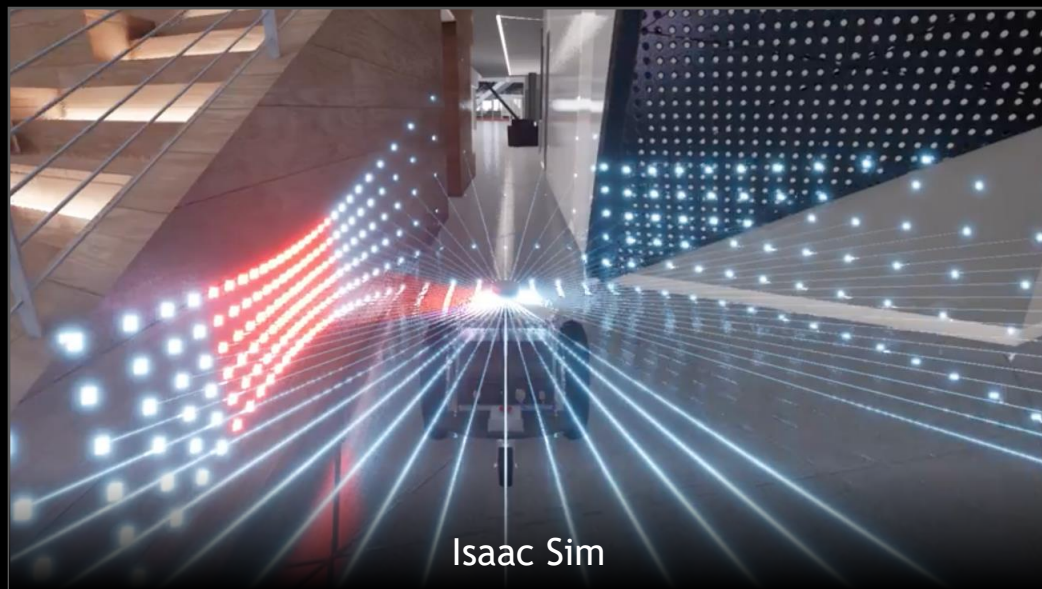
—ZDNet



NVIDIA DRIVE Constellation allows cars to drive billions of miles in virtual reality. Constellation consists of two different GPU servers. The first simulates the environment and what is detected by the car's many sensors — cameras, radar, and lidar. The second is the NVIDIA DRIVE Pegasus AI car computer that runs the complete AV software stack and processes the simulated detected data as if it were coming from a real car.

“The Robots Are Coming, Thanks to NVIDIA”

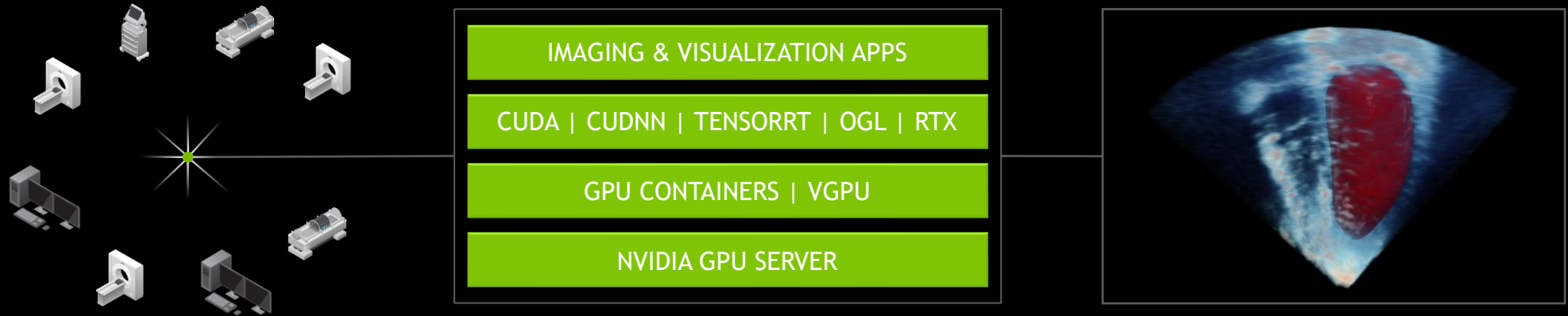
—FutureCar



The next chapter of AI is autonomous machines. We created a robotics platform called NVIDIA Isaac to accelerate the development and deployment of robotics across a broad range of industries. The Isaac SDK performs the important functions of robotics — perception, localization, navigation, and manipulation. We also created Isaac Sim, a virtual reality simulator where roboticists can create and train robots. Drop the software created in Isaac Sim into a robot with the Isaac SDK, and an intelligent machine is born.

“Medical Imaging at the Speed of Light: NVIDIA’s Clara Supercomputer”

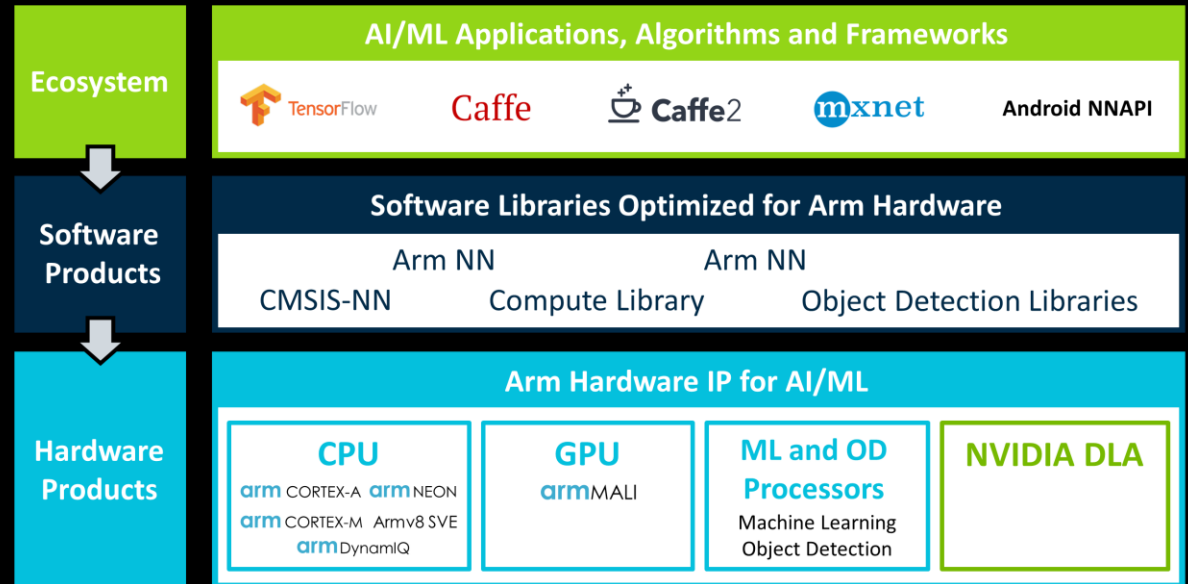
—ZDNet



Early detection is the most powerful weapon to treat disease. The latest breakthroughs of AI and computational imaging can help, but only if put into the hands of doctors using the 3 million medical instruments built a decade ago. We introduced Project Clara, NVIDIA’s medical imaging supercomputer, to do just that. With Clara, existing instruments will be supercharged with state-of-the-art image reconstruction, object detection and segmentation, and visualization capabilities.

“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.

“NVIDIA Stuns by Driving a Car in Real Life Through Virtual Reality”

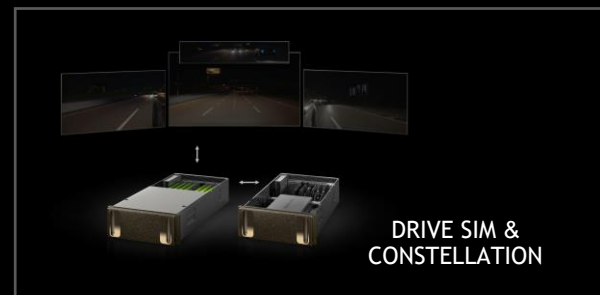
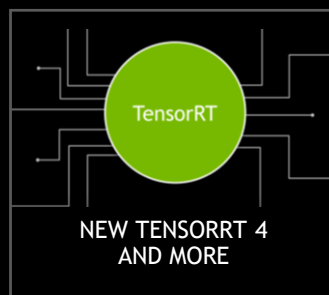
—Yahoo Finance



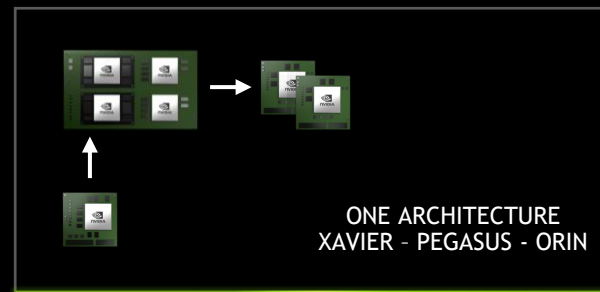
How do you drive a car remotely? Use virtual reality to teleport into the mind of an autonomous machine. Huh? As a final, mind-bending treat for the GTC audience, Jensen and NVIDIA's auto team demonstrated "Project Wakanda." A driver on the keynote stage entered virtual reality through NVIDIA Holodeck to teleport into a car at the back of the San Jose Convention Center. Seeing everything through the surround cameras of the car, he was able to remotely drive the car around a parked truck and safely into a parking space.

“What We Saw at GTC Was the Future”

—IT Business Edge



Kubernetes
On
NVIDIA
GPUs



GRAPHICS

AI

AUTO

NEW PLATFORMS

“NVIDIA CEO Jensen Huang: Tomorrowland of Technology”

—CNBC



Capping off an exciting week at GTC 2018, Jensen appeared on CNBC’s Mad Money for an interview with famed “NVIDIA” dog owner and stock market prognosticator Jim Cramer. The interview covered a broad set of topics, from our recently announced products to Jensen’s point of view on the future of computer graphics, AI, and autonomous cars.



*“NVIDIA Shows Why It’s
Leading in AI Mind
Expansion at GTC 2018”*

—Forbes

*“There’s no question that
NVIDIA remains one step
ahead.”*

—The Inquirer

*“The excitement
throughout the entire
session was amazing.”*

—Electronic Times

*“The company is about to
drop the pedal and hit
another gear in the AI race.”*

—Forbes

