SUPERCHARGED COMPUTING FOR THE DA VINCIS AND EINSTEINS OF OUR TIME

Twenty-five years ago, we set out to transform computer graphics.

Fueled by the massive growth of the gaming market and its insatiable demand for better 3D graphics, we've evolved the GPU into a computer brain at the intersection of virtual reality, high performance computing, and artificial intelligence.
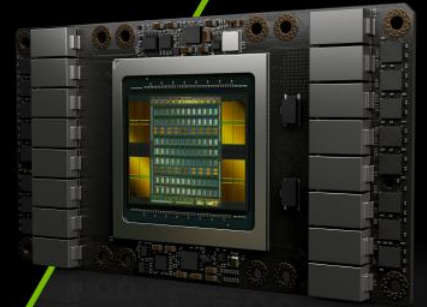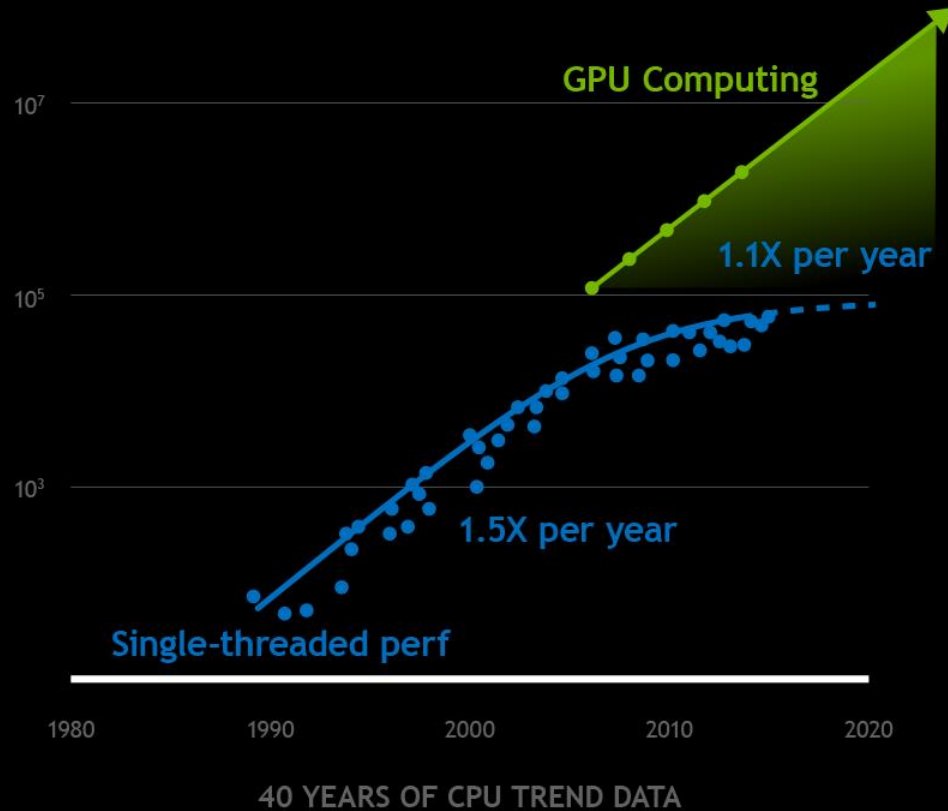
NVIDIA GPU computing has become the essential tool of the da Vincis and Einsteins of our time. For them, we've built the equivalent of a time machine.

Image: Scientists using GPU computing were able to "see" gravitational waves for the first time in human history.

A SUPERCHARGED LAW

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 optimized applications. The world is jumping on board — today, there are some 800,000 GPU developers.

GPU Computing

10^7

1.1X per year

10^5

10^3

1.5X per year

Single-threaded perf

1980   1990   2000   2010   2020

40 YEARS OF CPU TREND DATA

NVIDIA

GPU TECHNOLOGY CONFERENCE

THE EPICENTER OF GPU COMPUTING

GTC is ground zero of the GPU computing movement. This year's flagship event was a four-day gathering of 8,300 registered attendees who were offered more than 600 technical sessions. Three hundred reporters and analysts experienced first-hand NVIDIA's lineup of announcements. Now a global tour, GTCs are held around the world, from Washington to Munich to Tokyo.

NVIDIA DEFINES MODERN COMPUTER GRAPHICS

Our invention of the GPU in 1999 made real-time programmable shading possible, giving artists an infinite palette for expression. This year, the introduction of NVIDIA RTX™ ray-tracing technology fulfilled another vision of computer scientists, paving the way to new levels of art and realism in real-time graphics.

We've led the field of visual computing for decades.

A GIANT LEAP IN COMPUTER GRAPHICS

Real-time ray tracing has been the dream of computer scientists since it was first described nearly 40 years ago. NVIDIA RTX technology — a platform for real-time ray tracing — has brought it to life. Coupled with the Quadro® GV100, RTX is the greatest advance in computer graphics of the past 15 years, since our introduction of the programmable shader. NVIDIA has reinvented computer graphics, again.

CREATING AMAZING WORLDS

Today's blockbuster games are essentially big budget Hollywood films — artists work on titles for years, production value is a point of competition, and expectations are high. NVIDIA GPUs are the engines that make these games possible. NVIDIA GameWorks™ software allows developers to make games more photorealistic and immersive. And NVIDIA Ansel™ allows gamers to capture stunning in-game photography. Here, a scene from *Hellblade: Senua's Sacrifice*.

GEFORCE — THE WORLD'S LARGEST GAMING PLATFORM

Gaming is the world's largest entertainment industry. With 200 million gamers, NVIDIA GeForce® is its largest platform. GeForce GTX GPUs and the GeForce Experience™ application transform everyday PCs into powerful gaming machines.

G-SYNC

GEFORCE EXPERIENCE

MAX-Q DESIGN

GEFORCE GTX

HIGH-END PC GAMING FOR EVERYONE

One day, everyone will be a gamer. NVIDIA GeForce NOW™ is a cloud-based service that turns Macs and everyday PCs into virtual GeForce gaming machines, opening high-end gaming and blockbuster PC titles to millions of users for the first time.

NVIDIA IS REVOLUTIONIZING COMPUTING

In 2006, the creation of our CUDA® programming model and Tesla® GPU platform opened up the parallel-processing capabilities of the GPU to general-purpose computing.

A powerful new approach to computing was born.

POWERING THE WORLD'S FASTEST SUPERCOMPUTERS

GPU computing is the most accessible and energy-efficient path forward for HPC and datacenters. Today, NVIDIA powers the fastest supercomputers in the U.S. and Europe, as well as some of the most advanced systems under construction.

Japan's ABCI will come online in 2018 as the country's most powerful supercomputer and a global innovation platform for AI. And in the U.S., Lawrence Livermore and Oak Ridge National Labs will soon introduce next-generation NVIDIA-accelerated systems.

OAK RIDGE
National Laboratory

Lawrence Livermore
National Laboratory

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

AIST

東京工業大学
Tokyo Institute of Technology

ACCELERATING THE CODES OF SCIENTIFIC DISCOVERY

NVIDIA has been developing the CUDA programming model for over a decade. Today, it's the platform of choice for high-performance application developers, with support for more than 500 applications — including the top 15 HPC applications. From weather prediction and materials science to wind tunnel simulation and genomics, NVIDIA GPU computing is at the heart of HPC's most promising areas of discovery.
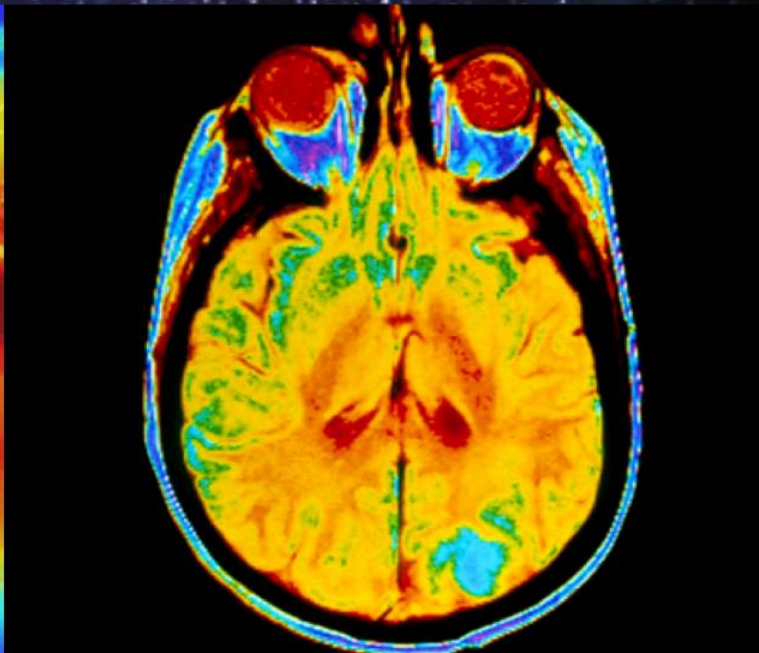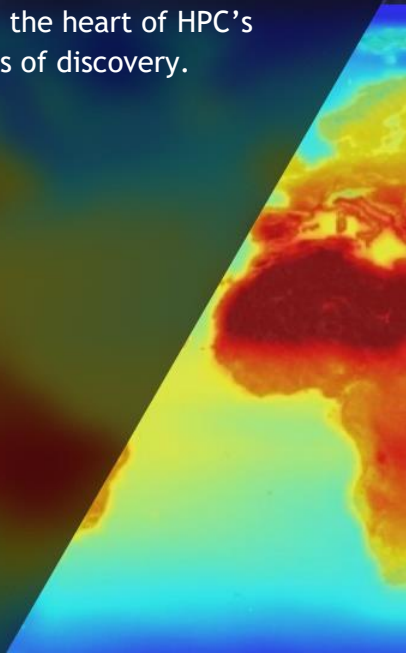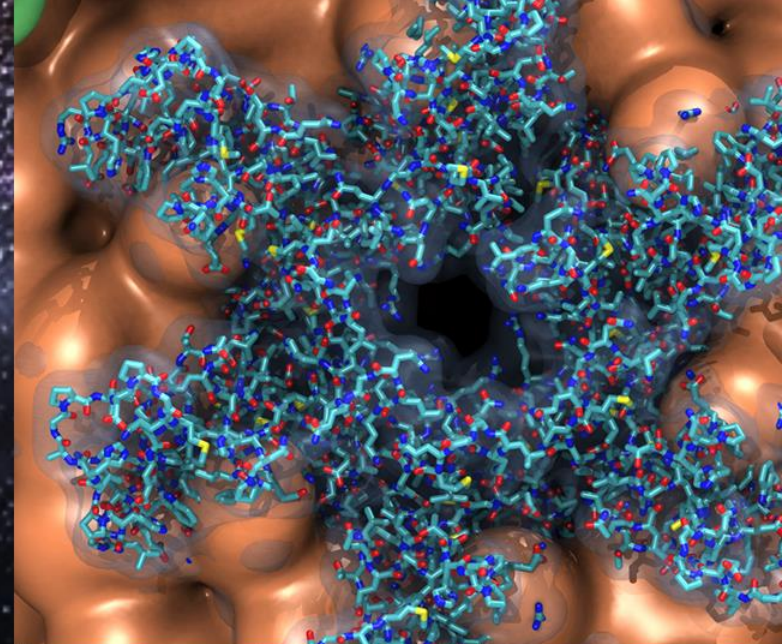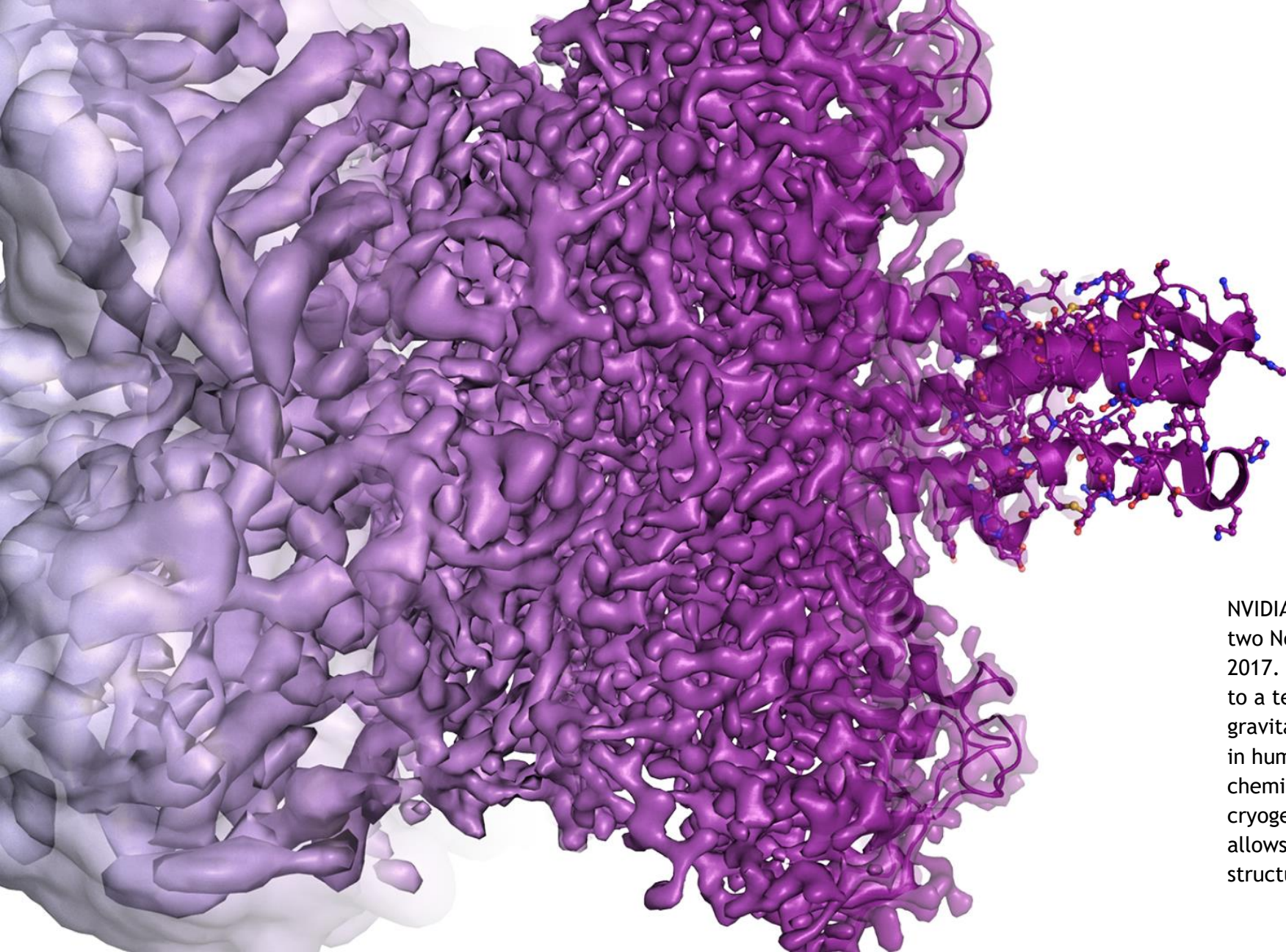
POWERING NOBEL PRIZE WINNERS

NVIDIA GPU computing played a role in two Nobel Prize-winning discoveries in 2017. The prize for physics was awarded to a team of scientists who detected gravitational waves for the first time in human history. And the prize for chemistry was awarded for work in cryogenic electron microscopy, which allows scientists to see detailed protein structures at the atomic level (*pictured*).

NVIDIA IGNITES
THE AI BIG BANG

Artificial intelligence is the use of computers to simulate human intelligence.

AI amplifies our cognitive abilities — letting us solve problems where the complexity is too great, the information is incomplete, or the details are too subtle and require expert training.

Learning from data — a computer's version of life experience — is how AI evolves. GPU computing powers the computation required for deep neural networks to learn to recognize patterns from massive amounts of data.

This new, supercharged mode of computing sparked the AI era.

THE ERA OF AI

The PC revolution put a computer in every home. The mobile era put a computer in every pocket. The cloud turned every mobile device into a supercomputer. The AI era will infuse intelligence into trillions of computing devices and be the single largest opportunity the industry has ever known. AI will spur a wave of social progress unmatched since the industrial revolution.

AI

CLOUD

MOBILE

PC

POWERING THE AI REVOLUTION

NVIDIA is advancing GPU computing for deep learning and AI at the speed of light. We create the entire stack. It starts with the most advanced GPUs and the systems and software we build on top of them. We integrate and optimize every deep learning framework. We work with the major systems companies and every major cloud service provider to make GPUs available in datacenters and in the cloud. And we create computers and software to bring AI to the edge, from self-driving cars to autonomous robots to medical devices.

ONE ARCHITECTURE

VOLTA EVERYWHERE

Volta, the world's most powerful GPU computing architecture, was built to drive the next wave of AI and HPC. Every major cloud service provider offers Volta instances, and every major computer maker offers Volta-based servers for on-premise datacenters. At GTC 2018, we supercharged the NVIDIA AI platform with the announcement of a "double-sized" 32GB Volta GPU.


EVERY CLOUD

Alibaba Cloud   aws   Baidu Cloud
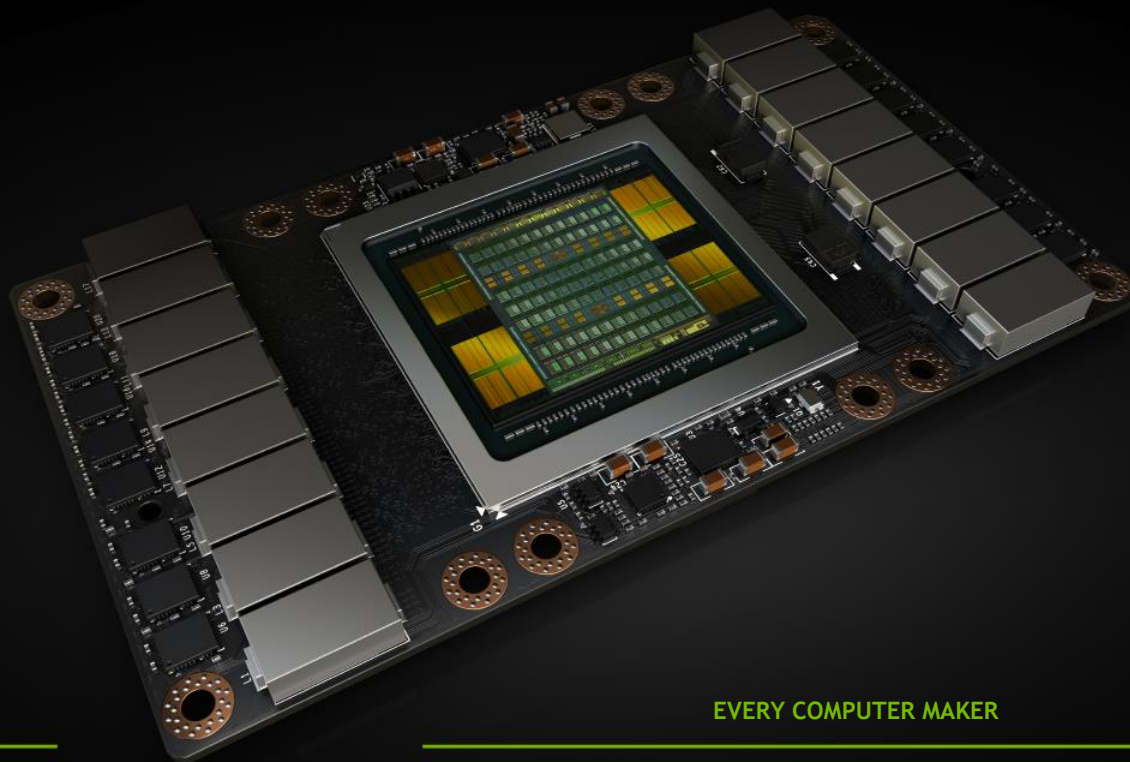
Google Cloud   Microsoft Azure   ORACLE   Tencent Cloud

EVERY COMPUTER MAKER

CISCO   CRAY   DELLEMC

Hewlett Packard Enterprise   IBM   Lenovo

THE LARGEST GPU IN THE WORLD

AI holds enormous promise, but it requires a massive amount of computing power. NVIDIA DGX-2™ is the first single server capable of delivering two petaflops of computational power — enough to replace 300 dual-CPU servers in today's hyperscale datacenters.

DGX-2 features NVSwitch™, a revolutionary GPU interconnect fabric that 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."
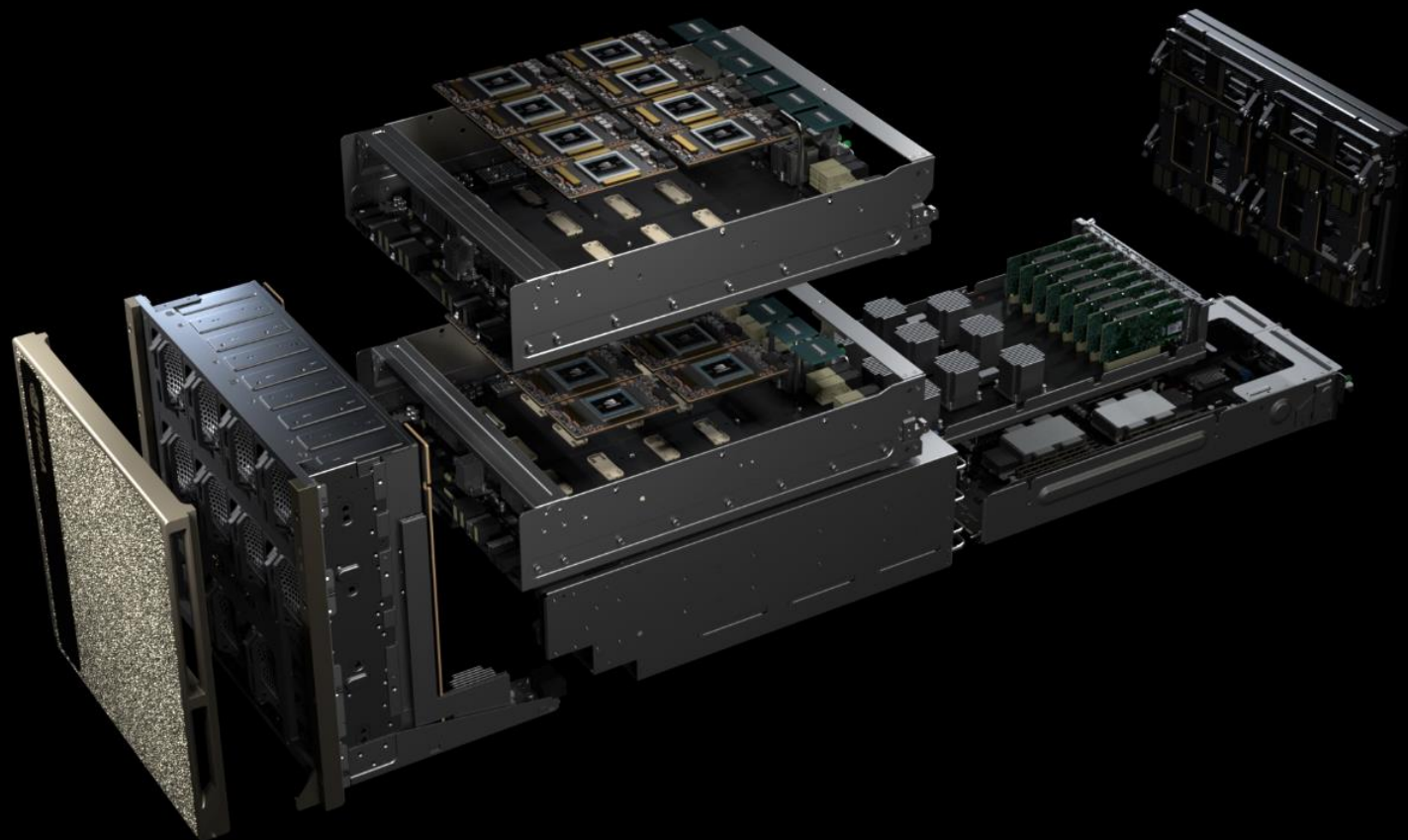
VOLTA
TAKES SATURNV
TO NEW HEIGHTS

NVIDIA's own SATURNV is one of the most powerful AI supercomputers in the world. It's also the fourth most energy efficient, based on the November 2017 Green500 list of supercomputers.

SATURNV is powered by 5,280 Volta GPUs, giving it a previously unimaginable FP16 performance of more than a half an exaflops on AI workloads.

A CAMBRIAN EXPLOSION OF AI

Since AlexNet, thousands of neural network 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 in the world. Maintaining them demands a balance of important factors: programmability, latency, accuracy, size, throughput, energy efficiency, and rate of learning.

CONVOLUTIONAL NETWORKS
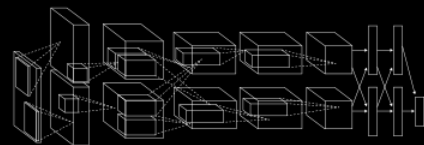
Encoder/Decoder · ReLu · BatchNorm
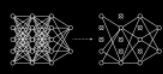
Concat · Dropout · Pooling

RECURRENT NETWORKS
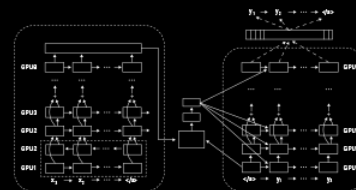
LSTM · GRU · Beam Search

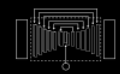WaveNet · CTC · Attention

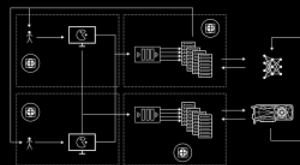## GENERATIVE ADVERSARIAL NETWORKS

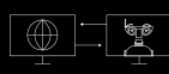3D-GAN · MedGAN · ConditionalGAN

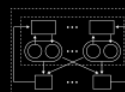Coupled GAN · Speech Enhancement GAN
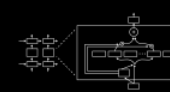
REINFORCEMENT LEARNING
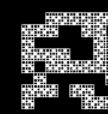
DQN · Simulation · DDPG

NEW SPECIES

Capsule Nets

Mixture of Experts · Neural Collaborative Filtering · Block Sparse LSTM

TRILLIONS OF INTELLIGENT THINGS

The programmable NVIDIA platform is designed for the complex universe of AI development and deployment. TensorRT™ 4, the latest version of our inference software, is integrated into Google's popular TensorFlow framework. Kaldi, the most popular framework for speech recognition, is now optimized for GPUs. And Kubernetes on NVIDIA GPUs allows orchestration of resources across multi-cloud GPU clusters. Hyperscale datacenters can save big money with NVIDIA inference acceleration.

NVIDIA INFERENCE ACCELERATION
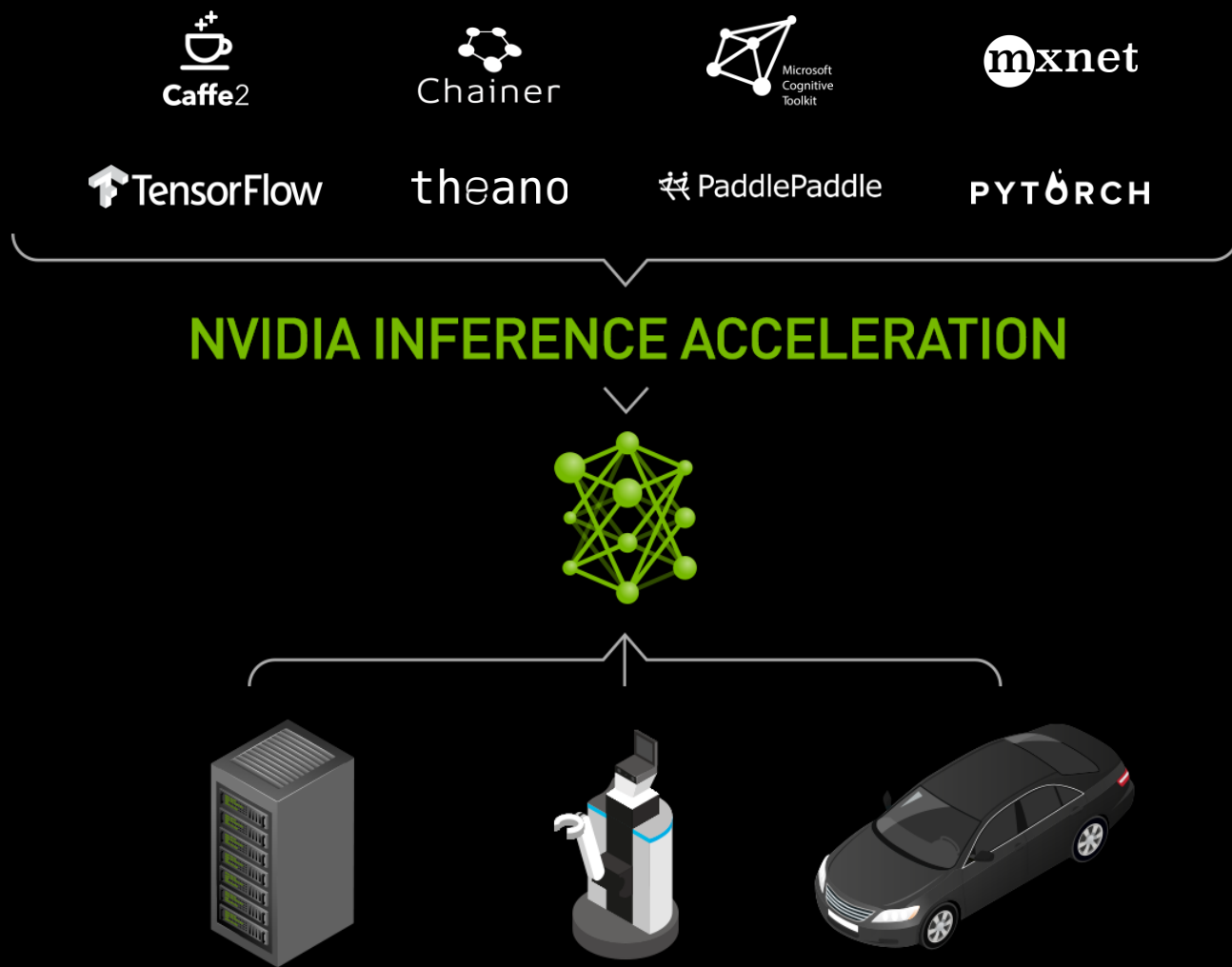
NVIDIA GPU CLOUD — ONE PLATFORM, RUN EVERYWHERE

The NVIDIA GPU Cloud registry gives developers access to GPU-optimized software stacks wherever they want it — on PCs, in the datacenter, or via the cloud. There are some 30 GPU-optimized containers for deep learning, HPC, HPC visualization, and analytics.

AI IS REVOLUTIONIZING EVERY INDUSTRY

The AI race is on. Deep learning breakthroughs no longer come from scientific and research labs alone. Today, in trillion-dollar industries like transportation, healthcare, and manufacturing, companies are using AI to transform the ways they do business. Self-driving cars, intelligent medical imaging systems, and autonomous factory robots have moved quickly from ideas to reality. And it's only the beginning.

REVOLUTIONIZING TRANSPORTATION

Transportation is a $10 trillion industry. Autonomous vehicles will change it forever, making our roads safer and our cities more efficient. More than 370 companies are using NVIDIA technology in their datacenters and vehicles. They range from car companies and suppliers, to mapping and sensor companies, to startups and research organizations.

NVIDIA DRIVE — FROM TRAINING TO SAFETY

Building an autonomous car is an extraordinary endeavor. To train the network, data from all over the world, covering every road condition, needs to be collected and labeled and fed into a DGX supercomputer. Simulation is used to expand the training set as well as cover dangerous or extreme scenarios that can't be captured on the road. The trained model is deployed on an in-car supercomputer, where it can tap into a sophisticated software stack covering everything from pedestrian detection to driver monitoring. At every step of the way, the most stringent standards are applied to ensure that safety is paramount.

1. COLLECT & PROCESS DATA

2. TRAIN MODELS

Cars · Pedestrians · Path
Lanes · Signs · Lights

3. SIMULATE

4. DRIVE

Cars · Pedestrians · Path
Lanes · Signs · Lights

SIMULATION — THE ROAD TO SAFE SELF-DRIVING CARS

Each year, 10 trillion miles are driven around the world. Test cars can eventually cover millions of miles, but we'll need billions to create safe and reliable self-driving cars.

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.

REVOLUTIONIZING HEALTHCARE

AI is transforming the spectrum of healthcare, from detection to diagnosis to treatment. The NVIDIA AI platform is the driving force. GE Healthcare has reinvented the CT, doubling image processing speeds by embedding GPU-powered AI in its new Revolution Frontier CT scanner. Nuance is helping radiologists use AI to speed their analysis of medical imaging by making pre-trained algorithms and vast imaging datasets available to them directly via its AI Marketplace.
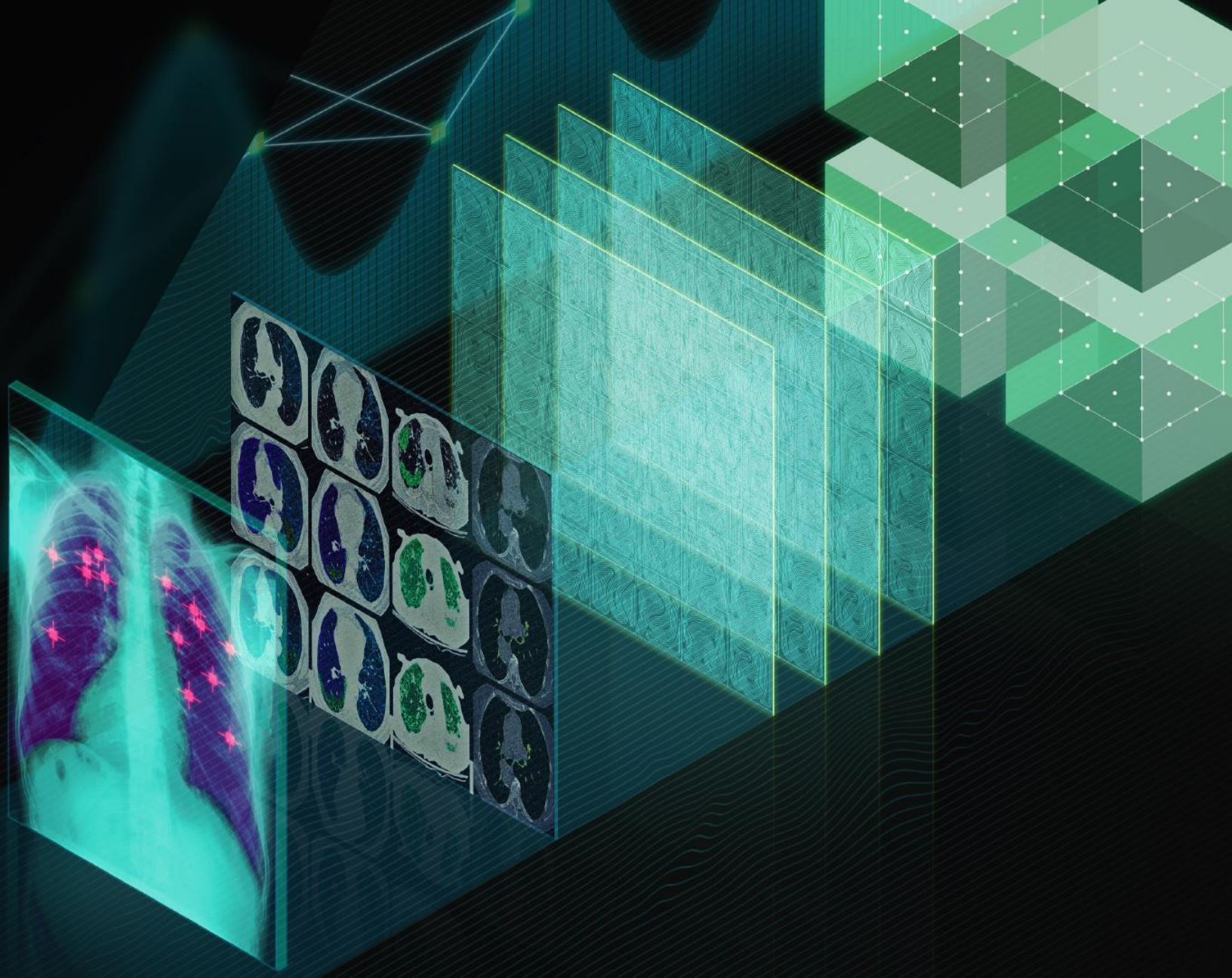
PROJECT CLARA — A MEDICAL IMAGING SUPERCOMPUTER

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. Project Clara, NVIDIA's medical imaging supercomputer in the cloud, can do just that. With Clara, existing instruments will be supercharged with state-of-the-art image reconstruction, object detection and segmentation, and visualization capabilities.
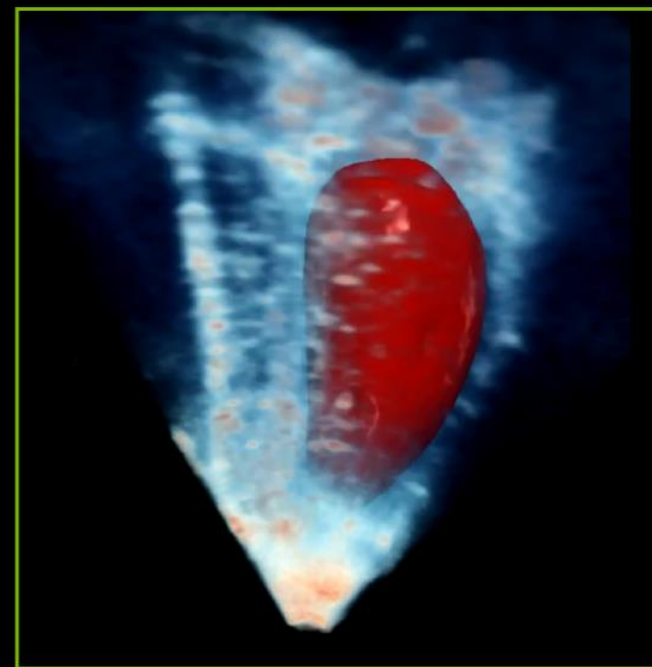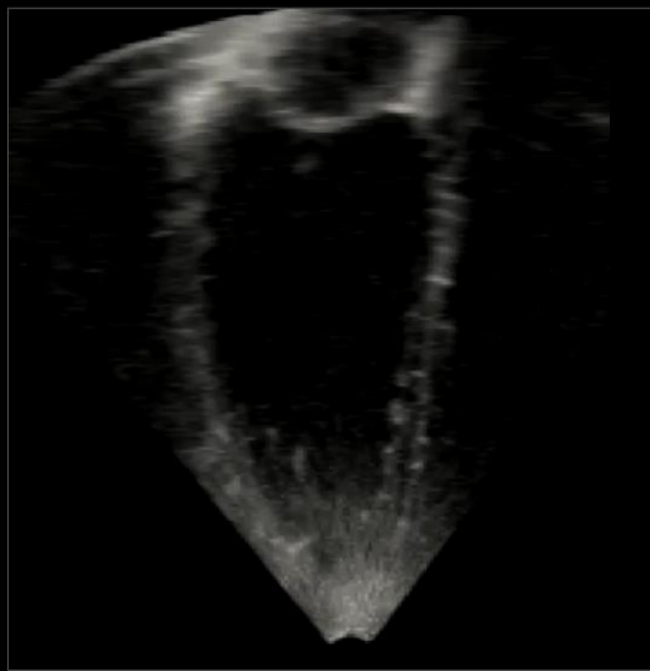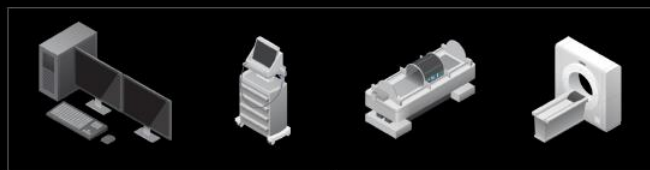

PROJECT CLARA

IMAGING AND VISUALIZATION APPS
CUDA | CUDNN | TENSORRT | OGL | RTX
GPU CONTAINERS | VGPU
NVIDIA GPU SERVER

REVOLUTIONIZING MANUFACTURING AND LOGISTICS

Deep learning and affordable sensors have created the conditions for "the automation of automation." NVIDIA Jetson™ TX2 delivers 1 teraflops of performance in a credit card-sized module. Such power will enable a new wave of manufacturing, drones that can inspect hazardous places, and robots that can deliver the millions of packages shipped every day.
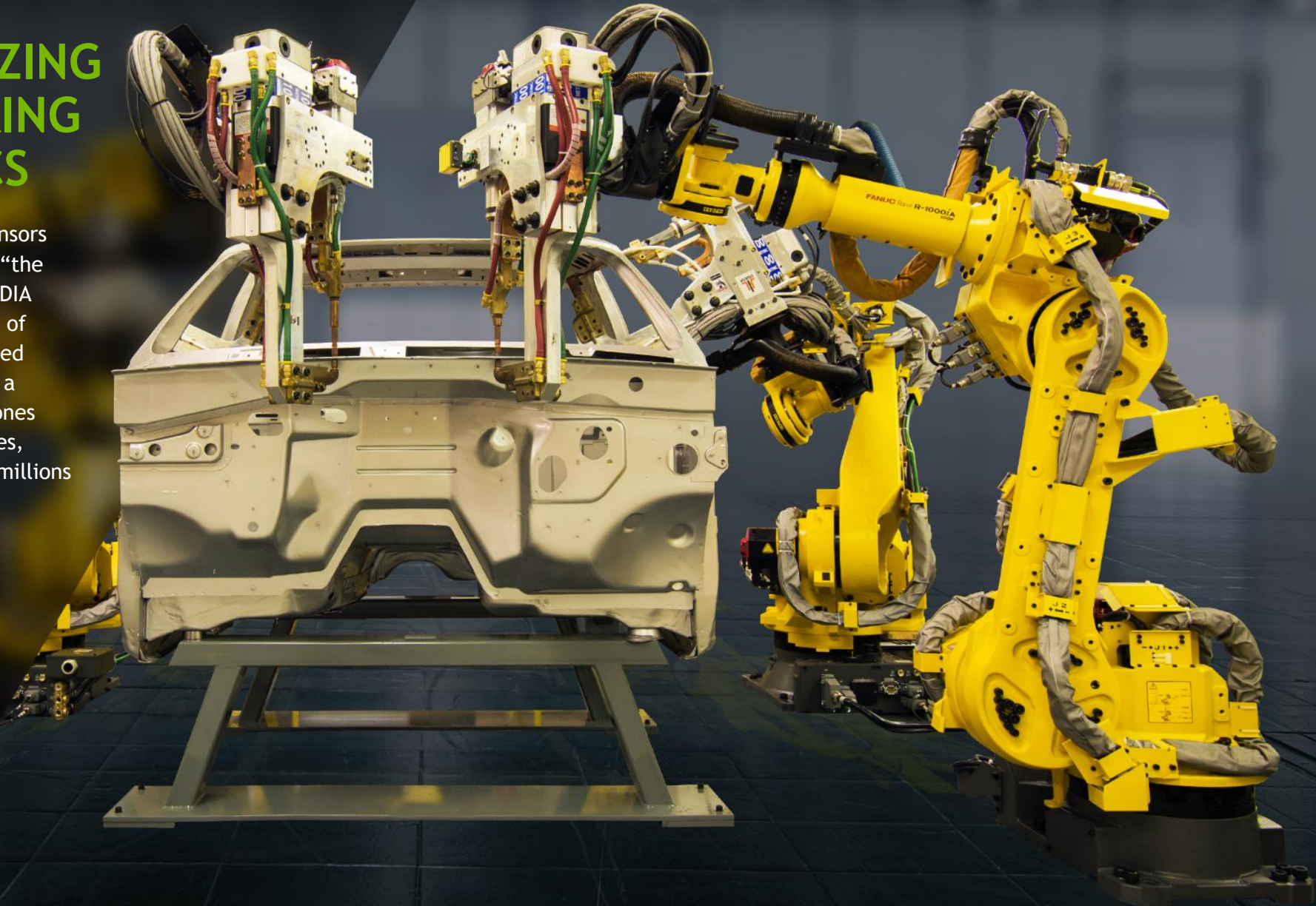
NVIDIA ISAAC — WHERE ROBOTS GO TO LEARN

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. Isaac Sim is 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.
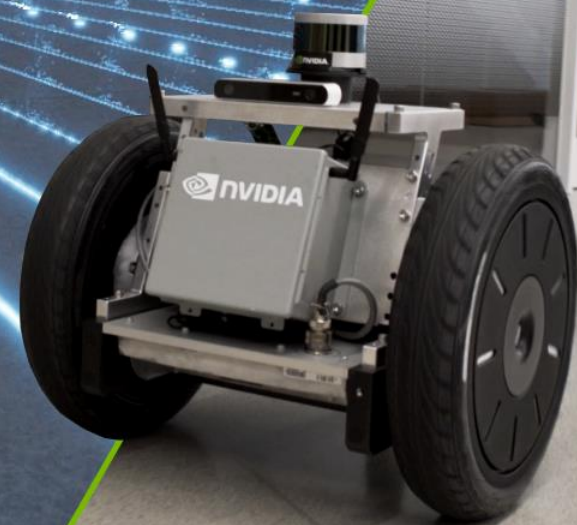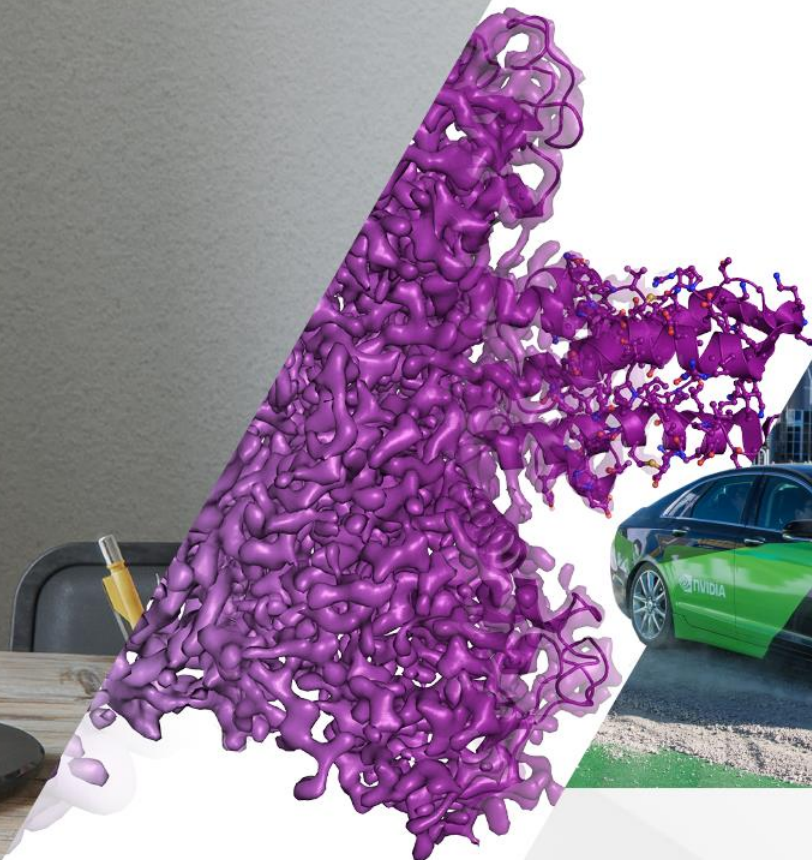
NVIDIA —
A LEARNING
MACHINE

NVIDIA has continuously reinvented itself over two decades.

Our invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics, and revolutionized parallel computing. More recently, GPU computing ignited the era of AI.

NVIDIA is a "learning machine" that constantly evolves by adapting to new opportunities that are hard to solve, that only we can tackle, and that matter to the world.

Nobel Prize Winning Cryo-EM

AI

HPC

GRAPHICS

OPERATING AT THE SPEED OF LIGHT

NVIDIA is united by a unique culture — the operating system of our learning machine.

We dream big. We take risks. We learn from our mistakes together. Speed and agility are the keys to our success. Craftsmanship is a discipline and passion. There are no org charts — the project is the boss.

These beliefs inform everything we do, from designing amazing products to building one of the world's great companies — a place where people can to do their life's work.

INSPIRED TO GIVE TO OUR COMMUNITIES

NVIDIA's people share a strong sense of corporate responsibility. Our philanthropic giving exceeded $6 million in 2017. To date, our NVIDIA Foundation's Compute the Cure initiative has directed more than $4 million to the fight against cancer. And our Techsplorer program, which introduces underserved youth to AI, has reached more than 5,800 students since it launched in 2017.

FORTUNE

DECEMBER 1, 2017   FORTUNE.COM
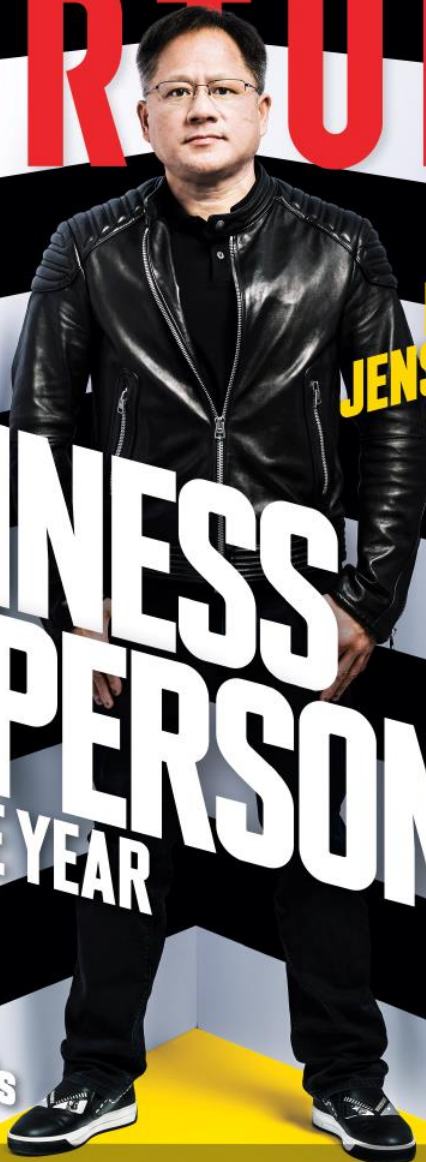
NVIDIA CEO
JENSEN HUANG
HOW HIS COMPANY IS POWERING THE AGE OF A.I.

BUSINESS PERSON
+ OF THE YEAR

19 MORE OF THE HOTTEST CEOs OF 2017

NVIDIA

"World's Best Performing CEOs"
— Harvard Business Review

"World's Most Admired Companies"
— Fortune

"World's Best CEOs"
— Barron's

"Most Innovative Companies"
— Fast Company

"Employees' Choice: Highest Rated CEOs"
— Glassdoor

"50 Smartest Companies"
— MIT Tech Review

Founded in 1993 | Jensen Huang, Founder & CEO | 12,000 employees | $9.7B in FY18