NVIDIA

@joshenders
Traffic engineering
@Pinterest,
Infrastructure
@comma_ai

Home Moments 99+ Notifications Messages Search Twitter Tweet

Josh Ende
@joshenders

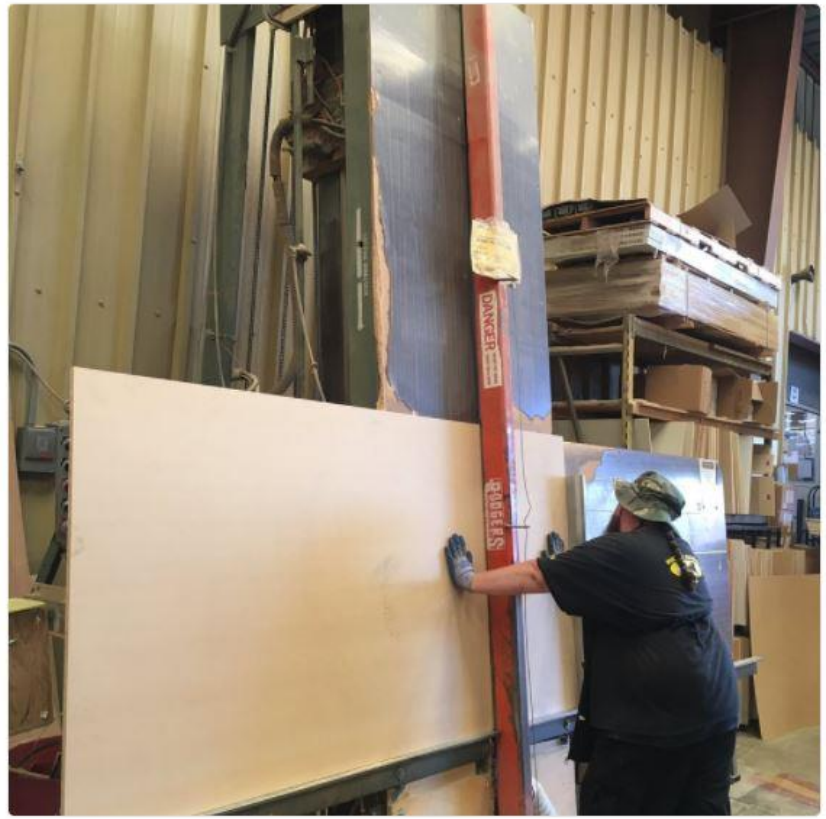
Traffic engineeri
Infrastructure @
@CloudFlare, @
@CalArts. Hacke
Human · ENFP

San Francisc
joshenders.c
Joined March

Josh Enders
@joshenders

Follow

Cutting some sturdy shelving for Comma's GPU cluster of tower servers



2:03 PM - 18 Jun 2017 from Discount Builders Supply

1 Retweet **17** Likes

2 1 17

A NEW ERA OF COMPUTING

PC INTERNET
WinTel, Yahoo!
1 billion PC users

MOBILE-CLOUD
iPhone, Amazon AWS
2.5 billion mobile users

AI & IOT
Deep Learning, GPU
100s of billions of devices

1995 2005 2015

HPC OR AI?

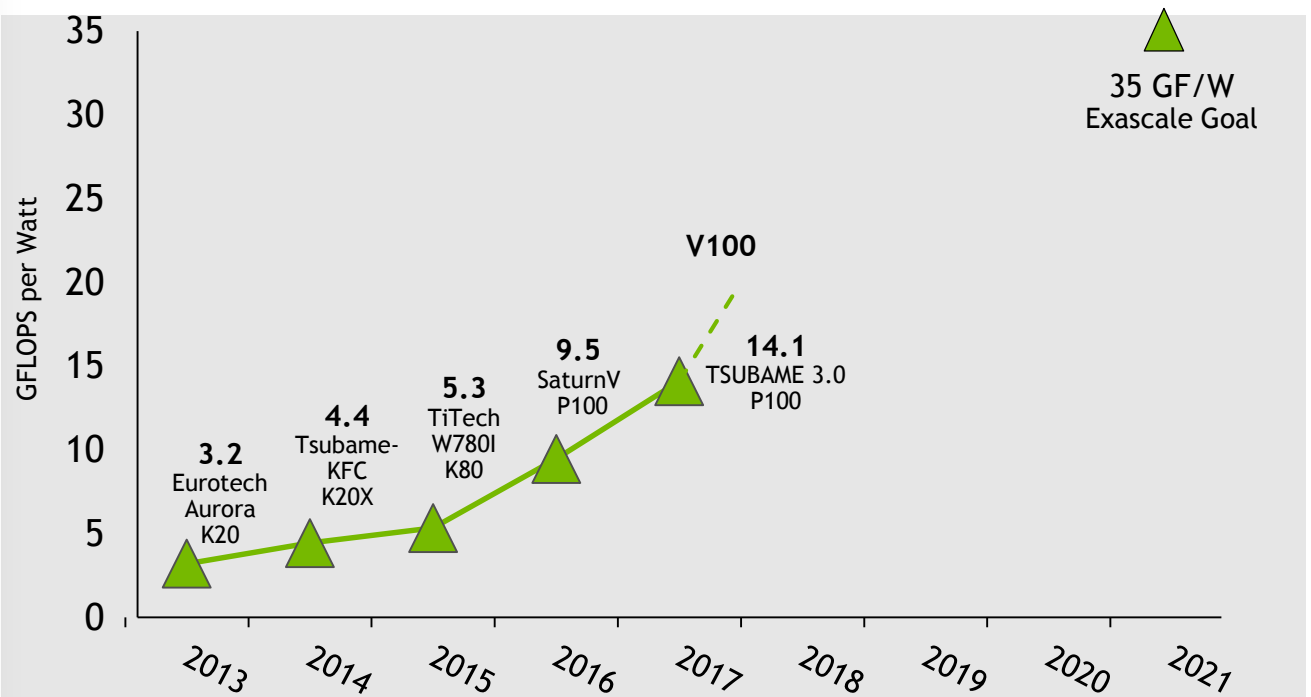
31	Facebook United States	NVIDIA DGX-1/Penguin Relion 2904GT, Xeon E5-2698v4 20C 2.2GHz/ E5-2650v4, Mellanox Infiniband EDR, NVIDIA Tesla P100 Facebook	60,512	3,307.0	4,896.5	350
32	NVIDIA Corporation United States	**DGX Saturn V** - NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 Nvidia	60,512	3,307.0	4,896.5	350
425	University of Oxford United Kingdom	**JADE** - NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 Bull, Atos Group	10,736	492.2	868.7	58

CAPEX OR OPEX?

13/13 Greenest Supercomputers
Powered by Tesla P100

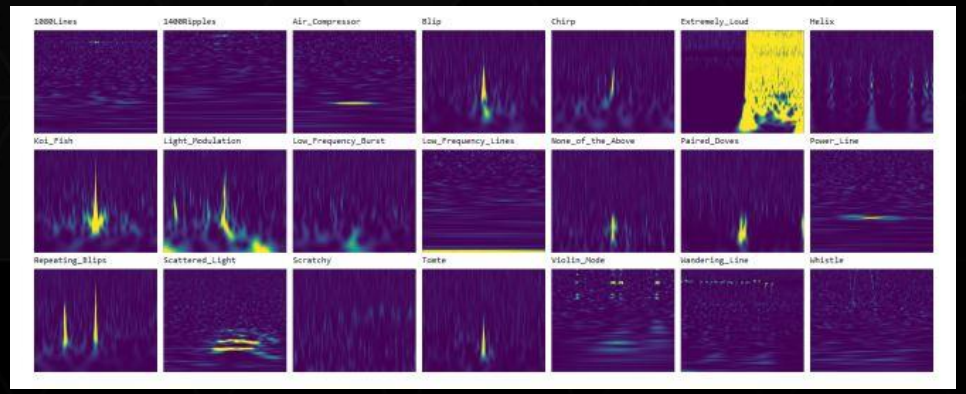
TSUBAME 3.0
Kukai
AIST AI Cloud
RAIDEN GPU subsystem
Piz Daint
Wilkes-2
GOSAT-2 (RCF2)
DGX Saturn V
Reedbush-H
JADE
Facebook Cluster
Cedar
DAVIDE

On Track To Meet Exascale Goal

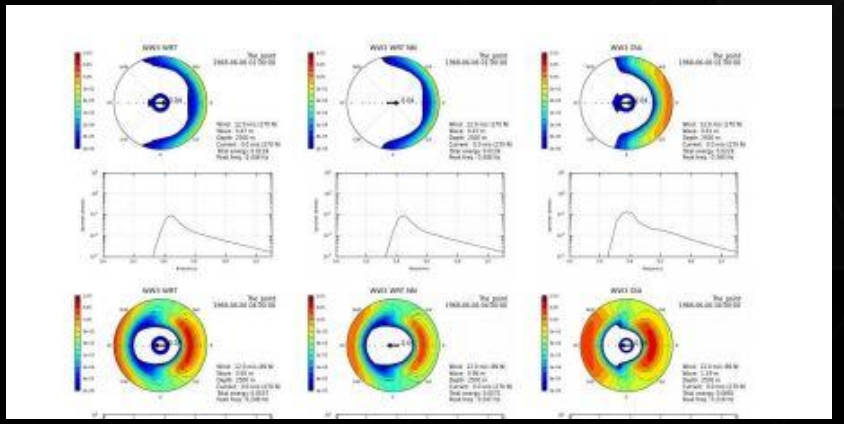


Top GPU Systems in Green500 Lists and NVIDIA Projections for V100

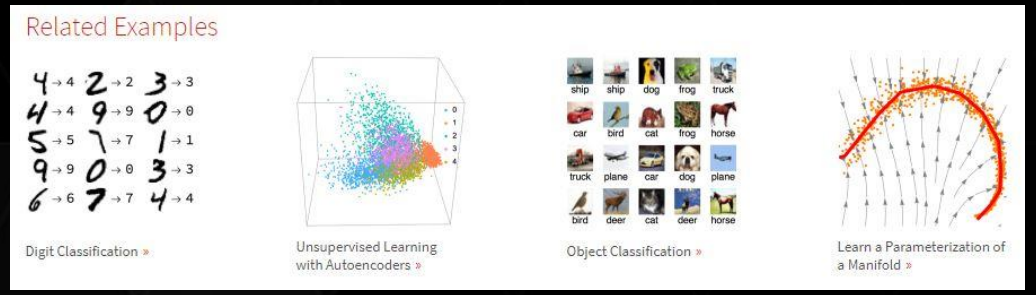
AI POWERING SCIENCE DISCOVERIES



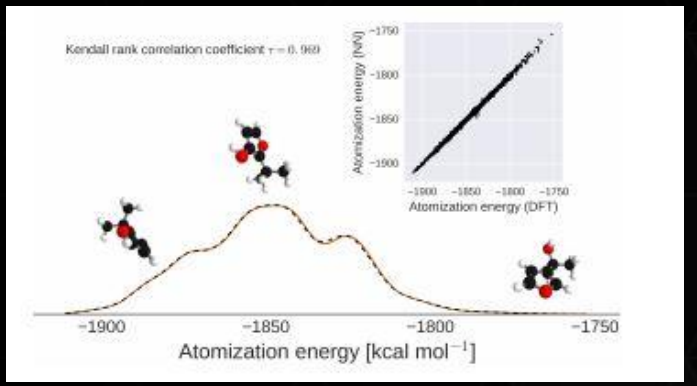
LIGO Deep Transfer Learning for Glitch Classification (NCSA)
https://arxiv.org/abs/1706.07446



Neural Networks for Operational Ocean Wave Prediction Models
http://www.sciencedirect.com/science/article/pii/S1877050914002816



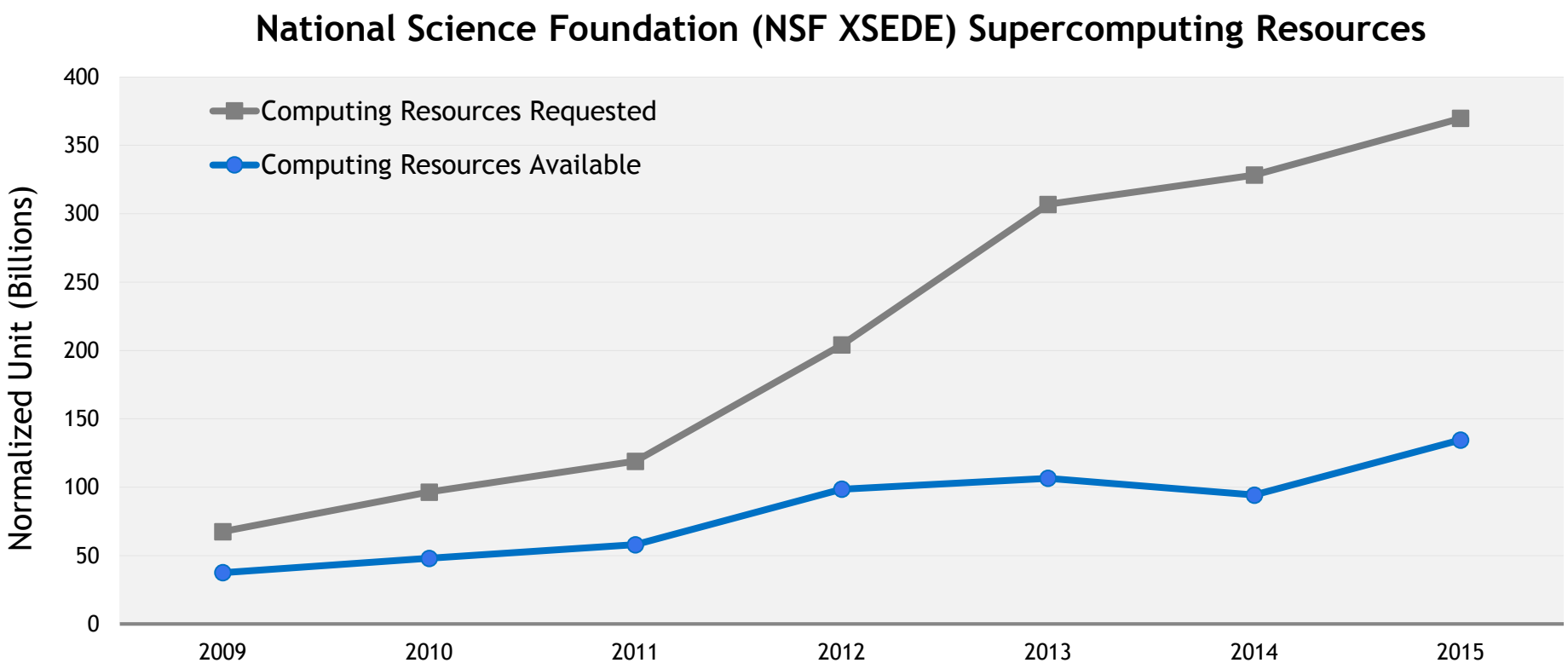
Wolfram 11 Neural Network library with NVIDIA cuDNN & GPUs
https://www.wolfram.com/language/11/neural-networks/



Quantum Chemical Insights from Deep Tensor Neural Networks
https://arxiv.org/pdf/1609.08259.pdf

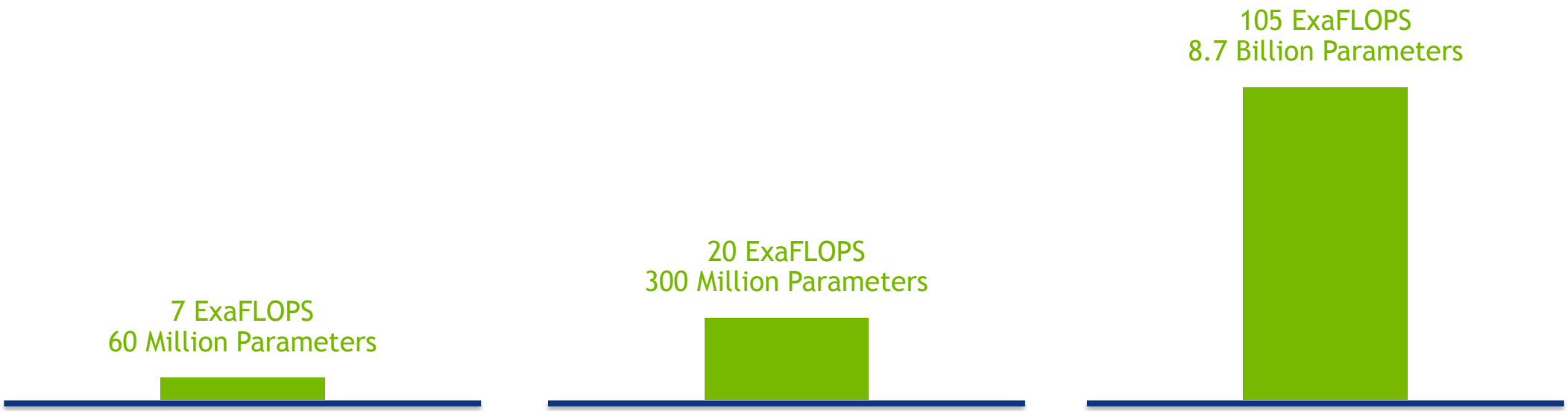
200B CORE HOURS OF LOST SCIENCE

Data Center Throughput is the Most Important Thing for HPC



Source: NSF XSEDE Data: https://portal.xsede.org/#/gallery
NU = Normalized Computing Units are used to compare compute resources across supercomputers and are based on the result of the High Performance LINPACK benchmark run on each system

AI MODEL COMPLEXITY IS EXPLODING



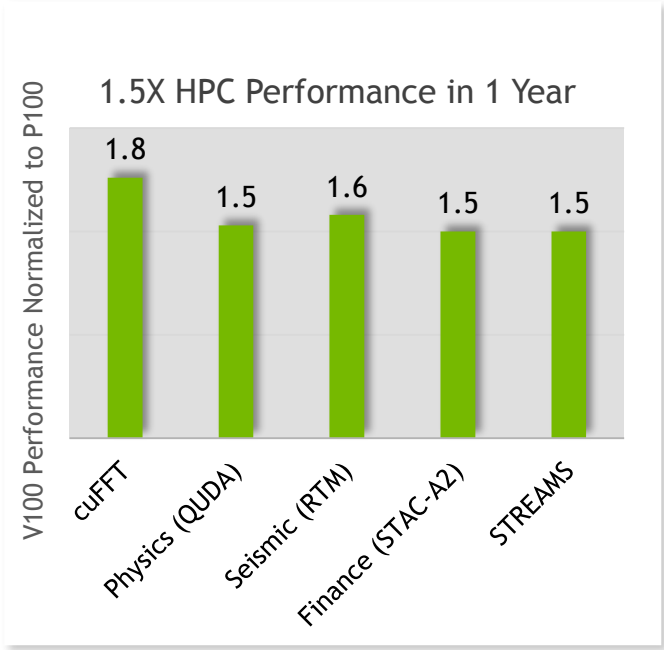
2015 — Microsoft ResNet

2016 — Baidu Deep Speech 2

2017 — Google NMT

ROAD TO EXASCALE

Volta to Fuel Most Powerful US Supercomputers

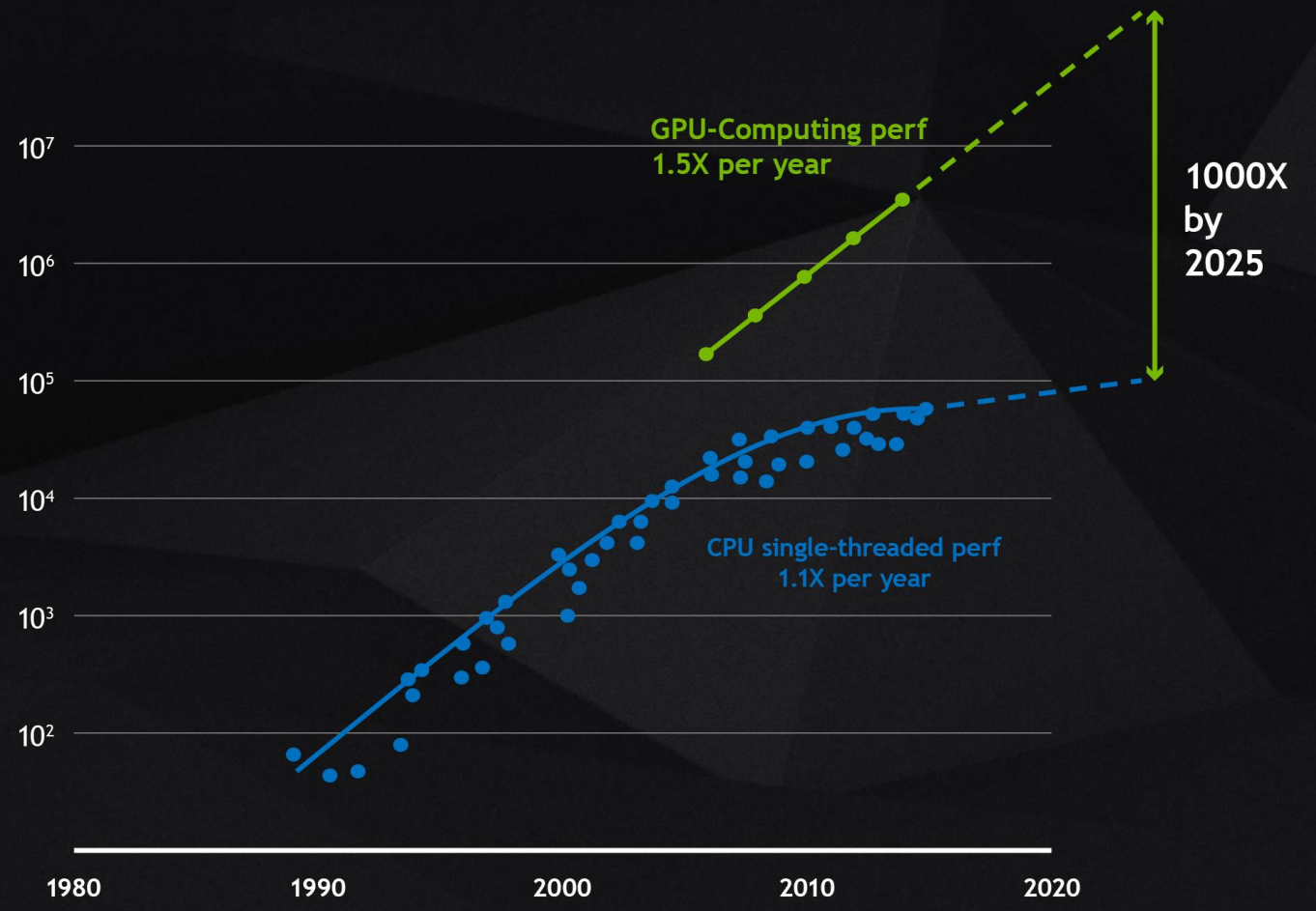


System Config Info: 2X Xeon E5-2690 v4, 2.6GHz, w/ 1X Tesla P100 or V100. V100 measured on pre-production hardware.

THE TIME FOR GPU COMPUTING HAS COME

The End of Road for General Purpose Processors and the Future of Computing

John Hennessy
Stanford University
March 2017



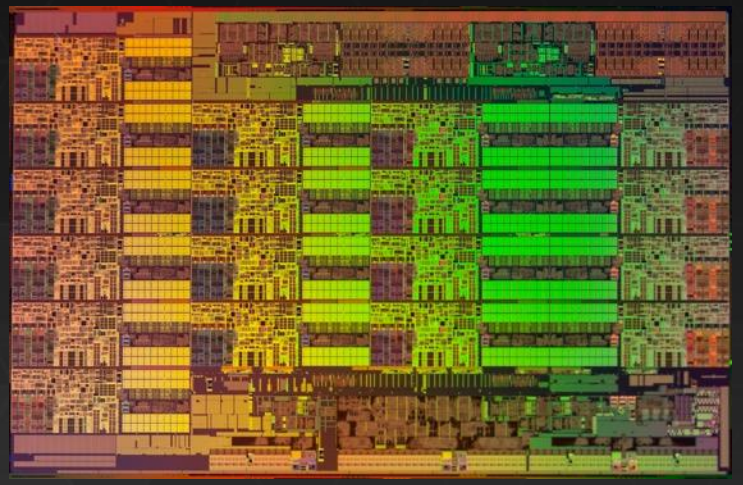
40 Years of Microprocessor Trend Data

Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

CPU
126 pJ/flop (SP)

Optimized for Latency

Deep Cache Hierarchy

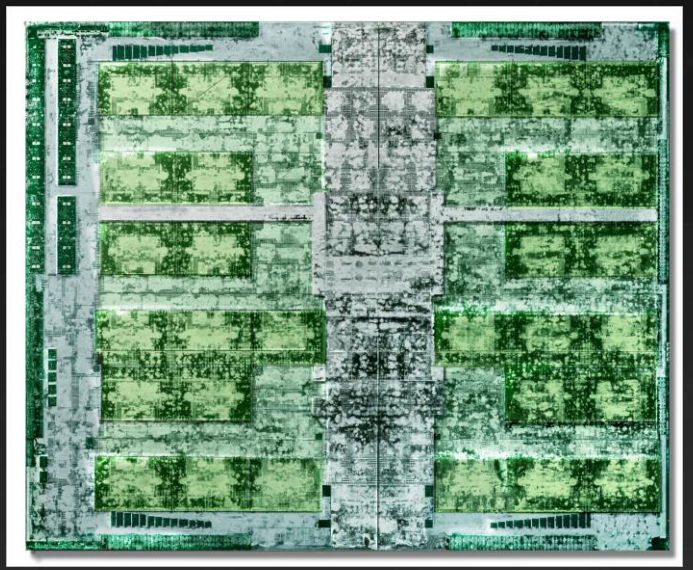


Broadwell E5 v4
14 nm

GPU
28 pJ/flop (SP)

Optimized for Throughput

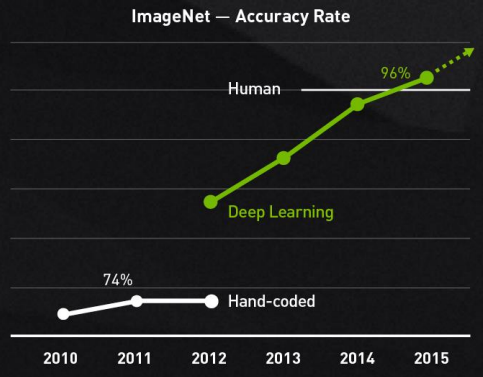
Explicit Management
of On-chip Memory



Pascal
16 nm

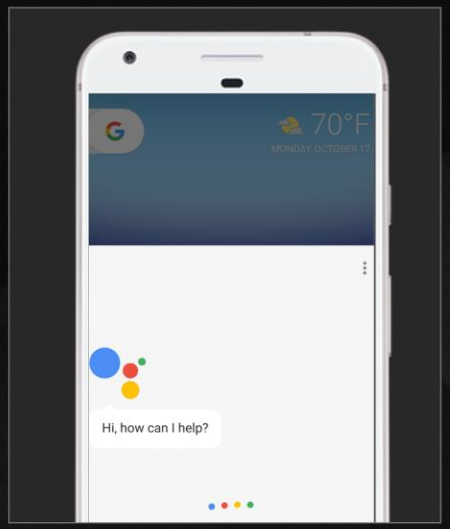
AI ACHIEVES “SUPERHUMAN” RESULTS

The big bang of modern AI set off a string of “superhuman” achievements. In 2015, Google and Microsoft both beat the best human score in the ImageNet challenge. DeepMind’s AlphaGo recorded its historic win over Go champion Lee Sedol in 2016 and, more recently, beat the best player in the world, Ke Jie. Breakthroughs in AI happen almost every day.

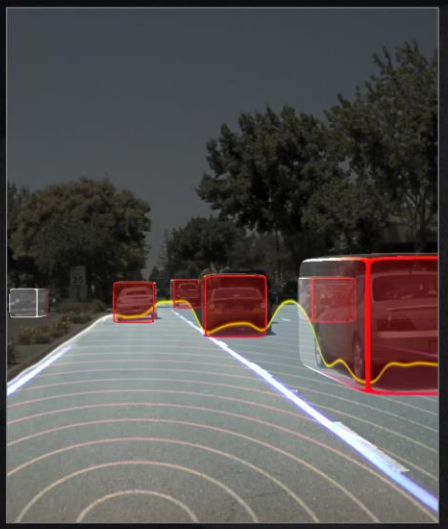


THE ERA OF AI

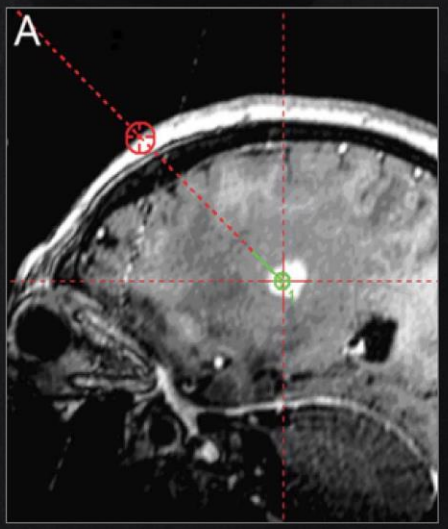
With deep learning, we can teach AI to do almost anything. New internet services, like Google Assistant, have learned speech from sound and provide a more natural way to access information. Self-driving cars use deep learning to recognize the space the car inhabits, the lanes in which it drives, and the objects it must avoid. In healthcare, neural networks trained with millions of medical images can find clues in MRIs that until now could only be found through invasive biopsies. These are just a few examples. AI will spur a wave of social progress unmatched since the industrial revolution.



"Ok Google"



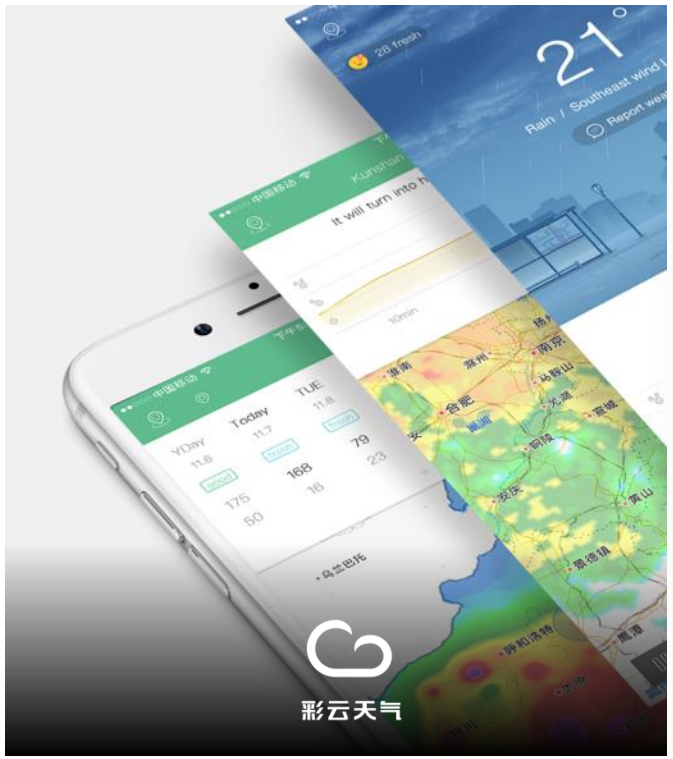
Self-Driving Cars



Medical Breakthroughs

DEEP LEARNING IS VITAL TO HPC

Monitoring Effects of Carbon and Greenhouse Gas Emissions



Minute-by-minute AI Weather Forecasting

92%

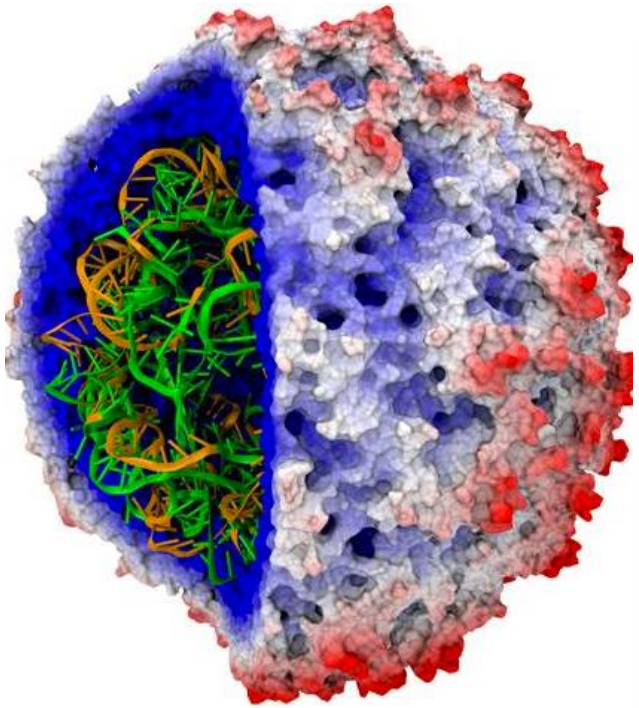
believe AI will impact their work

93%

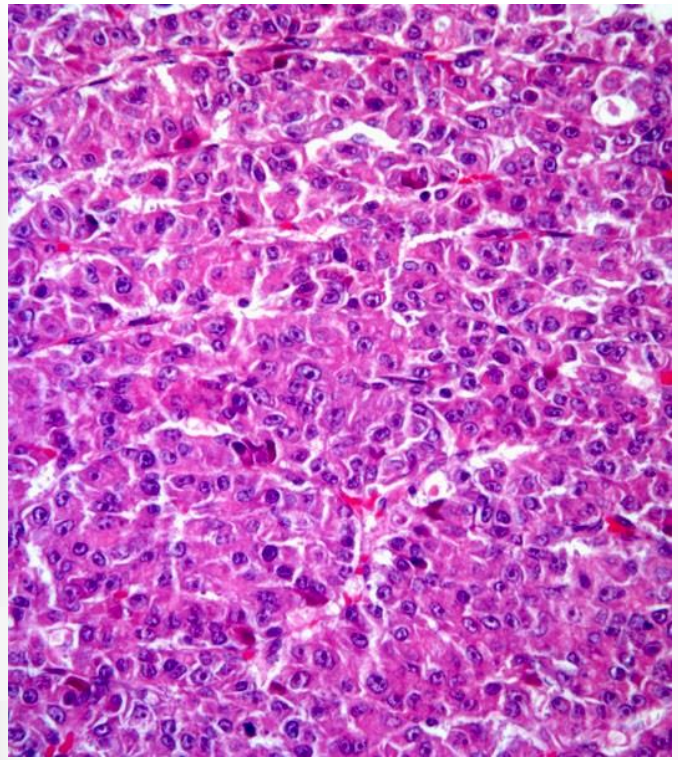
using deep learning seeing positive results

insideHPC.com Survey November 2016

DEEP LEARNING & SIMULATION FOR EXASCALE CANCER RESEARCH



Accelerate Discovery of Cancer Therapies



Predict Drug Response of Cancer Patients

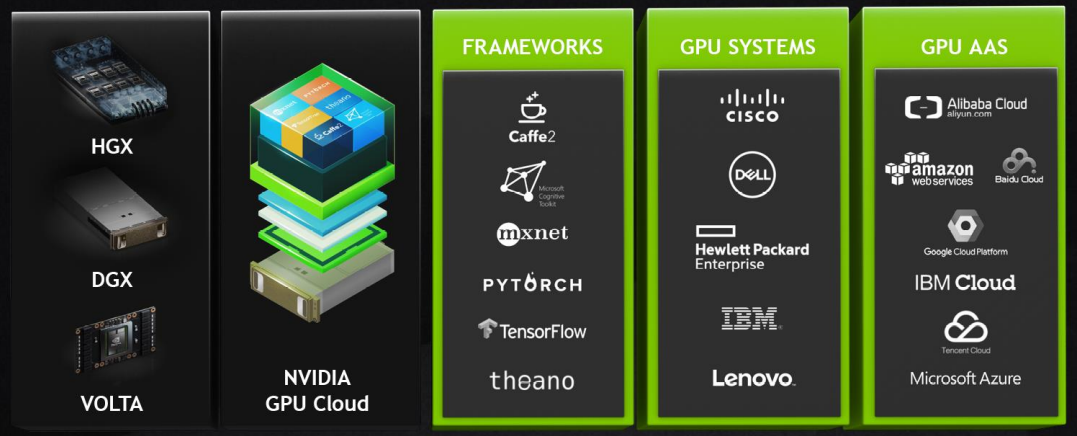
Automate Analysis of Treatment Effectiveness

June 2016 NCI Genomic Data Commons = 3PB Data

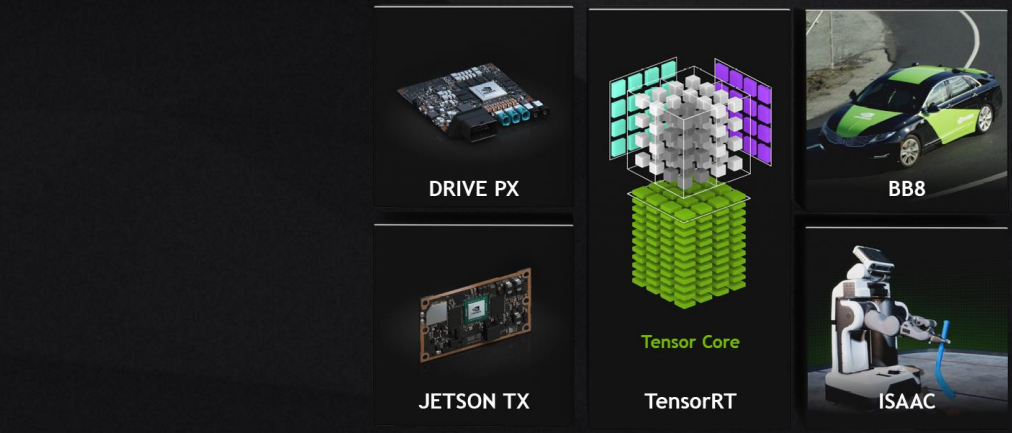
POWERING THE AI REVOLUTION

GPU computing is the most productive and pervasive platform for deep learning and AI. It begins 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 data centers and in the cloud. And we create computers and software to bring AI to edge devices, such as self-driving cars and autonomous robots.

AI in the Data Center and Cloud

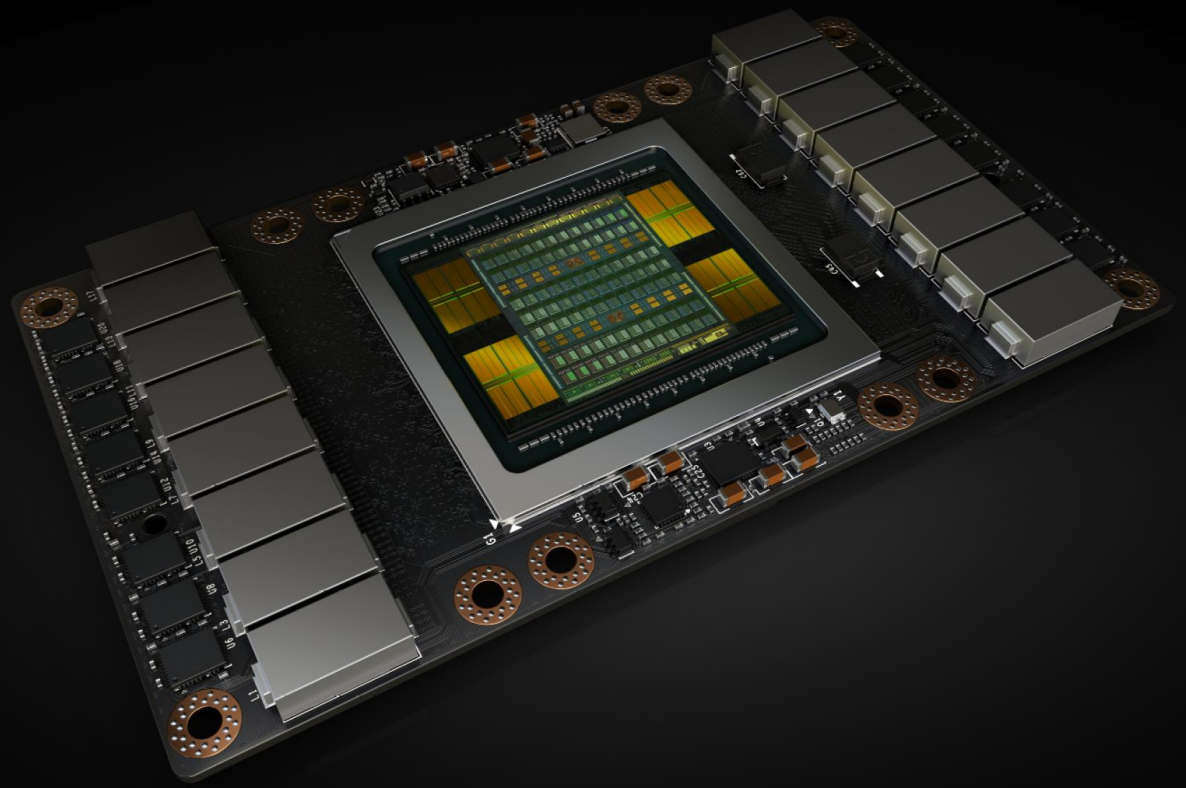


AI at the Edge



VOLTA — A GIANT LEAP FOR AI AND HPC

Volta, the world's most powerful GPU computing architecture, was built to drive the next wave of AI and HPC. Tesla V100, based on Volta, is the world's first 120 TeraFLOPS processor and brings extraordinary speed and scalability for AI inferencing and training, as well as for accelerating HPC and graphics workloads. Volta is the biggest chip on the planet — boasting 21 billion transistors and able to deliver the equivalent performance of 250 CPUs for deep learning.



TESLA V100

21B transistors
815 mm²

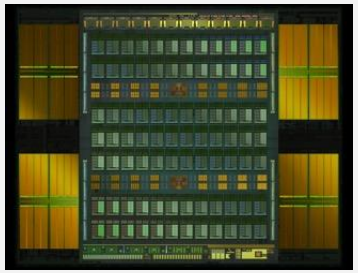
80 SM
5120 CUDA Cores
640 Tensor Cores

16 GB HBM2
900 GB/s HBM2
300 GB/s NVLink

*full GV100 chip contains 84 SMs

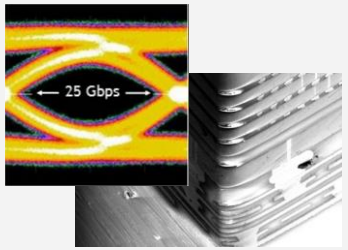
INTRODUCING TESLA V100

Volta Architecture



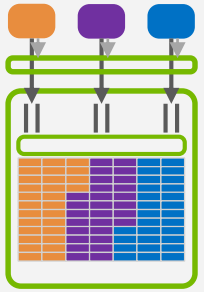
Most Productive GPU

Improved NVLink & HBM2



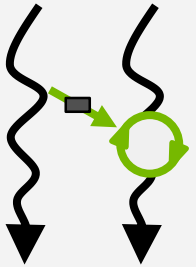
Efficient Bandwidth

Volta MPS



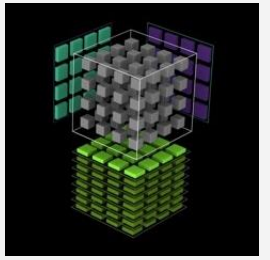
Inference Utilization

Improved SIMT Model



New Algorithms

Tensor Core



120 Programmable TFLOPS Deep Learning

The Fastest and Most Productive GPU for Deep Learning and HPC

NEW TENSOR CORE

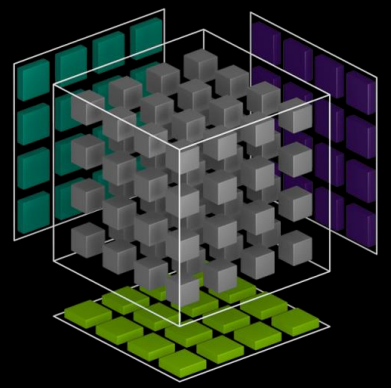
New CUDA TensorOp instructions & data formats

4x4 matrix processing array

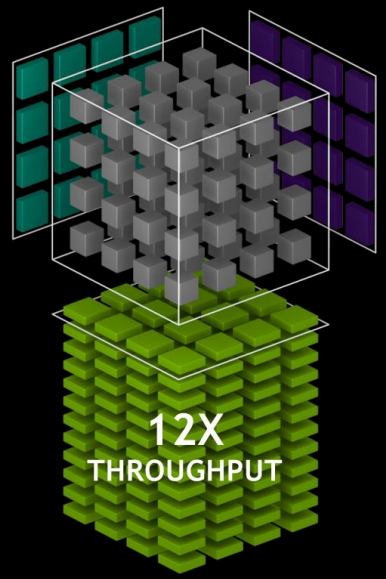
D[FP32] = A[FP16] * B[FP16] + C[FP32]

Optimized for deep learning

PASCAL



VOLTA TENSOR CORES



Activation Inputs Weights Inputs Output Results

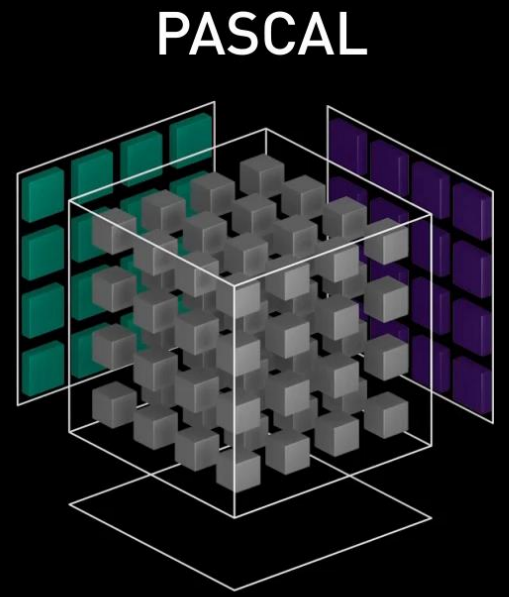
NEW TENSOR CORE

New CUDA TensorOp instructions
& data formats

4x4 matrix processing array

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

Optimized for deep learning



Activation Inputs Weights Inputs Output Results

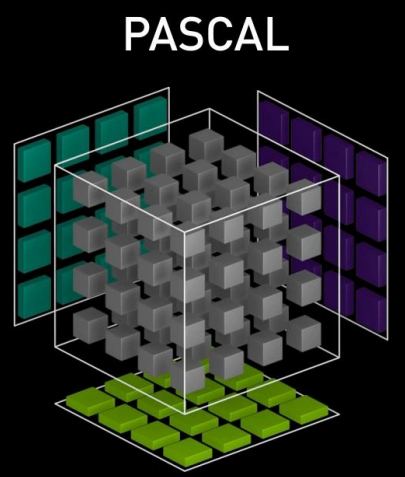
NEW TENSOR CORE

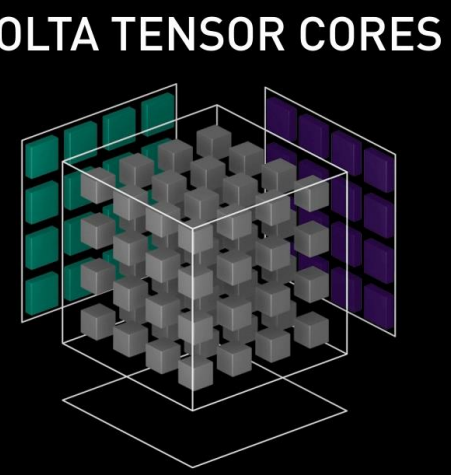
New CUDA TensorOp instructions & data formats

4x4 matrix processing array

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

Optimized for deep learning

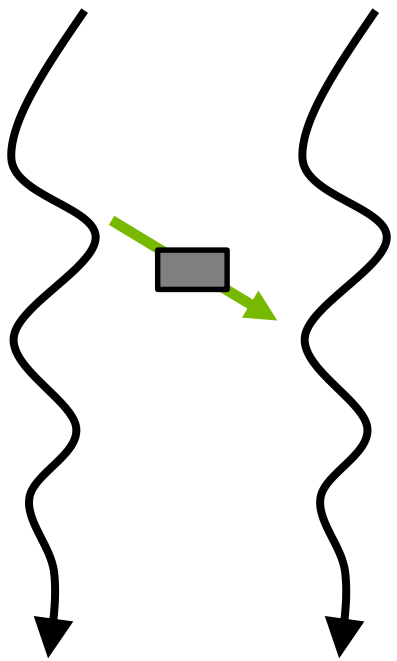




Activation Inputs Weights Inputs Output Results

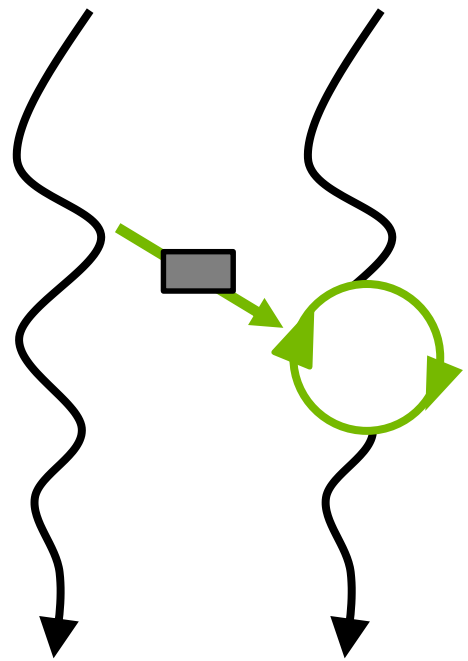
VOLTA: INDEPENDENT THREAD SCHEDULING

Communicating Algorithms



Pascal: Lock-Free Algorithms

Threads cannot wait for messages



Volta: Starvation Free Algorithms

Threads may wait for messages

STARVATION FREE ALGORITHM

Example

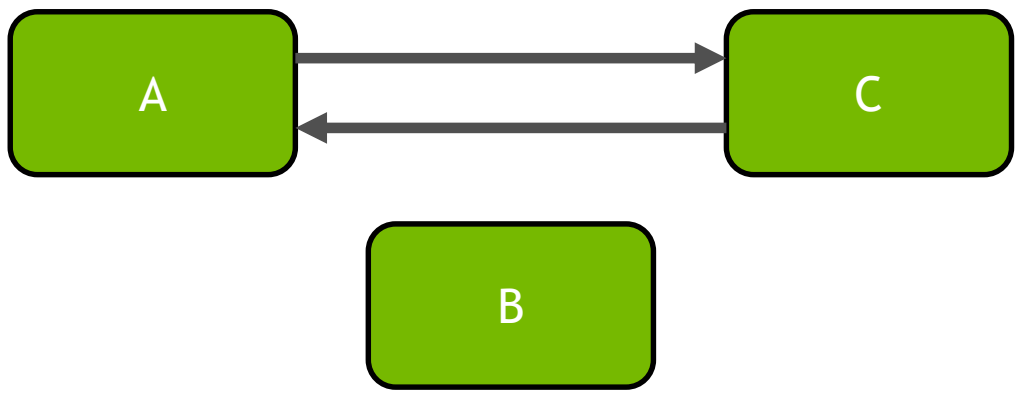
```
__device__ void insert_after(Node *a, Node *b)
{
    Node *c;
    lock(a); lock(a->next);
    c = a->next;

    a->next = b;
    b->prev = a;

    b->next = c;
    c->prev = b;

    unlock(c); unlock(a);
}
```

Doubly-Linked List with Fine Grained Lock



STARVATION FREE ALGORITHM

Example

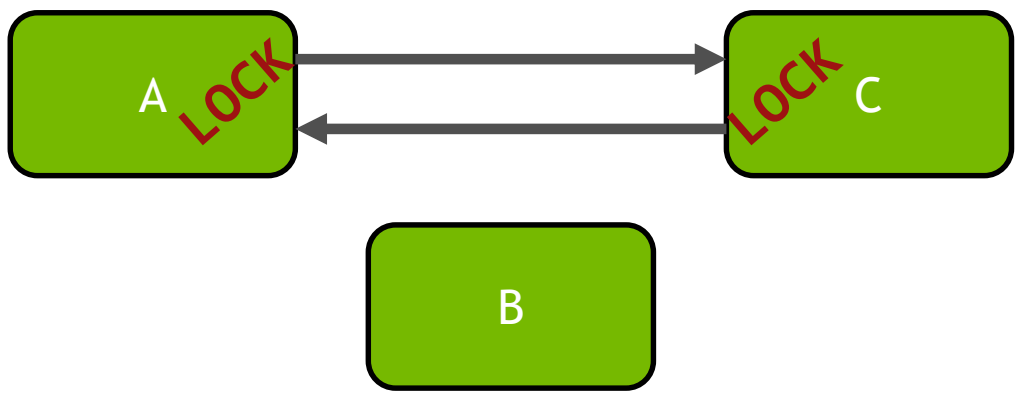
```
__device__ void insert_after(Node *a, Node *b)
{
    Node *c;
    lock(a); lock(a->next);
    c = a->next;

    a->next = b;
    b->prev = a;

    b->next = c;
    c->prev = b;

    unlock(c); unlock(a);
}
```

Doubly-Linked List with Fine Grained Lock



***Not shown: lock() implementation**

STARVATION FREE ALGORITHM

Example

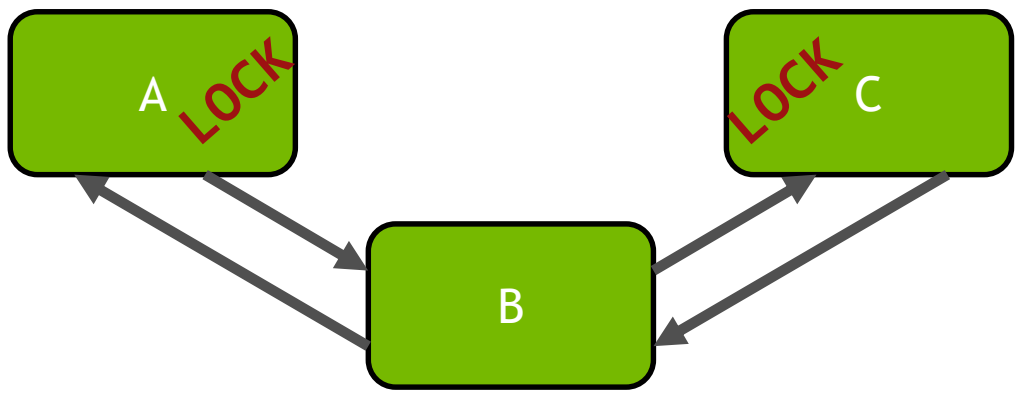
```
__device__ void insert_after(Node *a, Node *b)
{
    Node *c;
    lock(a); lock(a->next);
    c = a->next;

    a->next = b;
    b->prev = a;

    b->next = c;
    c->prev = b;

    unlock(c); unlock(a);
}
```

Doubly-Linked List with Fine Grained Lock



STARVATION FREE ALGORITHM

Example

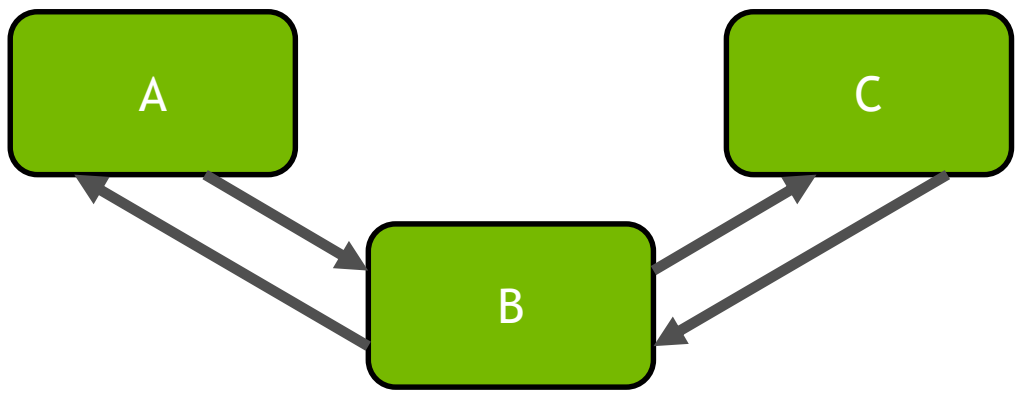
```
__device__ void insert_after(Node *a, Node *b)
{
    Node *c;
    lock(a); lock(a->next);
    c = a->next;

    a->next = b;
    b->prev = a;

    b->next = c;
    c->prev = b;

    unlock(c); unlock(a);
}
```

Doubly-Linked List with Fine Grained Lock



STARVATION FREE ALGORITHM

Example

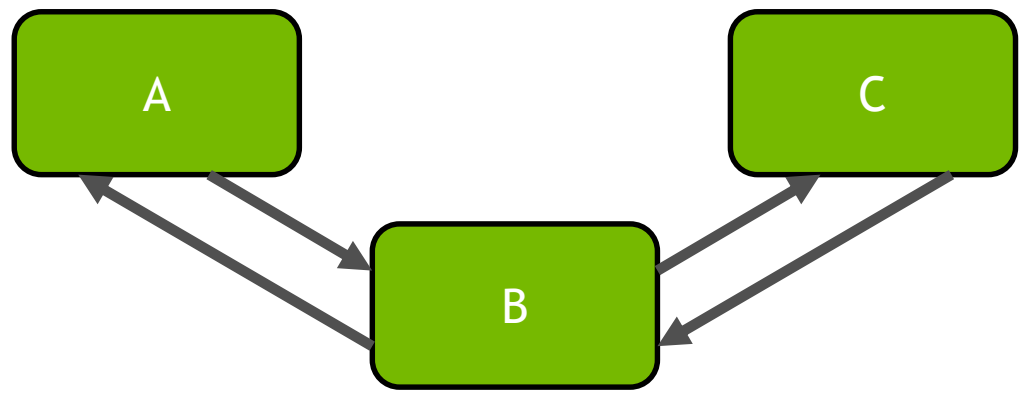
```
__device__ void insert_after(Node *a, Node *b)
{
    Node *c;
    lock(a); lock(a->next);
    c = a->next;

    a->next = b;
    b->prev = a;

    b->next = c;
    c->prev = b;

    unlock(c); unlock(a);
}
```

Doubly-Linked List with Fine Grained Lock



Tip! Volta can run 163,840 threads simultaneously

Minimize Contention!

SUPERCOMPUTERS PURPOSE-BUILT FOR AI

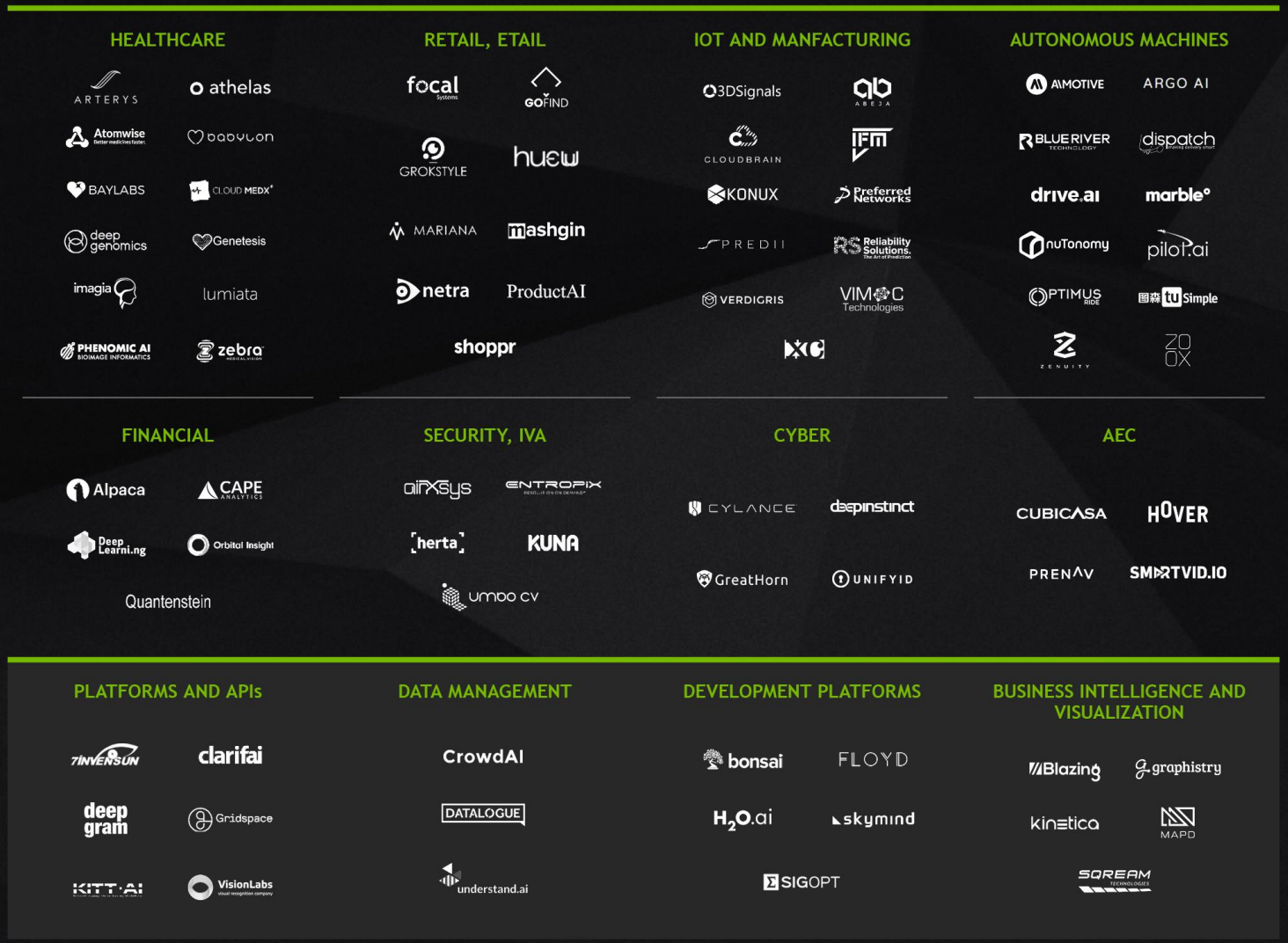
While deep learning holds enormous promise, it requires a massive amount of computing power. To arm data scientists in every organization, we created a lineup of AI supercomputers. NVIDIA DGX-1 with Volta, the world's first 1 PetaFLOPS computer, delivers the performance of up to 400 servers in a single box; while DGX Station packs 480 TeraFLOPS of computing power into a whisper-quiet workstation.

NVIDIA GPU CLOUD — ONE PLATFORM, RUN EVERYWHERE

The NVIDIA GPU Cloud gives AI developers access to our comprehensive deep learning software stack wherever they want it — on PCs, in the data center, or via the cloud. The containerized stack includes the latest deep learning frameworks, libraries, OS, and drivers. NVIDIA GPU Cloud makes it easier for developers to do deep learning training, experimentation, and deployment.

NVIDIA INCEPTION — 1,300 DEEP LEARNING STARTUPS

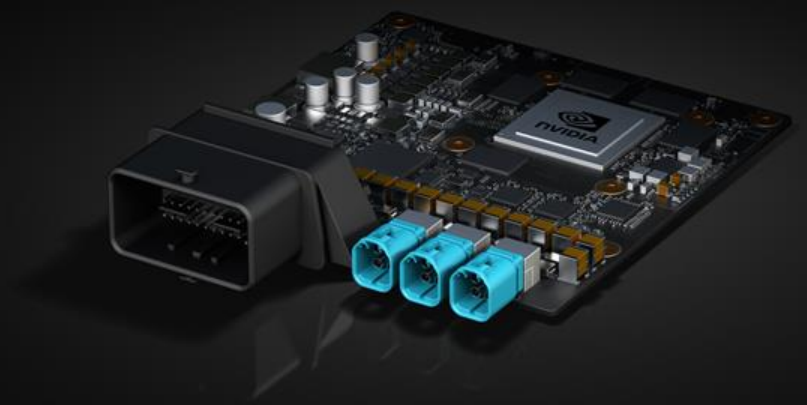
The NVIDIA Inception program nurtures more than 1,300 startups that are revolutionizing industries with advances in AI and data science. The program helps startups during critical stages of product development, prototyping, and deployment. At GTC, our annual developer conference, we awarded six AI startups with a total of $1.5 million to accelerate their work.



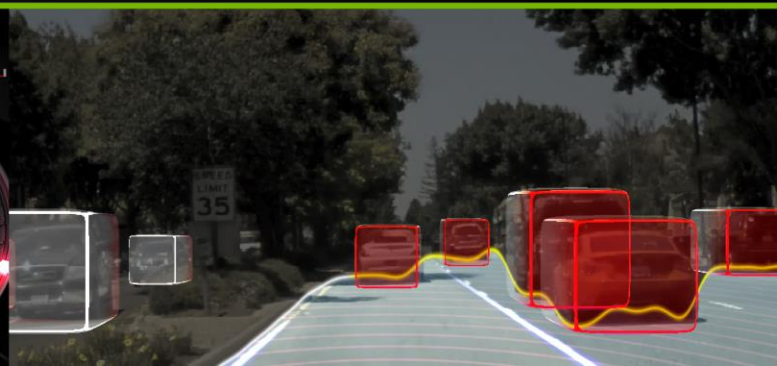
THE BRAIN OF AI CARS

Autonomous vehicles will modernize the $10 trillion transportation industry — making our roads safer and our cities more efficient. NVIDIA DRIVE™ PX is a scalable AI car platform that spans the entire range of autonomous driving.

Toyota recently joined some 225 companies around the world that have adopted the NVIDIA DRIVE PX platform for autonomous vehicles. They range from car companies and suppliers, to startups and research organizations.

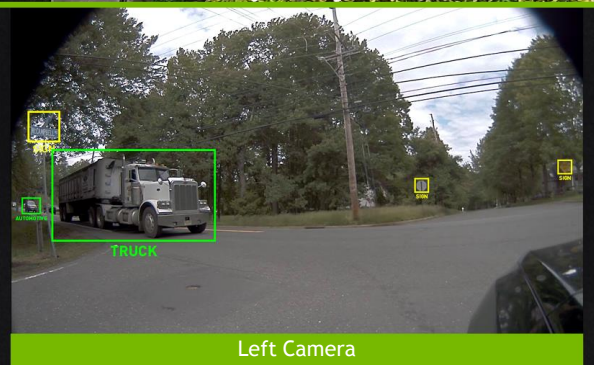


Mercedes-Benz
TESLA
TOYOTA
Audi
Baidu 百度
ZENRIN
TOMTOM
here
BOSCH
ZF



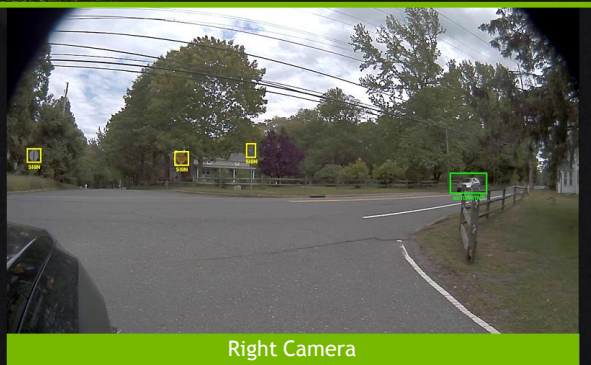
NVIDIA BB8 AI CAR — LEARNING BY EXAMPLE

Driving is a learned behavior that people do as second nature. Yet one that is impossible to program a computer to perform. Using all of the AI capabilities of NVIDIA DRIVE PX 2, our research AI car, BB8, watches humans drive, and has learned to drive in all kinds of conditions — on highways and dirt roads, through obstacle courses, at night, and in the rain. Processing data from multiple cameras, BB8 can even look both ways before safely crossing a busy road on its own.



Left Camera

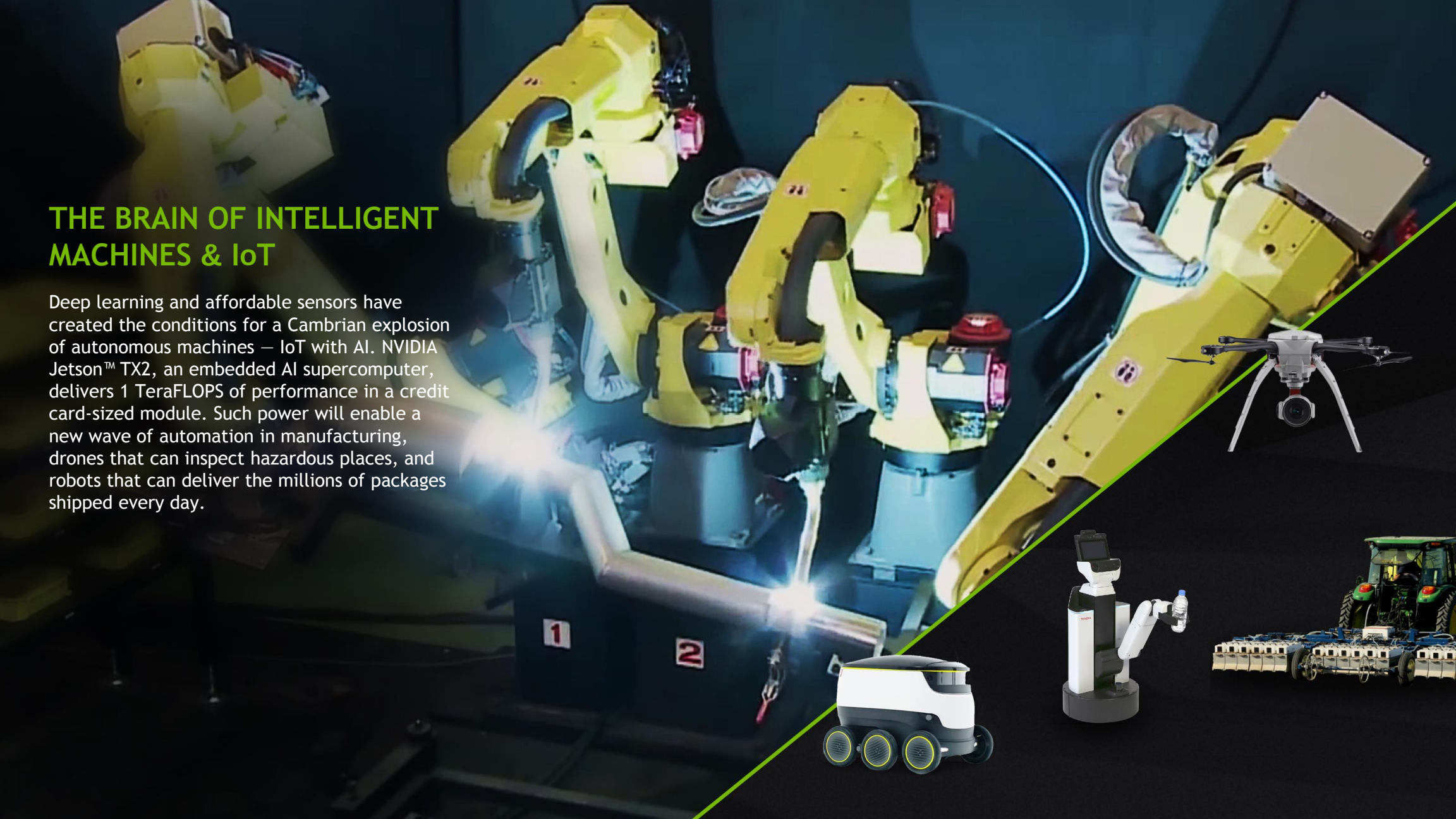
Center Camera



Right Camera

THE BRAIN OF INTELLIGENT MACHINES & IoT

Deep learning and affordable sensors have created the conditions for a Cambrian explosion of autonomous machines — IoT with AI. NVIDIA Jetson™ TX2, an embedded AI supercomputer, delivers 1 TeraFLOPS of performance in a credit card-sized module. Such power will enable a new wave of automation in manufacturing, drones that can inspect hazardous places, and robots that can deliver the millions of packages shipped every day.

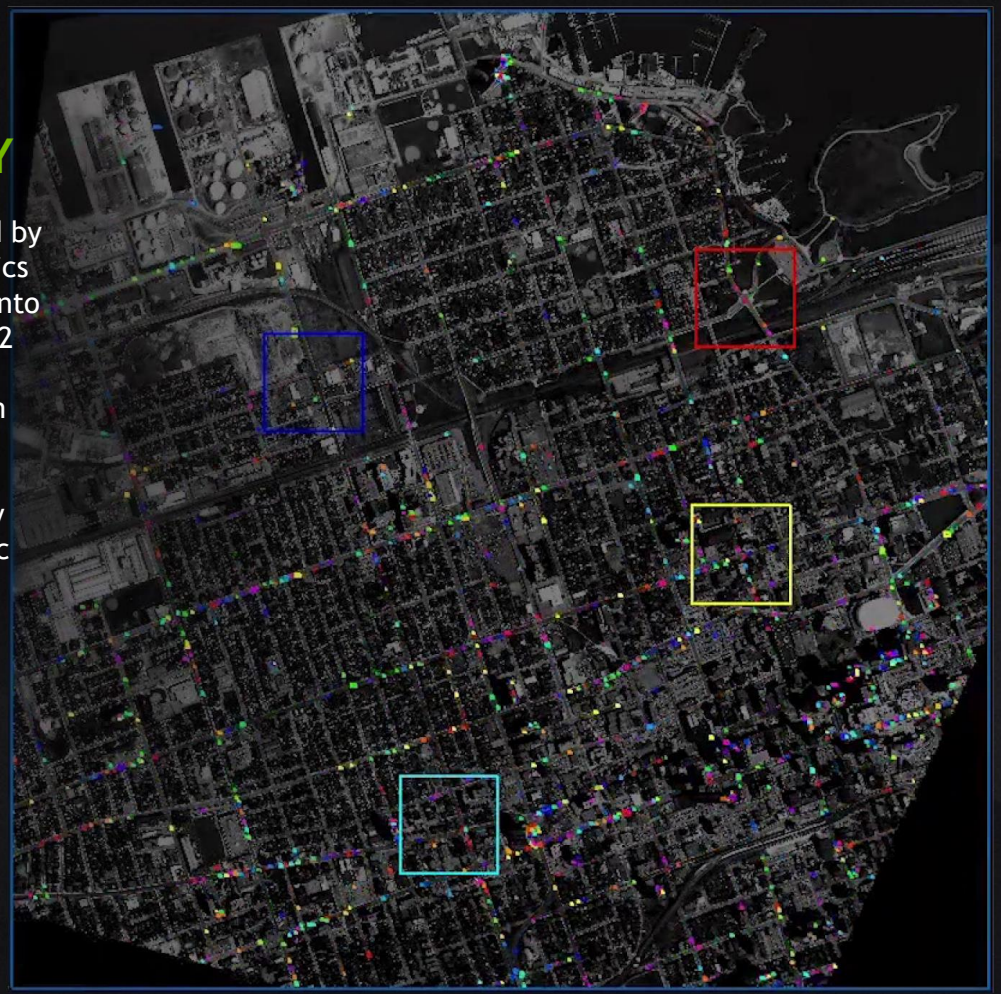


ISAAC — ACCELERATED LEARNING FOR A WORLD OF INTELLIGENT MACHINES

The Isaac robot simulator, an AI-based software platform, lets developers train robots in highly realistic, physics-based virtual environments and then transfer that knowledge to real-world units. Developers can set up extensive test scenarios using deep learning training, and then simulate them in minutes instead of months.

THE BRAIN OF THE AI CITY

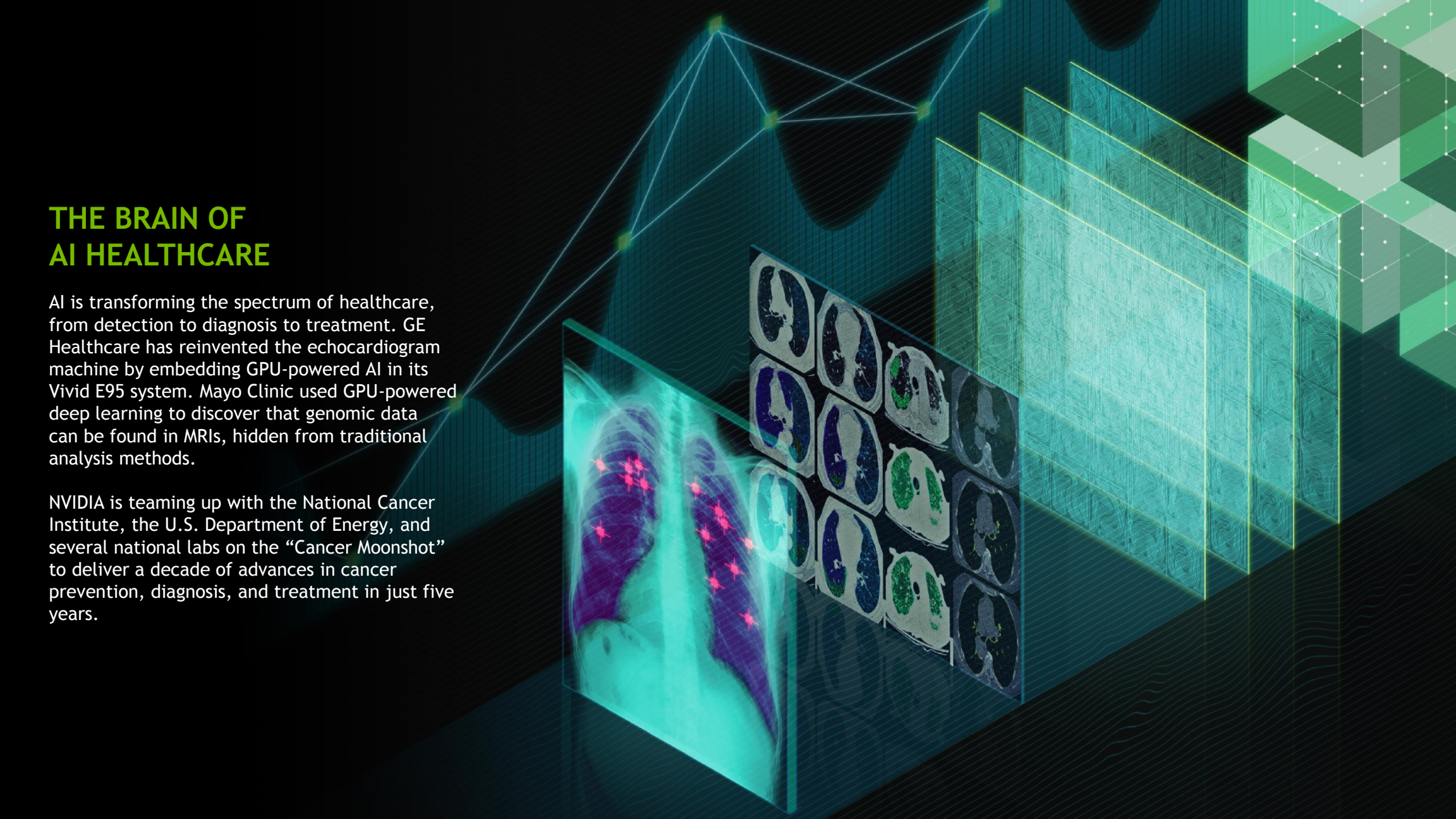
There will be 1 billion cameras in the world by 2020. AI will power intelligent video analytics that can turn this massive amount of data into safer, more efficient cities. With Jetson TX2 at the edge and Tesla GPUs in the cloud, NVIDIA Metropolis is an end-to-end platform that provides a foundation for the AI City. More than two dozen partners around the world have adopted Metropolis for a variety of applications, from public safety to traffic management to city services.



THE BRAIN OF AI HEALTHCARE

AI is transforming the spectrum of healthcare, from detection to diagnosis to treatment. GE Healthcare has reinvented the echocardiogram machine by embedding GPU-powered AI in its Vivid E95 system. Mayo Clinic used GPU-powered deep learning to discover that genomic data can be found in MRIs, hidden from traditional analysis methods.

NVIDIA is teaming up with the National Cancer Institute, the U.S. Department of Energy, and several national labs on the “Cancer Moonshot” to deliver a decade of advances in cancer prevention, diagnosis, and treatment in just five years.



NVIDIA