

AI & GPU ACCELERATED COMPUTING IN GOVERNMENT



Larry Brown Ph.D.
Federal Solutions Architecture Manager



NVIDIA

- > Founded in 1993
- > Jensen Huang, Founder & CEO
- > 11,000 employees
- > \$123B market cap; \$6.9B revenue in FY17

“World’s Most Admired Companies”

– Fortune

“50 Smartest Companies: #1”

– MIT Tech Review

“#3 Top CEO in the World”

– Harvard Business Review

“Most Innovative Companies”

– Fast Company

PERVASIVE USE OF GPUS IN GOVERNMENT

Geospatial | Intelligent Data & Video Analytics | Health & Human Services | Research



AI & DL | Data Analytics | High Performance Compute | Virtual Desktop Integration

THE ERA OF AI

The PC revolution put a computer in every home. The mobile era put a computer in every pocket, and then 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.

PC



MOBILE



CLOUD



AI



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™ is a scalable AI car platform that spans the entire range of autonomous driving, from traffic-jam pilots to robotaxis. More than 225 companies have adopted DRIVE, using NVIDIA AI technology in their data centers and vehicles. They range from car companies and suppliers, to mapping and sensor companies, to startups and research organizations.



Audi



Mercedes-Benz



TOYOTA

TESLA



Baidu

ZENRIN



TomTom



HERE

Autoliv



BOSCH



ZF



AI, ML AND DL

ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



MACHINE LEARNING

Machine learning begins to flourish.



DEEP LEARNING

Deep learning breakthroughs drive AI boom.



1950's

1960's

1970's

1980's

1990's

2000's

2010's

DEEP LEARNING USE CASES IN FEDERAL

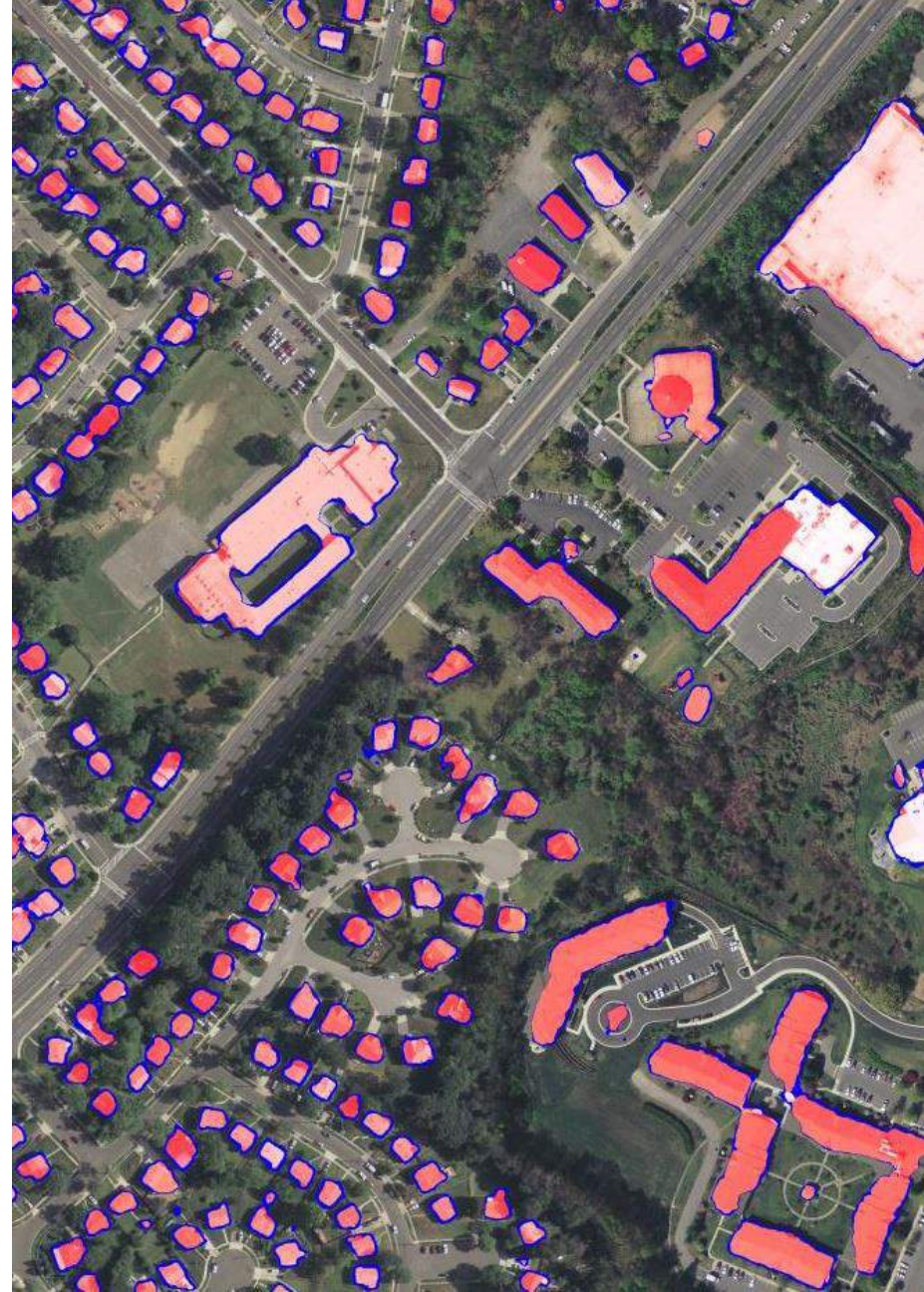
A 21ST CENTURY AI PLANNING TOOL

Understand distribution of **7B+ people**
for planning infra. & vital services

Process high-res imagery **in minutes** to
map urban dynamics and information
critical for emergency response



Urban Dynamics Institute

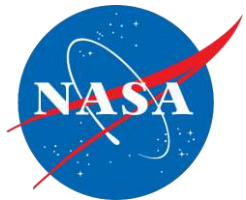


USING AI TO MONITOR EARTH'S VITALS

Record increases in global temperature, glacial retreat and rising sea levels

Using satellite imagery to measure the effects of carbon and greenhouse gas emissions

Help scientists to plan to protect ecosystems and farmers improve crop production



Ames Research Center

DeepSat



Agricultural growth near Qasr al Farafra, Egypt

MAPPING AN END TO POVERTY

Collecting Data on Global Poverty



CHALLENGE

1B people live on less than \$2 a day - ending poverty tops the list of the UN's sustainable development goals

Researching poverty is expensive, manual, dangerous and does not scale

SOLUTION

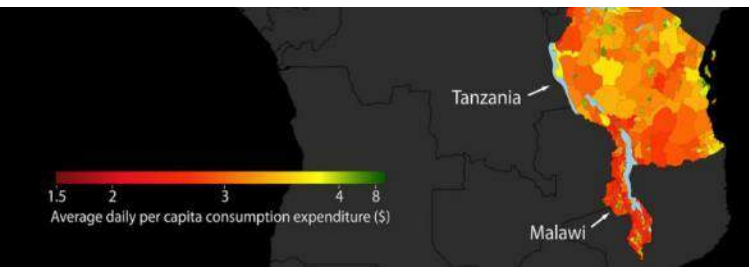
Big data and applied deep learning to show the distribution of wealth across the globe

'Earth at night' images determine useful features to indicate economic activity (such as city lights)

IMPACT

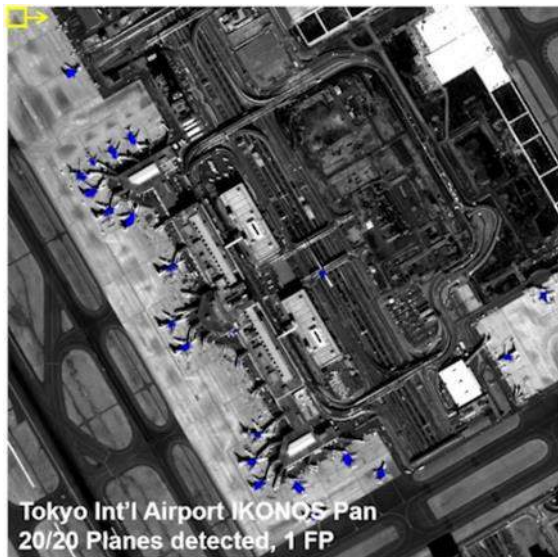
Scalable, safer and less-expensive way to collect data on poverty

Poverty maps will help world organizations locate those most in need of relief



DL FOR REMOTE SENSING

MEGA - Machine Learning for GEOINT Analytics



Automated Target Recognition



Real-time Ground Weather Intel



Activity Based Intelligence

DEEP LEARNING ON RADAR

Detect and recognize mobile ground objects from airborne platform

Speed & Accuracy are both critical.

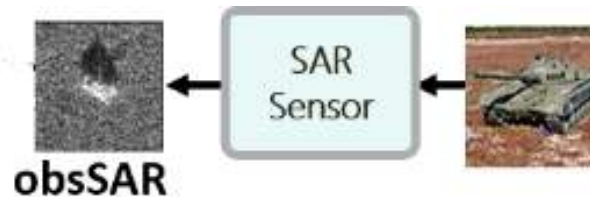
Trained on TitanX - deployed on Jetson.

	Nominal Competing Product	Deep Learning Analytics	Improvement Factor
Power (W)	>1000	~12	>80x
Weight (lbs)	>180	< 3.0	>60x
Size (cuft)	>30	~0.13	>200x
FAR (at canonical P_{ID})	50%	< 2%	>25x



TRACE

Target Recognition and Adaptation in Contested Environments



ANALYTICS PROTECTS & PREVENTS DELAYS

Tracks 150+ billion pieces of mail annually

Achieves near-immediate analysis of data from 213,000+ scanning devices

Reduced driving by 70m miles – 7m gallons of fuel, 70,000 tons of carbon



AI TO ACCELERATE CANCER RESEARCH

CANDLE is a common discovery platform, with the goal of achieving 10X annual increases in productivity for cancer researchers.

Unites the Department of Energy, National Cancer Institute with researchers at Oak Ridge, Lawrence Livermore, Argonne, and Los Alamos National Laboratories



DEEP LEARNING FOR CYBER SECURITY

Want more proactive approach than relying on signatures

Malconv results

Train set	Test set	Our model AUC	Byte n-grams AUC	PE-Header Network
Group B train (small)	Group A	98.5	98.4	97.7
Group B train (small)	Group B test	95.8	97.9	91.4
Group B train (large)	Group A	98.1	93.4	-
Group B train (large)	Group B test	98.2	97.0	-

Table 1: Performance of malware detection models on Group A and Group B test sets when trained on both the small (400,000 samples) and large (two million samples) Group B training sets.



Booz | Allen | Hamilton



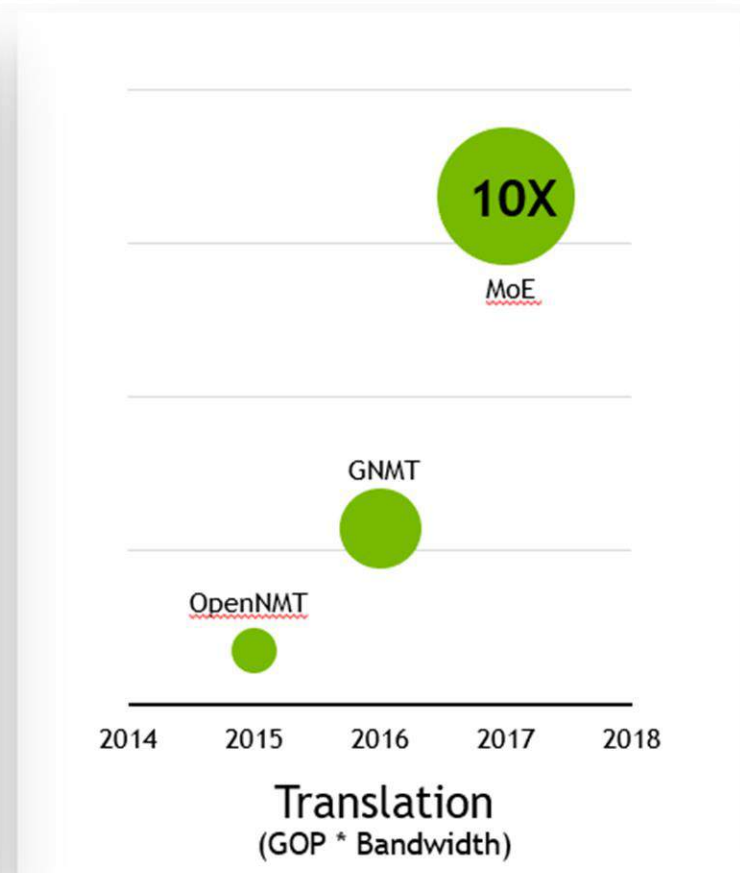
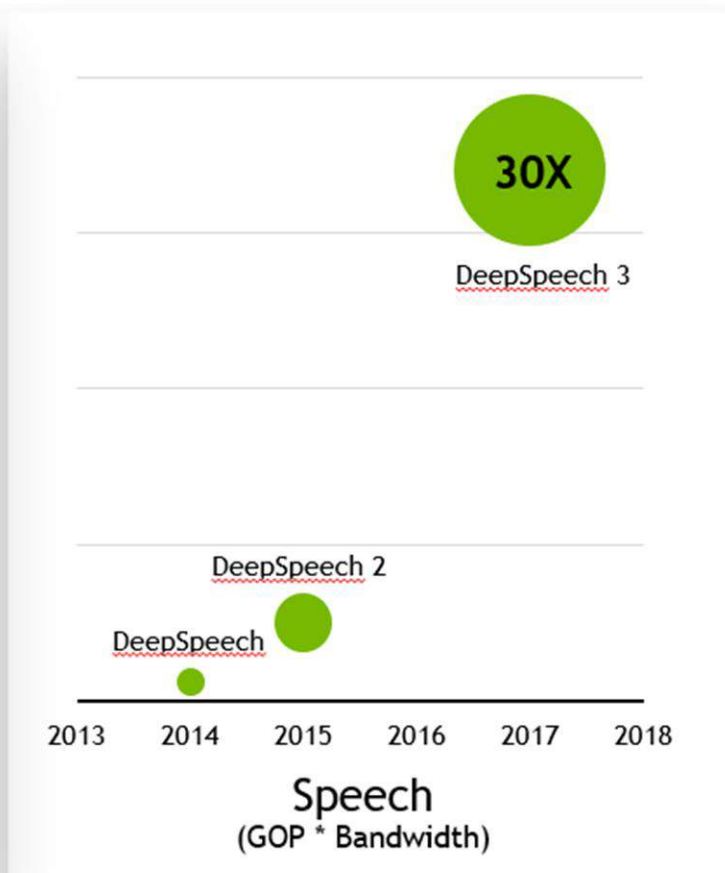
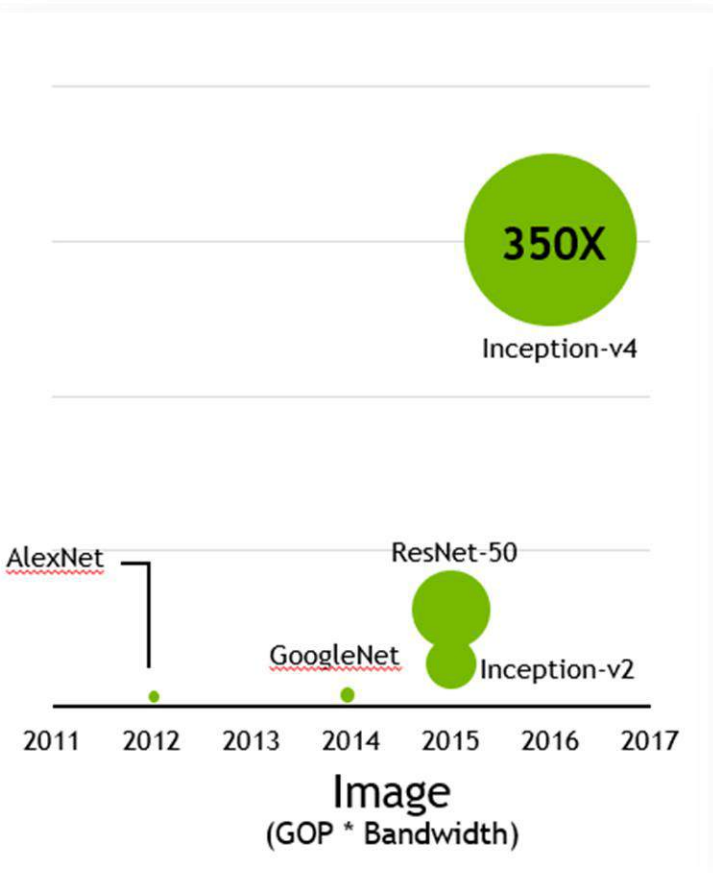
The Laboratory for Physical Sciences

University of Maryland

GPU - THE COMPUTE PLATFORM FOR AI

NEURAL NETWORK COMPLEXITY IS EXPLODING

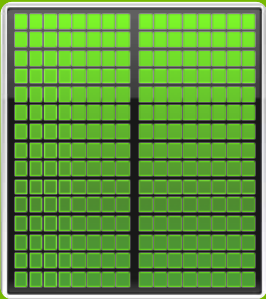
Large, More Compute Intensive



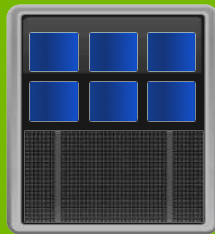
“It’s time to start planning for the end of Moore’s Law, and it’s worth pondering how it will end, not just when.”

*Robert Colwell
Retired Director, Microsystems Technology Office,
DARPA*

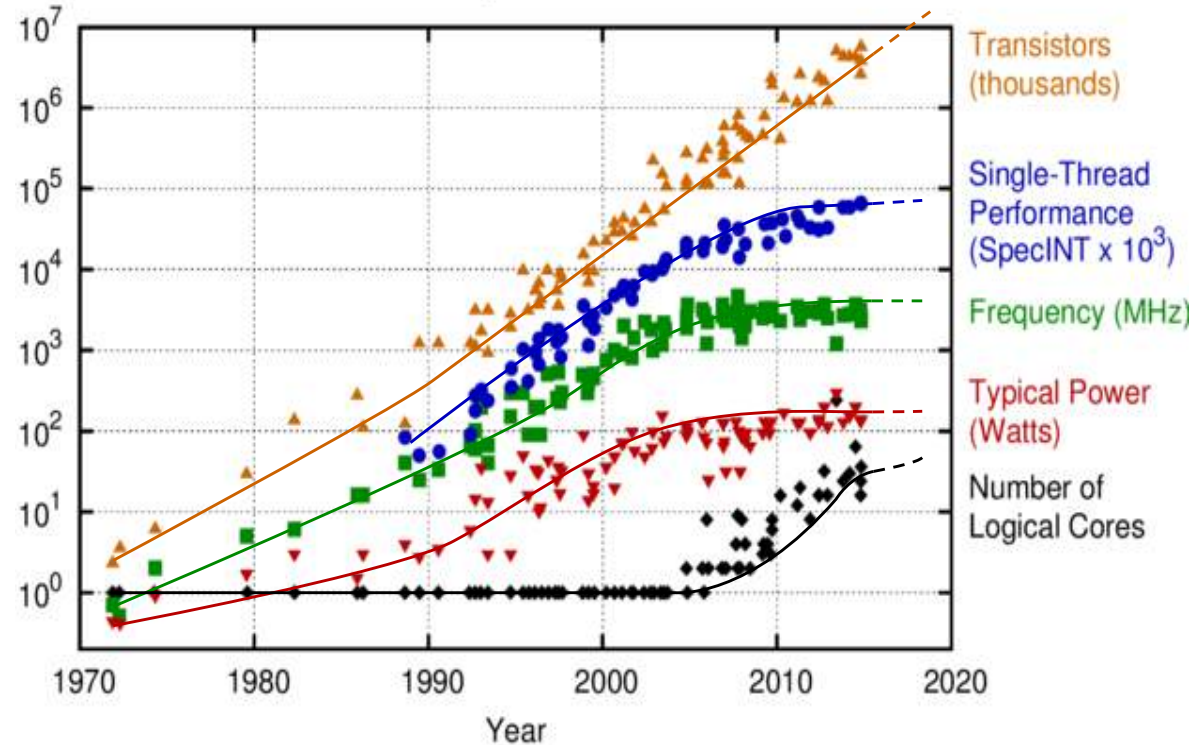
GPU Accelerator



CPU



40 Years of Microprocessor Trend Data



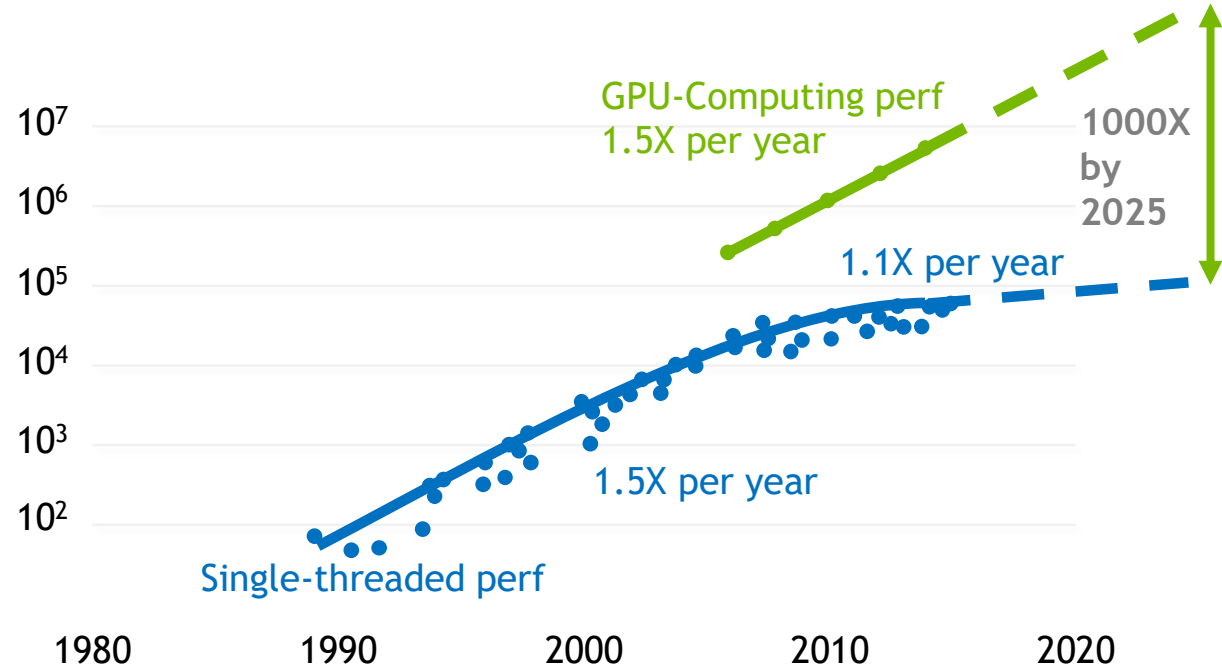
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

RISE OF GPU COMPUTING

10x-100x application performance gains

5x energy efficiency & low footprint

Next generation in-memory processing



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

PLATFORM BUILT FOR AI

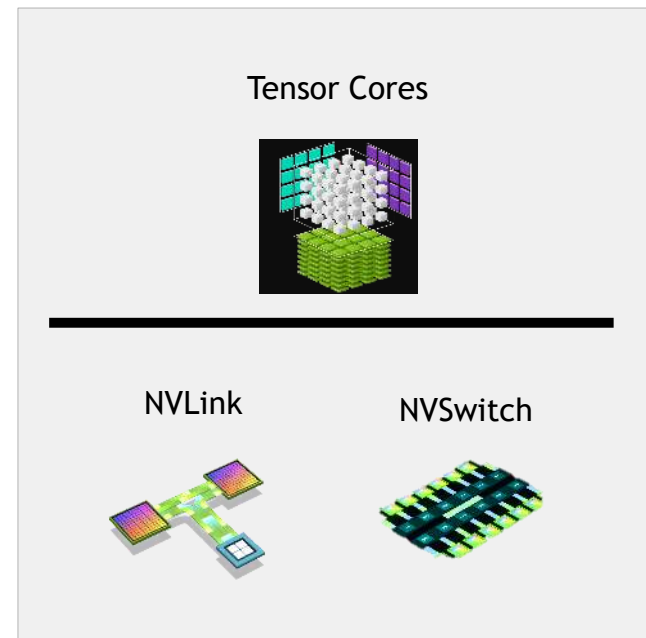
Accelerating Every Framework And Fueling Innovation



All Use-cases

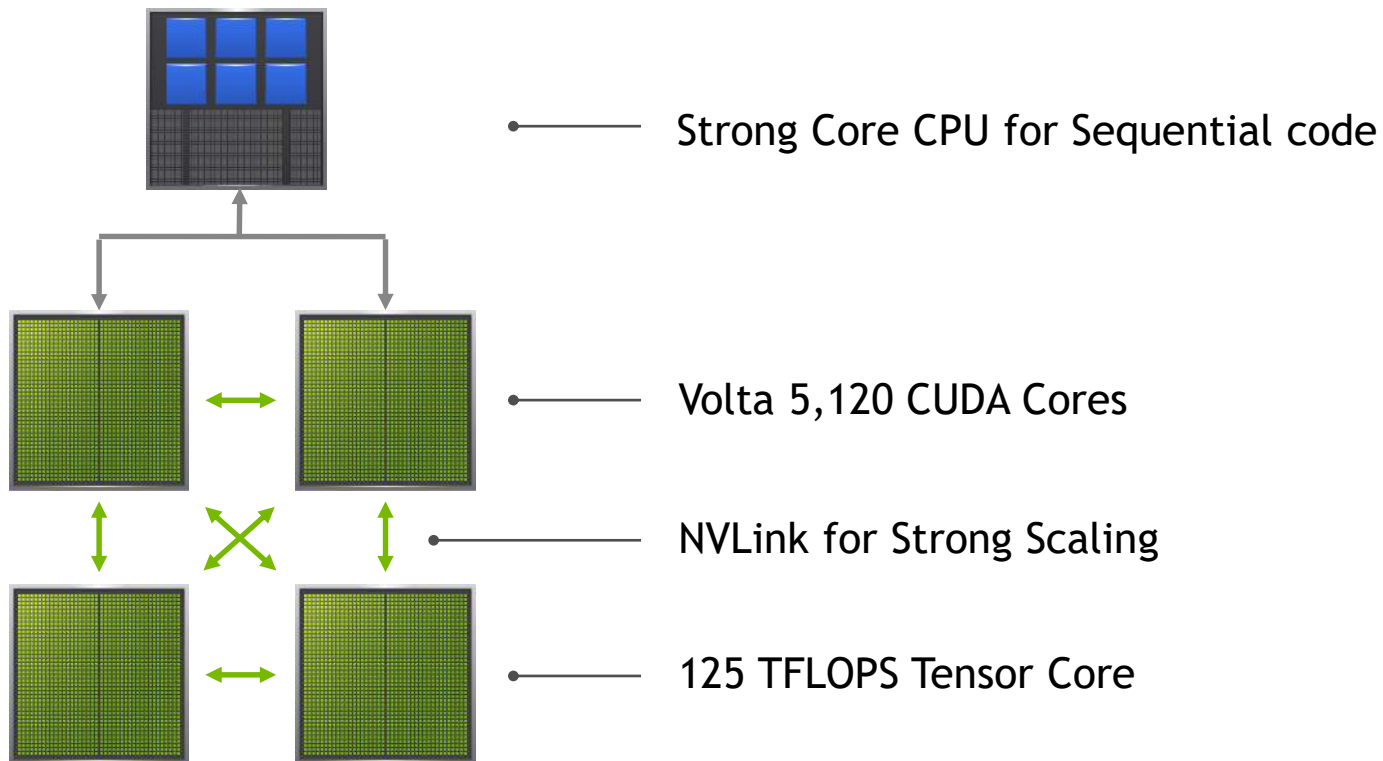


All Major Frameworks

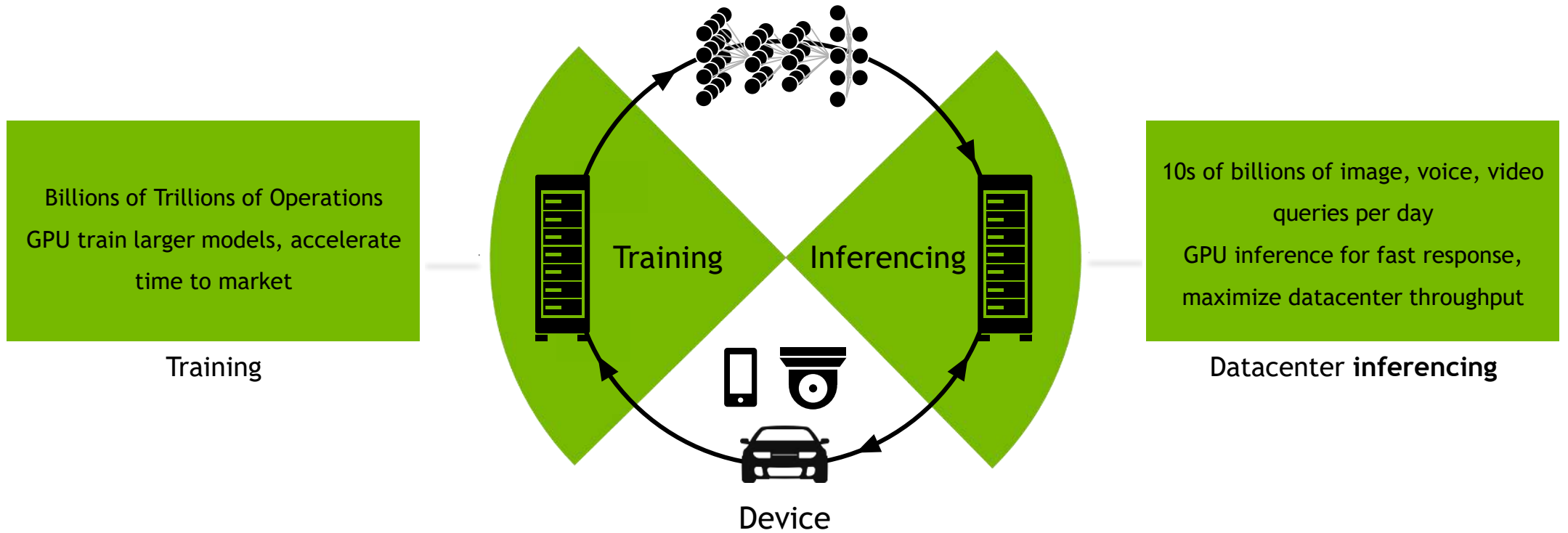


Volta Tensor Core, NVSwitch, NVLink

ARCHITECTING MODERN DATACENTERS



GPU DEEP LEARNING IS A NEW COMPUTING MODEL



TESLA V100 TENSOR CORE GPU

World's Most Advanced
Data Center GPU

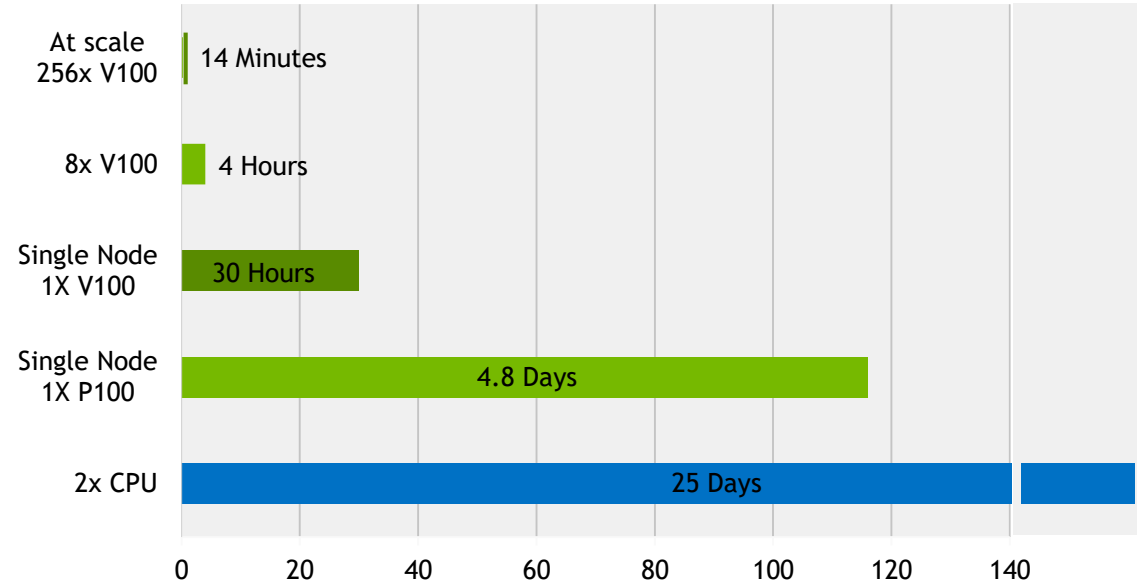
5,120 CUDA cores
640 NEW Tensor cores
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS
| 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
32 GB HBM2 @ 900GB/s |
300GB/s NVLink



TESLA PLATFORM ENABLES DRAMATIC REDUCTION IN TIME TO TRAIN



Relative Time to Train Improvements
(ResNet-50)



ANNOUNCING TESLA T4

WORLD'S MOST ADVANCED INFERENCE GPU

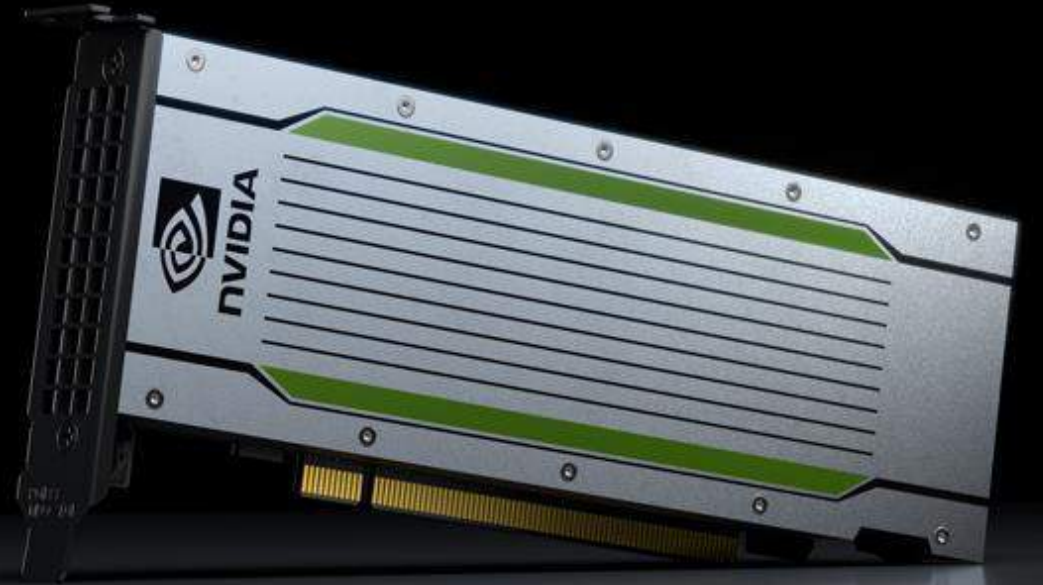
Universal Inference Acceleration

320 Turing Tensor cores

2,560 CUDA cores

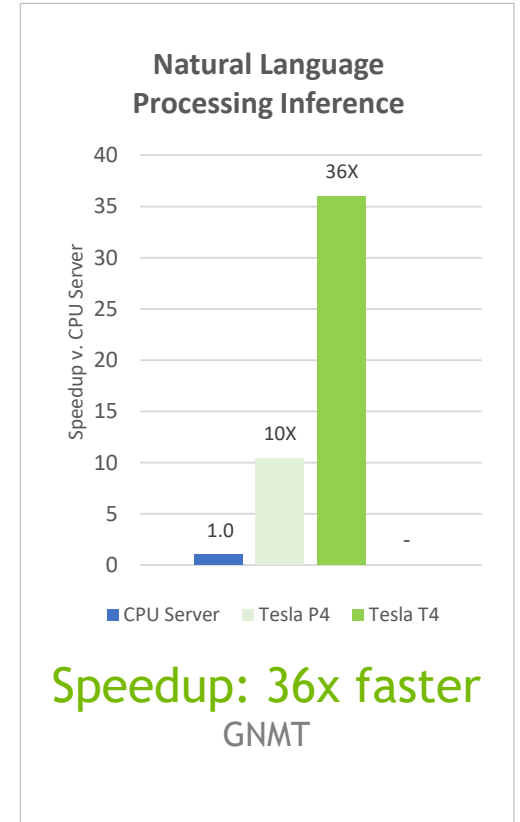
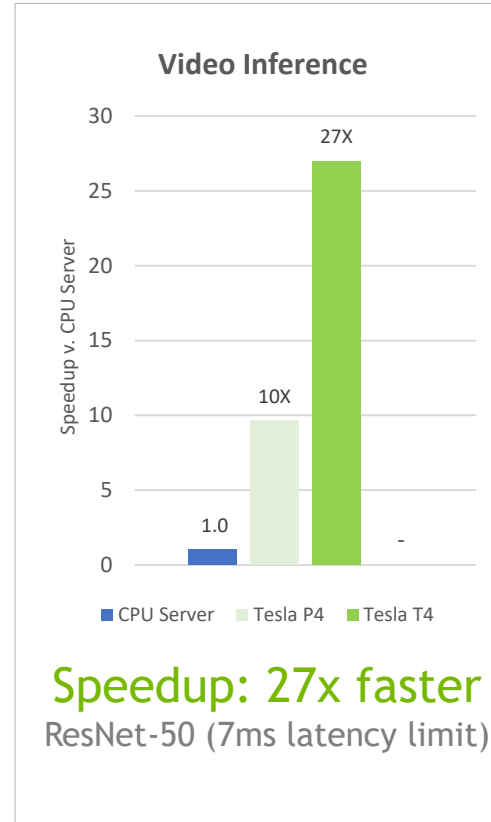
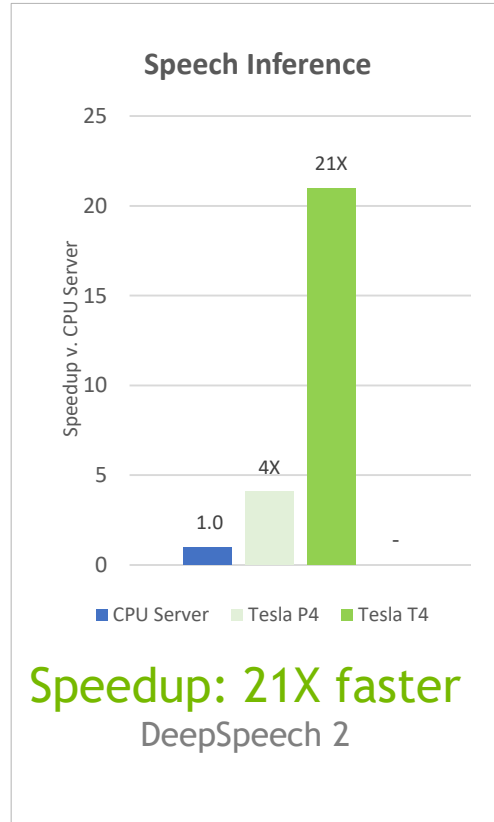
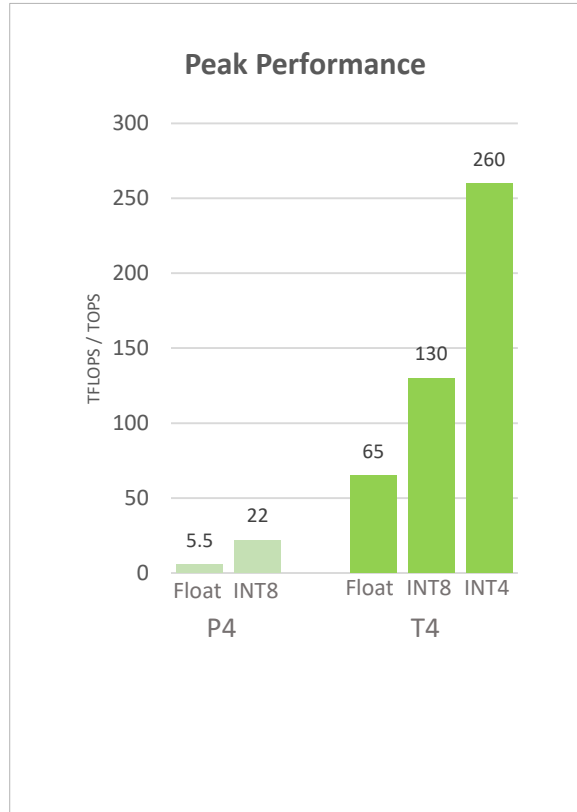
65 FP16 TFLOPS | 130 INT8 TOPS | 260 INT4 TOPS

16GB | 320GB/s

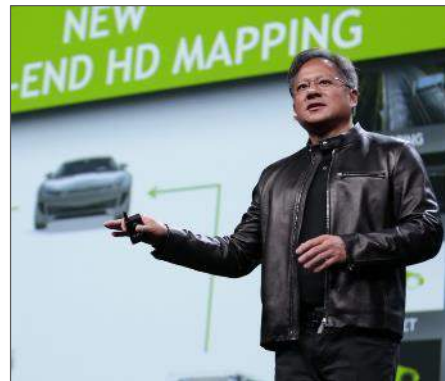


WORLD'S MOST PERFORMANT INFERENCE PLATFORM

Up To 36X Faster Than CPUs | Accelerates All AI Workloads



GPU TECHNOLOGY CONFERENCE



ADVANCE YOUR DEEP LEARNING TRAINING AT GTC

Don't miss the world's most important event for AI & GPU in Federal

Washington DC

October 23-24, 2018

