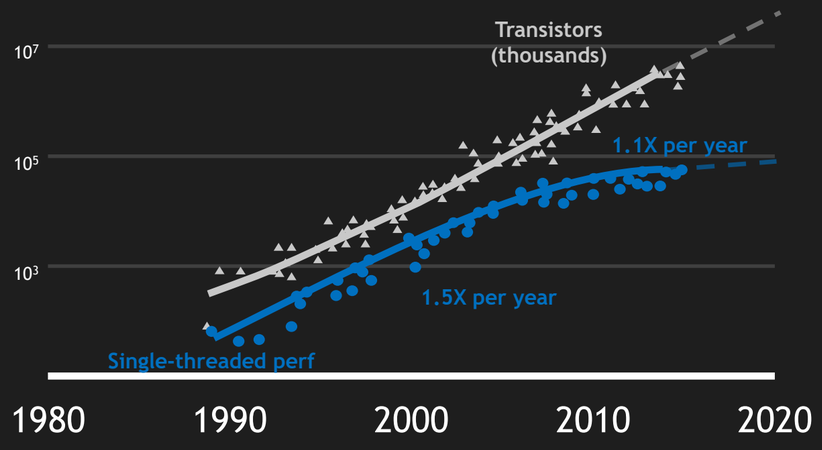
A NEW COMPUTING ERA

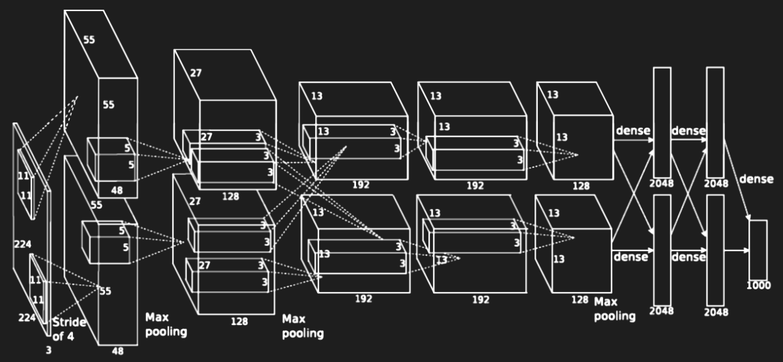
DAVID B. KIRK, FELLOW| NVIDIA AI Conference Singapore 2017

TWO FORCES DRIVING THE FUTURE OF COMPUTING



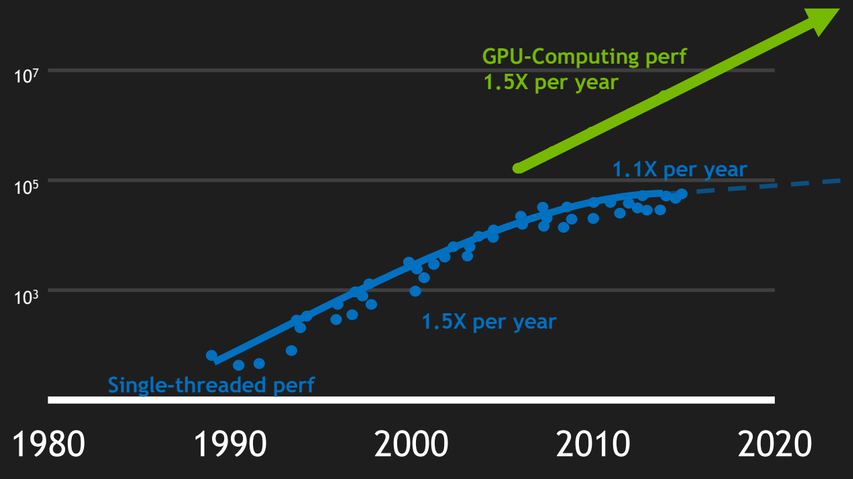
40 Years of CPU Trend Data

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



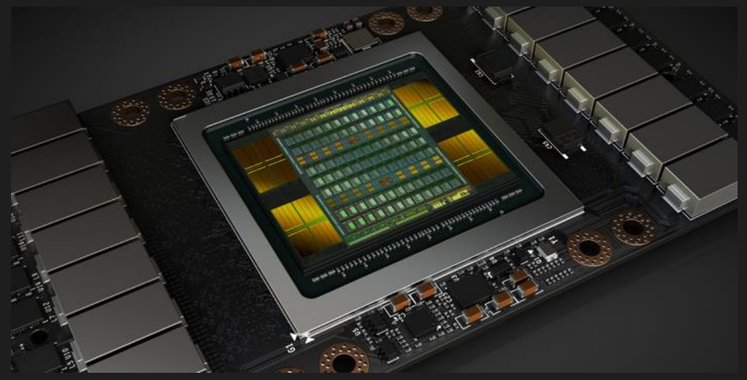
The Big Bang of Deep Learning

RISE OF NVIDIA GPU COMPUTING



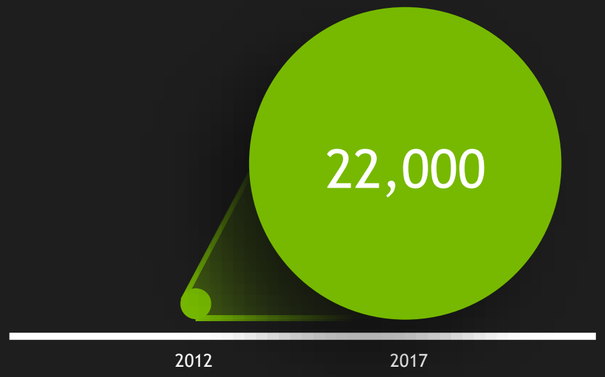
40 Years of CPU Trend Data

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

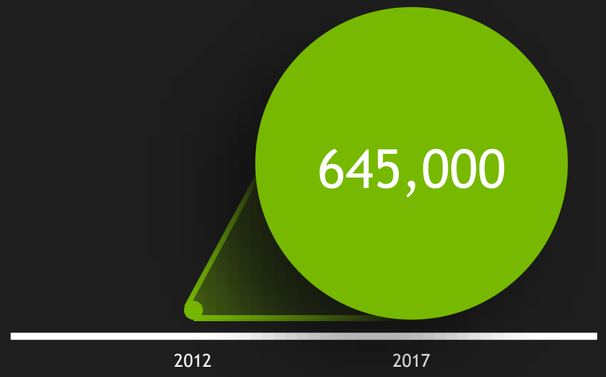


NVIDIA Volta GPU | 21B xtors | 120 TFLOPS

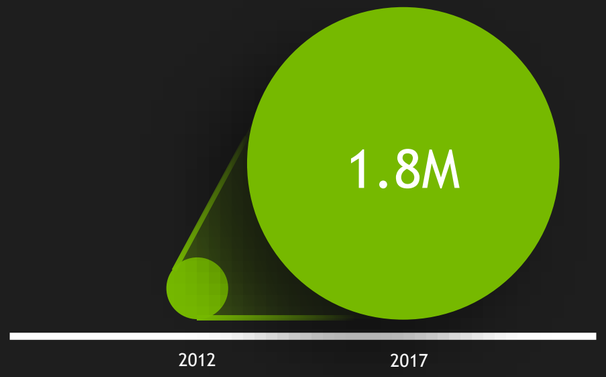
RISE OF NVIDIA GPU COMPUTING



Global GTC Attendees
10X in 5 Years



GPU Developers
15X in 5 Years



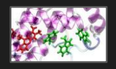
CUDA Downloads
5X in 5 Years

NVIDIA GPU-ACCELERATED APPLICATIONS

Computer Graphics

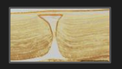
Scientific Computing

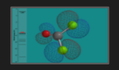
SIMULIA Abaqus Finite-Element Analysis

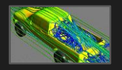

AMBER Molecular Dynamics

COSMO Climate Weather

ChaNGa Astrophysics


Schlumberger WG Seismic Processing


Gaussian Quantum Chemistry


ANSYS Fluent Computational Fluid Dynamics

PowerGrid Medical Imaging

Deep Learning

TensorFlow | Caffe2 | mxnet | Chainer

Microsoft Cognitive Toolkit | PYTORCH | theano | PaddlePaddle

Data Science

K-Means Clustering | Gradient Boosting

Generalized Linear Model | Support Vector Machine

ANALYTICS

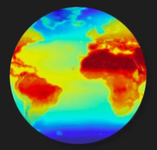
bonsai | graphistry | H_2O.ai | skymind

DATABASES

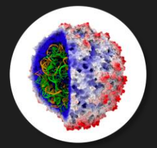
Blazing | kinetica | MAPD | SQREAM TECHNOLOGIES

NVIDIA CUDA

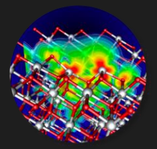
NVIDIA GPU-ACCELERATED INDUSTRIES



Weather



Molecular Dynamics



Materials Science

HPC

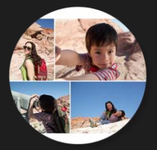
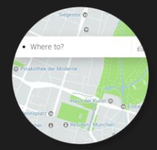


Photo Album

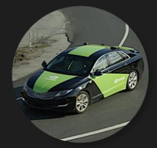
Recommendations

Social Media

INTERNET SERVICES



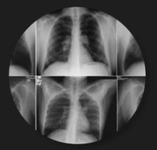
Ride Sharing



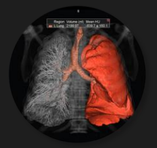
Self-Driving Cars

Design & Simulation

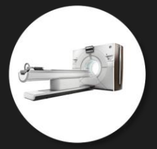
TRANSPORTATION



Radiology



Volumetric Rendering

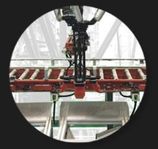


3D CT

MEDICAL IMAGING

AV Truck

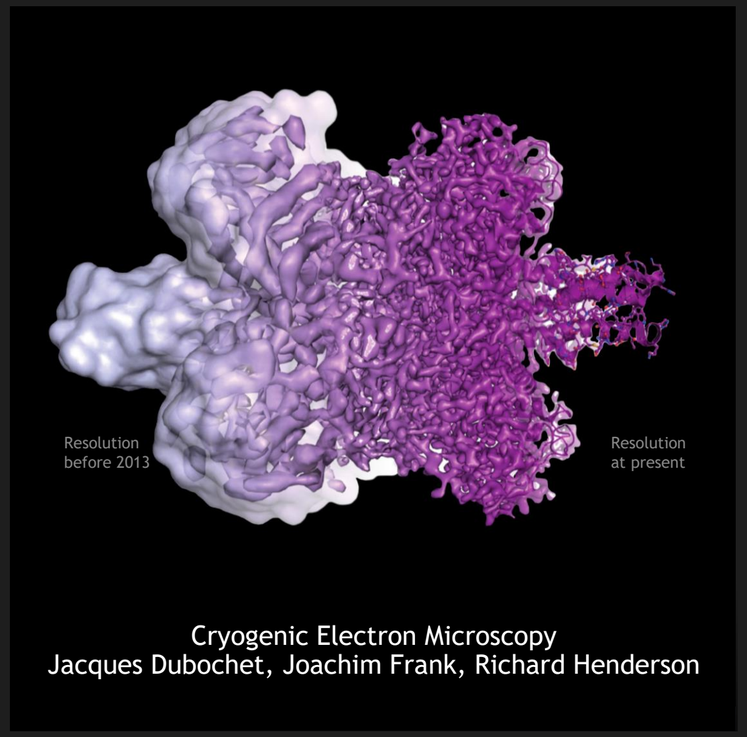
Route Management



Warehouse Automation

LOGISTICS

NVIDIA GPU ACCELERATES 2017 NOBEL PRIZES IN CHEMISTRY AND PHYSICS



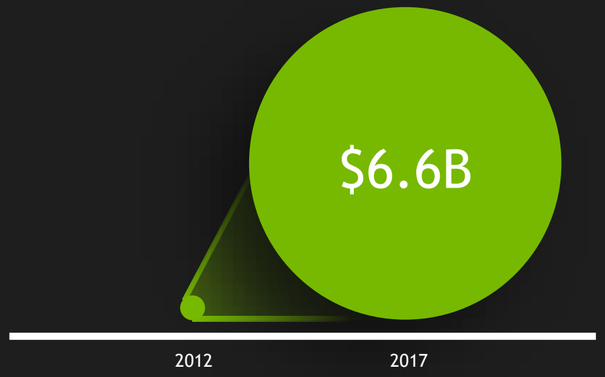
Cryogenic Electron Microscopy
Jacques Dubochet, Joachim Frank, Richard Henderson

Detection of Gravitational Waves
Rainer Weiss, Barry Barish, Kip Thorne

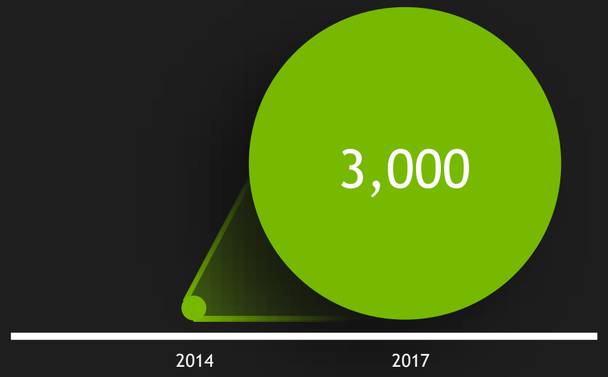
ANNOUNCING
NVIDIA HOLODECK
THE DESIGN LAB OF THE FUTURE
Photorealistic Models
Physically Simulated Interaction
Virtual Team Collaboration
GPU-accelerated AI
Early Access NOW
nvidia.com/holodeck

CREATE AND COLLABORATE IN NVIDIA HOLODECK

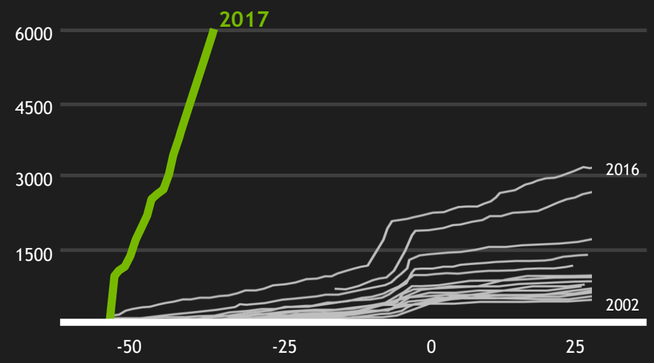
THE ERA OF AI



AI Startup Funding
12X in 5 Years



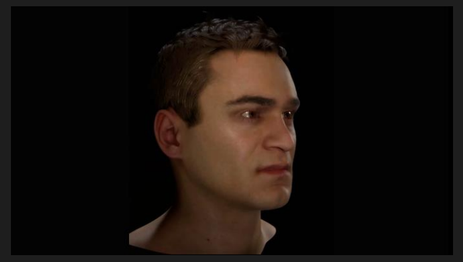
Deep Learning Papers Published
13X in 3 Years



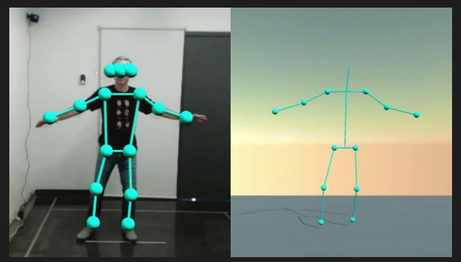
NIPS Registration

SOLVING THE UNSOLVABLE

NVIDIA
Interactive Ray Tracing



NVIDIA / Remedy
Audio-driven Facial Animation



WRNCH
Pose Estimation

University of Edinburgh
Character Animation

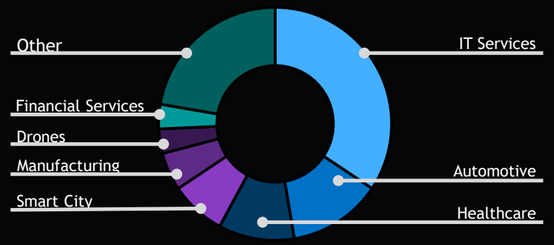
UC Berkeley / OpenAI
One-shot Imitation Learning

THE WORLD'S AI PLATFORM

Every Framework

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

NVIDIA Inception: 1,900 DL Startups

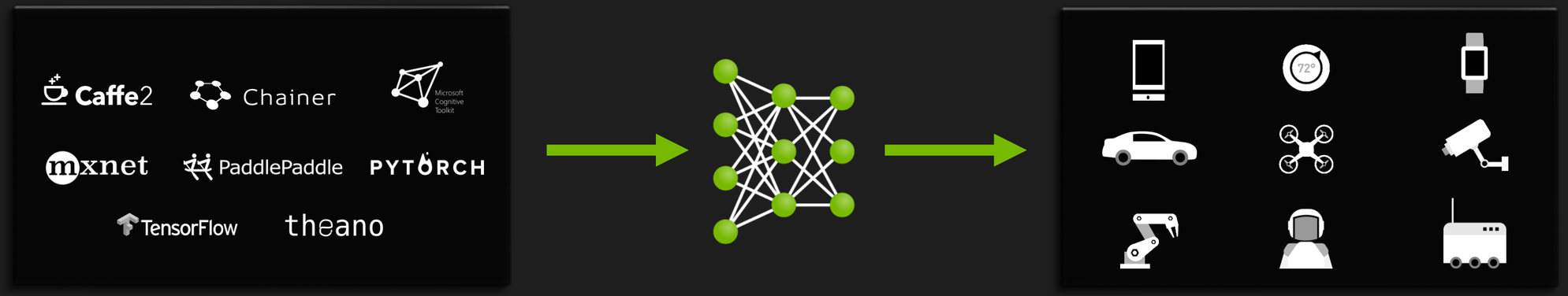


Every Cloud and Data Center

Alibaba Cloud · amazon web services · 百度云 cloud.baidu.com · Google Cloud Platform · Microsoft Azure · Tencent Cloud · cisco · DELL · Hewlett Packard Enterprise · HUAWEI · IBM · Lenovo

NVIDIA AI PLATFORM

AI INFERENCE IS THE NEXT GREAT CHALLENGE

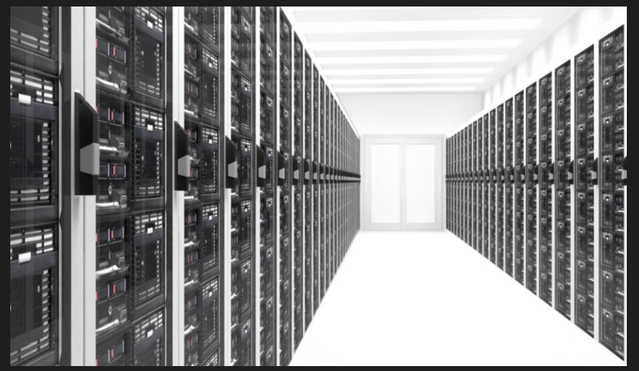


Training

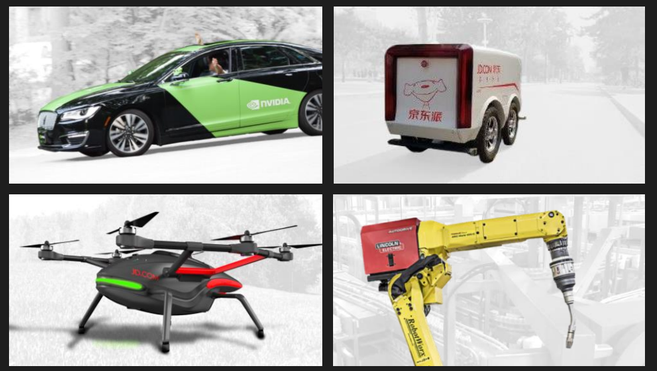
DNN Model

Inferencing

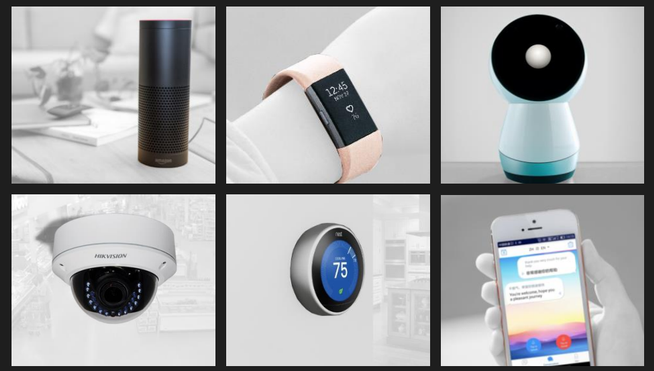
EXPLOSION OF INTELLIGENT MACHINES



20M Inference Servers



100s of Millions of Autonomous Machines



Trillions of IoT Devices

EXPLOSION OF NETWORK DESIGN

Convolution Networks

PRelu

ReLu

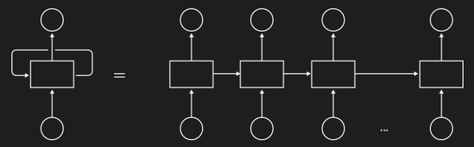
BatchNorm

Concat

Dropout

Pooling

Recurrent Networks



LSTM

GRU

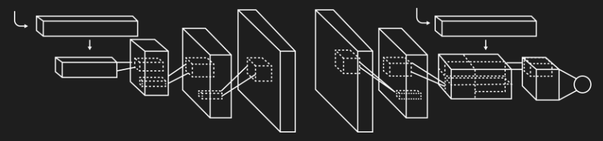
Highway

Projection

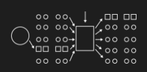
Embedding

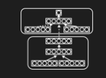
BiDirectional

Generative Adversarial Networks

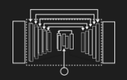


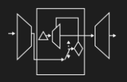
3D-GAN


Rank GAN


Conditional GAN

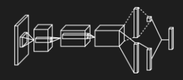
Coupled GAN

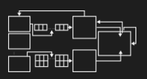

Speech Enhancement GAN


Latent space GAN

Reinforcement Learning

DQN


Dueling DQN


A3C

EXPLOSION OF NETWORK COMPLEXITY

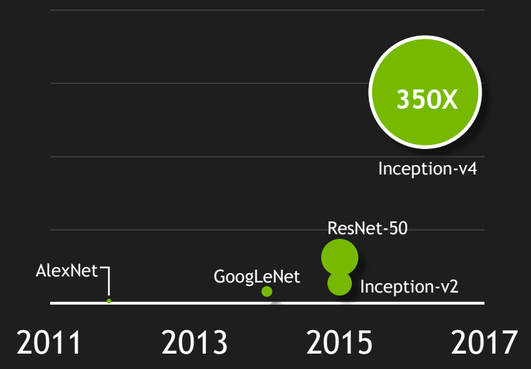
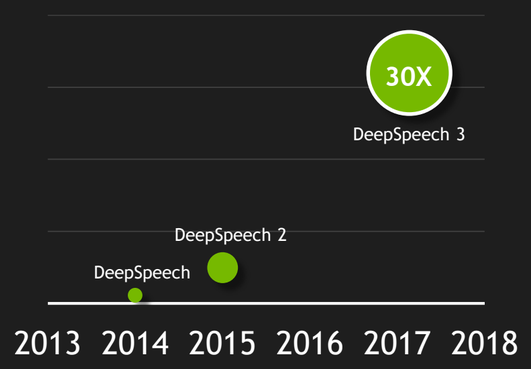
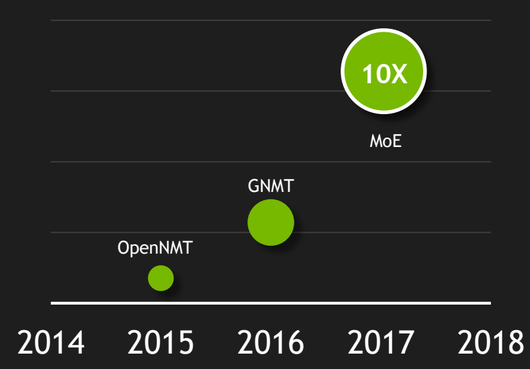


Image Network Complexity
GOPS * Bandwidth



Speech Network Complexity
GOPS * Bandwidth

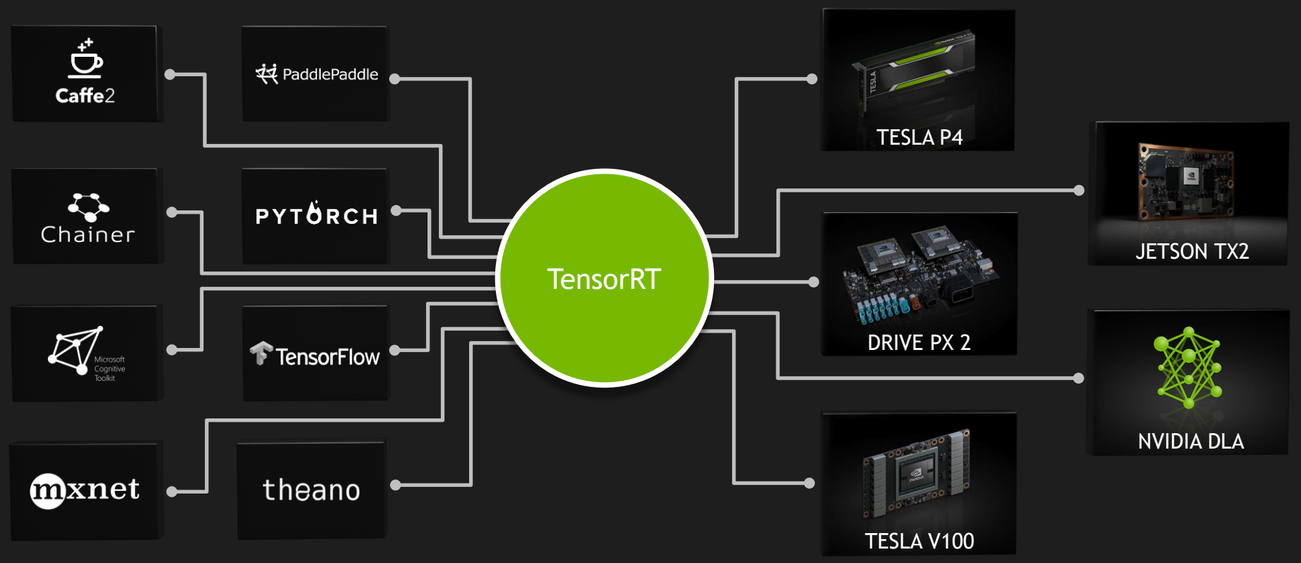


Translation Network Complexity
GOPS * Bandwidth

NEW
NVIDIA TENSORRT 3

PROGRAMMABLE INFERENCE ACCELERATOR

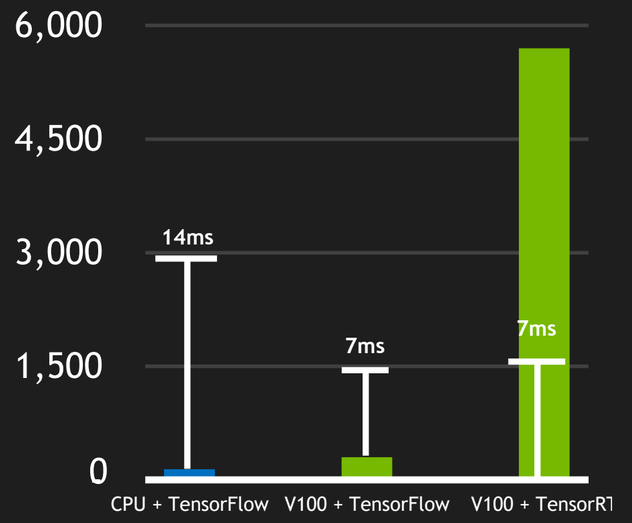
Compile and Optimize Neural Networks
Support for Every Framework
Optimize for Each Target Platform



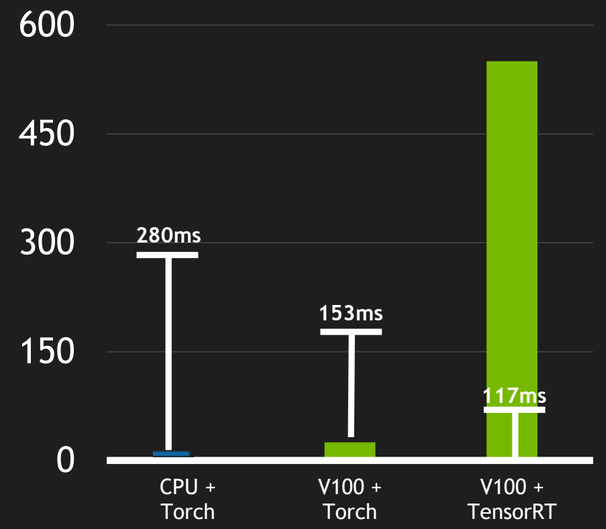
NEW
NVIDIA TENSORRT 3

PROGRAMMABLE INFERENCE ACCELERATOR

40x Speed-up on ResNet-50
140x Speed-up on OpenNMT
Half the Latency



Images/Sec (ResNet-50)



Sentences/Sec (OpenNMT)

NVIDIA TENSORRT
10X BETTER DATA CENTER TCO

1 NVIDIA HGX with 8 Tesla V100 GPUs
45,000 images / second
3 KWatts

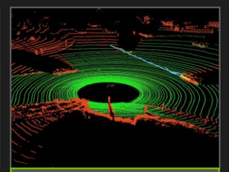
1/6 the Cost | 1/20 the Power | 4 Racks in a Box

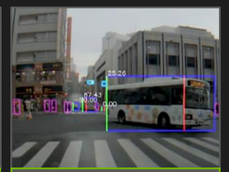
THE AUTONOMOUS VEHICLE REVOLUTION

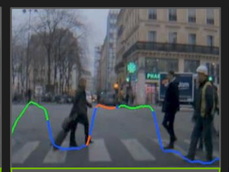
NVIDIA DRIVE
AV COMPUTING PLATFORM

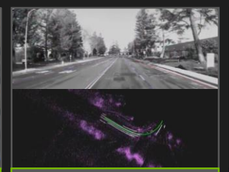
Sensor Fusion: RADAR, LIDAR, Camera
Deep Learning, CV, Parallel Computing
Diversity of Algorithms
ASIL-D Functional Safety
Fully Integrated into NVIDIA BB8

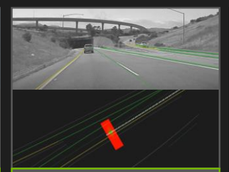
RADAR Fusion


LIDAR
Point-Cloud Processing


Camera Deep Learning
(Detection)


Camera Deep Learning
(Freespace)


Camera Computer Vision
(SLAM)


HD Map
(Localizing to Map)

Path Planning

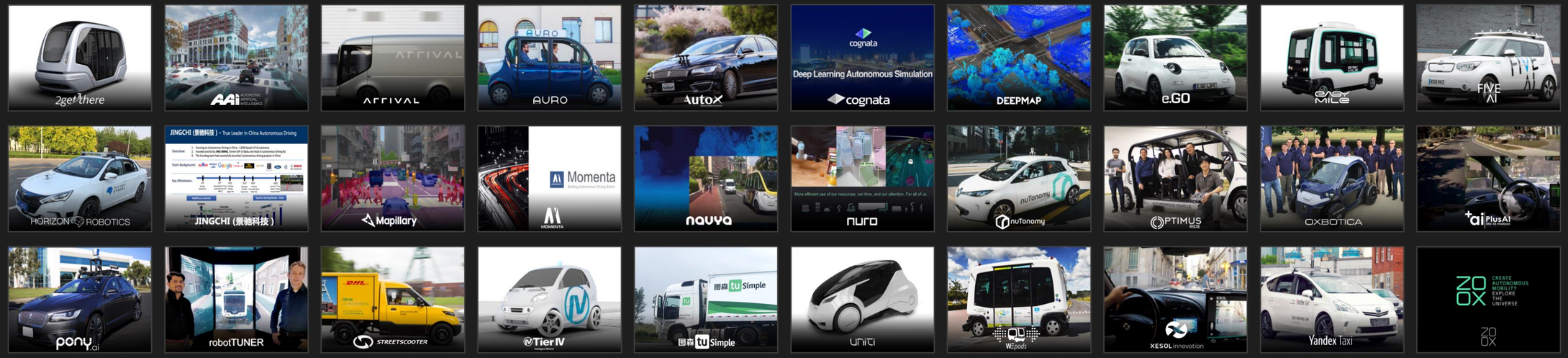
DRIVE AV

DRIVEWORKS SDK

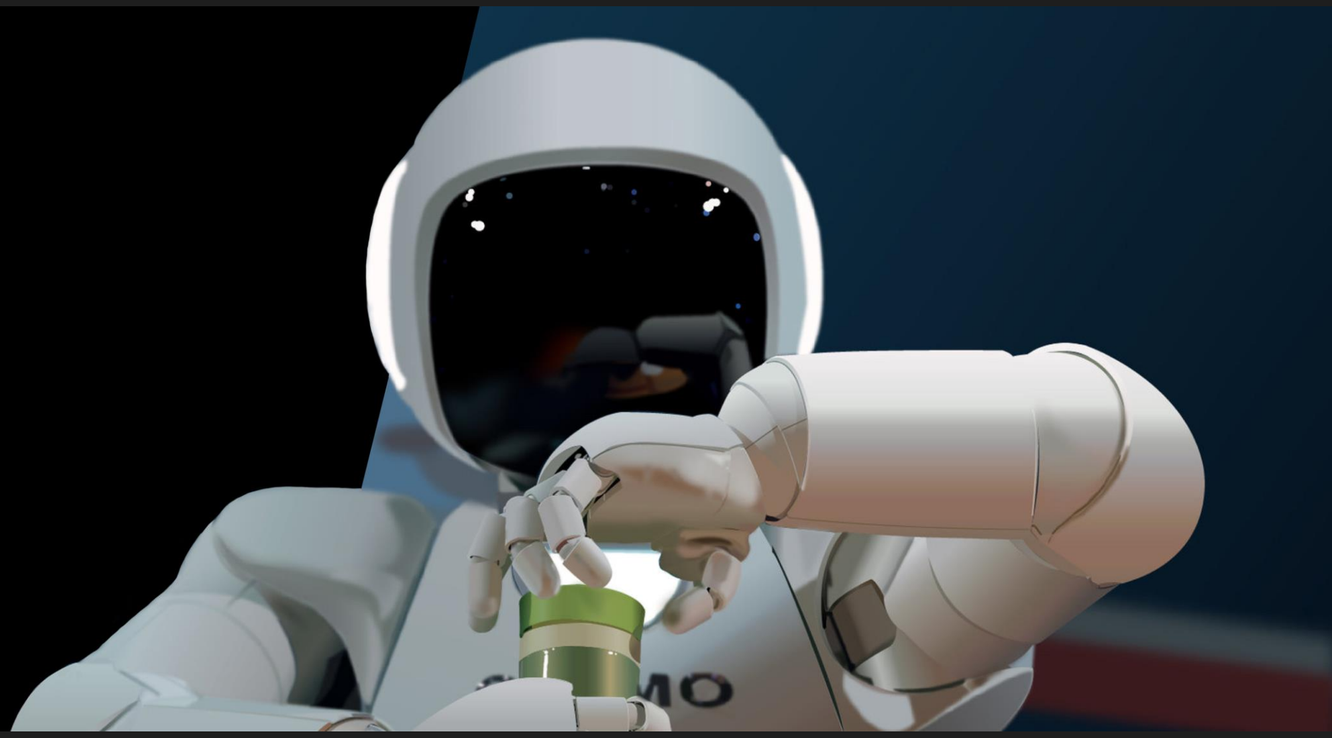
DRIVE OS

DRIVE PX – AI CAR COMPUTER

145 AV STARTUPS ON NVIDIA DRIVE



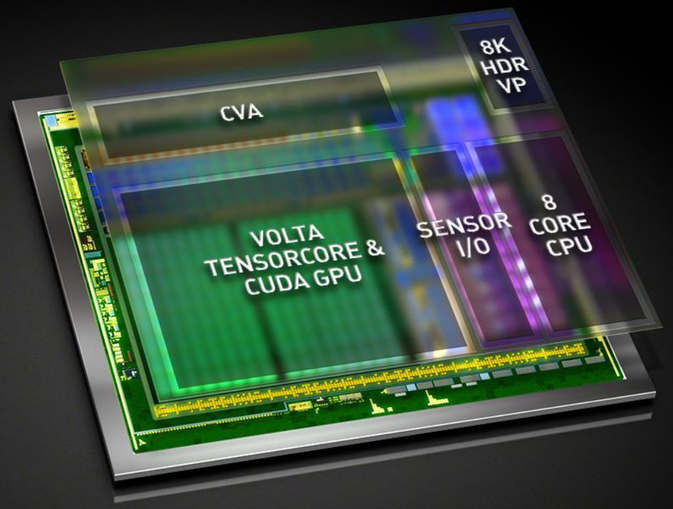
THE ERA OF AUTONOMOUS MACHINES



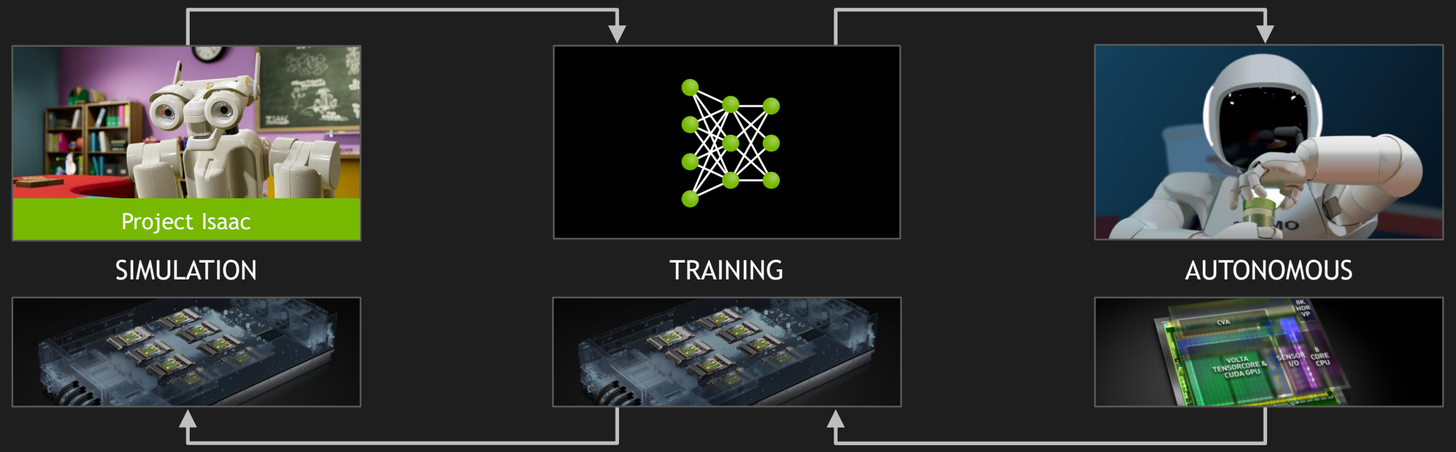
XAVIER

WORLD'S FIRST AUTONOMOUS MACHINE PROCESSOR

Deep Learning, CV, Parallel Computing
Rich High-Speed Sensor IOs
Extreme Energy Efficiency
30 TOPS at 30W



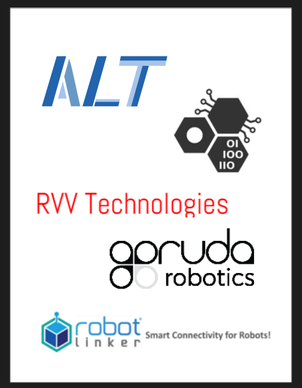
PROJECT ISAAC — AI ROBOT SIMULATOR



NVIDIA INCEPTION PARTNERS IN SEA

RECENTLY GROWN AI STARTUPS

Automotive



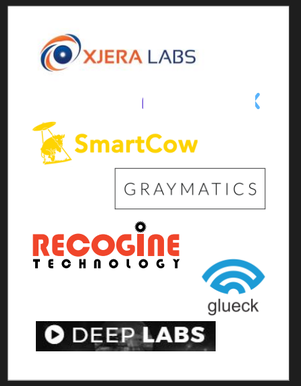
Drones / Robotics

Financial Services

Media / Entertainment

Retail / Etail

IT Services



Intelligent Video Analytics

Medical Imaging / Life Sciences

AI.PLATFORM@NSCC

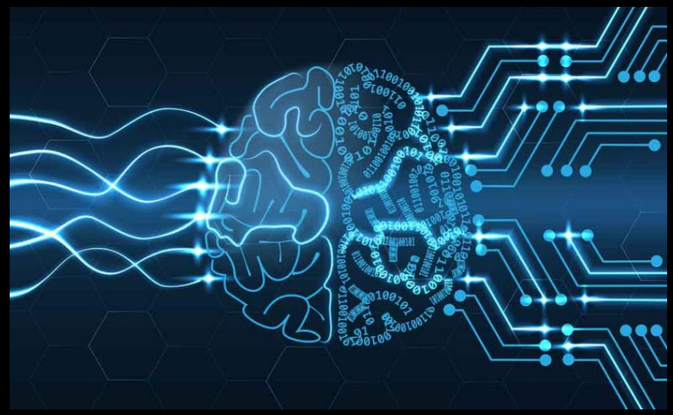
National Supercomputing Centre

Provides Computational Services for AI and Deep Learning Applications

Provides AI and Deep Learning Trainings

Provides Technical Expertise to Support AI Projects

AI SINGAPORE, AND OUR STAKEHOLDERS, WILL BE
THE MAIN USERS OF THIS FACILITIES FOR THE START



In collaboration with

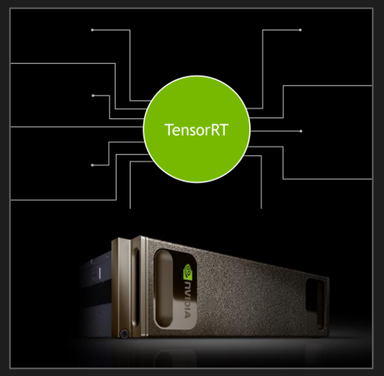
NVIDIA

Supported by

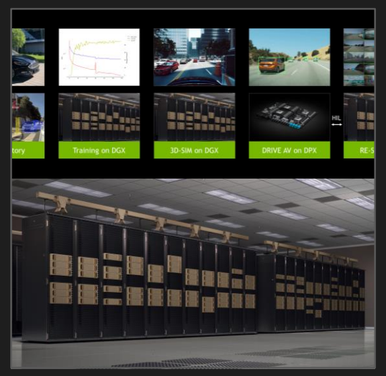
AI SINGAPORE

A NEW COMPUTING ERA

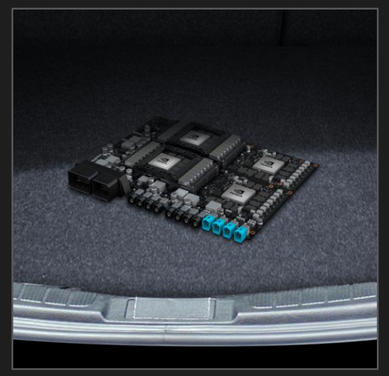
NVIDIA Holodeck
Design Lab of the Future



NVIDIA
The World's AI Computing Platform



NVIDIA DRIVE
Open AV Platform for the
Transportation Industry



NVIDIA DRIVE PX
Pegasus
Robotaxi AI Computer

Project Isaac
AI Robot Simulator

NVIDIA