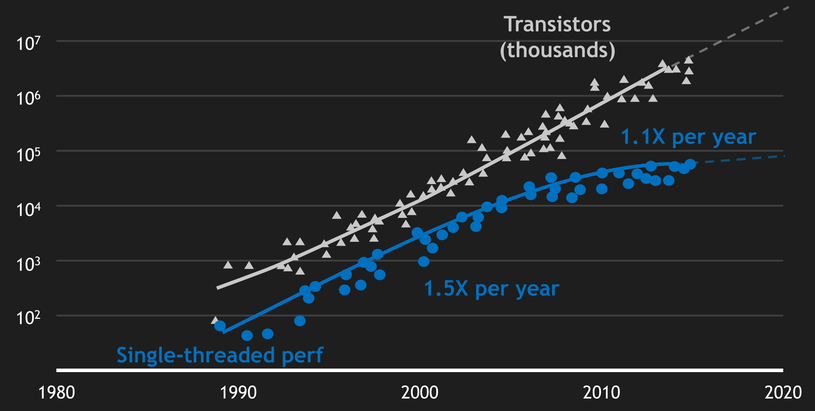
A NEW COMPUTING ERA

JENSEN HUANG, FOUNDER & CEO | GTC CHINA 2017

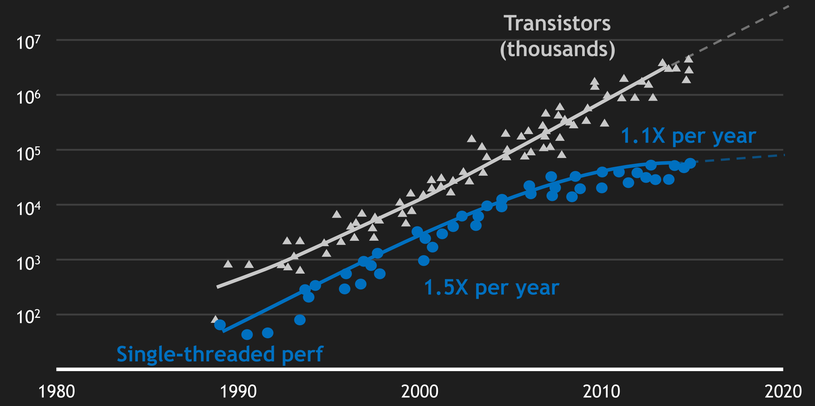
TWO FORCES DRIVING THE FUTURE OF COMPUTING



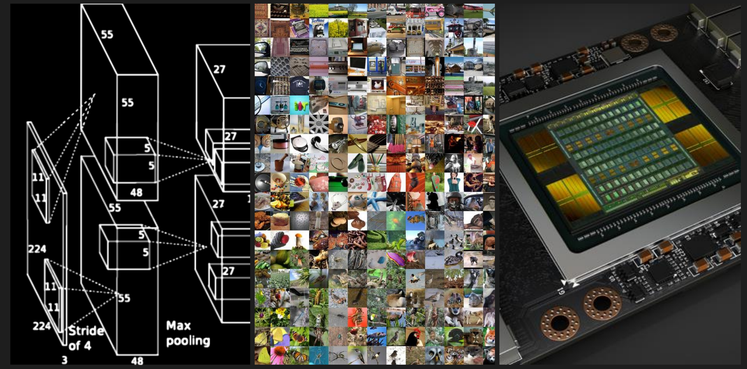
40 Years of Microprocessor Trend Data

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

TWO FORCES DRIVING THE FUTURE OF COMPUTING



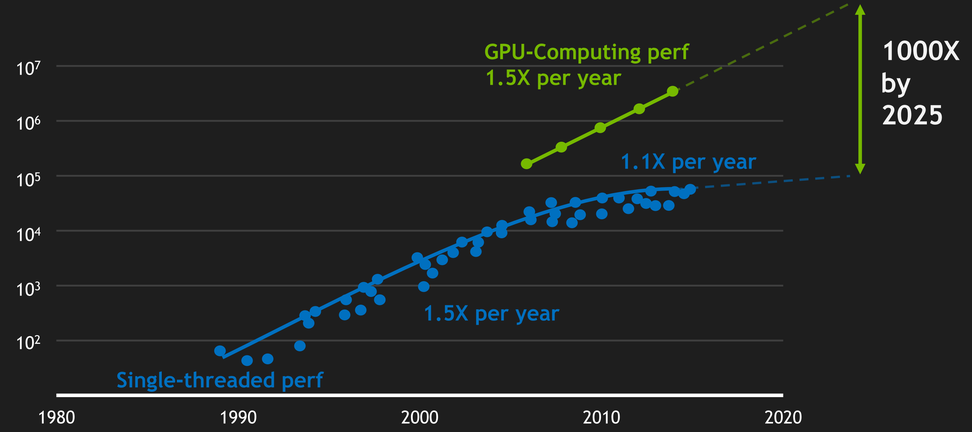
40 Years of Microprocessor Trend Data



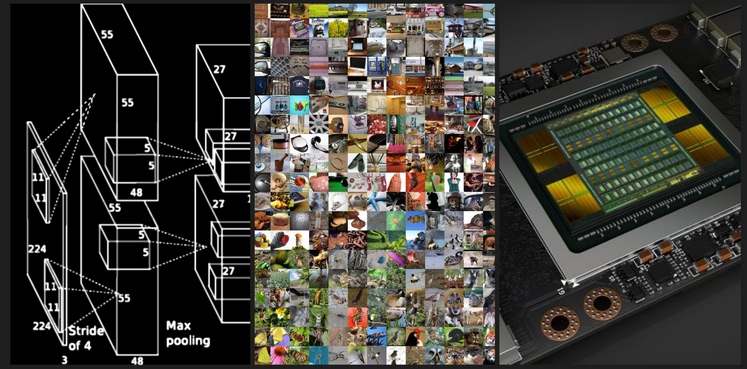
The Big Bang of Deep Learning

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 NVIDIA GPU COMPUTING



40 Years of Microprocessor Trend Data

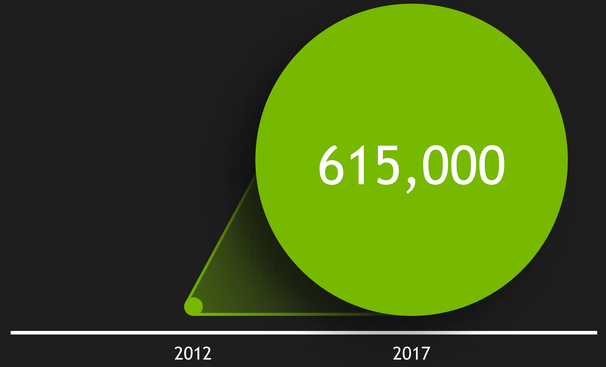


The Big Bang of Deep Learning

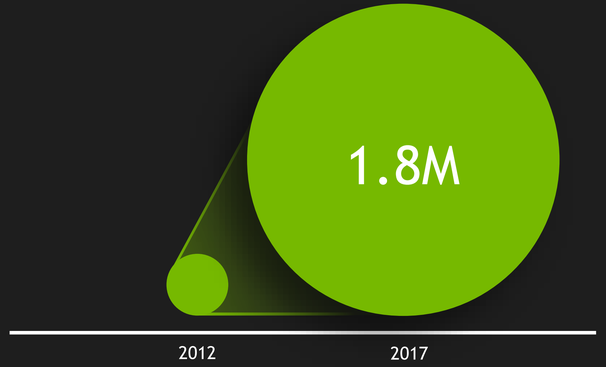
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

RISE OF NVIDIA GPU COMPUTING

Global GTC Attendees
10X in 5 Years



GPU Developers
14X in 5 Years

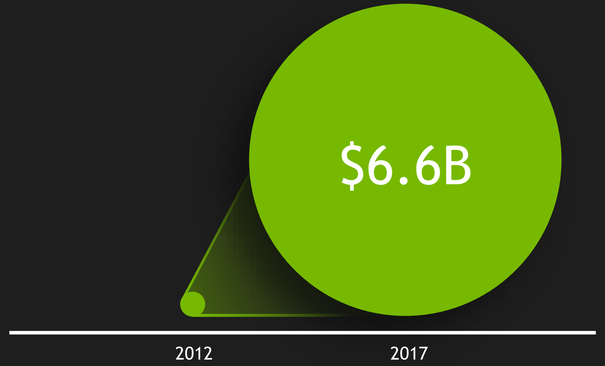


CUDA Downloads
5X in 5 Years, +800K in 1 Year

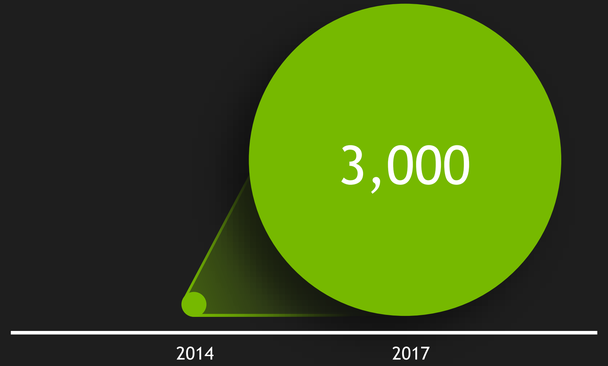
NVIDIA HOLODECK

Photorealistic Models
Physically Simulated Interaction
Virtual Team Collaboration
GPU-accelerated AI

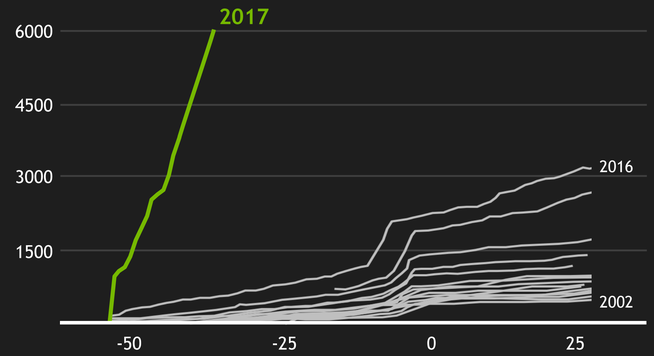
AI ON THE RISE



AI Startup Funding
12X in 5 Years



Deep Learning Papers Published
13X in 3 Years



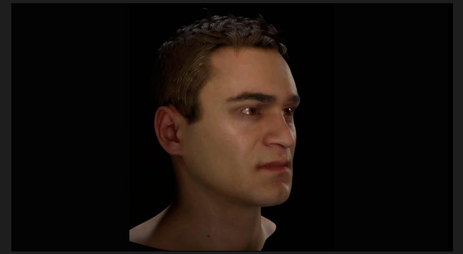
NIPS Registration

AMAZING ADVANCES IN AI

NVIDIA
Interactive Ray Tracing

AMAZING ADVANCES IN AI

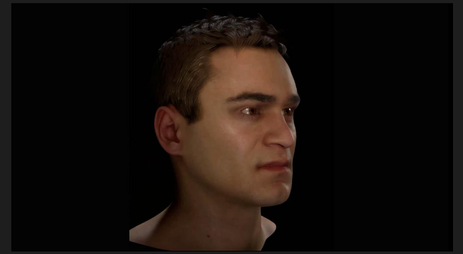
NVIDIA
Interactive Ray Tracing



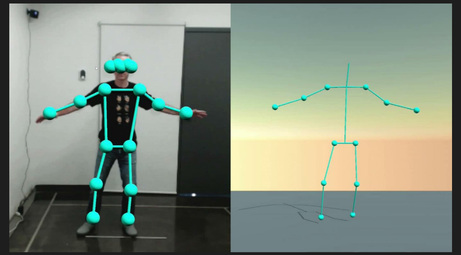
NVIDIA / Remedy
Audio-driven Facial Animation

AMAZING ADVANCES IN AI

NVIDIA
Interactive Ray Tracing



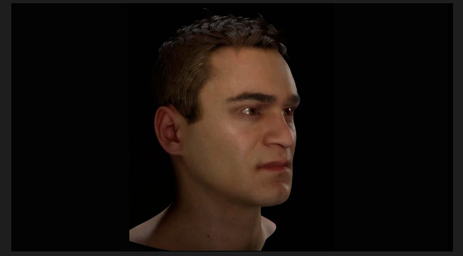
NVIDIA / Remedy
Audio-driven Facial Animation



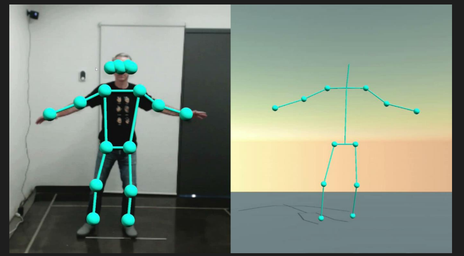
WRNCH
Pose Estimation

AMAZING ADVANCES IN AI

NVIDIA
Interactive Ray Tracing



NVIDIA / Remedy
Audio-driven Facial Animation

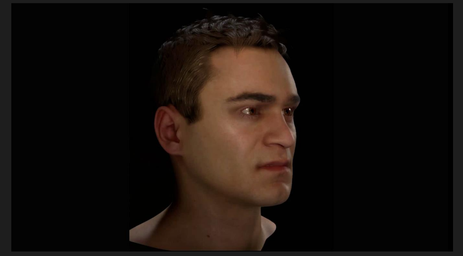


WRNCH
Pose Estimation

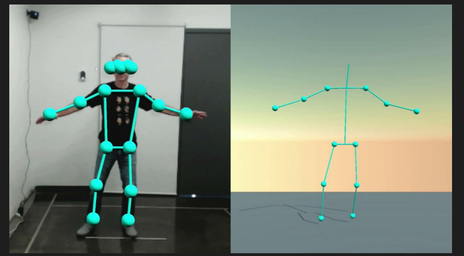
University of Edinburgh
Character Animation

AMAZING ADVANCES IN AI

NVIDIA
Interactive Ray Tracing

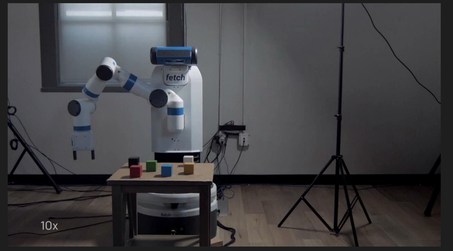


NVIDIA / Remedy
Audio-driven Facial Animation



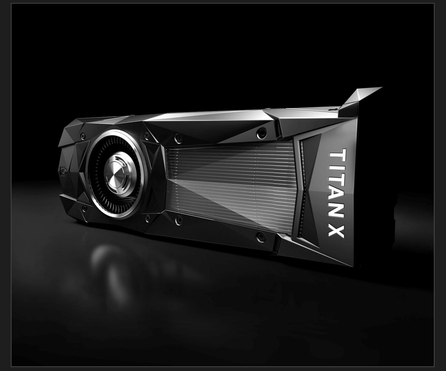
WRNCH
Pose Estimation

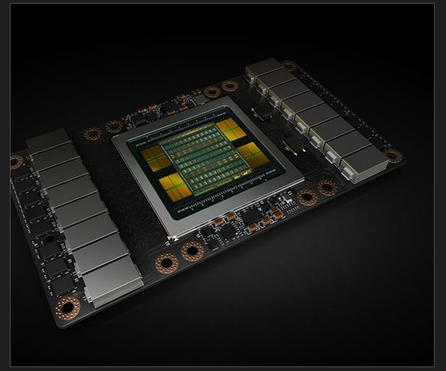
University of Edinburgh
Character Animation

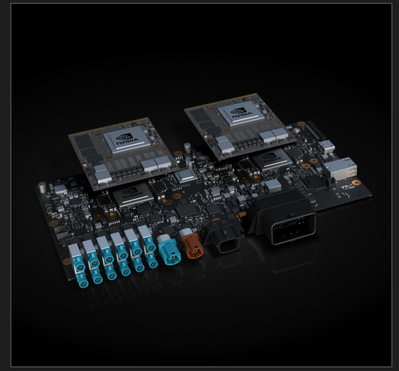


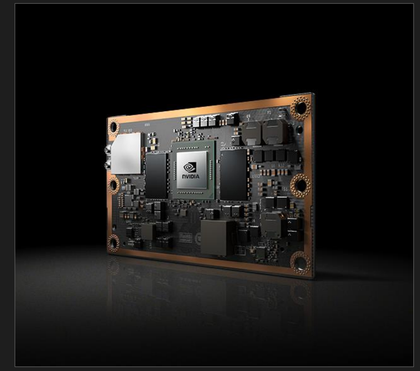
UC Berkeley / OpenAI
One-shot Imitation Learning

NVIDIA IS THE WORLD'S LEADING AI PLATFORM









ONE ARCHITECTURE — CUDA

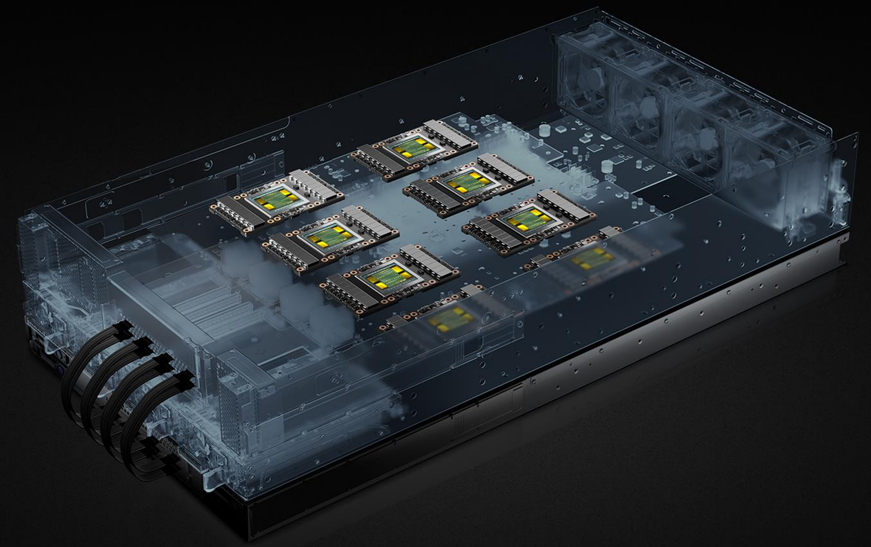
ANNOUNCING
CHINA TOP CSPs
ADOPT NVIDIA AI COMPUTING

2B USERS

Alibaba Cloud

百度云
cloud.baidu.com

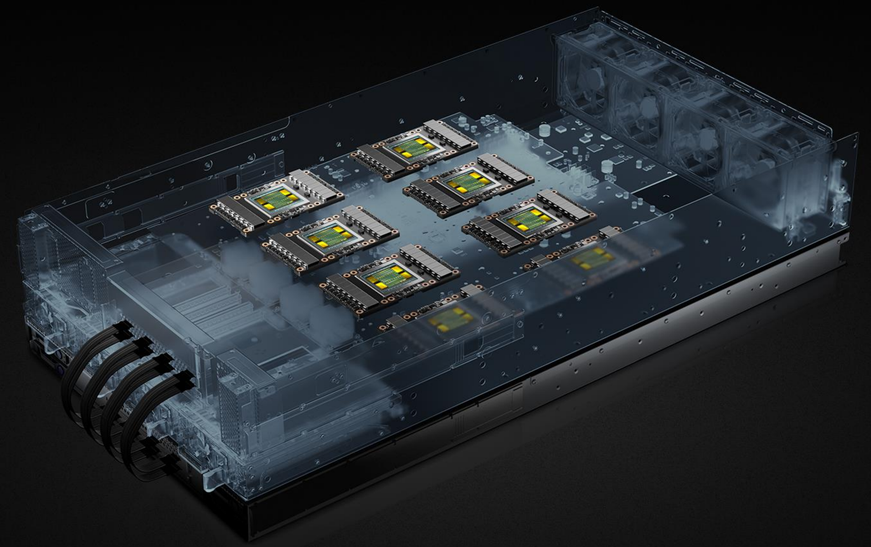
Tencent Cloud



ANNOUNCING CHINA ENTERPRISE IT LEADERS ADOPT NVIDIA AI COMPUTING

$200B CHINA IT MARKET

HUAWEI inspur Lenovo

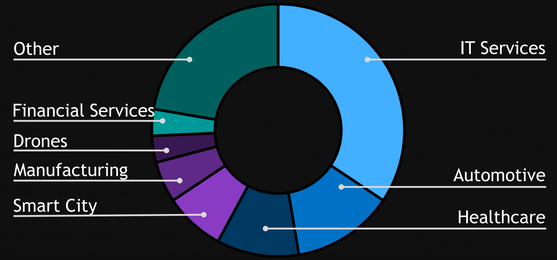


NVIDIA AI FOR DEVELOPERS EVERYWHERE

Every Framework

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

NVIDIA Inception: 1,900 DL Startups

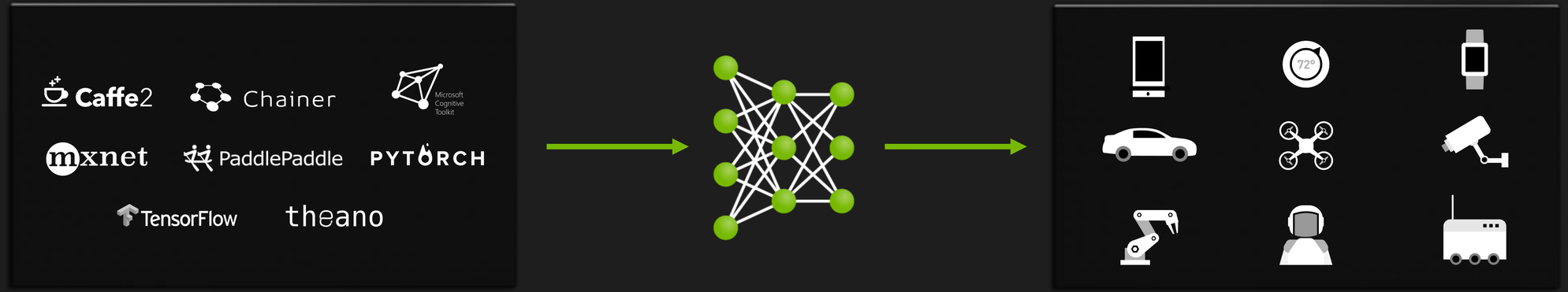


Every Cloud and Data Center

Alibaba Cloud · amazon web services · 百度云 cloud.baidu.com · Google Cloud Platform · Microsoft Azure · Tencent Cloud · cisco · hp · HUAWEI · IBM · inspur · Lenovo

NVIDIA AI PLATFORM

AI INFERENCE IS THE NEXT GREAT CHALLENGE



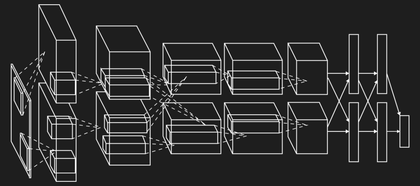
Training

DNN Model

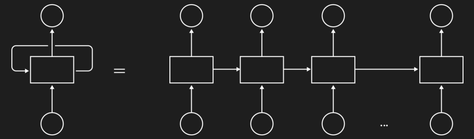
Inferencing

EXPLOSION OF NETWORK DESIGN

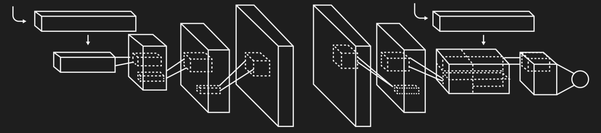
Convolution Networks

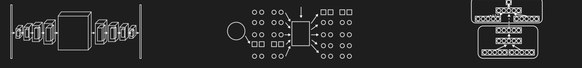


Recurrent Networks

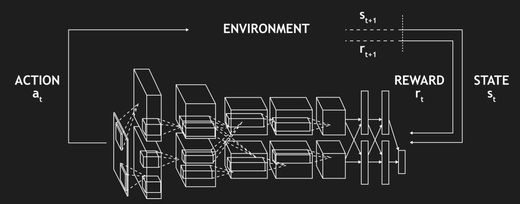


Generative Adversarial Networks



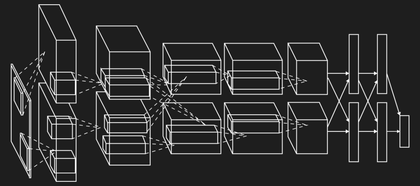


Reinforcement Learning



EXPLOSION OF NETWORK DESIGN

Convolution Networks



PRelu

ReLu

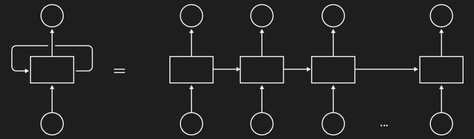
BatchNorm

Concat

Dropout

Pooling

Recurrent Networks



LSTM

GRU

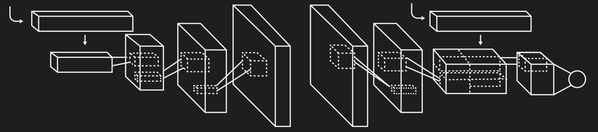
Highway

Projection

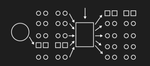
Embedding

BiDirectional

Generative Adversarial Networks

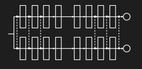


3D-GAN

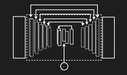


Rank GAN

Conditional GAN



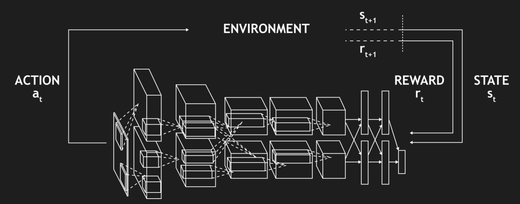
Coupled GAN



Speech Enhancement GAN

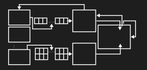
Latent space GAN

Reinforcement Learning



DQN

Dueling DQN



A3C

EXPLOSION OF NETWORK COMPLEXITY

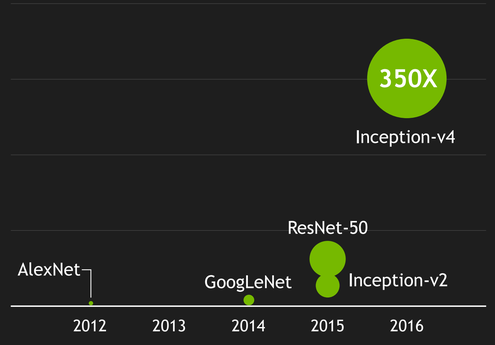
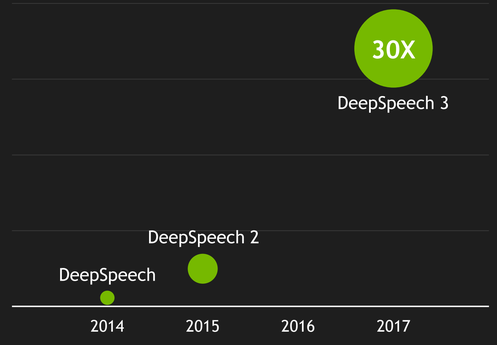
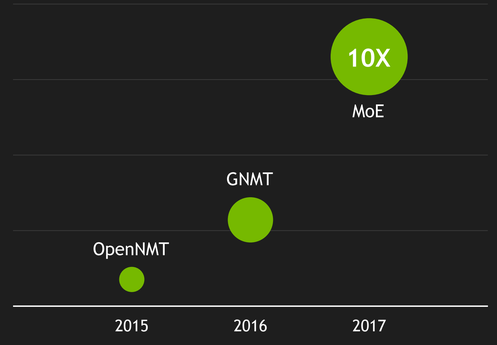


Image Network Complexity
GOPS * Bandwidth



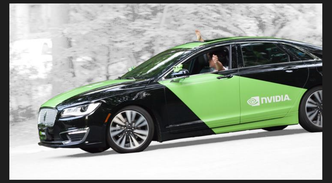
Speech Network Complexity
GOPS * Bandwidth



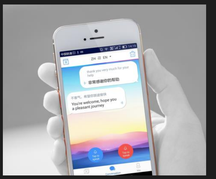
Translation Network Complexity
GOPS * Bandwidth

EXPLOSION OF INTELLIGENT MACHINES

20M Inference Servers



100s of Millions of Autonomous Machines

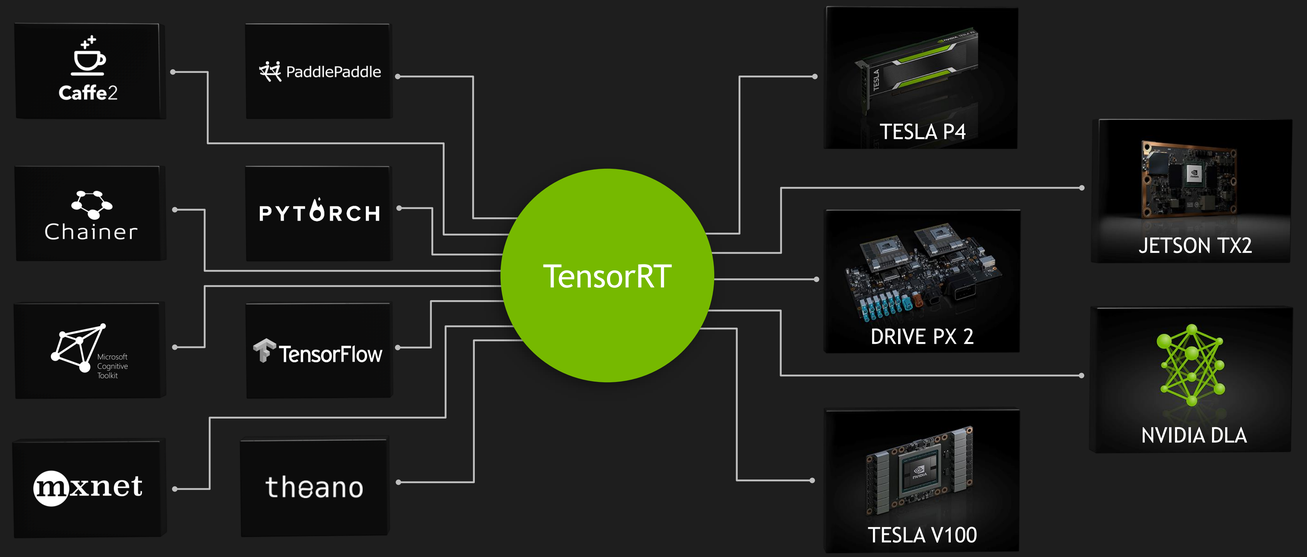


Trillions of IoT Devices

ANNOUNCING NVIDIA TENSORRT 3

PROGRAMMABLE INFERENCE ACCELERATOR

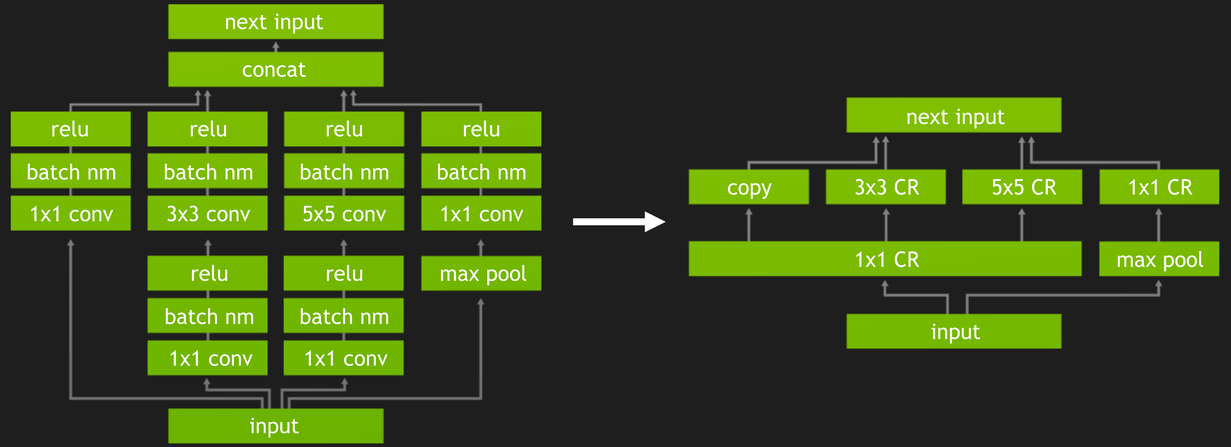
Compile and Optimize Neural Networks
Support for Every Framework
Optimize for Each Target Platform



ANNOUNCING NVIDIA TENSORRT 3

PROGRAMMABLE INFERENCE ACCELERATOR

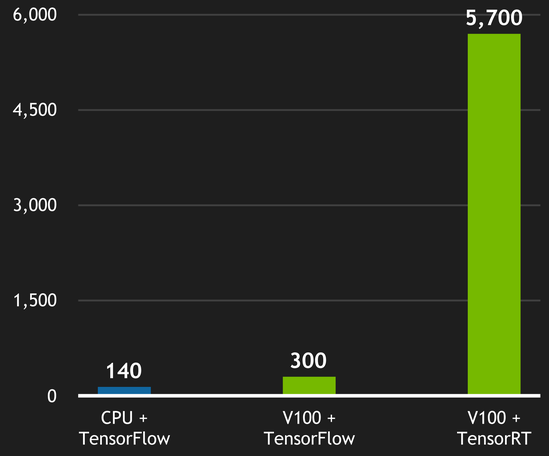
Weight & Activation Precision Calibration
Layer & Tensor Fusion
Kernel Auto-Tuning
Multi-Stream Execution



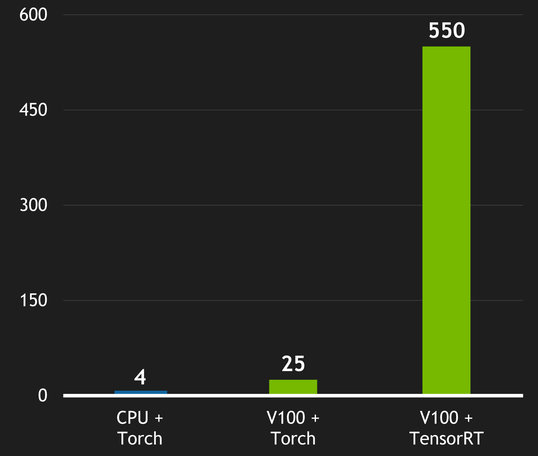
ANNOUNCING NVIDIA TENSORRT 3

PROGRAMMABLE INFERENCE ACCELERATOR

40x Speed-up on ResNet-50
140x Speed-up on OpenNMT



Images/Sec (ResNet-50)

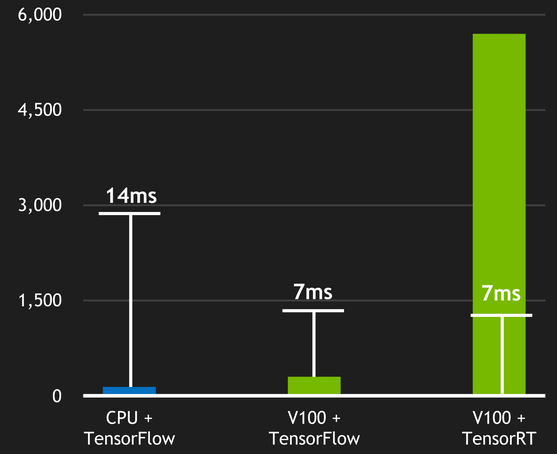


Sentences/Sec (OpenNMT)

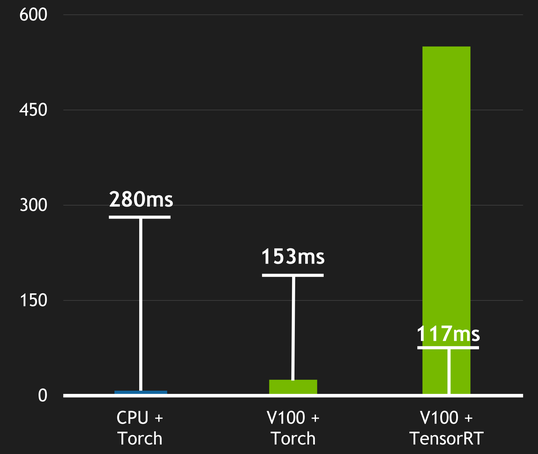
ANNOUNCING NVIDIA TENSORRT 3

PROGRAMMABLE INFERENCE ACCELERATOR

40x Speed-up on ResNet-50
140x Speed-up on OpenNMT



Images/Sec (ResNet-50)



Sentences/Sec (OpenNMT)

AI INFERENCE DRIVES DATACENTER TCO

160 CPU servers
45,000 images / second
65 KWatts

AI INFERENCE DRIVES DATACENTER TCO

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

4 Racks in a Box


Tensorflow on CPU
-Model: resnet_152_v1 [CPU]
-Graph: /models/frozen_resnet_v1_152.pb
-Size: [224,224,3]
INFERENCE
ON IMAGES
Images Per Sec : 0.171535
MPixels Per Sec: 0.0107209


Tensorflow on CPU
-Model: resnet_152_v1 [CPU]
-Graph: /models/frozen_resnet_v1_152.pb
-Size: [224,224,3]
Images Per Sec : 0.171535
MPixels Per Sec: 0.0107209

INFERENCE ON SPEECH


Listening ...
002:17 / 114:12

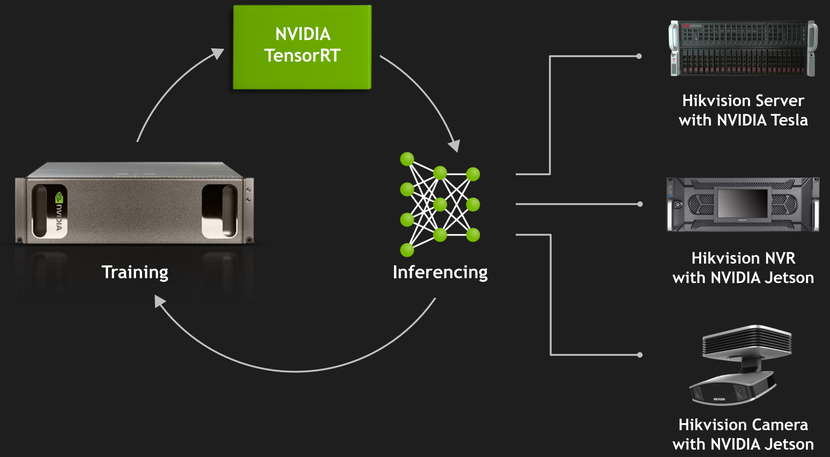
ANNOUNCING
CHINA CSPs ADOPT NVIDIA GPU-ACCELERATED INFERENCE PLATFORM

AI CITY

SMARTER, SAFER CITIES

1B cameras WW by 2020
Finding lost people
Improving traffic
Enhancing law enforcement

ANNOUNCING HIKVISION ADOPTS NVIDIA FOR AI CITY



NVIDIA AI CITY — SMARTER, SAFER CITIES IN CHINA

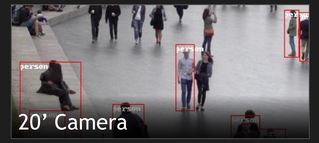
10X Larger Dataset for Face Recognition

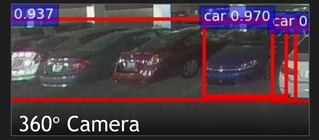
100s of Virtual Security Guards

15% Reduction in Peak Traffic

+90% Accuracy for Congestion Alerts

ADAPTING TO NEW USE CASES





Pre-Trained Network

Pre-Trained Network

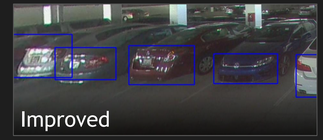
NVIDIA DIGITS

NVIDIA DGX Station

Optimized Network

NVIDIA TensorRT

NVIDIA Jetson



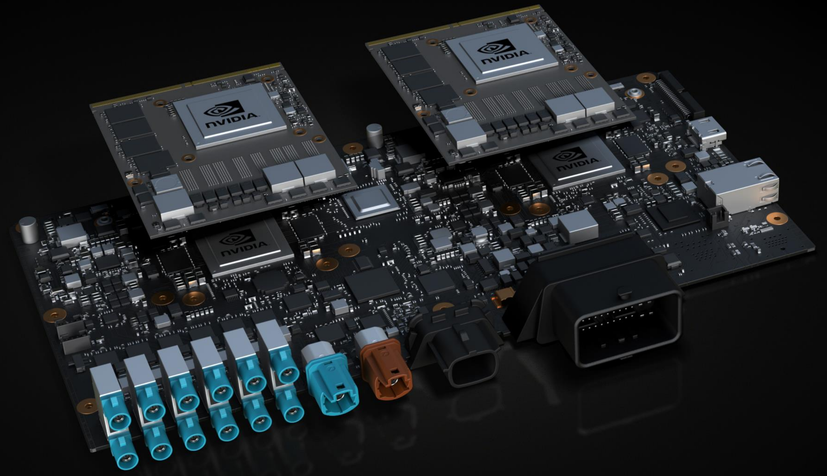
Optimized Network

THE AUTONOMOUS VEHICLE REVOLUTION

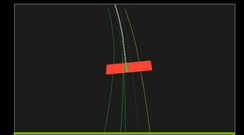
NVIDIA DRIVE

AV COMPUTING PLATFORM

Level 3 to Level 5
ASIL-D Functional Safety



Perception



Localization

Planning

DRIVE AV

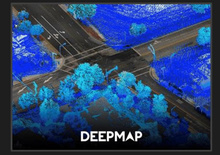
DRIVEWORKS SDK

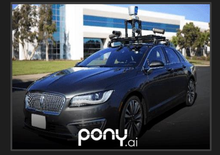
DRIVE OS

DRIVE PX

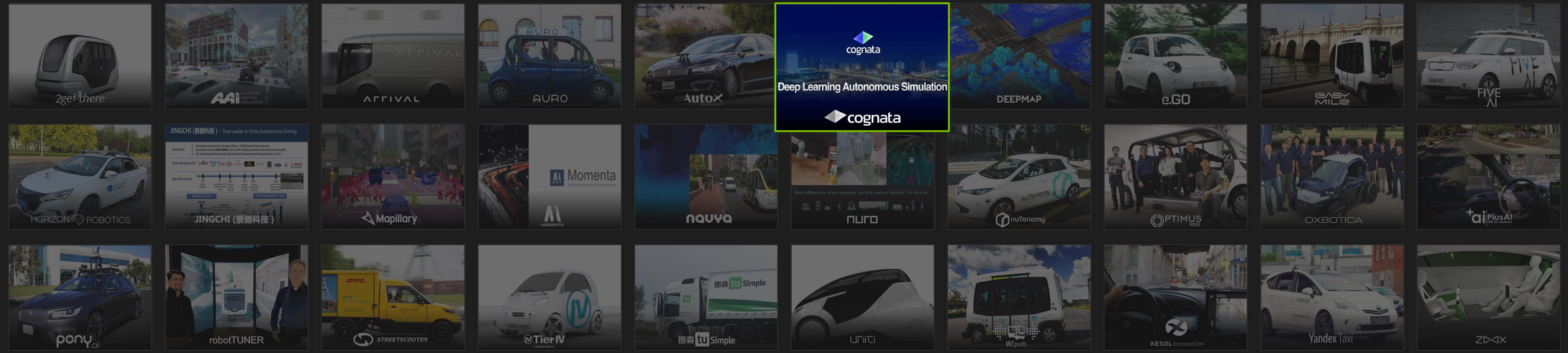

NVIDIA

145 AV STARTUPS ON NVIDIA DRIVE

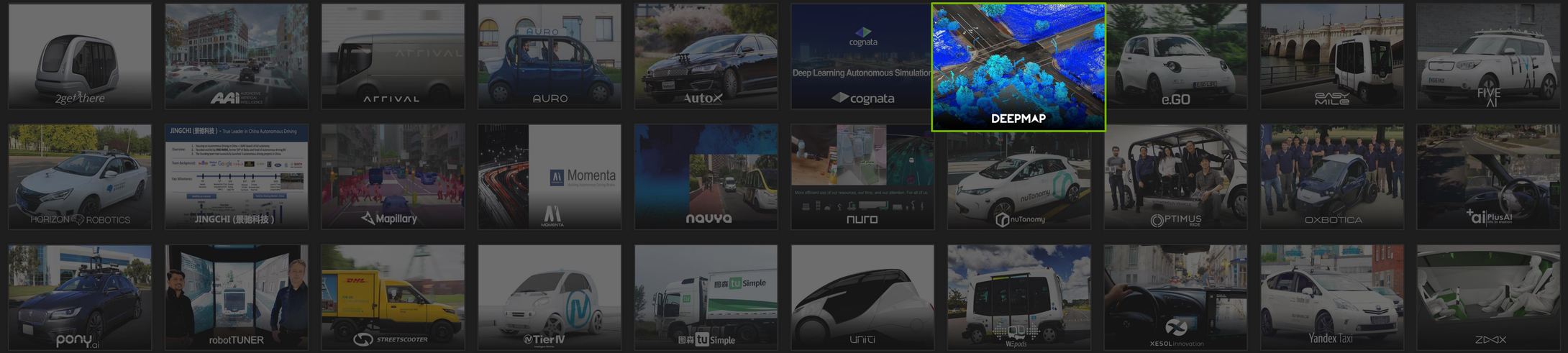




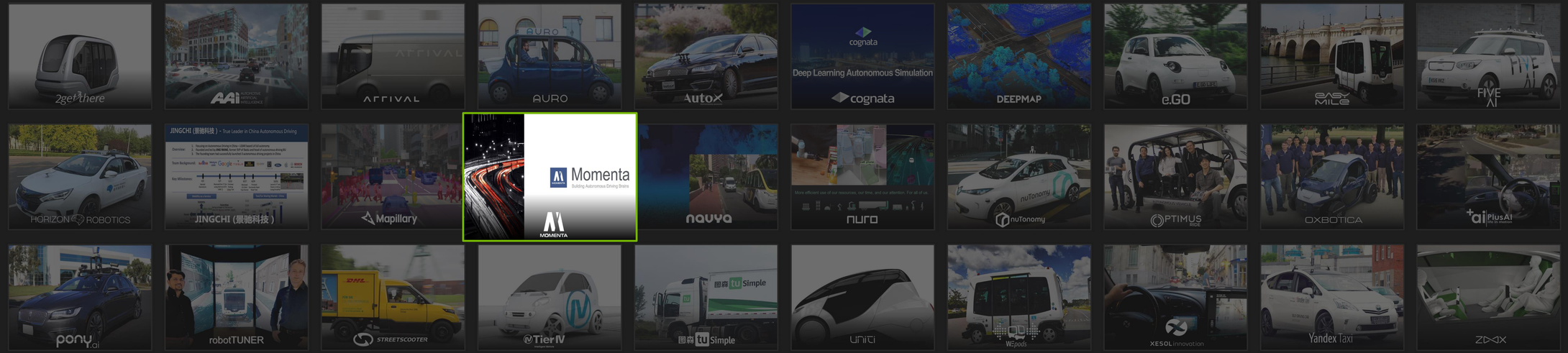
145 AV STARTUPS ON NVIDIA DRIVE



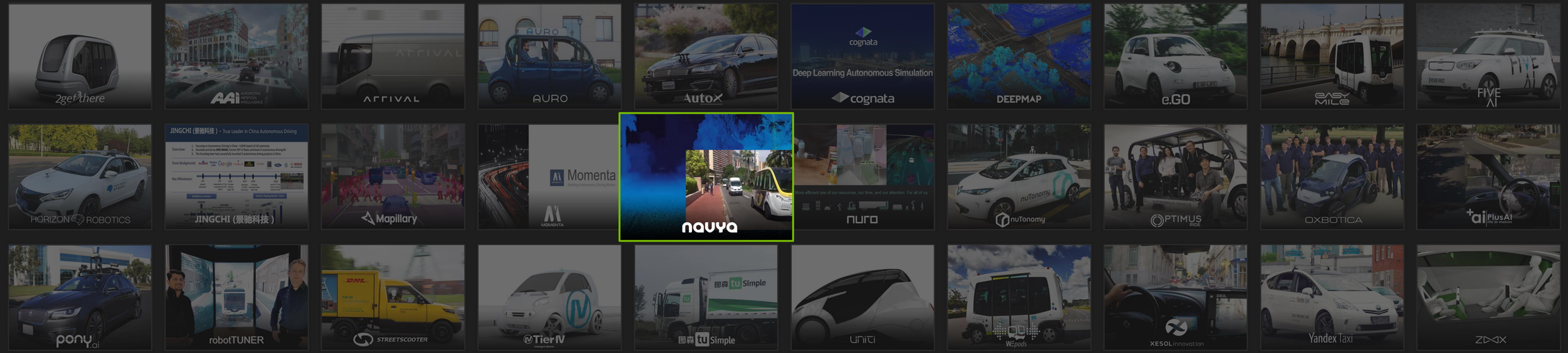
145 AV STARTUPS ON NVIDIA DRIVE



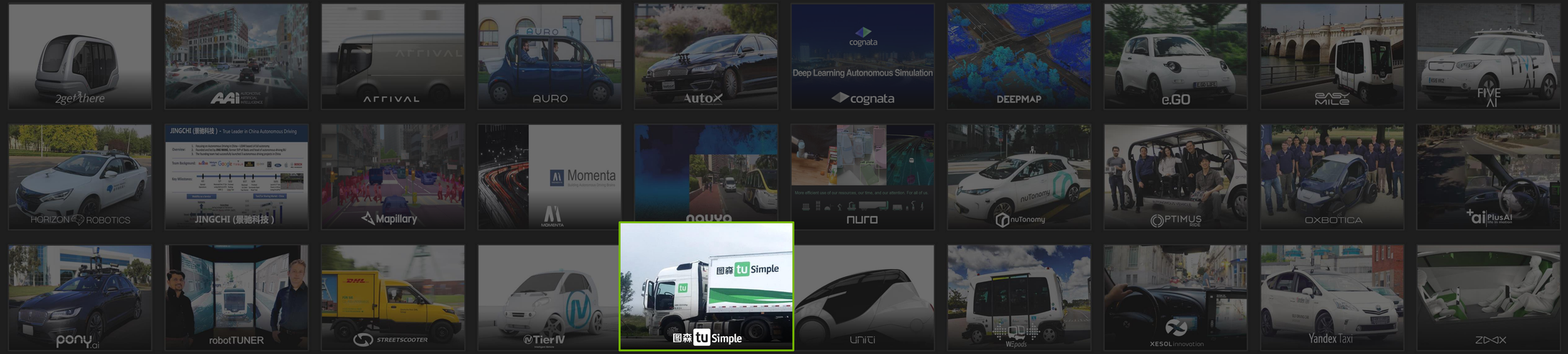
145 AV STARTUPS ON NVIDIA DRIVE



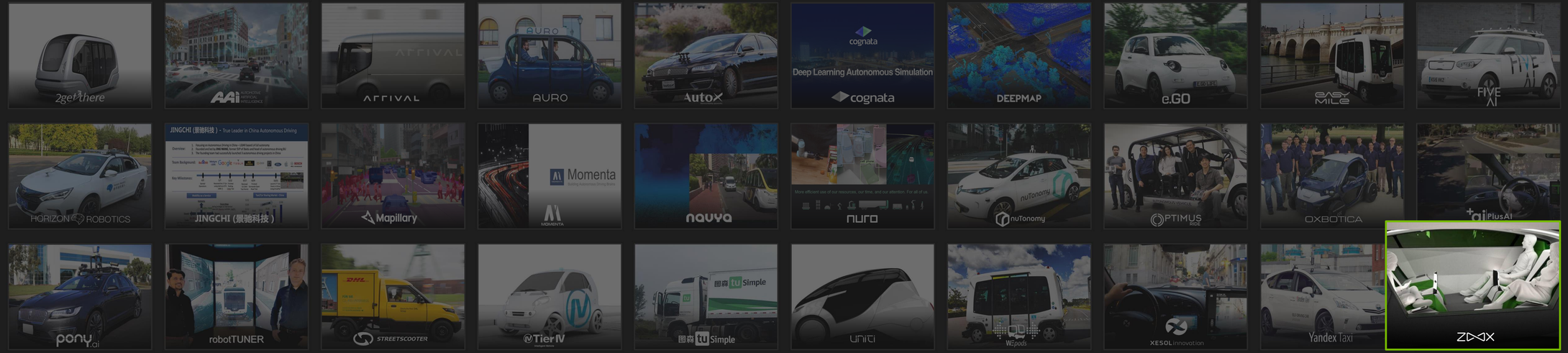
145 AV STARTUPS ON NVIDIA DRIVE



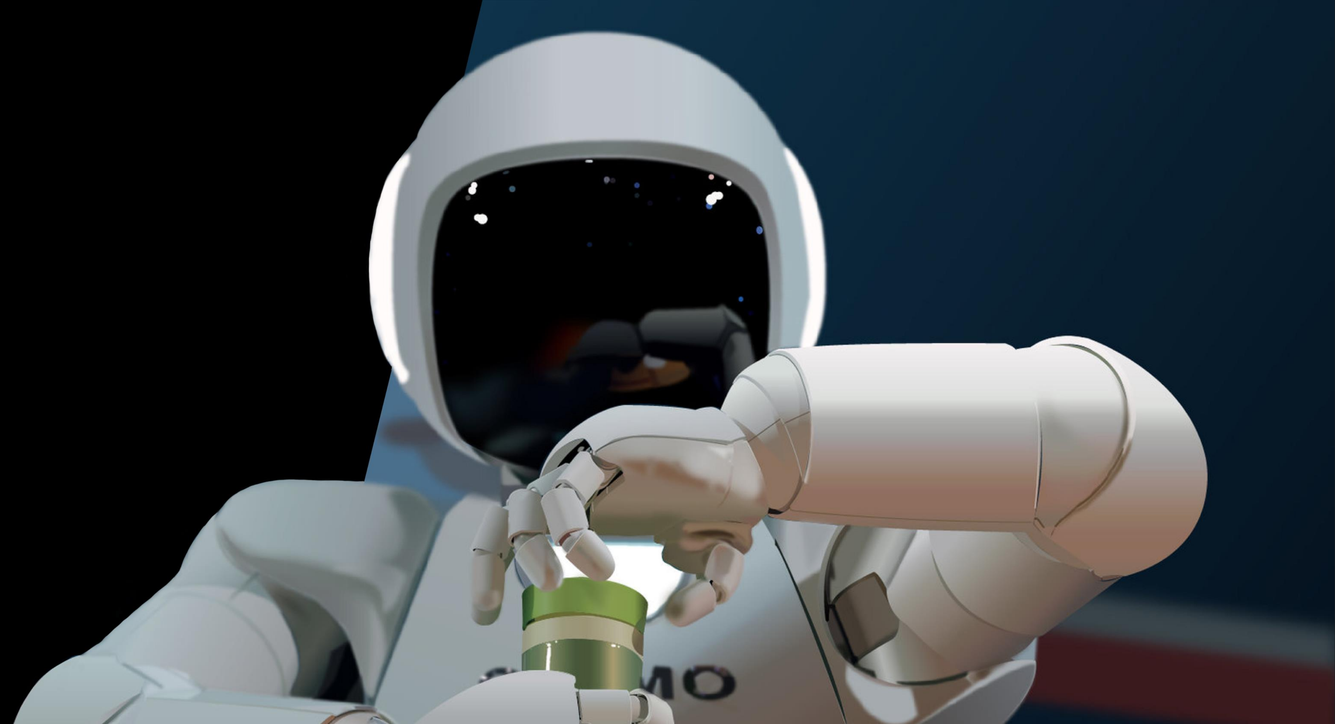
145 AV STARTUPS ON NVIDIA DRIVE



145 AV STARTUPS ON NVIDIA DRIVE



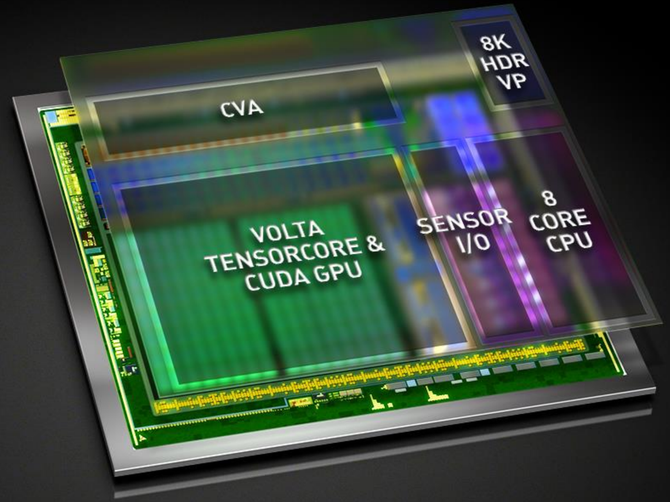
THE ERA OF AUTONOMOUS MACHINES



XAVIER

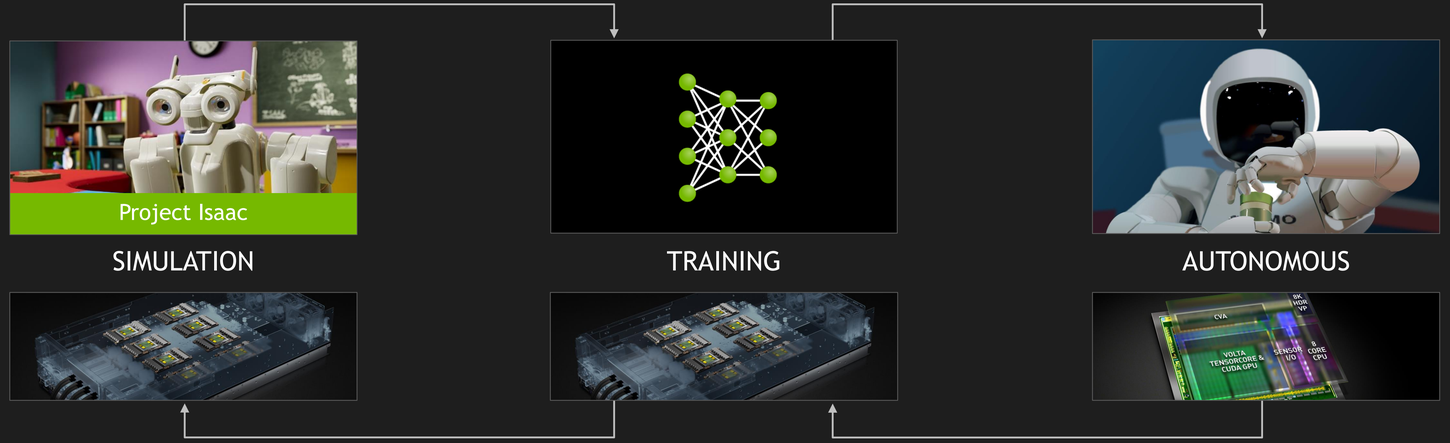
WORLD'S FIRST AUTONOMOUS MACHINE PROCESSOR

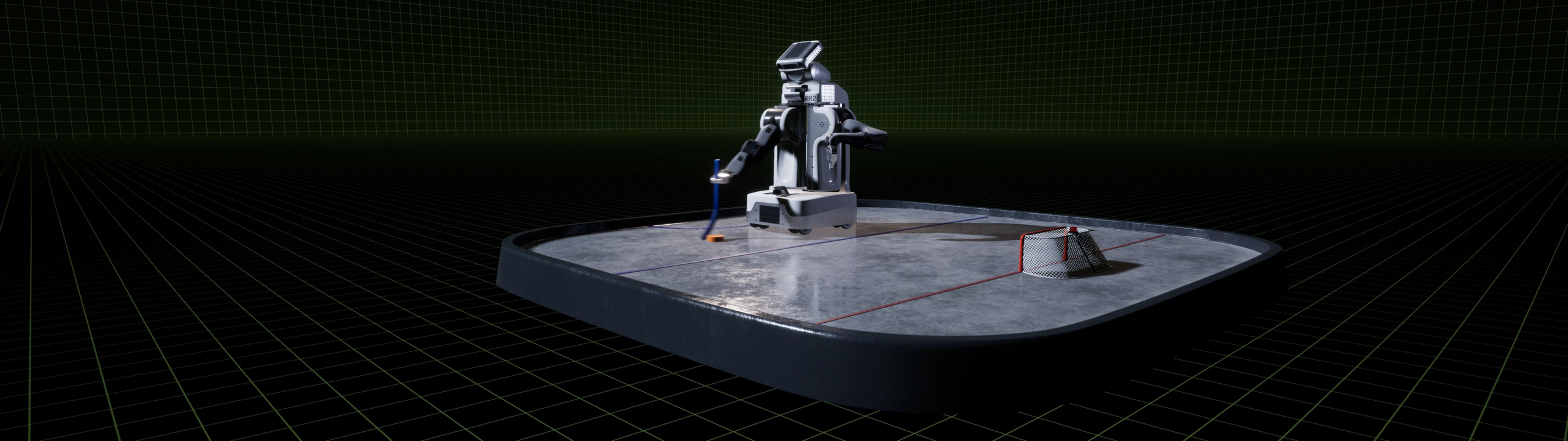
Supercomputing Performance
Extreme Energy Efficiency
Rich High-Speed Sensor IOs
30 TOPS of CV, Deep Learning, and Parallel Computing



ANNOUNCING
JD X SELECTS NVIDIA FOR AUTONOMOUS MACHINES

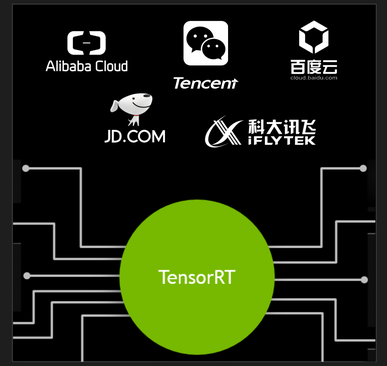
CREATING AUTONOMOUS MACHINES



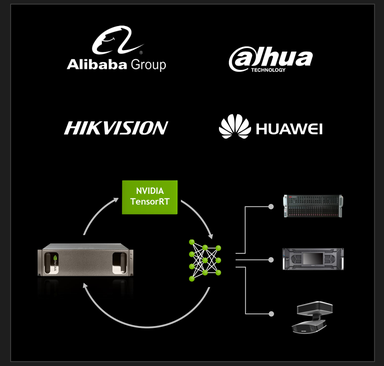

PR2

A NEW COMPUTING ERA

NVIDIA
The World's AI Computing Platform

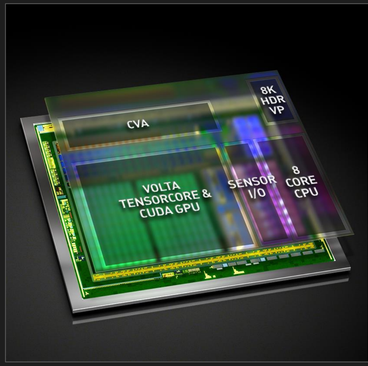


NVIDIA TensorRT
1st Programmable Inference Accelerator



NVIDIA TensorRT
AI Cities Platform

NVIDIA DRIVE
Open AV Platform for the Transportation Industry



Xavier
1st Autonomous Machine Processor


NVIDIA