A NEW COMPUTING ERA

JENSEN HUANG, FOUNDER & CEO | GTC CHINA 2017
TWO FORCES DRIVING THE FUTURE OF COMPUTING

40 Years of Microprocessor Trend Data

1.5X per year
1.1X per year

Transistors (thousands)

1.5X per year
1.1X per year

Single-threaded perf
TWO FORCES DRIVING THE FUTURE OF COMPUTING

The Big Bang of Deep Learning

40 Years of Microprocessor Trend Data

Transistors (thousands)

1.1X per year

1.5X per year

Single-threaded perf
RISE OF NVIDIA GPU COMPUTING

40 Years of Microprocessor Trend Data

- GPU-Computing perf: 1.5X per year
- Single-threaded perf: 1.5X per year
- 1.1X per year

1000X by 2025

The Big Bang of Deep Learning
RISE OF NVIDIA GPU COMPUTING

Global GTC Attendees
10X in 5 Years

2012: 22,000
2017: 615,000

GPU Developers
14X in 5 Years

2012: 22,000
2017: 1.8M

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

2012: 615,000
2017: 22,000
NVIDIA HOLODECK

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

- **AI Startup Funding**: $6.6B, 12X in 5 Years
- **Deep Learning Papers Published**: 3,000, 13X in 3 Years
- **NIPS Registration**: 2017
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
Tencent Cloud
ANNOUNCING CHINA ENTERPRISE IT LEADERS ADOPT NVIDIA AI COMPUTING

$200B CHINA IT MARKET
NVIDIA AI FOR DEVELOPERS EVERYWHERE

Every Framework
- Caffe2
- Chainer
- mxnet
- Paddle
- PyTorch
- TensorFlow
- theano

NVIDIA Inception: 1,900 DL Startups
- IT Services
- Financial Services
- Drones
- Manufacturing
- Smart City
- Automotive
- Healthcare
- Other

Every Cloud and Data Center
- Alibaba Cloud
- Google Cloud Platform
- Microsoft Azure
- Huawei Cloud
- IBM
- inspur
- Lenovo

NVIDIA AI PLATFORM
AI INFERENCE IS THE NEXT GREAT CHALLENGE

Training
- Caffe2
- Chainer
- mxnet
- PaddlePaddle
- PyTorch
- TensorFlow
- Theano

DNN Model

Inferencing
- Car
- Drone
- Robot
- Surveillance Camera
EXPLOSION OF NETWORK DESIGN

Convolution Networks

- PReLU
- ReLU
- BatchNorm

- Concat
- Dropout
- Pooling

- Projection
- Embedding
- BiDirectional

Recurrent Networks

- LSTM
- GRU
- Highway

- 3D-GAN
- Rank GAN
- Conditional GAN

- Coupled GAN
- Speech Enhancement GAN
- Latent space GAN

Generative Adversarial Networks

- Generator
- Discriminator
- Coupled GAN
- Speech Enhancement GAN
- Latent space GAN

Reinforcement Learning

- A3C
- Dueling DQN
- DQN
- 3D-GAN

- ENVIRONMENT
- ACTION
- REWARD
- STATE

- Embedding
- BiDirectional
- Concat
- Dropout
- Pooling

- Projection
- Embedding
- BiDirectional
EXPLOSION OF NETWORK COMPLEXITY

Image Network Complexity
GOPS * Bandwidth

Speech Network Complexity
GOPS * Bandwidth

Translation Network Complexity
GOPS * Bandwidth

- AlexNet
- GoogLeNet
- Inception v2
- Inception v4
- ResNet 50
- DeepSpeech 1
- DeepSpeech 2
- DeepSpeech 3
- MoE
- GNMT
- OpenNMT
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 INFECTION ACCELERATOR

40x Speed-up on ResNet-50
140x Speed-up on OpenNMT
ANNOUNCING NVIDIA TENSORRT 3
PROGRAMMABLE INFERENCE ACCELERATOR

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

Images/Sec (ResNet-50)  
CPU + TensorFlow: 14ms  
V100 + TensorFlow: 7ms  
V100 + TensorRT: 7ms

Sentences/Sec (OpenNMT)  
CPU + Torch: 280ms  
V100 + Torch: 153ms  
V100 + TensorRT: 117ms
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
INFERENCE ON IMAGES
INERENCE ON SPEECH
Listening ...
ANNOUNCING
CHINA CSPs ADOPT NVIDIA GPU-ACCELERATED INFERENCE PLATFORM

Alibaba Cloud
Tencent
Baidu Cloud
ecloud.baidu.com
JD.COM
iFlytek
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

Hikvision Camera with NVIDIA Jetson

Hikvision NVR with NVIDIA Jetson

Hikvision Server with NVIDIA Tesla

Training

Inferencing

NVIDIA TensorRT
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 → NVIDIA DIGITS → Optimized Network → NVIDIA Jetson

20° Camera → 360° Camera

Indoor Camera

NVIDIA DGX Station

NVIDIA TensorRT

Optimized Network

Improved
THE AUTONOMOUS VEHICLE REVOLUTION
NVIDIA DRIVE
AV COMPUTING PLATFORM
Level 3 to Level 5
ASIL-D Functional Safety
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 SELCTS NVIDIA FOR AUTONOMOUS MACHINES
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