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

- Transistors (thousands)
  - 1.1X per year
  - 1.5X per year

Single-threaded perf

The Big Bang of Deep Learning
RISE OF NVIDIA GPU COMPUTING

40 Years of CPU Trend Data

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

NVIDIA Volta GPU | 21B xtors | 120 TFLOPS

RISE OF NVIDIA GPU COMPUTING

- Global GTC Attendees: 10X in 5 Years
  - 2012: 22,000
  - 2017: 645,000

- GPU Developers: 15X in 5 Years
  - 2012: 22,000
  - 2017: 1.8M

- CUDA Downloads: 5X in 5 Years
  - 2012: 645,000
  - 2017: 1.8M
NVIDIA GPU-ACCELERATED APPLICATIONS

Computer Graphics

Scientific Computing

Deep Learning

Data Science

NVIDIA CUDA
NVIDIA GPU-ACCELERATED INDUSTRIES
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

$6.6B

3,000
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

NVIDIA Inception: 1,900 DL Startups

Other
Financial Services
Energy
Manufacturing
Smart City
Automotive
Healthcare
IT Services

Every Cloud and Data Center

NVIDIA AI PLATFORM
AI INFERENCE IS THE NEXT GREAT CHALLENGE
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
- 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

2011 2013 2015 2017

Speech Network Complexity
GOPS * Bandwidth

2013 2014 2015 2016 2017 2018

Translation Network Complexity
GOPS * Bandwidth

2014 2015 2016 2017 2018

- AlexNet
- GoogleNet
- Inception-v2
- ResNet-50
- DeepSpeech 2
- DeepSpeech 3
- GNMT
- MoE
- OpenNMT

Inception-v4
DeepSpeech

350X 30X 10X

30X 10X
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
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
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

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
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