A NEW COMPUTING ERA

Shanker Trivedi | Senior Vice President | Enterprise Business at NVIDIA
THE ERA OF AI
TWO FORCES DRIVING
THE FUTURE OF COMPUTING

Deep Learning Starts AI Revolution

RISE OF NVIDIA GPU COMPUTING

CUDA GPU
Parallel Domain-Specialized Accelerator
High Compute & Bandwidth
High-Throughput GPU Plus Low-Latency CPU

GPU Extends Performance Post-Moore’s Law

1.5X per year
1.1X per year

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
2012: 22,000
2017: 645,000
10X in 5 Years

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

CUDA Downloads
2012: 645,000
2017: 1.8M
5X in 5 Years
NVIDIA GPU ACCELERATES 2017 NOBEL PRIZES IN CHEMISTRY AND PHYSICS

Cryogenic Electron Microscopy
Jacques Dubochet, Joachim Frank, Richard Henderson

Resolution before 2013
Resolution at present

Detection of Gravitational Waves
Rainer Weiss, Barry Barish, Kip Thorne
AI — CUDA GPU’S NEXT KILLER APP

- **AI Startup Funding**: $6.6B, 10X in 5 Years
- **Deep Learning Papers Published**: 3,000, 10X in 3 Years
- **NIPS Registration**: 2017
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
- DeepSpeech 2
- DeepSpeech 3
- OpenNMT
- GNMT
- MoE


2014 2015 2016 2017

2015 2016 2017
THE WORLD’S AI PLATFORM

Every Framework

- Caffe2
- Chainer
- PaddlePaddle
- PyTorch
- TensorFlow
- Theano

NVIDIA Inception: 2,000 DL Startups

- IT Services
- Automotive
- Healthcare
- Smart City
- Manufacturing
- Drones
- Fin. Services
- Other

Every Cloud and Data Center

- Alibaba Cloud
- AWS
- Microsoft Azure
- Google Cloud
- Cisco
- Huawei
- Lenovo
- Dell
NVIDIA GPU CLOUD
GPU-ACCELERATED CLOUD PLATFORM
OPTIMIZED FOR DEEP LEARNING

Containerized in NV Docker
Optimization Across the Full Stack
Always Up-to-Date
Fully Tested and Maintained by NVIDIA
Coming this Month
Sign up now: www.nvidia.com/gpu-cloud
AI INFEERENCE IS THE NEXT GREAT CHALLENGE
EXPLOSION OF INTELLIGENT MACHINES

20M Inference Servers
100s of Millions of Autonomous Machines
Trillions of IoT Devices
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

Images/Sec (ResNet-50)

- CPU + TensorFlow: 140
- V100 + TensorRT: 5,700

Sentences/Sec (OpenNMT)

- CPU + Torch: 4
- V100 + TensorRT: 550

40x Speed-up on ResNet-50 | 140x Speed-up on OpenNMT
NVIDIA TENSORRT
10X BETTER
DATA CENTER TCO

160 CPU Servers
45,000 Images / Second
65 KWatts
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
NVIDIA VOLTA IN EVERY CLOUD, EVERY DATACENTER
NVIDIA TESLA DATACENTER MARKETS

$12B Market
HPC

80% of Apps by 2020
CSP TRAINING

20M Inference Servers
CSP INFEERENCE

$25B Market
PUBLIC CLOUD

600M Amazon Packages / Yr
INDUSTRIES

$3T IT Industry
ENTERPRISE
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
STATE-OF-THE-ART DRIVERLESS VEHICLES
NEW "PEGASUS"
ROBOTAXI DRIVE PX

320 TOPS CUDA TensorCore | 16x GMSL | 4x 10G | 8x 1G | 16x 100M | Auto-grade | ASIL D
500W | Late Q1 Early Access Partners
Supercomputing Data Center in Your Trunk
Size of a License Plate
NEW “PEGASUS”
ROBOTAXI DRIVE PX

<table>
<thead>
<tr>
<th>Component</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>320 TOPS CUDA TensorCore</td>
<td></td>
</tr>
<tr>
<td>16x GMSL</td>
<td></td>
</tr>
<tr>
<td>4x 10G</td>
<td></td>
</tr>
<tr>
<td>8x 1G</td>
<td></td>
</tr>
<tr>
<td>16x 100M</td>
<td></td>
</tr>
<tr>
<td>Auto-grade</td>
<td>ASIL D</td>
</tr>
<tr>
<td>500W</td>
<td>Late Q1</td>
</tr>
<tr>
<td>Early Access Partners</td>
<td></td>
</tr>
</tbody>
</table>

Supercomputing Data Center in Your Trunk
Size of a License Plate
THE ERA OF AUTONOMOUS MACHINES
10% of Manufacturing Tasks Are Automated
1M Pizzas Delivered Per Day by Domino’s
100M People 80+ Years Old
Ag Tech: 70% Increase in Farm Yields by 2050
600K Bridges to Inspect in the U.S.
300M Operations per Year WW
NVIDIA JETSON AUTONOMOUS MACHINE PLATFORM

Jetson TX2

JetPack SDK

DIGITS

Isaac Robot Simulator

Deep Learning Institute
ADAPTING TO NEW USE CASES

Pre-Trained Network → NVIDIA DIGITS → Optimized Network → NVIDIA TensorRT → NVIDIA Jetson

- 20° Camera
- Indoor Camera
- 360° Camera

Improved

Optimized Network
AI CITY
SMARTER, SAFER CITIES

1B cameras WW by 2020
Finding lost people
Improving traffic
Enhancing law enforcement
SAFE AND SMART CITIES IS AN AI PROBLEM

1B installed security cameras WW (2020)
30B frames per day

Challenging real-world conditions
Traditional video analytics not trustworthy

Image Classification

Accuracy

Human
Hand-coded CV
Deep Learning

Al achieves superhuman results
Al-driven intelligent video analytics
NVIDIA AI CITY PLATFORM ADOPTION

Industry’s first search by example

30x faster than real-time video synopsis

6x improvement for pedestrian detection in rain

5x speed-up for ALPR

10x speed-up in vehicle attribute classification

11x boost in investigation productivity

30x speed-up in people and attribute detection

World-leading object detection
A NEW COMPUTING ERA

NVIDIA AI
Volta in Every Cloud | NVIDIA GPU Cloud Registry | DGX-1 | DGX Station | TensorRT Programmable Inference Accelerator

NVIDIA AUTONOMOUS MACHINES
Jetson TX2 | DRIVE PX “PEGASUS” | Deep Learning Institute

NVIDIA AI CITY
Metropolis | AI-driven Intelligent Video Analytics