A NEW COMPUTING MODEL
JEN-HSUN HUANG, CO-FOUNDER & CEO | GTC 2016
LEAPS IN ADOPTION

2X GTC Attendees

2012: 2,350
2016: 5,500

2X Accelerated Systems, 96% of New Systems on NVIDIA

Nov 2013: 20
Nov 2014: 30
Nov 2015: 120

4X CUDA Developers, 10X in Hyperscale + Auto

2012: 30K
2016: 300K
5 THINGS

- A Toolbox
- A Deep Learning Chip
- A Deep Learning Box
- VR
- A Deep Learning Car
NVIDIA SDK
The Essential Resource for GPU Developers

- DEEP LEARNING
  - Deep Learning SDK
    - High-performance tools and libraries for deep learning

- SELF-DRIVING CARS
  - NVIDIA DriveWorks
    - Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

- VIRTUAL REALITY
  - NVIDIA VRWorks™
    - A comprehensive SDK for VR headsets, games and professional applications

Available Now

developer.nvidia.com
NVIDIA GAMEWORKS

Volumetric Lighting
Voxel Accelerated Ambient Occlusion
Hybrid Frustum Traced Shadows
Available Now

And other technologies such as:
Clothing, VXGI, Flex, Destruction
NVIDIA DESIGNWORKS

Adobe support of MDL
Siemens NX adopts Iray

and other technologies such as:
GL Extensions, GRID, GPU Direct for Video, Mosaic, VXGI, Warp and Blend
NVIDIA VRWORKS

Oculus Rift and HTC Vive integration
Epic, Max Play and Unity game engines
Available Now

- Multi-Res Shading
- VR SLI
- Context Priority
- Warp and Blend

And other technologies such as:
Direct Mode, GPUDirect for Video
NVIDIA COMPUTEWORKS

CUDA 8 — Available June

cuDNN 5 — Available April

nvGRAPH — Available June

IndeX plug-in for ParaView — Available May

And other technologies such as:
AMGx, cuSOLVER, cuSPARSE, OpenACC, NSIGHT, THRUST
NVIDIA DRIVeworks

SensorFusion  Detection  Localization  HD Maps

and other technologies such as:

Driving, Planning

JPL — Available Now
EAP — Available Q2’16
General release — Available Q1’17
Deep Learning SDK
DIGITS Workflow
VisionWorks
Jetson Media SDK

and other technologies such as:
Linux4Tegra, NSIGHT EE, OpenCV4Tegra, OpenGL, System Trace, Visual Profiler, Vulkan

NVIDIA JETPACK

GIE - GPU Inference Engine – Available May
Jetson TX1: 24 images/s/W
5 THINGS

NVIDIA SDK

A Deep Learning Chip

A Deep Learning Box

VR

A Deep Learning Car
A START OF A NEW PLATFORM

- Samsung, Oculus, HTC release headsets
- Google announces Jump VR camera platform
- Microsoft demonstrates Holoportation
- New York Times ships Cardboard to subscribers
- VR Startups Raise $1.3B in funding
IRAY VR
BREAKTHROUGH PHOTOREAL VR

Available starting in June
1. Design in 3ds Max

2. Download Iray for 3ds Max Plug-in

3. Download Android Viewer

4. Get VR HMD

Available in June
5 THINGS

NVIDIA SDK

A Deep Learning Chip

A Deep Learning Box

IRAY VR

A Deep Learning Car
AN AMAZING YEAR IN AI
A NEW COMPUTING MODEL

Traditional Computer Vision
Experts + Time

Deep Learning Object Detection
DNN + Data + HPC

Deep Learning Achieves “Superhuman” Results
$500B OPPORTUNITY OVER 10 YRS

Deep Learning Total Revenue by Segment

Deep Learning Software Revenue by Industry

IBM: “Cognitive business represents a $2T opportunity”

Cognitive opens new opportunities on top of traditional IT

Opportunity for decision-making support 2025

Traditional global IT spend 2018

~$2T

~$1.2T

NVIDIA GPU FOR HYPERSCALE

TESLA M40 + TESLA M4

10X Speed up | 20 images/s/W

Cloud Services Powered by AI
Figure 8: A “turn” vector was created from four averaged samples of faces looking left vs looking right. By adding interpolations along this axis to random samples we were able to reliably transform their pose.

“Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”

— Soumith Chintala, Facebook AI Research
Alec Radford & Luke Metz Indico Research
TESLA P100
THE MOST ADVANCED HYPERSONAL DATA CENTER GPU EVER BUILT

150B XTORS | 5.3TF FP64 | 10.6TF FP32 | 21.2TF FP16 | 14MB SM RF | 4MB L2 Cache
“FIVE MIRACLES”

- Pascal Architecture
- 16nm FinFET
- CoWoS with HBM2
- NVLink
- New AI Algorithms
GIANT LEAPS IN EVERYTHING

3x Compute
K-40 M40 P100 (FP32) P100 (FP16)

3x GPU Mem BW
K-40 M40 P100

5x GPU-GPU BW
K-40 M40 P100
“NVIDIA GPU is accelerating progress in AI. As neural nets become larger and larger, we not only need faster GPUs with larger and faster memory, but also much faster GPU-to-GPU communication, as well as hardware that can take advantage of reduced-precision arithmetic. This is precisely what Pascal delivers.”

Yann LeCun, Director of AI Research, Facebook

“This is a new era of computing. New approaches to the underlying technologies will be required for AI and cognitive. The combination of NVIDIA Pascal GPUs and IBM POWER accelerates Watson’s learning of new skills. Together, IBM and NVIDIA will advance the artificial intelligence industry.”

Dr. John Kelly III, SVP, Cognitive Solutions & IBM Research

“AI computers are like space rockets: The bigger the better. Pascal’s throughput and interconnect will make the biggest rocket we’ve seen yet.”

Andrew Ng, Chief Scientist, Baidu

“Microsoft is developing super deep neural networks that are more than 1000 layers. NVIDIA Tesla P100’s impressive horsepower will enable Microsoft’s CNTK to accelerate AI breakthroughs.”

Xuedong Huang, Chief Speech Scientist, Microsoft Research
5 THINGS

NVIDIA SDK

TESLA P100

A Deep Learning Box

IRAY VR

A Deep Learning Car
GPU-ACCELERATED DL FOR EVERY MARKET

Deep Learning in the Cloud

IBM: “Cognitive business represents a $2T opportunity”

Deep Learning for Enterprise

NVIDIA DGX-1
WORLD’S FIRST DEEP LEARNING SUPERCOMPUTER

Engineered for deep learning | 170TF FP16 | 8x Tesla P100 | NVLink hybrid cube mesh | Accelerates major AI frameworks
<table>
<thead>
<tr>
<th></th>
<th>DUAL XEON</th>
<th>DGX-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLOPS (CPU + GPU)</td>
<td>3 TF</td>
<td>170 TF</td>
</tr>
<tr>
<td>AGGREGATE NODE BW</td>
<td>76 GB/s</td>
<td>768 GB/s</td>
</tr>
<tr>
<td>ALEXNET TRAIN TIME</td>
<td>150 HOURS</td>
<td>2 HOURS</td>
</tr>
<tr>
<td>TRAIN IN 2 HOURS</td>
<td>&gt;250 NODES*</td>
<td>1 NODE</td>
</tr>
</tbody>
</table>

* Caffe Training on Multi-node Distributed-memory Systems Based on Intel Xeon Processor 3rd Family (unpublished).

Viewed from [here](https://example.com).
12x Speed-up in One Year

1.33 billion images/day
Persistent RNNs: Peak FLOPs at batch of 8

Add Model Parallelism over NVLINK

Compose with Data Parallelism

Strong scale to 32X more processors
NVIDIA DGX-1
WORLD’S FIRST DEEP LEARNING SUPERCOMPUTER
170TF | “250 servers in-a-box” | nvidia.com/dgx1
$129,000
Frameworks for Multi-GPU Pascal
Large-scale Deep Learning
Reinforcement Learning
Unsupervised and Transfer Learning
Natural Language Understanding
Autonomous Driving
Medical Applications
DEEP LEARNING FOR MEDICINE

NVIDIA Founding Technology Partner of MGH Center of Clinical Data Science

10B Medical images on DGX-1 to advance radiology, pathology, genomics
5 THINGS

NVIDIA SDK

TESLA P100

NVIDIA DGX-1

IRAY VR

A Deep Learning Car
AN AMAZING YEAR FOR SELF-DRIVING CARS

- Uber Enters the Race
- Tesla Model S Auto-pilot
- Toyota Research Institute
- Toyota invests $1B in AI Lab
- Volvo Drive me on Public Roads in 2017
- Audi, BMW, Daimler Buy HERE
- NHTSA: Computer Counts as Driver
- Honda, Nissan, Toyota Team Up
- GM Buys Cruise
- Tesla Model 3: 300K pre-orders
NVIDIA DRIVE PX
AI CAR COMPUTER

World’s first DL-powered car computing platform

One scalable architecture – from DNN training to cluster, infotainment, ADAS, autonomous driving, and mapping

Open platform
NVIDIA DRIVE PX PERCEPTION

Training on DGX-1

Driving with DriveWorks

NVIDIA DGX-1 NVIDIA DRIVE PX

NVIDIA DRIVENET
#1 accuracy score for KITTI car detection

<table>
<thead>
<tr>
<th>Method</th>
<th>Hard</th>
<th>Moderate</th>
<th>Easy</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA DRIVENet-H</td>
<td>83.76%</td>
<td>89.81%</td>
<td>90.92%</td>
<td>GPU @ 2.5 GHz (Python + C++)</td>
</tr>
<tr>
<td>sensekitti</td>
<td>79.99%</td>
<td>89.72%</td>
<td>91.42%</td>
<td>GPU @ 2.5 GHz (Python + C++)</td>
</tr>
<tr>
<td>3DP + RPN</td>
<td>78.38%</td>
<td>88.85%</td>
<td>90.14%</td>
<td>GPU @ 2.5 GHz (Python + C++)</td>
</tr>
<tr>
<td>Mono3D</td>
<td>78.96%</td>
<td>88.66%</td>
<td>92.33%</td>
<td>GPU @ 2.5 GHz (Matlab + C++)</td>
</tr>
<tr>
<td>3DOP</td>
<td>79.10%</td>
<td>88.64%</td>
<td>93.04%</td>
<td>GPU @ 2.5 GHz (Matlab + C++)</td>
</tr>
</tbody>
</table>
NEW END-TO-END HD MAPPING

Training on DGX-1

NVIDIA DGX-1

NVIDIA DRIVE PX

Driving with DriveWorks
PLATFORM FOR MAPPING THE WORLD

here

TOMTOM®

ZENRIN
WORLD’S FIRST AUTONOMOUS RACE CAR

Designed by Daniel Simon
2,200 lbs
Blazing fast
WORLD’S FIRST AUTONOMOUS CAR RACE

10 teams, 20 identical cars
DRIVE PX 2: The “brain” of every car
2016/17 Formula E season

ROBORACE
VR, AI, SELF-DRIVING CARS

NVIDIA SDK

TESLA P100

NVIDIA DGX-1

IRAY VR

HD MAPPING, AI DRIVING