LEAPS IN VISUAL COMPUTING
JEN-HSUN HUANG, CO-FOUNDER & CEO | GTC 2015
FOUR ANNOUNCEMENTS

A New GPU and Deep Learning

A Very Fast Box and Deep Learning

Roadmap Reveal and Deep Learning

Self-Driving Cars and Deep Learning
AMAZING YEAR IN VISUAL COMPUTING
10X GROWTH IN GPU COMPUTING

2008

150,000 CUDA Downloads

27 CUDA Apps

60 Universities Teaching

4,000 Academic Papers

6,000 Tesla GPUs

77 Supercomputing Teraflops
10X GROWTH IN GPU COMPUTING

<table>
<thead>
<tr>
<th>2008</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>150,000 CUDA Downloads</td>
<td>3 Million CUDA Downloads</td>
</tr>
<tr>
<td>27 CUDA Apps</td>
<td>10X GROWTH IN GPU COMPUTING</td>
</tr>
<tr>
<td>60 Universities Teaching</td>
<td></td>
</tr>
<tr>
<td>4,000 Academic Papers</td>
<td></td>
</tr>
<tr>
<td>6,000 Tesla GPUs</td>
<td></td>
</tr>
<tr>
<td>77 Supercomputing Teraflops</td>
<td></td>
</tr>
</tbody>
</table>
10X GROWTH IN GPU COMPUTING

2008
- 150,000 CUDA Downloads
- 27 CUDA Apps
- 60 Universities Teaching
- 4,000 Academic Papers
- 6,000 Tesla GPUs
- 77 Supercomputing Teraflops

2015
- 3 Million CUDA Downloads
- 319 CUDA Apps

10X GROWTH IN GPU COMPUTING
10X GROWTH IN GPU COMPUTING

2008

- 150,000 CUDA Downloads
- 27 CUDA Apps
- 60 Universities Teaching
- 4,000 Academic Papers
- 6,000 Tesla GPUs
- 77 Supercomputing Teraflops

2015

- 3 Million CUDA Downloads
- 319 CUDA Apps
- 800 Universities Teaching

10X GROWTH IN GPU COMPUTING
10X GROWTH IN GPU COMPUTING

2008

- 150,000 CUDA Downloads
- 27 CUDA Apps
- 60 Universities Teaching
- 4,000 Academic Papers
- 6,000 Tesla GPUs
- 77 Supercomputing Teraflops

2015

- 3 Million CUDA Downloads
- 319 CUDA Apps
- 800 Universities Teaching
- 60,000 Academic Papers
10X GROWTH IN GPU COMPUTING

2008

- 150,000 CUDA Downloads
- 27 CUDA Apps
- 60 Universities Teaching
- 4,000 Academic Papers
- 6,000 Tesla GPUs
- 77 Supercomputing Teraflops

2015

- 3 Million CUDA Downloads
- 319 CUDA Apps
- 800 Universities Teaching
- 60,000 Academic Papers
- 6,000 Tesla GPUs
- 54,000 Supercomputing Teraflops
TITAN X
THE WORLD’S FASTEST GPU

8 Billion Transistors
3,072 CUDA Cores
7 TFLOPS SP / 0.2 TFLOPS DP
12GB Memory
TITAN X FOR DEEP LEARNING

Training AlexNet

Days

- 16-core Xeon CPU
- TITAN
- TITAN Black cuDNN
- TITAN X cuDNN

Approximately 43 days for 16-core Xeon CPU, significantly faster times for TITAN and TITAN Black cuDNN.
TITAN X
THE WORLD’S FASTEST GPU

8 Billion Transistors
3,072 CUDA Cores
7 TFLOPS SP / 0.2 TFLOPS DP
12GB Memory

$999
FOUR ANNOUNCEMENTS

- A New GPU and Deep Learning
- A Very Fast Box and Deep Learning
- Roadmap Reveal and Deep Learning
- Self-Driving Cars and Deep Learning
A SHORT HISTORY OF DEEP LEARNING

Convolutional Neural Networks for Handwritten Digital Recognition
LECU, BOTTU, BENGIO, HAFFNER, 1998

ImageNet Classification with NVIDIA GPUs
KRIZHEVSKY, HINTON, ET AL., 2012


Accuracy %

2010 2011 2012 2013 2014

72% 74% 84%

DNN
CV
“Deep Image: Scaling up Image Recognition”

– Baidu: 5.98%, Jan. 13, 2015

“Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”

– Microsoft: 4.94%, Feb. 6, 2015


– Google: 4.82%, Feb. 11, 2015
THE BIG BANG
GPU-ACCELERATED DEEP LEARNING

START-UPS

Adobe · Alibaba.com · Baidu · Capio · clarifai · clarify · Dato · EMOTIENT
facebook · Flickr · Yahoo! · Google · enlitic · ersatz · EyeEm · herta · iFLYTEK
IBM · Microsoft · Nuance · Intelligent Voice · IQIYI · Letv · megvii · MetaMind
Nervana Systems · NVIDIA · Face++ · SenseTime · Sogou · Subaru · Vision\n
twitter · 北京文安 · zebra
DEEP LEARNING REVOLUTIONIZING MEDICAL RESEARCH

Detecting Mitosis in Breast Cancer Cells  
– IDSIA

Predicting the Toxicity of New Drugs  
– Johannes Kepler University

Understanding Gene Mutation to Prevent Disease  
– University of Toronto
Automated Image Captioning with ConvNets and Recurrent Nets

— Andrej Karpathy, Fei-Fei Li

Convolutional Neural Network

Recurrent Neural Network

a bird perched on a branch of a tree
DIGITS
DEEP GPU TRAINING SYSTEM FOR DATA SCIENTISTS

Design DNNs
Visualize activations
Manage multiple trainings
DIGITS

- Process Data
- Configure DNN
- Monitor Progress
- Visualize Layers
DIGITS DEVBOX

World’s fastest GPU
Max GPU out of a plug
Multi-GPU training & inference
DIGITS DEVBOX — EARLY RESULTS

“I’ve never seen AlexNet run this fast...TitanX is a monster, Crazy Fast”

DIGITS makes it way easier to design the best network for the job

— Simon Osindero
A.I. Architect

— Soumith Chintala
Research Engineer

flickr

facebook
DIGITS DEVBOX

Available May 2015
$15,000
FOUR ANNOUNCEMENTS

A New GPU and Deep Learning

A Very Fast Box and Deep Learning

Roadmap Reveal and Deep Learning

Self-Driving Cars and Deep Learning
GPU ROADMAP
Pascal 2.7x Memory Capacity
GPU ROADMAP
Pascal 4x Mixed Precision
GPU ROADMAP

Pascal 3x Bandwidth
PASCAL 10X MAXWELL

**Forward:**
- **Convolution (compute):** 4x (FP16)
- **Fully Connected (bandwidth):** 6x
- Mixed Precision

**Backward:**
- **Fully Connected (bandwidth):** 6x
- **Convolution (compute):** 4x
- Mixed Precision

**Weight Update (interconnect):**
- **Mixed Precision 3D Memory NVLINK**
- 5x
- 10x
- 2x

*Very rough estimates*
FOUR ANNOUNCEMENTS

A New GPU and Deep Learning
A Very Fast Box and Deep Learning
Roadmap Reveal and Deep Learning
Self-Driving Cars and Deep Learning
TODAY’S ADAS

SENSE
FPGA
CV ASIC

PLAN
CPU

ACT
WARN
BRAKE

SPEED LIMIT 35

Images of vehicles and road signs highlighting ADAS functionalities.
NEXT-GENERATION ADAS

- SENSE
  - FPGA
  - CV ASIC

- PLAN
  - CPU

- ACT
  - WARN
  - BRAKE
  - STEER
  - ACCELERATE

Image shows a car following the vehicle in front, with distances marked: 14.3 m and 31.2 m.
NVIDIA DRIVE PX SELF-DRIVING CAR COMPUTER

IMAGENET CHALLENGE

Accuracy %

2010 2011 2012 2013 2014

72% 74% 84% 74%

SENSE

FPGA CV ASIC

PLAN

CPU

ACT

WARN BRAKE STEER ACCELERATE

ACCELERATE
NVIDIA DRIVE PX SELF-DRIVING CAR COMPUTER

IMAGENET CHALLENGE

Accuracy %

- DNN: 74%, 84%
- CV: 72%

SENSE
- FPGA
- CV
- ASIC

PLAN
- CPU

ACT
- WARN
- BRAKE
- STEER
- ACCELERATE

DNN
NVIDIA DRIVE PX SELF-DRIVING CAR COMPUTER

IMAGENET CHALLENGE

Accuracy %

<table>
<thead>
<tr>
<th>Year</th>
<th>CV</th>
<th>DNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>72%</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>74%</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SENSE
- FPGA
- CV
- ASIC

PLAN
- DNN

ACT
- WARN
- BRAKE
- STEER
- ACCELERATE

ACCELERATE
PROJECT DAVE — DARPA AUTONOMOUS VEHICLE

DNN-based self-driving robot
Training data by human driver
No hand-coded CV algorithms

PROJECT LEADS
Urs Muller: Chief Architect, Autonomous Driving, NVIDIA
Yann LeCun: Director, AI Research, Facebook

IMAGENET CHALLENGE

<table>
<thead>
<tr>
<th>Year</th>
<th>CV</th>
<th>DNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>72%</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>74%</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DAVE IN ACTION
TRAINING DATA
225K Images
TEST DRIVE
No Training
TEST DRIVE
Partially Trained (52K images)
TEST DRIVE
Fully Trained (225K images)
<table>
<thead>
<tr>
<th></th>
<th>DAVE</th>
<th>AlexNet on DRIVE PX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Connections</td>
<td>3.1 Million</td>
<td>630 Million</td>
</tr>
<tr>
<td>Frames / Second</td>
<td>12</td>
<td>184</td>
</tr>
<tr>
<td>Connections / Second</td>
<td>38 Million</td>
<td>116 Billion</td>
</tr>
</tbody>
</table>

3,000x Faster
NVIDIA DRIVE PX SELF-DRIVING CAR COMPUTER

IMAGENET CHALLENGE

Accuracy %

2010 2011 2012 2013 2014

72% 74% 84%

SENSE
- FPGA
- CV
- ASIC

PLAN
- CPU

ACT
- WARN
- BRAKE
- STEER
- ACCELERATE

ACCELERATE
WELCOME TO THE FUTURE OF AUTOMOTIVE INNOVATION

SELF-DRIVING COMPUTING DEVELOPMENT PLATFORM

NVIDIA DRIVE™ PX is a powerful computer-on-a-chip designed to run the deep neural networks that will enable a car to see, think, and learn.

DRIVE PX features dual NVIDIA Tegra™ X1 processors and delivers 2.3 teraflops of performance. Twelve camera inputs enable a wide range of ADAS features to run simultaneously, including surround view, collision avoidance, pedestrian detection, lane-keep, lane departure, cross-traffic monitoring, and driver state monitoring.

DRIVE PX is available to automakers, Tier 1 suppliers, and research institutions working on developing systems that enable cars to drive themselves.

ORDER NOW

Available May 2015
$10,000
LEAPS IN VISUAL COMPUTING

TITAN X
The World’s Fastest GPU

DIGITS DevBox
GPU Deep Learning Platform

Pascal — 10x Maxwell
For Deep Learning

NVIDIA DRIVE PX
Deep Learning Platform for Self-Driving Cars