1. Three Requirements for Autonomous Driving: AI, Massive Parallel Processing and Powerful CPU Cores
2. GPU Computing (AI is also based on this technology)
3. Artificial Intelligence
   • Deep Learning Training and Inference
   • GPU widely used for Training
   • GPU with TensorRT Optimizer in lead for Inference
4. NVIDIA AI Autonomous Driving Platform
5. GTC2017 EU Announcement: New Drive PX for Level 5
6. (NVIDIA Positioning in the Autonomous Driving Industry)
7. Summary
1. Three Requirements for Autonomous Driving: AI, Massive Parallel Processing and Powerful CPU Cores
2. GPU Computing (AI is also based on this technology)
3. Artificial Intelligence
   • Deep Learning Training and Inference
   • GPU widely used for Training
   • GPU with TensorRT Optimizer in lead for Inference
4. NVIDIA AI Autonomous Driving Platform
5. GTC2017 EU Announcement: New Drive PX for Level 5
6. (NVIDIA Positioning in the Autonomous Driving Industry)
7. Summary
THREE REQUIREMENTS FOR AUTONOMOUS DRIVING: AI, MASSIVE PARALLEL PROCESSING AND POWERFUL CPU CORES

- **AI**
  - Super-Human Perception Capability by **Deep Learning**

- **CPU**
  - Massive Parallel Processing

- **GPU Computing** implementing various key Algorithms of OEMs, Tier1s and NVIDIA in Realtime

- **OpenRoadNet** (Segmentation for Free-Space)

- **HD Map**

- **DriveNet** (Detect Cars, Pedestrians, Traffic Signs etc)

- **LaneNet** (Detect Lanes)

- **Occupancy Grid(t)** (Map Objects around Self-Car)

  - Projected to several time units ahead

  - **Prediction**

- **PilotNet** (AI Path Planning)

- **Human Driving**

  - Based on Human Algorithm

  - Path Planning

  - Safety Cost Analysis

  - Distance to Obstacles/Lanes, Traffic Rules etc

- **Human Actuation Control**

- **Path Plan Decision**

To be Updated. This is from GTC2016 EU
1. Three Requirements for Autonomous Driving: AI, Massive Parallel Processing and Powerful CPU Cores

2. GPU Computing (AI is also based on this technology)

3. Artificial Intelligence
   • Deep Learning Training and Inference
   • GPU widely used for Training
   • GPU with TensorRT Optimizer in lead for Inference

4. NVIDIA AI Autonomous Driving Platform

5. GTC2017 EU Announcement: New Drive PX for Level 5

6. (NVIDIA Positioning in the Autonomous Driving Industry)

7. Summary
CPU PERFORMANCE GROWTH SATURATED
CANNOT BENEFIT FROM MOORE’S LAW ANYMORE
DUE TO LEAK CURRENT AND COMPLEX ARCHITECTURE

40 Years of CPU Performance Trend Data

One solution to continuously grow CPU performance was Multi CPU Core.

But Amdahl’s Law limits the efficiency of Multi CPU Core Processing.

\[ T_b = T_a \times \left( R + \frac{1 - R}{N} \right) \]

\[ \alpha = \frac{T_a}{T_b} = \frac{1}{R + \frac{1 - R}{N}} \]

N: Number of Processing Cores  
R: Ratio of Sequential Processing  
1 - R: Ratio of Parallel Processing  
\( \alpha \): Multi Core Efficiency  
(a =1 is equivalent to a single Processor)  
Ta: Single Processor Execution Time  
Tb: N Core Execution Time  

GPU can enjoy this linear Performance growth.
NVIDIA GPU PERFORMANCE CONTINUES TO GROW CAN STILL BENEFIT FROM MOORE’S LAW

NVIDIA Volta V100 GPU

FP32 Core: 5,120
FP64 Core: 2,560
21 Billion Transistors
120 TFLOPS

40 Years of CPU Performance Trend Data

NVIDIA GPU-ACCELERATED APPLICATIONS

Computer Graphics

Scientific Computing

Deep Learning

Data Science

NVIDIA CUDA
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
1. Three Requirements for Autonomous Driving: AI, Massive Parallel Processing and Powerful CPU Cores
2. GPU Computing (AI is also based on this technology)
3. Artificial Intelligence
   - Deep Learning Training and Inference
   - GPU widely used for Training
   - GPU with TensorRT Optimizer in lead for Inference
4. NVIDIA AI Autonomous Driving Platform
5. GTC2017 EU Announcement: New Drive PX for Level 5
6. (NVIDIA Positioning in the Autonomous Driving Industry)
7. Summary
DL TWO PHASES: TRAINING AND INFERENCE

TRAINING
Learning a new capability from existing data

INFERENCEn
Applying this capability to new data

Untrained Neural Network Model

Deep Learning Framework

TRAINING DATASET

Trained Model New Capability

NEW DATA

App or Service Featuring Capability

Trained Model Optimized for Performance
## 2013 TRAINING SPEED: GPU VS CPU

### Training Time for CNN

<table>
<thead>
<tr>
<th>Batch Size</th>
<th>Training Time (CPU)</th>
<th>Training Time (GPU)</th>
<th>GPU/CPU Acceleration</th>
</tr>
</thead>
<tbody>
<tr>
<td>64 Images</td>
<td>64s</td>
<td>7.5s</td>
<td>8.5X</td>
</tr>
<tr>
<td>128 Images</td>
<td>124s</td>
<td>14.5s</td>
<td>8.5X</td>
</tr>
<tr>
<td>256 Images</td>
<td>257s</td>
<td>28.5s</td>
<td>9.0X</td>
</tr>
</tbody>
</table>

- ILSVRC12 Supervision DNN
- 7-Layers (5-CNN, 2-FCN)
- Caffe Framework
- Training Time for 20-Iteration

CPU: Dual 10-Core Ivy Bridge
GPU: 1 Tesla K40
CPU Library: Intel MKL BLAS
GPU Library: cuBLAS

### Extrapolated to 1M-Images:
- CPU: 11.6 Days, GPU: 1.3 Days
2016: GPU TRAINING X60 IN 3 YEARS
ACHIEVED BY CUDNN LIBRARY, MULTI-GPU AND PASCAL ARCHITECTURE

Alexnet training throughput on:

CPU: 1xE5-2680v3 12 Co 2.5GHz 128GB System Memory, Ubuntu 14.04
M40 bar: 8xM40 GPUs in a node. P100: 8xP100 NVLink-enabled
2017: X12 ACCELERATION BY VOLTA GPU

- **CNN Training** (ResNet-50)
  - 8x K80
  - 8x P100
  - 8x V100

- **Multi-Node Training with NCCL 2.0** (ResNet-50)
  - 8x P100
  - 8x V100
  - 64x V100

- **LSTM Training** (Neural Machine Translation)
  - K80
  - P100
  - V100
AI INFERENCE IS THE NEXT GREAT CHALLENGE

Training

DNN Model

Inferencing

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

- Smartphone
- Clock
- Drone
- Security Camera
- Robot
- Smart Home Device
- Autonomous Vehicle
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

Weight & Activation Precision Calibration  |  Layer & Tensor Fusion
Kernel Auto-Tuning  |  Multi-Stream Execution
WHY WE SHOULD REDUCE MULTIPLICATION AND REDUCE CALCULATION PRECISION?

For 16b Int, MULT consumes x13 of ADD
32b FP ADD consumes x17 of 16b Int ADD
32b FP MULT consumes x80 of 16b Int ADD

<table>
<thead>
<tr>
<th>Operation</th>
<th>Energy, pJ</th>
<th>Relative cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>16b Int ADD</td>
<td>0.06</td>
<td>1</td>
</tr>
<tr>
<td>16b Int MULT</td>
<td>0.8</td>
<td>13</td>
</tr>
<tr>
<td>16b FP ADD</td>
<td>0.45</td>
<td>8</td>
</tr>
<tr>
<td>16b FP MULT</td>
<td>1.1</td>
<td>18</td>
</tr>
<tr>
<td>32b FP ADD</td>
<td>1.0</td>
<td>17</td>
</tr>
<tr>
<td>32b FP MULT</td>
<td>4.5</td>
<td>80</td>
</tr>
<tr>
<td>Register File, 1kB</td>
<td>0.6</td>
<td>10</td>
</tr>
<tr>
<td>L1 Cache, 32kB</td>
<td>3.5</td>
<td>58</td>
</tr>
<tr>
<td>L2 Cache, 256kB</td>
<td>30.2</td>
<td>500</td>
</tr>
<tr>
<td>on-chip DRAM</td>
<td>160</td>
<td>2667</td>
</tr>
<tr>
<td>DRAM</td>
<td>640</td>
<td>10667</td>
</tr>
<tr>
<td>Wireless transfer</td>
<td>60000</td>
<td>10000000</td>
</tr>
</tbody>
</table>

Table 1: Energy cost of common operations.

Artem Vasilyev, “CNN Optimization for Embedded Systems and FFT”
CS231n: CNN for Visual Recognition Course, Stanford, 2017
cuDNN Library reducing the number of MUL

- cuDNN: NVIDIA CUDA Library for CNN and other DNN features
  (cuDNNが自動で最適なアルゴリズムを決定)
  - CUDNN_CONVOLUTION_FWD_ALGO_IMPLICIT_GEMM
  - CUDNN_CONVOLUTION_FWD_ALGO_IMPLICIT_PRECOMP_GEMM
  - CUDNN_CONVOLUTION_FWD_ALGO_GEMM
  - CUDNN_CONVOLUTION_FWD_ALGO_DIRECT
  - CUDNN_CONVOLUTION_FWD_ALGO_FFT
  - CUDNN_CONVOLUTION_FWD_ALGO_FFT_TILING
  - CUDNN_CONVOLUTION_FWD_ALGO_WINOGRAD

Number of MUL Reduction for energy efficiency

- FFT (Fast Fourier Transform)
- WINOGRAD for small order CNNs
TensorRT automatically calibrate for FP32 to INT8 weight/activation conversion (1)

### High-throughput INT8 math

**DP4A - INT8 dot product**

- Requires `sm_61+` (Pascal TitanX, GTX 1080, Tesla P4, P40 and others).
- Four-way byte dot product accumulated in 32-bit result.

```
```
TensorRT automatically calibrate for FP32 to INT8 weight/activation conversion (2)

- NN Output signal distribution differs for DNN algorithms and layers
- Automatically minimize the difference between FP32, INT8 outputs
- Kullback-Leibler divergence (KL_divergence) is used as the metric

Q: How to optimize threshold selection?
- It’s always a tradeoff between range and precision of the INT8 representation.

Solution: Calibration
- Run FP32 inference on Calibration Dataset.
- For each Layer:
  - collect histograms of activations.
  - generate many quantized distributions with different saturation thresholds.
  - pick threshold which minimizes KL_divergence(ref_distr, quant_distr).
- Entire process takes a few minutes on a typical desktop workstation.
TensorRT automatically calibrate for FP32 to INT8 weight/activation conversion (3)

Calibration conducted to Six popular DNNs (e.g. ResNet, VGG, AlexNet)

- Degradation of DNN recognition precision: 0.01% - 0.45%
- Efficiency Improvement: x3.3 – x3.7 (Drive PX2 dGPU)

### Results - Accuracy

<table>
<thead>
<tr>
<th>NETWORK</th>
<th>FP32 Top1</th>
<th>FP32 Top5</th>
<th>INT8 Top1</th>
<th>INT8 Top5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>73.23%</td>
<td>91.18%</td>
<td>73.03%</td>
<td>91.15%</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>74.39%</td>
<td>91.79%</td>
<td>74.52%</td>
<td>91.64%</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>74.78%</td>
<td>91.82%</td>
<td>74.62%</td>
<td>91.82%</td>
</tr>
<tr>
<td>VGG-19</td>
<td>69.41%</td>
<td>88.76%</td>
<td>69.42%</td>
<td>88.96%</td>
</tr>
<tr>
<td>Googlenet</td>
<td>68.57%</td>
<td>88.83%</td>
<td>68.21%</td>
<td>88.67%</td>
</tr>
<tr>
<td>Alexnet</td>
<td>57.08%</td>
<td>80.06%</td>
<td>57.00%</td>
<td>79.96%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NETWORK</th>
<th>INT8 Top1</th>
<th>INT8 Top5</th>
<th>Calib. 5 batches</th>
<th>Calib. 10 batches</th>
<th>Calib. 50 batches</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td></td>
<td></td>
<td>73.03%</td>
<td>91.15%</td>
<td>73.02%</td>
</tr>
<tr>
<td>ResNet-101</td>
<td></td>
<td></td>
<td>74.52%</td>
<td>91.64%</td>
<td>73.10%</td>
</tr>
<tr>
<td>ResNet-152</td>
<td></td>
<td></td>
<td>74.62%</td>
<td>91.82%</td>
<td>74.38%</td>
</tr>
<tr>
<td>VGG-19</td>
<td></td>
<td></td>
<td>69.42%</td>
<td>88.96%</td>
<td>69.62%</td>
</tr>
<tr>
<td>Googlenet</td>
<td></td>
<td></td>
<td>68.21%</td>
<td>88.67%</td>
<td>68.12%</td>
</tr>
<tr>
<td>Alexnet</td>
<td></td>
<td></td>
<td>57.00%</td>
<td>79.96%</td>
<td>57.05%</td>
</tr>
</tbody>
</table>

TensorRT 2.1, all optimizations enabled. ILSVRC2012 validation dataset, batch = 25 images. Accuracy was measured on 500 batches which were not used for the calibration.

### Results - Performance

Performance of INT8 vs FP32, Titan X (Pascal)

Performance of INT8 vs FP32, DRIVE PX 2 (dGPU)
NEW NVIDIA TENSORRT 3
Programmable Inference Accelerator

Images/Sec (ResNet-50)

<table>
<thead>
<tr>
<th></th>
<th>CPU + TensorFlow</th>
<th>V100 + TensorFlow</th>
<th>V100 + TensorRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed-up</td>
<td>140</td>
<td>300</td>
<td>5,700</td>
</tr>
</tbody>
</table>

Sentences/Sec (OpenNMT)

<table>
<thead>
<tr>
<th></th>
<th>CPU + Torch</th>
<th>V100 + Torch</th>
<th>V100 + TensorRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed-up</td>
<td>4</td>
<td>25</td>
<td>550</td>
</tr>
</tbody>
</table>

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

TCO: Total Cost of Ownership
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

SAME THING APPLIES TO AUTO
1. Three Requirements for Autonomous Driving: AI, Massive Parallel Processing and Powerful CPU Cores
2. GPU Computing (AI is also based on this technology)
3. Artificial Intelligence
   • Deep Learning Training and Inference
   • GPU widely used for Training
   • GPU with TensorRT Optimizer in lead for Inference
4. NVIDIA AI Autonomous Driving Platform
5. GTC2017 EU Announcement: New Drive PX for Level 5
6. (NVIDIA Positioning in the Autonomous Driving Industry)
7. Summary
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
NVIDIA COMPETENCE: ONE-ARCHITECTURE
FROM SUPER COMPUTER TO AUTOMOTIVE SOC

Autonomous AI Processor Tegra
XAVIER
WORLD’S FIRST AUTONOMOUS MACHINE PROCESSOR

Deep Learning, CV, Parallel Computing
Rich High-Speed Sensor I/Os
Extreme Energy Efficiency
30 TOPS at 30W
XAVIER
AI SUPERCOMPUTER SOC

DRIVE PX 2
2 PARKER SoC + 2 PASCAL GPU
20 TOPS DL | 120 SPECINT | 80W

XAVIER
20 TOPS DL | 160 SPECINT | 20W
1. Three Requirements for Autonomous Driving: AI, Massive Parallel Processing and Powerful CPU Cores
2. GPU Computing (AI is also based on this technology)
3. Artificial Intelligence
   • Deep Learning Training and Inference
   • GPU widely used for Training
   • GPU with TensorRT Optimizer in lead for Inference
4. NVIDIA AI Autonomous Driving Platform
5. GTC2017 EU Announcement: New Drive PX for Level 5
6. (NVIDIA Positioning in the Autonomous Driving Industry)
7. Summary
“RIDE-HAILING INDUSTRY EXPECTED TO GROW EIGHTFOLD TO $285B BY 2030”

– Goldman Sachs
ROBOTAXI DEMANDS EXTREME COMPUTING

+ 10X camera resolution
+ Surround LIDAR point-cloud processing
+ Camera & LIDAR localization to HD map
+ Tracking all surrounding objects
+ New map generation
+ Sophisticated path planning & control
+ Algorithm diversity
+ Sensor & computing fail-operate
+ Excess computing capacity
STATE-OF-THE-ART DRIVERLESS VEHICLES
ANNOUNCING “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
ANNOUNCING
“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
1. Three Requirements for Autonomous Driving: AI, Massive Parallel Processing and Powerful CPU Cores
2. GPU Computing (AI is also based on this technology)
3. Artificial Intelligence
   • Deep Learning Training and Inference
   • GPU widely used for Training
   • GPU with TensorRT Optimizer in lead for Inference
4. NVIDIA AI Autonomous Driving Platform
5. GTC2017 EU Announcement: New Drive PX for Level 5
6. (NVIDIA Positioning in the Autonomous Driving Industry)
7. Summary
THE BRAIN OF AI CARS

NVIDIA DRIVE™ PX — scalable AI car supercomputer

Toyota, Tesla Motors, Audi, Daimler building AI cars with NVIDIA

Bosch and ZF auto suppliers have adopted DRIVE PX

HD mapping companies — Baidu, HERE, TomTom, ZENRIN — working with NVIDIA
HW2 based on NVIDIA Processors already in Mass-Production

*Functions activated after extensive validation and regulatory approval*

1) **Enhanced Autopilot**  
   automatic lane-change, transition/exit freeway

2) **Full Self-Driving Capability**  
   urban streets, dense freeways, park seek  
   *(requires Enhanced Autopilot)*
145 AUTONOMOUS VEHICLE STARTUPS ON NVIDIA DRIVE
SUMMARY

1. Three Requirements for Autonomous Driving: AI, Massive Parallel Computation and Powerful Multi-Core CPU

2. All these features integrated in Automotive Tegra SoCs, including Xavier

3. While CPU performance growth are saturating, GPU performance still benefits from Moore’s Law

4. AI Training: GPU Training performance is far exceeding that of other devices

5. AI Inferencing: TensorRT programmable inference accelerator and the progress in GPU architecture is making a big advantage

6. NVIDIA offering a scalable and open autonomous driving platform Drive PX

7. New Autonomous driving platform “Pegasus” for fully-autonomous driverless vehicle has been announced with 320 TOPS Performance