NEW NVIDIA PLATFORM FOR AI

Pedro Mario Cruz e Silva (pcruzesilva@nvidia.com) LinkedIn
Solution Architect Manager
Enterprise Latin America
Global Oil & Gas Team

NVIDIA
"GTC 2017: 'I AM AI' OPENING IN KEYNOTE"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=SUNPRR4O5ZA
LIFE AFTER MOORE’S LAW

40 Years of Microprocessor Trend Data

200B CORE HOURS OF LOST SCIENCE
Data Center Throughput is the Most Important Thing for HPC

National Science Foundation (NSF XSEDE) Supercomputing Resources

Source: NSF XSEDE Data: https://portal.xsede.org/#/gallery
NU = Normalized Computing Units are used to compare compute resources across supercomputers and are based on the result of the High Performance LINPACK benchmark run on each system.
THE ADVANTAGES OF GPU-ACCELERATED DATA CENTER
RISE OF GPU COMPUTING

APPLICATIONS
ALGORITHMS
SYSTEMS
CUDA
ARCHITECTURE

DEEP LEARNING
LEARNING FROM DATA
AND SOME BUZZ WORDS

ARTIFICIAL INTELLIGENCE
- Knowledge & Reason
- Learning
- Planning
- Communicating
- Perceiving

MACHINE LEARNING
- Learning from data
- Expert systems
- Handcrafted features

DEEP LEARNING
- Learning from data
- Neural networks
- Computer learned features
A NEW COMPUTING MODEL

TRAINING

Training Data

Input

“Label” Output

Trained Neural Network

INFERENCE

Input

Trained Neural Network

“Label” Output
A NEW COMPUTING MODEL
Outperform experts, facts, rules with software that writes software

Traditional Computer Vision
Experts + Time

Deep Learning Object Detection
DNN + Data + GPU

Deep Learning Achieves
“Superhuman” Results
“ACCELERATING EULERIAN FLUID SIMULATION WITH CONVOLUTIONAL NETWORKS”


Fig. 1: Smoke simulation using our system - our method is capable of fast and accurate simulation of the Euler Equations for incompressible fluid flow at interactive frame-rates. Videos can be found at: [http://cims.nyu.edu/~schlacht/CNNFluids.htm](http://cims.nyu.edu/~schlacht/CNNFluids.htm)
“ACCELERATING EULERIAN FLUID SIMULATION WITH CONVOLUTIONAL NETWORKS”
HTTPS://WWW.YOUTUBE.COM/WATCH?V=W71ZXKNIJFO
DEEP LEARNING SOFTWARE
POWERING THE DEEP LEARNING ECOSYSTEM

NVIDIA SDK accelerates every major framework

<table>
<thead>
<tr>
<th>COMPUTER VISION</th>
<th>SPEECH &amp; AUDIO</th>
<th>NATURAL LANGUAGE PROCESSING</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJECT DETECTION</td>
<td>VOICE RECOGNITION</td>
<td>RECOMMENDATION ENGINES</td>
</tr>
<tr>
<td>IMAGE CLASSIFICATION</td>
<td>LANGUAGE TRANSLATION</td>
<td>SENTIMENT ANALYSIS</td>
</tr>
</tbody>
</table>

![Images and icons for different frameworks and AI tasks](developer.nvidia.com/deep-learning-software)
DEEP LEARNING WORKFLOWS

**IMAGE CLASSIFICATION**
- 98% Dog
- 2% Cat
- Classify images into classes or categories
- Object of interest could be anywhere in the image

**OBJECT DETECTION**
- Find instances of objects in an image
- Objects are identified with bounding boxes

**IMAGE SEGMENTATION**
- Partition image into multiple regions
- Regions are classified at the pixel level

New in DIGITS 5
WHAT’S NEW IN DIGITS?

TENSORFLOW SUPPORT

Train TensorFlow Models Interactively with DIGITS

NEW PRE-TRAINED MODELS

Image Classification: VGG-16, ResNet50
Object Detection: DetectNet
NVIDIA TENSORRT PROGRAMMABLE INFERENCE ACCELERATOR
DEEP LEARNING APPLICATIONS
KLM's 235 social media service agents engage in 15K conversations a week, 24/7. To contend with the overwhelming volume of messages, KLM uses GPU-accelerated deep learning to predict the best response to an incoming message and shows it to a contact center agent for approval or personalization before sending it to the customer. The resulting time savings for KLM service agents means they can focus on customers with more pressing needs and handle a greater volume of questions while still maintaining a high degree of customer satisfaction.
10,000s of features make up today's fraudulent behavior. AI can detect patterns faster and more accurate than humans.

-Hui Wang, Senior Director of Global Risk Sciences, PayPal
SAP AI FOR THE ENTERPRISE

First commercial AI offerings from SAP

Brand Impact, Service Ticketing, Invoice-to-Record applications

Powered by NVIDIA GPUs on DGX-1 and AWS
MICROSOFT - "BUILD 2017: WORKPLACE SAFETY DEMONSTRATION"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=PL-C00M2CNI
"FUTURE OF AI CITIES ON DISPLAY AT ISC WEST 2017"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=ZHBV34KOWHM
"LIP READING SENTENCES IN THE WILD"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=5AOGZAUPILE
TRAINING SET

Features (Seismic)

Labels
AUTONOMOUS CARS
AN AMAZING YEAR FOR SELF-DRIVING CARS

- Uber Enters the Race
- Toyota Invests $1B in AI Lab
- Volvo Drive Me on Public Roads in 2017
- NHTSA: Computer Counts as Driver
- Tesla Model 3: 300K pre-orders
- Audi, BMW, Daimler Buy HERE
- Tesla Model S Auto-pilot
- Baidu Enters the Race
- Honda, Nissan, Toyota Team Up
- GM Buys Cruise
NEW AI DRIVING
"NVIDIA DRIVE AUTONOMOUS VEHICLE PLATFORM"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=0RC4RQYLTEU
"NVIDIA SELF-DRIVING CAR DEMO AT CES 2017"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=FMVWLR0X1SK
TESLA PLATFORM
TESLA V100
The Fastest and Most Productive GPU for AI and HPC

Volta Architecture

Tensor Core
125 Programmable TFLOPS Deep Learning

Improved NVLink & HBM2
Efficient Bandwidth

Volta MPS
Inference Utilization

Improved SIMT Model
New Algorithms
"DGX-1: WORLD’S FIRST DEEP LEARNING SUPERCOMPUTER IN A BOX"

HTTPS://WWW.YOUTUBE.COM/WATCH?V=FAZS4V2AOLI&T
NVIDIA® DGX-1™
Instant productivity — plug-and-play, supports every AI framework and accelerated analytics software applications

Performance optimized across the entire stack

Always up-to-date via the cloud

Mixed framework environments — baremetal and containerized

Direct access to NVIDIA experts
ANNOUNCING TESLA V100
GIANT LEAP FOR AI & HPC
VOLTA WITH NEW TENSOR CORE

21B xtors | TSMC 12nm FFN | 815mm²
5,120 CUDA cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS
NEW 120 Tensor TFLOPS
20MB SM RF | 16MB Cache
16GB HBM2 @ 900 GB/s
300 GB/s NVLink
New CUDA TensorOp instructions & data formats

4x4 matrix processing array


Optimized for deep learning
TENSOR CORE
4x4x4 matrix multiply and accumulate
# Tesla P100 vs Tesla V100

<table>
<thead>
<tr>
<th></th>
<th>Tesla P100 (Pascal)</th>
<th>Tesla V100 (Volta)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Memory</strong></td>
<td>16 GB (HBM2)</td>
<td>16 GB (HMB2)</td>
</tr>
<tr>
<td><strong>Memory Bandwidth</strong></td>
<td>720 GB/s</td>
<td>900 GB/s</td>
</tr>
<tr>
<td><strong>NVLINK</strong></td>
<td>160 GB/s</td>
<td>300 GB/s</td>
</tr>
<tr>
<td><strong>CUDA Cores (FP32)</strong></td>
<td>3584</td>
<td>5120</td>
</tr>
<tr>
<td><strong>CUDA Cores (FP64)</strong></td>
<td>1792</td>
<td>2560</td>
</tr>
<tr>
<td><strong>Tensor Cores (TC)</strong></td>
<td>NA</td>
<td>640</td>
</tr>
<tr>
<td><strong>Peak TFLOPS/s (FP32)</strong></td>
<td>10.6</td>
<td>15</td>
</tr>
<tr>
<td><strong>Peak TFLOPS/s (FP64)</strong></td>
<td>5.3</td>
<td>7.5</td>
</tr>
<tr>
<td><strong>Peak TFLOPS/s (TC)</strong></td>
<td>NA</td>
<td>120</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>300 W</td>
<td>300 W</td>
</tr>
</tbody>
</table>
ANNOUNCING NVIDIA SATURNV WITH VOLTA

40 PetaFLOPS Peak FP64 Performance | 660 PetaFLOPS DL FP16 Performance | 660 NVIDIA DGX-1 Server Nodes
THE VALUE OF GPU COMPUTING PLATFORM FOR AI
AI TO TRANSFORM EVERY INDUSTRY

HEALTHCARE

>80% Accuracy & Immediate Alert to Radiologists

INFRASTRUCTURE

50% Reduction in Emergency Road Repair Costs

IOT

>$6M / Year Savings and Reduced Risk of Outage
NEURAL NETWORK COMPLEXITY IS EXPLODING
Bigger and More Compute Intensive

Image (GOP * Bandwidth)
- AlexNet
- GoogleNet
- ResNet-50
- Inception-v2
- Inception-v4

Speech (GOP * Bandwidth)
- DeepSpeech
- DeepSpeech 2
- DeepSpeech 3

Translation (GOP * Bandwidth)
- OpenNMT
- GNMT
- MoE
WORLD’S MOST ADVANCED DATA CENTER GPU
NOW WITH 2X THE MEMORY

5,120 CUDA cores
640 NEW Tensor cores
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS | 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
32GB HBM2 @ 900GB/s | 300GB/s NVLink
FASTER RESULTS ON COMPLEX DL AND HPC

Up to 50% Faster Results With 2x The Memory

**FASTER RESULTS**

- **1.5X Faster Language Translation**
  - Neural Machine Translation (NMT)
  - 0.8 step/sec

- **1.5X Faster Calculations**
  - 3D FFT 1k x 1k x 1k
  - 1.2 step/sec
  - 2.5TF
  - 3.8TF

**HIGHER ACCURACY**

- **40% Lower Error Rate**
  - VGG-16
  - Accuracy (16 layers)
  - V100 16GB
  - V100 32GB

- **Accuracy (152 layers)**
  - RN-152
  - 512x512 res images

**HIGHER RESOLUTION**

- **4X Higher resolution**
  - V100 16GB
  - V100 32GB

- **GAN Image to ImageGen**
  - 1024x1024 res images

Unsupervised Image Translation

Input winter photo

AI converts it to summer

---

Dual E5-2698v4 server, 512GB DDR4, Ubuntu 16.04, CUDA9, cuDNN7 | NMT is GNMT-like and run with TensorFlow NGC Container 18.01 (Batch Size= 128 (for 16GB) and 256 (for 32GB) | FFT is with cufftbench 1k x 1k x 1k and comparing 2 V100 16GB (DGX1V) vs. 2 V100 32GB (DGX1V)

Neural Machine Translation (NMT)

- **1.5X Faster Language Translation**
  - Neural Machine Translation (NMT)
  - 0.8 step/sec

- **1.5X Faster Calculations**
  - 3D FFT 1k x 1k x 1k
  - 1.2 step/sec
  - 2.5TF
  - 3.8TF

**HIGHER ACCURACY**

- **40% Lower Error Rate**
  - VGG-16
  - Accuracy (16 layers)
  - V100 16GB
  - V100 32GB

- **Accuracy (152 layers)**
  - RN-152
  - 512x512 res images

**HIGHER RESOLUTION**

- **4X Higher resolution**
  - V100 16GB
  - V100 32GB

- **GAN Image to ImageGen**
  - 1024x1024 res images

Unsupervised Image Translation

Input winter photo

AI converts it to summer

---

Dual E5-2698v4 server, 512GB DDR4, Ubuntu 16.04, CUDA9, cuDNN7 | NMT is GNMT-like and run with TensorFlow NGC Container 18.01 (Batch Size= 128 (for 16GB) and 256 (for 32GB) | FFT is with cufftbench 1k x 1k x 1k and comparing 2 V100 16GB (DGX1V) vs. 2 V100 32GB (DGX1V)

Neural Machine Translation (NMT)

- **1.5X Faster Language Translation**
  - Neural Machine Translation (NMT)
  - 0.8 step/sec

- **1.5X Faster Calculations**
  - 3D FFT 1k x 1k x 1k
  - 1.2 step/sec
  - 2.5TF
  - 3.8TF

**HIGHER ACCURACY**

- **40% Lower Error Rate**
  - VGG-16
  - Accuracy (16 layers)
  - V100 16GB
  - V100 32GB

- **Accuracy (152 layers)**
  - RN-152
  - 512x512 res images

**HIGHER RESOLUTION**

- **4X Higher resolution**
  - V100 16GB
  - V100 32GB

- **GAN Image to ImageGen**
  - 1024x1024 res images

Unsupervised Image Translation

Input winter photo

AI converts it to summer

---

Dual E5-2698v4 server, 512GB DDR4, Ubuntu 16.04, CUDA9, cuDNN7 | NMT is GNMT-like and run with TensorFlow NGC Container 18.01 (Batch Size= 128 (for 16GB) and 256 (for 32GB) | FFT is with cufftbench 1k x 1k x 1k and comparing 2 V100 16GB (DGX1V) vs. 2 V100 32GB (DGX1V)

Neural Machine Translation (NMT)

- **1.5X Faster Language Translation**
  - Neural Machine Translation (NMT)
  - 0.8 step/sec

- **1.5X Faster Calculations**
  - 3D FFT 1k x 1k x 1k
  - 1.2 step/sec
  - 2.5TF
  - 3.8TF

**HIGHER ACCURACY**

- **40% Lower Error Rate**
  - VGG-16
  - Accuracy (16 layers)
  - V100 16GB
  - V100 32GB

- **Accuracy (152 layers)**
  - RN-152
  - 512x512 res images

**HIGHER RESOLUTION**

- **4X Higher resolution**
  - V100 16GB
  - V100 32GB

- **GAN Image to ImageGen**
  - 1024x1024 res images

Unsupervised Image Translation

Input winter photo

AI converts it to summer

---

Dual E5-2698v4 server, 512GB DDR4, Ubuntu 16.04, CUDA9, cuDNN7 | NMT is GNMT-like and run with TensorFlow NGC Container 18.01 (Batch Size= 128 (for 16GB) and 256 (for 32GB) | FFT is with cufftbench 1k x 1k x 1k and comparing 2 V100 16GB (DGX1V) vs. 2 V100 32GB (DGX1V)
## TESLA V100: CHOOSING BETWEEN 32GB & 16GB

<table>
<thead>
<tr>
<th>Application</th>
<th>Today’s Products</th>
<th>New Deployments</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL Training</td>
<td>P100, P40, V100 16GB</td>
<td>V100 32GB</td>
<td>Faster Result with Larger More Complex Models</td>
</tr>
<tr>
<td>Memory-Constrained HPC</td>
<td>K80, P100, V100 16GB</td>
<td>V100 32GB</td>
<td>Faster Results with Large Datasets</td>
</tr>
<tr>
<td>Compute-Bound HPC</td>
<td>K80, P100, V100 16GB</td>
<td>V100 16GB</td>
<td>Higher TCO</td>
</tr>
</tbody>
</table>
## V100 WITH 32GB HBM2

Maintain Form Factor Compatibility

<table>
<thead>
<tr>
<th>Form Factor</th>
<th>Performance</th>
<th>7.8TF DP, 15.7TF SP, 125TF FP16</th>
<th>7TF DP, 14TF SP, 112TF FP16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form Factor</td>
<td>32GB HBM2</td>
<td>32GB HBM2</td>
<td>7TF DP, 14TF SP, 112TF FP16</td>
</tr>
<tr>
<td>Performance</td>
<td>900GB/s</td>
<td>900GB/s</td>
<td>7TF DP, 14TF SP, 112TF FP16</td>
</tr>
<tr>
<td>Memory Size</td>
<td>NVLink</td>
<td>PCIe Gen3</td>
<td>32GB HBM2</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>Power</td>
<td>300W</td>
<td>250W</td>
</tr>
</tbody>
</table>
WORLD’S HIGHEST BANDWIDTH ON-NODE SWITCH

7.2 Terabits/sec or 900 GB/sec
18 NVLINK ports | 50GB/s per port bi-directional
Fully-connected crossbar
2 billion transistors | 47.5mm x 47.5mm package
ENABLES THE WORLD’S LARGEST GPU

16 Tesla V100 32GB Connected by New NVSwitch
2 petaFLOPS of DL Compute
Unified 512GB HBM2 GPU Memory Space
300GB/sec Every GPU-to-GPU
2.4TB/sec of Total Cross-section Bandwidth
2X HIGHER PERFORMANCE WITH NVSWITCH

- 2x DGX-1 (Volta)
- DGX-2 with NVSwitch
THE WORLD’S FIRST 2 PETAFLOPS SYSTEM

INTRODUCING NVIDIA DGX-2

THE WORLD’S MOST POWERFUL AI SYSTEM FOR THE MOST COMPLEX AI CHALLENGES

- DGX-2 is the newest addition to the DGX family, powered by DGX software
- Deliver accelerated AI-at-scale deployment and simplified operations
- Step up to DGX-2 for unrestricted model parallelism and faster time-to-solution
10X PERFORMANCE GAIN LESS THAN A YEAR

DGX-1, SEP’17

DGX-2, Q3‘18

PyTorch Stack: Time to Train FAIRSEQ

software improvements across the stack including NCCL, cuDNN, etc.
Container Orchestration for DL Training & Inference

KUBERNETES on NVIDIA GPUs

- Scale-up Thousands of GPUs Instantly
- Self-healing Cluster Orchestration
- GPU Optimized Out-of-the-Box
- Powered by NVIDIA Container Runtime
- Included with Enterprise Support on DGX
- Available end of April 2018
INFERENCION ON GPUS TODAY

BING Object Detection
60X Improved Latency

iFLYTEK Speech Recognition
10X Concurrent Requests Per Server

DARWIN AI NN Optimizations
1700X Faster Inference

VALOSSA Video Intelligence
6X Faster Video Processing

TRANSLATE
ALIBABA Neural Machine Translation
3X Requests Per Server

KLM Social Media Engagement
10X increase in customer responses
NVIDIA TESLA PLATFORM SAVES MONEY

Game-Changing Inference Performance

160 CPU Servers
45,000 images/sec
65 KWatts

1 HGX Server
45,000 images/sec
3 KWatts

SAME THROUGHPUT

1/20 THE SPACE

1/22 THE POWER

INFERENCE WORKLOAD:
Image recognition using Resnet 50

INFERENCE WORKLOAD:
Image recognition using Resnet 50

INFERENCE WORKLOAD:
Image recognition using Resnet 50
NVIDIA AI INFERENCE

30M HYPERSONE SERVERS

TensorRT 4
TensorFlow Integration
Kaldi Optimization
ONNX WinML

190X
IMAGE
ResNet-50 with TensorFlow Integration

50X
NLP
GNMT

45X
RECOMMENDER
Neural Collaborative Filtering

36X
SPEECH SYNTH
WaveNet

60X
SPEECH RECOG
DeepSpeech 2 DNN
TensorRT INTEGRATED WITH TensorFlow

Delivers 8x Faster Inference with TensorFlow + TRT

Available in TensorFlow 1.7
https://github.com/tensorflow

CPU: Skylake Gold 6140, 2.5GHz, Ubuntu 16.04; 18 CPU threads.
Volta V100 SXM; CUDA (384.111; v9.0.176);
Batch size: CPU=1, TF_GPU=2, TF-TRT=16 w/ latency=6ms

* Best CPU latency measured at 83 ms
NVIDIA TensorRT 4 RC NOW AVAILABLE

RNN and MLP Layers  •  ONNX Import  •  NVIDIA DRIVE Support

Maximize RNN and MLP Throughput

Optimize and Deploy ONNX Models

Support for NVIDIA DRIVE Xavier

Free download to members of NVIDIA Developer Program
developer.nvidia.com/tensorrt
GPU COMPUTING STACK CONTINUES TO EVOLVE

Beyond Moore’s Law
For HPC

Accomplished Platform
For AI Inference

Beyond Moore’s Law
For HPC

Accelerating Time To Solution
For AI Training

Compelling Platform
For AI Inference

10X In 6 Months

DGX-2
1.5 days

DGX-1 (Volta)
15 days

8X With TensorRT

ResNet-50 Inference @7ms on TensorFlow

GPU Accelerated Computing

Moore’s Law

Accelerating Time To Solution
For AI Training

18X In 5 Years

Measured performance of Amber, CHROMA, GTC, LAMMPS, MILC, NAMD, Quantum Espresso, SPECFEM3D

Beyond Moore’s Law
For HPC

GPU Computing Stack Continues to Evolve
NVIDIA SUPPORT PROGRAMS
What's New

**OptiX 5.0 Available** OptiX 5 SDK features built-in support for motion blur, deep-learning based denoiser and more.

**VRWorks 360 Video 1.1 Available** This release brings many improvements listed below including new calibration technologies.

**CUDA 9.1 Available** New NPP functions for augmentation, multi-GPU enhancement in cuFFT, cuBLAS updates for Volta GPUs and more.

**TensorRT 3 Available** New TensorFlow model optimization, faster mixed-precision for CNNs and RNNs used for vision, speech & NLP.

See More

Developer News

**AI Helps Farmers Distinguish Crop Data in Real Time**
April 9, 2018

**MIT Researchers Use AI to Capture Silent Speech**
April 9, 2018

**Drink up! Beer Tasting Robot Uses AI to Assess Quality**
April 6, 2018

**Researchers Develop AI System for License Plate Recognition**
April 5, 2018

**Boston University Researchers Use AI to Detect Kidney Disease**
April 4, 2018

See More

developer.nvidia.com

Join the NVIDIA Developer Program

Access everything you need to develop with NVIDIA products.

REGISTER NOW
NVIDIA DEEP LEARNING INSTITUTE

Hands-on self-paced and instructor-led training in deep learning and accelerated computing for developers

Request onsite instructor-led workshops at your organization: www.nvidia.com/requestdli

Take self-paced labs online: www.nvidia.com/dlilabs

Download the course catalog, view upcoming workshops, and learn about the University Ambassador Program: www.nvidia.com/dli

Deep Learning Fundamentals
Autonomous Vehicles
Medical Image Analysis
Genomics
Finance
Intelligent Video Analytics
Game Development & Digital Content
Accelerated Computing Fundamentals

More industry-specific training coming soon…

Caffe2
TensorFlow
PyTorch

Deep Learning
Fundamentals

Accelerated Computing
Fundamentals
NVIDIA HW GRANT PROGRAM

Titan X Pascal

- Scientific Computing
- HPC
- Deep Learning

Quadro P5000

- Scientific Visualization
- Virtual Reality

Jetson TX2 (Dev Kit)

- Robotics
- Autonomous Machines
NVIDIA INCEPTION PROGRAM

Accelerating AI startups with powerful GPU tools, tech, and deep learning expertise.

APPLY NOW

http://www.nvidia.com/object/inception-program.html
Pedro Mario Cruz e Silva (pcruzesilva@nvidia.com)
Solution Architect Manager
Enterprise Latin America
Global Oil & Gas Team
LinkedIn