ACCELERATING AI4EO

Carlo Nardone, Sr Solution Architect, NVIDIA

14/11/2018
NVIDIA
25 years long journey from computer gaming to AI Company

Computer Graphics

GPU Computing

Artificial Intelligence
RISE OF GPU COMPUTING

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

Transistors (thousands)

Single-threaded perf

GPU-Computing perf
1.5X per year

1000X by 2025

1.1X per year

1.5X per year
RISE OF GPU COMPUTING

THE BIG BANG IN MACHINE LEARNING

“A.I. is the new electricity”
Andrew Ng - Stanford U. & Baidu Research
A BIG BANG IN EO?

AI4EO: A.I. for Earth Observation!
POWERING THE AI REVOLUTION

NVIDIA is advancing GPU computing for deep learning and AI at the speed of light. We create the entire stack. It starts with the most advanced GPUs and the systems and software we build on top of them. We integrate and optimize every deep learning framework. We work with the major systems companies and every major cloud service provider to make GPUs available in datacenters and in the cloud. And we create computers and software to bring AI to the edge, from self-driving cars to autonomous robots to medical devices.
TESLA V100 32GB
World’s Most Advanced Data Center GPU
For Scale-up Environments

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
**TESLA V100 32GB**

World’s Most Advanced Data Center GPU
For Scale-up Environments

- 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
DGX-2: FULL NON-BLOCKING 16-WAY FABRIC
TESLA T4
WORLD’S MOST ADVANCED INFERENCE GPU

Universal Inference Acceleration
320 Turing Tensor cores
2,560 CUDA cores
65 FP16 TFLOPS  |  130 INT8 TOPS  |  260 INT4 TOPS
16GB  |  320GB/s
JETSON AGX XAVIER
World’s first AI computer for Autonomous Machines

AI Server Performance in 30W • 15W • 10W
512 Volta CUDA Cores • 2x NVDLA
8 core CPU
32 DL TOPS
NVIDIA DEEP LEARNING SOFTWARE STACK

**TRAINING**
- Training Data
- Data Management
- Training
- Model Assessment
- Trained Neural Network

**INFEERENCE**
- Data center
- Embedded
- Automotive
- TensorRT
- JETPACK SDK
- DriveWorks SDK

**NVIDIA DEEP LEARNING SDK and CUDA**

[developer.nvidia.com/deep-learning-software]
RAPIDS
A Geospatial use case

Geospatial Test:

100M (x,y) points
Convert latitude/longitude
0.01 mm accuracy

1 CPU core: 65 minutes
CPU multi-core: 13 minutes

RAPIDS + GPU: 2 seconds

Dell T7910 Xeon E5-2640V4x2 + NVIDIA Titan Xp
RAPIDS — ACCELERATING DATA SCIENCE
Open Software Stack for Python

Data Preparation ➔ Model Training ➔ Visualization

DASK

PYTHON

RAPIDS

CUDA

APACHE ARROW

CUML

CUDF

CUGRAPH

CUDNN

DEEP LEARNING FRAMEWORKS
THANK YOU!
cnardone @ nvidia.com
NVIDIA POWERS WORLD’S FASTEST SUPERCOMPUTERS
127 Systems in Top500 (+48% YoY) | 22 of Top 25 Greenest

ORNL Summit
World’s Fastest
27,648 GPUs | 144 PF

LLNL Sierra
World’s 2nd Fastest
17,280 GPUs | 95 PF

Piz Daint
Europe’s Fastest
5,704 GPUs | 21 PF

ABCI
Japan’s Fastest
4,352 GPUs | 20 PF

ENI HPC4
Fastest Industrial
3,200 GPUs | 12 PF
NVIDIA DGX-2: “WORLD’S LARGEST GPU”

1. NVIDIA Tesla V100 32GB
2. Two GPU Boards
   - 8 V100 32GB GPUs per board
   - 6 NVSwitches per board
   - 512GB Total HBM2 Memory interconnected by Plane Card
3. Twelve NVSwitches
   - 2.4 TB/sec bi-section bandwidth
4. Eight EDR Infiniband/100 GigE
   - 1600 Gb/sec Total Bi-directional Bandwidth
5. PCIe Switch Complex
6. Two Intel Xeon Platinum CPUs
7. 1.5 TB System Memory
8. 30 TB NVME SSDs
   - Internal Storage
NVIDIA SDK & LIBRARIES

CUSTOMER USECASES
- Speech
- Translate
- Recommender

APPS & FRAMEWORKS
- Python
- TensorFlow
- ONNX
- MXNet
- PyTorch
- Chainer
- RAPIDS
- TESLA UNIVERSAL ACCELERATION PLATFORM

MACHINE LEARNING | RAPIDS
- cuDF
- cuML
- cuGRAPH

DEEP LEARNING
- cuDNN
- cuBLAS
- CUTLASS
- NCCL
- TensorRT

SUPERCOMPUTING
- CuBLAS
- CuFFT
- OpenACC

CUDA

TESLA GPUs & SYSTEMS
- TESLA GPU
- VIRTUAL GPU
- NVIDIA DGX FAMILY
- NVIDIA HGX
- SYSTEM OEM
- CLOUD

CONSUMER INTERNET

INDUSTRIAL APPLICATIONS

SUPERCOMPUTING

Molecular Simulations
Weather Forecasting
Seismic Mapping

Speech
Translate
Recommender
Healthcare
Manufacturing
Finance

Amber
NAMD
ANSYS
+Sims

+550 Applications
NGC CONTAINER REGISTRY

Simple access to accelerated GPU software: ngc.nvidia.com
TRAINING VS INFEERENCE

TRAINING:
- Large N
- Forward and backward pass
- Error calculation

INFEERENCE:
- Smaller, varied N
- Forward pass
- "car" output
NVIDIA TENSORRT
Programmable Inference Accelerator

FRAMEWORKS
- Caffe2
- Chainer
- PyTorch
- TensorFlow
- mxnet
- Theano

GPU PLATFORMS
- TESLA P4
- JETSON TX2
- DRIVE PX 2
- TESLA V100
- NVIDIA DLA

TensorRT
Optimize
Runtime

developer.nvidia.com/tensorrt
## JETSON AGX XAVIER

<table>
<thead>
<tr>
<th></th>
<th>JETSON TX2</th>
<th>JETSON AGX XAVIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>256 Core Pascal</td>
<td>512 Core Volta</td>
</tr>
<tr>
<td>DL Accelerator</td>
<td>-</td>
<td>NVDLA x 2</td>
</tr>
<tr>
<td>Vision Accelerator</td>
<td>-</td>
<td>VLA - 7 way VLIW Processor</td>
</tr>
<tr>
<td>CPU</td>
<td>6 core Denver and A57 CPUs</td>
<td>8 core Carmel CPUs</td>
</tr>
<tr>
<td>Memory</td>
<td>8 GB 128 bit LPDDR4 58.4 GB/s</td>
<td>16 GB 256 bit LPDDR4x 137 GB/s</td>
</tr>
<tr>
<td>Storage</td>
<td>32 GB eMMC</td>
<td>32 GB eMMC</td>
</tr>
<tr>
<td>Video Encode</td>
<td>2x 4K @30 HEVC</td>
<td>2x 4K @ 60 / 4x 4K @30 HEVC</td>
</tr>
<tr>
<td>Video Decode</td>
<td>2x 4K @30 12 bit support</td>
<td>2x 8K @ 30 / 8x 4K @30 12 bit support</td>
</tr>
<tr>
<td>Camera</td>
<td>Up to 6 cameras CSI2 D-PHY 1.2 2.5 Gbps/lane</td>
<td>Up to 8 cameras CSI2 D-PHY 1.2 2.5 Gbps/lane</td>
</tr>
<tr>
<td>Mechanical</td>
<td>50mm x 87mm 400 pin connector</td>
<td>100mm x 87mm 699 pin connector</td>
</tr>
</tbody>
</table>

*New!*

**New!**

**New!**

**New!**

**x2**

**x2**

**x4**

**+2**
JETSON AGX XAVIER
20X PERFORMANCE IN 18 MONTHS

24x DL / AI

Jetson TX2: 1.3
Jetson AGX Xavier: 32

8x CUDA

Jetson TX2: 1.3
Jetson AGX Xavier: 11

2x CPU

Jetson TX2: 55
Jetson AGX Xavier: 112

2.4x DRAM BW

Jetson TX2: 58
Jetson AGX Xavier: 137

4x CODEC

Jetson TX2: 2
Jetson AGX Xavier: 8
JETSON AGX XAVIER
GPU WORKSTATION PERF • 1/10TH POWER

Al Inference Performance

<table>
<thead>
<tr>
<th></th>
<th>Core i7 + GTX 1070</th>
<th>Jetson AGX Xavier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resnet 50 Images/s</td>
<td>1200</td>
<td>1800</td>
</tr>
</tbody>
</table>

Al Inference Efficiency

<table>
<thead>
<tr>
<th></th>
<th>Core i7 + GTX 1070</th>
<th>Jetson AGX Xavier</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogleNet Images/s/W</td>
<td>40</td>
<td>100</td>
</tr>
</tbody>
</table>