MACHINE LEARNING WITH NVIDIA AND IBM POWER AI

July 2017

Joerg Krall
Sr. Business Development Manager
MFG EMEA
jkrall@nvidia.com
A NEW ERA OF COMPUTING

1995
PC INTERNET
WinTel, Yahoo!
1 billion PC users

2005
MOBILE-CLOUD
iPhone, Amazon AWS
2.5 billion mobile users

2015
AI & IOT
Deep Learning, GPU
100s of billions of devices
NVIDIA
“THE AI COMPUTING COMPANY”
TEN YEARS OF GPU COMPUTING

2006: CUDA Launched
2008: World’s First GPU Top500 System
2010: Fermi: World’s First HPC GPU
2012: Discovered How H1N1 Mutates to Resist Drugs
2014: Stanford Builds AI Machine using GPUs
2016: World’s First 3-D Mapping of Human Genome

Oak Ridge Deploys World’s Fastest Supercomputer w/ GPUs

AlexNet beats expert code by huge margin using GPUs

Google Outperforms Humans in ImageNet

World’s First Atomic Model of HIV Capsid

GPU-Trained AI Machine Beats World Champion in Go
AI, ML, DL……?
GPU COMPUTING HAS REACHED A TIPPING POINT

3x GPU Developers in 2 Years

- 2014: 120,000
- 2016: 400,000

13x Organizations Engaged with NVIDIA for DL

- Higher Ed
- Internet
- Healthcare
- Developer Tools
- Government
- Automotive
- Finance
- Others

- 2014: 1,549
- 2016: 19,439
THE EXPANDING UNIVERSE OF MODERN AI

THE BIG BANG
Big Data
GPU Algorithms

RESEARCH
Berkeley
Language Models
NYU

CORE TECHNOLOGY / FRAMEWORKS
facebook
Google
IBM Watson
clarifai

AI-as-a-PLATFORM
amazon web services
NVIDIA cuDNN

START-UPS
nervana
drive.ai

INDUSTRY LEADERS
SIEMEN
SALE

3,000+ AI START-UPS
$5B IN FUNDING
Source: Venture Scanner
TESLA PLATFORM POWERS LEADING DATA CENTERS FOR HPC AND AI
DEEP LEARNING
A NEW COMPUTING MODEL

Traditional Computer Vision

Domain experts design feature detectors
Quality = patchwork of algorithms
Need CV experts and time

Deep Learning Object Detection

DNN learn features from large data
Quality = data & training method
Needs lots of data and compute
DEEP LEARNING — A NEW COMPUTING MODEL
“Software that writes software”

LEARNING ALGORITHM
“millions of trillions of FLOPS”

“little girl is eating piece of cake”
SELF-DRIVING CARS ARE AN AI CHALLENGE

DEEP LEARNING
AI FOR SELF-DRIVING CARS

FREE SPACE DETECTION

CAR 3D DETECTION
AI IS EVERYWHERE

“Find where I parked my car”
“Find the bag I just saw in this magazine”
“What movie should I watch next?”
TOUCHING OUR LIVES

Bringing grandmother closer to family by bridging language barrier

Predicting sick baby’s vitals like heart rate, blood pressure, survival rate

Enabling the blind to “see” their surrounding, read emotions on faces
FUELING ALL INDUSTRIES

Increasing public safety with smart video surveillance at airports & malls

Providing intelligent services in hotels, banks and stores

Separating weeds as it harvests, reduces chemical usage by 90%
TESLA REVOLUTIONIZES DEEP LEARNING

GOOGLE BRAIN APPLICATION

<table>
<thead>
<tr>
<th></th>
<th>BEFORE TESLA</th>
<th>AFTER TESLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>$5,000K</td>
<td>$200K</td>
</tr>
<tr>
<td>Servers</td>
<td>1,000 Servers</td>
<td>16 Tesla Servers</td>
</tr>
<tr>
<td>Energy</td>
<td>600 KW</td>
<td>4 KW</td>
</tr>
<tr>
<td>Performance</td>
<td>1x</td>
<td>6x</td>
</tr>
</tbody>
</table>
DEEP LEARNING WORKFLOWS

**IMAGE CLASSIFICATION**

Classify images into classes or categories
Object of interest could be anywhere in the image

98% Dog
2% Cat

**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
GPU DEEP LEARNING IS A NEW COMPUTING MODEL

Billions of Trillions of Operations
GPU train larger models, accelerate time to market
GPU DEEP LEARNING IS A NEW COMPUTING MODEL

10s of billions of image, voice, video queries per day
GPU inference for fast response, maximize datacenter throughput

DATACENTER INFERENCE
GPU DEEP LEARNING IS A NEW COMPUTING MODEL

**TRAINING**

- Billions of Trillions of Operations
- GPU train larger models, accelerate time to market

**DATACENTER INFERENCING**

- 10s of billions of image, voice, video queries per day
- GPU inference for fast response, maximize datacenter throughput

**DEVICE INFERENCING**

- Billions of intelligent devices & machines
- Recognition, reasoning, problem solving
- GPU inference: real-time accurate response
NEW AI SERVICES POSSIBLE WITH GPU CLOUD

- **SPOTIFY**
  - SONG RECOMMENDATIONS

- **NETFLIX**
  - VIDEO RECOMMENDATIONS

- **YELP**
  - SELECTING COVER PHOTOS
DL INFERENCE ON GPU CLOUD COMPUTING

PINTEREST VISUAL SEARCH

TWITTER VIDEO CLASSIFICATION

KLM CUSTOMER SERVICE
## Deep Learning Requirements

<table>
<thead>
<tr>
<th>Deep Learning Needs...</th>
<th>Deep Learning Challenges</th>
<th>NVIDIA Delivers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scientists</td>
<td>Demand far exceeds supply</td>
<td>DIGITS, DLI Training</td>
</tr>
<tr>
<td>Latest Algorithms</td>
<td>Rapidly evolving</td>
<td>DL SDK, GPU-Accelerated Frameworks</td>
</tr>
<tr>
<td>Fast Training</td>
<td>Impossible -&gt; Practical</td>
<td>DGX-1, P100, P40</td>
</tr>
<tr>
<td>Deployment Platform</td>
<td>Must be available everywhere</td>
<td>TensorRT, P40, P4, Jetson, Drive PX</td>
</tr>
</tbody>
</table>
TESLA PLATFORM FOR HPC AND AI
ONE ARCHITECTURE BUILT FOR BOTH DATA SCIENCE & COMPUTATIONAL SCIENCE

AlexNet Training
8x P100 Faster than 128 Knights Landing Servers

GTC-P: Plasma Turbulence
8x P100 Faster than 64 Knights Landing Servers

Based on AlexNet Batch size 256, weak scaling up to 32 KNL servers, 64 & 128 estimated based on ideal scaling, Xeon Phi 7250 Nodes

GTC-P, Grid Size A, Systems: NVIDIA DGX-1, 8x P100, Intel KNL 7250 68 core Flat-Quadrant mode, Omnipath
TWO BIG INEFFICIENCIES WITH CPU NODES

Most of budget spent on non-compute overhead

Network overhead causes performance inefficiencies

Typical Data Center Budget

- Compute Servers, 39%
- Rack, Cabling
- Networking
- Non-compute, 61%
- Infrastructure

Sources: Microsoft Research on Datacenter Costs, Amber Simulations on SDSU Comet Supercomputer
WEAK NODES
Lots of Nodes Interconnected with Vast Network Overhead

STRONG NODES
Few Lightning-Fast Nodes with Performance of Hundreds of Weak Nodes
ONE PLATFORM BUILT FOR BOTH DATA SCIENCE & COMPUTATIONAL SCIENCE

Tesla Platform

Accelerating AI

Accelerating HPC
New GPU Architecture to Enable the World’s Fastest Compute Node

- Pascal Architecture: Highest Compute Performance
- NVLink: GPU Interconnect for Maximum Scalability
- CoWoS HBM2: Unifying Compute & Memory in Single Package
- Page Migration Engine: Simple Parallel Programming with Virtually Unlimited Memory Space
GIANT LEAPS
IN EVERYTHING
GPU DEEP LEARNING IS A NEW COMPUTING MODEL

Billions of Trillions of Operations
GPU train larger models, accelerate time to market
# TESLA P100 ACCELERATORS

<table>
<thead>
<tr>
<th></th>
<th>Tesla P100 with NVLink</th>
<th>Tesla P100 for PCIe</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compute</strong></td>
<td>5.3 TF DP · 10.6 TF SP · 21.2 TF HP</td>
<td>4.7 TF DP · 9.3 TF SP · 18.7 TF HP</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>HBM2: 732 GB/s · 16 GB</td>
<td>HBM2 16GB: 732 GB/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HBM2 12GB: 549 GB/s</td>
</tr>
<tr>
<td><strong>Interconnect</strong></td>
<td>NVLink (160 GB/s) + PCIe Gen3 (32 GB/s)</td>
<td>PCIe Gen3 (32 GB/s)</td>
</tr>
<tr>
<td><strong>Programmability</strong></td>
<td>Page Migration Engine Unified Memory</td>
<td>Page Migration Engine Unified Memory</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>300W</td>
<td>250W</td>
</tr>
</tbody>
</table>

*Interconnect Speed is measured bi-directional*
GPU DEEP LEARNING IS A NEW COMPUTING MODEL

10s of billions of image, voice, video queries per day
GPU inference for fast response, maximize datacenter throughput

DATACENTER INFERENCING
**TESLA P40**

Highest Throughput for Scale-up Servers

**P40**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># of CUDA Cores</td>
<td>3840</td>
</tr>
<tr>
<td>Peak Single Precision</td>
<td>12 TeraFLOPS</td>
</tr>
<tr>
<td>Peak INT8</td>
<td>47 TOPS</td>
</tr>
<tr>
<td>Low Precision</td>
<td>4x 8-bit vector dot product with 32-bit accumulate</td>
</tr>
<tr>
<td>Video Engines</td>
<td>1x decode engine, 2x encode engines</td>
</tr>
<tr>
<td>GDDR5 Memory</td>
<td>24 GB @ 346 GB/s</td>
</tr>
<tr>
<td>Power</td>
<td>250W</td>
</tr>
</tbody>
</table>

GoogLeNet, AlexNet, batch size = 128, CPU: Dual Socket Intel E5-2697v4

**4x Boost in Less than One Year**

- 8x M40 (FP32)
- 8x P40 (INT8)
TESLA P4

Maximum Efficiency for Scale-out Servers

40x Efficient vs CPU, 8x Efficient vs FPGA

- **CPU**
- **FPGA**
- **1x M4 (FP32)**
- **1x P4 (INT8)**

<table>
<thead>
<tr>
<th>P4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># of CUDA Cores</td>
<td>2560</td>
</tr>
<tr>
<td>Peak Single Precision</td>
<td>5.5 TeraFLOPS</td>
</tr>
<tr>
<td>Peak INT8</td>
<td>22 TOPS</td>
</tr>
<tr>
<td>Low Precision</td>
<td>4x 8-bit vector dot product with 32-bit accumulate</td>
</tr>
<tr>
<td>Video Engines</td>
<td>1x decode engine, 2x encode engine</td>
</tr>
<tr>
<td>GDDR5 Memory</td>
<td>8 GB @ 192 GB/s</td>
</tr>
<tr>
<td>Power</td>
<td>50W &amp; 75 W</td>
</tr>
</tbody>
</table>

AlexNet, batch size = 128, CPU: Intel E5-2690v4 using Intel MKL 2017, FPGA is Arria10-115
1x M4/P4 in node, P4 board power at 56W, P4 GPU power at 36W, M4 board power at 57W, M4 GPU power at 39W. Perf/W chart using GPU power
GPU SERVERS WITH NVLINK
S822LC FOR HPC: RECOMMENDED CONFIGURATION FOR POWERAI
2 SOCKET, 4 GPU SYSTEM WITH NVLINK

Required:
2 POWER8 10 Core CPUs
4 NVIDIA P100 ”Pascal” GPUs
256 GB System Memory
2 SSD storage devices
High-speed interconnect (IB or Ethernet, depending on infrastructure)

Optional:
Up to 1 TB System Memory
PCIe attached NVMe storage
POWERAI TAKES ADVANTAGE OF NVLINK BETWEEN POWER8 & P100 TO INCREASE SYSTEM BANDWIDTH

NVLink between CPUs and GPUs enables fast memory access to large data sets in system memory

Two NVLink connections between each GPU and CPU-GPU leads to faster data exchange
NVLink and P100 Advantage: Reducing Communication Time, Incorporating the Fastest GPU for Deep Learning

NVLink reduces communication time and overhead

Data gets from GPU-GPU, Memory-GPU faster, for shorter training times
Refer to “Measurement Config.” slide for HW & SW details, run with real data set
AlexnetOWT/GoogLenet use total batch size=1024, VGG-D uses total batch size=512, Incep-v3/ResNet-50 use total batch size=256
## NVIDIA DEEP LEARNING PARTNERS

<table>
<thead>
<tr>
<th>Graph Analytics</th>
<th>DL Frameworks</th>
<th>Enterprise DL</th>
<th>Enterprises</th>
<th>Data Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>blazegraph</td>
<td>Caffe</td>
<td>bonsai</td>
<td>IBM, SAP</td>
<td>Hortonworks</td>
</tr>
<tr>
<td>graphistry</td>
<td>Chainer</td>
<td>deepinstinct</td>
<td></td>
<td>splunk</td>
</tr>
<tr>
<td>kinetica</td>
<td>CNTK</td>
<td>SKYMIND</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAPD</td>
<td>TensorFlow</td>
<td>synerScope</td>
<td>accenture Technology Labs, deepsense.io, Spark,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>theano</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TELSA PASCAL FAMILY
END-TO-END PRODUCT FAMILY

TRAINING

FULLY INTEGRATED DL SUPERCOMPUTER

DGX-1

DATA CENTER

Tesla P4

Tesla P40

INFEERENCE

DATA CENTER

Tesla P4

Tesla P40

AUTOMOTIVE

Drive PX2

EMBEDDED

Jetson TX1

DESKTOP

Quadro GP100

Tesla P40 & P4
# TESLA PRODUCTS RECOMMENDATION

<table>
<thead>
<tr>
<th>PRODUCT</th>
<th>P100 NVLINK</th>
<th>P100 PCIE</th>
<th>P40</th>
<th>P4</th>
</tr>
</thead>
</table>
| **Target Use Cases** | • Highest DL training perf  
• Fastest time-to-solution  
• Larger “Model Parallel” DL model with 16GB x 8 | • HPC DC running mix of CPU and GPU workload  
• Best throughput / $ with mix workload | • Highest inference perf  
• Larger “Data Parallel” DL model with 24GB | • Low power, low profile optimized for scale out deployment  
• Most efficient inference and video processing |
<p>| <strong>Best Configs.</strong> | • 8 way Hybrid Cube Mesh | • 2-4 GPU/node (HPC) | • Up to 8 GPU/node | • 1-2 GPU/node |
| <strong>1st Server Ship</strong> | • Available Now | • Available Now | • OEM starting Oct’16 | • OEM starting Nov’16 |</p>
<table>
<thead>
<tr>
<th></th>
<th>K80</th>
<th>M40</th>
<th>M4</th>
<th>P100 (SXM2)</th>
<th>P100 (PCIE)</th>
<th>P40</th>
<th>P4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>2x GK210</td>
<td>GM200</td>
<td>GM206</td>
<td>GP100</td>
<td>GP100</td>
<td>GP102</td>
<td>GP104</td>
</tr>
<tr>
<td>PEAK FP64 (TFLOPs)</td>
<td>2.9</td>
<td>NA</td>
<td>NA</td>
<td>5.3</td>
<td>4.7</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>PEAK FP32 (TFLOPs)</td>
<td>8.7</td>
<td>7</td>
<td>2.2</td>
<td>10.6</td>
<td>9.3</td>
<td>12</td>
<td>5.5</td>
</tr>
<tr>
<td>PEAK FP16 (TFLOPs)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>21.2</td>
<td>18.7</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>PEAK TIOPs</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>47</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Memory Size</td>
<td>2x 12GB GDDR5</td>
<td>24 GB GDDR5</td>
<td>4 GB GDDR5</td>
<td>16 GB HBM2</td>
<td>16/12 GB HBM2</td>
<td>24 GB GDDR5</td>
<td>8 GB GDDR5</td>
</tr>
<tr>
<td>Memory BW</td>
<td>480 GB/s</td>
<td>288 GB/s</td>
<td>80 GB/s</td>
<td>732 GB/s</td>
<td>732/549 GB/s</td>
<td>346 GB/s</td>
<td>192 GB/s</td>
</tr>
<tr>
<td>Interconnect</td>
<td>PCIe Gen3</td>
<td>PCIe Gen3</td>
<td>PCIe Gen3</td>
<td>NVLINK + PCIe Gen3</td>
<td>PCIe Gen3</td>
<td>PCIe Gen3</td>
<td>PCIe Gen3</td>
</tr>
<tr>
<td>ECC</td>
<td>Internal + GDDR5</td>
<td>GDDR5</td>
<td>GDDR5</td>
<td>Internal + HBM2</td>
<td>Internal + HBM2</td>
<td>GDDR5</td>
<td>GDDR5</td>
</tr>
<tr>
<td>Form Factor</td>
<td>PCIE Dual Slot</td>
<td>PCIE Dual Slot</td>
<td>PCIE LP</td>
<td>SXM2</td>
<td>PCIE Dual Slot</td>
<td>PCIE Dual Slot</td>
<td>PCIE LP</td>
</tr>
<tr>
<td>Power</td>
<td>300 W</td>
<td>250 W</td>
<td>50-75 W</td>
<td>300 W</td>
<td>250 W</td>
<td>250 W</td>
<td>50-75 W</td>
</tr>
</tbody>
</table>
TESLA VOLTA IS COMMING ...
TESLA V100
THE MOST ADVANCED DATA CENTER GPU EVER BUILT

5,120 CUDA cores
640 NEW Tensor cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS
120 Tensor TFLOPS
20MB SM RF | 16MB Cache | 16GB HBM2 @ 900 GB/s
300 GB/s NVLink
NEW TENSOR CORE BUILT FOR AI
Delivering 120 TFLOPS of DL Performance

MATRIX DATA OPTIMIZATION:
Dense Matrix of Tensor Compute

TENSOR-OP CONVERSION:
FP32 to Tensor Op Data for Frameworks

VOLTA-TENSOR CORE
4x4 matrix processing array

\[ D[\text{FP32}] = A[\text{FP16}] \times B[\text{FP16}] + C[\text{FP32}] \]
Optimized For Deep Learning

VOLTA-OPTIMIZED cuDNN

ALL MAJOR FRAMEWORKS
REVOLUTIONARY AI PERFORMANCE

3X Faster DL Training Performance

Googlenet Training Performance (Speedup Vs K80)

- 8x V100 cuDNN7
- 8x P100 cuDNN6
- 4x M40 cuDNN3
- 1x K80 cuDNN2

Over 80x DL Training Performance in 3 Years

LSTM Training (Neural Machine Translation)

- 2X CPU: 15 Days
- 1X P100: 18 Hours
- 1X V100: 6 Hours

3X Reduction in Time to Train Over P100

Multi-Node Training with NCCL2.0 (ResNet-50)

- 8X P100: 18 Hours
- 8X V100: 7.4 Hours
- 64X V100: 1 Hour

85% Scale-Out Efficiency Scales to 64 GPUs with Microsoft Cognitive Toolkit

Neural Machine Translation Training for 13 Epochs | German >-English, WMT15 subset | CPU = 2x Xeon E5 2699 V4 | V100 performance measured on pre-production hardware.

ResNet50 Training for 90 Epochs with 1.3M images dataset | Cognitive Toolkit with NCCL 2.0 | V100 performance measured on pre-production hardware.
### SINGLE UNIVERSAL GPU FOR ALL ACCELERATED WORKLOADS

**Boosts all accelerated workloads**

<table>
<thead>
<tr>
<th>Workload</th>
<th>Vs P100</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPC</td>
<td>1.5X</td>
</tr>
<tr>
<td>AI Training</td>
<td>3X</td>
</tr>
<tr>
<td>AI Inference</td>
<td>3X</td>
</tr>
<tr>
<td>Virtual Desktop</td>
<td>2X</td>
</tr>
</tbody>
</table>

**V100 Universal GPU**
# TESLA V100 SPECIFICATIONS

<table>
<thead>
<tr>
<th>Compute</th>
<th>7.5 TF DP • 15 TF SP • 120 TF DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>HBM2: 900 GB/s • 16 GB</td>
</tr>
<tr>
<td>Interconnect</td>
<td>NVLink (up to 300 GB/s) + PCIe Gen3 (up to 32 GB/s)</td>
</tr>
<tr>
<td>Availability</td>
<td>DGX-1: Q3 2017</td>
</tr>
<tr>
<td></td>
<td>OEM : Q4 2017</td>
</tr>
</tbody>
</table>