GPUs & Deep Learning at scale

Robert Ober
Tesla Chief Platform Architect, NVIDIA
Exponential Performance Needed

Andrew Ng @ Baidu: DeepSpeech2 snapshot

deep + larger

90% → 99% recognition

10’s of ExaFlop (10^18) / cycle

~4TByte data / cycle

training:

8 hours = ~350 TFlop/s, 142MB/s

2 hours = ~1.5 PFlop/s, 568 MB/s
GPUs: Accurate results in real time

- Large NN
- Medium NN
- Small NN
- Traditional Model

CPU vs GPU performance comparison:
- 99%
- 95%
- 90%
- 85%

- 200ms Real Time Region
Matrix Math
Basis of Graphics

Transforms
Every:
  Pixel
  Shader
over and over…
Parallel Threads, SIMD
GPU Architecture

Two Main Components

Global memory: shared with host

Streaming Multiprocessors (SM)
  control units, registers,
  execution pipelines, caches
GPU Architecture
Streaming Multiprocessor (SM)

Many CUDA Cores per SM
Special-function units
  \( \cos/\sin/\tan, \text{ etc.} \)
Shared mem + L1 cache
Thousands of 32-bit registers
GPU Architecture

CUDA Core

Floating point & Integer unit
Fused multiply-add (FMA) instruction
don’t forget deep learning instructions
Logic unit
Move, compare unit
Branch unit
For DL
Matrix Math
Basis of Deep Learning

Convolution:
155M Parameters
Neurons
Layers
Iterations…

40 GOPs / Inference
Inference
Computational Load and Latency

An Analysis of Deep Neural Network Models for Practical Applications
Canziani, Culurciello, Paszke, May 2016
Throughput & Latency

25 : 1 in real time inference   FP32

- **Inferences/s vs Batch Size (bigger is better)**
  - E5-2698v3 Infer / sec
  - Pascal Infer / sec
  - 20x to 80x

- **Inference Latency Log (ms) vs Batch Size (smaller is better)**
  - E5-2698v3 Latency (ms)
  - Pascal Latency (ms)

Real Time Region

VGG19 Inference
1x E5-2698v3 FP32, IntelCaffe + MKL 2017   vs   1x Pascal FP32, GIE
Throughput & Latency

Updated: googLeNet  FP32, INT8

- GoogLeNet

E5-2690v4 12 core, @2.6GHz, 3.5GHz turbo, HT on, MKL 2017 gold release

TensorRT

- Inf/s vs Batch (bigger is better)
  - E5-2690v4 Inf/s
  - P4 Inf/s
  - P4 int8 Inf/s

- log (ms) vs Batch (small is better)
  - E5-2690v4 (ms)
  - P4 (ms)
  - P4 int8 (ms)

Real Time Region: 40x
HPC Scale
HPC Top 500
HPC Accelerated Clusters In the top 500

"Accelerators Will Be Installed in More than Half of New Systems"

Source: Top500.org
Source: Top 6 predictions for HPC in 2015
## Data center GPU

### Specialized Hardware

#### Updated

<table>
<thead>
<tr>
<th></th>
<th>P100 SXM2 + DGX-1</th>
<th>P100 + P40</th>
<th>P4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Form Factor</strong></td>
<td>Custom SXM2</td>
<td>PCIe Double Wide</td>
<td>PCIe ½ x ½</td>
</tr>
<tr>
<td><strong>Watts</strong></td>
<td>300</td>
<td>250</td>
<td>75, 50</td>
</tr>
<tr>
<td><strong>Single Precision FP 32</strong></td>
<td>10 Tflops (x8)</td>
<td>8-12 TFlops</td>
<td>5.5 TFlops</td>
</tr>
<tr>
<td><strong>Half Precision FP 16</strong></td>
<td>20 Tflops (x8)</td>
<td>16 TFlops</td>
<td>--</td>
</tr>
<tr>
<td><strong>Integer 8 Dot Product</strong></td>
<td>--</td>
<td>47 TOPs</td>
<td>22 TOPs</td>
</tr>
</tbody>
</table>
## Styles of Scale

**Depend on Use**

<table>
<thead>
<tr>
<th>Scale Up</th>
<th>A bit of Both</th>
<th>Scale Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node</td>
<td>Clusters of 8 Specialized Node</td>
<td>2 to 8 Specialized Node</td>
</tr>
<tr>
<td>GPU Interconnect</td>
<td>NVLINK Cube Mesh Fabric</td>
<td>PCIe + DMA Mailbox</td>
</tr>
<tr>
<td>Network</td>
<td>IB or 100GE RDMA</td>
<td>IB or ENet</td>
</tr>
<tr>
<td>Use</td>
<td>HPC + DL Training Model Parallel</td>
<td>HPC + DL Training Data Parallel + Inference</td>
</tr>
</tbody>
</table>
Deep Learning End to End
Towards a RESTful place

Source Dataset → IMPORT Format...
Curated Dataset → PREPROCESS clean, clip, label, Normalize, ...

MODEL ZOO → TRAIN

DEPLOY tune, compile + runtime → RESULT *
inference, prediction

INFERENCe & MICROSERVICES

REST API
# Datacenter Deployment

**Tools to Operationalize GPUs**

**Updated**

<table>
<thead>
<tr>
<th>Monitoring + Admin</th>
<th>Deployment</th>
<th>Resource Scheduling + Management</th>
<th>Frameworks + Libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCGM</td>
<td>Hypervisor Passthrough</td>
<td>DXO</td>
<td>All DL frameworks</td>
</tr>
<tr>
<td>Docker, AMI</td>
<td>Mesos</td>
<td></td>
<td>CUDA, CuDNN</td>
</tr>
<tr>
<td>TensorRT</td>
<td></td>
<td></td>
<td>CuBLAS</td>
</tr>
</tbody>
</table>
Redux

GPU Scale Deployment

Neural Networks
Accelerators
GPUs ↔ Deep Learning
Architectural Optimizations
Optimized Inference + REST
DCGM
$10^{18}$ FLOPs