S6474 - FROM WORKSTATION TO EMBEDDED: ACCELERATED DEEP LEARNING ON NVIDIA JETSON™ TX1

Julie Bernauer, 4/5/2016
GPU COMPUTING
GPU COMPUTING
A simple sum of two vectors (arrays) in C

```c
void vector_add(int n, const float *a, const float *b, float *c)
{
    for (int idx = 0 ; idx < n ; ++idx )
        c[idx] = a[idx] + b[idx];
}
```

GPU friendly version in CUDA

```c
__global__ void vector_add(int n, const float *a, const float *b, float *c)
{
    int idx = blockIdx.x*blockDim.x + threadIdx.x;
    if( idx < n )
        c[idx] = a[idx] + b[idx];
}
```
GPU ACCELERATED LIBRARIES
“Drop-in” Acceleration for Your Applications

Linear Algebra
FFT, BLAS, SPARSE, Matrix, cuSolver

Numerical & Math
RAND, Statistics

Data Struct. & AI
Sort, Scan, Zero Sum

Visual Processing
Image & Video
DEEP NEURAL NETWORKS AND GPUS
ACCELERATING INSIGHTS

“Now You Can Build Google’s $1M Artificial Brain on the Cheap”

Now You Can Build Google’s $1M Artificial Brain on the Cheap

Deep learning with COTS HPC systems, A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, B. Catanzaro ICML 2013
MODERN AI

2012: GOOGLE BRAIN

2016: AlphaGO
DEEP LEARNING EVERYWHERE

INTERNET & CLOUD
- Image Classification
- Speech Recognition
- Language Translation
- Language Processing
- Sentiment Analysis
- Recommendation

MEDICINE & BIOLOGY
- Cancer Cell Detection
- Diabetic Grading
- Drug Discovery

MEDIA & ENTERTAINMENT
- Video Captioning
- Video Search
- Real Time Translation

SECURITY & DEFENSE
- Face Detection
- Video Surveillance
- Satellite Imagery

AUTONOMOUS MACHINES
- Pedestrian Detection
- Lane Tracking
- Recognize Traffic Sign
NVIDIA GPU: THE ENGINE OF DEEP LEARNING

WATSON
- IBM

CHAINER
- Preferred Networks

THEANO
- Université de Montréal

MATCONVNET
- University of Oxford

TENSORFLOW
- Google

CNTK
- Microsoft

TORCH
- Facebook

CAFFE
- Berkeley

NVIDIA CUDA ACCELERATED COMPUTING PLATFORM
ACCELERATING DEEP LEARNING: CUDNN

- GPU-accelerated Deep Learning subroutines
- High performance neural network training
- Accelerates Major Deep Learning frameworks: Caffe, Theano, Torch

Caffe Performance

AlexNet training throughput based on 20 iterations, CPU: 1x E5-2680v3 12 Core 2.5GHz. 128GB System Memory, Ubuntu 14.04

developer.nvidia.com/cudnn
MULTI-GPU COMMUNICATION: NCCL

Collective library

- Research library of accelerated collectives that is easily integrated and topology-aware so as to improve the scalability of multi-GPU applications
- Pattern the library after MPI’s collectives
- Handle the intra-node communication in an optimal way
- Provide the necessary functionality for MPI to build on top to handle inter-node

[GitHub Link](https://github.com/NVIDIA/nccl)
NCCL EXAMPLE

All-reduce

#include <nccl.h>
ncclComm_t comm[4];
ncclCommInitAll(comm, 4, {0, 1, 2, 3});

foreach g in (GPUs) { // or foreach thread
    cudaSetDevice(g);
    double *d_send, *d_recv;
    // allocate d_send, d_recv; fill d_send with data
    ncclAllReduce(d_send, d_recv, N, ncclDouble, ncclSum, comm[g], stream[g]);
    // consume d_recv
}
NVIDIA DEEP LEARNING SDK
High Performance GPU-Acceleration for Deep Learning

DEEP LEARNING FRAMEWORKS

DEEP LEARNING
- cuDNN

MATH LIBRARIES
- cuBLAS
- cuSPARSE
- cuFFT

MULTI-GPU
- NCCL

COMPUTER VISION
- Image Classification
- Object Detection

SPEECH AND AUDIO
- Voice Recognition
- Language Translation

NATURAL LANGUAGE PROCESSING
- Recommendation Engines
- Sentiment Analysis
PLATFORM
AN END-TO-END SOLUTION

Data Scientist

Embedded platform

Solver
Network
Dashboard

Train

Model

Deploy

Classification
Detection
Segmentation
# CUDA FOR DEEP LEARNING DEVELOPMENT

<table>
<thead>
<tr>
<th>DEEP LEARNING SDK</th>
<th>TITAN X</th>
<th>DEVBOX</th>
<th>GPU CLOUD</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIGITS</td>
<td><img src="image" alt="TITAN X" /></td>
<td><img src="image" alt="DEVBOX" /></td>
<td><img src="image" alt="GPU CLOUD" /></td>
</tr>
<tr>
<td>cuDNN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cuSPARSE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cuBLAS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCCL</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TESLA FOR HYPERSONE SCALE DATACENTERS

**HYPERSCALE SUITE**

- **Deep Learning SDK**
- **GPU REST Engine**
- **GPU Accelerated FFmpeg**
- **Image Compute Engine**
- **GPU support in Mesos**

**TESLA M40**

POWERFUL: Fastest Deep Learning Performance

**TESLA M40**

LOW POWER: Highest Hyperscale Throughput
JETSON FOR INTELLIGENT MACHINES

JETSON SDK

JETSON TX1

<table>
<thead>
<tr>
<th>Feature</th>
<th>JETSON TX1 Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>1 TFLOP/s 256-core Maxwell</td>
</tr>
<tr>
<td>CPU</td>
<td>64-bit ARM A57 CPUs</td>
</tr>
<tr>
<td>Memory</td>
<td>4 GB LPDDR4</td>
</tr>
<tr>
<td>Storage</td>
<td>16 GB eMMC</td>
</tr>
<tr>
<td>Size</td>
<td>50mm x 87mm</td>
</tr>
<tr>
<td>Power</td>
<td>Under 10W</td>
</tr>
</tbody>
</table>

Additional Information:
- JETSON SDK includes various tools and filters for image processing.
- JETSON TX1 features a powerful GPU and efficient CPU for high-performance computing.
- 270M items sold/day and 10M users highlight the widespread adoption of JETSON technologies.
- 43% of sales are on mobile devices, indicating a strong market presence in the mobile space.
- The TX1 model includes a 1 TFLOP/s 256-core Maxwell GPU, 64-bit ARM A57 CPUs, 4 GB LPDDR4 with 25.6 GB/s of memory speed, 16 GB eMMC storage, and a size of 50mm x 87mm. The power consumption is under 10W.
NVIDIA DEEP LEARNING PLATFORM

DL FRAMEWORK (CAFFE, CNTK, TENSORFLOW, THEANO, TORCH...)

DEEP LEARNING SDK

TITAN X - DEVELOPERS

TESLA - DEPLOYMENT

AUTOMOTIVE - DRIVEPX

EMBEDDED - JETSON
USING THE GPU FOR DEEP LEARNING
Quickly design the best deep neural network (DNN) for your data

Train on multi-GPU (automatic)

Visually monitor DNN training quality in real-time

Manage training of many DNNs in parallel on multi-GPU systems
Four TITAN X GPUs with 12GB of memory per GPU

1600W Power Supply Unit
Ubuntu 14.04
NVIDIA-qualified driver
NVIDIA® CUDA® Toolkit 7.0
NVIDIA® DIGITS™ SW
Caffe, Theano, Torch, BIDMach
FOR THE DATACENTER: MULTI-GPU SERVERS

For accelerated training

Facebook
Big Sur Opencompute server
TESLA M40
World’s Fastest Accelerator for Deep Learning

8x Faster
Caffe Performance

Reduce Training Time from 8 Days to 1 Day

# of Days

Caffe Benchmark: AlexNet training throughput based on 20 iterations,
CPU: E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUDA Cores</td>
<td>3072</td>
</tr>
<tr>
<td>Peak SP</td>
<td>7 TFLOPS</td>
</tr>
<tr>
<td>GDDR5 Memory</td>
<td>12 GB</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>288 GB/s</td>
</tr>
<tr>
<td>Power</td>
<td>250W</td>
</tr>
</tbody>
</table>

Tesla M40
8x Faster Caffe Performance
THE DATA SCIENTIST JOURNEY MADE EASIER WITH GPUS
AN END-TO-END SOLUTION

Data Scientist

Embedded platform

Train

Dashboard

Solver

Network

Model

Deploy

Classification

Detection

Segmentation
SETTING UP DEEP LEARNING APPLICATIONS

Steps

• Data labelling / Curation
• Hyperparameter search
• Training
• Deployment
TRAINING EXAMPLE DIGITS ON THE AMAZON CLOUD
CLASSIFICATION EXAMPLE

Come to the lab:
Tuesday 3:00pm
L6139
A Tutorial on More Ways to Use DIGITS
EMBEDDED DEPLOYMENT
JETSON TX1 DEVKIT
JETSON LINUX SDK

Graphics

Deep Learning and Computer Vision

GPU Compute

Developer Tools

- OpenGL
- cuDNN
- Vulkan
- NVIDIA VISIONWORKS
- CUDA
- NVIDIA Nsight
- NVTX
- Debugger | Profiler | System Trace
EMBEDDED DEPLOYMENT
On Jetson TX1

Inference at 258 img/s
No need to change code
Simply compile Caffe and copy a trained .caffemodel to TX1
## GOOGLENET PERFORMANCE

### FP32 and FP16 inference

<table>
<thead>
<tr>
<th>Network: GoogLeNet</th>
<th>Batch Size</th>
<th>Tegra X1 (FP32)</th>
<th>Tegra X1 (FP16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inference Performance</td>
<td>1</td>
<td>33 img/sec</td>
<td>33 img/sec</td>
</tr>
<tr>
<td>Power</td>
<td></td>
<td>5.0 W</td>
<td>4.0 W</td>
</tr>
<tr>
<td>Performance/Watt</td>
<td></td>
<td>6.5 img/sec/W</td>
<td>8.3 img/sec/W</td>
</tr>
<tr>
<td>Inference Performance</td>
<td>128 (Titan X)</td>
<td>52 img/sec</td>
<td>75 img/sec</td>
</tr>
<tr>
<td>Power</td>
<td>64 (Tegra X1)</td>
<td>5.9 W</td>
<td>5.8 W</td>
</tr>
<tr>
<td>Performance/Watt</td>
<td></td>
<td>8.8 img/sec/W</td>
<td>12.8 img/sec/W</td>
</tr>
</tbody>
</table>

Server version (Tesla M40)
batch 1: ~130 img/sec - 120W / batch 128: ~850 img/sec - 225 W
CLASSIFICATION EXAMPLE

```
ubuntu@tegra-ubuntu:~$ 
```
CLASSIFICATION EXAMPLE

Come to the lab:

Tuesday 1:00pm

L6131 From Workstation to Embedded: Accelerated Deep Learning on NVIDIA Jetson TX1
OTHER USE CASES
IMAGE CAPTIONING

“Automated Image Captioning with ConvNets and Recurrent Nets”

—Andrej Karpathy, Fei-Fei Li
DRONES
DEEP LEARNING AT GTC
Deep Learning at NVIDIA, Monday 4/4

11:00am: Accelerate Deep Learning with NVIDIA's Deep Learning Platform

12:00pm: Hangout -- The DIGITS Roadmap

1:00pm: From Workstation to Embedded: Accelerated Deep Learning on NVIDIA Jetson TX1

3:00pm: A Tutorial on More Ways to Use DIGITS

4:00pm: Hangout - cuDNN -- Features, Roadmap and Q&A
DEEP LEARNING AT GTC
Frameworks Hands-on Labs

Wednesday 4/6
1:00pm: Introduction to CNTK
2:00pm: Machine Learning Using TensorFlow
2:00pm: BIDMach Machine Learning Toolkit
3:30pm: Applied Deep Learning for Vision and Natural Language with Torch7
3:30pm: Caffe Hands-on Lab

Thursday 4/7
9:30am: Chainer Hands-on: Introduction To Train Deep Learning Model in Python
9:30am: Deep Learning With the Theano Python Library
1:00pm: IBM Watson Developers Lab
THANK YOU

JOIN THE CONVERSATION
#GTC16  

JOIN THE NVIDIA DEVELOPER PROGRAM AT developer.nvidia.com/join