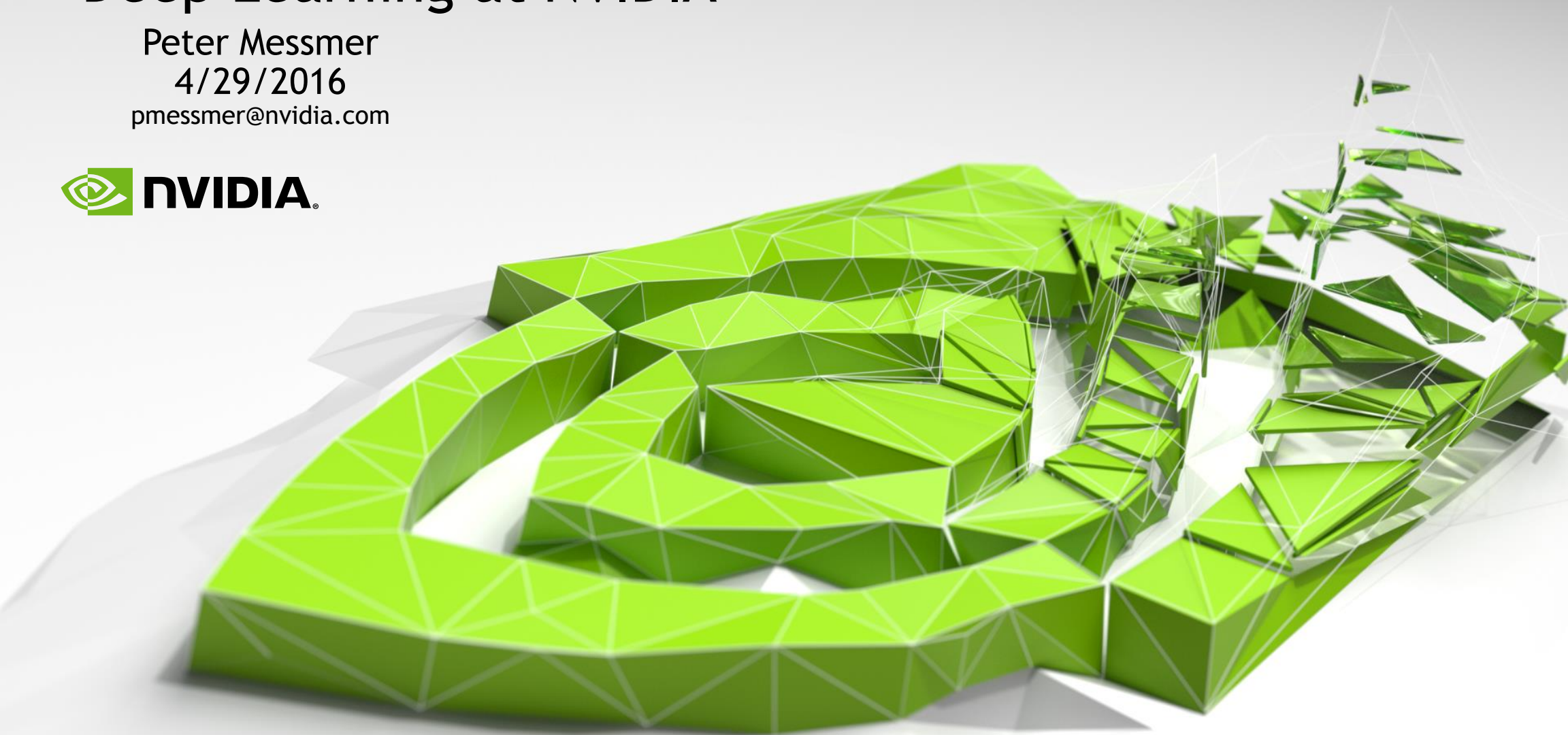


Deep Learning at NVIDIA

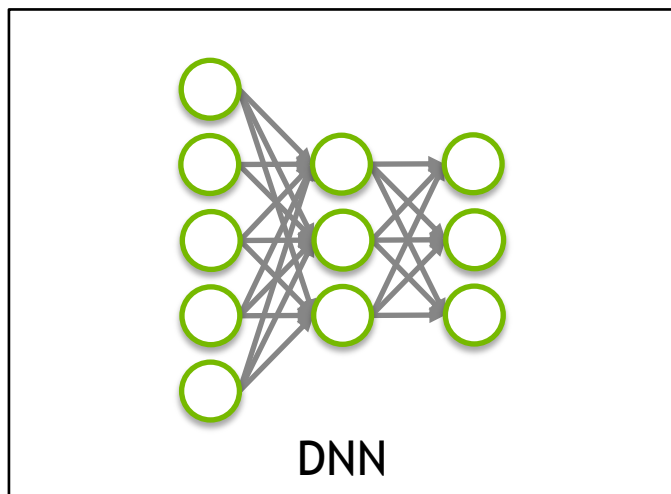
Peter Messmer

4/29/2016

pmessmer@nvidia.com



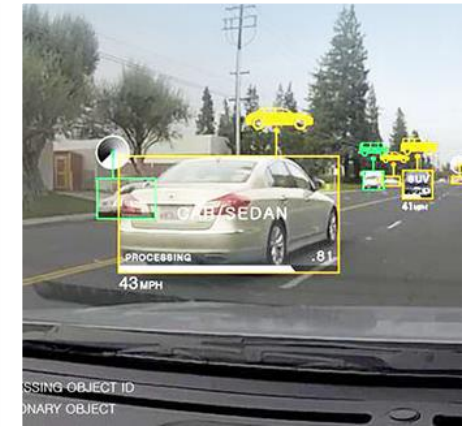
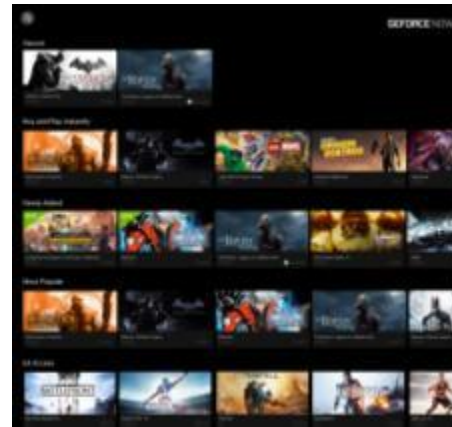
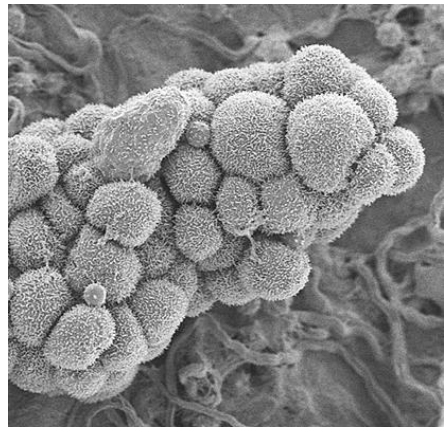
The Big bang in machine learning



“Google’s AI engine also reflects how the world of computer hardware is changing. (It) depends on machines equipped with GPUs... And it depends on these chips more than the larger tech universe realizes.”

WIRED

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

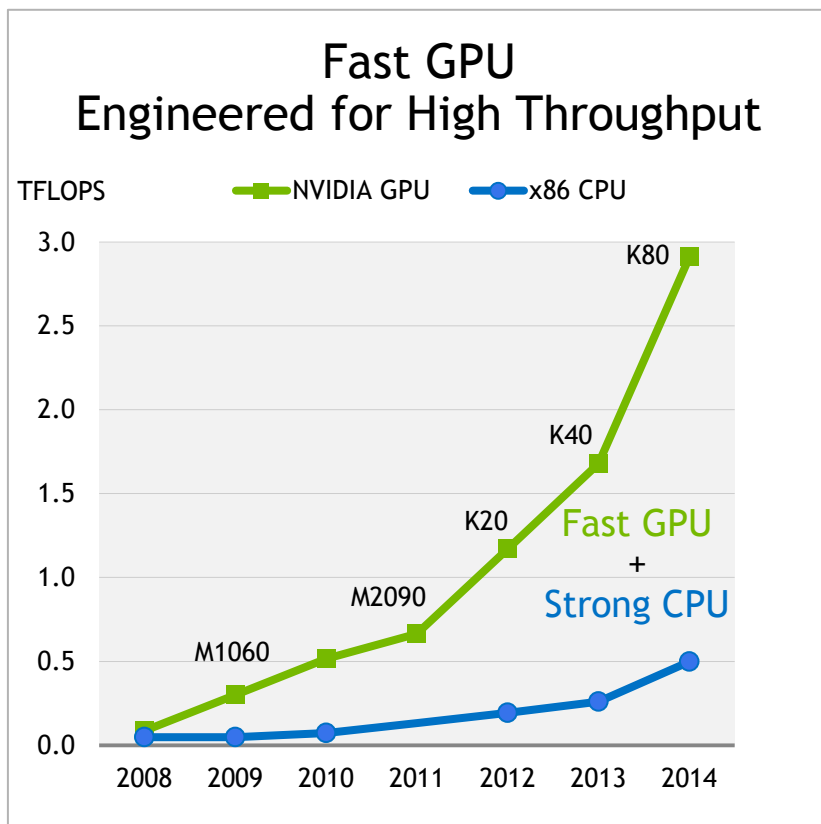
AUTONOMOUS MACHINES

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

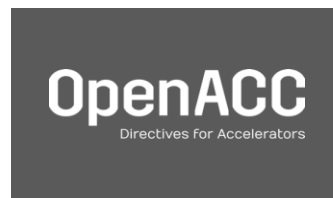


Tesla Accelerated computing platform

Focused on Co-Design from Top to Bottom



Productive Programming Model & Tools



Expert Co-Design



Accessibility



INTRODUCING TESLA P100

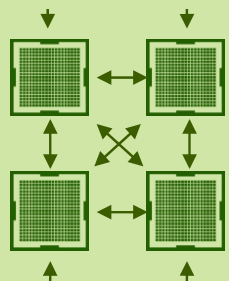
New GPU Architecture to Enable the World's Fastest Compute Node

Pascal Architecture



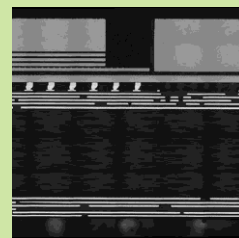
Highest Compute Performance

NVLink



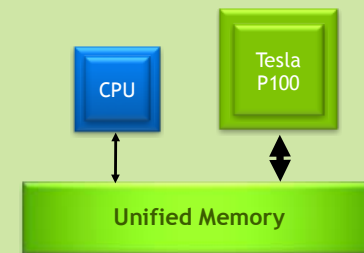
GPU Interconnect for Maximum Scalability

CoWoS HBM2

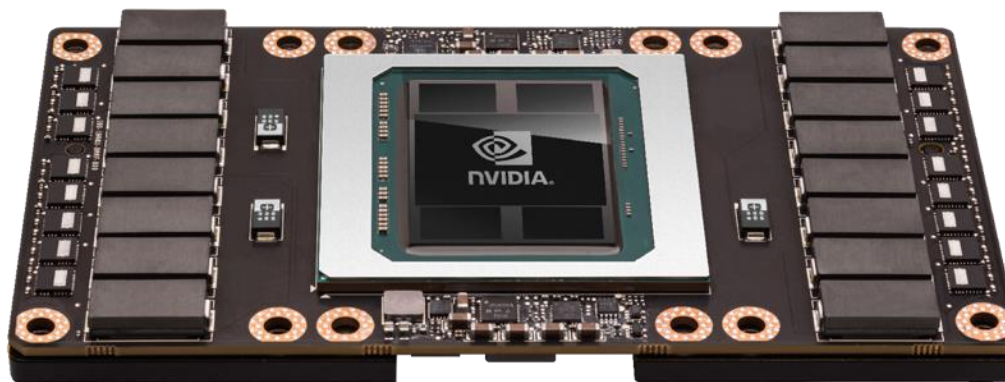


Unifying Compute & Memory in Single Package

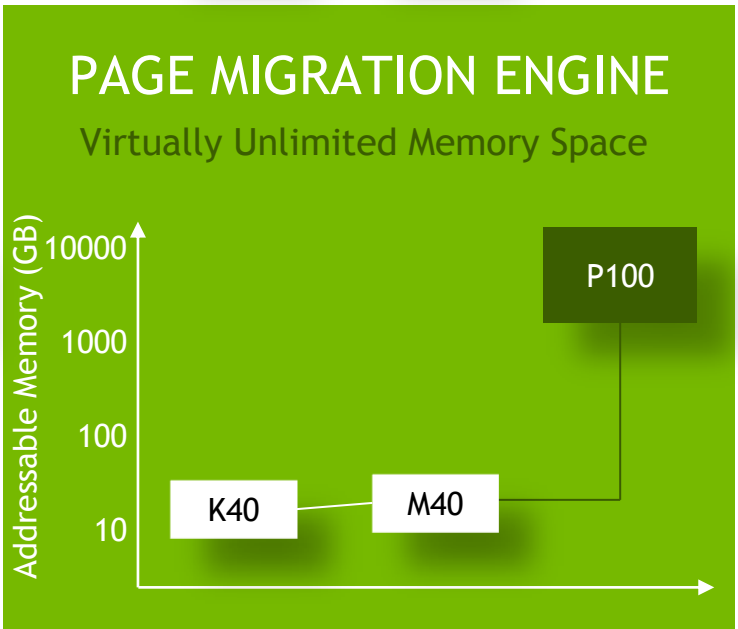
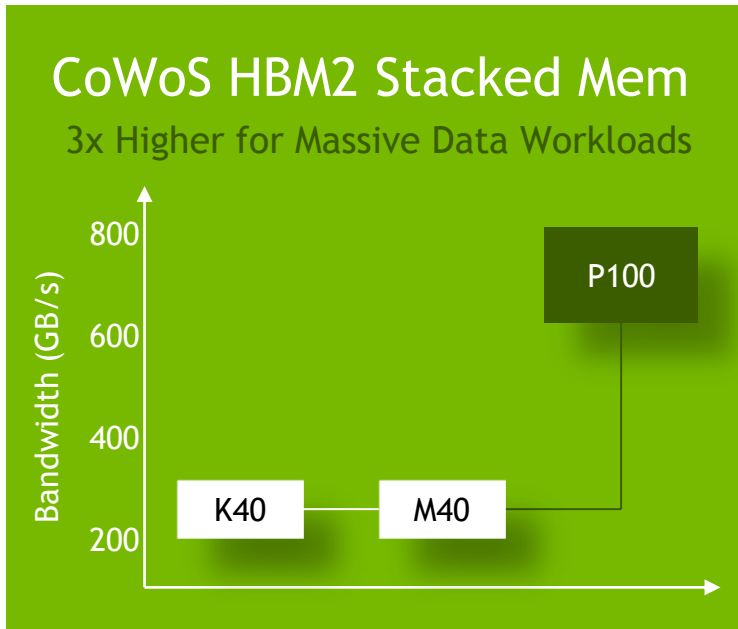
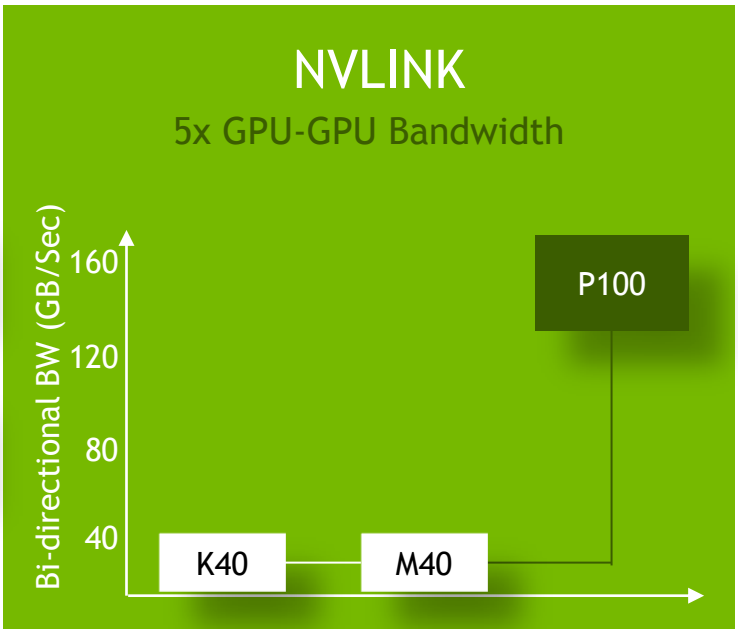
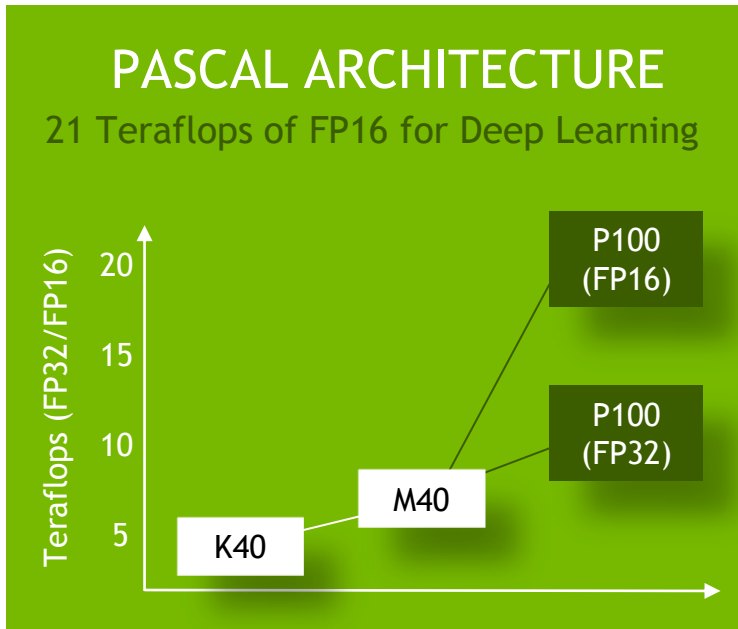
Page Migration Engine



Simple Parallel Programming with Virtually Unlimited Memory



Giant leaps in everything



nvidia DGX-1

world's first deep learning supercomputer



170 TFLOPS FP16

8x Tesla P100 16GB

NVLink Hybrid Cube Mesh

Accelerates Major AI Frameworks

Dual Xeon

7 TB SSD Deep Learning Cache

Dual 10GbE, Quad IB 100Gb

3RU - 3200W

NVIDIA SDK

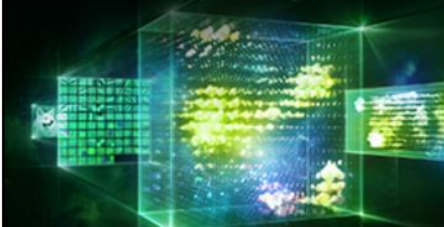
The Essential Resource for GPU Developers

NVIDIA SDK

DEEP LEARNING

Deep Learning SDK

High-performance tools and libraries for deep learning



SELF-DRIVING CARS

NVIDIA DriveWorks™

Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous



VIRTUAL REALITY

NVIDIA VRWorks™

A comprehensive SDK for VR headsets, games and professional applications



GAME DEVELOPMENT

NVIDIA GameWorks™

Advanced simulation and rendering technology for game development



ACCELERATED COMPUTING

NVIDIA ComputeWorks™

Everything scientists and engineers need to build GPU-accelerated applications



DESIGN & VISUALIZATION

NVIDIA DesignWorks™

Tools and technologies to create professional graphics and advanced rendering applications



AUTONOMOUS MACHINES

NVIDIA JetPack™

Powering breakthroughs in autonomous machines, robotics and embedded computing



ADDITIONAL RESOURCES

More resources for GPU Developers



NVIDIA Deep Learning SDK

High Performance GPU-Acceleration for Deep Learning

APPLICATIONS

IMAGENET





Image Classification Object Detection

COMPUTER VISION



Voice Recognition Translation

SPEECH AND AUDIO



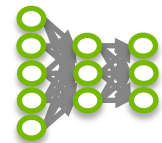
Recommendation Engines Sentiment Analysis

BEHAVIOR

FRAMEWORKS

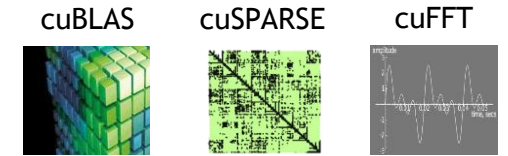


DEEP LEARNING SDK



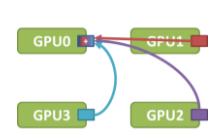
cuDNN

DEEP LEARNING



cuBLAS cuSPARSE cuFFT

MATH LIBRARIES



NCCL

MULTI-GPU

NVIDIA DIGITS

Interactive Deep Learning GPU Training System

Process Data

Job Information

Job Directory
/home/michaelo/digits
/jobs/20150311-171431-e0d8

Image Type
Color

Image Dimensions
256x256

Resize Mode
half_crop

Parse Folder (train/val)

Folder
http://sql/data/images/voc_cropped/

Number of categories
20

Training images
26759

Validation images
8917 (25.0%)

Create DB (train)

Input file
train.txt

DB Entries
26759

Configure DNN

Select Dataset

PASCAL VOC
ILSVRC 2012
MNIST Dataset

Solver Options

Training epochs
30

Validation interval (in epochs)
1

Batch size
100

Base Learning Rate
0.01

Show advanced learning rate options

Custom Network

```
{  
  "layer": {  
    "name": "conv1"  
    "type": "Convolution"  
    "bottom": "data"  
    "top": "conv1"  
    "param": {  
      "lr_mult": 1  
      "decay_mult": 1  
    }  
  }  
}
```

Pretrained model

Model Name
ImageNet

Create

Monitor Progress

Solver
solver.prototxt

Network (train/val)
train_val.prototxt

Network (deploy)
deploy.prototxt

Dataset
voc_cropped@256x256
Done Wed Mar 11, 05:16:57 PM

Image Size
256x256

Image Type
COLOR

Create DB (train)
26759 images

Create DB (val)
8917 images

Loss (train) **Loss (val)** **Accuracy (val)**

Epoch	Loss (train)	Loss (val)	Accuracy (val)
0.0	3.5	3.5	0
1.0	2.0	2.0	40
2.0	1.8	1.8	50
3.0	1.7	1.7	55
4.0	1.6	1.6	58
5.0	1.5	1.5	60
6.0	1.4	1.4	62
7.0	1.3	1.3	63
8.0	1.2	1.2	64
9.0	1.1	1.1	65
10.0	1.0	1.0	66

Visualize Layers

Predictions

8
3
0
6
4

Layer **Activations**

conv1

pool1

Good time for confluence of HPC and DL

NVIDIA Tesla Platform Will Help Swiss Scientists Solve Complex Problems

SAN JOSE, CA - GPU Technology Conference -- NVIDIA (NASDAQ: NVDA) today announced that **Pascal™ architecture**-based **NVIDIA® Tesla® GPU accelerators** will power an upgraded version of Europe's fastest supercomputer, the Piz Daint system at the Swiss National Supercomputing Center (CSCS) in Lugano, Switzerland. The upgrade is expected to more than double Piz Daint's speed, with most of the system's performance expected to come from its Tesla GPUs.

Piz Daint, named after a mountain in the Swiss Alps, currently delivers 7.8 petaflops of compute performance, or 7.8 quadrillion mathematical calculations per second. That puts it at No. 7 in the latest TOP500 list of the world's fastest supercomputers. CSCS plans to upgrade the system later this year with 4,500 Pascal-based GPUs.

CSCS to upgrade Piz Daint to Pascal GPUs
Simple process for basic applications

Shifter: Containers in HPC environments

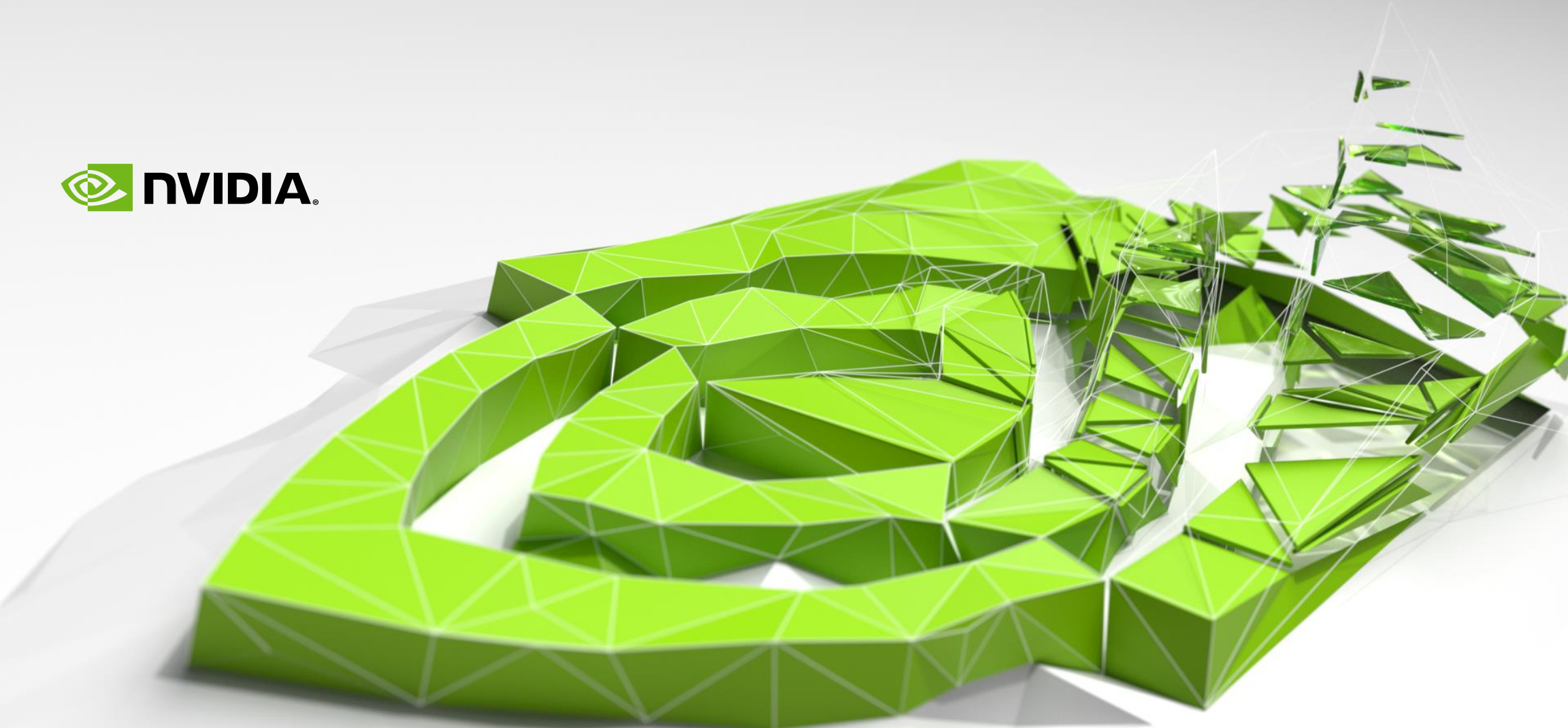
HPC Advisory Council Switzerland
Miguel Gila, CSCS
March 21, 2016

Practical use case: WLCG Swiss Tier-2



- CSCS operates the cluster Phoenix on behalf of CHiPP, the Swiss Institute of Particle Physics
- Phoenix runs Tier-2 jobs for ATLAS, CMS and LHCb, 3 experiments of the LHC at CERN and part of WLCG (Worldwide LHC Computing Grid)

CSCS working on containers for
WLCG data



nvGRAPH

Accelerated Graph Analytics

nvGRAPH for high performance graph analytics

Deliver results up to 3x faster than CPU-only

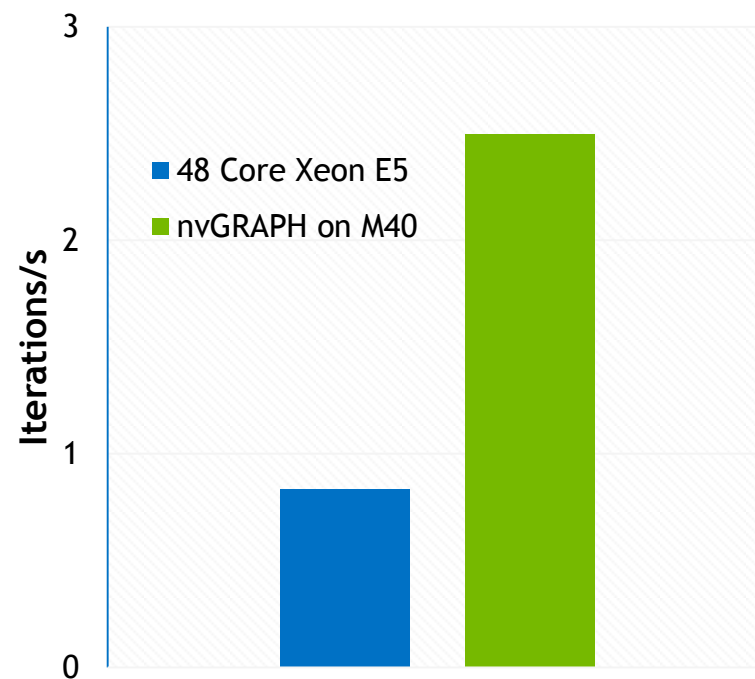
Solve graphs with up to 2.5 Billion edges on 1x M40

Accelerates a wide range of graph analytics apps:

PageRank	Single Source Shortest Path	Single Source Widest Path
Search	Robotic Path Planning	IP Routing
Recommendation Engines	Power Network Planning	Chip Design / EDA
Social Ad Placement	Logistics & Supply Chain Planning	Traffic sensitive routing

developer.nvidia.com/nvgraph

nvGRAPH: 3x Speedup



PageRank on Twitter 1.5B edge dataset

CPU System:
4U server w/ 4x12-core Xeon E5-2697 CPU,¹³
30M Cache, 2.70 GHz, 512 GB RAM



OPENACC

More Science, Less Programming

```
main()
{
  <serial code>
  #pragma acc kernels
  //automatically runs on
  GPU
  {
    <parallel code>
  }
}
```

SIMPLE

Minimum efforts
Small code modifications

POWERFUL

Up to 10x faster
application performance

PORTABLE

Optimize once,
run on GPUs and CPUs

FREE FOR ACADEMIA