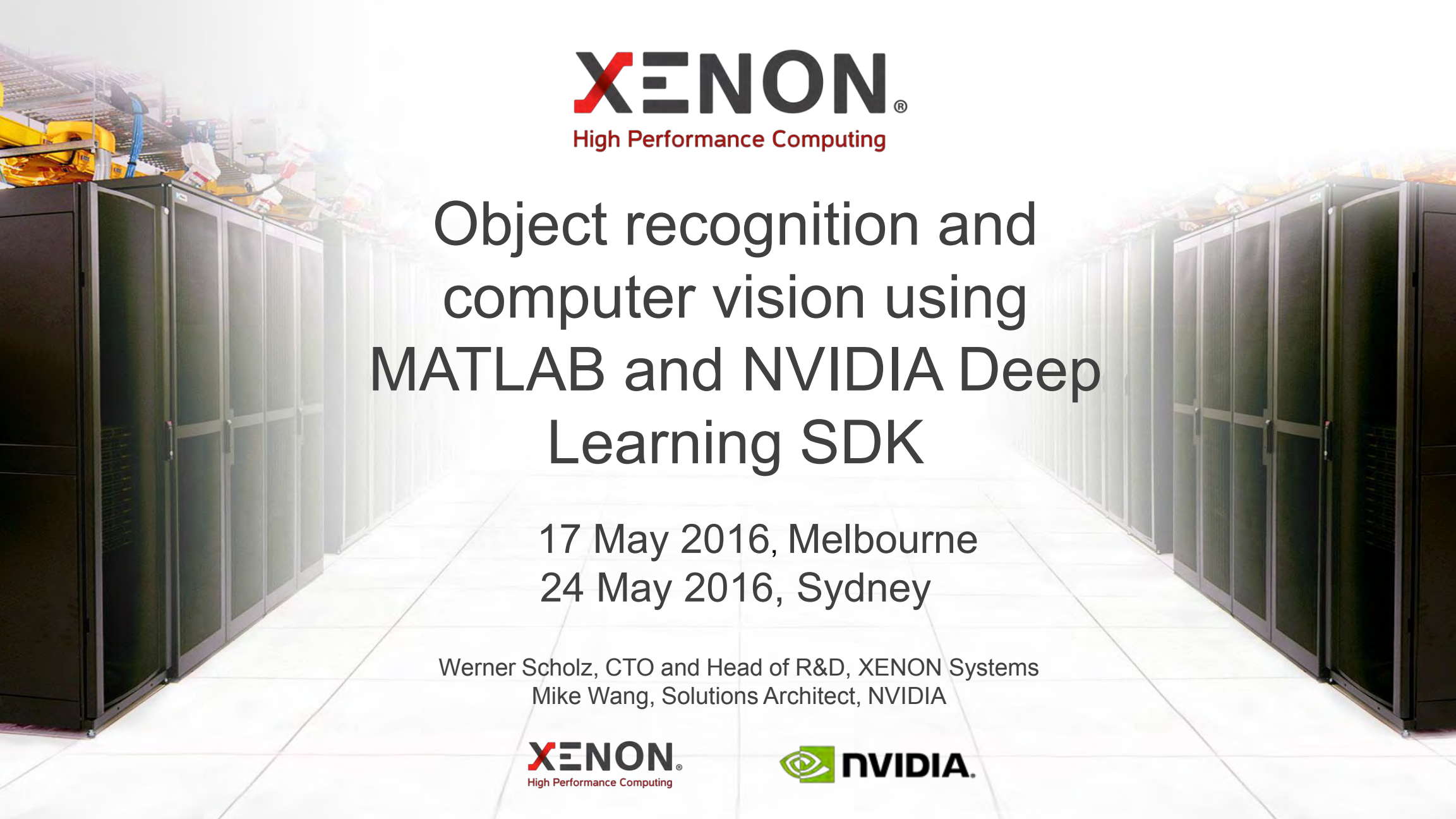


XENON®
High Performance Computing

Object recognition and computer vision using MATLAB and NVIDIA Deep Learning SDK

17 May 2016, Melbourne
24 May 2016, Sydney

Werner Scholz, CTO and Head of R&D, XENON Systems
Mike Wang, Solutions Architect, NVIDIA

XENON®
High Performance Computing

NVIDIA.

Outline

- Background of XENON System
- Object recognition - MATLAB interprets image contents
- How does it do it? Looking under the hood
- What is required? Software stack and installation
- NVIDIA Deep Learning SDK
- What does it run on?
- What happens next?

XENON Overview

Industries we work in:

XENON Solutions

- **Visual Workstations**
  XENON's Nitro visual workstations:
  Performance and Reliability for the most demanding graphics, engineering, digital arts worksloads and optimised for MATLAB.

- **GPU Computing**
  High performance **acceleration solutions for MATLAB** leveraging NVIDIA Tesla technology and the CUDA ecosystem

- **Virtualisation**
  End-to-end virtualisation solutions for compute, storage, networking, and desktop.
  Server and **VDI solutions** from high density servers to GPU enabled workstations and thin/zero client solutions.

XENON
High Performance Computing

XENON Milestones

Delivering world-class high performance computing solutions

1998
Telstra Digital Video Network
Supplied Video Server equipment

2000
Animal Logic
Render Farm delivered for the "Matrix" movie trilogy

2004
TataSky DTH India
Supplied Video Logging solution

2005
Thales Defence
Supplied Image Generators for the ASLAV simulators

2007
Victorian Partnership for Advance Computing
Supplied HPC Cluster for Advance Computing

2009
CSIRO
(10th in Green500 ranking)
Supplied Australia's first GPU Cluster ("Bragg")

2011
Australian Stock Exchange (ASX)
Supplied Infiniband equipment

2011
Fox Broadcasting USA
Supplied Video Logging solution

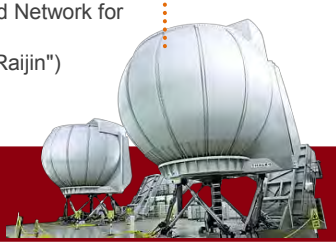
2012
Fujitsu
Supplied Infiniband Network for Australia's largest Supercomputer ("Raijin")

2013
Digital Cinema Systems
Upgraded 150 Independent Cinemas in ANZ region

2014
University of Queensland
FlashLite HPC cluster system for high data throughput

2015
Thales Defence
Delivered High Fidelity Solution for Australian Army Tiger Helicopter Simulator

2015
Futuris
HPC Linux Cluster in a Windows environment for crash test simulations



1990 2000 2010

XENON High Performance Computing

HPC GPU Cluster Bragg

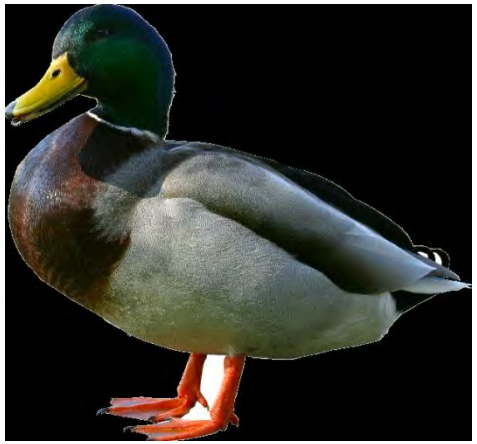
Large scale GPU deployment

Designed, delivered, and installed by XENON Systems

- 384x Tesla K20 GPUs
- 384 GPUs = 958,464 Thread Processors
- 2048x 2GHz Intel Xeon E5-2650 Cores
- 16.4TB DDR3-1600 System Memory
- 128TB SATA2 Local System Storage
- InfiniBand Interconnect FDR10 40Gb/s
- Linpack Result: 335Tflops (Double Precision)
- Peak power usage: 115 kW
- Currently #297 in Top500 and #24 in Green500 (Nov. 2015)

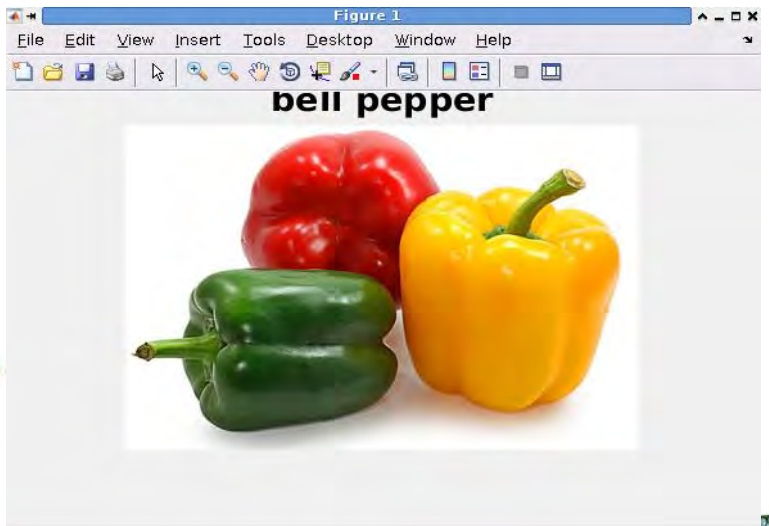
XENON
High Performance Computing

What's in an image?



An image says more than 1000 words…but what does it say?

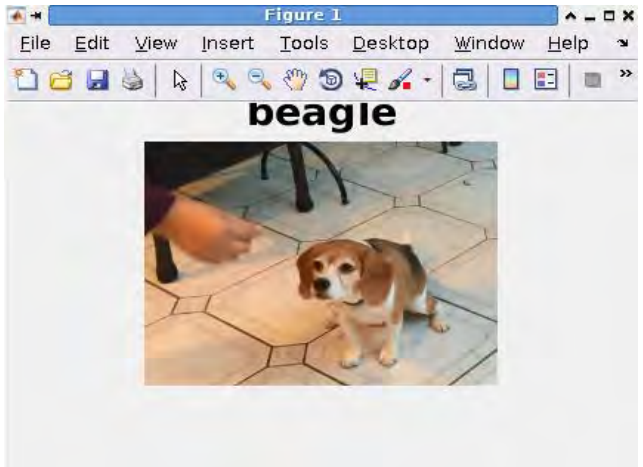
What's in an image?

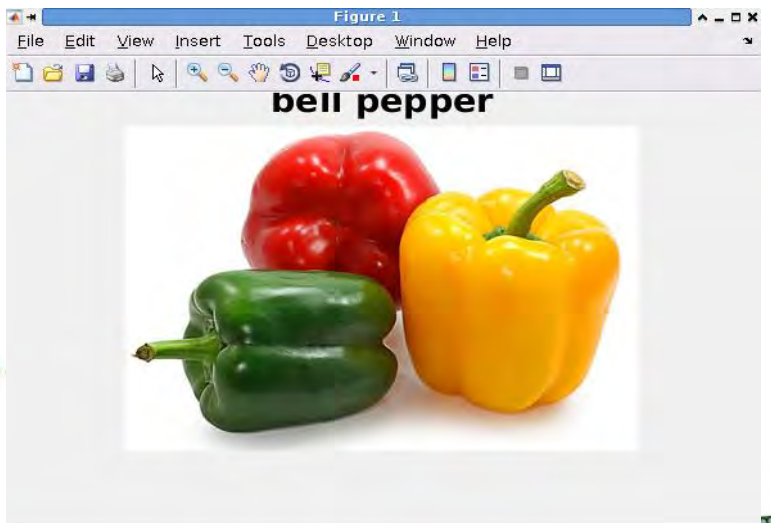


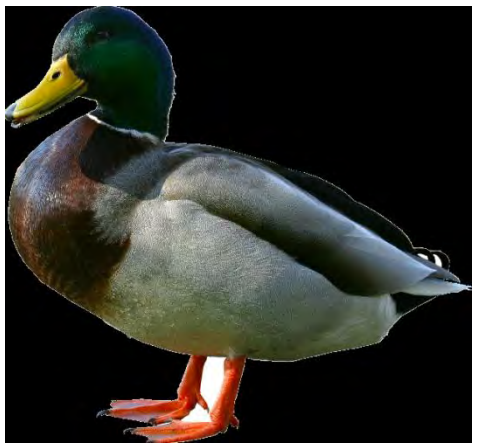
*)

*) Label determined based on inference by MATLAB using MatConvNet (see demo script below)

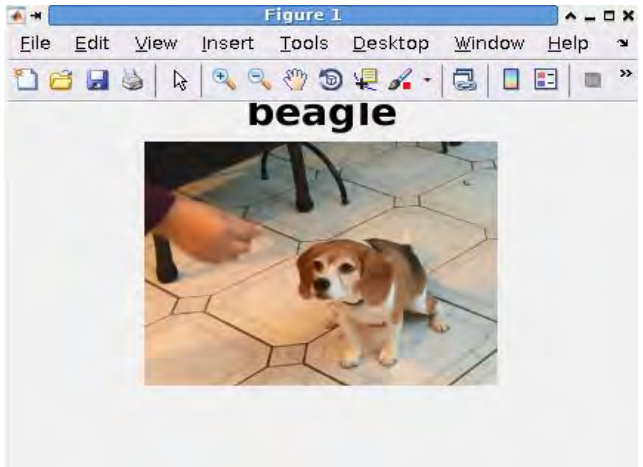
What's in an image?

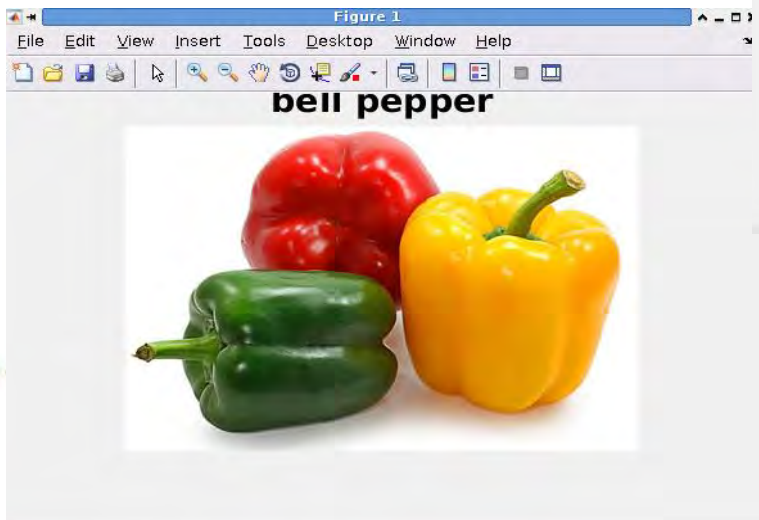


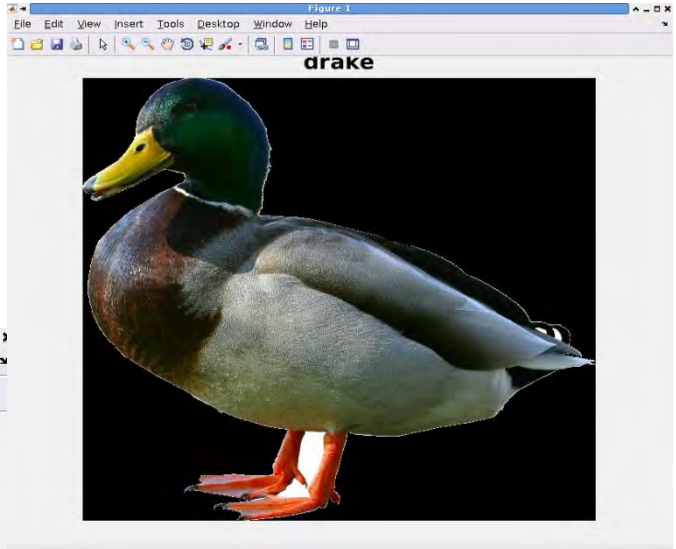




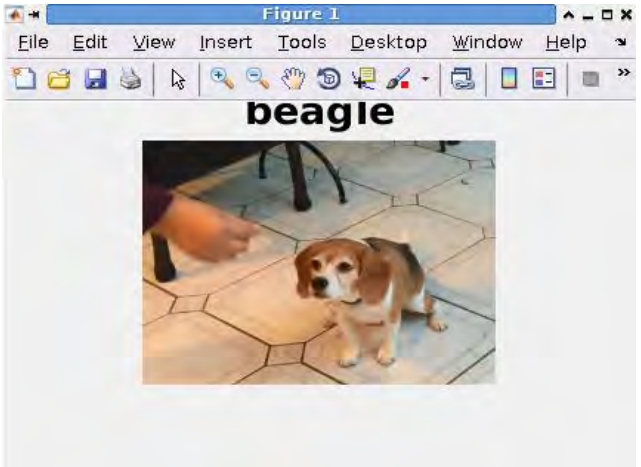
What's in an image?

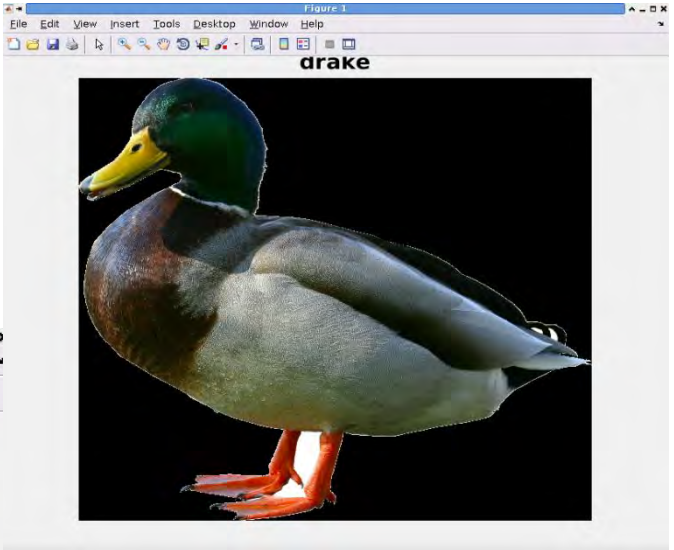


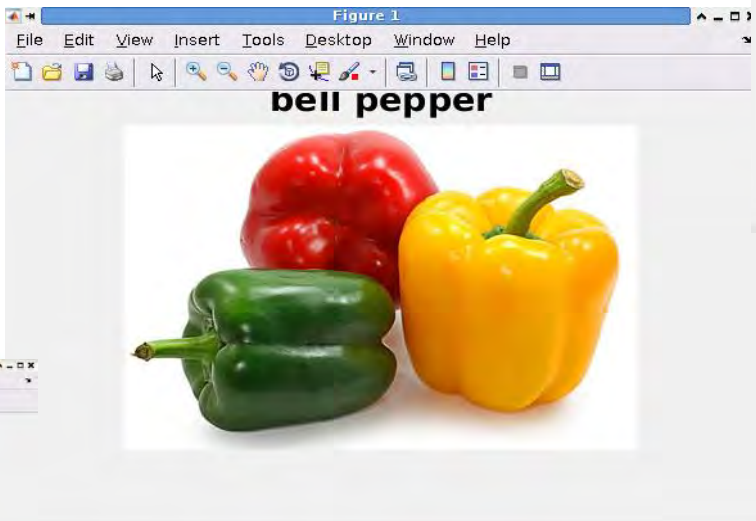


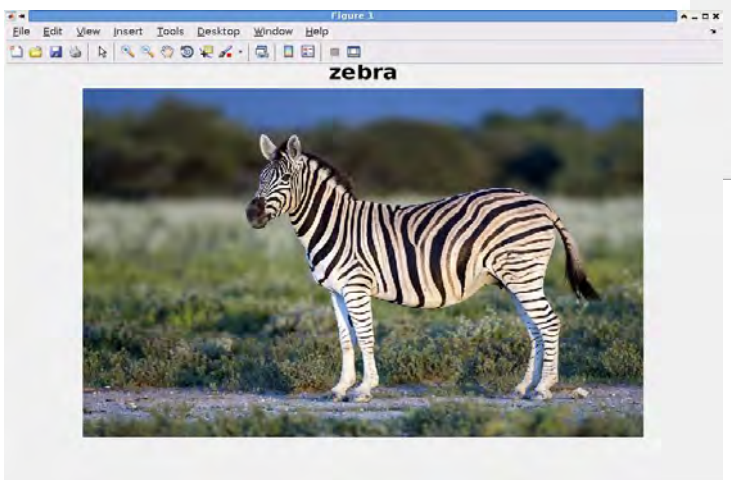


What's in an image?



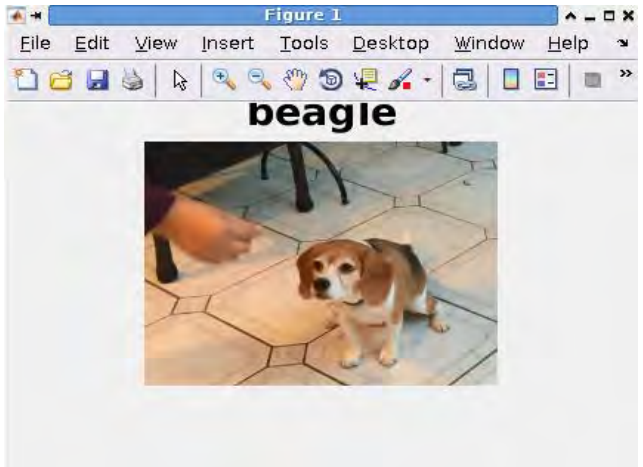


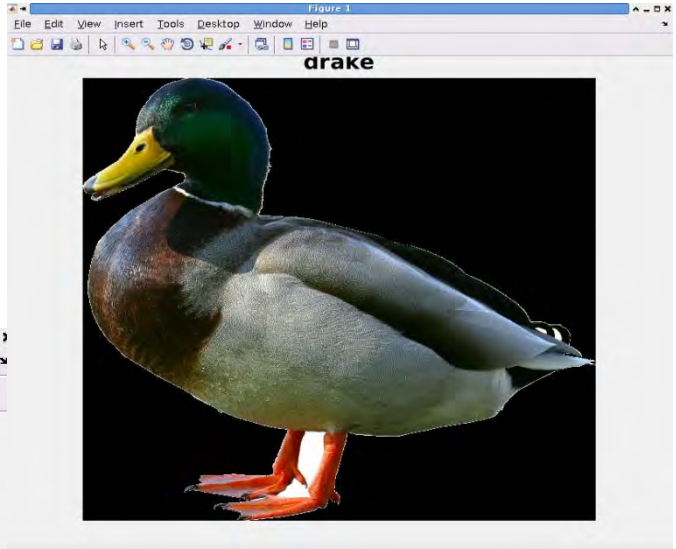


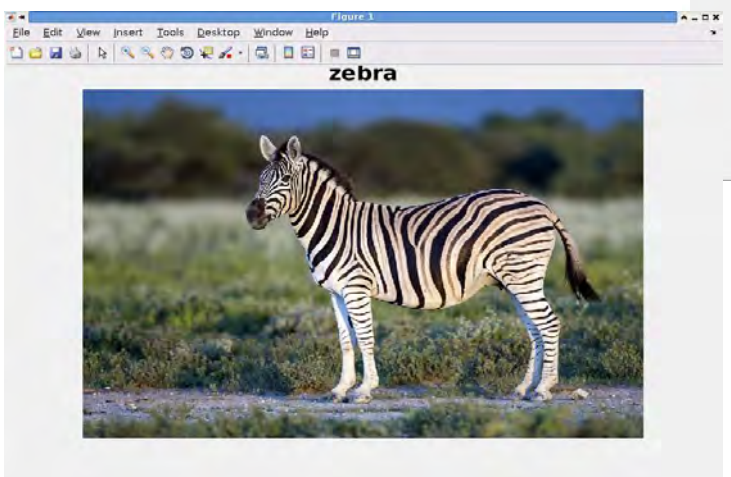


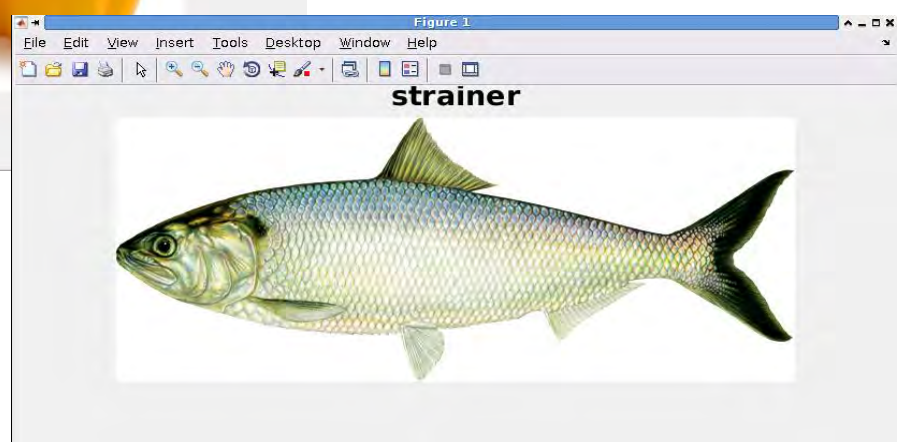
XENON
High Performance Computing

What's in an image?









XENON
High Performance Computing

How does MATLAB do it? Looking under the hood

- Design a Deep Neural Network
- Train the network
- Present new images to the network
- Be prepared to be surprised…

Every network is only as good as its training.

What is required?

- System with NVIDIA GPU
- OS (Ubuntu 14.04 is a commonly used platform)
- NVIDIA drivers
- NVIDIA cuDNN library
- MatConvNet library: MATLAB toolbox implementing Convolutional Neural Networks (CNNs) for computer vision applications
- MATLAB
- MATLAB Parallel Computing Toolbox™, Computer Vision System Toolbox™ and Statistics and Machine Learning Toolbox™
- A little bit of MATLAB code…

Object Recognition in 7 lines of MATLAB Code

```
% Download pretrained network from MatConvNet repository
urlwrite('http://www.vlfeat.org/matconvnet/models/imagenet-vgg-f.mat', 'imagenet-vgg-f.mat');

% Load the network
cnnModel.net = load('imagenet-vgg-f.mat');

% Set up MatConvNet
run(fullfile('/opt/matconvnet-1.0-beta20','matlab','vl_setupnn.m'));

% choose a test image and display it
im='pet_images/bell-peppers.jpg';
imshow(im);

% Predict its content using ImageNet trained vgg-f CNN model
label = cnnPredict(cnnModel,img);
title(label,'FontSize',20)
```

Ref: https://devblogs.nvidia.com/parallelforall/deep-learning-for-computer-vision-with-matlab-and-cudnn/

NVIDIA Deep Learning SDK

High Performance GPU-Acceleration for Deep Learning

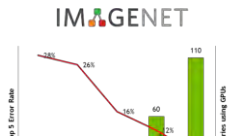


Image Classification

Object Detection

COMPUTER VISION

Voice Recognition

Language Translation

SPEECH AND AUDIO

Recommendation Engines

Sentiment Analysis

NATURAL LANGUAGE PROCESSING

Caffe

Chainer

DL4J

Deeplearning4j

Mocha.jl

julia

K

KERAS

Microsoft

CNTK

MatConvNet

mxnet

MINERVA

Purine

OpenDeep

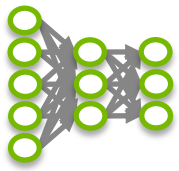
Pylearn2

TensorFlow

torch

theano

DEEP LEARNING FRAMEWORKS

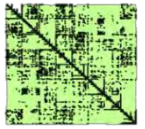


cuDNN

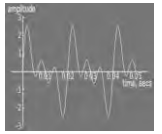
DEEP LEARNING

cuBLAS

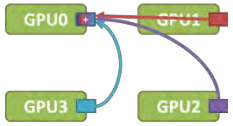
cuSPARSE



cuFFT



MATH LIBRARIES



NCCL

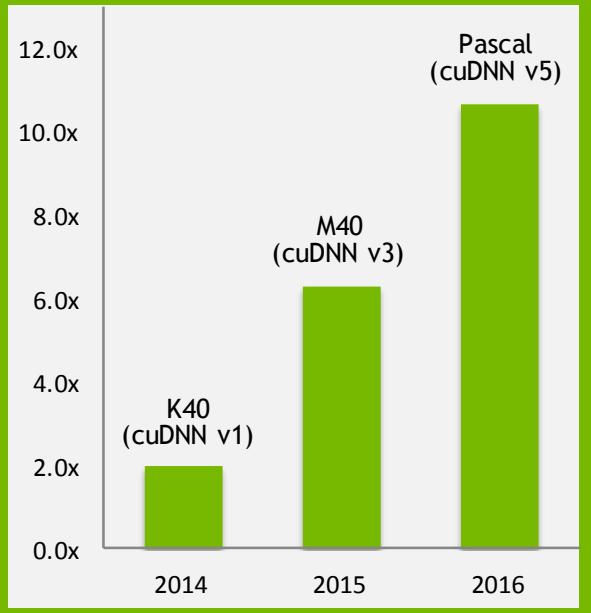
MULTI-GPU

NVIDIA cuDNN

Building blocks for accelerating deep neural networks on GPUs

- High performance deep neural network training
- Accelerates Deep Learning: Caffe, CNTK, Tensorflow, Theano, Torch
- Performance continues to improve over time

developer.nvidia.com/*cudnn*



AlexNet training throughput based on 20 iterations, CPU: 1x E5-2680v3 12 Core 2.5GHz.

"NVIDIA has improved the speed of cuDNN with each release while extending the interface to more operations and devices at the same time."

— Evan Shelhamer, Lead Caffe Developer, UC Berkeley

What's new in cuDNN 5?

Pascal GPU, RNNs, Improved Performance

LSTM recurrent neural networks deliver up to 6x speedup in Torch

Improved performance:

- Deep Neural Networks with 3x3 convolutions, like VGG, GoogleNet and ResNets
- 3D Convolutions
- FP16 routines on Pascal GPUs

Speedup for char-rnn RNN Layers

Speedup for DeepSpeech 2 RNN Layers

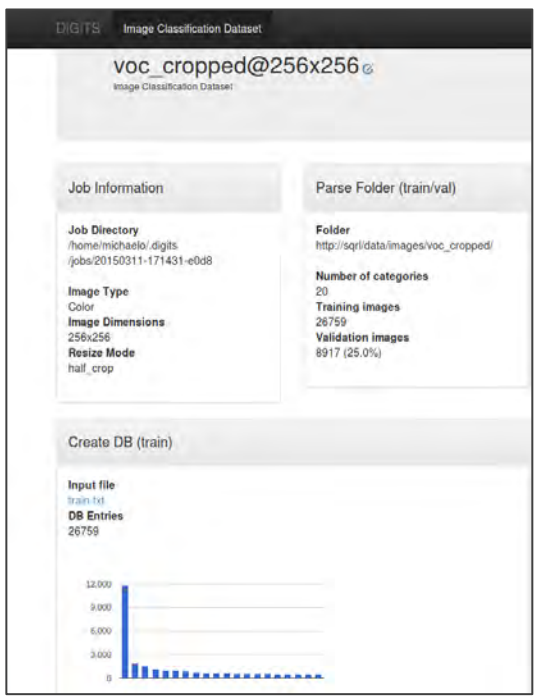
Performance relative to torch-rnn (https://github.com/jcjohnson/torch-rnn)
DeepSpeech2: http://arxiv.org/abs/1512.02595
Char-rnn: https://github.com/karpathy/char-rnn

NVIDIA DIGITS

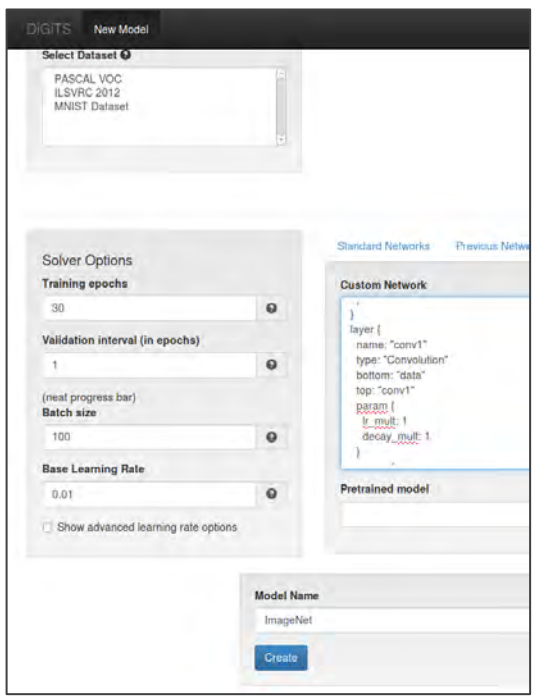
NVIDIA DIGITS

Interactive Deep Learning GPU Training System

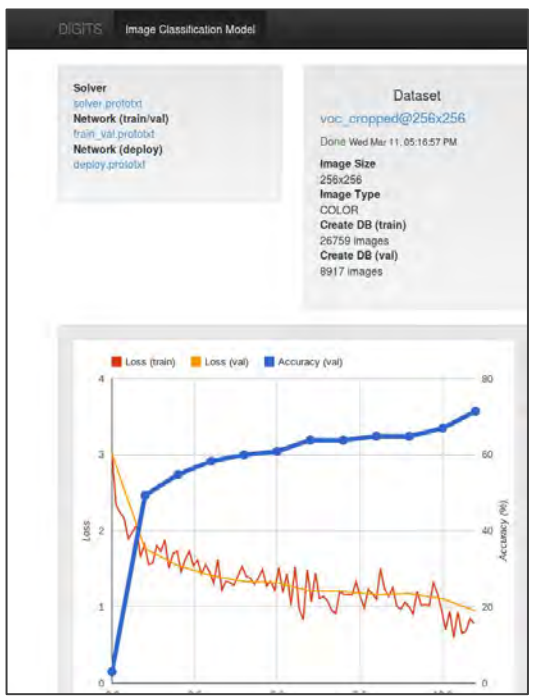
Process Data



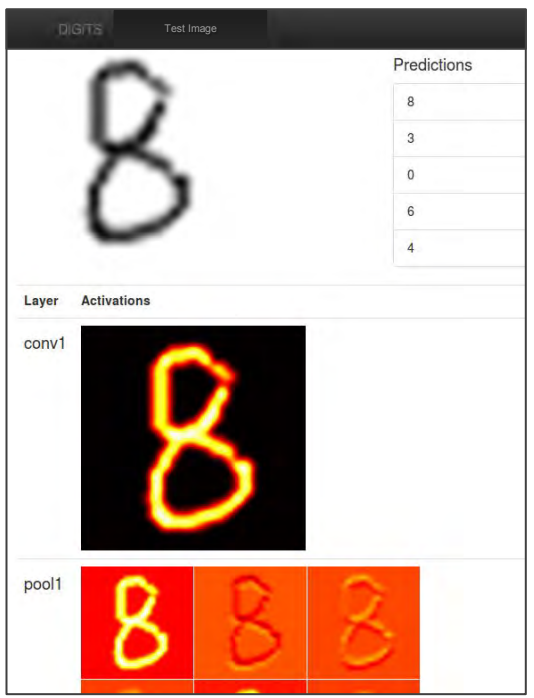
Configure DNN



Monitor Progress



Visualize Layers

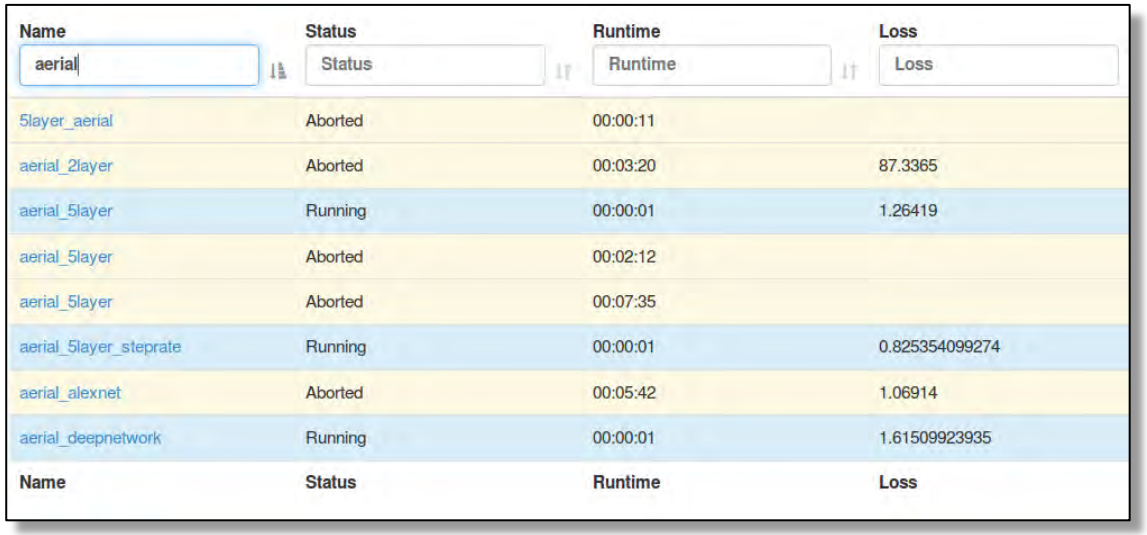


developer.nvidia.com/digits

NVIDIA DIGITS

Improves Deep Learning Training Productivity

- Train neural network models with Torch support (preview)
- Save time by quickly iterating to identify the best model
- Manage multiple jobs easily to optimize use of system resources
- Active open source project with valuable community contributions



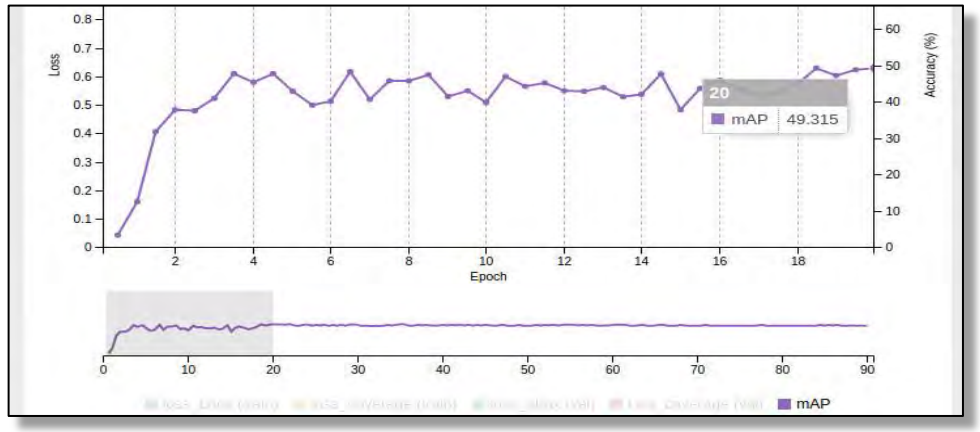
Name	Status	Runtime	Loss
aerial	Status	Runtime	Loss
5layer_aerial	Aborted	00:00:11	
aerial_2layer	Aborted	00:03:20	87.3365
aerial_5layer	Running	00:00:01	1.26419
aerial_5layer	Aborted	00:02:12	
aerial_5layer	Aborted	00:07:35	
aerial_5layer_steprate	Running	00:00:01	0.825354099274
aerial_alexnet	Aborted	00:05:42	1.06914
aerial_deepnetwork	Running	00:00:01	1.61509923935
Name	Status	Runtime	Loss

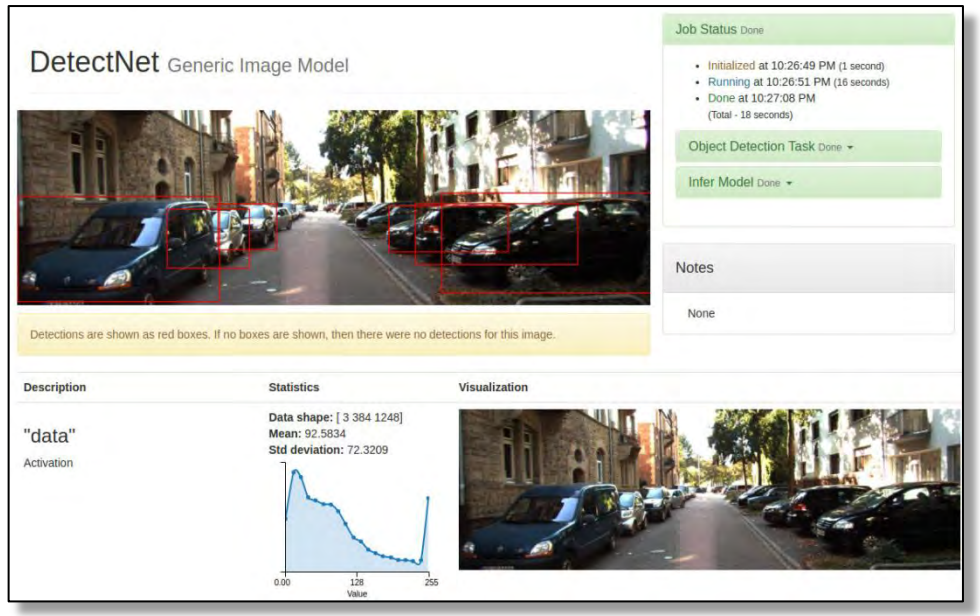
New Results Browser!

Preview DIGITS Future

Object Detection Workflow

- Object Detection Workflows for Automotive and Defense
- Targeted at Autonomous Vehicles, Remote Sensing





developer.nvidia.com/digits

What does it run on?

- XENON workstations with NVIDIA GPUs
- XENON DEVCUBE
  (see it in action at our stand!)
- XENON Radon rack servers
- 1U high density servers (up to 4 GPUs)
- 4U 8-GPU servers
- Custom configurations for your requirements
- and…

XENON
High Performance Computing

NVIDIA DGX-1 - The World's First Deep Learning Supercomputer in a Box

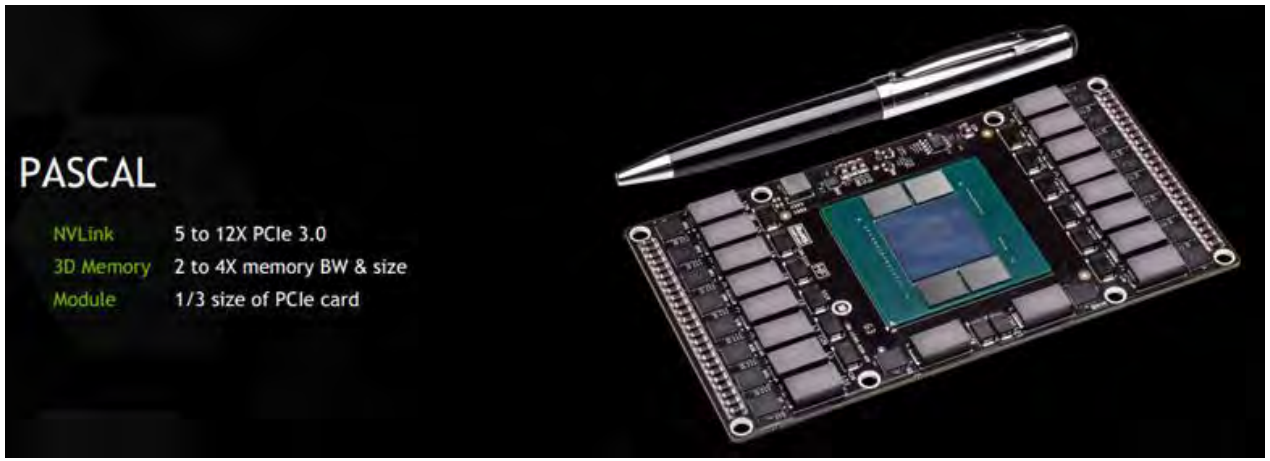
System Specifications
The NVIDIA DGX-1 system specifications include:
Up to 170 teraflops of half-precision (FP16) peak performance
Eight Tesla P100 GPU accelerators, 16GB memory per GPU
NVLink Hybrid Cube Mesh
7TB SSD DL Cache
Dual 10GbE, Quad InfiniBand 100Gb networking
3U - 3200W

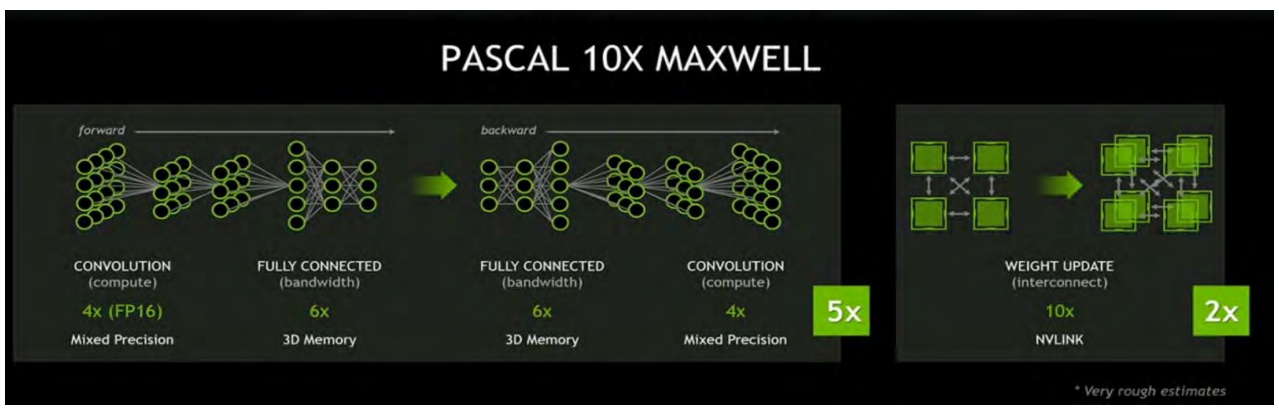
SYSTEM SPECIFICATIONS

- GPUs 8x Tesla GP100
- TFLOPS 42.5 Tflops double precision (FP64), 85 Tflops single precison (FP32), 170 Tflops half precision (FP16)
- GPU Memory 16 GB per GPU
- CPU Dual 16-core Intel Xeon E5-2698 v3 2.3 GHz
- NVIDIA CUDA Cores 28672
- System Memory 512 GB 2133 MHz DDR4 LRDIMM
- Storage 4x 1.92 TB SSD RAID 0
- Network Dual 10 GbE, 4 IB EDR
- Software Ubuntu Server Linux OS, DGX-1 Recommended GPU Driver

XENON
High Performance Computing

Pascal: The next generation GPU architecture





GPU Architecture	NVIDIA Fermi	NVIDIA Kepler	NVIDIA Maxwell	NVIDIA Pascal
GPU Process	40nm	28nm	28nm	16nm
Maximum Transistors	3.54 Billion (GTX 690)	7.08 Billion (Titan Z)	8.00 Billion (Titan X)	7.2 B (1080) 15.3 B (P100)
Maximum Die Size	294mm2	561mm2	601mm2	610mm2
Stream Processors Per Compute Unit	32 SPs	192 SPs	128 SPs	64 SPs
Maximum CUDA Cores	512 CCs	2880 CCs	3072 CCs	3584 CCs
FP32 Compute	2.08 TFLOPs(Tesla)	5.04 TFLOPs (Tesla)	6.84 TFLOPs (Tesla)	~10.6TFLOPs (Tesla)
FP64 Compute	0.66 TFLOPs (Tesla)	1.68 TFLOPs (Tesla)	0.21 TFLOPs (Tesla)	5.3 TFLOPs(Tesla)
Maximum VRAM	1.5 GB GDDR5	12 GB GDDR5	24 GB GDDR5	16 GB HBM2
Maximum Bandwidth	192 GB/s	336 GB/s	336 GB/s	720 GB/s

XENON High Performance Computing

XENON
High Performance Computing

NVIDIA

Thank You

XENON Systems Pty Ltd
10 Westall Road. Clayton South. Victoria 3169. Australia.
www.xenon.com.au

P +61 3 9549 1115 | **F** +61 3 9549 1199 | **E** info@xenon.com.au

A member of the XENON Technology Group
www.xtg.com.au