GPU TECHNOLOGY CONFERENCE

MELBOURNE | Oct.4-5, 2016

DEEP LEARNING AND ACCELERATED ANALYTICS: FASTER, BETTER RESULTS, UNIQUE INSIGHT

Werner Scholz, 4 Oct. 2016
XENON Systems, CTO and Head of R&D
werners@xenon.com.au

XENON High Performance Computing
www.xenon.com.au

NVIDIA

XENON SYSTEMS – WHO WE ARE

NVIDIA Elite Partner

Australian company established in 1996.

Direct Relationship with all major component manufacturers to lower cost and speed up support

In-house technical ability to build low volume custom designed servers

Focused on innovation and investment in Research & Development

Global Onsite hardware support and installation services in over 80 countries

Subsidiaries:

Mediaproxy
Global leader in compliance logging and transport stream monitoring for broadcast and TV industries.

XDT/Catapult
Software for film and post production industries.

XenOptics
Fibre automation solutions for SDN in data centres

XENON SOLUTIONS

XENON Nitro visual workstations

Performance and Reliability for the most demanding graphics, engineering, digital arts workloads and optimised for MATLAB.

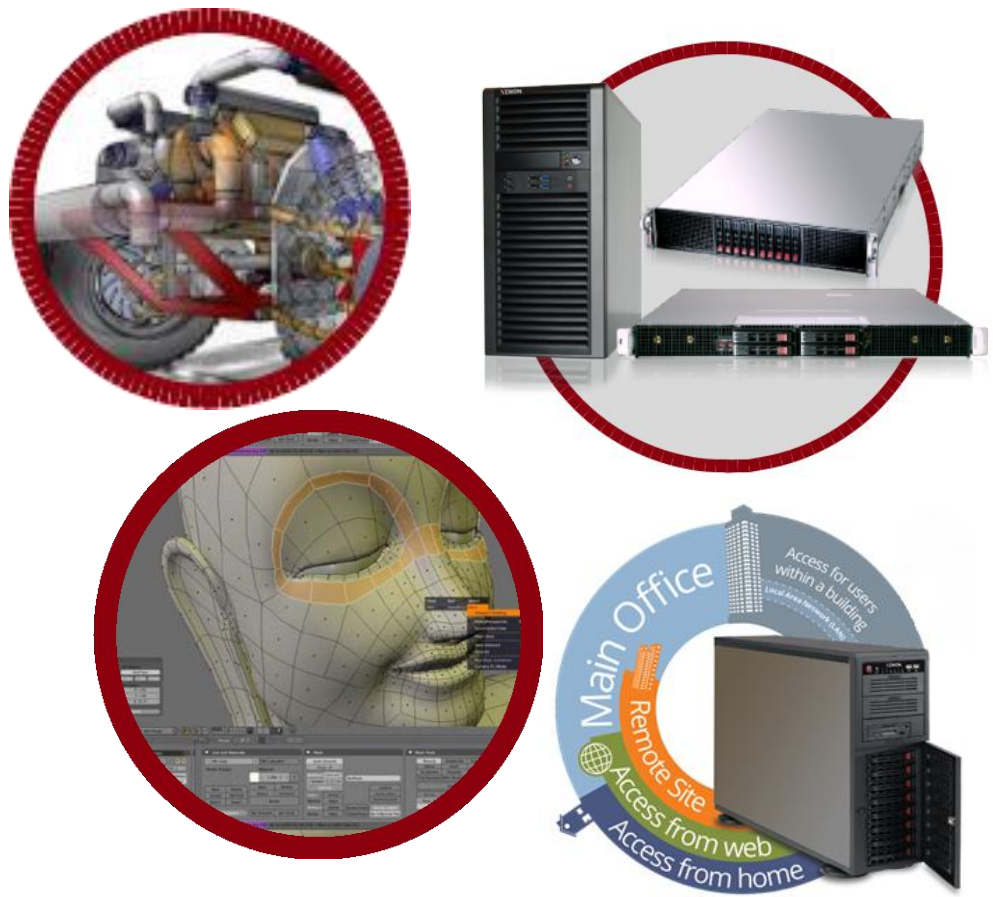
GPU Computing

High performance **acceleration solutions** 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 SYSTEMS – HISTORY

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
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 cluster for engineering workloads

2016
Garvan Institute
High performance Compute and Storage cluster, VDI solution

Walter and Eliza Hall Institute
Virtualised HPC cluster in private cloud, VDI

1990 2000 2010

CSIRO GPU CLUSTER “BRAGG”

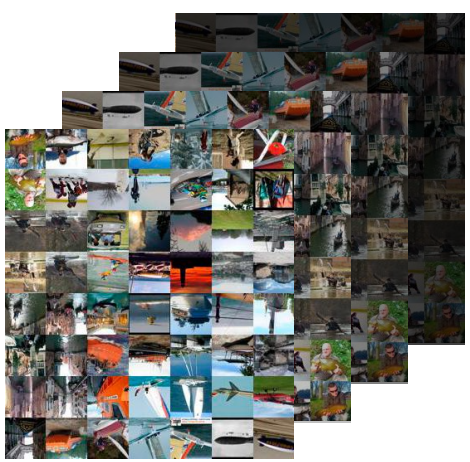
Designed and delivered by XENON Systems

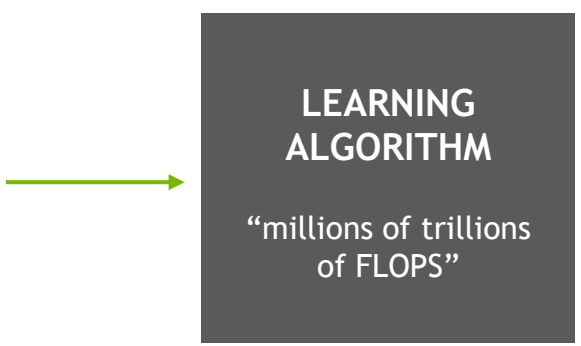
- 128 nodes
- 384x NVIDIA Tesla K20 GPUs

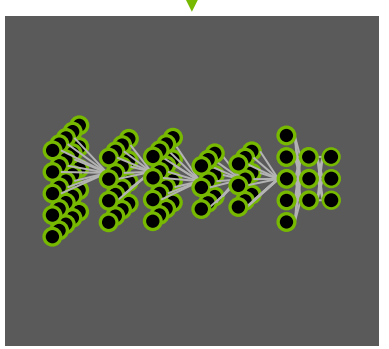
  (384 GPUs = 958,464 Thread Processors)
- 2048 CPU cores
- 16.4TB System Memory
- InfiniBand Interconnect FDR10 40Gb/s
- Linpack Result: 335Tflops (Double Precision)
- #260 in Top500 and #10 in Green500 (in 2013)

DEEP LEARNING — A NEW COMPUTING MODEL

"Software that writes software"





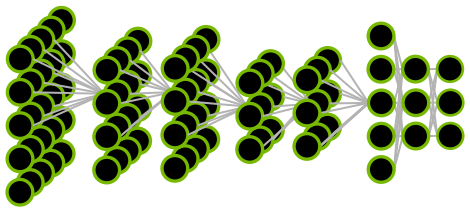


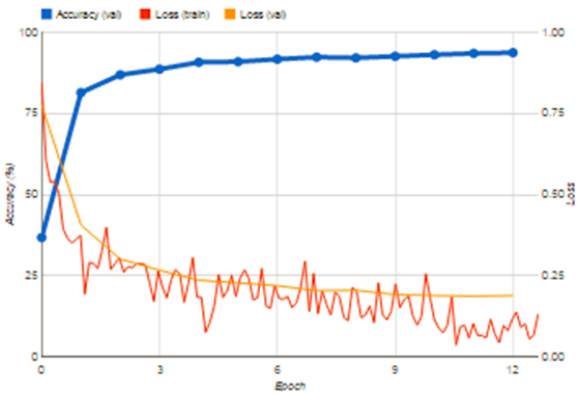
"little girl is eating
piece of cake"

OBJECT RECOGNITION

…in 7 Lines of Code

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

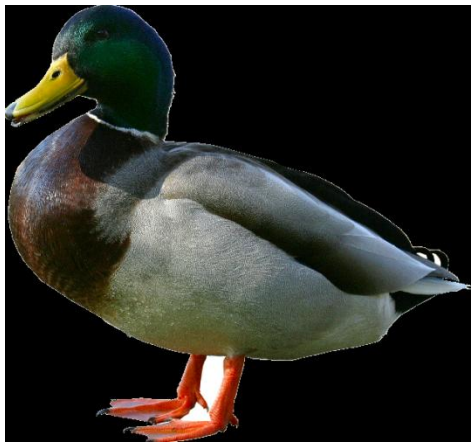


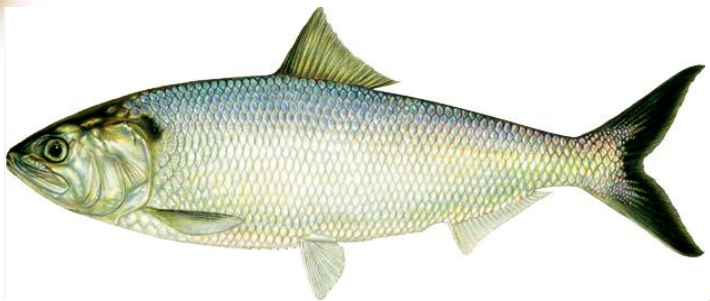


Every network is only as good as its training.

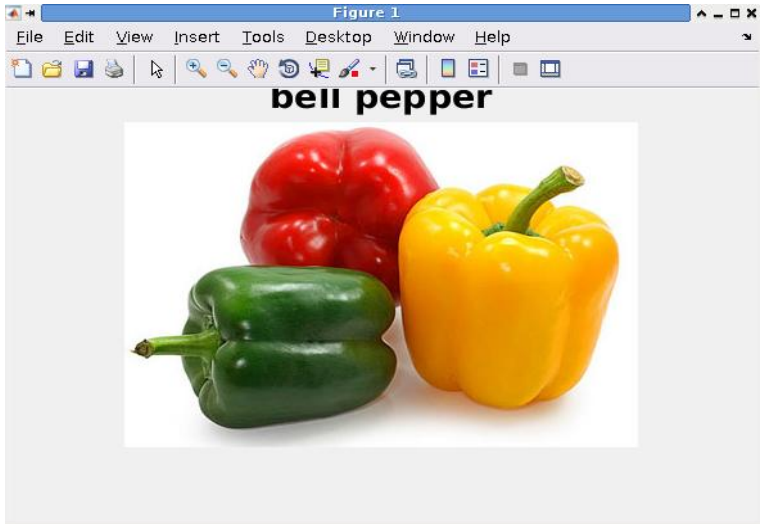
WHAT'S IN AN IMAGE?

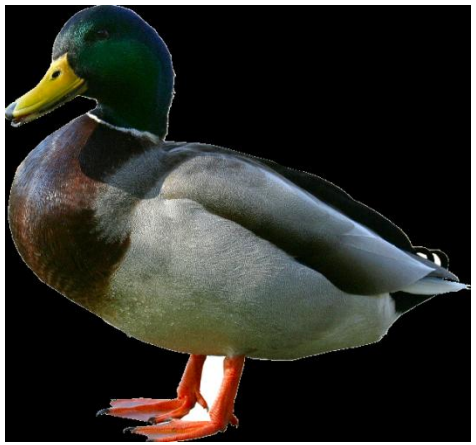
An image says more than a thousand words…but what does it say?

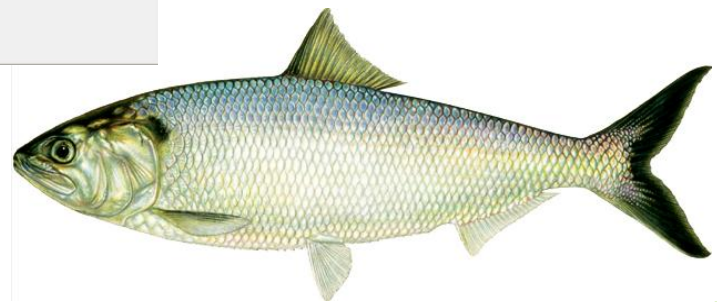




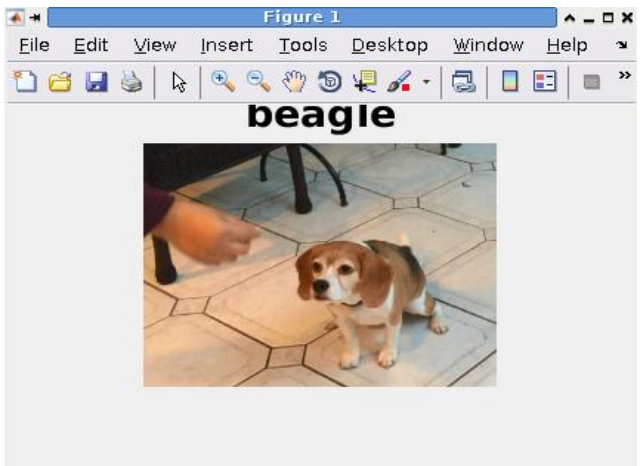
WHAT'S IN AN IMAGE?

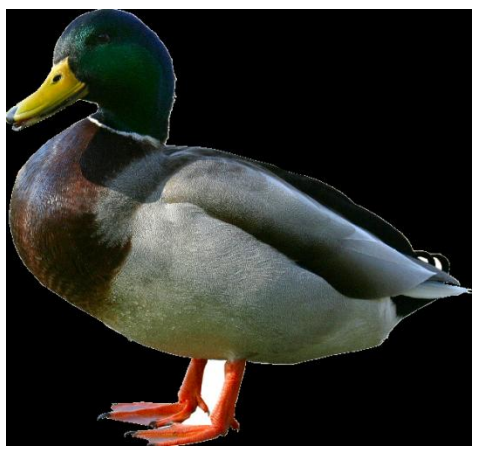


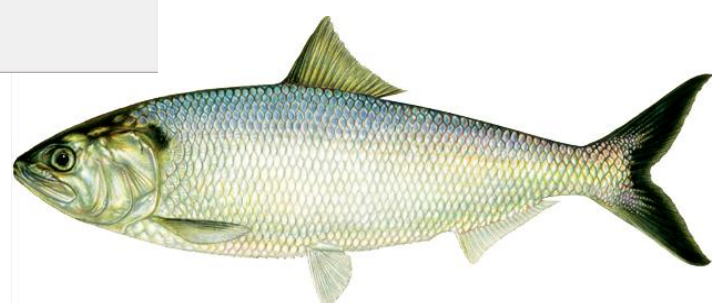




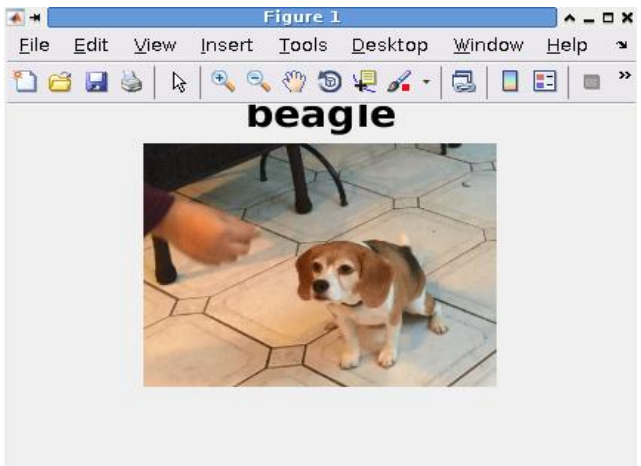
WHAT'S IN AN IMAGE?

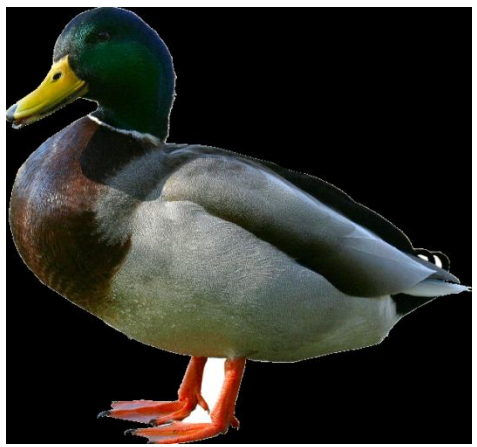


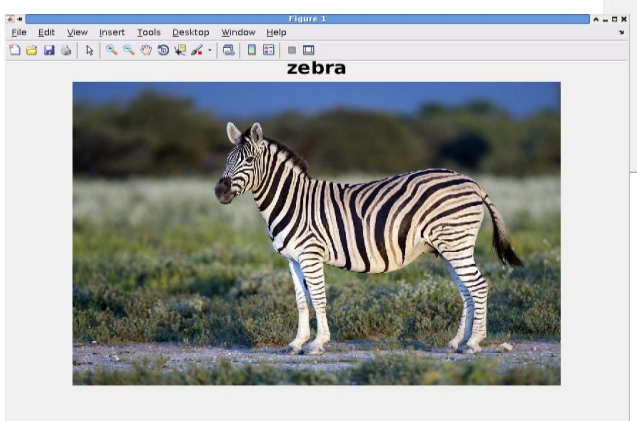


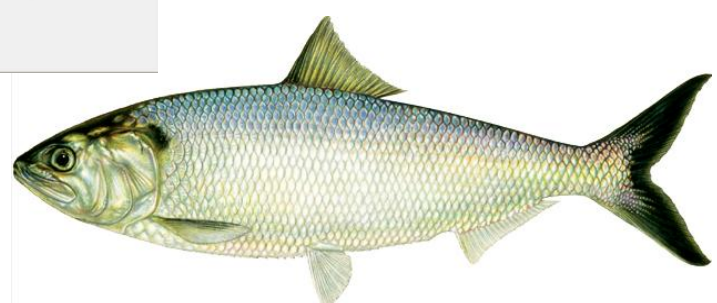


WHAT'S IN AN IMAGE?

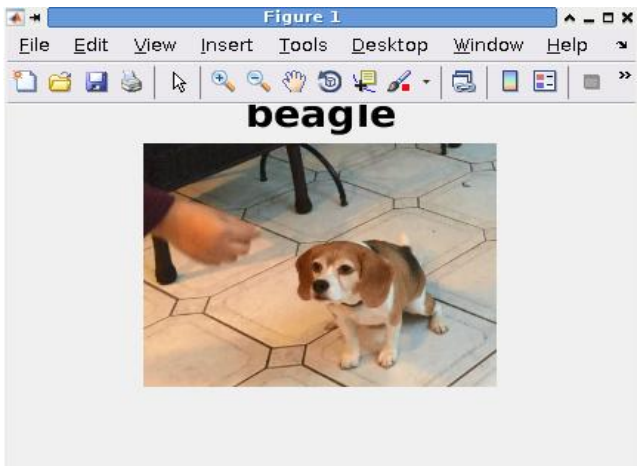


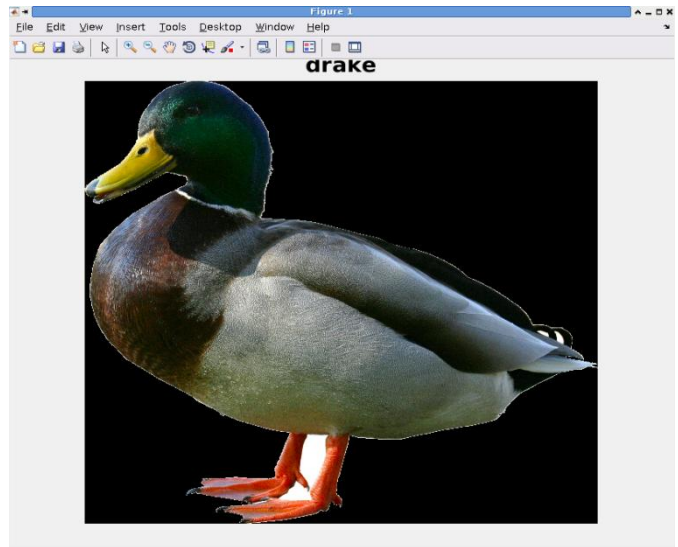


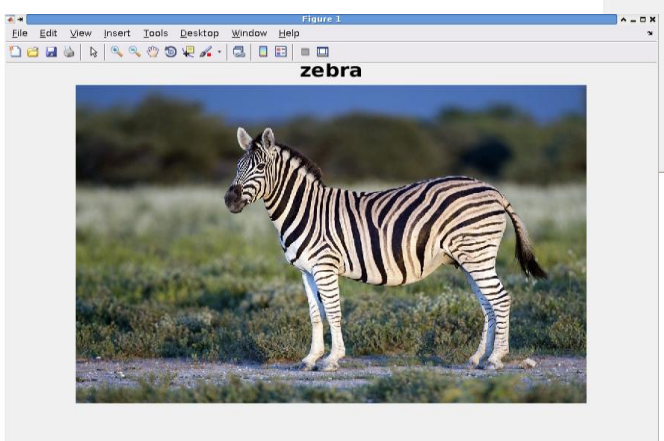


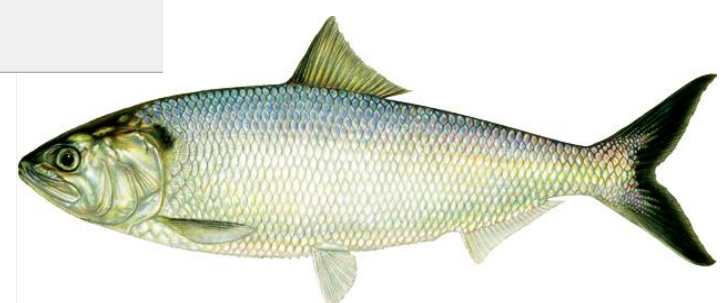


WHAT'S IN AN IMAGE?

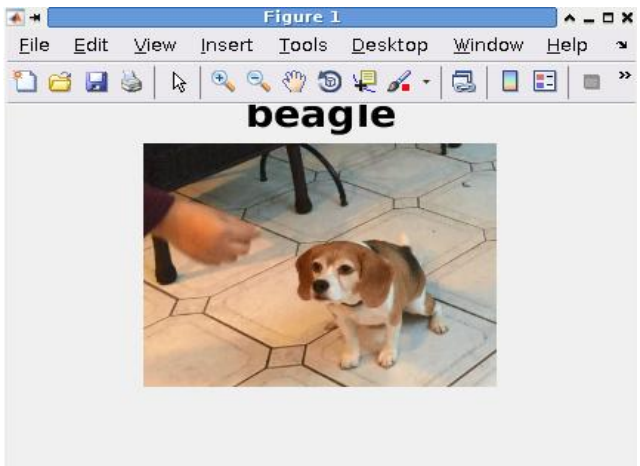


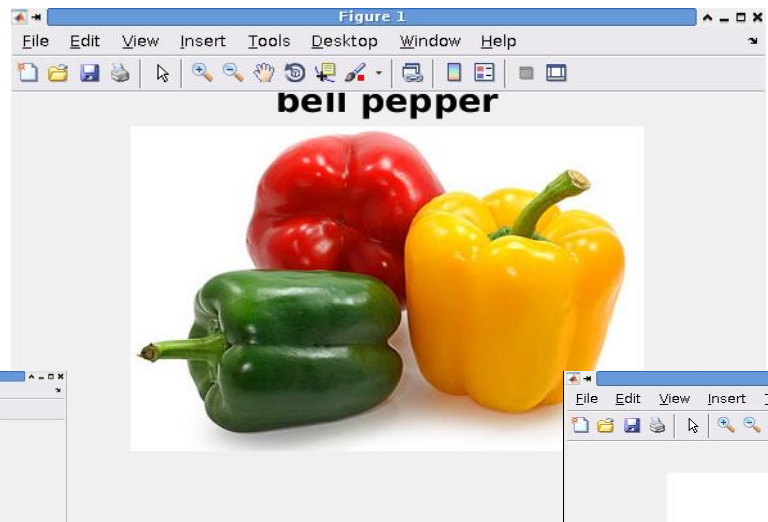


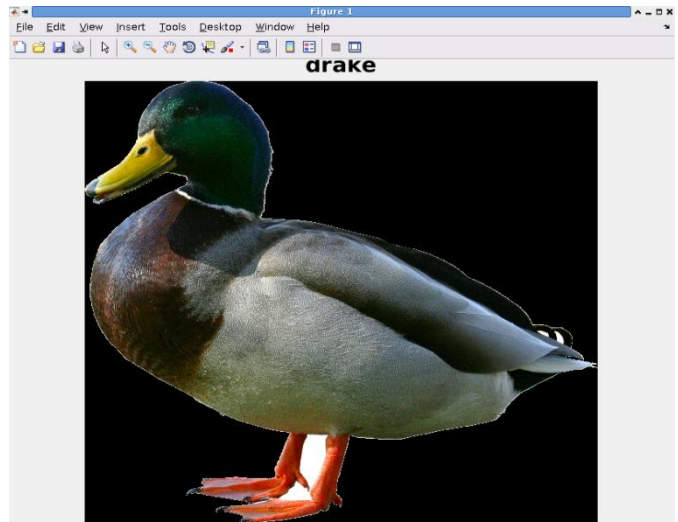


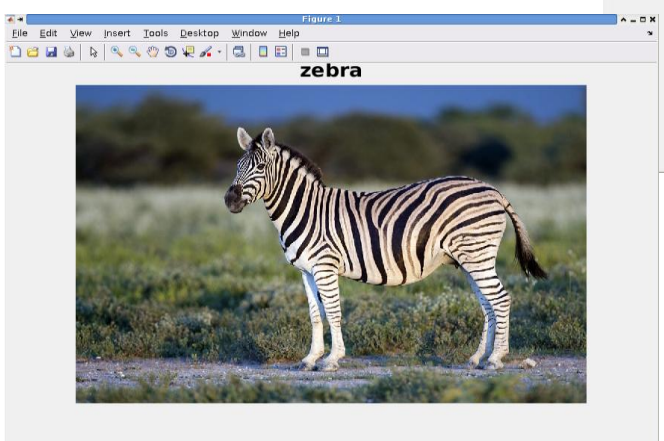


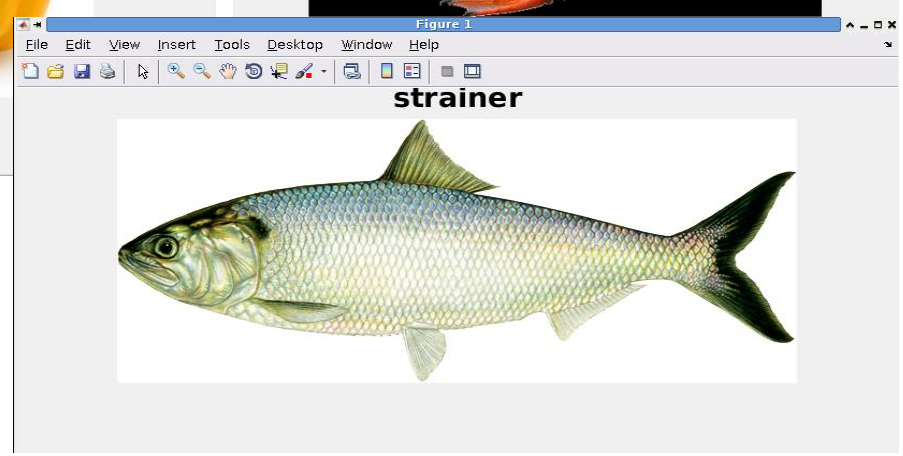
WHAT'S IN AN IMAGE?











XENON
High Performance Computing

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 and a little bit of MATLAB code…

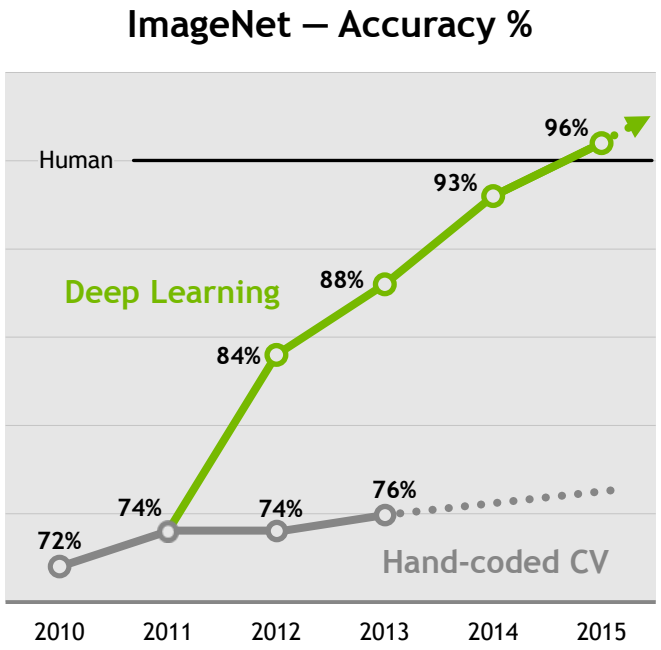
OBJECT RECOGNITION

…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/

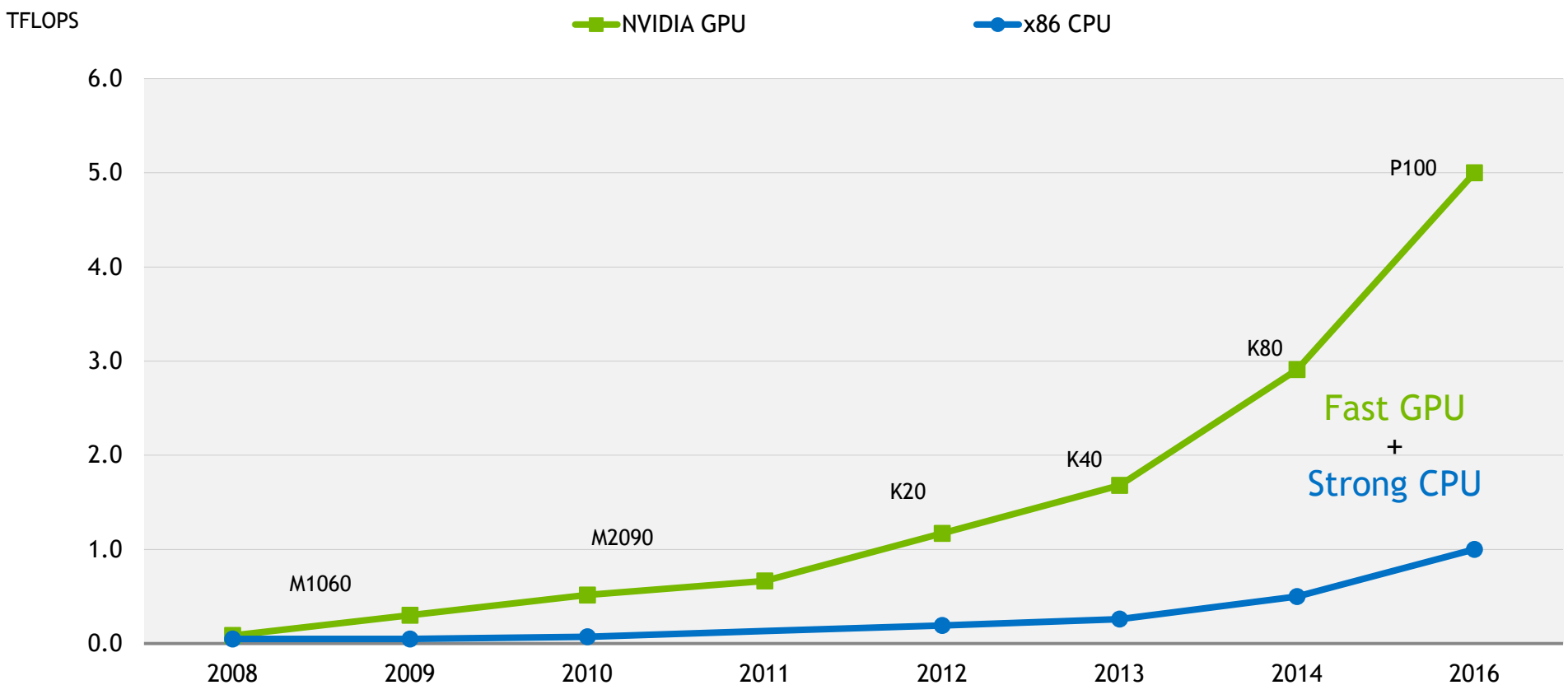
“SUPERHUMAN” RESULTS SPARK HYPERSCALE ADOPTION



Alibaba/Aliyun
Amazon
Baidu
eBay
Facebook
Flickr
Google
iFLYTEK
iQIYI
JD.com
Orange
Periscope
Pinterest
Qihoo 360
Shazam
Skype
Sogou
Twitter
Yahoo Supermarket
Yandex
Yelp

Cloud Services with AI Powered by NVIDIA

THE ADVANTAGES OF GPU-ACCELERATED DATA CENTER



NVIDIA DEEP LEARNING SDK

High Performance GPU-Acceleration for Deep Learning

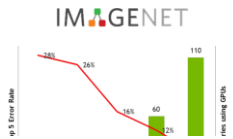


Image Classification

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

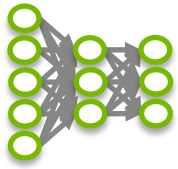
Pylearn2

TensorFlow

torch

theano

DEEP LEARNING FRAMEWORKS

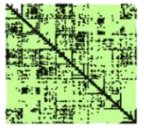


cuDNN

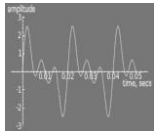
DEEP LEARNING

cuBLAS

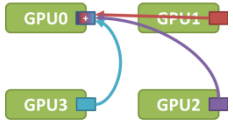
cuSPARSE



cuFFT



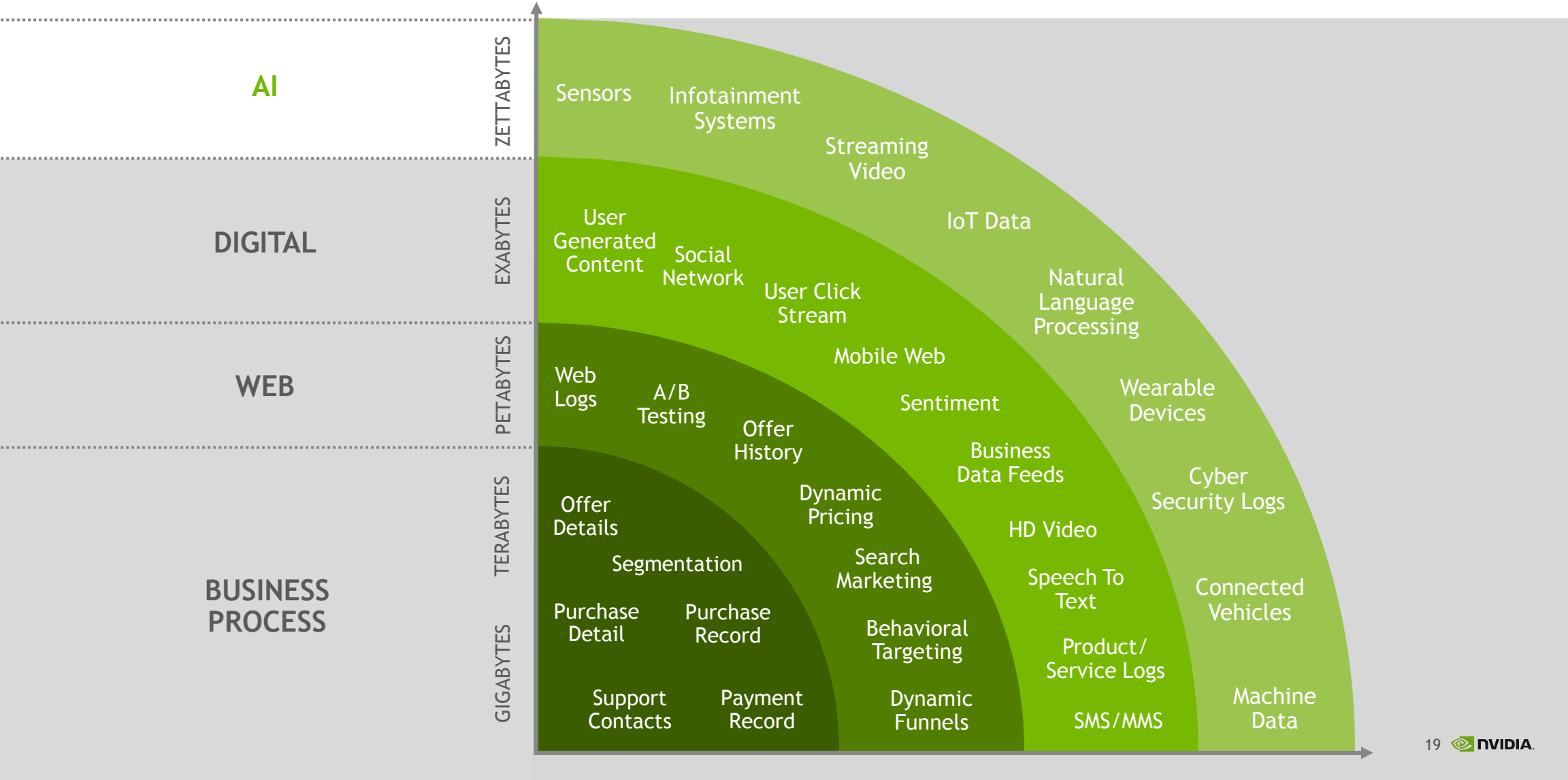
MATH LIBRARIES



NCCL

MULTI-GPU

DATA DELUGE TO DATA HUNGRY



GPU ACCELERATION OVERCOMES THE CHALLENGES OF SLOW COMPUTE ON ANALYTICS

Long response time constrains questions asked

ASK QUESTIONS YOU DON'T KNOW THE ANSWERS TO

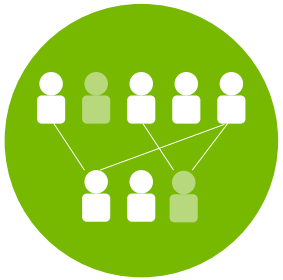
Issuing iterative queries becomes wearisome

EXPLORE FURTHER

Analyst creativity is impaired

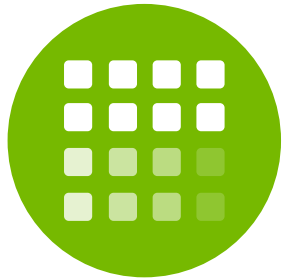
GO BEYOND WHAT'S BEING ASKED

WORKAROUNDS ARE NOT THE ANSWERS



Sampling misses the whole picture

EXPLORE THE OUTLIERS AND LONG-TAIL EVENTS



Pre-aggregation struggles at scale

RELY ON ACCURATE DATA

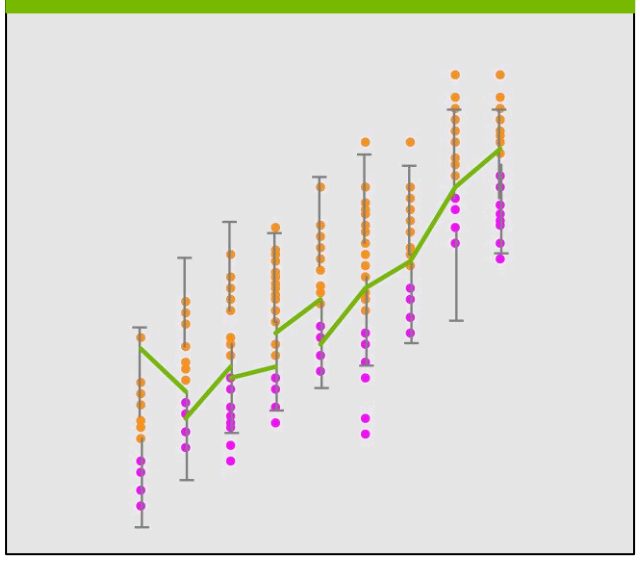
Scale out on CPU infrastructure has tremendous hidden costs

SCALE WITH A ROI

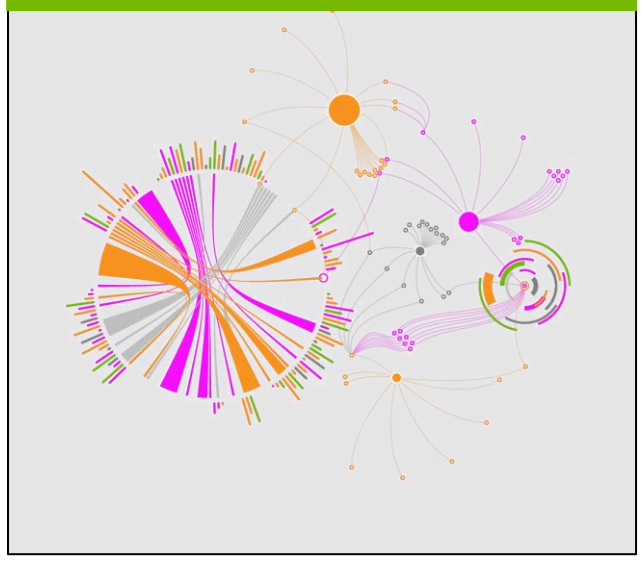
NVIDIA ACCELERATED ANALYTICS

GPUs in the Data Center

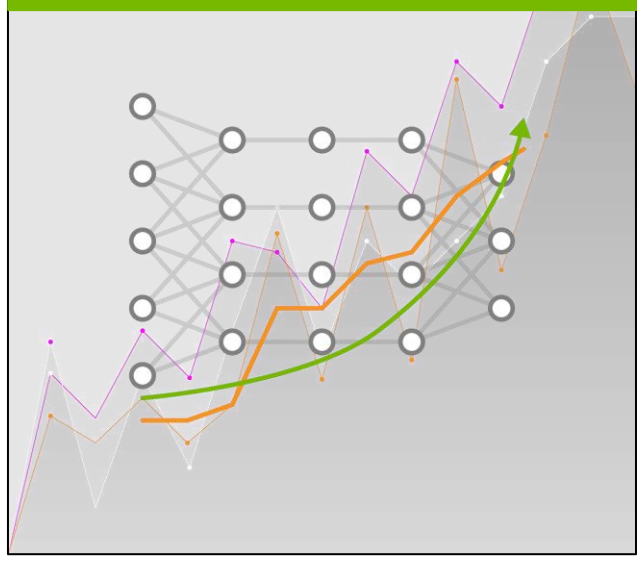
ANALYZE



VISUALIZE



AI-ACCELERATE



WHAT DOES IT RUN ON?

- XENON workstations with NVIDIA GPUs
- XENON DEVCUBE
- XENON Radon rack servers
  - 1U high density servers (up to 4 GPUs)
  - 4U 8-GPU servers
  - 4U 10-GPU servers (see it at our stand!)
- Custom configurations for your requirements
- and…

NVIDIA DGX-1

The Essential Tool for Data Scientists

AI massive opportunity

Data Scientist productivity is vital

NVIDIA is the choice for Deep Learning and AI-accelerated analytics

DGX-1 is fast, instantly productive

NVIDIA DGX-1

AI Supercomputer-in-a-Box

170 TFLOPS | 8x Tesla P100 16GB | NVLink Hybrid Cube Mesh

2x Xeon | 8 TB RAID 0 | Quad IB 100Gbps, Dual 10GbE | 3U — 3200W

DGX-1 INSTALLATIONS

Users in Academia, Research, Enterprise

OpenAI (San Francisco, USA): world's leading non-profit artificial intelligence research team

New York University, Stanford University, UC Berkeley (USA)

SAP (Germany, Israel): machine learning solutions for SAP's 320,000 customers

BenevolentAI (U.K.): accelerate drug discovery

Massachusetts General Hospital (USA): Advance health care with AI to improve the detection, diagnosis, treatment, and management of diseases

University of Reims Champagne-Ardenne (France): prevent or cure diseases affecting grapevines -and ensure the health of the region's most famous export: champagne

Pacific Northwest National Lab. (WA, USA): machine learning research

German Research Center for Artificial Intelligence (DFKI): advancing basic research in deep learning

Dalle Molle Institute for Artificial Intelligence (IDSIA; Switzerland): machine learning, operations research, data mining, and robotics.

and…

https://blogs.nvidia.com/blog/2016/08/15/first-ai-supercomputer-openai-elon-musk-deep-learning/
https://blogs.nvidia.com/blog/2016/09/28/sap-benevolentai-dgx-1/
http://www.lunion.fr/807432/article/2016-09-22/l-urca-premiere-universite-europeenne-a-s-equiper-d-un-serveur-dgx-1-pour-booste
http://www.pnnl.gov/science/highlights/highlight.asp?id=4431
https://www.top500.org/news/nvidia-expands-machine-learning-footprint-in-europe-previews-first-volta-gpu/
http://nvidianews.nvidia.com/news/nvidia-massachusetts-general-hospital-use-artificial-intelligence-to-advance-radiology-pathology-genomics

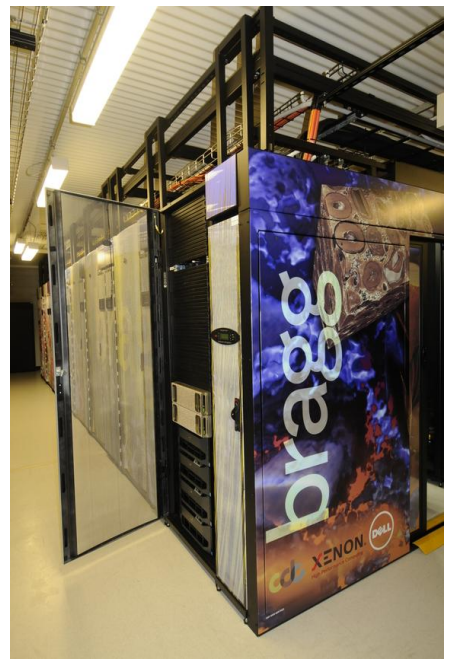
AUSTRALIA'S FIRST DGX-1 AT CSIRO

XENON Systems is NVIDIA's exclusive partner for DGX-1 in ANZ

XENON Systems delivered the first (two) DGX-1 in Australia to CSIRO on 13 Sept. 2016

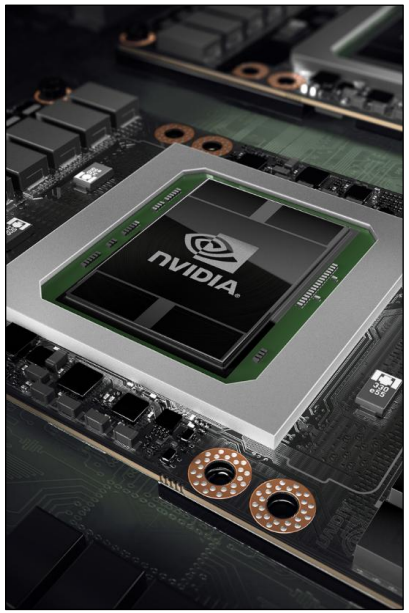
Expand research in areas such as

- Medical image analysis
- Nano-material modelling
- Genome analysis
- Astronomy
- Earth observation and analysis
- Health and biosecurity: Apply Deep Learning techniques to understand and model the effect of the environment on disease

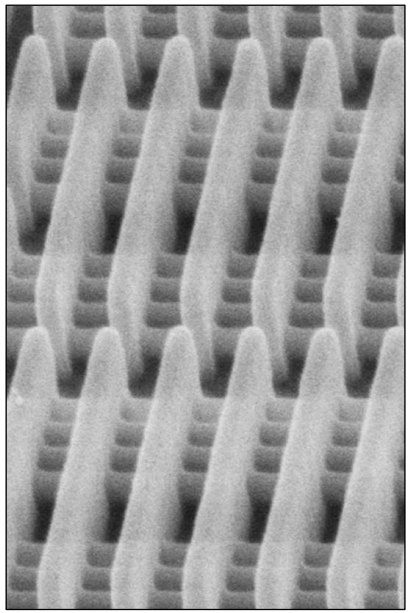


HPL Linpack results:
29 TFLOPS (DP per DGX-1)
59 TFLOPS (DP over Infiniband)
~10 GFLOPS/W

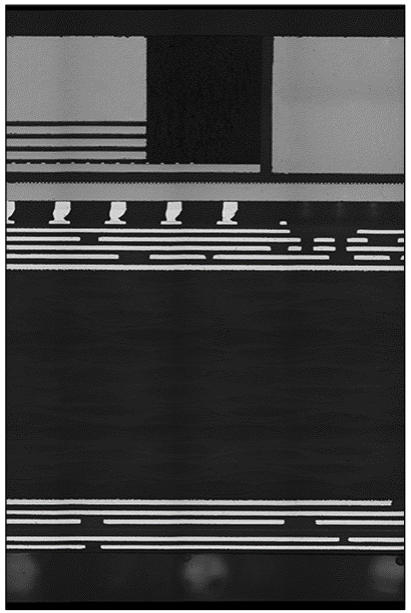
“FIVE MIRACLES”



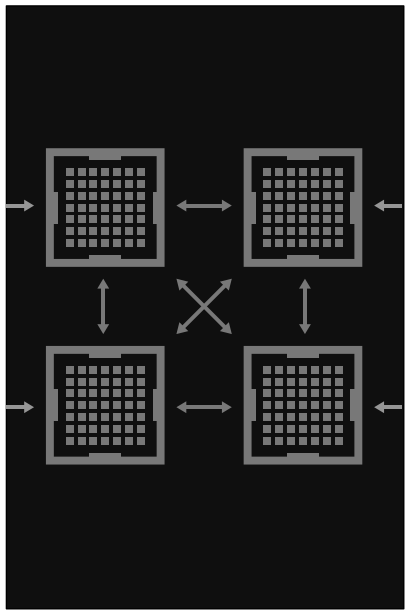
Pascal Architecture



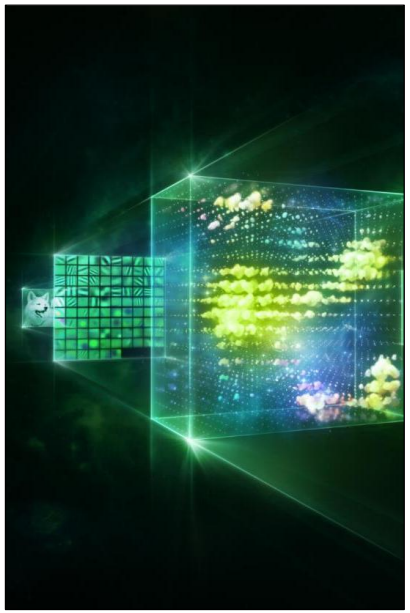
16nm FinFET



CoWoS with HBM2

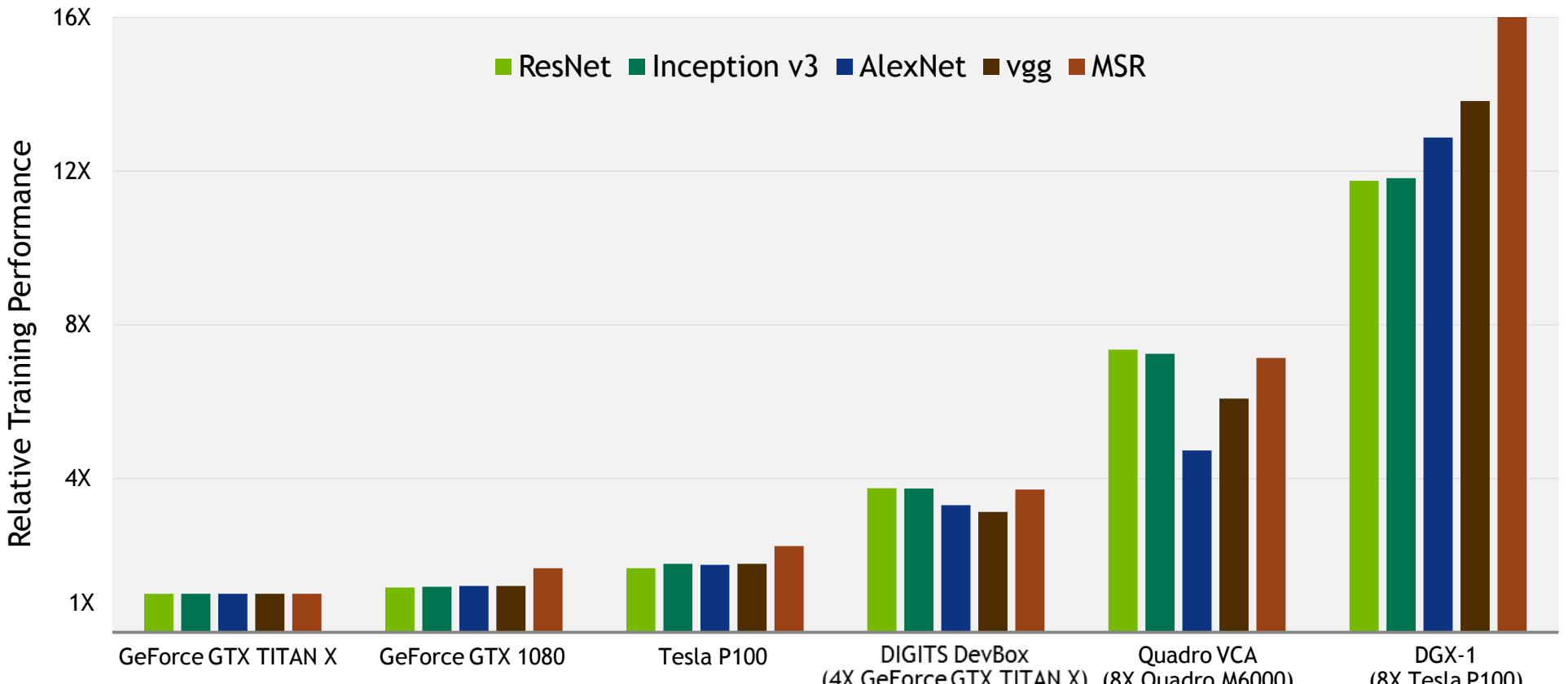


NVLink



New AI Algorithms

DGX-1 — A LEAGUE OF ITS OWN



Caffe on DeepMark. GeForce TITAN X and GTX 1080 system: Intel Core i7-5930K @ 3.5 GHz, 64 GB System Memory | Tesla P100 (SXM2) system: Dual CPU server, Intel E5-2698 v4 @ 2.2 GHz, 256 GB System Memory

DGX — THE ESSENTIAL TOOL
FOR DEEP LEARNING & ACCELERATED ANALYTICS

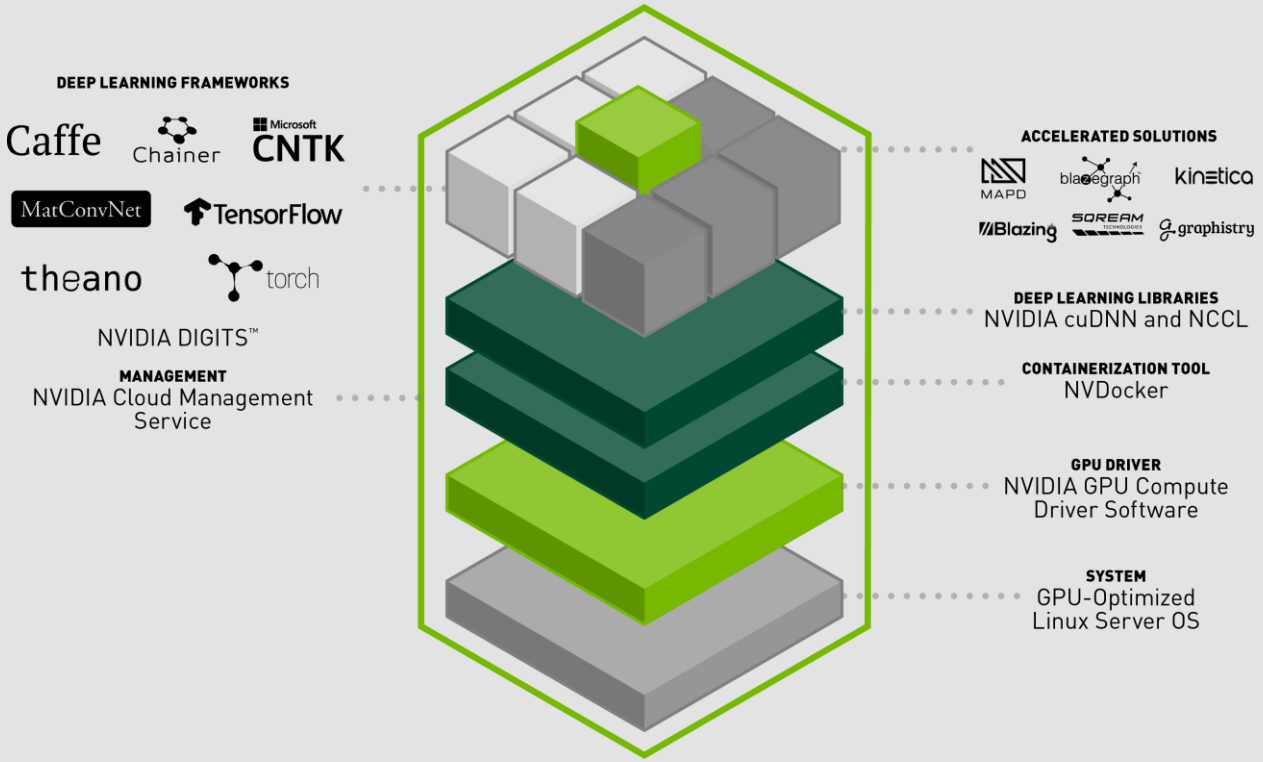
250 NODE AI & ANALYTICS SUPERCOMPUTER-IN-A-BOX

TIME TO INSIGHT 10-100X FASTER

100X MORE DATA IN MILLISECONDS

DGX STACK

Fully integrated Analytics and Deep Learning platform



Instant productivity — plug-and-play, supports every AI framework and accelerated analytics software applications

Performance optimized across the entire stack

Always up-to-date via the cloud

Mixed framework environments — baremetal and containerized

Direct access to NVIDIA experts

GPU-ACCELERATION HAS NO LIMITS

MapD

MapD is 55x to 1,000x faster than comparable CPU databases on billion+ row datasets

Kinetica

Hardware costs that are $^{1}/_{10}$ that of standard in-memory databases

BlazeGraph

200-300x speed-up

Graphistry

See 100x more data at millisecond speed

SQream

The supercomputing powers of the GPU combined with SQream's patented technology, results in up to 100 times faster analytics performance on terabyte-petabyte scale data sets

GPU TECHNOLOGY CONFERENCE

MELBOURNE | Oct.4-5, 2016

DEEP LEARNING AND ACCELERATED ANALYTICS: FASTER, BETTER RESULTS, UNIQUE INSIGHT

Werner Scholz, 4 Oct. 2016
XENON Systems, CTO and Head of R&D
werners@xenon.com.au

XENON High Performance Computing
www.xenon.com.au

NVIDIA