Deep Learning - Tools and Platforms for Today and Tomorrow

Werner Scholz, 15 Aug. 2017
XENON Systems, CTO and Head of R&D
werners@xenon.com.au

XENON
High Performance Computing
www.xenon.com.au

XENON SYSTEMS – WHO WE ARE

IBM and NVIDIA 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

XENON
High Performance Computing

- Defence
- Education
- Broadcast

- Scientific / Academic Research
- Oil and Gas
- Cloud

- Finance
- Telecommunication

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 server solutions

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

GPU Computing

High performance **acceleration solutions** leveraging NVIDIA Tesla technology and the CUDA ecosystem

Virtualisation

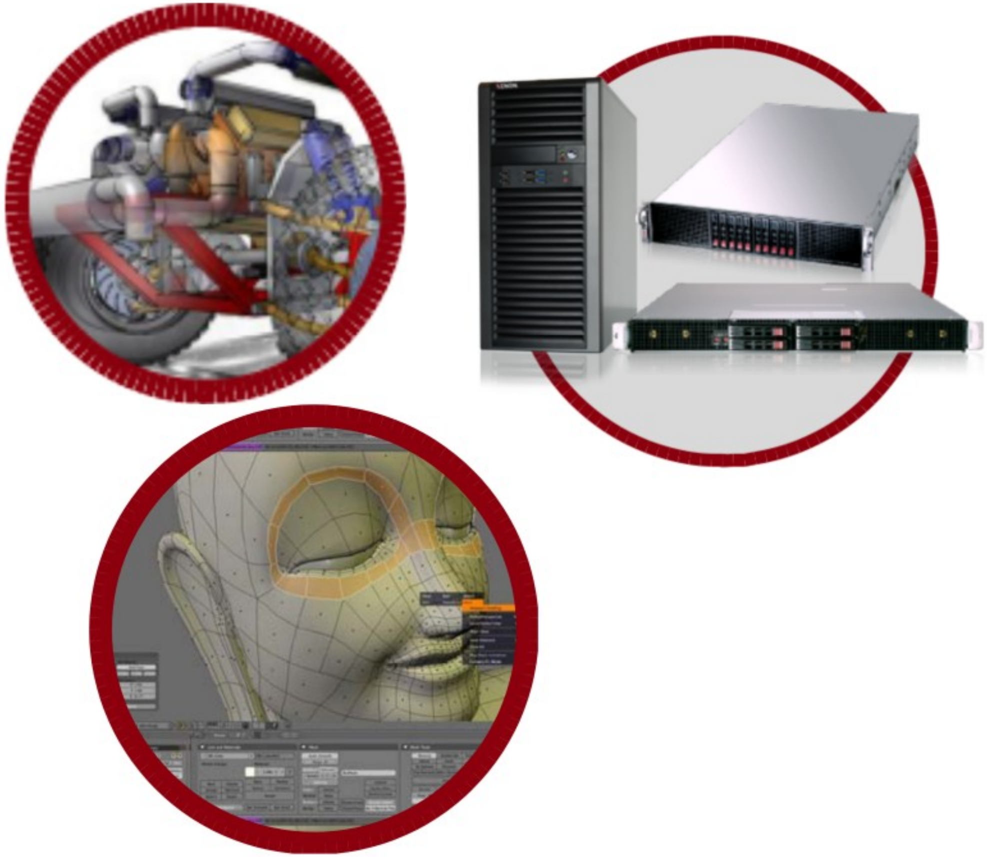
End-to-end virtualisation solutions for compute, storage, networking, and desktop.

Storage

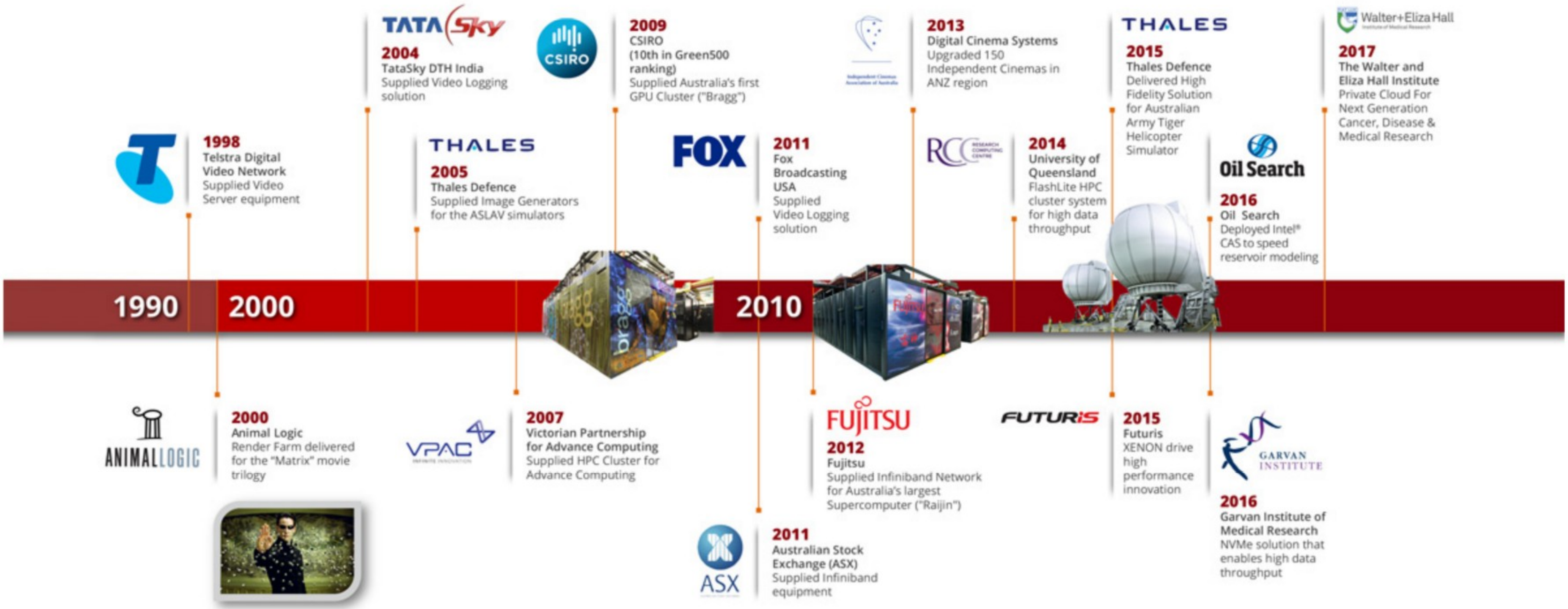
High performance parallel file systems,
e.g. IBM Spectrum Scale

Networking

High performance Infiniband and Ethernet 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 (10th in Green500 ranking)
Supplied Australia's first GPU Cluster ("Bragg")

2011
Fox Broadcasting USA
Supplied Video Logging solution

2011
Australian Stock Exchange (ASX)
Supplied Infiniband equipment

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
XENON drive high performance innovation

2016
Oil Search
Deployed Intel® CAS to speed reservoir modeling

2016
Garvan Institute of Medical Research
NVMe solution that enables high data throughput

2017
The Walter and Eliza Hall Institute
Private Cloud For Next Generation Cancer, Disease & Medical Research

CSIRO GPU CLUSTER “BRAGG”

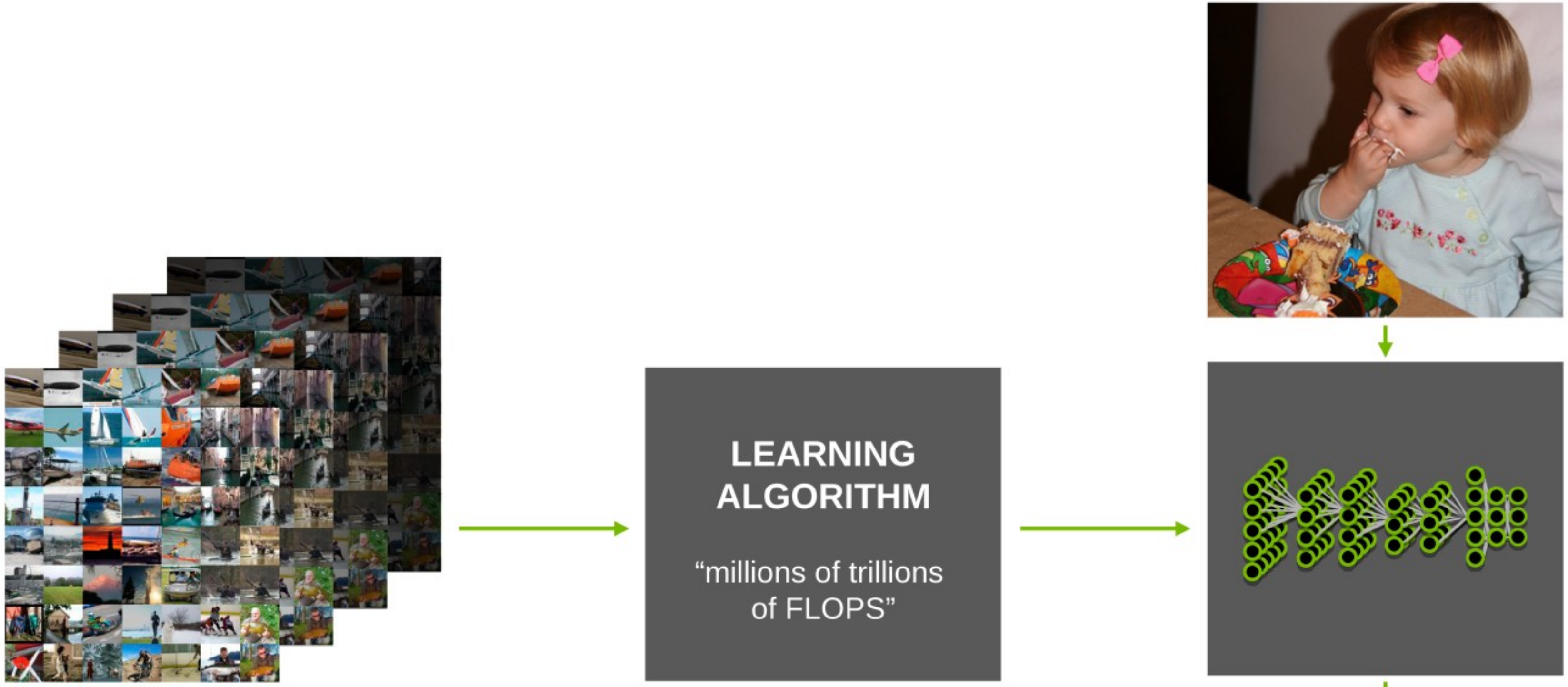
Designed and delivered by XENON Systems

- 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”



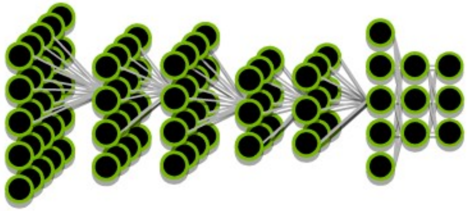
LEARNING ALGORITHM

“millions of trillions of FLOPS”

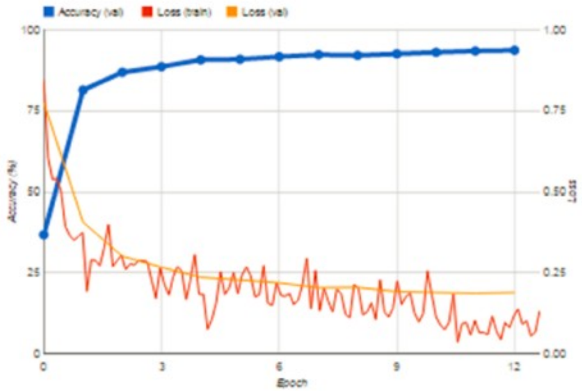
“little girl is eating piece of cake"

OBJECT RECOGNITION

…in 7 Lines of Code



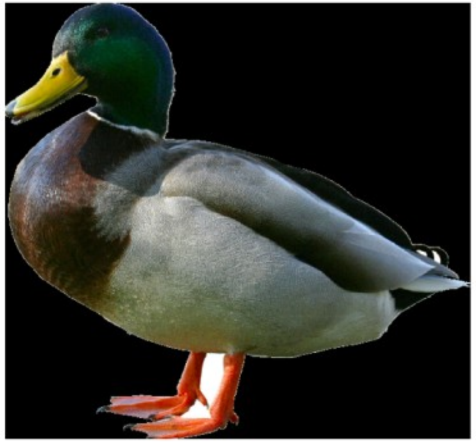
- Design a Deep Neural Network
- Train the network
- Present new images to the network
- Be prepared to be surprised…



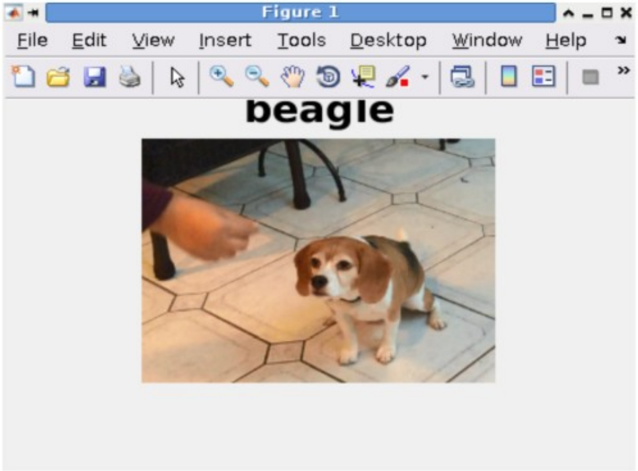
Every network is only as good as its training.

WHAT'S IN AN IMAGE?

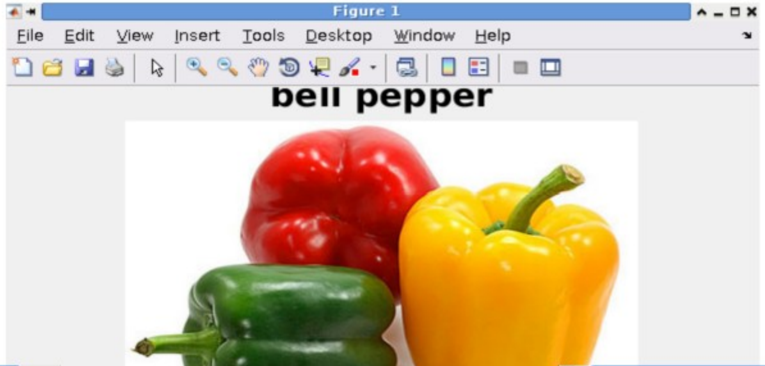
An image says more than a thousand words…but what does it say?



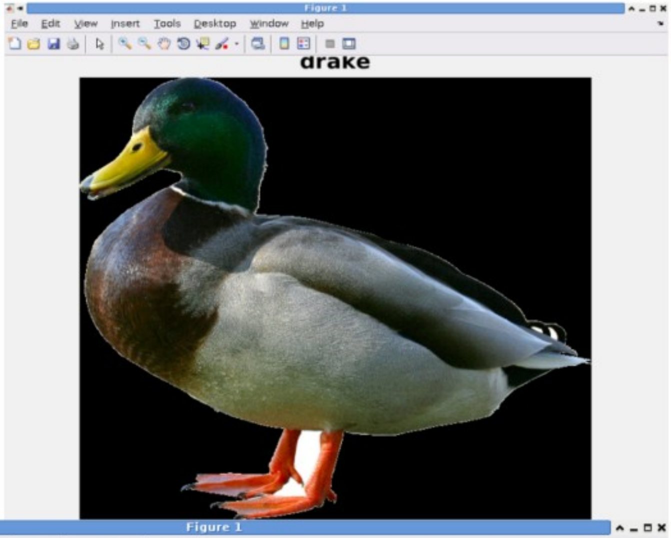
WHAT’S IN AN IMAGE?



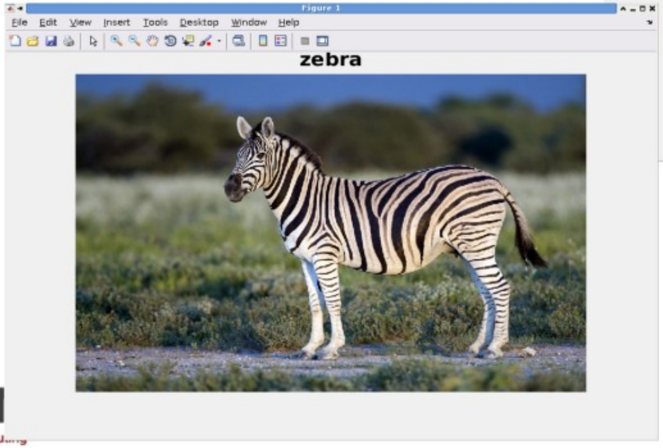
beagle



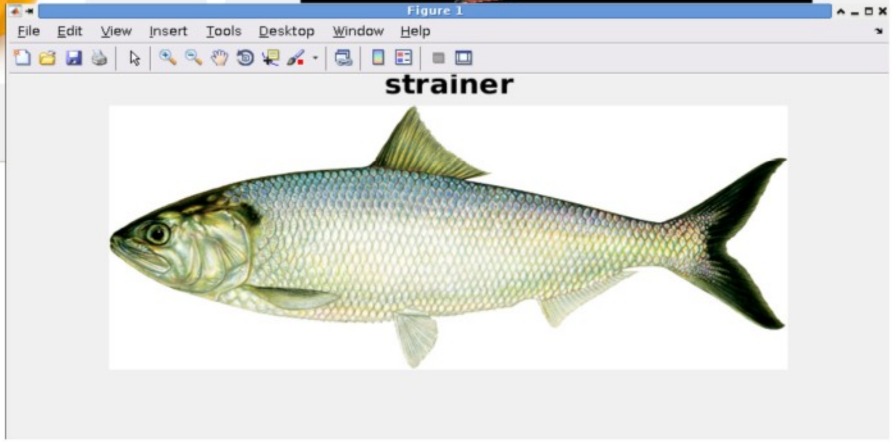
bell pepper



drake



zebra



strainer

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

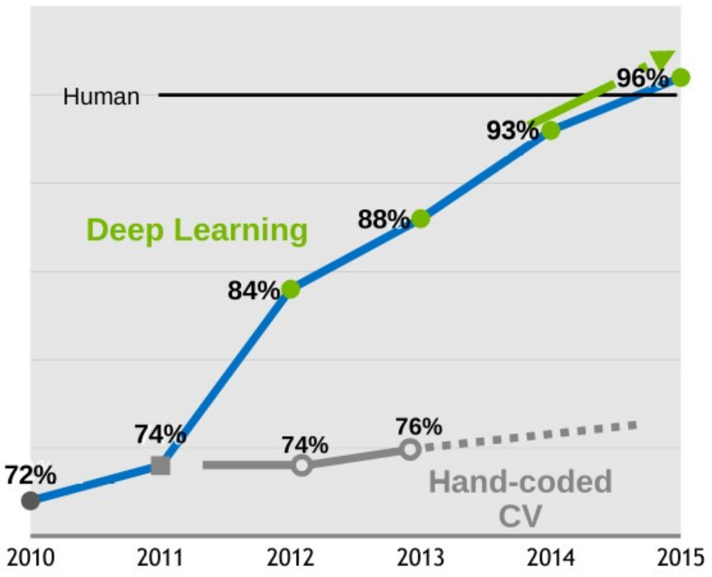
…in 7 lines of MATLAB Code

```
% 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

ImageNet — Accuracy %



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

ACCELERATED DEEP LEARNING TOOLS

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

Chainer

Mocha.jl julia

Microsoft CNTK

MatConvNet

MINERVA

mxnet

OpenDeep

Purine

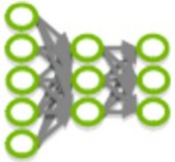
Pylearn2

TensorFlow

torch

theano

DEEP LEARNING FRAMEWORKS



cuDNN

DEEP LEARNING

cuBLAS

cuSPARSE

cuFFT

MATH LIBRARIES

GPU0
GPU1
GPU3
GPU2

NCCL

MULTI-GPU

DEEP LEARNING PLATFORMS - OVERVIEW

Workloads

- Dev and Test
- Training
- Inference

Technologies

- CPU
- GPU
- GPUs for DL (Tensor Cores), single prec., half prec.
- FPGA
- ASICS: TPU, etc.

On-premise

- GPU servers:
  IBM 822SL: Power8 + P100 + NVLINK
- PowerAI

Cloud

- CPU, GPU, FPGA instances
- HWaaS: Softlayer
- DLaaS: Watson, “Tensorflow”aaS

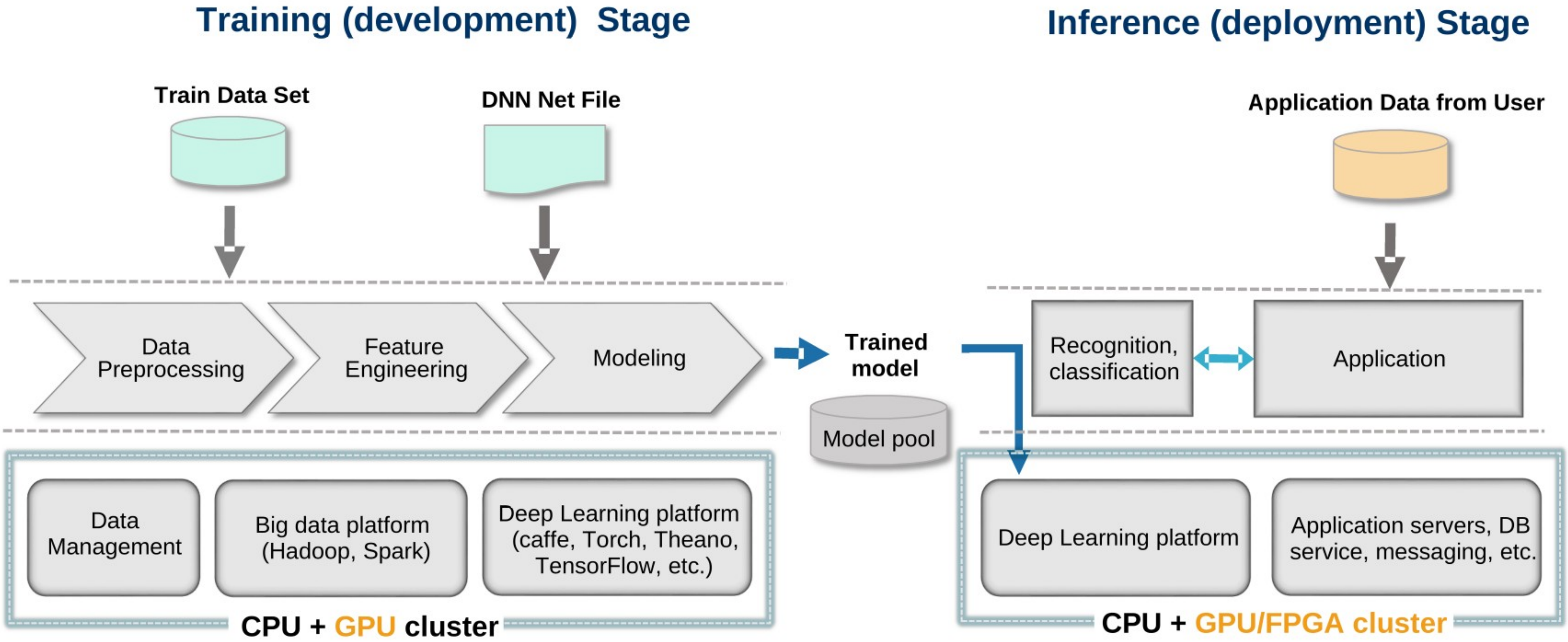
New Services

- AIvision
- DLInsight

Future

- CPU, GPU, FPGA instances
- Power9 + V100 + NVLINK2

TRAINING AND INFERENCE


Training (development) Stage
Inference (deployment) Stage
Train Data Set
DNN Net File
Application Data from User
Data Preprocessing
Feature Engineering
Modeling
Trained model
Model pool
Recognition, classification
Application
Data Management
Big data platform (Hadoop, Spark)
Deep Learning platform (caffe, Torch, Theano, TensorFlow, etc.)
CPU + GPU cluster
Deep Learning platform
Application servers, DB service, messaging, etc.
CPU + GPU/FPGA cluster

Ref: GTC 2017: Yonghua Lin (IBM Research): VisionBrain: Deep Learning Platform for Customized Visual Recognition in Cloud

DATA PREPARATION AND TRAINING

Training

- **Data intensive:** historical data sets
- **Resource intensive:** Input data sets need to be prepared for training
- **Compute intensive:** 100% accelerated
- **Development intensive:** Optimise the model for efficiency and size (possibly for deployment in much smaller devices on the edge)

Data prep

- Data storage
- Data labelling/classification
- Data trim/crop/resize/transform/transcode

Network design/optimisation

- Prebuilt networks
- Pretrained networks
- Optimisation

Training

- Data ingest
- Training
- Convergence visualisation, test
- Network export

On-premise

- IBM S822LC (“Minsky”)

Cloud

- CPU, GPU, FPGA instances
- IBM Bluemix
- HWaaS: IBM Softlayer
- DLaaS: Watson, “Tensorflow”aaS

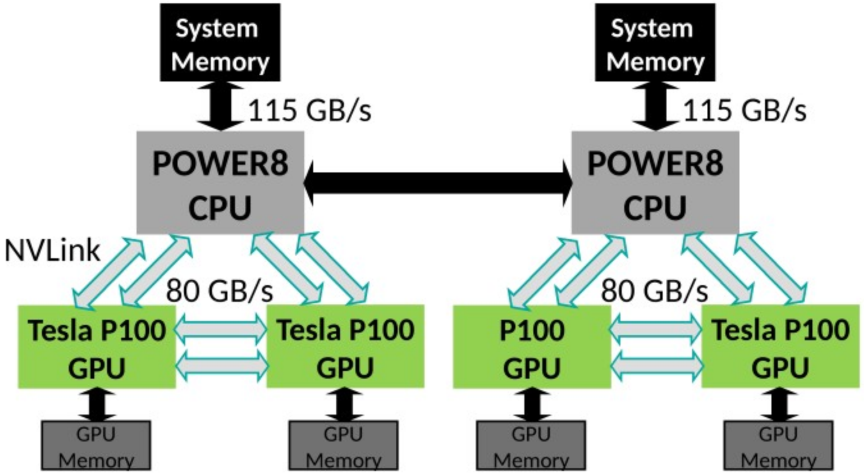
New Services

- AIvision

Future

- IBM Power9 + NVIDIA V100 + NVLINK2

Higher Performance with Power8 CPU-P100 GPU NVLink

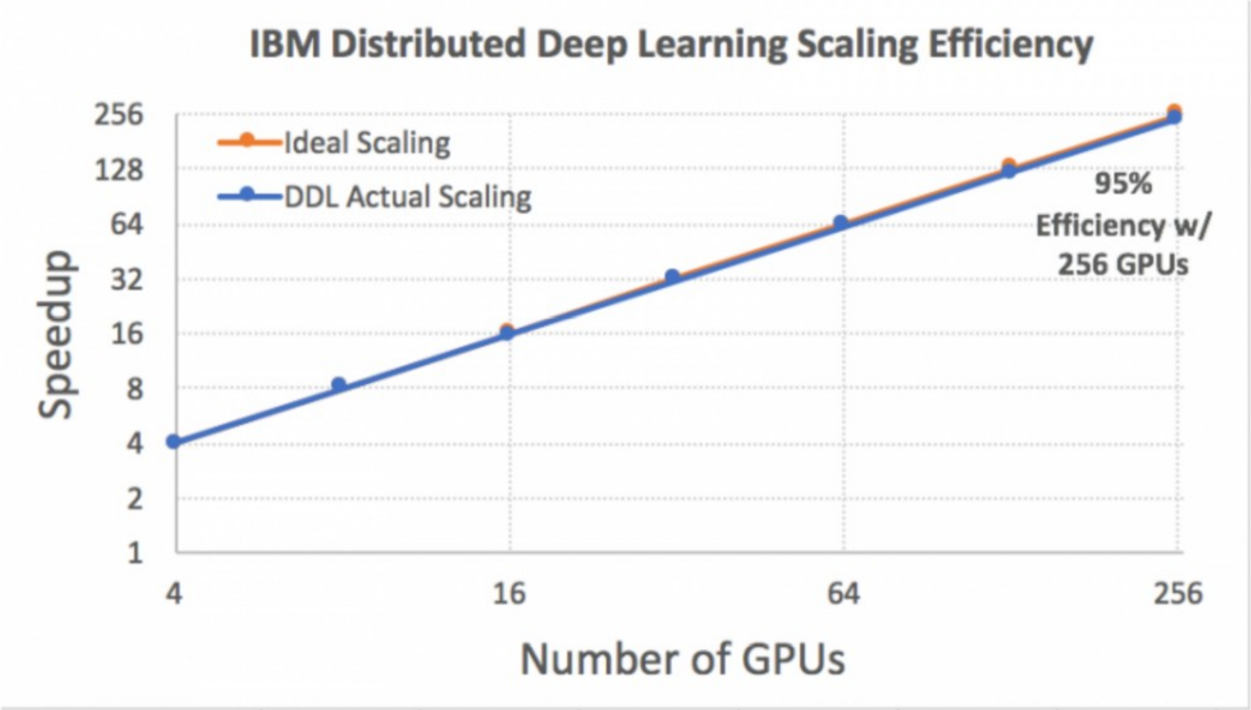


Minsky (S822LC for HPC): Recommended configuration for PowerAI

- 2 Socket, 4 GPU System with NVLink
- 2 POWER8 with NVLink
- Up to 1 TB System Memory
- 4 NVIDIA Tesla P100 GPUs
- 2 SSD storage devices
- High-speed interconnect (IB or Ethernet, depending on infrastructure)

- PowerAI leverages NVLink between CPUs and GPUs to enable fast memory access to large data sets in system memory
- Two NVLink connections between each GPU and CPU-GPU leads to faster data exchange
- Large NN models benefit the most

DISTRIBUTED DEEP LEARNING



Accelerate training by scaling out:

- 16 days on 1x S822LC

Parallel run

- 64 servers S822LC
- Infiniband fabric
- 256 NVIDIA P100 GPU accelerators
- Distributed Deep Learning (DDL) library

- ImageNet-1K data set using a ResNet-50 model
- 16 days reduced to 7 hours (60.6x speed-up): 95% efficiency

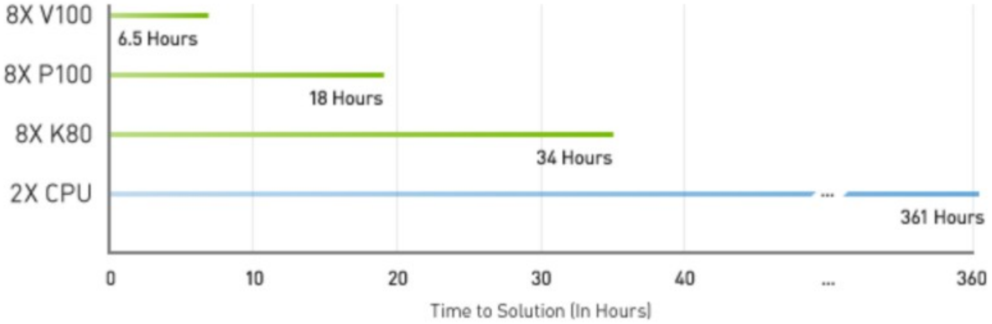
- ImageNet-22K data set using a ResNet-101 model
- 84% efficiency

Ref: https://www.ibm.com/blogs/systems/scaling-tensorflow-and-caffe-to-256-gpus/ , DDL: https://arxiv.org/abs/1708.02188

NVIDIA TESLA V100 (VOLTA ARCHITECTURE)

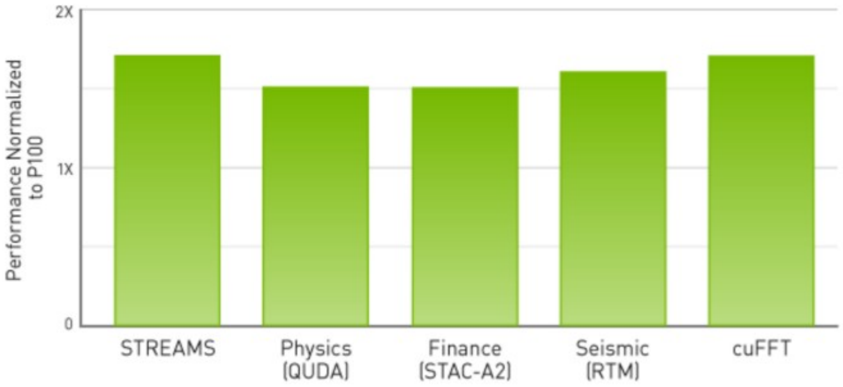
- TSMC 12nm FINFET process
- 21 Billion transistors
- >5000 compute units
- 15 TFLOPS DP
- 640 Tensor Cores
- 120 TFlops tensor operations
- 20MB register file
- 16MB cache
- 900 GB/s memory bandwidth
- 300 GB/s NVLINK2

3X Faster on Deep Learning Training



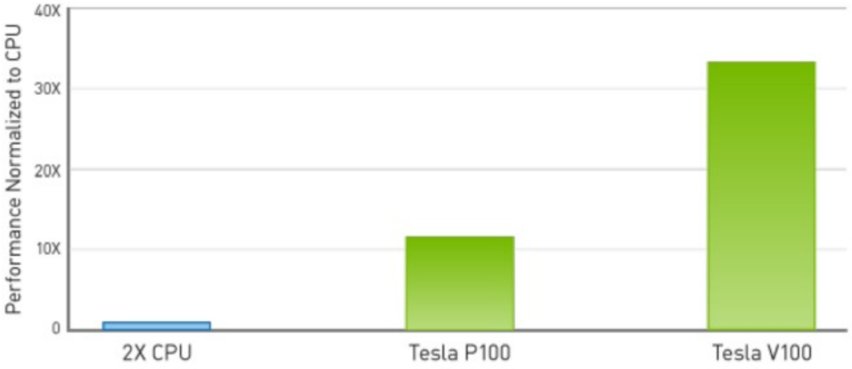
CPU Server: Dual Xeon E5-2699 v4, 2.6GHz | GPU Servers add 8X Tesla K80, Tesla P100 or Tesla V100 | V100 measured on pre-production hardware | Workload: NMT, 13 epochs to solution.

1.5X HPC Performance in One Year



CPU System: 2X Xeon E5-2660 v4 @ 2GHz | GPU System: NVIDIA® Tesla® P100 or V100 at 150W | V100 measured on pre-production hardware | Workload: ResNet-50

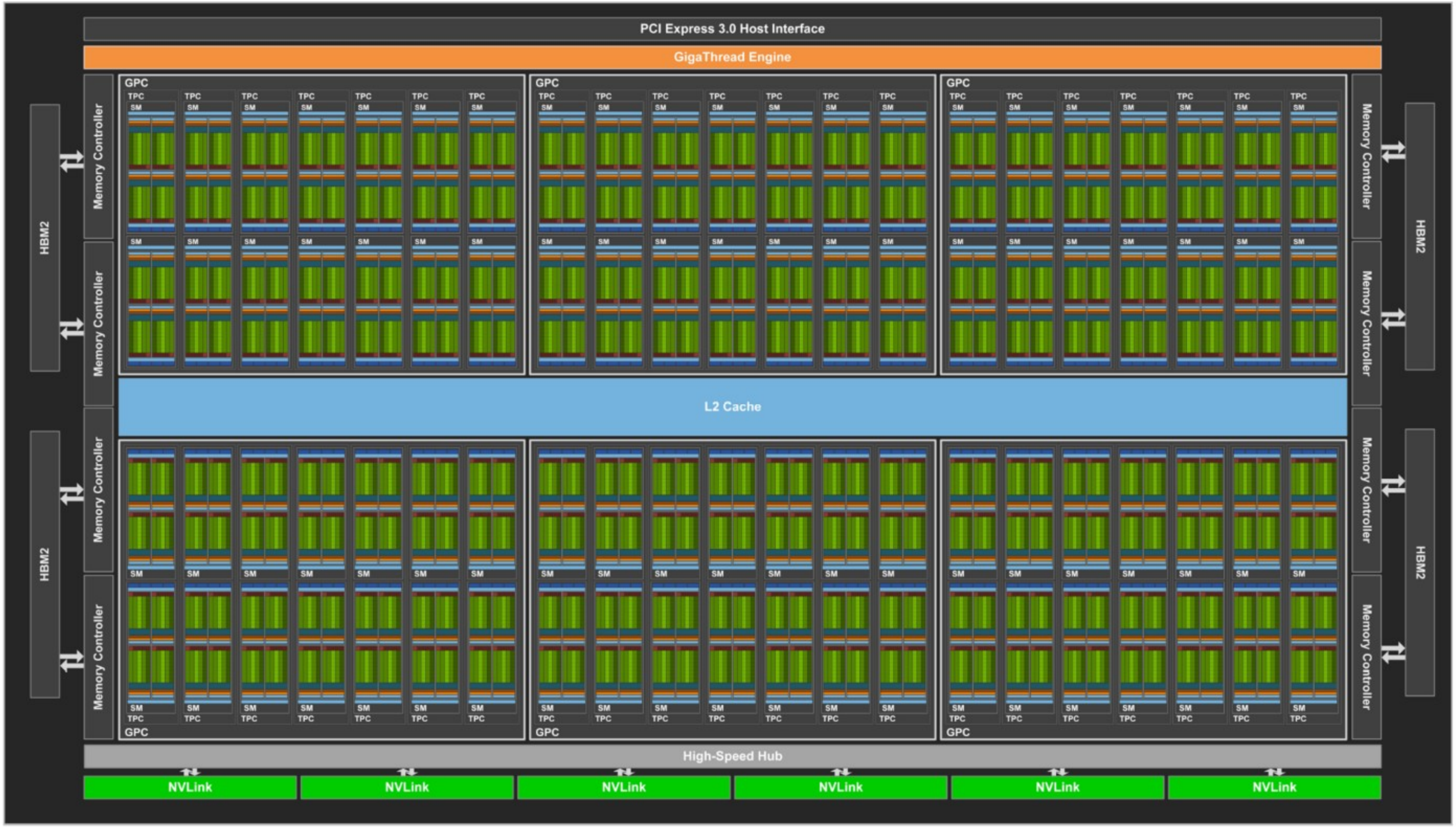
30X Higher Throughput than CPU Server on Deep Learning Inference



Workload: ResNet-50 | CPU: 2X Xeon E5-2660 v4, 2GHz | GPU: add 1X Tesla P100 or V100 at 150W | V100 measured on pre-production hardware.

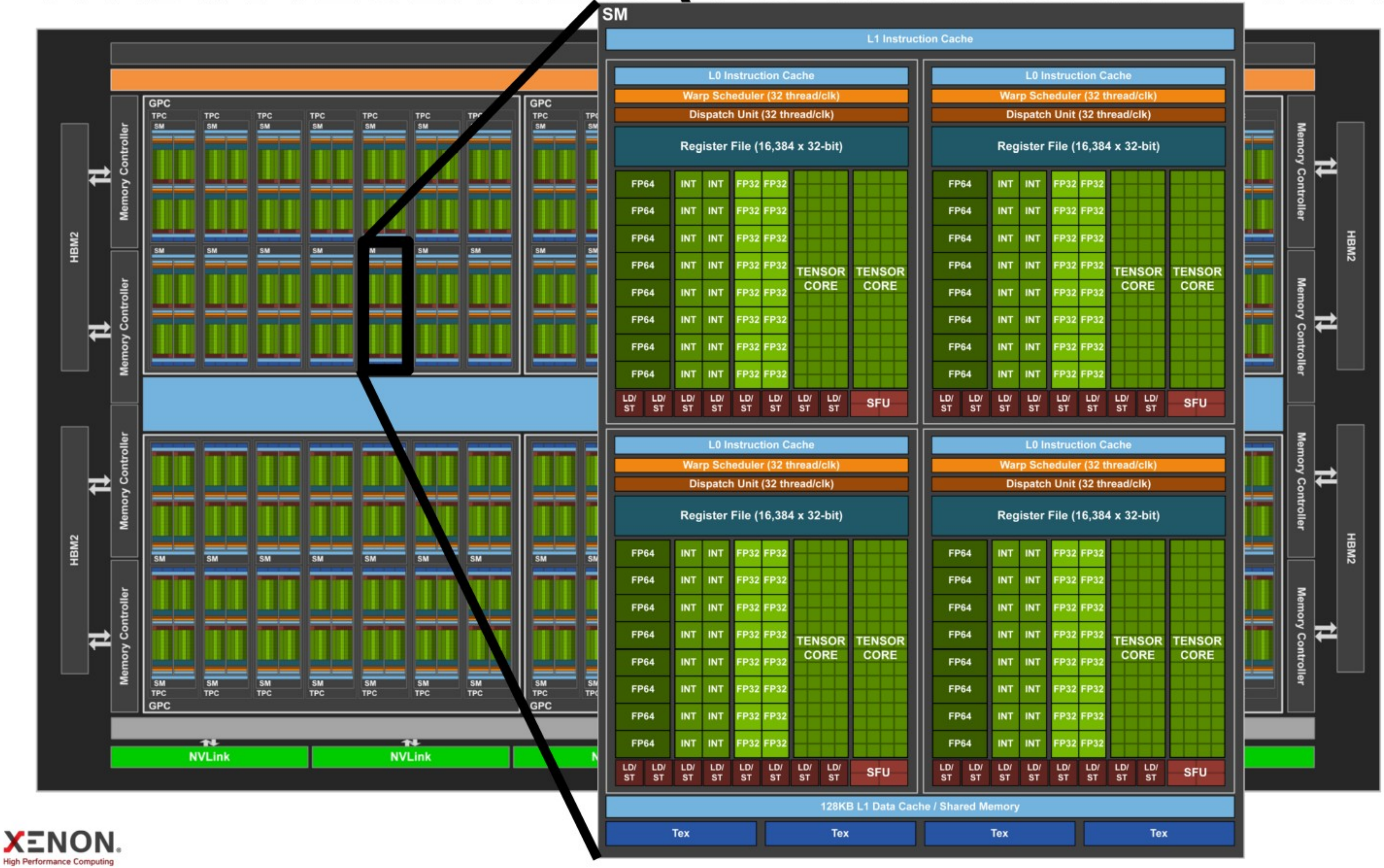
Ref.: https://www.nvidia.com/en-us/data-center/tesla-v100/

NVIDIA TESLA V100 (VOLTA ARCHITECTURE)

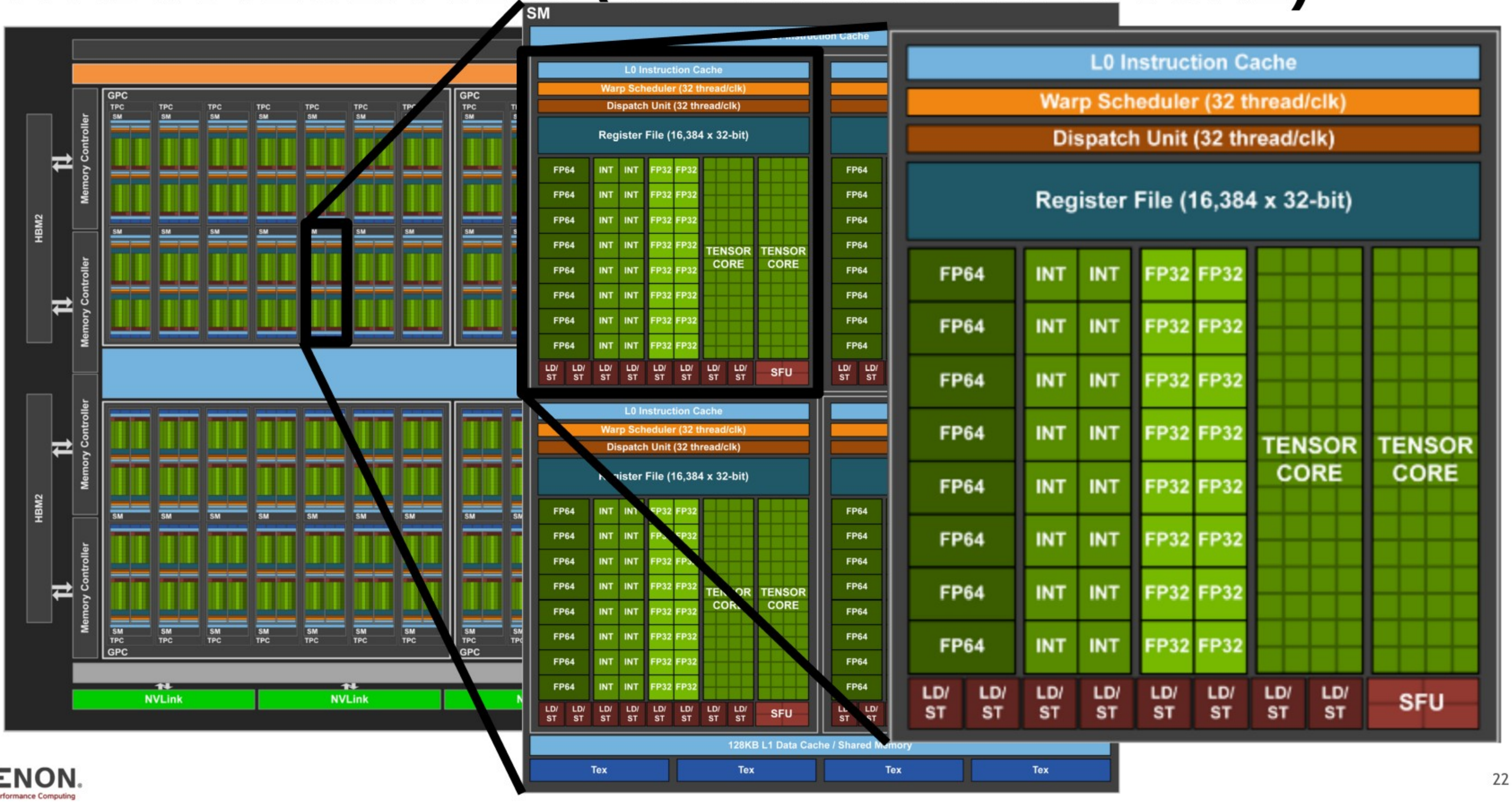


XENON. High Performance Computing

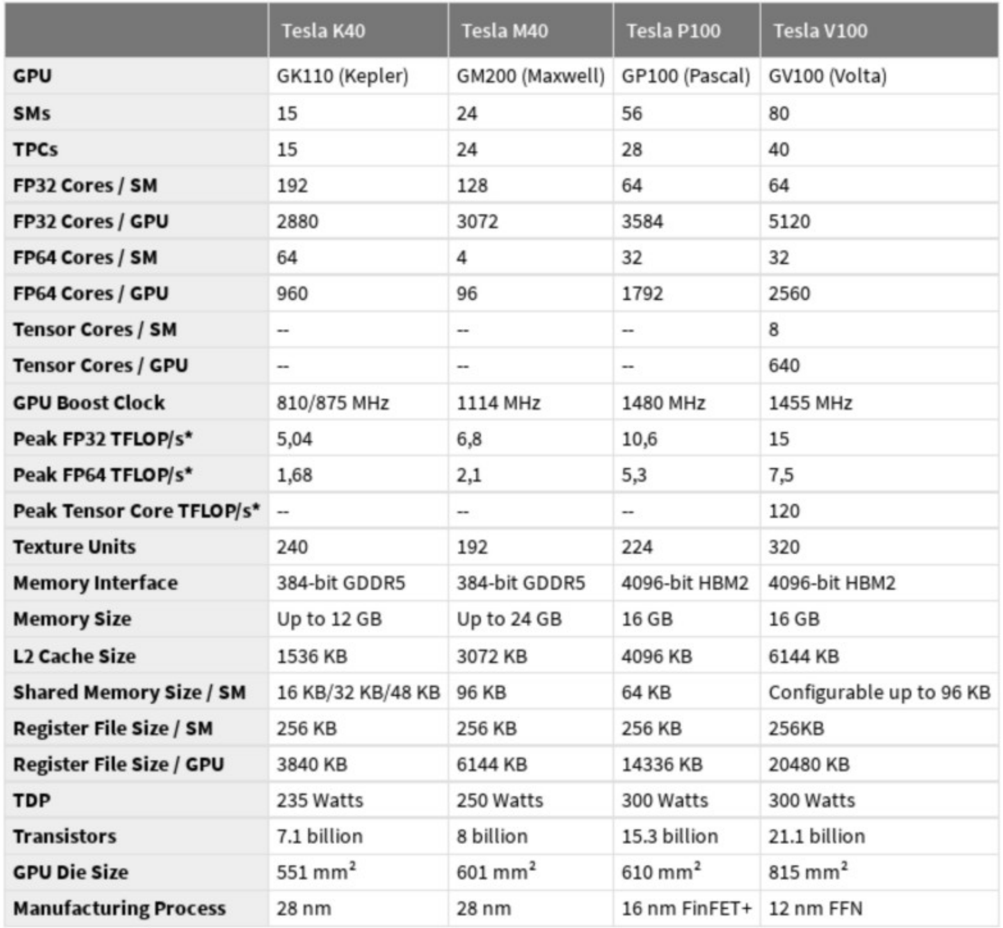
NVIDIA TESLA V100 (VOLTA ARCHITECTURE)


GPC
TPC
SM
Memory Controller
HBM2
NVLink
SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
FP64
INT
FP32
TENSOR CORE
LD/ST
SFU
128KB L1 Data Cache / Shared Memory
Tex

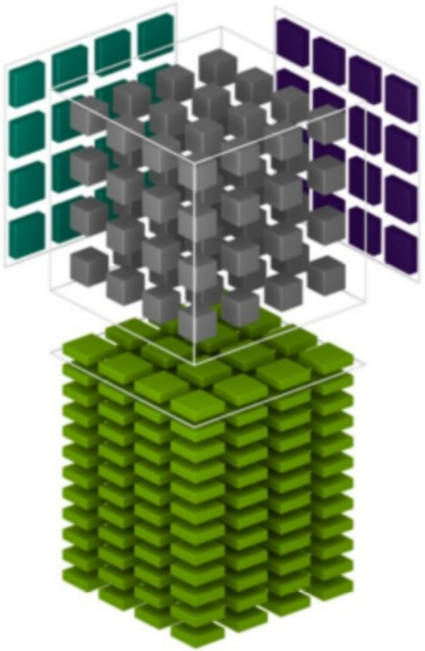
NVIDIA TESLA V100 (VOLTA ARCHITECTURE)

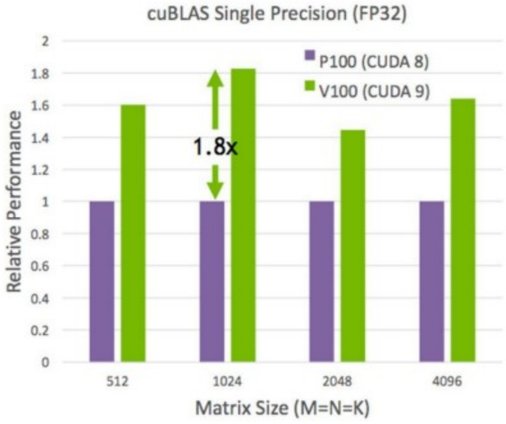


PERFORMANCE COMPARISON

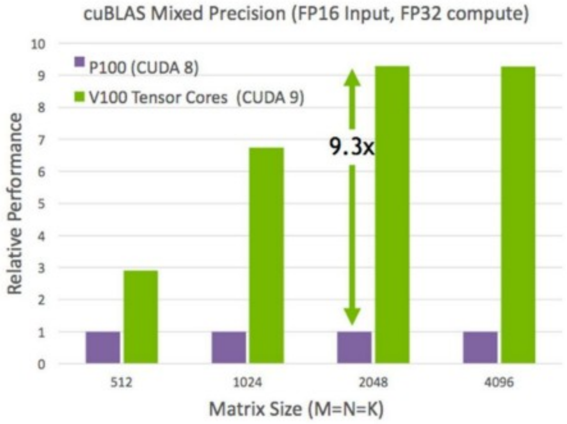


	Tesla K40	Tesla M40	Tesla P100	Tesla V100
GPU	GK110 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)	GV100 (Volta)
SMs	15	24	56	80
TPCs	15	24	28	40
FP32 Cores / SM	192	128	64	64
FP32 Cores / GPU	2880	3072	3584	5120
FP64 Cores / SM	64	4	32	32
FP64 Cores / GPU	960	96	1792	2560
Tensor Cores / SM	--	--	--	8
Tensor Cores / GPU	--	--	--	640
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz	1455 MHz
Peak FP32 TFLOP/s*	5,04	6,8	10,6	15
Peak FP64 TFLOP/s*	1,68	2,1	5,3	7,5
Peak Tensor Core TFLOP/s*	--	--	--	120
Texture Units	240	192	224	320
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2	4096-bit HBM2
Memory Size	Up to 12 GB	Up to 24 GB	16 GB	16 GB
L2 Cache Size	1536 KB	3072 KB	4096 KB	6144 KB
Shared Memory Size / SM	16 KB/32 KB/48 KB	96 KB	64 KB	Configurable up to 96 KB
Register File Size / SM	256 KB	256 KB	256 KB	256KB
Register File Size / GPU	3840 KB	6144 KB	14336 KB	20480 KB
TDP	235 Watts	250 Watts	300 Watts	300 Watts
Transistors	7.1 billion	8 billion	15.3 billion	21.1 billion
GPU Die Size	551 mm²	601 mm²	610 mm²	815 mm²
Manufacturing Process	28 nm	28 nm	16 nm FinFET+	12 nm FFN





cuBLAS Single Precision (FP32)



cuBLAS Mixed Precision (FP16 Input, FP32 compute)

Ref.: https://www.heise.de/newsticker/meldung/GTC-2017-Nvidias-Volta-Architektur-ein-Meilenstein-im-GPU-Design-3712206.html

IBM PowerAI Deep Learning Software Distribution

Deep Learning Frameworks

- Caffe
- NVCaffe
- IBMCaffe
- Torch
- TensorFlow
- Distributed TensorFlow
- Theano
- Chainer

Supporting Libraries

- OpenBLAS
- Bazel
- Distributed Communications
- NCCL
- DIGITS

Accelerated Servers and Infrastructure for Scaling

- Cluster of NVLink Servers
- Spectrum Scale: High-Speed Parallel File System
- Scale to Cloud

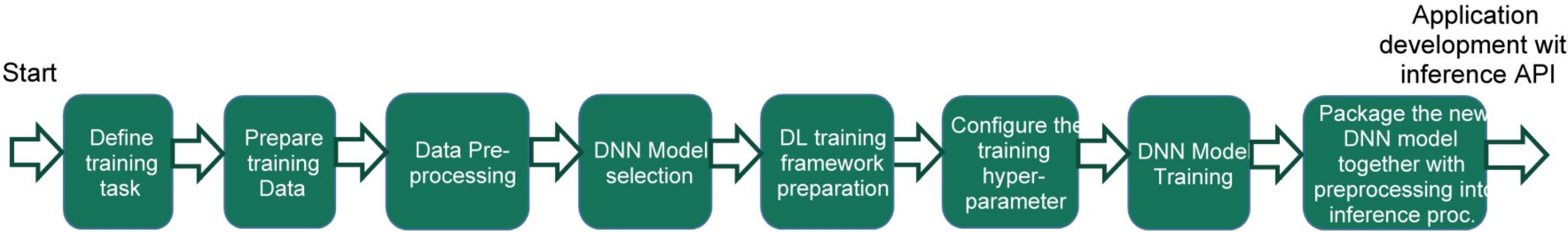
PowerAI: Making AI More Accessible to Developers

- AI Vision: Targeted at Application Developers
  - Custom application development tool aimed at Computer Vision workloads
- Data Extraction, Transformation and Preparation tool using Apache Spark
  - Powered IBM Spectrum Conductor with Spark
- DL Insight: Automated Model Tuning
  - Automatically tune hyper-parameters for models based on input data set using Spark-based distributed computing
  - Powerful and intuitive GUI—based developer tools that provide continuous feedback to quickly create and optimize deep learning models
- Distributed Deep Learning
  - HPC Cluster enabled distributed deep learning frameworks
  - Accelerated training with auto-distribution using Spark & HPC technology (TensorFlow & Caffe)

Multi-tenant, Enterprise-ready Deep Learning Platform for Data Scientists

Steps for Deep Learning Development

- Usually, developers need following steps to develop a DNN model and make it usable for applicati



Start

Define training task

Prepare training Data

Data Pre-processing

DNN Model selection

DL training framework preparation

Configure the training hyper-parameter

DNN Model Training

Package the new DNN model together with preprocessing int inference proc.

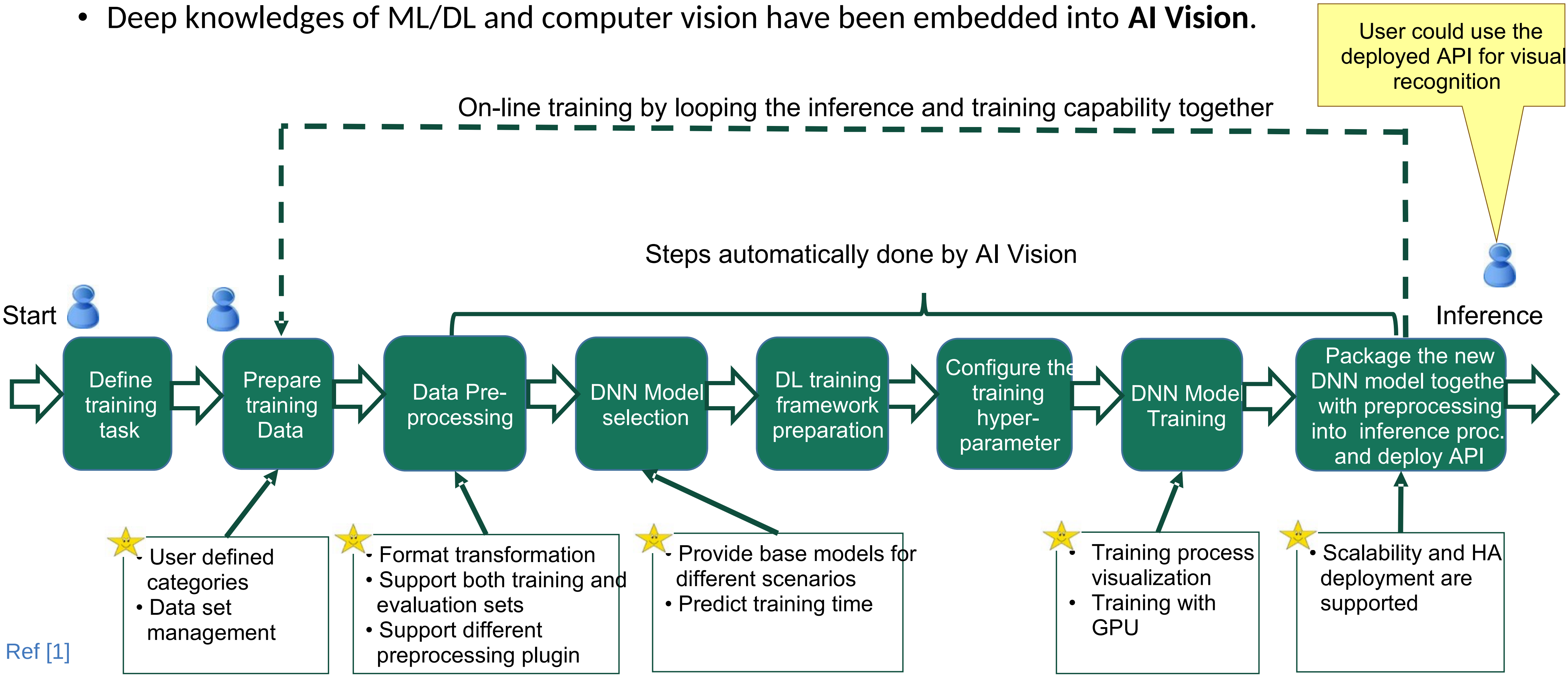
Application development with inference API

Most of enterprises are facing the challenges …

- No experience on DNN design and develop
- No experience on computer vision
- No experience on how to build a platform to support enterprise scale deep learning, including data preparation, training, and inference

AI Vision makes enterprise level DNN easier

- **AI Vision** automates the deep learning development cycles for developers.
- Deep knowledges of ML/DL and computer vision have been embedded into **AI Vision**.



User could use the deployed API for visual recognition

On-line training by looping the inference and training capability together

Steps automatically done by AI Vision

Start

Inference

Define training task

Prepare training Data

Data Pre-processing

DNN Model selection

DL training framework preparation

Configure the training hyper-parameter

DNN Model Training

Package the new DNN model togethe with preprocessing into inference proc. and deploy API

- User defined categories
- Data set management

- Format transformation
- Support both training and evaluation sets
- Support different preprocessing plugin

- Provide base models for different scenarios
- Predict training time

- Training process visualization
- Training with GPU

- Scalability and HA deployment are supported

Ref [1]

Image Classification example with **AI Vision**

I'm Aethopyga

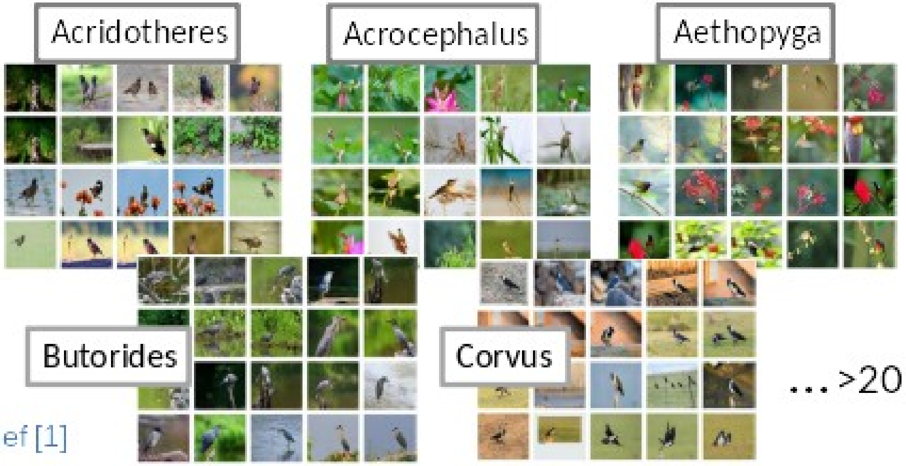
☹ Result on public cloud API :
white, red, yellow and teal bird

I'm Pycnonotus

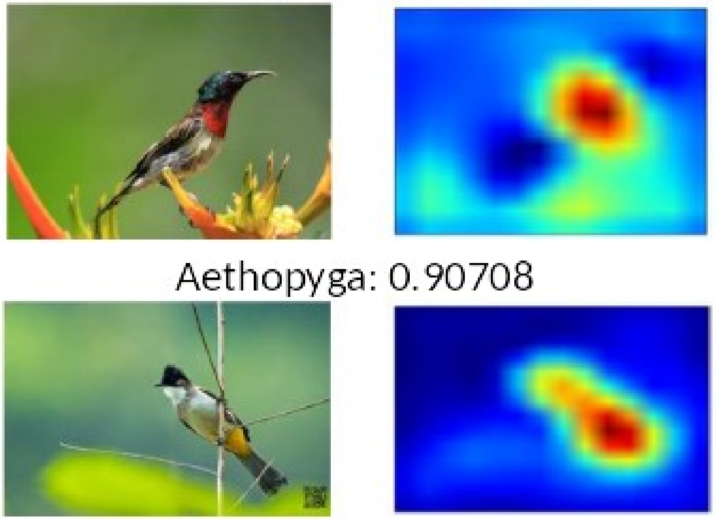
☹ Result on public cloud API :
white and black short beak bird

We need to get a new model to classify birds with professional knowledge.

User defines categories in AI Vision



…>20 categories

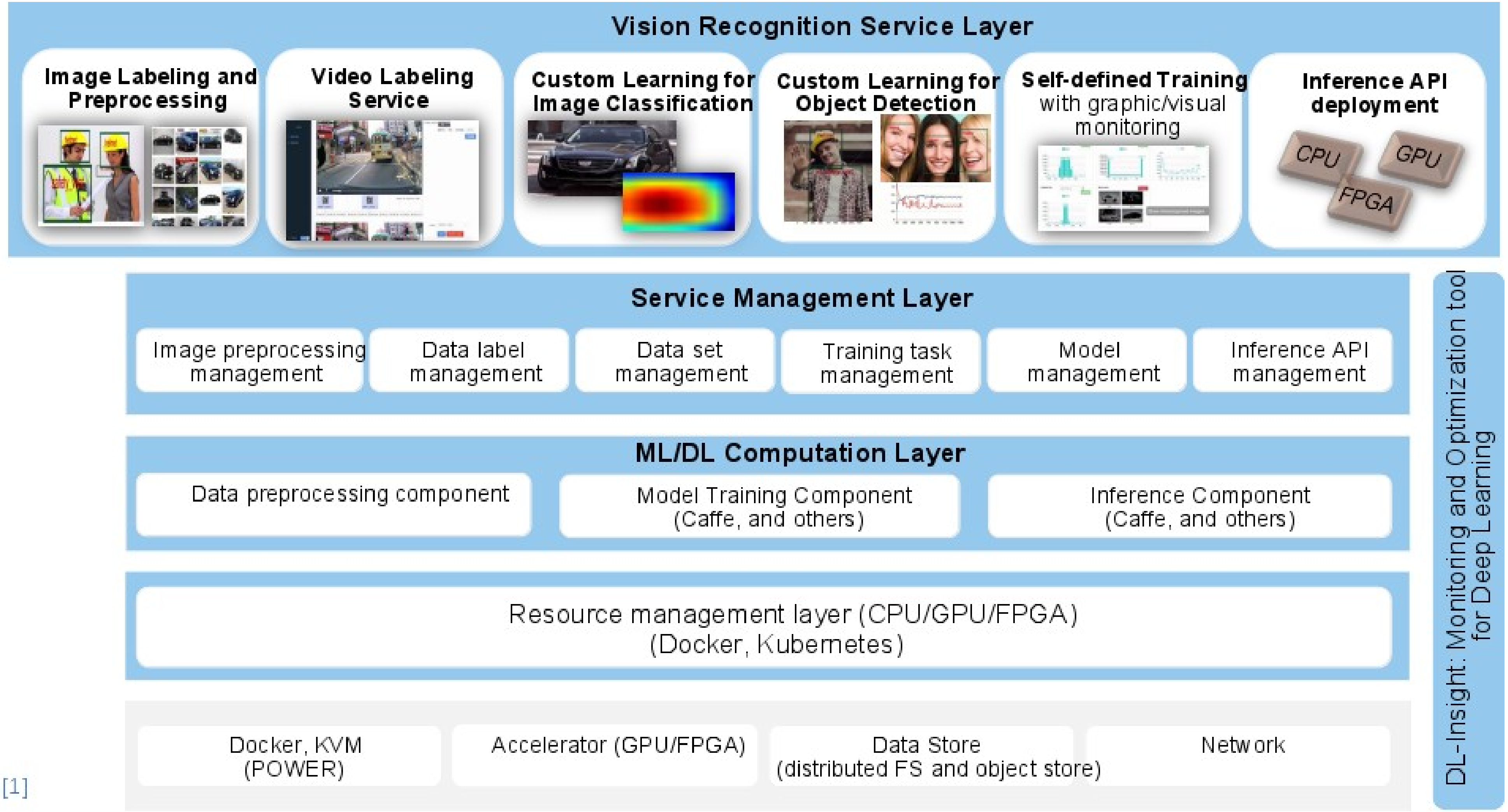


Aethopyga: 0.90708

Pycnonotus: 0. 99988

Ref [1]

AI Vision The Deep Learning Development Platform for image/video analysis



Vision Recognition Service Layer

Image Labeling and Preprocessing

Video Labeling Service

Custom Learning for Image Classification

Custom Learning for Object Detection

Self-defined Training with graphic/visual monitoring

Inference API deployment

Service Management Layer

Image preprocessing management

Data label management

Data set management

Training task management

Model management

Inference API management

ML/DL Computation Layer

Data preprocessing component

Model Training Component (Caffe, and others)

Inference Component (Caffe, and others)

Resource management layer (CPU/GPU/FPGA) (Docker, Kubernetes)

Docker, KVM (POWER)

Accelerator (GPU/FPGA)

Data Store (distributed FS and object store)

Network

DL-Insight: Monitoring and Optimization tool for Deep Learning

Ref [1]

CLOUD SOLUTIONS

Cloud

- CPU, GPU, FPGA instances
- IBM Bluemix
- HWaaS: IBM Softlayer
- DLaaS: Watson, “Tensorflow”aaS

Challenges

- Data locality
- Data sovereignty/privacy
- Network bandwidth
- Scaling performance
- GPU performance
- Software stack
- Cost

INFERENCE – USING DL MODELS

Deployment models

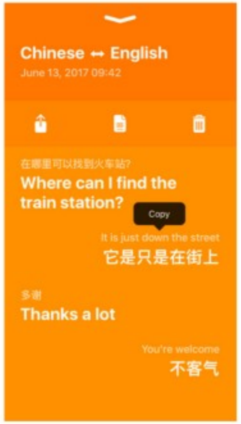
- Small, low power device on the edge
  e.g. mobile phone, CCTV camera, sensor, etc.

Cloud

- Device network connected
- “Phoning home”: Transfer data to server
- Run data through network
- Analyze result and make decisions
- Send result/action back to device

Examples

- Translation:
  e.g. iTranslate Converse



Chinese ↔ English
June 13, 2017 09:42

在哪里可以找到火车站?
Where can I find the train station?

It is just down the street
它是只是在街上

多谢
Thanks a lot

You're welcome
不客气

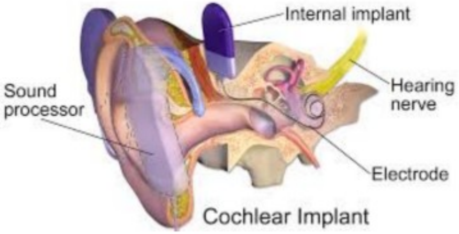
Embedded solutions

- Offload inference to edge device itself
- Required for off-line devices
- Faster response (avoids network latency)
- Sufficiently fast hardware required

Examples

- Autonomous cars
- In-phone Translator
- In-ear translator: e.g. Mymanu CLIK
- In-camera processing
- Cochlear implants: Machine Learning: “manual” implementation

Mymanu CLIK™



SUMMARY

Workloads

- Dev and Test
- Training
- Inference

Technologies

- CPU
- GPU
- GPUs for DL (Tensor Cores), single prec., half prec.
- FPGA
- ASICS: TPU, etc.

On-premise

- GPU servers:
  IBM 822SL: Power8 + P100 + NVLINK
- PowerAI

Cloud

- CPU, GPU, FPGA instances
- HWaaS: Softlayer
- DLaaS: Watson, “Tensorflow”aaS

New Services

- Aivision
- DLInsight

Future

- CPU, GPU, FPGA instances
- Power9 + V100 + NVLINK2

Getting Started with IBM PowerAI

- Visit the IBM Systems booth at the Tech Symposium to see a demo of IBM PowerAI Vision

- Download and install PowerAI for free on your existing S822LC for HPC server : http://ibm.biz/powerai
- Don’t have an S822LC for HPC?
  - POC/Test - 2 x IBM Minsky’s Available for Testing @ IBM Sydney Labs

- Videos to get started
  - Build a image classifier
    - http://www.youtube.com/watch?v=qHZRnswzqUI
  - Train models to analyze videos for Advanced Driver Assistant System
    - http://www.youtube.com/watch?v=beL9GTi9jjs
- Sample datasets
  - Download sample dataset for classifying breeds of dogs from stanford.edu

    http://vision.stanford.edu/aditya86/ImageNetDogs/images.tar

Thank you!

Werner Scholz, 15 Aug. 2017
XENON Systems, CTO and Head of R&D
werners@xenon.com.au

XEN
High Performanc
www.xeno