GPU ACCELERATED COMPUTING & DEEP LEARNING

- A Performance and Power Analysis

By Dipti Chaudhari

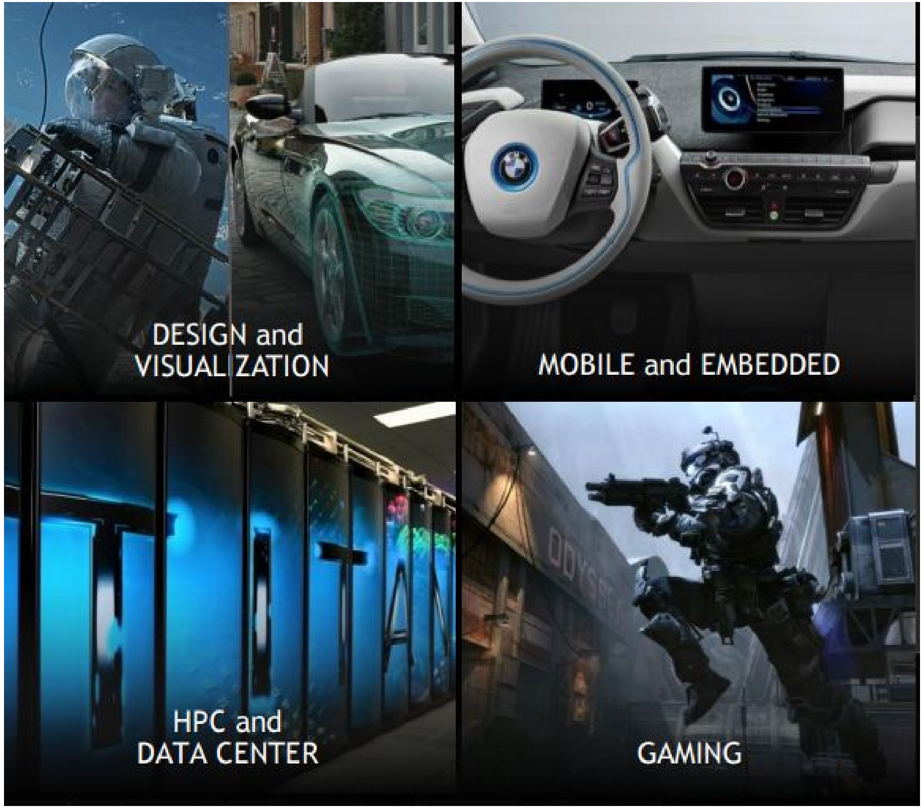
Agenda

- GPUs Past and Present
- What is Deep Learning?
- GPUs and DL
- DL in practice
- Scaling up DL
- Performance Analysis

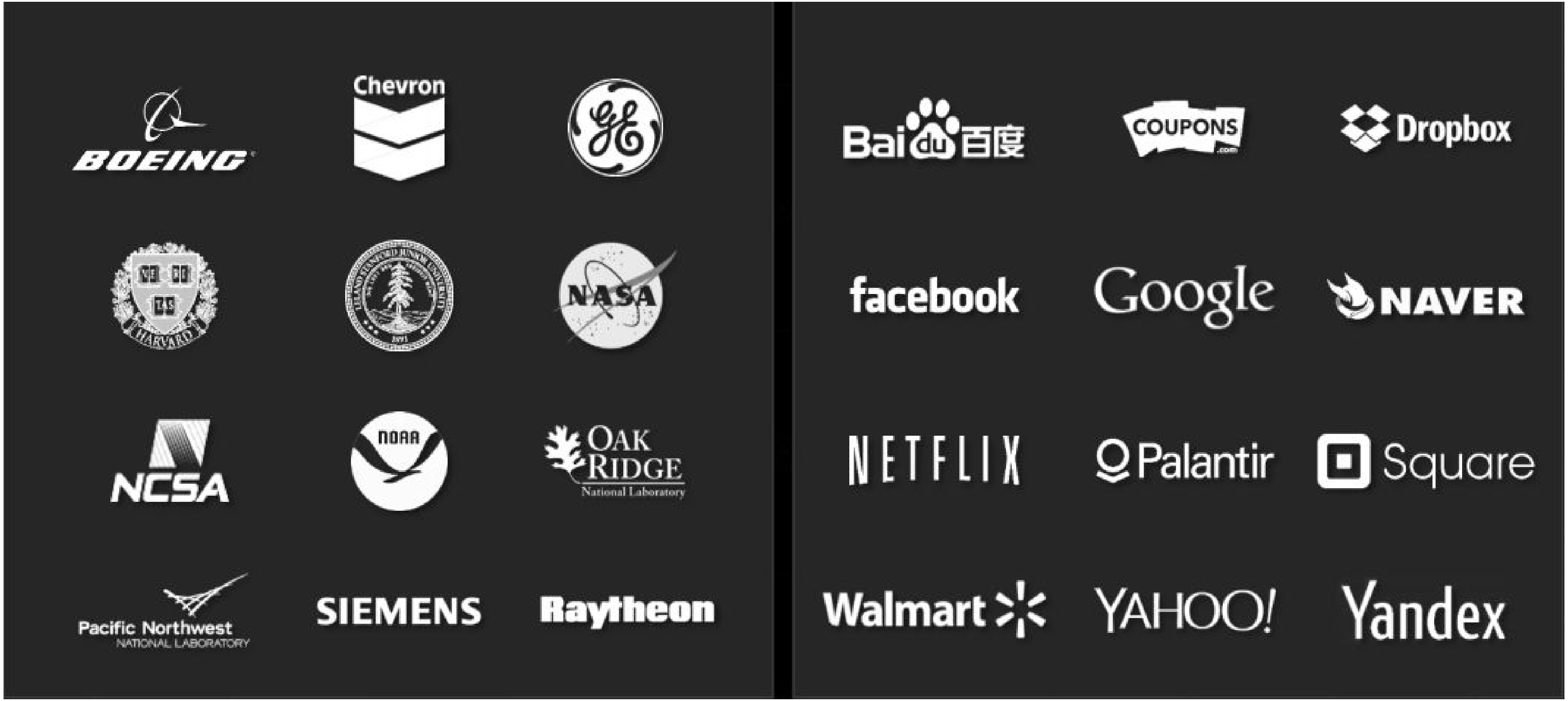
NVIDIA - INVENTOR OF THE GPU

- NVIDIA Invented the GPU in 1999
- Graphics support
- In 2007, NVIDIA launched the CUDA® programming platform
- Applications Areas

NVIDIA Platform



BEYOND HPC TO BIG DATA ANALYTICS



What is Deep Learning?

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

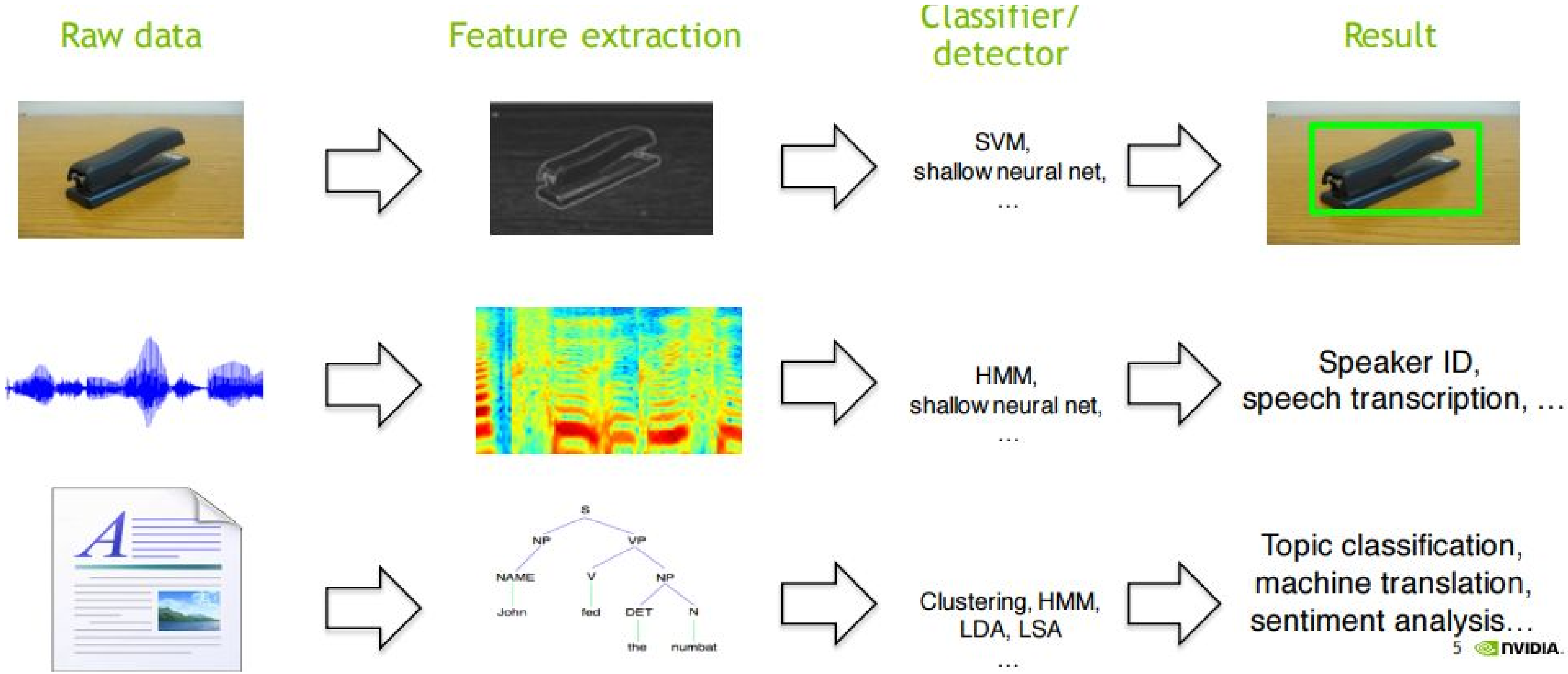
SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

AUTONOMOUS MACHINES

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

Traditional Machine Learning



Raw data

Feature extraction

Classifier/ detector

Result

SVM, shallow neural net, …

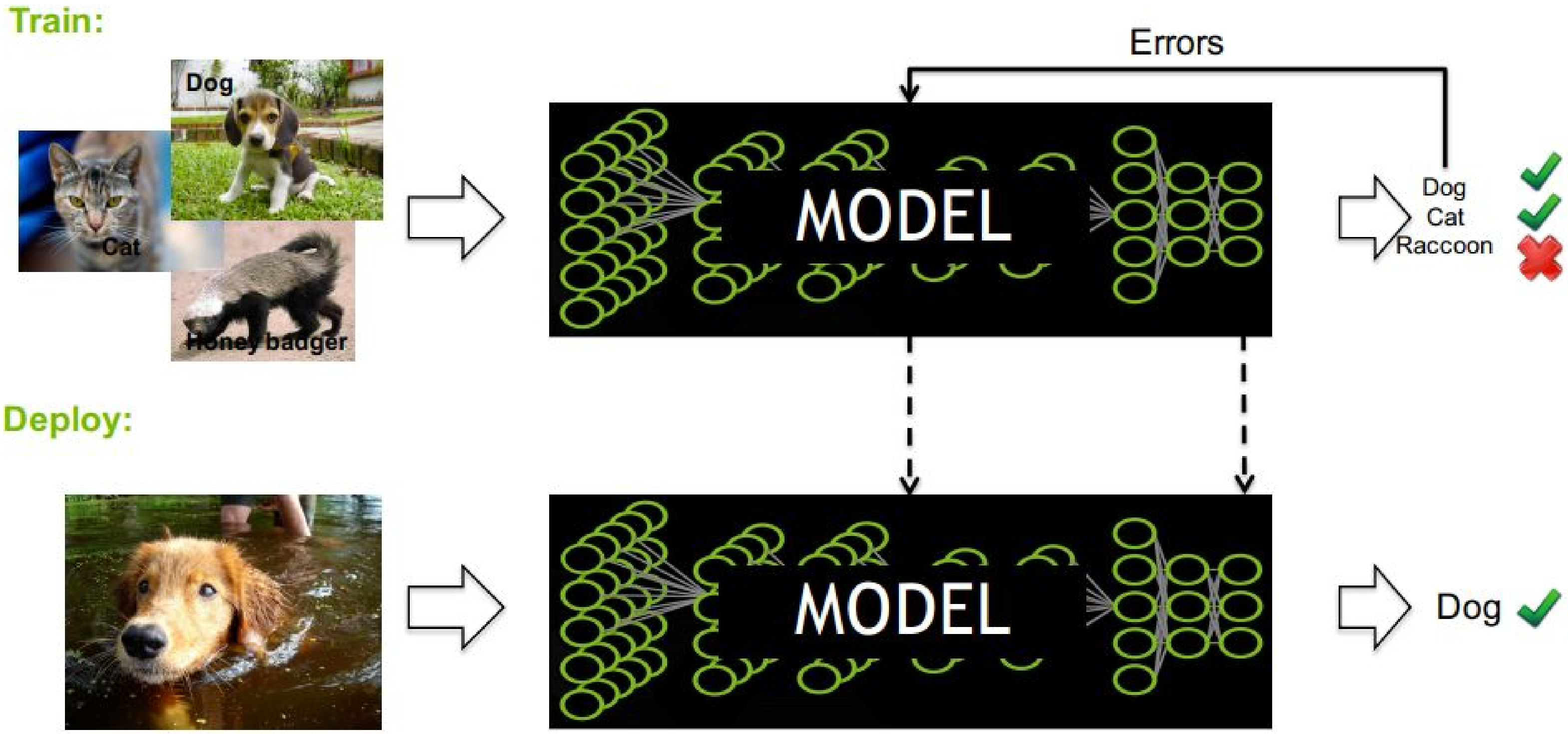
HMM, shallow neural net, …

Speaker ID, speech transcription, …

Clustering, HMM, LDA, LSA …

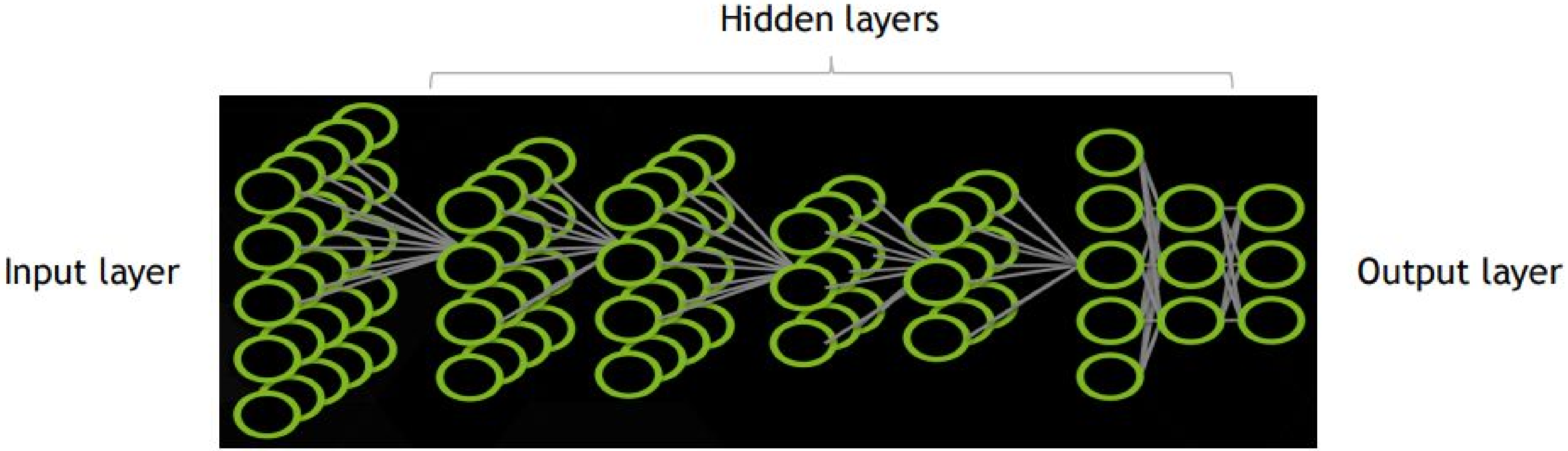
Topic classification, machine translation, sentiment analysis…

Deep learning approach



Artificial neural network

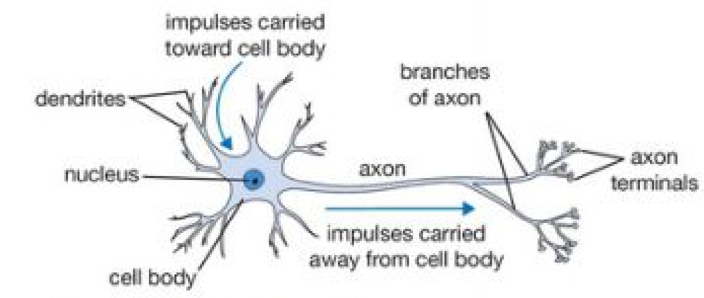
A collection of simple, trainable mathematical units that collectively learn complex functions



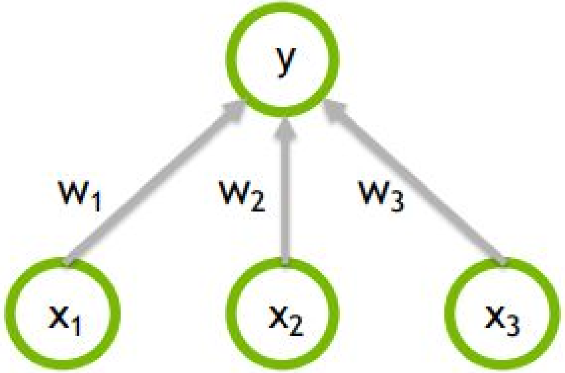
Given sufficient training data an artificial neural network can approximate very complex functions mapping raw data to output decisions

Artificial neurons

Biological neuron Vs Artificial neuron


impulses carried toward cell body
dendrites
branches of axon
nucleus
axon
axon terminals
impulses carried away from cell body
cell body

From Stanford cs231n lecture notes



$$y=F(w_1x_1+w_2x_2+w_3x_3)$$

$$F(x)=\max(0,x)$$

Deep neural network (dnn)

Identify face

Training data

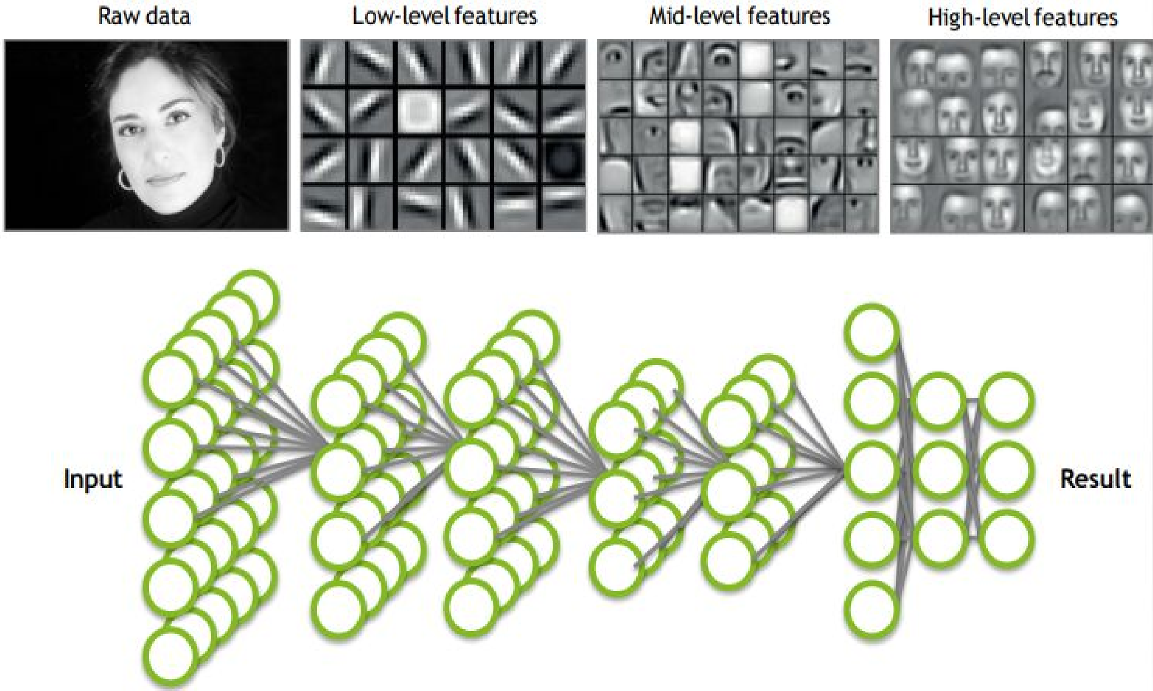
~10-100M images

Network architecture

~10 layers 1B parameter

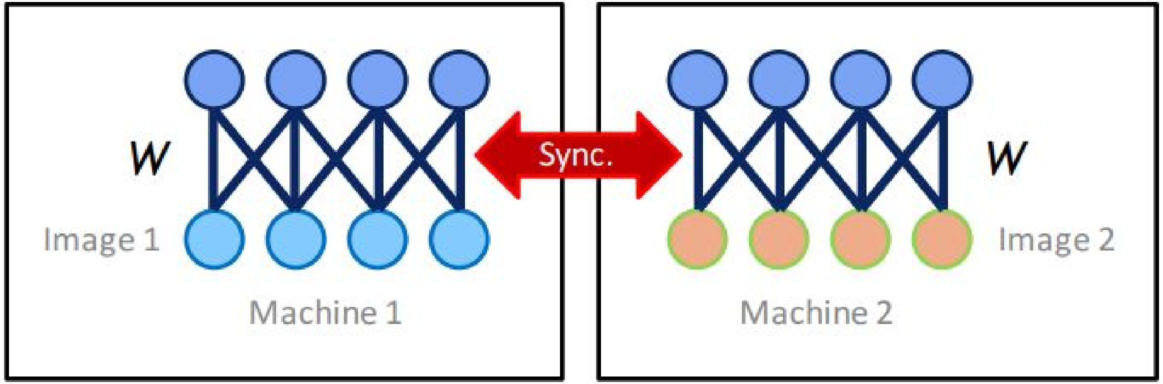
Learning algorithm

~30 Exaflops~30 GPU day

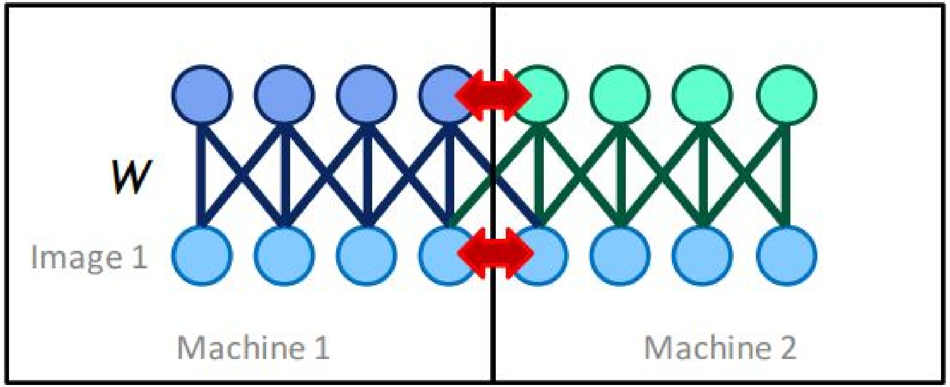


Scaling Deep Learning

- Data Parallelism



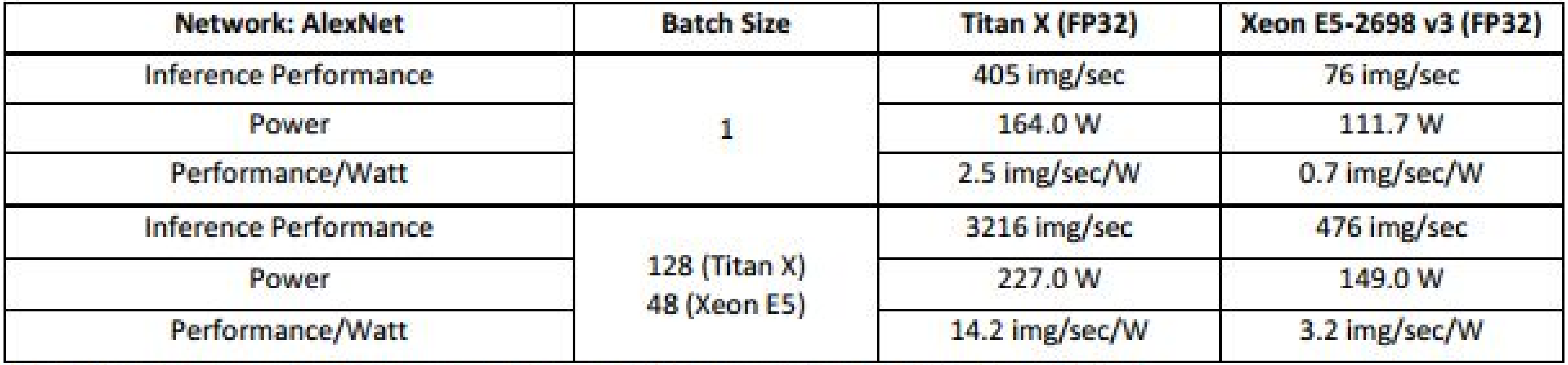
- Model Parallelism



PerFormance Analysis

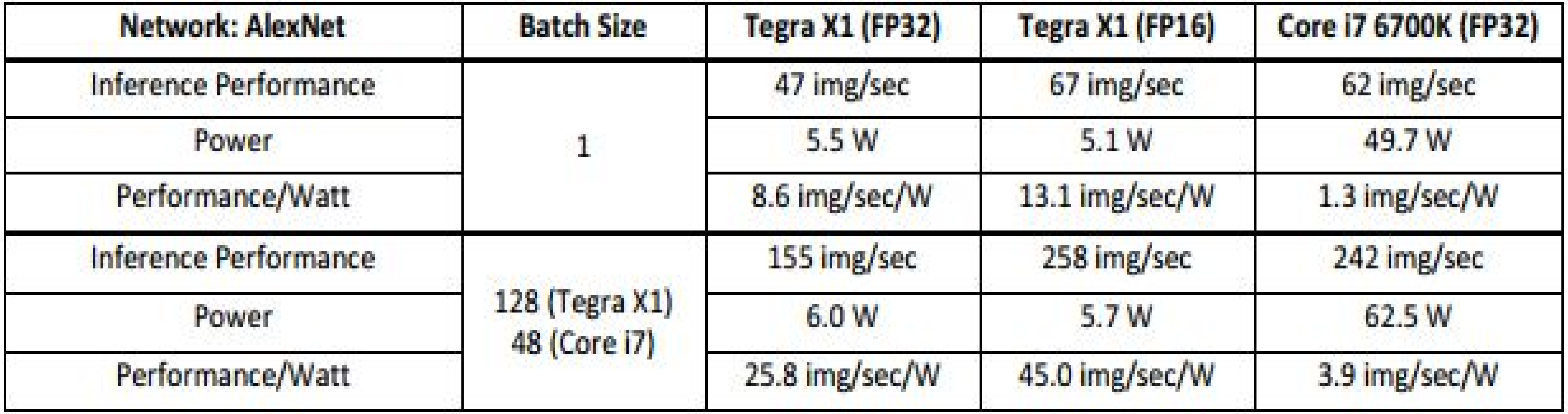
- Image Processing
- Caffe - Framework
- CuDNN - Library
- Intel Deep Learning Framework (IDLF) - Intel HW
- Small Power and High Power
- Batching and Non-batching

Power and Performance analysis (Powerful Infrastructure)



Network: AlexNet	Batch Size	Titan X (FP32)	Xeon E5-2698 v3 (FP32)
Inference Performance	1	405 img/sec	76 img/sec
Power		164.0 W	111.7 W
Performance/Watt		2.5 img/sec/W	0.7 img/sec/W
Inference Performance	128 (Titan X) 48 (Xeon E5)	3216 img/sec	476 img/sec
Power		227.0 W	149.0 W
Performance/Watt		14.2 img/sec/W	3.2 img/sec/W

Power and Performance analysis (less Power Infrastructure)



Network: AlexNet	Batch Size	Tegra X1 (FP32)	Tegra X1 (FP16)	Core i7 6700K (FP32)
Inference Performance	1	47 img/sec	67 img/sec	62 img/sec
Power		5.5 W	5.1 W	49.7 W
Performance/Watt		8.6 img/sec/W	13.1 img/sec/W	1.3 img/sec/W
Inference Performance	128 (Tegra X1) 48 (Core i7)	155 img/sec	258 img/sec	242 img/sec
Power		6.0 W	5.7 W	62.5 W
Performance/Watt		25.8 img/sec/W	45.0 img/sec/W	3.9 img/sec/W

?

Thank You For Listening :)