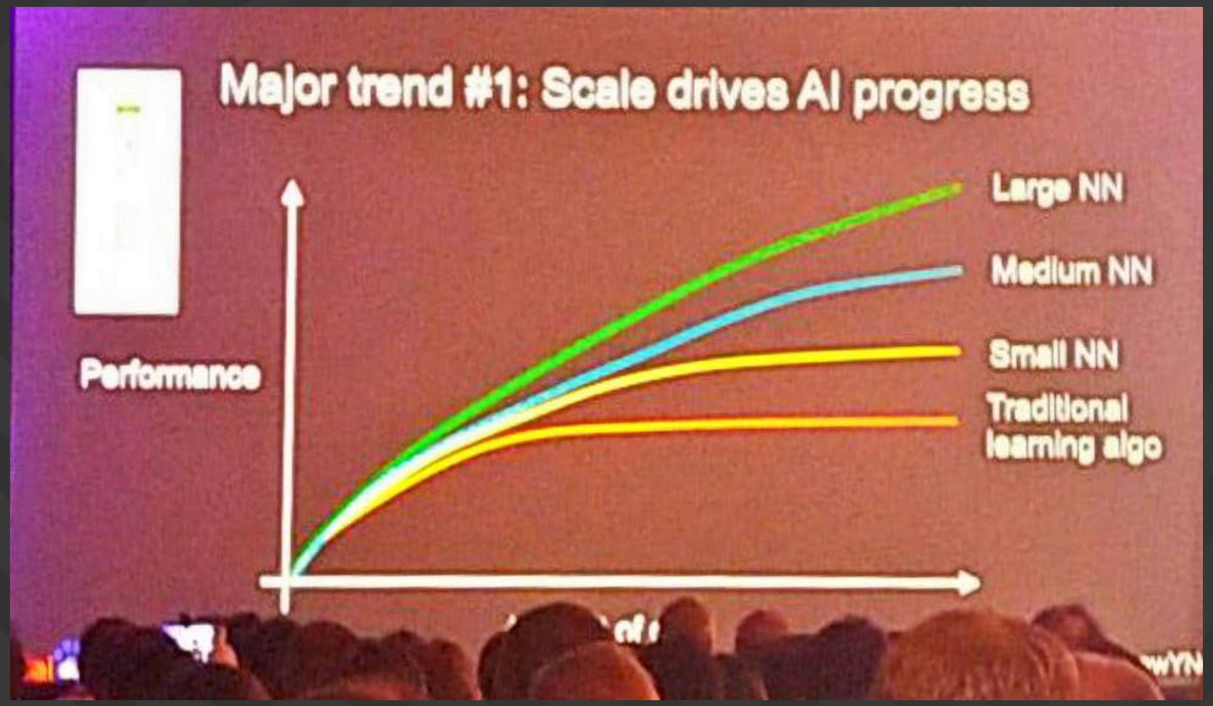
GPUs & Deep Learning at scale

Robert Ober
Tesla Chief Platform Architect, NVIDIA

Exponential Performance Needed

Andrew Ng @ Baidu: DeepSpeech2 snapshot



deeper + larger

90% → 99% recognition

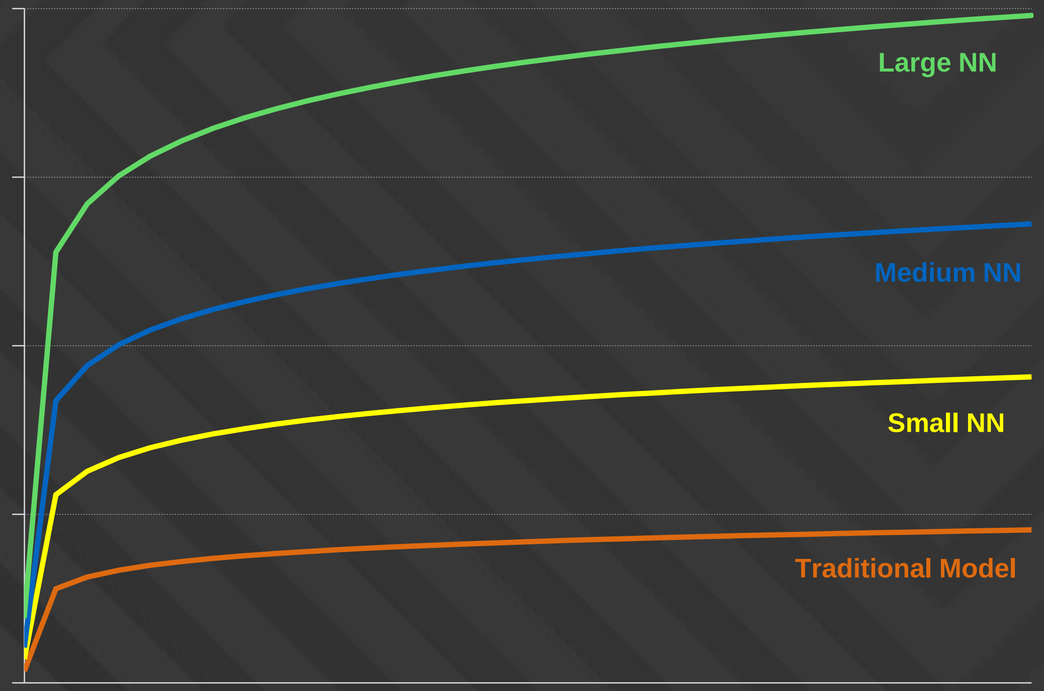
10’s of ExaFlop (10^18) / cycle

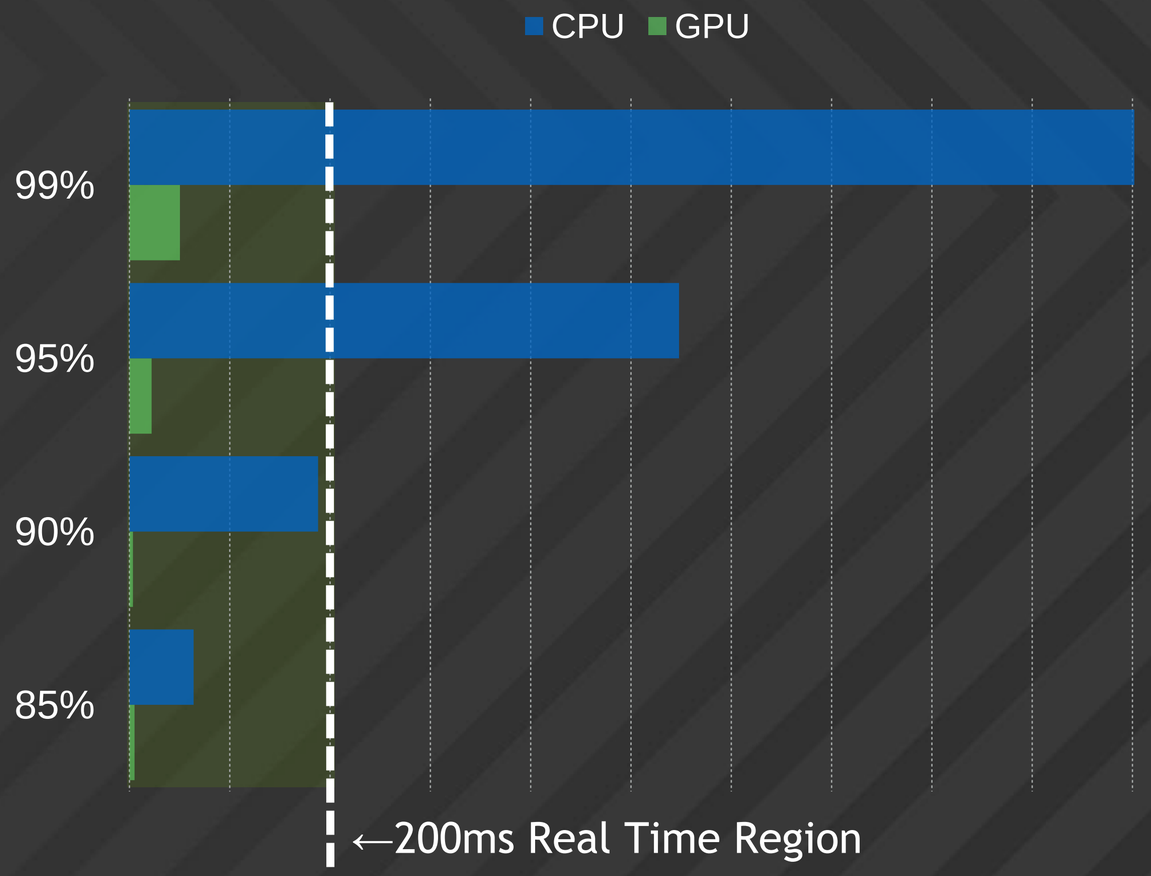
~4TByte data / cycle

training:

- 8 hours = ~350 TFlop/s, 142MB/s
- 2 hours = ~1.5 PFlop/s, 568 MB/s

GPUs : Accurate results in real time

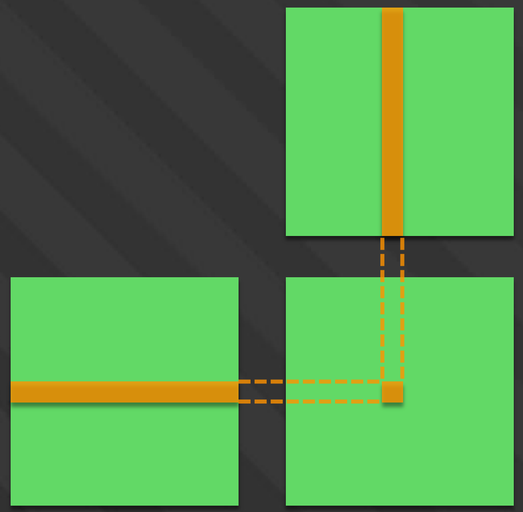




GPU 101

Matrix Math

Basis of Graphics



Transforms

Every:

- Pixel
- Shader
- over and over…

Parallel Threads, SIMD

GPU Architecture

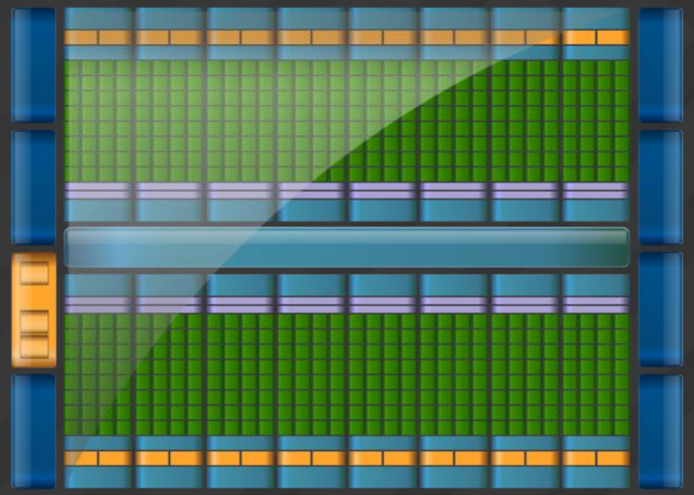
Two Main Components

Global memory: shared with host

Streaming Multiprocessors (SM)

control units, registers,

execution pipelines, caches



GPU Architecture

Streaming Multiprocessor (SM)

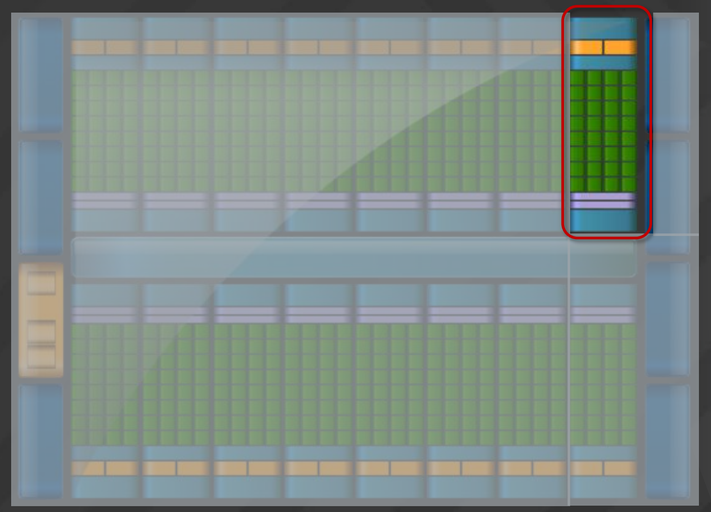
Many CUDA Cores per SM

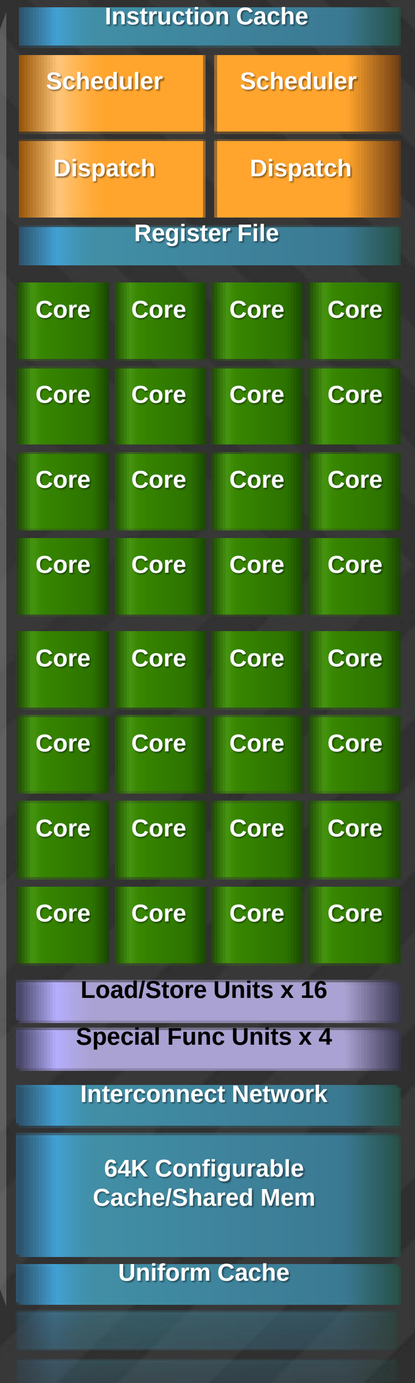
Special-function units

cos/sin/tan, etc.

Shared mem + L1 cache

Thousands of 32-bit registers





GPU Architecture

CUDA Core

Floating point & Integer unit

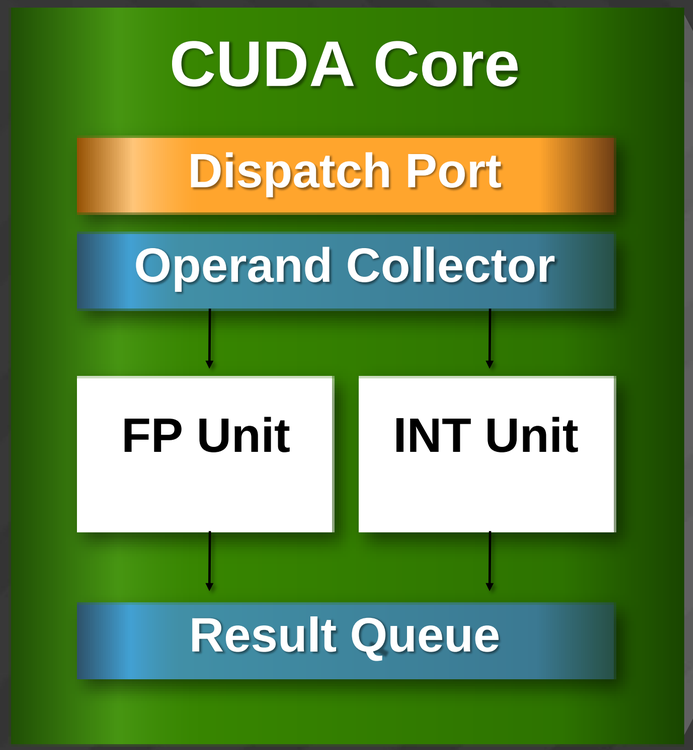
Fused multiply-add (FMA) instruction

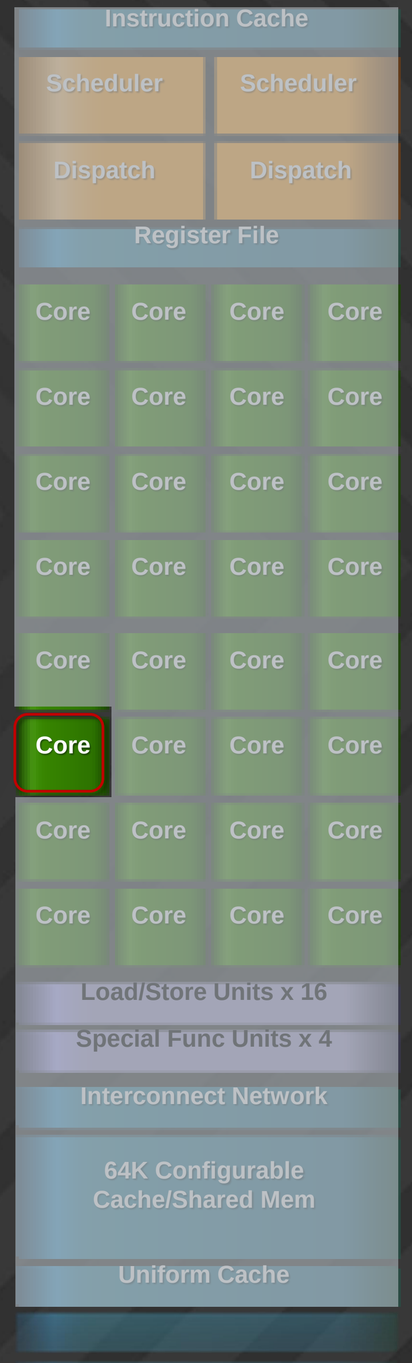
deep learning instructions

Logic unit

Move, compare unit

Branch unit

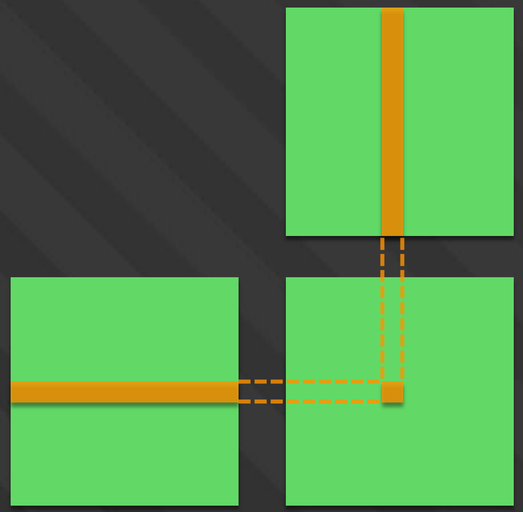


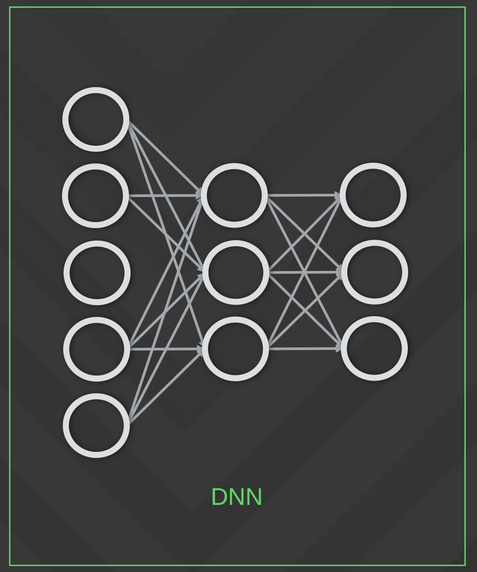


For DL

Matrix Math

Basis of Deep Learning





Convolution:

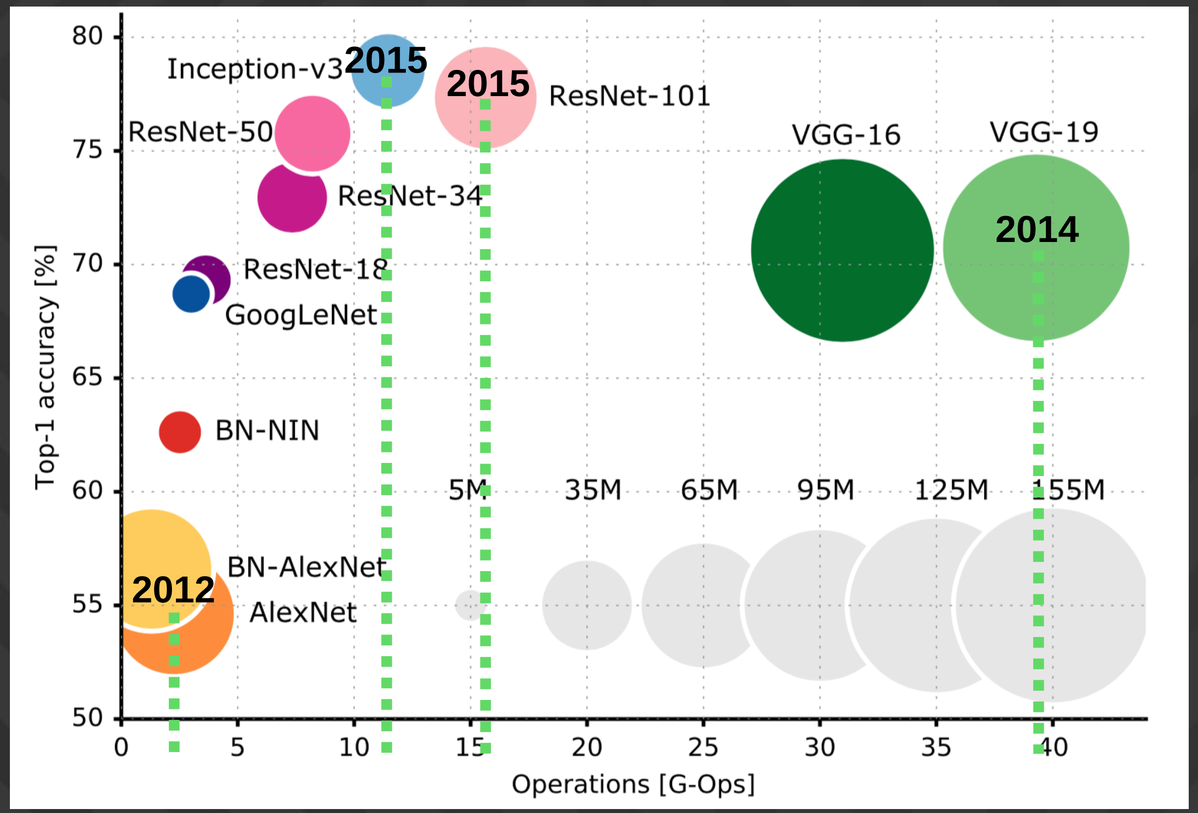
155M Parameters

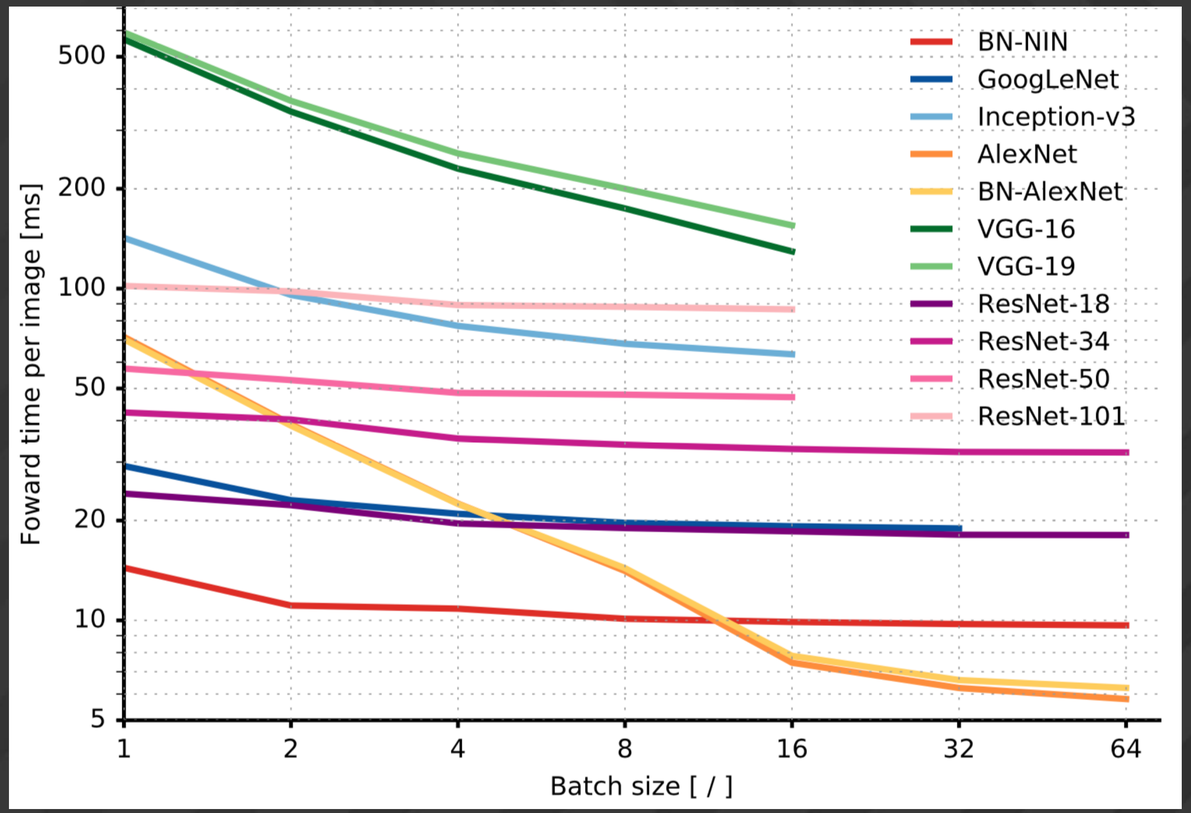
- Neurons
- Layers
- Iterations…

40 GOPs / Inference

Inference

Computational Load and Latency

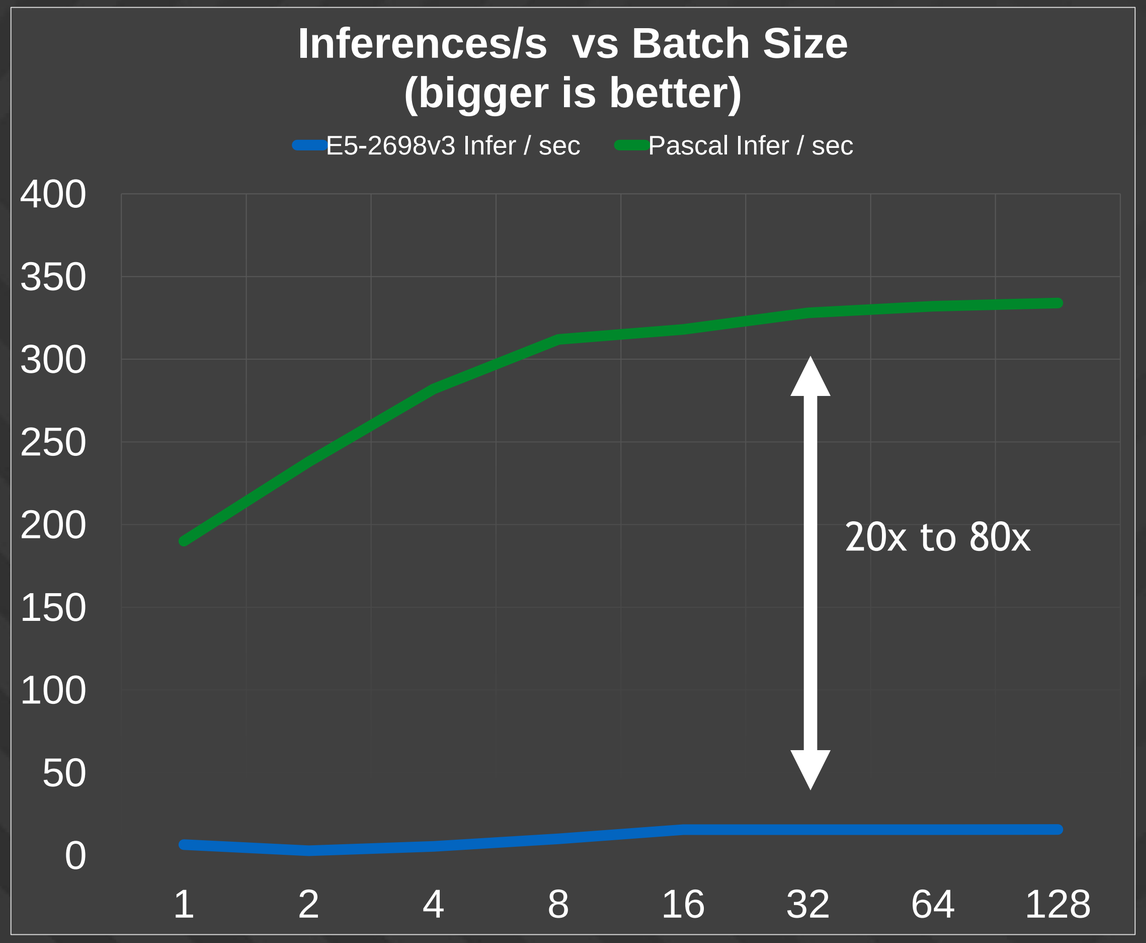


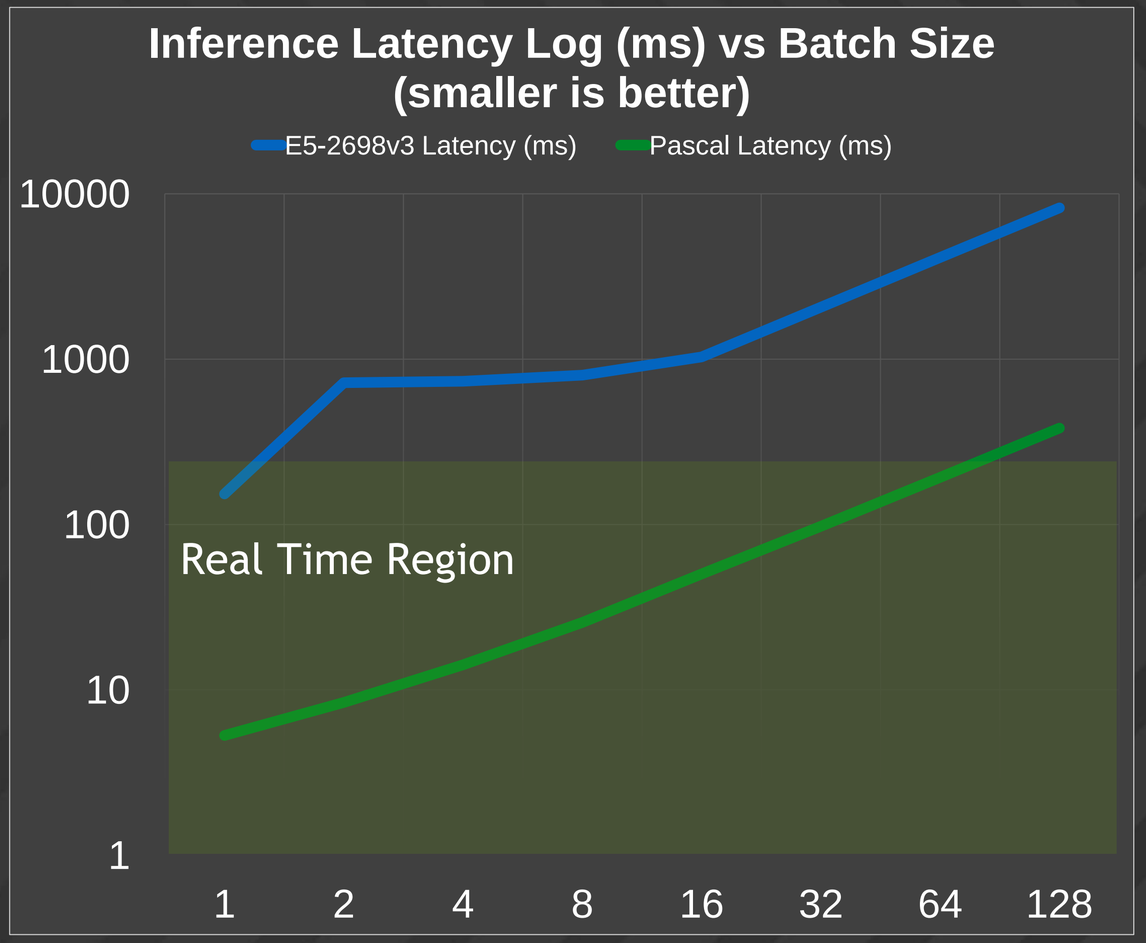


An Analysis of Deep Neural Network Models for Practical Applications
Canziani, Culurciello, Paszke, May 2016

Throughput & Latency

25 : 1 in real time inference FP32



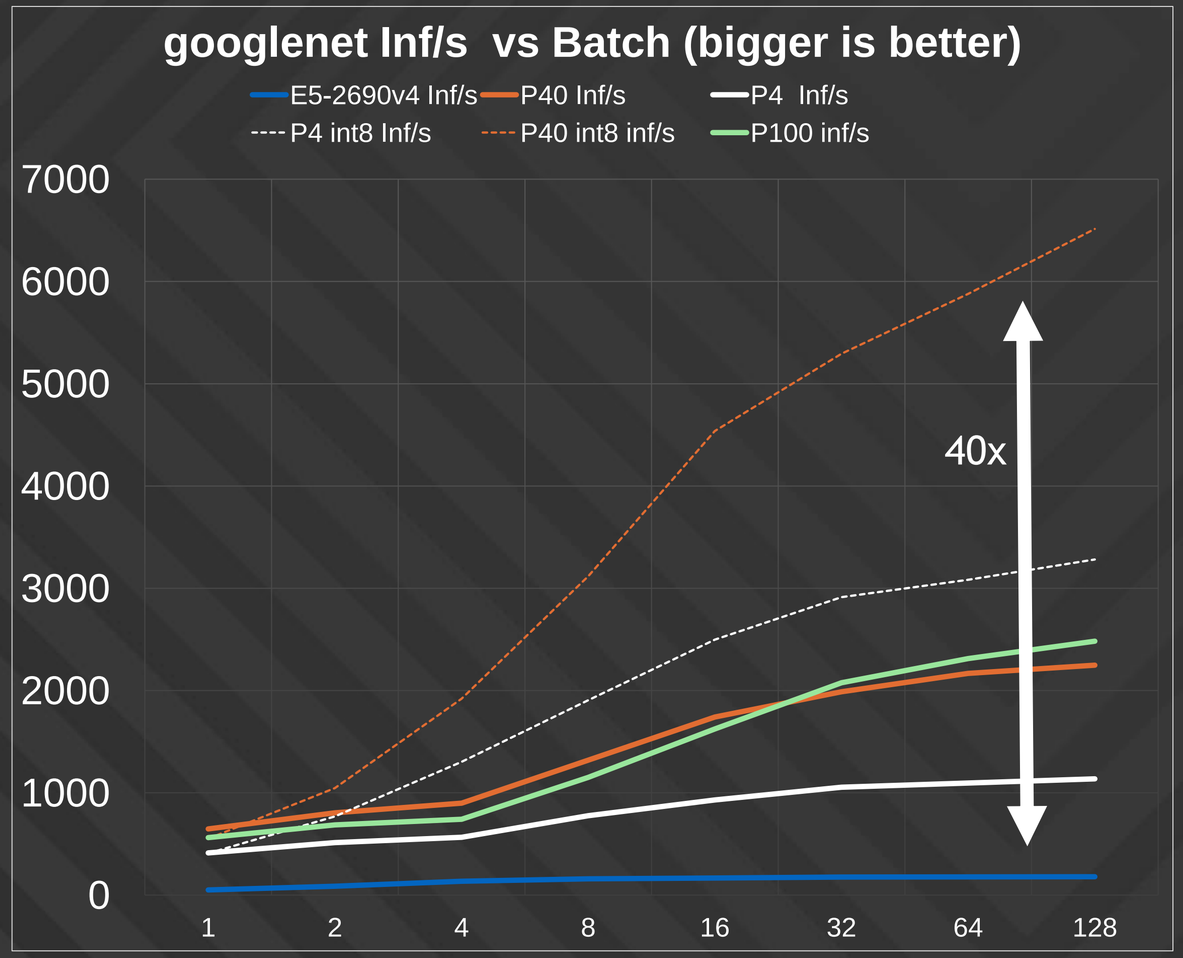


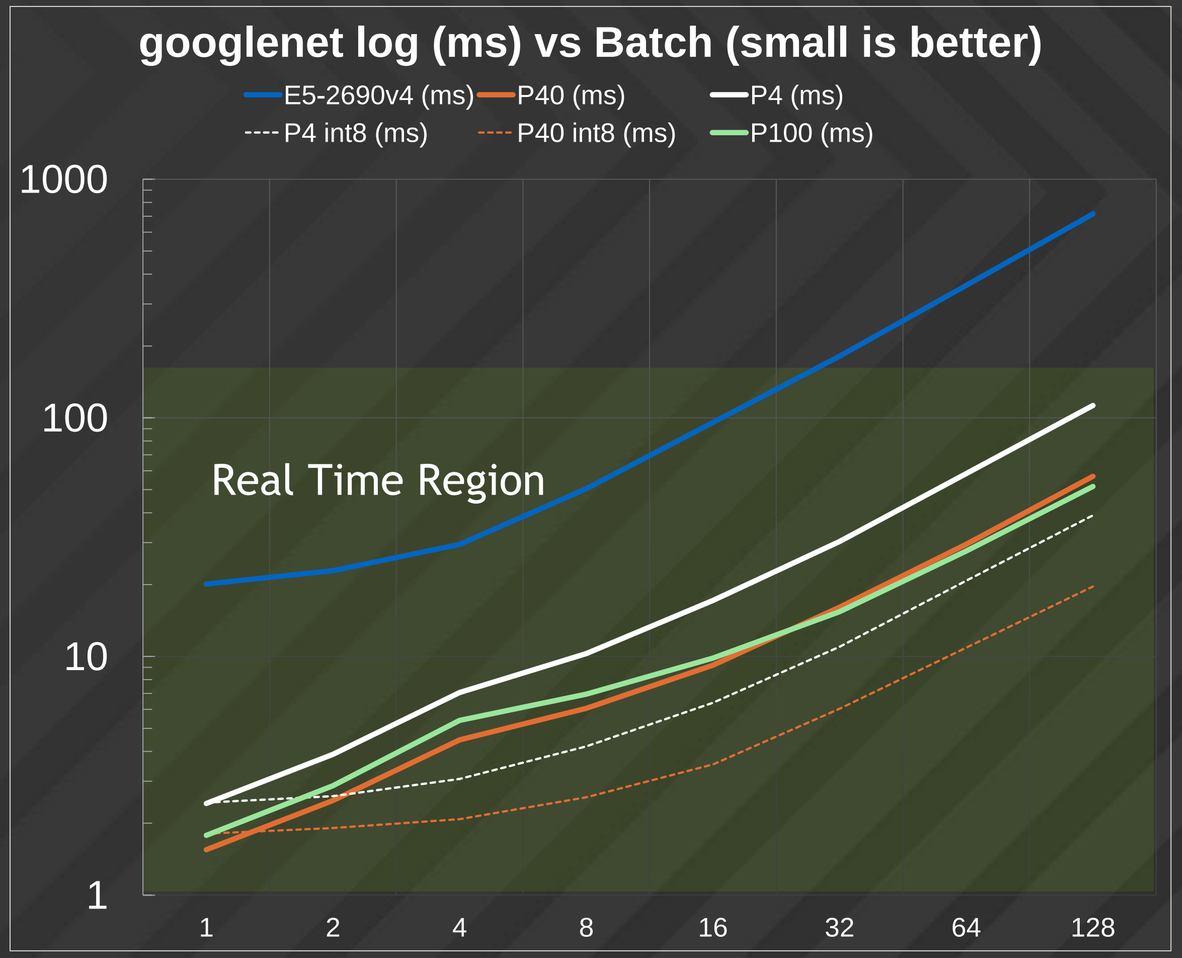
VGG19 Inference

1x E5-2698v3 FP32, IntelCaffe + MKL 2017 vs 1x Pascal FP32, GIE

Throughput & Latency

Updated: googLenet FP32, INT8



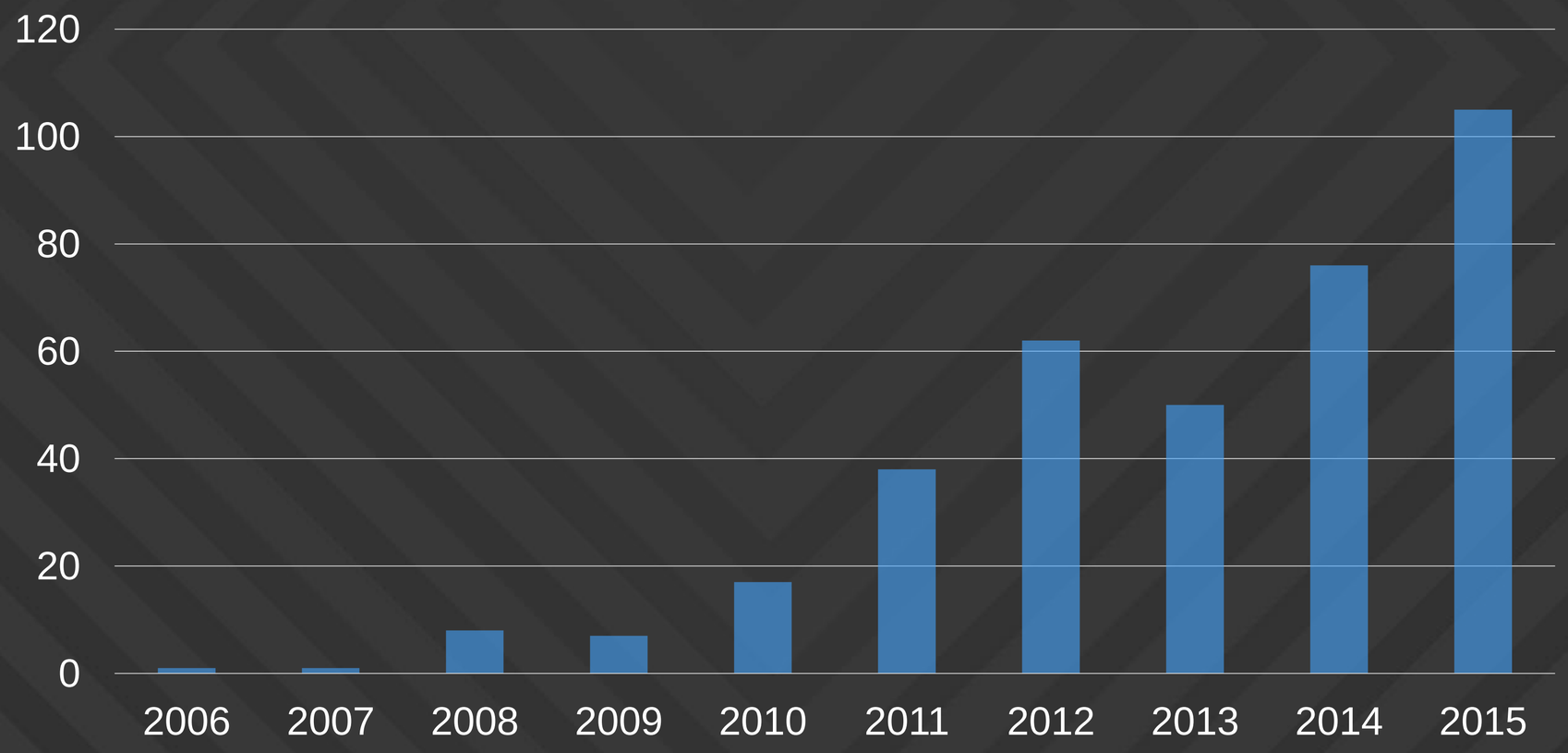


GoogLeNet
E5-2690v4 12 core, @2.6GHz, 3.5GHz turbo, HT on, MKL 2017 gold release
TensorRT

HPC Scale

HPC Top 500

HPC Accelerated Clusters In the top 500



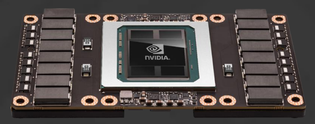
“Accelerators Will Be Installed in More than Half of New Systems”

Source: Top500.org

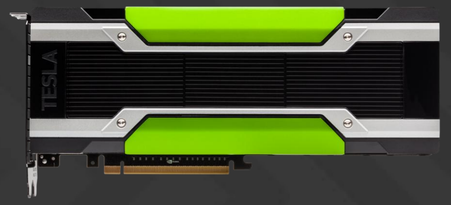
Source: Top 6 predictions for HPC in 2015

Data center GPU

Specialized Hardware
Updated



P100 SXM2
+ DGX-1



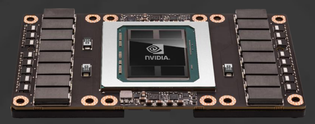
P100 + P40

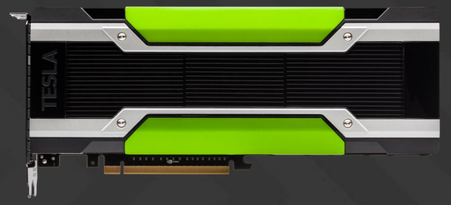
P4

Form Factor	Custom SXM2	PCIe Double Wide	PCIe ½ x ½
Watts	300	250	75, 50
Single Precision FP 32	10 Tflops (x8)	8-12 TFlops	5.5 TFlops
Half Precision FP 16	20 Tflops (x8)	16 TFlops	--
Integer 8 Dot Product	--	47 TOPs	22 TOPs

Styles of Scale

Depend on Use





	Scale Up	A bit of Both	Scale Out
Node	Clusters of 8 Specialized Node	2 to 8 Specialized Node	1 to 2 Standard Node
GPU Interconnect	NVLINK Cube Mesh Fabric	PCIe + DMA Mailbox	
Network	IB or 100GE RDMA	IB or ENet	Datacenter ENet
Use	HPC + DL Training Model Parallel	HPC + DL Training Data Parallel + Inference	Inference, Video Inference + Video

DL @Scale

Deep Learning End to End

Towards a RESTful place

Source Dataset

Curated Dataset

MODEL ZOO

REST API

IMPORT

Format…

PREPROCESS

clean, clip, label, Normalize, ..

TRAIN

DEPLOY

tune, compile + runtime

INFERENCE & MICROSERVICES

VISUALIZATION

SCORE + OPTIMIZE, VISUALIZATION

RESULT *

inference, prediction

Datacenter Deployment

Tools to Operationalize GPUs

Updated

Monitoring + Admin	Deployment	Resource Scheduling + Management	Frameworks + Libraries
DCGM	Hypervisor Passthrough	DXO	All DL frameworks
	Docker, AMI	Mesos	CUDA, CuDNN
	TensorRT		CuBLAS

Redux

GPU Scale Deployment

Neural Networks

Accelerators

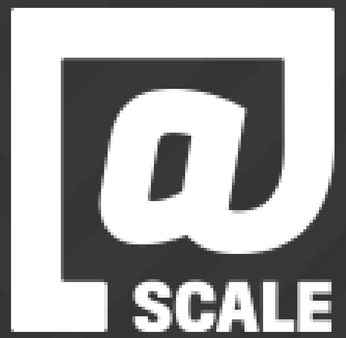
GPUs ⟷ Deep Learning

Architectural Optimizations

Optimized Inference + REST

DCGM

10^18 FLOPs


@
SCALE