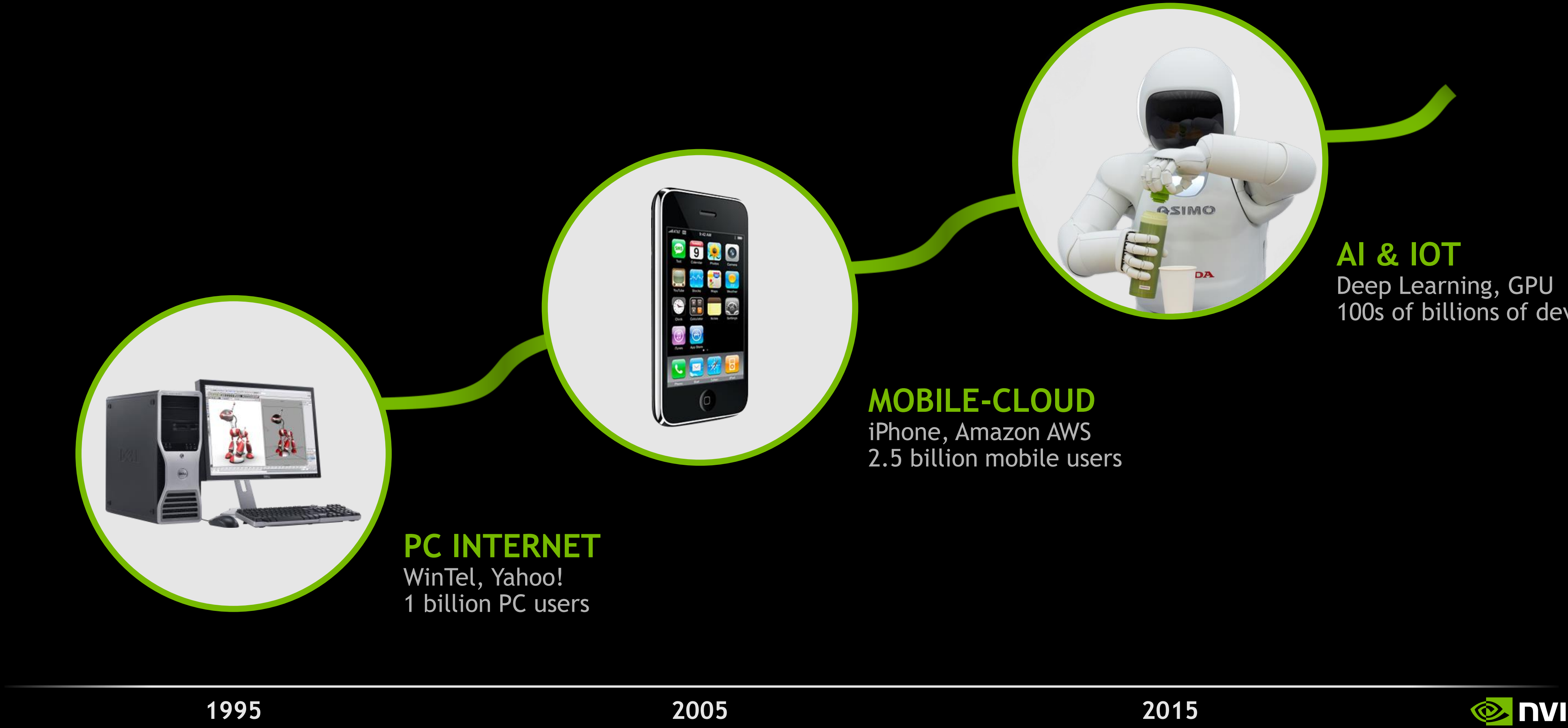
AI FOR INDUSTRY

GTC 2017

A NEW ERA OF COMPUTING



PC INTERNET
WinTel, Yahoo!
1 billion PC users

MOBILE-CLOUD
iPhone, Amazon AWS
2.5 billion mobile users

AI & IOT
Deep Learning, GPU
100s of billions of devices

1995

2005

2015

HOW A DEEP NEURAL NETWORK SEES

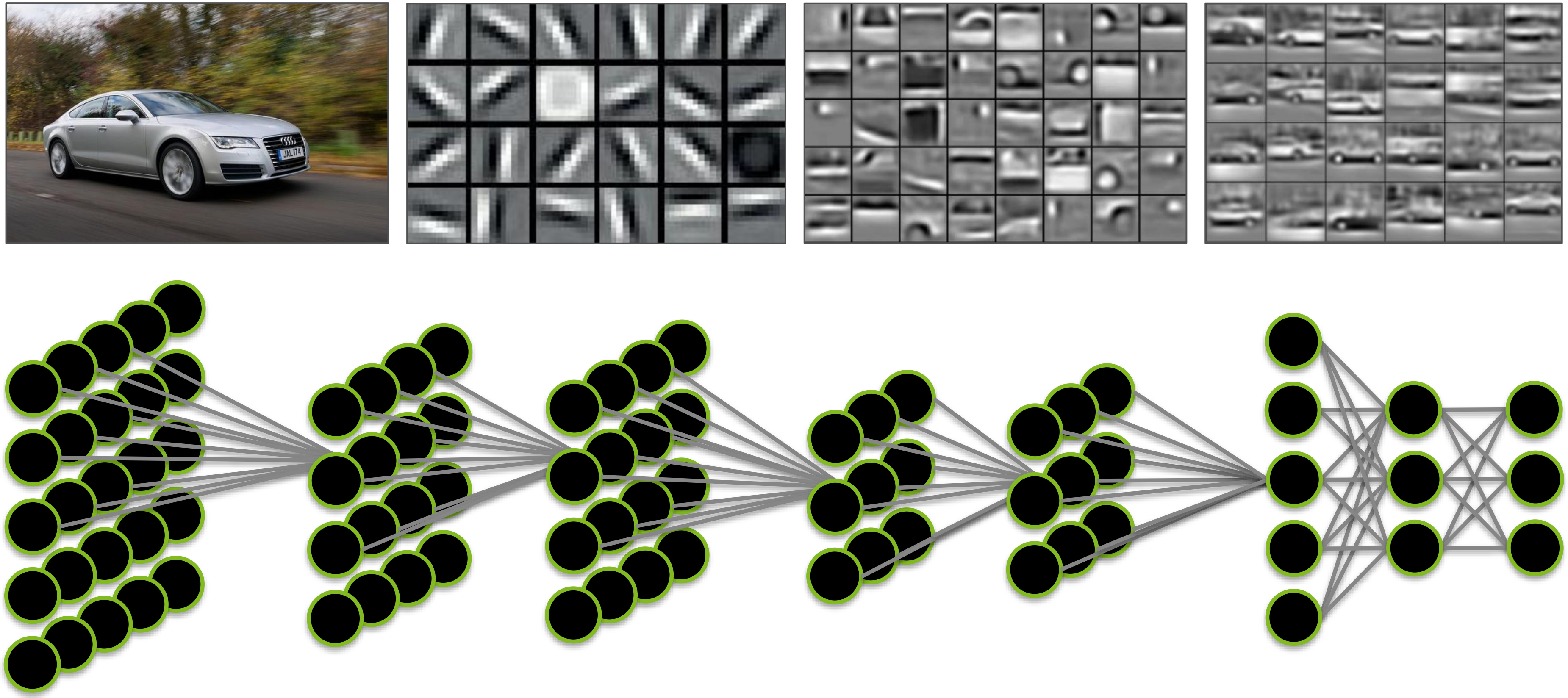


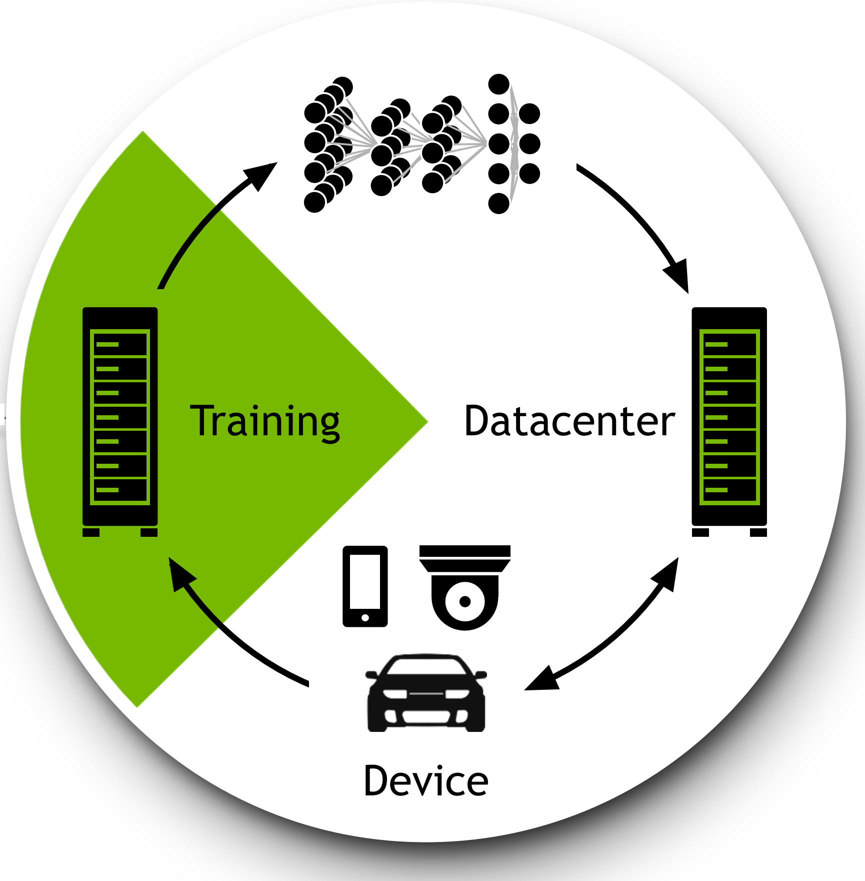
Image source: “Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks” ICML 2009 & Comm. ACM 2011. Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Ng.

GPU DEEP LEARNING IS A NEW COMPUTING MODEL

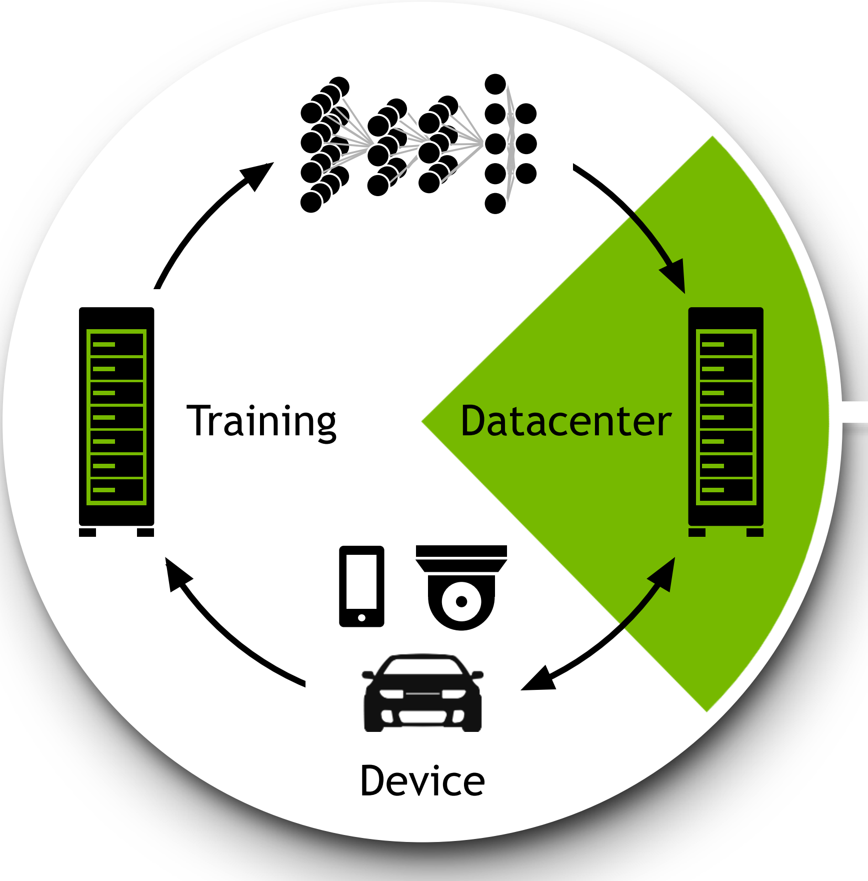
Billions of Trillions of Operations

GPU train larger models, accelerate time to market

TRAINING



GPU DEEP LEARNING IS A NEW COMPUTING MODEL



10s of billions of image, voice, video queries per day

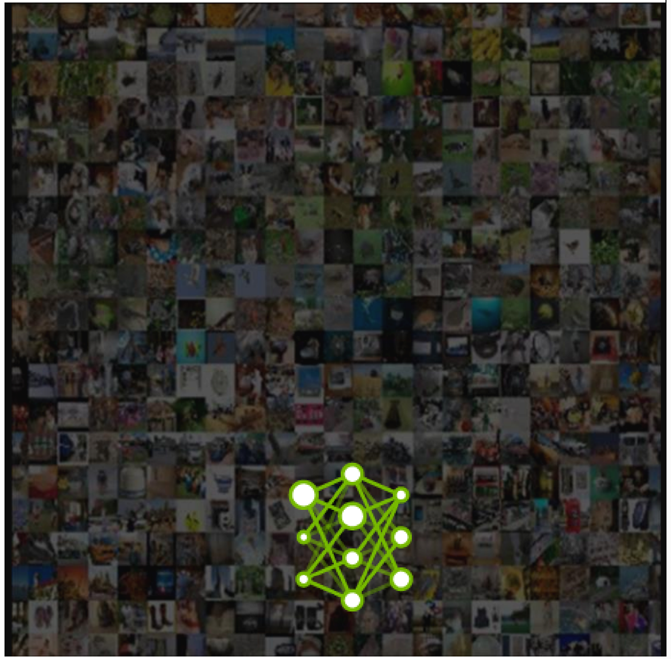
GPU inference for fast response, maximize datacenter throughput

DATACENTER INFERENCING

NEURAL NETWORK COMPLEXITY IS EXPLODING

To Tackle Increasingly Complex Challenges

7 ExaFLOPS
60 Million Parameters

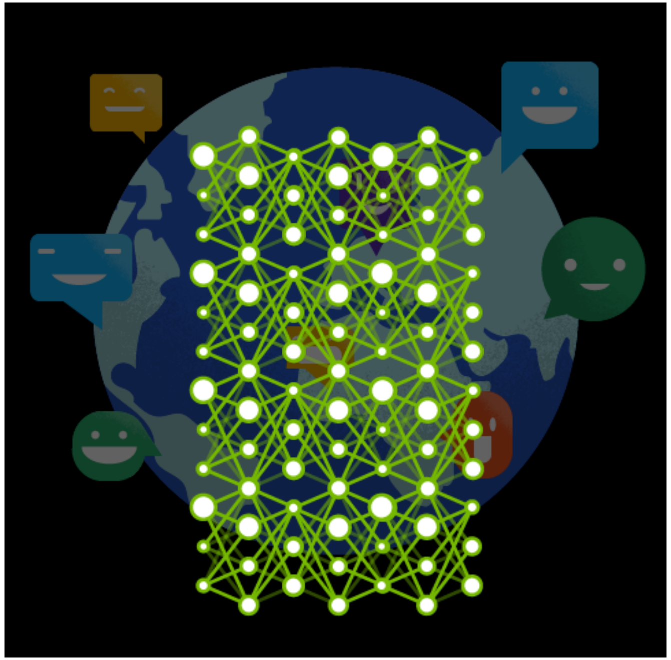


2015 - Microsoft ResNet
Superhuman Image Recognition

20 ExaFLOPS
300 Million Parameters

2016 - Baidu Deep Speech 2
Superhuman Voice Recognition

100 ExaFLOPS
8700 Million Parameters

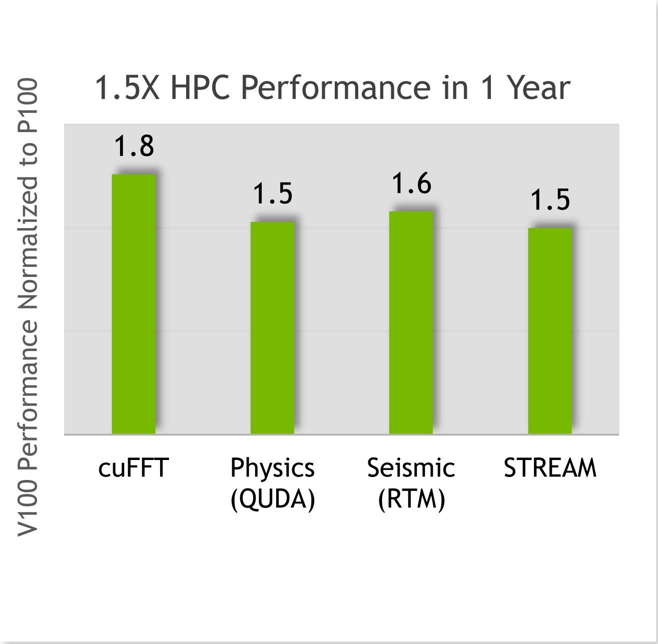


2017 - Google Neural Machine Translation
Near Human Language Translation

Summit
Supercomputer
200+ PetaFlops
~3,400 Nodes
10 Megawatts

ROAD TO EXASCALE

Volta to Fuel Most Powerful US Supercomputers



System Config Info: 2X Xeon E5-2690 v4, 2.6GHz, w/ 1X Tesla P100 or V100. V100 measured on pre-production hardware.

NVIDIA SATURN V

124 DGX-1 Deep Learning Supercomputers

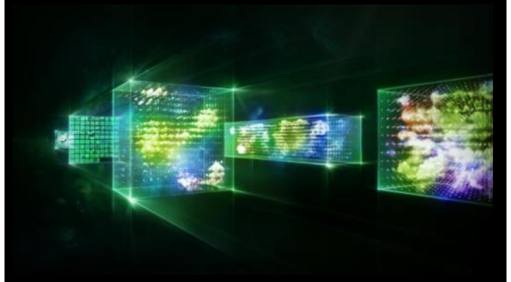
TESLA V100

THE MOST ADVANCED DATA CENTER GPU EVER BUILT

5,120 CUDA cores
640 NEW Tensor cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS
120 Tensor TFLOPS
20MB SM RF | 16MB Cache | 16GB HBM2 @ 900 GB/s
300 GB/s NVLink

NEW TENSOR CORE BUILT FOR AI

Delivering 120 TFLOPS of DL Performance



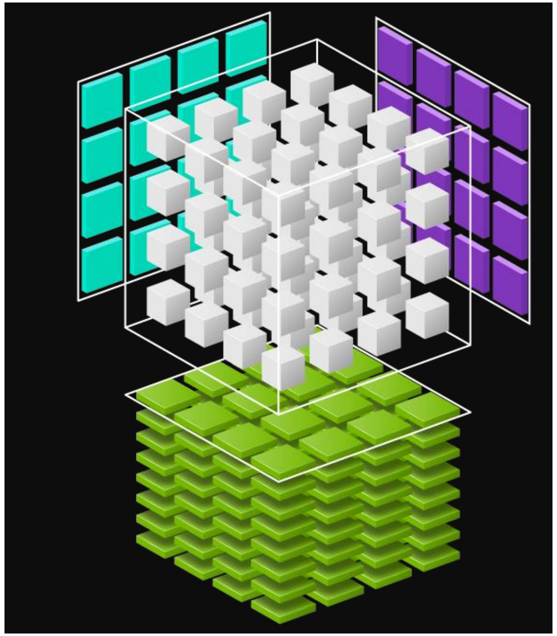
MATRIX DATA OPTIMIZATION:

Dense Matrix of Tensor Compute

TENSOR-OP CONVERSION:

FP32 to Tensor Op Data for Frameworks

VOLTA-OPTIMIZED cuDNN



VOLTA TENSOR CORE

4x4 matrix processing array

D[FP32] = A[FP16] * B[FP16] + C[FP32]

Optimized For Deep Learning

Caffe2

Microsoft Cognitive Toolkit

mxnet

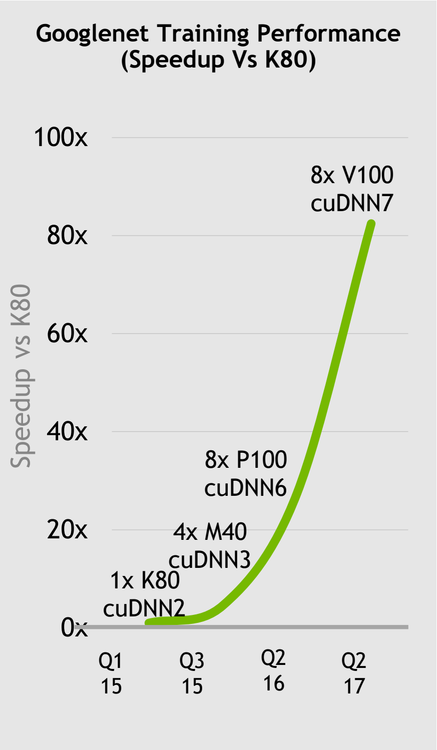
PYTORCH

TensorFlow

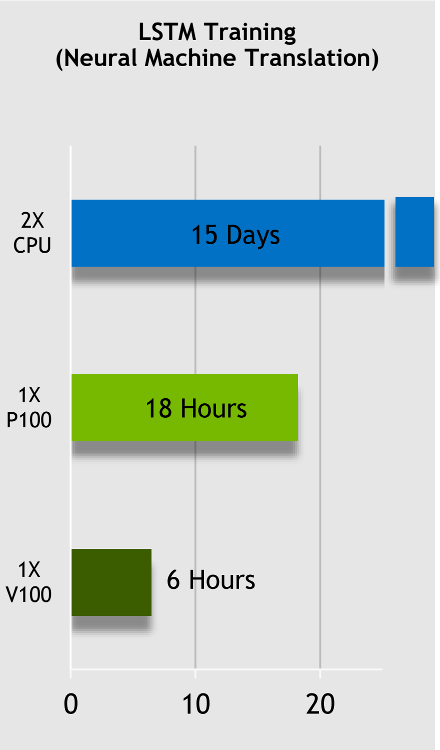
ALL MAJOR FRAMEWORKS

REVOLUTIONARY AI PERFORMANCE

3X Faster DL Training Performance

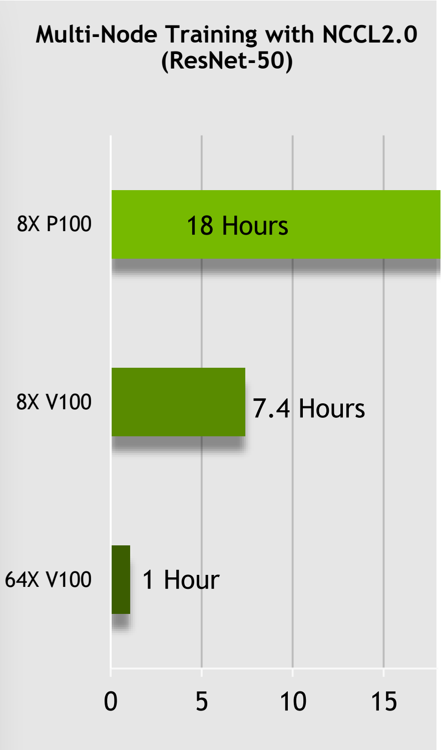


Over 80x DL Training Performance in 3 Years



3X Reduction in Time to Train Over P100

Neural Machine Translation Training for 13 Epochs |German ->English, WMT15 subset | CPU = 2x Xeon E5 2699 V4 | V100 performance measured on pre-production hardware.

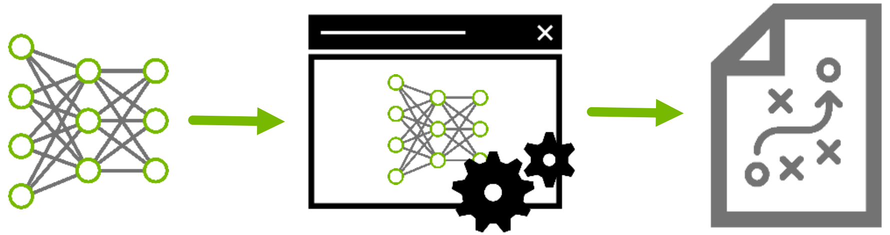


85% Scale-Out Efficiency Scales to 64 GPUs with Microsoft Cognitive Toolkit

ResNet50 Training for 90 Epochs with 1.28M images dataset | Using Caffe2 | V100 performance measured on pre-production hardware.

VOLTA DELIVERS 3X MORE INFERENCE THROUGHPUT

Low Latency performance with V100 and TensorRT

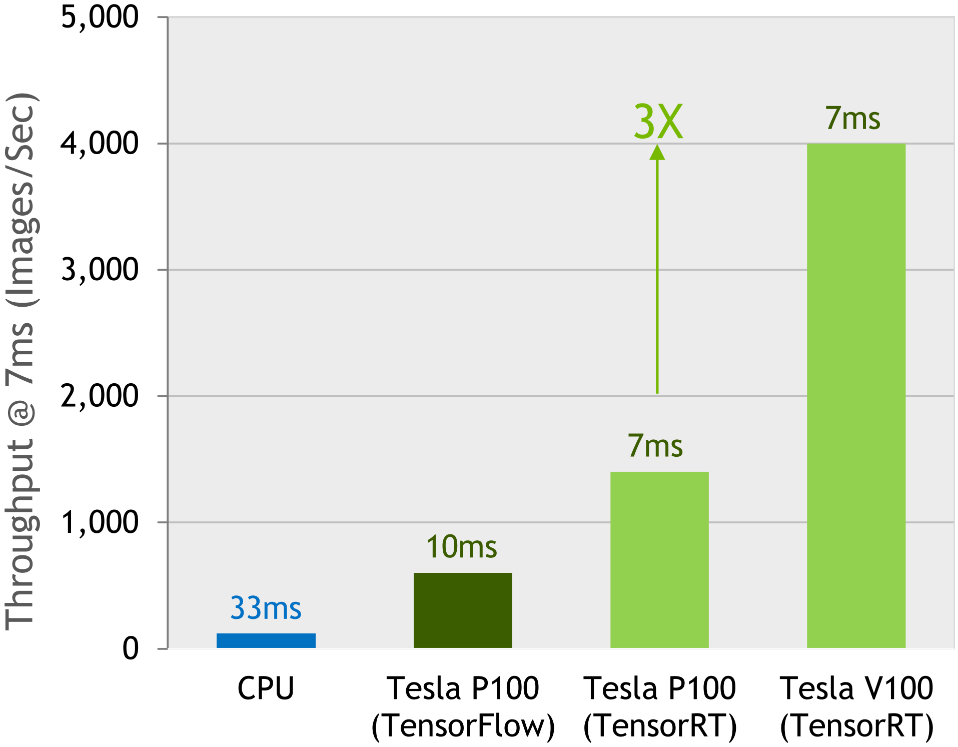


Trained Neural Network

TensorRT
Fuse Layers
Compact
Optimize Precision
(FP32, FP16, INT8)

Compiled Real-time Network

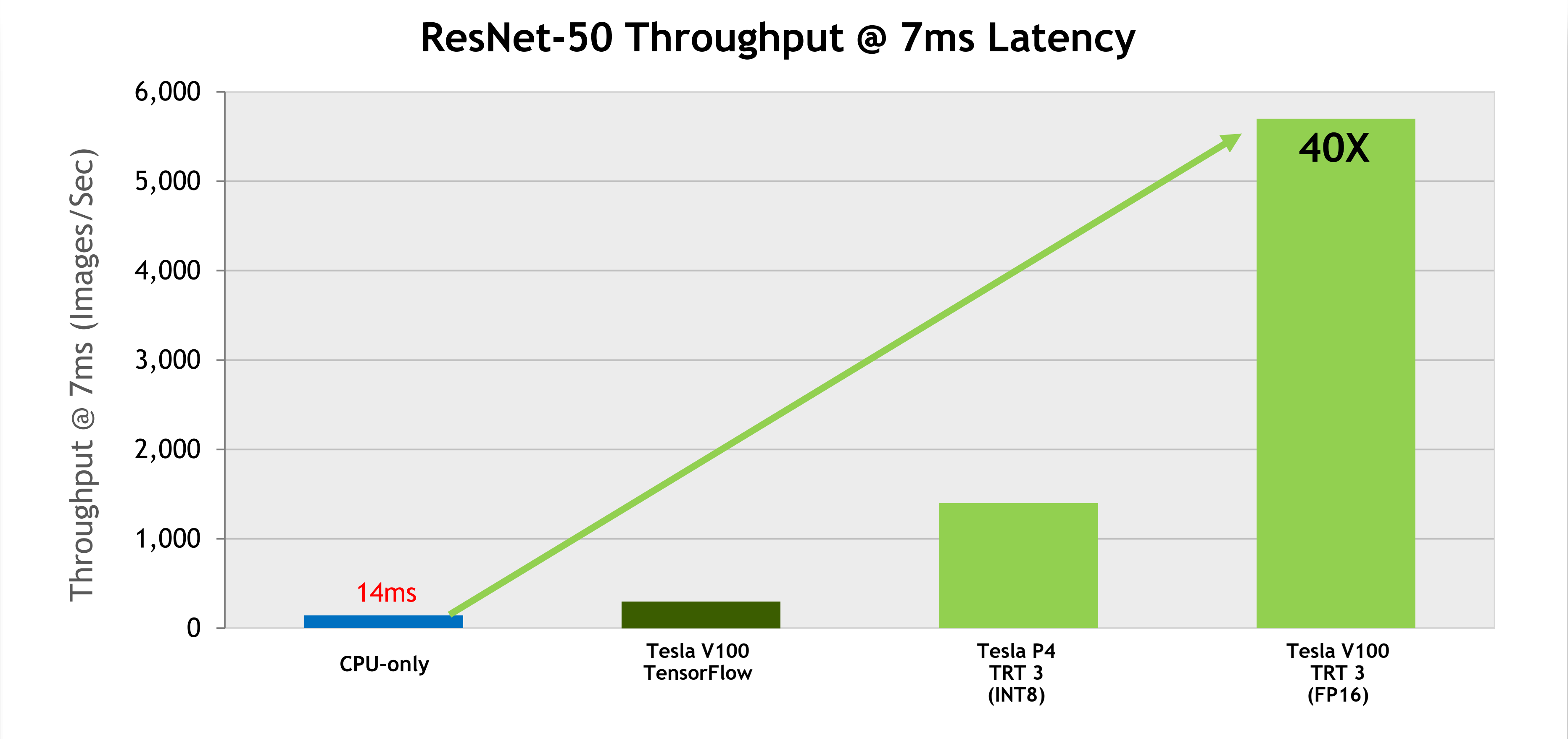
3x more throughput at 7ms latency with V100
(ResNet-50)



CPU Server: 2X Xeon E5-2660 V4; GPU: w/P100, w/V100 (@150W) | V100 performance measured on pre-production hardware.

NVIDIA TENSORRT 3

World’s Fastest Inference Platform



Workload ResNet-50| Data-set ImageNet | CPU: Skylake | GPU: Tesla P4 or Tesla V100

NVIDIA