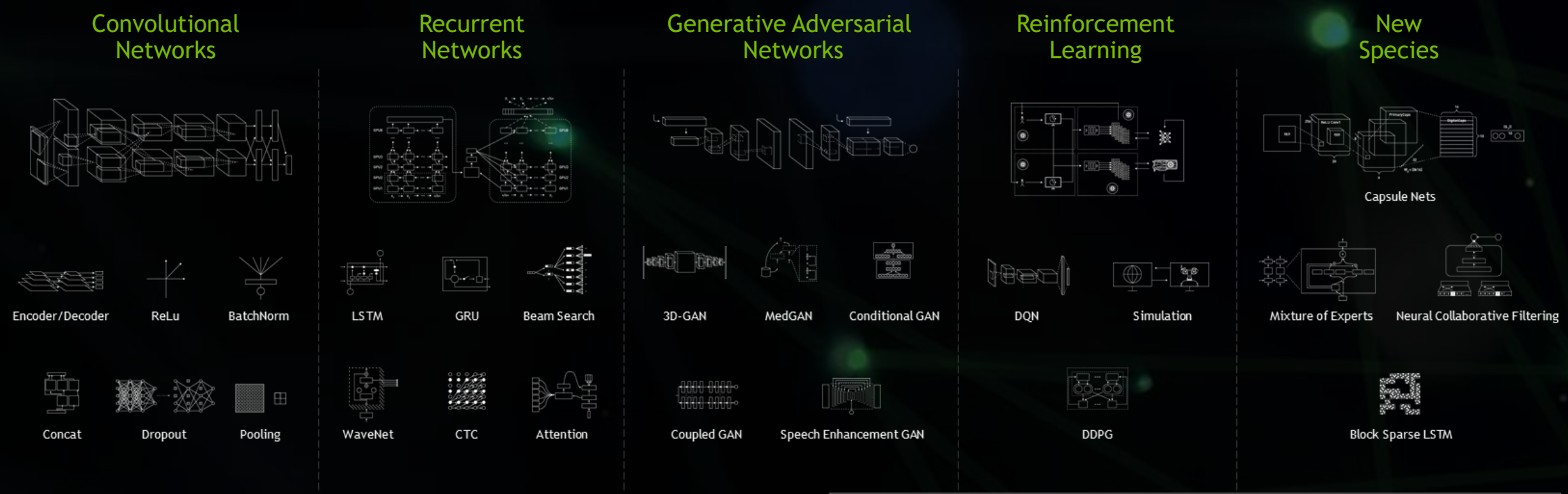
NVIDIA INFERENCE PLATFORM

September, 2018



There is a Cambrian explosion of neural networks.

Thousands of new networks over the past 5 years

Models are getting smarter

PROGRAMMABILITY
LATENCY
ACCURACY
SIZE
THROUGHPUT
ENERGY EFFICIENCY
RATE OF LEARNING

NEURAL NETWORK COMPLEXITY IS EXPLODING

Bigger and More Compute Intensive

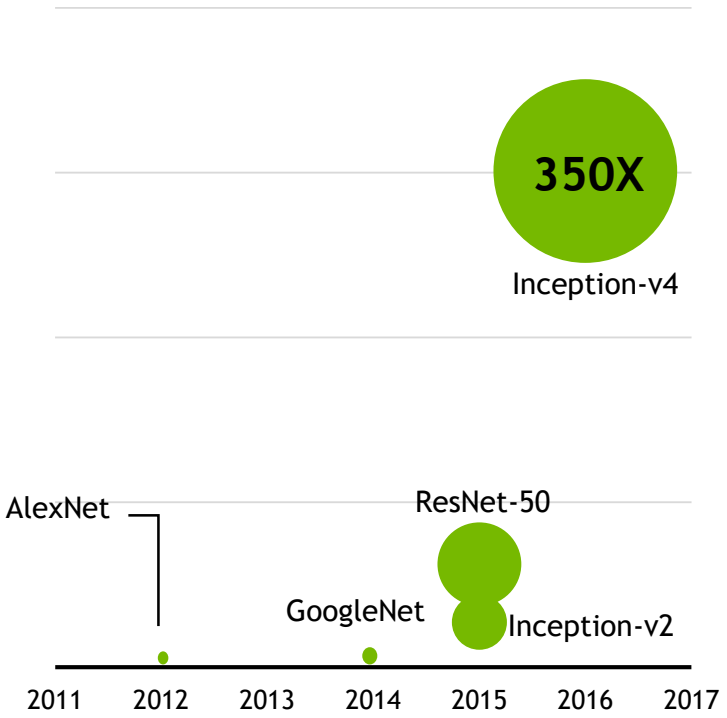
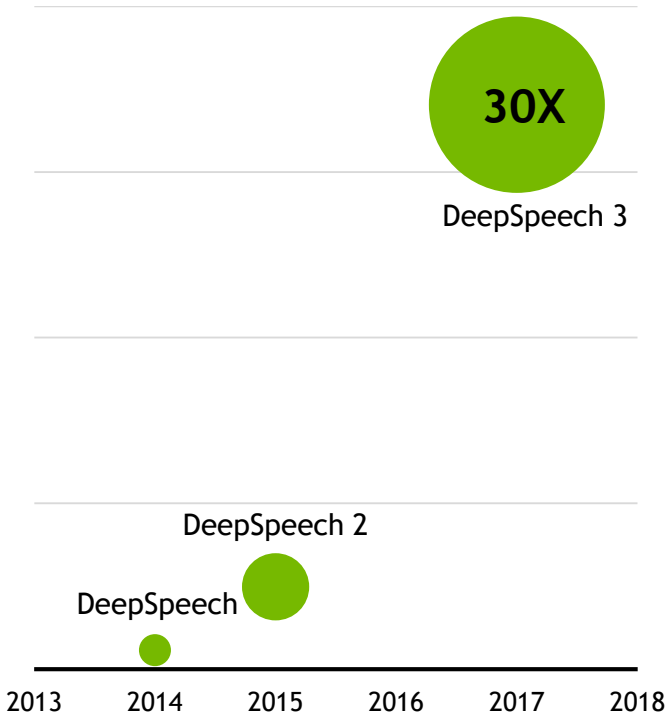
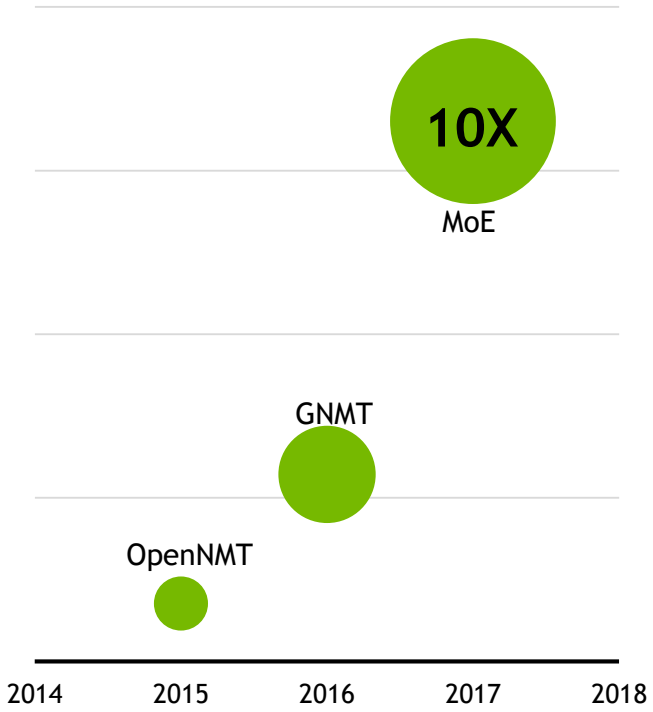


Image
(GOP * Bandwidth)

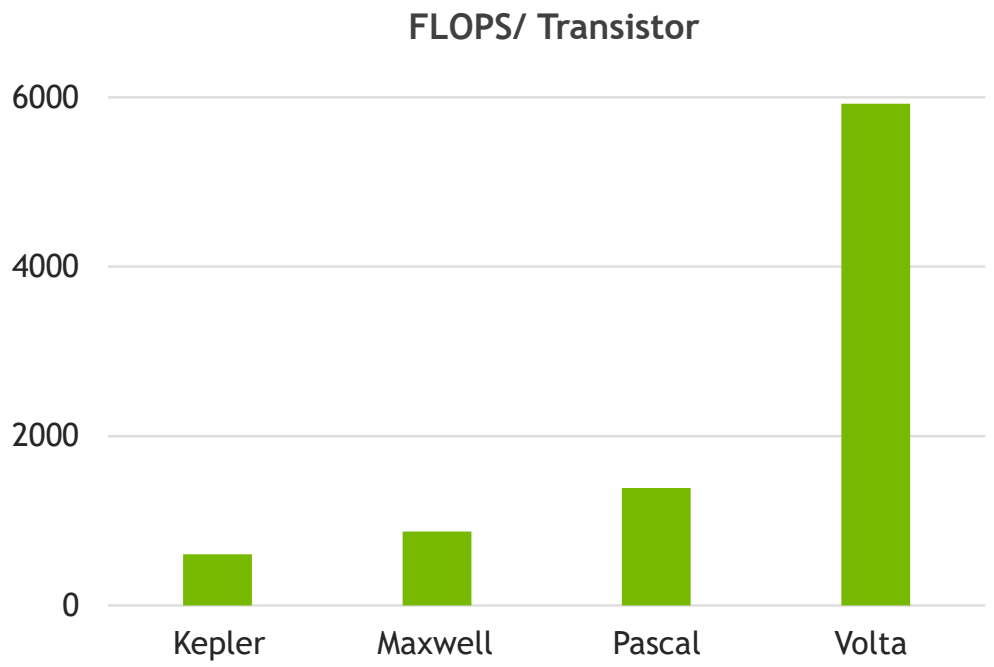


Speech
(GOP * Bandwidth)

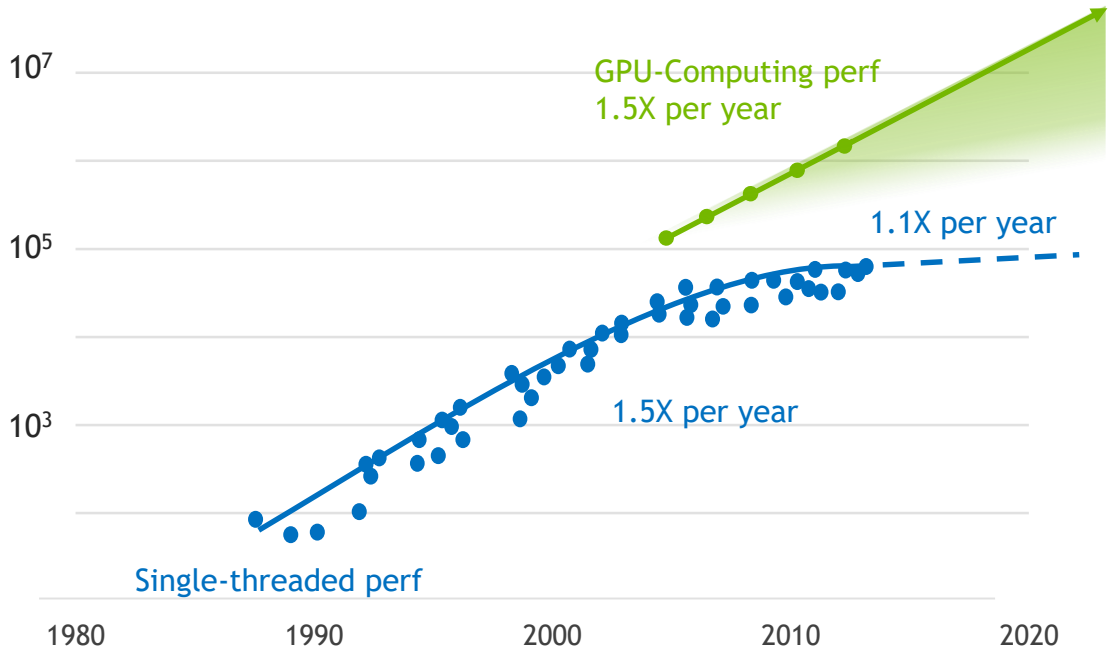


Translation
(GOP * Bandwidth)

RISE OF NVIDIA GPU COMPUTING



CUDA – Domain Specific Computing Architecture
10X in 5 Years



40 Years of CPU Trend Data

Original data up to the year 2010 collected and plotted by M. Horowitz,
F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

TESLA PLATFORM STACK

World's Leading Data Center Platform for Accelerating HPC and AI

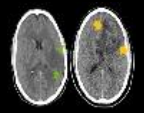
CUSTOMER USECASES


Speech

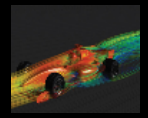
Translate


Recommender

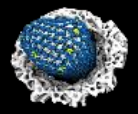
CONSUMER INTERNET


Healthcare

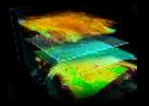

Manufacturing


Engineering

ENTERPRISE APPLICATIONS


Molecular Simulations


Weather Forecasting


Seismic Mapping

SUPERCOMPUTING

INDUSTRY FRAMEWORKS & APPLICATIONS

Caffe2 · Chainer · Microsoft Cognitive Toolkit · mxnet · PaddlePaddle · PYTORCH · TensorFlow

Amber · ANSYS · CHROMA · GROMACS FAST. FLEXIBLE. FREE. · LAMMPS · NAMD · SIMULIA · VASP

+550 Applications

NVIDIA SDK & LIBRARIES

cuBLAS · cuDNN · cuFFT · cuSPARSE · DeepStream · NCCL · TensorRT · OpenACC

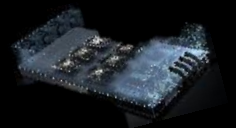
CUDA

TESLA GPUs & SYSTEMS


TESLA GPU


NVIDIA DGX STATION


NVIDIA DGX-1

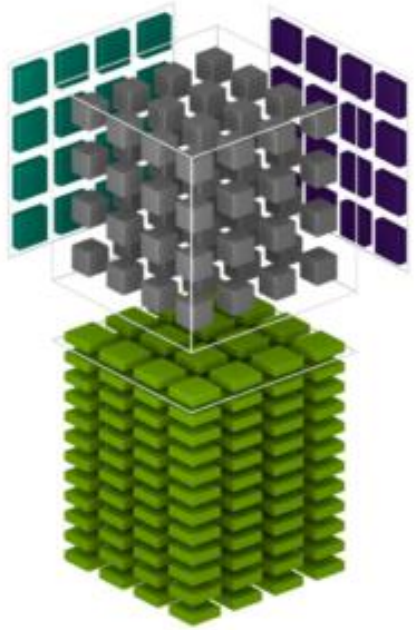

NVIDIA HGX-1

SYSTEM OEM: DELL · Hewlett Packard Enterprise · IBM

CLOUD: amazon web services · Google Cloud Platform · Microsoft Azure

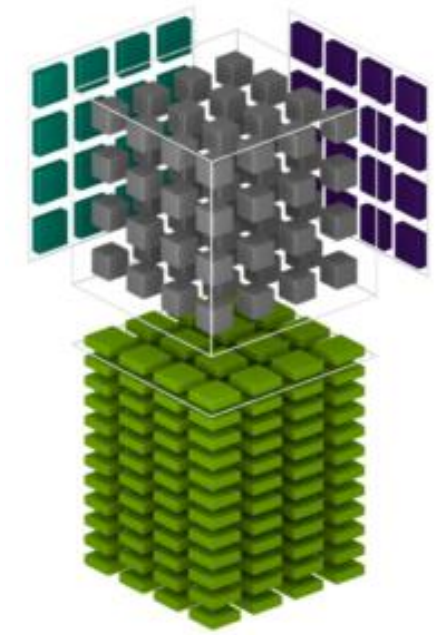
NVIDIA

VOLTA TENSOR CORE



FP16
8x CUDA core
125 TFLOPs

TURING TENSOR CORE



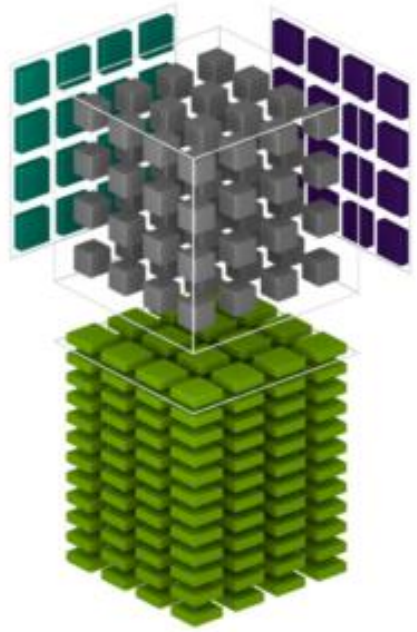
	FP16	Int8	Int4
T4	**8x** CUDA core	**16x** CUDA core	**32x** CUDA core
75W	**65+** TFLOPs	**130+** TOPs	**260+** TOPs

SW at Launch TensorRT, Libraries CUTLASS Open Source Tensor Library, CUDA

T4 and Turing performance projections and specifications are preliminary and subject to change without notice.

NVIDIA T4: NEXT GENERATION INFERENCE WITH TURING TENSOR CORES

Turing Tensor Cores



65 TFLOPs FP16

130 TOPs INT8

260 TOPs INT4

Programmable Acceleration at scale

Inference
Next Generation Inference With Turing Tensor Cores for FP16, INT8, INT4, INT1

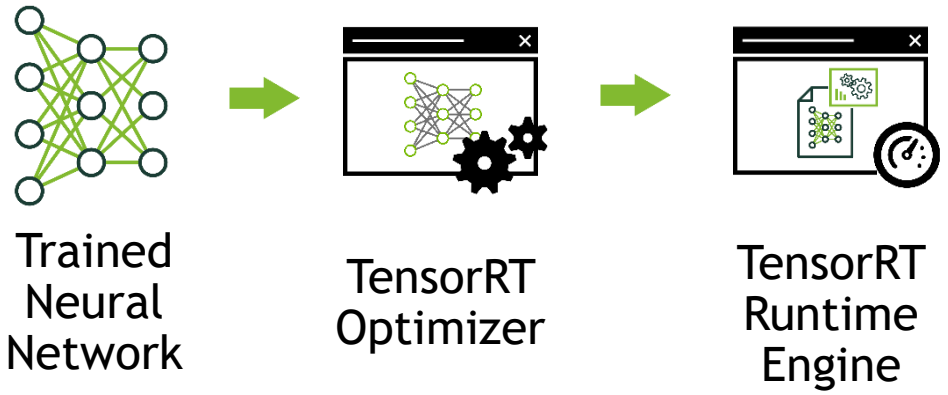
Video & Graphics
2x User density vs P4
2x Video Decode capability vs P4

DL Training
Entry level training SKU with Turing Tensor Cores

T4 and Turing performance projections and specifications are preliminary and subject to change without notice.

NVIDIA TensorRT

Deep Learning Inference Optimizer and Runtime

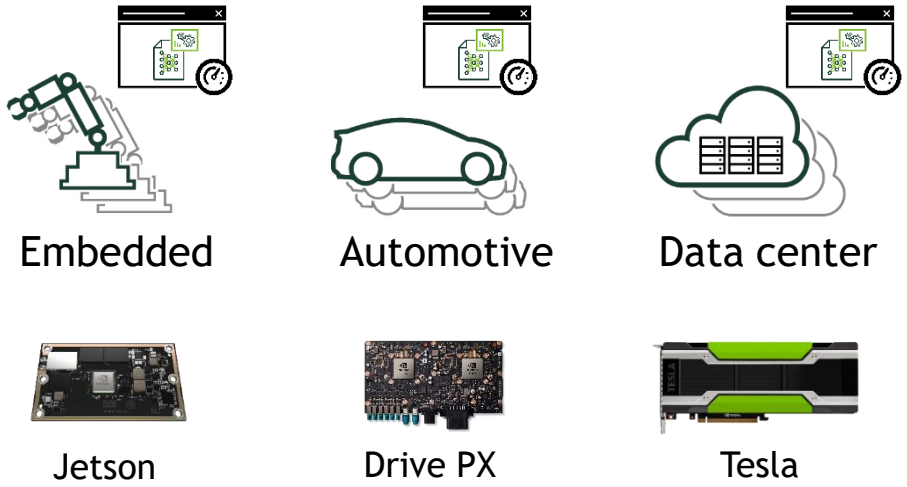


Optimize neural networks and Deploy in production environments

Maximize inference throughput for latency-critical services in production

Deploy faster, more responsive and memory efficient applications with INT8 and FP16 optimizations

Accelerate models trained in any framework with ONNX support and native framework integrations



developer.nvidia.com/tensorrt

WIDELY ADOPTED

Baidu 百度
BOSCH
Continental
DAIMLER
DarwinAI
DEEPMAP
Ford
here
科大讯飞 iFLYTEK
IBM
JD.COM
金山云
AURORA
MAN
navya
SAP
SEARIDGE TECHNOLOGIES
SK telecom
图森 tuSimple
verizon
VOLVO
wrnch AI
欢聚时代 YY Inc.
ZENRIN
ZF
ZOOX

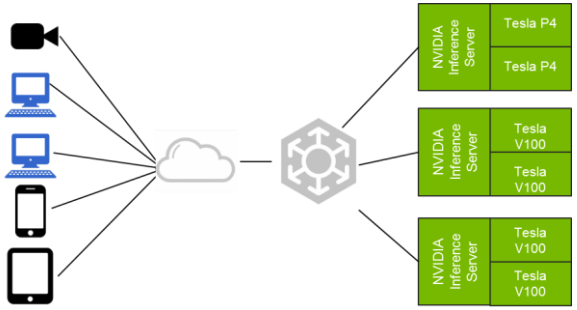
ANNOUNCING TensorRT 5

Turing Support • Inference Server • Optimizations & APIs

World's Most Advanced Inference Accelerator

Up to 50x faster inference for apps such as translation using mixed precision on Turing Tensor Cores

TensorRT Inference Server



Maximize GPU utilization by executing multiple models from different frameworks on a node via API

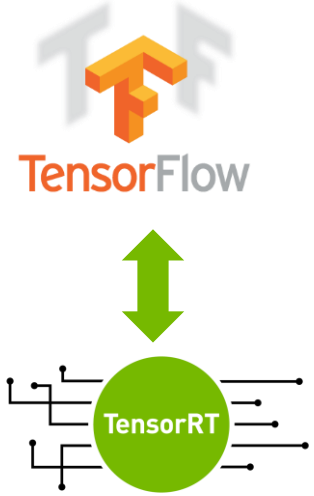
New optimizations & flexible INT8 APIs

Achieve highest throughput at low latency with newly optimized operations, INT8 workflows, and support for Win and CentOS

Free download to members of NVIDIA Developer Program soon at developer.nvidia.com/tensorrt

TensorRT INTEGRATED WITH TensorFlow

Speed up TensorFlow inference with TensorRT optimizations



Speed up TensorFlow model inference with TensorRT with new TensorFlow APIs

Simple API to use TensorRT within TensorFlow easily

Sub-graph optimization with fallback offers flexibility of TensorFlow and optimizations of TensorRT

Optimizations for FP32, FP16 and INT8 with use of Tensor Cores automatically

```
# Apply TensorRT optimizations
trt_graph = trt.create_inference_graph(frozen_graph_def,
                output_node_name,
                max_batch_size=batch_size,
                max_workspace_size_bytes=workspace_size,
                precision_mode=precision)

# INT8 specific graph conversion
trt_graph = trt.calib_graph_to_infer_graph(calibGraph)
```

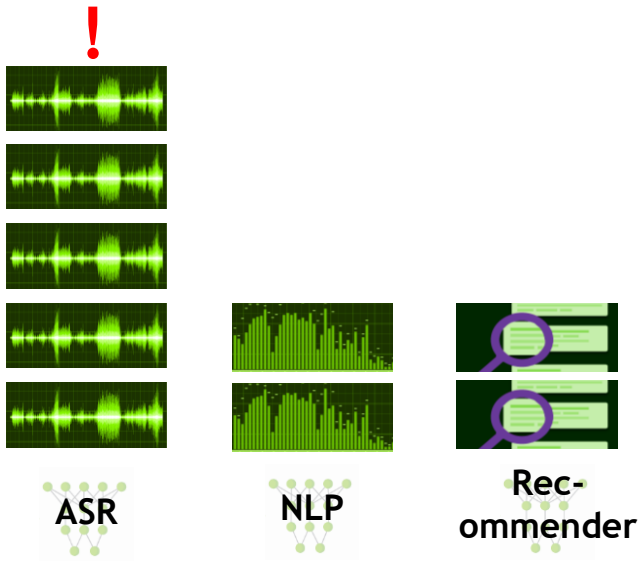
Available from TensorFlow 1.7

https://github.com/tensorflow/tensorflow

INEFFICIENCY LIMITS INNOVATION

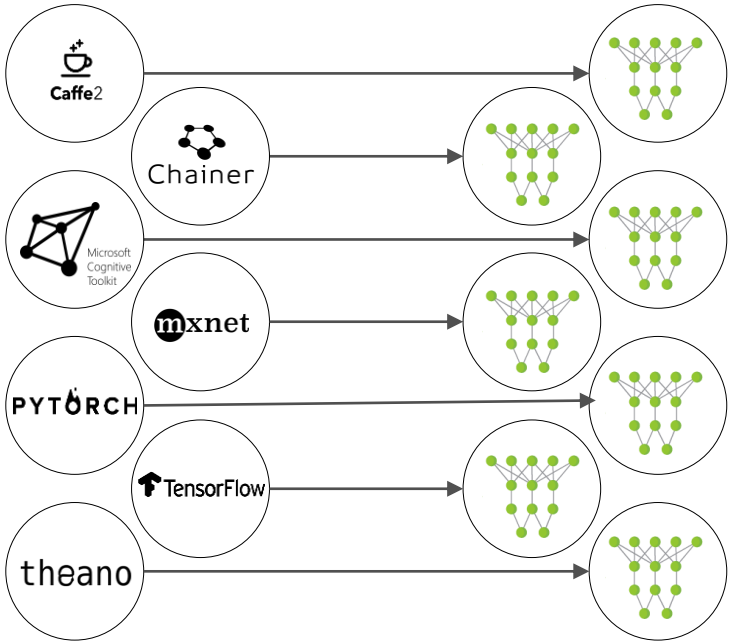
Difficulties with Deploying Data Center Inference

Single Model Only



Some systems are overused while others are underutilized

Single Framework Only



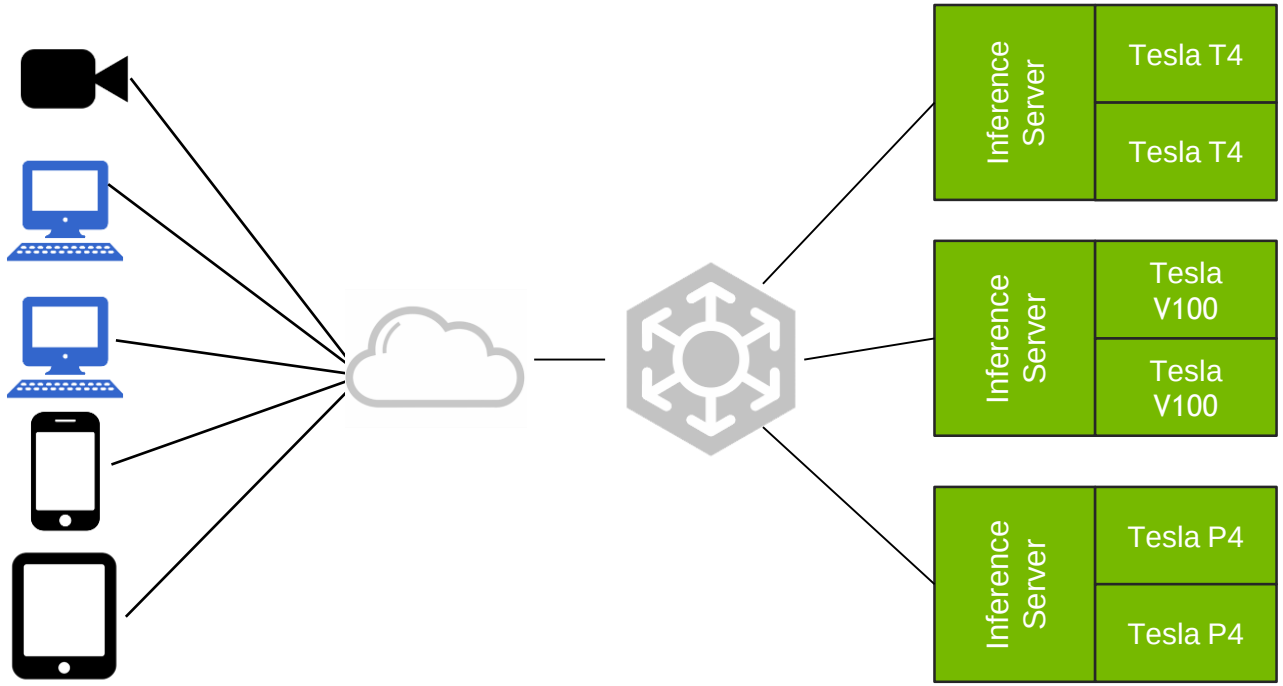
Solutions can only support models from one framework

Custom Development

Developers need to reinvent the plumbing for every application

NVIDIA TensorRT INFERENCE SERVER

Containerized Microservice for Data Center Production



Maximize real-time inference performance of GPUs

Quickly deploy and manage multiple models per GPU per node

Easily scale to heterogeneous GPUs and multi GPU nodes

Integrates with orchestration systems and auto scalers via latency and health metrics


NVIDIA®