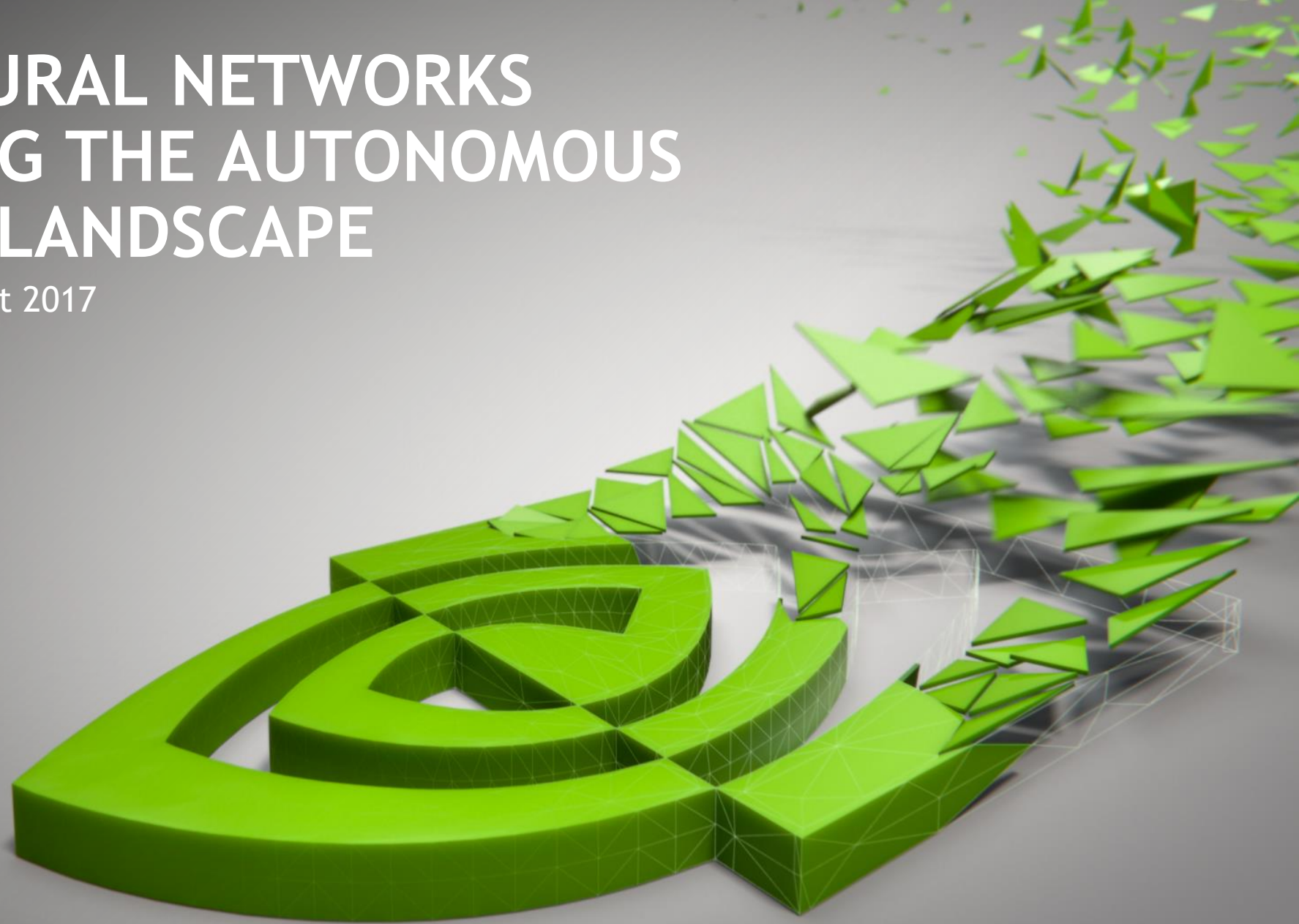
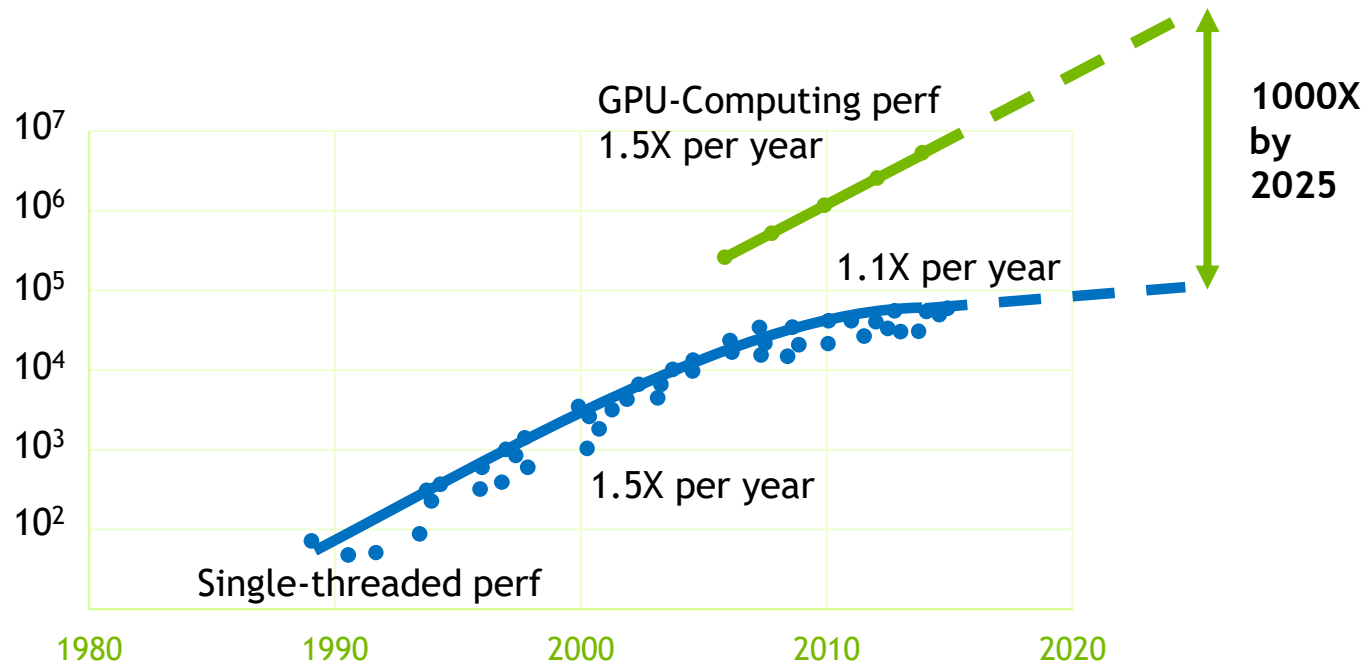
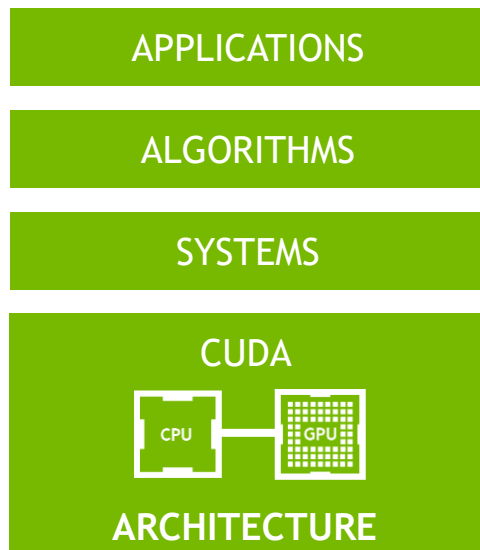


DEEP NEURAL NETWORKS CHANGING THE AUTONOMOUS VEHICLE LANDSCAPE

Dennis Lui | August 2017



THE RISE OF GPU COMPUTING



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

THE EXPANDING UNIVERSE OF MODERN AI

"THE BIG BANG"

Big Data
GPU
Algorithms

RESEARCH



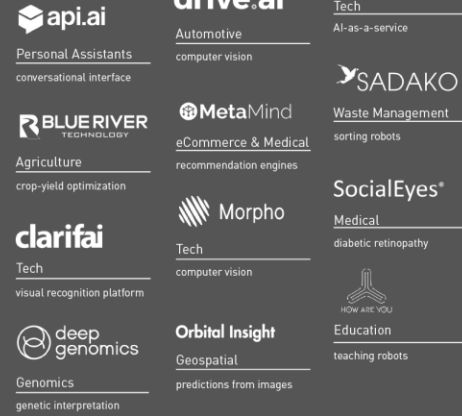
CORE TECHNOLOGY / FRAMEWORKS



AI-as-a-PLATFORM



START-UPS



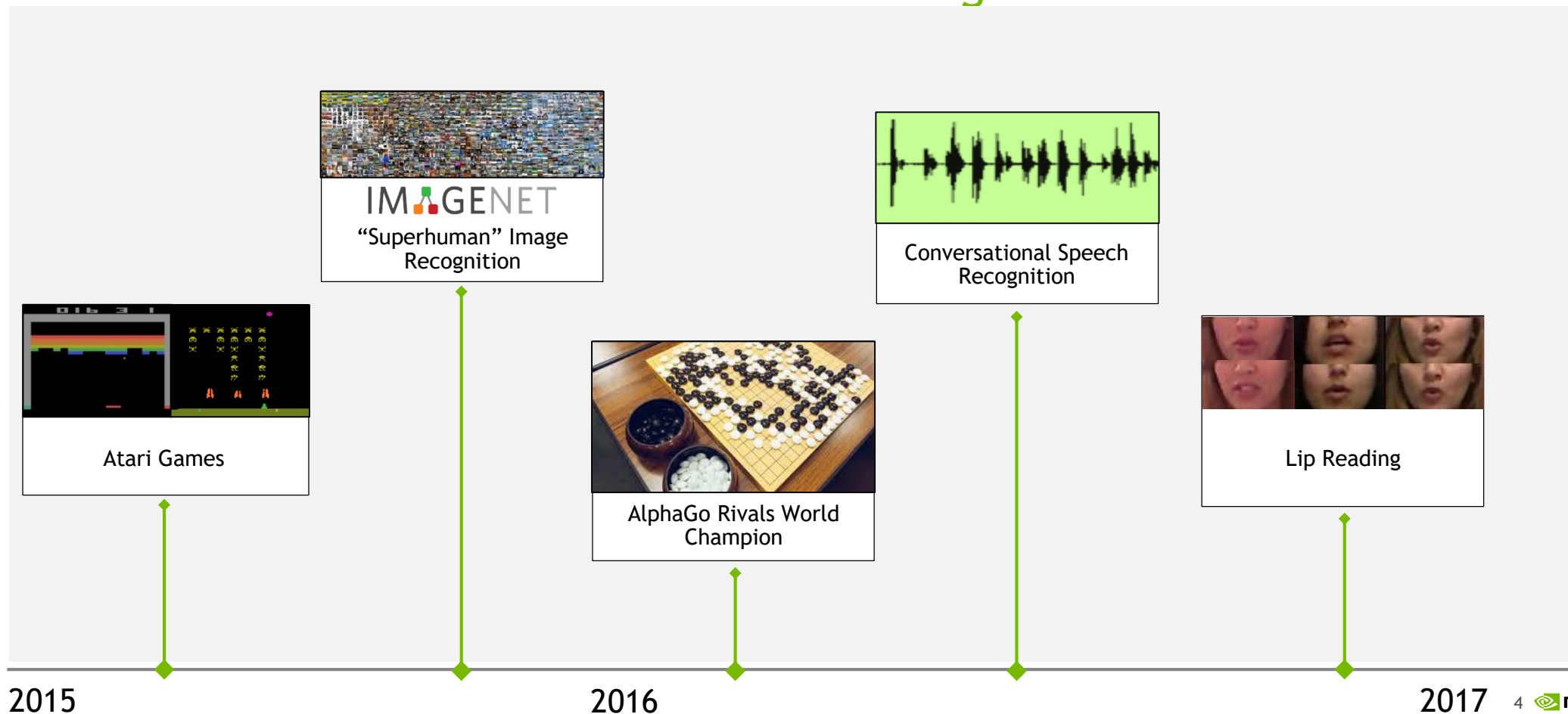
3,000+ AI START-UPS
\$5B IN FUNDING
 Source: Venture Scanner

INDUSTRY LEADERS



AI BREAKTHROUGHS

Recent Breakthroughs



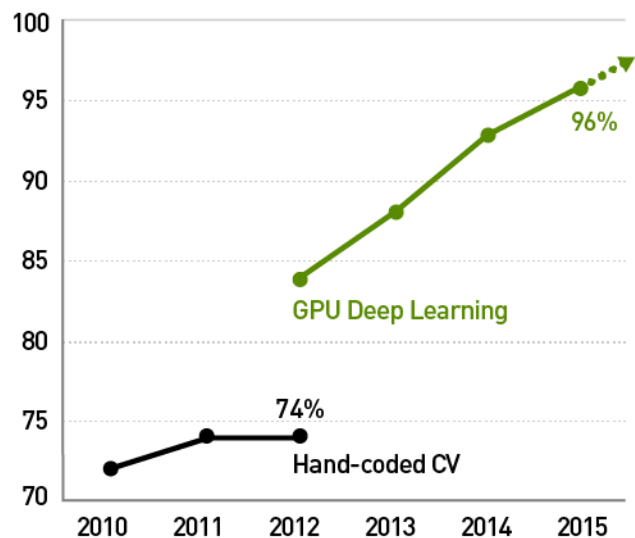
2015

2016

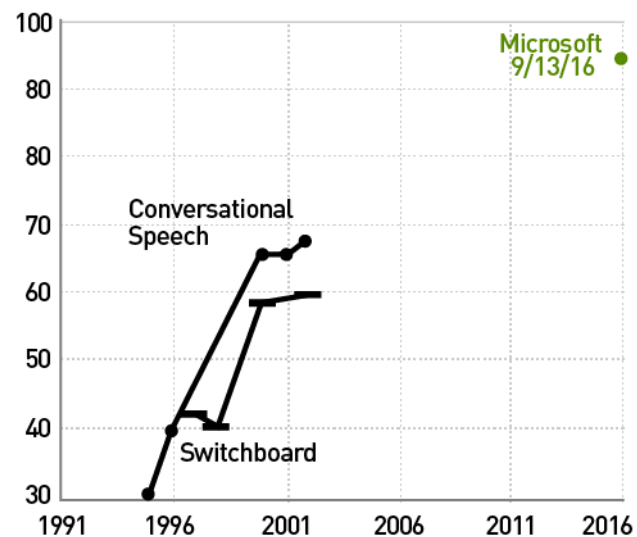
2017

AI IMPROVING AT AMAZING RATES

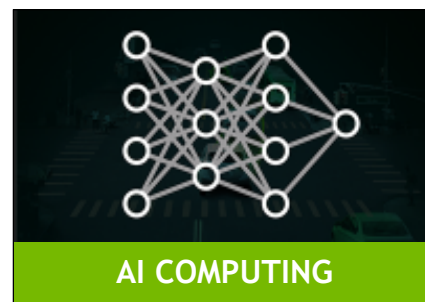
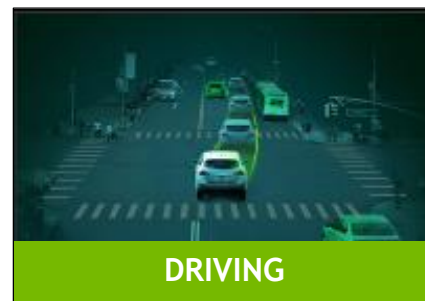
IMAGENET ACCURACY



SPEECH RECOGNITION ACCURACY

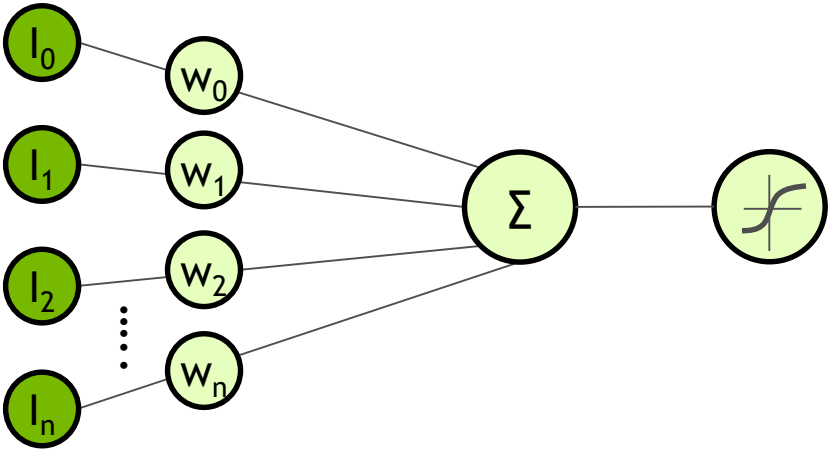
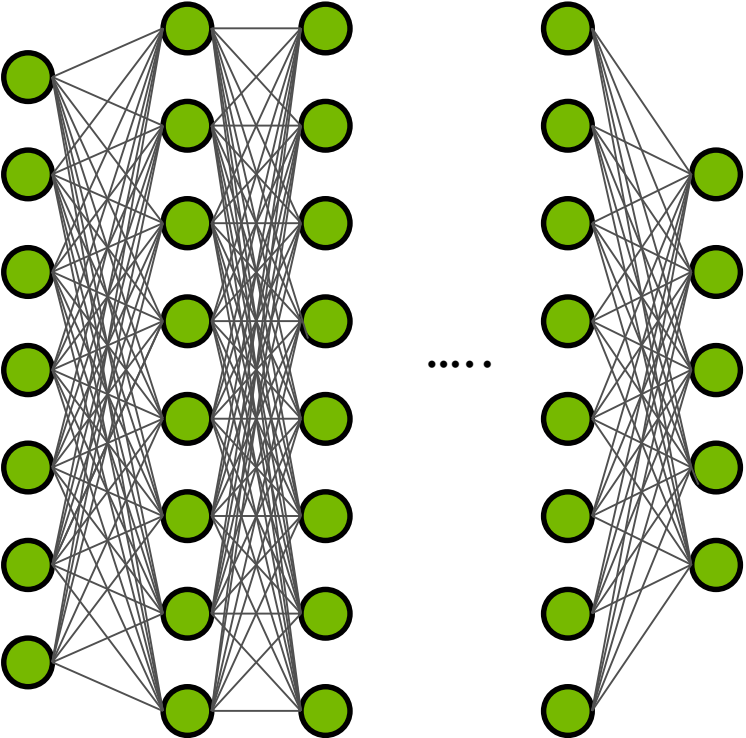


AI IS THE SOLUTION TO SELF DRIVING



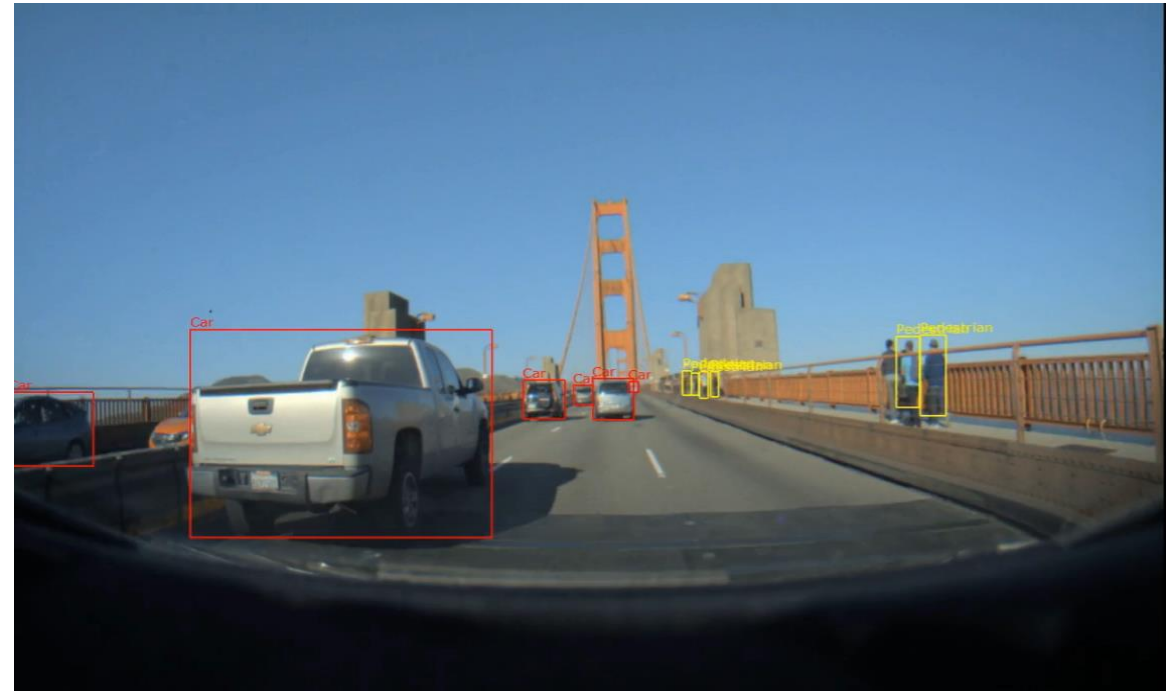
DEEP LEARNING FOR AUTONOMOUS DRIVING

DEEP NEURAL NETWORK



MULTICLASS OBJECT DETECTION & CLASSIFICATION NETWORK

Description	Demonstrates NVIDIA's proprietary deep neural network (DNN) to perform object detection
Types Detected/ Color Code	Red: Cars Cyan: Trucks Green: Traffic Signs (Detection Only) Blue: Bicycles Yellow: Pedestrians



LANE DETECTION NETWORK

Description	<p>Demonstrates NVIDIA's proprietary deep neural network (DNN) to perform lane detection on the road</p> <p>Detects ego-lane by showing the boundaries of the left and right lane, and in some cases, is able to show the left and right boundaries of adjacent lanes as well</p>
Color Code	<p>Red: Ego-lane left Green: Ego-lane right Yellow: Left adjacent lane Blue: Right adjacent lane</p>



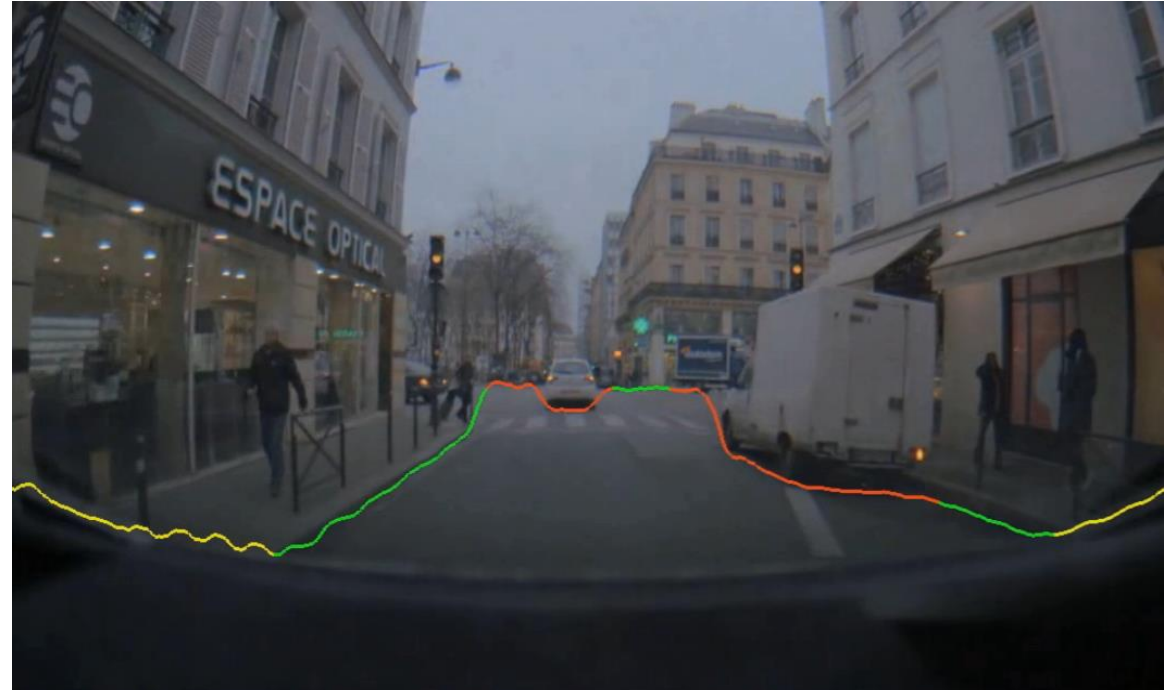
FREE SPACE DETECTION NETWORK

Description

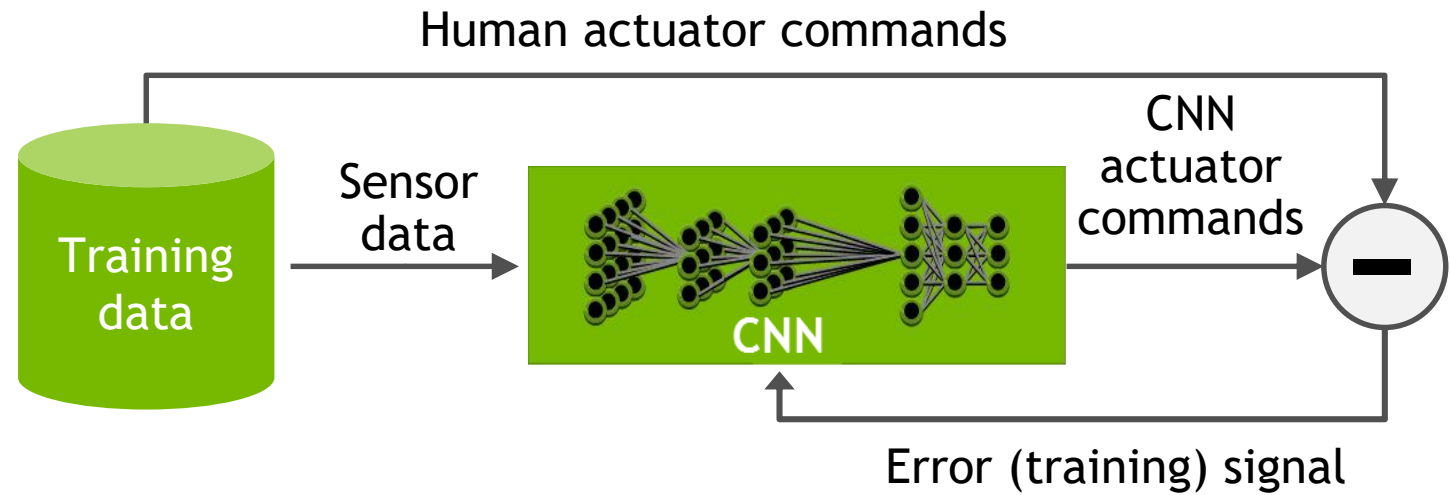
Demonstrates NVIDIA's proprietary deep neural network (DNN) to detect free space in front of the vehicle.

Color Code

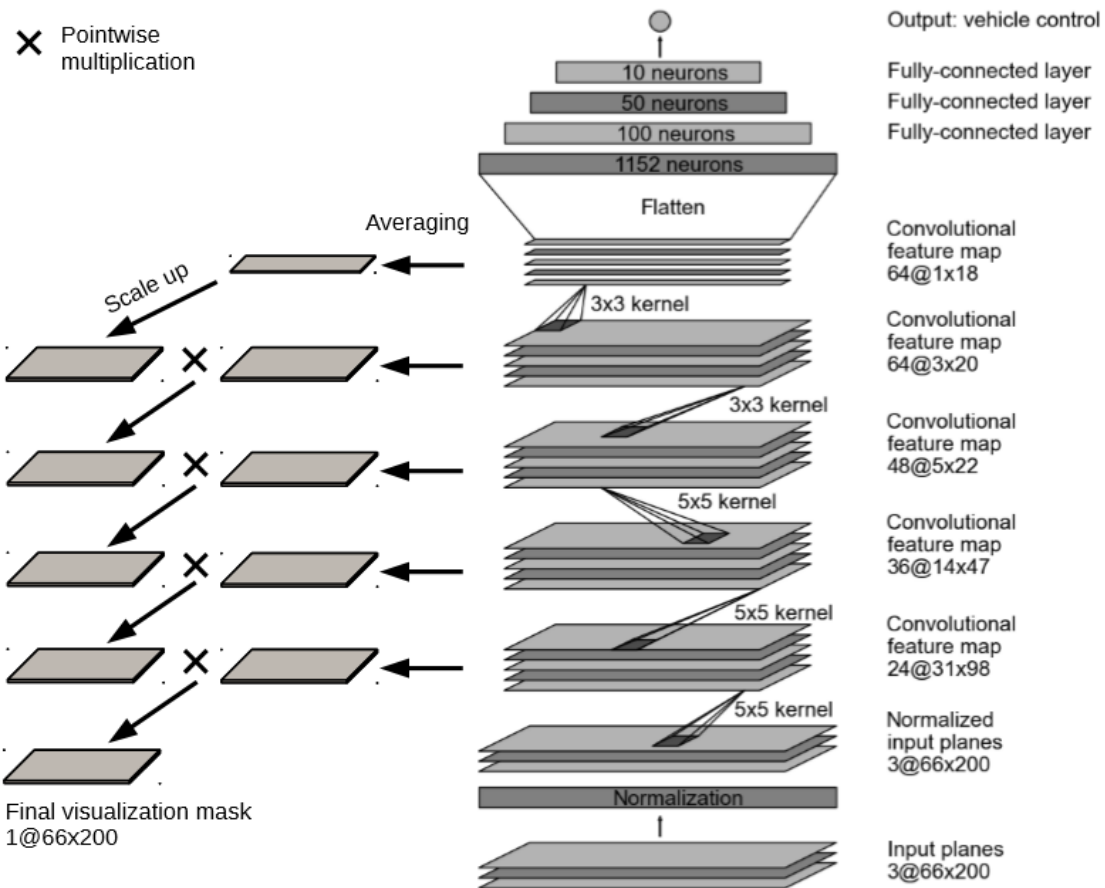
Red: cars
Green: Curbs
Blue: Pedestrians
Yellow: Others



END-TO-END AUTONOMOUS DRIVING NETWORK

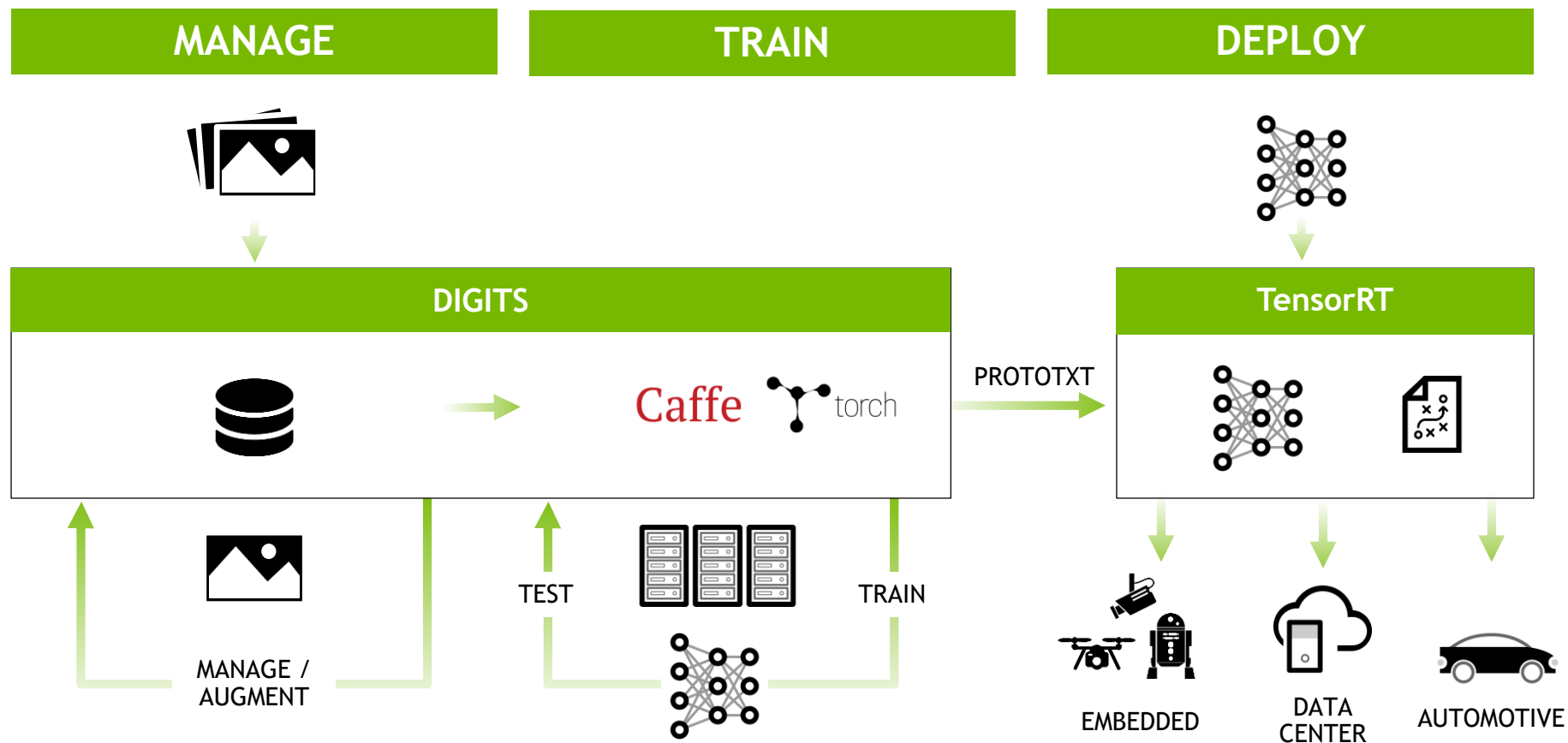


END-TO-END AUTONOMOUS DRIVING NETWORK



GPU DEEP LEARNING COMPUTING MODEL

A COMPLETE DEEP LEARNING PLATFORM



DEEP LEARNING PLATFORMS

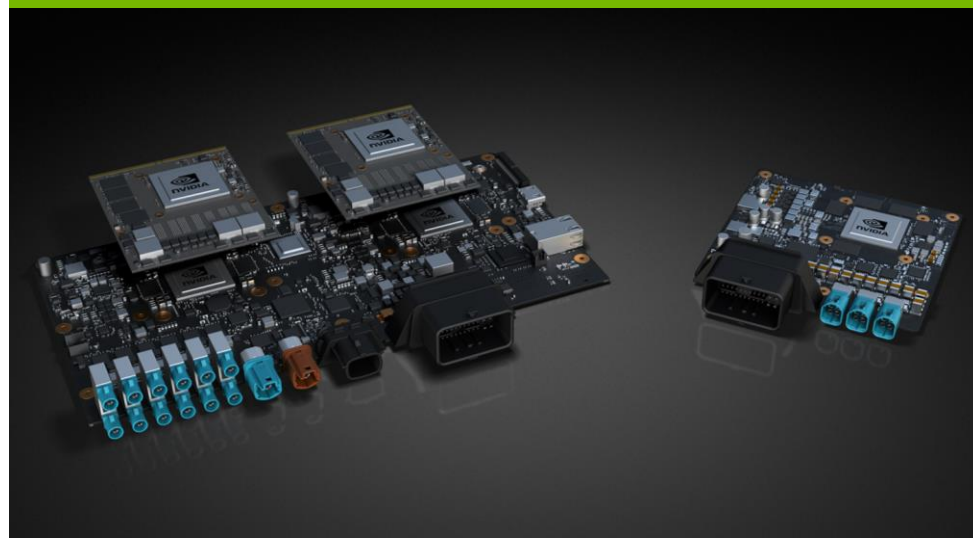
From Training to Development and Production

TRAINING



Nvidia DGX-1
with Tesla V100 (DGX-1V)

DEPLOY



DRIVE PX 2
2 PARKER + 2 PASCAL GPU
20 TOPS DL
120 SPECINT
80W

ONE
ARCHITECTURE



XAVIER
30 TOPS DL
160 SPECINT
30W

TENSOR CORE

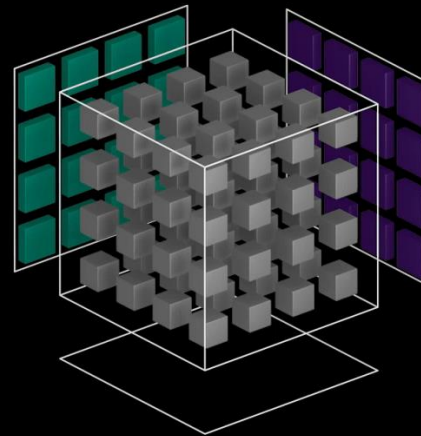
CUDA TensorOp instructions & data formats

4x4 matrix processing array

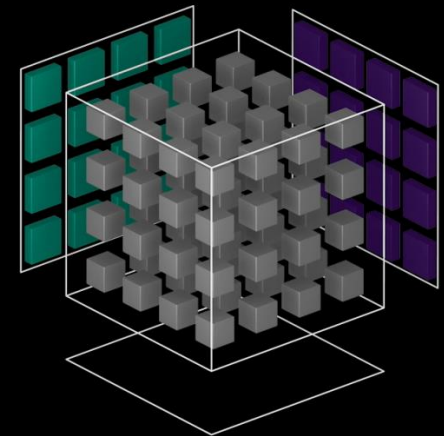
$D[\text{FP32}] = A[\text{FP16}] * B[\text{FP16}] + C[\text{FP32}]$


Optimized for deep learning

PASCAL



VOLTA TENSOR CORES



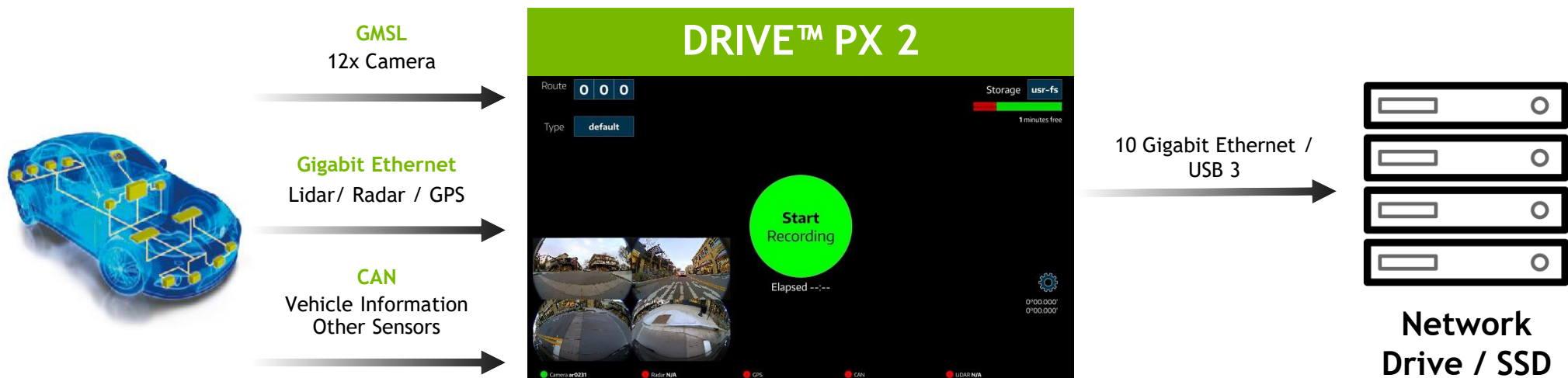
 Activation Inputs

 Weights Inputs

 Output Results

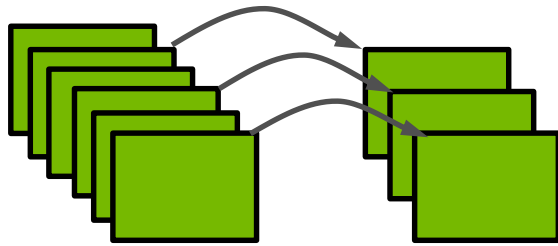
DATASET CREATION

DATA ACQUISITION



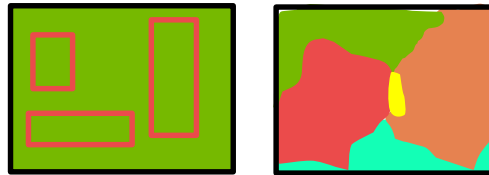
DATASET CREATION

DATA CURATION



Filter and keep data of interest

DATA ANNOTATION



Bounding boxes, per pixel labeling

START FROM TRAINED NETWORK

May reduce required data size

TRAINING DEEP NEURAL NETWORKS

NVIDIA DIGITS

Interactive Deep Learning GPU Training System

MANAGE DATA

MANAGE DATA

CONFIGURE NEURAL NET

CONFIGURE NEURAL NET

MONITOR TRAINING

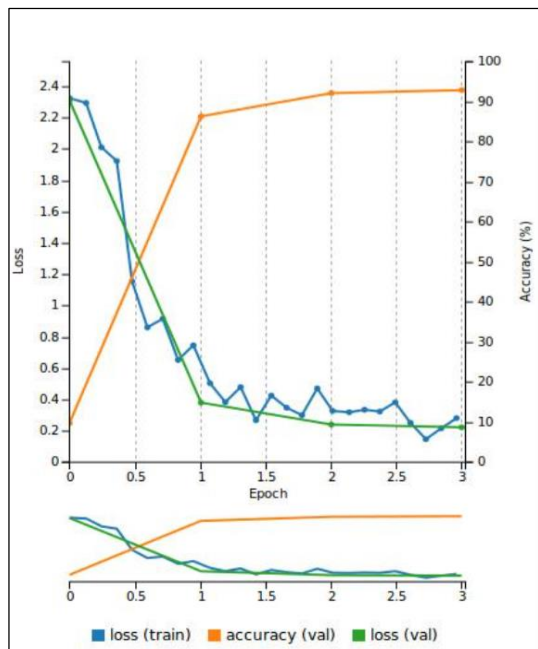
MONITOR TRAINING

VISUALIZE RESULTS

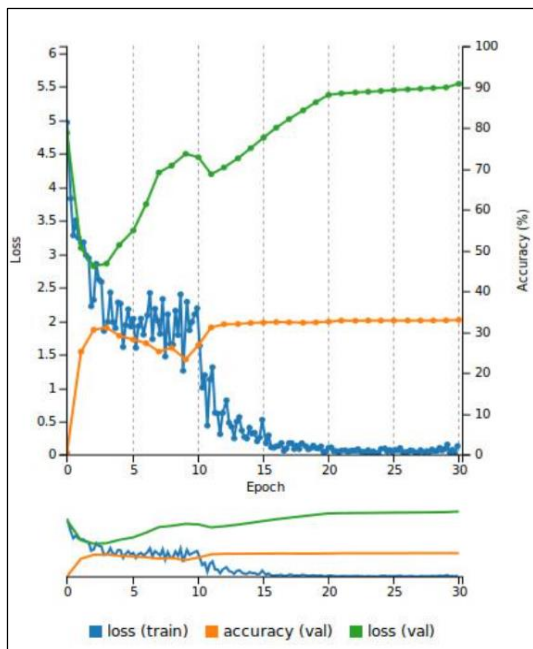
VISUALIZE RESULTS

NVIDIA DIGITS

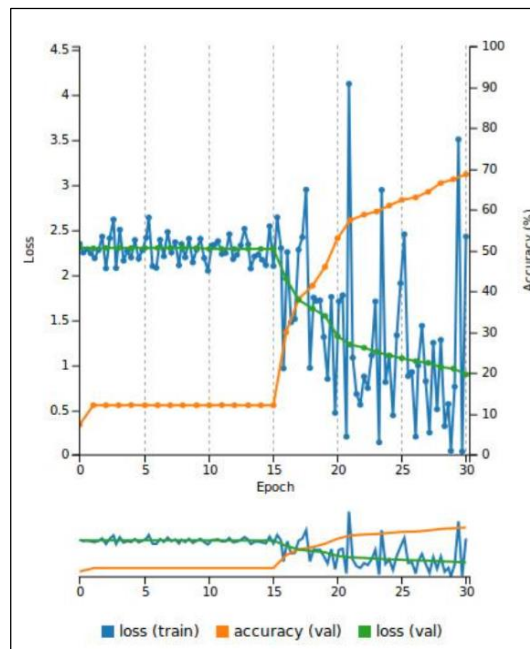
Monitor Training



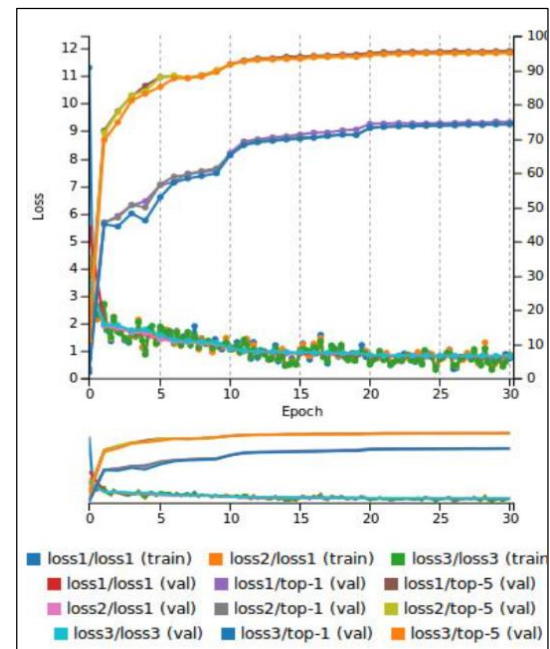
WELL BEHAVED



OVERFIT



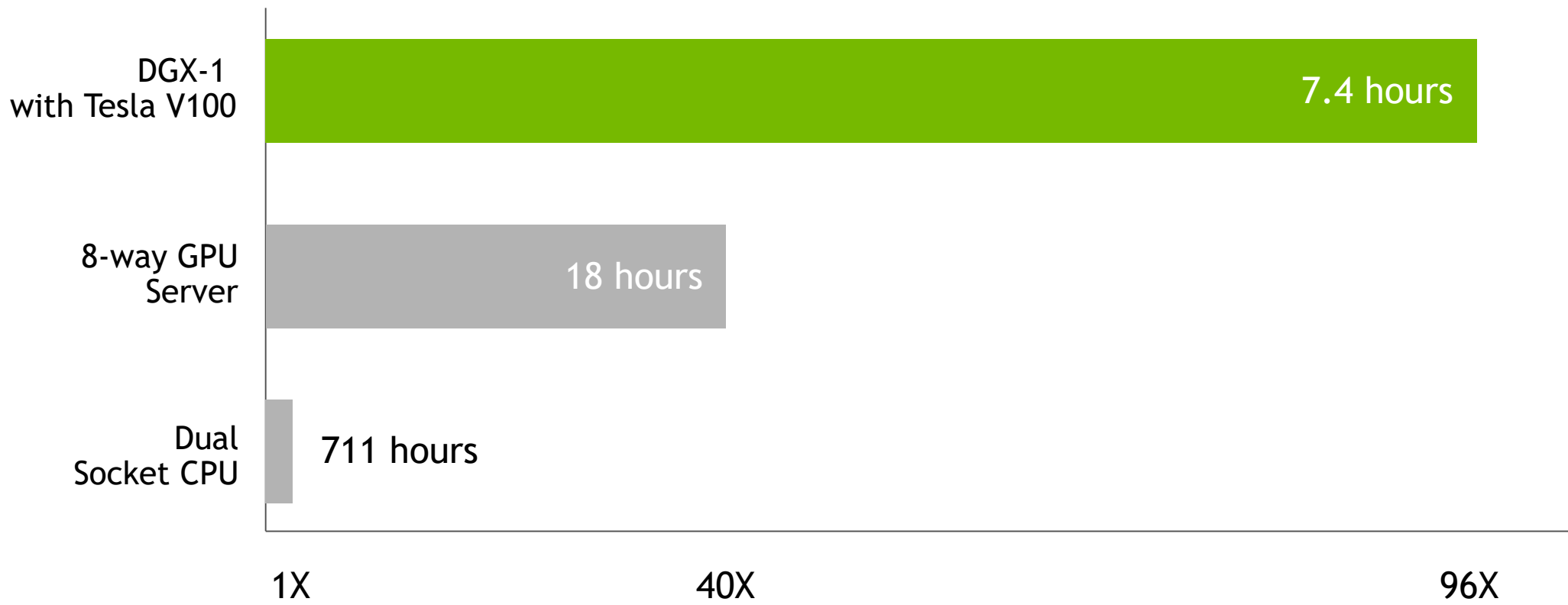
ILL BEHAVED



MANY OUTPUTS

DNN TRAINING

Iterate and Innovate Faster



Workload: ResNet50, 90 epochs to solution | CPU Server: Dual Xeon E5-2699 v4, 2.6GHz

DNN INFERENCE OPTIMIZATIONS

DNN INFERENCE OPTIMIZATIONS

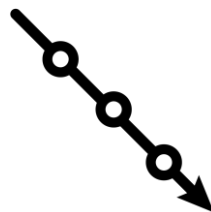
HARDWARE ACCELERATIONS

Specialized instructions for deep learning operations



PRUNING

Prune down the network size (neurons + connections) to reduce inference time



TensorRT

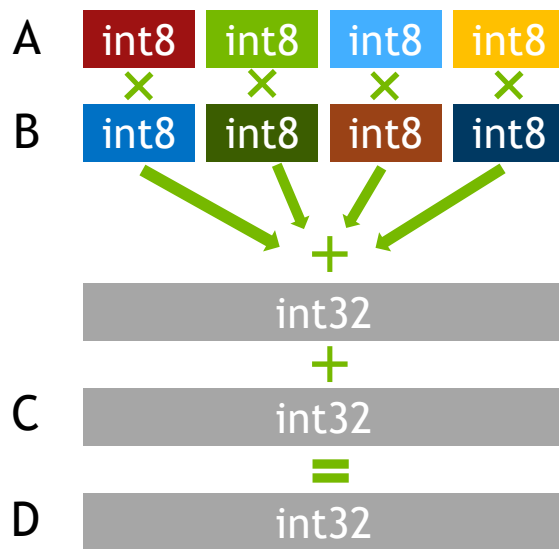
Accelerated neural network inference engine



HARDWARE ACCELERATIONS

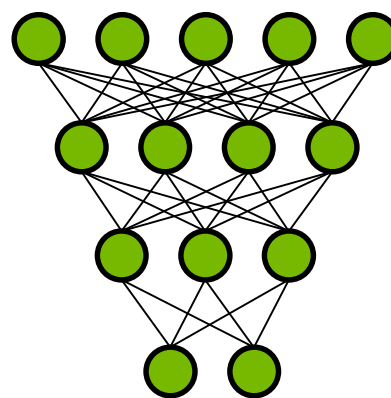
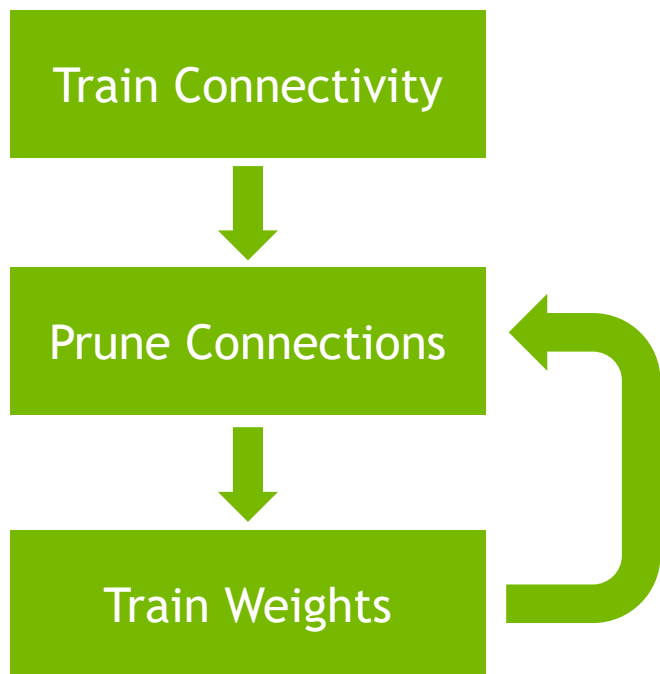
Specialized Instruction for Deep Learning Operations

DP4A



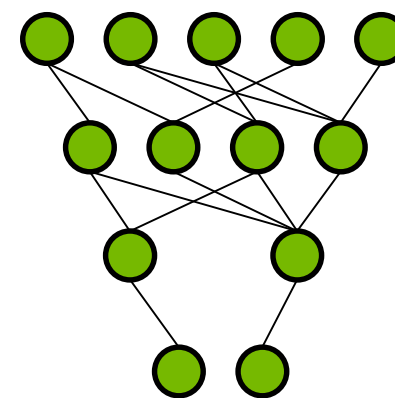
INT8 dot product

PRUNING



Before Pruning

Pruning Synapses →
Pruning Neurons →



After Pruning

PRUNING

NETWORK	TOP-1 ERROR	TOP-5 ERROR	PARAMETERS	COMPRESSION RATE
LeNet-300-100 Ref	1.64%	-	262K	12x
LeNet-300-100 Pruned	1.59%	-	22K	
LeNet-5 Ref	0.80%	-	431K	12x
LeNet-5 Pruned	0.77%	-	36K	
AlexNet Ref	42.78%	19.73%	61M	9x
AlexNet Pruned	42.77%	19.67%	6.7M	
VGG-16 Ref	31.50%	11.32%	138M	13x
VGG-16 Pruned	31.34%	10.88%	10.3M	

TensorRT

High-performance framework makes it easy to develop GPU-accelerated inference

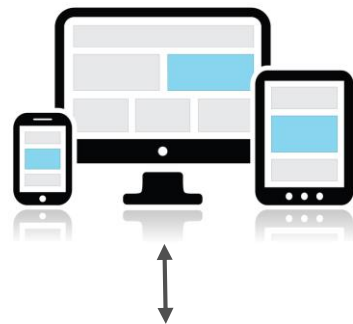
Production deployment solution for deep learning inference

Optimized inference for a given trained neural network and target GPU

Solutions for Hyperscale, ADAS, Embedded

Supports deployment of fp32,fp16,int8* inference

* int8 support will be available from v2



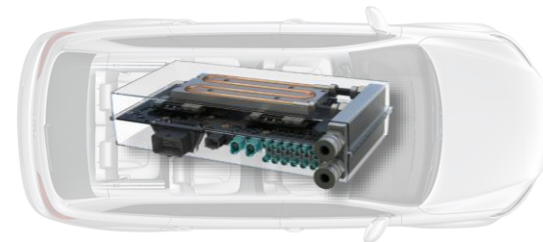
TensorRT for Data Center

Image Classification	Object Detection	Image Segmentation
----------------------	------------------	--------------------



TensorRT for Automotive

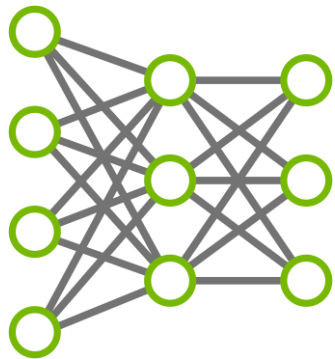
Pedestrian Detection	Lane Tracking	Traffic Sign Recognition
----------------------	---------------	--------------------------



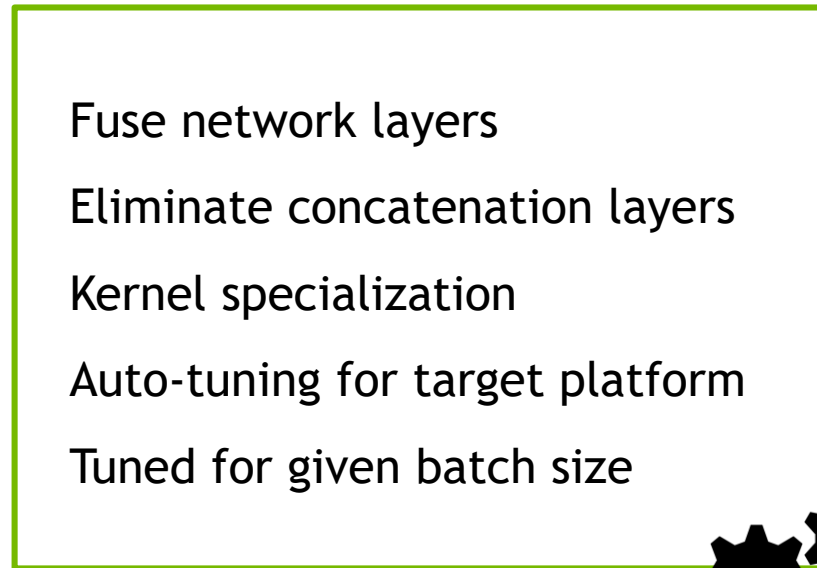
NVIDIA DRIVE PX 2

TensorRT

Optimizations



TRAINED
NEURAL NETWORK

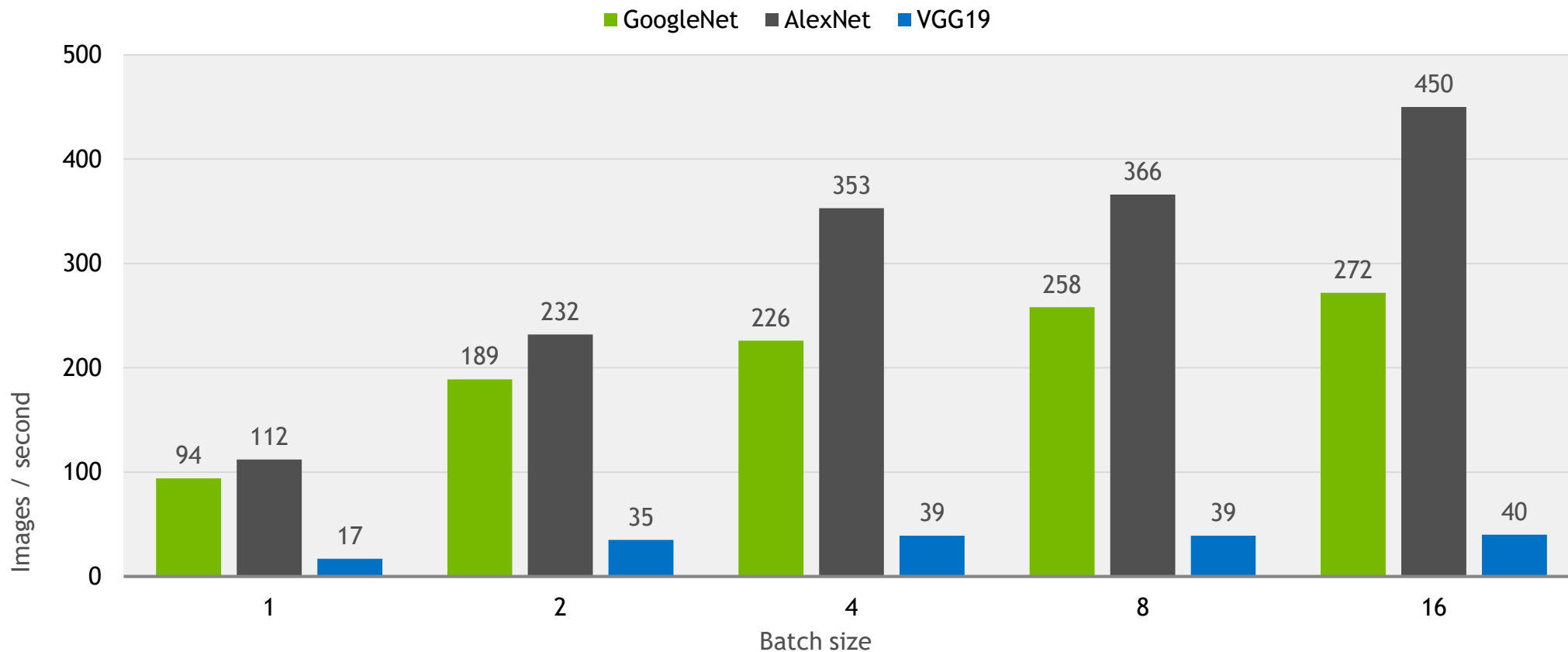


OPTIMIZED
INFERENCE
RUNTIME



TensorRT – iGPU (FP16)

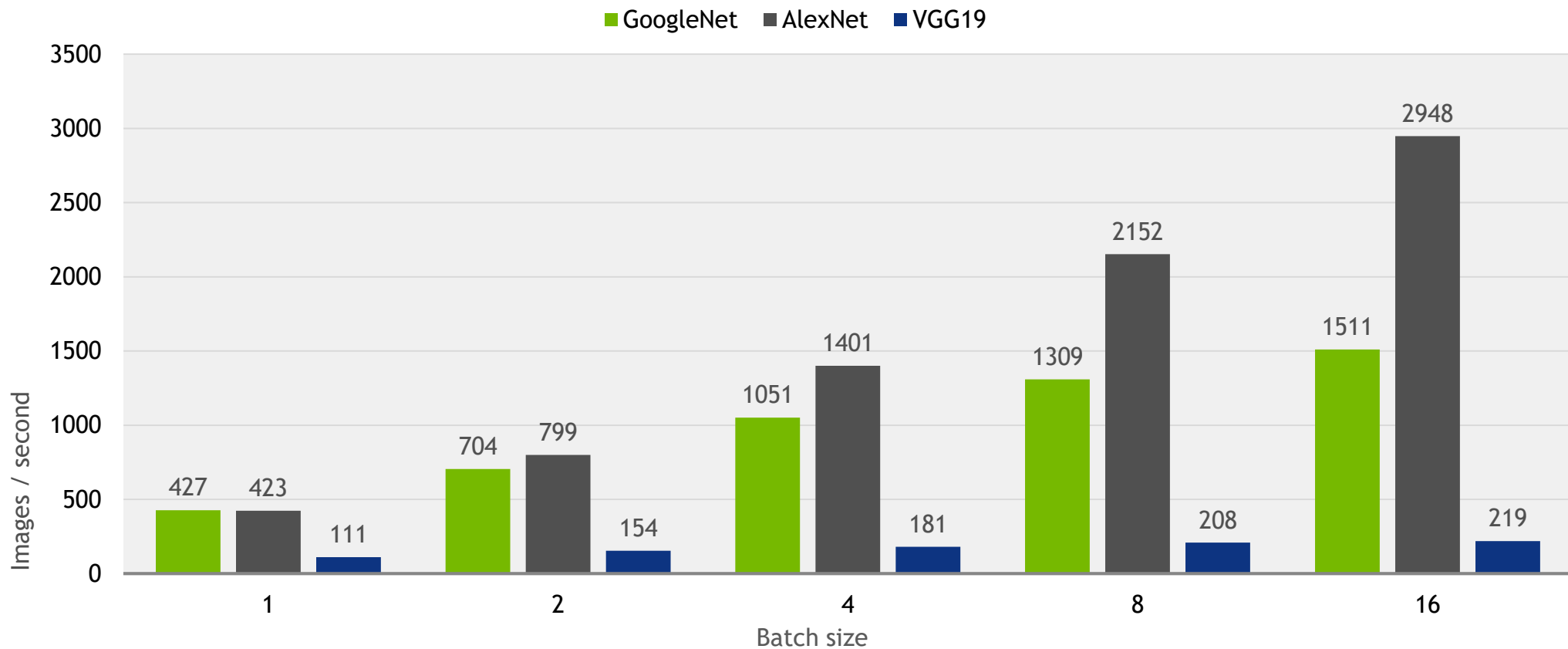
GoogleNet, AlexNet, VGG19



Configuration:
HW (DRIVE PX 2 iGPU@1275 MHz), SW (PDK ALPHA 2.0, TensorRT 1.0RC), GoogleNet & VGG19 Input Image Resolution (224x224)
AlexNet Input Image Resolution (227x227)

TensorRT – dGPU (INT8)

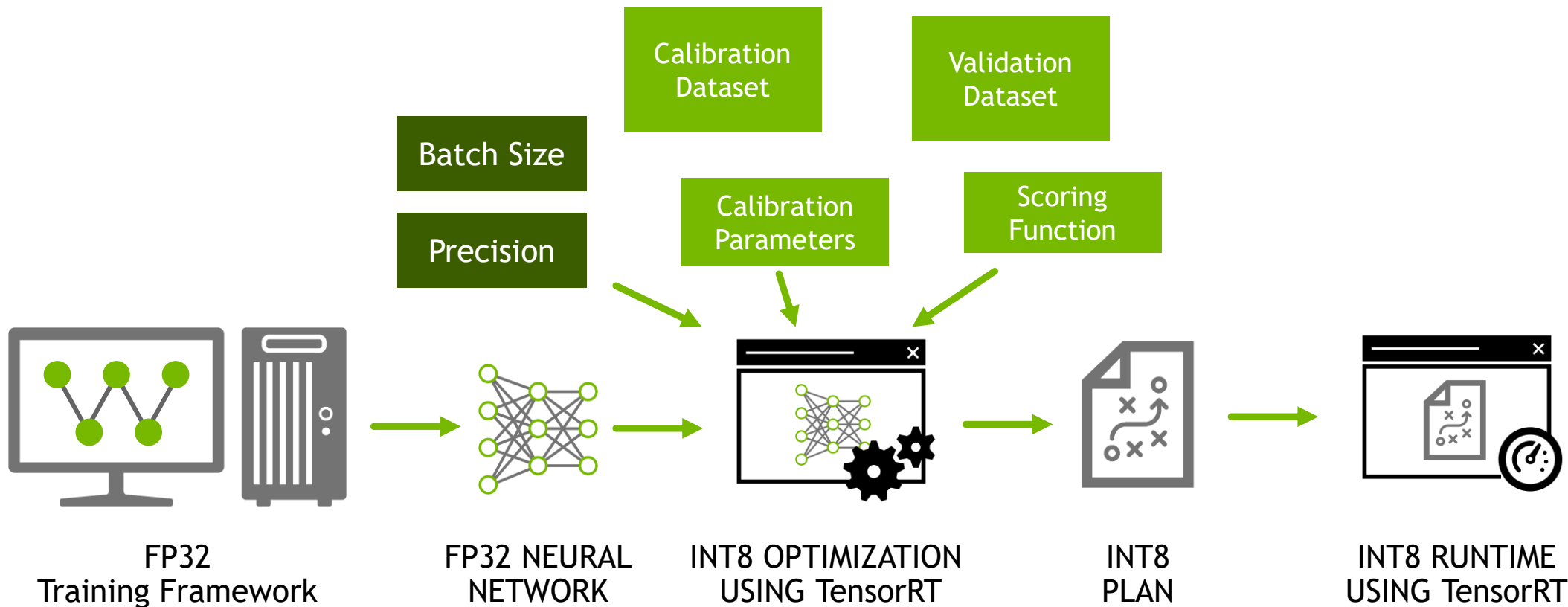
GoogleNet, AlexNet, VGG19



Configuration:
HW (DRIVE PX 2 dGPU@1290 MHz), SW (PDK ALPHA 2.0, TensorRT 2.0EA), GoogleNet & VGG19 Input Image Resolution (224x224)
AlexNet Input Image Resolution (227x227)

TensorRT

INT8 Workflow



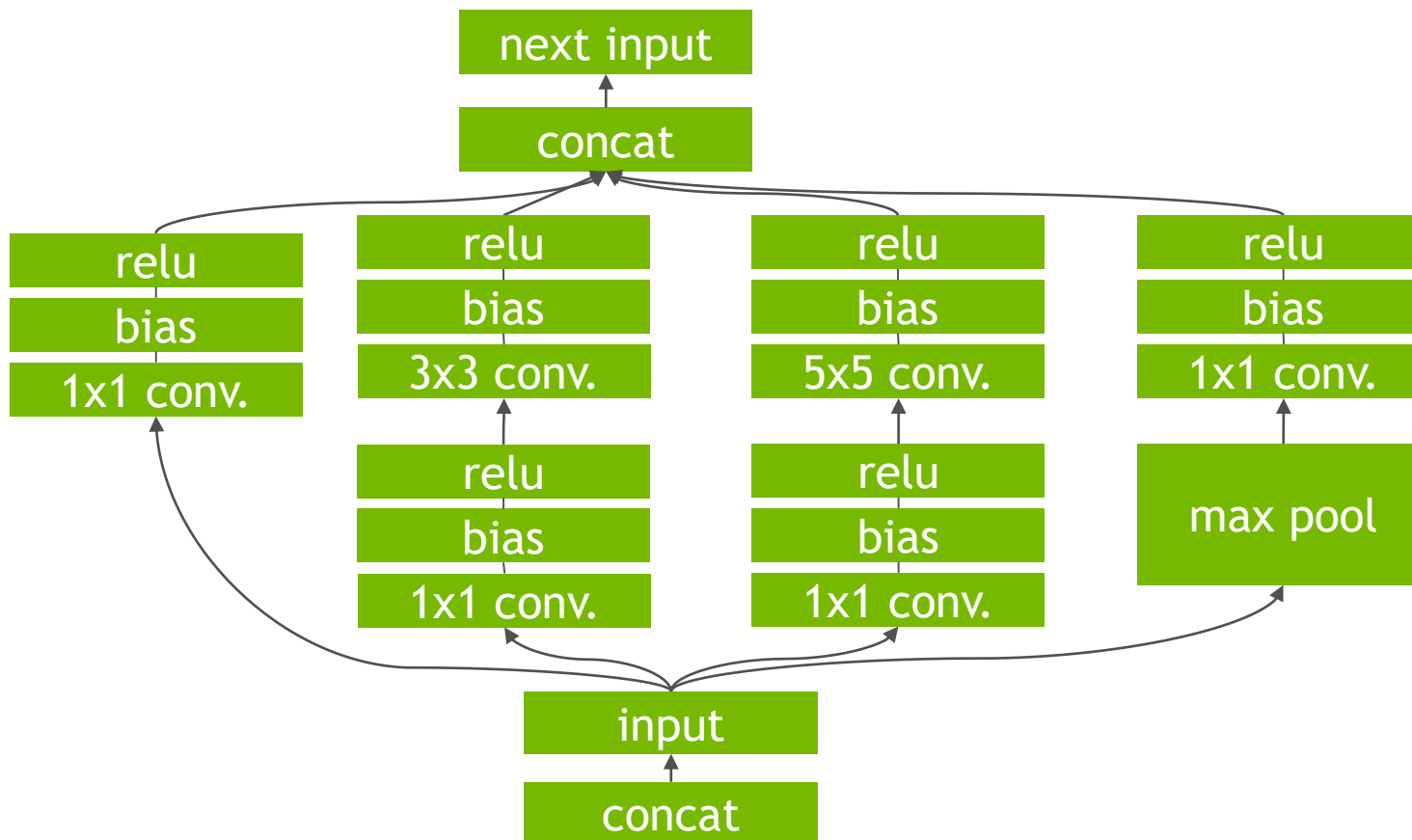
TensorRT

8-bit Inference: Top-1 Accuracy

NETWORK	FP32 TOP1	INT8 TOP1	DIFFERENCE	PERF GAIN
AlexNet	57.22%	56.96%	0.26%	3.70x
GoogLeNet	68.87%	68.49%	0.38%	3.01x
VGG	68.56%	68.45%	0.11%	3.23x
Resnet-152	75.18%	74.56%	0.61%	3.42%

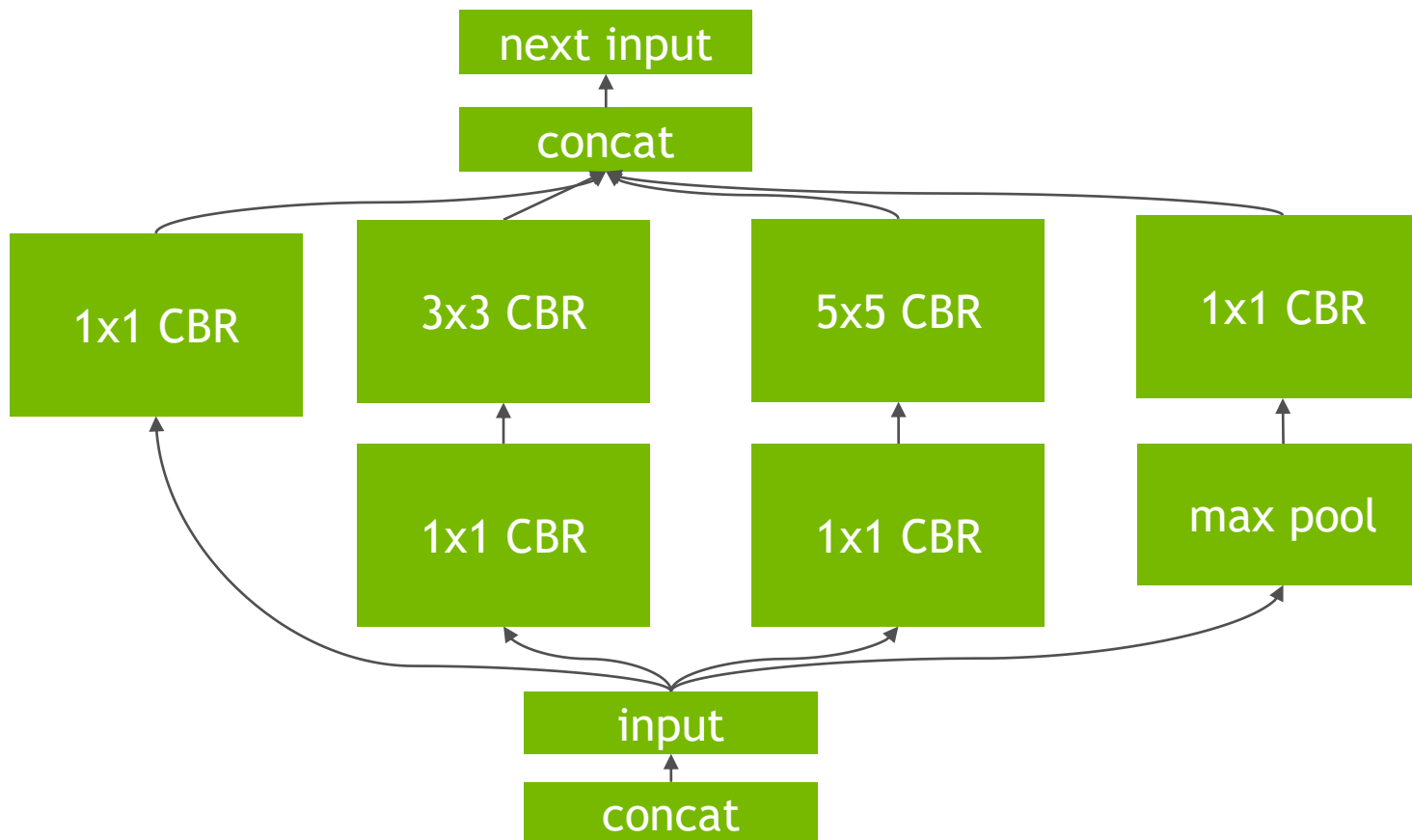
TensorRT – GRAPH OPTIMIZATION

Unoptimized Network



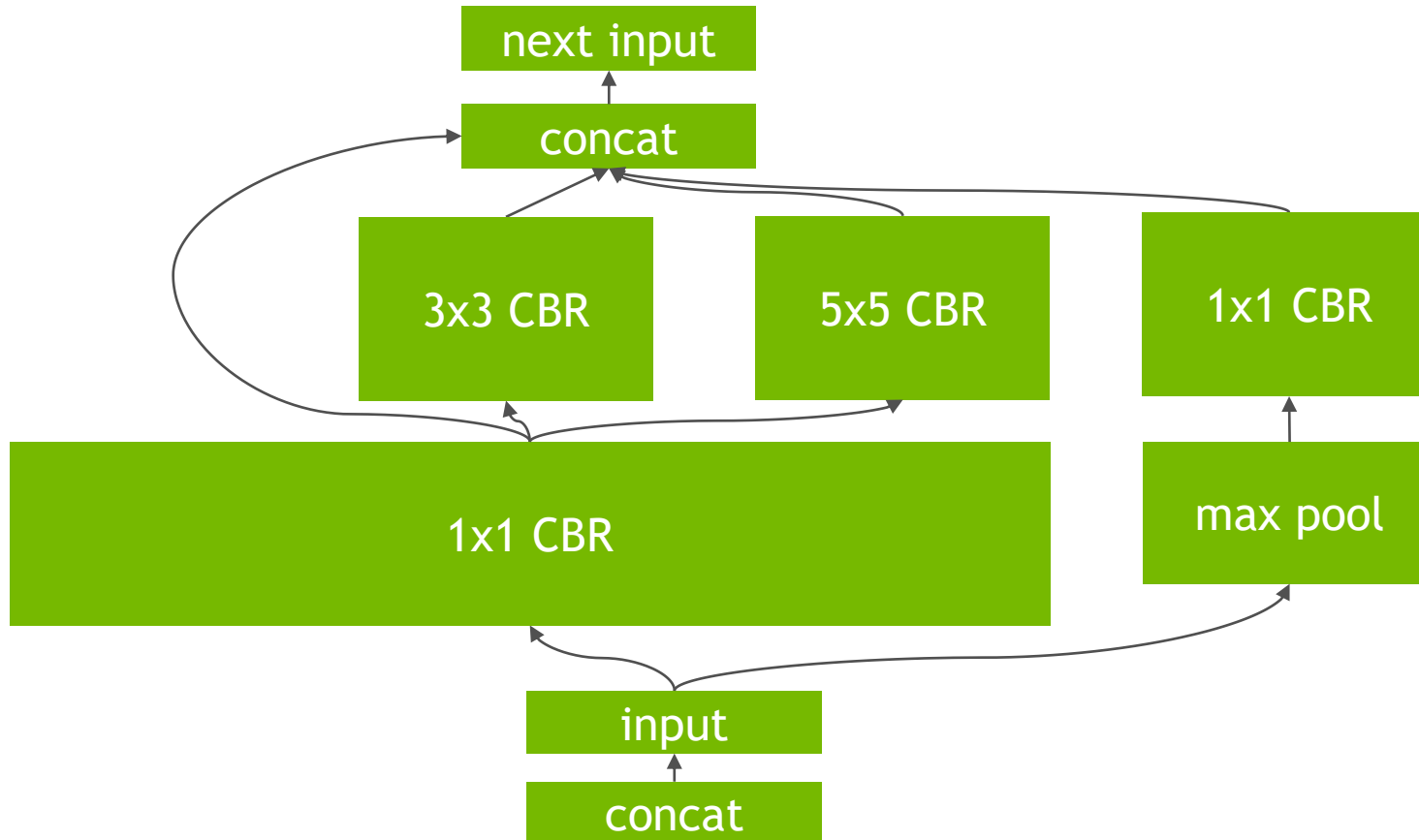
TensorRT – GRAPH OPTIMIZATION

Vertical Fusion



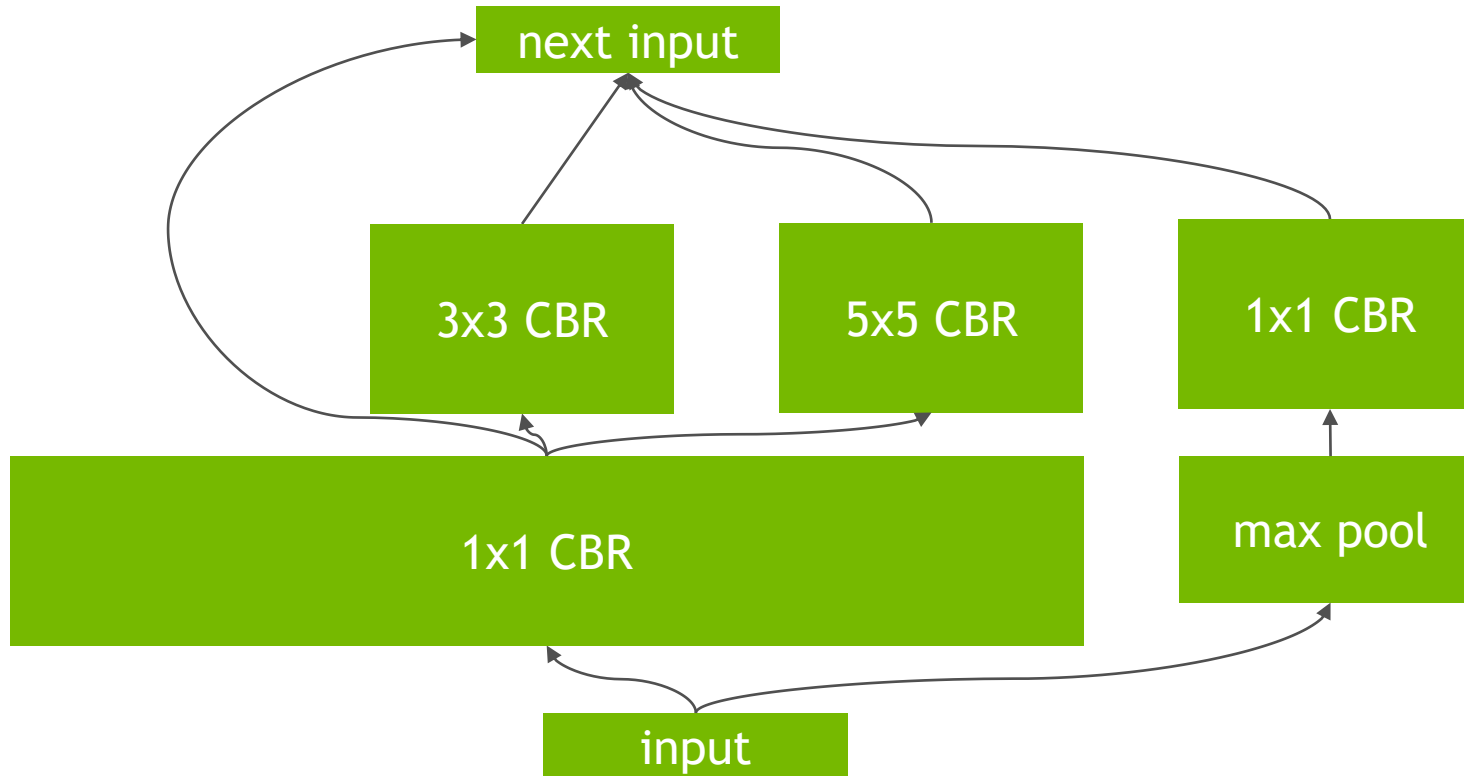
TensorRT – GRAPH OPTIMIZATION

Horizontal Fusion



TensorRT – GRAPH OPTIMIZATION

Concat Elision



AUTONOMOUS DRIVING CHALLENGES

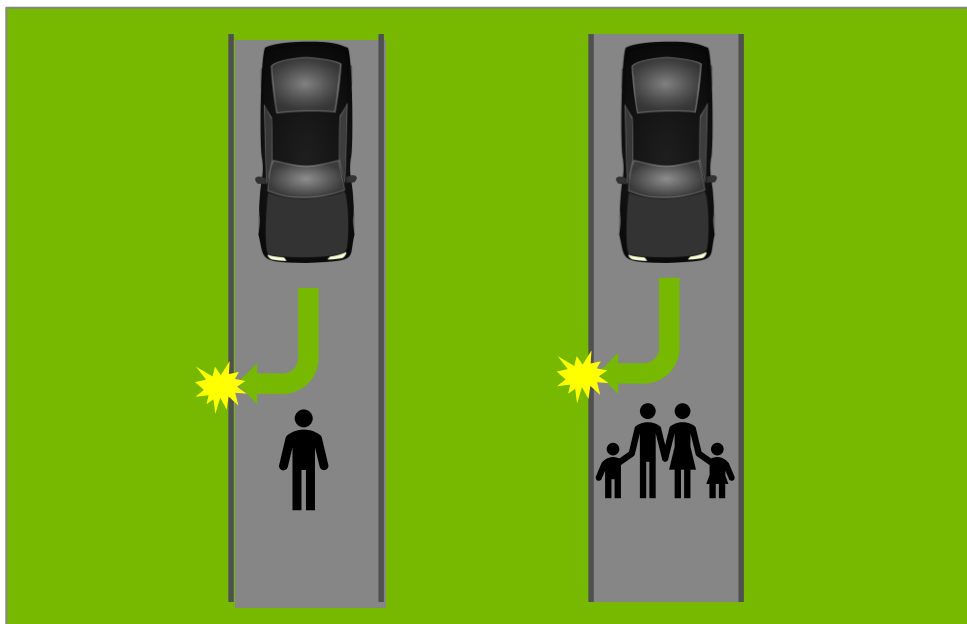
AUTONOMOUS DRIVING

Challenges



AUTONOMOUS DRIVING

Challenges



PUTTING IT ALL TOGETHER

RESOURCES

NVIDIA DRIVE Platform

<https://developer.nvidia.com/drive>

NVIDIA DGX-1

<http://www.nvidia.com/object/volta-architecture-whitepaper.html>

NVIDIA DIGITS

<https://www.nvidia.com/en-us/data-center/dgx-1/>

Volta Architecture Whitepaper

<https://developer.nvidia.com/digits>

TensorRT

<https://developer.nvidia.com/tensorrt>

QUESTIONS?

