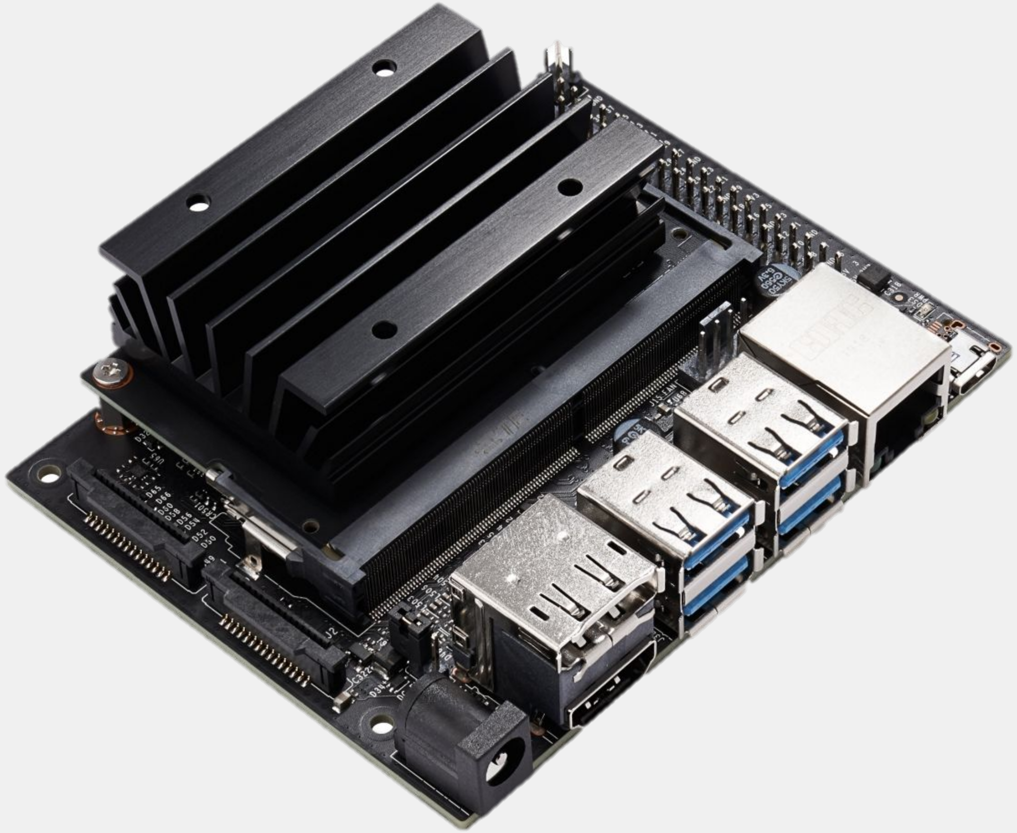
AUGMENTED STARTUPS
Computer Vision | AI | Robotics

TensorRT Optimization NVIDIA JETSON



Overview

AUGMENTED STARTUPS
Computer Vision | AI | Robotics

- Starting with Jetson
- Setting up AI Jetson
- Basics of Computer Vision
- Object Detection and Its Application
- Object Detection on custom dataset

PRO

- Model optimization using TensorRT
- Introduction to DeepStream
- DeepStream multiple camera synchronization

- Real-life challenges
- Number plate recognition on Jetson
- Human Pose estimation
- Face Recognition and Attendance system

Content

- About TensorRT
- Why TensorRT
- Model optimization using TensorRT
- Factors involved in model optimization

TensorRT

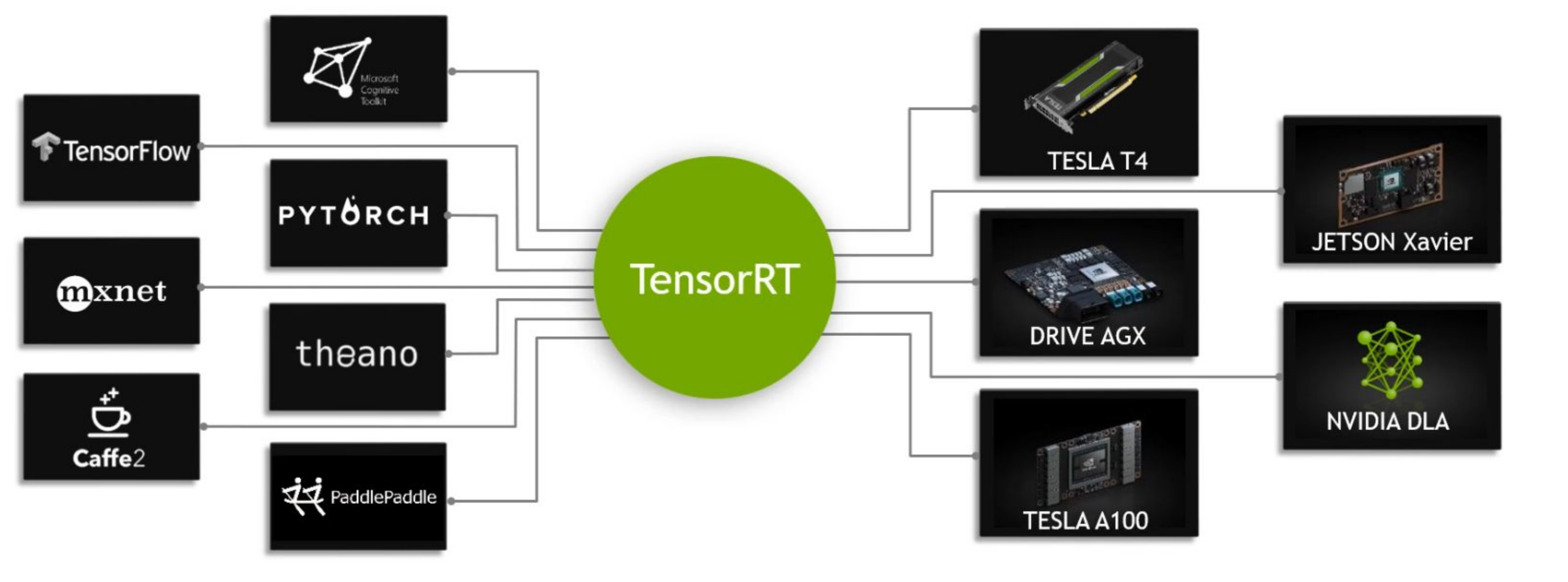
- High Performance SDK for DL inference
- Introduced by NVIDIA
- Built on CUDA
- Supportive for Real-time applications
- Compatible with all NVIDIA devices

Why TensorRT

- Best Inference Framework for NVIDIA GPUs
  - Speed and Memory
- 4 to 5 times faster inference
- Platform portability

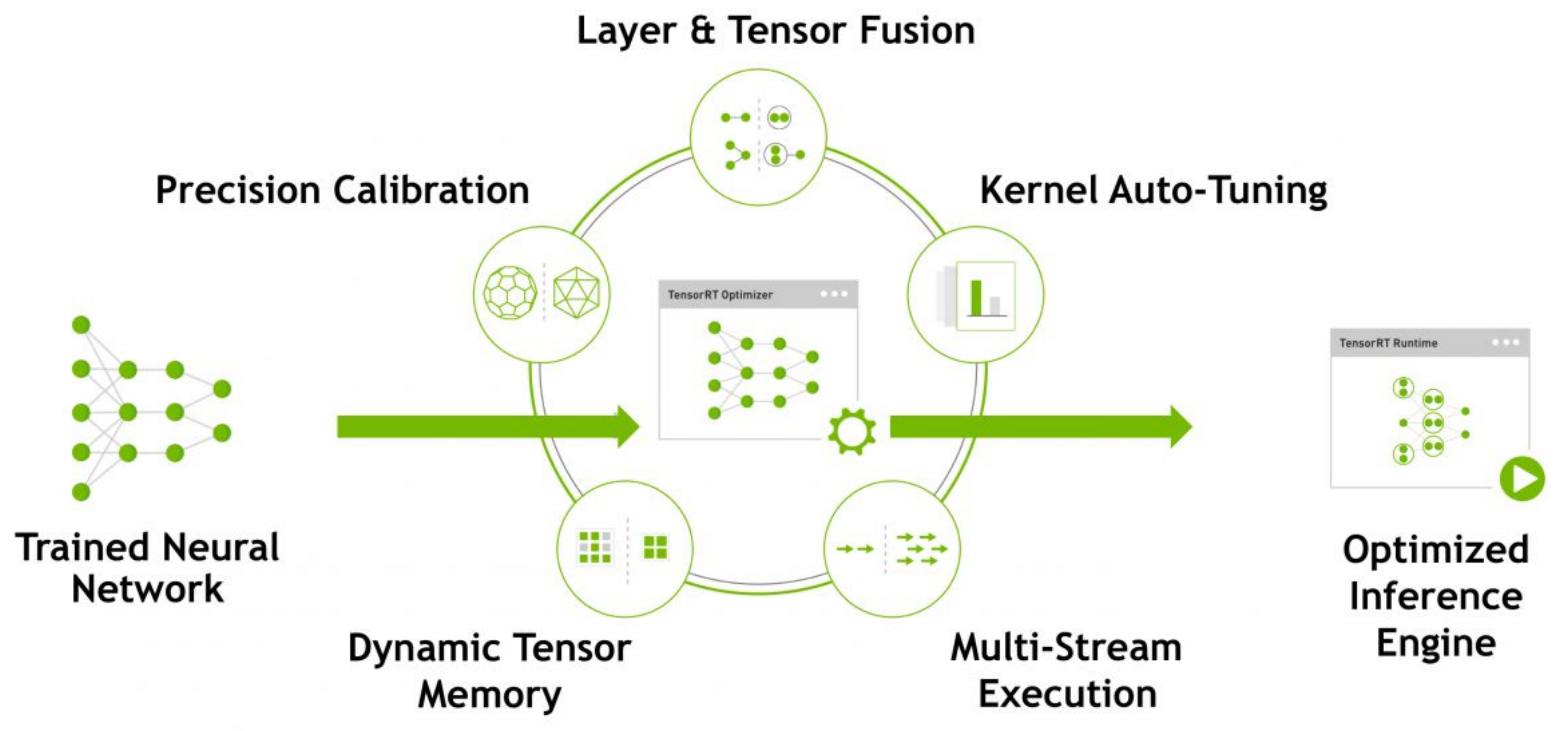
From Many to One

- Compatible with all NVIDIA GPU Devices



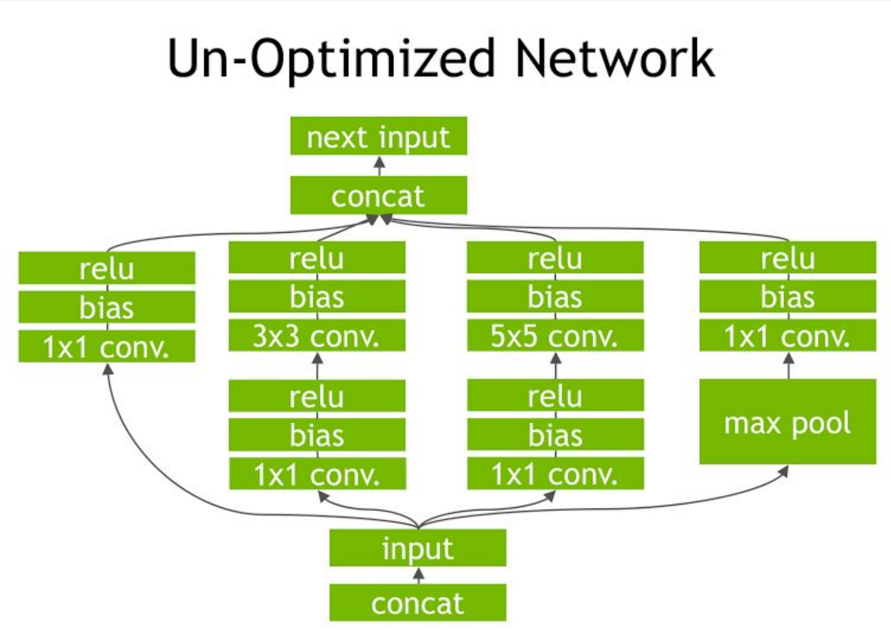
TensorRT Optimization

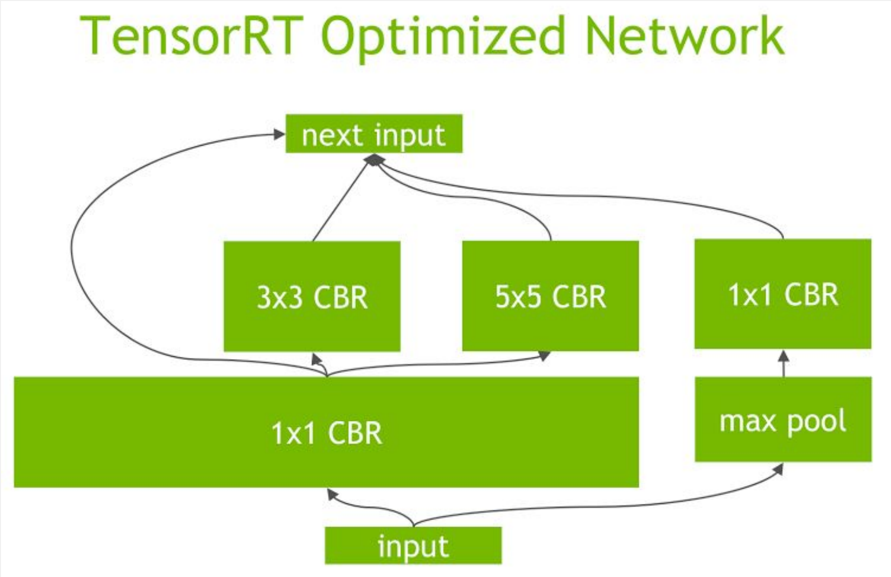
- TRT implements 5 technologies for optimization



Layer and Tensor Fusion

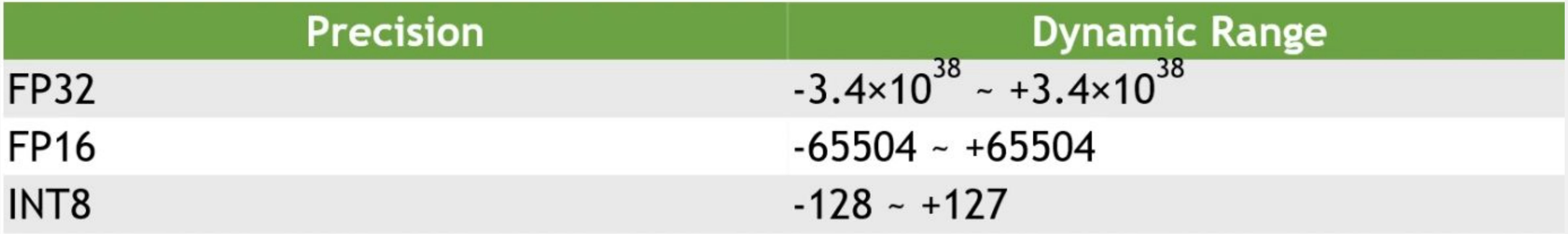
- Less Kernel Launch
- Better memory usage
- Combine Sequential Kernels
- Combine same Kernels
  - Common input but different weights





Precision Calibration

- DNN models are trained at FP32
- Converted to FP16 or INT8
- Lower memory reduces latency
  - Smaller size, higher throughput



Precision	Dynamic Range
FP32	$-3.4\times10^{38} \sim +3.4\times10^{38}$
FP16	-65504 ~ +65504
INT8	-128 ~ +127

Kernel Auto-Tuning

- Avoid execution of multiple algorithms
- Choose the optimal kernel
  - batch size, filter-size etc.
- Kernel selection is based on target platform

Dynamic Tensor Memory

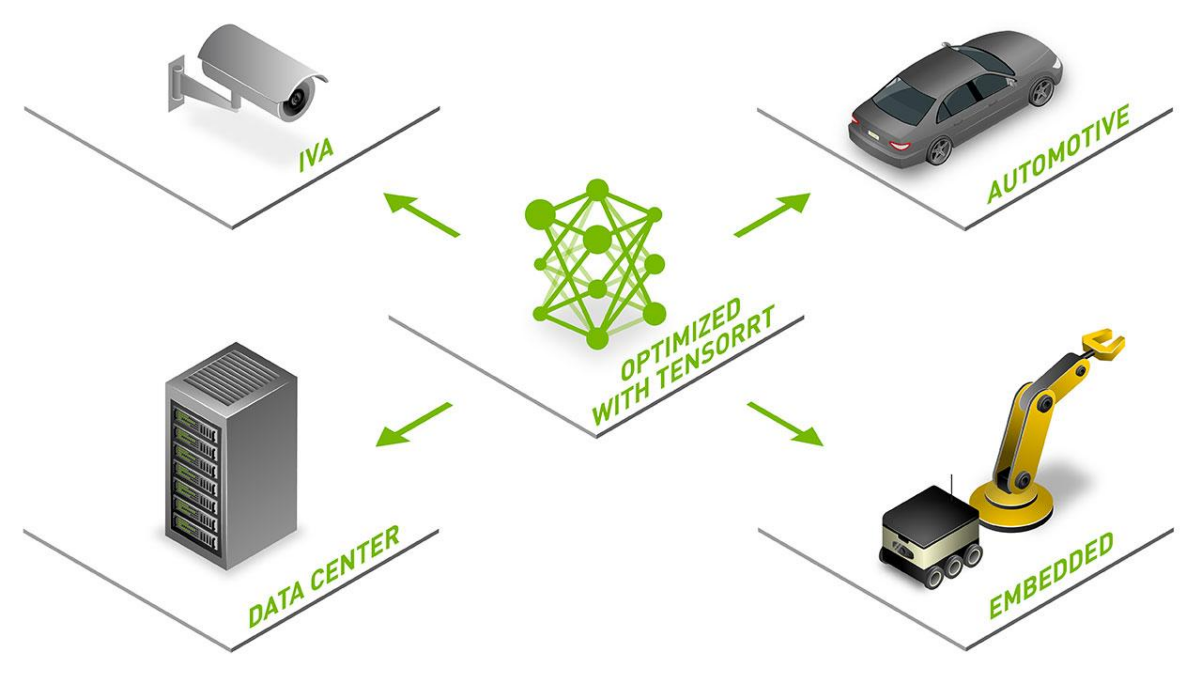
- Improves memory reuse
- Allow memory for the duration of usage
- Reduces memory footprint

Multi-Stream Execution

- Execute multiple inputs in parallel
- Parallel execution is done by mean of CUDA

Accelerate Inference with TRT

- TRT optimize and deploy various applications
  - Data center, automotive environment etc.
- Integrated with application-specific SDKs
  - DeepStream, Merlin, Maxine etc.



Thank You