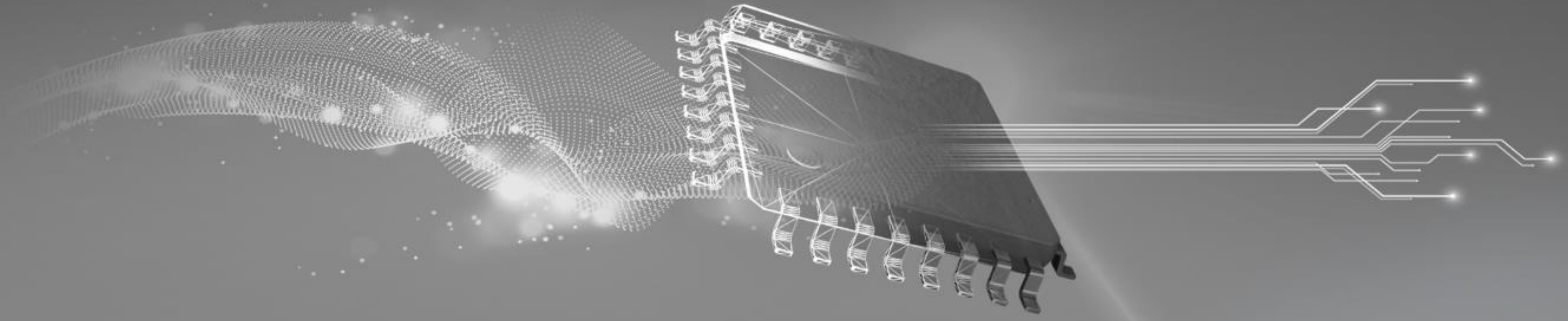
TI TECH DAYS



Deep Learning with Jacinto™ 7 SoCs: TDA4x

Tarkesh Pande

ADAS Business Unit

Agenda

- Introduction to Jacinto™ 7 processors

- Three step process
  - Network selection
  - Optimization
  - Deployment

- TI deep learning tools

- Open source frameworks

Introducing – Jacinto 7 processors

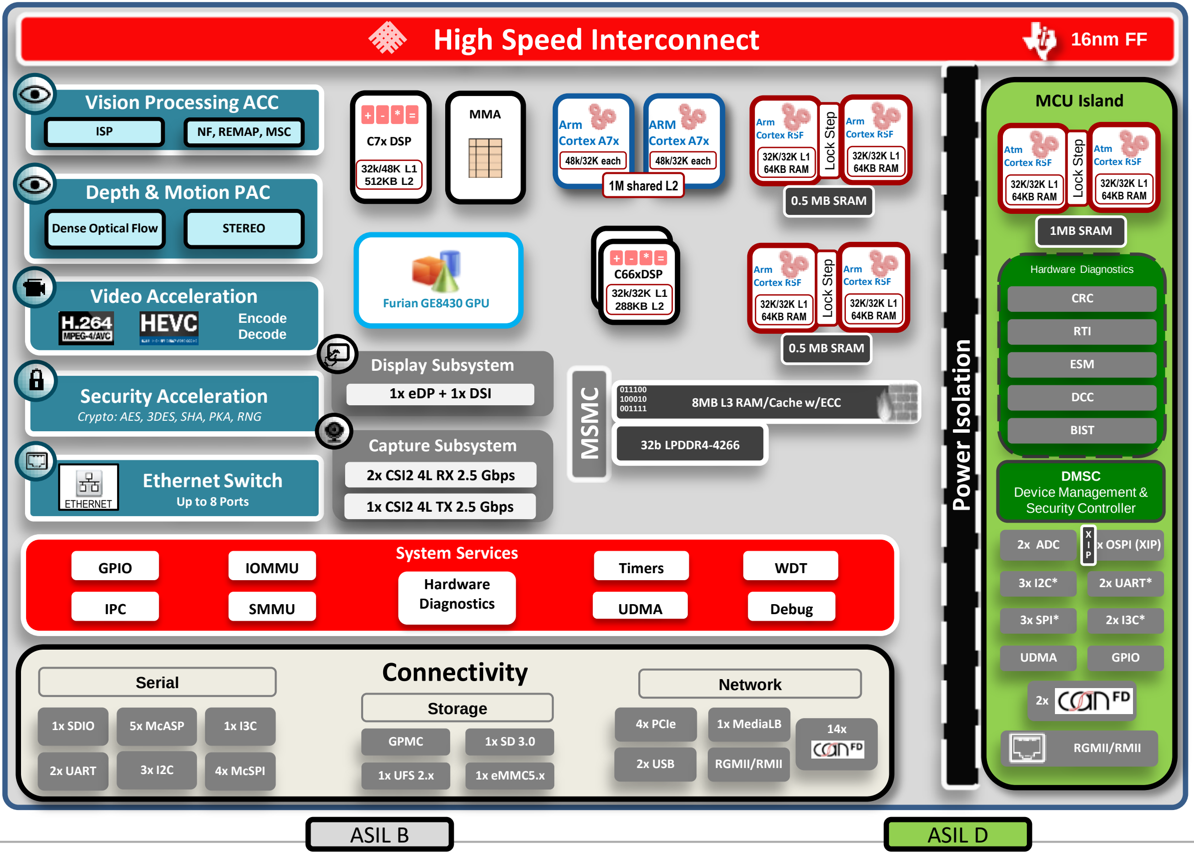
Accelerating key functions lowers power

- DSP for Computer Vision
- Vision Processing
- Video, Graphics
- Deep Learning

Industry's most efficient DL architecture

- Enables passively-cooling designs
- 90% utilization of deep learning accelerator due to smart memory system

Automotive Quality-ready process technology



Deep learning system performance on TDA4VM

5 simultaneous deep learning algorithms on 3x 1MP camera each @ ~16 fps

- Parking spot detection
- Vehicle detection
- Semantic segmentation
- Motion segmentation
- Depth estimation

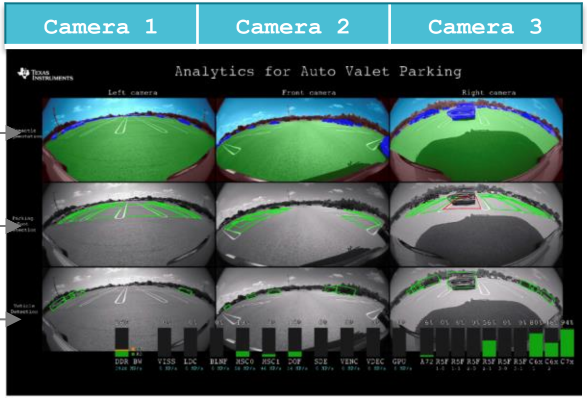
Resource loading

A72:	6%	**C7x+MMA:**	94%
1xR5:	56%	**DOF:**	18%
5xR5s:	free	**DDR BW:**	26%

Semantic segmentation

Parking spot detection

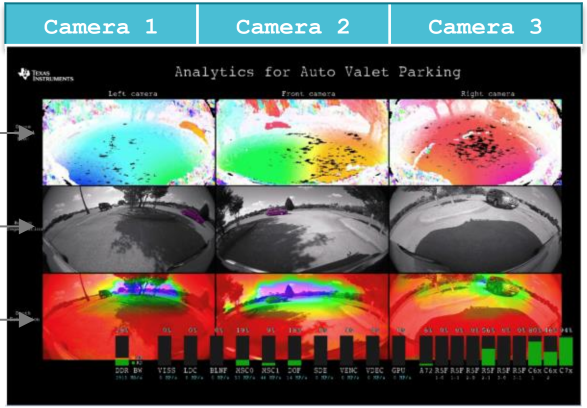
Vehicle detection



Dense optical flow

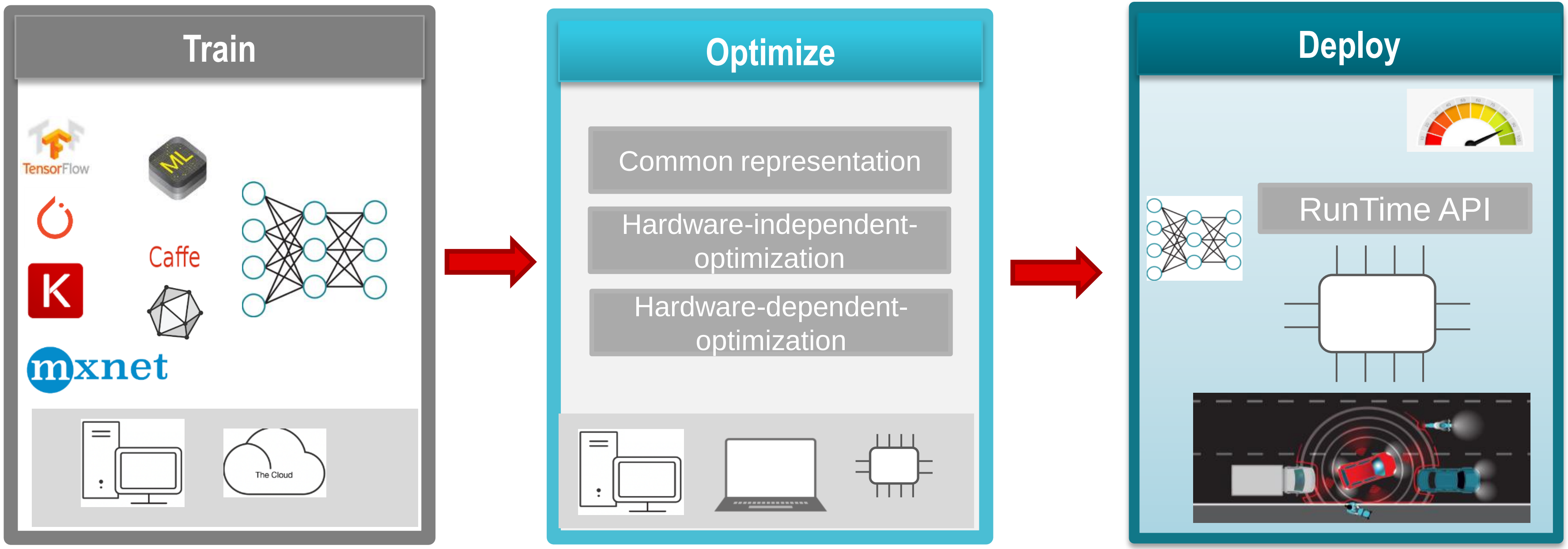
Motion segmentation

Depth estimation



Deep learning inference at the edge

TI Deep Learning (TIDL) library provides custom tools for high-performance fixed-point inference on TDA4x hardware accelerators

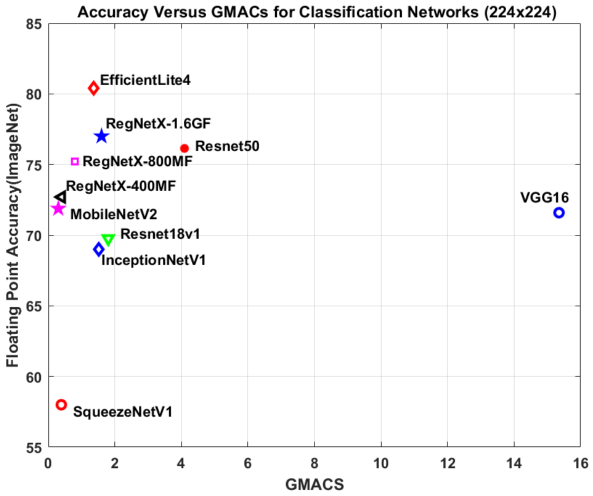


Training and network selection

- Choose and train network architecture that meet metrics
  - Accuracy and performance
  - Complexity is proportional to performance (time for execution)
- New architectures and ideas evolve quickly
  - Want to minimize time for test and deployment

Training and network selection

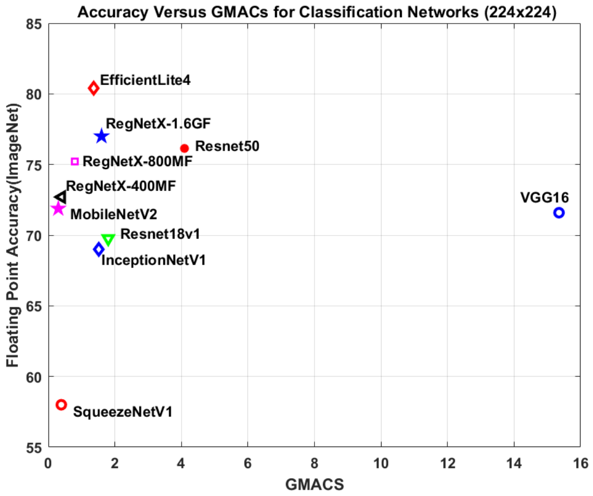
- Choose and train network architecture that meet metrics
  - Accuracy and performance
  - Complexity is proportional to performance (time for execution)
- New architectures and ideas evolve quickly
  - Want to minimize time for test and deployment



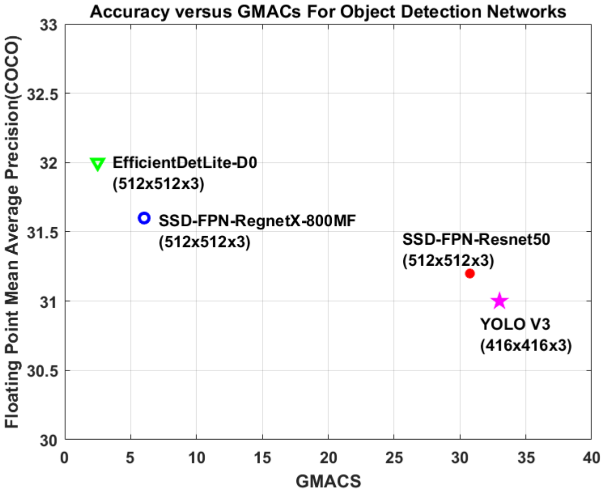
Classification

Training and network selection

- Choose and train network architecture that meet metrics
  - Accuracy and performance
  - Complexity is proportional to performance (time for execution)
- New architectures and ideas evolve quickly
  - Want to minimize time for test and deployment

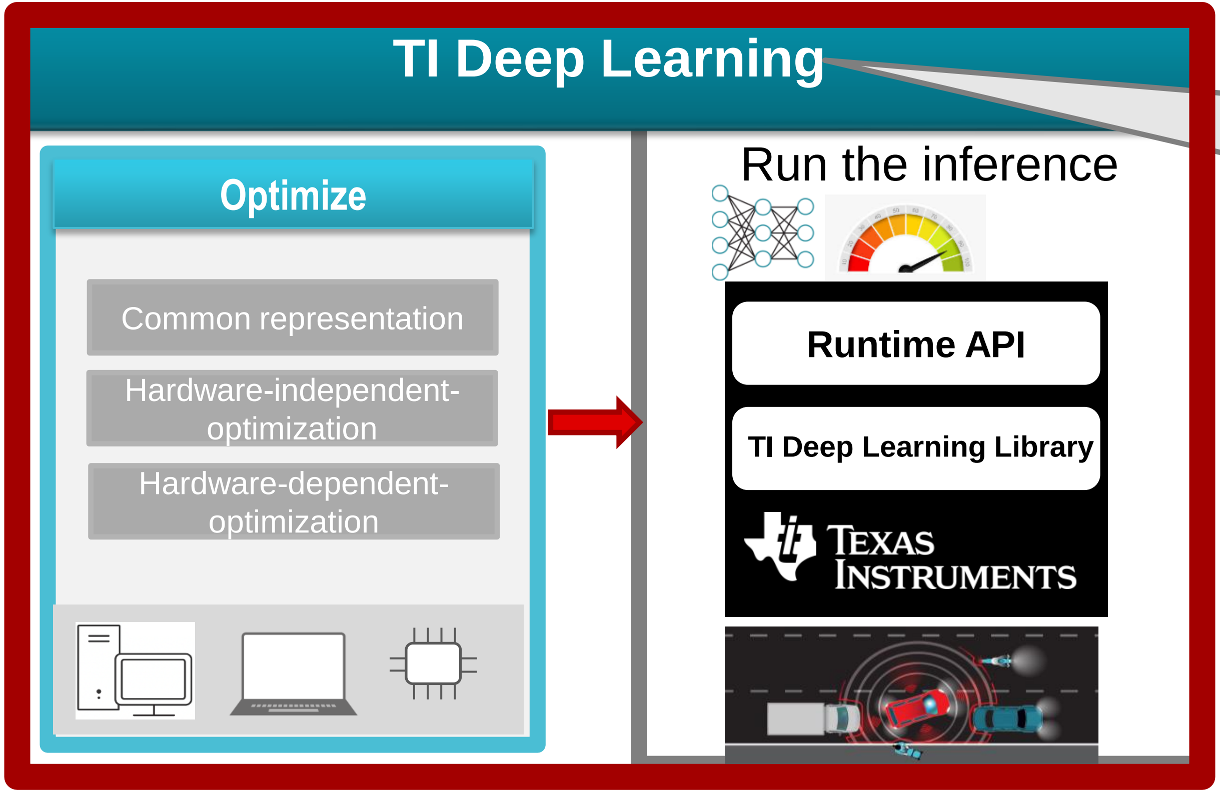


Classification



Object Detection

TI deep learning (TIDL)



- Enable high-performance inference at the edge on TI SoC

TI deep learning (TIDL) - model optimizer

Model Optimizer

Common representation

Hardware independent optimization

Hardware dependent optimization

Code generation

PC based tools for :

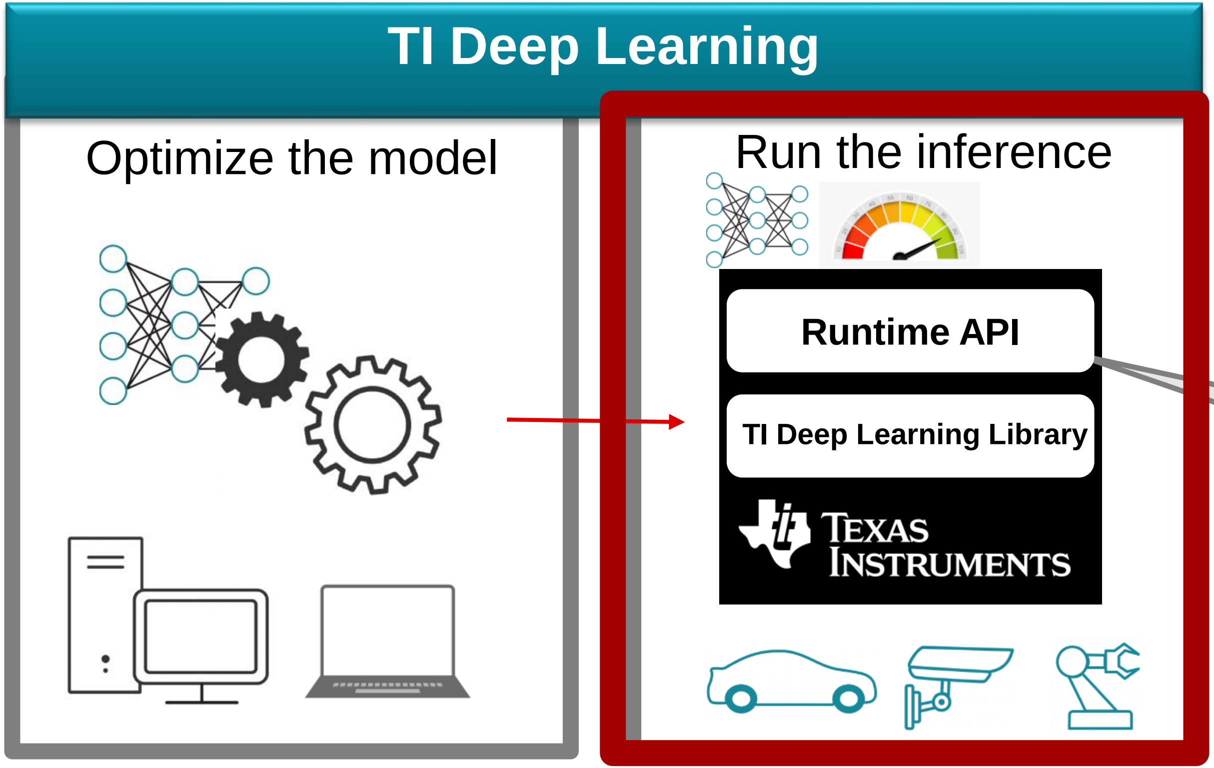
- ➢ Model optimizations
- ➢ Performance estimation
- ➢ Accuracy benchmarking

TensorFlow mxnet Caffe TensorFlowLite

- Quantization
- Fusion
- Constant folding
- simplifications, dead code elimination
- Layout transformation

- Block based processing
- Vectorization
- Memory management

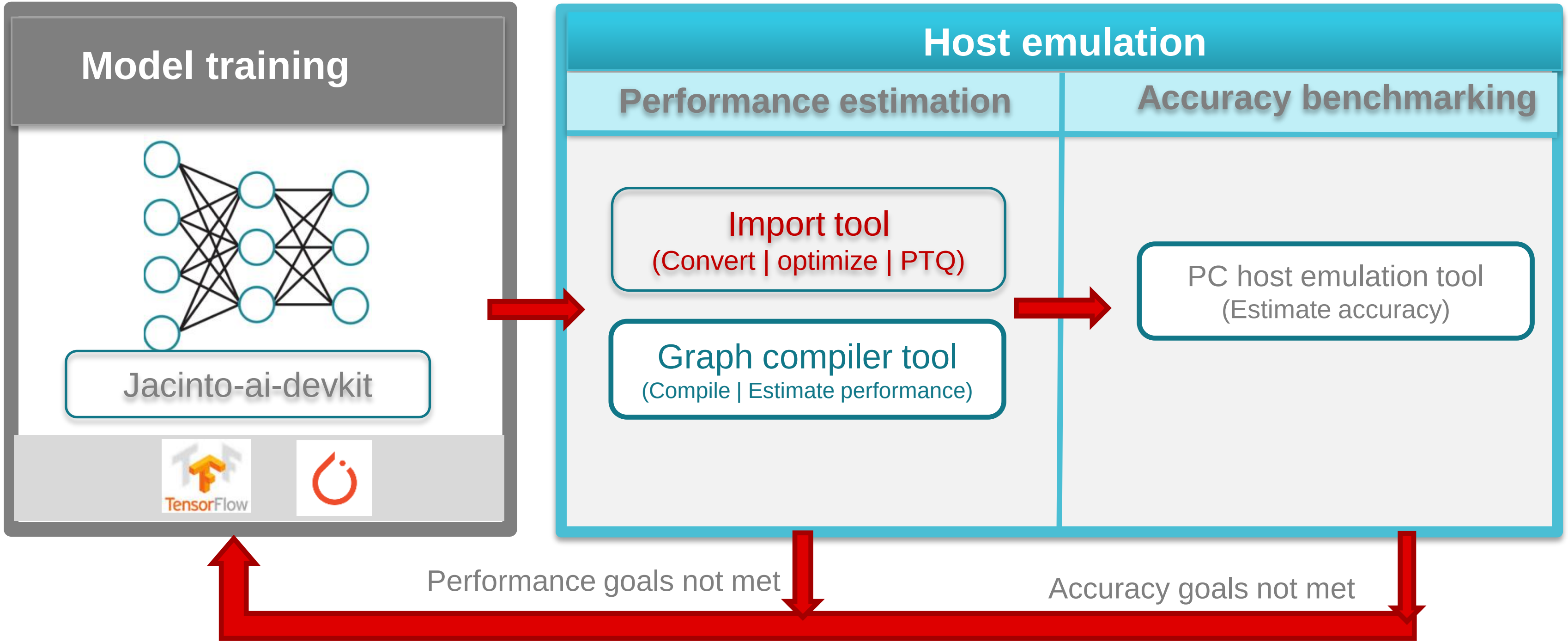
TI deep learning (TIDL)



Comprise of:

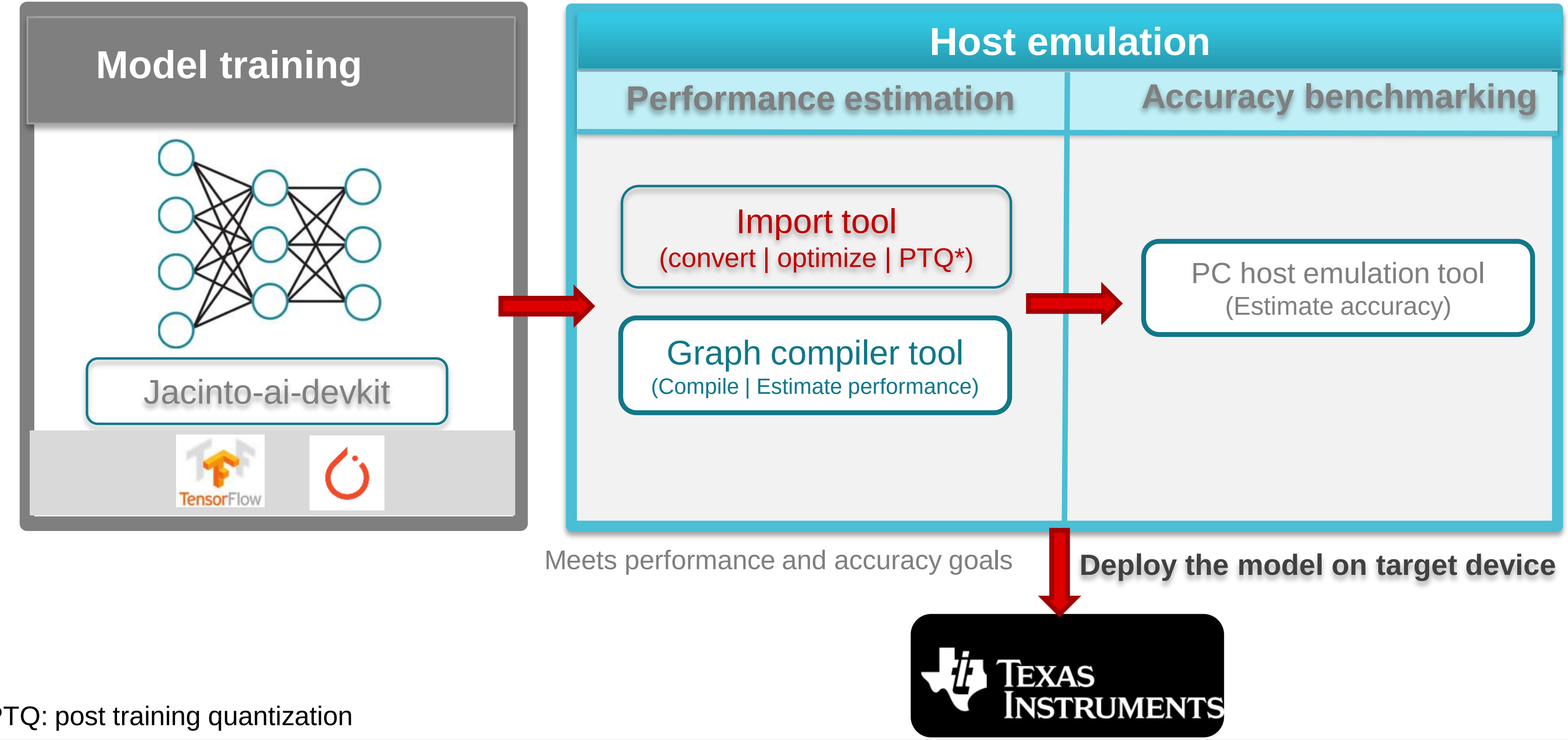
- PC based tools for model optimizations and performance estimation
- Runtime API and hand-optimized library for efficient inference execution on TI SoC

Network model development flow using TIDL tools



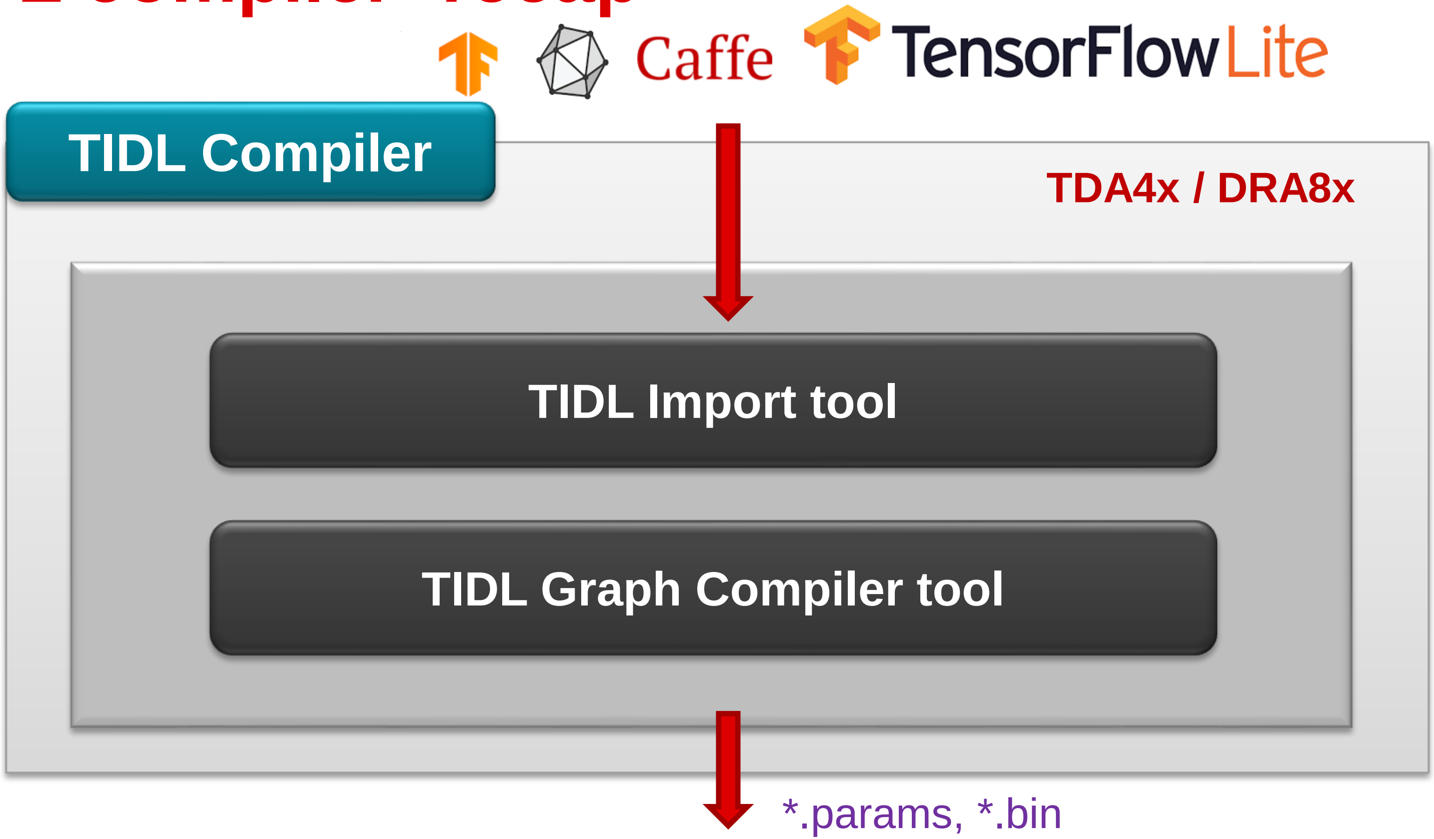
*PTQ: post training quantization

Network model development flow using TIDL tools

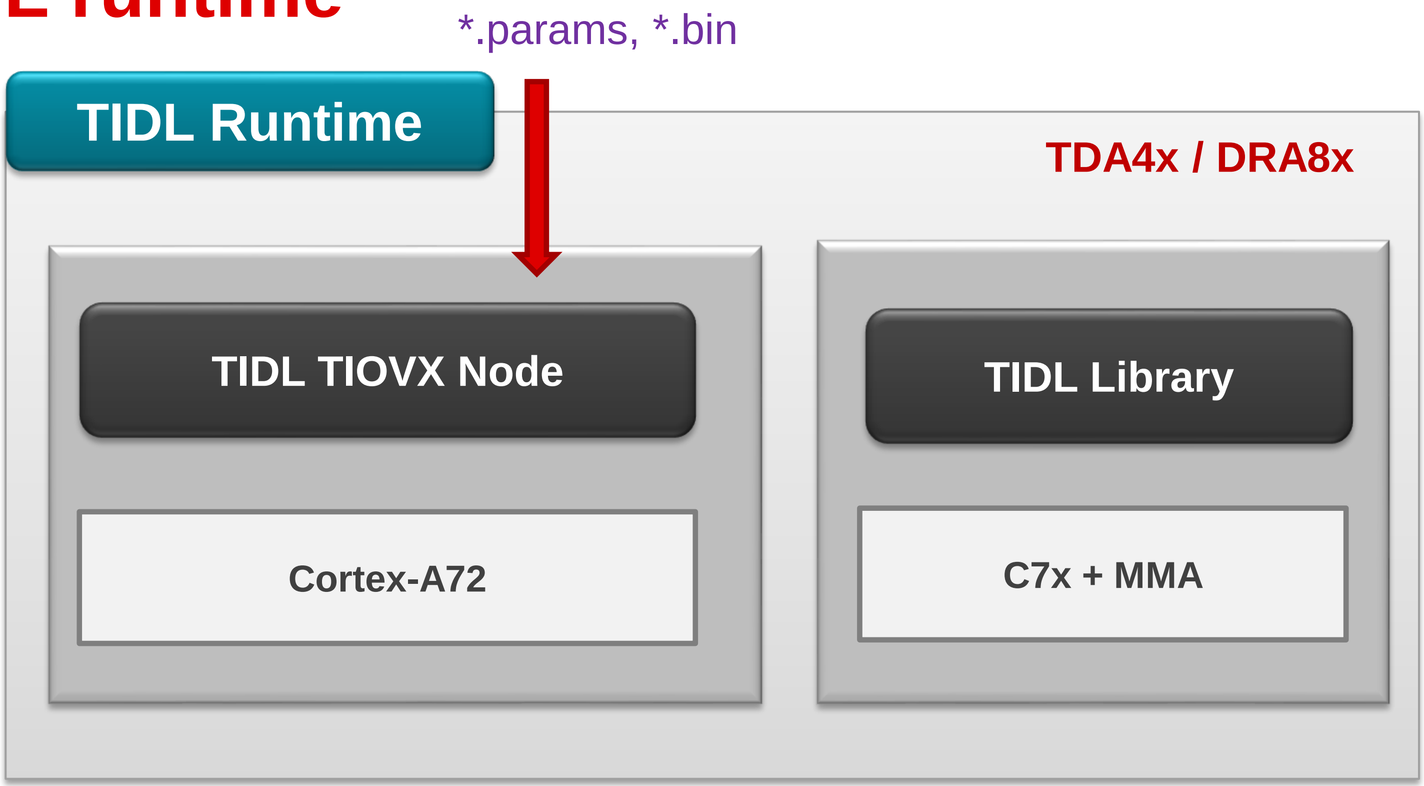


*PTQ: post training quantization

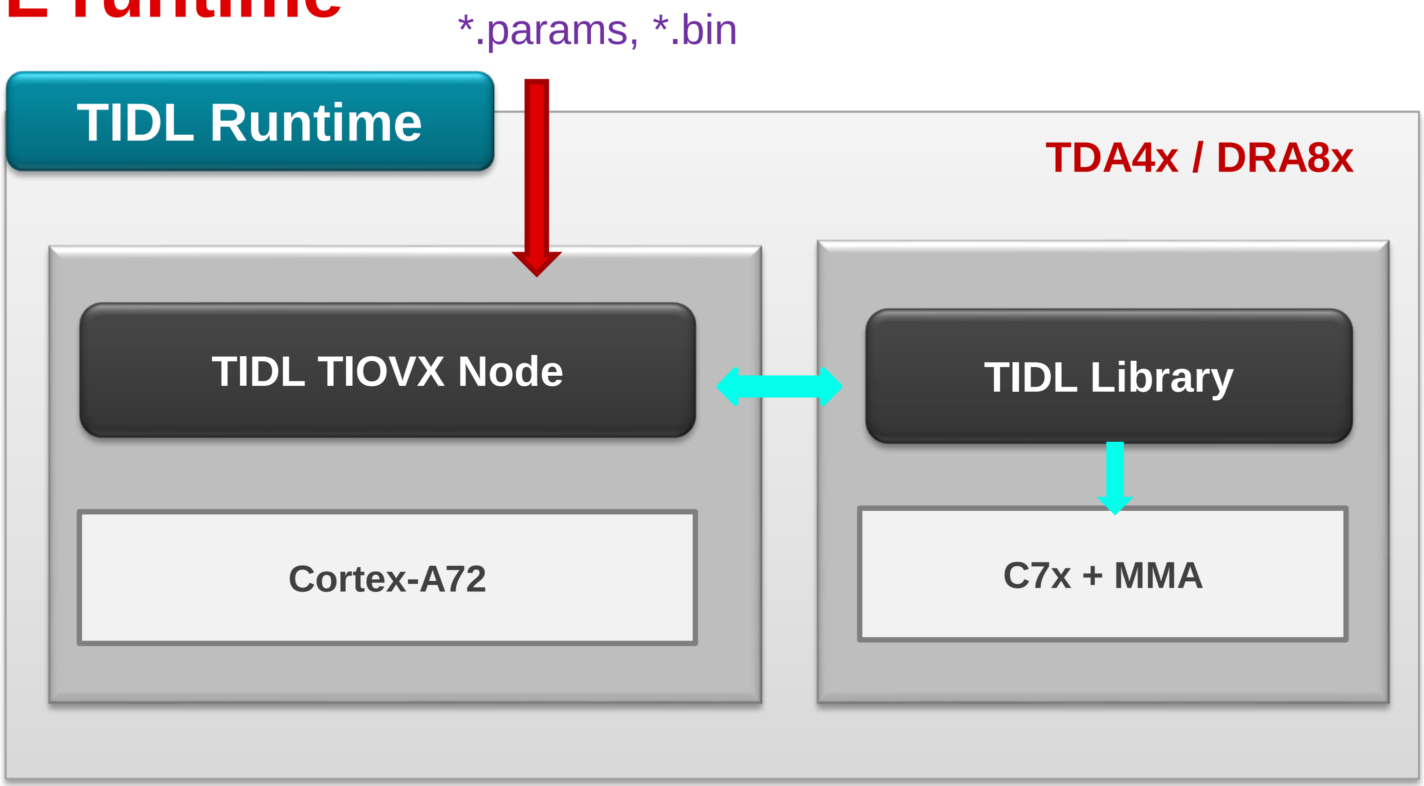
TIDL compiler- recap



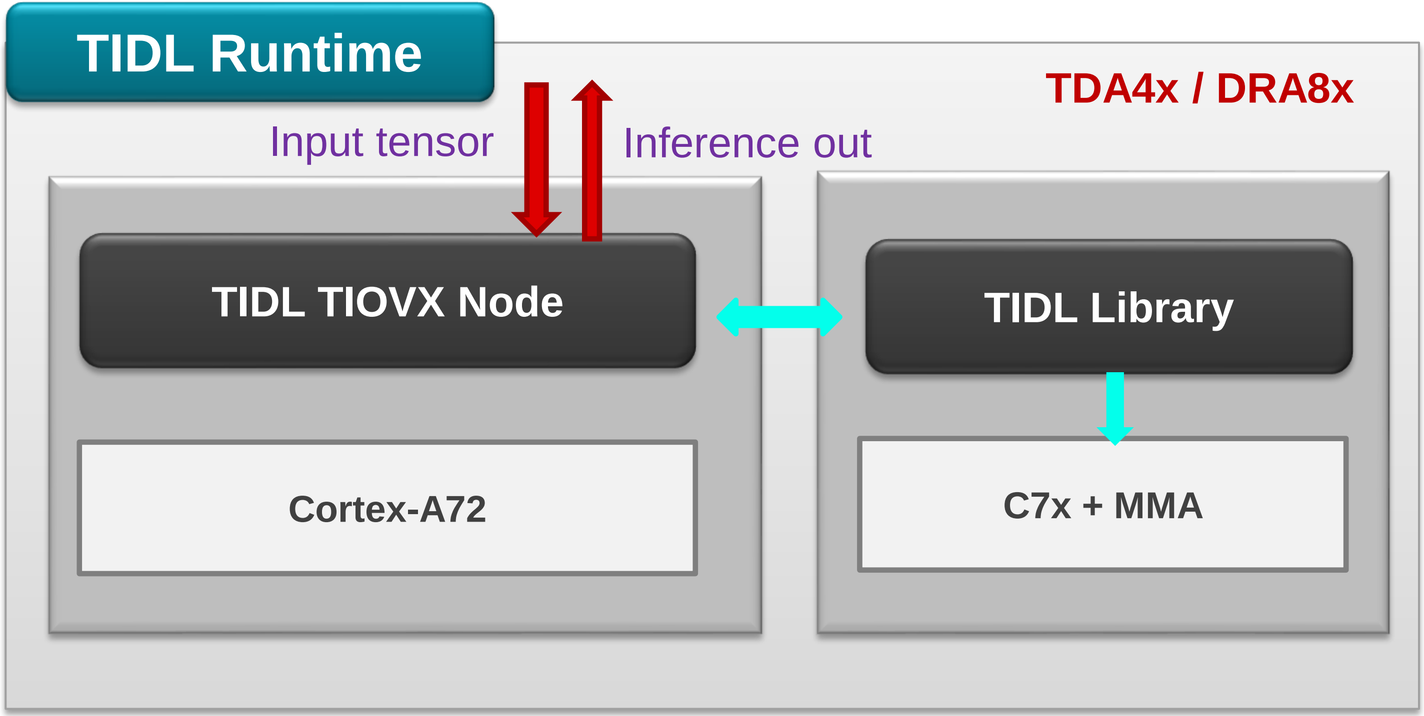
TIDL runtime



TIDL runtime



TIDL runtime



TIDL accelerate all operators you rely on

TIDL features:

- Accelerates all operators commonly used by CNN vision models on our deep learning accelerator cores
- Out-of-box support for 35+ pre-trained CNN models
- List continues to grow

Popular operators supported include:

- Convolution
- Pooling
- Element Wise
- Inner-Product
- Soft-Max
- Bias Add
- Concatenate
- Scale
- Batch Normalization
- Re-size
- Arg-max
- Slice
- Crop
- Flatten
- Shuffle Channel
- Detection output
- Deconvolution/Transpose convolution

Refer to latest Processor SDK user guide document for complete list of accelerated operators and tested models:

https://software-dl.ti.com/jacinto7/esd/processor-sdk-rtos-jacinto7/latest/exports/docs/tidl_j7_01_02_00_09/ti_dl/docs/user_guide_html/md_tidl_layers_info.html

Open source run-time support

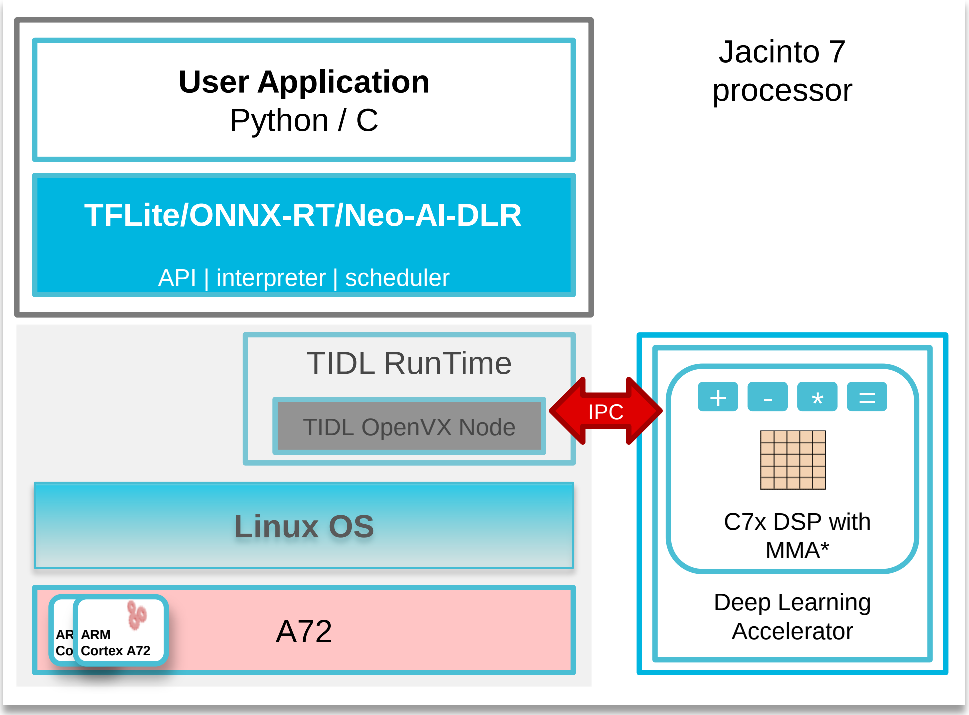
- Work in Progress

Training framework		Optimizer / Compiler / Graph format		RunTime Engine
TensorFlow	❯	TensorFlowLite™ format	❯	TFLite RunTime
	❯	ONNX™ format	❯	ONNX-RunTime
TensorFlowLite	❯	TVM / SageMaker Neo™ compiler	❯	Neo-AI-DLR

Now accelerate all your models with open source API

- Ease of use by providing Open Source programming interface
- TIDL unsupported operators run on Cortex-A core

TensorFLow Lite models	›	TFLite RunTime
ONNX models	›	ONNX-RunTime
TVM/SageMaker Neo compiler	›	Neo-AI-DLR



*MMA: Matrix Multiplication Accelerator (Tensor Processing Unit)

Summary

- TIDL minimizes time to deployment of Deep Learning Networks onto TI SOCs
- Aim for a seamless user experience
- New features and support constantly added