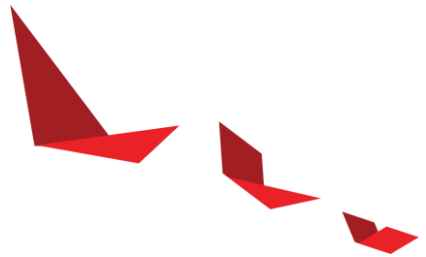
XILINX

Xilinx ML Solutions

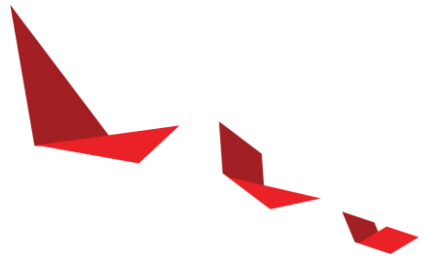
Jon Cory
Xilinx ML Specialist

Agenda



- How are We Different?
- Vitis and Vitis-AI
- Vitis-AI design flow
- Deployment
  - Edge
  - Cloud
- Getting started
- Not just CNNs..

Agenda



- How are We Different?
- Vitis and Vitis-AI
- Vitis-AI design flow
- Deployment
  - Edge
  - Cloud
- Getting started
- Not just CNNs..

How are We Different?

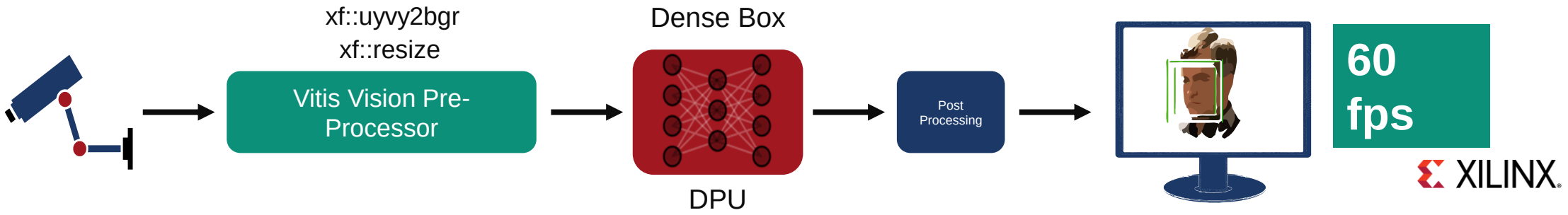
	Competing GPU (21 TOPS INT8)		ZCU104 (2 * B4096 @ 300MHz V1.4.1, 2.46TOPS INT8)	
Model	Efficiency Batch=1	Efficiency Batch>1	Efficiency Thread = 1	Efficiency Thread > 1
Inceptionv4(299x299), 24.5G	28.2%	32.8%	30.1%	58.4%
Resnet-50(224x224), 7.74G	19.9%	29.8%	24.5%	44.7%
SSD Mobilenet-V1(300x300), 2.47G	5.7%	9.3%	9.3%	33.1%
VGG-19(224x224), 39.28G	10.8%	31.2%	29.4%	54.0%

2-3x Efficiency Improvement

Chip Down Design

- Form Factor Control
- BOM Control (PPAP, etc.)
- EMI/EMC
- Cost Control
- AECQ100
- Scalable Portfolio
- Pin Compatible Packages

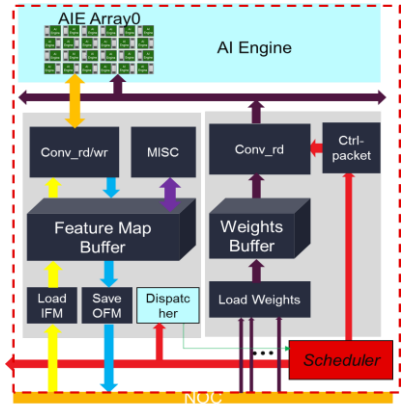
Custom, Flexible, Low Latency Acceleration, Fusion, & Interfaces



XILINX

Performance on Versal AI Core Series (AI Edge use case)

87% Higher Performance, 19x Lower Latency Than Nvidia Jetson AGX Xavier



ResNet50 v1.5	Versal AI Core (96 AIE, 3DPU, 32T)		NVidia AGX Xavier (32T)
Performance	1567 FPS	87% Higher →	879 FPS
Latency	7.6 ms	19x Lower →	145.7 ms

Our Differentiators

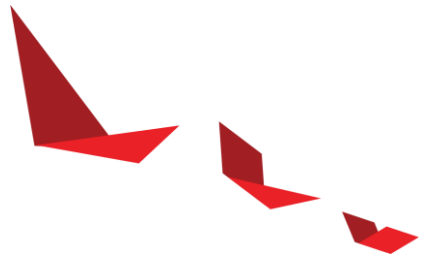
- AI Engines and DPU design for high compute efficiency
- Cacheless memory hierarchy for determinism and low latency
- High bandwidth IO to remove IO bottlenecks

NETWORK	BATCH SIZE	PERF (img/sec)	LATENCY (ms)
ResNet-50	1	358	2.8
ResNet-50	2	508	3.9
ResNet-50	4	634	6.3
ResNet-50	8	717	11.2
ResNet-50	16	767	20.9
ResNet-50	32	841	38.0
ResNet-50	64	869	73.6
ResNet-50	128	879	145.7

AGX Xavier performance

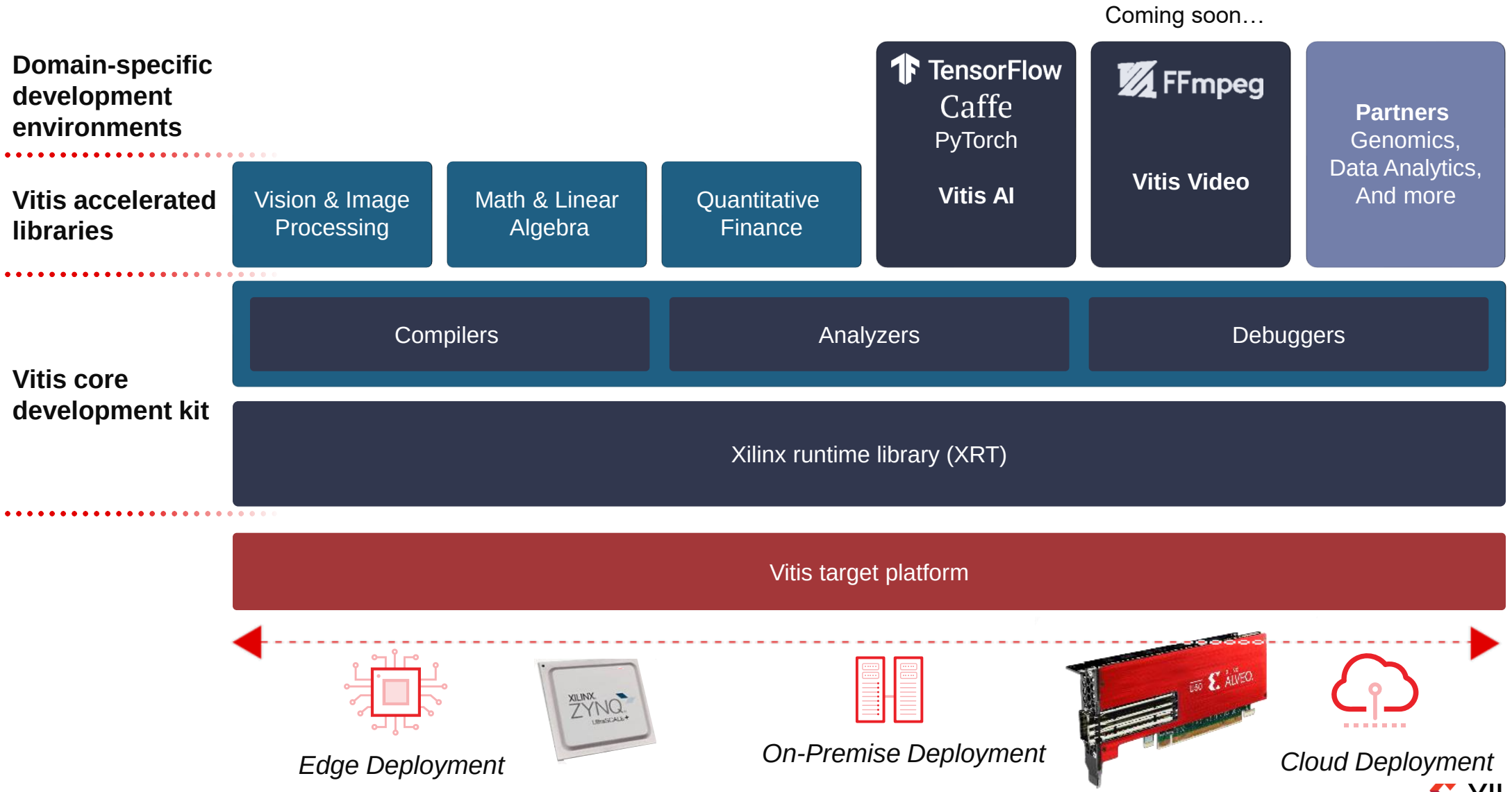
XILINX

Agenda



- How are We Different?
- Vitis and Vitis-AI
- Vitis-AI design flow
- Deployment
  - Edge
  - Cloud
- Getting started
- Not just CNNs..

Vitis Unified Software Platform



Vitis AI: Deep Learning Acceleration

Frameworks

TensorFlow Caffe PyTorch tvm

Vitis AI models

Vitis AI development kit

AI Optimizer | AI Quantizer | AI Compiler | AI Profiler | AI Library

Xilinx runtime library (XRT)

Deep Learning Processing Unit (DPU)

CNN DPU | LSTM DPU | MLP DPU

XILINX

Develop: Use Extensive, Open Source Libraries

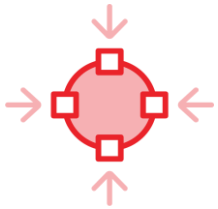
GitHub

Domain-Specific Libraries

Vision & Image

Quantitative Finance

Data Analytics & Database

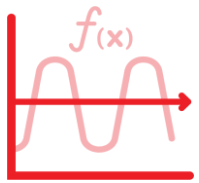


Data Compression

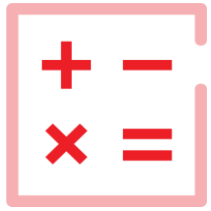
Data Security

Partner Libraries

Common Libraries

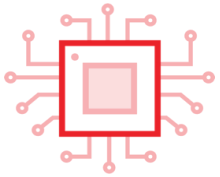


Math



Linear Algebra

Statistics



DSP

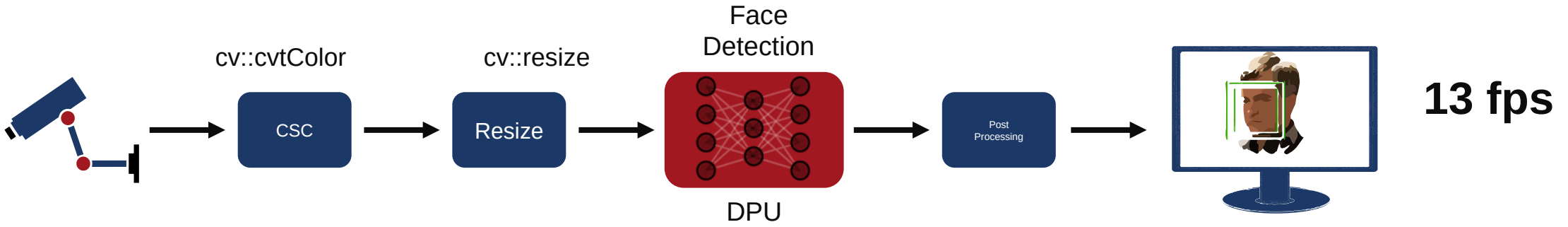
Data Management

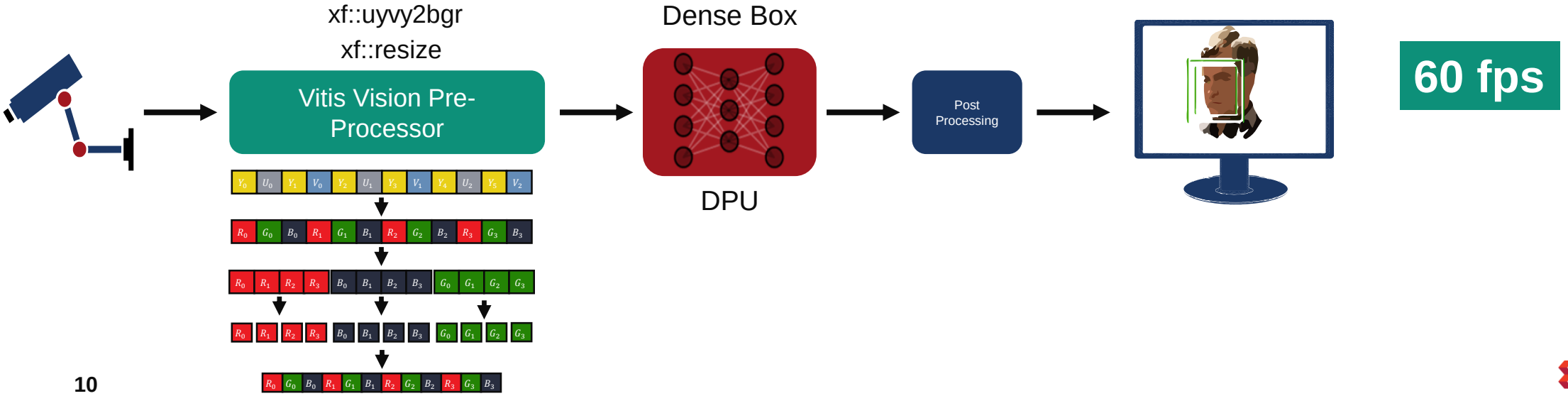
500+ functions across multiple libraries for performance-optimized out-of-the-box acceleration

XILINX

Preprocess Acceleration with Vitis Vision Library

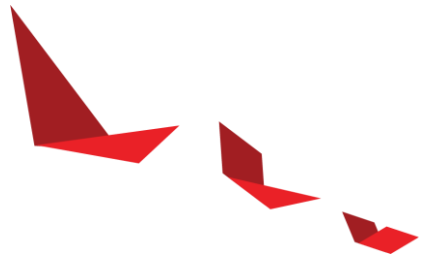
› Design captures 1080p60 camera data, runs a neural network inference, and displays the results on a monitor.





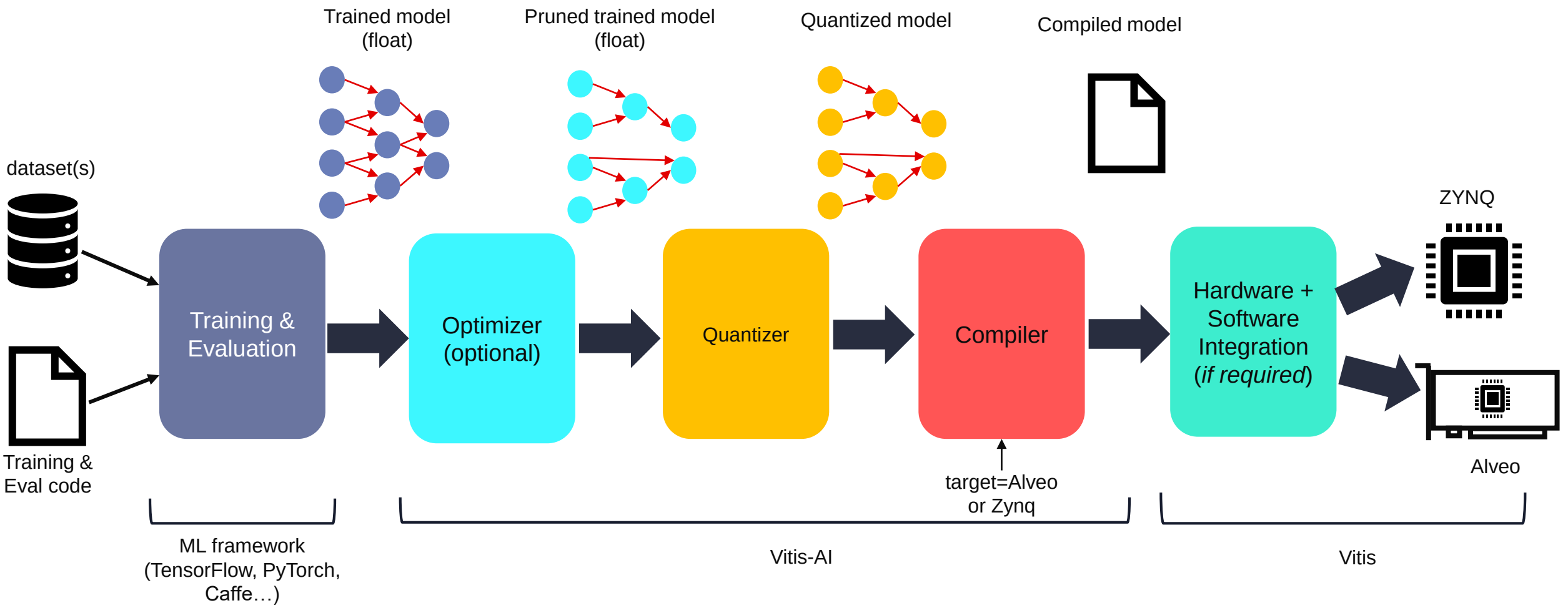
XILINX.

Agenda



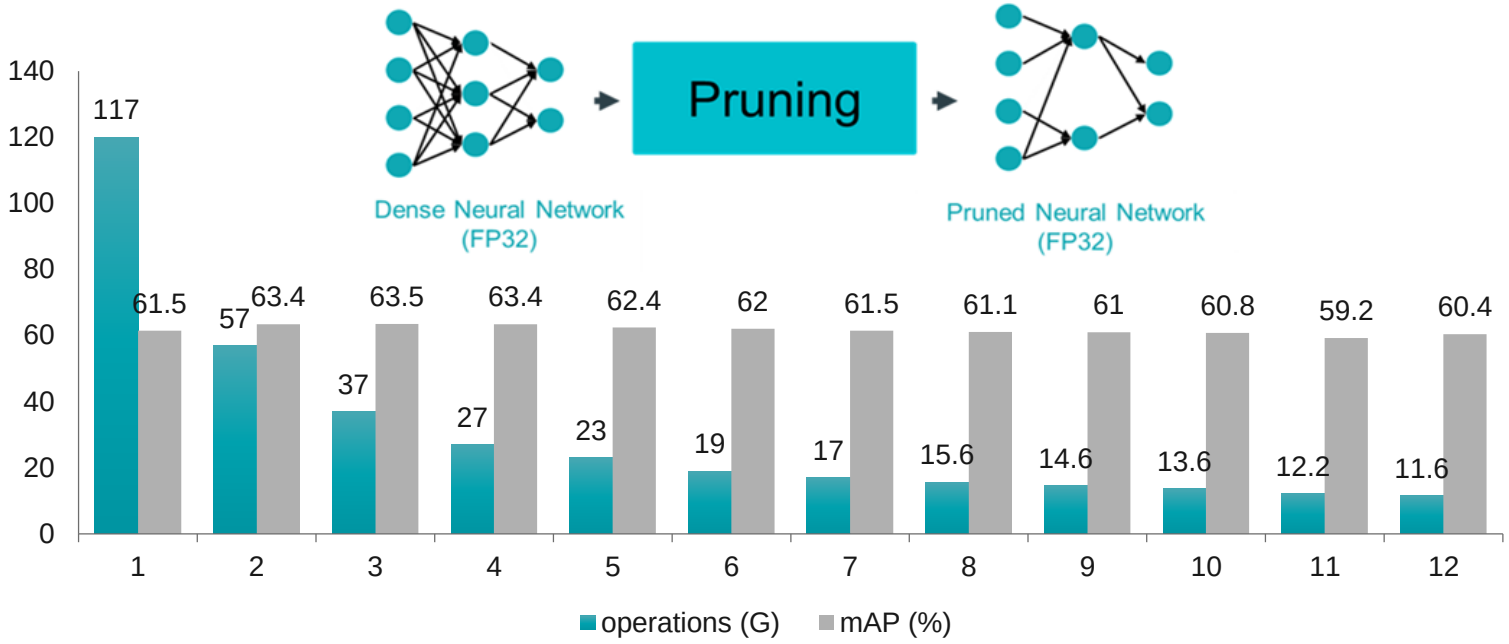
- How are We Different?
- Vitis and Vitis-AI
- Vitis-AI design flow
- Deployment
  - Edge
  - Cloud
- Getting started
- Not just CNNs..

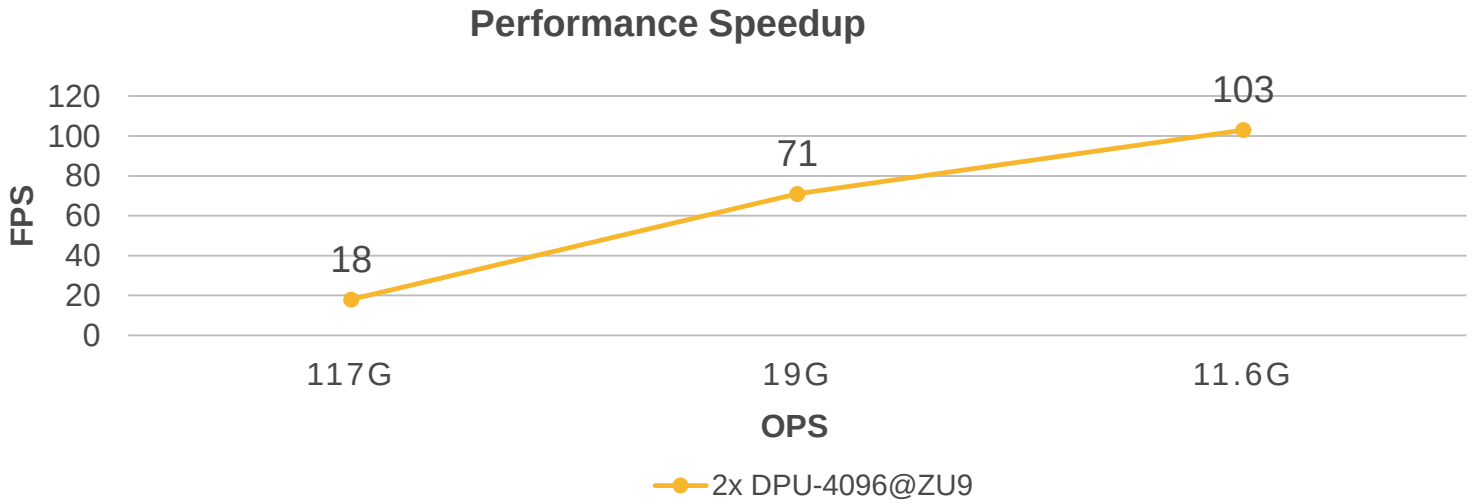
Unified Edge/Cloud Development Flow



XILINX

Vitis AI Optimizer

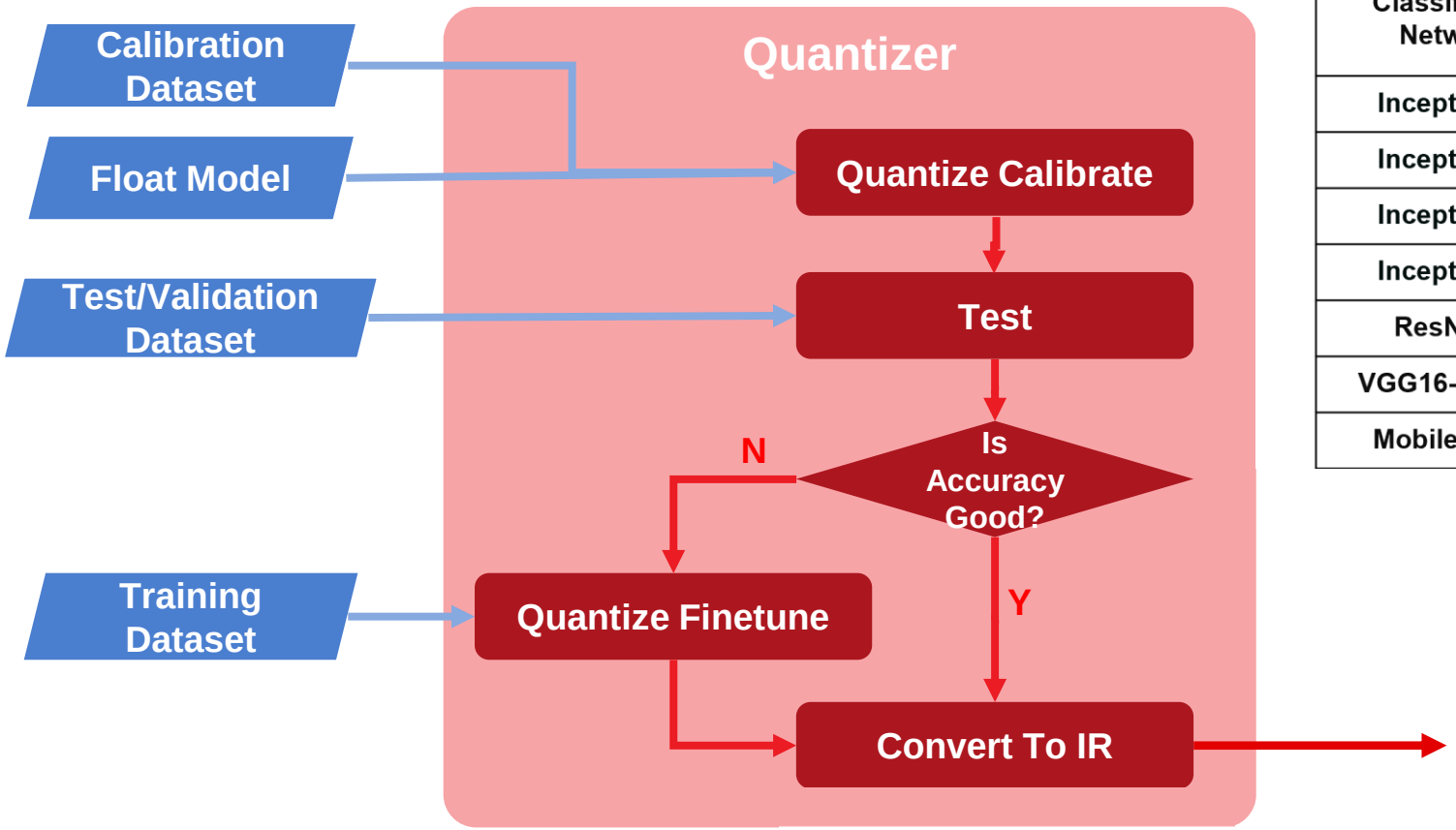




SSD+VGG @ Surveillance 4 Classes

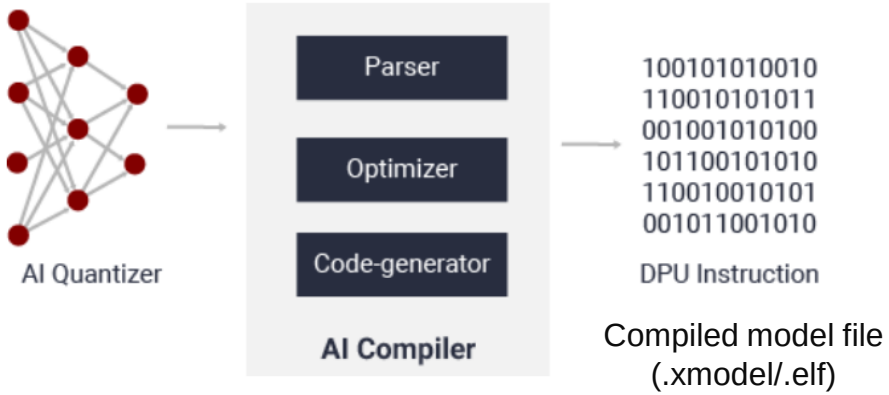
XILINX

Vitis AI Quantizer and Compiler



Classification Networks	Float		After quantized finetune			
	Top1	Top5	Top1	ΔTop1	Top5	ΔTop5
Inception_v1	66.90%	87.68%	66.62%	-0.28%	87.58%	-0.10%
Inception_v2	72.78%	91.04%	72.40%	-0.38%	90.82%	-0.23%
Inception_v3	77.01%	93.29%	76.56%	-0.45%	93.00%	-0.29%
Inception_v4	79.74%	94.80%	79.42%	-0.32%	94.64%	-0.16%
ResNet-50	74.76%	92.09%	74.59%	-0.17%	91.95%	-0.14%
VGG16-3fc-float	70.97%	89.85%	70.74%	-0.23%	89.79%	-0.06%
MobileNet_v1	70.61%	89.63%	69.71%	-0.90%	89.06%	-0.57%

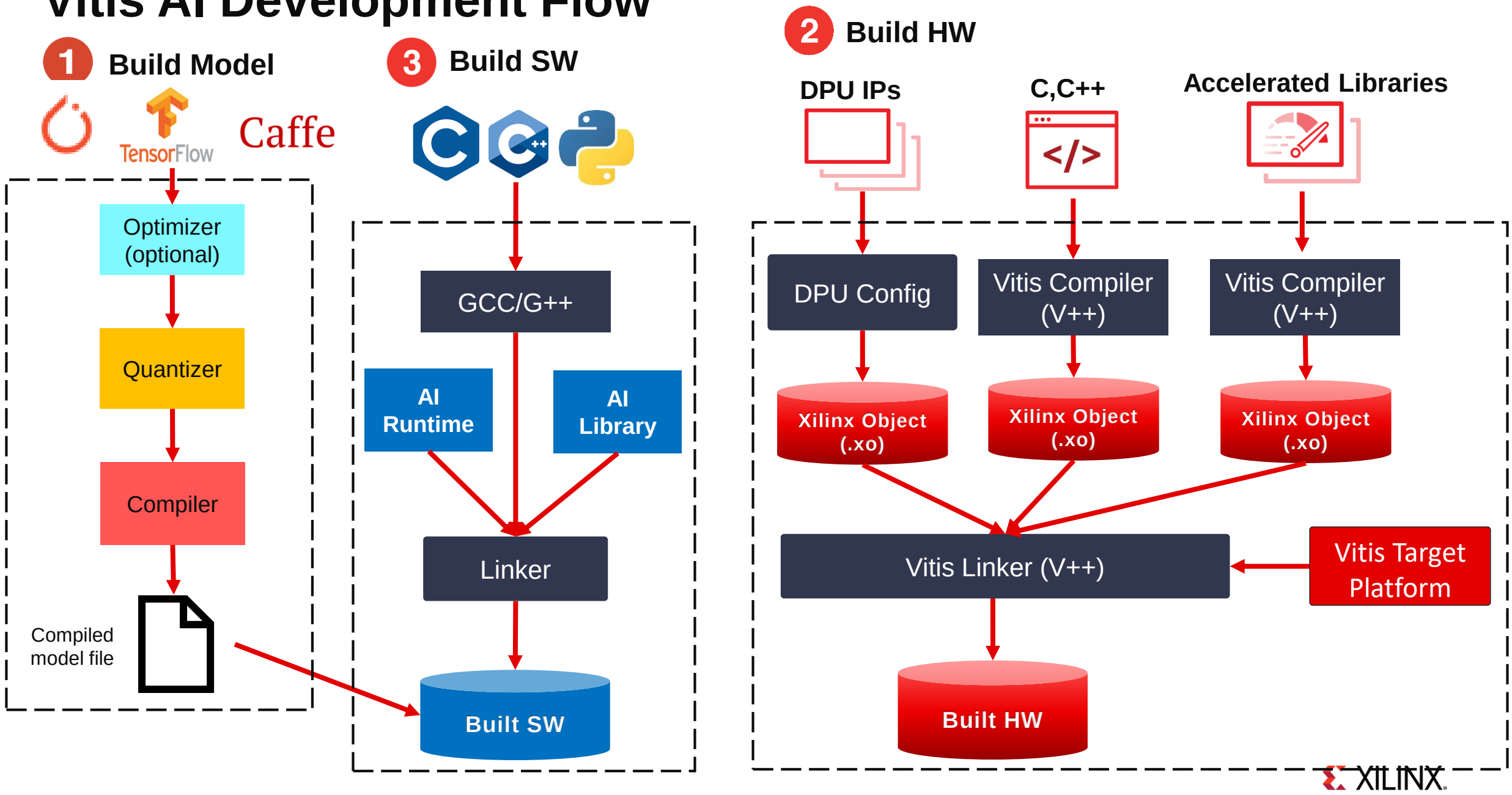
Vitis AI Compiler



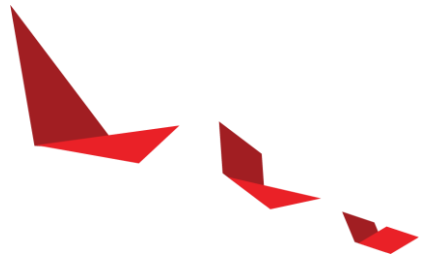
Compiled model file (.xmodel/.elf)

XILINX

Vitis AI Development Flow



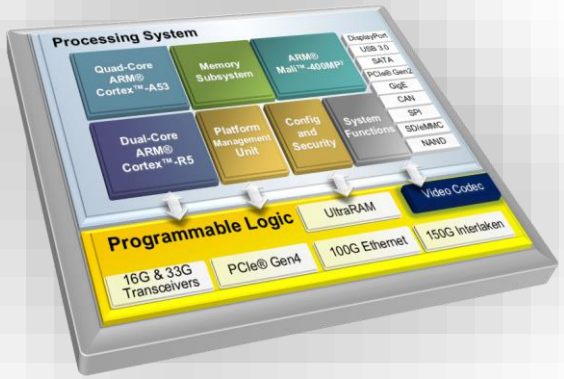
Agenda



- How are We Different?
- Vitis and Vitis-AI
- Vitis-AI design flow
- Deployment
  - Edge
  - Cloud
- Getting started
- Not just CNNs..

Target Platforms

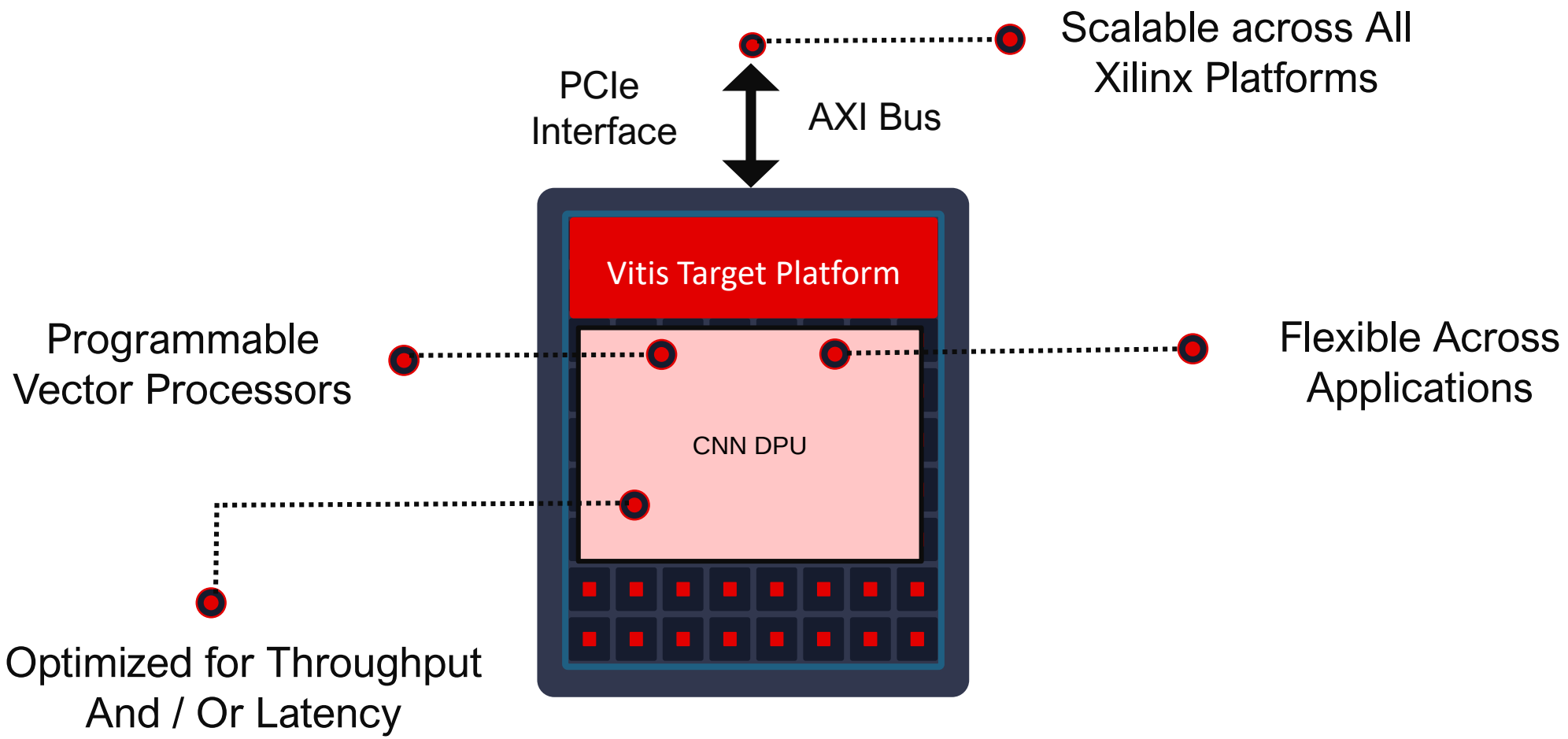
- Edge
  - Zynq family with built-in ARM processors
    - Zynq MPSoC: Quad Cortex-A53 + dual Cortex-R5
    - Zynq-7000: Dual Cortex-A9 for Zynq-7000
  - H.264/265 Video Codecs
    - ideal for streaming video
  - ARM Mali-400 GPU and DisplayPort outputs
  - Multiple high-speed interfaces
    - PCIe, USB3, Gigbit Ethernet, SATA, MIPI…



- Cloud/DataCenter
  - Xilinx provides the Alveo Family of PCI Express accelerator cards
  - Vitis-AI 1.2 currently supports
    - U50, U50LV, U280, U200, U250
    - U25 support coming soon
  - Users can download the Xilinx provided DPU configurations directly over the PCIe connection
  - No hardware development necessary
  - ..or can mix Xilinx DPU IP with custom logic

XILINX

Deep Learning Processing Unit (DPU)



CNN DPUs for 16/28nm

HIGH

THROUGHPUT

LOW

LATENCY

HIGH

Throughput optimized (DDR)

- Alveo U200 / U250
- Optimized for 100% INT8 models and ResNet/Inception Models
- Optimized for pruned models
- High Throughput at medium latency

Throughput optimized (HBM)

- Alveo U280 / U50
- Optimized for ResNet/Inception models and HBM devices
- Low-cost and high power efficiency solution

Latency optimized

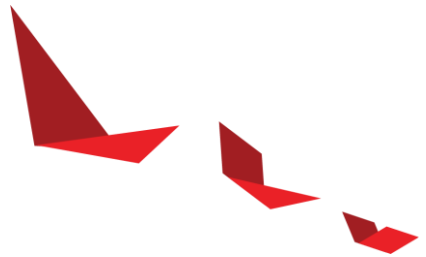
- Alveo U280 /U50
- Optimized for MobileNet-like models and HBM devices
- Extreme low latency with high-resolution images

Scalability optimized (DDR)

- Zynq-7000, ZU+ MPSoC
- Support popular CNN models
- Scalable with power & cost

XILINX.

Agenda



- How are We Different?
- Vitis and Vitis-AI
- Vitis-AI design flow
- Deployment
  - Edge
  - Cloud
- Getting started
- Not just CNNs..

How to get Vitis-AI

- Clone or download it from GitHub
- Vitis-AI tools are provided as Docker containers
  - Includes:
    - ML frameworks
    - Vitis-AI tools
      - Default version does not include Optimizer (pruning)
  - Two versions available
    - CPU only: Does not require a GPU card.
    - GPU support: Includes CUDA 10 and CuDNN 7 – accelerates quantization phase and allows training to be done from inside Vitis-AI.
  - Either download prebuilt Docker images or build them yourselves
    - All Docker recipes are provided

Everything available from GitHub: https://github.com/Xilinx/Vitis-AI

Vitis-AI Model Zoo

- A collection of pre-trained models ready to be deployed
  - Includes both trained floating-point model and quantized models
  - TensorFlow, PyTorch, Caffe and Darknet models
- Wide range of model types and applications
  - Classification, Segmentation, Object detection
  - license plate detection, face detection, ADAS vehicle, pedestrian detect..
- All models include test, demo and evaluation code
- Performance numbers provided for each model for different target platforms
  - ZCU102, ZCU104, U50, U280…

Performance on U50 lv10e

Measured with Vitis AI 1.2 and Vitis AI Library 1.2

▼ **Click here to view details**

The following table lists the performance number including end-to-end throughput and latency for each model on the `Alveo U50` board with 10 DPUv3E kernels running at 275Mhz in Gen3x4:

No.	Model	Name	Frequency (MHz)	E2E throughput - fps(Multi Thread)
1	resnet50	cf_resnet50_imagenet_224_224_7.7G	247.5	802.46
2	resnet18	cf_resnet18_imagenet_224_224_3.65G	247.5	1934.48
3	Inception_v1	cf_inceptionv1_imagenet_224_224_3.16G	247.5	1536.64
4	Inception_v2	cf_inceptionv2_imagenet_224_224_4G	247.5	1313.99
5	SqueezeNet	cf_squeeze_imagenet_227_227_0.76G	247.5	3451.05
6	ssd_pedestrian_pruned_0_97	cf_ssdpedestrian_coco_360_640_0.97_5.9G	247.5	755.24
7	refinedet_pruned_0_8	cf_refinedet_coco_360_480_0.8_25G	247.5	273.79
8	refinedet_pruned_0_92	cf_refinedet_coco_360_480_0.92_10.10G	247.5	574.76
9	refinedet_pruned_0_96	cf_refinedet_coco_360_480_0.96_5.08G	247.5	795.12
10	ssd_adas_pruned_0_95	cf_ssdadas_bdd_360_480_0.95_6.3G	247.5	818.22
11	ssd_traffic_pruned_0_9	cf_ssdtraffic_360_480_0.9_11.6G	247.5	570.84
12	VPGnet_pruned_0_99	cf_VPGnet_caltechlane_480_640_0.99_2.5G	275	658.99
13	FPN	cf_fpn_cityscapes_256_512_8.9G	247.5	552.17
14	SP_net	cf_SPnet_aichallenger_224_128_0.54G	275	1706.95
15	Openpose_pruned_0_3	cf_openpose_aichallenger_368_368_0.3_189.7G	220	39.68

https://github.com/Xilinx/Vitis-AI/tree/master/AI-Model-Zoo

Vitis-AI tutorials

XILINX

Vitis AI Tutorials

Tutorial	Description
Quantization and Pruning of AlexNet CNN trained in Caffe with Cats-vs-Dogs dataset (UG1336)	Train, prune, and quantize a modified version of the AlexNet convolutional neural network (CNN) with the Kaggle Dogs vs. Cats dataset in order to deploy it on the Xilinx® ZCU102 board.
MNIST Classification using Vitis™ AI and TensorFlow (UG1337)	Learn the Vitis AI TensorFlow design process for creating a compiled ELF file that is ready for deployment on the Xilinx DPU accelerator from a simple network model built using Python. This tutorial uses the MNIST test dataset.
CIFAR10 Classification using Vitis AI and TensorFlow (UG1338)	Learn the Vitis AI TensorFlow design process for creating a compiled ELF file that is ready for deployment on the Xilinx DPU accelerator from a simple network model built using Python. This tutorial uses the CIFAR-10 test dataset.
Using DenseNetX on the Xilinx DPU Accelerator (UG1340)	Learn about the Vitis AI TensorFlow design process and how to go from a Python description of the network model to running a compiled model on the Xilinx DPU accelerator.
Freezing a Keras Model for use with Vitis AI (UG1380)	Freeze a Keras model by generating a binary protobuf (.pb) file.
Quantize in fixed point some custom CNNs and deploy	

- Examples that show complete flow from training to running on an eval board
  - Written by Xilinx ML Specialists

https://github.com/Xilinx/Vitis-AI-Tutorials

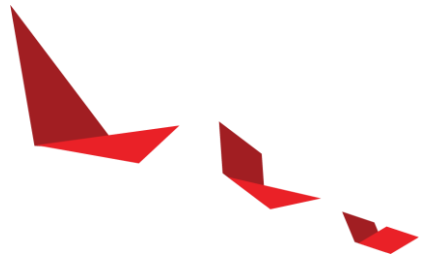
Vitis-AI training course

- Online or instructor-led
- 16 hours of training with 5 hands-on labs
- Covers entire Vitis-AI design flow:

https://xilinxprod-catalog.netexam.com/Certification/45382/developing-ai-inference-solutions-with-the-vitis-ai-platform

XILINX

Agenda



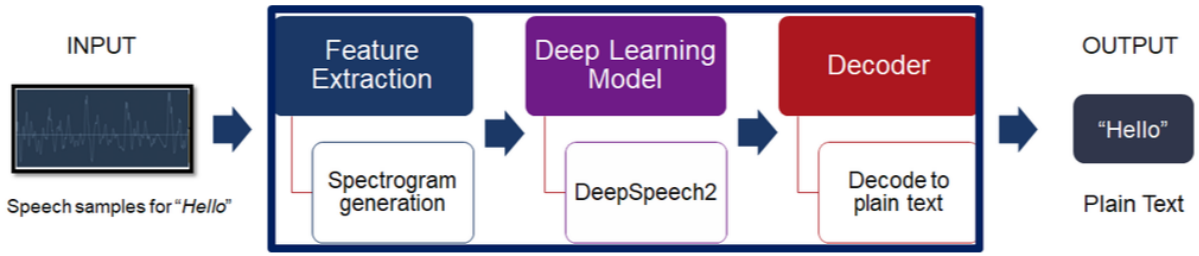
- How are We Different?
- Vitis and Vitis-AI
- Vitis-AI design flow
- Deployment
  - Edge
  - Cloud
- Getting started
- Not just CNNs..

Vitis Data Analytics library

- New library in Vitis 2020.1
- Introduces ML algorithms other than just convolutional neural networks:
  - Decision Trees
  - Random Forest
  - Logistic Regression
  - Linear Support Vector Machine
  - Naïve Bayes
  - K-Means clustering

DeepSpeech2 model on Zynq

- Xilinx Engineering have implemented DeepSpeech2 on Zynq MPSoC
- Software-only and hardware-accelerated versions available
- Conversion time is faster than speech duration time.
  - Allows for realtime speech-to-text



Vitis AI for Point Cloud 3D Detection - PointPillars

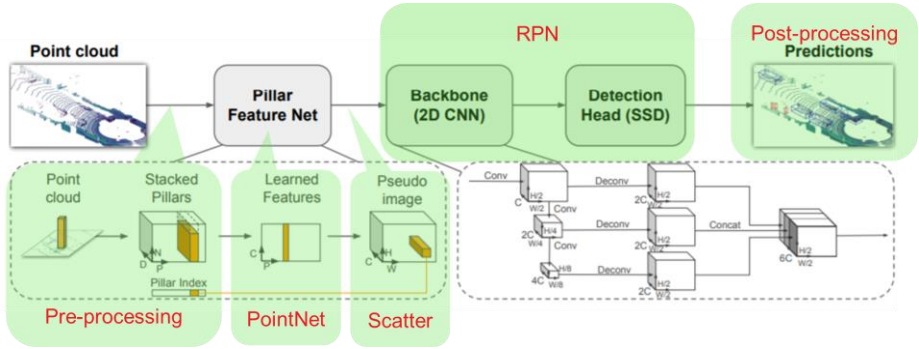
Demo specification

- Model: Pointpillars
- Framework: Pytorch
- Dataset: Kitti, 64-channel, 1~2Mpoints/sec
- 25fps (40ms latency), 1x DPU B4096 @ 300MHz on ZCU102
- Demo available for early access customers
- General access in Vitis AI 1.3

XILINX VITIS™ AI

Features highlight

- Algorithm & SW optimized for better performance acceleration
  - Pruning applied and optimized in model structure
  - 7X computing reduced: 70Gops float → 10Gops pruned
  - 13X end-to-end performance speed-up
- Detect multiple classes: vehicle, bicycle and people



XILINX

Demo Video

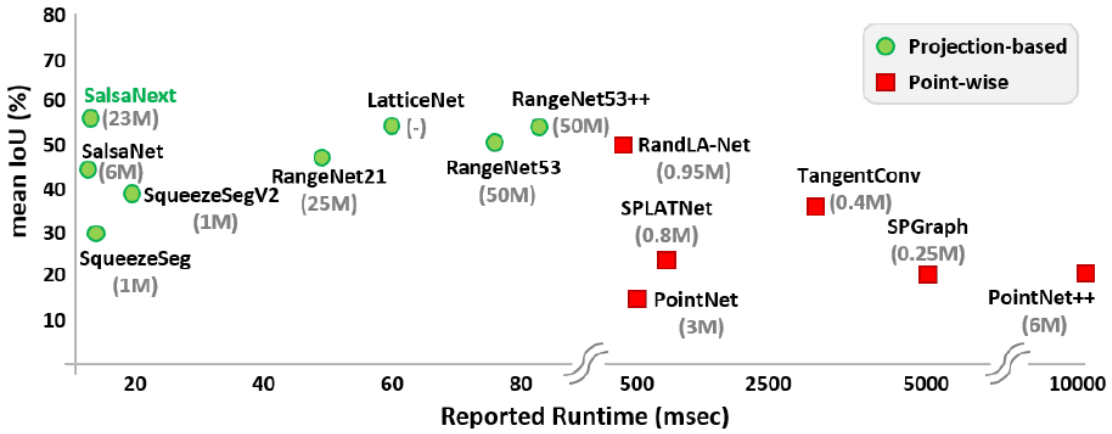
Real-time Multi-class 3D Point Cloud Object Detection - Powered by Vitis AI

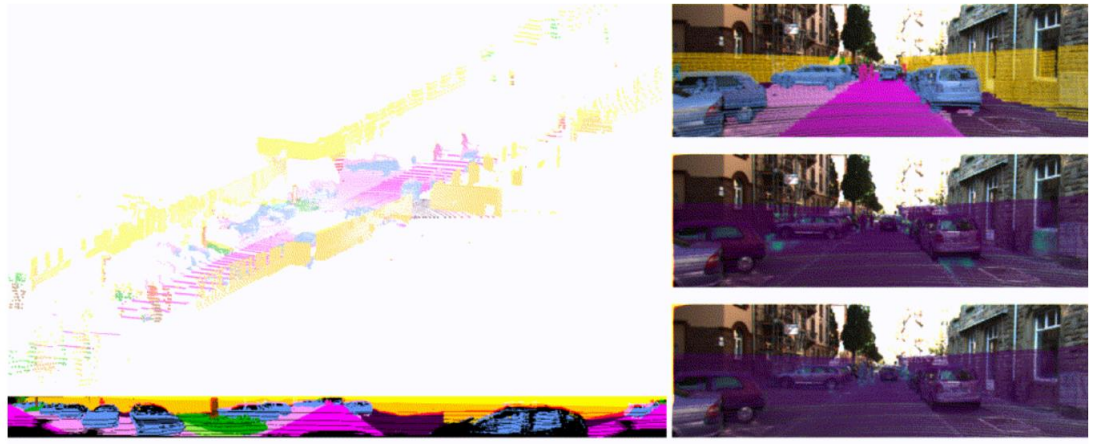
Vitis AI for Point Cloud 3D Segmentation - SalsaNext

Point cloud Segmentation

- Model: SalsaNext
- Framework: Pytorch
- Dataset: Semantic Kitti, 64-channel
- DPU for SalsaNext: DPU supports all layers; post-processing on CPU
- Model pruned 80Gops -> 20Gops, *21fps on ZCU102 (still under optimization)
- General access in Vitis AI 1.3

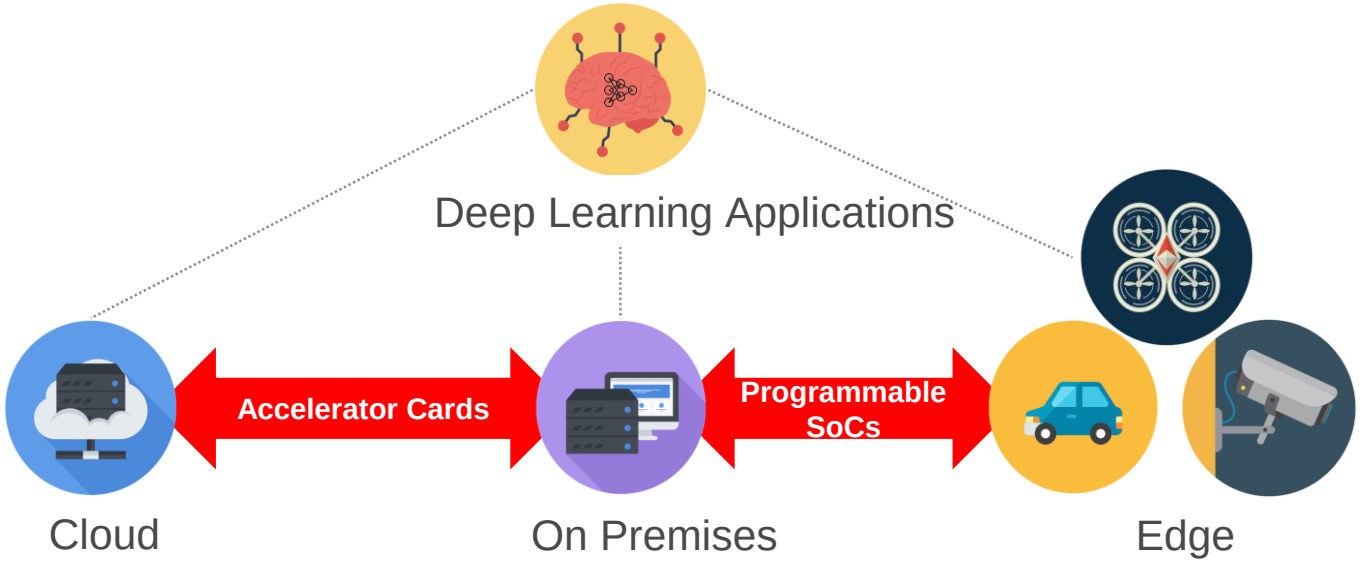
https://arxiv.org/abs/2003.03653
https://github.com/TiagoCortinhal/SalsaNext





Summary

- Xilinx provides an easy path to machine learning inference
  - Complete toolset to go from trained neural network to deployment
    - Network compression via pruning and quantization
    - Hardware accelerator DPU IP
    - Software stack, APIs, examples, tutorials
- Programmable logic provides the perfect solution to *whole application* acceleration
  - ..its not just the ML network that needs to be accelerated
- Easy to scale up/down between Edge and Cloud/DC with our unified design flow and hardware platforms
  - Chip-down *and* Accelerator cards



Thank You