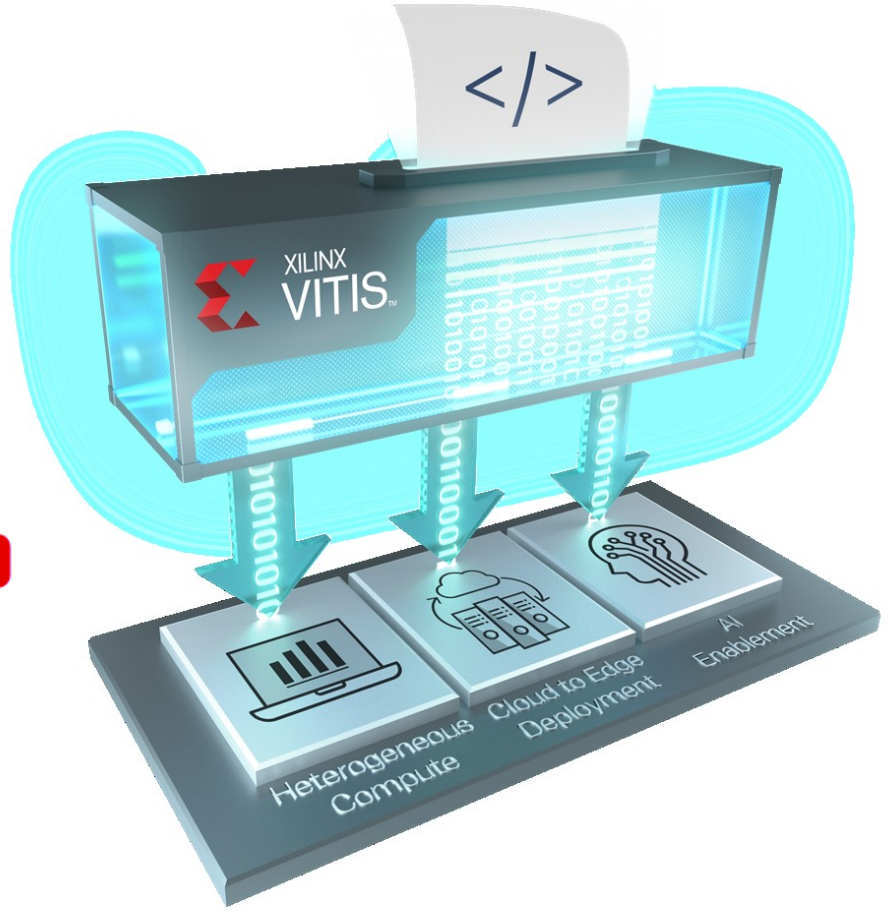
Accelerating AI in Intelligent Video Sur

Accelerating AI Camera Development with Xilinx VITIS

Maxime Rocca
XILINX - Tech Sales Lead EMEA

12th International Business Forum All-over-IP 2019

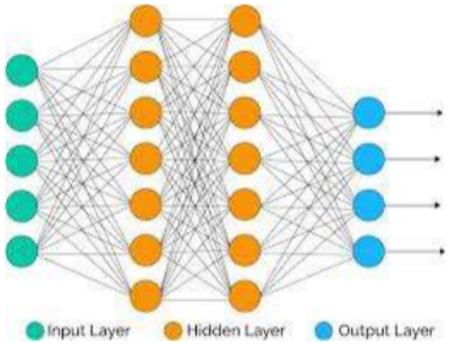
November 20–22, 2019 | Moscow | Sokolniki Exhibition and Convention Center, pavilion 4, 4.1



[Credit : Quenton Hall | XILINX AI System Architect for Industrial, Scientific,

© Copyright 2019 Xilinx

Defining the Problem



= 43 TOPS

?

1080p Object Detection (SSD) @ 30 FPS

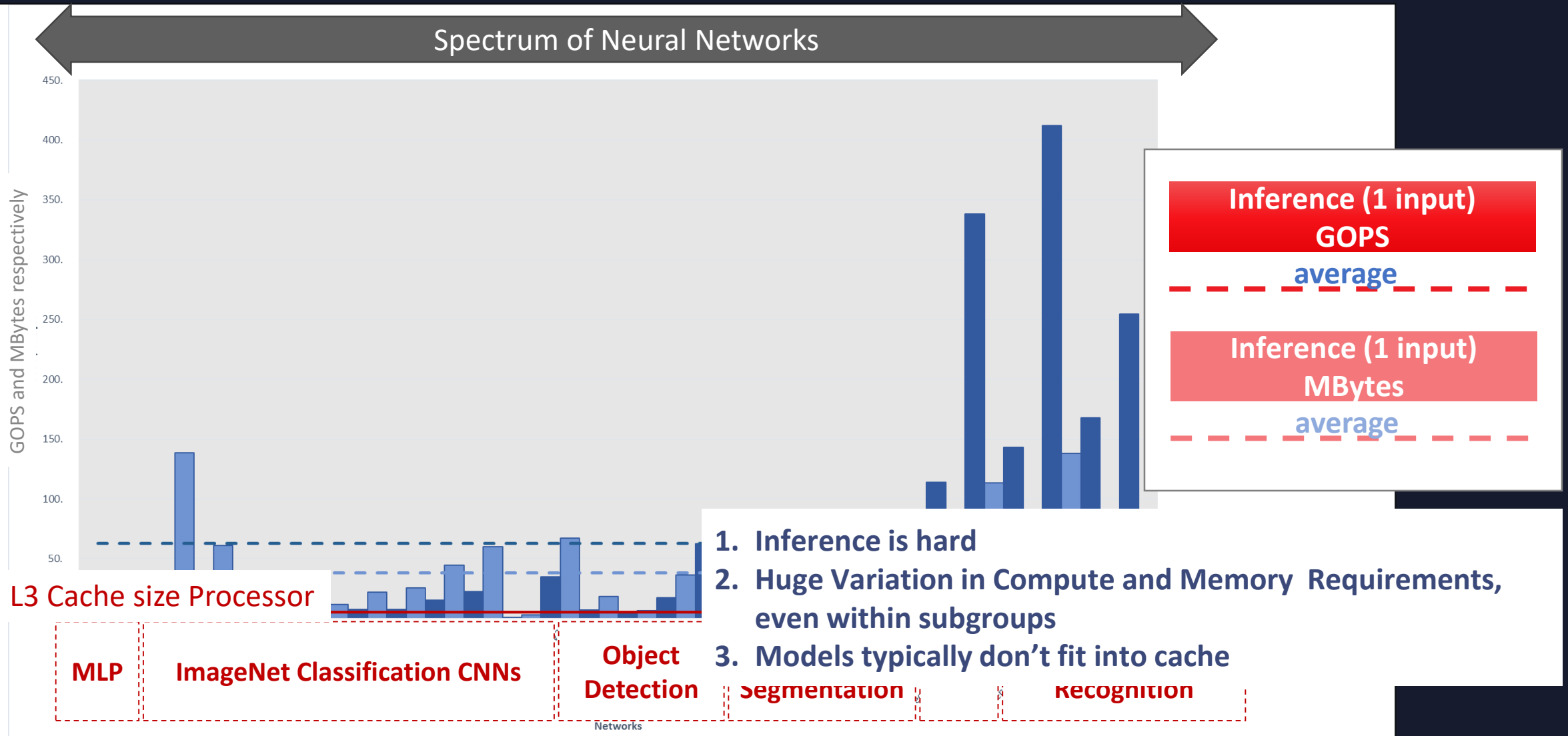
< 10W, < 50 ms latency, <$50

Jon Cory, ML Live Detroit

© Copyright 2019 Xilinx

Inference Compute and Memory
Across a Spectrum of Neural Networks

*architecture independent
**1 image forward
*** batch = 1
**** int8

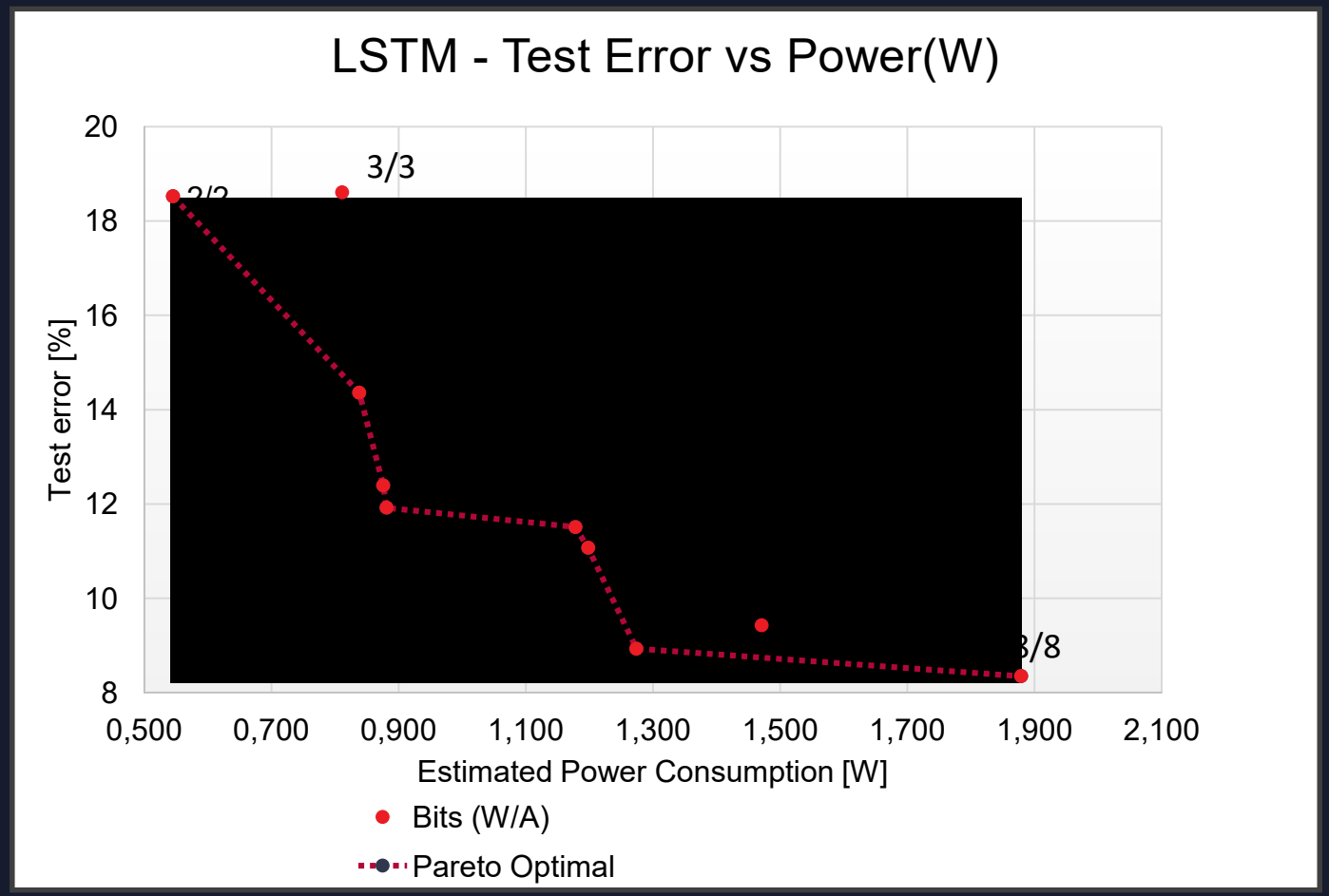


1. Inference is hard
2. Huge Variation in Compute and Memory Requirements, even within subgroups
3. Models typically don't fit into cache

© Copyright 2019 Xilinx

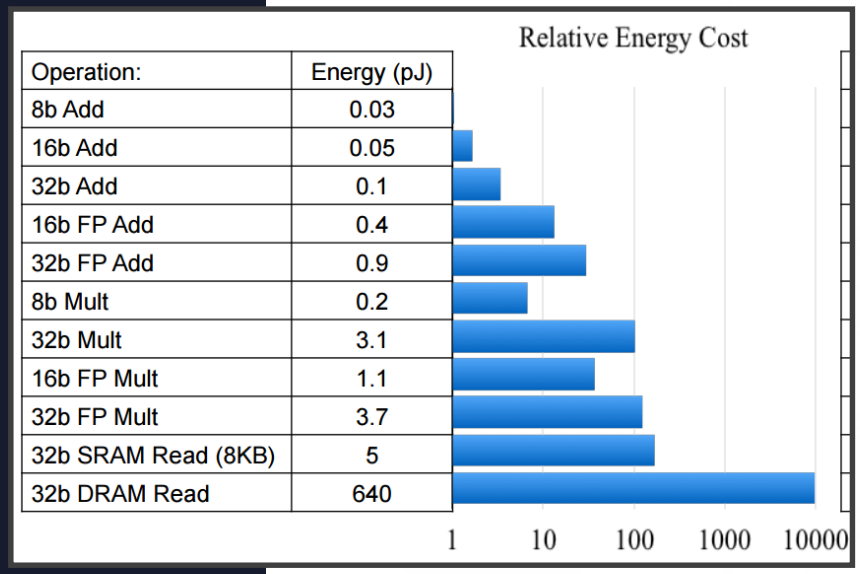
Reducing Precision Inherently Saves Power

FPGA



Target Device ZU7EV ● Ambient temperature: 25 °C ● 12.5% of toggle rate ● 0.5 of Static Probability ● Power reported for PL accelerated block only

ASIC



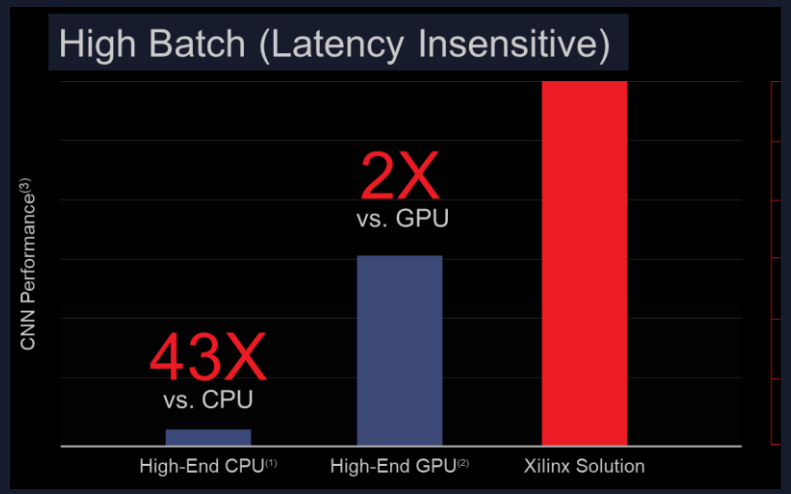
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

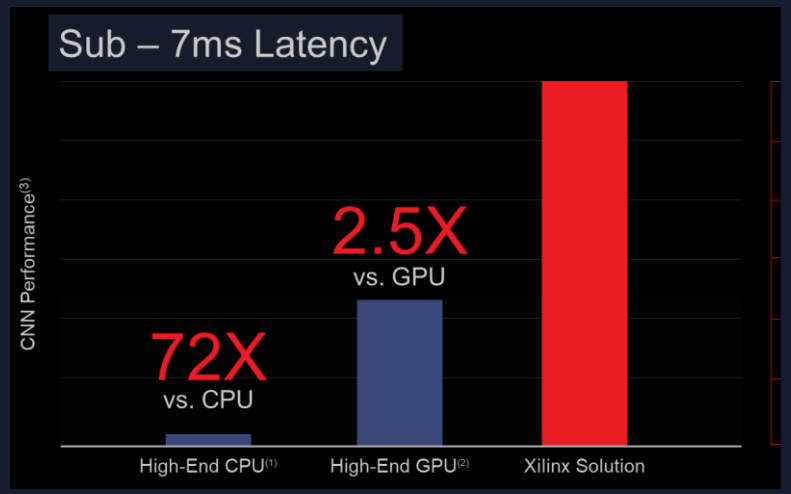
Source: Bill Dally
Network Summit,

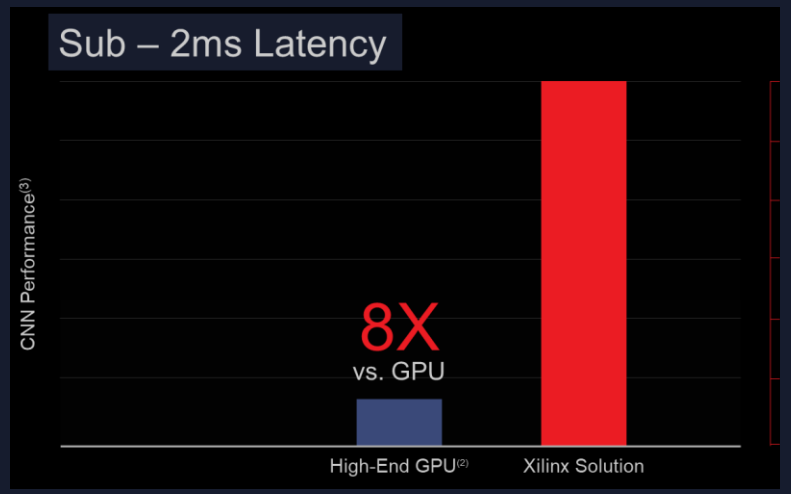
Rybalkin, V., Pappalardo, A., Ghaffar, M.M., Gambardella, G., Wehn, N. and Blott, M. "FINN-L: Library Extensions and Design Trade-off Analysis for Variable Precision LSTM Networks on FPGAs."

© Copyright 2019 Xilinx

AI Compute Compared to CPUs and GPUs



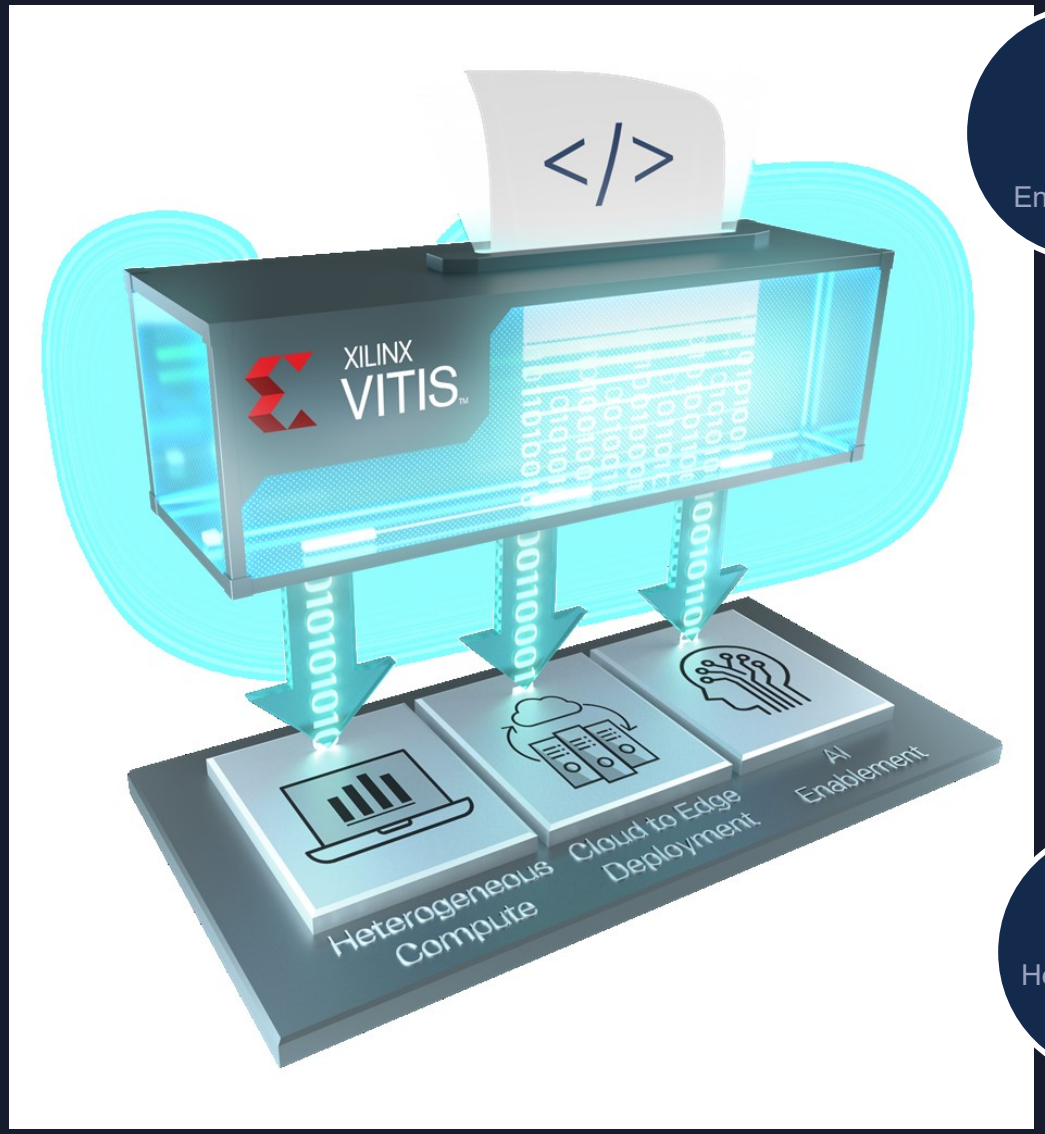




© Copyright 2019 Xilinx

VITIS Unified Software Platform

- Available in November
- Standards, Open
- Free!



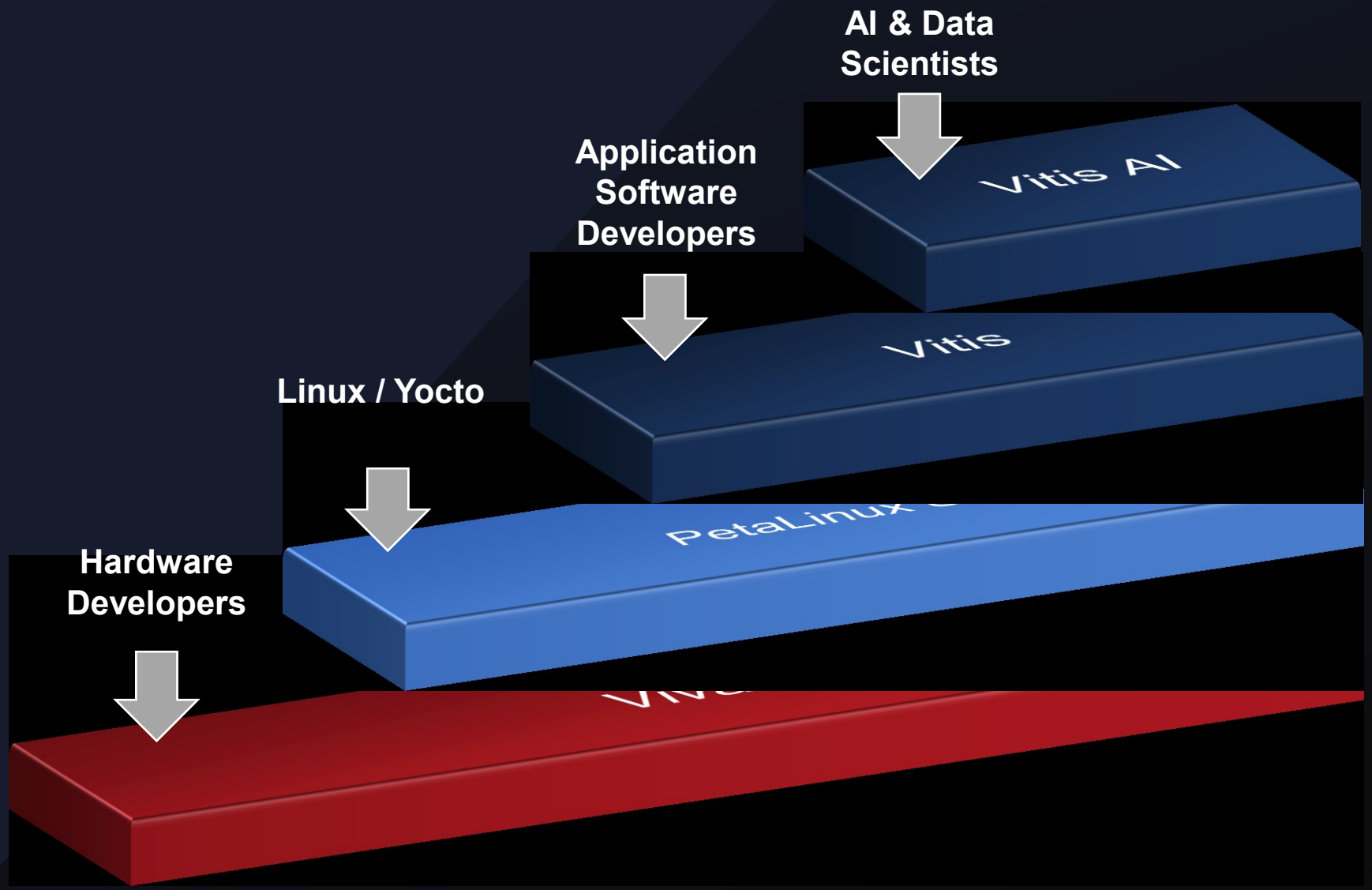
AI Enablement

Cloud to Edge

Heterogeneous Computing

© Copyright 2019 Xilinx

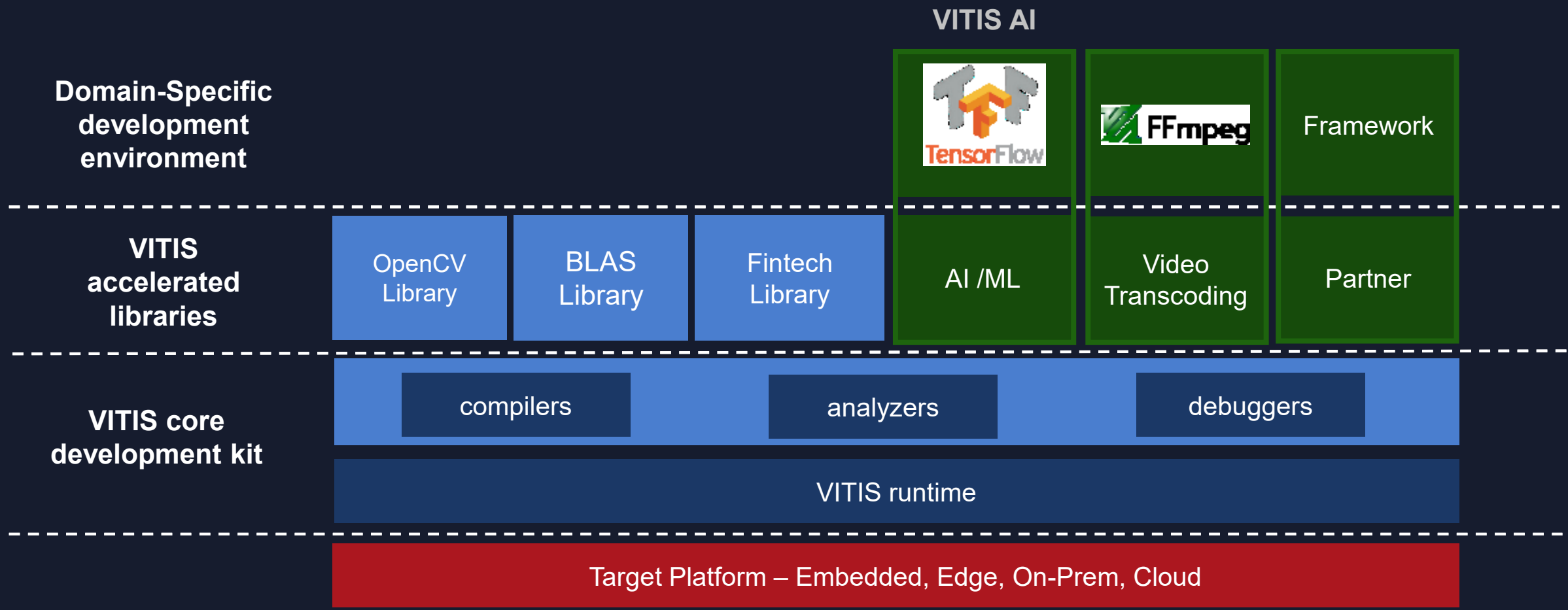
Development Platforms for ALL Developers



OPEN SOURCE

© Copyright 2019 Xilinx

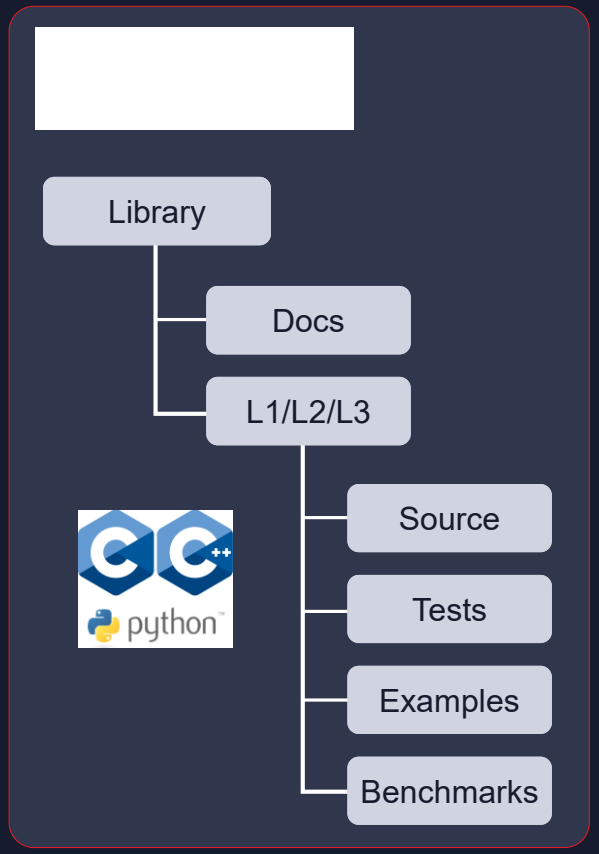
VITIS: Unified Software Platform

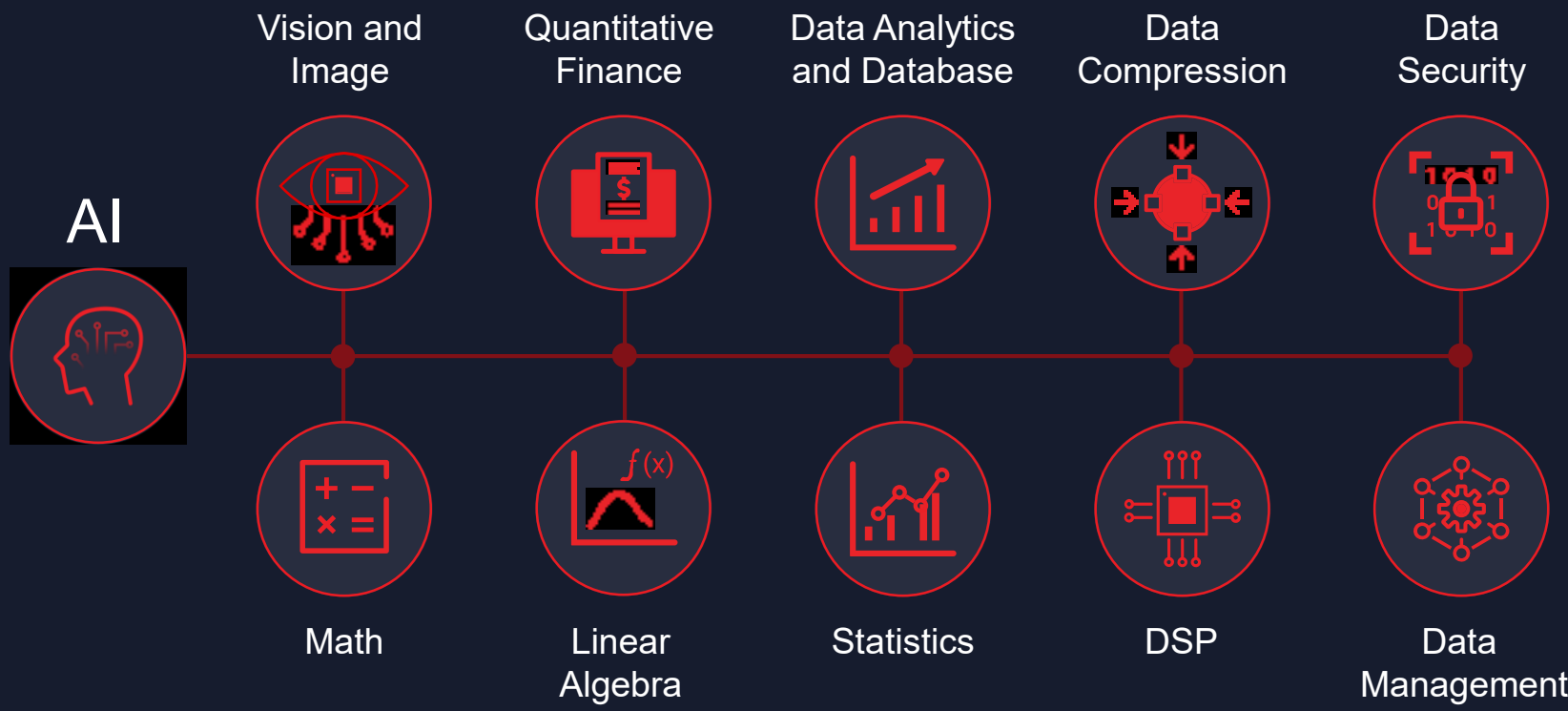


© Copyright 2019 Xilinx

VITIS Accelerated Libraries

- Open-source, performance-optimized libraries offering out-of-the-box acceleration
- Seamless deployment at the edge, on-premises or in the cloud





© Copyright 2019 Xilinx

VITIS Platforms Enable Composable Acceleration

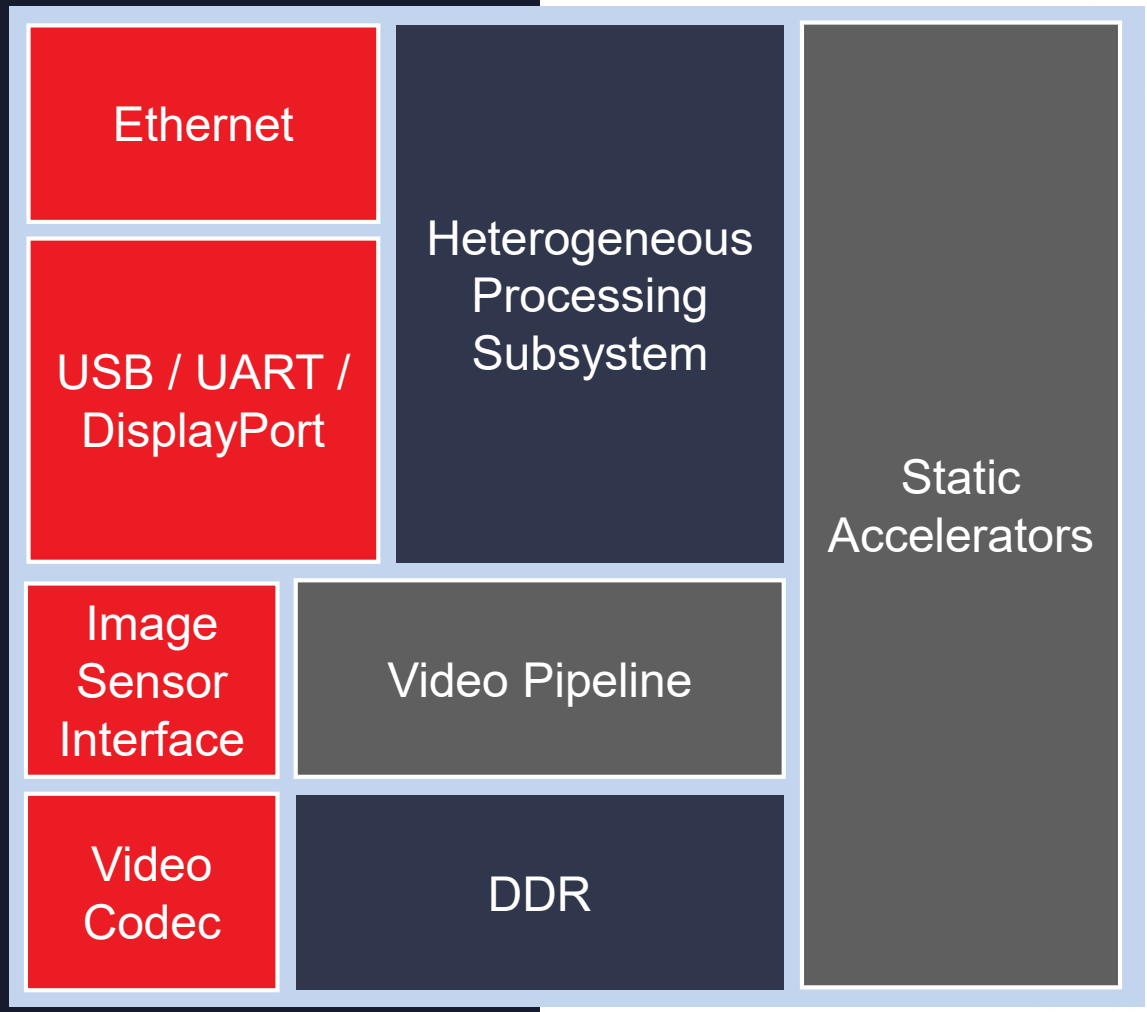
> Static hardware

- Fixed interfaces and acceleration components
  - Custom SoC paradigm
- End-user can deploy custom models trained for specific use cases – classification, detection, segmentation, etc.
  - Supporting Python, C, C++ APIs
- The perfect paradigm for Machine Vision, Retail Analytics, Smart Cities and other Smart Camera applications

> Dynamic hardware

- All of the above, plus:
- Dynamic acceleration components
  - Eg: customized pre/post processing
    - RTL, IP, HLS OpenCL, C, C++
- The perfect paradigm for advanced acceleration models where user customization is desired
  - Reconfigurable SoC paradigm

Smart Camera Platform



Concept credit: Parker Holloway

© Copyright 2019 Xilinx

VITIS AI

Frameworks

Caffe | Darknet | Keras | dmlc mxnet | PyTorch | TensorFlow

Models

Xilinx Pre-optimized | Open Source | Custom

Classification | Detection | Semantic Segmentation | RNN | Multi-Layer Perceptron | Multi-task Multi-Model Multi-Stream

Vitis AI Development Kit

AI Optimizer

AI Quantizer

AI Compiler

Xilinx Runtime library (XRT)

AI Profiler

Domain Specific Architectures

CNN DPU | LSTM DPU | MLP DPU

Platforms

ZCU102

ZCU104

Ultra96

Custom Hardware

Alveo

HUAWEI

aws marketplace | NIMBIX | Microsoft Azure

Cloud

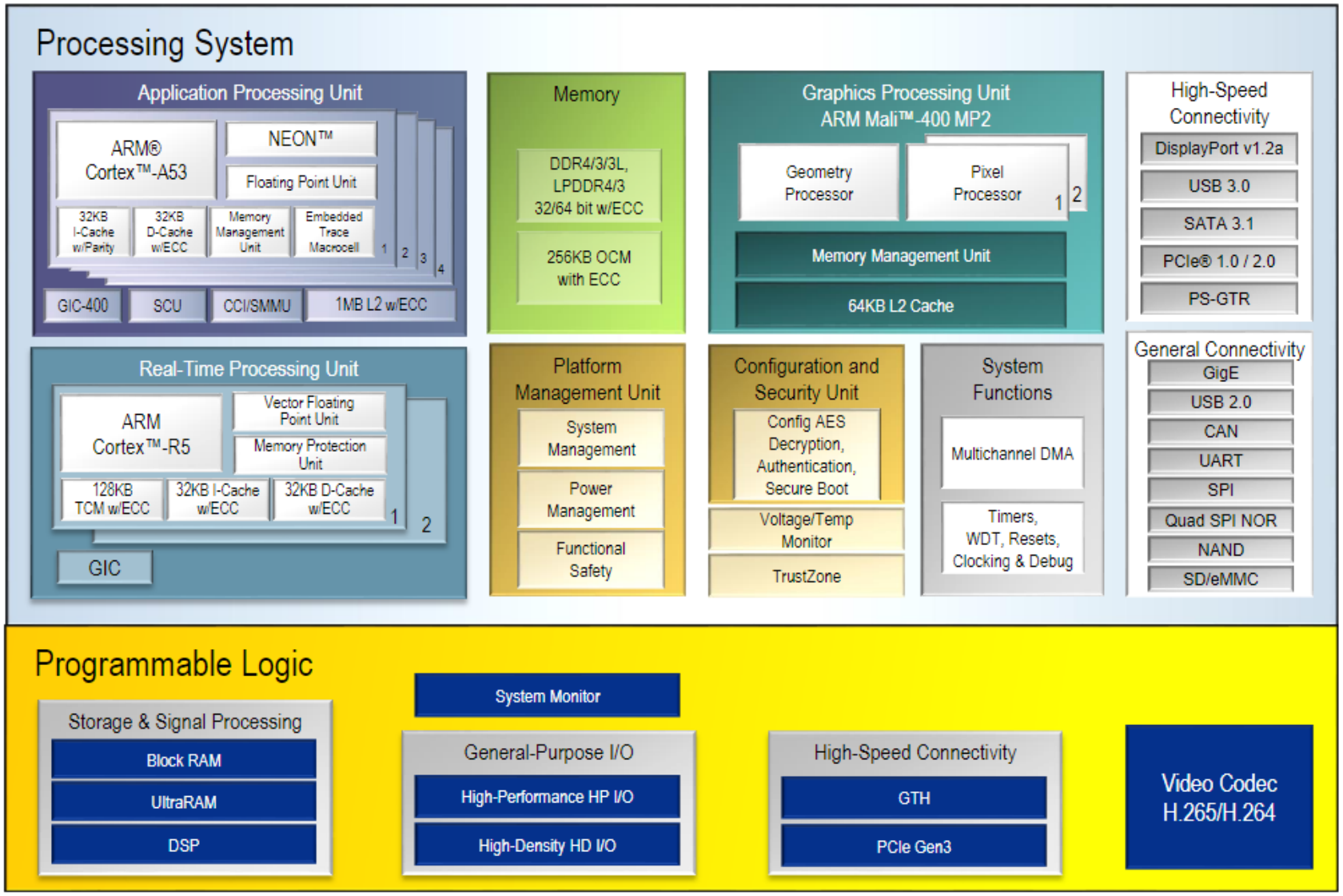
© Copyright 2019 Xilinx

So What About the Hardware?

Xilinx Smart Camera Hardware Solutions

© Copyright 2019 Xilinx

Zynq® UltraScale+™ MPSoCs: EV Block Diagram



© Copyright 2019 Xilinx

Embedding Scalability Without Compromise

- Common CPU and Infrastructure scales from GOPs to TOPs
- Dual and Quad Core Options
- Metadata + JPEG forwarding for bandwidth limited networks
- High-Bandwidth, High-Resolution Machine Vision
- Any sensor interface: MIPI, SLVS, SLVS-EC, USB, HDMI, DP, LVDS, TSN

Zynq® UltraScale+™ MPSoCs: EG Devices

		Device Name[1]	ZU2EG	ZU3EG	ZU4EG	ZU5EG	ZU6EG	ZU7EG	ZU9EG	ZU11EG	ZU15EG	ZU17EG	ZU19EG
Processing System (PS)	Application Processor Unit	Processor Core	**Quad-core** ARM® Cortex™-A53 MPCore™ up to 1.5GHz										
		Memory w/ECC	L1 Cache 32KB I / D per core, L2 Cache 1MB, on-chip Memory 256KB										
	Real-Time Processor Unit	Processor Core	**Dual-core** ARM Cortex-R5 MPCore™ up to 600MHz										
		Memory w/ECC	L1 Cache 32KB I / D per core, Tightly Coupled Memory 128KB per core										
	Graphic & Video Acceleration	Graphics Processing Unit	Mali™-400 MP2 up to 667MHz										
		Memory	L2 Cache 64KB										
	External Memory	Dynamic Memory Interface	x32/x64: DDR4, LPDDR4, DDR3, DDR3L, LPDDR3 with ECC										
		Static Memory Interfaces	NAND, 2x Quad-SPI										
	Connectivity	High-Speed Connectivity	PCIe® Gen2 x4, 2x USB3.0, SATA 3.1, DisplayPort, 4x Tri-mode Gigabit Ethernet										
		General Connectivity	2xUSB 2.0, 2x SD/SDIO, 2x UART, 2x CAN 2.0B, 2x I2C, 2x SPI, 4x 32b GPIO										
	Integrated Block Functionality	Power Management	Full / Low / PL / Battery Power Domains										
		Security	RSA, AES, and SHA										
		AMS - System Monitor	10-bit, 1MSPS – Temperature and Voltage Monitor										
PS to PL Interface			12 x 32/64/128b AXI Ports										
Programmable Logic (PL)	Programmable Functionality	System Logic Cells (K)	103	154	192	256	469	504	600	653	747	926	1,143
		CLB Flip-Flops (K)	94	141	176	234	429	461	548	597	682	847	1,045
		CLB LUTs (K)	47	71	88	117	215	230	274	299	341	423	523
	Memory	Max. Distributed RAM (Mb)	1.2	1.8	2.6	3.5	6.9	6.2	8.8	9.1	11.3	8.0	9.8
		Total Block RAM (Mb)	5.3		4.5	5.1	25.1	11.0	32.1	21.1	26.2		34.6
		UltraRAM (Mb)	-										36.0
	Clocking	Clock Management Tiles (CMTs)	3										11
	Integrated IP	DSP Slices											
		PCI Express® Gen 3x16											
		150G Interlaken	-										4
		100G Ethernet MAC/PCS w/RS-FEC	-		-	-	-	-	-	2	-		4
		AMS - System Monitor	1		1	1	1	1	1	1	1		1
	Transceivers	GTH 16.3Gb/s Transceivers	-	-	16	16	24	24	24	32	24	44	44
		GTY 32.75Gb/s Transceivers	-	-	-	-	-	-	-	16	-	28	28
	Speed Grades	Extended[2]	-1 -2 -2L		-1 -2 -2L -3				-1 -2 -2L -3				
		Industrial	-1 -1L -2										

Performance scaling by more than one order of magnitude with no architecture changes!

Notes:
1. For full part number details, see the Ordering Information section in DS891, *Zynq UltraScale+ MPSoC Overview*.
2. -2LE (Tj = 0°C to 110°C). For more details, see the Ordering Information section in DS891, *Zynq UltraScale+ MPSoC Overview*.

© Copyright 2019 Xilinx

Scalable Video Compression

- Multi-resolution multi-stream encoding
- AI Appliances
- Multi-sensor cameras
- Smart City
- Smart Retail
- Smart Advertising
- 4k@60 Encode
- Any sensor interface: MIPI, SLVS, SLVS-EC, USB, HDMI, DP, LVDS, TSN

Zynq® UltraScale+™ MPSoCs: EV Devices

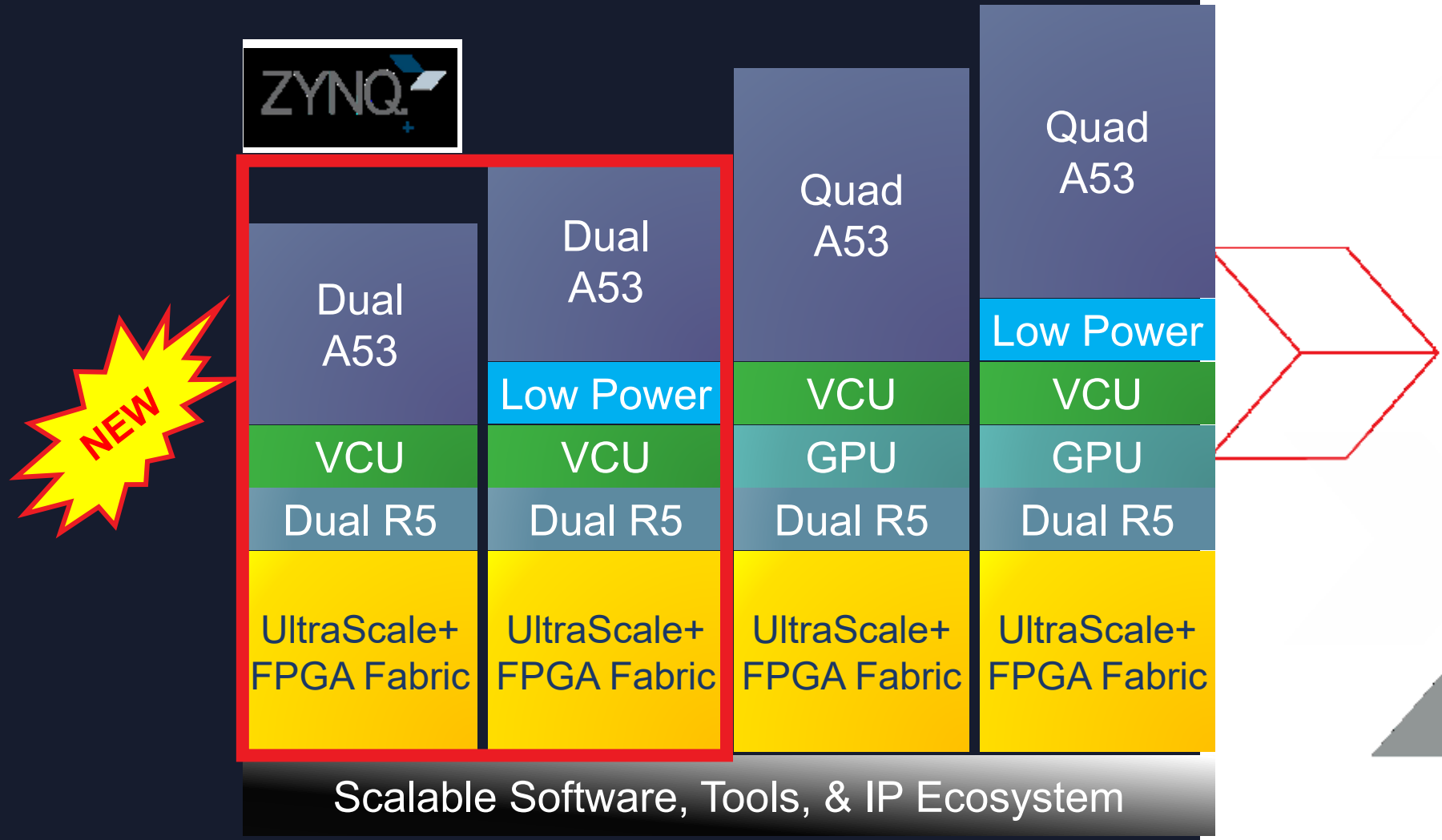
		Device Name[1]	ZU4EV	ZU5EV	ZU7EV
Processing System (PS)	Application Processor Unit	Processor Core	**Quad-core** ARM® Cortex™-A53 MPCore™ up to 1.5GHz		
		Memory w/ECC	L1 Cache 32KB I / D per core, L2 Cache 1MB, on-chip Memory 256KB		
	Real-Time Processor Unit	Processor Core	**Dual-core** ARM Cortex-R5 MPCore™ up to 600MHz		
		Memory w/ECC	L1 Cache 32KB I / D per core, Tightly Coupled Memory 128KB per core		
	Graphic & Video Acceleration	Graphics Processing Unit	Mali™-400 MP2 up to 667MHz		
		Memory	L2 Cache 64KB		
	External Memory	Dynamic Memory Interface	x32/x64: DDR4, LPDDR4, DDR3, DDR3L, LPDDR3 with ECC		
		Static Memory Interfaces	NAND, 2x Quad-SPI		
	Connectivity	High-Speed Connectivity	PCIe® Gen2 x4, 2x USB3.0, SATA 3.1, DisplayPort, 4x Tri-mode Gigabit Ethernet		
		General Connectivity	2xUSB 2.0, 2x SD/SDIO, 2x UART, 2x CAN 2.0B, 2x I2C, 2x SPI, 4x 32b GPIO		
	Integrated Block Functionality	Power Management	Full / Low / PL / Battery Power Domains		
		Security	RSA, AES, and SHA		
		AMS - System Monitor	10-bit, 1MSPS – Temperature and Voltage Monitor		
PS to PL Interface			12 x 32/64/128b AXI Ports		
Programmable Logic (PL)	Programmable Functionality	System Logic Cells (K)	192	256	504
		CLB Flip-Flops (K)	176	234	461
		CLB LUTs (K)	88	117	230
	Memory	Max. Distributed RAM (Mb)	2	3.5	2
		Total Block RAM (Mb)		5.1	
		UltraRAM (Mb)			
	Clocking	Clock Management Tiles (CMTs)			
	Integrated IP	DSP Slices			
		Video Codec Unit (VCU)			
		PCI Express® Gen 3x16			
		150G Interlaken			
		100G Ethernet MAC/PCS w/RS-FEC		-	
		AMS - System Monitor	1	1	1
	Transceivers	GTH 16.3Gb/s Transceivers	16	16	24
		GTY 32.75Gb/s Transceivers	-	-	-
	Speed Grades	Extended[2]	-1 -2 -2L -3		
		Industrial	-1 -1L -2		

Scalable Inference With Integrated Video Compression

© Copyright 2019 Xilinx

New Scalable Platforms for Smart Cities, Retail

New Product Options for Smart City, Retail applications, based on the **XCZU4EV-1SFVC784I**



© Copyright 2019 Xilinx

CNN DPU Overlay IP

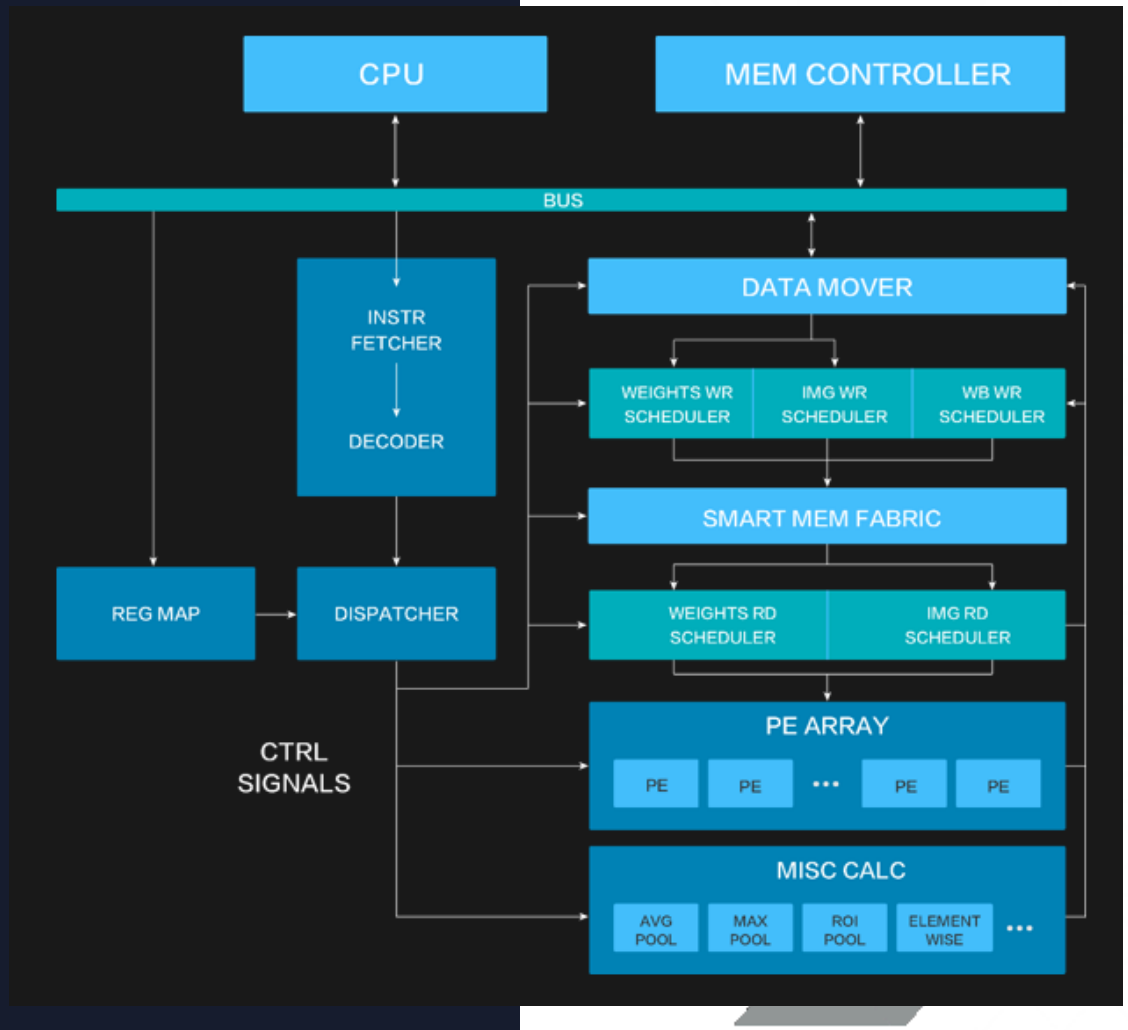
> Instruction Set

- >> Tensor based instructions
- >> Up to 268,435,456 MACs/instruction

> Hardware Architecture

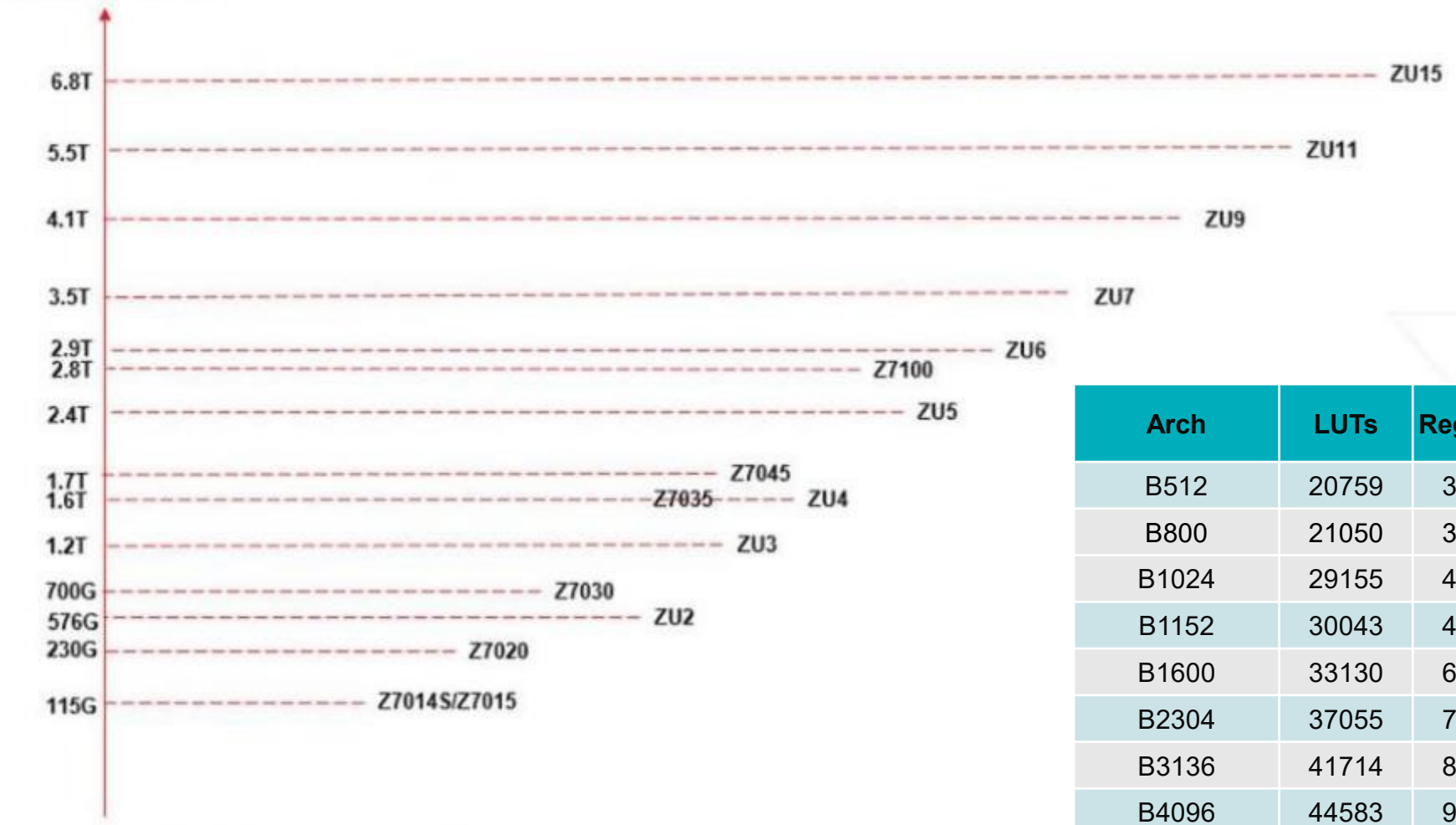
- >> Multi-dimension parallelism
- >> Configurable and extensible
- >> Highly optimized micro-architecture
  - – Smart instruction merging and splitting
  - – On-chip memory pool with high internal bandwidth
  - – Intelligent DMA engine optimized for CNN patterns

> Simple C/C++/Python API



© Copyright 2019 Xilinx

High Computational Cost? No Problem…



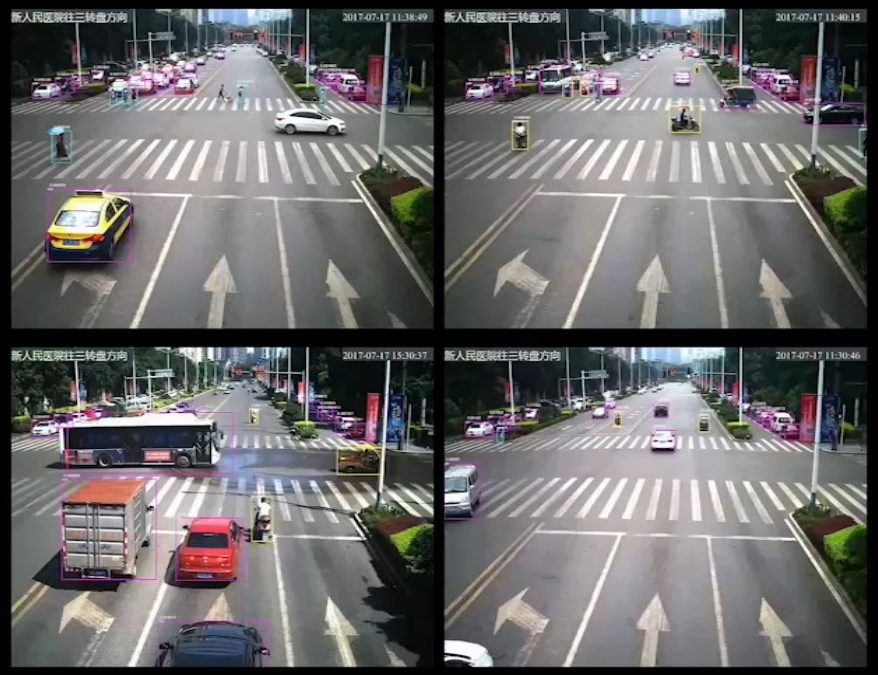
Arch	LUTs	Registers	BRAM	DSP
B512	20759	33572	69.5	66
B800	21050	33752	87	102
B1024	29155	49823	101.5	130
B1152	30043	49588	117.5	146
B1600	33130	60739	123	202
B2304	37055	72850	161.5	290
B3136	41714	86132	203.5	394
B4096	44583	99791	249.5	514

Please leverage the Vivado IP wizard for DPU configuration specific resource utilization numbers.

© Copyright 2019 Xilinx

Acceleration on Hardware – VGG-SSD

Before Compression
18FPS@117GOPs

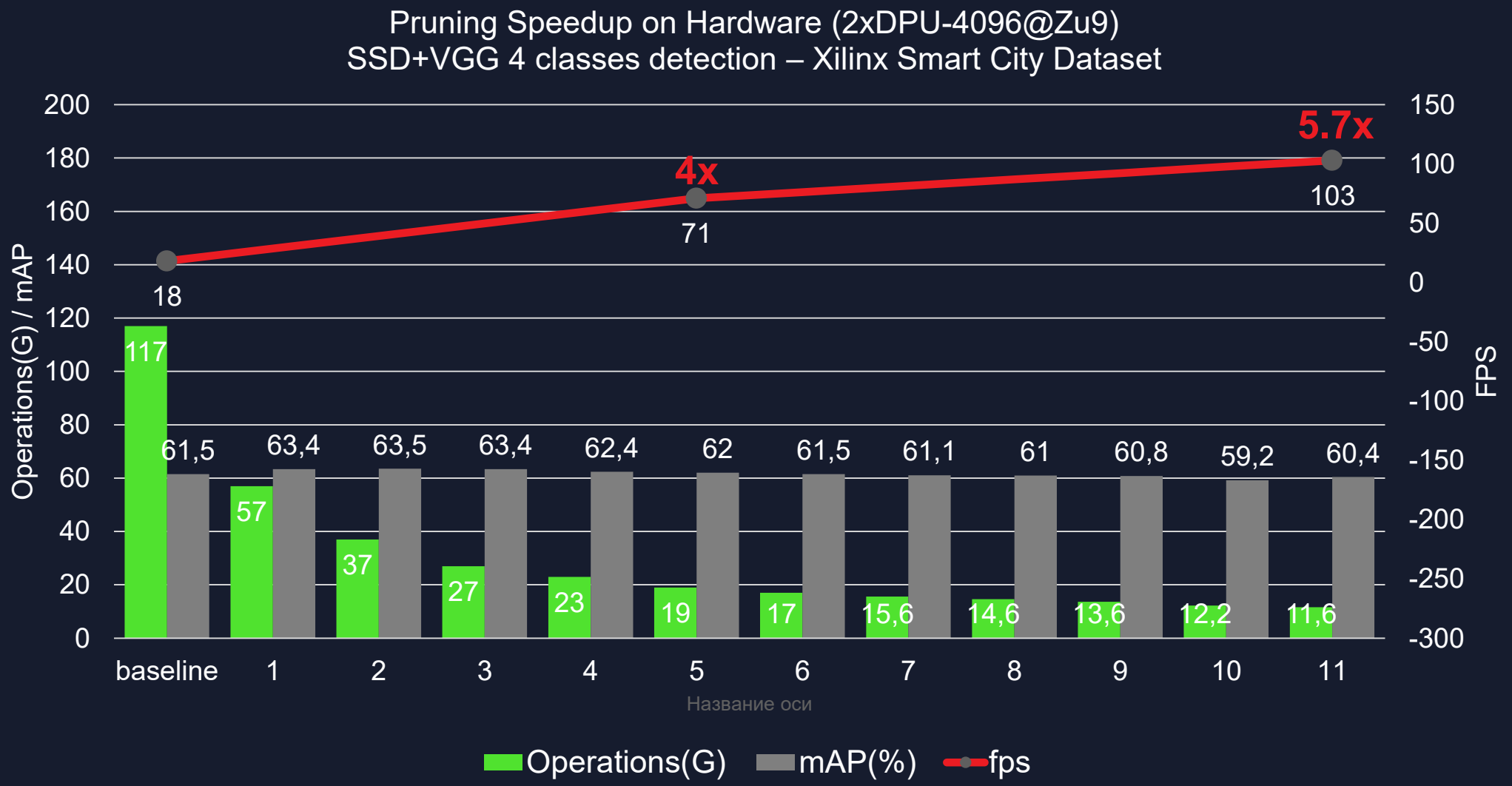


After 10X Compression
103FPS@11.6GOPs

XILINX

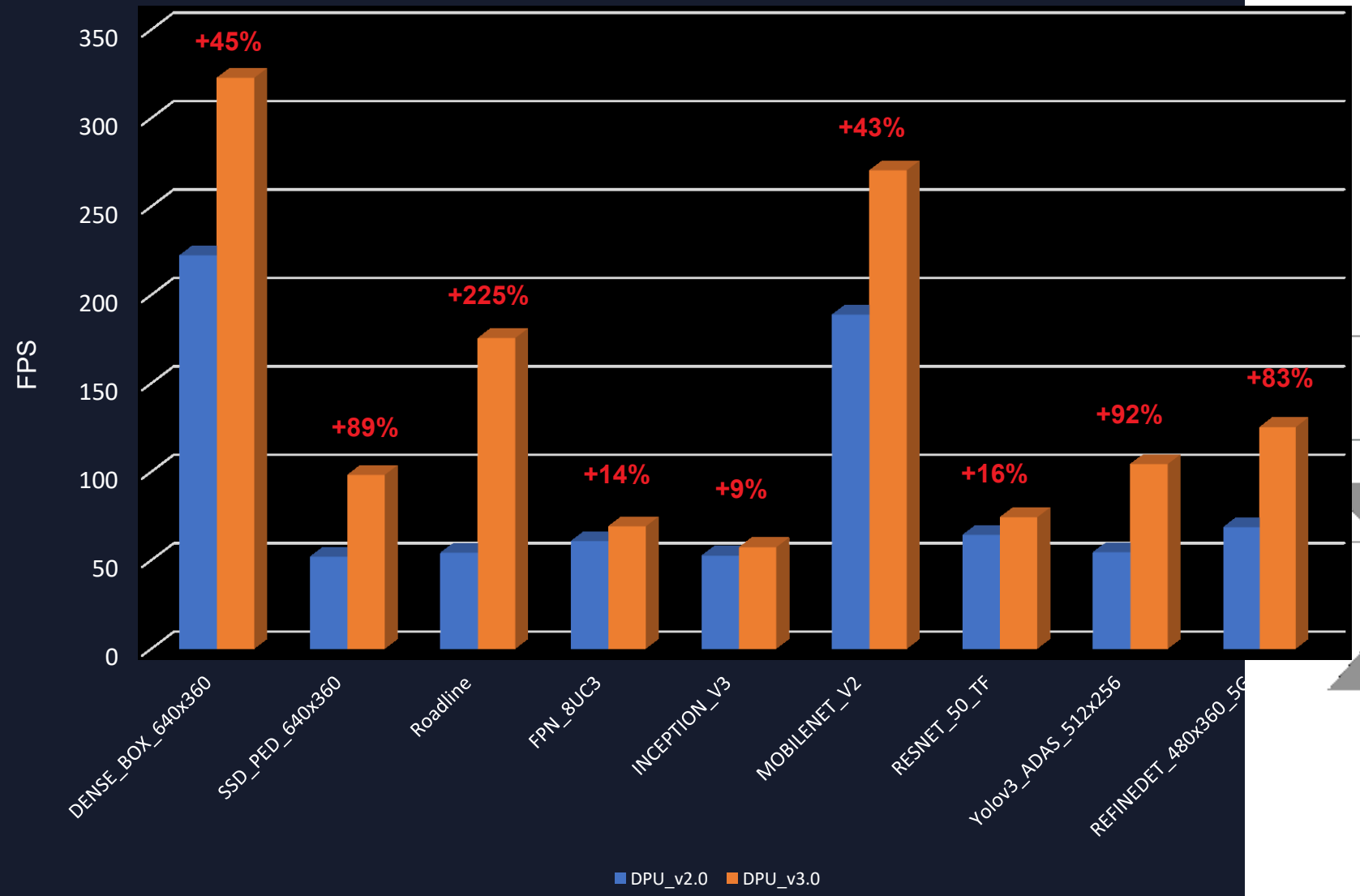
© Copyright 2019 Xilinx

Pruning Reduces Computational Cost



© Copyright 2019 Xilinx

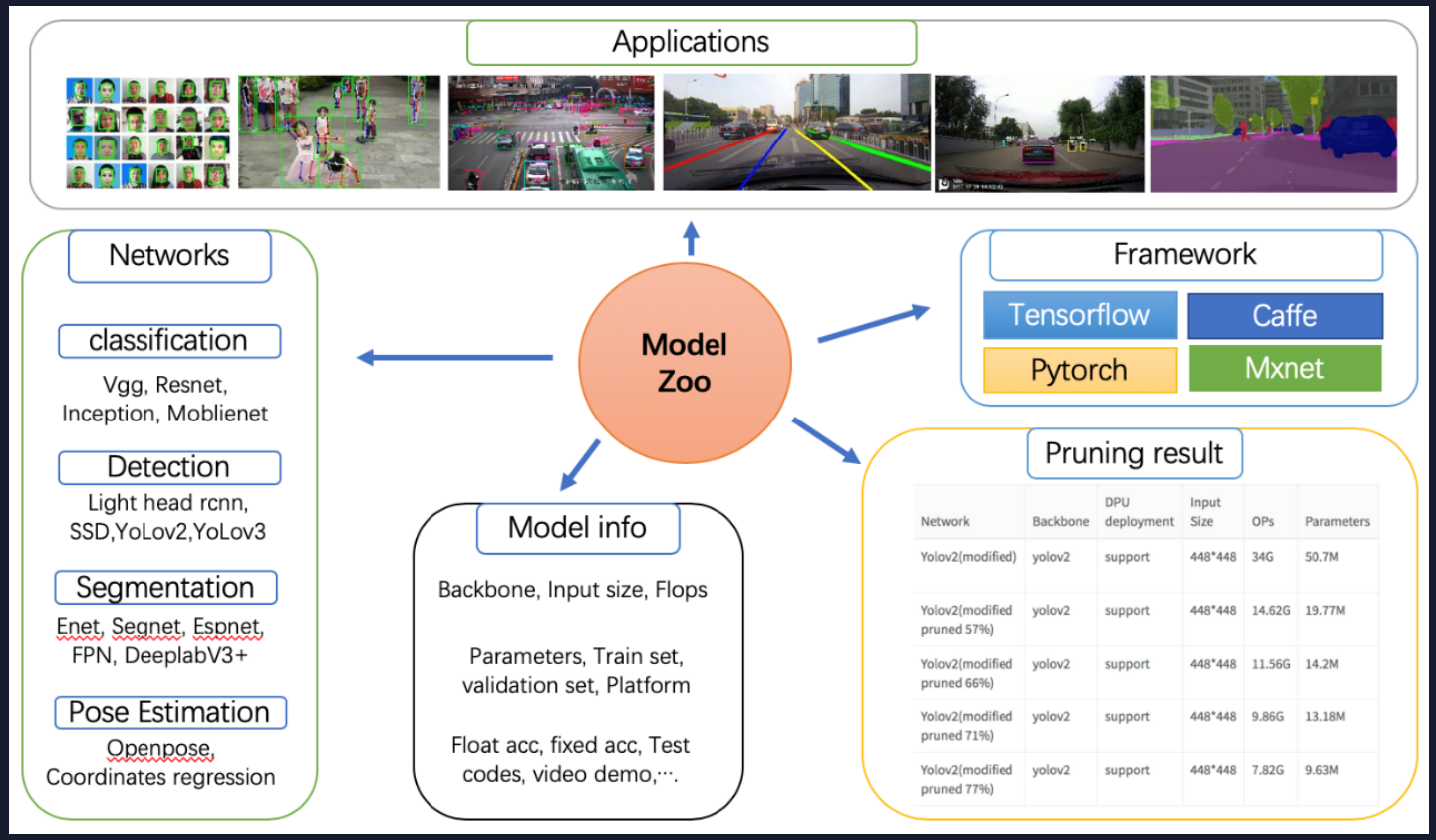
Recent DPU Performance Improvements



© Copyright 2019 Xilinx

VITIS AI Model Zoo

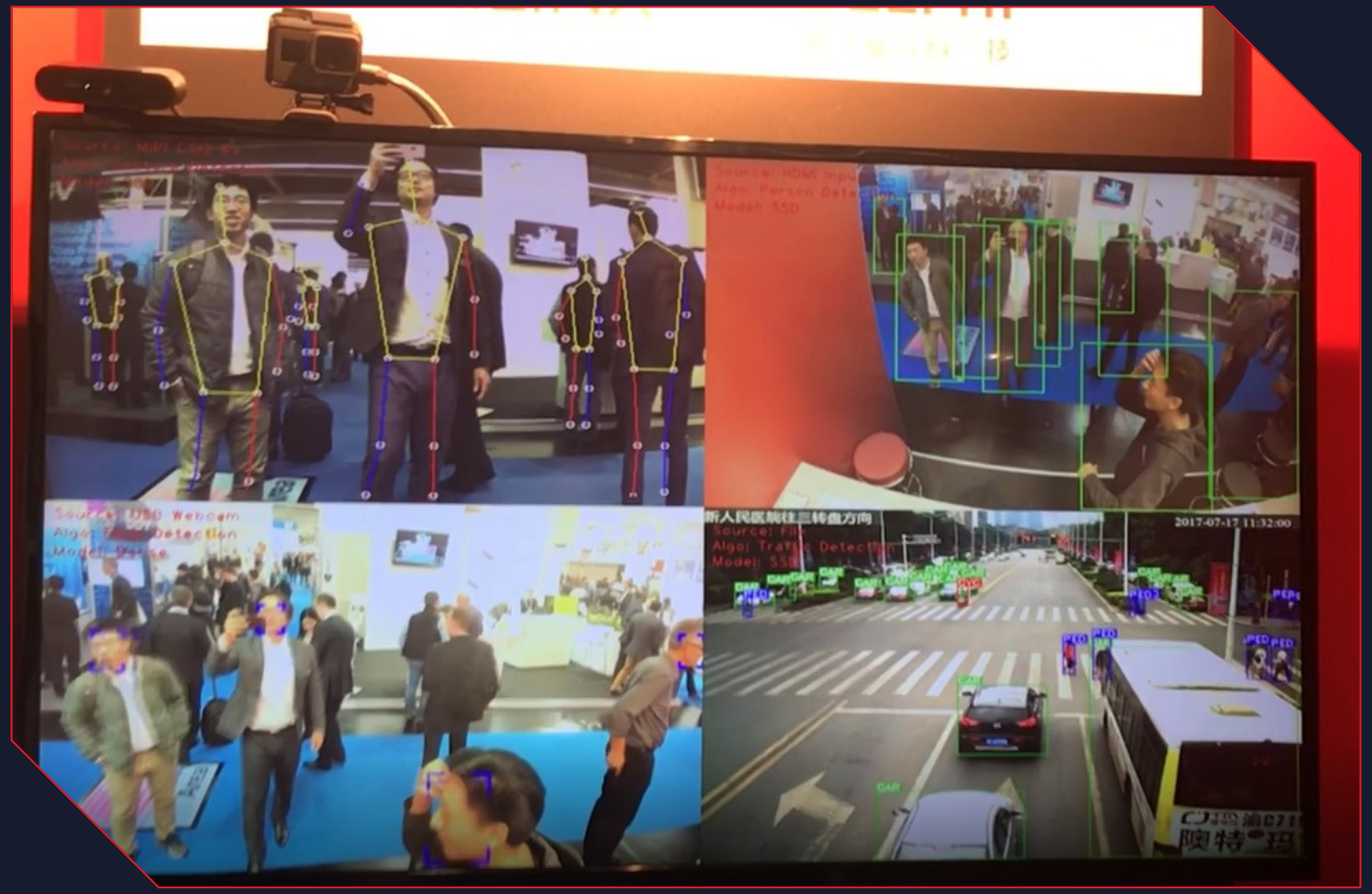
- ✓ Open for all users
- ✓ Leveraging mainstream frameworks and networks
- ✓ Deployable and re-trainable



Application	Module
Face	Face detection
	Landmark Localization
	Face recognition
	Face attributes recognition
Pedestrian	Pedestrian Detection
	Pose Estimation
	Person Re-identification
Video Analytics	Object detection
	Pedestrian Attributes Recognition
	Car Attributes Recognition
	Car Logo Detection
	Car Logo Recognition
	License Plate Detection
	License Plate Recognition
ADAS/AD	Object Detection
	3D Car Detection
	Lane Detection
	Traffic Sign Detection
	Semantic Segmentation
	Drivable Space Detection

© Copyright 2019 Xilinx

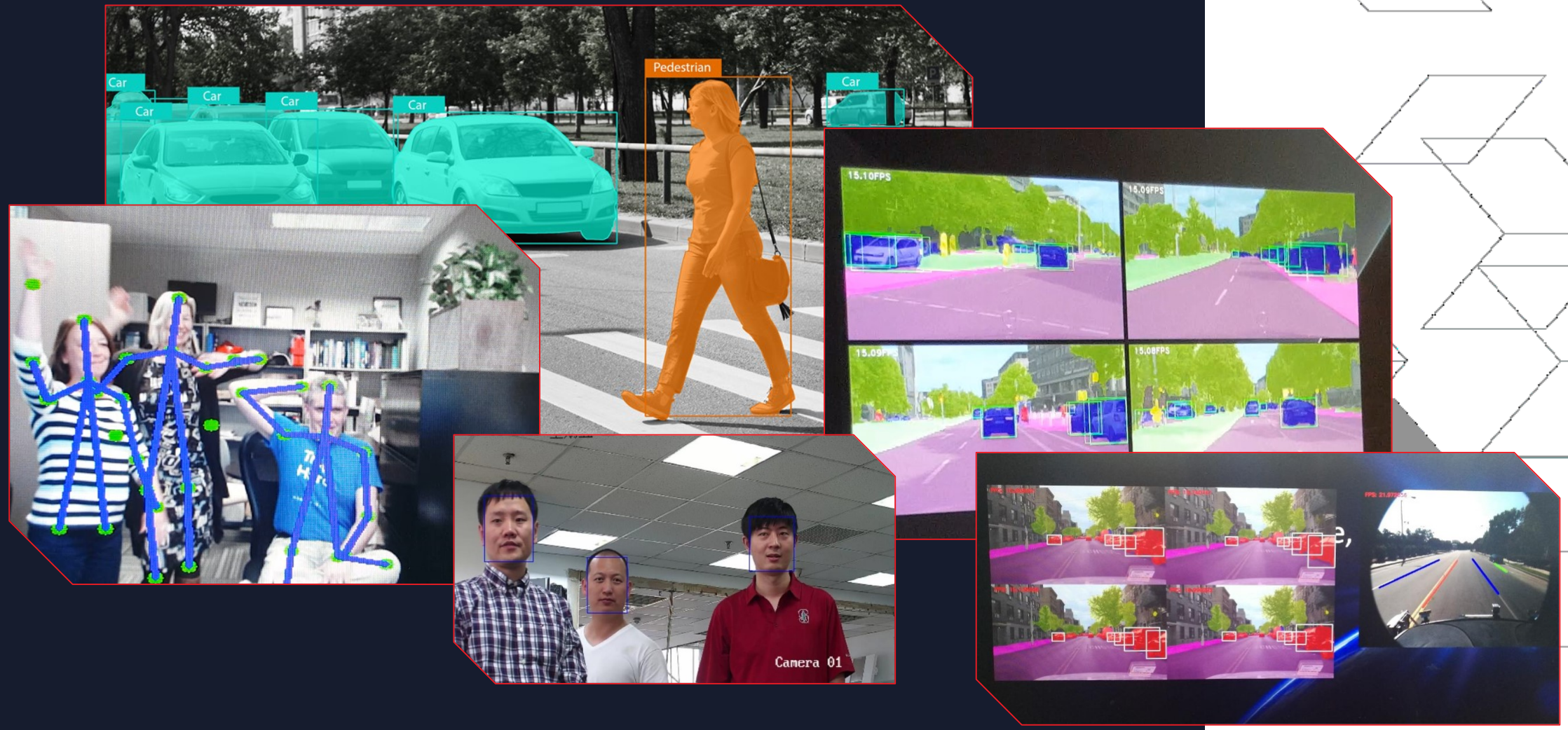
Single-chip Deployment of Multiple Models



- ✓ Multi-task
- ✓ Multi-mode
- ✓ Multi-frame
- ✓ Cascaded
- ✓ One or mo U instan
- ✓ Custom lay
- ✓ Graph segm tion
- ✓ One bitstre orts mar

© Copyright 2019 Xilinx

Edge Deployment of Custom Models



© Copyright 2019 Xilinx

AI Tutorials

482 lines (481 sloc) | 16.7 KB | Raw | Blame | History

CIFAR-10 image classification with TensorFlow
Part 1 - Training and evaluation

This Jupyter NoteBook will explain how to build a miniVGGNet CNN for classifying the CIFAR-10 dataset using the TensorFlow layers API, then how to train and evaluate it. The complete python code (cifar10_train.py and miniVGGNet.py) can be found in this GitHub repo.

The MiniVGGNet CNN was developed by Adrian Rosebrock and looks like this:

First, we import the necessary Python pa

In [ ]:
```
import os
import sys
import shutil
import numpy as np

import tensorflow as tf
```

Now we create some directories for th
already exist, we delete them and recrea

In [ ]:
```
SCRIPT_DIR = os.getcwd()

INFER_GRAPH = 'inference_graph.p
CHKPT_FILE = 'float_model.ckpt'
```

Freeze the Floating-Point Model

Now that we have saved the trained parameters of our network as a checkpoint and graph, we need to 'freeze' it by converting all the variables into constants and stripping out the training nodes to leave just the nodes that we need for deployment.

Luckily, TensorFlow provides a script called 'freeze_graph.py' which will do this for us..

In [ ]:
```
# delete previous results
!rm -rf ./freeze
!mkdir ./freeze

!freeze_graph --input_graph=./chkpts/training_graph.pb \
              --input_checkpoint=./chkpts/float_model.ckpt \
              --input_binary=true \
              --output_graph=./freeze/frozen_graph.pb \
              --output_node_names=dense_1/BiasAdd
```

Quantize the floating-point model

This is the first step that involves the Xilinx™ DNNDK toolkit. We will now quantize the floating point model by running DECENT_Q.
The quantization process has a calibration phase which requires approximately 1000 images files in a format that is compatible with openCV.
This section of python code will fetch the MNIST dataset as numpy arrays and the convert the test dataset (10k arrays) into 10k .png image files and write them into a separate folder.
It will also create a text file in that same folder which lists the images - this is also required for calibration.

In [ ]:
```
import os
import shutil
import numpy as np
import cv2

from keras.datasets import mnist

SCRIPT_DIR = os.getcwd()
CALIB_DIR = os.path.join(SCRIPT_DIR, 'calib_dir')
```

XILINX

Edge AI Tutorials

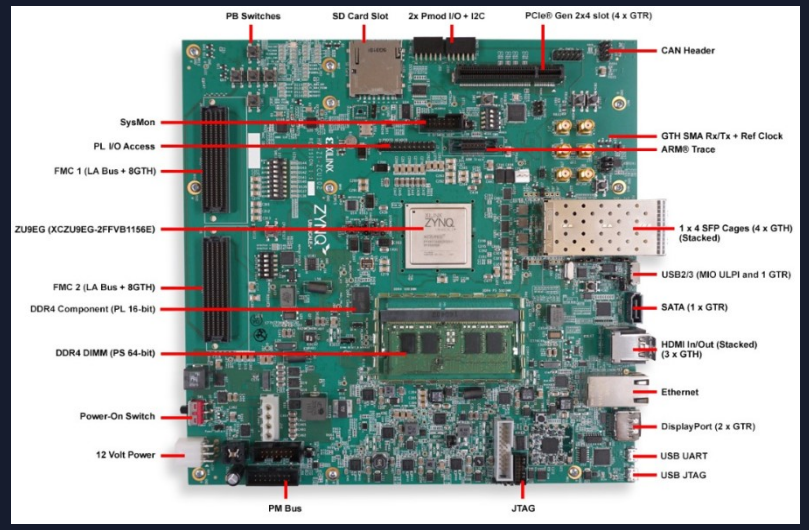
Tutorial	Description
CIFAR10 Caffe Tutorial (UG1335)	Train, quantize, and prune custom CNNs with the CIFAR10 dataset using Caffe and the Xilinx® DNNDK tools.
Cats vs Dogs Tutorial (UG1336)	Train, quantize, and prune a modified AlexNet CNN with the Kaggle Cats vs Dogs dataset using Caffe and the Xilinx DNNDK tools.
ML SSD PASCAL Caffe Tutorial (UG1340)	Train, quantize, and compile SSD using PASCAL VOC 2007/2012 datasets with the Caffe framework and DNNDK tools, then deploy on a Xilinx ZCU102 target board.
DPU Integration Lab (UG1350)	Build a custom system that utilizes the Xilinx Deep Learning Processor (DPU) IP to accelerate machine learning algorithms.
Yolov3 Tutorial with Darknet to Caffe Converter and Xilinx DNNDK (UG1334)	Use the Yolov3 example, which converts the Darknet model to Caffe model and uses the DNNDK tool chain for quantization, compilation, and deployment on the FPGA.
MNIST Classification with TensorFlow (UG1337)	Learn the DNNDK v3.0 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 with TensorFlow (UG1338)	Learn the DNNDK v3.0 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.
Freezing a Keras model for use with DNNDK (UG1380)	Freeze a Keras model by generating a binary protobuf (.pb) file.
Deep Learning with custom GoogleNet and ResNet in Keras and Xilinx DNNDK TF 3.0 (UG1381)	Quantize in fixed point some custom CNNs and deploy them on the Xilinx ZCU102 board, using Keras and the Xilinx DNNDK 3.0 tool chain based on TensorFlow (TF).

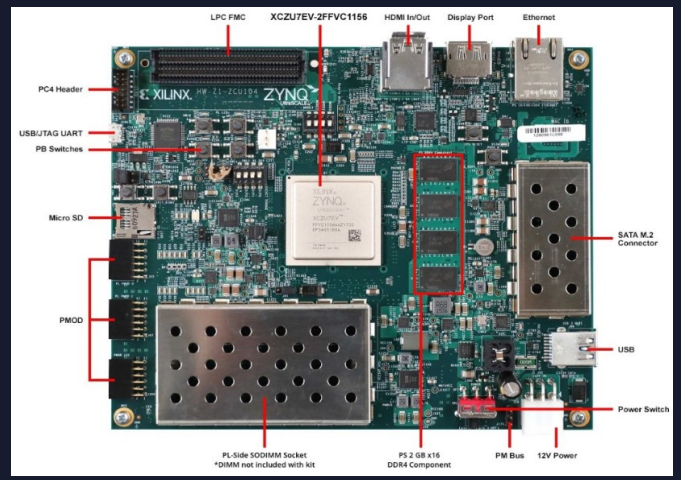
Copyright© 2019 Xilinx

© Copyright 2019 Xilinx

Slide credit: Clayton Cameron

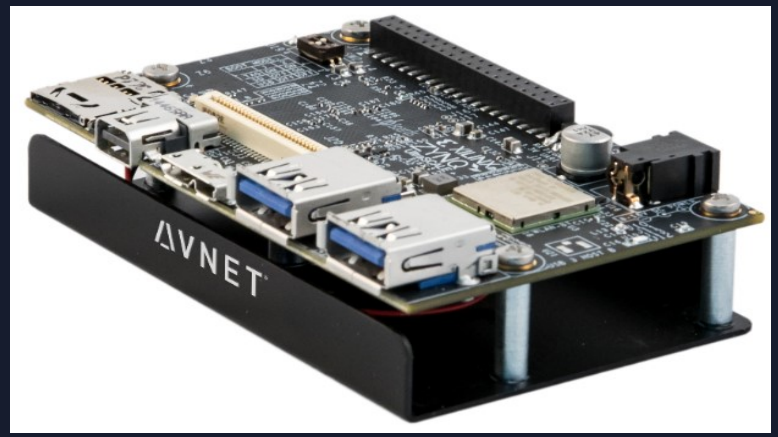
AI Developer Hub

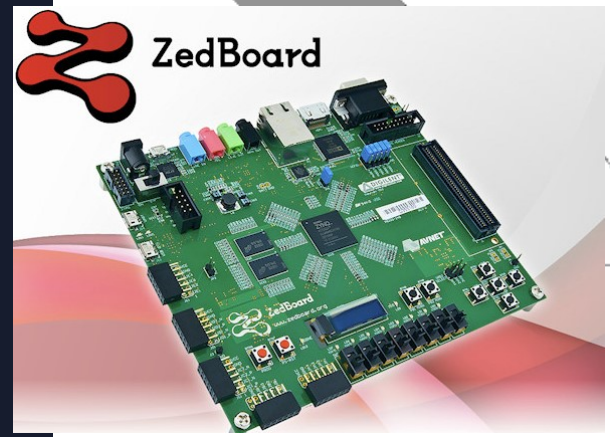




Edge AI Evaluation Boards

Product	Documentation	Image Download	DNNDK Version	File Size	MD5 Checksum
ZCU102 Kit	ZCU102 User Guide (UG1182)	petalinux-user-image-zcu102-zynqmp-sd-20190802.img.gz	V3.1	311 MB	1d67b1380ffef0da18d923deb92b5f6b
		xilinx-zcu102-prod-dpu1.4-2018.3-desktop-buster-2019-04-24.img.zip	v3.0	657 MB	d49eab4d293d8d1af40fcc369e1c4f53
		2018-12-04-zcu102-desktop-stretch.img.zip	v2.08	571 MB	d0d5faf8ece80b96f5591d09756d5a5d
ZCU104 Kit	ZCU104 User Guide (UG1267)	petalinux-user-image-zcu104-zynqmp-sd-20190802.img.gz	V3.1	311 MB	a238b286651fbb308cec923664980b34
		xilinx-zcu104-prod-dpu1.4-desktop-buster-2019-04-23.img.zip	v3.0	655 MB	503661dd1ee4549a562775034b95d0c8
		2018-12-04-zcu104-desktop-stretch.img.zip	v2.08	571 MB	ada2420c4afbd89efdeea741e0917e26
Avnet Ultra 96	Ultra 96 User Guide	petalinux-user-image-ultra96-zynqmp-sd-20190802.img.gz	v3.1	537 MB	aee8464d0fd52776567f28a221c80c71
		xilinx-ultra96-prod-dpu1.4-desktop-buster-2019-05-31.img.zip	v3.0	576 MB	c9c6a5f5a772077abc8fffde6ea8f3db
		xilinx-ultra96-desktop-stretch-2018-12-10.img.zip	v2.08	566 MB	c5d2422063213b4bc4c18a3223c6adc8
Avnet Zedboard	Zedboard User Guide	xilinx-zedboard-dnndk3.1-image-20190812.zip	v3.1	315 MB	36706f19ec949ad964f6ab328f874e88





Slide credit: Clayton Cameron

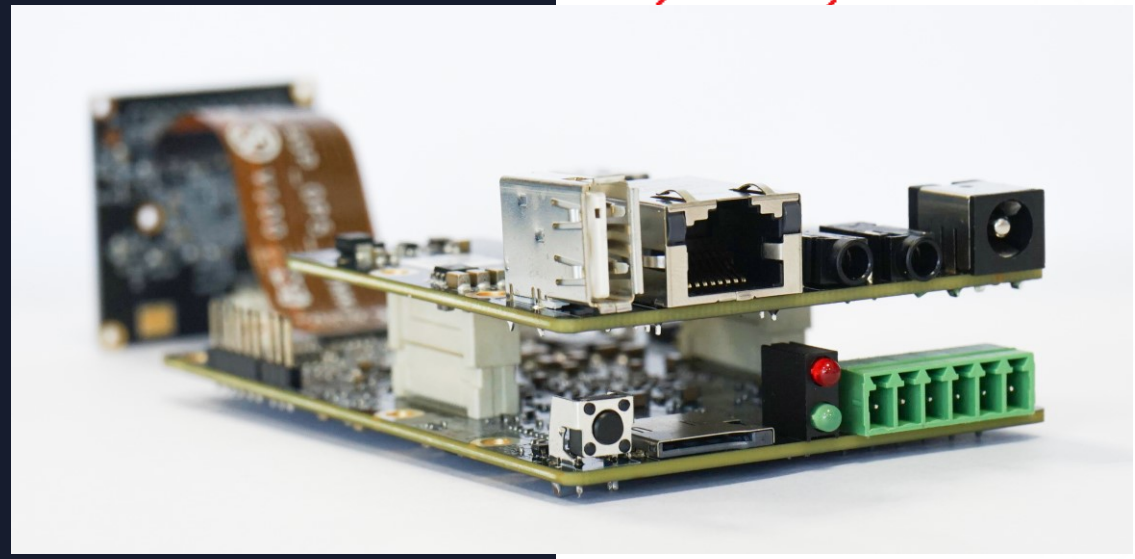
© Copyright 2019 Xilinx

Xilinx Smart Camera Demo Platform

Onsemi AP1302
MIPI Image Signal Processor
13MP@30, 1080p@120
HDR, Clarity+, Bayer, JPEG, YUV, Gamma, Face Detection

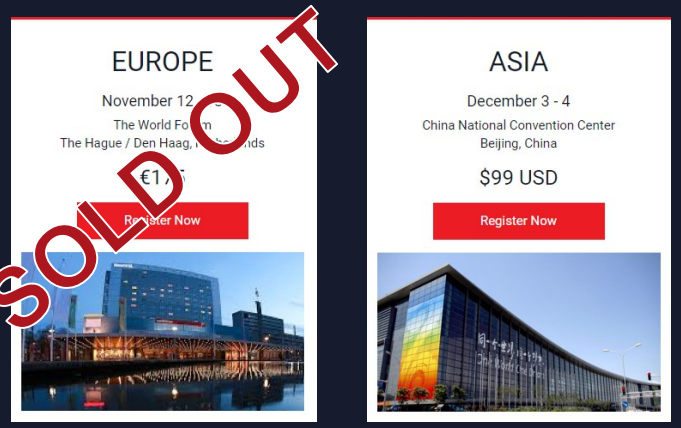
Onsemi AR1335 13MP
1/3” Progressive Scan
4208x3120, 30fps

- eMMC
- DisplayPort Output
- microSD
- Debug UART
- HTTP, RTSP, TCP/IP, UDP
- H.265/H.265/ MJPEG
- USB 2.0
- Audio Codec



© Copyright 2019 Xilinx

Next Steps & Conclusion



- Learn more about VITIS at XDF2019
  - www.xilinx.com/XDF2019
- Evaluate AI inference deployment and test-drive Xilinx pre-optimized models:
  - https://www.xilinx.com/products/design-tools/ai-inference/ai-developer-hub.html#edge
- Visit our booth at SPS Nuremberg, November 26-28, 2019

- VITIS available for download
  - https://www.xilinx.com/products/design-tools/vitis.html
- Contact your local Xilinx sales team for details of new ZU4EV domain-specific variants for Smart City and Retail apps

© Copyright 2019 Xilinx

© Copyright 2019 Xilinx