

Introducing the Vitis Unified Software Platform





WELCOME

All Developers

XDF 2019

Learn Directly From Experts

WELCOME

All Developers

76 HOURS OF TECHNICAL SESSIONS
IN 6 TRACKS

11 LABS
AVAILABLE FOR 20 HOURS

69 DEMOS
18 Xilinx Demos, 39 Partners,
12 Alveo Demos (Partners)





**Heterogeneous
Compute**



Cloud to Edge

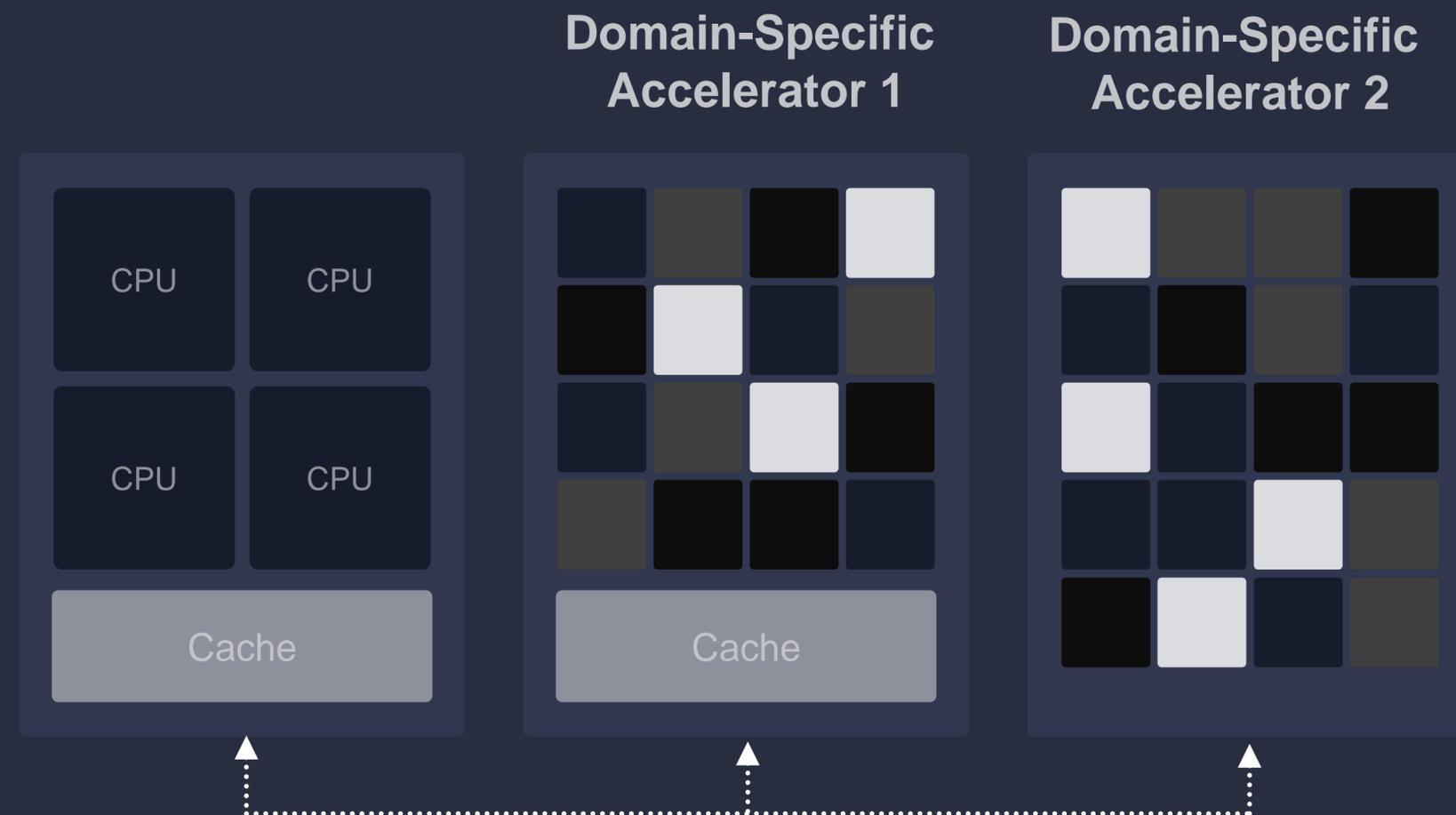


AI Proliferation

Industry Trend: Heterogeneous Compute



Engines Customized to Accelerate Specific Domains



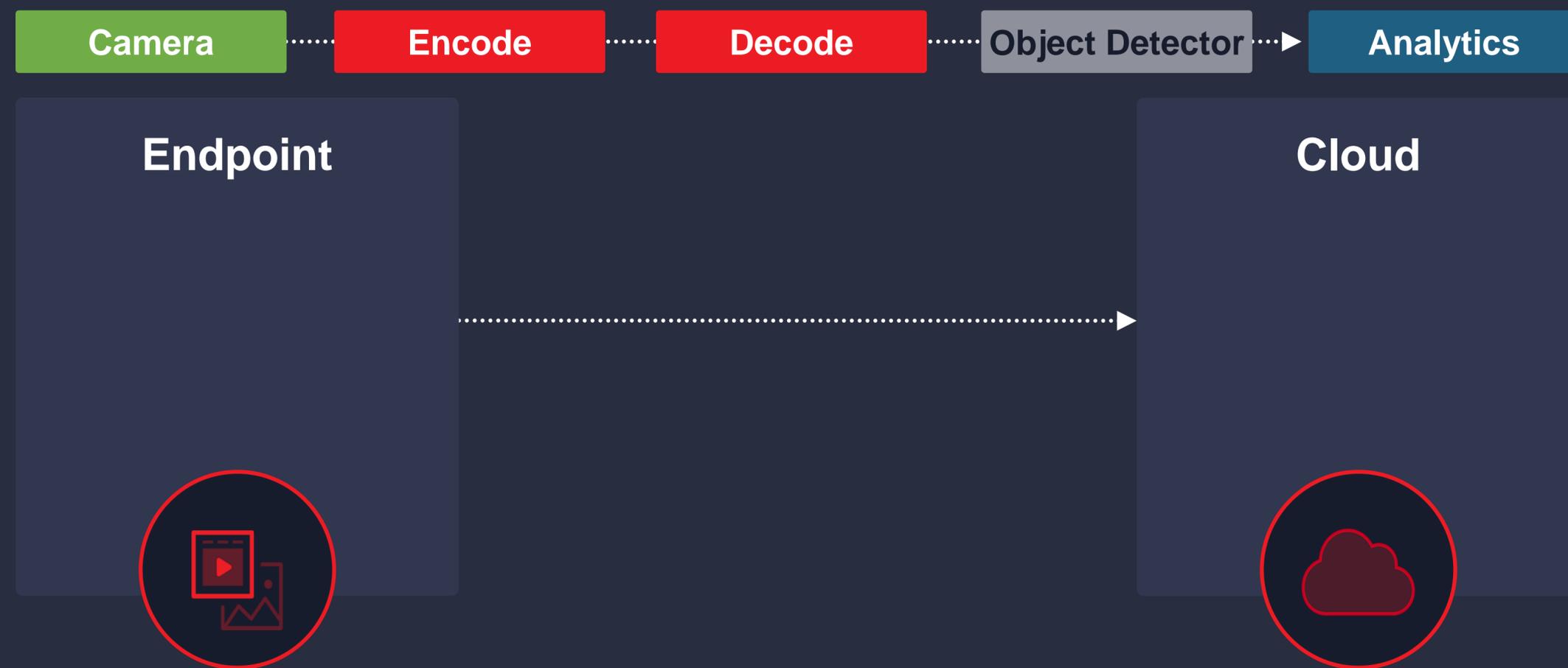
Key Challenge

Programming & integration of Adaptive Acceleration Engines

Industry Trend: Cloud to Edge



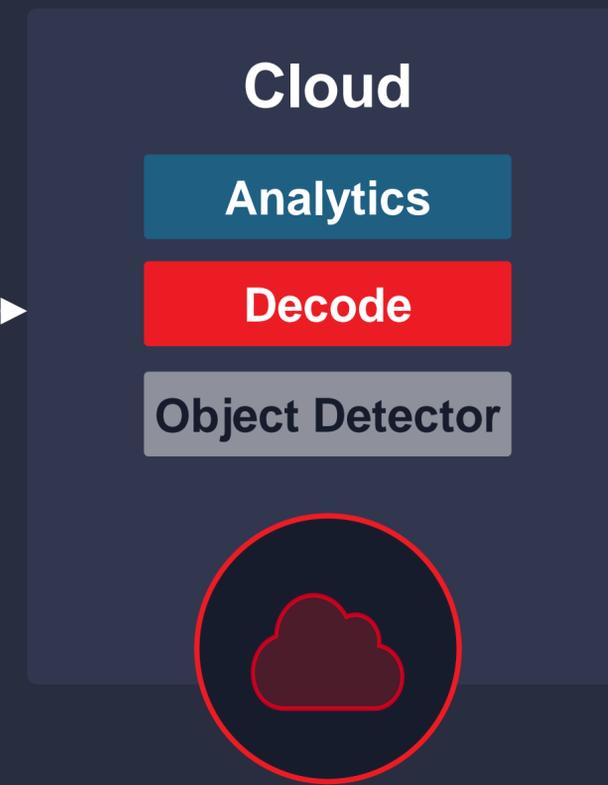
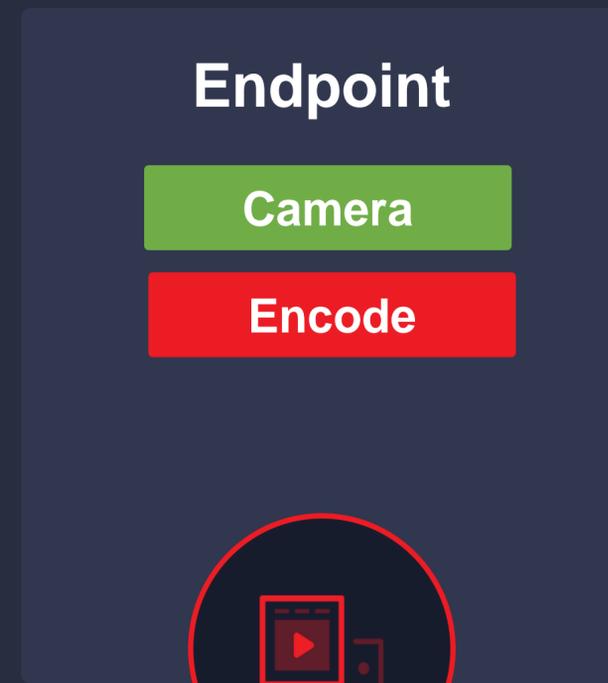
Applications are often split between cloud and edge



Industry Trend: Cloud to Edge



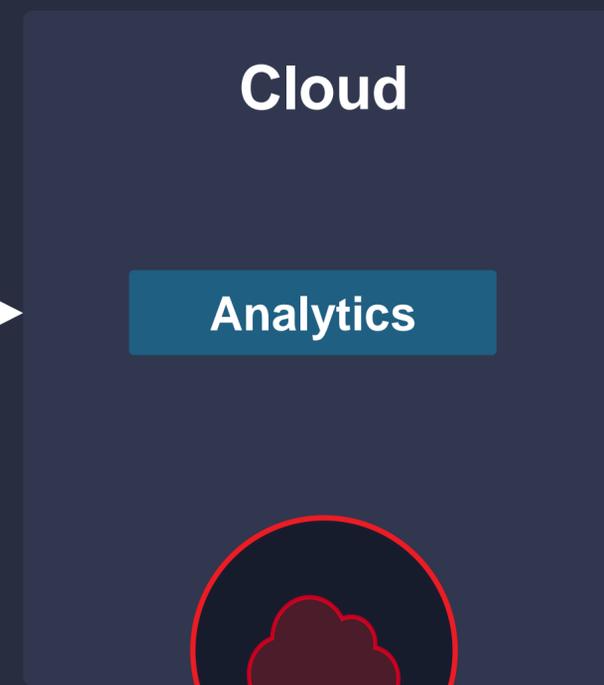
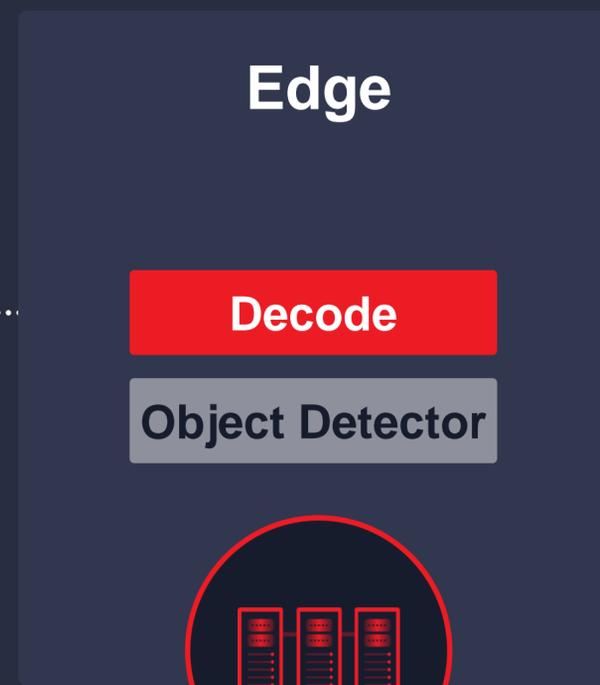
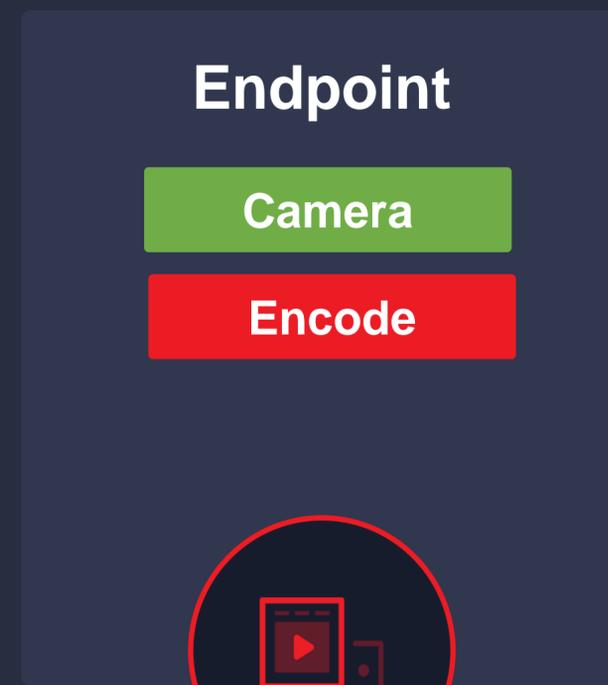
Applications are often split between cloud and edge



Industry Trend: Cloud to Edge



Applications are often split between cloud and edge

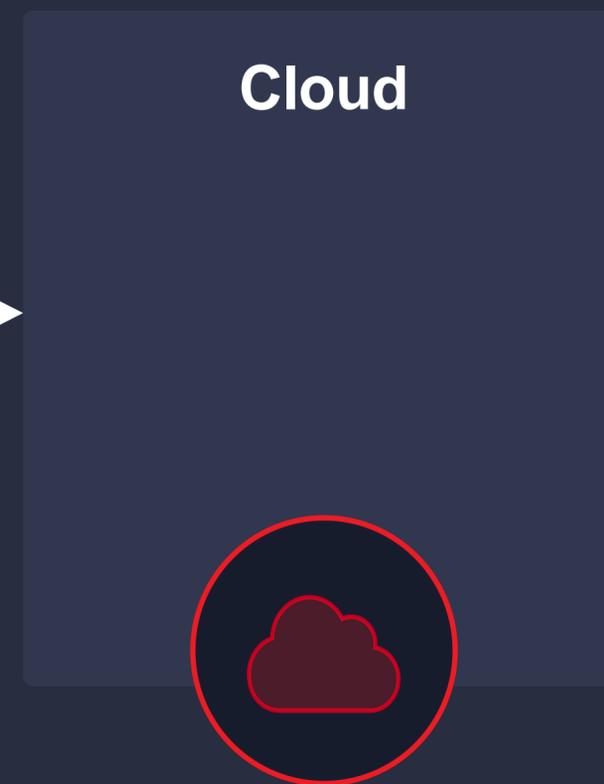
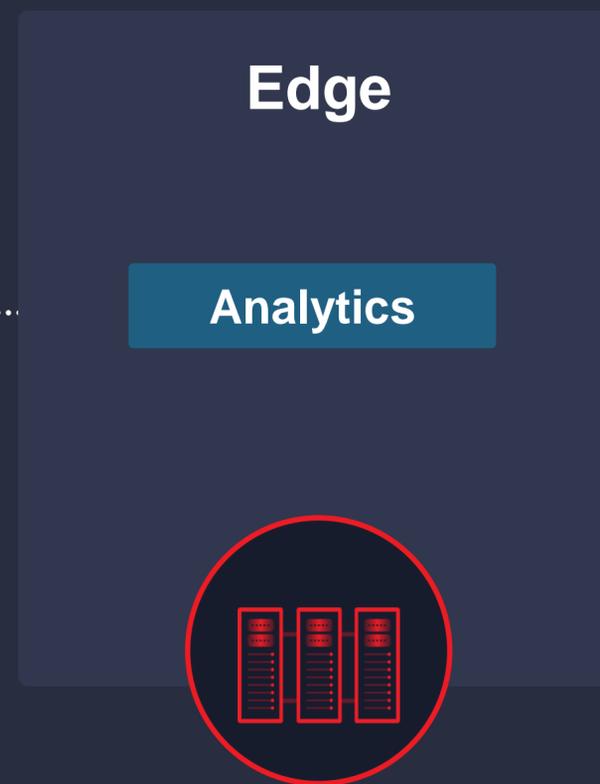
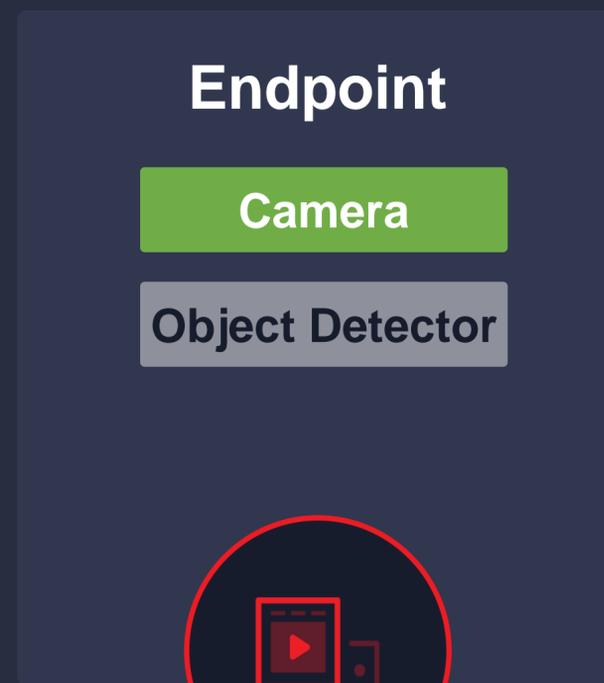


Key Challenge
Need for Retargetability

Industry Trend: Cloud to Edge



Applications are often split between cloud and edge



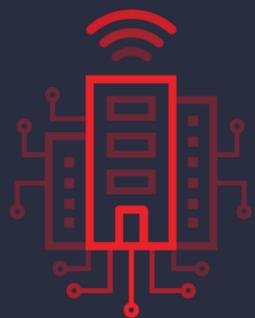
Key Challenge
Need for Retargetability

Industry Trend: AI Proliferation



AI Proliferation

AI is being used in many applications



Smart City



Smart Retail



Autonomous Driving



Security



Genomics



Video Analytics



Healthcare



Finance

Key Challenge

Acceleration and Integration of the Whole Application



**Heterogeneous
Compute**



Cloud to Edge



AI Proliferation



**Heterogeneous
Compute**

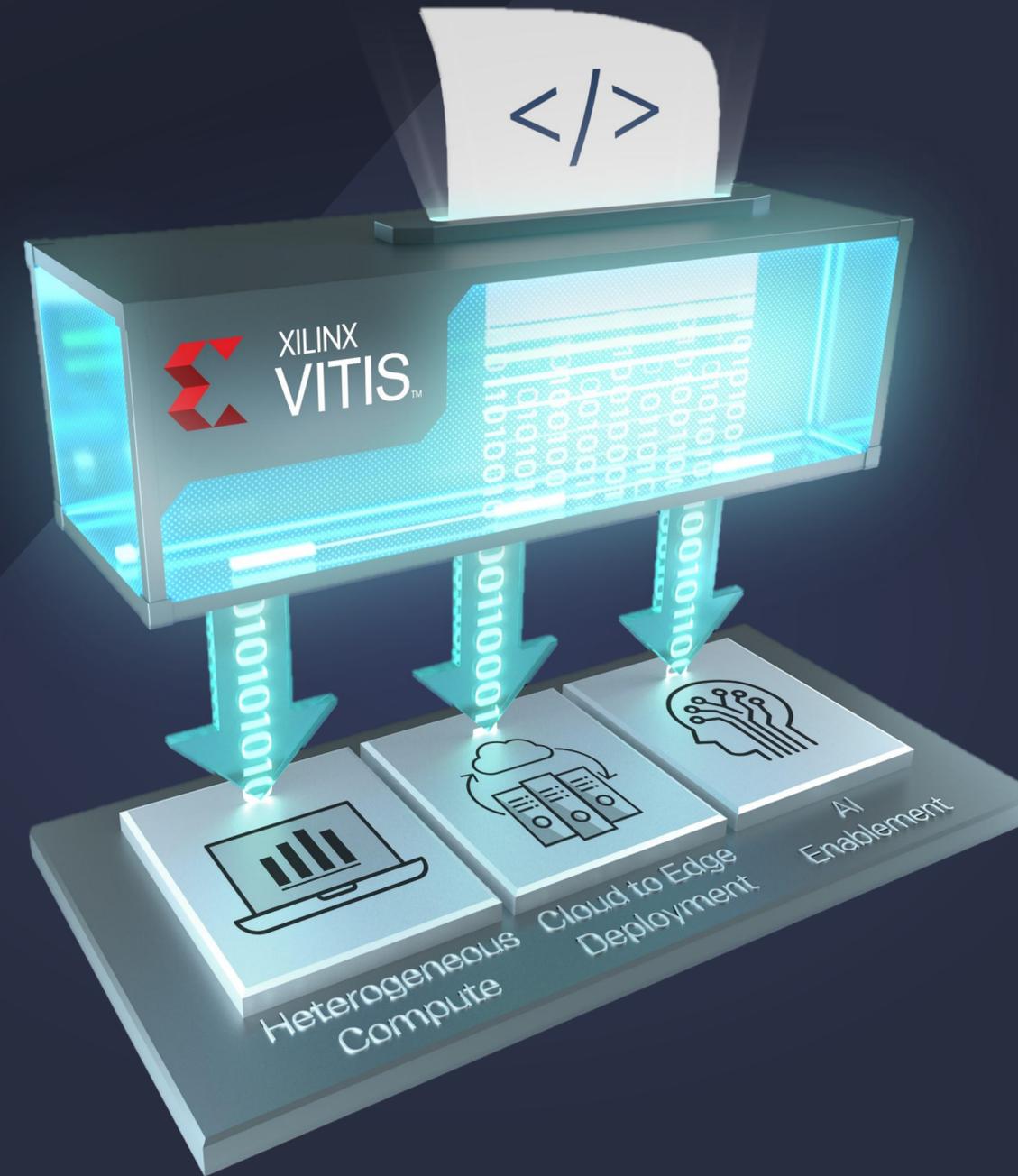


Cloud to Edge



AI Proliferation

Vitis Unified Software Platform



Platform Transformation

Vitis Unified Software Platform
Adaptable & Programmable

#DEVELOPERS



Vivado

2012



OS and Firmware SDK



SDSoC, Embedded



SDAccel, Data Center (FaaS, Alveo)



AI inference Acceleration



Vivado

2019



Enables all Developers to Build and Deploy to All Platforms



Build



Embedded Developers



Enterprise Application Developers



Enterprise Infrastructure Developers



Data & AI Scientists



XILINX
VITIS™



Deploy



Zynq



Ultrascale

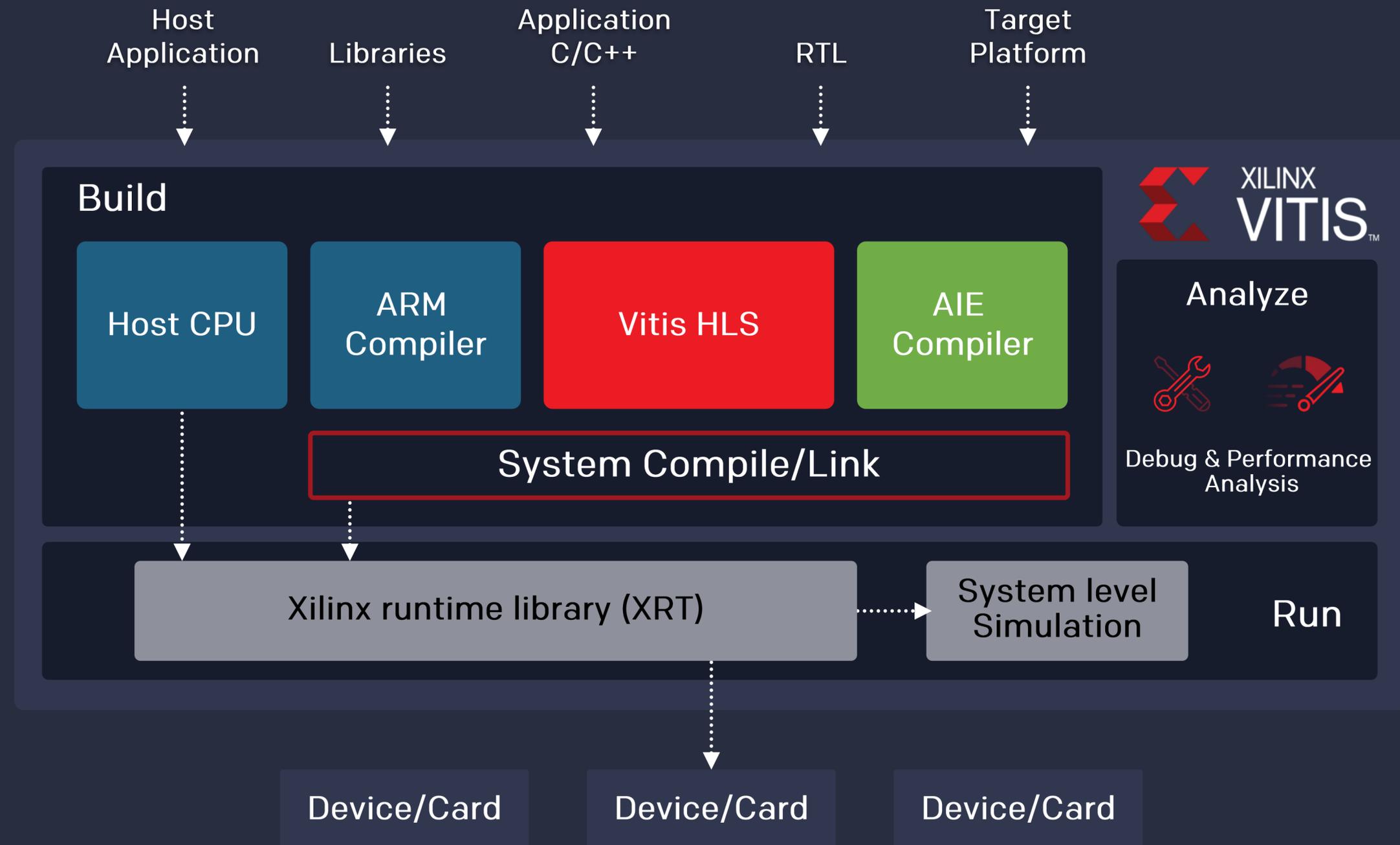


Alveo



Data Center Rack

Build Comprehensive Development Tool Suite



Build

Comprehensive Development Tool Suite



400+ functions across multiple libraries
Open-Source, performance-optimized out-of-the-box acceleration

Domain-Specific Libraries



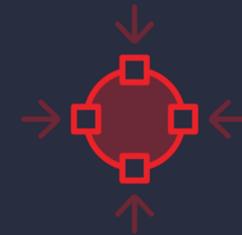
Vision &
Image



Finance



Data Analytics &
Database

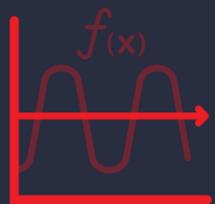


Data Compression

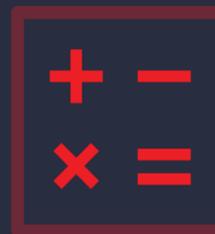


Data Security

Common Libraries



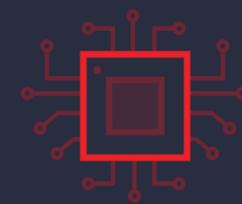
Math



Linear Algebra



Statistics



DSP



Data Management

Deploy
**Embedded
Deployment**

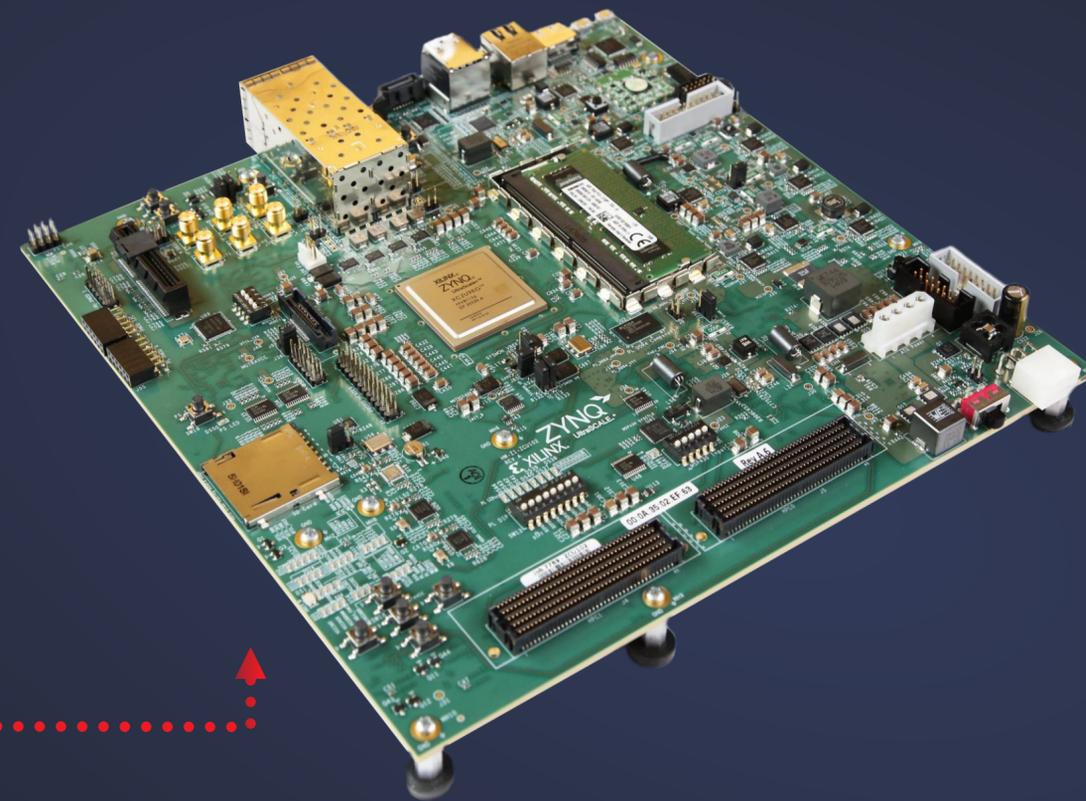
Application



Executable



Runtime



Deploy
**Single Server
Deployment**

Application



Executable



Runtime



Deploy Scale Out Deployment

Executable

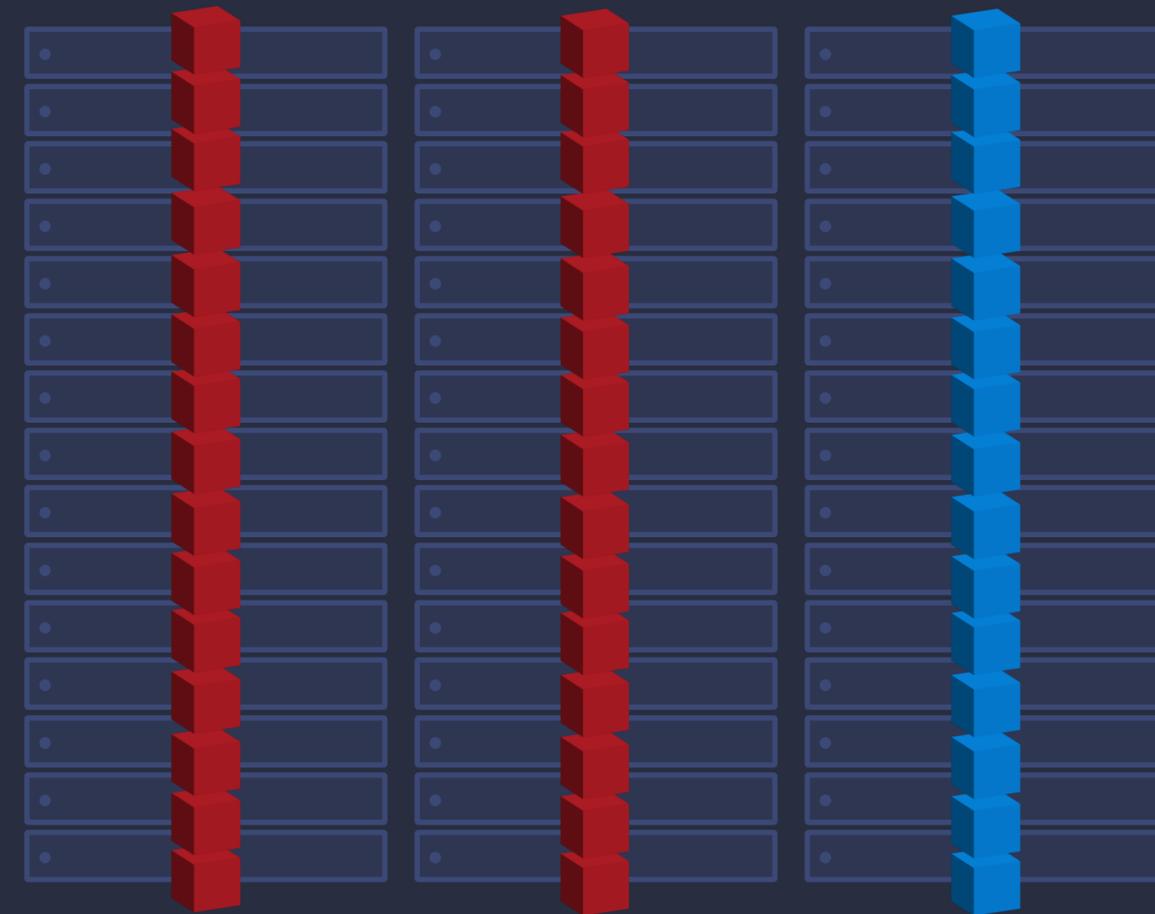
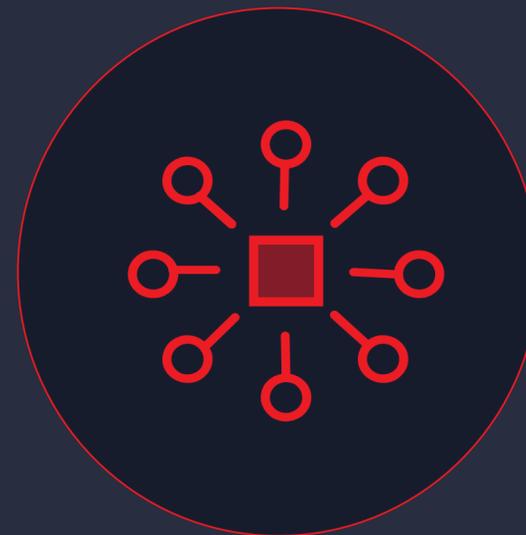


Runtime

Xilinx Docker
Registry



Scale Out



Vitis AI: From TensorFlow to Implementation in Minutes



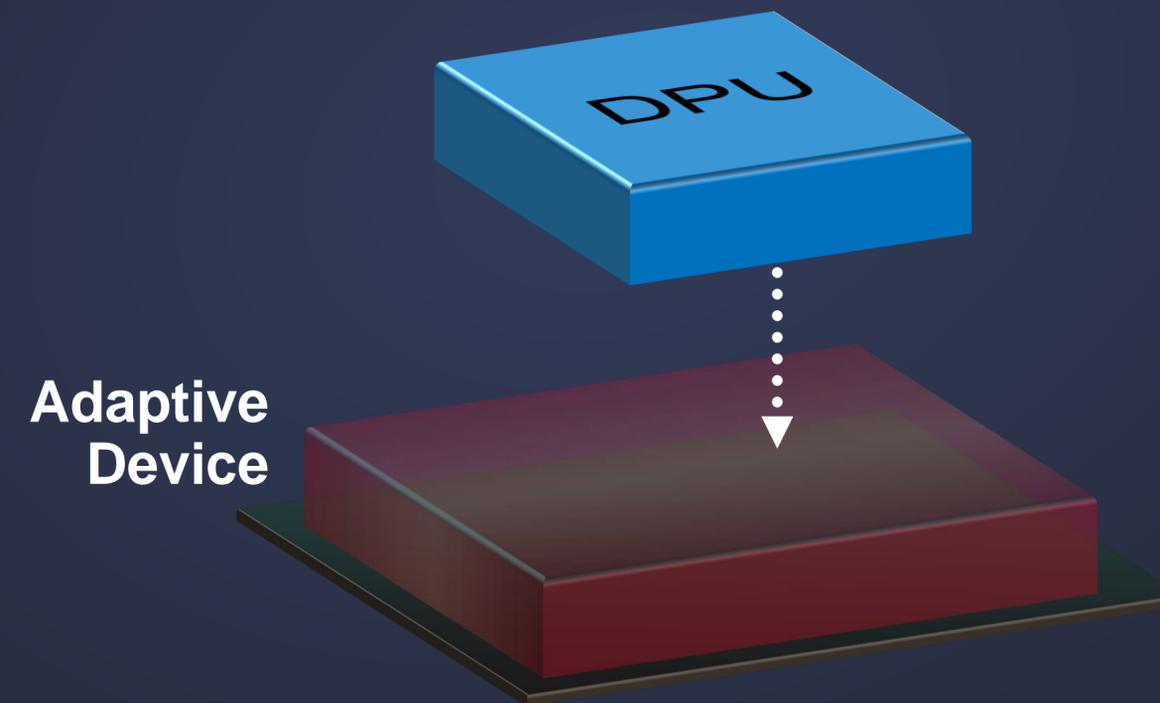
DNN Processing Unit (DPU)



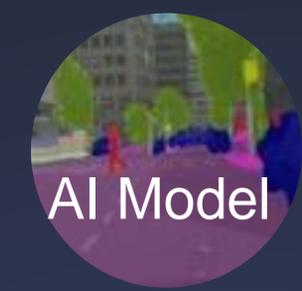
Direct Framework Compilation



Minutes of Compile Times



Vitis AI: From TensorFlow to Implementation in Minutes



DNN Processing Unit (DPU)

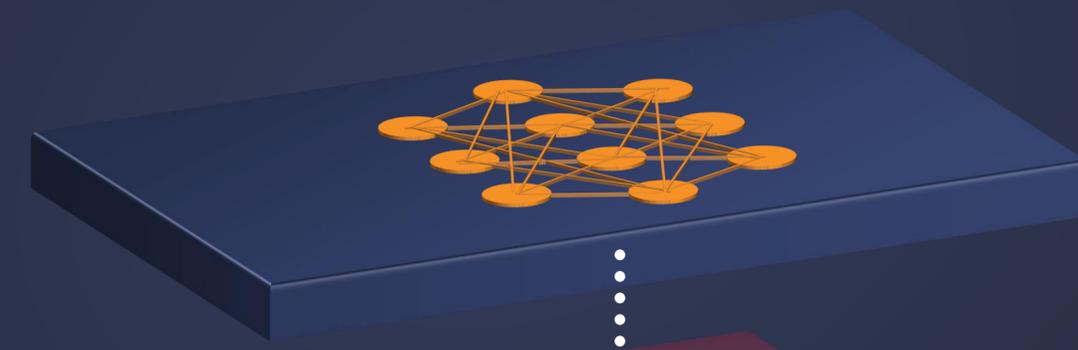


Direct Framework Compilation

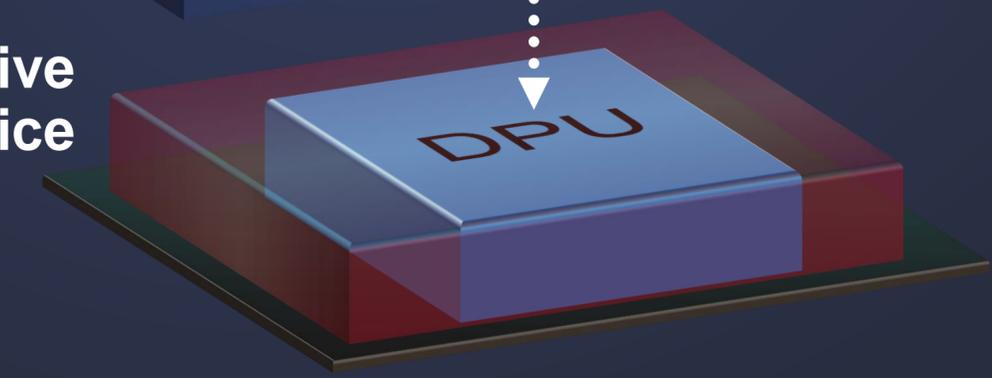


Minutes of Compile Times

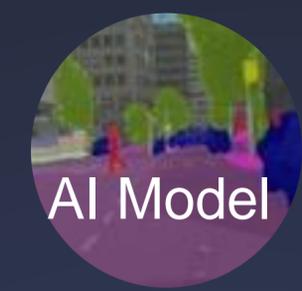
Vitis AI



Adaptive Device



Vitis AI: From TensorFlow to Implementation in Minutes



DNN Processing Unit (DPU)



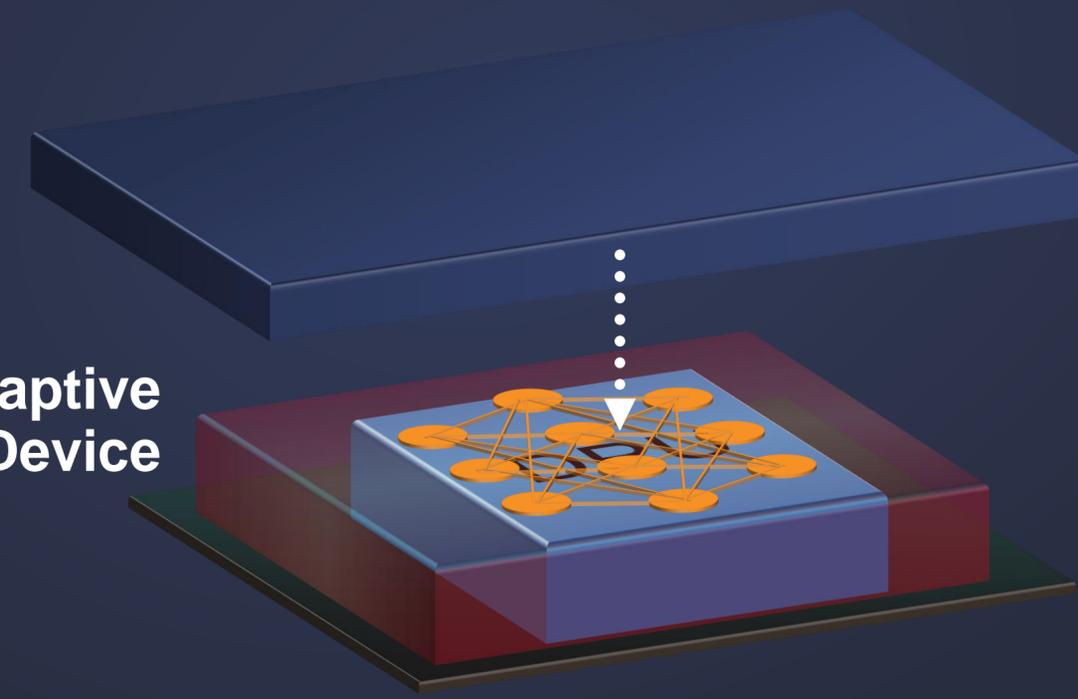
Direct Framework Compilation



Minutes of Compile Times

Vitis AI

Adaptive Device



Enabling AI

Frameworks

TensorFlow

Caffe

PyTorch

Vitis AI Models



Vitis AI Development Kit

AI Optimizer

AI Quantizer

AI Compiler

AI Profiler

Vitis drivers & runtime (XRT)

DSA

CNN DPU

LSTM DPU

MLP DPU

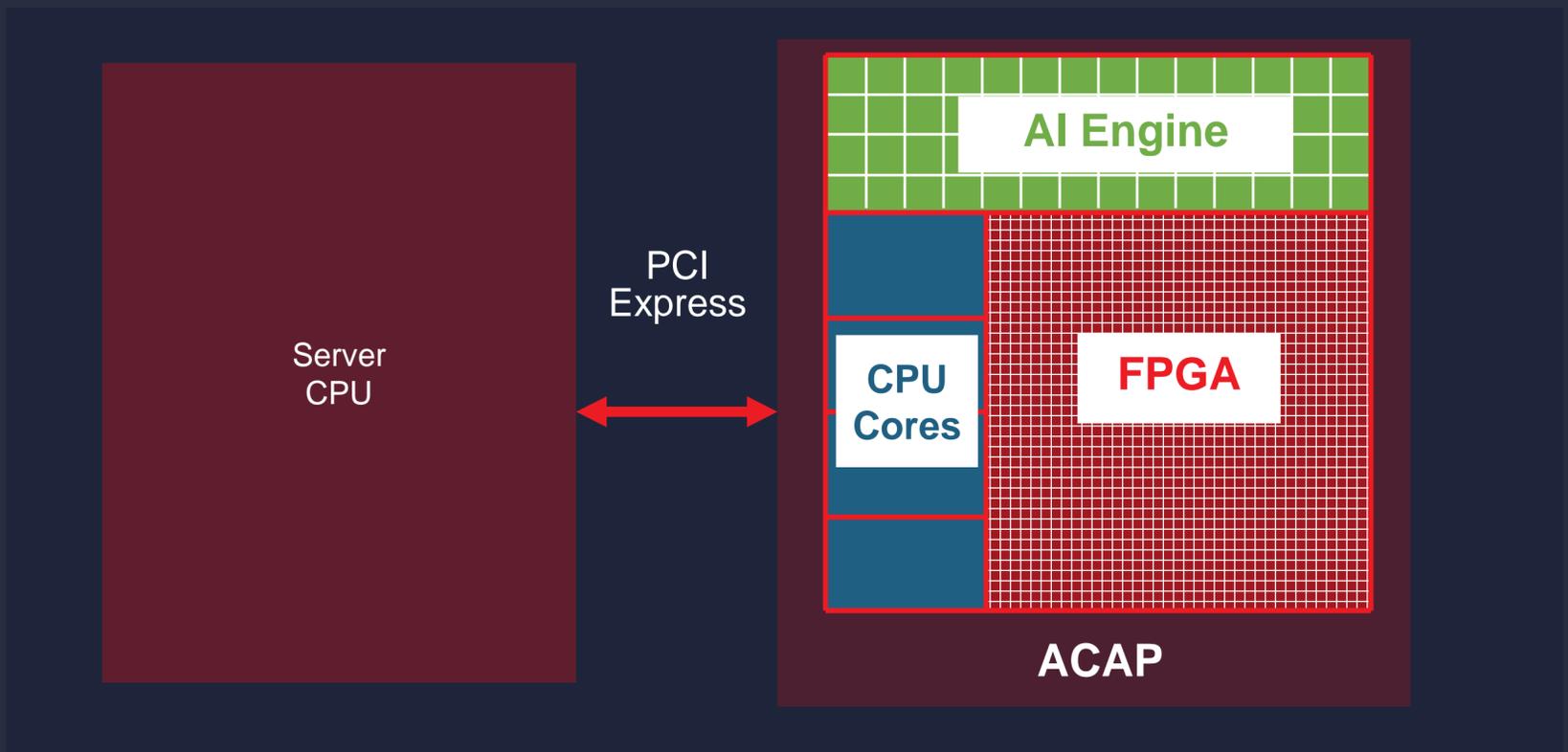
30+ pretrained, optimized reference models

Performance improvement up to 10-20x

Tensor based ISA for true software programmability



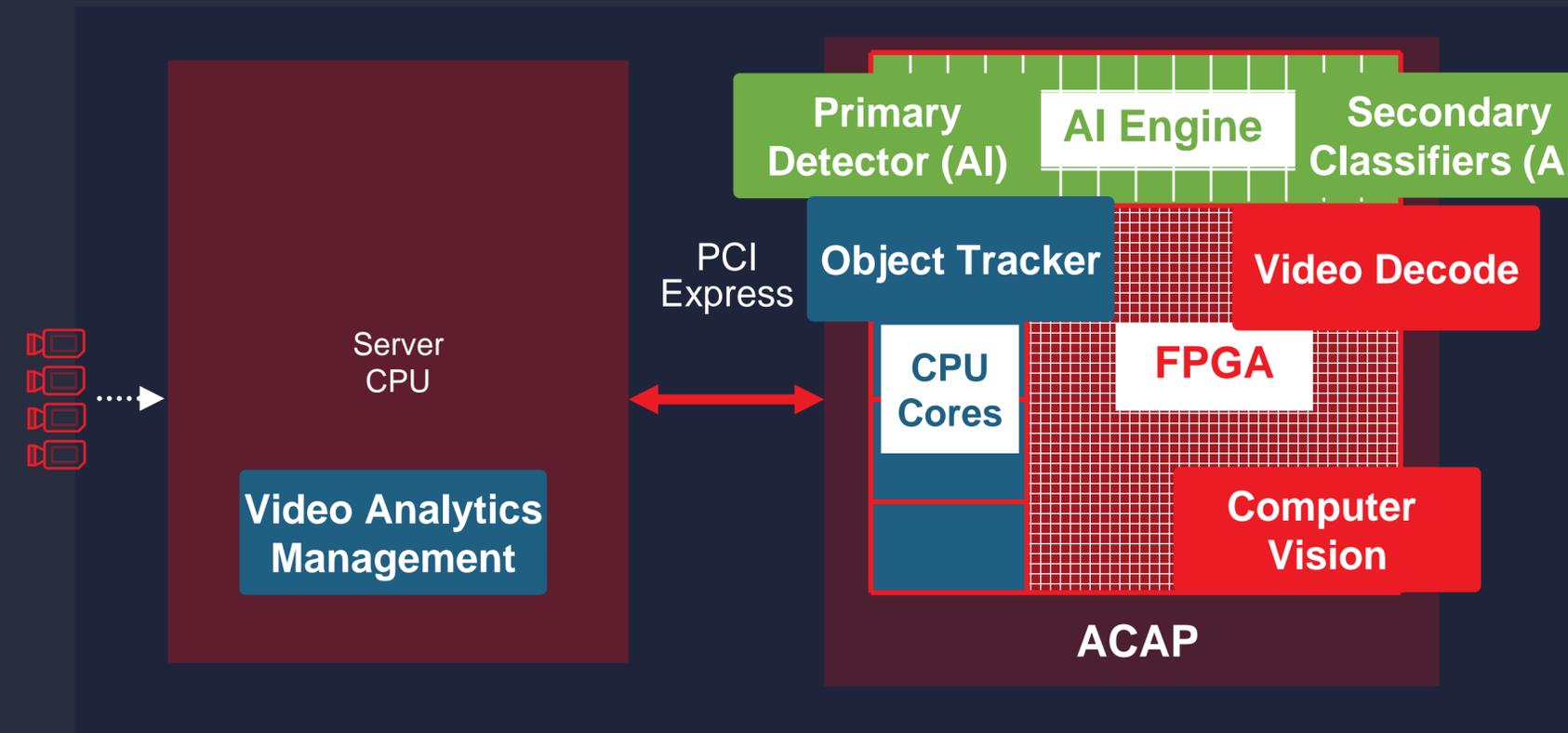
Vitis Enables Whole App Acceleration



Smart City Example

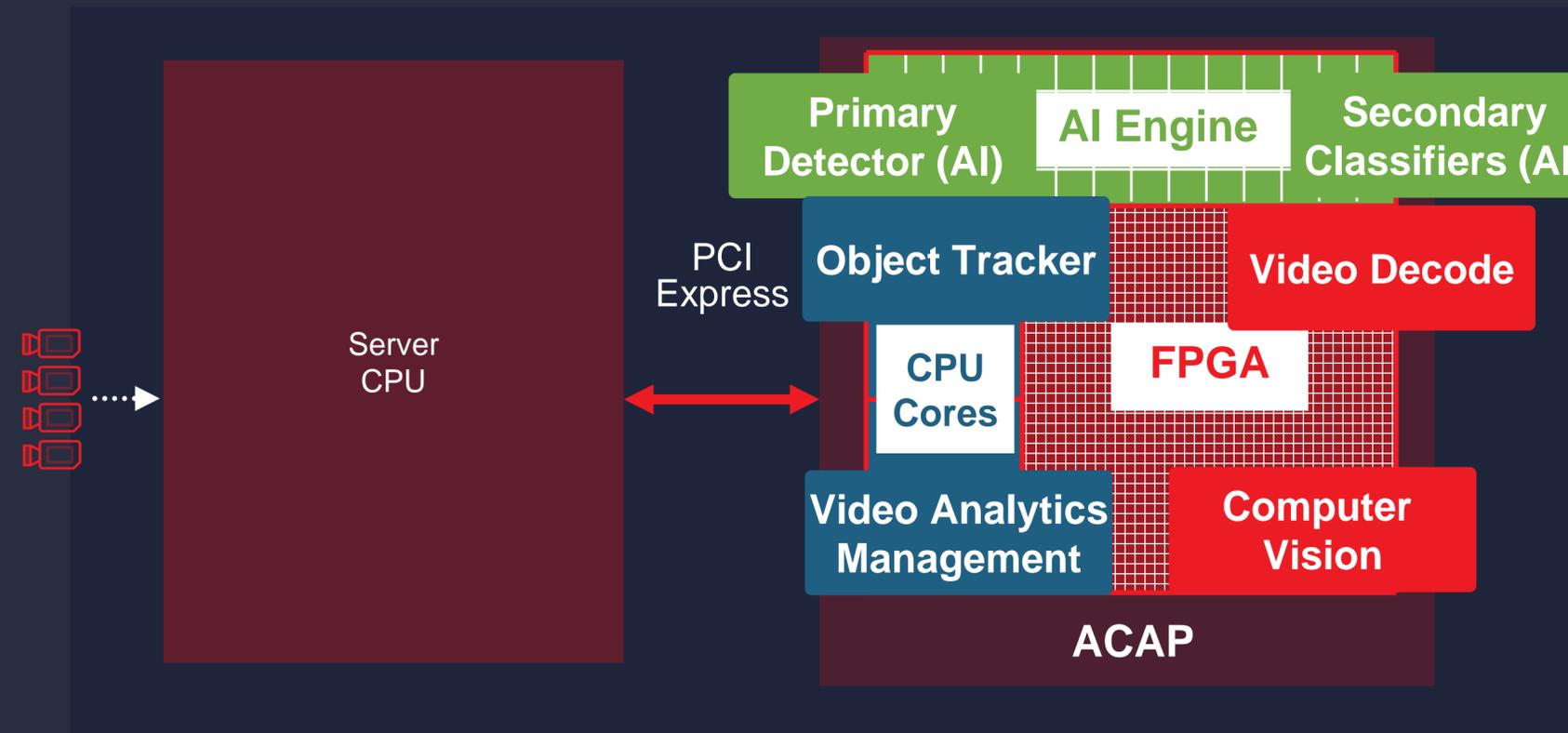


Vitis Enables Whole App Acceleration



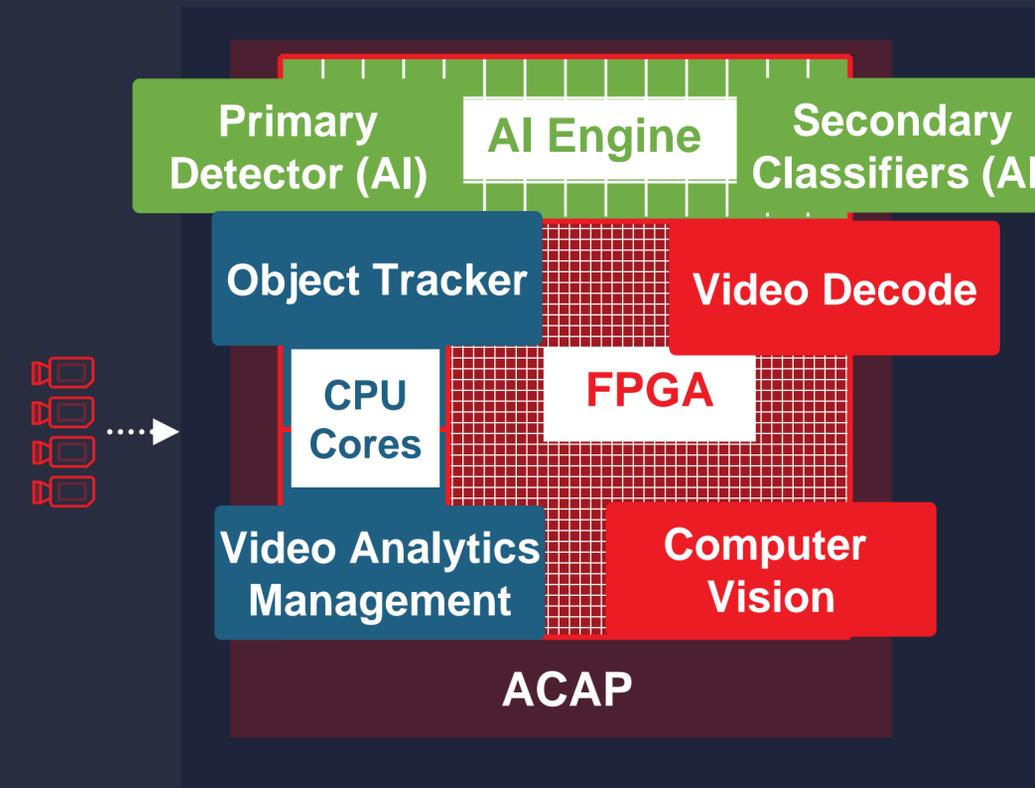
**Smart City
Example**

Vitis Enables Whole App Acceleration



Smart City
Example

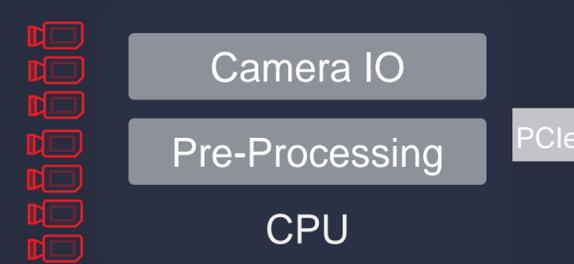
Vitis Enables Whole App Acceleration



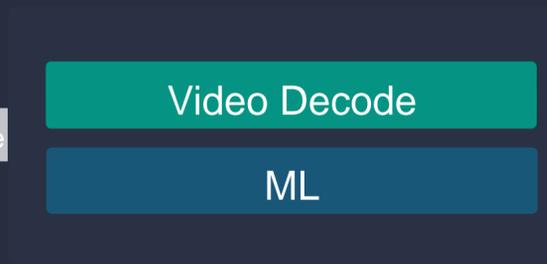
Smart City
Example

Impact of Whole Application Acceleration

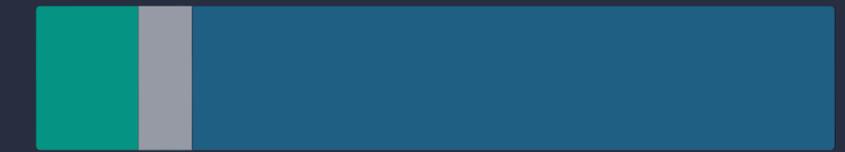
7 Channels 1080p



12nm GPU



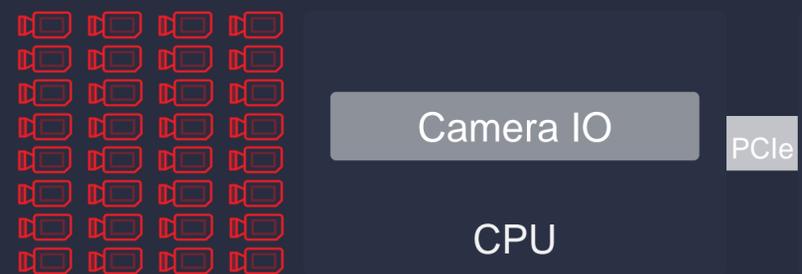
Decode Detect + Classify



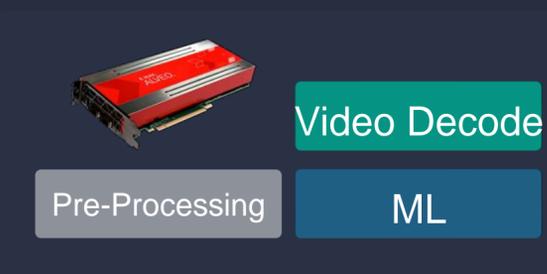
➤ 4.5x throughput

➤ 1/6 latency

32 Channels 1080p



Versal ACAP



Decode Detect + Classify



Launching *developer.xilinx.com*

Beta Site Now Available

Tutorials, Articles & Projects

Accessible from a Single Location

Learn Directly from Vitis Experts

Model	FPS
FPN	64.29
Unet-Full	14.27
Unet-Lite	36.25

Using data gathered throughout this tutorial, we can compare the performance of the ZCU102 vs. the GTX1080ti graphics from section 4.1. Albeit, this isn't a fair comparison for two reasons:

1. We are comparing an embedded ~20W device with a 225W GPU
2. The ZCU102 execution time includes reading/preparing the input and displaying the output whereas the GPU measures inference time of the models

That said, this still provides some data points which are useful to garner further understanding. The following chart shows measured on the ZCU102 vs. the GTX1080ti.

Model	ZCU102 (FPN) Display@15	Floating Point GTX1080ti
Enet	~60	~35
ESPNet	~40	~95
FPN	~65	~65
Unet-Full	~15	~30
Unet-Lite	~35	~70

What is perhaps a bit more useful than comparing raw FPS, however, is to compare FPS/W (performance/Watt) as this is performance is achievable for a certain power cost. Bear in mind, this is still not a fair comparison due to reason 2, but the little more in this light. In reality the advantage is even more pronounced if only the DPU throughput is considered.

In order to perform this comparison, ~20W was measured on the ZCU102 board during forward inference, and the nvidia-s during forward inference of each of the models as part of section 4.1. The comparison between the two can be seen in the

YOLC: Input Image → Color Conv → Resize → Scale → To CNN

GoogleNet: Input Image → Resize → Mean Subtraction → To CNN

Sep 24, 2019 3:11:58 PM

Integrating optimized RTL Kernels into Accelerated Applications using Vitis

Ted Ennis

Sep 24, 2019 2:57:45 PM

Profiling and Accelerating C++ Applications and Algorithms

Christophe Charpentier

$$\text{similarity}(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}}$$

Sep 24, 2019 1:13:57 PM

Cosine Similarity Using Xilinx Alveo

Alvin Clark, Kumar Deepak, Liang Ma

Sep 23, 2019 3:57:27 PM

Task-level parallelism and pipelining in HLS (fork-join and beyond)

Frédéric Rivoallon

before, we'll configure the library (in the hardware, via templates in our hardware source aware algorithm is not equivalent to listing 3.21 in standard OpenCV.

Listing 3.21: Example 8: Bilateral Resize an

Sep 23, 2019 3:31:04 PM

Get Moving with Alveo: Example 8 Pipelining Operations with OpenCV

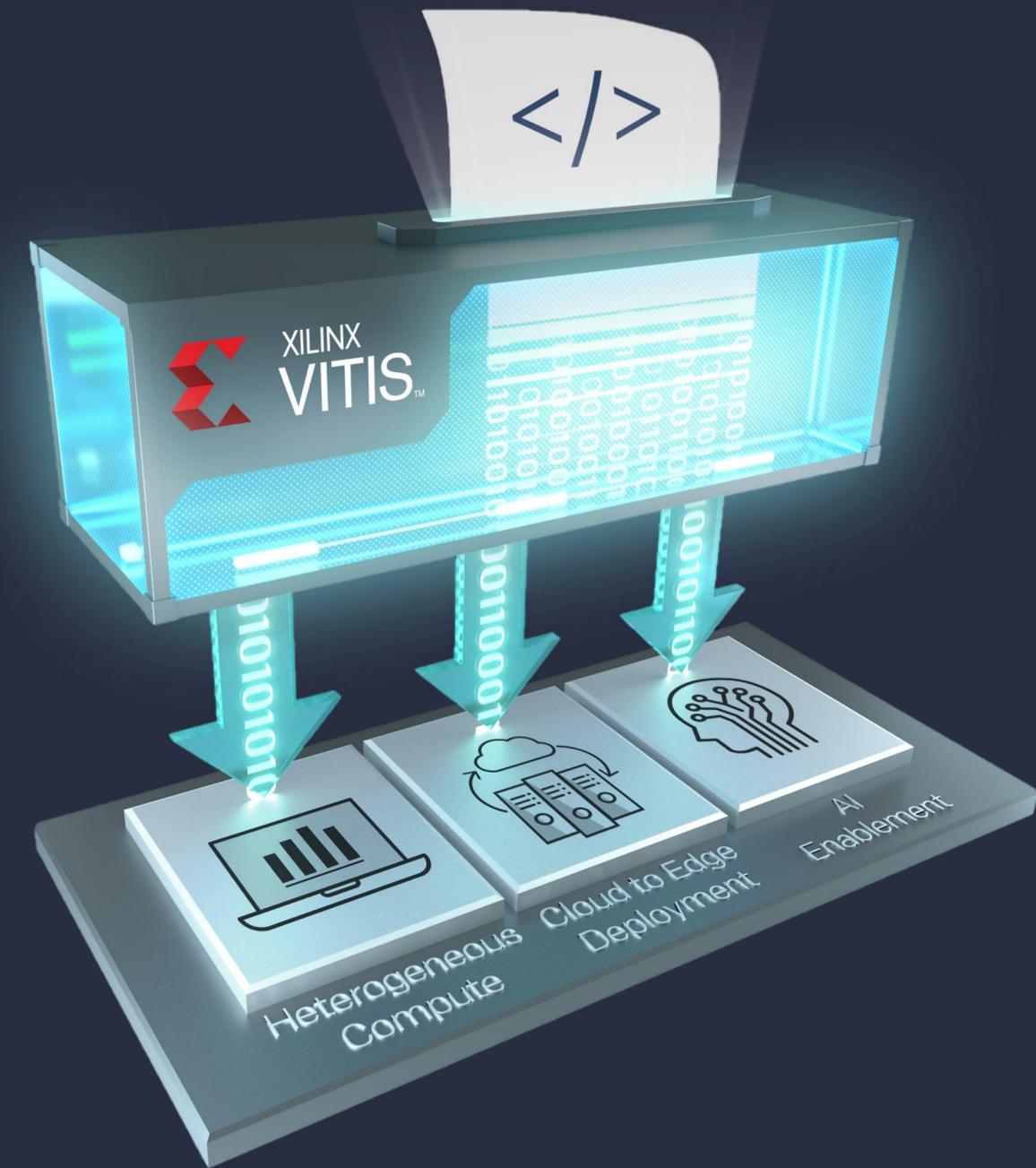
Rob Armstrong

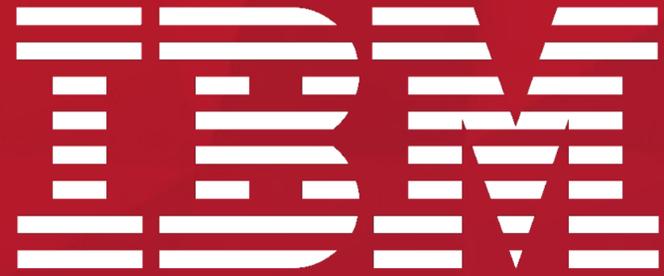
Sep 23, 2019 3:30:44 PM

Get Moving with Alveo: Example 7 Image Resizing with OpenCV

Rob Armstrong

30+ expert articles & projects (and growing)





Sumit Gupta

Vice President of Product
AI/ML & HPC





Tom Eby

SVP & GM, Compute & Networking Business Unit,
Micron Technology



Development Platforms for ALL Developers

- Unified
- Open Source Libraries

Free!

