

altera<sup>TM</sup>

An Intel Company

# FPGA AI Suite

Jean-Michel Vuillamy – June 27, 2024

# FPGA AI Suite Enables Custom AI Hardware

## Faster Time to Market

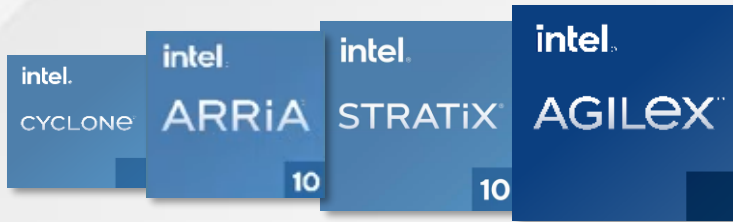
- Create FPGA inference accelerator from TensorFlow/PyTorch/Keras etc.,
- Adapt quickly to new Networks, Precisions

## Microseconds Matter: Examples →

- Medical Image Recognition
- Industrial Inspection
- Defense: Radar classification
- Speech Recognition
- Large physics experiments

## Delivering Hardware Customization with Integrated AI

- Low-end to high-end Altera FPGAs
- Low latency Inline Performance
- Security



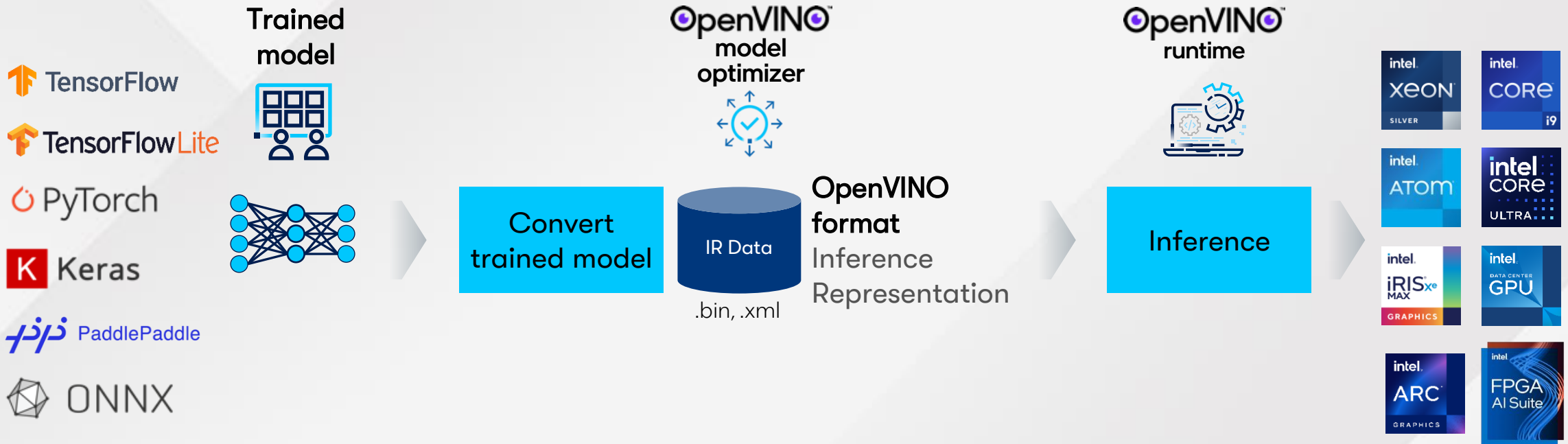
Enabling low latency AI in a wide range of embedded, edge, and data center applications

# OpenVINO™ Toolkit: Deploy on Diverse Hardware

1 | MODEL

2 | OPTIMIZE

3 | DEPLOY



Download:

<https://software.intel.com/content/www/us/en/develop/tools/openvino-toolkit.html>  
<https://docs.openvino toolkit.org/latest/index.html>

**Note:**

- The OpenVINO™ toolkit plugin is provided by the Intel® FPGA AI Suite. Users need to download the OpenVINO toolkit software and the Intel FPGA AI Suite OpenVINO plugin.
- The Intel FPGA AI Suite supports OpenVINO LTS releases only.
- The FPGA plugin is no longer available on the OpenVINO toolkit website and is provided as part of the Intel® FPGA AI Suite.

# FPGA AI Suite for Intel® FPGAs

## ❑ Customizable

- Customizable soft IP designed to meet stringent performance, accuracy, low energy inference, and area requirements [specify FPGA resources - # of Logic Elements, DSP, Memory]

## ❑ Flexible system topologies

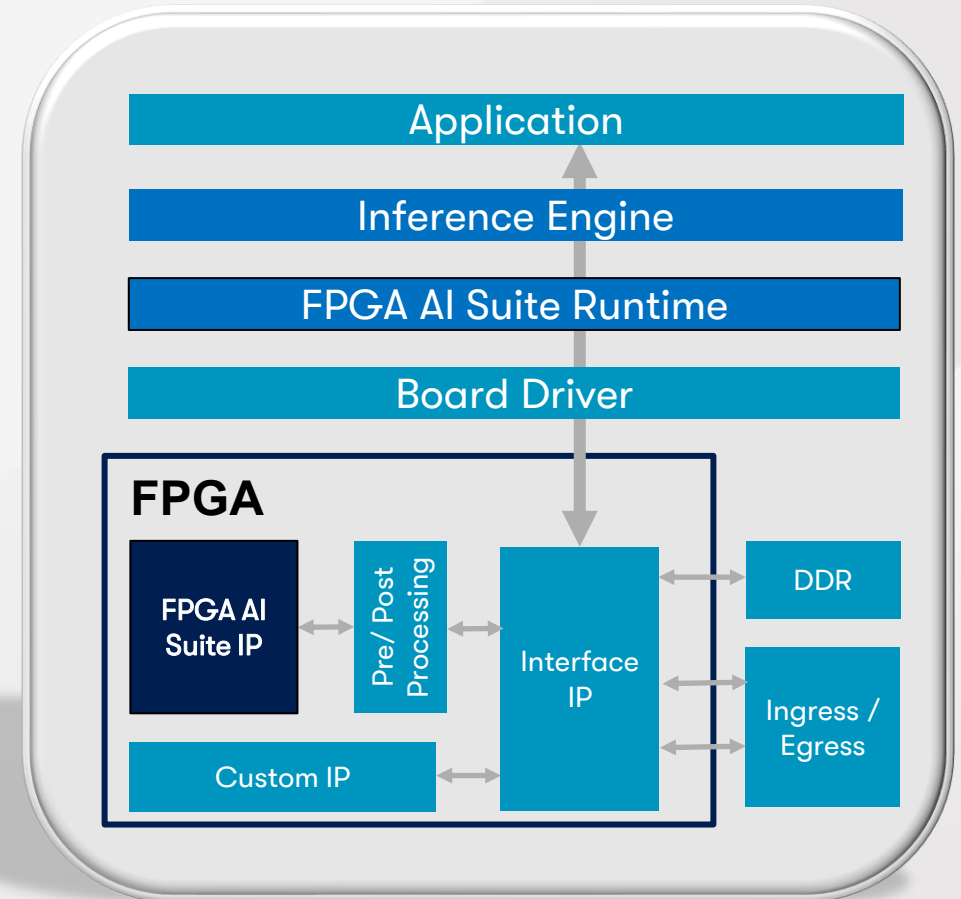
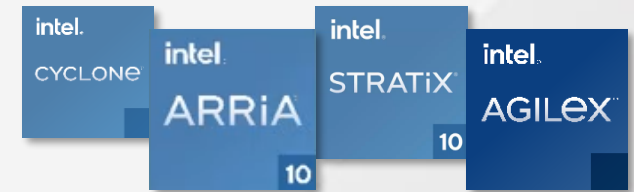
- Flexible deployment across multiple FPGA families
  - (Agilex™ / Arria® 10 / Cyclone® 10 GX / Stratix® 10)
- Supports multiple host processors
  - Intel® Core™, Intel® Xeon®, and ARM processor

## ❑ Open frameworks

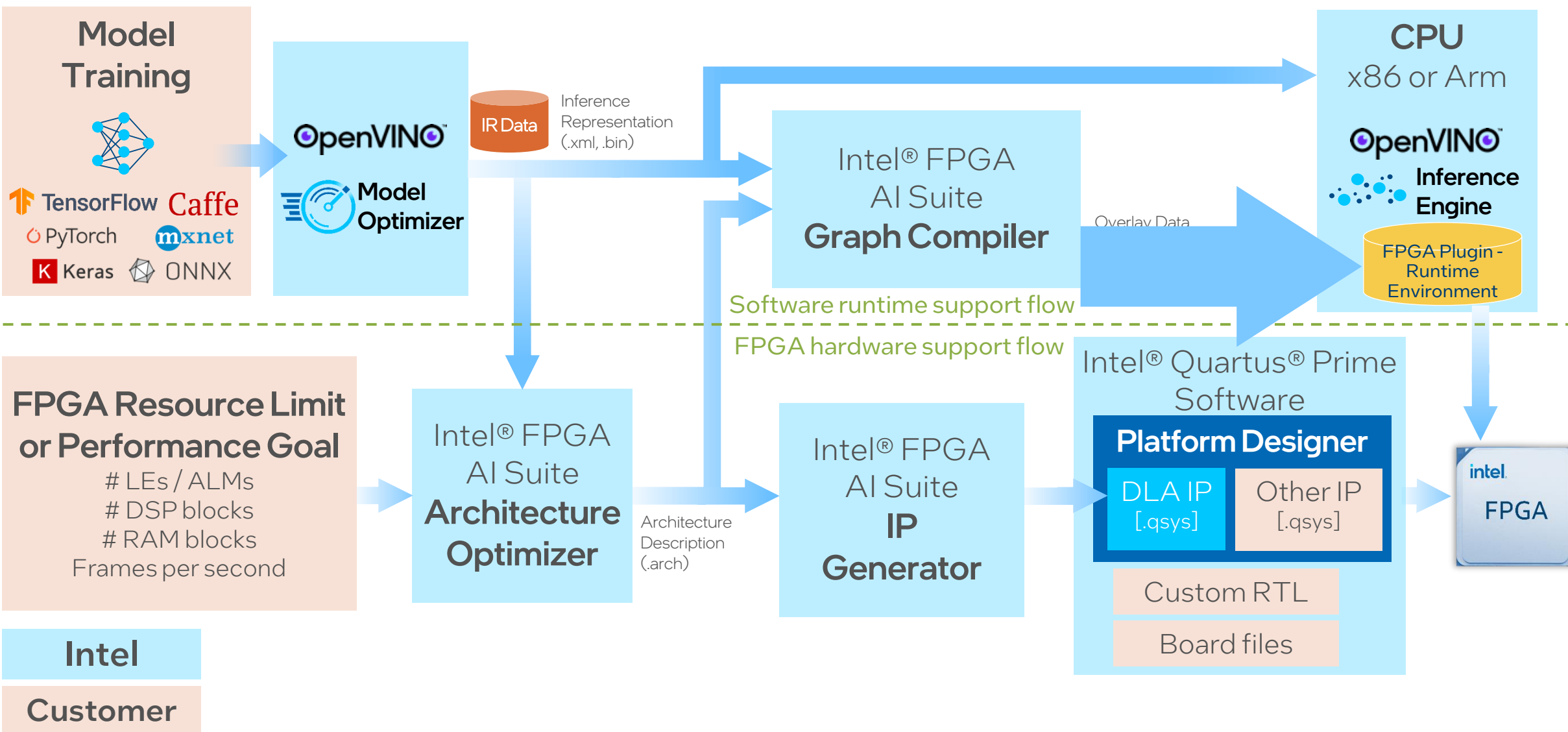
- Optimized for major deep learning frameworks (TensorFlow, PyTorch etc.,) to RTL via Intel's OpenVINO™ toolkit

## ❑ Tested Networks, Layers and Activation Functions

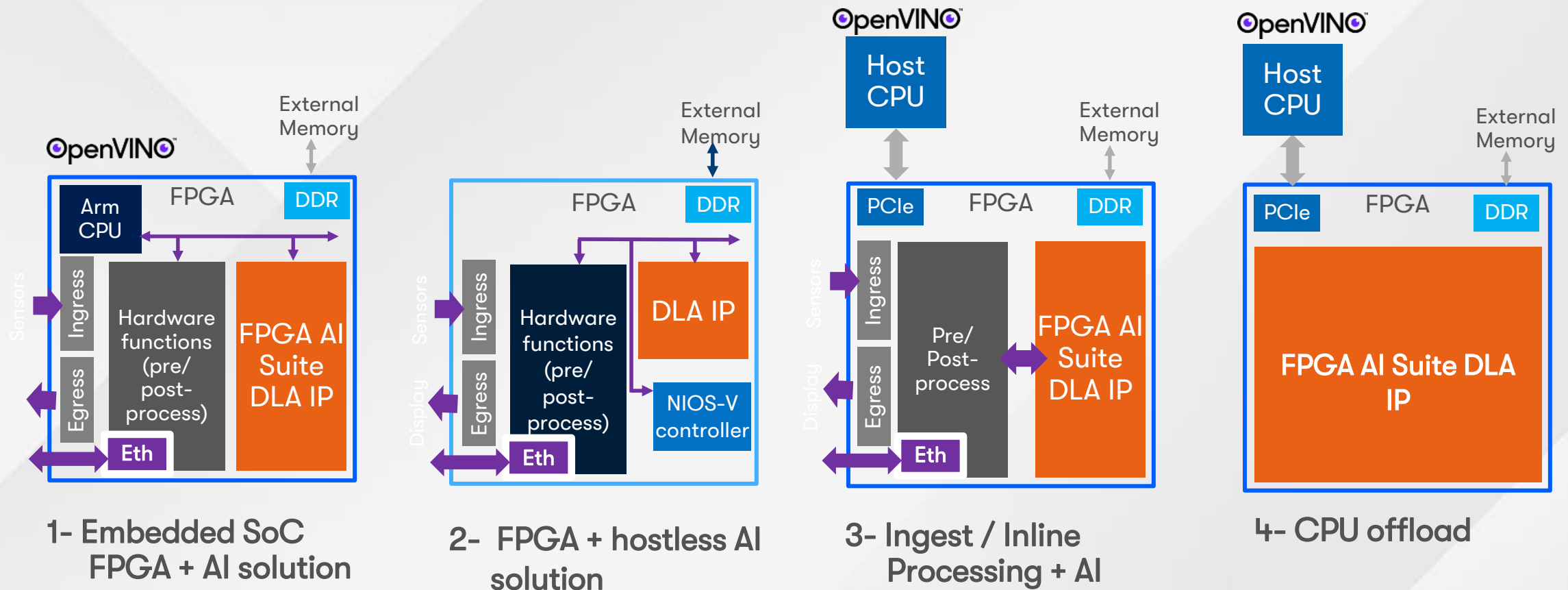
- ResNet-50, MobileNet v1/v2/v3, YOLO v3, TinyYOLO v3, VGG-16/19, Unet, SqueezeNet v1.1, Inflated 3D (I3D), Multilayer perceptrons (MLPs)
- 2D Conv, 3D Conv, Depthwise, Scale-Shift, Deconv, ReLU, pReLU, Leaky ReLU, Clamp, Round\_Clamp, H-sigmoid, H-swish, Max Pool, Avg Pool, SoftMax, EltWise Mult, BatchNorm, Fully Connected, Clamp, pReLU, SoftMax



# Intel® FPGA AI Suite Development Flow



# Intel® FPGA AI Suite Enables “+AI” Architectures



Above examples are generic and meant to illustrate the flexibility of FPGA AI Suite. Applications may utilize a soft host CPU (such as RISC-V based Nios® V processor in the FPGA fabric) and/or other in-line, streaming sources such as HDMI.

# Resnet-50 & MobileNetv2

ResNet-50		MobileNet v2		ALM	M20K	DSP	Batch Size
Arria 10	Agilex 7	Arria 10	Agilex 7				
Frames / Second		Frames / Second					
29	38	48	61	12K	167	31	1
34	44	56	71	14K	233	55	1
41	54	68	87	17K	258	87	1
53	70	88	112	22K	380	167	1
111	89	111	143	28K	524	295	1
127	168	211	270	53K	890	602	1
228	301	378	484	95K	1780	1166	2

## Configuration

- The Intel® Arria® 10 FPGA performance numbers presented herein are measured using an Intel Arria 10 FPGA Programmable Acceleration Card (PAC), incorporating an Intel Arria 10 FPGA 1150 (speed grade 2) with DLA IP @FP11 data type.
- The Intel® Agilex™ FPGA performance numbers presented herein are high confidence estimates based on a smaller set of results measured on an Intel Agilex FPGA AGF014 1.4M (speed grade 2) with DLA IP @FP11 data type. Actual results may vary based on device utilization.
- Dataset: ImageNet ILSVR2012 @ 224x224
- The host is an Intel® Xeon® processor E5-1650 v3 @ 3.5 GHz w/ 132 GB RAM.

**Last Updated:** 10/13/2021; Version 2021.2  
**Note:** These metrics are continuously under improvement

Resources in Arria 10 1150			Resources in Agilex AGF014		
ALM	M20K	DSP	ALM	M20K	DSP
427200	2713	1518	487,200	7110	4510

# Newer Families Further Expand FPGA AI

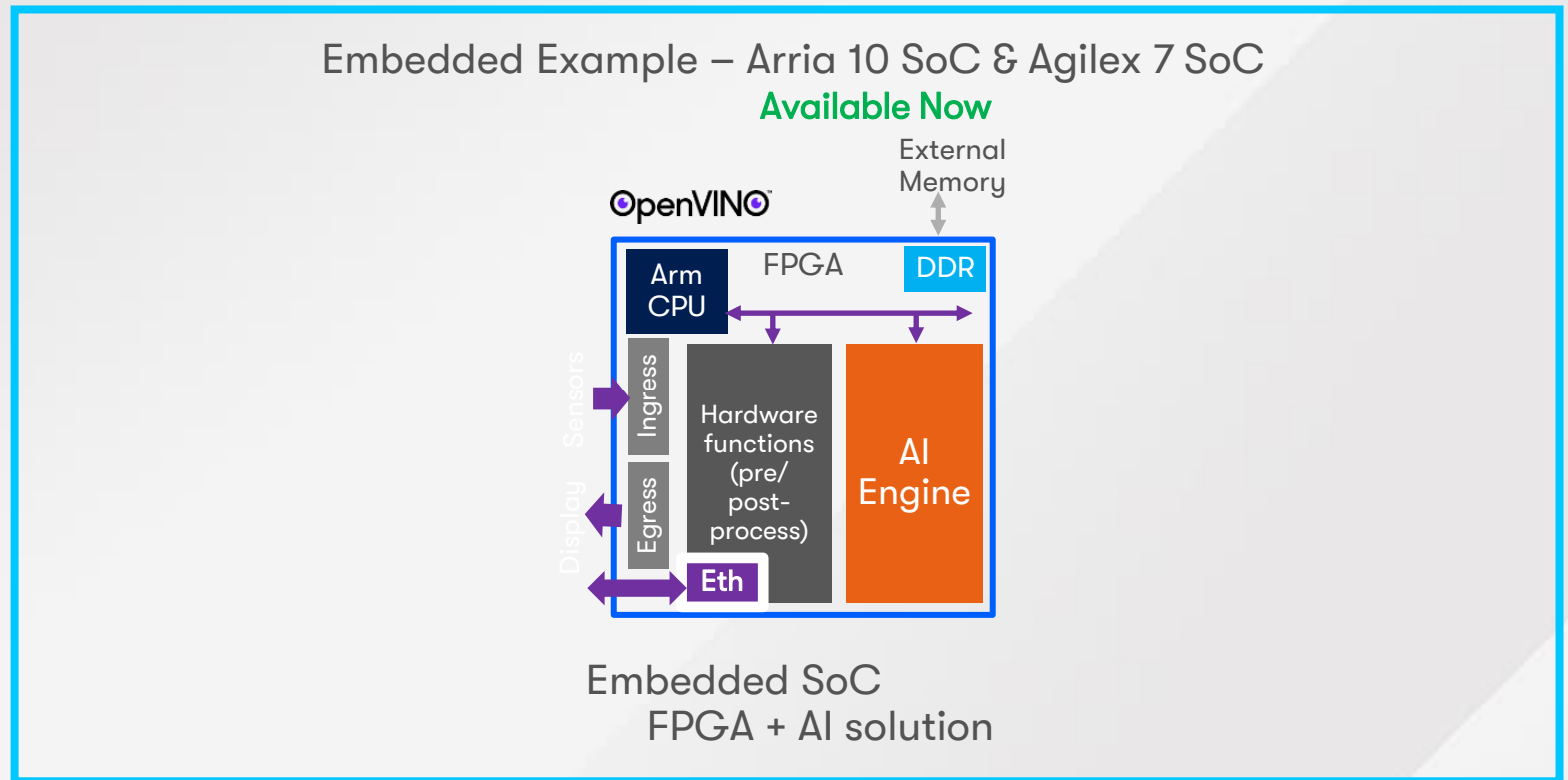
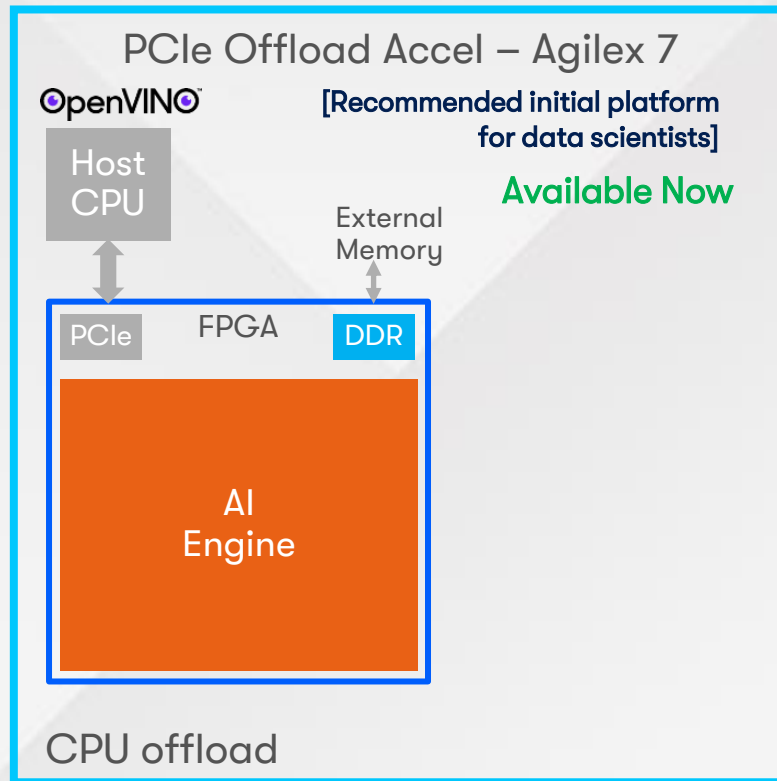
- Order of Magnitude Improvement in AI and DSP compute density

## Intel® Agilex™ 5 FPGA and SoC FPGA

- Offers 2 TOPS to 56 TOPS for INT8 across E-Series and D-Series families with Enhanced DSP with AI Tensor Blocks
  - Multi-core Arm processors of Dual-core A55 @ 1.5 GHz and Dual-core A76 @ 1.8GHz
  - Small package sizes: 15 mm x 15 mm and 23 mm x 23 mm

Applications	Multiplier	Capabilities per DSP Block		Improvement*
		Earlier Intel Agilex Devices	Enhanced DSP with AI Tensor Block*	
AI, Signal Processing	INT8	4 OPS	20 OPS	5X
	INT9	4 Multipliers	6 Multipliers	50%
Signal Processing	16-bit Complex Multiplier	Needs 2 DSP Blocks	1 DSP Block	2X

# Intel® FPGA AI Suite Design Examples



Note: Boards are not included in the FPGA AI Suite.

Note: Design examples are development platforms to get started and adapt to custom platforms or partner boards.

# Intel® FPGA AI Suite Design Examples

PCIe Offload Accel – Agilex 7  
[Recommended initial platform for data scientists]  
**Available Now**



**Terasic\* DE10\*-Agilex Development Board  
with Intel® Xeon®/Intel® Core™ CPU**

**Hardware required:**

- x86 Host system with PCIe slot (Gen 3, x16)
- [Terasic DE10-Agilex Board](#)

Note: Boards are not included in the FPGA AI Suite.

Note: Design examples are development platforms to get started and adapt to custom platforms or partner boards.

Embedded Example – Arria 10 SoC  
**Available Now**

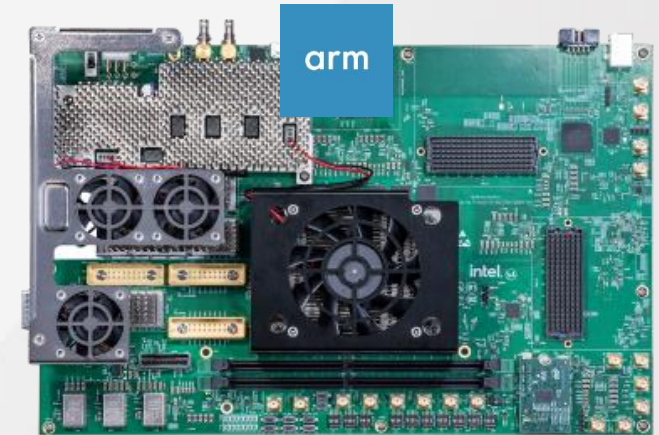


**Intel® Arria 10 SoC Development Kit  
with embedded Arm CPU as host**

**Hardware required:**

- [Intel Arria 10 SoC Development Kit](#)

Embedded Example – Agilex 7 SoC  
**Available Now**



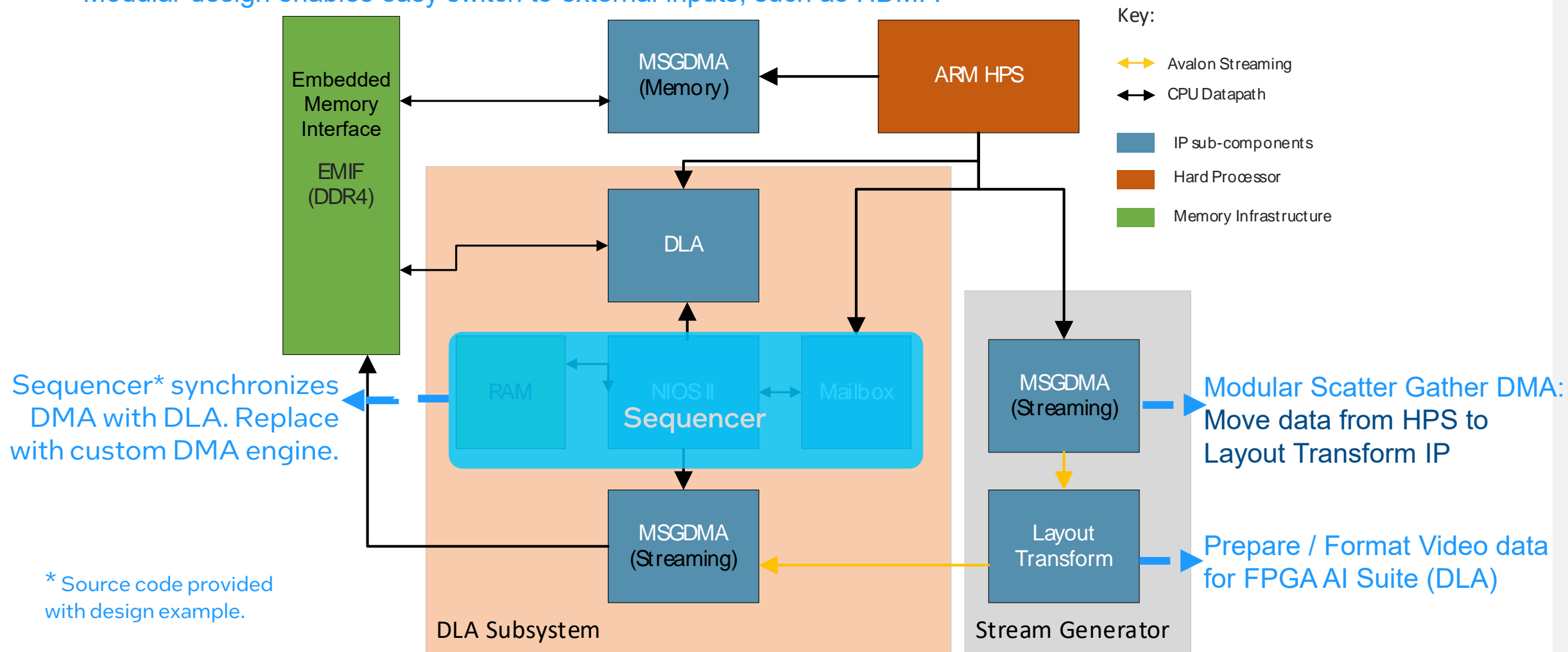
**Intel® Agilex® 7 FPGA I-Series Transceiver-SoC  
Development Kit (4x F-Tile)  
with embedded Arm CPU as host**

**Hardware required:**

- [Intel Agilex® 7 FPGA I-Series Transceiver-SoC Development Kit \(4x...](#)

# FPGA SoC Example Design: Streaming

- ARM HPS Linux application\* emulates streaming video with MSGDMA streaming mode.
- Modular design enables easy switch to external inputs, such as HDMI .



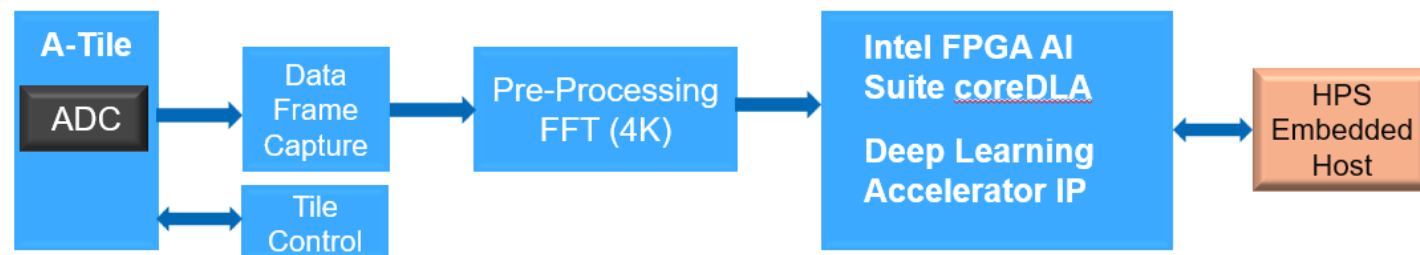
# RF Signal EagleNet Classification Example Design

## ■ Dataset Collection

- Sample Rate 48GSPS
- DDC x32 decimation mode
- NCOs: CNCO=3.75GHz, FNCO=0
- Size of captured segment: 4096 I/Q samples
- Number of waveform classes: 7
- ADC uncalibrated

## ■ Equipment

- Signalhound VSG60A
- Stratix-10 AX Dev.Kit – RevA/ES1



Description	Folder name	Number of waveforms
Carrier at 3.85GHz, AM modulated with 2MHz sine data	Am_2M	800
Sine tone at 3.85GHz	Cw	800
Carrier at 3.85GHz, FM modulated with 10MHz maximum deviation sine data	Fm_sine_10M	1,000
Background noise of ADC when signal generator in RF off	No_signal	1,000
OFDM with 32 carriers, each modulated QAM16. Overall BW=20MHz	Ofdm_32car_qam16_20M	1,000
Digital modulation QPSK with symbol rate 12.5MSPS	Qpsk_rect_12p5M	1,000
Chirp of 40MHz with period of 10us	Ramp_10us_40M_m8dbfs	1,000

# Summary

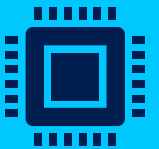


Data scientists or machine learning engineers can use end-to-end Intel AI software →

- 1) Development - data ingest, training
- 2) Deployment- FPGA-based inference with Intel® FPGA AI Suite and OpenVINO™ toolkit



FPGA AI Suite generates right-sized AI for your latency/ power/ performance needs



FPGA engineers can integrate the deep learning FPGA inference IP with the Intel® Quartus® Prime Software and Platform Designer

# Call to Action

## ❖ **Download Intel FPGA AI Suite.**

- Step through the Getting Started Guide and run pretrained images on hardware
- Modify design example in Intel Quartus
- Run limited inferences in hardware without license
- Purchase license to run unlimited inferences

# Thank You

