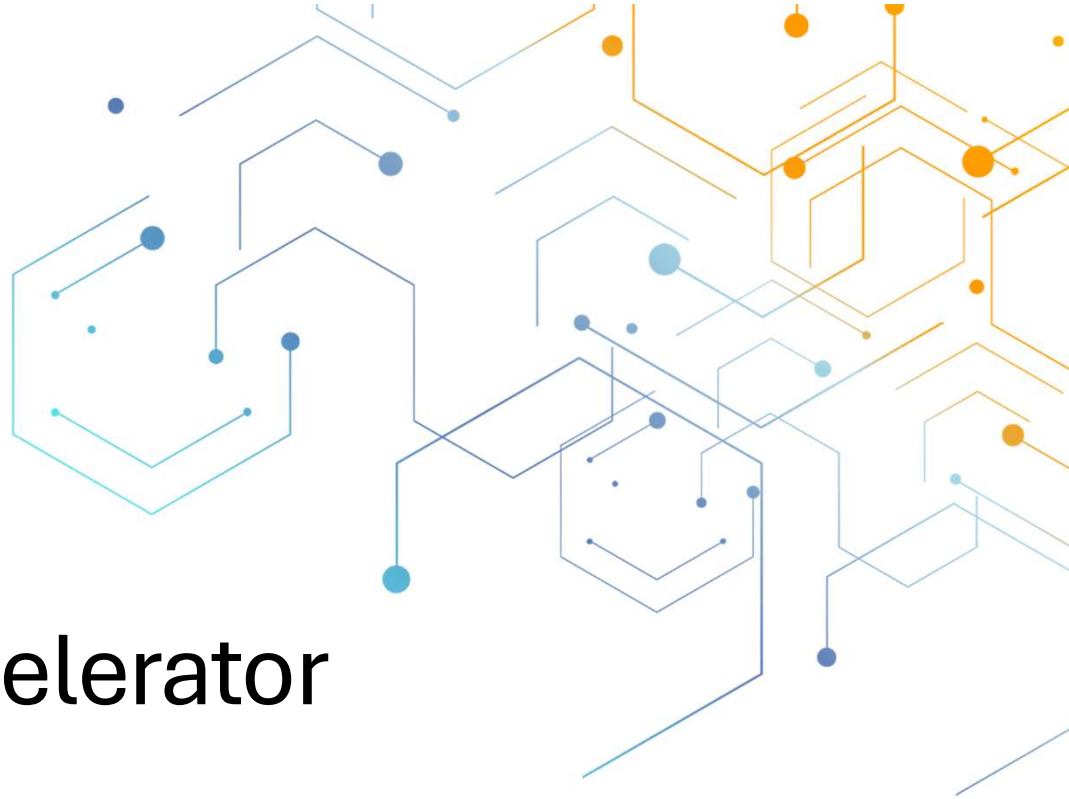


Why Your Next AI Accelerator Should Be an FPGA

What is an FPGA?

“A Silicon Device that Can Be Programmed, and Reprogrammed to Contain Any Desired Circuit or Function”

About Efinix

- Founded in 2012
- Headquartered in Cupertino CA
- 200+ Employees
- Global Footprint
- Over 50M Units Shipped

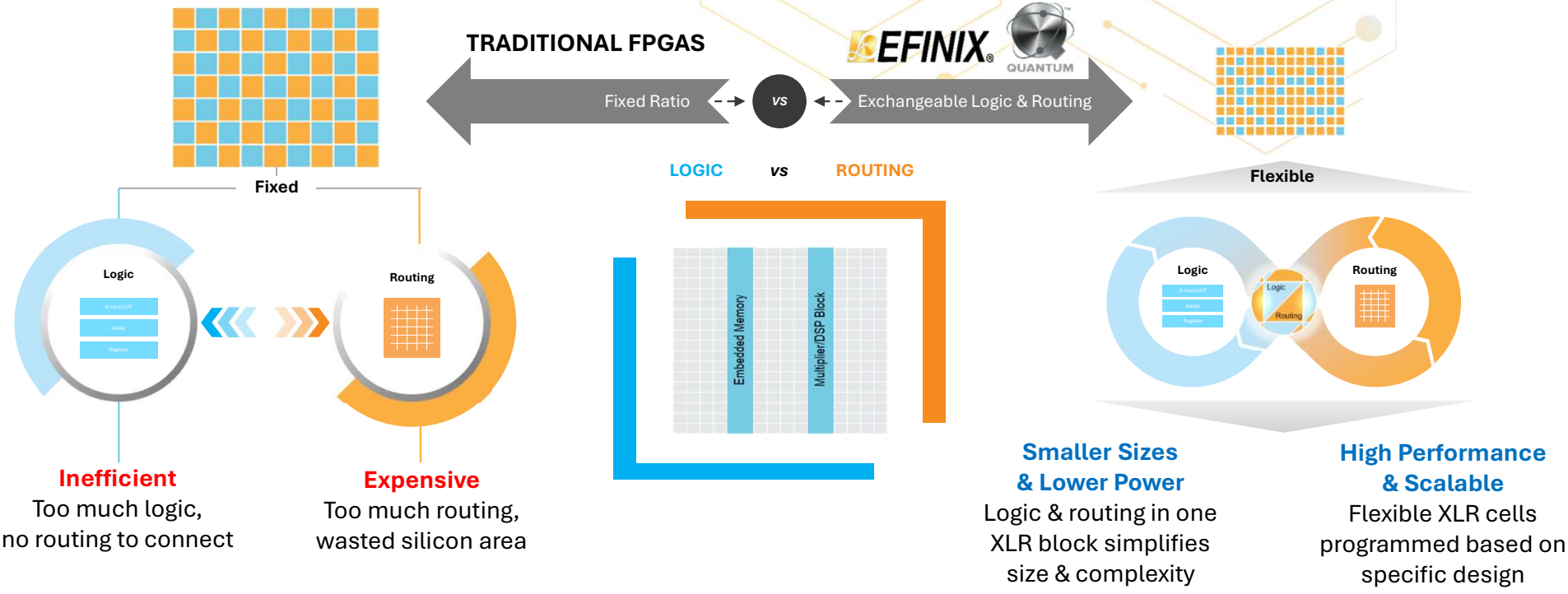
Trion FPGAs
4k – 120k LE Connectivity optimized

Topaz FPGA
50k – 325k LE High Volume Value optimized

Titanium FPGA
35k – 2M LE Performance optimized

Efinity IDE Software
Simple and intuitive GUI

Patented FPGA Architecture



Inefficient
Too much logic, no routing to connect

Expensive
Too much routing, wasted silicon area

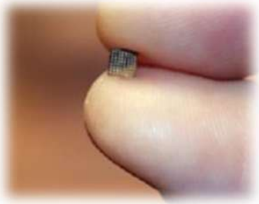
Smaller Sizes & Lower Power
Logic & routing in one XLR block simplifies size & complexity

High Performance & Scalable
Flexible XLR cells programmed based on specific design

Small Form Factor

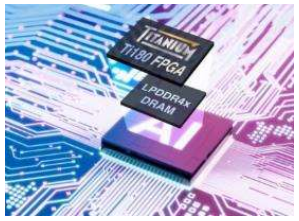
Low Power and Small Form Factor for Cost Sensitive Edge Applications

Ti60W64
3.5x3.4mm
WLCSP package



Ti60F100S3F2
5.5x5.5mm
256 Mb HyperRAM +
16 Mb Flash

Ti180J484D1
15x15mm
2 Gb LPDDR4x
3000 Mbps



Ti135N567D2F4
16x16 mm
2 Gb LPDDR4X
16 Gbps SerDes

FPGA Value Proposition

- Low Power with Hardware Performance
  - Performance of a Parallel Hardware Architecture
  - Low Power for Edge Applications
- Ultimate Flexibility
  - Reprogrammable, Software Defined Architecture
  - Adapt to Changing Standards Even After System Deployment
- Cost Effective
  - Zero Risk and NRE
  - Highly Integrated for Low BOM Costs

Edge Constraints

In the Data Cener, AI is the Mission
At The Edge, It's a Feature to Achieve the Mission

- Low Power
- Harsh Environments
- Diverse & Evolving Interfaces
- Fast Deterministic Response

The Shift to Edge AI

AI Belongs Where Data is Created.

- Real Time Response
- Enhanced Privacy
- Sensor Fusion
- Pre-Processing
- Local Decision Making

FPGAs vs GPUs For Edge AI

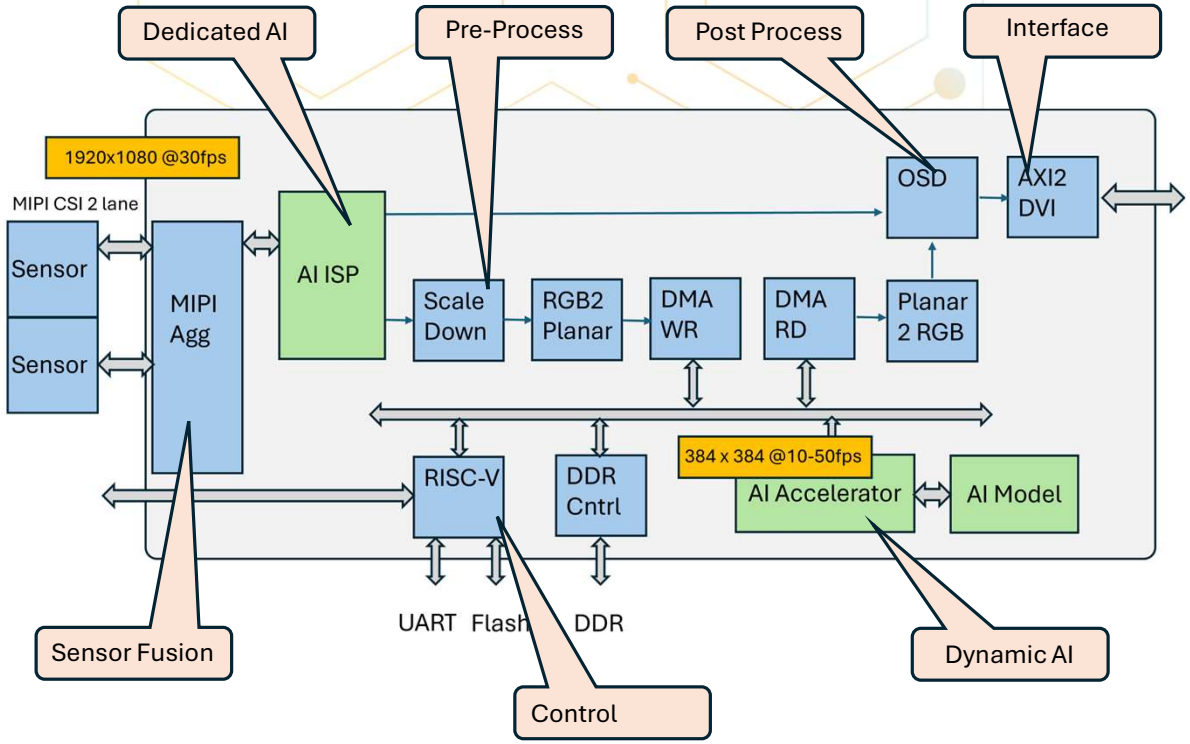
Metric	FPGA	GPU
Power Consumption	Single Digit Watts	10s of Watts
Latency	Deterministic	Non-Deterministic
Reconfigurability	Fully Hardware Reconfigurable	Software Reconfiguration
Best Use Case	Edge AI Inference	Cloud Training

FPGAs vs Custom Silicon

Metric	FPGA	Custom Silicon
Time to Market	Weeks	Years
NRE	No NRE	Millions of Dollars
Reconfigurability	Flexible for Evolving Models	Fixed Hardware Architecture
Per Unit Cost	Medium	Low

AI in Highly Integrated Designs

- Not All AI Models Share the Same Performance Requirements
- AI Imposes System Requirements Beyond the Actual AI Model
- Diverse Sensors Require Different Treatment Under AI
- AI Needs To “Plug In” Seamlessly In a Larger System Design



FPGAs Provide the Perfect Sand Box For AI Integration

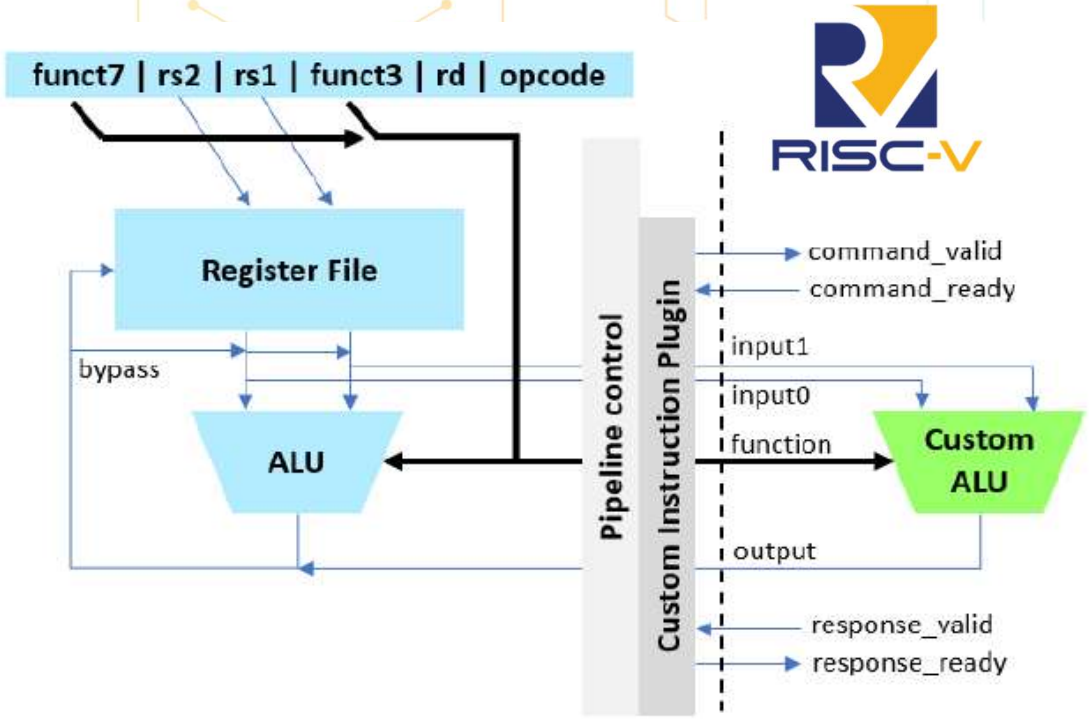
Harnessing the Power of FPGAs

FPGAs Deliver Ultimate Flexibility to Tradeoff Performance, Power, Time to Market and Cost in Highly Integrated Designs

But How Do We Make That Flexibility Available to Designers?

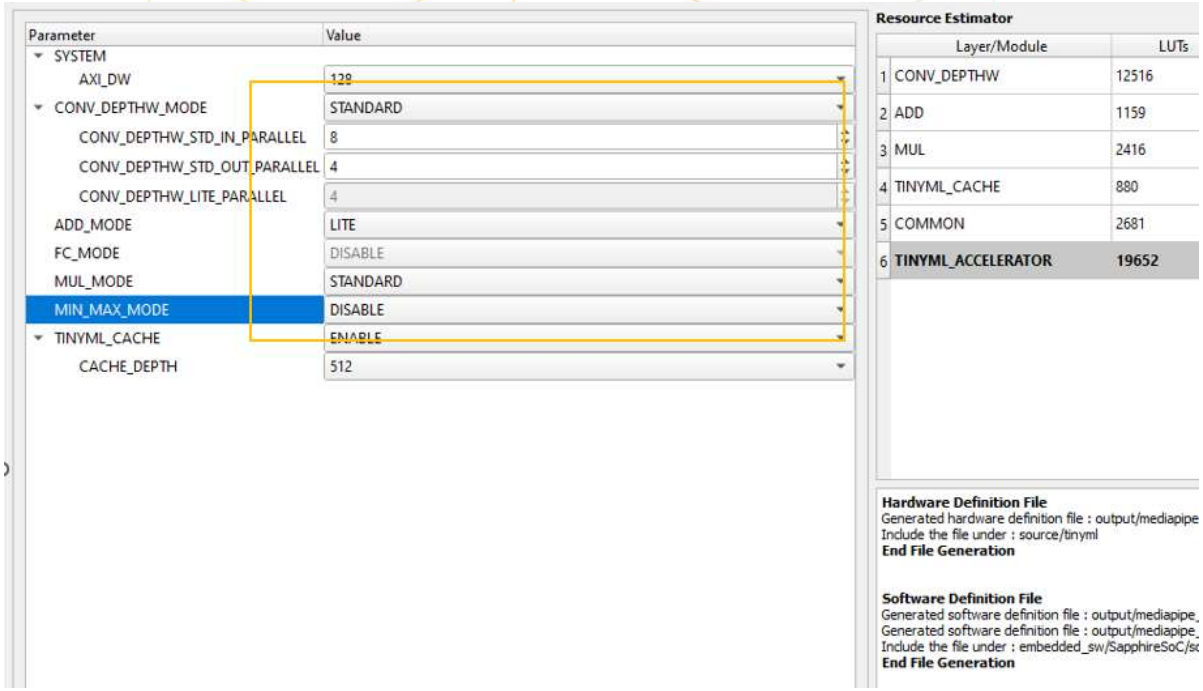
A Software Centric Approach

- Not All RISC-V Instructions are Defined
- A Custom ALU Can Be Defined To Implement Undefined Instructions
- Custom ALU Can Absorb AI Primitives at Hardware Speed
- This is a Perfect Application Fit for FPGAs



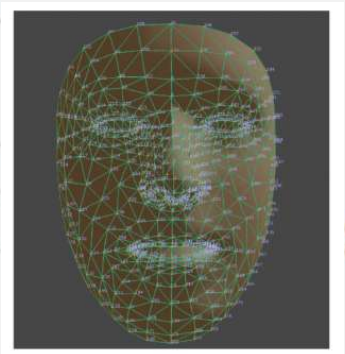
Accelerator Generator

- Automatically Analyzes Model
- Presents List of Available Accelerators
- Graphically Configures Accelerators
- Automatically Generates Project Files



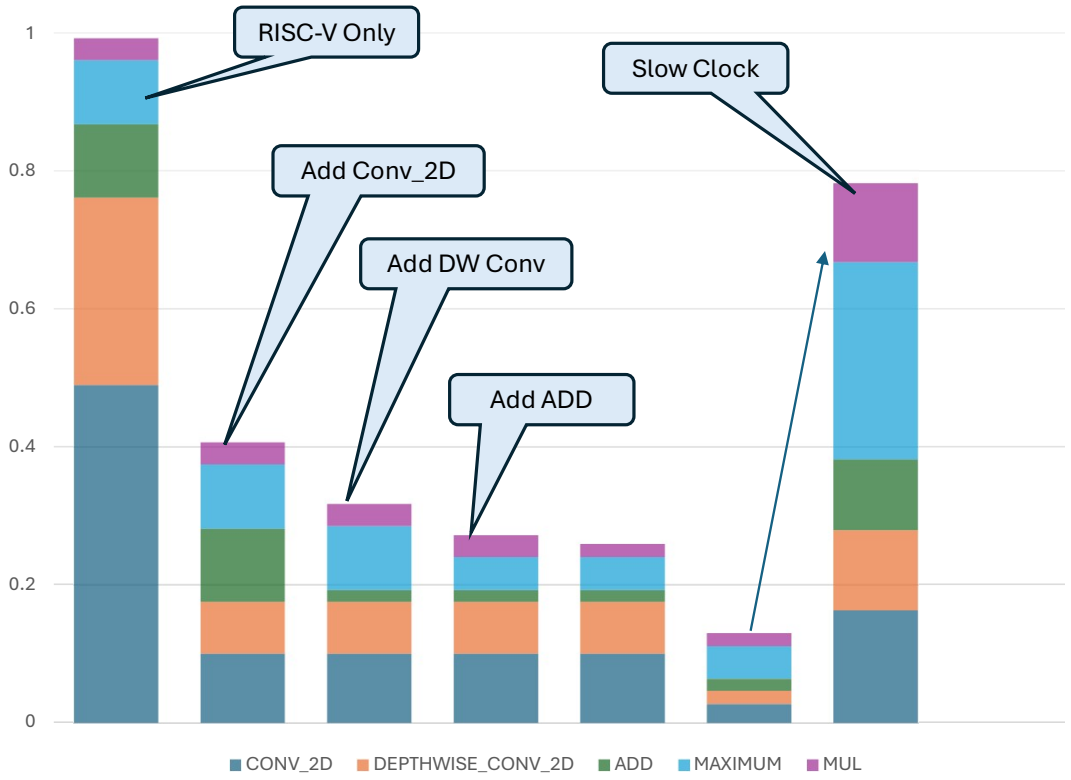
Design Process Example

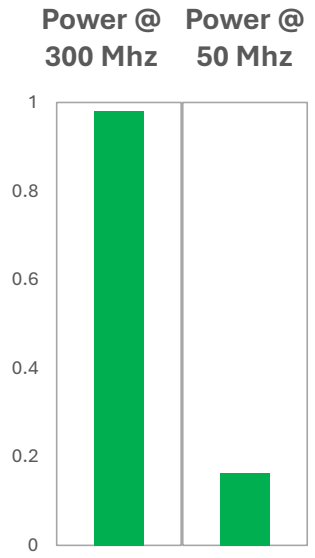
MediaPipe Face Mesh



Pre-trained Network, Run on RISC-V & Profiled

Layer/OP	Latency (%)
CONV_2D	49%
DEPTHWISE_CONV_2D	27%
ADD	11%
MAXIMUM	9%
MUL	3%
PAD	<1%
MAX_POOL_2D	<1%





Added Performance Can Be Traded Off Against Power By Slowing Clock

Quick Start Design Examples

Visit https://github.com/Efinix-Inc/tinyml/ to Access Open-source Examples.

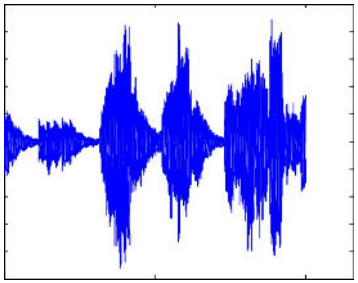
Explore Ready-to-run Designs:

- YOLO Person Detection
- MobiletNetV1 Person Detection
- MediaPipe Face Landmark Detection

See Edge AI in Action on the Ti60F225 Development Kit!

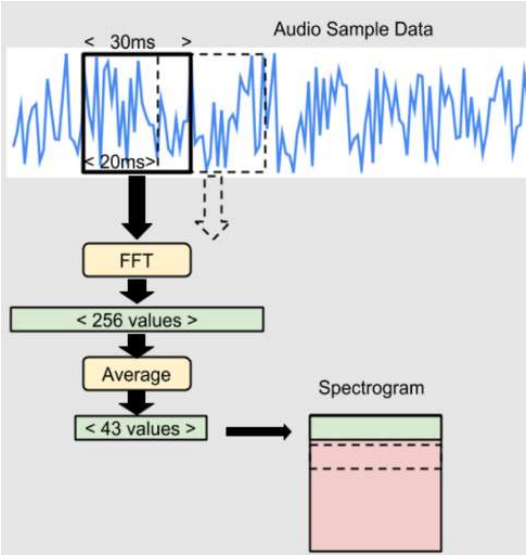
- Frequency: 300 MHz
- Power : ~300 mW

Application: Identify Wakewords



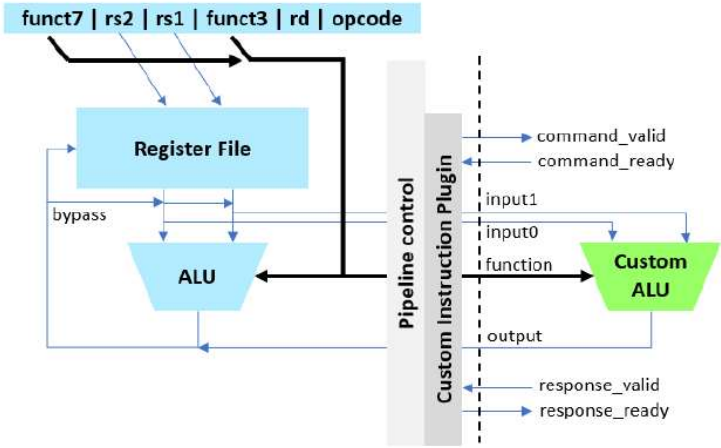
Audio Input

Pre-processing In Software
On a Soft RISC-V Core



Audio Pre-processing

AI Acceleration in a Custom
ALU In the FPGA Fabric



Resources & Performance

Resource Utilization in Trion Family with 20K LE

Building Block	19728 Logic Elements	204 Memory Blocks
RISC-V SoC	7474	187
Instruction Accelerators	4563	12

Inference Time (RISC-V @100MHz)

Application	Model Storage Option	Inference time (ms)	
		Pure SW	With Lite Accelerator
Keyword Spotting	Tensor data in SPI Flash	999	376

Power Consumption with Accelerator: ~124.8mW

Human Presence Detection

Model Card	
Model Name	MoiblenetV1
Model Type	Classification
Input Format	NCHW (1 or 3 Channels)
Output	2 (Person , Not Person)
FLOPs (M)	15

Hardware Deployment

Device: Titanium with 60K LE

System Frequency(MHz) : 300

Resource Utilization (RISC-V): 20%

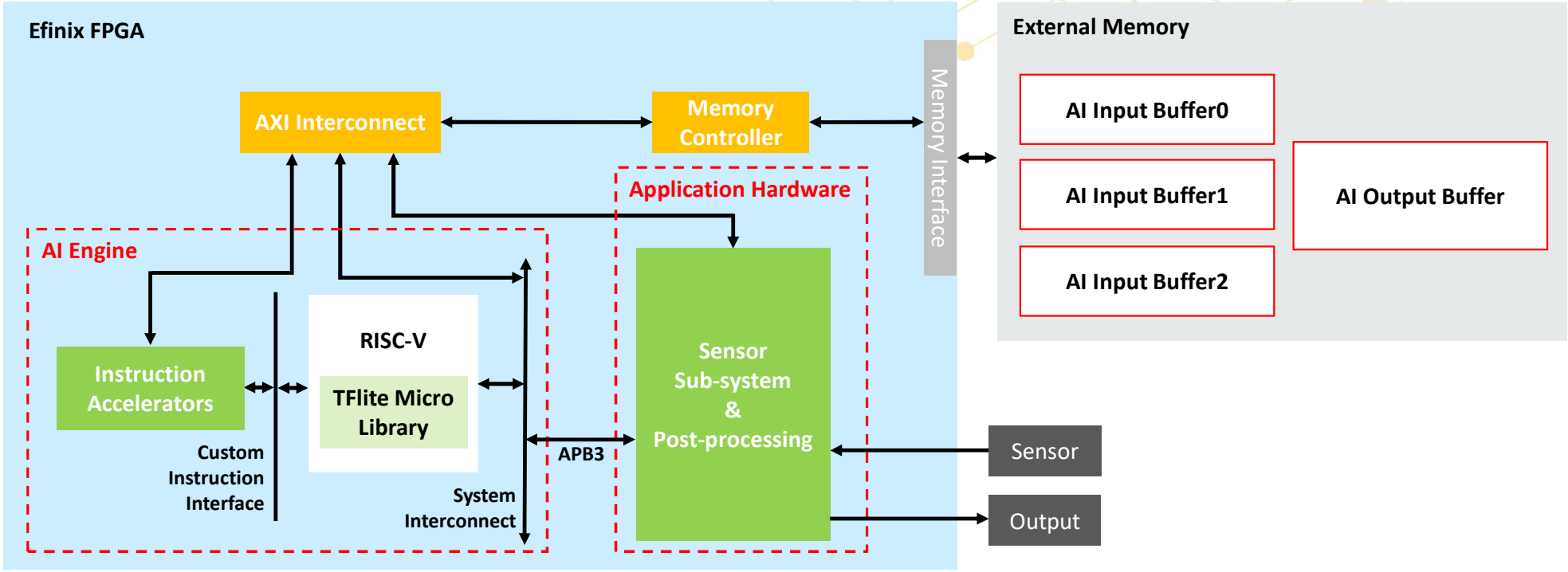
Input: 96x96x1

Output: 2 classes (person, not person)

Performance (FPS) : 6

Power: 320 mW

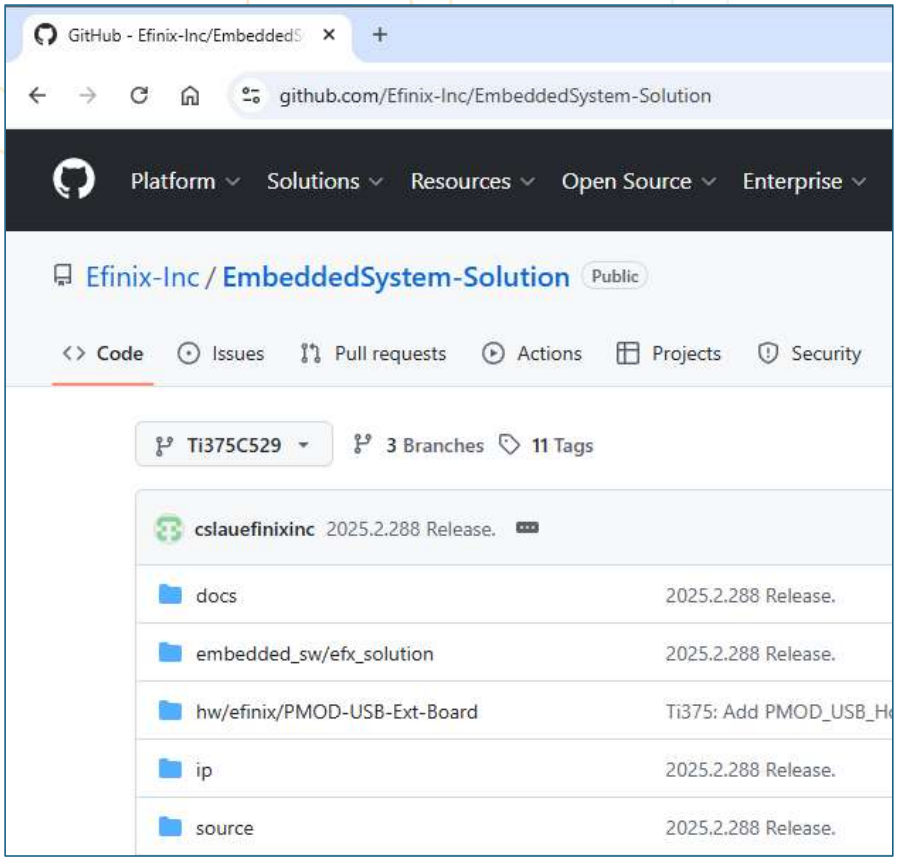
Ease of Integration



GitHub Reference Platforms

https://github.com/Efinix-Inc/EmbeddedSystem-Solution

https://github.com/Efinix-Inc/tinyml/



The Pros and Cons of Instruction Acceleration

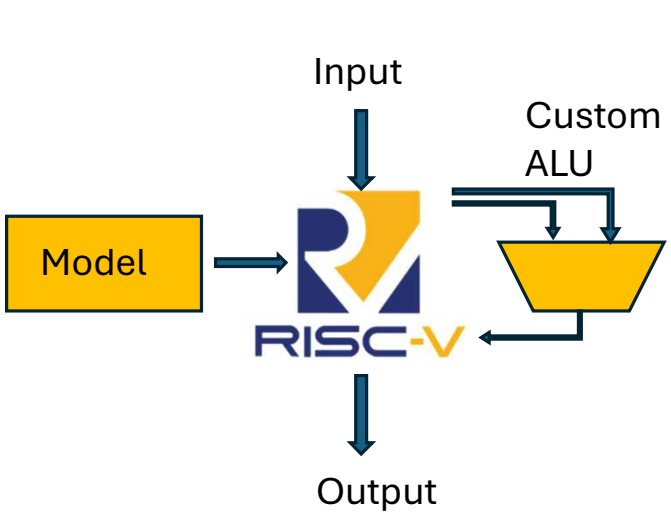
Benefits

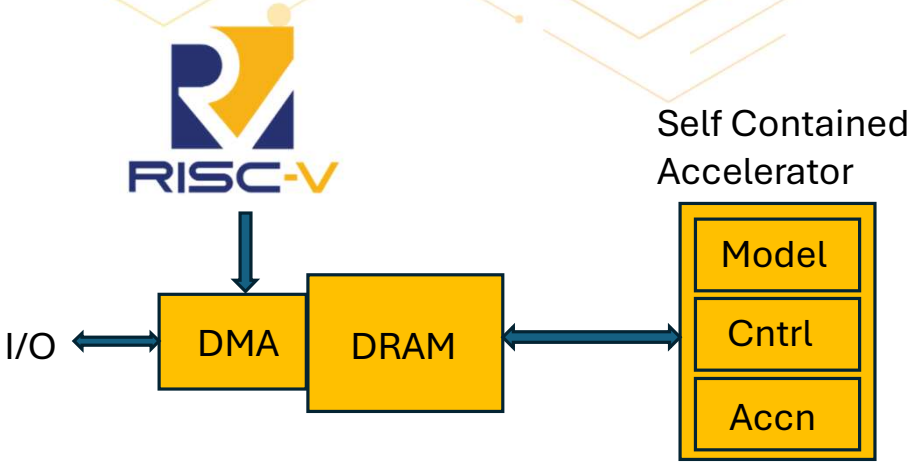
- Intuitive Software Flow
- Turnkey Acceleration
- Fast Time to Market
- Flexibility
- Scalable Acceleration

Drawbacks

- Non-Deterministic Latency
- Performance Limitation
- Limited Model Complexity

Rearchitecting the Acceleration



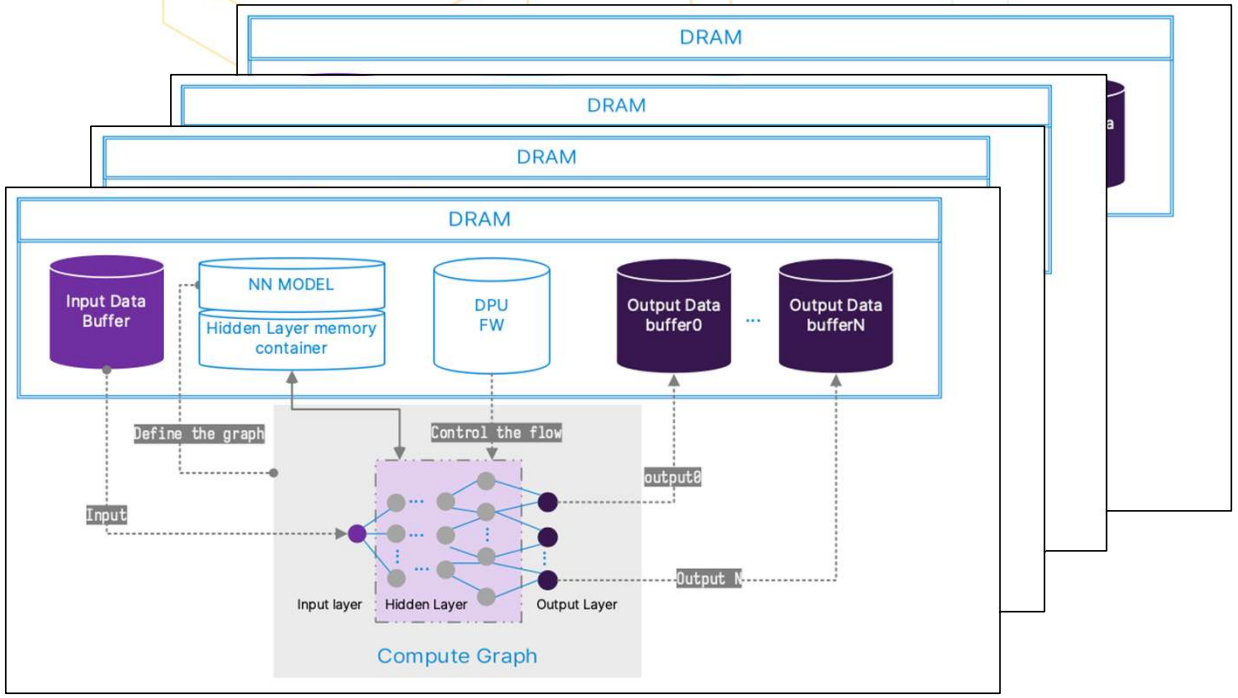


Efinix AI Engine Accelerator

High Performance and Optimal Efficiency

Dedicated Hardware Accelerators

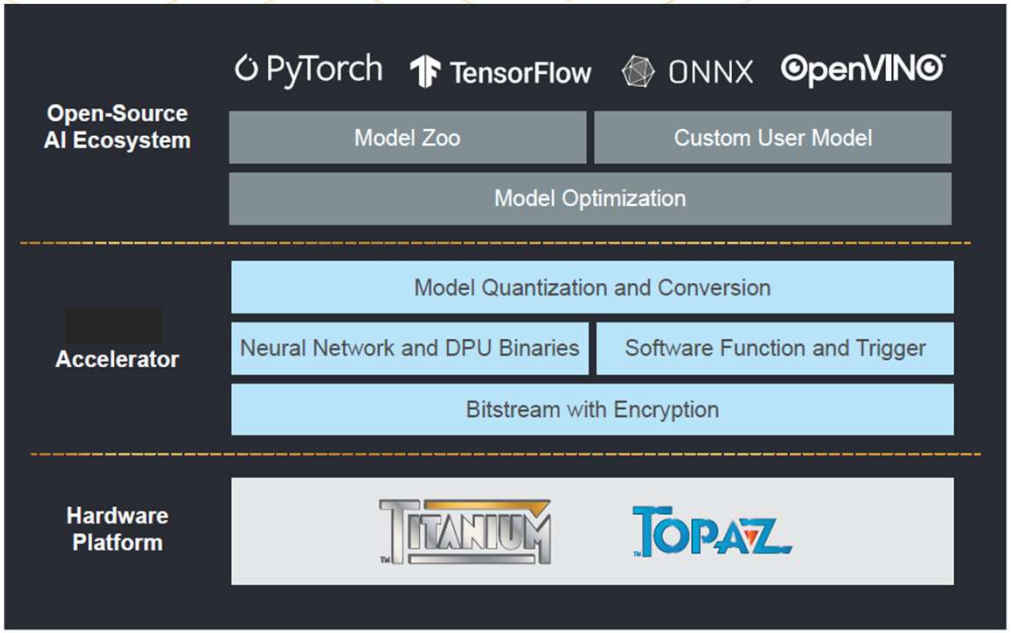
- **Scalable Architecture**
  - Up to 4 Acceleration Units
  - Delivering 1.6 TOPS INT8
- **Optimized Efficiency**
  - Efficient Implementation on FPGA Fabric



Turnkey AI Accelerator for FPGAs

Powerful, Full-Featured Engine for Deep Neural Networks

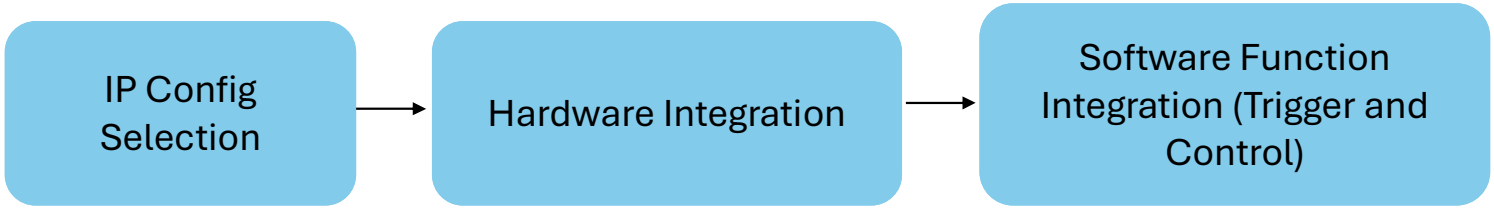
- **Framework Support**
  TensorFlow, PyTorch, Caffe, and ONNX
- **Model Flexibility**
  Rich Library of Accelerators
- **Rich Model Zoo**
  ~100 Pre-Trained and Ported Models



Scalable and Flexible

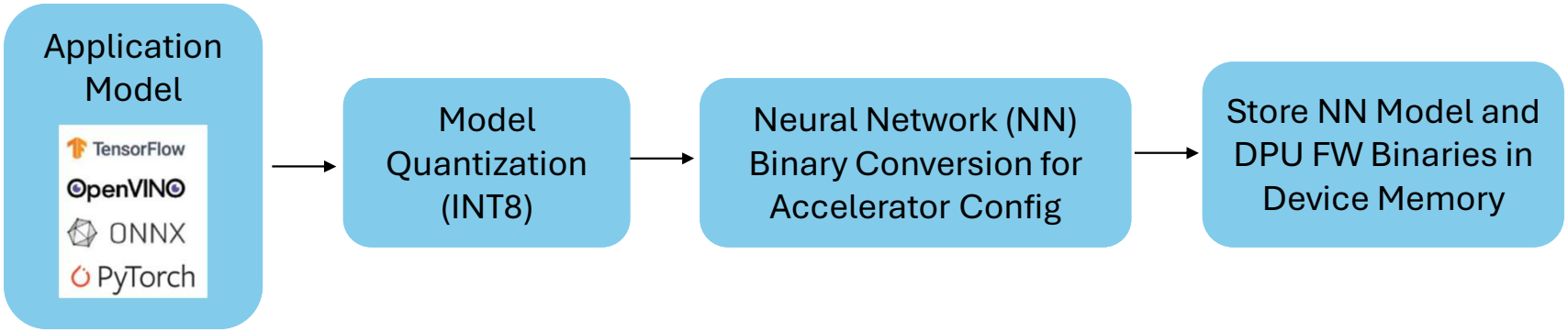
- Select Accelerator Configuration Based On:
  - Resource
  - Performance
  - Power

- Integrate With Customer's System with Proper Handshake.



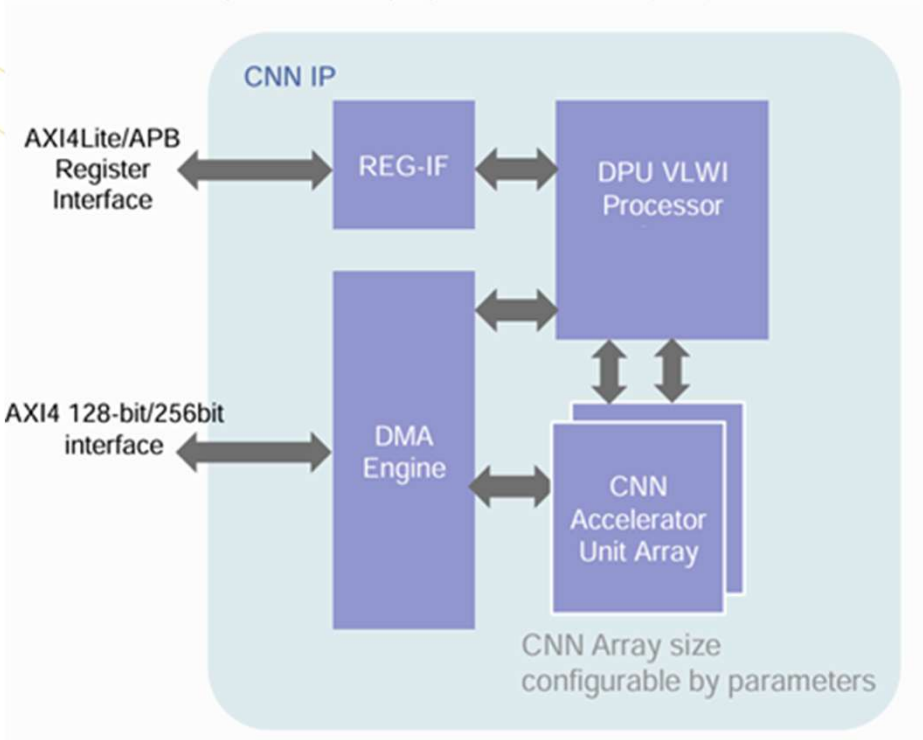
Model Conversion

- **Proprietary Model Conversion and Quantization Tool**
  - Quantization of Trained Model (Float32) to INT8 Used by IP
  - Quantization Error <1%
  - Creation of Binaries and Bitstream for Device Memory.

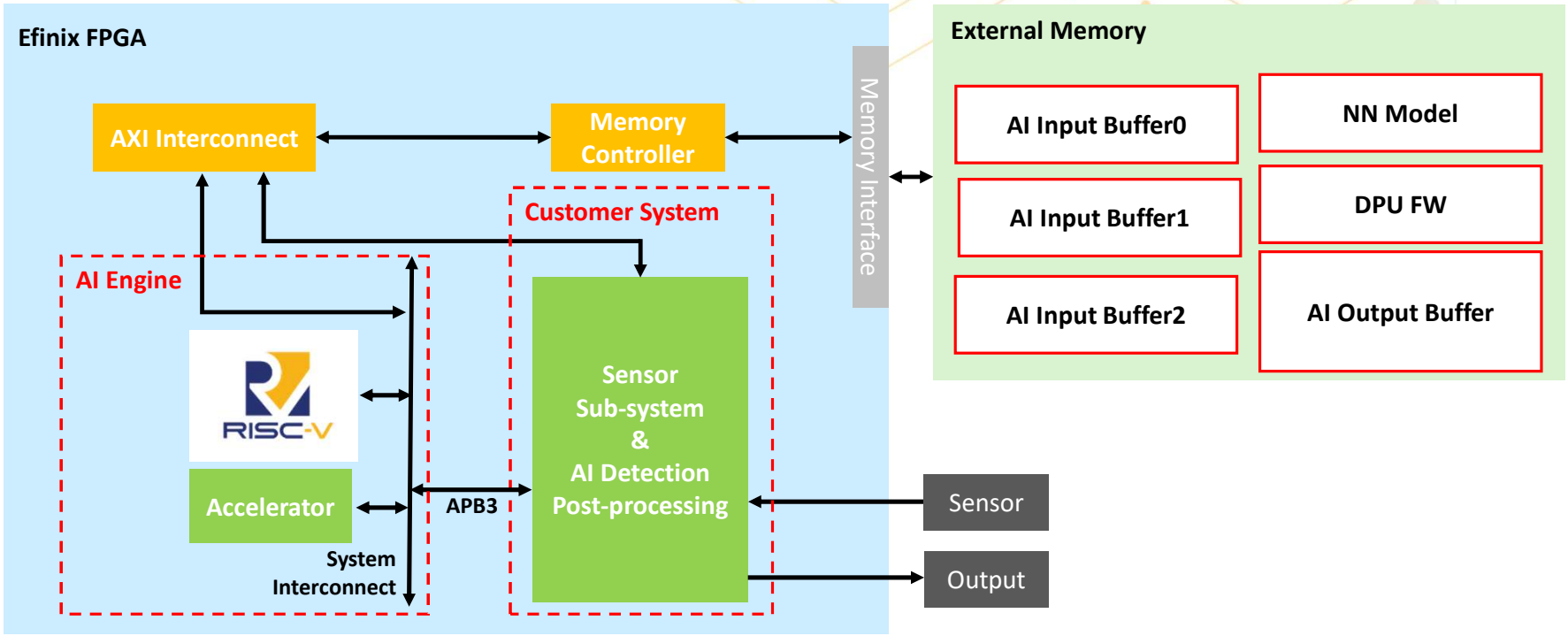


Ease of Integration

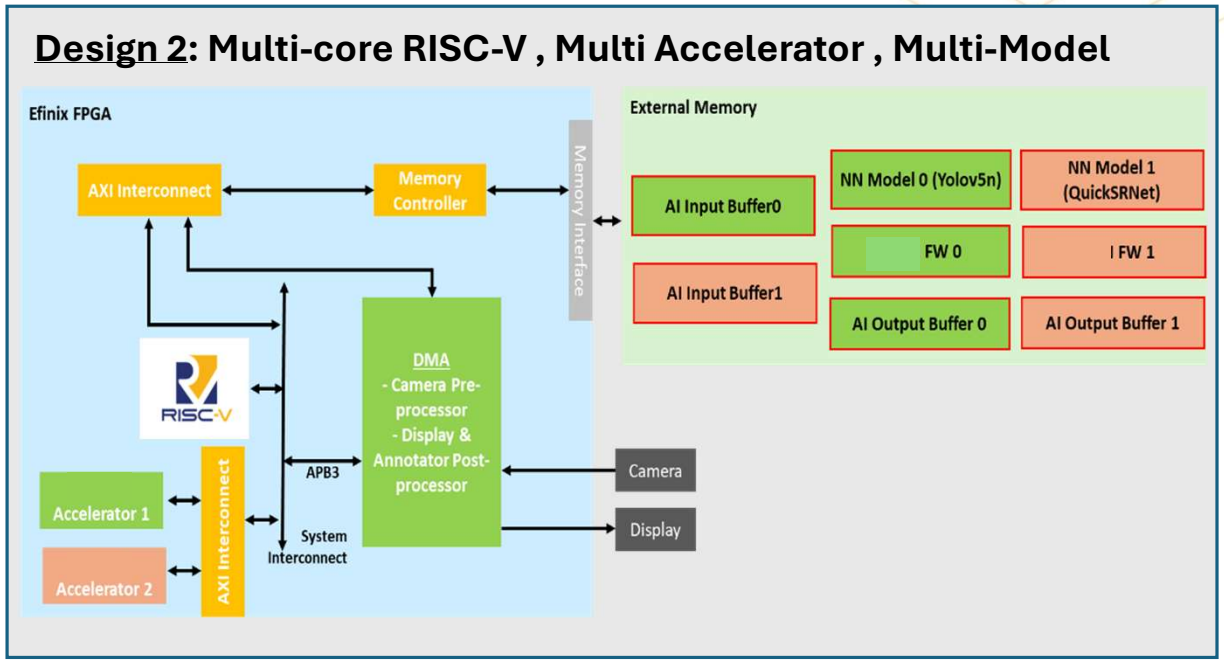
- AXI4Lite/APB3 Register Interface with Simple Start/Stop/Status Query.
- Embedded DPU Processor Can Process All NN Layers Independently
- Output Stored in DRAM Buffers.
- Minimal API for Control and Trigger.



Same Standardized Integration



Multi Model Scalability



- Each RISC-V Core Can Control and Trigger With Separate Buffer
- Both Accelerators Run Independently
- Supports Running Different Models

Scaling for Power and Performance

Benchmark on Titanium Ti180 FPGA @ 200MHz:

Configuration	Yolov5n Performance	AI Engine Power	BlazeFace Performance
Small (1xAccelerator)	~25 fps	**0.7 W**	167 fps
Medium (2xAccelerators)	~43 fps	**1.2 W**	~200 fps
Big (4xAccelerators)	~62 fps	**1.7 W**	~208 fps

Super Resolution Use Case

Model Card	
Model Name	QuickSRNet-pruned
Model Type	Super Resolution
Input Format	NCHW (1 or 3 Channels)
Output	Enhanced/Scaled Image
Scaling Factors	1.5x, 2x, 3x, 4x
Param (M)	~0.1-0.5
FLOPs (G)	~0.05-0.3

Hardware Deployment

Device: Ti180J484

IP Config : Small (1xCNN)

Frequency(MHz) : 200

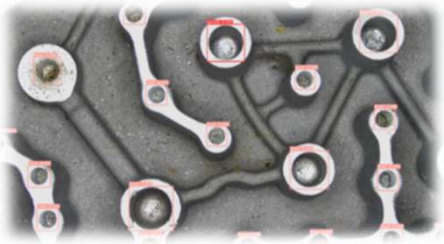
Input: 640x512x1 (Grayscale)

Output: 1280x1024x1 (Grayscale)

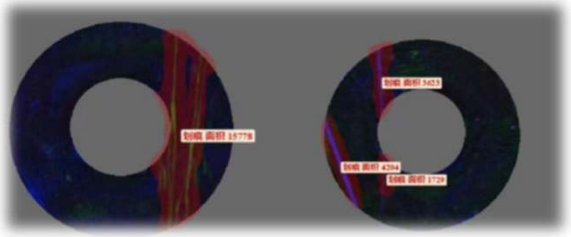
Performance (FPS): 30

Power (AI Engine): 505 mW (0.85 V)

Intelligent Camera Use Case



Defect Detection



Defect Segmentation

Model Card	
Model Name	Yolov5l-seg
Model Type	Object Detection + Segmentation
Input Format	NCHW (1 or 3 Channels)
Output	Detection based on classes and Instance Segmentation
Param (M)	46.5
FLOPs (G)	109.1

Hardware Deployment

Device: Ti180J484

IP Config: Big (4xCNN)

Frequency(MHz): 200

Input: 1080x1080x3 (RGB)

Output: 10 class detect detection + segmentation

Performance (FPS): 2

Power: 1.7 W

Scalable AI Acceleration

Architecture	Model		Hardware	
	Applications	Architecture	Performance	System Power
RISC-V Alone	Human Presence Detection	MobilenetV1	**1 FPS**	**160 mW**
RISC-V With Custom Instruction Acceleration	Smart Goggles	Tiny YOLOv3	**2 FPS**	**300 mW**
RISC-V with 1 EFX_AIE Accelerator	Super Resolution	QuickSRNet-pruned (2x)	**30 FPS**	**1.6 W**
RISC-V with 4 EFX_AIE Accelerators	Object Detection	YOLOv5n	**62 FPS**	**3.1 W**

Summary

FPGAs Deliver…

A Scalable Approach to Deploying Edge AI Models

A Future Proof Platform That Can Be Updated Even After Deployment

Fast Time to Market With Zero Risk and NRE

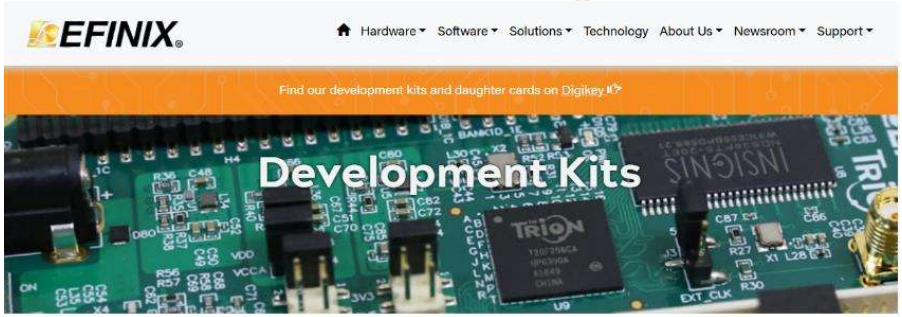
Low Power With Highly Parallel, Deterministic Performance

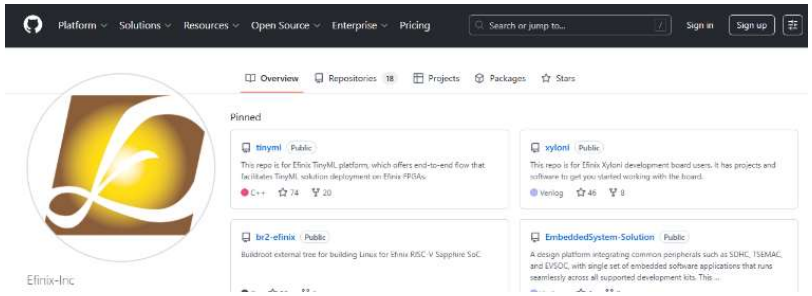
High Integration for Low BOM Costs

How to Get Involved

Register on the Efinix Support Site / Download the Free Tools

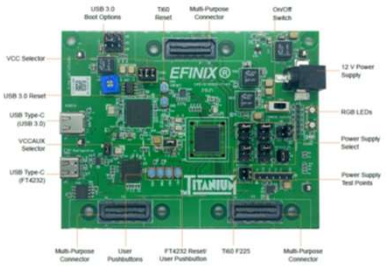
Buy An Efinix Development Kit and Start Designing



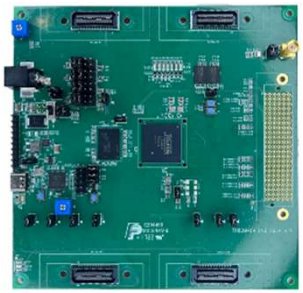


Visit the Efinix Web and Github Sites For Quick Start Examples

Titanium Development Kits



Ti60F225C-DK



Ti180J484C-DK

Ti375C529C-DK

Ti375C1156C-DK

Daughter Cards

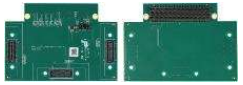
EFX_DC_GPIO_B
LVDS expansion

EFX_TI60_2x30_IFB
Dual MIPI to DSI

EFX_DC_CAM_FPC15_B
Raspberry PI Con

EFINIX_IFB_PICAMX2
Dual Raspberry PI Con

EFX_FMC_DDR3_GBIO
FMC DDR3 and GPIO



EFX_GPIO_FL_DC_C
FMC to QSE

EFX_RGMII
Ethernet card

Come See Us…

FPGA Horizons Worcester MA – April 29/30

Embedded Vision Summit – May 11/13

Embedded World North America – September 22/24

Thank You

Embedded Vision Summit
May 11-13, 2026

The premier conference for innovators incorporating computer vision and physical AI in products

Why Attend?:

- First rate speakers with practical knowledge.
- Relevant exhibitors who can help with your challenges.

15% off until April 10th with discount code:
26EVSUM-WEBINAR

Register here: embeddedvisionsummit.com

99%
of attendees would recommend the event

"The Embedded Vision Summit offers a special and unique opportunity to foster knowledge sharing across the edge AI ecosystem."
—Dan Teodorescu, Director of Product Architecture, Alcon

The Edge AI and Vision Alliance

Empower product creators to harness physical AI and vision

The Alliance inspires and empowers system and solution creators to design better products that perceive and understand.

Accelerate technology supplier success

The Alliance helps companies that provide physical AI and vision building-block technologies, services and solutions to grow their businesses through leads, partnerships, and insights.

Empowering Product Creators to Harness Physical AI and Vision

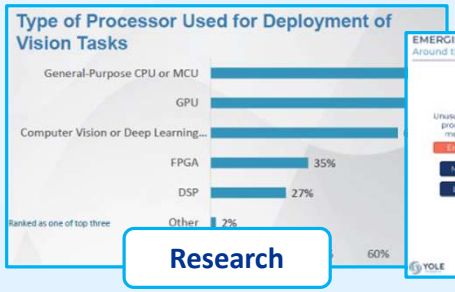
Alliance Mission

The Alliance inspires and empowers system and solution creators to design better products that perceive and understand:

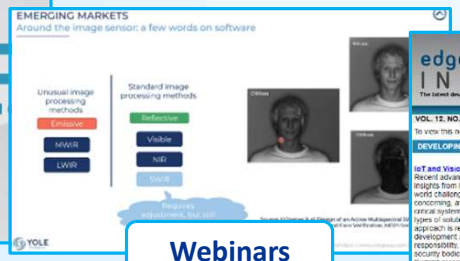
Expand and sharpen your skills

Stay up to date

Connect with key technology suppliers



Research



Webinars

Newsletter

Embedded Vision Summit

Register for updates at www.edge-ai-vision.com