Lightning Talk:
A Path to Scalable Low Power AI
Efinix TinyML Flow for Titanium FPGAs

Joachim Mueller
Efinix

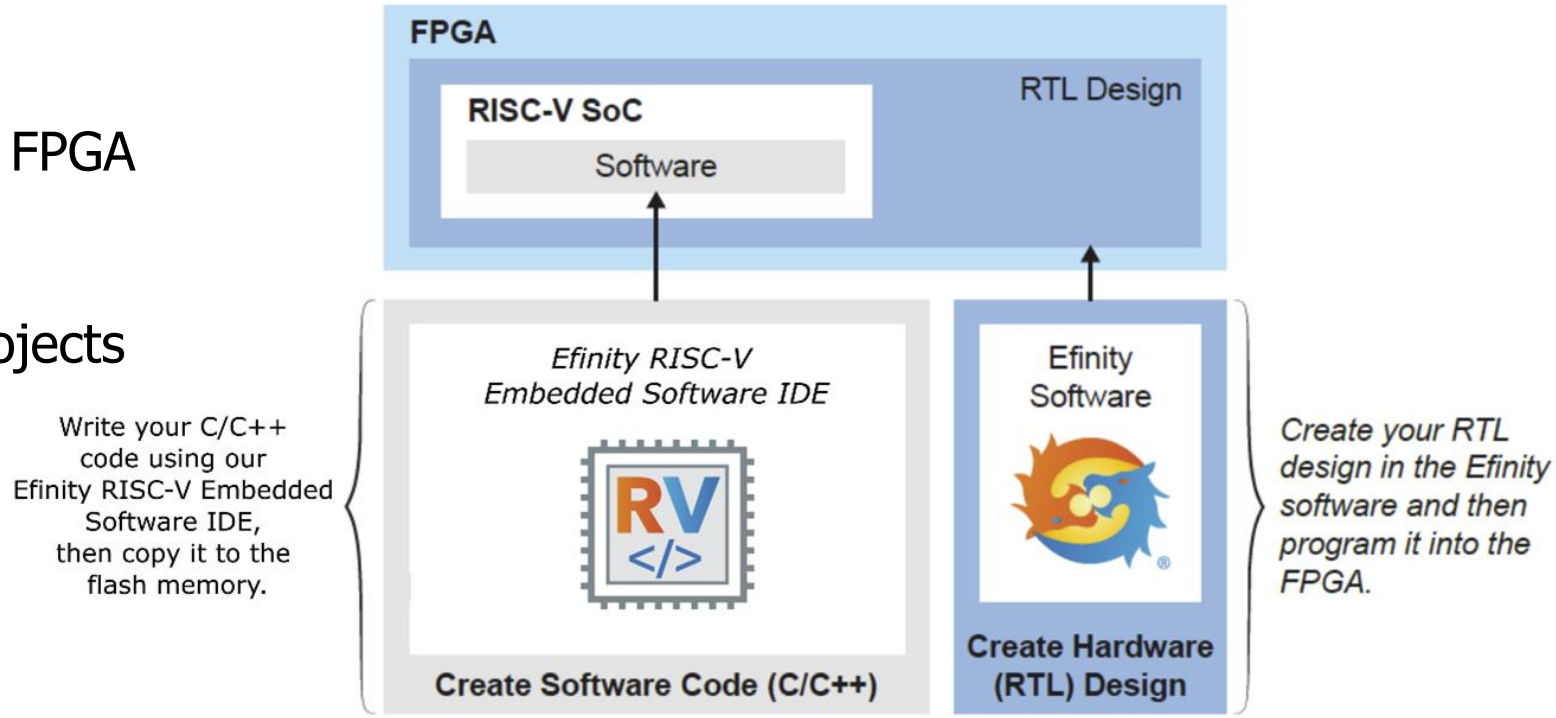
Prerequisites: Software and Tools

Efinity Design Environment

- Configurable Sapphire RISC-V® in Efinity IP Catalog
- Create RTL Design
- Implement and program the FPGA

Efinity RISC-V IDE

- Eclipse IDE for managing projects
- Create Software Code
- Debug and Build

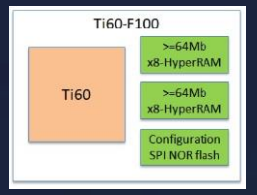


Prerequisites: High-Performance Titanium FPGA Family

TSMC 16 nm Process

QUANTUM
Quantum Compute Accelerated

Innovative Package Options



System in Package

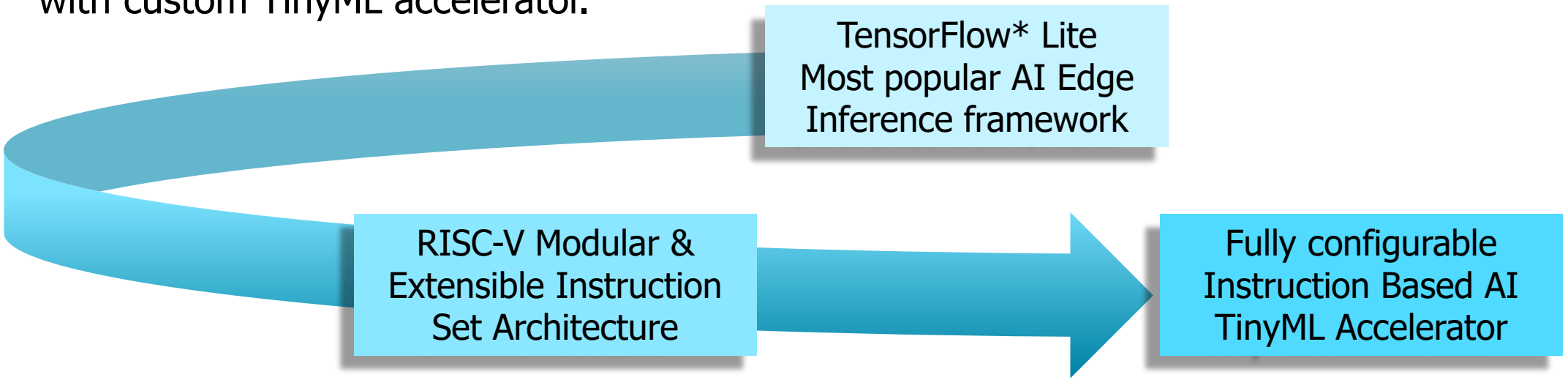
			Ti35	Ti60	Ti90	Ti120	Ti180	Ti240	Ti375	Ti550	Ti750	Ti1000
LE (k)			36	62	92	123	176	237	370	533	727	969
10K RAM (Mb)			1.53	2.62	7.34	9.80	13.11	19.37	27.53	39.65	54.07	72.09
DSP			93	160	359	478	640	946	1344	1936	2640	3520
PLLs			4	4	10	10	10	10	10	10	10	10
MIPI DPHY 1.5G, support CSI-2/DSI			Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
MIPI DPHY x4 2.5Gbps Hard IP			-	-	X4	X4	X4	X3	X3	X3	X3	X3
LPDDR4/X			-	-	X32	X32	X32	2 X32	2 X32	2x72	2x72	2x72
RISC-V hardwired								x	x	x	x	x
16 Gbps Serdes			-	-	X8	X8	X8	X16	X16	X16	X16	X16
25.8 Gbps Serdes			-	-	-	-	-	-	-	X8	X8	X8
PCIE Gen4 (16G) hard controller			-	-	1x PCIE Gen4x4	1x PCIE Gen4x4	1x PCIE Gen4x4	2x PCIE Gen4x4	2x PCIE Gen4x4	2x PCIE Gen4x8	2x PCIE Gen4x8	2x PCIE Gen4x8
Packages (HSIO/HVIO/LPDDR4/MIPI /S SERDES)												
W64	0.4mm	3.5x3.4mm	-	34/0			-	-	-		-	-
F100 (SIP)	0.5mm	5.5x5.5mm	61/0	61/0			-	-	-		-	-
F225	0.65mm	10x10mm	140/23	140/23				-	-		-	-
M361	0.65mm	13x13mm	-	-	110/20/16/2/0	110/20/16/2/0	110/20/16/2/0	-	-		-	-
L484	0.8mm	18x18mm	-	-	116/27/0/4/0	116/27/0/4/0	116/27/0/4/0	-	-		-	-
M484	0.8mm	18x18mm	-	-	116/27/32/4/0	116/27/32/4/0	116/27/32/4/0	-	-		-	-
A484 (SIP)	0.65mm	15x15mm	-	-	210/54/16/1/0	210/54/16/1/0	210/54/16/1/0	-	-			
M529	0.8mm	19x19mm	-	-	210/48/32/0/0	210/48/32/0/0	210/48/32/0/0	-	-			
N484**	0.65mm	15x15mm	-	-	-	-	-	60/29/32/1/8	60/29/32/1/8	x/S	x/S	
M676**	0.8mm	21x21mm	-	-	-	-	-	130/60/32/2/12	130/60/32/2/12	x/S	x/S	
N900**	0.8mm	25x25mm	-	-	-	-	-	172/60/32x2/1/16	172/60/32x2/1/16	x/S	x/S	x/S
P1156	1.0mm	35x35mm	-	-	-	-	-	256/118/32x2/3/16	256/118/32x2/3/16	x/S	x/S	x/S

** subject to change

F100 (SIP) Flash + 256Mbit
A484 (SIP) 2Gb LPDDR4

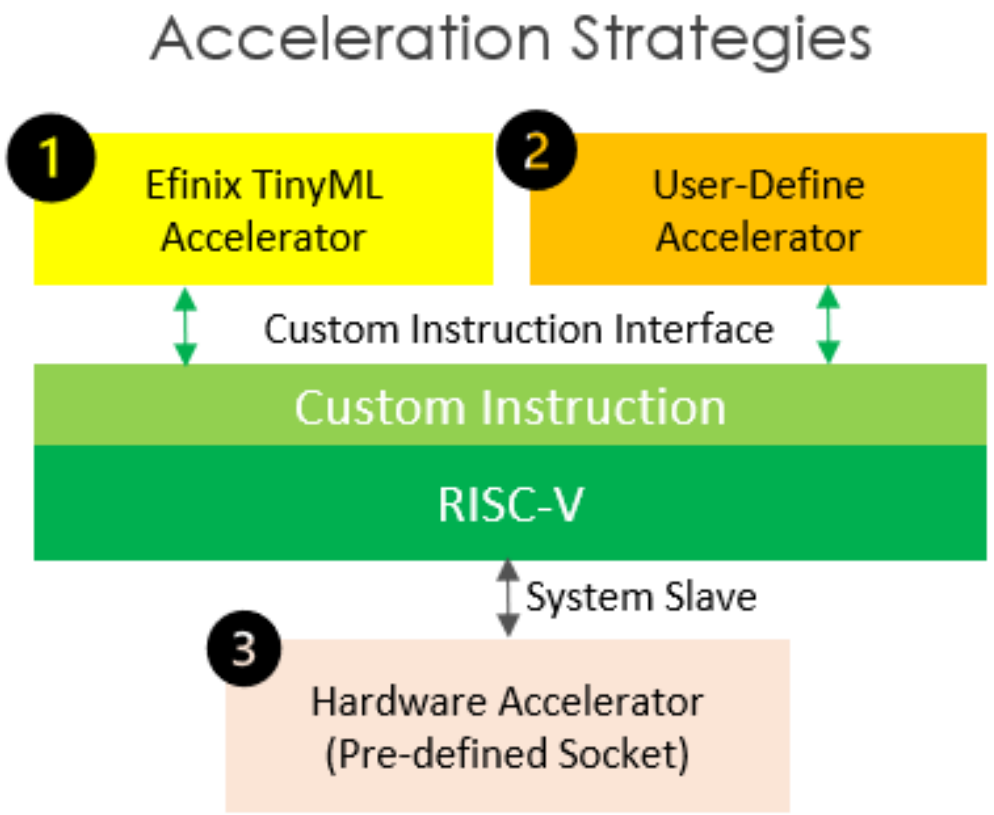
High Level Summary

Efinix offers a platform
- based on an open-source TensorFlow Lite Micro C++ library
- running on Efinix Sapphire RISC-V SoC
- with custom TinyML accelerator.



*TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

EFINIX TinyML Key Concepts



Advantages of Efinix TinyML Platform:

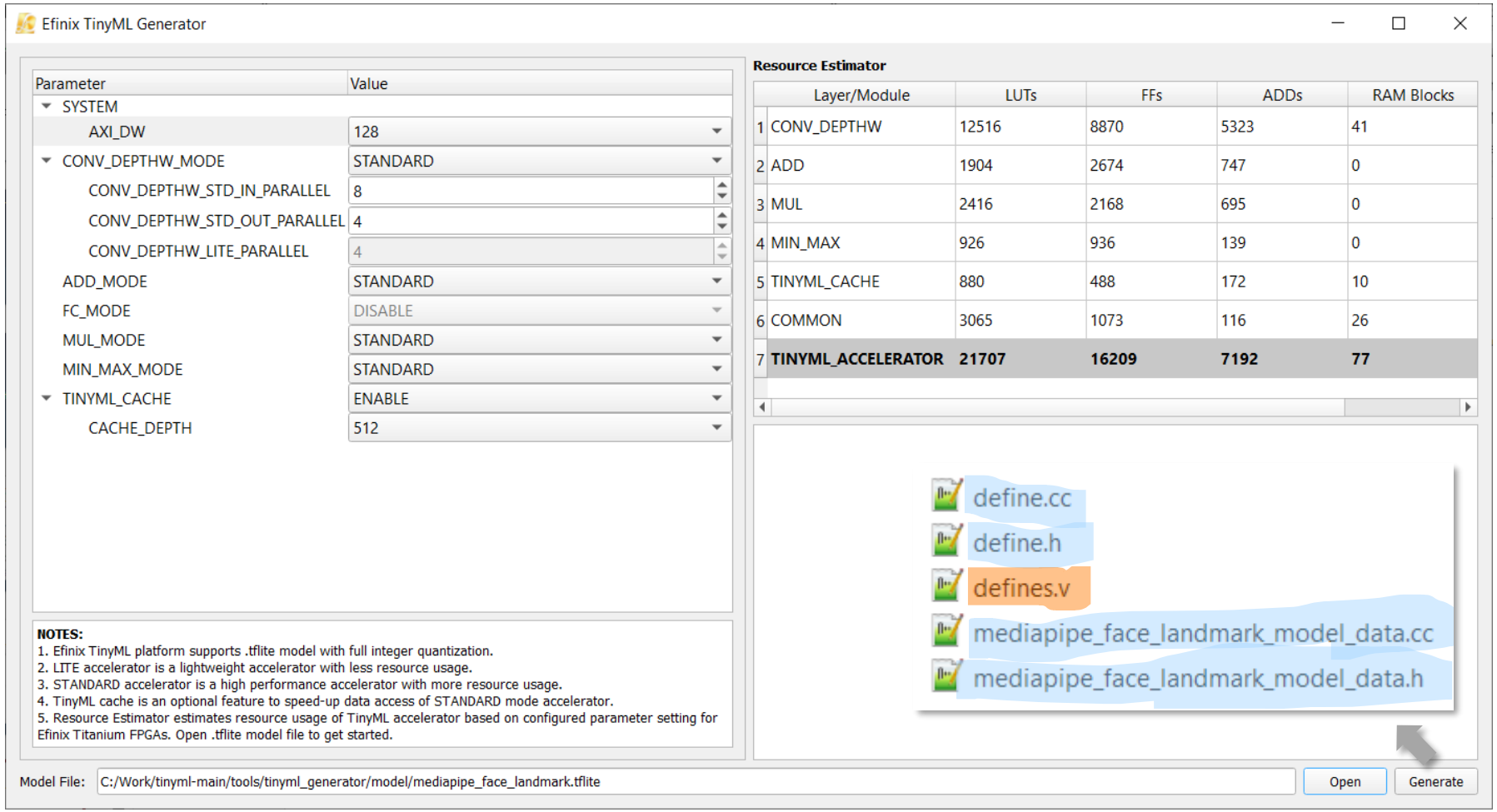
- All AI inferences supported by TFLite Micro library maintained by open-source community

Configurable RISC-V SoC
- Efinix TinyML Accelerator
- optional user-defined accelerator
- hardware accelerator socket

Multiple acceleration options
- performance-and-design-effort ratio
- speed-up overall deployment

*TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

Efinix TinyML Generator



Example: Face Landmark

- Reads tflite model files
- Select Acceleration Options
  - STANDARD / LITE / DISABLE
- Resource Estimator shows the expected utilization in the FPGA
- Generate project specific files to include in **FPGA project** and **RISC-V SW Project**
- **Acceleration achieved: x 20**
- **Power for AI < 500 mW**

Conclusion

- Standard Path
  - TensorFlow* – TensorFlow Lite* – TensorFlow Lite Micro Library
- Open RISC-V® Architecture
- Free Efinix TinyML Generator
- Free Efinity Software for FPGA Design Implementation
  - Free FPGA project examples for Quick Start on github
    - TinyML Hello World Design for Ressource Estimation
    - TinyML Vision for Real Time Application (Video In and Out)

- Free Efinix RISC-V® IDE to compile and run the RISC-V® code
  - Example projects and Model Zoo for Quick Start on github

*TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

References

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

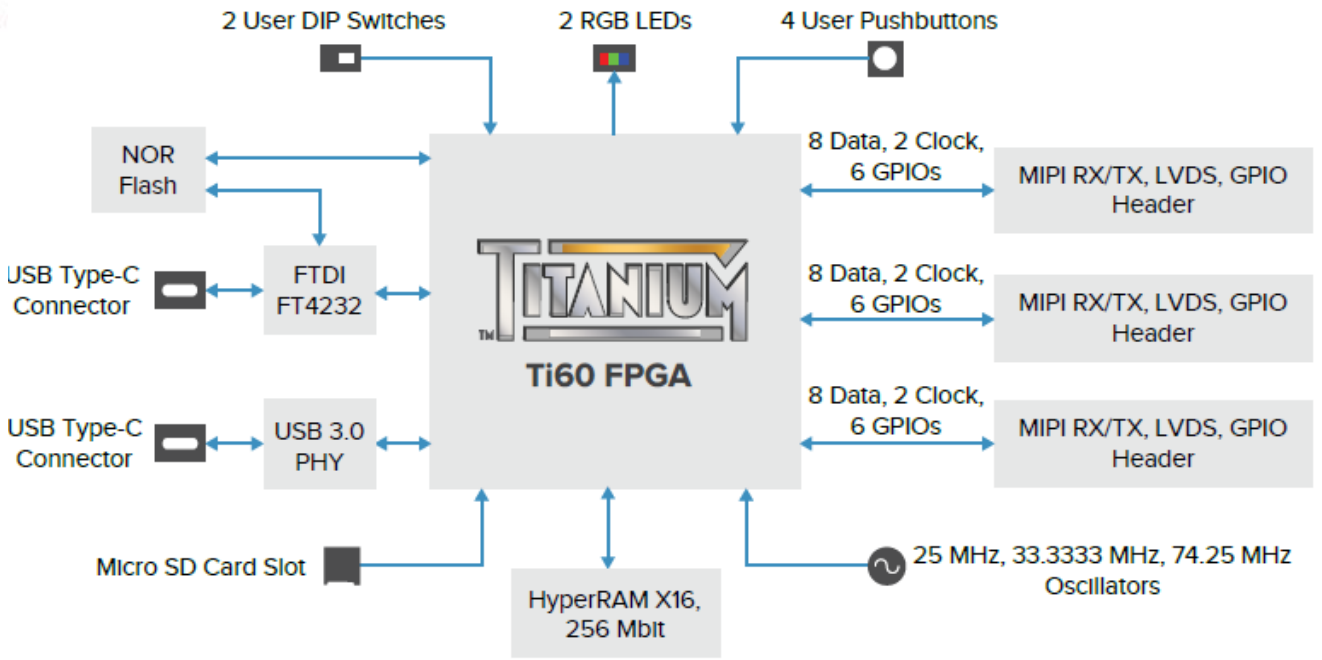
www.tensorflow.org/lite/microcontrollers/library
www.risc-v.org

Thank You!

Joachim Mueller
FAE Manager, Europe
joachimm@efinixinc.com

Supplemental Information

Development Kit Ti60F225 for TinyML Evaluation



Ti60F225

Development Kit Ti180J484C-DK for TinyML

Main Features:

- Efinix Ti180J484 FPGA in a 484-ball FineLine BGA package
- LPDDR4 or LPDDR4x 256 Mbit x16 bits memory:
  - Up to 2.0 Gbps double-data rate (LPDDR4x)
  - Up to 1.6 Gbps double-data rate (LPDDR4)
- Two 256 Mbit SPI NOR flash memories
- Four MIPI, LVDS, GPIO high-speed QSE connectors
  - to attach daughter cards included in the kit or your own custom cards
- Micro-SD card slot
- FPGA mezzanine card (FMC) with low pin-count connector (LPC)
- USB Type-C connector to configure the development board
- 33.33, 50, and 74.25 MHz oscillators for Ti180 PLL input
- User LEDs and switches:
  - 6 LEDs, 2 pushbutton switches

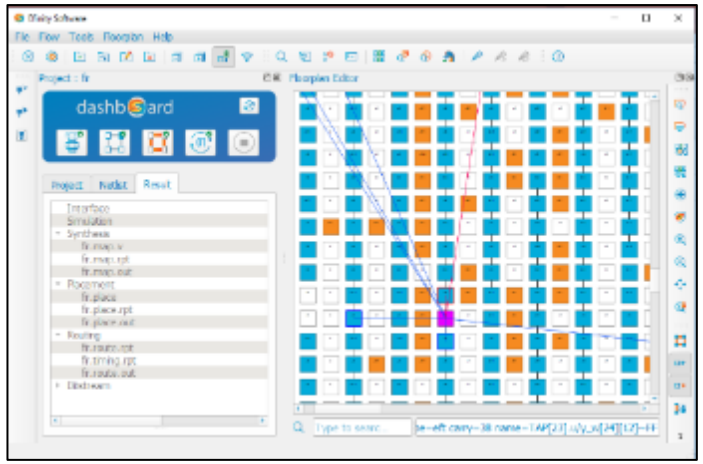
Ordering Code: TI180J484C-DK

Efinity® Integrated Development Suite

- Standard RTL-to-bitstream FPGA development tool (free of charge)
- Core and interface concept enables system integration
- Windows & Linux

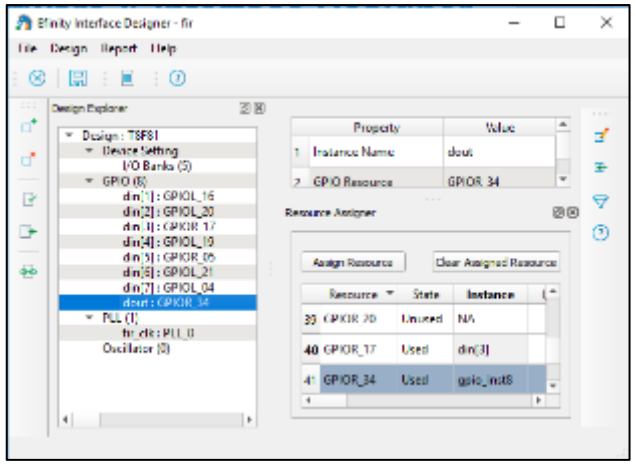
Core Designer

FPGA Core Fabric
(Synthesis, Place and Route)



Interface Designer

Subsystems
(I/O, DDR, MIPI, etc.)



Design Entry → Synthesis → Place & Route → Timing Analysis → Debugging → Bitstream Generation

Key Concepts

Explore Efinix Model Zoo + Post-Training Quantization

- Jupyter Notebook
- Google Colab
- TensorFlow Lite Converter

Explore Efinix TinyML Generator

- Generate model data files based on TensorFlow Lite model (.tflite)
- Customize Efinix TinyML Accelerator

Run TinyML Hello World on Efinix Development Board

- RISC-V with Efinix TinyML Accelerator
- Optional user-defined accelerator
- Pre-defined hardware accelerator socket
- Inference on static input data

Optimization Iterations

Explore Efinix Domain-Specific TinyML Framework

- RISC-V with Efinix TinyML Accelerator
- Optional user-defined accelerator
- Pre-defined hardware accelerator socket
- Domain-specific I/O support
- Inference on real-time captured data

Create Your Own TinyML Solution Using Efinix Design Flow

- Leverage on Efinix TinyML design flow, reference design, and framework to get started
- Quick deployment of TinyML solutions

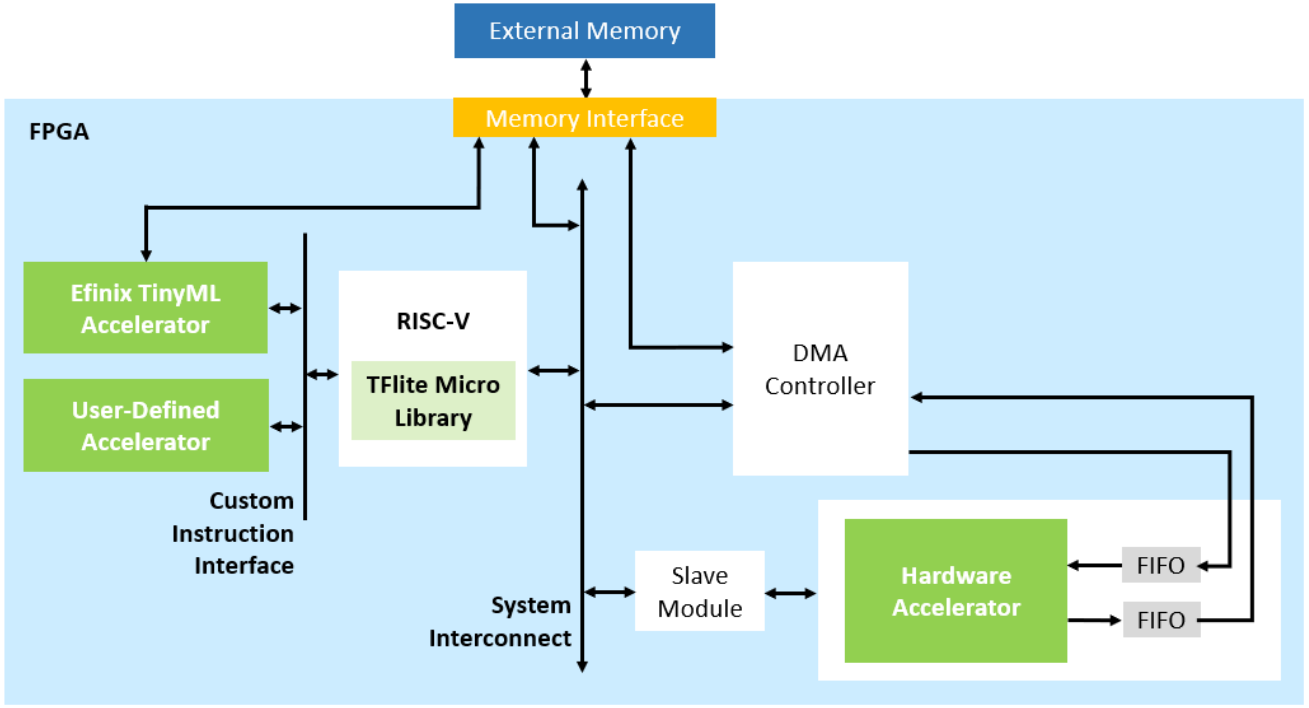
*TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

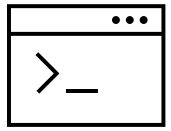
Model Zoo - Examples

Application	Framework	Trained Model format	Model	Input Size	Dataset	Quality Target (%)
Person Detection	Darknet	.cfg, .weights	Yolo	96x96x3	COCO(person)	20.30 (mAP@0.5)
Person Detection	Tensorflow	.pb	MobilenetV1	96x96x1	Visual Wake Words	84.0
Image Classification	Tensorflow	.h5	ResNet	32x32x3	CIFAR10	85.0
Keyword Spotting	Tensorflow	.pb	DS-CNN	49x10x1	Speech Commands	90.0
Face Landmark	Tensorflow	.pb	MediaPipe	192x192x3	Charade	468(3D Landmark)
Anomaly Detection	Tensorflow	.h5	Deep AutoEncoder	1x640	ToyADMOS	0.85 (AUC)

All examples are converted to tflite. For conversion flow details please refer to https://github.com/Efinix-Inc/tinyml/blob/main/docs/model_conversion.md

Model Evaluation with TinyML Hello World Design





Print Results to Terminal

TinyML Hello World Design is a FPGA Implementation for estimating performance enhancement:

- Hello World runs AI inference on static input data
  - No camera or display needed
  - Deploy to Ti60 or Ti180 Development Kit
  - Print information to a terminal

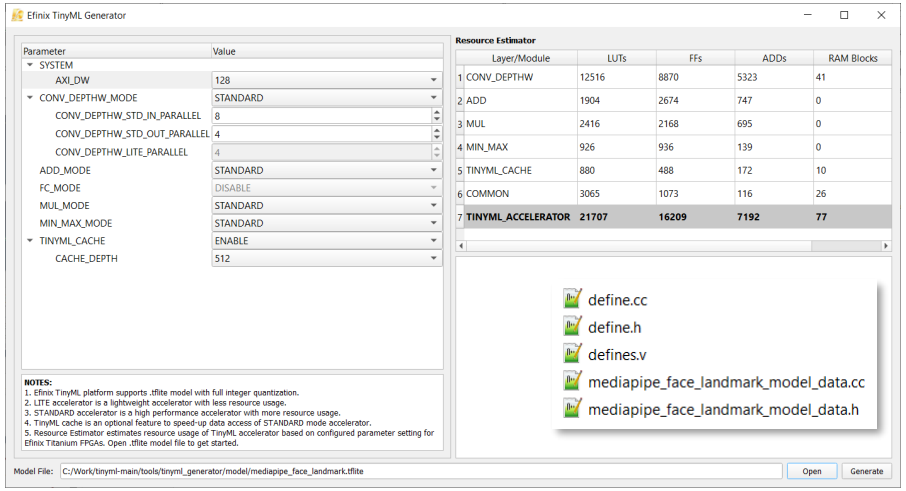
Evaluation Strategy:
For evaluation generate project files with all accelerators included (set STANDARD).

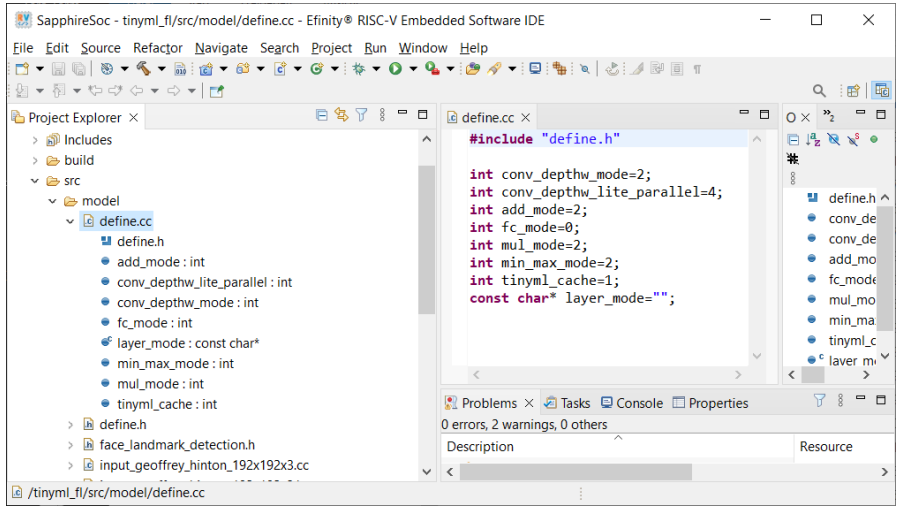
- Enable or disable accelerators in software
  - Quick iterations to compare cost and performance

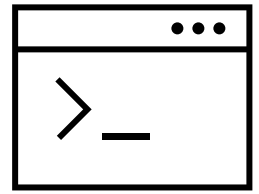
Provided Hello World Design includes the accelerators to support all examples in Efinix' Model Zoo (next page)

- Hex File for FPGA programming allows Quick Start
  - No need to implement the FPGA design for first steps

Evaluation Iterations







Acceleration options in TinyML Generator

- All „Standard" for max acceleration
- Implement FPGA (Hello World Design)
- Deploy Design to Development Kit
- Connect Terminal

Run Efinity RISC-V IDE Hello World Example

- Use define and data files created by TinyML Generator
- Load and Run the RISC-V application
- Set Accelerator on/off, recompile, run
  - Parameter per layer in define.cc

Print Results (Time) to Terminal

Face Landmark Example Evaluation Results

Minimum and maximum Inference Time

No accelerator enabled: 9105 ms

All accelerators in Standard Mode: 479 ms

Resource Estimator

	Layer/Module	LUTs	FFs	ADDs	RAM Blocks	DSP Blocks
1	CONV_DEPTHW	12516	8870	5323	41	48
2	ADD	1904	2674	747	0	8
3	MUL	2416	2168	695	0	7
4	MIN_MAX	926	936	139	0	0
5	TINYML_CACHE	880	488	172	10	0
6	COMMON	3065	1073	116	26	0
7	**TINYML_ACCELERATOR**	**21707**	**16209**	**7192**	**77**	**63**

TinyML Generator Resource Estimation (Face Landmark)

Performance Impact per Layer, Off versus Standard

Time	RefTime	Acceleration
7006 ms	479 ms	14.63
1415 ms	479 ms	2.95
842 ms	479 ms	1.76
1280 ms	479 ms	2.67
943 ms	479 ms	1.97

All numbers and estimations are snapshots and subject to change with versions and implementation options

Face Landmark Example Evaluation Results

Impact on dynamic power dissipation of RISC-V and TinyML Accelerators

RISCV + all Standard Accelerators for Face Landmark 498 mW

Same in Titanium Low Power Device 377 mW

Resource Estimator

	Layer/Module	LUTs	FFs	ADDs	RAM Blocks	DSP Blocks
1	CONV_DEPTHW	12516	8870	5323	41	48
2	ADD	1904	2674	747	0	8
3	MUL	2416	2168	695	0	7
4	MIN_MAX	926	936	139	0	0
5	TINYML_CACHE	880	488	172	10	0
6	COMMON	3065	1073	116	26	0
7	**TINYML_ACCELERATOR**	**21707**	**16209**	**7192**	**77**	**63**

TinyML Generator Output (Face Landmark)

Estimated per Module

Ti60	Ti60L
177 mW	135 mW
38 mW	28 mW
13 mW	8 mW

All numbers and estimations are snapshots and subject to change with versions and implementation options

Go Live with TinyML Vision FPGA Design

Based on Hello World Design estimation results and accelerator selection

- Generate files for target implementation (TinyML Generator)
- Build your custom FPGA design starting from TinyML Vision
  - TinyML Vision Example on github (Efinity Project)
    - Similar to Hello World Design, but
      - Prepared for Video input and output
      - Deploy to Ti60 or Ti180 ´Development Kit
        - Camera, Display or HDMI Video Adapter included
    - TinyML Vision is ready to use with models in Model Zoo
    - Adapt to your target configuration and to your target hardware

*TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

Questions?

joachimm@efinixinc.com