Delivering High Performance, Low Power Edge-AI Applications with the SiMa.ai ONE Platform

Edge AI and Vision Alliance Technical Webinar
October 2024

Agenda

- Why Edge AI?
- SiMa.ai MLSoC & Markets
- MLPerf Benchmarks
- MLSoC™ Hardware & Palette™ Software
- Developer Journey
- Model-SDK Walkthrough
- GStreamer Pipeline Demo
- Application Deployment Demo
- Palette™ Edgematic
- MLSoC™ Modalix
- Q & A

Why Edge AI?

Latency & Throughput

- Real time processing of time-sensitive data
- Performance per Watt

Privacy & Security

- Data and models kept local with full transparency & control
- Raw data analytics and raw data backups

Reliability & Cost

- Controlled and affordable allocation of resources & TCO
- Deterministic processing

Edge AI Challenges & Solution

ORIGIN

Initial compute built for data centers or servers with high performance. Power not a constraint for data centers.

Software rigidity, hand coded model development.

CHALLENGE

Platforms at the edge need high performance and low power.

Require software flexibility to enable ever changing models and customer pipelines and 10,000+ customers.

SiMa.ai MLSoC

Software Centric Purpose Built System On Chip for ML at the Edge.

Complete ML Application Pipeline Acceleration at the highest FPS / Watt.

SiMa.ai Focus: Embedded Edge AI

Active in Multiple Vertical Markets

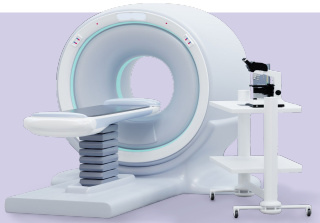
Smart Vision

Robotics + Industrial 4.0

ADAS

Government

Industrial Drones



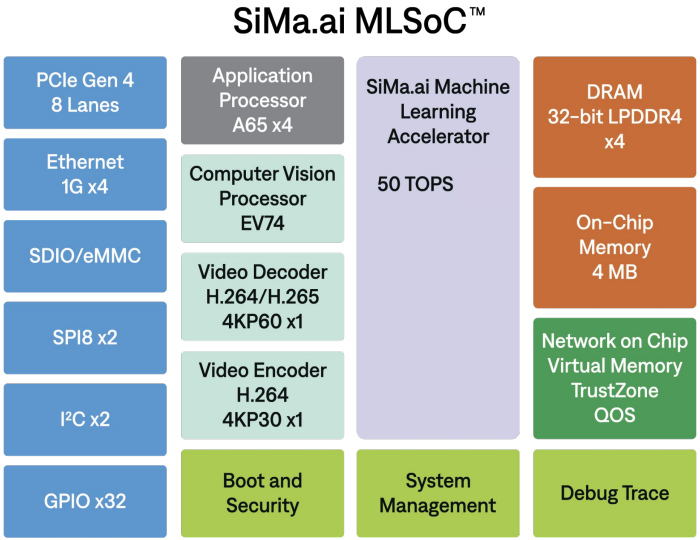
Healthcare

Agriculture

Pharma

MLSoC™ - Purpose-built for the Edge

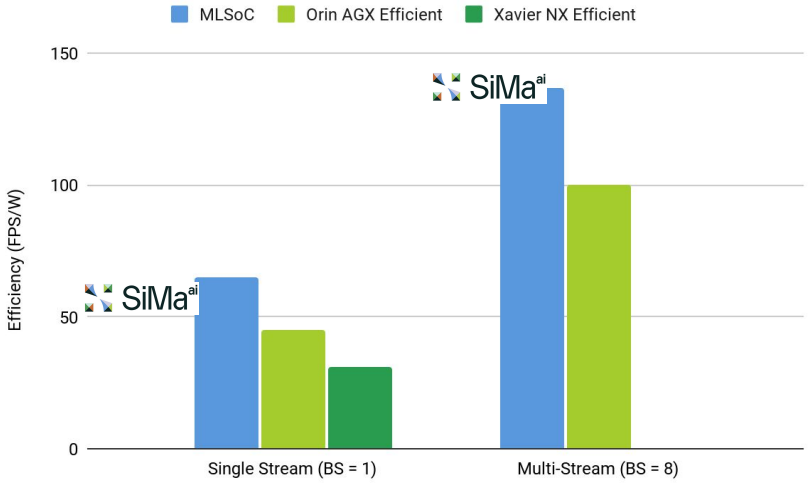
Palette™
ML SOFTWARE DONE RIGHT

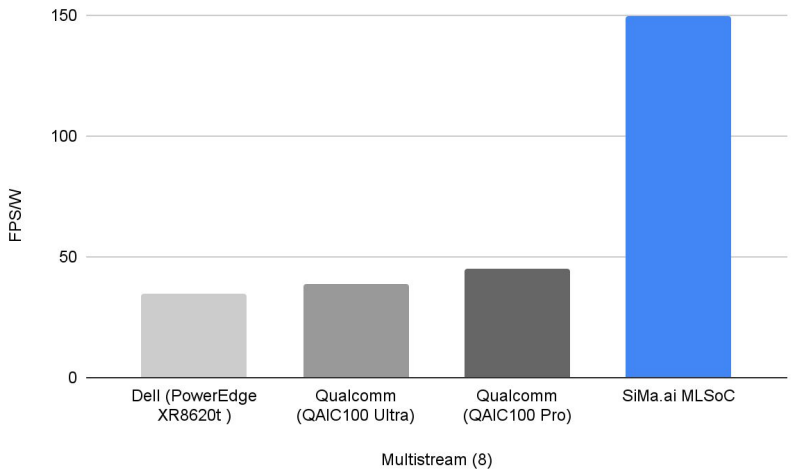


First **software-centric** purpose-built MLSoC that runs end to end edge ML applications

MLPerf: SiMa.ai delivers advantage over incumbent

*SiMa.ai MLSoC (N16) compiled results beats Orin (8nm) on both **performance and power in the closed edge category***

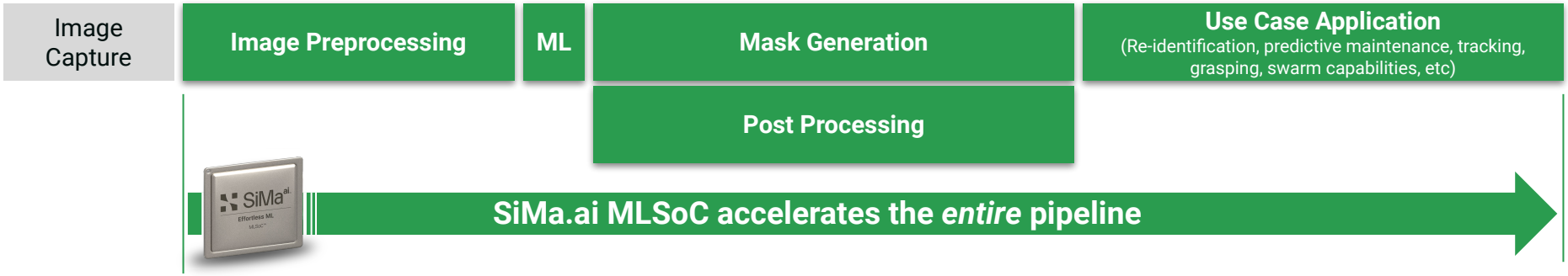


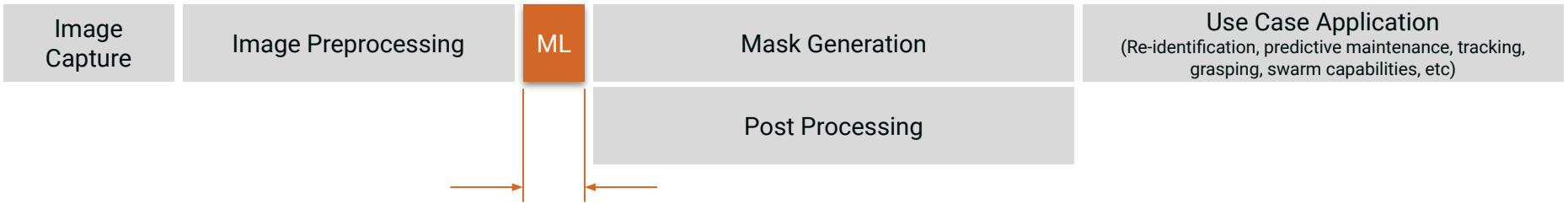


1 Camera
1.4x better performance vs Orin
2.1x Xavier

8 Cameras
1.37x better performance vs Orin
Xavier data not published

Accelerating the **entire** pipeline and application on a **single** chip

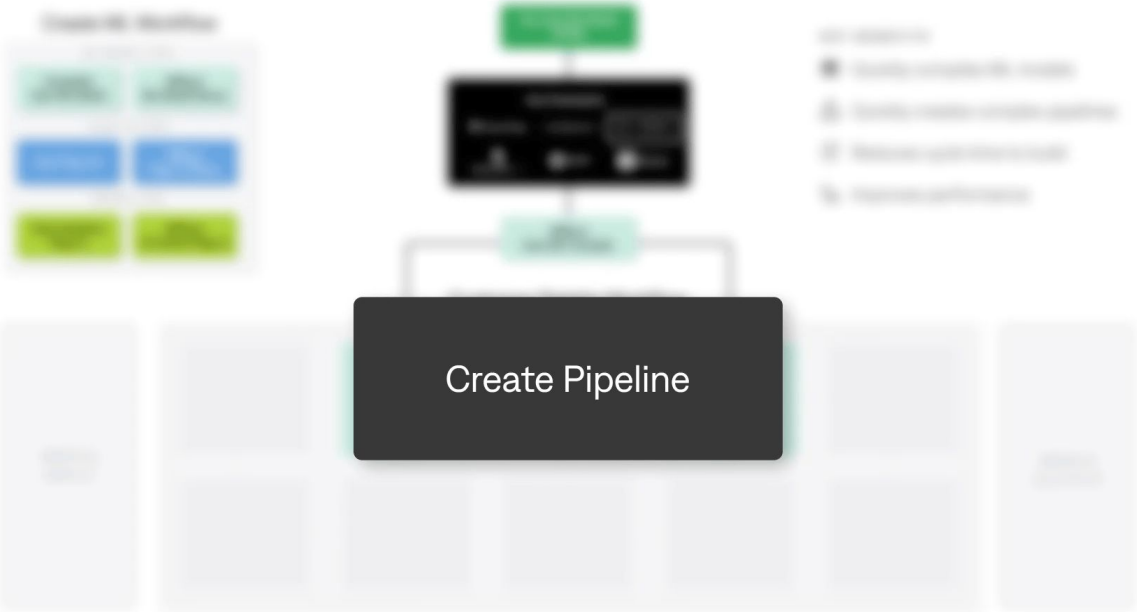




Competing ML accelerators only support **this** part of the pipeline

Accelerate entire application with Palette™

ML SOFTWARE DONE RIGHT



Palette™ - Software Platform

Containerized Development Package

CLI and Benchmarking

Create | Remote FW Update | Manage Devices | Deploy

...

GStreamer Optimized Plugins

Machine Learning, Computer Vision and DSP, HW Encode/Decode

Develop on Linux and Windows

C++ Co-Processing APIs

Palette™

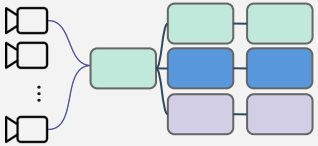
ML SOFTWARE DONE RIGHT

Easy install

Examples and tutorials

Patented SW Static Scheduling

Multiple input streams Multiple models

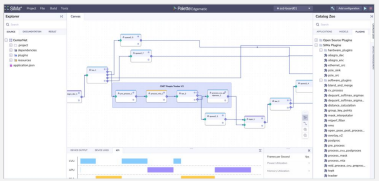


ModelSDK

Python APIs

Quantize, Calibrate, Test & Compile

Palette Edgematic



CNNs, Transformers, LLMs & LMMs

Scalable Architecture

Deployment Modes

MLSoC Standalone

MPK Application

SiMa.ai Application Development Interfaces

GStreamer

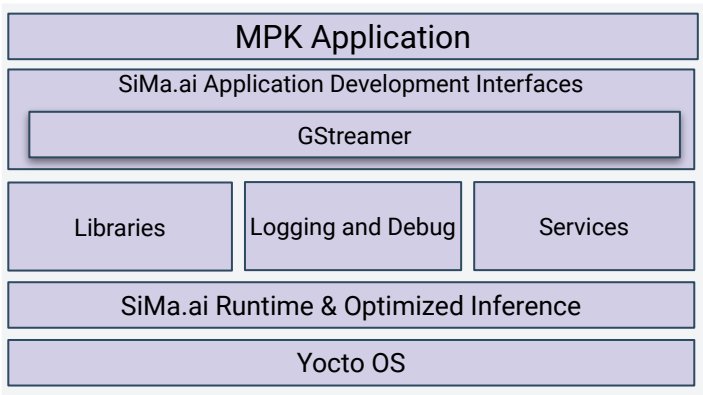
Libraries | Logging and Debug | Services

SiMa.ai Runtime & Optimized Inference

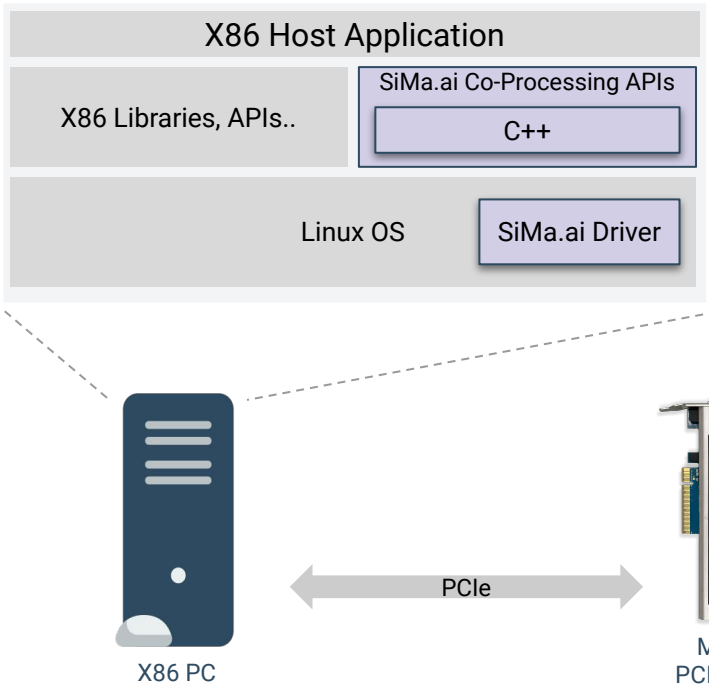
Yocto OS

Deployment Modes

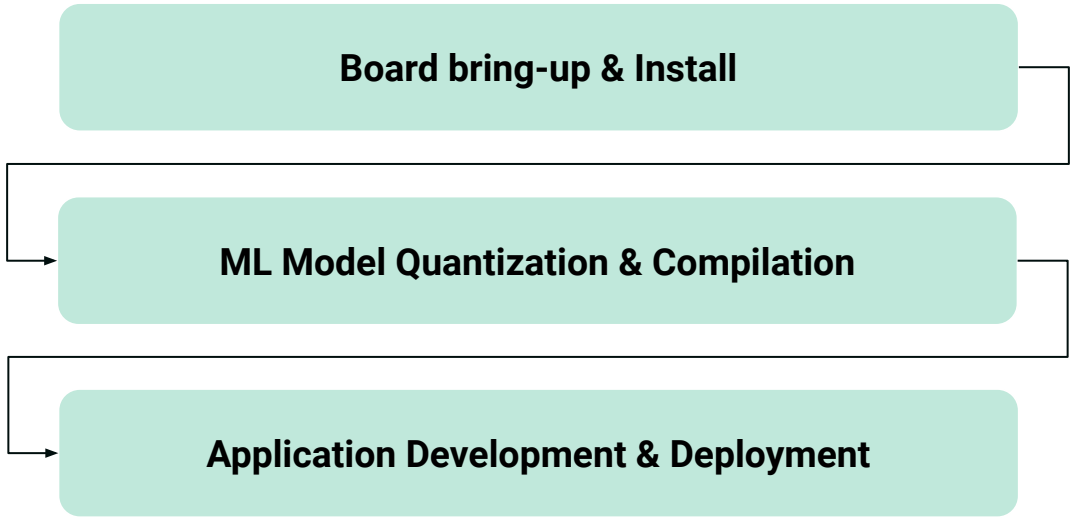
MLSoC Standalone



MLSoC Co-Processor

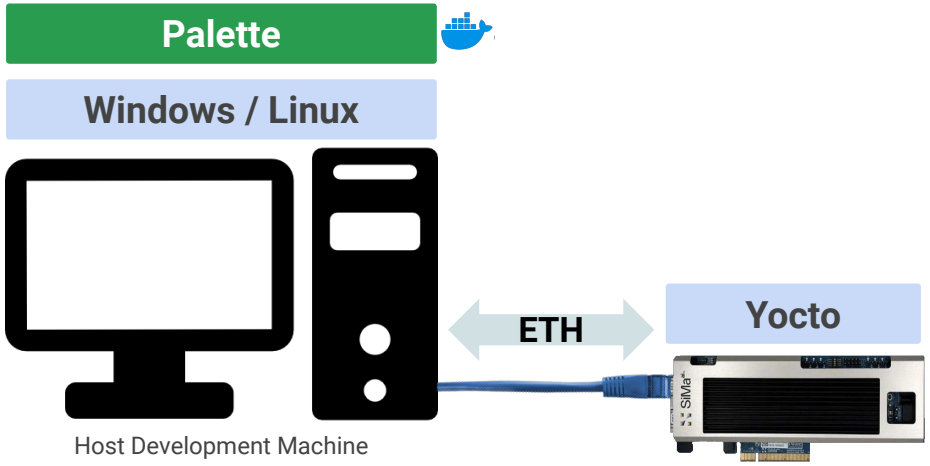


Development Journey Overview

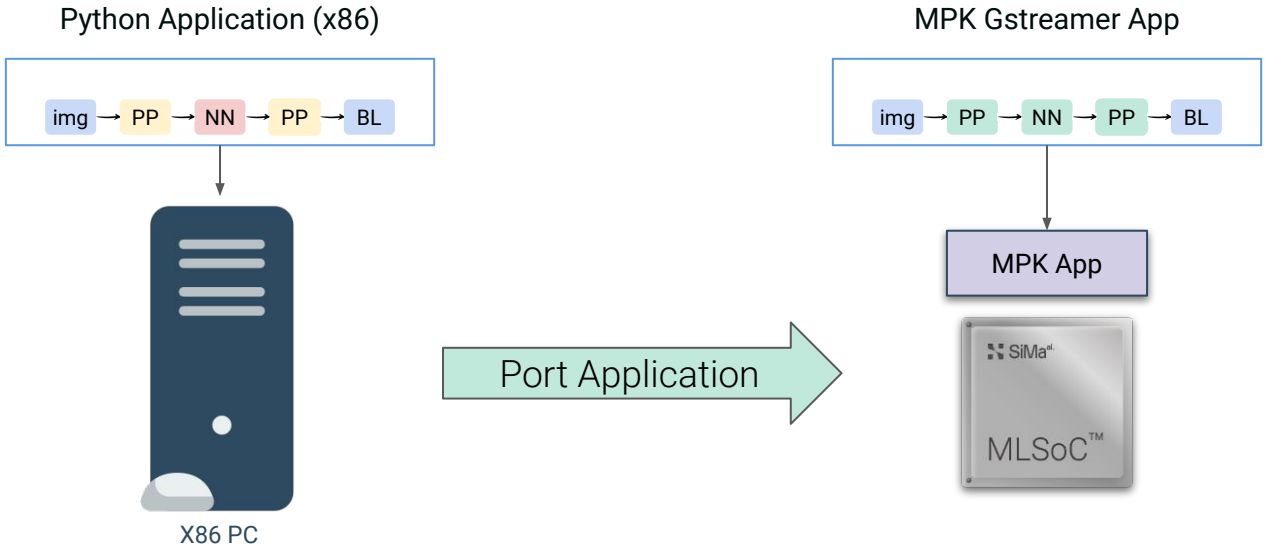


Development Journey Overview

Board bring-up & Install



Reference Application



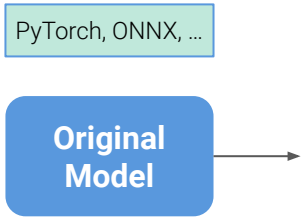
ONNX RUNTIME

OpenCV

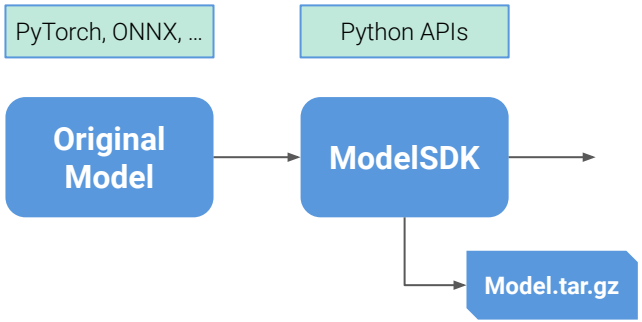
gstreamer

SiMa

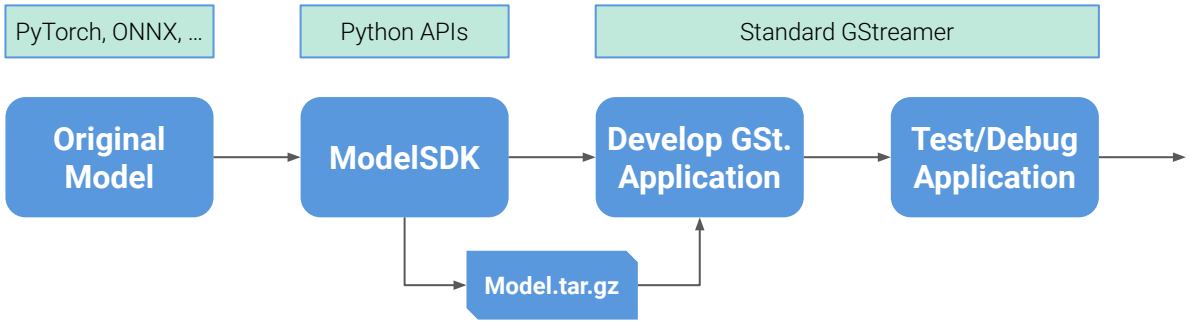
Application Development Journey



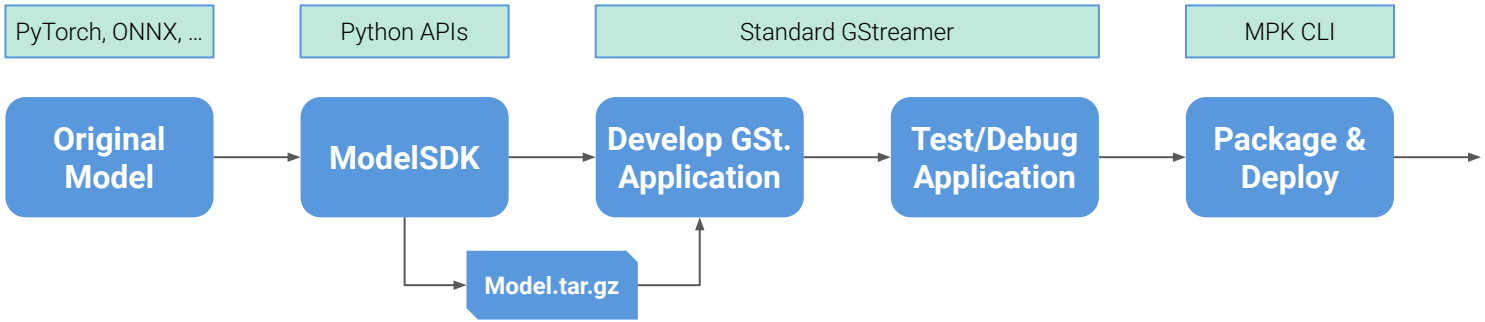
Application Development Journey



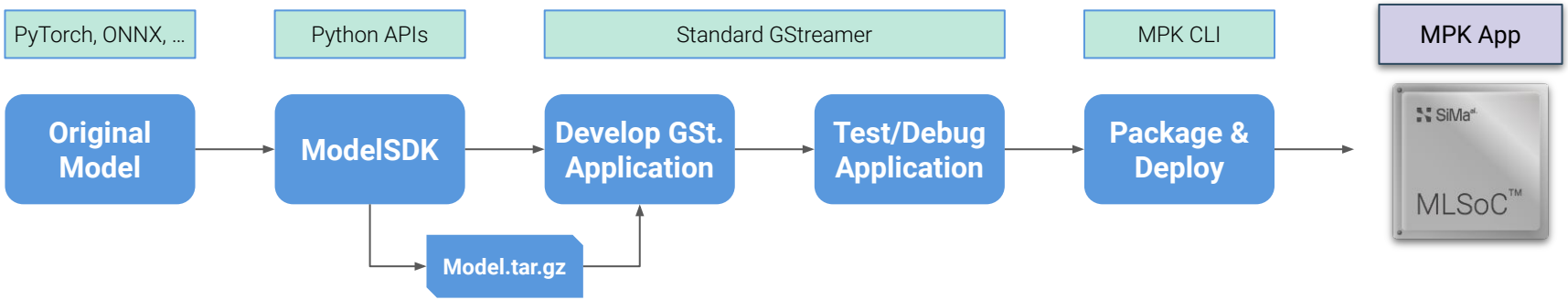
Application Development Journey



Application Development Journey



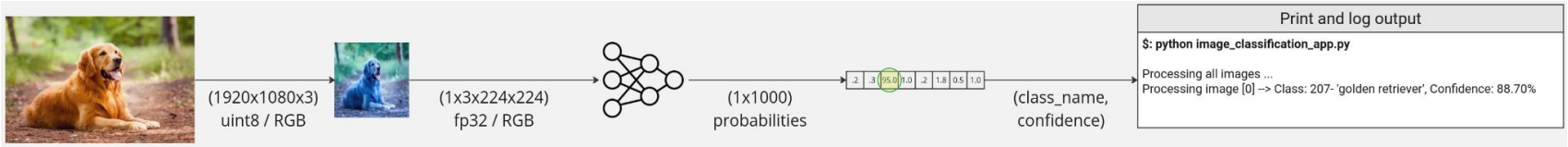
Application Development Journey



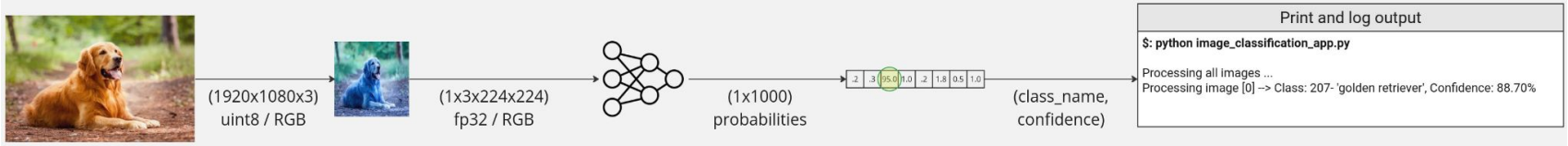
Typical ML Application

Typical ML Application

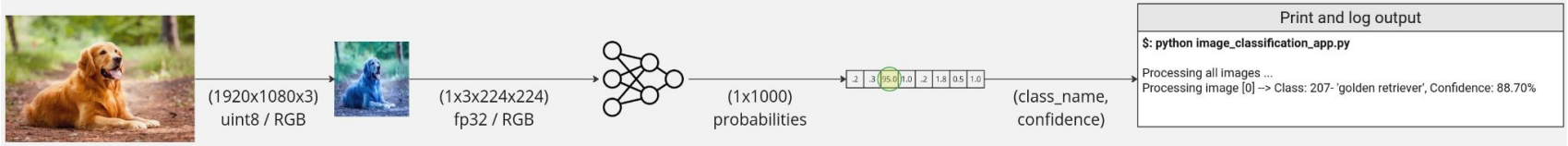
Typical ML Application

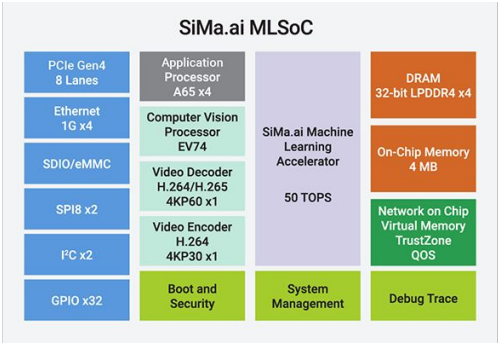


Typical ML Application

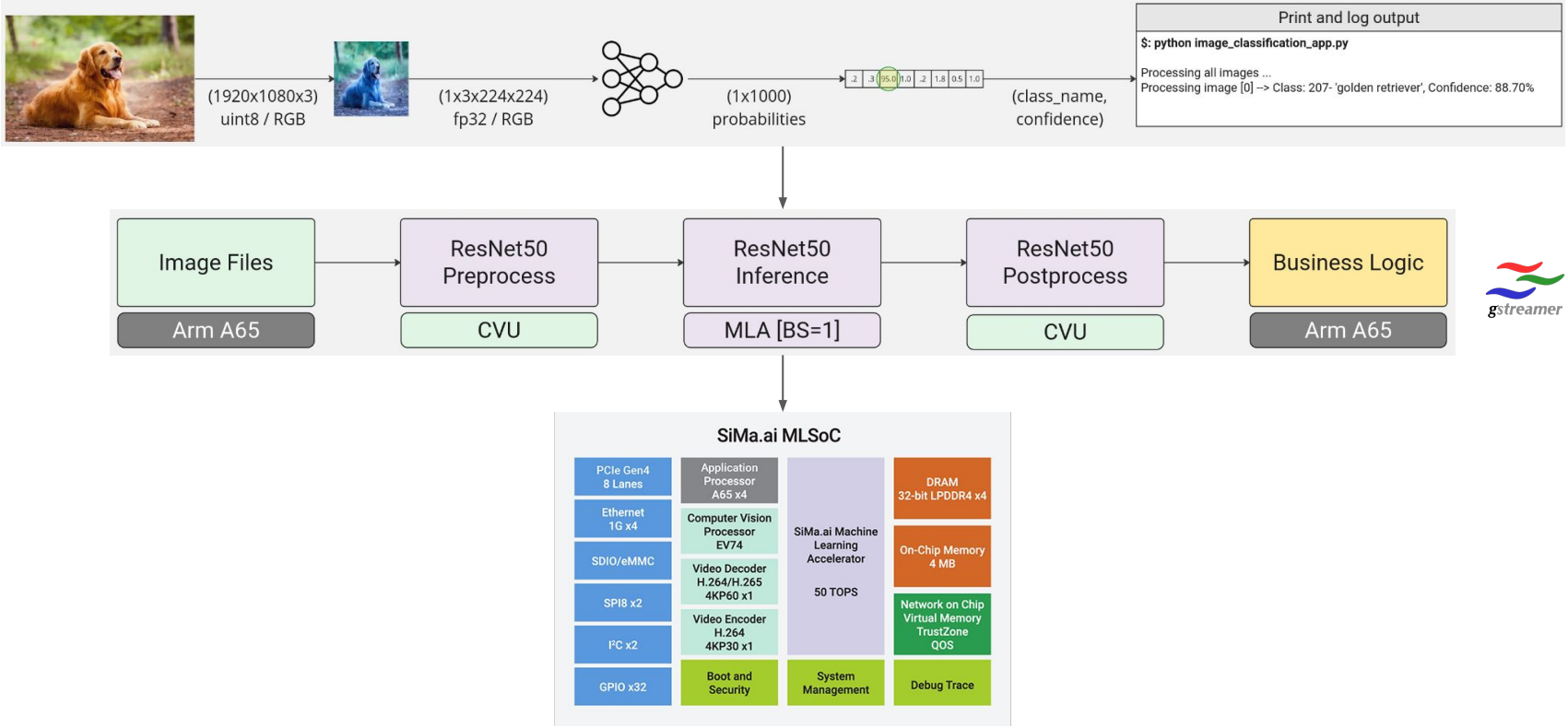


Typical ML Application





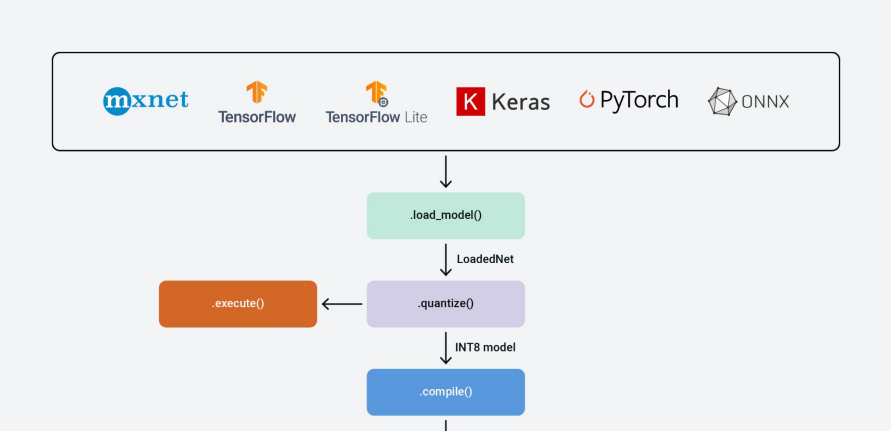
Typical ML Application



Model-SDK Introduction & Demo

The main functions of the Model-SDK are:

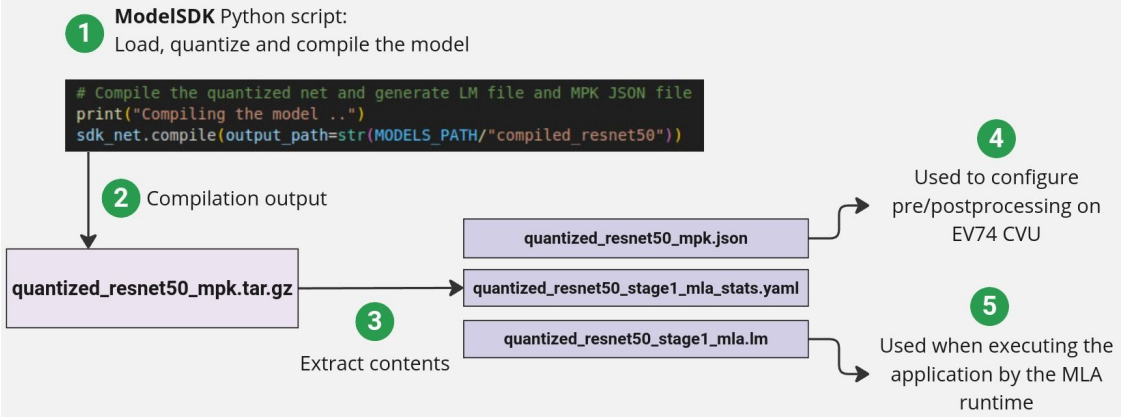
- **Loading** a floating-point model from one of the supported frameworks and formats - (PyTorch, Keras HDF5, TensorFlow, ONNX, and TFLite).
- **Quantizing** the floating-point model into an 8-bit integer (int8) format or a 16-bit (int16) format.
- **Evaluating** the quantized model.
- **Compiling** the quantized model into an inference model that can be executed on an MLSoC device.



Model-SDK Introduction & Demo

The output of the Model-SDK:

- **Compressed** package that contains:
  a. Metadata JSON (.json)
  b. Statistics file (.yaml)
  c. The compiled model (.lm)



Palette - Model SDK Demo

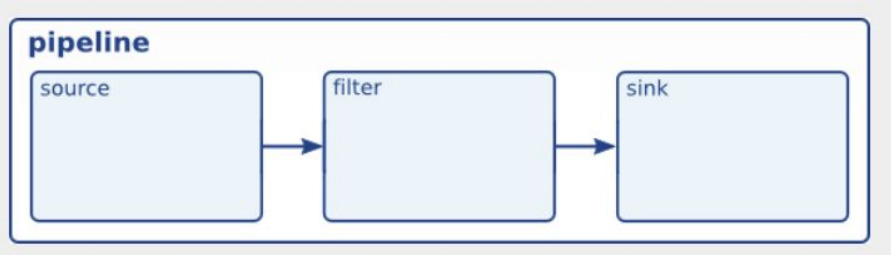
Code from (requires developer access):

[ModelSDK - Compiling ML Models — Documentation documentation (sima.ai)]

GStreamer Introduction

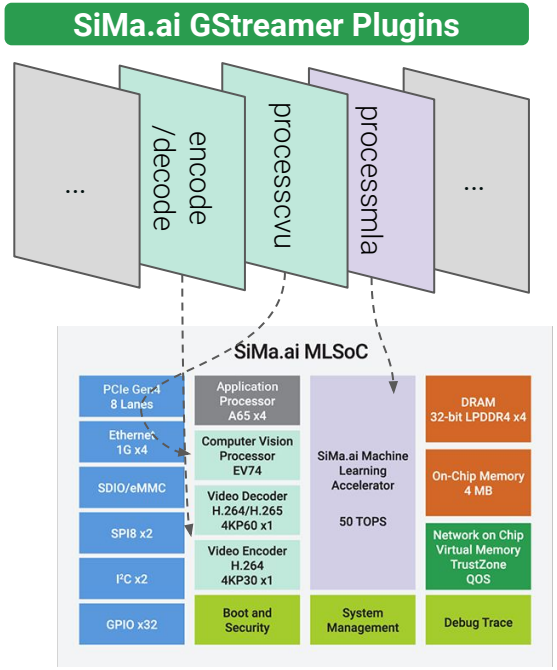
GStreamer is a pipeline-based multimedia framework that links together a wide variety of media processing systems to complete complex workflows.

- **Modular Design:** GStreamer plugin architecture allows easy integration of different processing elements.
- **Flexibility:** Supports a wide range of input and output formats, making it versatile for various ML tasks.
- **Efficiency:** Optimized for performance, making it suitable for real-time applications on resource-constrained devices.
- **Scalability:** Can handle complex workflows, enabling the construction of advanced ML pipelines with ease.

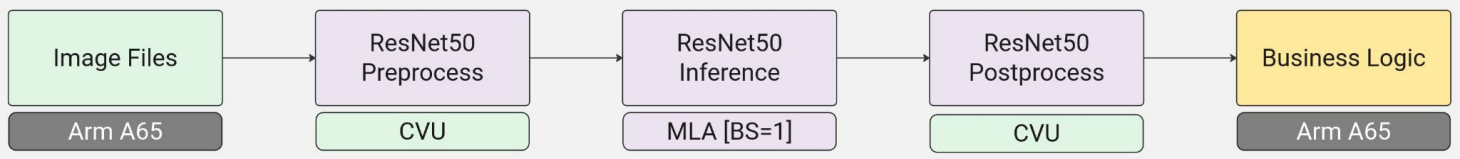


SiMa.ai GStreamer Plugins

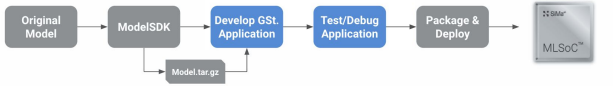
SiMa.ai provides a **library of GStreamer plugins and their source code** that aid developers in constructing an optimized end-to-end application on SiMa's MLSoC



GStreamer Application Development & Demo




```
export LD_LIBRARY_PATH="${SIMA_PLUGINS_DIR}"
gst-launch-1.0 -v --gst-plugin-path="${SIMA_PLUGINS_DIR}" \
simaaisrc mem-target=1 node-name="my_image_src" location="${SAMPLE_IMAGE_SRC}" num-buffers=1 ! \
simaaiprocesscvu source-node-name="my_image_src" buffers-list="my_image_src" config="${PREPROC_CVU_CONFIG_JSON}" ! \
simaaiprocessmla config="${INFERENCE_MLA_CONFIG_JSON}" ! \
simaaiprocesscvu source-node-name="mla-resnet" buffers-list="mla-resnet" config="${DETESSDEQUANT_CVU_CONFIG_JSON}" ! \
argmax_print config="${ARGMAX_PRINT_CONFIG_JSON}" ! \
fakesink
```

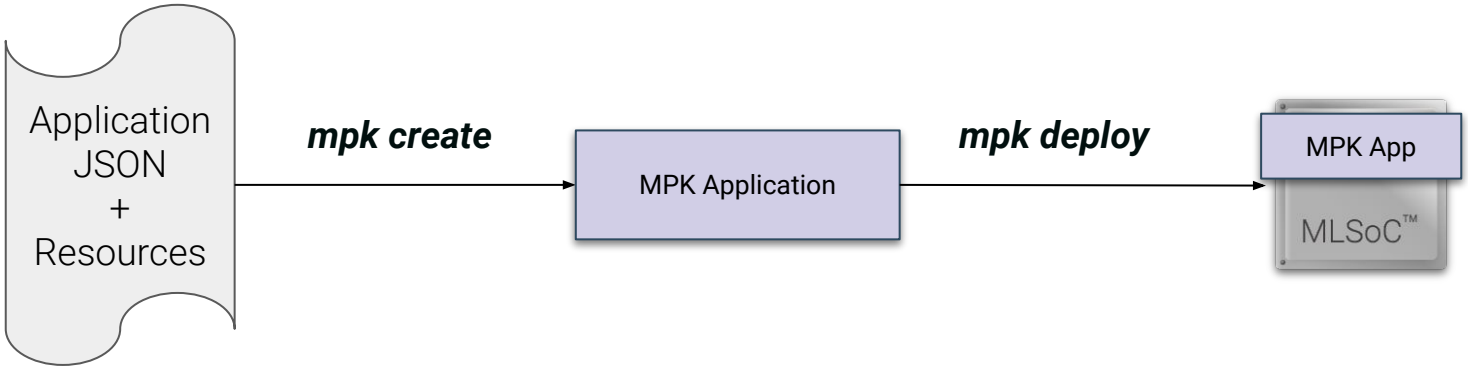


Palette - GStreamer Demo

Code from (requires developer access):
Developing End-to-End Applications on MLSoC (GStreamer)

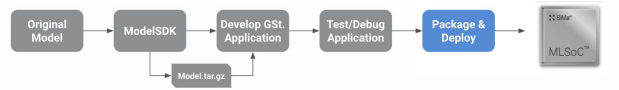
MPK Packaging & Deploy Demo

An **mpk application** is used to conveniently **package and deploy** an application to MLSoC target devices using MPK CLI tools included with Palette software. An MPK Application consists of a package that configures, and executes a GStreamer application at runtime.



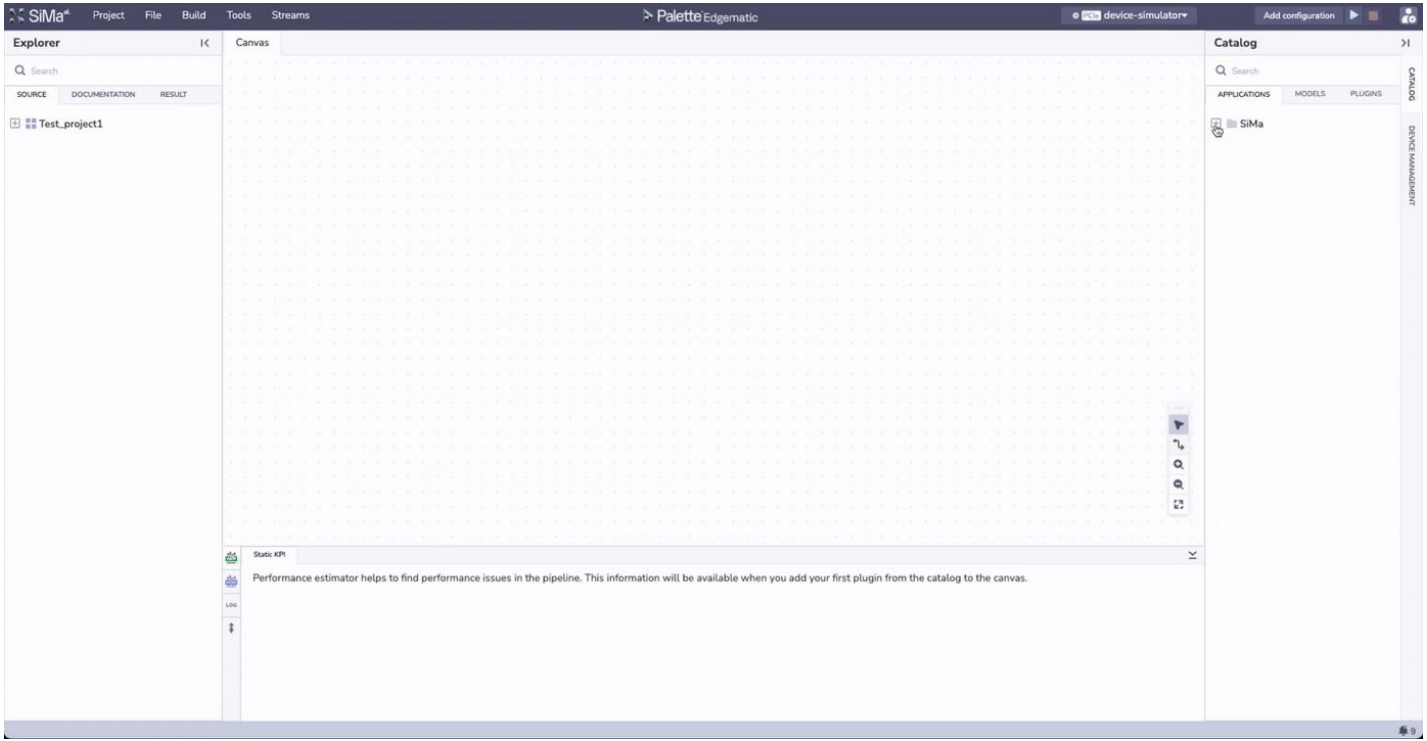
Palette - MPK Packaging Demo

Code from (requires developer access):
Developing End-to-End Applications on MLSoC (MPK)



Palette™ - Edgematic

Evaluate and run **complete** pipelines from inside Edgematic



Introducing SiMa.ai MLSoC™ Modalix

ONE Platform to support CNNs, Transformers, LLMs, LMMs and GenAI at the Edge

Modalix: Breakthrough Multi-Modal Edge AI Product Family

First Multi-modal edge SoC

MLSoC Modalix delivers GenAI at the edge at the highest performance/watt

Enhanced Processor and Peripherals

Beefed up processor complex and memory. On-chip ISP pipeline and MIPI, supporting RAW including true-color and infrared

ONE Platform for complete ML pipelines

Seamless software support across MLSoC & MLSoc Modalix simplifies development and optimizes TCO

Range of TOPs

25-200 TOPs device family in commercial and industrial grades optimizes performance and cost

Improved power efficiency

Multiple voltage domains further reduce power consumption by turning off unused features

GenAI and beyond

Hardware innovations in on-chip Machine Learning Accelerator power LLM, LMM and legacy CNN computations

ONE Platform for Edge AI: Performance per Watt, **ANY** Model, **ANY** Modality

Multi-modal

MANY modalities, models >10x in FPS/W

Edge AI your way

TOPs, power, modality, framework

Palette

Streamlined Edge-ML Deployment

The Edge AI and Vision Alliance is a partnership of ~100 leading edge AI and vision technology and services suppliers, and solutions providers

The Alliance provides high-quality technical educational resources for product developers

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

The Alliance enables edge AI and vision technology providers to grow their businesses through leads, partnerships, and insights

For membership, email us: membership@edge-ai-vision.com

Join us at the Embedded Vision Summit
May 20-22, 2025—Santa Clara, California

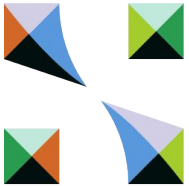
The only industry event focused on practical techniques and technologies for system and application creators

- *"Awesome! I was very inspired!"*
- *"Fantastic. Learned a lot and met great people."*
- "*Wonderful speakers and informative exhibits!*"

Summit highlights:

- **Inspiring keynotes** by leading innovators
- High-quality, practical **technical, business and product talks**
- Exciting **demos**, **tutorials** and **expert bars** of the latest applications and technologies

Visit www.embeddedvisionsummit.com for updates



SiMa ai™

Thank You!
Q & A

Learn more at: https://sima.ai/mlsoc/
Blogs: https://sima.ai/blog/