Machine Learning System on Chip MLSoC™ Modalix Product Family

Product Brief (PRELIMINARY)

Overview

Modalix family is the first multimodal machine learning system-on-chip (MLSoC) capable of executing generative artificial intelligence (GenAI) inference, computer vision, and machine learning (ML) inferencing in complex pipelines orchestrated by the on-chip application (APU) and computer vision unit (CVU). The Modalix family is available in several configurations to deliver 25 to 200 TOPS in an incredibly compact and low-power 25mmx25mm 1369 ball FCBGA package.

Each Modalix SoC includes on-die LPDDR5* memory interfaces, multiple 10G ethernet, multiple MIPI digital camera interfaces, PCIe, video encode-decode, and security blocks. The modules are brought together with an internal and secure network on chip (NoC).

Building AI solutions with Modalix is simplified using the SiMa.ai ONE Platform including the Palette software. Palette integrates the programming of the APUs, CVU, and ML accelerators into complete pipelines to target applications from computer vision in industrial automation to drones, robotics, and autonomous vehicles, and incorporating next-generation models such as transformer-based, LLM, LMM, and GenAI.
Highlights

Compute Engines

- Application Processor Unit (APU)
- Computer Vision Unit (CVU)
- Machine Learning Accelerator (MLA)
- Image Signal Processor (ISP)

Application Development

- Supports generative AI use-cases with high-performance, low-power, safe and secure ML inferencing
- Best-in-class multimodal model interference efficiency
- Supports any ML framework (PyTorch, ONNX, Keras, TensorFlow, etc.)
- Innovative Palette™ Software suite for ease of development, covering the entire product life-cycle from model optimization to application development and deployment.
- Comprehensive support for Models and Plug-ins
- Secure boot and trusted execution environment

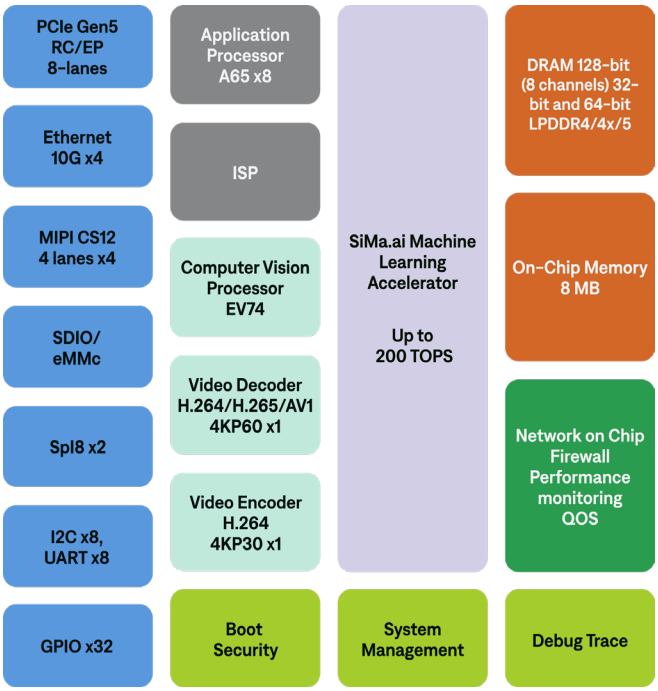
Peripherals

- MIPI CSI2 in 4 x 4 configuration
- 4 x 10Gb Ethernet
- 8 x PCIe Gen5 root-complex and endpoint

Target Industries

- Smart vision
- Drones
- Robotics (including AGV and AMR)
- Industry 4.0
- Automotive
- Smart Retail
- Healthcare
- Military & Government

Modalix chip form-factor: FCBGA 1369 balls; 25mm x 25mm



MLSoC Modalix Functional Blocks and Features

The MLSoC Modalix contains the following high-level functions:

- Machine Learning Accelerator (MLA) - provides 25-200 Tera Operations per Second (25-200 TOPS) for neural network computation and enhanced hardware for faster GenAI computations with increased accuracy, BF16 in hardware, improved DMA bandwidth and dual-voltage support.
- Application Processing Unit (APU) - a cluster of eight Arm Cortex A65 dual-threaded processors operating at 1.5 GHz to deliver up to 32k Dhrystone MIPS.
- Video encoder/decoder - supports the MJPEG, H.264 and H.265 compression standards, AOMedia Video 1 (AV1) with support for main/high/professional profiles, 4:2:0 pixels and 8-bit precision. The encoder supports H.264 at rates up to 4KP30 while the decoder supports H.264/265 at rates up to 4KP60.
- Computer Vision Unit (CVU) - consists of a 1GHz four-core Synopsys ARC EV74 video processor supporting up to 720 16-bit GOPS.
- Image Signal Processor (ISP) - Arm C-71 running at 1.2GHz. RAW 8, 10, 12, 14, 16, 20, 22, and 24-bit inputs from CFA image sensor. Supports RGGB, RCCG, RCCB, RCCC, and RGBIR color formats. Supports 24-bit Wide Dynamic Range (WDR).
- High-speed I/O subsystem - provides four 10-Gigabit Ethernet ports plus a PCIe Gen5 8-lane interface usable as root-complex or endpoint, and with bifurcation capability.
- DRAM interface system (DIS) - supporting eight 32-bit LPDDR5, LPDDR4 and LPDDR4x with support for x32 and x64 LPDDR chips. Target speed 6400 Mbps (LPDDR5) providing an effective theoretical bandwidth of 102 GB/s across all the DDR channels.
- Boot and security unit (BSU) - provides secure key storage in an eFuse memory, and key management. Supports decrypting and authentication of the boot image as well as providing a security API to the user code.

Software-First Development Environment

Compiling an ML-trained model to target particular hardware can be challenging if the software toolchain and hardware are not co-designed. SiMa.ai software-first approach includes carefully defined intermediate representations (including TVM Relay IR), along with novel compiler optimization techniques. This software architecture enables SiMa.ai to support a wide range of frameworks (e.g., TensorFlow, PyTorch, ONNX, etc.), and compile over 250+ models, thus providing customers with an effortless experience and world-class performance-per-watt results. SiMa.ai Palette™ software runs seamlessly on the MLSoC and MLSoC Modalix.

About SiMa.ai

SiMa.ai is the software-centric, embedded edge machine learning system-on-chip (MLSoC) company. SiMa.ai's hardware to software stack flexibly adjusts to any framework, network, model, sensor, or modality all in ONE Platform. Edge ML applications that run completely on the SiMa.ai MLSoC see a tenfold increase in performance and energy efficiency, bringing higher fidelity intelligence to ML use cases spanning computer vision to generative AI, in minutes. With SiMa.ai, customers unlock new paths to revenue and significant cost savings to innovate at the edge across industrial manufacturing, retail, aerospace, defense, agriculture, and healthcare. SiMa.ai was founded in 2018, has raised $270M and is backed by Fidelity Management & Research Company, Maverick Capital, Point72, MSD Partners, VentureTech Alliance and more. For more information, visit www.SiMa.ai

SiMa Technologies, Inc.
333 West San Carlos St,
Suite 1100
San Jose CA 95110.
mlsoc@sima.ai

SiMa.ai India Private Limited
Bagmane Tech Park Unit 02
2nd Floor, B Wing, Laurel Building
C V Raman Nagar, Bengaluru, Karnataka - 560093