SiMa.ai Machine Learning System on Chip

MLSoCTM Modalix Product Family | Product Brief

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 a custom highly programmable ML accelerator assisted by the on-chip Arm application processor (APU) and a Digital Signal Processor (DSP) which serves as a Computer Vision Unit (CVU). The Modalix chip delivers 50 Tera Operations per Second (TOPS) in an incredibly compact and low-power 25mm x 25mm 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 that support applications ranging from computer vision in industrial automation to drones, robotics, and autonomous vehicles, while also enabling next-generation models such as transformer-based architectures, LLMs, LMMs, and GenAI. **HTEC's deep expertise in AI/ML engineering and embedded edge systems played a key role in helping SiMa.ai refine and optimize these capabilities, ensuring the delivery of a streamlined and highly efficient development experience for customers.**

Highlights

Compute Engines

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

Peripherals

- MIPI CSI-2 in 4 × 4 configuration
- 4 × 10Gb Ethernet
- 8 x PCIe Gen5 root-complex and endpoint

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 PaletteTM 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

Target Industries

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

MLSoC Modalix Functional Blocks and Features

The MLSoC Modalix contains the following high-level functions:

Machine Learning Accelerator (MLA)

Provides 50 Tera Operations per Second 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 1.4 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. Both encoder and decoder support H.264/265 at rates up to 4KP60 × 1.

Computer Vision Unit (CVU)

Consists of a 1 GHz four-core Synopsys ARC EV74 video processor supporting up to 720 16-bit GOPS.

Image Signal Processor (ISP)

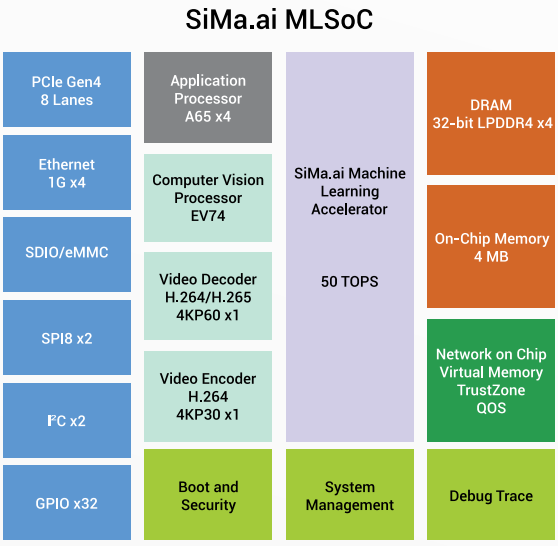
Arm C-71 running at 1.2 GHz. RAW 8, 10, 12, 14, 16, 20, 22, and 24-bit inputs from CFA image sensor. Supports RGGB, RCCG, RCCB, RCCC, and RGBIR color formats. Supports 24-bit Wide Dynamic Range (WDR).

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.

MODALIX CHIP FORM-FACTOR:

FCBGA 1369 balls;
25mm x 25mm



DRAM interface system (DIS)

Supporting 4 × 32 LPDDR5 chips or 2 × 64 LPDDR5 chips. Target speed 6400 Mbps (LPDDR5) providing an effective theoretical bandwidth of 102 GB/s across all 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's software-first approach includes a full SDK with a highly optimized compiler that supports running any ML network and ML framework, along with Python and GStreamer APIs that enable effortless development and deployment for customers. This software architecture enables SiMa.ai to support a wide range of frameworks (e.g., TensorFlow, PyTorch, ONNX, etc.) and compile over 250+ models, 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. Because delivering such a sophisticated software stack for the MLSoC requires deep expertise, **SiMa.ai partnered with HTEC, whose strong background in AI/ML engineering and embedded edge systems makes them an ideal strategic collaborator.**

Configuring MLSoC™ Modalix Solutions for Even Higher Performance

MLSoC Modalix is designed to operate in cluster mode, with the ability to combine two or four chips in configurations offering up to 100 TOPS and 200 TOPS performance, respectively, should your application demand it. For example, two MLSoC Modalix chips can be combined on a single board to create a 100 TOPS module, while a 200 TOPS module—most frequently in a PCIe FHFL card—can be created by combining four MLSoC Modalix chips. Please contact a SiMa.ai representative to learn about the system configurations for 100 TOPS and 200 TOPS at www.sima.ai/contact-us.

About SiMa.ai

SiMa.ai is a leader in Physical AI, delivering a purpose-built, software-centric platform that brings best-in-class performance, power efficiency, and ease of use to Physical AI applications. Focused on scaling Physical AI across robotics, automotive, industrial automation, aerospace & defense, smart vision, and healthcare, SiMa.ai is led by seasoned technologists and backed by top-tier investors. Headquartered in San Jose, California. Learn more at www.SiMa.ai.

About HTEC

HTEC Group Inc. is a global AI-first provider of strategic software and hardware embedded design and engineering services, specializing in Advanced Technologies, Financial Services, MedTech, Automotive, Telco, and Enterprise Software & Platforms. HTEC has a proven track record of helping Fortune 500 and hyper-growth companies solve complex engineering challenges, drive efficiency, reduce risks, and accelerate time to market. HTEC prides itself on attracting top talent and has strategically chosen the locations of its 20+ excellence centers to enable this. Learn more at www.htec.com.