MLSoC™ - An overview

Hot Chips 35, August 28-29, 2023

Srivi Dhruvanarayan, Victor Bittorf
Our Vision: Effortless machine learning for the embedded edge

Run any computer vision application, any network, any model, any framework, any sensor, any resolution.

Wide variety of end2end CV applications; Customers can customize applications for performance or power; Sensor connections; Accuracy and common operator support

Any

10x

Application

Efficient ML handling: Tile architecture. Supports batch processing, large tensors, concurrent models and fast context switching.

Scheduling: All operations efficiently pipelined and scheduled across Decoder, CVU, MLA and Arm App processor with fast context switching

Performance

Data handling: On chip decoder/encoder, compliant with AVC, HEVC, and MJPEG

Pre/Post processing: Dedicated CVU and application processor blocks enabling data to stay on chip for the entire pipeline

Efficient static scheduling: Maximize compute while minimizing data movement; High utilization of MLA Tiles

Low power

Quantization scheme: Patented low-power scheme that retains accuracy

Fully INT8 inference: 100% of total compute dedicated for INT8 inference

Patented cache usage: System power significantly reduced due to lower DDR traffic by retaining activations inside the SoC

Low code productization; Ready to use models, production ready platforms and dev kits

Pushbutton
# SiMa.ai key innovations

## Highly flexible ML accelerator
- **Fully programmable MLA**
- SW controlled data movement, scheduling & synchronization
- Supports full complement of CV applications at lowest power

## ANY and 10x
- **Secure, self-contained SoC**
- Seamless heterogeneous compute
- Optimized end-to-end SoC pipeline
- Enables ML capabilities for legacy apps; future proofs applications
- Highly optimized building blocks enable best in class end-to-end performance

## Pushbutton
- **Simple to develop & deploy**
- Effortless customer integration
- Low code customer evaluation
- Low code ML
- Customers can develop & deploy applications without having to understand details of HW

## Vision Development Platform (VDP)
- Allows customers to build apps without having to write code

---

**Effortless customer integration**

**Low code customer evaluation**

**Low code ML**

Customers can develop & deploy applications without having to understand details of HW

**Vision Development Platform (VDP)**

Allows customers to build apps without having to write code
Purpose built for ML edge at embedded edge

**Acquire ANY Data**
- PCIe Gen4, 1GB Ethernet
- Dedicate: I2C, SPI, GPIO, SDIO

**10X ML Processing**
- Video, CV & ML Processors
- ML Accelerator - 50 TOPS INT8
- Quad Vector DSP - 600 GOPS
- HW video encode/decode
- 16GB LPDDR4

**Decide, Control & Update**
- Quad A65E ARM8.3
- Dedicated secure boot processor

**MLSoC™ Machine Learning System-on-Chip**

- TSMC - 16nm
- TDP - 15-20W
- Typical ML Workloads, CV Pipelines - 8-10W
Silicon Overview - 10x Performance for CV Processing

CV Processor

400 GFLOPS
600 INT16 GOPS

Key Benefits:
- Quad Vector DSPs optimize pre/post processing
- Each Vector DSP has 512b VLIW processor

Multicore Complex
- 4 EV74 Vector DSP cores @ 800MHz
- Coherent, symmetric multiprocessing

Shared Memory
  - Total 2MB local storage
  - 8 banks, each 32K x 64bit

System bus masters
  - Two AXI3 masters for CPU instruction fetch and data traffic
  - Four AXI3 masters for integrated DMA

Debug Features
  - SiMa debug register group
  - ARConnect debug architecture
  - Interrupt Distribution Unit

EV74 DSP Core: RISC with SIMD
- Level 1 Caches
  - L1 instruction 64KB, 4 ways
  - L1 data 64KB
  - Dedicated 2KB coherency lookup (four 256x25)

Debug Features
  - Trace memory, 4KB
  - APB slave access

Vector Memory 256KB
Silicon Overview - 10x Performance for ML Processing

Machine Learning Accelerator

50 INT8 TOPS

10x10 Array

Decompression Engine
DMA
Decompression Engine
DMA
Decompression Engine
DMA
Decompression Engine
DMA

60GB/s B/W

16GB LPDDR4

1 GHz Tile + L1 256KB

DMA Instruction Manager SRAM ARM M4

50 INT8 TOPS
SiMa.ai MLA innovations

- Unique matrix processor architecture
- Microarchitecture optimizations
- Novel convolution algorithm
- Compiler driven statically scheduled mesh
- ML compiler: Radically new approach and design
- SW managed memory hierarchy
Flexible Application Processor
- Quad A65 ARM8.3 ISA + FPU @1 GHz
- I$:32K/D$:32K/L2:128K/L3:512K
- Coherent Cache
- Coherent Mesh Network
- Secure Boot Core
- Safety, Security, Debug

Key Benefit:
- Entire application, not just ML
- Embedded appliance size

Eliminate
- PC and x86 Blade Servers
- Multi-Purpose HW
- Nonsecure OS & Client SW

Silicon Overview - 10x Performance for Application Processing
MLPerf: SiMa.ai delivers advantage over NVIDIA

SiMa.ai MLSoC (N16) compiled results unseat Orin (8nm) on both performance and power

Purpose built
Low power + small area

General purpose (GPU)
High power + large area

SiMa.ai MLSoC
100% of total TOPS dedicated to power optimized INT8 inference; 35MB of dedicated MLA memory

NVIDIA
37% of total TOPS dedicated to power optimized INT8 inference; 9.2MB of GPU cache*

1 Camera
1.4x Orin
2.1x Xavier

8 Cameras
1.37x Orin
Xavier data not published

Link to MLPerf results
SiMa.ai MLSoC vs. Nvidia Xavier + Orin (FPS/W)

DLA+GPU configuration at 15W

Average across 23 models: SiMa.ai is 9x Xavier and 4x Orin on FPS/W
Customer pipelines are a lot more complex

Image Capture

Image Preprocessing

ML

Mask Generation

Post Processing

Use Case Application
(Re-identification, predictive maintenance, tracking, grasping, swarm capabilities, etc)

What Sima.ai can execute

What competing ML accelerators execute
Summary

- The **only** startup company with performance and flexibility roadmap for customers
- **Extensible** silicon and software architecture
- Leverages open source software with broad innovations to deliver a complete software and platform solution to our customers
- **First time right silicon!** Demonstrated engineering execution on both silicon and software
- **Improved time-to-market.** Production ready boards and software building blocks that customers can integrate readily into their platforms