MX3 Edge AI Accelerator

The MemryX MX3 Edge AI Accelerator is architected for simplicity and scalability, and an ideal solution for efficiently running AI models in a variety of application environments including industrial, automotive, and IoT.

MemryX uses an innovative pure dataflow architecture, with programmable hardware that best mirrors the data-driven design of AI algorithms. Dataflow architecture is designed for streaming inputs from cameras or other sensors. MemryX's proprietary 1-click compilation and mapping software optimizes the performance of any AI model(s) on MX3 without requiring any retraining or software tuning. Such ease-of-use greatly accelerates development cycles and offers seamless upgradability for Edge AI.

Supplementing or offloading AI inference processing from an Application Processor to a dedicated AI accelerator maximizes system efficiencies while minimizing costs. Using seamless chip-to-chip connectivity, multi-MX3 chiplet configurations enable scalability to any desired level of AI performance and/or model size. MX3 supports all major AI frameworks, and full interoperability across application processors and host operating systems.

Innovative Edge AI Architecture

At-memory computing for the highest throughput and lowest energy. Pure Dataflow architecture optimized for streaming inputs from cameras or other sensors.

Simplicity

1-click performance and power optimization of trained models with no retraining, hand-tuning, or quantization necessary. And with just one click, the majority of models execute with >50% chip utilization.

Scalability

The exact same software is used for multiple small AI networks in one MX3 chiplet or large model(s) across multiple chiplets. Any number of MX3 chiplets can be used to scale to the desired level of performance, latency, and/or model size(s).

Deterministic

Low Latency (Batch=1) pipelined dataflow design with no internal control planes or control logic to limit input data streaming. All memory is included on the chiplet, fully mitigating any system bottlenecks and placing no processing requirements on the host. With no bottlenecks, AI performance is highly consistent/deterministic.

Broad Support

Supports all popular AI software frameworks (e.g. PyTorch, ONNX, Tensorflow, Tensorflow-Lite, Keras), Application Processors (Arm, x86, Risc-V), and operating systems (Ubuntu, Ycoto, Android, Windows, etc.). Connects to USB or PCI-E I/O ports.

Markets

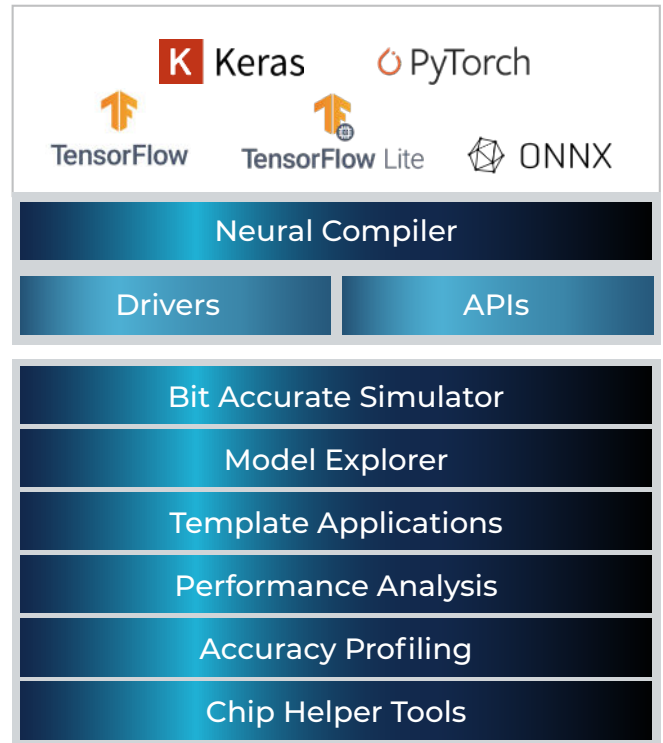| Industrial 4.0 & Robotics	Automotive	IoT	Metaverse	Smart Vision Systems	Computing Devices

memryX

The MemryX Software Developer Hub (SDK) is a comprehensive software solution including a 1-click compiler, bit-accurate simulator, and a suite of utility and runtime tools to ease development. MemryX's compiler has been tested on over a thousand standard and custom AI models for computer vision and sensor processing in a variety of applications (classification, object detection, segmentation, pose & depth estimation, OCR, GAN, etc.)

K Keras   ○ PyTorch
TensorFlow   TensorFlow Lite   ◇ ONNX

Neural Compiler

Drivers   **APIs**

Bit Accurate Simulator

Model Explorer

Template Applications

Performance Analysis

Accuracy Profiling

Chip Helper Tools

SDK Benefits

- Supports all common frameworks

- Support for multiple HW and OS platforms

- Bit Accuracy that aligns with the deterministic HW platform performance

- Scalable for any # of models across any number of MX3 chips

- Available code samples, tutorials, user guides, and Model Zoo

MX3 Accelerator IC Specifications

Performance TOPS	>5 TFLOPs / chip (>50% utilization)
Model size	10M parameters (8-bit) / chip
Performance (Batch=1) *Out-of-the Box Standard Model**	SSDMobilenet (224×224): 1403 fps (x1) SqueezDet (353×1248): 185 fps (x1) YOLOv7 tiny (416×416): 375 fps (x2) CenterNet (ResNet) (512×512): 64 fps (x4) YOLOv4 (416×416): 15 fps (x8)
Operations	Activations: 16-bit floating point (default) Weights: 4/8/16-bit (default = 8 bit)
Framework Support	TensorFlow, Tensorflow-lite, PyTorch, ONNX, Keras
Operating System Support	Linux (Ubuntu, Yocto, etc.), Android, Windows
Interface	USB 3.2 Gen1 x1 (5 Gbps) PCIe Gen3 x2
Power Consumption	1-2W Average
Temperature	-40C – 105C
Package Size	9mm x 9mm FCCSP

* Model as publicly available with no pruning, compression, or re-training. Performance is AI core and can vary based on host interface.

memryX

1600 Huron Parkway, 2nd Floor, Ann Arbor, MI 48109, USA

MemryX, Inc. is an AI semiconductor startup headquartered in Ann Arbor, MI USA with branches in Taipei and Hsinchu, Taiwan. We develop a highly scalable and innovated AI accelerator that offers high performance, low power, and customer ease of implementation for embedded Edge AI vision-based applications and real-time processing.

www.MemryX.com