Real-time Megapixel Vision with X1

Real-time Megapixel Vision requires high bandwidth support for complex deep learning models operating with small batch sizes in real time. InferX X1 is designed to solve these problems.

InferX™ X1 Edge Inference Accelerator

The InferX™ X1 Edge Inference Accelerator is **optimized for the processing of real-time megapixel vision workloads**.

These workloads are characterized by **deep networks with many feature maps and multiple operator types**. Also, model accuracy targets may require the use of mixed precisions, including INT8, INT16 and BF16. These workloads also require low latency batch size = 1inference processing.

The X1 dynamic tensor processor array **offers ASIC speed and efficiency while providing model flexibility**, through the use of reconfigurable control logic technology, to quickly adopt and deploy new Edge AI model technologies via field updates, thus future-proofing designs. The accelerator architecture of the X1 makes it easy to support processing of multiple data types including high resolution cameras.

The X1 is supported by the **InferX Edge Inference SDK**, which provides both model compiler and runtime software. The model compiler converts models expressed in TensorFlow Lite or TorchScript and compiles them to operate directly on the X1 accelerator.

X1P1

X1M

X1

FEATURES

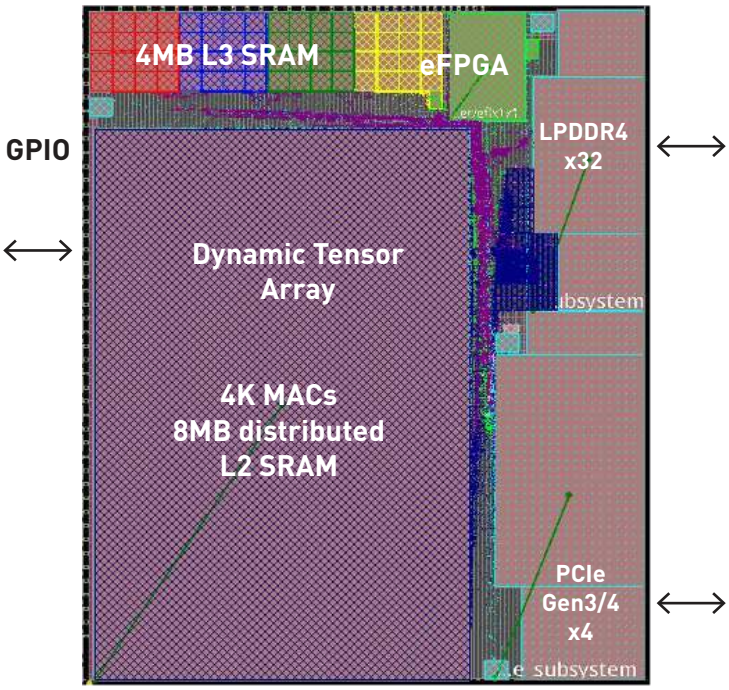
- High performance 4K element dynamic tensor processor array
- Optimized for tough, megapixel image processing models
- Designed for low latency B=1 inference processing
- Dynamic architecture future-proofs designs
- Higher throughput from less hardware/$/power
- INT8, INT16, BFloat16 support—can mix between layers
- Programmed via TensorFlow Lite/TorchScript

X1 Architecture

The InferX™ X1 contains a dynamic tensor processor array with 4K MAC units and 12 Megabytes of on-chip SRAM. The X1 also includes connectivity to external LPDDR4 DRAM for model weight, configuration and internal activation storage and Gen3/4 PCI Express for connectivity to a host processor.

The X1 Edge Inference accelerator approach supports a choice of host architecture (x86, Arm), operating system and system features including easy integration of various sensor input types (cameras, IR, Ultrasonic, RF, etc.) and communication standards (Ethernet, USB, Wi-Fi, etc.).

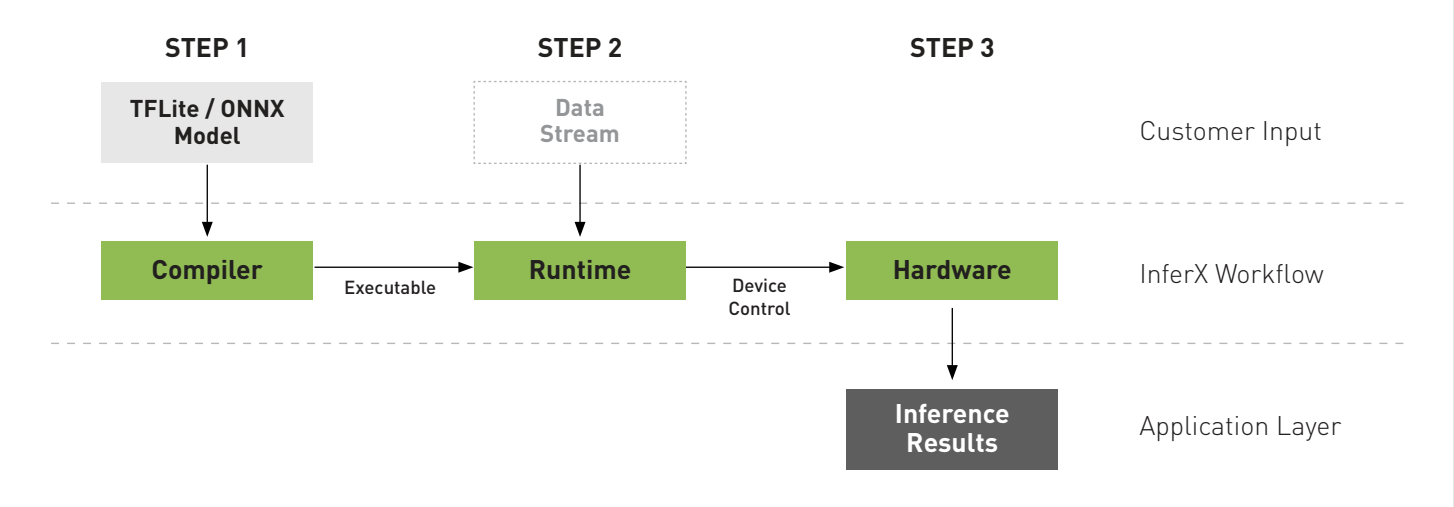
The InferX X1 accelerator is designed to support both existing and future AI/ML models through the dynamic tensor processor array. The inherent reconfigurability of this architecture provides future-proofing for customers whose edge inferencing workloads are continuing to evolve and improve.



InferX Edge Inference SDK

The InferX compiler takes as input TensorFlow Lite / TorchScript models and directly outputs a binary execution plan for the X1.

The InferX Runtime controls the execution of the model through a simple command structure designed for ease of use.



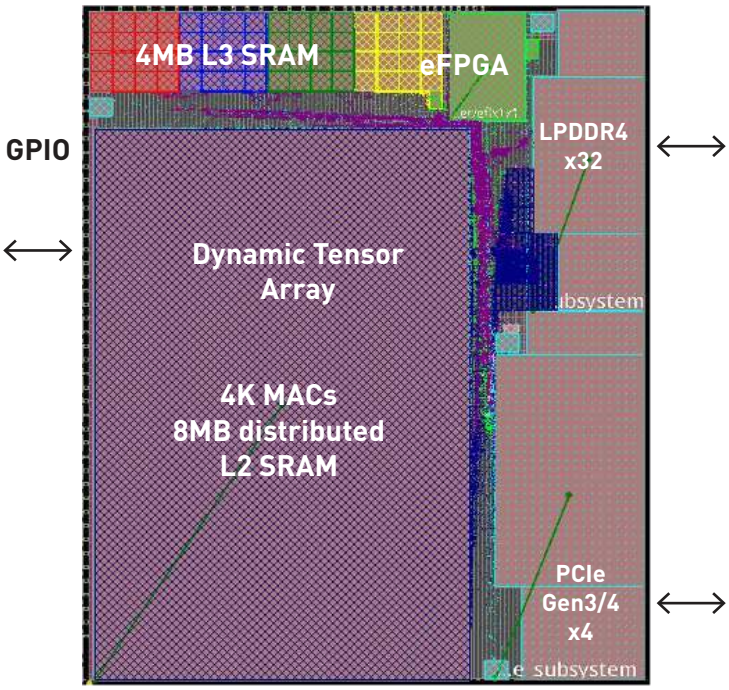
To Learn more please contact us: info@flex-logix.com or visit www.flex-logix.com

X1 Architecture

The InferX™ X1 contains a dynamic tensor processor array with 4K MAC units and 12 Megabytes of on-chip SRAM. The X1 also includes connectivity to external LPDDR4 DRAM for model weight, configuration and internal activation storage and Gen3/4 PCI Express for connectivity to a host processor.

The X1 Edge Inference accelerator approach supports a choice of host architecture (x86, Arm), operating system and system features including easy integration of various sensor input types (cameras, IR, Ultrasonic, RF, etc.) and communication standards (Ethernet, USB, Wi-Fi, etc.).

The InferX X1 accelerator is designed to support both existing and future AI/ML models through the dynamic tensor processor array. The inherent reconfigurability of this architecture provides future-proofing for customers whose edge inferencing workloads are continuing to evolve and improve.

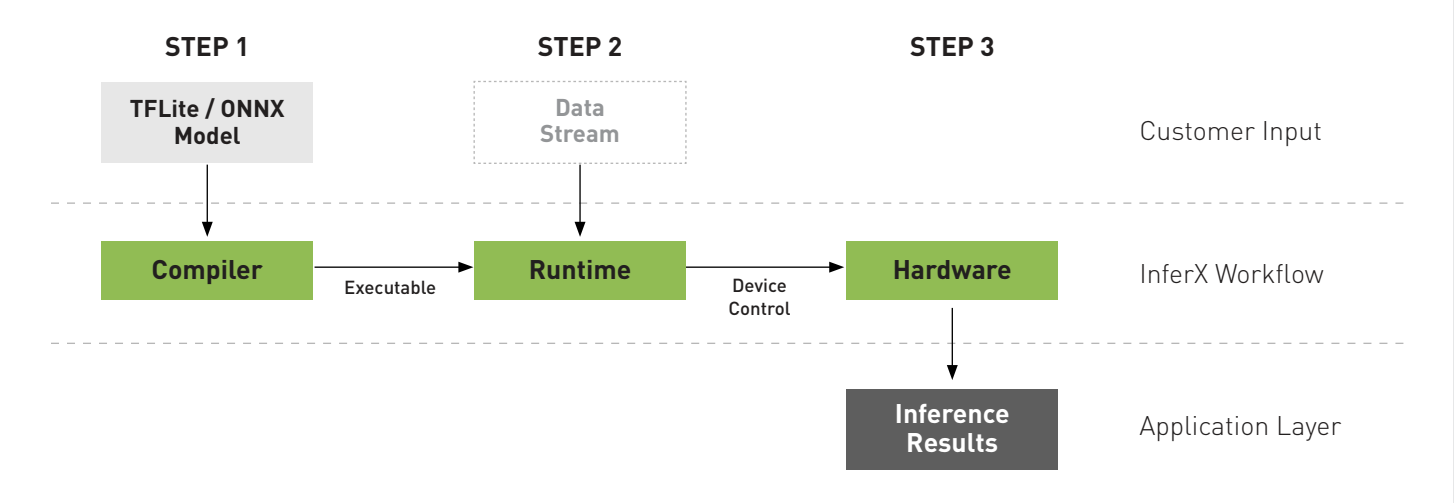


InferX Compiler & Runtime

The InferX X1 HW is designed around existing/future AI/ML models for future proofing.

The InferX compiler automatically groups layers where it improves throughput and power. The InferX Compiler takes as input TensorFlow Lite / Pytorch/ONNX models and directly outputs a binary execution plan for the X1.

The InferX Runtime controls the execution of the model through a simple command structure designed for ease of use.



To Learn more please contact us: info@flex-logix.com or visit www.flex-logix.com