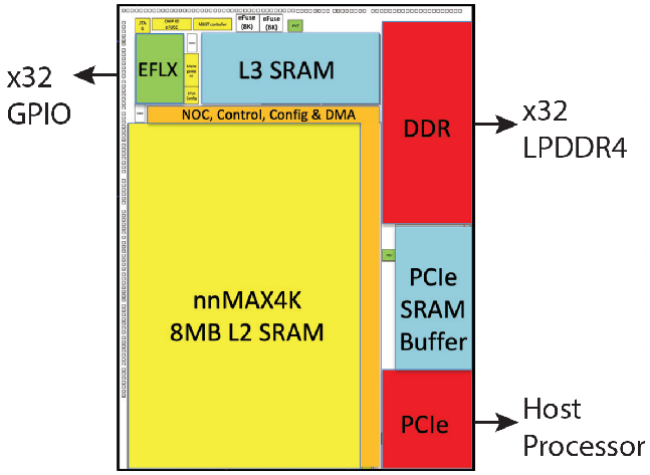
InferX™ X1

Edge Inference Coprocessor

flexlogix® AI + eFPGA

Datacenter Speed at Edge Power/Price

Block Diagram, Specifications



Package	21x21 mm BGA
Frequency	933MHz
MACs	4096
On-Chip SRAM	8MB
DRAM Interface	x32 LPDDR4
Host Interfaces	x4 PCIe Gen3, x32 GPIO
Chip Worst-Case TDP	13.5W

Features

Compiler support for models defined in Tensorflow Lite and ONNX

INT8 computation at full 933MHz, with BF16 operation at 2 cycles per MAC

INT8 and BF16 can be mixed layer by layer

On the fly Winograd transformation hardware, with INT8 convolutions performed at INT12 mode to ensure no loss of precision

Unique architecture based on inference optimized eFPGA

8MB of on-chip SRAM

Layer reconfiguration in less than 8 µsec

Background loading of weights during computation

GPIO x32 connectors

Benefits

•Leverage existing machine learning ecosystem
•Lowers the effort of application development

•InferX X1 allows you to achieve maximum precision at high throughput

•Models can achieve the optimum combination of precision and speed. Layers that need more precision can generate 16 bit activations (eg, speech, image upscaling)

•2.25x acceleration for supported convolutions
•Minimizes DRAM bandwidth by storing original weights
•No loss of precision

•Neural models map easily and efficiently
•Hardware resources can be reconfigured for maximum utilization

•Inference activations can be kept local, reducing power

•Higher hardware utilization

•Higher hardware utilization

•For connection to hosts without PCIe

InferX™ X1

Edge Inference Coprocessor

flexlogix® AI + eFPGA

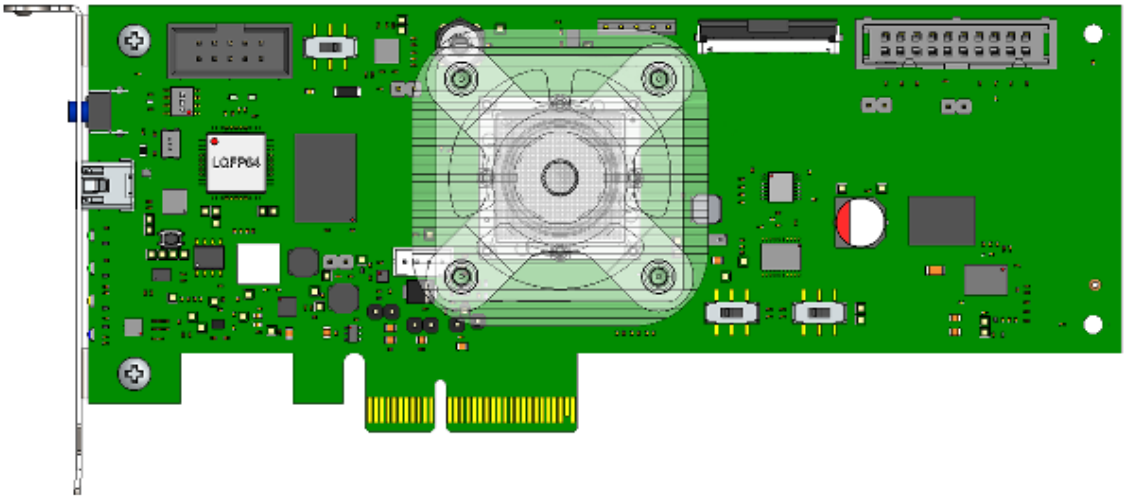
Product Overview

InferX X1 performs inference faster than existing edge inference chips, and is comparable in performance to datacenter inference cards. The X1 is optimized for the real-time input streams fundamental to edge applications, and operates in INT8 or BF16 precision over a batch size of 1 for minimum latency.

At the center of the X1 is the nnMAX acceleration engine. The programmable logic in the nnMAX rewire internal MAC clusters to create an optimal dataflow path for layers' convolutional kernels, with intermediate activations stored in local SRAM or passed directly to the next layer. Complementing the nnMAX inference acceleration engine, the X1 contains a PCIe x4 Gen3 interface to connect to a host, a x32 GPIO interface for those hosts without PCIe, and a x32 DDR interface for a single LPDDR4 DRAM.

The nnMAX compiler takes as input models defined in TensorFlow Lite and ONNX and directly outputs a binary execution plan for the X1. InferX X1 will be as available as standalone chips, or on PCIe boards with single and multi-chip configurations. Performance modeling via the nnMAX compiler is available for evaluation now.

HHHL InferX 1X1 PCIe Board, Available Q4 2020



X1 Streaming Performance Estimates (Batch=1)

Model	Precision	Input Size	Latency (ms)	Throughput
YOLOv3	INT8	1440	108.6	9.2
YOLOv3	INT8	608	18.5	54.1
ResNet-50	INT8	1440	50.8	19.7
ResNet-50	INT8	224	3.7	272.4