Product Brief

InferX™ X1
Edge Inference Co-Processor

flexlogix® Inference CoProcessor

Near Data Center Throughput at Edge Inference Chip Power/Price

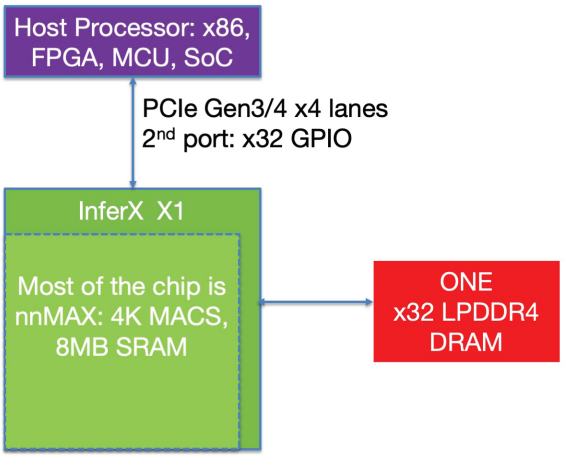
InferX X1 has throughput higher than existing edge inference chips and close to Data Center class inference cards. X1 is optimized for large models or large images at the small batch sizes required by edge applications. InferX X1 has a simple but powerful architecture: most of the chip is nnMAX inference acceleration engine with 4 lane PCIe Gen3 interface to a host processor, a x32 GPIO interface for hosts without PCIe or for operation of two X1 in tandem, and a x32 DDR interface to connect to a single LPDDR4 DRAM. InferX X1 will be available as a chip and on a PCIe card with both single and double X1 configurations. InferX X1 is programmed with TensorFlow Lite and ONNX.

Features	Benefits
nnMAX Compiler supports **Tensorflow Lite and ONNX**	Leverage existing ecosystems with a wide selection of tools. InferX X1 is easy to program. Performance modelling available now.
Numerics: **INT 8x8, 16x8** run at full 1.067GHz Numerics: **BFloat 16x16** takes 2 cycles per MAC BFloat16 accumulation is done at BF24 precision. On Chip **Hardware converts INT to BF** and back **INT8/16 and BF16 can be mixed layer by layer**	Layers that need more precision can generate 16 bit activations. Speech and some layers need the dynamic range of floating point. InferX X1 allows you to achieve maximum precision at high throughput. FP32 models can quickly be converted to BF16 for rapid evaluation. Models can achieve the optimum combination of precision and speed.
Winograd Transformation Hardware for INT8	3x3 Convolutions of Stride 1 are accelerated by 2.25x. To minimize DRAM bandwidth, weights are converted into Winograd form on the fly entering the MAC clusters. Winograd MACs are done in 12 bit mode to ensure no loss of precision. Models like YOLOv3 run 1.7x faster.
Unique architecture based on inference optimized eFPGA	Neural networks are data flow graphs. eFPGA is a data flow architecture. Neural models map easily and efficiently.
InferX uses an array of nnMAX MAC clusters with SRAM	Most inference compute is kept local using on-chip SRAM lowering power.
Clusters of 64 nnMAX connected by patented XFLX/ArrayLinx nonblocking interconnects reconfigured each layer	InferX reconfigures for each layer of the model to directly connect RAM to MAC clusters to programmable logic for activation then back to SRAM. **While running each layer the data path is like an ASIC.** SRAM bandwidth is utilized efficiently minimizing DRAM bandwidth. **Multiple layers can run at the same time eliminating the need to store large intermediate activations reducing DRAM bandwidth.**
SRAM bandwidth utilization is very high	Reduces DRAM bandwidth: DRAM memory references take 100x the energy of nnMAX MAC clusters directly connected to local SRAM.
Weights are loaded for the next layer in the background	Hardware utilization is high even at batch=1. Lower cost, higher speed.

InferX X1 Architecture, Specs & Benchmarks

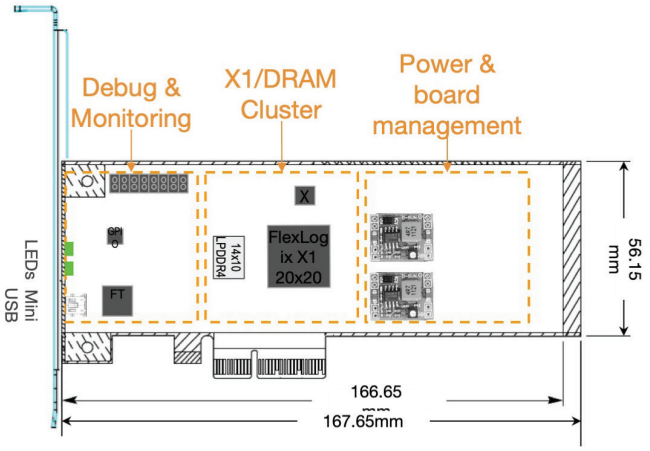
flexlogix® Inference CoProcessor

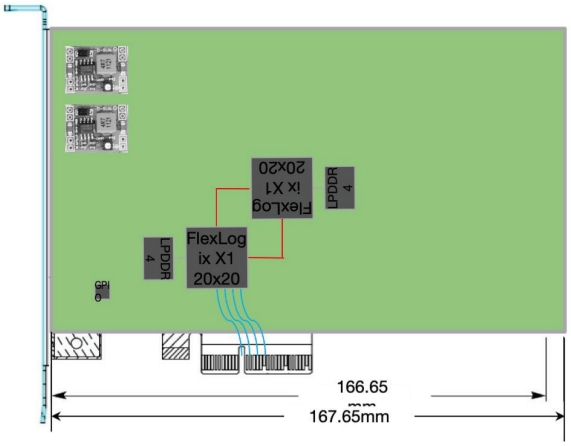
InferX X1 Chip



	InferX X1
Target Market	Edge Devices, Gateways, Low End Servers
Tape-out, TSMC16FFC	Q3/2019
TOPS (MACs x 1.067GHz x 2)	8.5
On-chip compute/RAM	4K MACs, 8MB SRAM
Interfaces	x4 Lanes PCIe Gen3/4 & x32 GPIO
Typical power varies by Model (25C, 0.8V, TT process)	~2.2W (ResNet-50)
X32 LPDDR4 DRAM	1
PCIe Card TDP (adds power of DRAM, regulators)	~9.6W (YOLOv3)

InferX X1 PCIe Cards: Single X1 HHHL and Double X1 FHHL





InferX X1 Benchmarks

Our nnMAX Compiler is available now for performance estimation using Tensorflow Lite models with INT8. Shortly we will add support for ONNX and BFloat16. Multi-layer configurations are now supported; the algorithm is likely to improve over time. InferX X1 is optimized for large models and/or large images: this is what most customers want to run in order to achieve the highest prediction accuracy. YOLOv3 does 227 Billion MACs for a 2 Megapixel image while MobileNetv2 was written to minimize MACs to run on processors: it does ~300 Million MACs per image. InferX X1 still outperforms alternatives even on these small models with image sizes not relevant to today's applications.

Model	Image Size	Source	Frames/Sec.	DRAM Bandwidth	MAC Utilization
YOLOv3	2048x1024	nnMAX Compiler	12.7	6.6 GB/sec	71%
YOLOv3	608x608	nnMAX Compiler	66.7	5.9 GB/sec	65%
YOLOv2	2048x1024	Spreadsheet	34.2	6.6 GB/sec	69%
InceptionV4	299x299	nnMAX Compiler	116	7.8 GB/sec.	31%
ResNet50 v1	224x224	Spreadsheet	363	11.2 GB/sec	22%
ResNet50 v1	2048x1024	Spreadsheet	26.4	7.3 GB/sec	67%
ResNet101 v2	224x224	nnMAX Compiler	181	10.7 GB/sec	43%
MobileNetv2	224x224	nnMAX Compiler	772	10 GB/sec	6%

All benchmarks are at batch size = 1