FLEX LOGIX DISCLOSES REAL-WORLD EDGE AI INFERENCE BENCHMARKS SHOWING SUPERIOR PRICE/PERFORMANCE FOR ALL MODELS

InferX X1 sampling expected Q3 2020

MOUNTAIN VIEW, Calif. – April 9, 2020 – Flex Logix® Technologies, Inc., the leading supplier of embedded FPGA (eFPGA) IP, architecture and software, today announced real-world benchmarks for its InferX™ X1 edge inference co-processor, showing significant price/performance advantages when compared to Nvidia's Tesla T4 and Xavier NX when run on actual customer models. These details were presented at today's Linley Spring Processor Conference by Vinay Mehta, Flex Logix's inference technical marketing manager, which can be viewed at this link.

The InferX X1 has a very small die size: 1/7th the area of Nvidia's Xavier NX and 1/11th the area of Nvidia's Tesla T4. Despite being so much smaller, the InferX X1 has latency for YOLOv3, an open source model that many customers plan to use, similar to Xavier NX. On two real customer models, InferX X1 was much faster, as much as 10x faster in one case.

In terms of price/performance as measured by streaming throughput divided by die size, InferX X1 is 2-10x better than Tesla T4 and 10-30x better than Xavier NX.

"Customers expect that they can use performance on ResNet-50 to compare alternatives. These benchmarks demonstrate that the relative performance of an inference accelerator on one model does not apply to all models," said Geoff Tate, CEO and co-founder of Flex Logix. "Customers should really be asking each vendor they evaluate to benchmark the model that they will use to find out the performance they will experience. We are doing this for customers now and welcome more – we can benchmark any neural network model in TensorFlow Lite or ONNX."

The InferX X1 is completing final design checks and will tape-out soon with sampling expected in Q3 2020 as a chip and as a PCIe board.

About InferX X1

Flex Logix's InferX X1 edge inference co-processor provides excellent inference efficiency, delivering more throughput on tough models for less $, less watts. It has been optimized for what the edge needs: large models and large models at batch=1. InferX X1 offers throughput close to data center boards that sell for thousands of dollars, but does so at single digit watts and at a fraction of the price. InferX X1 is programmed using TensorFlow Lite and ONNX: a performance modeler is available now.

InferX X1 is based on Flex Logix's nnMAX™ architecture integrating 4 tiles for 4K MACs and 8MB L2 SRAM. InferX X1 connects to a single x32 LPDDR4 DRAM. Four lanes of PCIe Gen3 connect to the host

processor; a x32 GPIO link is available for hosts without PCIe. Two X1's can work together to increase throughput up to 2x.

About Flex Logix

Flex Logix provides solutions for making flexible chips and accelerating neural network inferencing. Its eFPGA platform enables chips to be flexible to handle changing protocols, standards, algorithms and customer needs and to implement reconfigurable accelerators that speed key workloads 30-100x compared to processors. Flex Logix's second product line, nnMAX, utilizes its eFPGA and interconnect technology to provide modular, scalable neural inferencing from 1 to >100 TOPS using a higher throughput/$ and throughput/watt compared to other architectures. Flex Logix is headquartered in Mountain View, California.

####

MEDIA CONTACTS
Kelly Karr
Tanis Communications
kelly.karr@taniscomm.com
+408-718-9350