FLEX LOGIX ANNOUNCES WORKING SILICON OF FASTEST AND MOST EFFICIENT AI EDGE INFERENCE CHIP

InferX X1 brings AI to the masses; rivaling NVIDIA Xavier at 1/7th size and 10-100x better price/performance

MOUNTAIN VIEW, Calif. – October 20, 2020 – Flex Logix® Technologies, Inc. today announced availability of its InferX™ X1, the industry's fastest AI inference chip for edge systems. InferX X1 accelerates performance of neural network models such as object detection and recognition, and other neural network models, for robotics, industrial automation, medical imaging, gene sequencing, bank security, retail analytics, autonomous vehicles, aerospace and more. InferX X1 runs YOLOv3 object detection and recognition 30% faster than NVIDIA's industry leading Jetson Xavier and runs other real-world customer models up to ten times faster.

Many customers plan to use YOLOv3 in their products in robotics, bank security and retail analytics because it is the highest accuracy object detection and recognition algorithm. Additional customers have custom models they have developed for a range of applications where they need more throughput at lower cost. Flex Logix has benchmarked models for these applications and demonstrated to these customers that InferX X1 provides the needed throughput and lower cost. Sampling to early engagement customers will begin soon with broader sampling in Q1, and production parts will be available in the second quarter of 2021.

The InferX X1 silicon area is 54mm^2 which is 1/5th the size of a US penny and is much smaller than NVIDIA's Jetson Xavier at 350mm^2. InferX X1's high-volume price is as much as 10 times lower than NVIDIA's Xavier NX, enabling high-quality, high-performance AI inference for the first time to be implemented in mass market products selling in the millions of units.

InferX X1's software makes it easy to adopt. The InferX Compiler takes models in TensorFlow Lite or ONNX to program the InferX X1.

"Customers with existing edge inference systems are asking for more inference performance at better prices so they can implement neural networks in higher volume applications. InferX X1 meets their needs with both higher performance and lower prices," said Geoff Tate, CEO and

co-founder of Flex Logix. "InferX X1 delivers a 10-to-100 times improvement in inference price/performance versus the current industry leader."

"The technology announced by Flex Logix is a game changer and will significantly expand AI applications by bringing inference capabilities to the mass market," said Mike Gianfagna, principal at Gforce Marketing Inc. and SemiWiki contributor. "Other inference solutions can't compete with the price/performance that Flex Logix has achieved, and this is going to be a major disruptor in a market that is already forecast to grow exponentially in the future."

"TIRIAS Research believes that wide deployment of machine learning inference both inside and outside the data center has just begun," said Kevin Krewell, Principal Analyst at Tirias Research. "The key to deploying inference in volume and to a broad range of industries will be power- and cost-efficient silicon solutions that can be deployed in various edge device form factors. We believe Flex Logix is on track to deliver such a solution with its InferX X1 accelerator and boards."

"The InferX performance results are impressive and exceed the throughput of other products in its class," said Linley Gwennap, principal analyst at The Linley Group. "Customers should find the combination of efficient performance and relatively low price to be attractive for a range of intelligent edge devices."

Based on multiple Flex Logix proprietary technologies, the InferX X1 features a new architecture that achieves more throughput from less silicon area. Flex Logix's XFLX™ double density programmable interconnect is already used in the eFPGA (embedded FPGA) that Flex Logix has supplied for years to multiple customers including Dialog, Boeing, Sandia National Labs, and Datung Telecommunications. This is combined with a reconfigurable Tensor Processor consisting of 64 1-Dimensional Tensor Processors that are reconfigurable to efficiently implement the wide range of neural network operations. Because reconfiguration can be done in microseconds, each layer of a neural network model can be optimized with full-speed data paths for each layer.

InferX X1 mass production chips and software will be available Q2 2021. Customer samples and advance Compiler and Software Tools will be available in Q1 2021. Customers with Neural Network Models in TensorFlowLite or ONNX with volume applications in 2021 may contact Flex Logix now for performance benchmarking, early sampling and tool access, and detailed specifications and pricing.

More information is also available at www.flex-logix.com.

Technology Details and Specifications:

- High MAC utilization up to 70% for large models/images translates into less silicon area/cost
- 1-Dimensional Tensor Processors (1D TPUs) are a 1D systolic array

    - 64 byte input tensor
    - 64 INT8 MACs
    - 32 BF16 MACs
    - 64 byte x 256 byte weight matrix
    - One dimensional systolic array produces an output tensor every 64 cycles using 4096 MAC operations
- Reconfigurable Tensor Processor made up of 64 1D TPUs per X1
    - TPUs can be configured in series or in parallel to implement a wide range of tensor operations; this flexibility enables high performance implementation of new operations such as 3D convolution
    - Programmable interconnect provides a full speed, non-contention data path from SRAM through the TPUs to SRAM
- eFPGA programmable logic implements high speed state machines that control the TPUs and implement the control algorithms for the operators
- Each layer of a model is configured exactly as needed; reconfiguration for a new layer takes just microseconds
- DRAM traffic bringing in the weights and configuration for the next layer occurs in the background during compute of the current layer; this minimizes compute stalls
- Combining two layers in one configuration (layer fusion) minimizes DRAM traffic delays.
- Minimal memory keeps cost down: LPDDR4x DRAM, 14MB total SRAM
- x4 PCIe Gen 3 or Gen 4 provides rapid communication with the host
- 54 mm^2 die size in 16nm process
- 21 x 21 mm flip-chip Ball Grid Array package

Availability and Pricing

The InferX X1 is sampling soon to selected customers and production is expected in the second quarter of 2021. Pricing ranges based on configuration and volumes from $34 - $199.

Commercial Pricing (1,000 units – 1 million units)

Speed	Commercial Cost
933 MHz	$199-$69
800 MHz	$149-$49
667 MHz	$124-$39
533 MHz	$99-$34

Industrial and aerospace temperature range versions are also available.
See www.flex-logix.com or contact info@flex-logix.com for details on pricing.

About Flex Logix

Flex Logix provides industry-leading solutions for making flexible chips and accelerating neural network inferencing. Its InferX X1 is the industry's fastest and most-efficient AI edge inference accelerator that will bring AI to the masses in high-volume applications, surpassing competitor's performance at 1/7th size and 10x lower price. Flex Logix's 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 is headquartered in Mountain View, California. For more information, visit https://flex-logix.com

####

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