NEWS RELEASE

Flex Logix Launches InferX™ X1 Edge Inference Co-Processor That Delivers Near-Data Center Throughput at a Fraction of the Power and Cost

Delivers up to 10 Times the Throughput Compared to Existing Inferencing Edge ICs

Mountain View, Calif., April 10, 2019 – Flex Logix® Technologies, Inc. today announced that it has leveraged its core patent-protected interconnect technology from its embedded FPGA (eFPGA) line of business combined with inference-optimized nnMAX™ clusters to develop the InferX™ X1 edge inference co-processor. Unveiled today in a presentation at the Linley Processor Conference in Santa Clara, the Flex Logix InferX X1 chip delivers high throughput in edge applications with a single DRAM, resulting in much higher throughput/watt than existing solutions. Its performance advantage is especially strong at low batch sizes which are required in edge applications where there is typically only one camera/sensor.

InferX X1's performance at small batch sizes is close to data center inference boards and is optimized for large models which need 100s of billions of operations per image. For example, for YOLOv3 real time object recognition, InferX X1 processes 12.7 frames/second of 2 megapixel images at batch size = 1. Performance is roughly linear with image size: so frame rate approximately doubles for a 1 megapixel image. This is with a single DRAM.

InferX X1 will be available as chips for edge devices and on half-height, half-length PCIe cards for edge servers and gateways. It is programmed using the nnMAX Compiler which takes Tensorflow Lite or ONNX models. The internal architecture of the inference engine is hidden from the user.

InferX supports integer 8, 16 and bfloat 16 numerics with the ability to mix them across layers, enabling easy porting of models with optimized throughput at maximum precision. InferX supports Winograd transformation for integer 8 mode for common convolution operations which accelerates throughput by 2.25x for these functions while minimizing bandwidth by doing on-chip, on-the-fly conversion of weights to Winograd mode. To ensure no loss of precision, Winograd calculations are done with 12 bits of accuracy.

"The difficult challenge in neural network inference is minimizing data movement and energy consumption, which is something our interconnect technology can do amazingly well," said Geoff Tate, CEO of Flex Logix. "While processing a layer, the datapath is configured for the entire stage using our reconfigurable interconnect, enabling InferX to operate like an ASIC, then reconfigure rapidly for the next layer. Because most of our bandwidth comes from local SRAM, InferX requires just a single DRAM, simplifying die and package, and cutting cost and power."

Added Tate, "Our on-chip Winograd conversion further reduces bandwidth due to weight loading because weights are 1.8x larger in Winograd format. Our mixed numerics capability enables customers to use integer 8 where practical, but falls back to floating point as needed for achieving the desired prediction accuracy. This combination of features allows for high prediction accuracy, high throughput, low cost, and low power edge inference."

Two High-Growth Businesses: One Innovative Interconnect Technology

Flex Logix has emerged as a market leader in the eFPGA market, with customers such as MorningCore/Datang Telecom, DARPA, Boeing, Harvard, Sandia, SiFive RISC-V, and many more designing chips based on this platform. The new nnMAX neural inference engine leverages the same core interconnect technology used in eFPGA combined with multiplier-accumulators optimized for inference and aggregated into clusters of 64 with local weight storage for each layer.

In neural inference, computation is dominated by trillions of operations (multiplies and accumulates), typically using 8-bit integer inputs and weights, and sometimes 16-bit integer or 16-bit bfloat floating point. It is possible to mix these numerics layer by layer as needed to achieve target precision. The technology Flex Logix has developed for eFPGA is also ideally suited for inference because eFPGA allows for re-configurable data paths and fast control logic for each network stage. SRAM in eFPGA is reconfigurable as needed in neural networks where each layer can require different data sizes; and Flex Logix interconnects allow reconfigurable connections between SRAM input banks, MAC clusters, and activation to SRAM output banks at each stage.

The result is an nnMAX tile of 1024 MACs with local SRAM, which in 16nm has ~2.1 TOPS peak performance. nnMAX tiles can be arrayed into NxN arrays of any size, without any GDS change, with varying amounts of SRAM as needed to optimize for the target neural network model, up to to >100 TOPS peak performance.

High MAC utilization means less silicon area/cost, and low DRAM bandwidth means fewer DRAMs, less system cost and less power.

InferX is programmed using TensorFlow Lite and ONNX, two of the most popular inference ecosystems.

Availabiity

nnMAX is in development now and will be available for integration in SoCs by Q3 2019. InferX X1 will tape-out in Q3 2019 and samples of chips and PCIe boards will be available shortly after. For more information, prospective customers can go to www.flex-logix.com to review the slides presented today at the Linley Processor Conference and/or contact info@flex-logix.com for further details of nnMAX and InferX under NDA.

About Flex Logix

Flex Logix, founded in March 2014, 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 faster than Microsoft Azure processing in the Cloud. eFPGA is available for any array size on the most popular process nodes now with increasing customer adoption. Flex Logix's second product line, nnNMAX, utilizes its eFPGA and interconnect technology to provide modular, scalable neural inferencing from 2 to >100 TOPS using 1/10th the typical DRAM bandwidth, resulting in much lower system power and cost. Having raised more than $25 million of venture capital, Flex Logix is headquartered in Mountain View, California, and has sales rep offices in China, Europe, Israel, Japan, Taiwan and throughout the USA. More information can be obtained at http://www.flex-logix.com or follow on Twitter at @efpga.

####

PRESS CONTACT:
Kelly Karr
Tanis Communications, Inc.
kelly.karr@taniscomm.com
+408-718-9350