FLEX LOGIX ANNOUNCES nnMAX IP DELIVERS HIGHER-THROUGHPUT/$ AND HIGHER THROUGHPUT/WATT FOR KEY DSP FUNCTIONS

Cheng Wang unveiled key performance benchmarks at today's virtual Linley Processor Forum, proving that nnMAX is not only superior for AI inference, but also for key DSP functions

MOUNTAIN VIEW, Calif. – April 7, 2020 – Flex Logix® Technologies, Inc., the leading supplier of embedded FPGA (eFPGA) IP, architecture and software, today unveiled key benchmarking information around its new nnMAX™ architecture, showing how it can effectively be used for DSP acceleration for key functions. For FIR (finite impulse response) filters, nnMAX is able to process up to 1 Gigasamples per second with hundreds and even thousands of "taps" or coefficients. FIR filters are widely used in a large number of commercial and aerospace applications. Cheng Wang, Flex Logix's senior VP engineering and co-founder, disclosed these benchmarks and more at today's online Linley Processor Forum in a presentation titled "DSP Acceleration using nnMAX." His full presentation can be viewed at this link.

"Because nnMAX is so good at accelerating AI inference, customers started asking us if it could also be applied to DSP functions," said Geoff Tate, CEO and co-founder of Flex Logix. "When we started evaluating their models, we found that it can deliver similar performance to the most expensive Xilinx FPGAs in the same process node (16nm), and is also faster than TI's highest-performing DSP – but in a much smaller silicon area than both those solutions. nnMAX is available now for 16nm SoC designs and will be available for additional process nodes in 2021."

About nnMAX

NMAX is a general purpose Neural Inferencing Engine that can run any type of NN from simple fully connected DNN to RNN to CNN and can run multiple NNs at a time. It has demonstrated excellent inference efficiency, delivering more throughput on tough models for less $, less watts.

nnMAX is programmed with TensorFlow Lite and ONNX. Numerics supported are INT8, INT16 and BFloat16 and can be mixed layer by layer to maximize prediction accuracy. INT8/16 activations are processed at full rate; BFloat16 at half rate. Hardware converts between INT and BFloat as needed layer by layer. 3×3 Convolutions of Stride 1 are accelerated by Winograd hardware: YOLOv3 is 1.7x faster, ResNet-50 is 1.4x faster. This is done at full precision. Weights are stored in non-Winograd form to keep memory bandwidth low. nnMAX is a tile architecture any throughput required can be delivered with the right amount of SRAM for your model.

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