The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

Flex Logix Moves Into Chips

InferX Coprocessor Targets Neural-Network Inference in Edge Servers

By Mike Demler (April 15, 2019)

After four years of creating embedded-FPGA intellectual-property (IP) cores, Flex Logix is entering the AI-chip business. The InferX X1 is an AI coprocessor designed to connect to a host CPU through its PCIe interface. The chip integrates four enhanced versions of the NNMax tiles the company introduced last year in an IP core, as Figure 1 shows. The company will continue licensing FPGA IP and will selectively license the NNMax IP, but it expects InferX chips and boards to quickly become its largest revenue source. It plans to manufacture the X1 in 16FFC technology at TSMC and to ship samples around the end of the year, making production likely in late 2020.

The InferX X1 chip targets edge gateways, low-end servers, robotics, surveillance cameras, and other high-performance edge devices. The company plans to sell it in a chip package as well as on a half-height half-length PCIe card and a full-height half-length card containing two chips. The X1 has a four-lane PCIe Gen3/4 interface; it also has 32 GPIO ports, allowing it to connect to MCUs that lack PCIe. The GPIO ports additionally enable customers to connect two or four chips in parallel.

Although Flex won't tape out the X1 until 3Q19, it has run cycle-accurate gate-level simulations using ResNet-50, Yolo v3, and other popular neural networks. For a fast-fast process corner in a device operating at maximum temperature and voltage while running Yolo v3, the simulations predict 7.1W for worst-case power consumption. Running the less complex ResNet-50, typical power consumption falls to 2.2W. At a 1.0GHz clock rate, the four-tile accelerator delivers approximately 8.5 trillion operations per second (TOPS). By comparison, the ResNet performance per watt is 9x that of NovuMind's NovuTensor, which also works as a PCIe-connected AI coprocessor.

Computational Tiles That Float

For the InferX X1 coprocessor, Flex Logix enhanced the computational-tile architecture in its NMax512 IP core (see *MPR 11/5/18,* "Flex Logix Spins Neural Accelerator"). Each of the four tiles includes EFLX programmable logic and 16 NNMax clusters per tile, twice the number in the IP version.

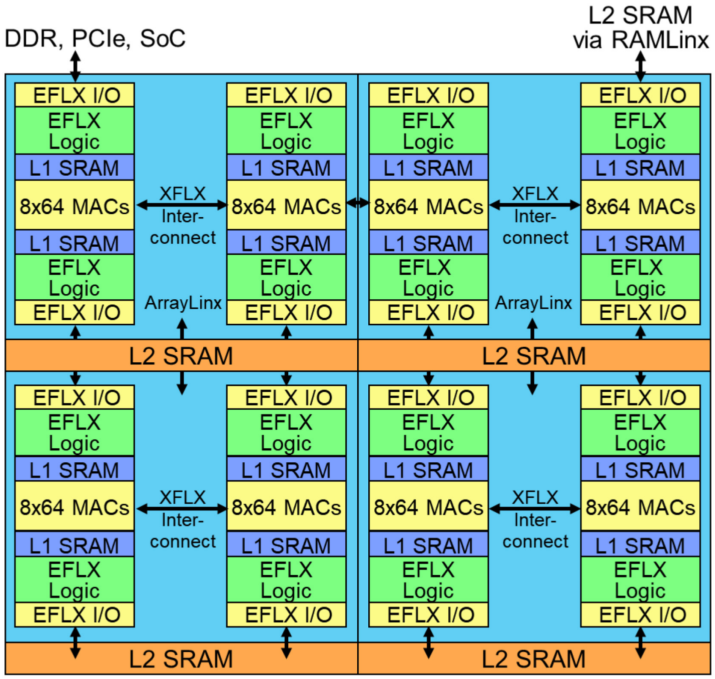

Figure 1. Flex Logix InferX X1 coprocessor. The chip integrates four enhanced NNMax compute tiles, each combining FPGA logic with 1,024 multiply-accumulators, SRAMs, and reconfigurable interconnect. The company plans to manufacture the chip in 16FFC technology at TSMC. Running at 1.0GHz, it delivers 8.5 trillion operations per second.

Price and Availability

Flex Logix withheld pricing for the InferX X1 products, but we expect the standalone chip to sell for approximately $30 and the single-chip PCIe card to sell for approximately $300. The company plans to ship sample quantities around the end of the year, and we expect it to reach volume production in late 2020. For more details, access *www.flex-logix.com/inference/*.

Each cluster is a 64-multiply-accumulate (MAC) systolic array comprising 1,024 MACs per tile. The MACs are like the hard-logic DSP blocks often embedded in FPGAs, combining the power-performance advantage of custom logic with programmable gates for flexibility. Each cluster also integrates 256KB SRAMs for local weight storage, which Flex calls level-zero (L0) memory. The X1 integrates a total of 8MB of L2 SRAM as well.

The MACs handle 16-bit x 8-bit signed or unsigned multipliers and a 32-bit accumulator. Users can configure them to deliver 8x8 and 16x8 integer results in a single cycle, but 16x16 integer and Bfloat16 operations are half-rate. Bfloat16 support is new in the InferX X1. It provides the same dynamic range as IEEE 754 single-precision (FP32) floating point, but in a 16-bit format that reduces memory bandwidth and storage. In Bfloat16 mode, the accumulator is 24 bits. The MACs allow mixed-resolution operations layer by layer. Hardware converts activations from Bfloat16 to INT8/16, or vice versa, between layers as necessary.

The Trouble With Transforms

Convolutions can be up to 90% of a neural network's operations. In the brute-force direct method, they use wide MAC units to perform large numbers of dot-product operations (see *MPR 7/9/18,* "Arm Dot Products Accelerate CNNs"). Although convolutions came from signal-processing theory, neural-network developers can employ some of the same DSP techniques to accelerate the calculations, such as digital filters and Fast Fourier Transforms (FFTs).

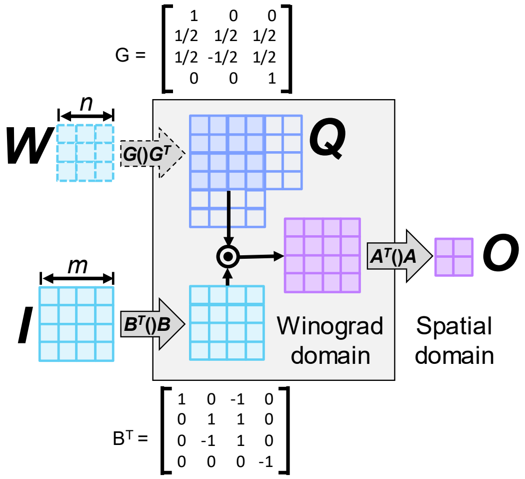

Figure 2. Winograd transform. The technique reduces the total number of MAC operations compared with direct convolution, but it requires transforming the input-image tiles (I) and weight filters (W) to 4x4 matrices. The output (O) produces a 2x2 feature map in pixel space.

Recently, many researchers have adopted a technique created by Stanford professor Terry Winograd. Compared with the direct convolution method, the Winograd transform reduces the number of MAC operations by applying sparse-matrix transformations to input image tiles and weight filters. For example, Winograd converts four 3x3 convolutions (or 4x3x3=36 multiplications) with a stride of one to a single 4x4 convolution with a stride of two, as Figure 2 shows. Applying a Winograd transform for that convolution requires only 16 multiplications, reducing the number of operations by 2.25x. The InferX compiler automatically employs Winograd on applicable layers, and the chip includes hardware that applies the transformations and detransformations to INT8 operations.

As always, however, there's no free lunch. Converting the input tiles involves four additions or subtractions. If the pixel values exceed one-fourth of the maximum INT8 value (i.e., 2^6=64), the four additions will cause a 1-bit overflow (4x64=256). Likewise, four additions of 128 or greater (4x2^7=2^9=1,024) result in a 2-bit overflow. To avoid overflow/underflow while maintaining the same precision, the INT8 MACs must switch to INT10.

Difficulties can also arise in the weight-filter transformation, since it introduces two matrix multiplications that each multiply weights by 0.5. The combination of the two matrix operations multiplies some weights by 0.25 (an implicit two-bit shift) and adds or subtracts as many as nine weights. The weight transformation effectively increases the dynamic range such that maintaining the same precision requires up to 12-bit MACs, data paths, and memories. Therefore, fixed INT8 MACs can't perform Winograd transforms without losing precision. But to boost throughout and computational efficiency, Flex Logix optimized its design for Winograd. Its 4x4 filters employ up to INT12 resolution, completing the transformation without losing accuracy but adding die area relative to 8x8-bit MAC units.

On-the-Fly Configurability

The InferX X1 uses the ArrayLinx mesh interconnect to link the NNMax tiles in a 2x2 array. Activation, pooling, and other nonconvolution layers run in the FPGA. The FPGA blocks comprise 1,960 six-input LUTs per tile. The programmable LUTs and XFLX interconnect allow on-the-fly reconfiguration of the compute-tile functions, data paths, and SRAM read/write ports for each layer. Each NNMax tile includes 736 I/Os, which are configurable for connection to the internal SoC bus, DRAM, and PCIe. The X1 has a single 32-bit LPDDR4 interface.

In the NNMax architecture, feature maps and weights flow through a three-level memory hierarchy. To minimize storage requirements, the DRAM holds weights in their original (non-Winograd) format. The RAMLinx interconnect reads the weights from DRAM into the L2 SRAM. Activations (i.e. feature maps) flow directly from the L2, or from DRAM (for activations too large to store on chip), into the NNMax clusters. Each tile can couple with up to 4MB of L2 SRAM, and all L2 SRAMs in the NNMAX array are accessible through the ArrayLinx interconnect. After MAC operations, the outputs flow to the FPGA to calculate the next set of activations, which flow back to the L2, and so on.

The chip converts weights to Winograd format on the fly, expanding them by a factor of 1.8x (4x4/3x3) without writing them back to memory. While one layer executes, the chip prefetches weights for the next layer from the L2 SRAM, performs the Winograd tranfromation, and writes the output to the L1 SRAM in the NNMax tiles. The L1s are shadow memories for the L0s. The tiles also have a shadow memory that prefetches configurations for the next layer.

To initiate processing of a new neural-network layer, the tiles first use the configuration bits to set up the new operations, then they load the prefetched weights from the L1 into each cluster's L0. The configuration process for these layers takes on the order of 1,000 cycles, or about 1.0 microseconds. But it's orders of magnitude faster than programming times for a traditional commercial FPGA. Flex's unique FPGA design accelerates reconfiguration by writing an entire row of LUTs in parallel rather than by using the serial bit-stream process of conventional FPGAs.

By pipelining layer operations in its data-flow architecture, InferX reduces DRAM bandwidth. Figure 3 shows an example for the Yolo v3 network. Layer 0 reads in a 6MB image frame, but its operations produce a 64MB output. By using a section of NNMax clusters for layer 0 and another section for layer 1, InferX can configure the XFLX interconnect to keep the data on chip, sending the outputs directly between NNMax clusters.

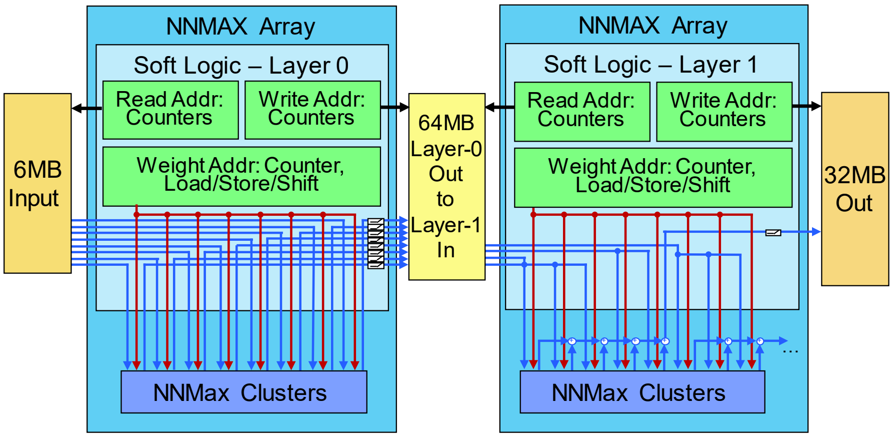

Figure 3. InferX pipeline operations. The data-flow architecture can keep data on chip rather than writing results from each layer to DRAM. The programmable logic and configurable interconnect enable the NNMax tiles to process different neural-network layers simultaneously.

Winning on Efficiency

By offering InferX, Flex Logix aims to compete with edge-inference chips, such as Intel's Myriad X (see *MPR 12/17/18,* "Intel Gains Myriad Customers"). According to its tests running GoogLeNet, the InferX X1 can classify 266 images per second (IPS), equivalent to 3.3x Myriad X's throughput and 3.6x the power efficiency. The Flex chip requires an external host processor, however, whereas Myriad is a complete SoC.

InferX fares even better compared with NovuMind's coprocessor, which powers the NovuTensor PCIe card (see *MPR 11/26/18,* "NovuMind Relieves Tensor Headaches"). Its single-batch ResNet-50 performance is 363 IPS, equivalent to 3x NovuTensor's. Although the NovuMind product consumes 15W, the Flex product consumes only 5.0W (worst case) on that model. On the more challenging Yolo v3, the X1 can process 12.7fps at 2Mpixel resolution, requiring 4.0W (typical).

The company supports the X1 with its NNMax compiler, which handles TensorFlow Lite and ONNX models. The compiler's front end includes performance estimation: it computes latency, MAC utilization, and DRAM bandwidth per neural-network layer and per model. Its back end translates neural-network models to the programmable architecture, doing place and route, pipelining, retiming, and binary generation.

Plotting a New Strategy

Many AI startups have launched with a strategy to offer both licensable cores and SoCs, as the former seems a quick path to revenue that can help finance an expensive tapeout. But pursuing both business models while avoiding competition with customers is a challenge, and the dueling strategies place a strain on a startup's engineering resources. Cambricon is one apparently successful exception (see *MPR 3/5/18,* "Cambricon Leads China Into AI Chips"), but that Chinese startup only serves companies in its home market, where the government is pouring billions of dollars into developing an AI industry.

Flex Logix will continue licensing its embedded-FPGA IP, but it plans to emphasize the chip business because of the even higher revenue potential. It will selectively license NNMax IP to customers in complementary markets and to those with long and expensive design cycles, such as automotive suppliers.

The company must now deliver on its latest promises. Its embedded-FPGA business is growing, with customers building chips in 40nm to 12nm TSMC technologies. Less than one year ago, Flex published a roadmap that would take it to 7nm (see *MPR 7/23/18,* "Flex Logix Diversifies eFPGA"), but it's now charting a new course.

Additional investment is a priority. The company's current $12 million in venture-capital funding is sufficient for an IP business, but taking a 16nm chip to production will require much more. Fortunately, the FPGA IP is delivering increasing revenue, and in 2Q19, the company expects to close a Series C round that will be sufficient to bring the InferX X1 to market. The simulations are impressive, and the tile architecture allows scaling to higher performance. We expect investors will find the design attractive, but in the fast-moving AI market, much could change before it reaches production in late 2020. ♦