FLEX LOGIX ACCELERATES DSP

By Mike Demler (April 27, 2020)

Since AlexNet burst onto the scene in 2012, convolutional neural networks (CNNs) have become popular for image classification and object recognition. Although using convolution for AI is a relatively recent trend, French mathematician Joseph Fourier developed the method for extracting a function's fundamental components more than 200 years ago. Fourier introduced convolution in his eponymous transform, which signal-processing engineers later adapted for digital filters. Recently, Flex Logix customers recognized the common mathematical foundation shared by CNNs and DSPs, leading the embedded-FPGA developer to evaluate the performance of its NNMax deep-learning accelerators (DLAs) on finite-impulse-response (FIR) filters.

Flex developed NNMax as licensable intellectual property (IP), which it offers as a tile comprising systolic multiply-accumulate (MAC) arrays, memory blocks, programmable logic, and the company's XFLX configurable interconnect (see *MPR 11/5/18,* "Flex Logix Spins Neural Accelerator"). In a 16FFC process at TSMC, the 1,024-MAC-unit tile consumes 4.5mm^2 and runs at 933MHz. In this configuration, NNMax performs single-cycle 8x8- and 16x8-bit MACs, delivering 1.9 trillion operations per second (TOPS). Bfloat16 and INT16 MACs achieve half that rate.

Flex instantiates four NNMax tiles in the InferX AI coprocessor (see *MPR 4/15/19,* "Flex Logix Moves Into Chips"). Along with 4,096 MAC units, InferX integrates 8MB of total L2 SRAM distributed among the NNMax clusters, a 4MB L3 SRAM, a 32-bit LPDDR4 interface, and a four-lane PCIe interface for connecting to a host processor. The company plans to sample it in 3Q20.

MAC operations typically consume 90% or more of the cycles in a CNN, but FIR filters employ the same operations. In a CNN, each artificial neuron multiplies its inputs by a set of weights (also called a filter), adds the result, and feeds the sum forward as an input to the next layer. FIR filters are similar, comprising a series of MAC operations called filter taps. Each stage in the filter multiplies the input signal (s) by a filter coefficient (h), combines the result with the output of the previous filter tap, and sends the sum to the next tap. The mathematical equation for an N-stage FIR filter is the following:

$$y(n) = h(0)s(n) + h(1)s(n-1) + \ldots + h(N-1)s(n-N-1)$$

NNMax is well suited to FIR operations because it's arranged in rows of 32 MAC units, with each row connecting to a small SRAM that can hold either neural-network weights or filter coefficients. Flex designed the NNMax SRAM to store 10-bit unsigned or 11-bit signed coefficients, but users can combine two of these values to represent 16-bit integer or Bfloat16 parameters. Although each NNMax cluster can execute a filter with up to 64 taps, the XFLX interconnect allows user configuration of larger filters by cascading multiple clusters or multiple tiles. The system DRAM can store several filter-configuration files, and loading one into on-chip SRAM takes less than two microseconds.

The achievable filter performance depends on the number of taps per NNMax tile and the coefficient precision. At a 1GHz sample rate, a single NNMax cluster can handle a 16-tap filter composed of real INT16 or Bfloat16 coefficients. Reducing the clock speed to 500MHz doubles the number of supported taps. Because they must compute both real and imaginary components, complex INT16/Bfloat16 filters run at half the speed and support one-fourth as many taps as real filters with the same cluster resources.

Flex compared its estimated NNMax FIR-filter performance with that of Ceva's new XC16 DSP (see *MPR 3/23/20,* "Ceva XC16 Stays Ahead of 5G Rollout"). It showed a 20% advantage in operations per second, but that comparison is unfair. Although the XC16's vector cores have 256 MAC units, those units represent a small portion of its 1,280-bit SIMD/VLIW engine, which Ceva designed for 5G radios. The XC16 is a complete C-programmable DSP with an ISA that includes FIR filters as well as many other signal-processing functions, so an accelerator such as NNMax isn't comparable.

Designers needing a dedicated programmable DSP can choose from numerous Cadence, Ceva, and Synopsys products. NNMax requires connection to a host CPU, and it's more suitable for designs that require a single DSP function, such as an audio filter. Flex Logix plans to develop a DSP compiler that maps Matlab output to NNMax. The DLA's ability to serve as both a DLA and a DSP accelerator will be attractive to some customers, and the initial performance looks promising. But for NNMax to be useful beyond FIR filters, the company must develop a more comprehensive function library, which will likely require new architectural features as well. ♦