IP-core of the accelerator of tensor calculations based on the systolic array

Osipenko P.N., Uvkin I.V.

LLC IVA Technologies. 109316, Moscow, Volgogradskiy Avenue, 43, bild.3

At present, the problem of calculating artificial neural networks is solved, in most cases, using graphics accelerators. An alternative solution is dedicated processors such as Google's TPU series of processors, which, however, are not intended for distribution. This report describes an original configurable IP-block of tensor calculations to accelerate the computation of artificial neural networks (IVA_TPU), intended for embedding in the SoC.

The characteristics of the IP block and the parameters of the developed chips based on it are given. The description of the programming design-flow of IP block IVA_TPU is given.

The IP block IVA_TPU, developed by Hitech LLC, is designed to speed up the calculation of artificial neural networks when embedded in the SoC.

The IP block EVA_TPU contains the following blocks: Control Unit, Memory System, Shared Memory, Vector Engine, and Matrix Engine.

Standard AXI4 and AXE-Light buses are used for integration in SoC. The block diagram of the IP block is shown in figure 1.

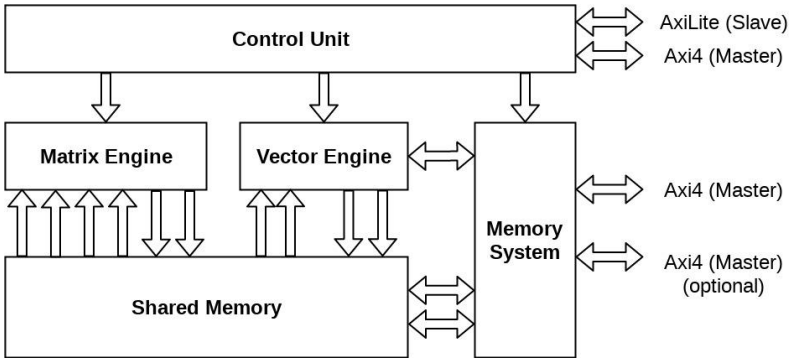

Figure 1. Block diagram of the IP block IVA_TPU

The main computing load is placed on the matrix processor block (Matrix Engine), which is built as a systolic array of processor elements. The size of the systolic array is a configurable parameter and can reach 65K processor elements, each of which is capable of performing a multiply-accumulate operation with data in int8 format every clock cycle.

To work with the IP block IVA_TPU, a software package has been developed that includes a compiler and a set of utilities that allow you to accept trained neural networks in Tensor Flow format as input. Work is underway to provide support for the PyTorch, Keras, ONNX, and MXNet network formats.

Developed an FPGA version of the IP block IVA_TPU, which is accessible remotely. Potential users use the FPGA boards to evaluate the performance parameters of their neural networks.

To evaluate technical solutions, two 28nm SoCs were developed with the size of a systolic array of 64*64 (mobile processor) and 128*128 processor elements (server processor). The estimated operating frequency of the IVA_TPU block is 1GHz, which gives a peak performance of 8tops for the mobile processor and 32tops for the server processor. Production of chip samples is expected by the end of 2020.

Demo modules in COM Express and PCIe-card format are being developed. Production of demonstration modules is expected in Q1 2021.