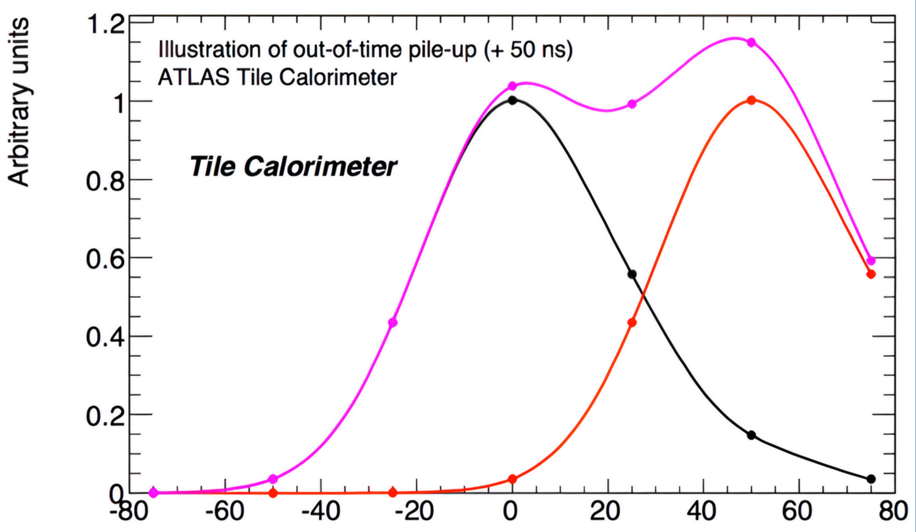
Versal ACAP processing for ATLAS-TileCal signal reconstruction

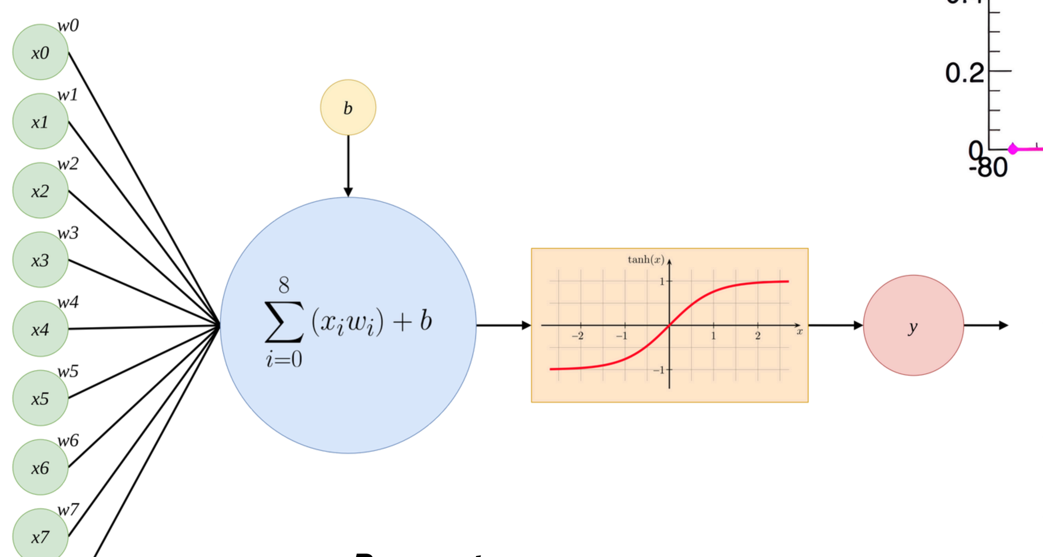
Francisco Hervás, Luca Fiorini, Alberto Valero, Hector Gutiérrez
IFIC (Universitat de València - CSIC)
HIGH-LOW TED2021-130852B-I00

INTRODUCTION

Particle detectors in accelerators generate large amounts of data that need processing and analysis. A challenge arises with signal pile-up, where multiple particles generate signals in the same sensor during collisions. This overlap complicates identifying individual signals, leading to the loss of several energy pulses.



Signal pile-up

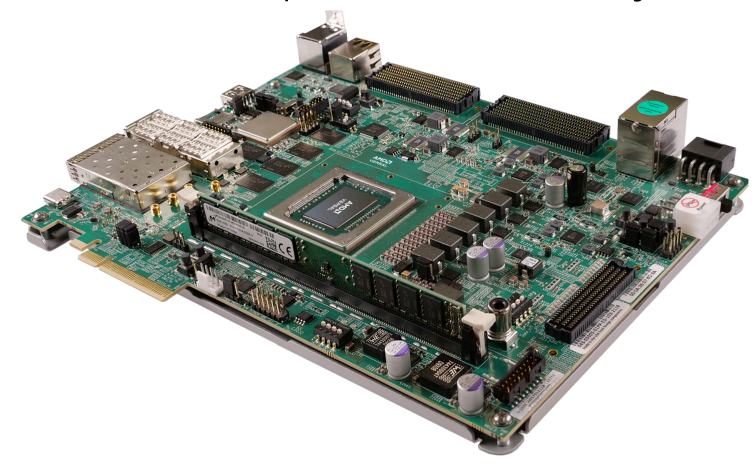


Perceptron

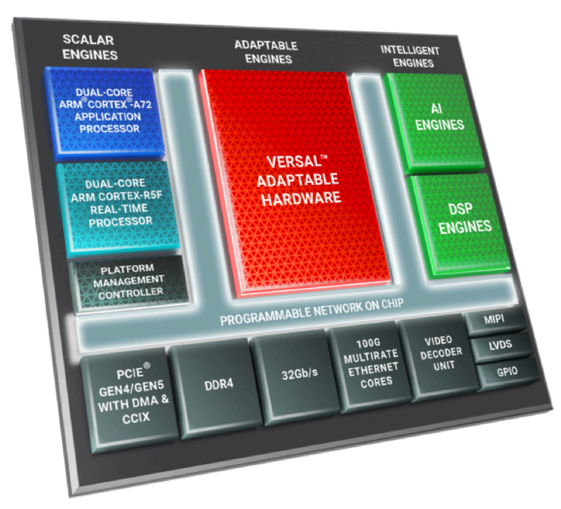
Deep learning algorithms are required to filter signal pile-up and detect individual energy pulses. A Perceptron, a type of Artificial Neural Network (ANN), consists of a single neuron that performs a weighted sum of inputs with a bias, then passes the result through an activation function.

METHODS

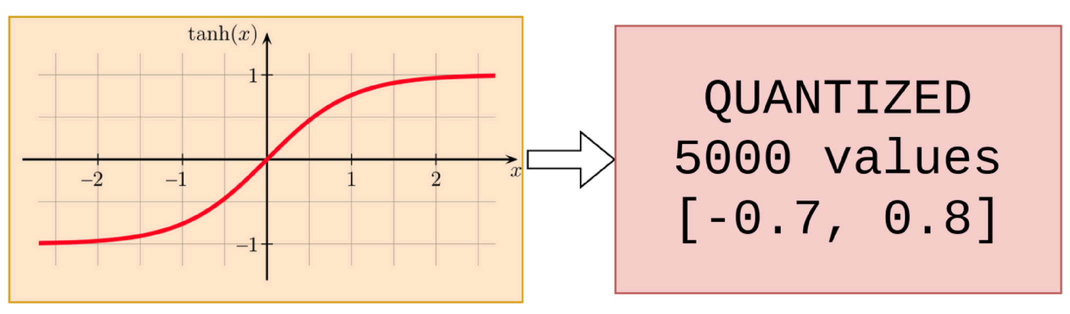
Field-Programmable Gate Arrays (FPGAs) are ideal for implementing deep learning algorithms because they perform parallel mathematical calculations, unlike traditional CPUs, which operate sequentially.The system is implemented using the Versal AI Core VC1902 [3], a System on Chip (SoC) that combines a traditional CPU (Processing System), FPGA (Programmable Logic), and transceivers for external communication. This integration enhances overall performance. The device is integrated in the VCK190 [4] development board made by AMD-Xilinx.



VCK190 Development Board



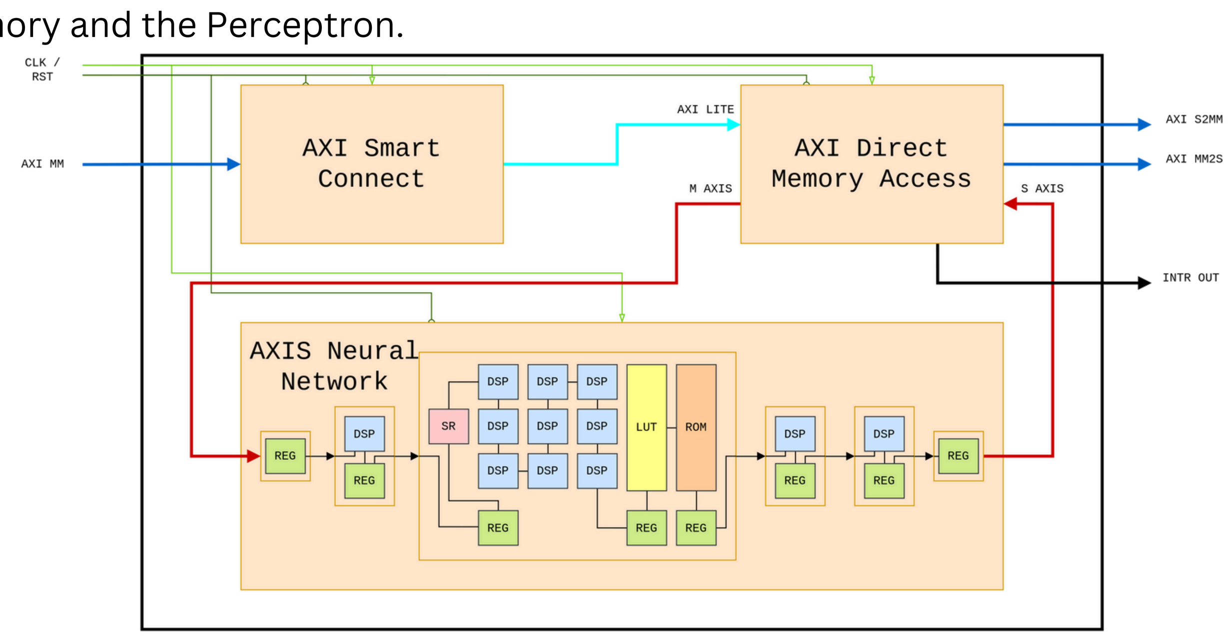
Versal ACAP Architecture



Hyperbolic tangent quantization

To implement the Perceptron's non-linear activation function, a hyperbolic tangent (tanh(x)) is required. However, implementing non-linear functions on an FPGA is resource-intensive. To address this, a quantized version of tanh(x) is used via a Lookup Table (LUT) for efficiency, with 5000 discrete values in the range of -0.7 to 0.8 [1].

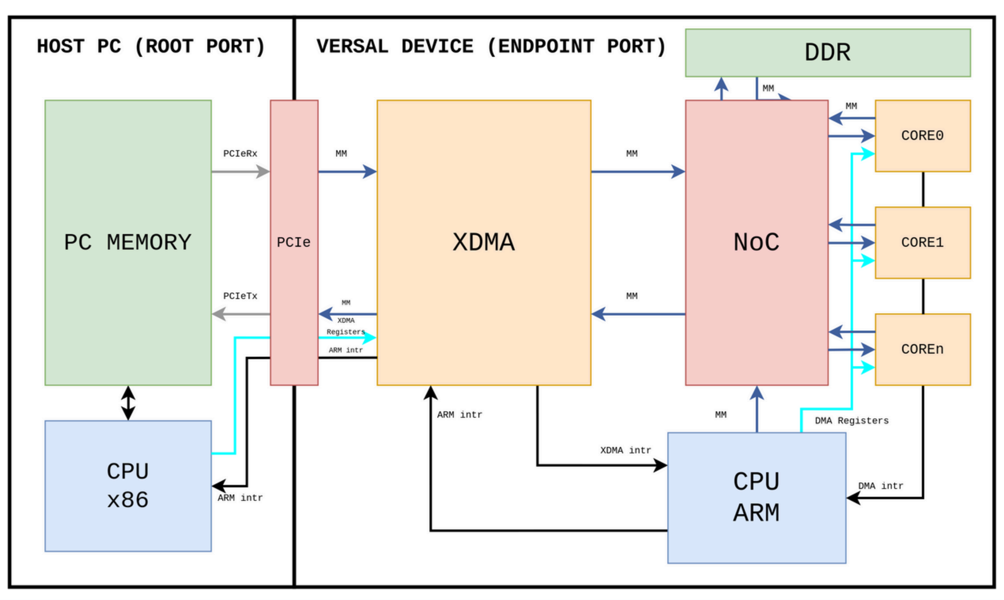
The firmware for the system is developed in VHDL, a Hardware Description Language used to define the logic connecting the FPGA's internal circuits. It features a processing unit based on the Perceptron algorithm, utilizing registers, Digital Signal Processor (DSP) slices, Lookup Tables (LUTs), and Read-Only Memories (ROM). The replicated Neural Network Core includes the Perceptron unit along with an AXI [2] Smart Connect, which translates between AXI Memory Map and AXI Lite protocols, and an AXI Direct Memory Access (DMA) unit that facilitates data transfer between DDR memory and the Perceptron.



Neural Network Core

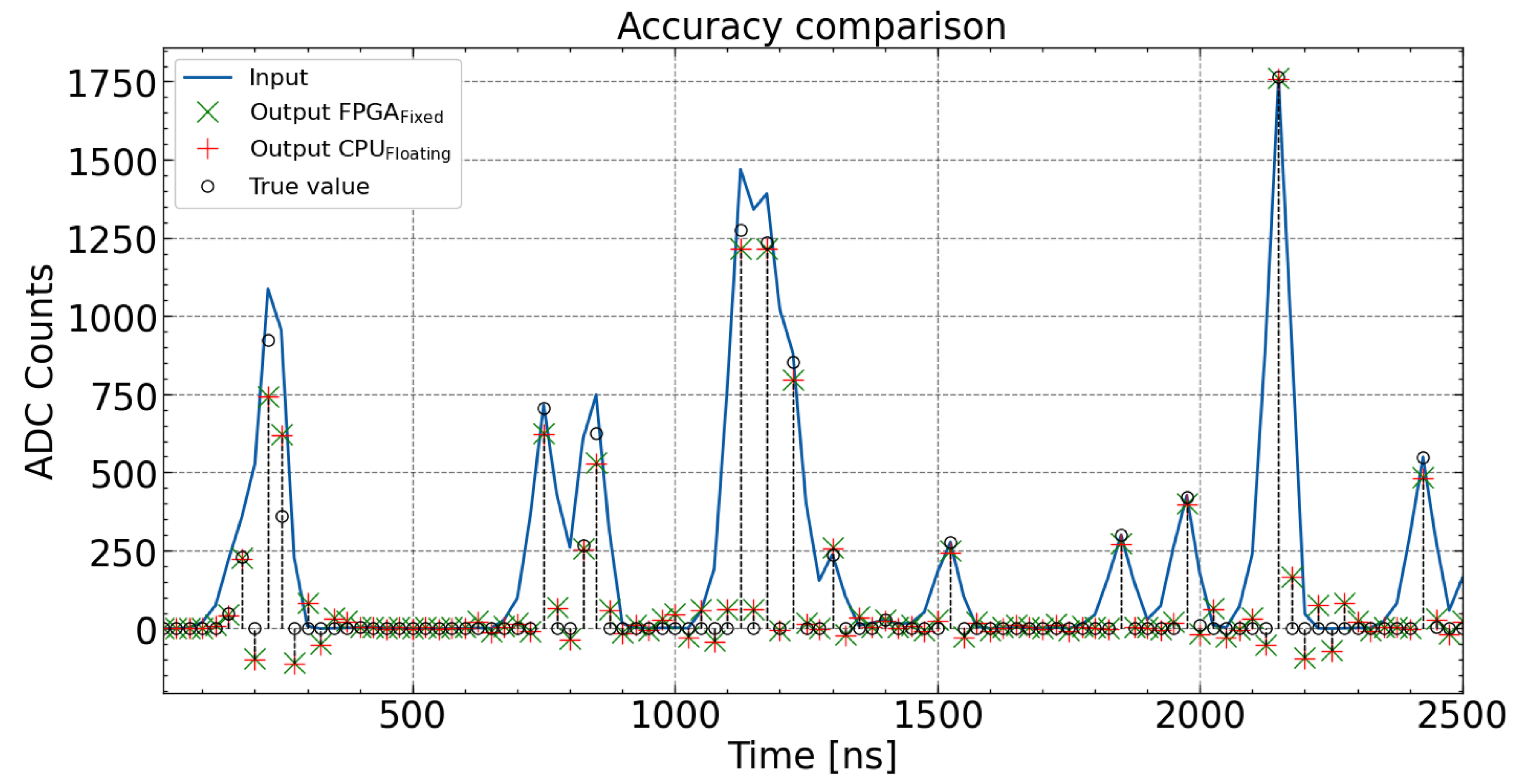
The Neural Network Cores will be distributed across the FPGA (PL side) of the VC1902, creating a network of parallel Perceptrons that efficiently process a large volume of events, thereby reducing processing time and power consumption. The overall system comprises several components: Neural Network Cores, DDR memory for buffering, a high-bandwidth Network on Chip (NoC), an ARM CPU for managing internal processes, an XDMA for PCIe data transfer, and a host PC with an x86 CPU and memory. Together, these elements facilitate effective data handling and processing across the Versal platform.

Complete system integration

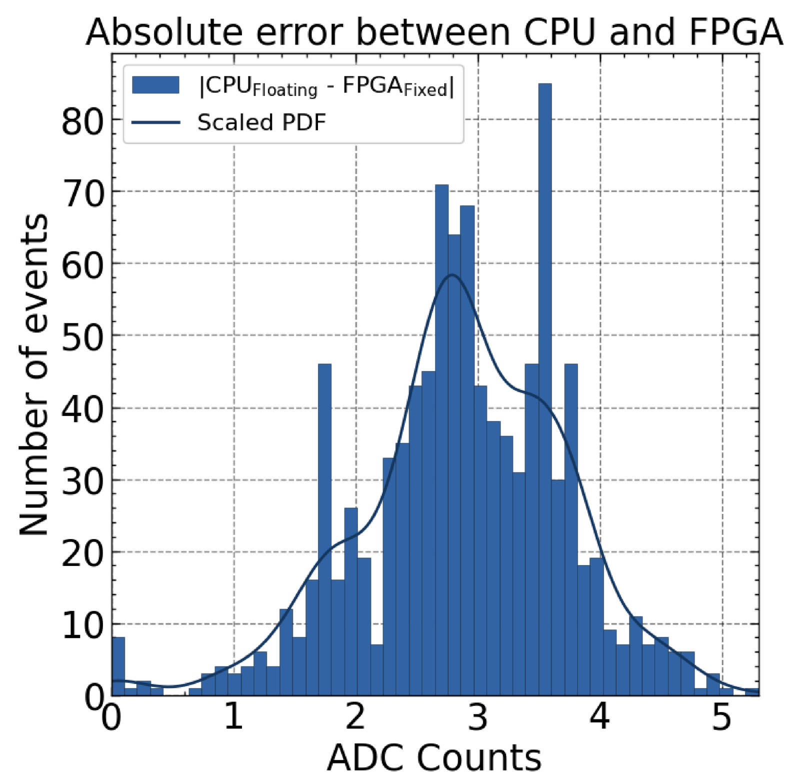


RESULTS

The results obtained for these experiment are shown in the following figures. The first one shows the accuracy comparison between the output data from the FPGA coded with Fixed Point, where the number of bits used for the integer and for the fraction part are limited, and the output data from the CPU coded with Floating Point, where the accuracy is better. Both results are compared with the input signal of the detector and with the true value, that is the value that should be expected to be the true energy of the detector.



Event accuracy comparison



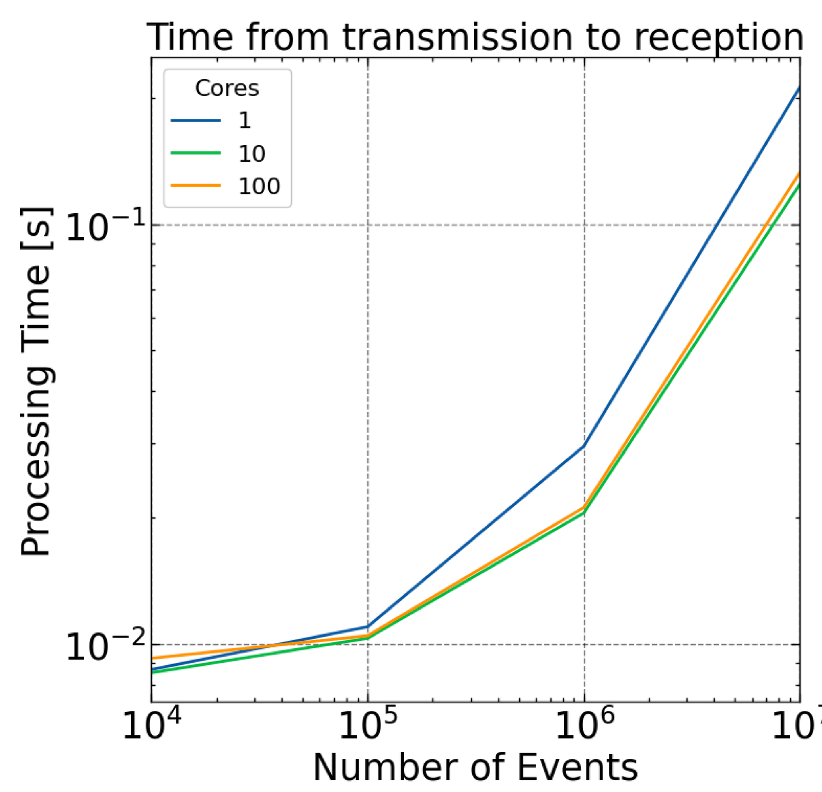
Absolute error between CPU and FPGA

The following figure represents the number of events that is in every bin of error of the histogram for a 1000 sample test.

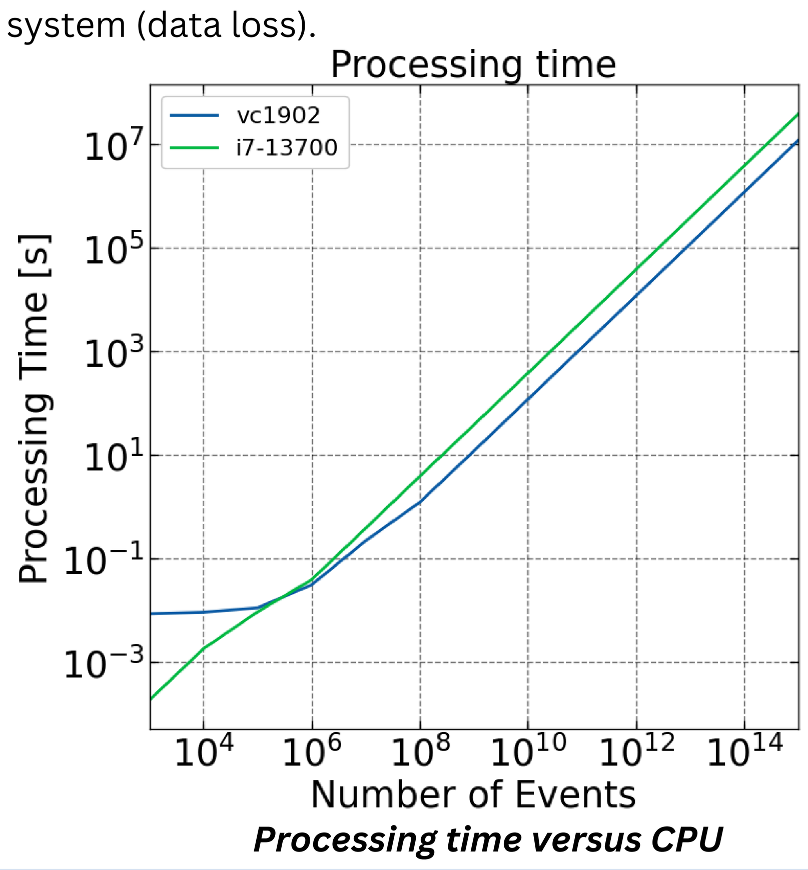
The error is calculated by the difference in the result in CPU (Floating Point) and the FPGA (Fixed Point). There should be appreciated that the maximum error achieved is 5 ADC Counts, and the typical error is 3 ADC Counts.

The following figure shows the time elapsed during the transmission of the data from the host PC to the device, which includes the processing time. This time is obtained from different scenarios, changing the number of Neural Network Cores used in each one.

There should be appreciated that using 10 cores improves the processing time in relation to using 1 core, but increasing to 100 cores does not improve practically nothing. This happens because the number of optimal cores is going to depend by the number of acceses that the PL side has to the DDR memory, made over the NoC. In the case of the VC1902, this number of access is 14. Having more of this access implies channel multiplexing in the PL, which means introducing backpressure to the system (data loss).



Transmission time over multiple cores



Processing time versus CPU

The following figure shows the time elapsed during the processing of the data, made for 1 core in the FPGA and for the CPU. For this test a large amount of data has been used. The result for more than 10^{12} events has been extrapolated from previous results because the time used for the processing scales exponentially.

It can be appreciated that for less than 10^6 events, the CPU is better because the access to memory is integrated in the device, and it is not necessary to access an external peripheral unlike the FPGA, that has an intermediate buffer for accessing the data. However, increasing the number of events makes this time insignificant, compared with the time gained in the processing. For 10^{12} events, the time is improve by 7 hours in favor of the FPGA.

REFERENCES

[1] Ortiz Arciniega, J. L., Carrió, F., & Valero, A. (n.d.). FPGA implementation of a deep learning algorithm for real-time signal reconstruction in radiation detectors under high pile-up conditions.

[2] ARM, AMBA AXI Protocol Specification

[3] Versal Adaptive SoC Technical Reference Manual (AM011)

[4] VCK190 Evaluation Board User Guide (UG1366)