FPGA based Accelerator Design, Assignment 1

Monsoon 2021, IIIT-H
Suresh Purini and Padmini Gopalakrishnan

1. What is the FPGA used on the Amazon F1 instance? List down the important available hardware resources such as the number of LUTs, Flip Flops, DSPs, BRAM block, device memory size etc. Similarly list down the host CPU configuration like processor, clock frequency, main memory size, cache size etc.

2. Find the PCIe bandwidth to the FPGA device on the F1 instance using the XRT command *xbutil*.

3. For the *vadd*, *wide_vadd* programs from the Vitis Tutorials repository, plot the following metrics in the form of graph.

   (a) For increasing vector sizes ($N = 2^{10}, 2^{11}, 2^{12}, \cdots$), find the kernel computation time and total communication time. Plot CPU vector addition time with FPGA vector addition time (include both computation and communication cost).

   (b) Repeat the above, by replacing the add operation with floating-point multiplication.

4. Repeat the above problem using the Vitis tutorial Example 05 (https://github.com/Xilinx/Vitis-Tutorials/blob/2021.1/Hardware_Acceleration/Introduction/) wherein we overlap computation and communication. Also, for CPU vector addition parallelize the vector addition using OpenMP pragma. Compare FPGA performance with the parallelized CPU vector addition (refer Example 06 on the Tutorial).

5. Assume that you have a database of N vectors of size 256 each. Each element of the vector is a floating-point number. Given a query vector q, we need to find the vector which is closet to the query vector with respect to the cosine similarity measure. Assume the all the vectors are normalized. Design a system which maximizes the query throughput. You should provide a detailed analysis of latency, throughput, hardware resource utilization etc.

   Extra Credit: Perform a roofline analysis. Check if your final version is compute-bound or memory bound.