Qishen Li's Scribe

This paper talk about Near-DRAM Acceleration using FPGA and HLS to accelerate the process of neural networking. Different from homogeneous system, heterogeneous system could accelerate specific applications more effectively. The accelerator the paper talks about could access to memory directly, dealing with the data processing prior to the main processor. It could release the pressure of the main memory. But heterogeneous structure also means it demands high DRAM bandwidth for the computation and therefore the NDA conception comes out in the paper.

The conception of NDA is pretty similar to the previous "Tetris" paper. The accelerator could be in the same chip as DRAM. They could in the different layers of chip. Because the physical distance between DRAM and accelerator is much closer than traditional ways such as PCI-E or something else, it could have very high performance. Also, because they are in the same structure, the cost of bandwidth could be reduced a lot. This design also provides high level of parallelism since for every chip it could have one accelerator, the effect of reducing the process time could be higher when there are lots of DRAM chips for a single memory.

For the combination of FPGA and HLS, the accelerator could be customized for specific applications. The paper uses Stratix V and DDR3 DIMMs for the experiment. The speed is much faster than Intel Xeon. Also, because of FPGA, the customer could customize the different types of accelerator for different applications and this characteristic could save the cost and enhance the overall performance.

But the problem the professor mentioned is the comparing target. Comparing with Xeon's performance is not persuasive because for nowadays neural network uses GPU not CPU to accelerate the data. I also have a question for this paper. The paper only talks about FPGA solution, what about ASIC solution? Could it gain much faster speed or not? I think this paper needs more experiment for certifying the effectiveness of this accelerator.

Discussions:

During the class, the class mentioned that paper did not say what the data buffer in the DDR4 was originally intended for before they repurposed it. Also, the class also suspects whether it is simple enough for the C++

code to transfer to the hardware. This paper may need further research.

For the discussion on Canvas, Casey suspects that it may not be practical because it needs sometime for FPGA to translate C/C++ language and it is not convenient to change the accelerator immediately. Also, Casey cares about the specific cost of the solution mentioned in the paper. Mark & MengJu Yu asked Why can increasing the number of memories not increase the bandwidth of channels. Sam & Karpur asked why this NDA structure is desirable & beneficial.

Material:
After the discussion to professor Bahar, the No.6 & No.7 paper in the reference is the good extension material for this paper which introduces NDA & ConTutto structure in details.