Contribution ID: **71** Type: **not specified**

Co-Design for Efficient & Adaptive ML

Wednesday 27 September 2023 09:00 (45 minutes)

Beyond the well-known highlights in computer vision and natural language, AI is steadily expanding into new application domains. This Pervasive AI trend requires supporting diverse and fast-moving application requirements, ranging from specialized I/O to fault tolerance and limited resources, all the while retaining high performance and low latency. Adaptive compute architectures such as AMD FPGAs are an excellent fit for such requirements but require co-design of hardware and ML algorithms to reap the full benefits. In this talk, we will cover a breadth of co-design techniques, including their merits and challenges, from streaming dataflow architectures to quantization, from sparsity to full circuit co-design. By combining such techniques, we can enable nanosecond-latency and performance in the hundreds of millions of inferences per second. The proliferation of this technology is enabled via open-source AMD tools such as FINN, Brevitas and LogicNets, as well as the AMD-FastML collaborative project QONNX.

Presenter: UMUROGLU, Yaman

Session Classification: Invited Talks

Track Classification: Invited Talks