Contribution ID: **83** Type: **not specified**

Intel® FPGA AI Suite and AI Tensor Blocks: Empowering Real-time, Low-Latency, and Low-Power Deep Learning Inference with Intel FPGAs

Thursday 28 September 2023 13:00 (2 hours)

This two-part tutorial presents an update on Intel HLS flow and the Intel FPGA AI Suite. In the first part, we will have a 30-minute update on how the latest oneAPI tool flow for IP authoring works. In the second part we will present Intel FPGA AI Suite and groundbreaking AI Tensor Blocks newly integrated into Intel's latest FPGA device families for deep learning inference. These innovative FPGA components bring real-time, low-latency, and energy-efficient processing to the forefront, supported by the inherent advantages of Intel FPGAs, including I/O flexibility, dynamic reconfiguration, and long-term support. We delve into the Intel FPGA AI Suite, demonstrating its flexibility in achieving scalable performance and seamless integration with industry-leading frameworks like TensorFlow and PyTorch, facilitated by Quartus Prime Software. Moreover, we highlight the game-changing role of AI Tensor Blocks in enhancing deep learning inference performance. This tutorial offers both theoretical insights and practical experiences, equipping participants to leverage these advancements and revolutionize FPGA-based AI applications.

Presenters: AHMAD, Jahanzeb; DEMIRSOY, Suleyman