Contribution ID: **52** Type: **Oral (16mins + 4 mins)**

Leveraging Vendor Tools for AI Acceleration

Wednesday, March 6, 2024 11:50 AM (20 minutes)

Several large vendors have been expanding their ML deployment tooling to allow for easy deployment of machine learning models on processing devices. AMD Xilinx has developed a toolkit for accelerating ML calculations on their FPGAs by utilizing either dedicated "AI Engine"(AIE) hardware or an openly available IP block known as "Deep Learning Processing Units"(DPUs). Vitis AI is actively maintained, with regular version releases. Google has created extensions to TensorFlow to compile models to programs that can run on dedicated accelerators, such as those provided by Coral.ai, while also allowing for other deployment methods. This is known as TensorFlow Lite, part of the TensorFlow ecosystem and TFX deployment pipelines. This toolchain is able to compile for ARM, Xilinx, and GPUs. Presented here is the use of one of these toolchains to develop a laser focal position controller at LBNL's BELLA facility, including a discussion of future plans for easing control and deployment needs for such a system.

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, under Award Number(s) DE-SC0021680 and Prime Contract No. DE-AC02-05CH11231.

Primary Keyword

AI-based controls

Secondary Keyword

MLOps

Tertiary Keyword

Primary authors: EINSTEIN-CURTIS, Joshua (RadiaSoft LLC); COOK, Nathan (RadiaSoft LLC)

Co-authors: BERGER, Curtis (Lawerence Berkeley National Lab); VAN TILBORG, Jeroen (Lawerence Berkeley National Lab); EDELEN, Jonathan (RadiaSoft LLC); NAGLER, Rob (RadiaSoft LLC); BARBER, Samuel (Lawerence Berkeley National Lab); COLEMAN, Stephen (RadiaSoft LLC)

Presenters: EDELEN, Jonathan (RadiaSoft LLC); EINSTEIN-CURTIS, Joshua (RadiaSoft LLC)

Session Classification: Infrastructure / Deployment Workflows

Track Classification: Infrastructure / Deployment Workflows