Machine Learning Accelerator Engineer: FPGA Hardware and Software Stack

Type of Position: Full time/Intern

Achronix is a privately-held fabless semiconductor company based in Santa Clara, California, that offers high-performance FPGA solutions. Our history is one of pushing boundaries, creating new markets, and offering innovative solutions to our customer's most challenging problems. Our target industries include 5G wireless infrastructure, network switching, as well as edge device and datacenter compute acceleration. Our product offerings include embedded FPGA fabrics IPs, high-performance and high-density packaged FPGAs with hardened system-level interfaces, data center and HPC hardware accelerator boards, and best-in-class EDA software. The Achronix CAD Environment (ACE) is the software tool used by our customers to synthesize, map, place-and-route, and program our FPGAs. Achronix Software Engineers research and develop novel and computationally hard software algorithms to enable our programmable hardware. ACE is the window to our FPGA technology.

Job Description/Responsibilities

You will work on a small team driving the development of a complete software and hardware stack for an FPGA-based machine learning inference accelerator card reference platform. You will adapt existing open-source and university software when possible, and develop new software from scratch as needed, to assemble a complete full-stack end-to-end software solution. You will work closely with sales and marketing, systems engineers, EDA tool developers, and FPGA architects. The resulting system will support diverse internal and external use models including FPGA fabric optimization, accelerator micro-architecture exploration, memory subsystem design, place & route software verification, system prototyping, pre-sales demonstrations, and customer deployment and scaling.

Prior experience is required working with an existing open-source or proprietary machine learning accelerator architecture such as OpenTPU, NVDLA, Eyeriss, or VTA. You must have a background in compiler hacking on one or more of TVM, Glow, Halide, Spatial, XLA, CLANG, LLVM, or GCC. Experience desired with compiler Intermediate Representations (IRs) and back-ends; JIT compilers; kernel-mode and user-mode Windows, Unix, or embedded systems runtime environments and device drivers. Familiarity is desired with machine learning frameworks such as TensorFlow, PyTorch, Caffe2, Keras and MXNet; domain-specific languages such as Halide and Spatial; and with common DNN models such as AlexNet, ResNet50, Inception, YOLO, RNN, and LSTM.

Skills:

- 2+ years of work or educational experience in machine leaning accelerator architectures, micro-architectures, and compilers.
- Skilled practitioner in C++, Python, and Verilog. Familiarity with System-C, System Verilog, HLS, Catapult-C, or Chisel.
- Experience in two of more of the following categories:
    - Machine learning accelerators such as OpenTPU, NVDLA, Eyeriss, and VTA.
    - Machine learning frameworks such as TensorFlow, PyTorch, Caffe2, Keras, or MXNet.
    - Common DNN models such as AlexNet, ResNet50, Inception, YOLO, RNN, orLSTM.
    - Domain-specific languages such as Halide or Spatial.
    - Compilers such as TVM, Glow, MLIR, Halide, LLVM, or GCC.
    - Embedded system runtime environments and device drivers.

Education:

MS or Ph.D. in Computer Science, Computer Engineering, Electrical Engineering, Applied Math, or Physics