Inference Software Developer

Flex Logix develops industry-leading AI inference engines and we are the #1 provider of eFPGA solutions. Our InferX X1 is the industry's fastest and most-efficient AI edge inference accelerator that brings AI to high-volume applications, surpassing competitor's performance at 1/7th size and much lower price. InferX X1 is available as a chip, PCIe board and M.2 board. InferX X1 is working and running YOLOv3 today and starts production shipments this summer.. Our Inference Compiler is easy to use (we take in neural network models in TensorFlowLite and ONNX) and our APIs for Infer X1 allow rapid integration by the customer. Our architecture is covered by dozens of patents and is highly differentiated giving us a sustainable competitive advantage.

We recently completed a $55M funding round.

Flex Logix is seeking Inference Software Developers to join our team developing state-of-the-art compilers and run time software for InferX X1. This is an exceptional opportunity to develop the technology that breaths life into AI inferencing solutions targeting systems in medical, industrial, automotive and other Enterprise edge applications.

Responsibilities

Part of the excellent team responsible for our InferX Model Compiler: a DNN Model-to-binary flow:

- Expanding functionality of our Model Compiler, written in modern C++, for support of additional
  capabilities, in particular:

- Parsing of TensorflowLite/ONNX/other DNN model description languages to our internal model format

- Mapping of required computations from DNN model in TFLite/ONNX to Verilog RTL code, running on
  EFLX eFPGA inside the InferX X1 chip.

This is a software developer role but you need to understand computer architecture and digital logic, as Model Compiler produces Verilog code for eFPGA, which controls nnMAX computational blocks and memory connections.

Read more about InferX on our Inference page at www.flex-logix.com

EXPERIENCE AND SKILL REQUIRED

BSCS/MSCS/BSEE/MSEE/BSCE/MSCE with courses in software/computer architecture/digital logic & 3+ years of relevant industry experience. AI/NN expertise is not required but experience is preferred in TensorFlowLite, Multi-core programming, and/or Windows and Linux, C++ and FPGA synthesis tools such as Synopsys Synplify.

Understanding of software/hardware development engineering practices, people with previous internships and/or projects are preferred.

Must be very smart and very motivated, must be a quick learner, proactive and curious.
Must be passionate about being part of an aggressive, venture-backed startup team that is changing the

way chips are architected, designed, and programmed

Must be entrepreneurial, innovative problem solver and willing to work hard.

Must live in Silicon Valley or Austin TX.. Strong preference for US citizenship or permanent residency ("green card"); will consider candidates with current H1-B visas who are willing to transfer promptly.