# AI Inference Software Architect

Flex Logix has finished the hardware design and is fabricating it's first Inference Accelerator Co-Processor, InferX X1, which is based on our nnMAX Inference IP. We will have chips and PCIe boards this autumn. Our software team is preparing our Inference Model Compiler to be ready to run deep neural network models on X1.
We have begun architecting the follow-on chip. InferX has industry-best inference efficiency: more inference throughput per $ and per watt. We excel on larger models and megapixel images, but can run any neural network.

## RESPONSIBILITIES

Part of the small but excellent team responsible for our nnMAX Model Compiler: a DNN (Deep Neural Network) Model-to-binary flow - we are looking for a software architect-level position to expand architecture of our Model Compiler, written in modern C++, for addition of functionality for support of additional capabilities, in particular:

- Parsing of TensorflowLite/ONNX/other DNN model description languages to our internal model format, support of custom-defined operators
- Consider numerous parameters (memory bandwidth, memory access pattern, memory & compute resource allocation, etc.) to arrive at an optimal computation strategy for each new operator
- Mapping of each operator and its computational strategy to Verilog RTL code, running on EFLX eFPGA inside nnMAX chip and controlling Flex computation & memory engines.

This is a software architect/developer role but your activities will include integration with Flex-Logix hardware-based computation architecture for DNN as well as representation of computational architecture in software abstractions.

## EXPERIENCE AND SKILL REQUIRED

Expertise in developing software compilers for one or more hardware-accelerated computational engines, preferably for DNN training or inference
Abilities to take complex problems and come up with efficient, innovative solutions
Experience with modern AI frameworks and inference engines – TensorFlow, PyTorch, TfLite, ONNX
Experience with Verilog and hardware computation architectures
Experience with FPGA synthesis tools such as Synopsys Synplify is a plus
Very good grasp of proper software/hardware development engineering practices, modeling, design, representation in documentation, excellent communication skills.
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 and supporting software are architected, designed, and programmed.
Must be entrepreneurial, innovative problem solver and willing to work hard.
Must live in Silicon Valley. Strong preference for US citizenship or permanent residency ("green card"); will consider candidates with current H1-B visas who are willing to transfer promptly.

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