InferX™ Edge Inference SDK

Deep Learning Software Suite

flexlogix AI + eFPGA ®

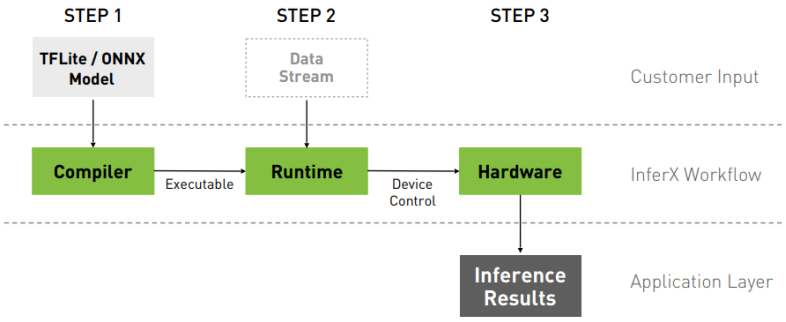
Datacenter Acceleration at Edge Power and Price

Using the InferX Edge Inference SDK is simple and easy; our compiler finds an optimal hardware mapping for each layer of your neural network without making any modifications to your model, and the compiled binary can then run on any InferX X1 chip or board using our runtime environment.

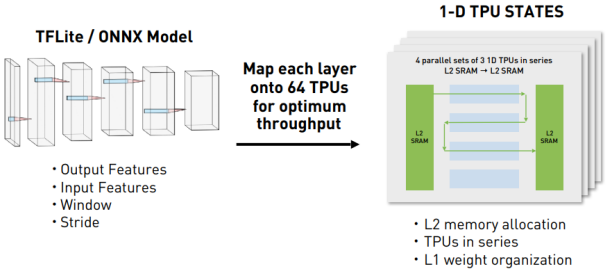
Workflow Overview

ONNX

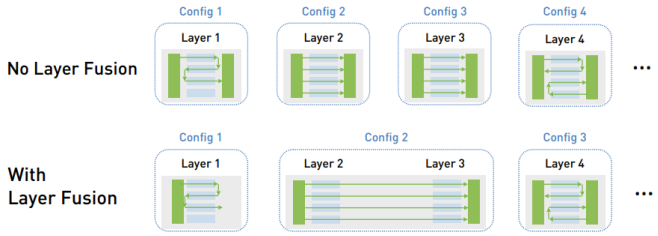
TensorFlow Lite



Compiler



The first part of the process is to parse the model and determine the set of configuration states for the X1.



Next, the compiler fuses configurations in order to maximzime compute utilization and minimize latency of inference.

Runtime

Features

Support for models defined in Tensorflow Lite and ONNX

Precompile-time performance analysis

Ubuntu, CentOS, Windows driver support

Benefits

- Leverage existing machine learning ecosystem
- Lowers the effort of application development
- Determine performance bottlenecks early on
- Wide end-device coverage