

Graph-level Optimizations

NNVM IR

Tensor Optimizations

TVM IR

JIT Compilation

Runtime API

Prgramming Interface

VTA ISA

Hardware Implementation

VTA Architecture

