Dynamic Neural Accelerator® F-series

EDGECORTIX

DNA-F-Series – Low latency AI Inference Engine for ALVEO Edge Servers

INTRODUCTION

DNA-F series, part of the Dynamic Neural Accelerator IP family, is the ideal solution for low-latency and power efficient Deep Learning compute engine for neural network inference on FGPAs with streaming data.

Complemented by the **EDGECORTIX MERA^{TM}** dataflow compiler stack, this dedicated AI engine enables deep learning engineers to use Xilinx ALVEO^{TM} series of FPGA boards as drop-in replacement for standard CPUs or GPU, without leaving the comfort zone of standard deep learning frameworks like PyTorch or TensorFlow. DNA-F100 and DNA-F200 are dedicated bitstreams optimized for ALVEO U50 that provide significantly lower batch-1 inference latency and higher power efficiency compared to other general-purpose hardware.

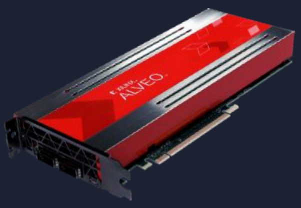
PRODUCT HIGHLIGHTS

- Dataflow architecture-based AI inference engine
- Optimized for streaming data (batch size 1)
- Integer 8-bit precision optimized - post-training quantization
- Easily deploy existing neural network models developed for CPUs or GPUs with DNA-F-series bitstream on the ALVEO U50
- Native support for PyTorch machine learning framework
- Built-in hardware interpreter and simulator
- MERA compiler automatically optimizes scheduling of deep neural network operators between host CPU and ALVEO FPGA.
- XILINX Vitis support

XILINX

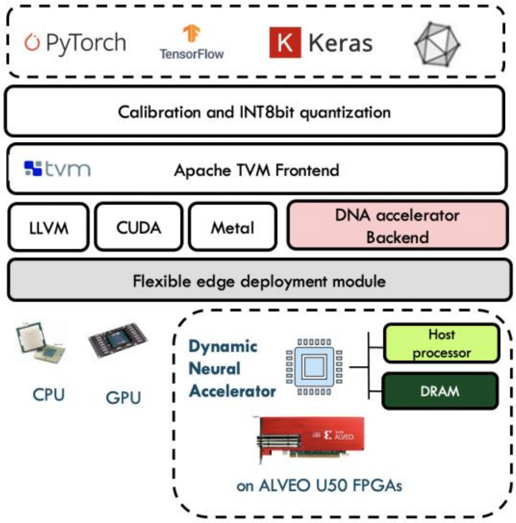
SOLUTION BRIEF

XILINX ALVEO™



- Seamlessly replace or complement CPUs/GPUs
- Ultra low latency
- High energy efficiency

Compile with EDGECORTIX MERA Compiler



PyTorch TensorFlow Keras

Calibration and INT8bit quantization

tvm Apache TVM Frontend

LLVM CUDA Metal DNA accelerator Backend

Flexible edge deployment module

CPU GPU

Dynamic Neural Accelerator

Host processor

DRAM

on ALVEO U50 FPGAs

Adaptable. Intelligent.

SOLUTION OVERVIEW

EDGECORTIX DNA-F series Deep Learning computer engine is optimized for the XILINX ALVEO^{TM} U50 FPGA board and is shipped as ready to use bitstreams. Docker containers are provided for easy deployment of neural network inference on NIMBIX cloud or on premises. Comes with MERA^{TM} compiler stack pre-provisioned with docker containers, enabling seamless compilation and execution of standard or custom developed convolutional neural networks (CNN).

FEATURES

Diverse Operator Support

> Standard and depth-wise convolutions

> Stride and dilation

> Symmetric/asymmetric padding

> Max pooling, average pooling

> ReLU, ReLU6, LeakyReLU, and H-Swish

> Upsampling and Downsampling

> Residual connections, split etc.

Drop-in Replacement for GPUs

> PyTorch and TensorFlow supported

> No need for retraining

> Supports high-resolution inputs

INT8 bit Quantization

> Post-training quantization

> Support for ML framework built-in quantizer

> High accuracy

PERFORMANCE COMPARISON WITH STANDARD CNNs*

	MOBILENET-V2	RESNET 50	MOBILENET v2-SSD	MOBILENET v2-SSD	Tiny Yolo v3	Yolo V3
Input size	224x224	224x224	640x480	1920x1080	416x416	416x416
DNA-F100 290 MHz	4.3 ms	12.2 ms	12.3 ms	27 ms	14.9 ms	46.2 ms
DNA-F200 300 MHz	2.6 ms	7.7 ms	9.2 ms	23 ms	11.3 ms	40.7 ms

* Performance numbers are based on end-to-end batch-1 inference latency (milliseconds) measured with U50 on-premise. Performance measured on Nimbix cloud may slightly vary

TAKE THE NEXT STEP

Learn more about EDGECORTIX Dynamic Neural Accelerator® IP series and the MERA™ compiler (www.edgecortix.com)
Ask about DNA IP licensing or sales dnap-ip@edgecortix.com

Dynamic Neural Accelerator, MERA, EDGECORTIX are registered trademarks of Edgecortix Inc.

XILINX