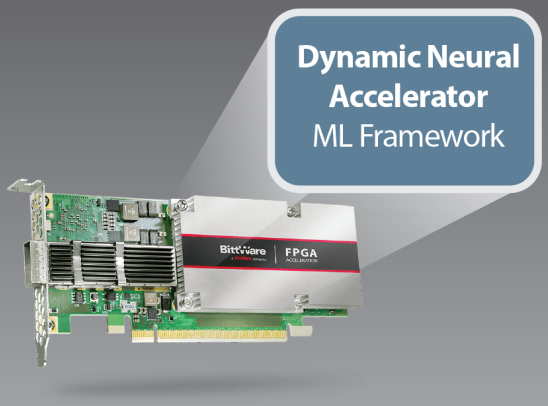
BittWare
a molex company

Partner Framework

EDGECORTIX

Dynamic Neural Accelerator
ML Framework



High-Efficiency Inference Using Standard ML Frameworks

EdgeCortix Dynamic Neural Accelerator (DNA), is a flexible IP core for deep learning inference with high compute capability, ultra-low latency and scalable inference engine on BittWare cards featuring Agilex FPGAs.

Specially optimized for inference with streaming and high resolution data (Batch size 1), DNA is a patented reconfigurable IP core that, in combination with EdgeCortix's MERA™ software framework, enables seamless acceleration of today's increasingly complex and compute intensive AI workloads, while achieving over 90% array utilization.

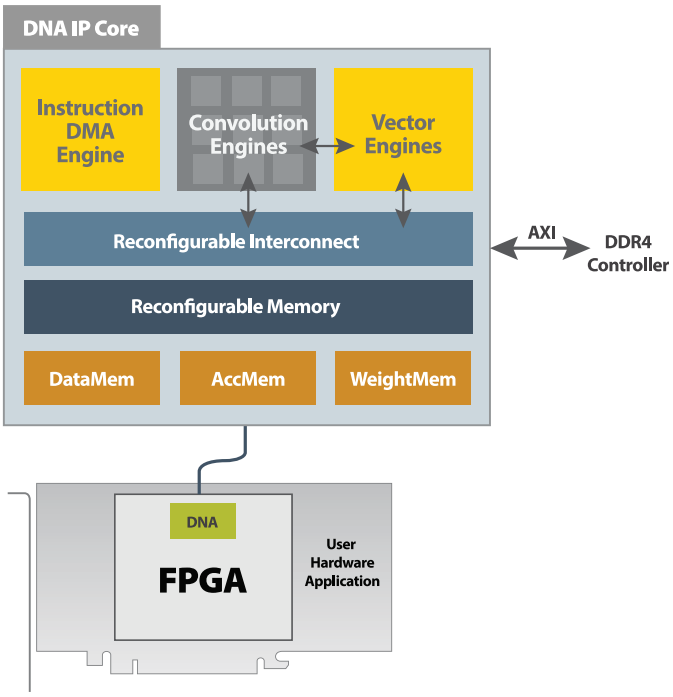
Complemented by the MERA framework that provides an integrated compilation library and runtime, this dedicated IP core enables software engineers to use the Bittware IA-840F and IA-420F FPGA cards as drop-in replacements for standard CPUs or GPUs, without leaving their comfort zone of standard frameworks like PyTorch and TensorFlow. DNA bitstreams for Agilex provides significantly lower inference latency on streaming data with 2X to 6X performance advantage compared to competing FPGAs, and better power efficiency compared to other general purpose processors.

key features

Up to **20 TOPS @ 400 MHz**

INT8 Inference (99% of FP32 Accuracy)

50+ models tested with MERA framework



Features

Ultra-low Latency AI inference IP Core:

- Up to 24576 MACs and dedicated vector engine for non-convolution operations @ 400 MHz
- Data-flow array based architecture with optimization for INT8 parameters and activations
- Patented runtime reconfigurable interconnect

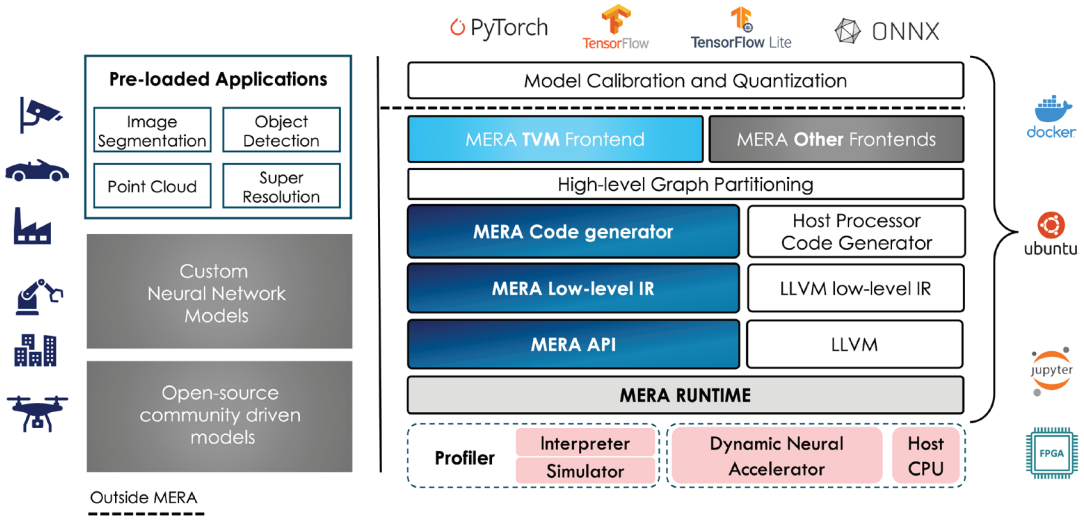
Robust Open-sourced MERA software framework:

- MERA compiler 1.0 exploits multiple forms of parallelism and maximizes compute utilization
- Native support for Pytorch and TensorFlow Lite models
- Built-in profiler in MERA framework
- Integrated with open-sourced Apache TVM

Dynamic Neural Accelerator
Framework

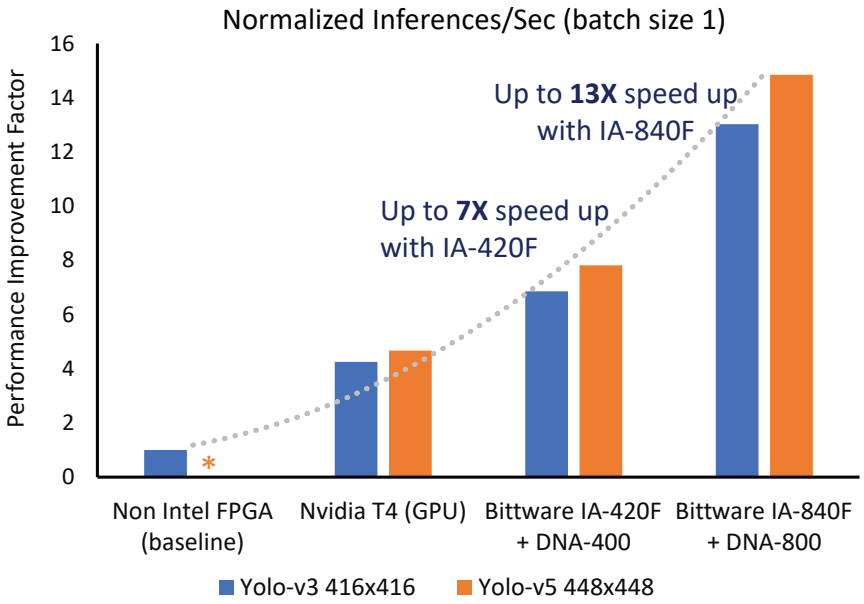
Product Description

EdgeCortix deep learning compute engines as part of the DNA IP Core is optimized for the Bittware IA-840F and IA-420F cards and is shipped as ready to use bitstreams. The EdgeCortix solution suite comes with MERA™ framework that can be installed from a public pip repository, enabling seamless compilation and execution of standard or custom convolutional neural networks (CNN) developed in industry-standard frameworks.



MERA consists of the compiler and software tool-kit needed to enable deep neural network graph compilation and inference using the integrated DNA bitstreams. Having built-in support for the open-source Apache TVM compiler framework, it provides the tools, APIs, code-generator and runtime needed to deploy a pre-trained deep neural network after a simple calibration and quantization step. MERA supports models to be quantized directly in the deep learning framework such as Pytorch or TensorflowLite.

Inference Performance Comparison



Comparing the inference performance of Edgecortix DNA running on BittWare cards with Intel Agilex FPGAs with the baseline of native AI acceleration IP on non-Intel FPGA and Nvidia T4 GPU, for state-of-the-art object detection models. Baseline of Yolov3 uses a solution that runs at 333 MHz on the non-Intel FPGA. DNA-400 and DNA-800 are new Intel Agilex optimized IP from EdgeCortix running at 350 MHz.

* Yolo-v5 did not run natively on the non Intel FPGA.

** DNA-400 /800 results are projected performance under real-time settings.

*** GPU numbers benchmarked with PyTorch 1.8 on AWS g4dn.2xlarge

All numbers are benchmarked at batch size 1

Dynamic Neural Accelerator
Framework

Detailed Feature List

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:

- Python and C++ interfaces
- PyTorch and TensorFlow-lite supported
- No need for retraining
- Supports high-resolution inputs

INT8 bit Quantization:

- Post-training quantization
- Support for deep learning framework built-in quantizers
- Maintains High accuracy

FPGA Card Options

IA-420F-0006	BittWare IA-420F card powered by EdgeCortix® Dynamic Neural Accelerator
IA-840F-0014	BittWare IA-840F card powered by EdgeCortix® Dynamic Neural Accelerator

Looking for a different card? Ask us about other compatible card options.

To learn more, visit **www.BittWare.com**

Rev 2022.5.6 | May 2022

BittWare
a molex company