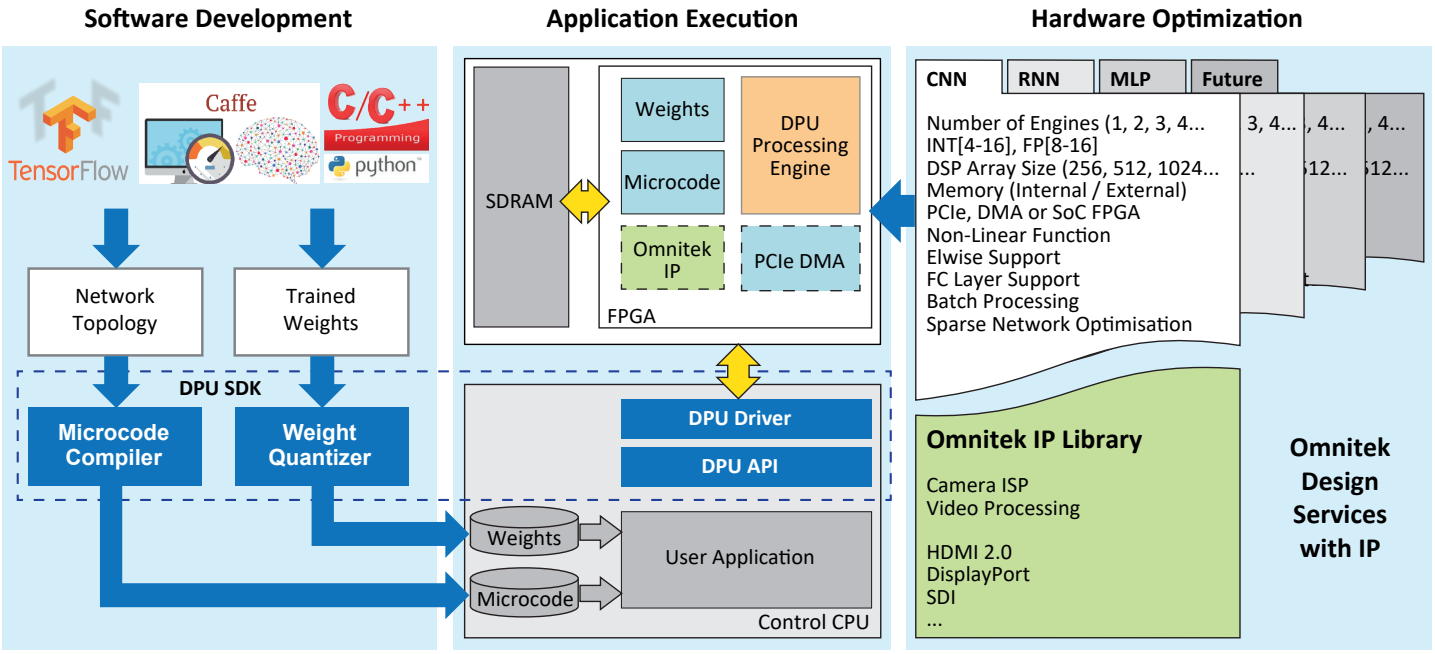
Omnitek DPU

Omnitek

High-Performance FPGA-based Engine for Deep Neural Networks



The Omnitek DPU (Deep Learning Processing Unit) is a configurable IP core built from a suite of FPGA IP comprising the key components needed to construct inference engines suitable for running Deep Neural Networks (DNNs) used for a wide range of Machine Learning applications, plus an SDK supporting the development of applications which integrate the DPU functionality. These can be targeted for a range of devices including small FPGAs with an embedded processor control for edge devices, or a PCI Express card with a large FPGA for data centre applications.

The Omnitek DPU can be programmed by creating a model of a chosen neural network in C/C++ or Python using standard frameworks such as TensorFlow. The SDK provides an API to enable the DPU inference to be integrated into a user application. The DPU SDK Compiler converts the model into microcode for execution by the Omnitek DPU. A quantizer optimally converts the weights and biases into the selected reduced precision processing format.

Implementation in today's high-performance FPGAs makes the Omnitek DPU not only fast but also highly adaptable. The architecture Omnitek has developed for its DPU ensures world-class performance across different neural network topologies (including CNNs, RNNs/LSTMs and MLPs) by adapting the FPGA design to optimise for a given workload using a range of novel architecture features, making optimal use of the FPGA's resources and running at the highest possible speed.

For evaluation purposes, Omnitek provides an example of GoogLeNet together with a C Model that works with the same microcode. This performs inference on 224x224 images using 8-bit integer processing at over 5,300 inferences per second when running in a Xilinx UltraScale+ VU9P-3 PCI Express card with a demonstration application run on a PC under Linux.

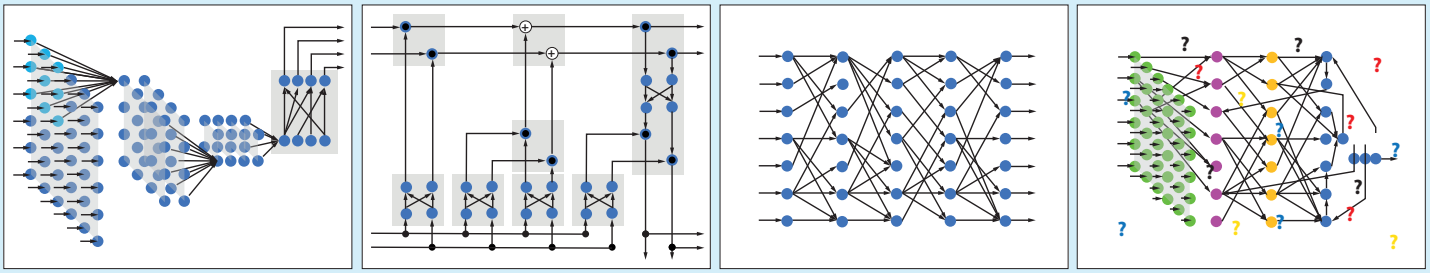
Key Features

- Faster than any alternative DNN running on an equivalent FPGA. Out-performs GPUs for a given power or cost budget.
- Fully software programmable in C/C++ or Python via standard frameworks such as TensorFlow
- Highly efficient FPGA use for optimum performance, cost and power
- Highly flexible:
  - Able to optimise architecture for the application workload
  - Able to adopt novel topologies and optimisation techniques as they emerge from industry and academia
- Suitable for either Data Centre (FPGA) or Embedded (FPGA SoC) applications

Applications

- Autonomous driving
- Object detection
- Smart security cameras
- Language translation
- Upscaling video to 8K
- Medical image analysis
- User interaction in virtual reality
- Big data statistical analysis

Software Programmable, Hardware Optimised for DNN Topology



Advantages of FPGA in AI systems

The highly flexible nature of the Omnitek DPU results from the choice of FPGA as the delivery platform. FPGAs offer massively parallel DSP blocks, distributed memory storage and reconfigurable logic which are ideal for neural network processing.

FPGAs offer many benefits over GPUs, ASICs or ASSPs for machine learning applications, including:

- High performance per watt
- Low latency and smaller batch size
- Hardware optimised for network topology
- Future-proofed technology:

  Easy to reprogram to accommodate novel network features and meet the demands of new applications

  Code written for one FPGA is readily transportable e.g. to the latest, more powerful device

  Easy integration with other IP such as video/vision functions to create a complete system on chip
- Speedy time to market

R&D Programme

To further the development of FPGA platforms for Deep Neural Networks, Omnitek engaged in active research in neural network algorithms and their optimum implementation on FPGA and other novel hardware architectures.

This work is being carried out in conjunction with Oxford University, via the Omnitek Oxford University Research Scholarship.

Research results are being continually fed into our DPU product development program.

	Supported now	In development
DNN Types	CNN	RNN/LSTM, MLP
Development Frameworks	TensorFlow	Caffe
Internal Data Format	INT8	FP8 - FP11
Target FPGA Families	Xilinx Kintex UltraScale+ Xilinx Virtex UltraScale+ Xilinx Zynq UltraScale+ MPSoC	To be announced
Network Model Primitives	Convolutional matrix multiply, Simple (fully connected) matrix multiply, ReLU, concatenation, max pool, average pool, batch normalization, softmax bias add, elwise add	Sparse matrix multiply (fully connected), elwise multiply, leaky ReLU, arbitrary non-linear activation

Omnitek

UK Head Office
Intec 3, Level 1
Wade Road
Basingstoke
Hampshire
RG24 8NE

Tel: +44 (0)1256 345900
Fax: +44 (0)1256 345901

Email: consultancy@omnitek.tv

About Omnitek

Omnitek is a world leader in the design of intelligent video and vision systems based on programmable FPGAs and SoCs. Through the supply of expert design services with highly optimised FPGA intellectual property cores covering high-performance video/vision and AI / machine learning, Omnitek can provide cost-optimised solutions to a broad range of markets.

Omnitek reserves the right to change specifications without notice. Refer to the Omnitek web site for the latest specifications and further information:

www.omnitek.tv

ISO 9001:2015

Published 12/18

Version 2