SOLUTION BRIEF

Programmable Solutions
FPGA Video and Vision

Deep Learning Processing Logic (DLPL) FPGA IP

High-Performance FPGA-based Engine for Deep Neural Networks

Summary

The Intel Deep Learning Processing Logic (DLPL), 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 DLPL 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.

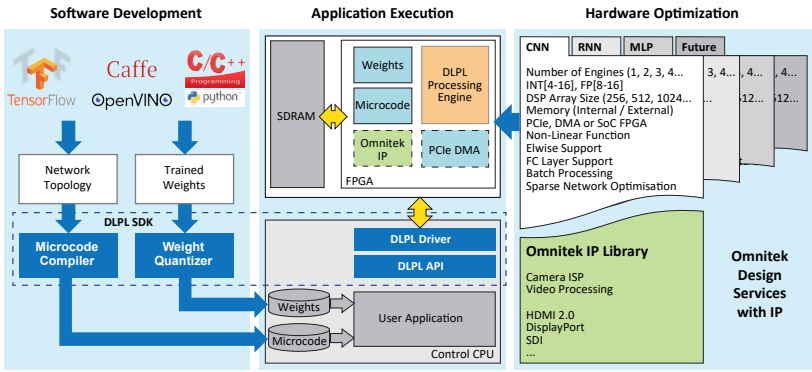

Figure 1. Typical Implementation of the DLPL Pipeline IP

The Intel DLPL 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 DLPL instance to be integrated into a user application. The DLPL SDK Compiler converts the model into microcode for execution by the Intel DLPL. A quantizer optimally converts the weights and biases into the selected reduced precision processing format.

Implementation in today's high-performance FPGAs makes the Intel DLPL not only fast but also highly adaptable. The architecture Intel has developed for its DLPL 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.

The DLPL design employs a novel mathematical framework which is highly compatible with the architecture of the Arria 10 and Stratix 10 architectures, leading to excellent resource efficiency and great compute density. Using a combination of low-precision fixed point maths and floating-point maths the DLPL is able to achieve this very high compute density with no loss of accuracy compared with that of the original FP32 model.

For evaluation purposes, Intel can provide an example of GoogleNet, ResNet-50 or VGG16 together with a C Model that works with the same microcode. This performs inference on 224×224 images running in Intel's Arria 10 GX 1150 device on the Arria 10 GX Development Kit 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

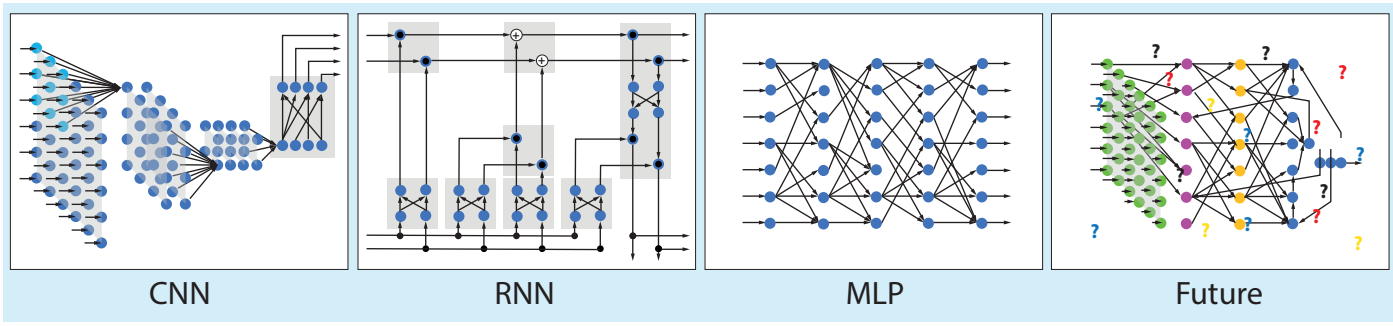

Figure 2. DPU DNN diagram

Advantages of FPGA in AI systems

The highly flexible nature of the Intel DLPL 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
- Speedy time to market
- 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

R&D Programme

To further the development of FPGA platforms for Deep Neural Networks, Intel 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 Intel 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 OpenVINO
Target FPGA Families	Arria 10 GX	Stratix 10 GX
Primitives	Convolutional Matrix Multiply, Fully Connected Matrix Multiply (simple and sparse), ReLU, Concatenation, Max Pool, Average Pool, Batch Normalization, Distribution Shift (DSConv), Softmax, Bias add, Element-wise add, Depth-wise Separable Convolutions	Arbitrary Non-linear Activation, Deconvolution/Transpose convolution

Document Number: 618539-0.1