An FPGA CNN for Intelligent Video/Vision Systems

Roger Fawcett, CEO Omnitek

XDF Frankfurt, 10th December 2018

Omnitek

XILINX

Alliance Program (Premier)
Accelerator Program

Omnitek

- A world leader in the design of intelligent video and vision systems based on programmable FPGAs and SoCs
  - IP & design services for AI and Video / Vision applications
  - Highly skilled staff with significant DSP expertise
  - Chip sets for ASSP replacement
  - 50 staff
  - 35% annual invoice growth over 4 years, profitable
  - 60%+ orders growth
  - 2 x Queens Awards for Enterprise in 2018 (Innovation & International Trade)

ISO 9001:2015

iabm Silver Member

SMPTE

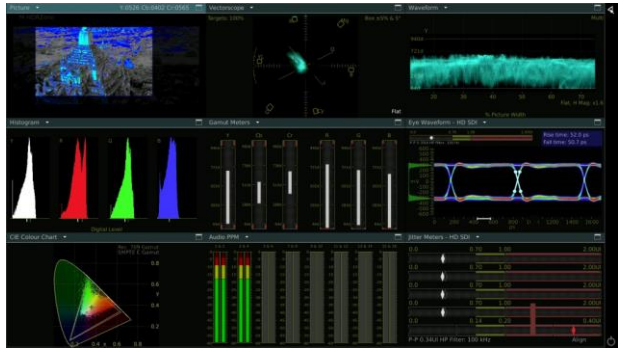
Omnitek

Broad Range of Video Vision Markets Adopting AI

Projectors
Displays

Surgical Robotics
Medical Imaging

Automotive
ADAS



Broadcast

Aerospace
Defence

Smart Boards
Interactive Displays

AR / VR
Consumer Electronics

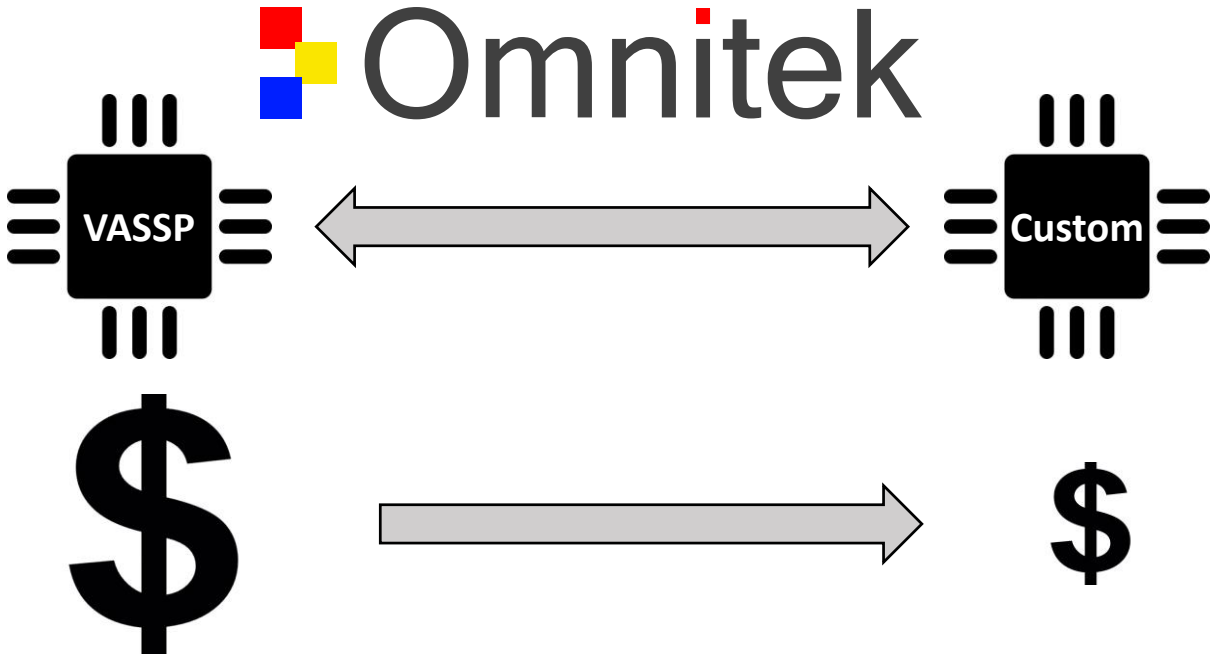
Professional Video &
Audio Equipment

Omnitek

The Omnitek Proposition

We create cost-optimised semiconductor devices
for clients in competitive intelligent video/vision markets
which enable them to differentiate their products
and bring them to market rapidly.

- Differentiated products
- Optimised for cost
- Rapid time to market
- No expertise required



Omnitek

Omnitek IP / Technology: 180 IP Cores

AI

- **DPU CNN Processor**

Video Processing

- **OSVP Suite**:

Scaler, Deinterlacer, chroma resampler, color processing, noise reduction, deblock

- **Warp Processor**
- **ISP**
- **2D Graphics**
- **HDR Tone Mapping**
- **MPEG2**
- **Image Stitch**
- **SDI Analysis T&M**
- **3D Colour Processing**

Connectivity

- **HDMI 2.0**
- **V-by-One**
- **SDI**
- **SDI Gearbox**
- **Audio embed / extract**
- **Def-Stan-0082 VoIP**
- **PCIe DMA**
- **Javelin H.265 AV over IP**

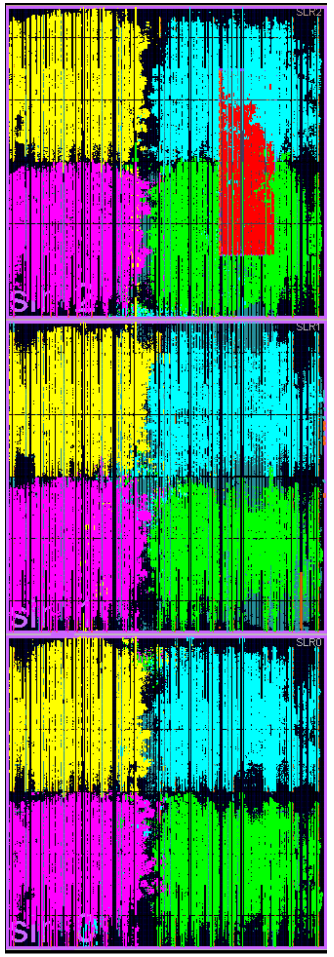
Computer Vision / AI

- **3D Depth Map**
- **Object Tracking**
- **VR & AR IP**

Omnitek

Omnitek DPU

Deep-Learning Processing Unit

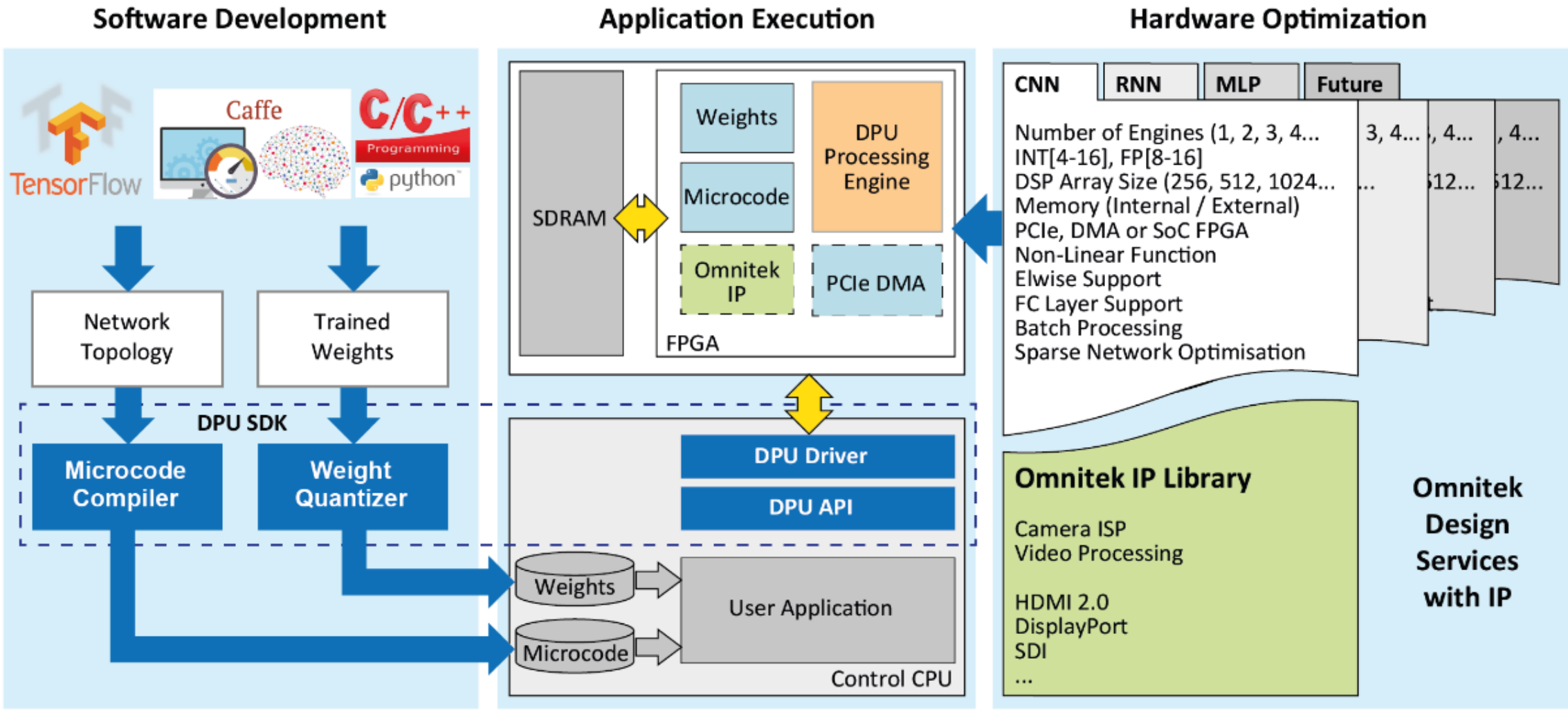


- IP Core and Software Framework (Overlay)
- Software Programmable / Hardware Optimised
- 800 MHz DSP performance
- 90% DSPs used: 2 x INT8 MACs per cycle
- 23.9% LUTs
- 86% Efficiency
- System level integration through video/vision IP & design services

Omnitek

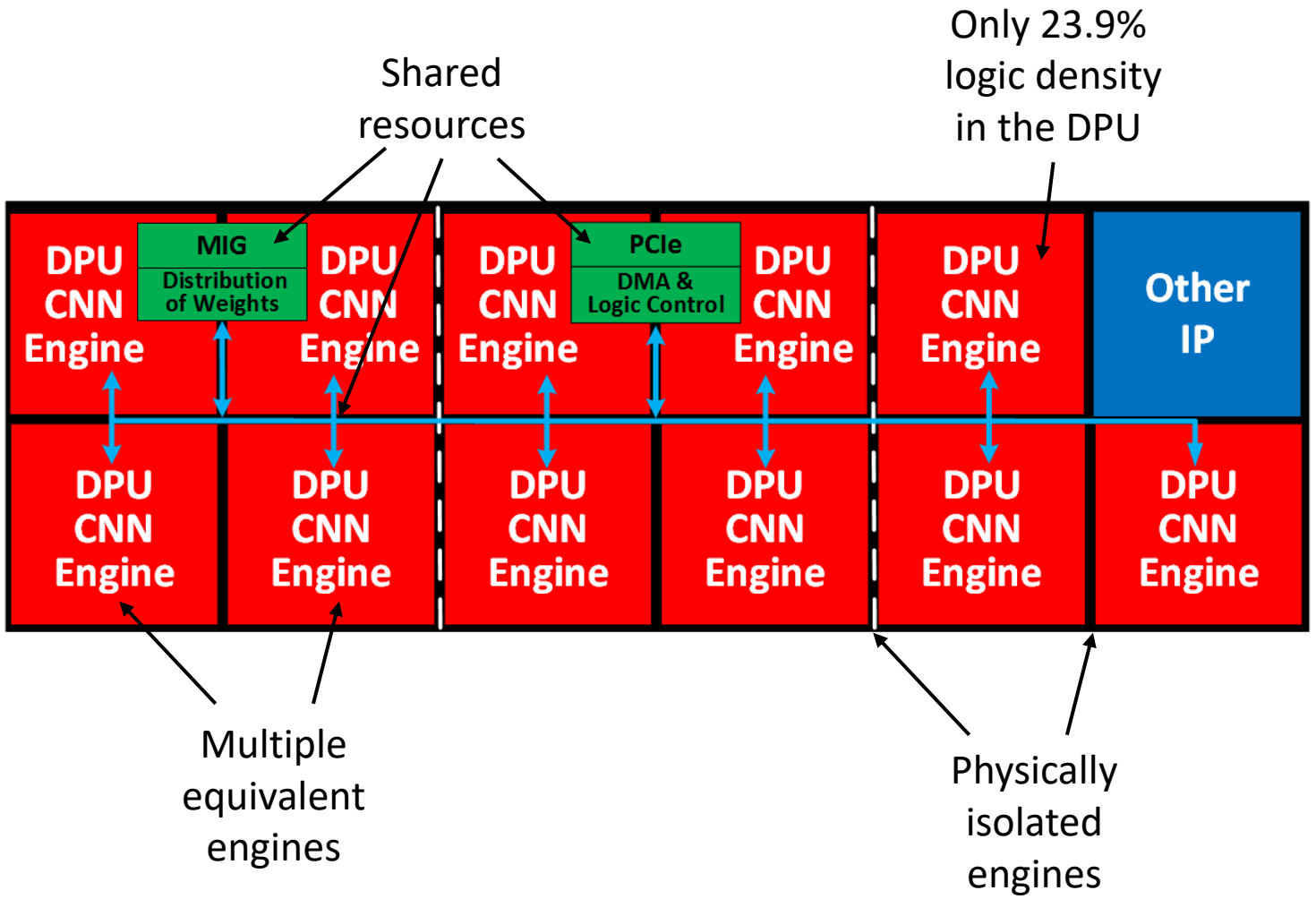
Omnitek DPU

Deep-Learning Processing Unit

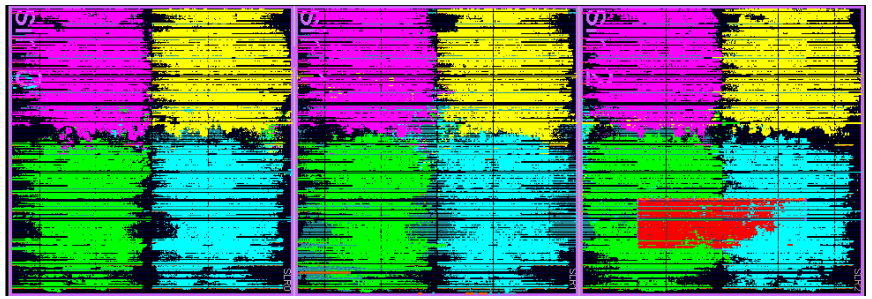


Omnitek

Designed For Scalability & System-On-Chip Integration

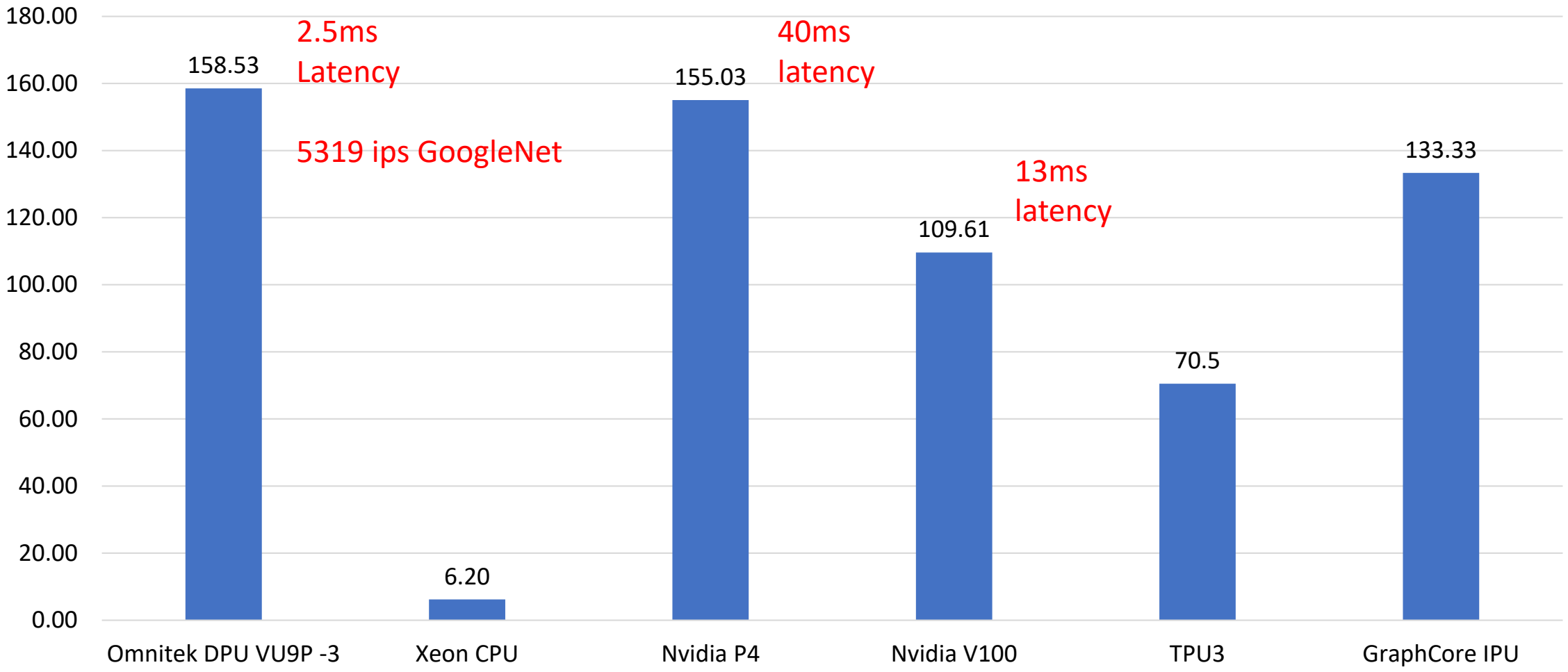


- Highly scalable
- Easy to add other IP



Omnitek

GOPs/W Performance



Google TPU3
Power consumption has been estimated at 200W.
The overall peak performance is 92 TOP/s
Although no figures have been published for the TPU3, the TPU1 paper indicates that CNN1 (a typical CNN) operates at 14.1 TOPs compared to a similar peak performance to the TPU3.
GraphCore IPU
Power has been given at 150W
Overall peak performance has been given as around 100 TOP/s per device.
The only efficiency figure supplied is around 20% when training ResNet. We don't know how this varies across different CNNs or during inference, but will take this as a typical figure. Certainly, it appears that each processing element is required to pause computation while data is being transferred.

Omnitek

Latest Developments

Constantly changing Landscape:

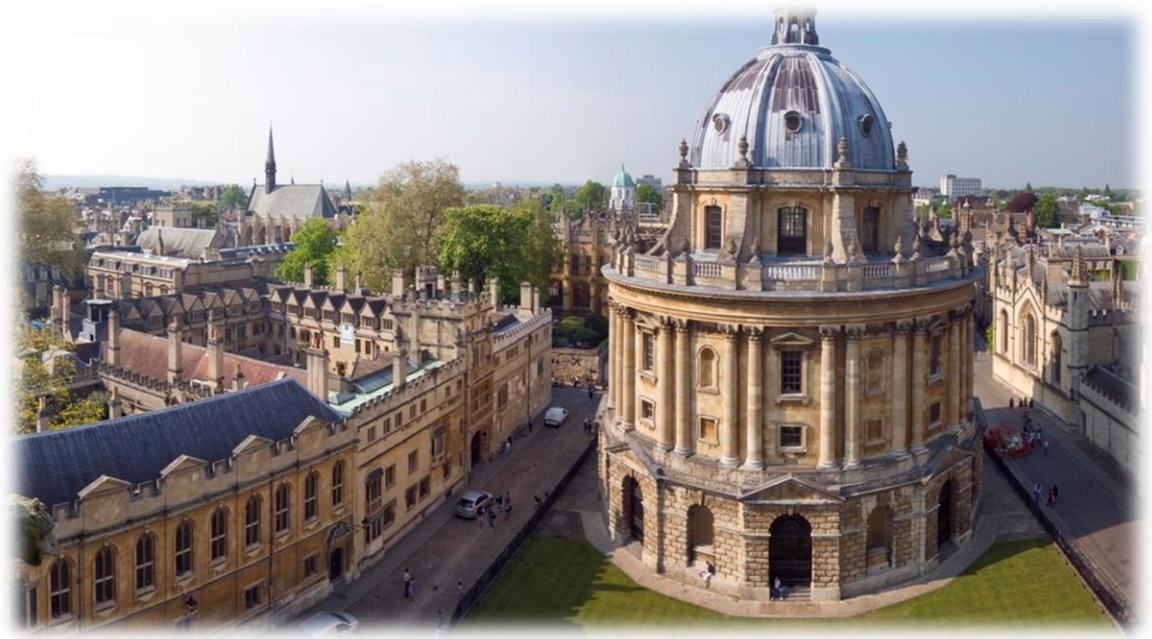
▪ New Silicon architectures	e.g. Versal
▪ New CNN topologies	e.g. MNasNet, AI Upscale
▪ New Data types / quantisation	e.g. DBConv (Omnitek)
▪ Integration into complete SoC systems	e.g. Smart Camera

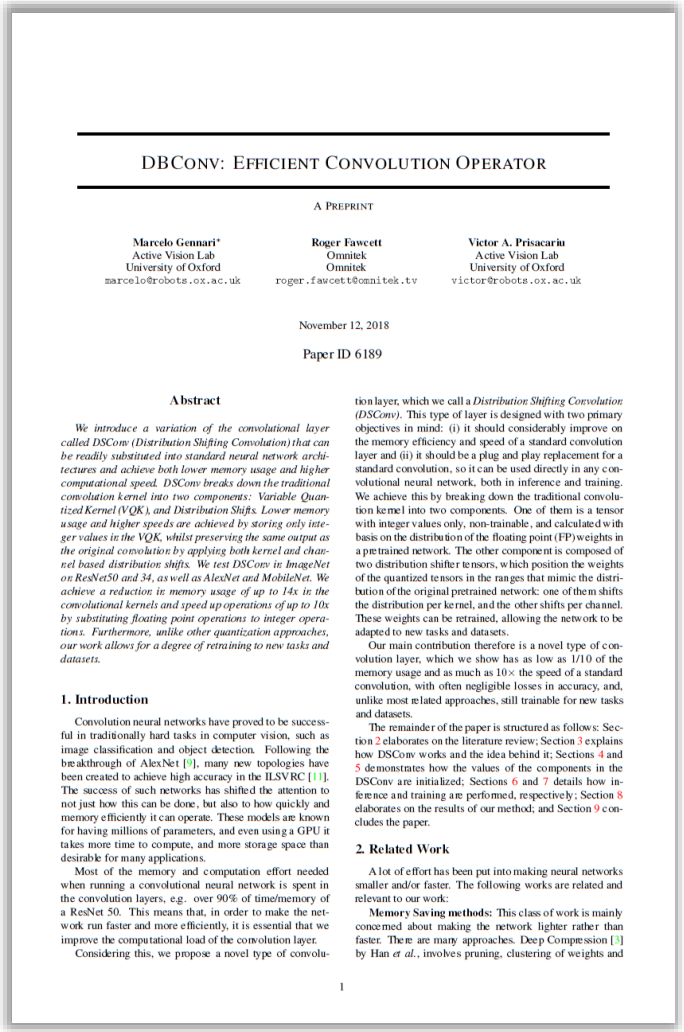
Requires continuous innovation and RTL design by Omnitek

Omnitek

Oxford University Research Partnership

- DPhil (PhD) research Scholarship
- Novel mathematical representations
- Optimum AI Architectures for FPGAs





DBConv: Efficient Convolution Operator

A Preprint

Marcelo Gennari*
Active Vision Lab
University of Oxford
marcelo@robots.ox.ac.uk

Roger Fawcett
Omnitek
Omnitek
roger.fawcett@omnitek.tv

Victor A. Prisacariu
Active Vision Lab
University of Oxford
victor@robots.ox.ac.uk

November 12, 2018

Paper ID 6189

Abstract

We introduce a variation of the convolutional layer called DSConv (Distribution Shifting Convolution) that can be readily substituted into standard neural network architectures and achieve both lower memory usage and higher computational speed. DSConv breaks down the traditional convolution kernel into two components: Variable Quantized Kernel (VQK), and Distribution Shifts. Lower memory usage and higher speeds are achieved by storing only integer values in the VQK, whilst preserving the same output as the original convolution by applying both kernel and channel based distribution shifts. We test DSConv in ImageNet on ResNet50 and 34, as well as AlexNet and MobileNet. We achieve a reduction in memory usage of up to 14x in the convolutional kernels and speed up operations of up to 10x by substituting floating point operations to integer operations. Furthermore, unlike other quantization approaches, our work allows for a degree of retraining to new tasks and datasets.

1. Introduction

Convolution neural networks have proved to be successful in traditionally hard tasks in computer vision, such as image classification and object detection. Following the breakthrough of AlexNet [9], many new topologies have been created to achieve high accuracy in the ILSVRC [11]. The success of such networks has shifted the attention to not just how this can be done, but also to how quickly and memory efficiently it can operate. These models are known for having millions of parameters, and even using a GPU it takes more time to compute, and more storage space than desirable for many applications.

Most of the memory and computation effort needed when running a convolutional neural network is spent in the convolution layers, e.g. over 90% of time/memory of a ResNet 50. This means that, in order to make the network run faster and more efficiently, it is essential that we improve the computational load of the convolution layer.

Considering this, we propose a novel type of convolution layer, which we call a *Distribution Shifting Convolution (DSConv)*. This type of layer is designed with two primary objectives in mind: (i) it should considerably improve on the memory efficiency and speed of a standard convolution layer and (ii) it should be a plug and play replacement for a standard convolution, so it can be used directly in any convolutional neural network, both in inference and training. We achieve this by breaking down the traditional convolution kernel into two components. One of them is a tensor with integer values only, non-trainable, and calculated with basis on the distribution of the floating point (FP) weights in a pretrained network. The other component is composed of two distribution shifter tensors, which position the weights of the quantized tensors in the ranges that mimic the distribution of the original pretrained network: one of them shifts the distribution per kernel, and the other shifts per channel. These weights can be retrained, allowing the network to be adapted to new tasks and datasets.

Our main contribution therefore is a novel type of convolution layer, which we show has as low as 1/10 of the memory usage and as much as 10× the speed of a standard convolution, with often negligible losses in accuracy, and, unlike most related approaches, still trainable for new tasks and datasets.

The remainder of the paper is structured as follows: Section 2 elaborates on the literature review; Section 3 explains how DSConv works and the idea behind it; Sections 4 and 5 demonstrates how the values of the components in the DSConv are initialized; Sections 6 and 7 details how inference and training are performed, respectively; Section 8 elaborates on the results of our method; and Section 9 concludes the paper.

2. Related Work

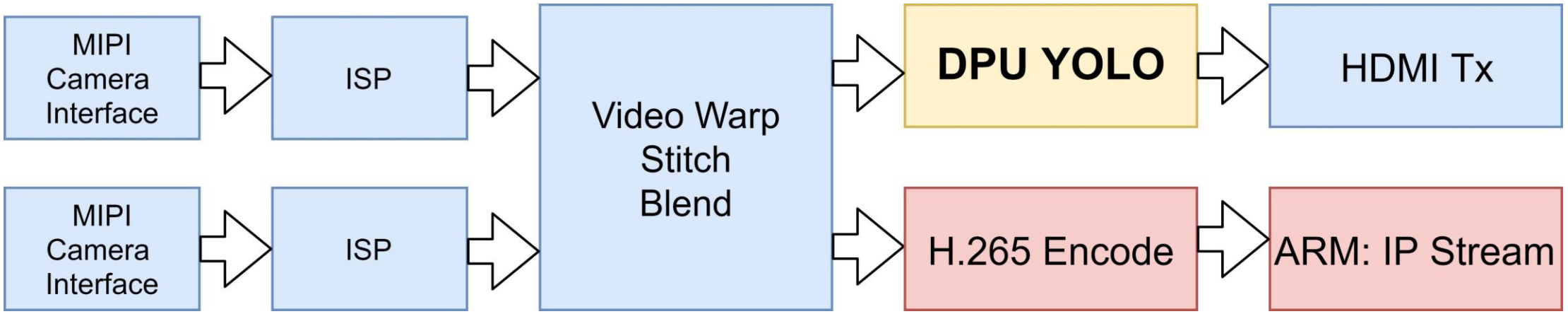
A lot of effort has been put into making neural networks smaller and/or faster. The following works are related and relevant to our work:

Memory Saving methods: This class of work is mainly concerned about making the network lighter rather than faster. There are many approaches. Deep Compression [3] by Han *et al.*, involves pruning, clustering of weights and

1

Omnitek

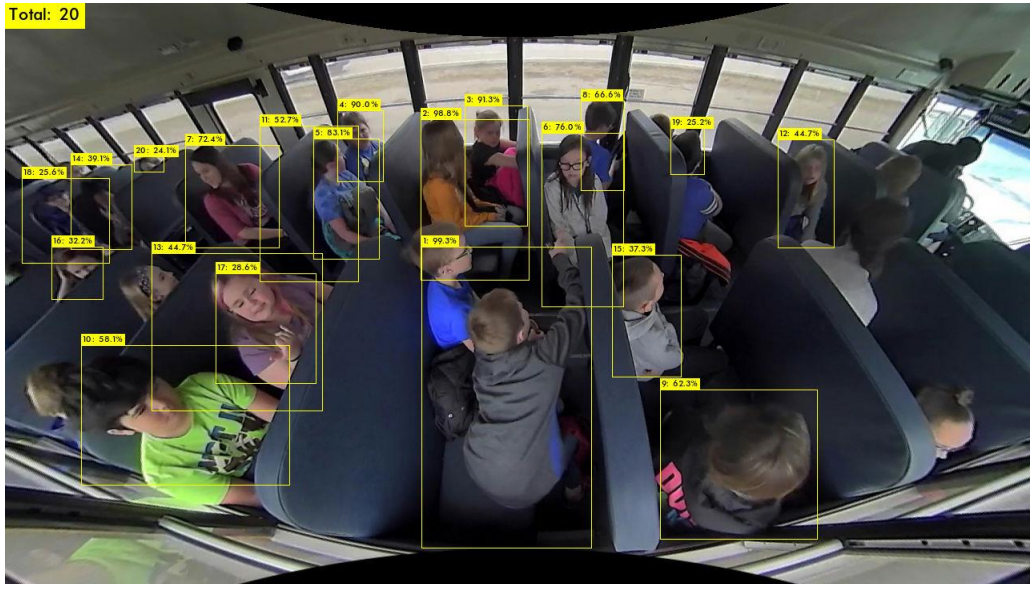
AI Enabled Design: Smart Camera: ZU5

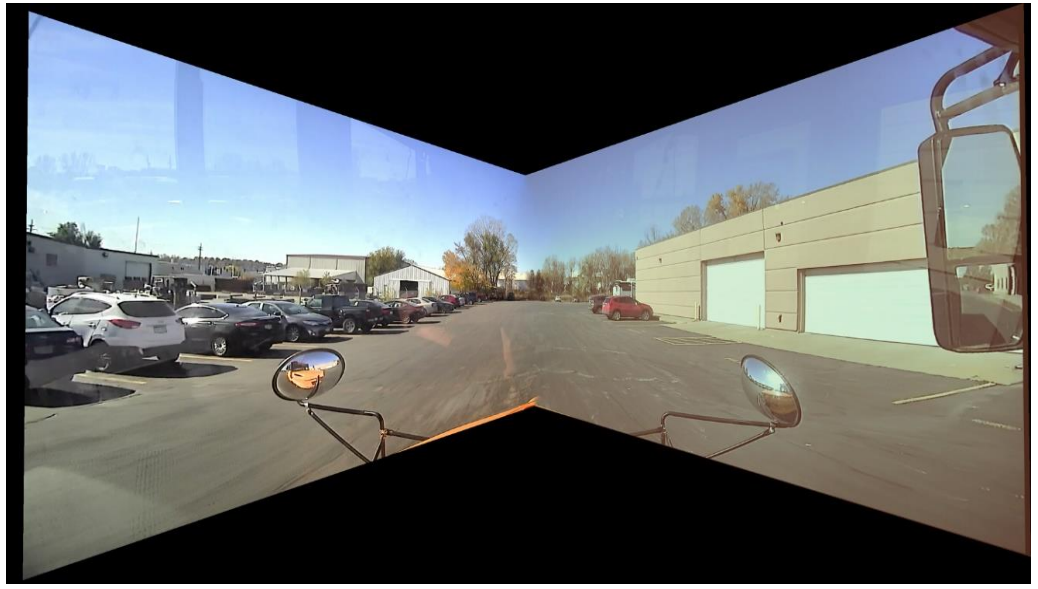


- MIPI Interface
- Camera ISP
- Image Warp and Stitch
- Scaler
- H.265 Streaming over IP

- Count People
- Posture detection
- Gesture recognition
- Object detection

Omnitek





Omnitek

AI Upscaler: New Architecture For Optimum Performance: Adapted RAISR Algorithm

Original **Bicubic** **AI Upscale**

- CNN, however with some architectural differences
- New Optimised RTL Design

Omnitek

Wrap-up

Omnitek DPU

- Leading CNN Inferencing on an FPGA
- Performance, Cost and Power advantages for all designs
- Software programmable
- Delivered on a platform that supports

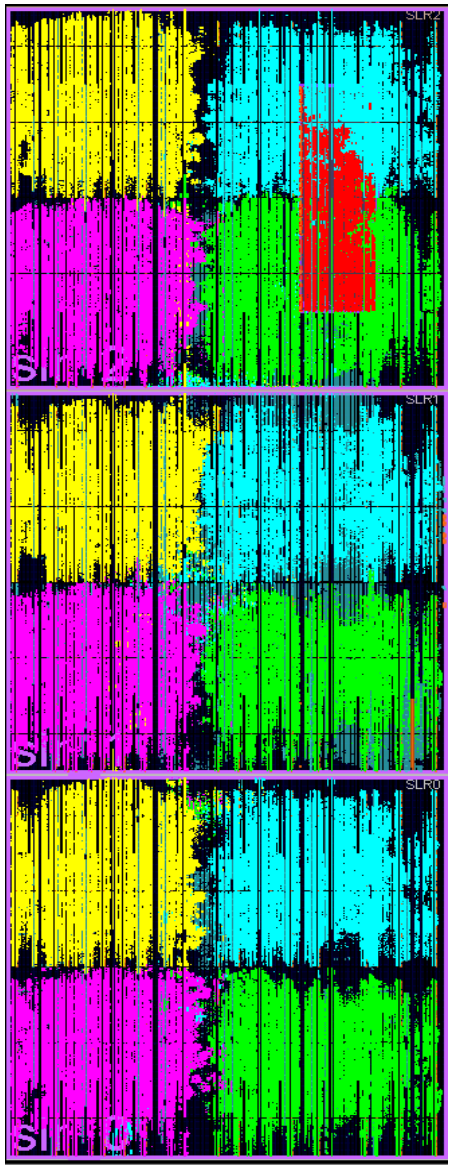
- Continuous Research:
  - Novel AI architectures
  - System Integration

Omnitek

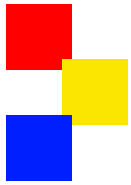
- Differentiated products
- Optimised for cost
- Rapid time to market
- No expertise required

Learn more…

- Visit us here at booth #16
- www.Omnitek.tv



Omnitek



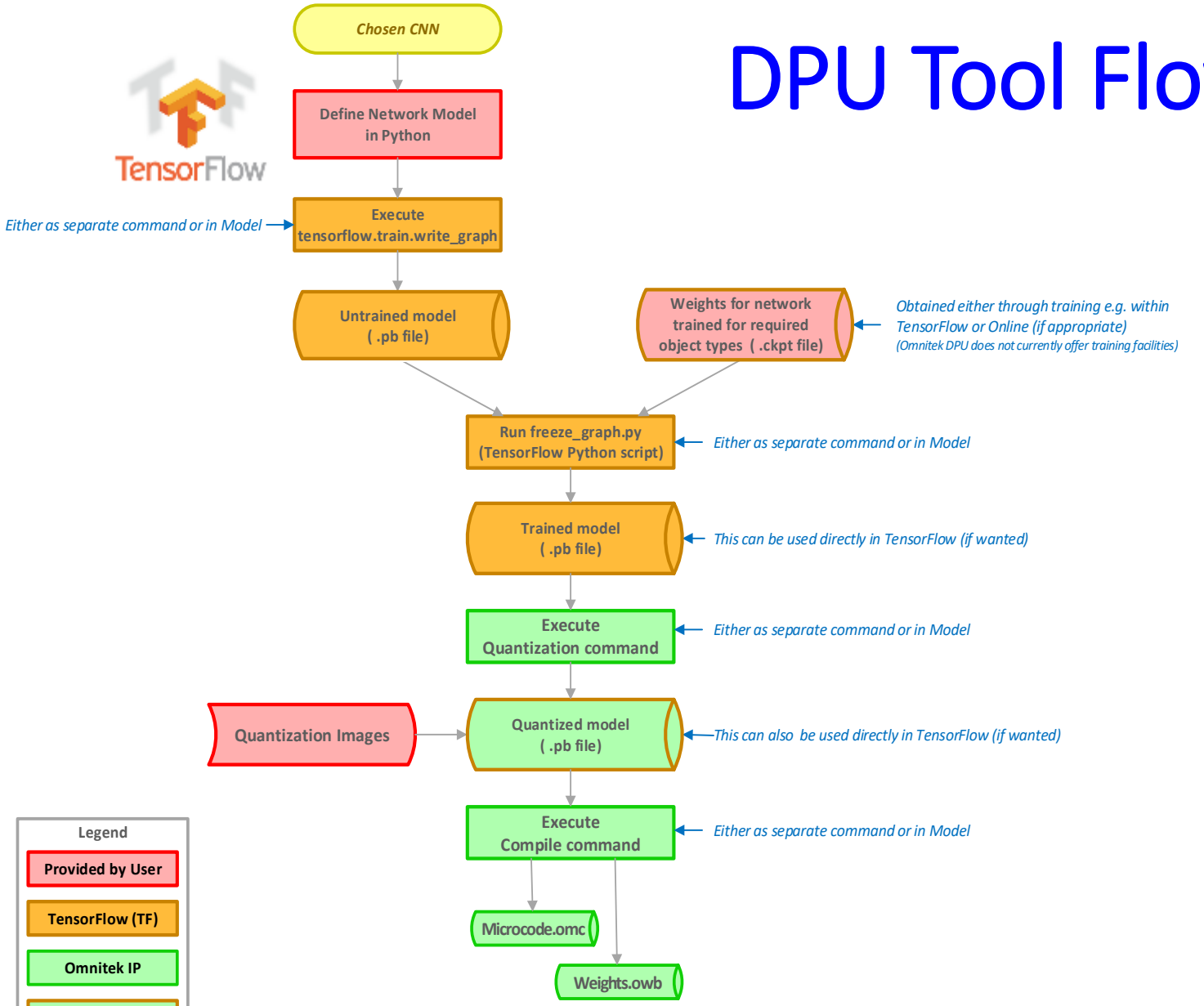
Omnitek

Intelligent Video/Vision Systems

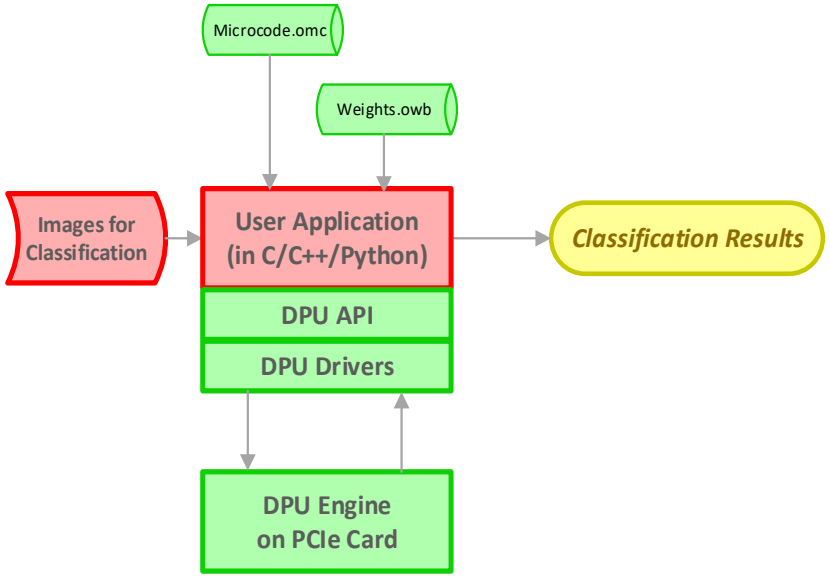
Omnitek

Can we make a better single slide for the DPU SDK?

DPU Tool Flow



Design Creation Via Python in Tensor Flow



Runtime Integration via DPU API

Omnitek