An FPGA CNN for Intelligent Video/Vision Systems

Roger Fawcett, CEO Omnitek
XDF Frankfurt, 10th December 2018
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)
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
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 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 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
Designed For Scalability & System-On-Chip Integration

- Highly scalable
- Easy to add other IP

Multiple equivalent engines

Physically isolated engines

Only 23.9% logic density in the DPU

Shared resources

DPU CNN Engine

DPU CNN Engine

DPU CNN Engine

DPU CNN Engine

DPU CNN Engine

Other IP

MIG
Distribution of Weights

PCle
DMA & Logic Control

DPU CNN Engine

DPU CNN Engine

DPU CNN Engine

DPU CNN Engine

DPU CNN Engine
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 TOP/s 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.
Latest Developments

Constantly changing Landscape:

- New Silicon architectures
  e.g. Versal
- New CNN topologies
  e.g. M NasNet, 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
Oxford University Research Partnership

- DPhil (PhD) research Scholarship
- Novel mathematical representations
- Optimum AI Architectures for FPGAs
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
AI Upscaler: New Architecture For Optimum Performance: Adapted RAISR Algorithm

- CNN, however with some architectural differences
- New Optimised RTL Design
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
Intelligent Video/Vision Systems
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

Microcode.omc

Images for Classification

User Application
(in C/C++/Python)

DPU API

DPU Drivers

DPU Engine
on PCIe Card

Classification Results

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