technology
from seed

Hybrid Dot-Product Calculation for Convolutional Neural Networks in FPGA

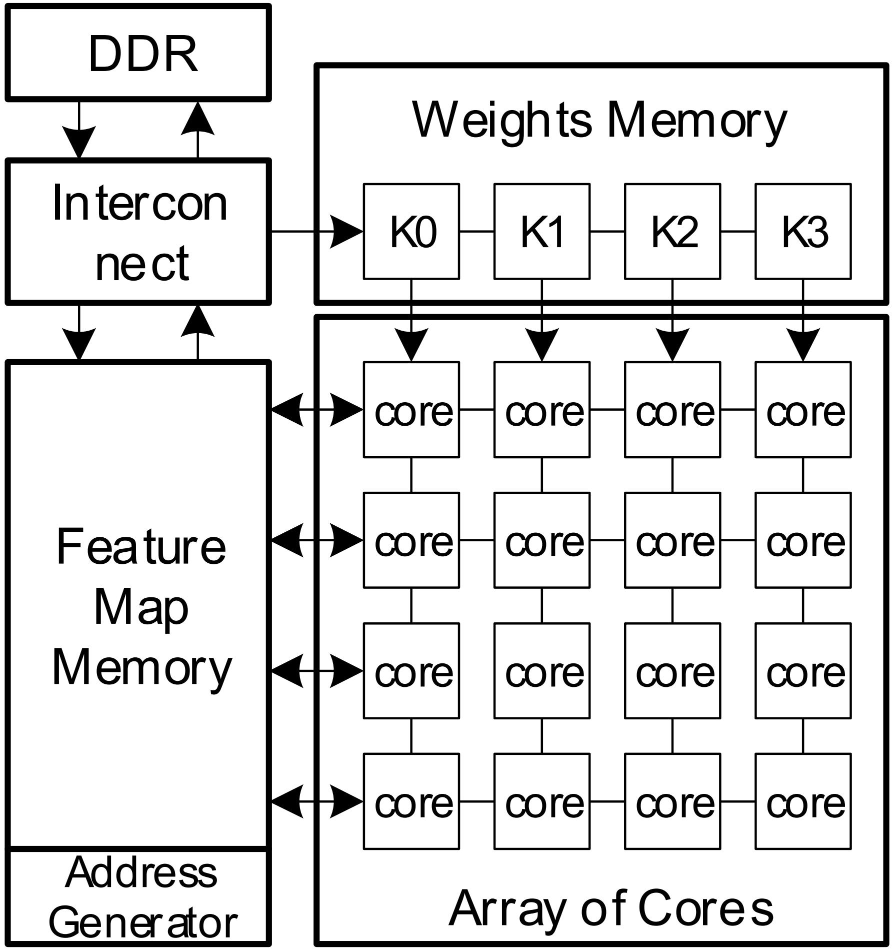
Mário Véstias	INESC-ID/ISEL/IPL
Rui P. Duarte	INESC-ID/IST/UL
José T. de Sousa	INESC-ID/IST/UL
Horácio Neto	INESC-ID/IST/UL

Convolutional Neural Networks for Embedded Computing

- CNNs are very good on many AI applications, like image classification;
- If applied near the image sensor, avoids information communication to server for processing;
- CNNs have high computational complexity and high memory bandwidth requirements;
- Efficient hybrid CNNs are obtained using different fixed-point scales for different layers;

Hybrid Dot-Product Calculation to Support Hybrid Cores

Baseline Architecture



Hybrid Dot-Product Calculation

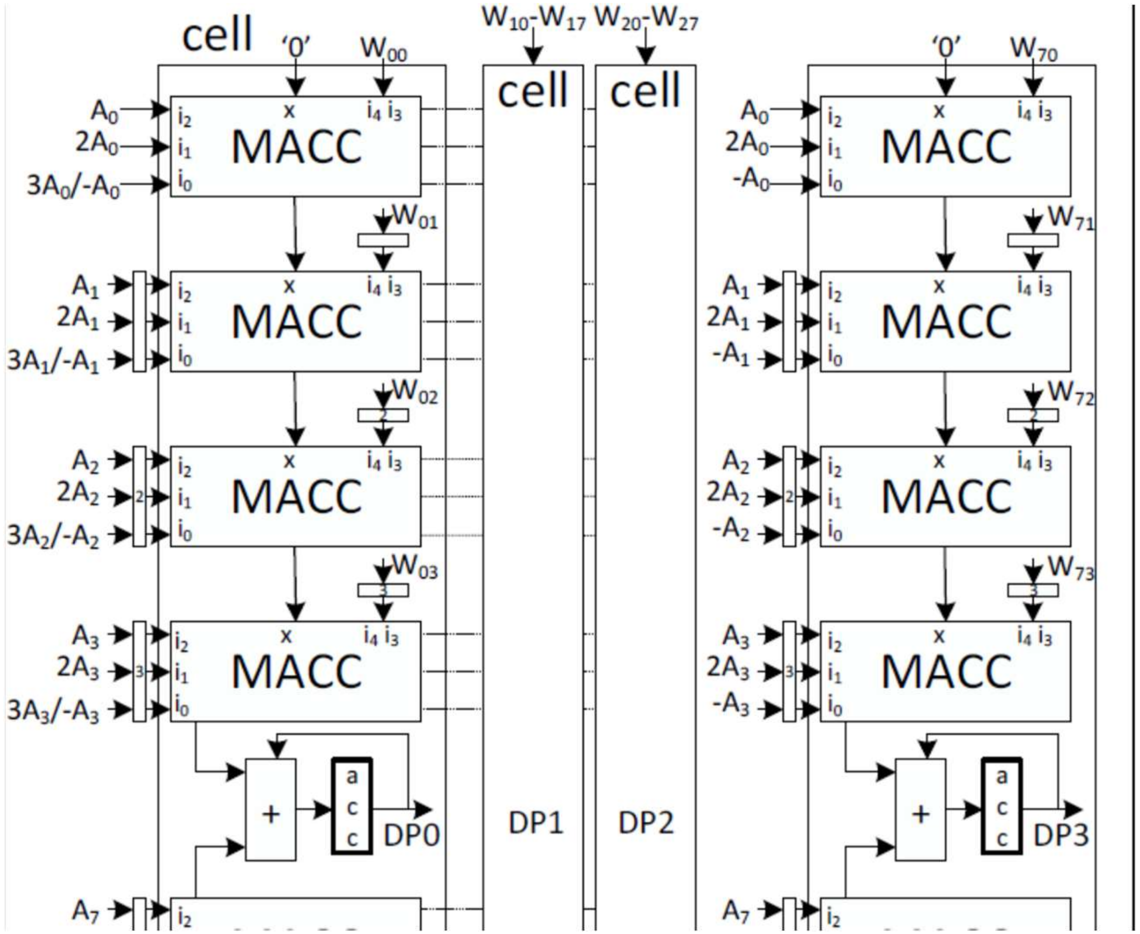
- Case study: activations x weights: 8 x 8 and 8 x 2
- Eight 8-bit activations (64 bits) are run in parallel in each core;
- 32 2-bit weights (64 bits) are read in parallel:
  - W00-W07, W10-W17, W20-W27, W30-W37
- Four dot-products are generated in parallel

$$DPj = \sum_{i=0}^{i=7} A_i \times W_{ji}$$

- For 8-bit weights, a single dot-product is generated

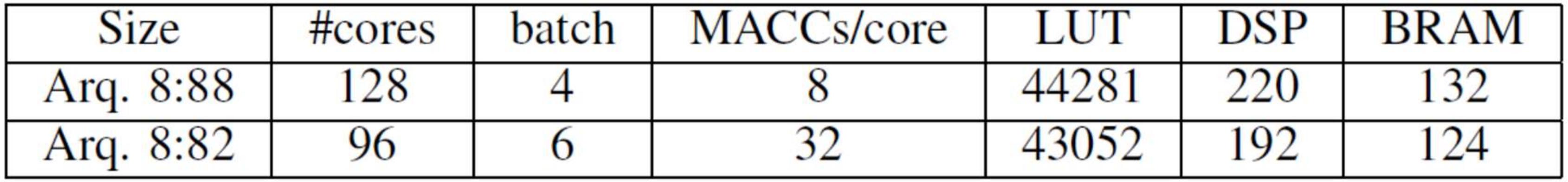
$$A \cdot W = DP3 \times 2^6 + DP2 \times 2^4 + DP1 \times 2^2 + DP0$$

Hybrid Core (8x8, 8x2)



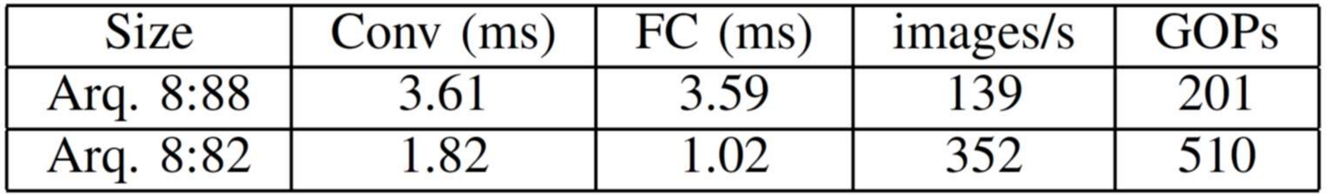
Results

Size of the complete architecture:



Size	#cores	batch	MACCs/core	LUT	DSP	BRAM
Arq. 8:88	128	4	8	44281	220	132
Arq. 8:82	96	6	32	43052	192	124

AlexNet mapped in the proposed architecture



Size	Conv (ms)	FC (ms)	images/s	GOPs
Arq. 8:88	3.61	3.59	139	201
Arq. 8:82	1.82	1.02	352	510

Conclusions and Future Work

- The hybrid architecture proposed in this work supports the execution of layers with different weight sizes;
- With 25% more resources per core, the performance of the hybrid architecture running AlexNet increases by about 2.5X;

- The core is being extended to support other fixed-point sizes.