Towards An Efficient Accelerator for DNN-based Segmentation on FPGA

Imperial College London

Shuanglong Liu, Wayne Luk

s.liu13@ic.ac.uk

FPL, 10 September 2019

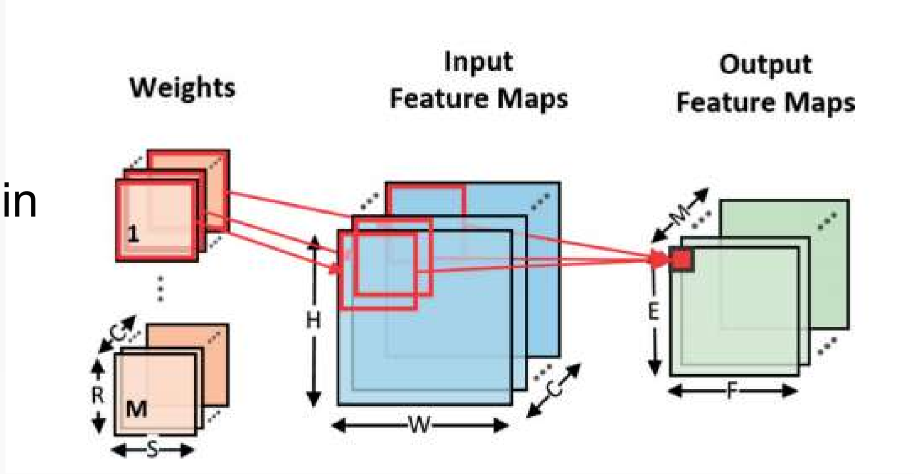
Imperial College London

Motivation

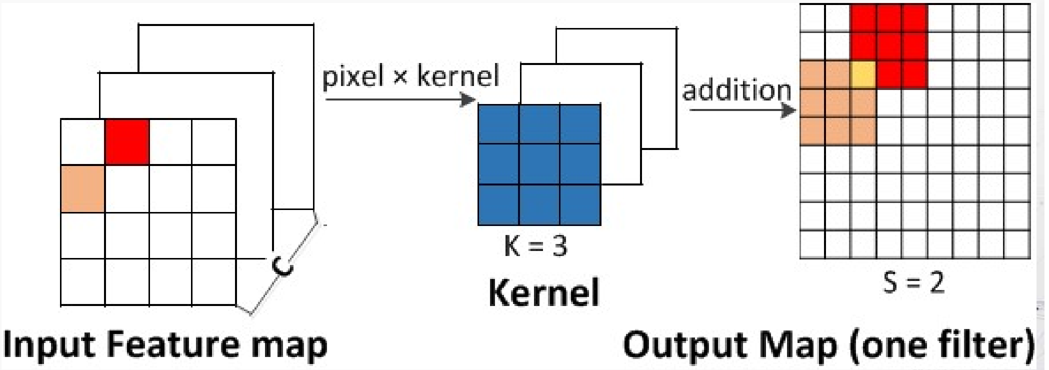
❖ **Challenges:**

? How to efficiently map deconv in conv accelerator;

? How to achieve real-time response for larger networks;



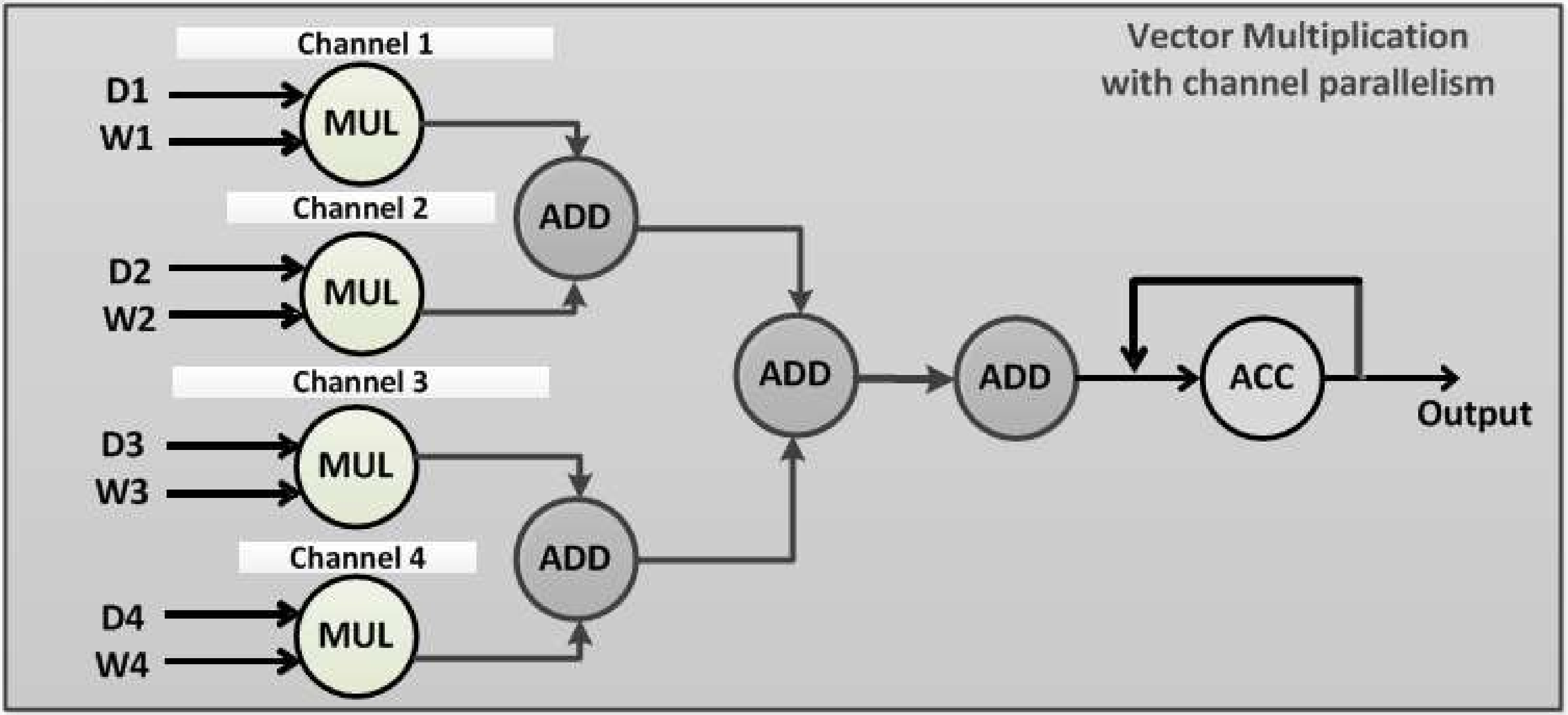
CONV



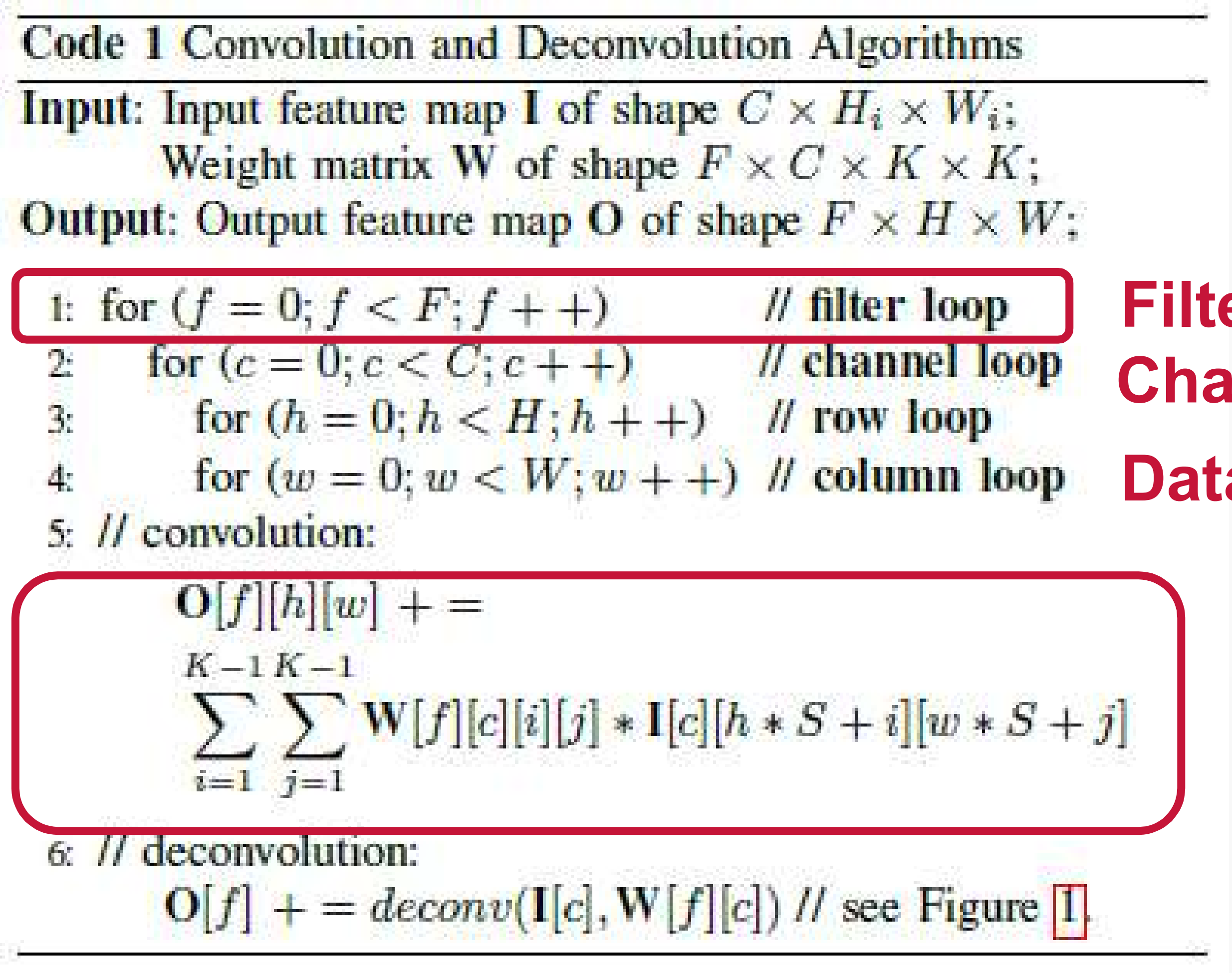
DECONV

Imperial College London

Design 1. Uniform Architecture



Design 2. Parallelism Exploration



```
Code 1 Convolution and Deconvolution Algorithms
Input: Input feature map I of shape $C \times H_i \times W_i$;
       Weight matrix W of shape $F \times C \times K \times K$;
Output: Output feature map O of shape $F \times H \times W$;
for ($f = 0; f < F; f++$)          // filter loop
    for ($c = 0; c < C; c++$)      // channel loop
        for ($h = 0; h < H; h++$)  // row loop
        for ($w = 0; w < W; w++$)  // column loop
// convolution:
        $\mathbf{O}[f][h][w] += \sum_{i=1}^{K-1} \sum_{j=1}^{K-1} \mathbf{W}[f][c][i][j] * \mathbf{I}[c][h*S+i][w*S+j]$
// deconvolution:
        $\mathbf{O}[f] += deconv(\mathbf{I}[c], \mathbf{W}[f][c])$ // see Figure [1]
```

Filter parallelism

Channel parallelism

Data parallelism

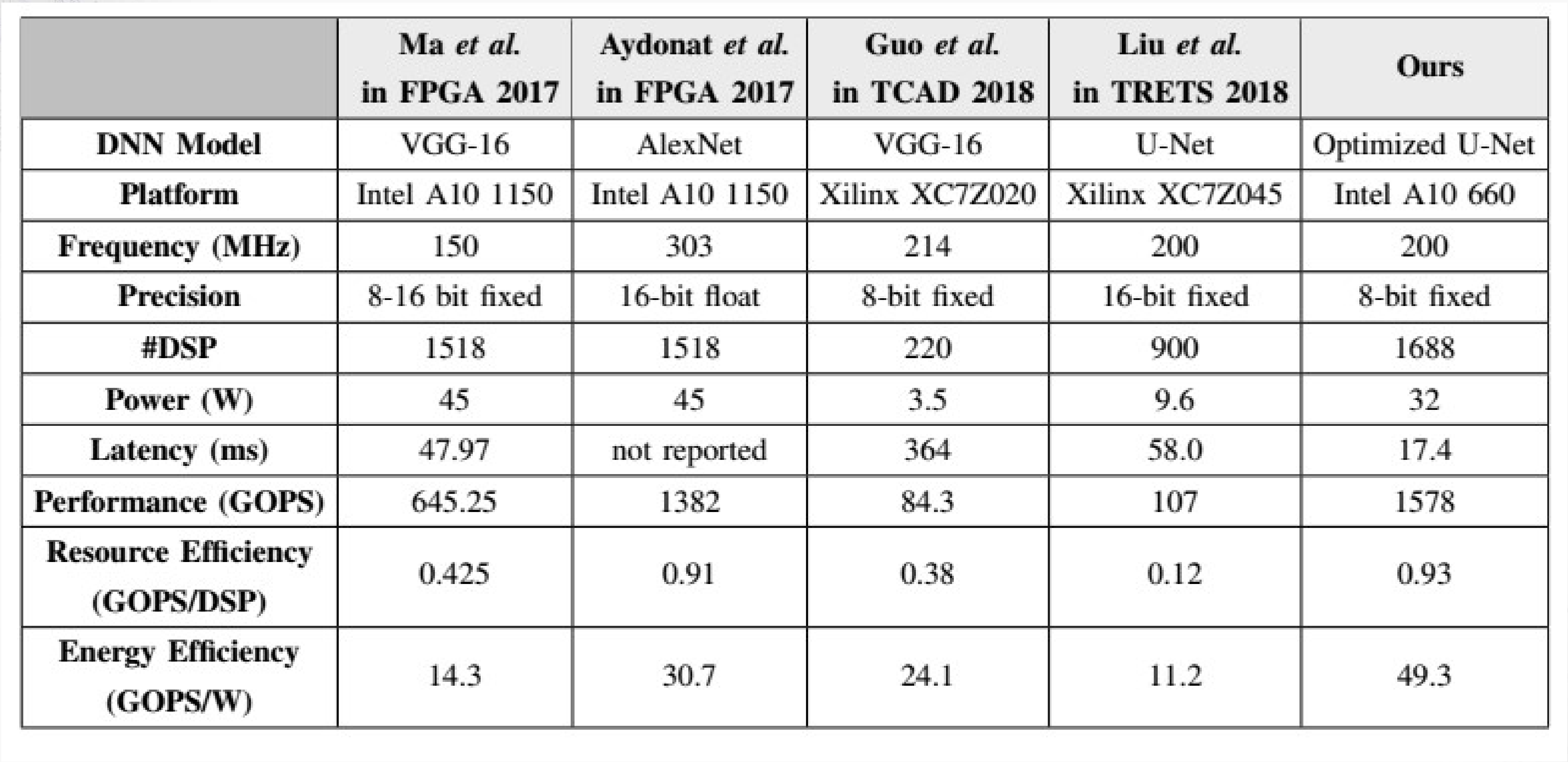
Unroll dot product

- Workload imbalance
- Computation Inefficiency

Optimizations

- Input Reshaping
- Layer Fusion
- DSP Configuration
- Model Compression

Evaluation: FPGA Accelerator vs. Prior Work



	Ma *et al.* in FPGA 2017	Aydonat *et al.* in FPGA 2017	Guo *et al.* in TCAD 2018	Liu *et al.* in TRETS 2018	Ours
DNN Model	VGG-16	AlexNet	VGG-16	U-Net	Optimized U-Net
Platform	Intel A10 1150	Intel A10 1150	Xilinx XC7Z020	Xilinx XC7Z045	Intel A10 660
Frequency (MHz)	150	303	214	200	200
Precision	8-16 bit fixed	16-bit float	8-bit fixed	16-bit fixed	8-bit fixed
#DSP	1518	1518	220	900	1688
Power (W)	45	45	3.5	9.6	32
Latency (ms)	47.97	not reported	364	58.0	17.4
Performance (GOPS)	645.25	1382	84.3	107	1578
Resource Efficiency (GOPS/DSP)	0.425	0.91	0.38	0.12	0.93
Energy Efficiency (GOPS/W)	14.3	30.7	24.1	11.2	49.3

- 1578 GOPS, 0.93 GOPS/DSP and 49.3 GOPS/W.
- 57 frames per second with a power consumption of 32 W.

Thanks

Welcome to the poster for more!