FORTH
INSTITUTE OF COMPUTER SCIENCE

An Experimental Analysis of the Opportunities to Use FPGA Multiprocessors for On-board Satellite Deep Learning Classification of Spectroscopic Observations from Future ESA Space Missions

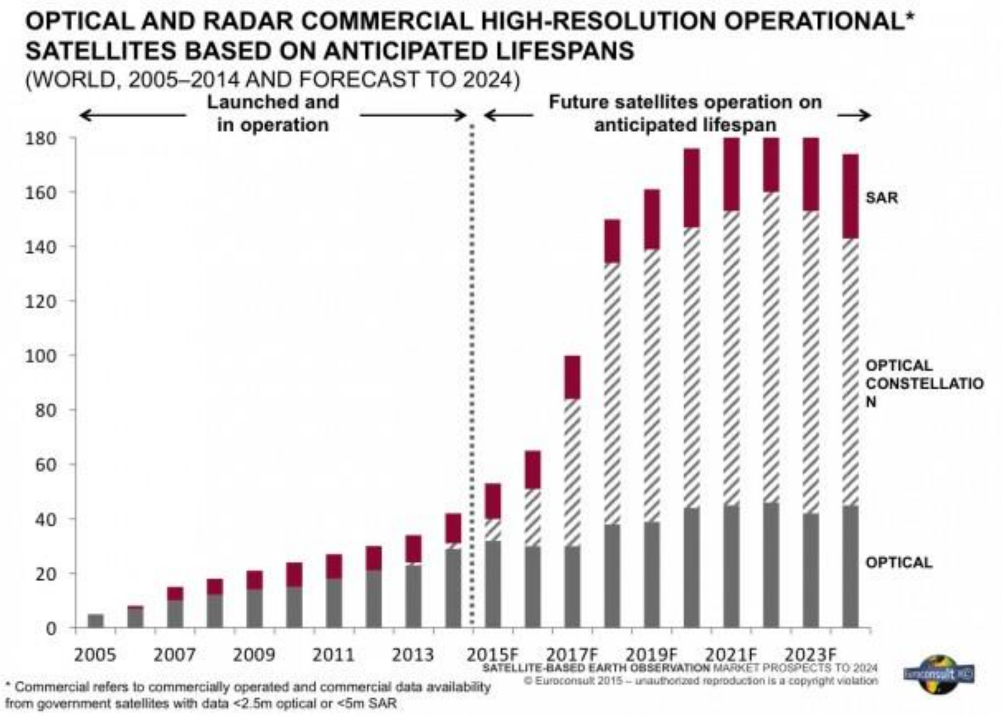
I. Kalomoiris, G. Pitsis, **Grigorios Tsagkatakis**, A. Ioannou,
C. Kozanitis, A. Dollas, P. Tsakalides, M. GH Katevenis

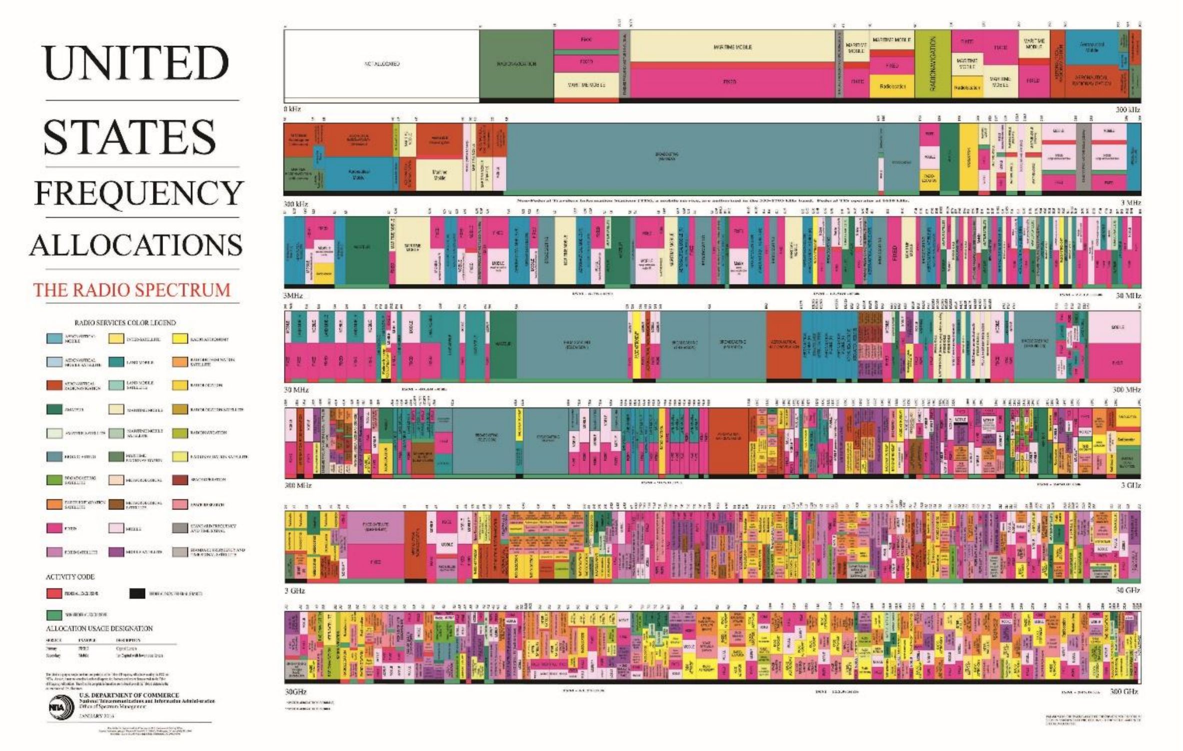
EURO EXA
FUNDED BY THE EUROPEAN UNION
DEDALE
European Commission
HORIZON 2020

Outline

- Introduction
- Optimization of Memory Footprint and Requirements
- FPGA Architectures
- Experimental Results
- Conclusions and Future Work

Motivation


OPTICAL AND RADAR COMMERCIAL HIGH-RESOLUTION OPERATIONAL* SATELLITES BASED ON ANTICIPATED LIFESPANS
(WORLD, 2005–2014 AND FORECAST TO 2024)
Launched and in operation
Future satellites operation on anticipated lifespan
180
160
140
120
100
80
60
40
20
0
2005
2007
2009
2011
2013
2015F
2017F
2019F
2021F
2023F
SAR
OPTICAL CONSTELLATIO N
OPTICAL
SATELLITE-BASED EARTH OBSERVATION MARKET PROSPECTS TO 2024
© Euroconsult 2015 – unauthorized reproduction is a copyright violation
* Commercial refers to commercially operated and commercial data availability from government satellites with data <2.5m optical or <5m SAR

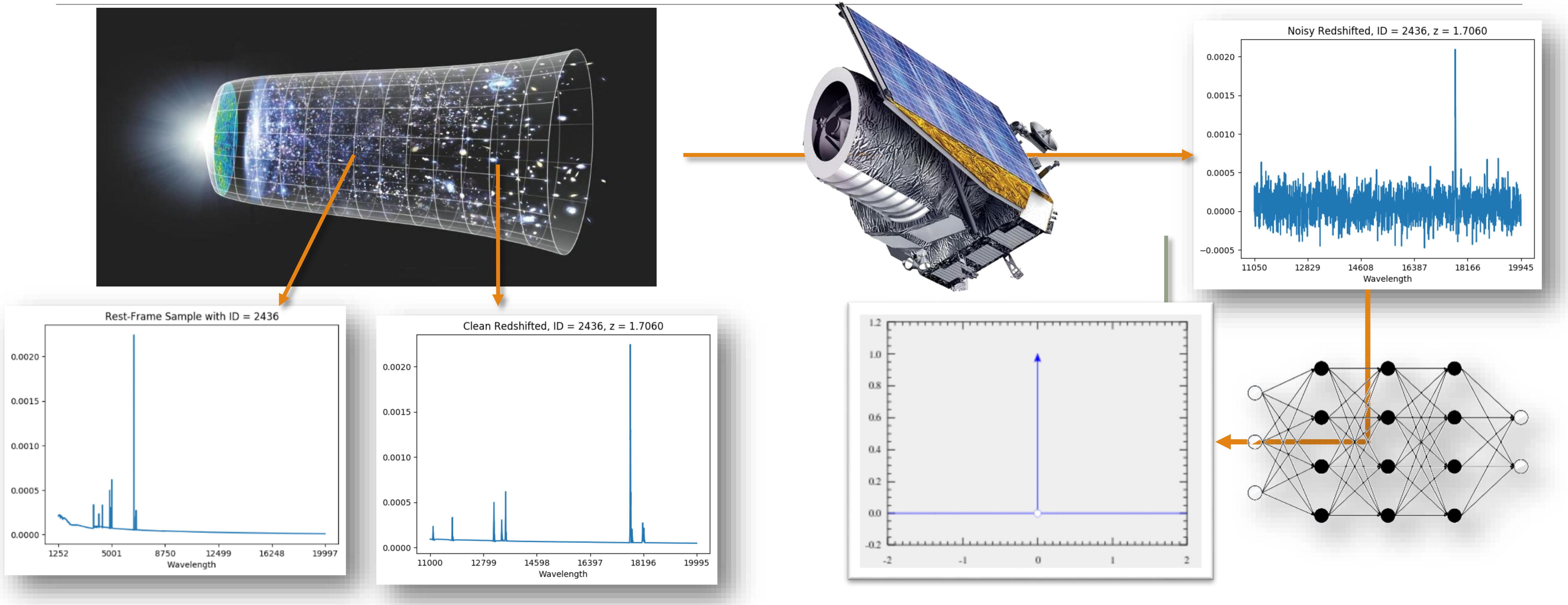

UNITED
STATES
FREQUENCY
ALLOCATIONS
THE RADIO SPECTRUM

Motivation

- Exponential growth of data (Copernicus 1.8 Petabytes/day)
- Limited growth in downlink bandwidth
- Proper management of these amount of data
- Deep Learning has become the driving force in AI
  - Convolutional Neural Networks (machine vision, speech recognition)
- FPGA-Based CNN accelerators (HPC, embedded applications)

FORTH
INSTITUTE OF COMPUTER SCIENCE

Analysis of EUCLID observations



R. Stivaktakis, G. Tsagkatakis, B. Moraes, F. Abdalla, J.-L. Starck, P. Tsakalides, "Convolutional Neural Networks for Spectroscopic Redshift Estimation on Euclid Data," EEE Transactions on Big Data: Special Issue on Big Data from Space, 2018

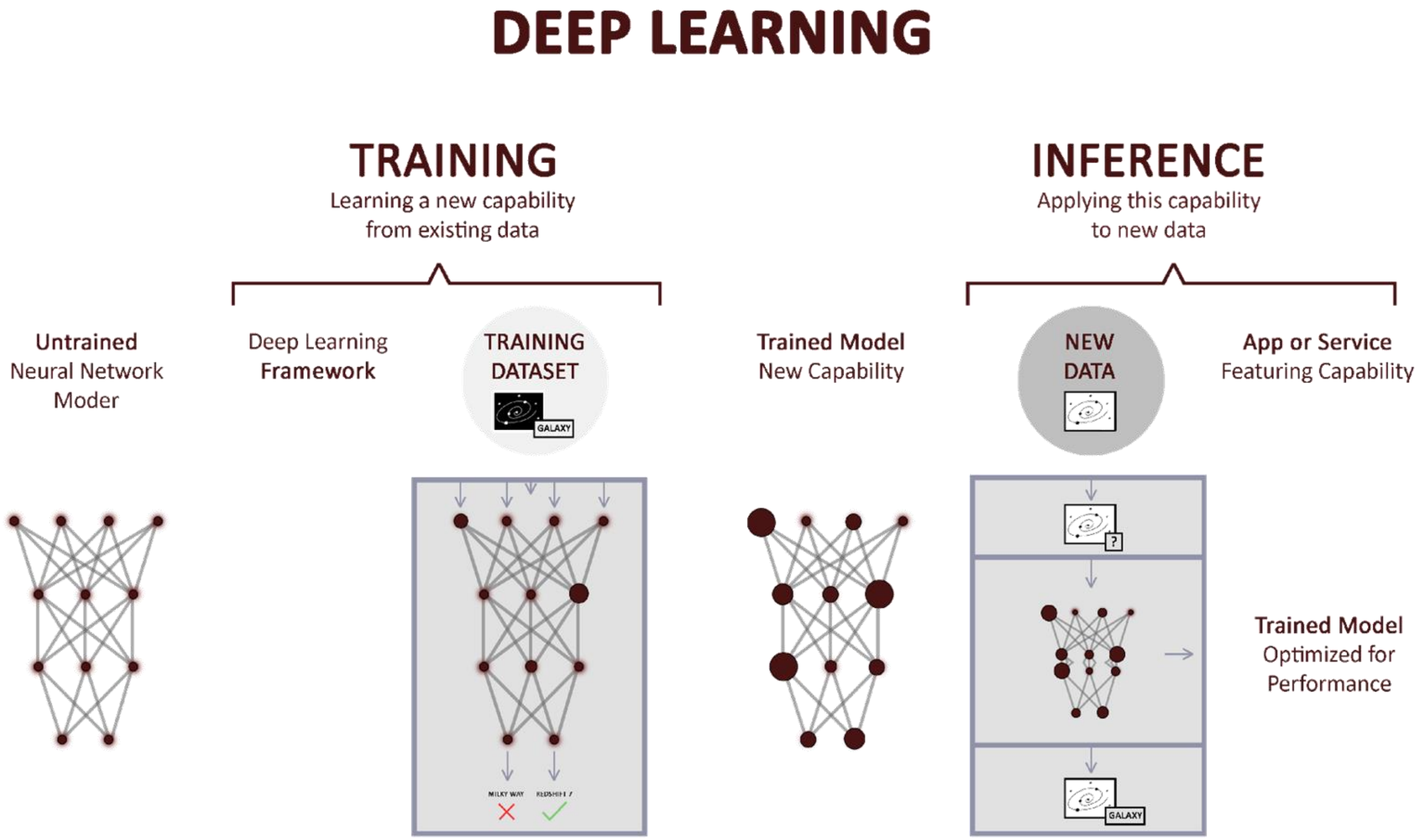
Deep Neural Networks

Training

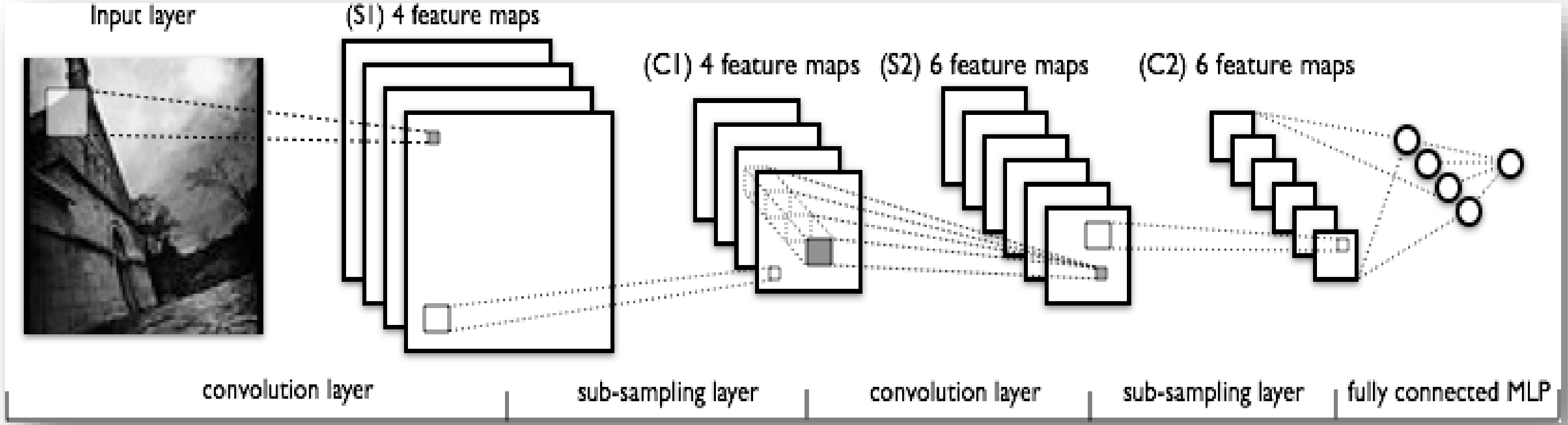
- ➢Large Dataset (Big Data)
- ➢Learning Features
- ➢Distributed processing

Inference

- ➢New observations
- ➢Run-time requirements



Convolutional Neural Networks



(Convolution + Subsampling) + () … + Fully Connected

LeCun, Yann, and Yoshua Bengio. "Convolutional networks for images, speech, and time series." The handbook of brain theory and neural networks 3361.10 (1995): 1995.

FORTH INSTITUTE OF COMPUTER SCIENCE

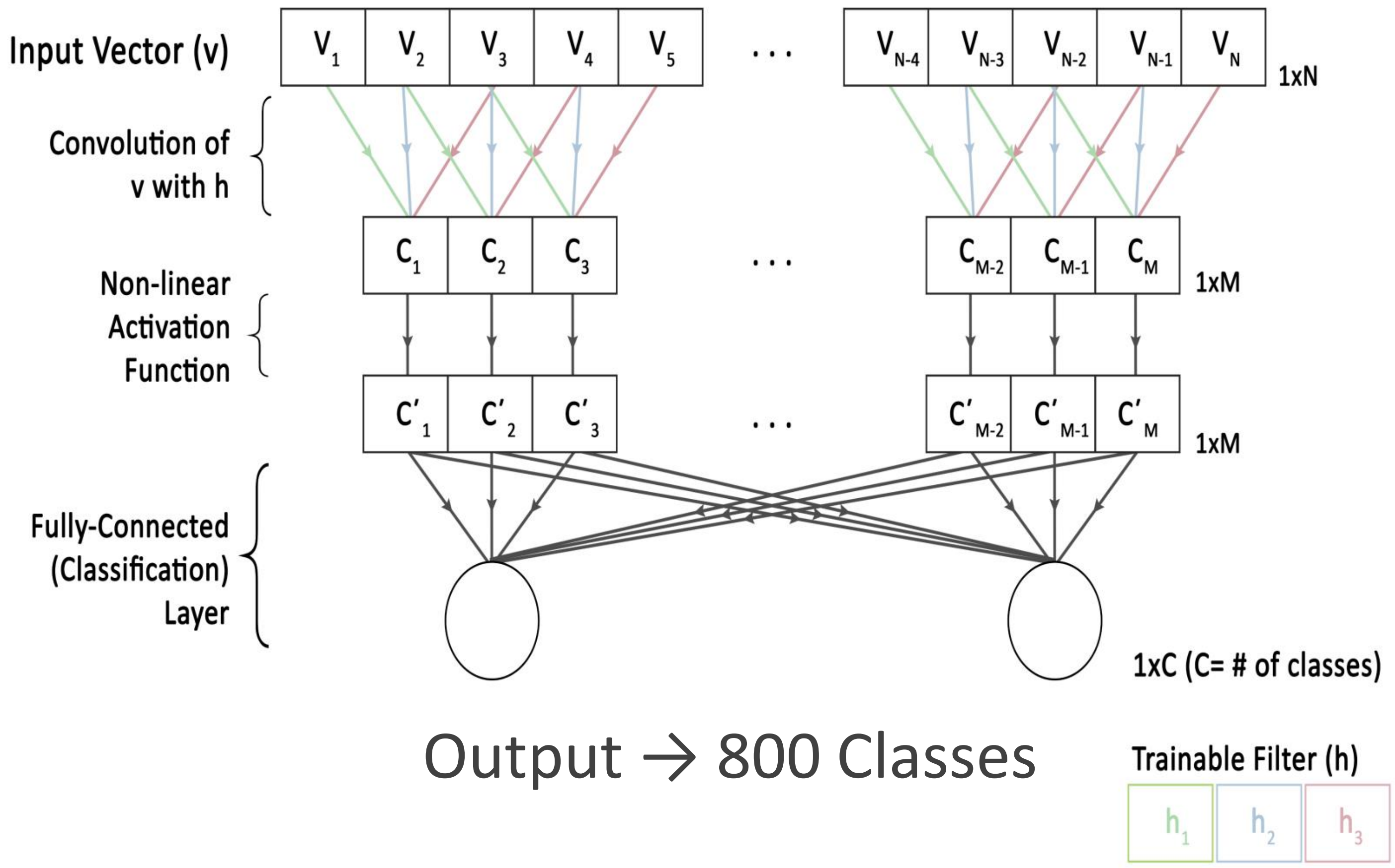
Spectroscopic Red-Shift Estimation with CNNs

- ESA Euclid specifications
- Training on NVIDIA GPU

Network consists of:

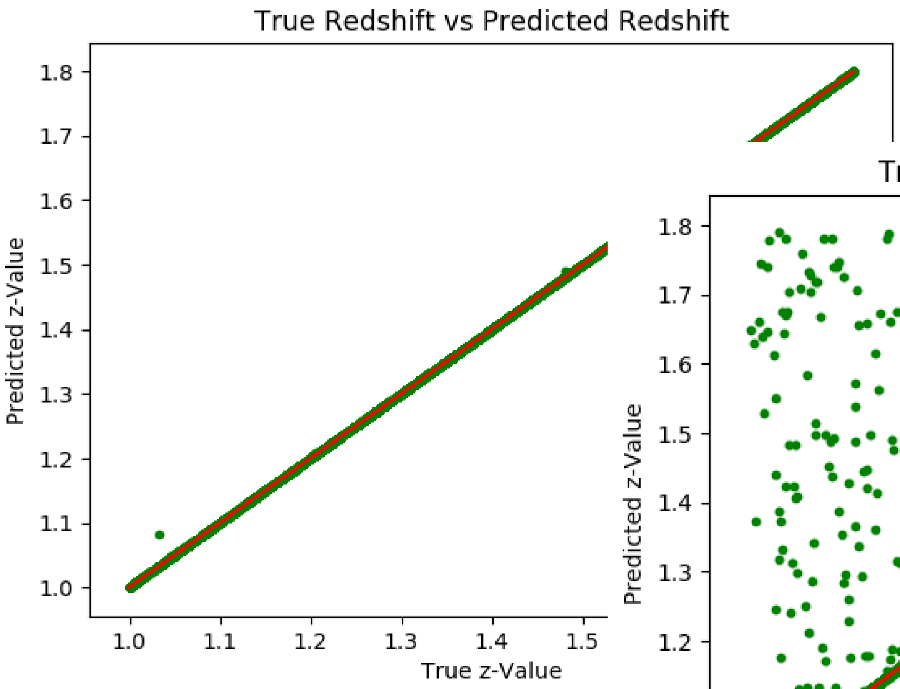
- Three Convolutional Layers
- Activations (ReLU)
- A Fully-Connected Layer

Input → 1x1800

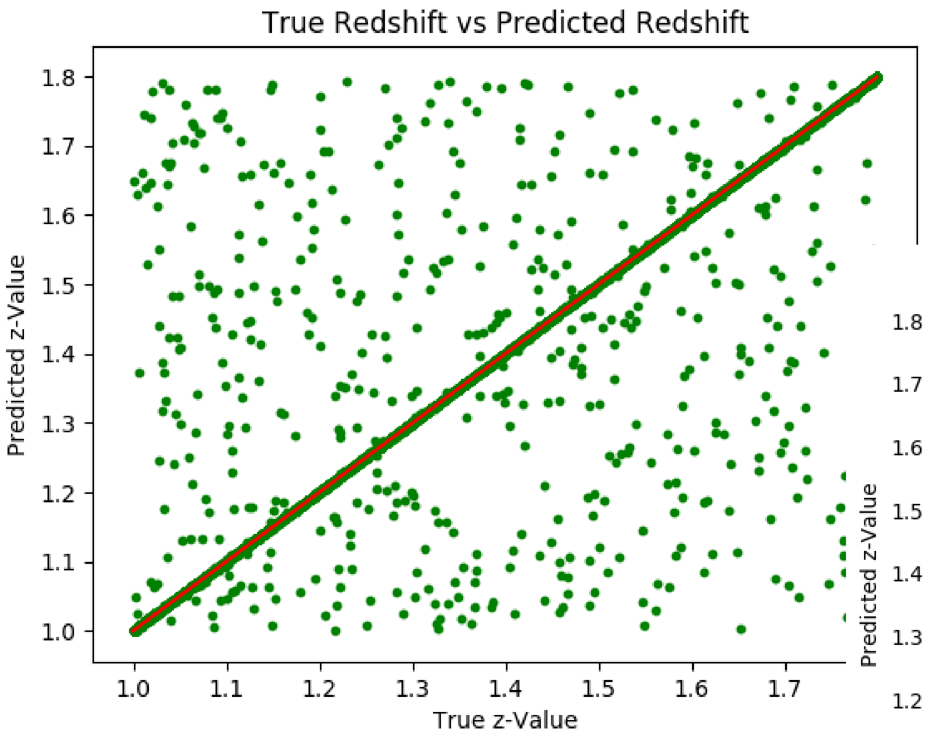


Output → 800 Classes

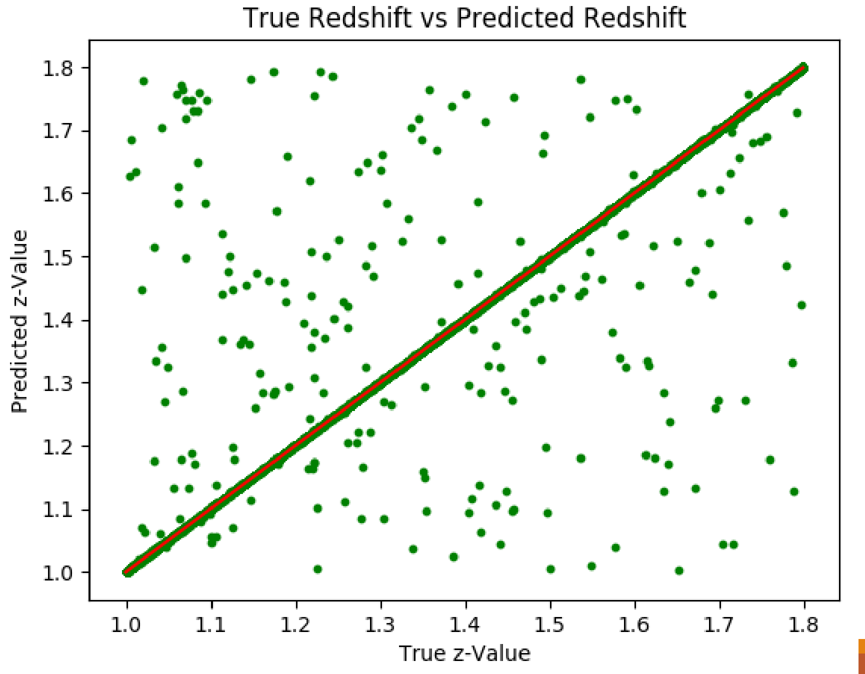
Prediction accuracy



Clean - 400K



Noisy – 400K



Noisy – 4M

Contributions

Explore methods to reduce

- Memory Footprint of weights → on-chip memory requirements
- Redundancy of CNN models → throughput & energy/input

Hardware design and realization

- Realization in single and quad FPGA architecture
- Comparison with GPU and CPU implementation

Inference

CNN was originally implemented and trained on TensorFlow

Implementation of inference in MATLAB creating an CNN-Toolbox

Objective

- Quantify performance gap between TensorFlow and Matlab
- Error = deviation of MATLAB vs TensorFlow

Data type (IEEE)	Error rate (%)
Double float	0
Single float	0.02
Half float	0.04

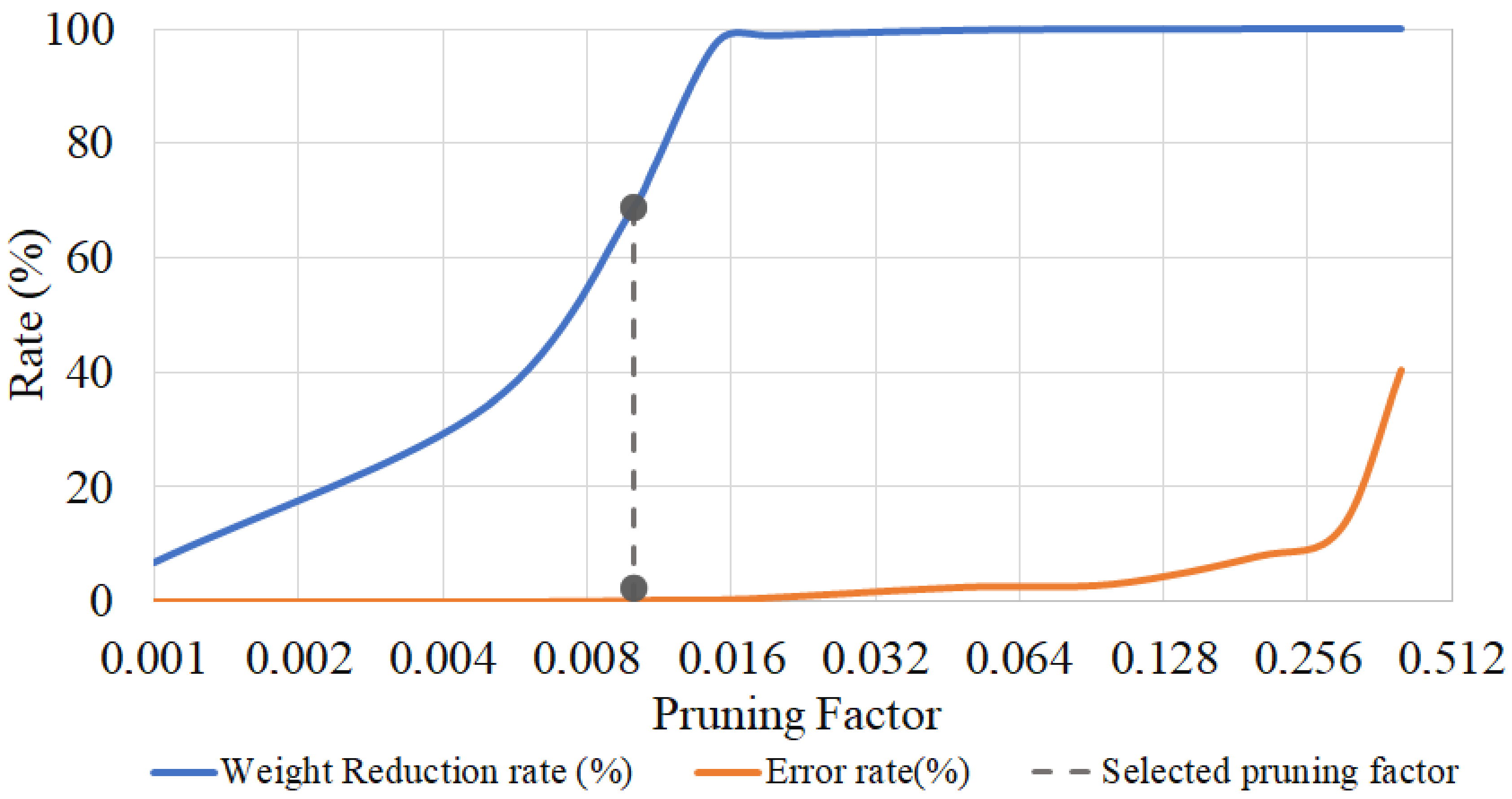
Memory Footprint

Layer	#Weights	Kernels size	Footprint
conv1	144	(16,8)	1.1 KB
conv2	2,064	(16,16,8)	16.1 KB
conv3	2,064	(16,16,8)	16.1 KB
fc	22,771,200	(800,28464)	173.7 MB

Methods to reduce FC Memory Footprint

- Single Float Precision
- Quantization with Codebook

Weight value pruning



Weight Quantization with Codebook

Lloyds Algorithm

- ➢ Group weights in k centroids and store them in a codebook
- ➢ Each weights stores the index of centroid
- ➢ Reduce memory footprint $(\log_2 k)$

codebook

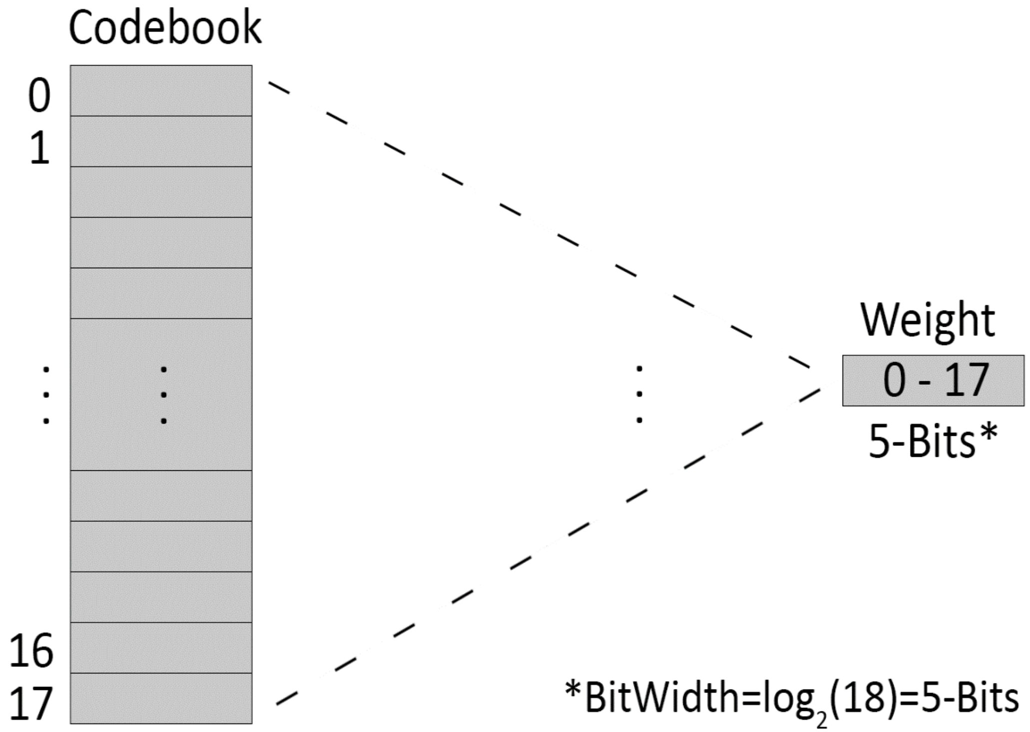
0	1	2	3
-0.54	-0.01	0.23	0.58

0.6	0.54	0.04	0.01
0.02	-0.04	-0.63	-0.54
0.05	-0.03	-0.02	0.03
0.3	0.23	-0.04	-0.05

Clustering → 4 centroids

cluster index

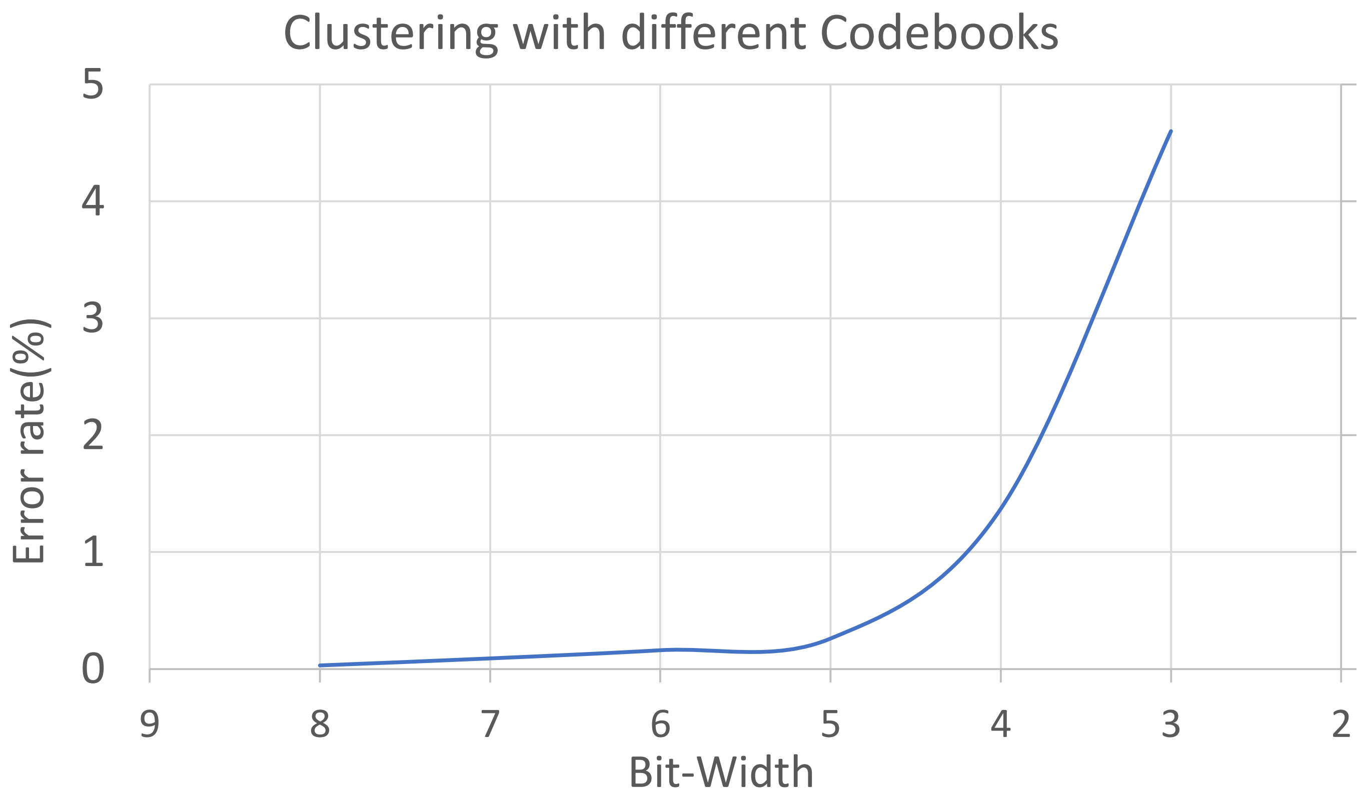
3	3	1	1
1	1	0	0
1	1	1	1
2	2	1	1



*BitWidth=$\log_2(18)$=5-Bits

Comparison of different codebooks

#Centroids	Bit-Width	Error rate(%)	Compression
256	8	0.03	8x
128	7	0.09	9.1x
64	6	0.16	10.7x
32	5	0.26	12.8x
16	4	1.37	16x
8	3	4.6	21.3x



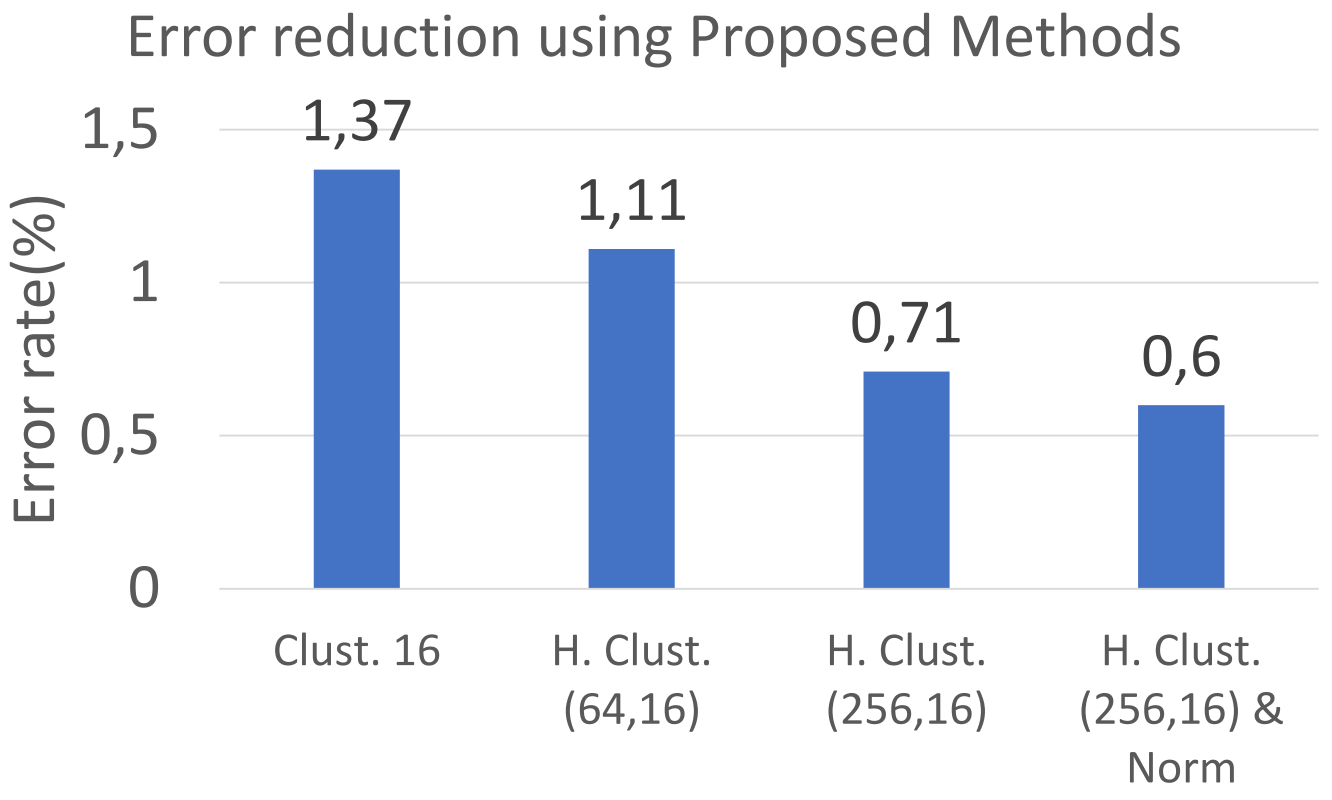
Methods for better clustering

Hierarchical Clustering : Clust. (n, k)

- Clustering with n centroids (256)
- Clustering with k centroids (16)

Inverse Density : (sigmoid function)

- Normalize Initial Codebook
- Importance => Resolution



Results using proposed optimizations

Method	Bits/Weight	Error rate(%)	Compression rate
Clust. 16	4	1.37	16x
H.Clust. 16	4	0.60	16x
P.C.& H.Clust. 16	2.5	0.62	25.6x
Q.C.& H.Clust. 16	1.75	0.76	36.57x
P.C.& H.Clust. 18 & SLC WB-12	1.3	0.5	49.24x
Q.C.& H.Clust. 18 & SLC WB-8	1.17	0.8	54.73x

Initial FC weights => Compressed FC weights

173.7 MB => 10.86 MB

Platforms used

ZCU-102

System Logic Cells	Block RAM	DSP Slices	High Performance Ports
600 K	4 MB	2520	4

QFDB

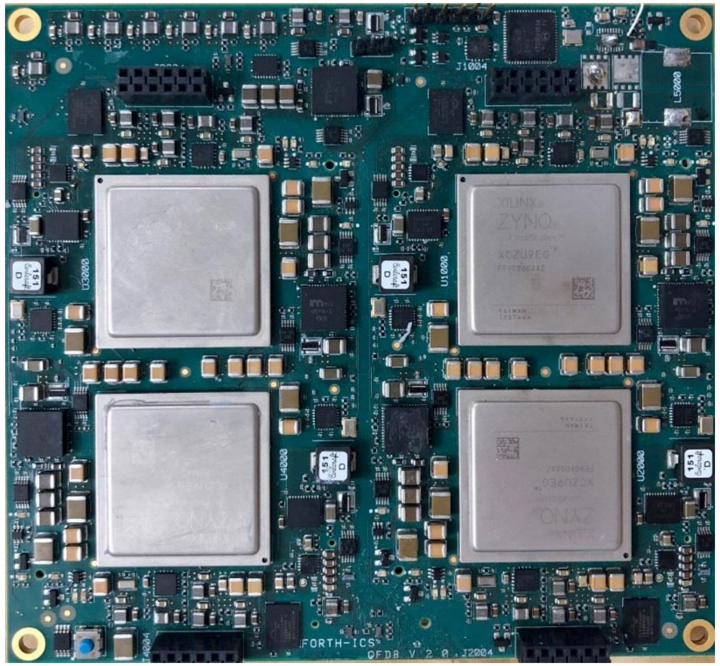
- Quad FPGA board, designed and implemented at FORTH
- 4 ZCU-102 in parallel

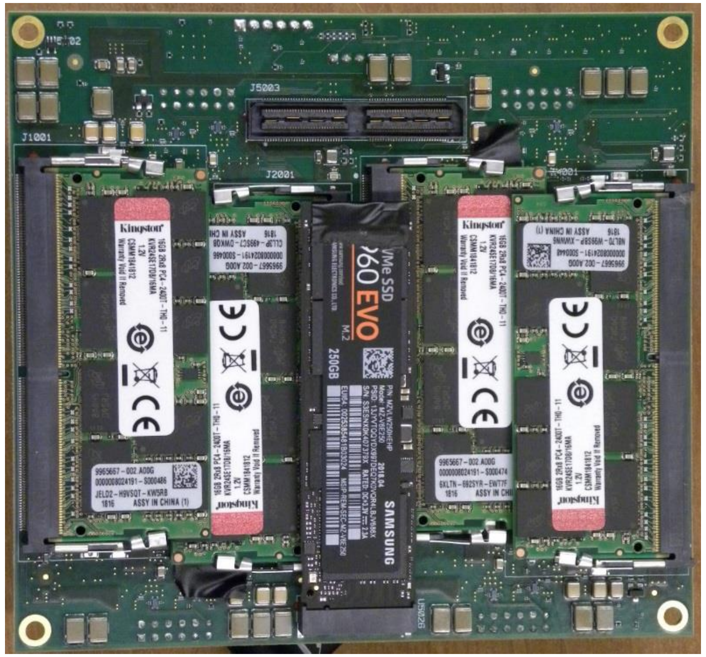
The Quad FPGA Daugher Board (QFDB)

FORTH
INSTITUTE OF COMPUTER SCIENCE

Based on Zynq Ultrascale+

- 4x ARM A53 (up to 1333MHz)
- Gen2 PCIe x4
- Programmable Logic
- DSP cores
- 3MBytes of SRAM
- 16 transceivers @16Gbps max.





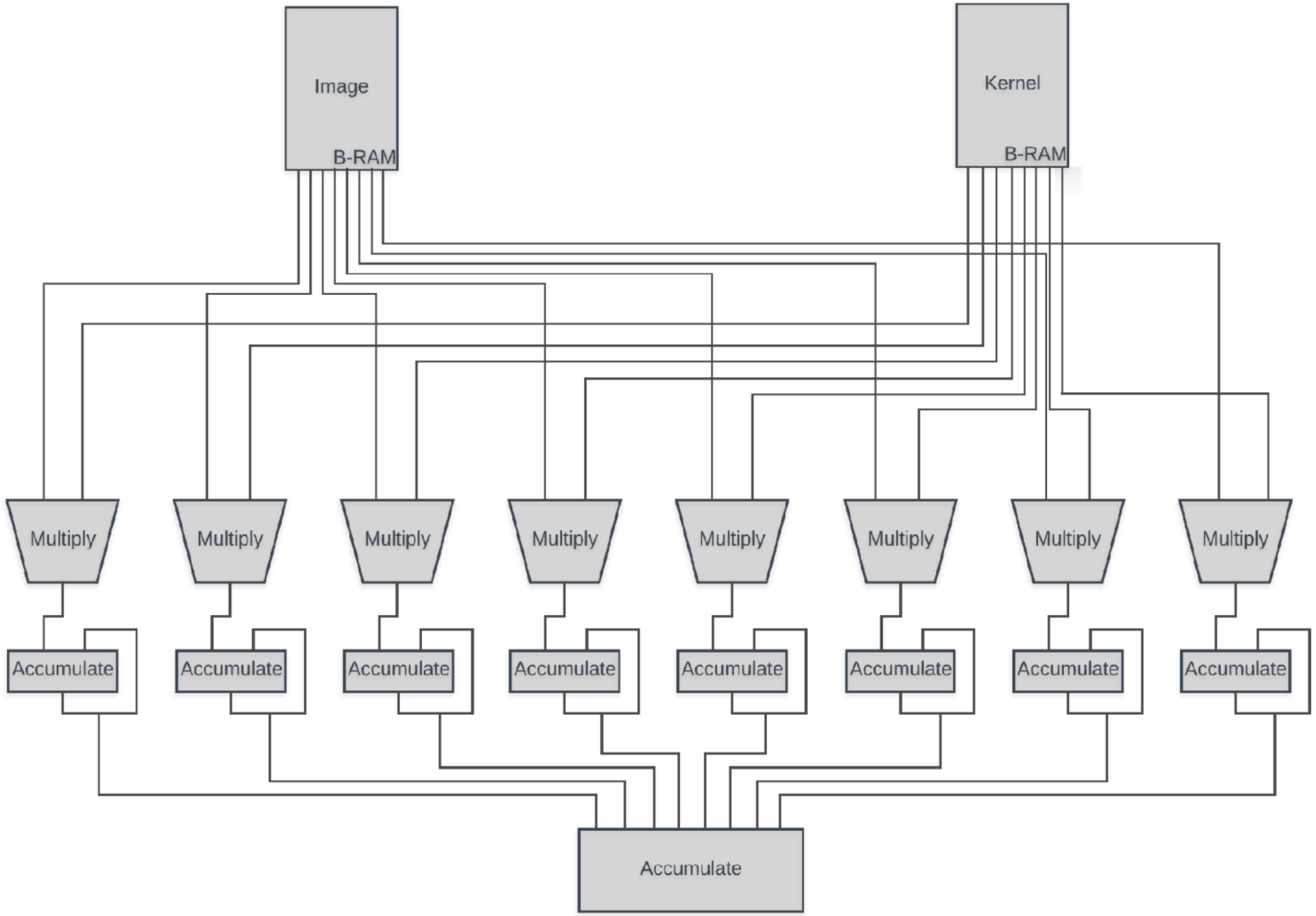
DDR4 SO-DIMMs

16Gbytes modules @ DDR4-PC2133

High density (120x130x25mm)

Convolution Module

- Convolution operation
  - Multiply and Accumulate
- 8 MAC operation in a cycle
- Shift Image for every kernel
- Data stored in B-RAM



Fully Connected Module

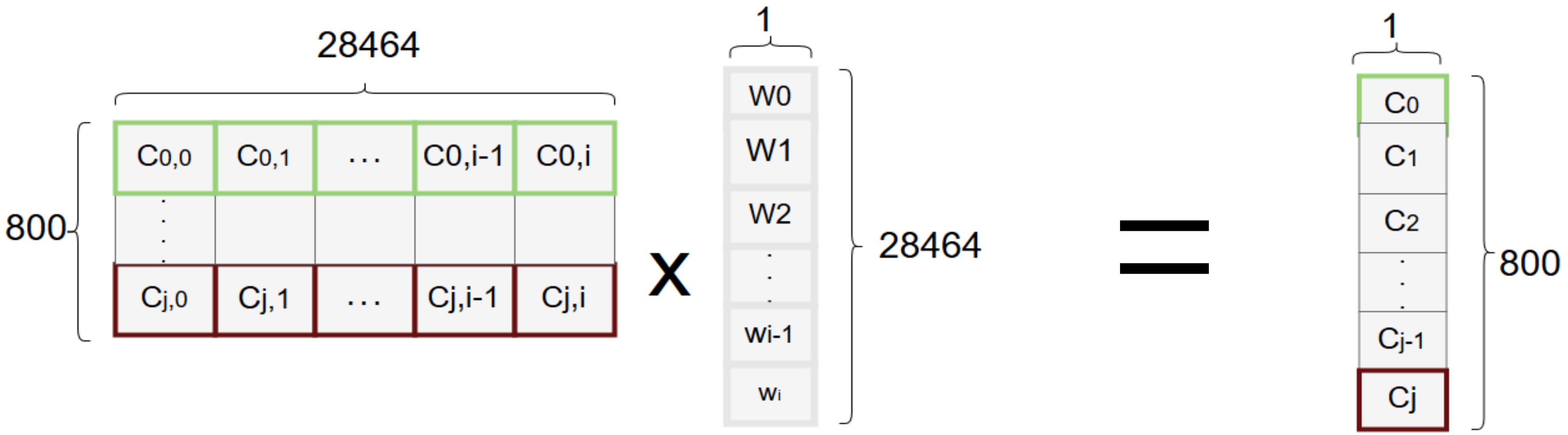
INSTITUTE OF COMPUTER SCIENCE

Matrix Multiplication

- Multiply and Accumulate

1 MAC operation in a cycle

Streaming data for every input spectrum



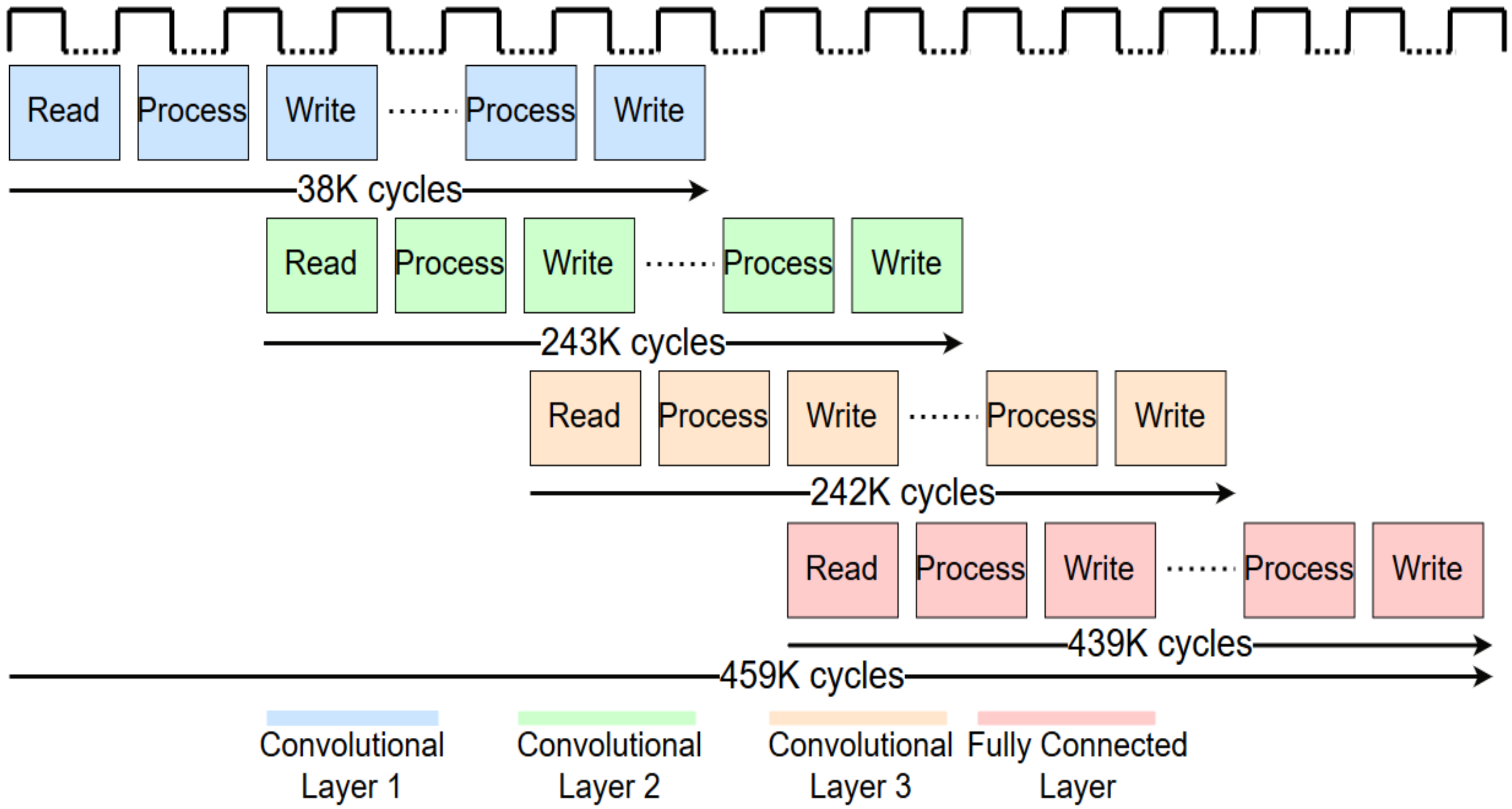
Pipelining Conv and FC Module

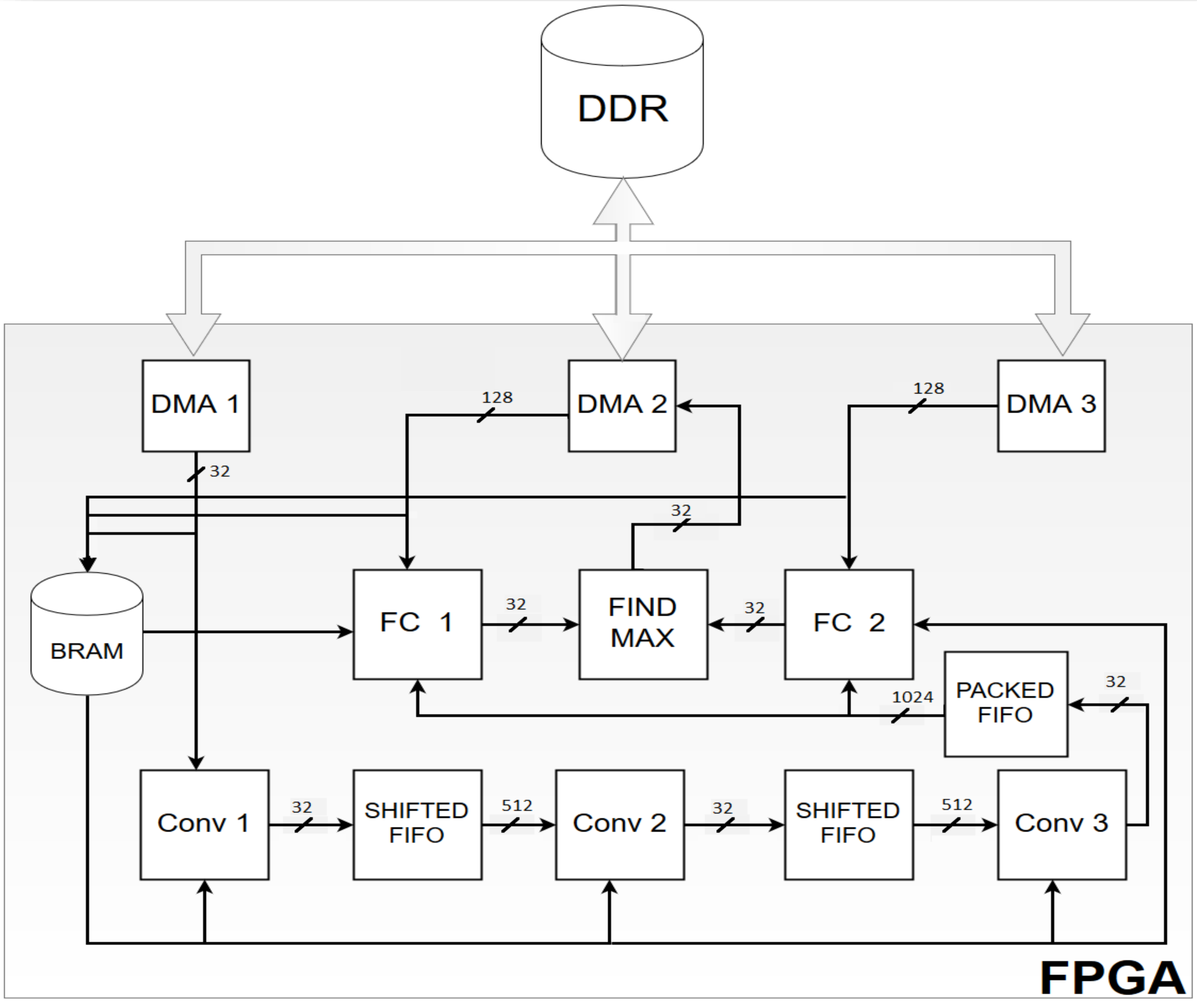
Convolutional Transformations

Fully Connected Transformations

2 Instances (400 classes)

	FLOPS	MAC / cycle
CONV	15.1M	40 (peak)
FC	45.6M	64
CNN	60.7M	104(peak)




DDR
DMA 1
DMA 2
DMA 3
128
128
32
32
BRAM
FC 1
FIND MAX
FC 2
32
32
PACKED FIFO
1024
32
Conv 1
32
SHIFTED FIFO
512
Conv 2
32
SHIFTED FIFO
512
Conv 3
FPGA

FPGA Architecture Enhancements

Batching 2 spectra instead of one → Doubling Calculations per sec

Resource Optimizations

- Custom Loop Unroll

Peak Performance => 416 GFLOPS

ZCU-102 (250 MHz)	FLOPS	MAC / cycle	GFLOPS
CONV	30.2M	80 (peak)	28.7
FC	91.2M	128	51.9
CNN	121.4M	208 (peak)	66.1

QFDB (250 MHz)	FLOPS	MAC / cycle	GFLOPS
CONV	120.8M	320 (peak)	57.4
FC	364.8M	512	104
CNN	485.6M	832 (peak)	265

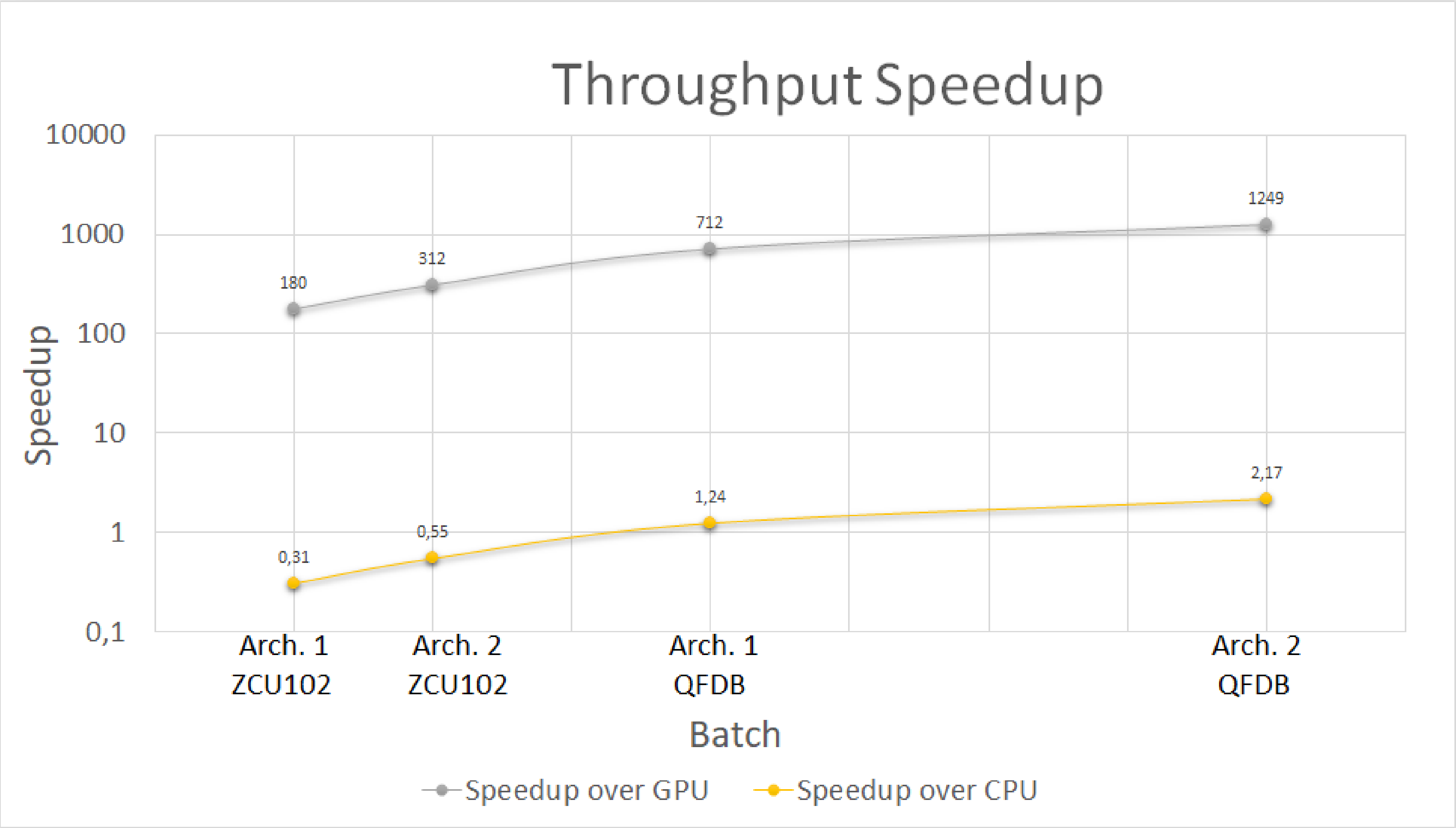
Comparison with CPU and GPU

INSTITUTE OF COMPUTER SCIENCE

Architectures v1 and v2 ported on ZCU-102 and QFDB (*for 10K spectra)

	FPGA Architecture		Intel i-7 7700HQ	Nvidia K2200
	ZCU-102	QFDB		
Clock Frequency (MHz)	250	250	3800	1124
Throughput (Spectra/s)	1084	**4334**	3.47	2000
Latency (s)	0.003	**0.003**	7.6	0.06
GFLOPS	66.1	**265**	0.21	122.5
TDP (Watt)	11.8	47.3	100	300
Energy Consumption (Joule)*	108.8	109.1	288K	1500
Spectra/Joule	91.9	**91.6**	0.035	6.66

Speedup and Efficiency



Speedup and Efficiency

- Latency and Throughput speedup
- Energy and Power Efficiency

	ZCU-102 vs CPU	**ZCU-102 vs GPU**
Latency speedup	2533x	20x
Throughput speedup	312x	0.55x
Energy Efficiency	2286x	11.9x

	QFDB vs CPU	**QFDB vs GPU**
Latency speedup	2533x	**20x**
Throughput speedup	1249x	**2.17x**
Energy Efficiency	2286x	**11.9x**

Conclusions and Future Work

Conclusions

- ✓ Compression of CNN weights → x54 less memory @ 0.8 error
- ✓ FPGA: through (x2) and energy (x10) speedup over GPU
- ✓ Flexibility and reconfigurability
- ✓ Train ↔ run-time : S/C ↔ G/S

Future Work

- ➢ Hardware Implementation of Pair-Compression and SLC
- ➢ Scale up to more FPGAs -> Mezzanine 8-QFDB
- ➢ Port to a space rad-hard FPGA

THANK YOU !!


WALL-E


Workshop on
Computational Intelligence
in Remote Sensing & Astrophysics
17-19 July 2019
FORTH