

Índice

Abstract	v
Resumen	vii
Resum	ix
List of Acronyms	xi
List of Figures	xv
List of Tables	xvii
1. Introduction	1
1.1. Multi-FPGA Systems and Their Challenges	4
1.2. Objectives of the Thesis	6
1.3. Structure of the Thesis	7
1.4. List of Publications	7
2. Background and State of the art	9
2.1. Convolutional Layers	10
2.2. FPGA as Inference Accelerators	12
2.3. Scaling to Multi-FPGA	14
3. Design of the Multi-FPGA System	17
3.1. Infrastructure and Physical Design	17
3.2. FPGA Design	21
3.3. CNN Accelerator – HLSinf	26
3.3.1. Input and Output Modules	27
3.3.2. Storage Modules	28
3.3.3. Compute Modules	29
3.3.4. Efficiency Results and Resources Overview	30
3.4. Architecture and Communication	33
3.4.1. Topology	33
3.4.2. Shared Distributed Memory Support	36
3.4.2.1. Evaluation	37
3.4.3. Streaming Support	40
3.4.3.1. Receive Mode	43
3.4.3.2. Transmit Mode	45
3.4.3.3. Evaluation	47
3.5. Scalability	48
3.6. Conclusions	50
4. Software Stack and Mapping Algorithms	51
4.1. Software Stack	51
4.1.1. Linux Driver XDMA	52
4.1.2. High-Level Abstraction Libraries	53

4.1.3. Python wrapper	57
4.1.4. User Applications	59
4.2. Mapping Methodologies	60
4.2.1. Partitioning Strategies	61
4.2.2. Partition-Oriented Mapping Algorithm	64
4.2.2.1. Evaluation	67
4.2.3. Streaming and Partitioning Algorithm	69
4.2.3.1. Evaluation	73
4.3. Conclusions	74
5. Conclusions	77
Bibliography	81