

Contents

Acknowledgements	iii
Abstract	v
Resumen	vii
Resum	ix
Contents	xi
List of Figures	xvii
1 Introduction	1
1.1 Motivation	6
1.2 Objectives	9
1.3 Organisation of this document	11
2 GPU-Enabled Serverless Workflows for Efficient Multimedia Processing	13
2.1 Introduction	14
2.2 Related Work	15
2.3 Architecture of the Serverless Processing Platform	17
2.3.1 Components	17
2.3.2 Integration of SCAR with AWS Batch	19
2.3.3 Functions Definition Language for Serverless Workflows	20
2.4 Serverless Workflow for Multimedia Processing	21

2.5	Results and Discussion	24
2.5.1	Analysis of the <i>Lambda-Batch</i> Execution Mode	25
2.5.2	GPU and CPU Comparison for Video Processing	28
2.5.3	AWS Batch Auto Scaling	30
2.5.4	Workflow Execution	31
2.6	Conclusions and Future Work	33
3	Serverless Workflows for Containerised Applications in the Cloud Continuum	35
3.1	Introduction	36
3.2	Related Work	38
3.3	Components to Support Serverless Workflows along the Cloud continuum	41
3.3.1	SCAR: Serverless Scientific Computing in Public Clouds	42
3.3.2	OSCAR: Open-Source Serverless Computing for Data-Processing Applications	43
3.3.3	Functions Definition Language for Data-Driven Serverless Workflows	46
3.4	Use Case: Mask Wearing Detection via Anonymised Deep Learning Video Processing	49
3.4.1	Optimal Resource Allocation for the Lambda Function	52
3.4.2	Case Study Design	54
3.4.3	Results and Discussion	55
3.5	Conclusions	60
4	A serverless gateway for event-driven machine learning inference in multiple clouds	63
4.1	Introduction	64
4.2	Related Work	66
4.3	Components and architecture	69
4.3.1	DEEPaaS API	70
4.3.2	Web-based Scientific Gateway	71
4.3.3	Serverless Frameworks	73
4.3.4	Architecture	75
4.4	Use Cases	76
4.5	Results	79
4.6	Discussion	88
4.7	Conclusions	89
5	Rescheduling Serverless Workloads Across the Cloud-to-Edge Continuum	91
5.1	Introduction	92
5.2	Related Work	94

5.3	Proposed Architecture	95
5.3.1	Resource Manager	98
5.3.2	Rescheduler	99
5.3.3	Delegation Mechanism	100
5.4	Use Case: Serverless Fire Detection Across the cloud-to-edge continuum . .	103
5.4.1	Case Study Design	105
5.4.2	Results and Discussions	107
5.5	Conclusions and Future Work	111
6	Discussion of the Results	113
6.1	Summary	113
6.2	Successful use cases	117
6.2.1	AI-SPRINT project	117
6.2.2	OSCAR as a backend for RDataFrame	120
6.3	Scientific contributions	121
6.4	Research projects	122
7	Conclusions	125
7.1	Future work	127
	Bibliography	129