Multimedia Cloud Computing Systems

Mohsen Amini Salehi • Xiangbo Li

Multimedia Cloud Computing Systems

Springer

Mohsen Amini Salehi
Computing and Informatics
University of Louisiana at Lafayette
Lafayette, LA, USA

Xiangbo Li
Twitch Interactive Inc.
San Francisco, CA, USA

ISBN 978-3-030-88450-5 ISBN 978-3-030-88451-2 (eBook)
https://doi.org/10.1007/978-3-030-88451-2

© Springer Nature Switzerland AG 2021, corrected publication 2022
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my parents, Naser Amini Salehi and Sedighe Afchangi, who taught me love to learn and learn to love! And to my wife, Bahar, who helped me practicing those.

Preface

From the emergence of Cloud technology in 2007–2010, the goal of cloud providers has been mitigating the burden of system administration and maintenance via consolidating servers and forming centralized datacenters. Since then, several progresses have been made in the computing world. Importantly, the sunset of Moore's law during the past decade has shifted the heterogeneous computing landscape from "consistently heterogeneous" systems that include different forms of CPU to "inconsistently heterogeneous" systems where domain-specific processors, a.k.a. Application-Specific Integrated Circuits (ASICs), are used. The emergence of GPU, TPU, FPGA, and most recently, Data Processing Unit (DPU) is an example of this shift in the heterogeneous computing landscape. Datacenters with inconsistently heterogeneous machines will be the bedrock of next-generation cloud computing systems. We foresee that the increasing desire for high-level services in certain areas, such as Machine Learning, data analytics, mapping, and multimedia streaming alongside with the ASIC technology in cloud datacenters will give rise to the idea of domain-specific cloud systems in the remaining years of the current decade. While first generation of cloud systems eased the job of system administrators, the second generation aims at mitigating the burden for programmers and solution architects by providing them with high-level abstractions, APIs, and services within a certain context.

With this vision in mind, we observe a shortage for a platform to enable the goals of a domain-specific cloud. Accordingly, the goal of authoring this book is to overcome this shortage and lay the foundation for the platform of domain-specific clouds. Throughout the book, we particularly concentrate on multimedia streaming as the context and develop the Multimedia Streaming Cloud (MSC) platform. The platform will be cloud native and based on the serverless computing paradigm that will provide an isolated and low-overhead framework for the next generation of cloud-scale computing in the particular domain of multimedia streaming. The reason that we concentrate on the multimedia cloud is the popularity of these services in the current and future of the computing and network industry. Moreover, Dr. Amini and his research team in the High Performance Cloud Computing (HPCC) research group (http://hpcclab.org) have accumulated more than 6 years

of knowledge and experience in cloud-based multimedia streaming that provides us with the vision needed to develop such platform and author this book. The second author of this book, Dr. Xiangbo Li, is an experienced research scientist and practitioner in cloud-based multimedia streaming. He was the first graduate of the HPCC Lab in 2016, where he worked on on-demand video transcoding using Cloud Services and under Dr. Amini's supervision laid the foundation of this research track in the HPCC Lab. Dr. Li is now an experienced software engineer and researcher on video platforms at the Twitch Interactive, a subsidiary of Amazon.com, Inc., where he worked on projects such as Amazon Interactive Video Service (AWS IVS) and several other projects.

This book targets audience in both academia and industry. In particular, researchers and practitioners in the Cloud and multimedia areas, solution architects, cloud programmers, and college students can benefit from the content of this book. We hope that this book can push the boundaries of knowledge in computing, help democratizing multimedia streaming, and make multimedia content a more useful tool to improve the life of human beings.

Lafayette, LA, USA Mohsen Amini Salehi
July 2021

The original version of this book was revised: Affiliation of Author Xiangbo Li has been corrected. The correction to this book is available at https://doi.org/10.1007/978-3-030-88451-2_11

Acknowledgments

This book is the result of more than 6 years of research and development in the High Performance Cloud Computing (HPCC) research laboratory. A postdoctoral associate, Dr. Chavit Denninnart, and multiple Ph.D. and Master's students have helped and contributed significantly in developing the vision needed for authoring this book. We like to particularly acknowledge the contributions of Dr. Mahmoud Darwich, Mr. Davood Ghatreh-Samani, Mr. Ali Mokhtari, Mr. Vaughan Veillon, Mr. James Gentry, Yamini Joshi, Brad Landreneau, and our research visitor Shangrui Wu.

Authoring this book and conducting its research projects is supported by the National Science Foundation CAREER award titled, "CAREER: Developing a Flexible Serverless Multimedia Streaming Cloud Platform" (Award No. CNS-2047144).

Contents

1 Introduction 1
1.1 Overview 1
1.2 Multimedia Streaming and Cloud Computing 2
1.3 The Essence of This Book 4
1.4 Characteristics of the Multimedia Streaming Cloud (MSC) 5
1.4.1 Quality of Experience (QoE) in the MSC Platform 5
1.4.2 Robustness of the MSC Platform 5
1.4.3 Function-as-a-Service and Serverless Computing in the MSC Platform 6
1.5 A Bird-Eye View of the Multimedia Streaming Cloud (MSC) Platform 7
References 9

2 Demystifying Multimedia Streaming Workflow 11
2.1 Overview 11
2.2 Video Streaming Types 13
2.3 Video Transcoding 15
2.3.1 Bit Rate 15
2.3.2 Resolution 15
2.3.3 Frame Rate 16
2.3.4 Codec 16
2.4 Video Delivery 17
2.4.1 Packaging 17
2.4.2 Encryption 17
2.4.3 Streaming Protocols 17
2.5 Content Delivery Network (CDN) 19
2.6 Video Playback 20
2.7 Summary 20
References 21

3 Multimedia Cloud: Designing a Special-Purpose Cloud Platform for Interactive Multimedia Streaming 23
3.1 Overview 23
3.2 Characterizing the Multimedia Streaming Cloud (MSC) Environment 24
3.2.1 Stakeholders of MSC 25
3.2.2 Characteristics of Multimedia Streaming Tasks 25
3.2.3 Uncertainty 27
3.3 Architecture of the Multimedia Streaming Cloud (MSC) Platform 28
3.3.1 Object-as-a-Service (OaaS) Abstraction in the MSC Platform 31
3.3.2 Enabling Object-as-a-Service (OaaS) Abstraction in the MSC Platform 32
3.3.3 Enabling Live Object Migration 33
3.3.4 Single Pane of Glass to Objects in MSC 34
3.4 Summary 35
References 35

4 Applications of Multimedia Clouds 39
4.1 Overview 39
4.2 Multimedia Streaming Types 40
4.2.1 On-Demand Multimedia Streaming 40
4.2.2 Live Multimedia Streaming 44
4.2.3 Live-to-VOD Streaming 45
4.2.4 Differences in Processing Live and VOD Streaming 45
4.3 Basic Services Offered by MSC 46
4.3.1 Multimedia Content Transcoding 46
4.3.2 Video Packaging 48
4.3.3 Analytical Services of Multimedia Streaming 48
4.4 Advanced Services Offered by MSC 50
4.4.1 Smart (AI-Based) Multimedia Streaming Services 50
4.4.2 Augmented Reality (AR) and Virtual Reality (VR) Streaming 50
4.4.3 Holographic Multimedia Streaming 51
4.4.4 360° Multimedia Streaming 52
4.5 Summary 53
References 53

5 Computing Infrastructure for Multimedia Streaming Clouds (MSC) 57
5.1 Overview 57
5.2 Virtualization Platforms for MSC 58
5.2.1 Case-Study: Appropriate Virtualization Platform for Multimedia Processing Using FFmpeg 60

5.3 Heterogeneous Computing for Multimedia Streaming Clouds (MSC) 63
5.3.1 Heterogeneous Resource Provisioning in MSC 65
5.3.2 Case-Study: Performance Analysis of Video Transcoding Operations on Heterogeneous Cloud VMs 66
5.4 Performance-Cost Trade-Off of Multimedia Processing on Heterogeneous Cloud VMs 76
5.4.1 Modeling Performance Versus Cost Trade-Off of Transcoding Tasks on Heterogeneous VMs 77
5.4.2 Case Study of the Cost Performance Trade-Off Model 79
5.5 Scheduling of Multimedia Segments on Heterogeneous Machines in MSC 82
5.5.1 QoS-Aware Multimedia Task Scheduler 83
5.5.2 Self-Configurable Heterogeneous VM Provisioner 87
5.6 Case-Study: Making Use of Heterogeneous Computing in Live-Streaming Industry 92
5.7 Summary 93
References 94

6 Service Reuse in Multimedia Clouds 97
6.1 Overview 97
6.2 Is Function Aggregation (Merging) Beneficial? A Case-Study on Video Transcoding 99
6.2.1 Introducing Video Benchmark Dataset 100
6.2.2 Case-Study: Benchmarking Execution-Time of Video Transcoding Tasks 101
6.2.3 Analyzing the Impact of Merging Video Tasks 102
6.3 Predicting the Execution-Time Saving of Aggregating Functions 105
6.3.1 A Model to Predict Execution-Time Saving 105
6.3.2 Gradient Boosting Decision Tree (GBDT) to Predict the Execution-Time Saving 106
6.3.3 Performance Evaluation of the Execution-Time Saving Predictor 107
6.4 Function Aggregation in the Admission Control Unit of MSC 111
6.5 Task Similarity Detection 112
6.5.1 Categories of Mergeable Tasks 112
6.5.2 Detecting Tasks of Similar Functions 113
6.6 Identifying Merging Appropriateness 114
6.6.1 Overview 114
6.6.2 Evaluating the Impact of Merging 115
6.6.3 Positioning Aggregated Tasks in the Scheduling Queue 116

6.7 Adapting Merging Based on the Oversubscription Level 119
6.7.1 Overview 119
6.7.2 Quantifying Oversubscription in the MSC Platform 120
6.7.3 Adaptive Task Merging Aggressiveness 121
6.8 Summary and Discussion 121
References 122

7 Low-Latency Delivery Networks for Multimedia Streaming 125
7.1 Overview 125
7.2 Content Delivery Networks (CDN) 128
7.3 Peer to Peer (P2P) Networks 129
7.4 Fog Delivery Networks (FDN) Versus Content Delivery Networks (CDN) 131
7.4.1 Federated Fog Delivery Networks (F-FDN) 132
7.4.2 Efficient Operation of F-FDN 135
7.4.3 Introducing Different Streaming Delivery Methods for Evaluation 137
7.4.4 Evaluation of Stream Delivery Methods 139
7.5 Streaming Protocols 144
7.6 Case-Study: Low-Latency Streaming in Practice 145
7.7 Summary 147
References 148

8 Other Aspects of Multimedia Clouds 153
8.1 Domain-Specific Billing 153
8.2 Networking of the Multimedia Streaming Clouds 154
8.3 Security of Multimedia Streaming 155
8.3.1 Privacy 155
8.3.2 Digital Rights Management 156
8.4 Storage Service for Multimedia Contents 157
8.4.1 Cloud Storage for Multimedia Streaming 159
8.5 Summary 161
References 161

9 Prototype Implementation of the MSC Platform 163
9.1 Overview 163
9.2 Serverless Computing Paradigm in Practice 163
9.3 A Use Case for the MSC Platform 164
9.4 Characteristics of Multimedia Stream Processing 165
9.5 Architecture of the MSC Prototype Implementation 165
9.5.1 Media Repository 166
9.5.2 Service Repository 167
9.5.3 Request Ingestion 167
9.5.4 Task Admission Control 168
9.5.5 Task Queue 168
9.5.6 Task Scheduler 168

9.5.7 Task Execution Time Estimator 169
9.5.8 Execution Engine 169
9.5.9 Provisioning Manager and Elasticity Manager 171
9.5.10 Stream Manager 171
9.5.11 Media Caching 172
9.6 Performance Evaluation 172
9.6.1 Experimental Setup 172
9.6.2 Evaluating Task Execution Unit Configurations 173
9.6.3 Evaluating Scheduling Policy 174
9.7 Summary 177
References 177

10 Future of Multimedia Streaming and Cloud Technology 179
10.1 Overview 179
10.2 Application-Specific Integrated Circuits (ASICs) in Domain-Specific Clouds 179
10.3 Efficient Scheduling of Functions on Heterogeneous Machines ... 180
10.4 Supporting Both Live and On-Demand Multimedia Streaming on the Same Underlying Resources 181
10.5 Blockchain Technology for Multimedia Streaming 182
10.6 Reuse and Approximate Computing of Functions in Multimedia Clouds and Other Domain-Specific Clouds 182
10.7 Machine Learning for Multimedia Processing 183
10.8 Avoiding Bias in MSC and Other Domain-Specific Clouds 184
10.9 Dependability of Cloud-Based Multimedia Streaming 185
References 186

Correction to: Multimedia Cloud Computing Systems C1