A Generalizable Framework for Automated Cloud Configuration Selection

Supervisors: Adam Barker & Yuhui Lin

Jack Briggs - 140011358

MSc Data-Intensive Analysis

2019-06-06

Abstract

Outline of the project using at most 250 words

Declaration

I declare that the material submitted for assessment is my own work except where credit is explicitly given to others by citation or acknowledgement. This work was performed during the current academic year except where otherwise stated. The main text of this project report is NN,NNN* words long, including project specification and plan. In submitting this project report to the University of St Andrews, I give permission for it to be made available for use in accordance with the regulations of the University Library. I also give permission for the title and abstract to be published and for copies of the report to be made and supplied at cost to any bona fide library or research worker, and to be made available on the World Wide Web. I retain the copyright in this work.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 Cloud Computing

Cloud computing is an ever-growing field that now ranges from Infrastructure-as-a-service (IaaS) to Software-as-a-service(SaaS). Services under cloud computing are characterised by their ability to offer access to a shared pool of highly elastic on-demand computing resources that offer broad network access [1,2]. Cloud services as an industry has had an explosive growth, and it has been predicted that 83% of enterprise (Companies with 1000+ employees) workloads will be in the cloud by 2020 [3], with 41% run on public cloud platforms such as Amazon AWS and Microsoft Azure. Services offered range from various levels and forms of abstractions, from directly provisioning Virtual Machines (VMs) or storage services, allowing users full control over their cloud infrastructure, to deploying 'serverless' containers, where the actual managing of the hardware is instead handled by the cloud provider.

The appeal is obvious, with cloud services allowing organizations and developers to utilize a diverse range of computational resources on demand without any up-front commitment or cost [4]. This can lead to both significant cost-savings as well as improved revenue through better customer experiences and enabling risk-free experimentation [5]. Academics, too, are utilizing the available services as volumes of data grow impractical to store and analyse on local machines [6,7]. This includes large-scale collaborative projects involving huge data sets hosted on the cloud such as the 1000 Genomes Project[1] or Common Crawl[2].

[1] https://aws.amazon.com/1000genomes
[2] commoncrawl.org

1.1.2 Cloud optimization

A wide range of applications are now deployed on cloud machines or make use of objects stored on them, from large-scale data analytics jobs mentioned to media-streaming servers such as Netflix or Twitch [8]. The resource dependencies of these applications similarly vary widely, from the CPU dependent data analysis tasks to network-heavy streaming services. Virtual machines offered by different cloud providers vary in terms of memory amounts of number and speed of virtual CPUs (vCPUs), and each application's performance will have different relationships with these options. While medium-length video transcoding operations will benefit primarily from faster processing speeds, data analysis tasks involving large datasets may find a more cost-effective option in prioritising VMs with a local solid-state drive (SSD) offering high I/O performance.

1.1.3 Benefits

It is desirable for both users and providers to maximise the optimize purchased cloud configurations to best serve the needs of their applications. Users or developers who fail to do this risk paying far more than they need to for the same performance. A given data analysis task can cost around 3.4 times as much on an average configuration compared to the optimal available option [9]. Even serverless frameworks simply shift the burden of optimization from the users to the cloud providers. For cloud providers too, efficient deployment across available Virtual Machines frees up extra resources available for other purposes or other customers. In addition, energy-related costs make up to 42% of managing a data-centre, and the ability to idle inactive resources would lead to a significant reduction both in energy cost and environmental impacts. [10].

1.1.4 Difficulties

Search space

Heterogeneity

Variation

Application range

Objective measure

1.2 Aims and Objectives

1.3 Contributions

1.4 Dissertation Outline

Chapter 2

Literature Survey

2.1 Cloud prevalence

2.2 Cloud variability

2.3 Optimization methods

2.3.1 CherryPick

2.3.2 PARIS

2.3.3 Ernest

2.3.4 Daleel

2.3.5 OTHERS, LOOK UP

2.3.6 Exhaustive search

2.4 Benchmarks

2.4.1 Cloudsuite

2.4.2 vBench

2.4.3 OTHERS, LOOK UP

2.5 Infrastructure-as-code

2.5.1 Terraform

Chapter 3

Requirements specification

3.1 Use-case

3.2 Requirements

3.3 Optional Requirements

Chapter 4

Design

4.1 Numerical Optimization

4.2 Modular approach

4.3 System Architecture

4.4 Searcher

4.4.1 Bayesian Optimization

4.5 Selector

4.5.1 Exact vs. Closest Match

4.6 Deployer

4.6.1 VM Provisioner

4.6.2 Docker Deployer

4.6.3 Ping server

4.6.4 Simulated Deployment

4.7 Interpreter

Chapter 5

Implementation

5.1 General usage

5.2 Driver

5.3 Searcher

5.3.1 Spearmint

5.4 Selector

5.4.1 Exact Match

5.5 Deployer

5.5.1 VM Provisioner

5.5.2 Terraform

5.5.3 Docker deployer

5.5.4 Cloudsuite

5.5.5 Ping servers

5.6 Interpreter

5.6.1 Sysbench

5.6.2 Cloudsuite

Chapter 6

Evaluation

6.1 Functionality

6.2 Results Analysis

6.2.1 Exhaustive search

6.2.2 Bayesian Optimization

Cross-provider

Concurrent Jobs

Chapter 7

Discussion

7.1 Functionality

7.2 Testing

7.3 Future work

Chapter 8

Critical Appraisal

Chapter 9

Conclusion

Bibliography

[1] G. Pallis, "Cloud Computing: The New Frontier of Internet Computing," *IEEE Internet Computing*, vol. 14, no. 5, pp. 70–73, sep 2010. [Online]. Available: http://ieeexplore.ieee.org/document/5562494/

[2] P. M. Mell and T. Grance, "The NIST definition of cloud computing," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep., 2011. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

[3] Intricately, "2019 Intricately Cloud Market Share Report," Intricately, Tech. Rep., 2019. [Online]. Available: https://content.intricately.com/hubfs/Cloud-Market-Share-Report.pdf

[4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," *EECS Department, University of California, Berkeley*, pp. 1–25, feb 2009. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

[5] B. Power and J. Weinman, "Revenue growth is the primary benefit of the cloud," *IEEE Cloud Computing*, vol. 5, no. 4, pp. 89–94, jul 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8436080/

[6] G. B. Berriman, E. Deelman, G. Juve, M. Rynge, and J. S. Vöckler, "The application of cloud computing to scientific workflows: A study of cost and performance," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1983, 2013. [Online]. Available: http://dx.doi.org/10.1098/rsta.2012.0066

[7] A. Ruiz-Alvarez and M. Humphrey, "An automated approach to cloud storage service selection," in *Proceedings of the 2nd international workshop on Scientific cloud computing - ScienceCloud '11*. San Jose, California, USA: ACM Press, 2011, p. 39. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1996109.1996117

[8] K. Bilal and A. Erbad, "Impact of multiple video representations in live streaming: A cost, bandwidth, and QoE analysis," in *Proceedings - 2017 IEEE International Conference on Cloud Engineering, IC2E 2017*, 2017, pp. 88–94. [Online]. Available: https://ieeexplore.ieee.org/ielx7/7922463/7923761/07923791.pdf?tp={&}arnumber=7923791{&}isnumber=7923761{&}ref=aHR0cHM6Ly9pZWVleHBsb3JlLmllZWUub3JnL2Fic3RyYWN0L2RvY3VtZW50Lzc5MjM3OTE=

[9] O. Alipourfard, H. H. Liu, J. Chen, S. Venkataraman, M. Yu, M. Zhang, Y. University, H. Harry Liu, J. Chen, S. Venkataraman, M. Yu, and M. Zhang, "CherryPick: Adaptively Unearthing the Best Cloud Configurations for Big Data Analytics," in *Proceedings of the 14th USENIX Conference on Networked Systems Design and Implementation*, 2017, pp. 469–482. [Online]. Available: https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/alipourfardhttps://dl.acm.org/citation.cfm?id=3154669

[10] A. Berl, E. Gelenbe, M. Di Girolamo, G. Giuliani, H. De Meer, M. Q. Dang, and K. Pentikousis, "Energy-Efficient Cloud Computing," *The Computer Journal*, vol. 53, no. 7, pp. 1045–1051, sep 2010. [Online]. Available: https://academic.oup.com/comjnl/article-lookup/doi/10.1093/comjnl/bxp080

Appendices

Testing Summary

User Manual