Location-aware Cloud Service Brokering in Multi-cloud Environment

Hui Ma[1], Sven Hartmann[2], Gang Chen[1], and Tao Shi[3]

[1] Victoria University of Wellington, Wellington, New Zealand
{hui.ma, aaron.chen}@ecs.vuw.ac.nz
[2] Clausthal University of Technology, Clausthal-Zellerfeld, Germany,
{sven.hartmann}@tu-clausthal.de
[3] Qingdao Agricultural University, Qingdao, China
{tao.shi}@ecs.vuw.ac.nz

Abstract. More and more enterprises are increasingly gaining technical and economic benefits from the global cloud marketplace. A multi-cloud environment makes it possible to orchestrate the access and management of multiple cloud resources at the global scale. Cloud service brokers can be regarded as the intermediate to discover, integrate, aggregate, customize, and optimize cloud services, e.g. Infrastructure as a Service (IaaS), from different cloud providers. However, the existing cloud brokering schemes rarely consider the optimization techniques of deploying cloud services with respect to various criteria, such as cost and performance. In this tutorial, we present some of our works using AI based approaches to the multi-cloud service brokering problem, i.e. selecting and leasing virtual machines (VMs) for cloud users with minimal cost and network latency. Experimental studies using real-world datasets show that our approaches with problem-tailored solution representation, efficient performance measurement, and smart solution initialization strategies can significantly outperform many existing approaches proposed in the literature.

General Information

Length of the Tutorial.

The tutorial is planned for a duration of 1 hour 45 minutes.

Level and Prerequisites.

The tutorial is intended to be on an introductory level assuming only some general knowledge in areas such as cloud and service computing and artificial intelligence.

Intended Audience.

The tutorial addresses service and cloud computing researchers, practitioners and students who are interested in AI approaches to resource allocation in multi-cloud for the deployment of software applications and services.

Importance of this proposed tutorial to the ICSOC audience

Many cloud providers provide resources. However, there are still not effective approaches to properly choose resources in cost efficient way. This tutorial provides the fundamentals of multi-cloud brokering, covering their computational development, technical capabilities, and the roles of services, cloud, and edge computing in a cyber-physical space from a social-technical perspective. The tutorial will provide clear evidence that cost-efficient cloud service brokering play an ever-increasingly essential and critical role in supporting software application and service deployment, interdisciplinary research in cloud computing, artificial intelligent, and information systems. This tutorial will also further investigate new best practices and directions.

About the Presenters.

Hui Ma (BEng, MInfSci, PhD) is associate professor in the School of Engineering and Computer Science at Victoria University of Wellington, New Zealand. Her major research interests are conceptual modeling, distributed databases, service composition, service deployment and resource allocation in clouds. She has published over 140 fully refereed articles in internationally recognized journals and conferences. She served as PC chair for several international conferences, including DEXA2016, APCCM 2017, ER2017, APCCM2018 and DEXA2018. She served at local co-chair of two international conferences (AI2018 and CEC2019) in the area of evolutionary computation and artificial intelligence.

Sven Hartmann (MSc, PhD, DSc) was full professor at Massey University, New Zealand. Since 2008 he is a full professor of computer science and chair for databases and information systems at Clausthal Technical University, Germany. There he is also serving as academic dean at the Faculty of Mathematics, Informatics and Mechanical Engineering. Sven has more than 150 publications, including venues like ACM TODS, ACM SIGMOD, VLDB, IEEE ICDE, and ER. He served as a PC member for more than 80 conferences, including 10 times as PC chair. His research interests include database systems, big data engineering, conceptual modeling, data-aware optimization and aeronautical informatics.

Gang Chen (MSc, PhD) is senior lecturer in the School of Engineering and Computer Science at Victoria University of Wellington. His research interests include evolutionary computation, reinforcement learning, multi-agent systems, and cloud and service computing. He has more than 130 publications, including leading journals and conferences in machine learning, evolutionary computation, and distributed computing areas, such as IEEE TPDS, IEEE TEVC, JAAMAS, ACM TAAS, IEEE ICWS, IEEE SCC. He is serving as the PC member of many prestigious conferences, including ICLR, ICML, NeurIPS, IJCAI, and AAAI, and co-chair for Australian AI 2018 and CEC.

Tao Shi (MSc, PhD) is currently working as assistant professor at Qingdao Agricultural University. He obtained his PhD in the School of Engineering and Computer Science, Victoria University of Wellington, New Zealand. His main

research interests include cloud computing and distributed system. His research focuses on resource management and combinatorial optimization in clouds.

Previous Presentations.

The presenters have given tutorials at various conferences such as ICDE2009, ER2009, APSCC2009, iiWAS2009, WISE2008, and invited talks to ADC 2022. Further, Hui has given an invited talk at the Women in Services Computing workshop at the IEEE World Congress on Services 2019, on Web Information System engineering and service computing.

Content Outline

The outline of our tutorial are as the following:

- **Multi-cloud brokering, current status, and research challenges.** (15 minutes)
  We will introduce the fundamental concepts of multi-cloud brokering, discuss variations of the problem that have been investigated in research and practice, and present related multi-cloud frameworks with illustrative examples.
- **Optimizing the location-aware composite application deployment for multi-cloud.** (25 minutes)
  We will introduce a new type of composite application deployment problem that jointly considers both the performance optimization and budget control in multi-cloud at the global scale. To tackle this, we discuss the development of evolutionary computation approaches such as a hybrid genetic algorithm (GA) techniques, featuring new design of domain-tailored service clustering, repair algorithm, solution representation, population initialization, and genetic operators.
- **Optimizing the application replication and deployment for multi-cloud.** (25 minutes)
  We will study the global-wide cloud application replication and deployment problem considering the application average response time, including particularly application execution time and network latency, subject to the budgetary control and present a genetic algorithm based approach with domain-tailored solution representation, fitness measurement, and population initialization.
- **Optimizing the elastic application deployment for multi-cloud.** (20 minutes)
  We will study the development of machine-learning methods such as deep reinforcement learning (DRL) techniques to dynamically select and lease virtual machines to process online requests at the global scale. We will study the usefulness of deep Q-networks to minimize the average network latency. We will also investigate the use of penalty-based reward function for effective budget control.

– **Conclusion and future directions** (10 minutes)
– **Q&A** (10 minutes)

Literature

The work presented in the tutorial is based on our journal publications [2, 4, 6, 13, 15] on the subject, and various conference publications [1, 3, 5, 7–12, 14].

References

1. Chen, Y., Shi, T., Ma, H., and Chen, G. Automatically design heuristics for multi-objective location-aware service brokering in multi-cloud. In *2022 IEEE International Conference on Services Computing (SCC)* (2022), pp. 206–214.
2. Poaka, V., Hartmann, S., Bochinski, M., and Seggelke, N. New architectural design of the runtime server for remote vehicle communication services. *SAE Intl. J. Connected and Automated Vehicles 3* (2020), 19–26.
3. Shi, T., Chen, G., and Ma, H. Multi-objective container consolidation in cloud data centers. In *The Australasian Joint Conference on Artificial Intelligence (AJCAI)* (2018), pp. 783–795.
4. Tan, B., Ma, H., Mei, Y., and Zhang, M. Evolutionary multi-objective optimization for web service location allocation problem. *IEEE Transactions on Services Computing (TSC) 14*, 2 (2021), 458–471.
5. Ma, H., Schewe, K., Thalheim, B., and Wang, Q. Cloud warehousing. *J. Universal Comp. Sci. 17* (2012), 1183–1201.
6. Ma, H., Schewe, K., Thalheim, B., and Wang, Q. A formal model for the interoperability of service clouds. *Service Oriented Computing and Applicationn (SOCA) 6* (2012), 1–17.
7. Ma, H., Schewe, K., and Wang, Q. An abstract model for service provision, search and composition. In *IEEE APCSC* (2009), pp. 95–102.
8. T. Shi, H. Ma, and Chen, G. Energy-aware container consolidation based on PSO in cloud data centers. In *IEEE Congress on Evolutionary Computation (CEC)* (2018), pp. 1–8.
9. T. Shi, H. Ma, and Chen, G. A genetic-based approach to location-aware cloud service brokering in multi-cloud environment. In *IEEE International Conference on Services Computing (SCC)* (2019), pp. 146–153.
10. T. Shi, H. Ma, and Chen, G. A seeding-based GA for location-aware workflow deployment in multi-cloud environment. In *IEEE Congress on Evolutionary Computation (CEC)* (2019), pp. 3364–3371.
11. T. Shi, H. Ma, and Chen, G. Seeding-based multi-objective evolutionary algorithms for multi-cloud composite applications deployment. In *IEEE International Conference on Services Computing (SCC)* (2020), pp. 240–247.
12. T. Shi, H. Ma, Chen, G., and S. Hartmann. Location-aware and budget-constrained application replication and deployment in multi-cloud environment. In *IEEE IEEE International Conference on Web Services (ICWS)* (2020), pp. 110–117.
13. T. Shi, H. Ma, Chen, G., and S. Hartmann. Location-aware and budget-constrained service deployment for composite applications in multi-cloud environment. *IEEE Transactions on Parallel and Distributed Systems (TPDS) 31* (2020), 1954–1969.

14. T. Shi, H. Ma, Chen, G., and S. Hartmann. Location-aware and budget-constrained service brokering in multi-cloud via deep reinforcement learning. In *Service-Oriented Computing (ICSOC)* (2021), p. 756–764.
15. T. Shi, H. Ma, Chen, G., and S. Hartmann. Cost-effective web application replication and deployment in multi-cloud environment. *IEEE Transactions on Parallel and Distributed Systems (TPDS) 33*, 8 (2022), 1982–1995.