QoS Assurance for Service Function Placement in Edge Computing: State of The Art, Issues and Challenges

Briytone Mutichiro, Younghan Kim*
*Soongsil University

briyt.mutichiro@dcn.ssu.ac.kr, *younghak@ssu.ac.kr

Abstract

The deployment of IoT/cloud-enabled smart services still suffers from several quality of services (QoS) management issues such as latency sensitivity for interactive applications and real-time big data analytics. In this paper, we propose to present a novel approach for the management of QoS in edge service placement. Accordingly, we introduce an overview of QoS requirements in fog computing environment for several IoT/cloud-enabled use cases, to identify QoS methods and models. The synthesis of the requirements review allows us to present a solution service placement based on QoS assurance in edge

Ⅰ. Introduction

Edge computing promises to inspire the development of a range of applications and services characterized by ultra-low latency and high QoS due to its dense geographical distribution and wide support for mobility, location awareness and real-time interactions [1]. Generally, the internet has always operated on a best-effort basis from the perspective of QoS, with the responsibility of guaranteeing QoS falling on the infrastructure provider. The requirements and deployment mechanisms differ across the QoS spectrum for both cloud and edge computing as edge needs the incorporation of new components to provide the support platform for emerging applications and vertical services. In [2] they define QoS as the satisfaction that is attained by a service in comparison to others, allowing consumers to determine the best service that best suits their requirements. Due to the heterogeneity and complexity of edge, there are no uniform QoS standards and protocols for edge computing [3] from any standardization body, and infrastructure providers ought to define QoS in a generic fashion, that creates an agnostic system that may guarantee QoS to user requests and applications.

Multiple factors impact performance and application QoS during runtime in the edge cluster. Varying workloads and dynamic network conditions affect performance [4], resulting in application service downtime, unavailability and unsatisfactory QoS. Resource management is made difficult by user requirements complexity in terms of request-resource matching [5], service interdependencies and service competition for both sharable and non-sharable resources. Optimal scheduling mechanisms in edge clusters that [6] address over/under resource allocation to ensure higher values in QoS attributes like scalability and availability. Finally, function selection for each service request, is aimed at optimizing the overall QoS criteria of the resultant composite service. This requires service QoS monitoring and can be facilitated through strategies such as QoS prediction and QoS estimation [5] of various QoS attributes and contextual factors. Ksentini et. al [6] conduct a qualitative study on QoS requirements of emerging services provided by fog computing systems, which highlights QoS requirements for each IoT/cloud enabled use case. The focus of this study slightly differs from our work as our focus is on how to model the edge system for QoS assurance. Our work is different in the content and research issues.

The rest of this work is organized follows: In section II we present a brief on QoS modelling, followed by issues in section III. We conclude with a discussion on issues and challenges in sections IV and V.

Ⅱ. QoS Modelling in Edge Computing

We reviewed different literature with a focus on constraints adopted in the formulation of service placement problems. We observed that most studies considered network related constraints specifically latency, bandwidth, reliability, and availability. We present in Table. 1, a summary of our findings. The most frequent QoS constraints are latency and node capacity.

III. Issues and Research Challenges

Optimal resource scheduling among edge nodes requires a scheduling mechanism that can guarantee responses [6] upon the uncertainty of the runtime environment.

Table 1. Summary

QoS Constraints						
Latency	Bandwidth	Availability	Reliability	Workload	Capacity	Service
[4][6][8] [9][10] [11][12]	[4][8] [9][10] [12]	[10] [11] [12]	[8] [11]	[6] [10] [11]	[6][8] [9][10] [11][12]	[9] [10]

Supporting the QoS constraints requires the pre-knowledge of execution environment such as underlying infrastructure and the configuration of application components on the network [5]. An adaptation technique is fully advantageous, if it does not require relying on the knowledge provided by previous experiences.

Reliability in an application even during a fault e.g. losing control over edge nodes or undetermined latency needs to be high. This requires elements of decentralized control or off-line detection and recovery.

The role of data and application scenarios has been mostly overlooked. Specifically, to improve the performance of a machine learning algorithms and the need to learn more parameters and the training data. This demonstrates the importance of data on the development of AI.

IV. Conclusion

In this paper, we focused on QoS assurance and edge platforms. A qualitative study on QoS modelling was presented, highlighting various works and current state of the art based on constraints and methods. We concluded with a brief review of current issues and challenges that future research work should address.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-00946,Development of Fast and Automatic Service recovery and Transition software in Hybrid Cloud Environment)

References

[1] H. Yang and K. Younghan, "Design and Implementation of High-Availability Architecture for IoT-Cloud Services," MDPI Sensors (Basel, Switzerland), vol. 19, no. doi:10.3390/s19153276, 2019.

[2] M. Ahmed, L. Liu, J. Hardy and B. Yuan, "An Efficient Algorithm for Partially Matched Web Services Based on Consumer's QoS Requirements," in IEEE/ACM 7th International Conference on Utility and Cloud Computing, London, 2014, pp. 859-864, doi: 10.1109/UCC.2014.140.

[3] G. Pavlou and I Psaras, " The troubled journey of QoS: From ATM to content networking, edge computing and distributed internet governance, " Computer Communications, vol. 13, no. ISSN 0140-3664, https://doi.org/10.1016/j.comcom.2018.07.006, pp. 8-12, 2018.

[4] A. Brogi, S. Forti and F. Paganelli, "Probabilistic QoS-aware Placement of VNF chains at the Edge," CoRR, vol. abs/1906.00197, https://dblp.org/rec/journals/corr/abs-1906-00197, 2019

[5] Shangguang Wang, Yali Zhao, Lin Huang, Jinliang Xu, Ching-Hsien Hsu. QoS prediction for service recommendations in mobile edge computing. Journal of Parallel and Distributed Computing. Volume 127. 2019. Pages 134-144. doi.org/10.1016/j.jpdc.2017.09.014.

[6] H. A. Alameddine, S. Sharafeddine, S. Sebbah, S. Ayoubi and C. Assi, "Dynamic Task Offloading and Scheduling for Low-Latency IoT Services in Multi-Access Edge Computing," IEEE Journal on Selected Areas in Communications, vol. 37, no. 3, p. 668- 682, Mar. 2019.

[7] A. Ksentini, M. Jebalia and S. Tabbane, "IoT/cloud-enabled smart services: A review on QoS requirements in fog environment and a proposed approach based on priority classification technique," International Journal of Communication Systems, vol. 34, 2019.

[8] R. Xavier, L. Z. Granville, B. Volckaert and F. De Turck, "Elastic Resource Allocation Algorithms for Collaboration Applications," J. Netw. Syst. Manag., vol. 25, no. no. 4, pp. 699-734, 2017.

[9] A. Leivadeas, G. Kesidis, M. Ibnkahla and I. Lambadaris, "VNF placement optimization at the edge and cloud," Futur. Internet, vol. 11, pp. 1-23, 2019.

[10] C. Yan-Ting, "Mobility-aware service function chaining in 5G wireless networks with mobile edge computing," in IEEE International Conference on Communications, 2019.

[11] Z. Zhong and R. Buyya, "A cost-efficient container orchestration strategy in kubernetes-based cloud computing infrastructures with heterogeneous resources," ACM Transactions on Internet Technology (TOIT) , vol. 20 (2), pp. 1-24, 2020.

[12] N. Kherraf, S. Sharafeddine, C. M. Assi and A. Ghrayeb, "Latency and Reliability-Aware Workload Assignment in IoT Networks With Mobile Edge Clouds," IEEE Transactions on Network and Service Management, Vols. 16, no. 4, no. doi: 10.1109/TNSM.2019.2946467, pp. 1435-1449, 2019.