July 2017

Edge-as-a-Service: Towards Distributed Cloud Architectures

Blesson VARGHESE [a,1], Nan WANG [a], Jianyu LI [a] and Dimitrios S. NIKOLOPOULOUS [a]

[a] *School of Electronics, Electrical Engineering and Computer Science Queen's University Belfast, UK*

Abstract. We present an Edge-as-a-Service (EaaS) platform for realising distributed cloud architectures and integrating the edge of the network in the computing ecosystem. The EaaS platform is underpinned by (i) a lightweight discovery protocol that identifies edge nodes and make them publicly accessible, (ii) a scalable resource provisioning mechanism for offloading a number of workloads from the cloud on to the edge for servicing multiple user requests, and (iii) a dynamic auto-scaling mechanism for efficiently managing resources allocated to a workload. We validate the feasibility of EaaS on an online game use-case to highlight the improvement in the QoS of the application hosted on the cloud-edge by demonstrating low overheads in our platform, reducing data traffic to the cloud and application latency, and minimising Service Level Objective (SLO) violation rates.

Keywords. edge computing, distributed cloud, resource discovery, resource management, resource scaling

1. Introduction

The upcoming Internet-of-Things paradigm has motivated the need for designing and developing distributed cloud architectures for meeting future computing challenges. Up to 50 billion devices are forecast to be connected to the Internet by 2020 generating 50 trillion gigabytes of data[2]. Centralised cloud architectures will become untenable because the frequency and latency of communication between user devices (for example, smartphones and tablets) and geographically distant cloud servers will significantly increase beyond that which can be handled by existing communication and computing infrastructure [1]. The Quality-of-Service (QoS) will therefore be adversely affected [3].

Distributed cloud architectures not only make use of computing resources concentrated in cloud data centers, but also tap into resources that may be located elsewhere in the network [4]. One viable location is the edge of the network so that data generated on user devices are processed closer to the source to reduce network traffic. This may be possible by using spare computing resources that may be available on local routers, mobile base stations and switches or on dedicated nodes comprising low cost and low

[1] Corresponding Author. E-mail: varghese@qub.ac.uk; Web: www.blessonv.com

[2] `http://www.gartner.com/newsroom/id/3165317`