Latency Analysis of Cloud Infrastructure for Time-Critical IoT Use Cases

Kartinah Zen
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Sarawak, Malaysia
kartinah@unimas.my

Saju Mohanan
Department of Information Technology
University of Technology and Applied Sciences
Muscat, Oman
saju.mohanan@hct.edu.om

Seleviawati Tarmizi
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Sarawak, Malaysia
swati@unimas.my

Noralifah Annuar
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Sarawak, Malaysia
anoralifah@unimas.my

Najm Us Sama
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Sarawak, Malaysia
najmussama@gmail.com

2022 Applied Informatics International Conference (AiIC) | 978-1-7281-8428-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/AiIC54368.2022.9914601

***Abstract*— The time-critical Internet of Things (IoT) use cases such as driverless cars and robotic surgical arms need high bandwidth and low latency for real-time intelligent data processing and trained machine learning inference. Latency in real-time processing is influenced by many factors such as artificial intelligence (AI) computing algorithm, device processing capabilities, the frameworks, and also the distance from the cloud infrastructure. However, the geographical distance between the data origin and data processing is one of the major factors contributing to the network latency for time-critical IoT use cases. In this paper, we analyzed the latency from a particular client point based on the live data generated by their cloud data centers. The experiments were done through the big three cloud vendors, which are Microsoft Azure, Amazon Web Services (AWS), and Google Cloud Platform (GCP). As a result, a time-critical IoT low latency approach is proposed in this paper.**

Keywords*— *Latency, IoT, Edge Computing, Cloud Computing, RTT

I. Introduction

The Internet of Things (IoT) involves many aspects of life and consists of intelligent machines that are interacting with other machines, things, the environment, and infrastructure. The IoT refers to the connected digital and physical things that are embedded with sensors, actuators, and other technologies to do business transactions and human needs and to make things easier. The IoT is changing the way of life in various fields, economy and technological aspects. IoT will give a great impact on the economy and society by transforming many business transactions into digital transactions. According to Statista Research department's projection [1], it is predicted that the total number of connected devices in the world by 2025 will be approximately 75.44 billion.

The data volume is a problem in IoT-based cloud computing. By 2025, the total data volume of connected IoT devices worldwide is expected to reach 79.4 Zettabytes [2]. However, sending such data to the remote cloud data center may lead to a heavy burden on the backbone network, which results in undesired latency. This is true especially when the IoT data involves the machine and deep learning, which involves processing massive amounts of data in real-time. As an example, Google Cloud Speech is powered by a machine learning framework, TensorFlow. Deep learning is a promising method along with machine learning approaches, to solve many problems in IoT with large datasets. Since the model training is resource-based, it is normally required to send the data to the remote cloud with sufficient resources to be processed. Data processing at the cloud is one of the existing approaches. The main drawback of this approach is its high latency and unreliability. Even though the cloud is not a mandatory platform to run IoT applications, the cloud can provide a hybrid IT infrastructure for IoT device management, monitoring, data analytics, and visualization.

In many research, edge computing architecture has been introduced to solve this issue. With the IoT network's technological growth and advance, most data are available at the network edge [3]. Edge Computing as a new computing paradigm, triggers to move data from the remote cloud data center to the local edge of the network and end devices, enabling local processing of the user data. The machine learning method is usually adopted and used to extract relevant useful information from the generated data in the IoT edge network [4].

This paper, analyses the round trip time (RTT) of major vendor cloud computing from one point of data origin to different locations. The results show that the high latency for certain locations limits the cloud computing capabilities to process time-critical IoT use cases data. Therefore, as stated by most recommendations, it is best to allocate the location of their cloud server close to the IoT end devices to reduce the latency for the time-critical use cases. Furthermore, we suggest either fog or edge architecture to be adopted. Both computing provides the same functionality for the data analytic

978-1-7281-8428-9/22/$31.00 ©2022 IEEE