Orchestrating Quality of Service in the Cloud of Things Ecosystem

Prem Prakash Jayaraman, Karan Mitra, Saguna Saguna, Tejal Shah, Dimitrios Georgakopoulos, Rajiv Ranjan

Abstract— Cloud of Things (CoT) is a vision inspired by Internet of Things (IoT) and cloud computing where the IoT devices are connected to the clouds via the Internet for data storage, processing, analytics and visualization. CoT ecosystem will encompass heterogeneous clouds, networks and devices to provide seamless service delivery, for example, in smart cities. To enable efficient service delivery, there is a need to guarantee a certain level of quality of service from both cloud and networks perspective. This paper discusses the Cloud of Things, cloud computing, networks and new quality of service management research issues arising due to realisation of CoT ecosystem vision.

Keywords— Internet of Things, Cloud computing, Cloud of Things, Sensing-as-a-Service, Quality of Service, Network

I. INTRODUCTION

Sensing-as-a-service is a recent phenomenon driven by the tremendous growth in the areas of Internet of Things (IoT) and cloud computing providing data-, analytics- and visualization-as-a-Service. The term Internet of Things (IoT) collectively describes technologies and research disciplines that enable the Internet to reach out into the world of physical objects. Technologies like RFID, short-range wireless communications, real-time localization and sensor networks have become increasingly pervasive thus, making the IoT a reality. According to the recent Gartner report [1], it is estimated that IoT will grow to 26 billion units by 2020, excluding PCs, tablets and smartphones (40 billion things including tablets and smartphones). The revenue as a result of this growth is estimated to be approximately $1.9 trillion. The IoT will fuel a paradigm shift of a “truly connected world” in which everyday objects become inter-connected and smart with the ability to communicate many different types of information with one another. Cloud computing on the other hand allows IT-related resources (e.g., CPU, applications, network and storage) to be provided as virtualized services to the customers under a usage-based payment model. Using cloud computing, customers (e.g., SMEs, governments, universities) can leverage these virtualized services on-the-fly; ensuring that they do not have worry about the infrastructure details such as, where these resources are hosted or how they are managed.

Cloud computing and Internet of Things (IoT) are two technologies that are already becoming part of our daily lives and are attracting significant interest from both industry and academia. The Cloud of Things (CoT) is a vision inspired from the IoT paradigm where everyday devices namely “smart objects” are fully connected to the network and integrated with the cloud(s) for data storage, processing, analytics and visualization. There can be several applications (possibly as-a-service model) that can leverage CoT. For example, smart healthcare, smart homes, smart manufacturing, assisted living, digital agriculture and logistics, to name a few.

We expect that the CoT paradigm will soon give rise to the development of innovative and novel applications that are composed from IoT devices operated by multiple providers (for clouds, sensors, and data) and are hosted on public/private cloud environments. Consider a disaster management application that detects and alerts the citizens regarding floods. This application may detect the level of flooding using data obtained from numerous IoT devices such as water flow sensors placed on the roads, weather, IP cameras owned by the municipalities and social media (e.g., Twitter and Facebook). In a typical CoT ecosystem, these sensors will be owned and operated by multiple providers such as councils, government bodies, private organisations and individuals. The disaster management application needs to deliver a certain level of quality of service (QoS) composed of both functional and non-functional components. The functional components may include the cost to obtain data from multiple providers, the rate at which data is produced, the types of sensors, the analytics required for detection and altering of events, application response time (based on network, processing delay), and the overall functional infrastructure to host the service(s). The non-functional components may include the accuracy of event detection, timing/deadlines such as the maximum time within which an event is detected and responders are dispatched, cost of event detection, data quality and sensor calibration issues.

CoT is a relatively new concept and poses significant challenges; and raises a number of questions from both technological and societal perspectives. For example, how can we build the CoT infrastructure? How can we provide efficient citizen-centric services via CoT? How can we justify the cost of building such infrastructure? To build a CoT infrastructure, one the crucial challenge is ensuring applications’ QoS. For example, to ensure novel and disruptive services in a smart city environment such as augmented reality and disaster management, there is a need to provision QoS in the form of low response time and high throughput. While these citizen-centric services require stringent response times in the order of tens of milliseconds, current generation of cloud resource management techniques can only guarantee low level QoS metrics such as ensure high resource availability (e.g. guarantee 99.99% CPU availability).

In this context, one of the major research question is how would the future CoT application represent and ensure quality of service (QoS)? We expect that the future CoT applications will soon be challenged with QoS requirements. To illustrate this let us consider Remote Healthcare CoT Application (e.g. the MOVE Lab at Newcastle University) for aged patients.
This application uses body sensors, RFID readers and accelerometers to monitor events such as: (i) changes in heart rate and sugar levels; (ii) administration of medications on time; and (iii) detecting incidents such as falling. In the future, users of such systems say paying $100 per month for the service will expect 1) events such as heart attack, falls are detected (from IoT devices) within x milliseconds, and (2) alerts are automatically sent to the doctors, caregivers, and emergency ambulance teams within y minutes of event detection.

As mentioned previously, the CoT will be the efficient combination of IoT and cloud computing. Therefore, the applications and services will be hosted on the clouds and will be heavily dependent on the service level agreements (SLAs) provided by cloud providers. Current approaches to define QoS and the corresponding SLAs for CoT are in infancy and are not addressed in the state-of-the-art research. Given the inherent complexities of CoT introduced by various dimensions of IoT namely, the physical layer (devices), infrastructure layer (networking, processing, storage) and analytics layers (analytic algorithms to draw insights from data), defining QoS for IoT is a hard problem.

This paper takes the initial steps in presenting our vision of a future CoT ecosystem that is able to satisfy the multi-dimensional QoS constraints imposed by IoT and clouds. Our aim in this paper is to propose the design requirements of a conceptual CoT system that is able to: (i) consume data from multiple IoT data providers; (ii) invoke analytics services from third-party providers; and (ii) deploy and deliver the composed applications via the cloud(s) to the end-user while satisfying the multi-dimensional QoS constrains imposed by networks and clouds. Our vision of a QoS driven CoT ecosystem will have multiple SLAs’s to suit the needs of the end-user. For example, the accuracy of the event detection could be 90% for some user applications while other user application would be happy with 80% accuracy. This constraint could be directly related to the cost for the service. Our vision aims to closely follow the most popular pay for usage cloud model.

II. CLOUD OF THINGS

Figure 1 depicts a conceptual overview of the Cloud of Things ecosystem. The physical layer comprises things such as smart sensing devices, human sensors, connected cars, and smartphones. In the physical layer, the sensors and actuators (real and virtual) allow the capture and change of certain aspects of the physical world. The cloud layer is responsible for modelling and representation of the physical entities as virtual entities. The cloud layer also has the application layer that is composed of IoT services, application and business processes that make use of the virtual entities and their virtual representation to control/monitor/detect state changes in the physical world. Consider the query “Provide the indoor temperature in Room 1.23” or “Set light level in Room 2.57 to 15”. To support this query, the interactions and associations between the physical layer, the virtual entity and the IoT application needs to be modelled. For example, the association will contain the information that Sensor 456 provides the indoor temperature of Room 1.23 and the measurement unit is Celsius. In our vision of CoT, the virtual entities has functions to interact with the IoT device subsystem as well as functionality for discovering and looking up IoT devices that can provide information about virtual entities. Furthermore, it contains all the functionality needed for managing associations, as well as dynamically finding new associations and monitoring their validity, e.g., due to the mobility of IoT devices that can impact the state of the virtual entities.

III. CLOUD OF THINGS: APPLICATION USE CASES

CoT can be used in variety of applications within the smart city concept such as smart remote healthcare, smart homes and home automation, smart buildings. In this paper, we describe three scenarios to motivate and illustrate the need for CoT in the real world.

1. Healthcare-as-a-service: The healthcare sector has witnessed a tremendous growth in terms of using a wide variety of sensors to monitor the health of patients both at home and at the hospitals. There are a number of sensors and devices that can be used to sense, collect, store and process data related to different physical and environmental aspects of a patient and provide different healthcare services. The large-scale data produced from these services requires efficient storage, processing and timely response (in near real time) using the Internet and cloud computing. The main goal of Healthcare-as-a-Service built using CoT will be to provide seamless monitoring of the patients’ health at the hospital or at their homes in order to prevent an emergency and provide timely care. Such a system will include a large number of sensors and devices on the patients or in their environment (homes and offices). These will be connected to a gateway for transmitting the data to a storage facility on the cloud. There could also be additional nodes between the sensor and the gateway for pre-processing the sensor data and then forwarding it to the gateway. Further, there will be an application running on the cloud to detect the behavior of the
patients and alert the healthcare professionals on the patient’s situation in real-time.

Based on scenario 1, we present a detailed use-case for remotely monitoring the elderly at home to provide timely healthcare support. Marie is an elderly woman who lives alone at home. She has a number of health problems but prefers to live at her home for as long as she can. A healthcare system should be able to detect a number of situations for example, fall detection or if she suffers from a heart attack, if her behavior is abnormal on in the day, if she sleeps too long, if she leaves the door open at night or she goes out of her house at night. In this case, there are a number of sensors to monitor Marie’s vital physiological parameters such as body temperature, heart rate, breathing rate, blood oxygen level and blood pressure. Also, Marie wears an accelerometer sensor for measuring her body motion activities like sitting, standing, walking, sleeping and a location badge to detect her location both inside and outside her home. Marie also wears a microphone and an emergency bracelet to communicate with healthcare professionals in an emergency situation for example, if she slips and falls. In addition, Marie’s smart home is equipped with sensors to detect opening and closing of doors and windows, room temperature, video camera and location sensors which work in conjunction with her location badge. All these sensing devices are connected to the gateway either directly or via additional actuators that process and transmit the data to the gateway. The gateway then sends this information to the cloud via the available network. In some cases, the same devices may be directly connected to the Internet as services. Data from all sensors is monitored and the application running on the cloud alerts the healthcare professionals if any situation is detected where the patient needs immediate attention.

2. Smart homes and home automation: Smart home applications are the highest-ranking IoT applications as more companies are actively involved in building smart home applications as compared to any other application in the field of IoT. Applications can range from providing basic home automation like such as intelligently operating and controlling heating, ventilation, doors, windows, lights, home appliances, watering systems, etc to home security related applications where sensors can detect if someone has entered the home when the owner is away and sound an alert. These applications are complex and need to process different types of sensor data in order to assist homeowners in their daily life. The sensors and devices send information via a home gateway to the cloud or process such information on the local cloud at the user’s home. Smart homes equipped with sensors can also detect the behavior of family members and create an environment that is in synchronization with their activities and mood. For example, switching on or dimming the lights, opening or closing curtains, informing about when a washer has finished its cycle and when the oven is to be switched on or off.

Based on scenario 2, we consider a use case where John and Olivia are a couple living with their children in a smart home in Sweden. The different sensors within their home help them to control and efficiently manage the different objects, devices and appliances they interact with. Such a smart home also helps them to better synchronize their lives for example when they have to cook a meal and perform other chores in the house like washing the laundry. They are also able to remotely monitor different aspects of their home. If Alice is out grocery shopping her smart home application can switch on and preheat the oven at home just in time for her to reach home and start cooking her fish. In the meantime, John can be notified just in time to bring the children back home from their play date when the fish is cooked and the oven is switched off. Further, the heating and ventilation in the bedrooms is accordingly adjusted to prepare the rooms for the family to sleep.

Further based on scenario 2, we consider another use case for energy efficient homes. The same sensors used to facilitate a couple live their lives seamlessly can also be used to make their home energy efficient. For example, when the family is not at home the heating and ventilation turns minimum levels and is turned back to ambient levels when they are back at home. Similarly, the different lights and appliances are switched on and off based on the presence of the family members. John and Olivia can be notified of the best and most energy efficient cycles for different appliances such as dishwasher, washer and dryer. Their smart home application can also turn on the appliances when the electricity prices are low for example at night.

These scenarios highlight the need of CoT where applications process the large volumes of data produced from the many IoT sensors and devices. These CoT applications need to communicate efficiently with the many sensors in order to perform and deliver timely results.

IV. QoS RESEARCH CHALLENGES AND FUTURE DIRECTIONS

The current mechanisms for QoS provisioning and supporting service level agreements [3][4][5][6] in IoT and clouds have major limitations. In the light of CoT, these mechanisms will have to be radically reconsidered or even reinvented to meet the challenges posed by upcoming CoT applications. Figure 1 captures the complexities of CoT applications from the physical device (data collection) to virtual layer (storage and processing) to application layer (delivery). However, QoS guarantee for the CoT is expectedly challenging, and an emerging discipline. This is due to the shortage of a standardized, end-to-end approaches for QoS assurance (between the end user, IoT devices and the cloud), the complexity of the integration of different layers (see figure 1 and 2), and the presence of a plethora of QoS constraints and parameters at each layer. We expect that the traditional way of QoS assurance will not be sufficient. For instance, we will soon be looking at satisfying requirement such as “detect/notify events within 5 minutes of occurrence” rather than the traditional model to “guarantee 99.99% CPU availability”.

Multidimensional QoS metrics: Several efforts specific to wireless sensor networks (WSN) have appeared in the literature [4][6][8][9]. The focus of these efforts is to understand and model the network communication characteristics of a WSN to define the corresponding network level QoS. However, from the perspective of CoT applications, network communication is an important but a smaller component. Parameters related to clouds (e.g., number of I/O operations and CPU throughput), devices (battery), network type, and application would have to be used in conjunction with network parameters (bandwidth, delay, and jitter) to determine the overall QoS of the CoT.
A. Multidimensional and a complex problem that needs to be observed from the equation, from end-to-end QoS is a cost. This includes (but not limited to): sampling rate, timeliness of the IoT data used for decision support and the physical layer is related directly to the quality and the messages to the end-user within acceptable limits. The QoS of the entire ecosystem needs to support delivery of emergency such emergency messages will need to be identified and the applicability of traffic demands. For example, emergency message from a requires adaptive QoS constraints that can deal with all kinds of traffic demands. To realise the larger vision of CoT, in particular the sensing-as-a-service paradigm, it is important to understand and carefully develop QoS metrics and corresponding SLA’s that take into consideration the complexity introduced by each layer (as depicted in figure 2). We define CoT QoS metric as:

\[ QoS = f(cloud\ QoS, network\ QoS, location, battery, ..., N) \] (1)

where \( N \) represents the total number of parameters. As can be observed from the equation, from end-to-end QoS is a multidimensional and a complex problem that needs to be carefully addressed.

A. Cloud of Things: IoT Physical Layer

**QoS semantics:** The IoT physical layer (see figure 2) comprises heterogeneous sets of devices ranging from wireless sensor networks, body area networks, and other virtual sensors (e.g., weather and flood services). With technological advancements, this space is going to get further crowded with billions of devices each with different capabilities and QoS, parameters and constraints. The heterogeneity among the IoT requires adaptive QoS constraints that can deal with all kinds of traffic demands. For example, emergency message from a sensor is more important than a regular monitoring message. Such emergency messages will need to be identified and the entire ecosystem needs to support delivery of emergency messages to the end-user within acceptable limits. The QoS of the physical layer is related directly to the quality and the timeliness of the IoT data used for decision support and automation. This includes (but not limited to): sampling rate, transmission speed, data quality, coverage, mobility, location and cost.

The sampling rate determines the rate at which a sensor measures an observed phenomenon (e.g., 5 Hz). Different applications require different sampling rate based on their criticality. The transmission speed is network dependent and refers to the actual rate at which data is sent to the cloud layer from the physical layer. This is influenced by network topology and the device connectivity (e.g. Bluetooth, WiFi and 4G).

Data quality is a complex metric [11] that needs to be carefully considered for future adoption and sustainability of CoT applications. However, similar to variation in quality of real-world things (e.g., quality of cotton), the need for high quality data satisfied by metrics such as high accuracy (recently calibrated) and minimal error is very application dependent.

The coverage at the device layers identifies the extent to which the sensor data covers the observer phenomenon geographically.

**Mobility** is another key QoS metric as the devices at the physical layer are inherently mobile. Mobility enables greater coverage but also introduces further challenges that impact other metrics such as location, quality and transmission speed. Finally, cost is a summative metric that relies on the previously identified metrics. For example, the cost of IoT data that is 99% accurate could be much higher when compared to data that is 70% accurate. It is clear from the above discussion that QoS will enable CoT systems to cater to the needs of different applications hence reaching a wider audience. Moreover, our vision of CoT of the future is an open world where multiple devices and analytics applications (owned and operated by independent providers) can be fused together to create innovative CoT applications/services.

Some of these metrics identified may be trivial for a single device per se. However, the notion of CoT is for multiple devices to collaborate and achieve a goal that is otherwise impossible to do it alone. Consider the example, a temperature sensor measuring temperature in Fahrenheit. When this information is shared with the power company that accepts data in Celsius, this leads to incorrect data. The question here is then, who should agree on the formats? Is it the job of the human or can the devices auto-negotiate the semantics? There is need to develop novel methods to map and represent the QoS into a language that can be used by the whole CoT ecosystem.

B. Cloud of Things: Cloud Layer

**Distributed QoS-aware CoT data processing:** In a CoT paradigm, the devices will be connected to the clouds for data storage, processing, analytics and visualization (see figure 1 and 2). Cloud computing offers virtually infinite set of resources (CPU, memory, storage and I/O) at reasonable costs based on a multi-tenant and pay-as-you-model. Clouds enable applications to be hosted in a highly distributed manner across multiple datacenters around the globe. At the same time, clouds ensure that the applications are highly available and scalable. These characteristics have made clouds to be widely adopted by private and public sector industries for reliable and efficient service delivery.

With the emergence of IoT, it is expected that 40 yottabytes of data will be generated by the end of 2020, [12]. Therefore, requiring large-scale data (big data) processing at multiple cloud datacenters. This is a challenging problem as most of the cloud data processing and analytics tools such as MapReduce are optimized to run in a single datacenter [12]. Further, there exist no comprehensive mechanisms where large datasets spanning multiple datacenters can be efficiently processed. This can mainly be attributed to the bottlenecks in network communication medium, and cloud and network infrastructure that can be congested and/or may not provide sufficient throughput to transfer large datasets in a timely manner for QoS-aware data processing and analytics.
Therefore, there is a need to address the challenge of distributed QoS-aware data processing in CoT ecosystems.

**Cross-layer cloud and network QoS monitoring:** A typical cloud system comprises three layers: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) [13]. Each layer has a specific set of metrics. For example, for SaaS: application throughput, application and CPU utilization; for PaaS: system up time, number of requests served; and for IaaS: Memory and CPU utilization. Cross-layer cloud QoS monitoring is a challenging task [15]. This is based on the assertion that there are no standard formats and APIs for cloud QoS monitoring. Currently, each cloud provider such as Amazon EC2, Google Compute Engine, and Microsoft Azure provide their own set of APIs. On top of that only Amazon provides a limited set of QoS metrics for cloud monitoring. Companies such as CloudHarmony [14] provide cloud monitoring information for multiple cloud providers but do not provide metrics for comparison and analysis of cloud providers. Further, most of research in this domain does not aim to collect a large pool of data for numerous metrics belonging to each cloud layer. Alhamazani et al. [15] has recently tried to address this challenge and built a service for cross-layer cloud QoS data collection. However, they do not provide methods for QoS analysis of various clouds and recommending methods for application hosting.

End-to-End QoS also involves network QoS in the form of propagation delay or network latency and throughput. One of the bottlenecks to ensure high network QoS is the network latency. In that, if an application requests some data from the CoT application hosted on the cloud, this data may involve some time to arrive back to the application. This is due to the distance (in hops) between the application and the cloud datacenter (assuming no cloud processing delay). For instance, the latency will be shorter (in tens of milliseconds) if the application running on device present in Europe (connected via Ethernet) requests data from European datacenter instead of South-Asian datacenter where the latency can be in the order of hundreds of milliseconds. These factors necessitate large-scale cross-layer cloud QoS monitoring and network QoS monitoring for QoS-aware CoT ecosystem.

Guaranteeing performance SLAs for future CoT applications requires clear understanding of important performance metrics across cloud-hosted big data processing frameworks (e.g., Apache Kafka, Apache Storm, Apache Hadoop, etc.) and hardware resources (CPU, Storage, and Network). The problem is further complicated due to the fact that the QoS metrics at CoT device layer, CoT application layer, cloud-based big data processing frameworks layer and cloud-based hardware resource layer are not necessarily the same [18][19]. For example, the key QoS metrics are (i) network jitter and delay at the IoT sensor layer; ii) event detection delay and decision making delay at CoT application level; iii) throughput and latency in distributed messaging queuing systems (e.g., Apache Kafka), iv) response time in batch processing systems (e.g., Apache Hadoop), (v) response time for processing top-k queries in transactional systems (e.g., Apache Hive); (vi) read/write latency and throughput for the file system of big data cluster; and vii) utilization and energy efficiency for CPU resources. Therefore, it is not yet clear how i) these QoS metrics could be defined and formulated coherently across layers and ii) the various QoS metrics should be combined to give a holistic view of the of the data flows across multiple IoT sensors, big data software frameworks and hardware resources. In order to ensure application-level performance SLAs/QoS there is also need to monitor workload metrics (data volume, data velocity, data variety and sources, and types and mix of search queries) across big data processing frameworks such that appropriate workload characterization models could be developed. The hard challenge is how to collect and integrate monitoring data from all the big data processing frameworks and hardware resources for administrators to easily track and understand application level SLAs/QoS without the need to understand the complexity of the whole platform.

**QoS-aware cloud and network selection and orchestration:** Over the past decade, we have witnessed a significant increase in the usage of smartphones, sensors, and applications. These technologies will become an integral part of the CoT ecosystem, thereby bootstrapping novel applications in the areas of emergency management, healthcare, industrial informatics, and transportation and logistics. It is evident that for these classes of applications, data would be sent (sensor data from heart rate monitor) and retrieved (processed response) while the users (e.g., emergency personal and patients) are on-the-move. Mobility has its inherent challenges, for example, a mobile device may connect to different access networks, for example, 3G and WiFi; where each network offers different latencies for upstream data (from the device) to downstream data from the clouds. This creates several challenges: first, there is a need to select a best network as mobile devices can connect to heterogeneous access networks leading to stochastic network QoS; and second, QoS-aware cloud selection as the CoT applications and data will be hosted in multiple datacenters. Therefore, we consider end-to-end QoS provisioning in CoT as a joint optimization problem where both network and cloud QoS should be optimized together.

A large body of work exists in the area of cloud selection. For example, authors in [16] present analytic hierarchy process based method for cloud ranking and selection. However, there is a dearth of literature that considers joint cloud and network selection. Authors in [17] proposed a system for cloud and network selection while users are on the move in heterogeneous access networks. However, these methods are also limited. For example, when a set of applications running on multiple devices selects a particular cloud for application data processing, all the requests may automatically gets transferred to that cloud datacenter, creating burden on servers running on that datacenter, leading to overprovisioning on one datacenter and under provisioning on another datacenters. Further, this may also trigger VM
migrations and data replication across multiple datacenters leading to an inefficient CoT ecosystem. This necessitates development of novel cloud orchestrators that are QoS and mobility-aware to for efficient QoS provisioning in CoT.

V. CONCLUSION

The rising trends in Cloud of Things can prove to be a disruptive technology enabling seamless and QoS-aware services across multiple application domains that can be the part of future smart cities. CoT application systems have the potential to provide a competitive edge to businesses and society. However, the intersection between IoT, cloud and big data analytics systems is still relatively untapped. Techniques for orchestrating the new breed of CoT applications (e.g., disaster management, smart home, remote healthcare, and smart energy grids etc.) in the cloud need to be highly elastic and intelligent. These applications generate massive, heterogeneous and fast streams of data that require guarantee on real-time application-level QoS (aQoS) constraints such as, minimise event detection delay and maximise application availability, in order to extract significant value.

Quite simply, existing state-of-the-art in cloud-based big data processing frameworks (BDPF) such as Apache Hadoop and Apache Storm and the orchestration platforms (e.g. Apache YARN, Amazon Elastic Map Reduce) are computationally challenged and have no support for meeting aQoS constraints. They do not scale well when subjected to unpredictable workload (changes in data volume, data velocity, data variety, query types) and disruptive events (e.g., failures, delay in detection, contention, and overloading) that could potentially degrade the application’s performance.

In the future the research community will need to investigate and propose elastic orchestration algorithms that extend the capabilities of existing BDPF to intelligently exploit the cloud computing resources (CPU, Storage, and Network) for processing massive, heterogeneous and fast streams of CoT data such that aQoS constraints are constantly achieved under uncertain workload (IoT sensor behaviours), resource and network conditions.

REFERENCES


