Emerging Edge Computing Technologies for Distributed IoT Systems

Ali Alnoman, Shree Krishna Sharma, Waleed Ejaz and Alagan Anpalagan

***Abstract*—The ever-increasing growth of connected smart devices and Internet of Things (IoT) verticals is leading to the crucial challenges of handling the massive amount of raw data generated by distributed IoT systems and providing timely feedback to the end-users. Although existing cloud computing paradigm has an enormous amount of virtual computing power and storage capacity, it might not be able to satisfy delay-sensitive applications since computing tasks are usually processed at the distant cloud-servers. To this end, edge/fog computing has recently emerged as a new computing paradigm that helps to extend cloud functionalities to the network edge. Despite several benefits of edge computing including geo-distribution, mobility support and location awareness, various communication and computing related challenges need to be addressed for future IoT systems. In this regard, this paper provides a comprehensive view on the current issues encountered in distributed IoT systems and effective solutions by classifying them into three main categories, namely, radio and computing resource management, intelligent edge-IoT systems, and flexible infrastructure management. Furthermore, an optimization framework for edge-IoT systems is proposed by considering the key performance metrics including throughput, delay, resource utilization and energy consumption. Finally, a Machine Learning (ML) based case study is presented along with some numerical results to illustrate the significance of ML in edge-IoT computing.**

I. Introduction

The next generation of Information and Communication Technology is characterized by the ubiquity of smart devices and machines that perform intelligent functions by autonomously sensing, analyzing, and exchanging information via the Internet. From E-health, smart homes and intelligent transportation to industrial manufacturing and supply chain, Internet of Things (IoT) is intended to provide humanity with easier, safer and more intelligent lifestyle. However, the rapid growth of IoT applications has increased the number of connected 'Things' to unprecedented levels. The number of connected devices is forecasted to reach about 125 billion (IHS Markit) by 2030, and Machine-to-Machine (M2M) communications, which constitutes a large proportion of IoT applications, is expected to occupy almost 45% of the entire network traffic by 2022 [1].

This remarkable increase in the number of connected devices needs to be accompanied by an equivalent increase in resource provisioning to avoid any sort of service disruption. Although the existing cloud computing paradigm is highly capable of handling the massive amount of data, it is not suitable for distributed IoT systems due to the potentially incurred delays [3]. For this reason, providing computing, storage, and communication functionalities at the network edge helps not only in reducing the end-to-end delay, but also can alleviate the burdens on cloud-servers and backhaul links. Furthermore, due to the physical proximity of edge devices with end-users, edge computing can support distributed IoT applications that require location awareness and higher Quality of Service (QoS) [3, 4].

In contrast to the conventional IoT architecture, where storage and computing operations are mostly performed in the cloud-center, distributed IoT systems incorporate nodes/-gateways/servers at the network edge to fulfill heterogenous IoT requirements with less delay and energy consumption. However, edge nodes might not always have sufficient computing and storage resources to process the massive amount of IoT data; therefore, the cooperation between edge and cloud entities is indispensable to take the best of both computing paradigms [3]. In addition to providing computing capabilities at the vicinity of IoT users, edge devices can perform various pre-processing tasks such as data classification and filtration, service-level agreement ranking, and parameter measurements before involving the central-cloud.

One of the crucial challenges in IoT systems is the limited radio resources required to provide reliable connectivity to the massive number of devices. Herein, one of the envisioned solutions to cope with the scarcity of radio resources is to exploit and integrate all available communications, caching and computing resources and Radio Access Technologies (RATs) such as 5G, LTE and WiFi by using efficient resource allocation schemes. In addition, harnessing large numbers of low-power small-cell Base Stations (SBSs) can improve the cellular network capacity by allowing spatial frequency reuse over small geographical areas. However, dealing with systems characterized by such resource heterogeneity requires sophisticated management and control schemes. To this end, implementing softwarized and virtualized platforms such as Software-Defined Networking (SDN) and Network Function Virtualization (NFV) technologies can significantly ease and automatize the entire network control [5].

Most existing research works have focused on centralized IoT systems without providing a high-level coordination among the distributed communication and computing entities. Furthermore, SDN, NFV, big data analytics and artificial intelligence are usually introduced as application-specific technologies/platforms rather than being adopted within a fundamental

A. Alnoman and A. Anpalgan are with the Department of Electrical and Computer Engineering, Ryerson University, 350 Victoria St., Toronto, Canada (Email: ali.alnoman@ryerson.ca, alagan@ee.ryerson.ca), S. K. Sharma (corresponding author) is with the SnT, University of Luxembourg, L-1855, Luxembourg (Email: shree.sharma@ieee.org), W. Ejaz is with the Department of Applied Science & Engineering at the Thomson River University, British Columbia, Canada (Email: walid.ejaz@ieee.org.)

optimization framework. In this work, we aim to present comprehensive insights on distributed IoT systems by taking into account the challenges that encounter both radio and computing elements. Moreover, effective potential solutions that foster adaptivity, elasticity, and self-learning capabilities are also introduced. The main contributions of this paper are highlighted below:

- Introduce the main practical challenges facing distributed IoT computing systems and highlight the potential solutions.
- Provide a classification of the emerging technologies in distributed IoT systems into different sub-categories along with their relevant discussions.
- Propose an optimization framework to tackle various system-level aspects such as computing, delay, scheduling, and energy consumption.
- Present a Machine Learning (ML)-based case study for efficient IoT device clustering in the context of edge-IoT system optimization.

II. Computing in Distributed IoT Systems: Challenges and Potential Solutions

The ever-increasing number of IoT devices, and the heterogeneous nature of their demands pose many practical challenges, especially in regard with the system management and resource provision. Edge computing can help to resolve these challenges by exploiting the physical proximity with IoT devices towards supporting context-awareness, data filtration, and on-demand resource provision at the network edge. In this section, we categorize the main challenges in edge-IoT systems into three main domains, and present potential enabling solutions as listed in Table I.

A. Heterogeneity of IoT Systems

The ubiquity of IoT devices in a variety of applications diversifies the Quality of Experience (QoE) requirements. One of the important QoE parameters is the end-to-end delay experienced by IoT devices. While some IoT applications such as climate monitoring can tolerate up to several minutes of delay, other IoT applications including autonomous driving and biomedical sensors can tolerate only a few milliseconds. In addition, machine-type communications is generally characterized with a bursty and low data-rate transmission, and with the massive number of connected machines, traffic burstiness can overload or even crash the cellular network; in particular, the mobility management units. For instance, according to real-time measurements in a wideband network, the duration between 90% of subsequent cell congestion occurrences lasts for less than 13 minutes, and about 90% of these occurrences last for only 1.2 seconds [7]. Hence, there is an indispensable need to devise flexible radio resource allocation schemes that can adapt to such burstiness and temporary variations in edge-IoT networks. In the context of cellular IoT networks, the Baseband Unit (BBU) pool that supports the architecture of Heterogeneous Cloud Radio Access Networks (H-CRANs) can provide significant assistance since all network resources are virtualized and managed by a unified controller.

From the information security perspective, the limited computing resources of IoT devices may lead to serious security challenges especially when tasks are offloaded to the remote servers. Thus, it is essential harness the powerful edge computing resources and support IoT devices with ML and self-organizing capabilities to reduce malicious attacks. Also, it is important for IoT devices to consider the potential delay and energy consumption when making an offloading decision [6].

B. Resource Management

The spatial and temporal variations of IoT devices' demands necessitate the bi-directional load sharing mechanisms between cloud and edge servers to maximize the computing resource utilization and reduce the resource wastage. In a similar context, caching popular contents at the network edge is considered a promising solution to reduce service delay and load at the central-cloud. Content caching along with data aggregation and analysis can also help to reduce redundant data storage and transmission. Since IoT devices in general have more demand on the uplink (e.g., IoT sensors), efficient frequency allocation schemes need to be developed to accommodate the massive number of connected IoT devices. To this end, hybrid multiple access schemes help to combine the merits of both the scheduled and random multiple access schemes depending on the real-time system parameters [8].

C. Performance Coordination

Integrating a wide variety of services, devices, and RANs in the IoT paradigm imposes many challenges in regard with QoE provision, compatibility, load sharing, and network-wide synchronization. In addition, high complexity of IoT systems necessitates the need for distributed network intelligence to provide edge nodes with decision-making capabilities and to establish a stand-alone IoT environment [9]. For instance, although cloud-servers have more powerful capabilities, the round-trip time between IoT devices and the cloud might not satisfy the desired QoE for delay-sensitive applications. Therefore, intelligent coordination among different fog layers on one hand, and between fog layers and the cloud-center on the other hand, must be maintained to minimize the latency experienced by the end-users.

Another major issue in the implementation of distributed IoT computing is standardization. Unlike Mobile Edge Computing (MEC), which was standardized by the European Telecommunications Standards Institute (ETSI), IoT community is still facing challenges in making global IoT standards towards better flexibility and interoperability [2]. IoT standardization involves the development of technical standards in regard with its architecture, protocols, identification, and security. On the communication side, standardization seems more encouraging as the 3GPP Release 13 has already revealed the Narrowband IoT technology to enable low-power wide area networking for IoT systems.

III. Classification of Emerging Edge-IoT Technologies

Since IoT systems inherently integrate both cellular and computing functions, several technologies should be consid-

TABLE I: Challenges and Potential Solutions for Edge-IoT Systems

Domain	Challenges	Potential solutions
Heterogeneity of IoT s	Delay	Intelligent task offloading mechanisms
	Mobility	BBU pool, distributed fog nodes, and Ad-hoc fogs
	Security and privacy	IoT authentication, access control, ML-based malware detection, pseudonymization techniques, and secured task offloading
Resource management	Insufficient computing resources	Bi-directional resource sharing between edge and cloud servers
	High demand on a particular content	In-network caching
	Redundant data transmission	Data aggregation and analysis
	More demand on the uplink	Scheduling techniques
Performance coordination	Various computing entities	Cooperative hierarchical architectures (e.g., fog-to-cloud and cloud-to-fog)
	Multiple service providers	Standardization and utilization of compatible infrastructures and platforms
	Interoperability	Network slicing, cross-layer optimization, and load sharing

ered to enable the joint operation of radio and computing infrastructures. Herein, we classify the enabling technologies for edge-IoT systems into three categories, as highlighted in Fig. 1, and provide their brief discussions below.

A. Coordinated Resource Management

1) Communication Resources: Towards improving both spectral and energy efficiencies, and reducing the overhead of information exchange among the cellular nodes, the emerging C-RAN architecture can maintain efficient coordination among the coexisting RATs in the unified SDN-based BBU pool. For instance, 5G, LTE, WiFi, Coordinated Multi-Point (CoMP), millimeter wave (mmWave), massive-MIMO, and Non-Orthogonal Multiple Access (NOMA) technologies can operate concurrently in different network layers and nodes with less communication overhead [8]. Moreover, Device-to-Device (D2D) communications can reduce burdens on both cellular and computing infrastructures by allowing mobile IoT devices to use out-of-band frequencies and benefit from the available computing resources of nearby devices [9].

2) Computing Resources: The hierarchical computing operations including the cloud-to-fog and fog-to-cloud paradigms helps to improve the utilization of computing resources by leveraging intelligent task offloading decisions. It is also expected in future computing networks that every mobile user with computing capability can take part in the global computing process. The concept of consumer-as-a-provider is one of the implementations that allow user devices to share their available computing capabilities with ambient devices, or further, with the fog and cloud servers. Furthermore, the volatile vehicular and drone-based fogs can play a role in providing on-demand computing services for adjacent IoT users with reduced latency.

3) Joint Communication and Computing Resources: Providing SBSs with cloud-like computing capabilities can transform those SBSs from just radio access nodes into the so-called 'smart SBSs'. In LTE-based systems, the small-cell cloud enhanced eNodeB paradigm is a practical implementation of smart SBSs wherein both cellular and edge functionalities can be attained [10]. On one side, SBSs in this paradigm are connected with the cellular core via backhaul links; on the other side, SBSs are associated with mobile devices to provide computing services. By exploiting the accessibility of network-wide content statistics, SBSs can achieve accurate content caching which helps to reduce both latency and communication overhead. Therefore, both network providers and cloud operators will have to collaborate in order to optimize the allocation of backhaul, frequency, and computing resources. Furthermore, the interplay between cloud and edge providers can support a broad diversity of IoT applications that have various QoE requirements [3].

B. Intelligent Edge-IoT Systems

The enormous amount of data collected from IoT devices form an important source of data-sets that can train and improve the accuracy of ML-based schemes. In the context of IoT, efficient ML schemes can assist in a variety of computing aspects such as content caching, task offloading, and device clustering that empower the decentralized operation of IoT systems. In the following, we discuss the important components of intelligent edge-IoT systems highlighted in Fig. 2.

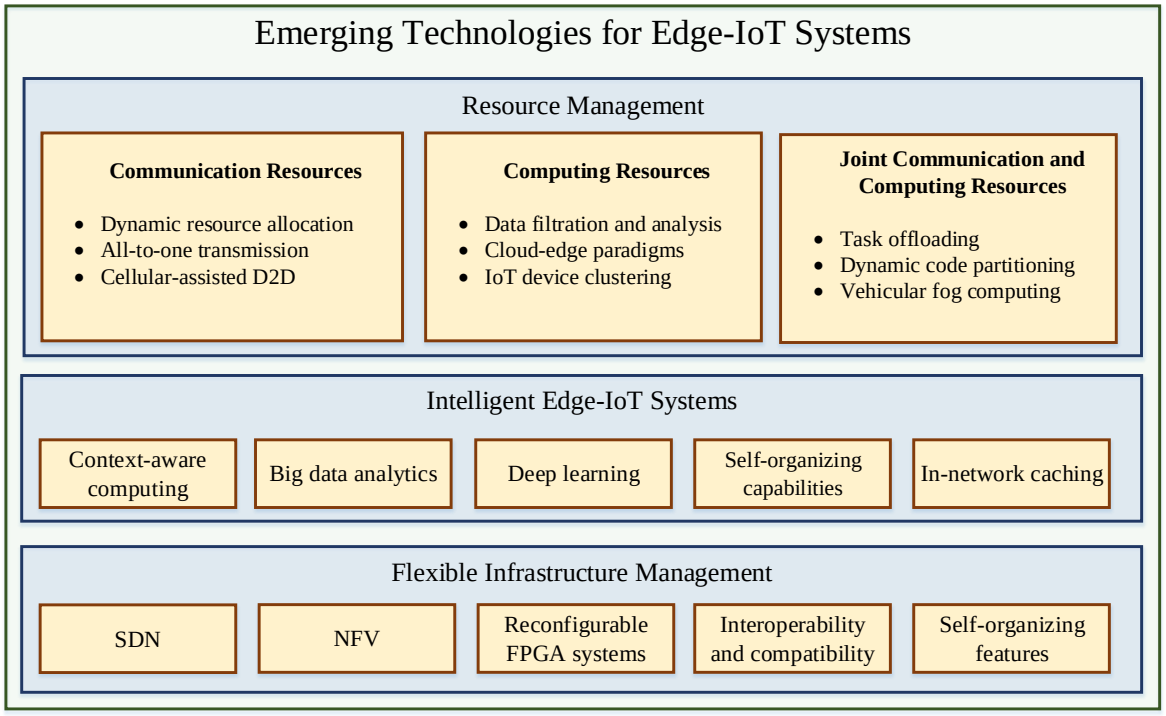


Fig. 1: Classification of emerging edge computing technologies for distributed IoT systems.

1) Content Caching: The recent advances in the caching-enabled cellular architectures will help network elements to carry out a comprehensive reasoning and prediction about network conditions and resources. In addition, the existence of SBSs in the vicinity of mobile users assists in providing accurate spatio-temporal statistics about popular contents; as a result, more accurate content caching and better usage of the limited storage capacity can be achieved.

2) Big Data Analytics: Performing computing tasks at the network edge simplifies the evaluation and analysis of IoT data through aggregation, filtering, and pre-processing prior to sending those data to the distant cloud. As a result, less communication overhead and processing delay can be achieved, and that is important for computing-hungry applications especially those equipped with signal processing functions [4]. Aided by the emerging ML techniques, big data analytics can be harnessed to make accurate predictions of content popularity, cache these contents in the network edge, and provide timely feedback to end-users without adding burdens on backhaul resources [3, 11]. However, the complicated data analytic algorithms that are suitable for the powerful cloud need to be simplified in order to match with the resource- and computing-constrained edge-devices [9].

3) Machine/Deep Learning: The decentralized operation of distributed IoT systems can be realized via the application of efficient ML techniques that provide IoT devices and service providers with self-organizing and self-healing features. To this end, several ML-based approaches (e.g., Neural Networks (NNs), Support Vector Machines (SVMs), K-Means and Linear Regression) [13] are available to make intelligent decisions in several network aspects including load balancing, fault management, and adaptive resource allocation. Also, it should be noted that training data have to be carefully determined to achieve the best ML performance. For instance, applications that have stringent delay requirements such as autonomous driving can be guided using real-time information. On the other hand, feedback made for applications that are not delay-sensitive can be made based on long-term statistics [12]. To achieve a higher level of accuracy, ML can be upgraded to the so-called 'deep learning' such as deep NNs that contain more hidden layers and mathematical operations to enhance the feature extraction performance.

From the data security perspective, several ML techniques including supervised, unsupervised, and Reinforcement Learning (RL) can be applied. For instance, SVMs and NNs are supervised learning techniques that can be used in network intrusion detection. Multivariate correlation analysis is an unsupervised learning technique that can be used to detect the denial of service (DoS) attacks. Also, RL-based techniques such as Q-learning can be used to improve authentication and malware detection in IoT systems [14].

4) IoT Device Clustering: To avoid redundant data transmission, IoT devices can be classified and clustered according to predetermined criteria such as functionality or geographic location. For instance, home sensors can be processed by a central home controller to detect unusual data (e.g., air pollution) and forward these data to the homeowner or specialized authorities. Moreover, within a neighborhood, data collected from the ambient sensors can be processed and verified by a neighborhood controller (e.g., nearby edge device). That way, redundant data transmission from thousands of devices can be avoided allowing only abstracted meaningful information

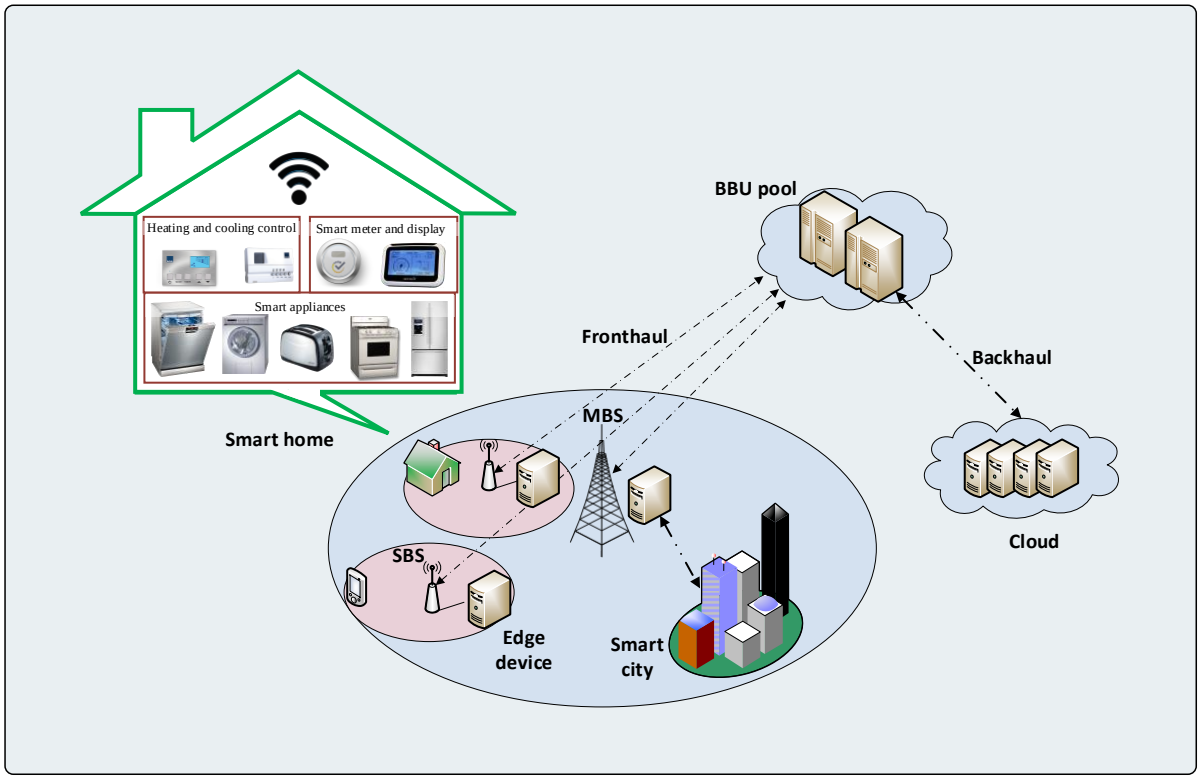


Fig. 2: Integrated cellular-computing architecture for edge-IoT systems. The cellular part of the system is represented by the MBS, SBSs, and the BBU pool, whereas the computing part includes edge devices and the central-cloud.

through the cloud-center.

The clustering process can be managed by edge devices that virtually cluster ambient IoT devices, extract correlated features of these clusters, and make necessary decisions such as predicting future behavior of devices or cache the popular contents for each particular cluster. A case study will be presented later in Section V to illustrate the effectiveness of IoT device clustering on the resource utilization and QoE performance.

C. Flexible Infrastructure Management

Due to the unprecedented density and heterogeneity of the connected IoT devices, network management can be quite challenging. This necessitates the implementation of flexible management technologies to maintain network adaptability, reconfigurability, and scalability. In the following, we discuss some of the key enablers for flexible infrastructure management.

1) SDN: As mentioned earlier, the layered fog-cloud architecture is a promising solution to achieve efficient resource management. However, this paradigm requires flexible packet forwarding over the participating fog layers on one side, and between fog layers and the cloud layer on the other side. The SDN technology can effectively facilitate this process and reduce the inherent complexity of the bi-directional fog-cloud operation. In addition, the SDN-based cellular core which adopts OpenFlow controllers and protocols can further ease the data forwarding process by facilitating control, mobility management, authentication and network virtualization. For instance, when an IoT device roams between different fog nodes, the SDN technology helps to calculate the delay cost of Virtual Machine (VM) migration, and then decides on whether to trigger the VM migration or not based on the cost evaluation [15].

2) NFV: By virtualizing multiple network functions on shared hardware, the NFV technology helps not only to reduce the complexity of both network administration and IoT system management [5], but also to improve system scalability by allowing resource sharing based on runtime needs [10]. Examples of NFV technology include the Content Delivery Network (CDN) and Platform as a Service (PaaS) paradigms that allow users to access particular contents on shared machines.

3) Flexible Radio: To maintain flexible radio resource provision, it is essential for cellular networks to conduct elastic frequency allocation strategies that can adapt to the real-time demands of IoT devices. Here, implementing the cognitive radio technology that supports adaptivity in resource allocation and network tier association (e.g., macro, pico, femto cells), can play an important role in realizing the goals of scalability and flexibility in IoT systems, especially when equipped with efficient ML capabilities.

IV. Proposed Optimization Framework for Edge-IoT Systems

Herein, we present an integrated cellular-computing architecture which is comprised of multiple cellular elements, namely, SBSs, MBS, BBU pool, and backhaul links, whereas

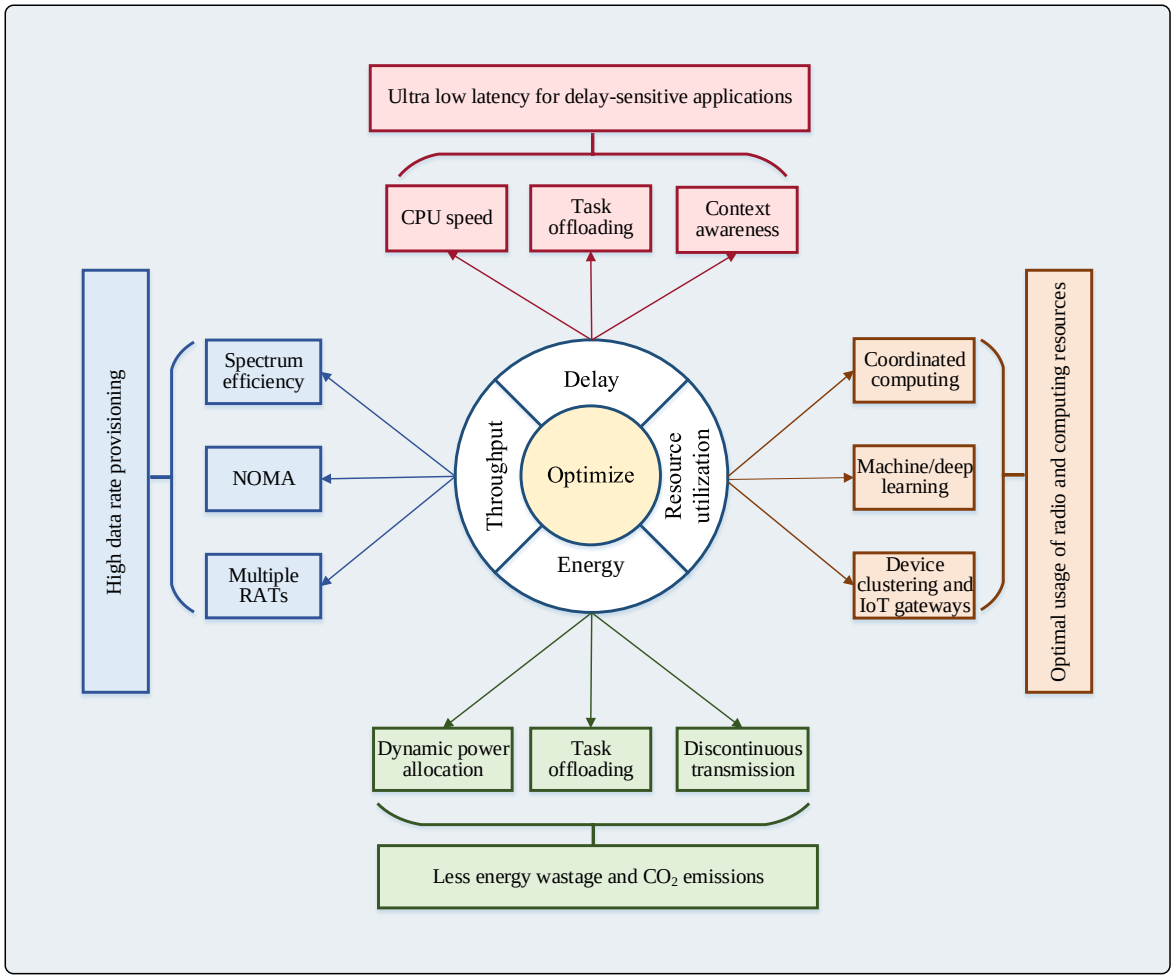


Fig. 3: Proposed optimization framework for distributed IoT systems considering delay, resource utilization, energy and throughput as the optimization objectives. Potential solutions are identified for each objective to effectively reach the desired optimization goals.

the computing elements are represented by the cloud and edge devices as depicted in Fig. 2.

As depicted in Fig. 3, the proposed optimization framework aims to achieve the following four objectives: (i) ultra-low latency for delay-sensitive IoT applications, (ii) optimal usage of available radio and computing resources, (iii) less energy consumption and CO_2 emissions, and (iv) high data rate provisioning. The first objective can be achieved via optimal CPU speed determination, smart offloading decisions, and context awareness. In the second objective, resource utilization can be enhanced by using SDN technology, coordinated computing, and efficient device clustering mechanisms. The third objective deals with energy efficiency which is a major concern in IoT systems. Energy consumption can be optimized using different techniques such as dynamic power allocation, energy-aware task offloading, and discontinuous transmission. Maximizing data rate is the fourth objective and can be fulfilled using several enabling technologies such as NOMA, distributed small cells, and multi-RAT network planning.

Since IoT systems constitute a mixture of both computing and communication networks, data security and user privacy are inherently involved in the aforementioned enabling technologies. For instance, IoT devices have to recognize the secure fog-servers to avoid identity-based attacks such as spoofing. Moreover, IoT resources in the edge nodes must be secured by undertaking efficient authentication and access control strategies. Coordinated computing schemes can provide IoT devices with enhanced defensive strategies benefiting from the powerful computing capabilities of the cloud and edge processors. To this end, ML-based data security techniques such as authentication, access control, and malware detection need to be implemented to strengthen IoT defense strategies against potential threats [14]. In the following, we briefly describe some promising solutions to improve the four performance metrics that are highlighted in Fig. 3.

1) **Delay**: Minimizing delay is one of the key objectives in IoT systems, especially for delay-sensitive applications such as autonomous driving and health-monitoring, in which delay consequences can be critical. Therefore, decisions on whether to process tasks using the on-device processor or to offload tasks to the edge-device have to be carefully made based on accurate estimation of the round-trip time and the amount of required CPU cycles that satisfy the task completion deadline set by IoT applications. Moreover, the arrival rate of incoming tasks to edge-devices must be incorporated in such decisions to avoid overloading server buffers.
2) **Resource utilization**: Establishing a high-level coordi-

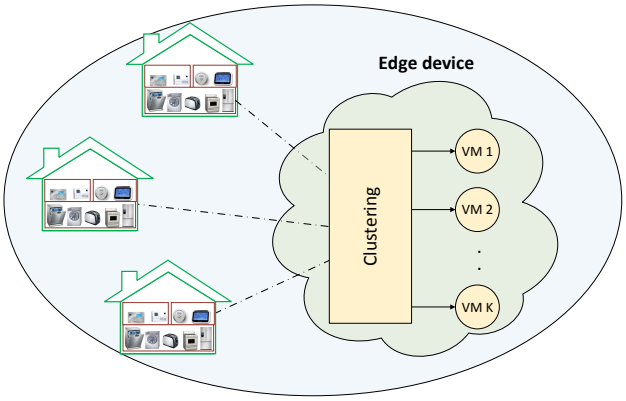
nation among IoT devices, edge-devices, and the central-cloud, is crucial to take full advantage of all available computing resources. From the communication point of view, the SDN-based BBU pool helps to optimally allocate frequency resources to IoT devices by taking the bandwidth and interference constraints into account. In the context of smart IoT systems, it is important to support IoT gateway systems with ML-based algorithms and protocols to dynamically adapt to the large-scale heterogeneity of IoT devices, and to automatize the process of registering newly added IoT devices in the network database. Moreover, the formation of volatile on-demand fog nodes and device clustering can optimize resource usage in edge-IoT systems.

3) **Energy**: As a major challenge in future communication and computing systems, energy has to be addressed not only to prolong the lifetime of the on-device batteries, but also to reduce the excessive amounts of CO_2 emission. In this direction, some promising solutions include efficient power allocation, short-range transmission via SBSs, CPU cycle optimization, task offloading, discontinuous transmission, and hardware sleeping mechanisms.
4) **Throughput**: Providing IoT devices with the powerful edge computing capabilities must be accompanied by a high data-rate provisioning to avoid undesired latencies. To this goal, exploiting the spatial frequency reuse of SBSs over small areas can increase the per-user data rate. In addition, the recent advances in communication technologies including NOMA, massive MIMO, 5G, LTE-A and WiFi can diversify the supply of frequency resources, and thus improve the spectral efficiency.

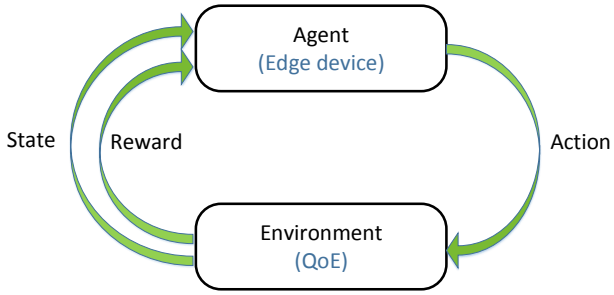
V. Use Case Study

Towards avoiding the congestion of backhaul links and saving both computing and radio resource, IoT devices can be virtually clustered at the network edge forming virtual IoT groups. Herein, we present a case study to demonstrate the significance of RL-assisted solutions in forming IoT device clusters. In such a scenario, RL-based data classification and aggregation can help not only to save the computing resources but also to support a scalable computing paradigm that can adapt to sudden changes in data volumes and traffic variations. Fig. 4 demonstrates the proposed case study.

We consider a distributed IoT system where smart home devices and sensors offload their tasks to the edge-devices for processing as shown in Fig. 4(a). The clustering is managed by the edge-device which is located at the vicinity of IoT devices and possesses K servers (VMs). The edge-device aims to maximize each VM's utilization by aggregating IoT tasks under the constraint of VM capacity and task completion deadline of IoT devices. Fig. 4(b) illustrates the RL-based scheme where the edge-device is responsible for taking the increment and decrement action on the number of IoT devices associated with each VM. The edge-device then monitors the QoE that results from the taken action; in particular, monitors the delay experienced by IoT devices. The number of states in



(a) System layout.

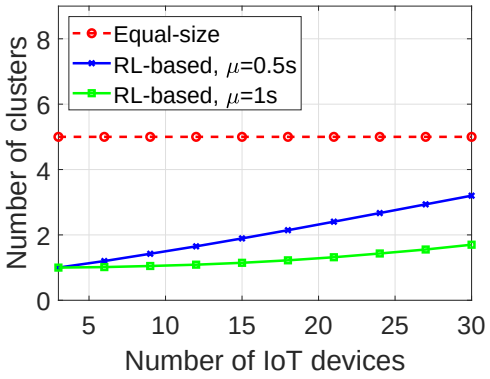


(b) RL-based learning mechanism.

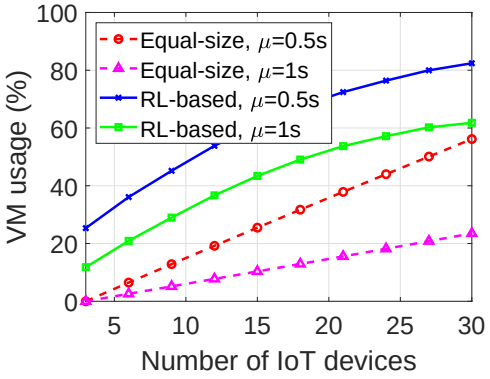
Fig. 4: Illustration of the proposed virtualized IoT device clustering.

the system corresponds to the number of IoT devices whereby the agent (edge-device) moves through these states and takes the required actions.

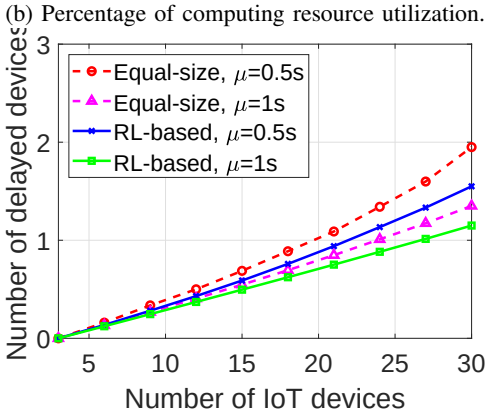
Each IoT device is assumed to have a particular packet size and a task completion deadline to accomplish the task. The packet size of IoT devices is uniformly distributed between 500 kB and 4 MB. The task completion deadline is assumed to differ among IoT devices as follows; the first group has a deadline range between $100 - 900$ ms (i.e., mean delay $\mu = 0.5$ s), whereas the second group has a deadline range of $500 - 1500$ ms (i.e., mean delay $\mu = 1$ s). The edge device has $K = 5$ VMs each of which has a processing capacity of 500 MHz. The aim of this study is to investigate the optimal number of IoT devices associated with each VM such that less VMs are to be used while satisfying the deadline requirement of IoT devices. To achieve this goal, Q-learning, which is an RL-based technique, is used to adaptively cluster IoT devices into available VMs. Two actions are considered during the Q-learning process, namely, 'increment' and 'decrement', where the reward of incrementing and decrementing the number of cluster members is $+5$ and -1, respectively. Thereby, increasing the number of IoT devices per cluster is always preferred; however, since the VM capacity is limited, cluster members can experience more delay, and the task completion deadline will be violated. To tackle the aforementioned problem, when any IoT device is considered 'delayed', the RL-based controller changes the reward of incrementing and decrementing the number of members into -10 and 5, respectively. The used values for the discount factor and learning rate in the



(a) Number of required clusters under different schemes.



(b) Percentage of computing resource utilization.



(c) Average number of delayed IoT devices.

Fig. 5: Performance of the proposed IoT device clustering method.

RL-based scheme are 0.9 and 0.1, respectively.

Fig. 5 depicts the system performance using the RL-based and equal-size clustering schemes. In the 'equal-size' clustering scheme, IoT devices are evenly distributed among all available VMs in the system regardless of the data size and task completion deadline. Unlike the 'equal-size' clustering scheme which lacks QoE awareness, the RL-based scheme monitors the QoE experienced by IoT devices. It can be observed in Fig. 5(a) that the RL-based scheme provides better performance in regard with reducing the amount of required computing resources represented by the number of VMs. The 'equal-size' clustering scheme utilizes all available VMs at all the time and that keeps the number of used resources fixed at 5, while the RL-based scheme is capable of increasing the percentage of VM utilization by filling up VMs more quickly compared to the other scheme and this explains the non-linear behaviour of the RL-based scheme as shown in Fig. 5(b). An interesting observation in this study is that despite all VMs are utilized in the 'equal-size' scheme, the RL-based scheme can provide better QoE performance (less delayed IoT devices) as observed in Fig. 5(c). This is due to the diversity of IoT devices' demands in regard with the data size and task completion deadline which requires QoE-aware and adaptive resource allocation mechanism. In other words, due to the variety of task sizes and task completion deadlines among IoT devices, some IoT clusters might have higher processing demands compared to others, and that results in some clusters being satisfied while others not. Also, it can be noted that having more stringent delay requirements (i.e., $\mu = 0.5$ s) necessitates the utilization of more computing resources (VMs), and increases the number of delayed devices since the task completion deadline can be exceeded more easily.

VI. Conclusions

Edge computing along with the central-cloud constitute a powerful computing paradigm for the practical implementation of distributed IoT systems. However, there still exist several challenges from both the communication and computing perspectives. Various technologies including cooperative resource management, ML, context-aware computing, and flexible infrastructure management have emerged in this direction as promising solutions. This article presented a comprehensive view on the existing research issues, and introduced potential solutions along with an optimization framework taking into account the key performance metrics in edge-IoT systems.

The upcoming IoT era calls for serious efforts in the direction of ML, SDN, and NFV to establish self-organized and self-resilient computing systems that can cope with the heterogeneity of IoT services. Furthermore, since both communication and computing parties are inherently integrated in IoT systems, system optimization must consider the constraints imposed by the limitations of cellular networks such as insufficient resources, spatio-temporal variations of wireless links, and other issues including energy consumption and cost. In other words, a high-level coordination among the cellular and edge nodes is essential to fulfill the deployment of efficient edge-IoT computing systems.

References

[1] A. Al-Fuqaha, *et al.*, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347-2376, 4th Quart. 2015.

[2] S. Chen, *et al.*, "A Vision of IoT: Applications, Challenges, and Opportunities With China Perspective," *IEEE Internet of Things J.*, vol. 1, no. 4, pp. 349-359, Aug. 2014.

[3] S. K. Sharma and X. Wang, "Live Data Analytics With Collaborative Edge and Cloud Processing in Wireless IoT Networks," *IEEE Access*, vol. 5, pp. 4621-4635, March 2017.

[4] N. Abbas, *et al.*, "Mobile Edge Computing: A Survey," *IEEE Internet of Things J.*, vol. 5, no. 1, pp. 450-465, Feb. 2018.

[5] F. van Lingen, *et al.*, "The Unavoidable Convergence of NFV, 5G, and Fog: A Model-Driven Approach to Bridge Cloud and Edge," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 28 – 35, Aug. 2017.

[6] X. Meng, W. Wang, and Z. Zhang, "Delay-Constrained Hybrid Computation Offloading with Cloud and Fog Computing," *IEEE Access*, vol. 5 PP. 21355-21367, Sep. 2017.

[7] X. Zhou, *et al.*, "Toward 5G: When Explosive Bursts Meet Soft Cloud," *IEEE Network*, vol. 28, no. 6, pp. 12 – 17, Dec. 2014.
[8] A. Alexiou, "Wireless World 2020: Radio Interface Challenges and Technology Enablers," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 46 – 53, Mar. 2014.
[9] Y. Sahni, *et al.*, "Edge Mesh: A New Paradigm to Enable Distributed Intelligence in Internet of Things," *IEEE Access*, vol. 5, pp. 16441-16458, Aug. 2017.
[10] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, "Communicating While Computing: Distributed Mobile Cloud Computing Over 5G Heterogeneous Networks," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 45 – 55, Nov. 2014.
[11] E. Zeydan, "Big Data Caching for Networking: Moving from Cloud to Edge," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 36 – 42, Sept. 2016.
[12] Y. Mao, *et al.*, "Stochastic Joint Radio and Computational Resource Management for Multi-User Mobile-Edge Computing Systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5994 – 6009, Sept. 2017.
[13] C. Jiang, *et al.*, "Machine Learning Paradigms for Next-Generation Wireless Networks," in *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98-105, Apr. 2017.
[14] L. Xiao and X. Wan and X. Lu and Y. Zhang and D. Wu, "IoT Security Techniques Based on Machine Learning: How Do IoT Devices Use AI to Enhance Security?," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 41 – 49, Sep. 2018.
[15] X. Sun and N. Ansari, "EdgeIoT: Mobile Edge Computing for the Internet of Things," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 22 – 29, Dec. 2016.

Authors' Biographies

Ali Alnoman received his B.Sc. and M.Sc. degrees in Electrical Engineering from the University of Baghdad, Iraq, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering at Ryerson University, Toronto, Canada. His research interests include energy efficiency and resource allocation in HetNets, IoT, and cloud computing. During 2012-2014, he worked as a faculty member at Ishik University, Erbil, Iraq. He also served as a technical program committee member in the IEEE Vehicular Technology Conference VTC2017-Fall in Toronto.

Shree Krishna Sharma is currently a Research Scientist at the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg. Prior to this, he held postdoctoral research positions at the University of Western Ontario, Canada and at the SnT after obtaining his Ph.D. degree from University of Luxembourg in 2014. He has published more than 80 technical papers in scholarly journals and international conferences, and has over 1500 google scholar citations. His current research interests include 5G and beyond wireless, IoT, machine learning and edge computing. He is Senior Member of IEEE and an Associate Editor of IEEE Access.

Waleed Ejaz is an Assistant Professor in the Department of Applied Science & Engineering at Thompson Rivers University, Kamloops, BC, Canada. Previously, he held academic and research positions at Ryerson University, Carleton University, and Queenâs University in Canada. He received the Ph.D. degree in Information and Communication Engineering from Sejong University, Republic of Korea. He has co-authored over 90 papers in prestigious journals and conferences, and 3 books. His current research interests include Internet of Things (IoT), energy harvesting, 5G and beyond networks, and mobile edge computing. He is a registered Professional Engineer in the province of British Columbia, Canada.

Alagan Anpalagan received the BASc, MASc, and PhD degrees in electrical engineering from the University of Toronto, Canada. He is a Full Professor in ELCE Department at Ryerson University, Canada where he served as Associate Chair for Graduate Studies, EE Program Director, and Graduate Program Director. He has co-authored four edited books including Design and Deployment of Small Cell Networks, Powerline Communication Systems for Smart Grids, and Handbook on Green Information and Communication Systems. He was a recipient of IEEE Canada J.J. Ham Outstanding Engineering Educator Award (2018), Ryerson YSGS Outstanding Contribution to Graduate Education Award (2017), and IEEE M.B. Broughton Central Canada Service Award (2016). He is a Fellow of IET and a Fellow of the Engineering Institute of Canada.