HybridIoT: Integration of Hierarchical Multiple Access and Computation Offloading for IoT-Based Smart Cities

Li Ping Qian, Yuan Wu, Bo Ji, Liang Huang, and Danny H. K. Tsang

Abstract

The Internet of Things (IoT) is an emerging technology that proffers to connect massive smart devices together and to the Internet. On the basis of IoT, a smart city is endowed with real-time monitoring, ubiquitous sensing, universal connectivity, and intelligent information processing and control. An IoT-based smart city can offer various smart services to citizens and administrators, thus improving the utilization of public resources regarding transportation, healthcare, environment, entertainment, and energy. The integration of transmitting, computing, and caching is having a profound impact on the development of flexible and efficient IoT in smart cities. However, with the introduction of ultra dense networking (UDN) and mobile edge computing (MEC), we have to carefully consider a joint problem across the physical layer and MAC layer to enable the efficient transmission, computation, and caching of big IoT data generated by massive IoT devices distributed in a city. In doing so, efficient multiple access and computation offloading should be addressed in the physical layer and MAC layer, respectively. In this article, we propose a scalable and sustainable IoT framework that integrates UDN-based hierarchical multiple access and computation offloading between MEC and cloud to support the smart city vision. The proposed integrated framework can substantially reduce the end-to-end delay and energy consumption of computing data from massive IoT devices. Numerical comparison results are presented to show the efficiency of the proposed framework. In addition, we discuss a number of open research issues in implementing the proposed framework.

Introduction

The Internet of Things (IoT) paradigm enables universal interconnections among a tremendous number of ubiquitous smart devices and objects (e.g., smart cars, wearable devices, smartphones, tablet computers, industrial and utility components) via a network of networks anywhere at any time [1]. With the rapid advancement of IoT technology, the IoT-based smart city has received significant attention and is emerging as a promising technology integrated with ubiquitous sensing, universal networking, intelligent information processing, and real-time control [2]. The key goal of the IoT-based smart city is to efficiently utilize public resources, thus offering a broad range of intelligent applications, including smart metering, smart manufacturing, smart home, automatic driving, and health monitoring. In the context of the IoT-based smart city, ubiquitously connected IoT devices are deployed to monitor the physical world in people's daily lives in real time by collecting and uploading their local sensed contents such as images, videos, and textual data [3]. Due to the limited spectrum-computation resource faced with the unprecedented IoT traffic volume, current wireless cellular networks are becoming incapable of guaranteeing the efficient transmitting, computing, and caching of big IoT data in smart cities.

To cope with these challenges, ultra dense networking (UDN) [4] and mobile edge computing (MEC) [5] are emerging as promising technologies for IoT. UDN increases the network capacity and extends the network coverage to accommodate the 1000× capacity delivery of IoT traffic through deploying ultra-dense small cell base stations (BSs) [4]. MEC provides cloud computing, resource caching, and networking capabilities as well as an IT service environment at the edge of the radio access network (RAN) (e.g., beside the small cell BS) in close proximity to mobile devices, which offers ultra low-latency and high-bandwidth context-aware services [6]. It is envisioned that when MEC-enabled UDN is integrated into IoT, enormous potential benefits can be brought to various smart city applications. However, many challenges remain unsolved, such as multiple access and computation offloading.

MEC-enabled small cell BSs play the role of accessing, computing, and caching various IoT data in MEC-enabled UDN. With the ever growing popularity of IoT devices, the big IoT data generated from IoT devices are delivered to MEC-enabled small cell BSs. It is a potential bottleneck for massive access to the small cell BS when the limited spectrum resources are orthogonally allocated to the devices. Thus, new multiple access techniques are essential to meet the heterogeneous demands on low latency, high reliability, massive connectivity, and high throughput through enabling massive IoT devices to effectively share spectrum resources in UDN. MEC works

This work was supported in part by the National Natural Science Foundation of China under Project 61379122, Project 61572440, and Project 61502428, in part by the Zhejiang Provincial Natural Science Foundation of China under Project LR16F010003 and Project LR17F010002, and in part by the Open Research Fund of the National Mobile Communications Research Laboratory, Southeast University (No. 2019D11).

Digital Object Identifier:
10.1109/MNET.2019.1800149

Li Ping Qian is with Zhejiang University of Technology and the National Mobile Communications Research Laboratory, Southeast University, Yuan Wu (corresponding author) and Liang Huang are with Zhejiang University of Technology; Bo Ji is with Temple University; Danny H. K. Tsang is with Hong Kong University of Technology and Science.

0890-8044/19/$25.00 © 2019 IEEE

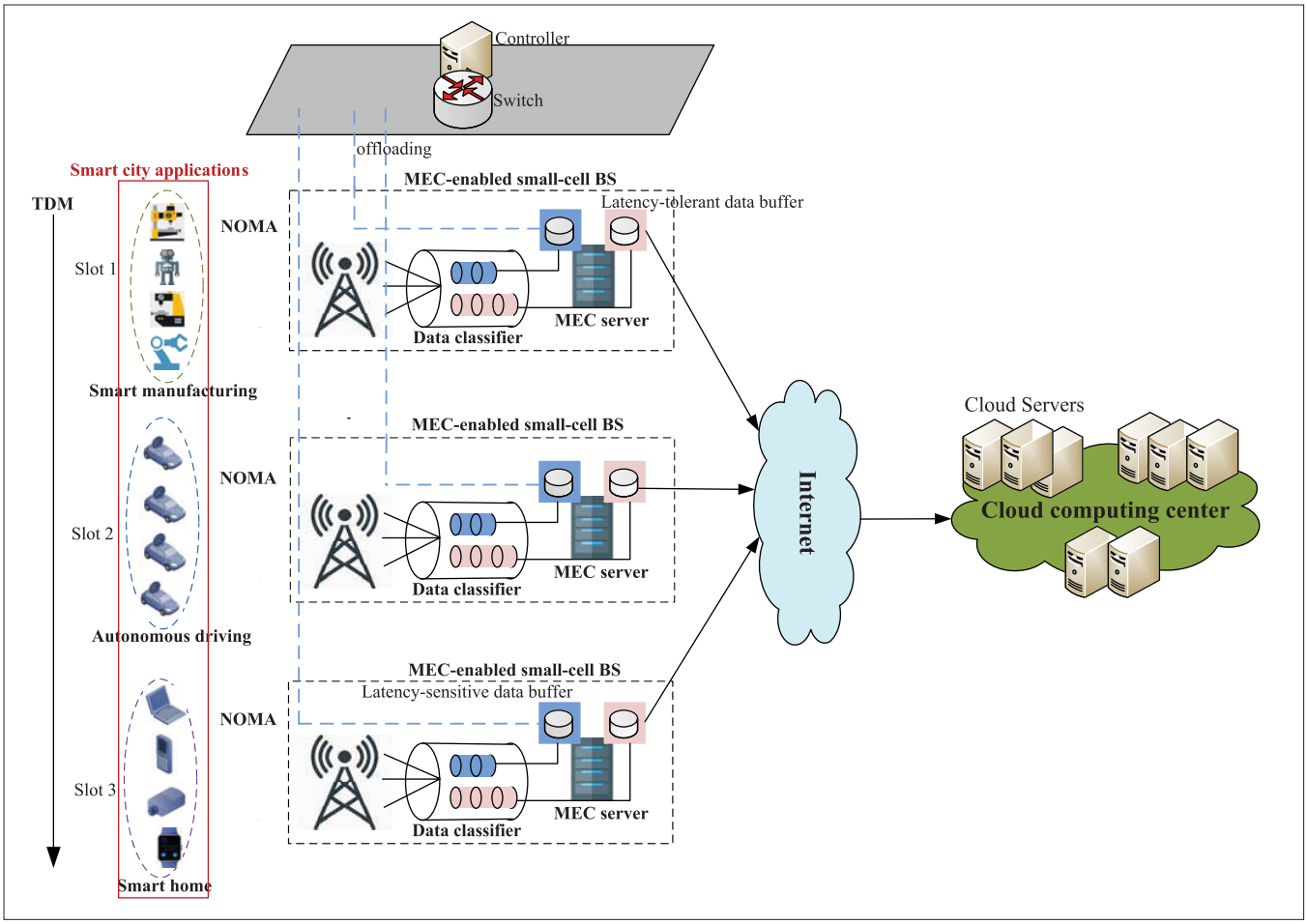

FIGURE 1. Architecture of HybridIoT.

as a smart "brain" for computing and caching big IoT data at the network edge. However, it is difficult to process a flood of big IoT data due to the limited computation resource provisioning in MEC. Although delivering big IoT data from IoT devices to the remote cloud center via the Internet suffers severe transmission delay, the cloud poses flexible and efficient resource provisioning for data processing [8]. With the diversity of big IoT data, the latency-tolerant IoT data can be offloaded to the remote cloud via the Internet for improving the computation and caching efficiency in MEC. Therefore, it is necessary to design an efficient big data transmission and computation architecture to explore the valuable information from IoT devices in IoT-based smart cities.

In this article, we address the aforementioned challenges arising in IoT-based smart cities from the physical layer and medium access control (MAC) layer of IoT networks, such as multiple access of massive IoT devices and computation offloading of big IoT data. To this end, a scalable and sustainable IoT architecture (i.e., HybridIoT) is proposed to efficiently transmit, compute, and cache big IoT data for various smart city applications by leveraging UDN-based hierarchical multiple access and computation offloading between MEC and cloud. The procedure for UDN-based hierarchical multiple access is provided. Specifically, non-orthogonal multiple access (NOMA), which allows multiple simultaneous transmissions to be active on the same frequency-time resource [7], is developed as a promising multiple access technique to provide the direct access of IoT devices to one small cell BS. Furthermore, to reduce transmission collisions among small cells due to frequency reuse, all small cell BSs operate in time-division multiplexing (TDM) mode, in which every small cell BS performs data transmission/reception at individual time slots. Then, based on the service requirements of IoT applications and the computation resource constraints of MECs, big IoT data are classified into latency-sensitive data (e.g., online gaming and augmented reality) and latency-tolerant data (e.g., web, chat, email) by the data classifier equipped at a small cell BS. In Internet Engineering Task Force (IETF) RFC 2474, the data classification is carried out according to the 6-bit differentiated services code point (DSCP) field in the IP header of each data packet. In particular, the application is identified as delay-sensitive if the DSCP decimal value belongs to the set of {8, 10, 14, 18, 22, 24, 28, 34, 36, 38}. Otherwise, the application is identified as delay-tolerant. The latency-tolerant data is forwarded to the remote cloud via the Internet, and the latency-sensitive data is offloaded among MECs to minimize the end-to-end computation delay or computation cost. The end-to-end computation delay and computation cost are analyzed and verified for HybridIoT. Finally, open research topics are discussed.

The remainder of this article is organized as follows. We describe an overview of recent lit-

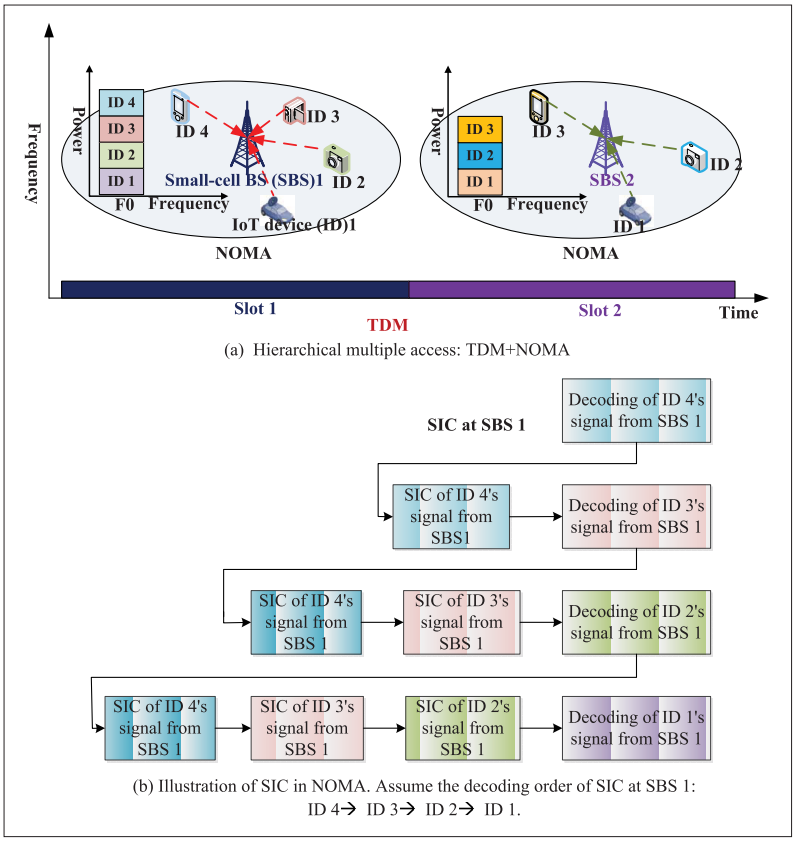

FIGURE 2. Hierarchical multiple access in HybridIoT and an illustration of performing NOMA.

erature on IoT. We present the IoT architecture (i.e., HybridIoT) proposed for smart cities, which integrates UDN-based hierarchical multiple access and computation offloading between MEC and cloud. Some performance of HybridIoT is analyzed and evaluated through simulations. Finally, we close the article with conclusions and discussions on future research.

Overview of Related Work on IoT

As the basis of the smart city concept, the advancement of IoT technology plays a significant role in IoT-based smart city realization. In this section, we briefly introduce some past and present efforts on IoT for the efficient transmission, computation, and caching of big IoT data.

Computation Architecture for IoTs

In order to meet the computing and caching requirements of big IoT data, various computing paradigms have been proposed, such as cloud computing, fog computing, and MEC. In the framework of cloud computing, the remote cloud center provides flexible and efficient resource provisioning for data processing and caching [8]. Thus, most data need to be transferred from the distributed IoT devices to the cloud center via the Internet. However, big data transmission via the Internet may not be feasible or economical due to the limited network resource, such as bandwidth, energy, and time. Therefore, cloud computing may not be applied to latency-sensitive IoT applications. In order to meet the ultra low-latency demand, fog/edge computing and MEC have been proposed to enable computing services to reside at the edge of the network as opposed to servers in the cloud center [1]. In fog computing, near-user fog servers compose a distributed computing system to carry out a substantial amount of caching, networking, and computing. MEC is mainly oriented to RANs in close proximity to mobile devices. MEC servers integrated into mobile BSs provide computing and caching capacities for mobile services at the edge of RANs.

Cloud/fog computing and MEC have been widely developed for IoT. A typical architecture of integrating cloud/fog computing and MEC for IoT applications is introduced in [9]. The distributed IoT devices use cloud/fog computing or MEC depending on the types of their applications, contents, and services.

Multiple Access Technologies for IoTs

Efficient multiple access techniques have been significantly studied for a long time in wireless systems, including cellular networks and emerging IoT networks. Compressive random access has been studied for uplink access in traditional IoTs based on the idea of compressive sensing [10]. Since compressive random access needs to detect distinct pilot signals from all active IoT devices at the access point before performing access authorization, the uniqueness of pilot signals makes it difficult to satisfy the requirement of massive connectivity for IoT devices. Orthogonal multiple access (OMA) techniques have been proposed for multi-user connectivity in past and current cellular networks, such as time-division multiple access (TDMA) in the second generation (2G), code-division multiple access (CDMA) in 3G, and orthogonal frequency-division multiple access (OFDMA) in 4G cellular systems [7]. However, the conflict between the scarcity of radio resource and the large number of IoT devices renders OMA unsuitable for emerging IoT.

Along with the introduction of UDNs into IoT, NOMA has been actively investigated as a potential alternative to conventional OMA due to its superior benefits in spectral and energy efficiency, massive connectivity, and low transmission latency. By means of NOMA, the technical challenges of IoTs, such as massive connectivity and utmost network capacity, can be partially fulfilled in the context of UDN.

Emerging Communication Technologies for IoT

The widespread proliferation of IoT devices leads to the exponential increase of IoT data. IoT networks have evolved to meet various demands for low power consumption, massive connectivity, low latency, and high data rate. To this end, various potential communication technologies in addition to multiple access have been actively studied for IoT networks. The UDN has been proposed as a promising and scalable solution to accommodate massive access of IoT devices to RANs through deploying ultra dense small cell BSs [4]. A low-power wide area network (LPWAN) radio technology standard, narrowband IoT, has been developed by the Third Generation Partnership Project (3GPP) for the IoT, which enables a wide range of IoT devices to be connected using cellular telecommunications bands with low power consumption [11]. Driven by the large available bandwidth provisioning, millime-

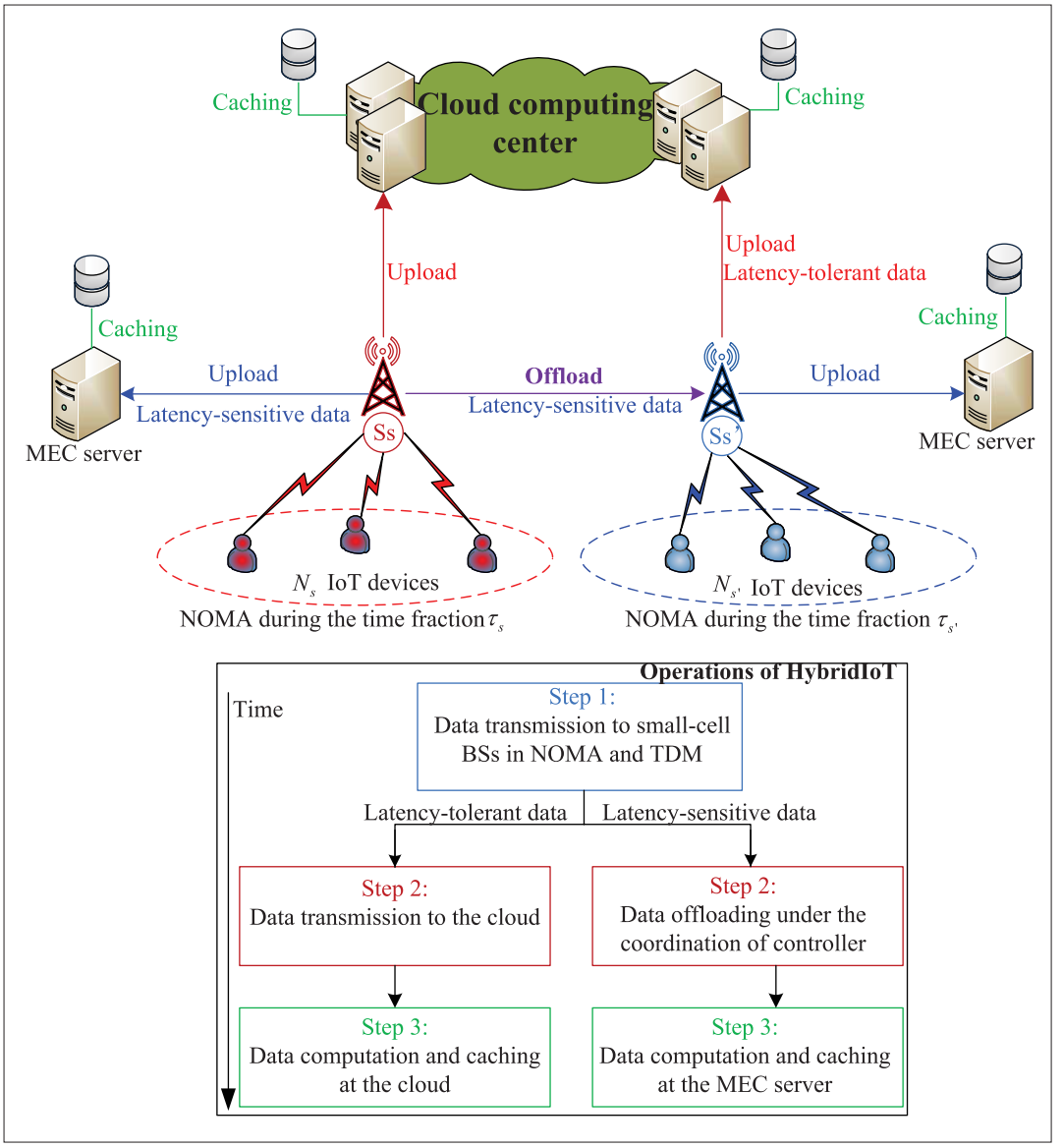

FIGURE 3. The details of implementing HybridIoT.

ter-wave communications [12] and massive multiple-input multiple-output communications [13] have been explored to support multi-gigabit data transmission in the Internet of Vehicles, which has been the most visible and familiar example of IoT. A full-duplex transmission framework has been proposed for IoT [14] because it can potentially double the spectral efficiency by performing simultaneous transmission and reception on the same frequency band.

Apart from the aforementioned communication technologies, other networking technologies may also be imperative for IoT. As a new type of network architecture, software-defined networking (SDN) has been explored to perform flexible resource allocation for better scalability and controllability in IoT [9] by introducing the separation of the data and control planes, and the direct programmability of network control with service virtualization. Cloud RAN (C-RAN) has been developed as another disruptive and emerging technology to facilitate IoT deployment by applying the centralization and virtualization of network functions to RANs [15].

HybridIoT Design

In this section, we propose an IoT framework that leverages UDN-based hierarchical multiple access and computation offloading between MEC and cloud to support the IoT-based smart city vision. We name this framework HybridIoT. The challenges, design, and implementation details of this framework are described as follows.

Challenges

Serving massive heterogeneous IoT devices for transmitting, computing, and caching is not trivial in the IoT-based smart city. The challenges mainly come from the following aspects.

The first one is the aspect of multiple access technology. Due to the scarcity of radio resource, IoT devices suffer severe co-channel and adjacent interference along with the denser and denser deployment of small cells in UDNs when the massive access of devices to IoT is realized through the spectrum multiplexing among IoT devices and small cells. To this end, heterogeneous access methods should be considered from the view-

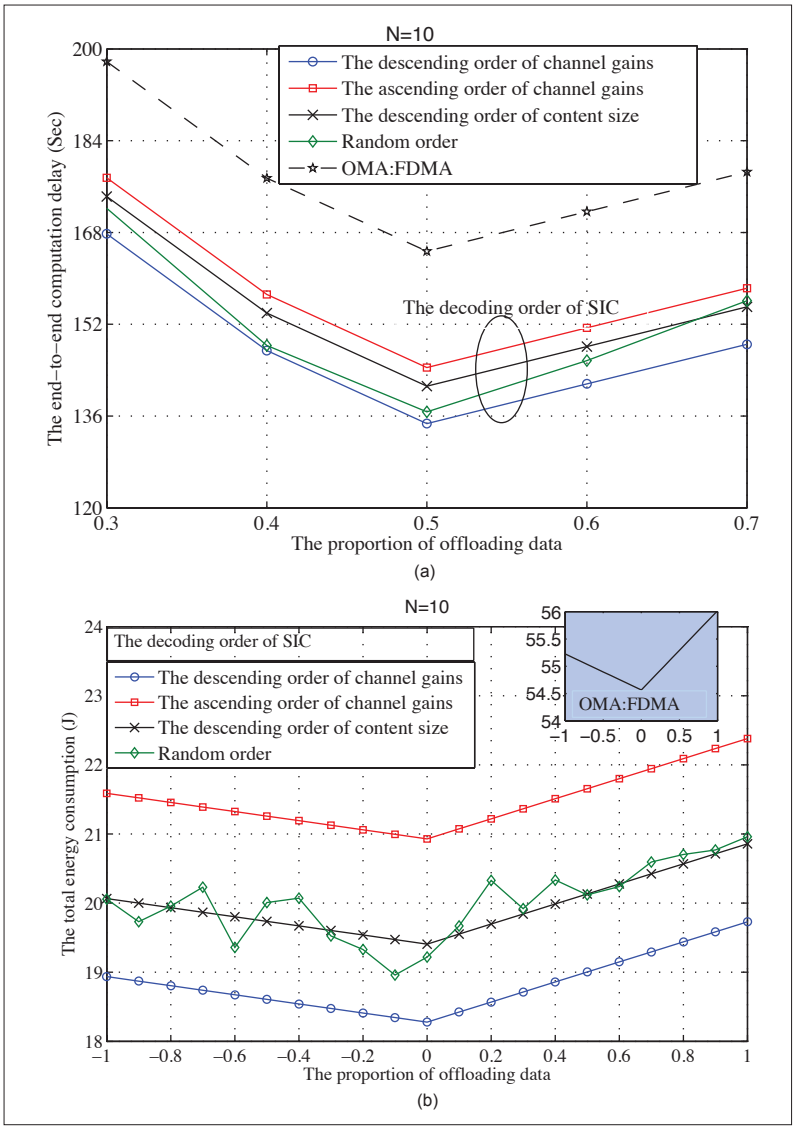

FIGURE 4. Performance evaluation of HybridIoT with different proportion of offloading between two small cell BSs: a) the end-to-end computation delay; b) the total energy consumption. Here, the negative proportion means the data is offloaded from small cell 1 to small cell 2, and vice versa.

point of IoT devices and small cells. The second one is the aspect of computing and caching. The IoT applications can be divided into two types in terms of latency sensitivity: latency-sensitive applications and latency-tolerant applications. Latency-sensitive applications include but are not limited to safety applications in an autonomous driving system, monitoring, free-viewpoint video service, and mobile augmented reality, which have strict latency requirements within the range of hundreds of milliseconds to even seconds. In order to reduce latency, one approach is to reduce the transmission delay through transmitting the data generated from the latency-sensitive applications to MEC servers at the network edge for computing and caching. The latency-tolerant applications mainly cover the non-real-time content delivery, such as web browsing, imaging, messaging, and file transfers. With the massive deployment of IoT devices, the latency-tolerant data generated from IoT devices is also high in volume. It may be inefficient for MEC servers to compute and cache these data with limited caching and computation resource. For this, these data can be forwarded by small cell BSs to the remote cloud with flexible and sufficient resource provisioning. Due to the imbalance of computing and caching capability between MECs and the cloud, we need to take into account computation and caching offloading between MECs and the cloud depending on the diversity and workload of IoT applications in each cell.

Therefore, in this article, we consider the physical layer issue and MAC layer issue together for IoT networks, and propose a novel IoT framework that integrates hierarchical multiple access as well as computation and caching offloading based on data retrieval in the UDN context. Based on the cooperation among MECs and the cloud, this integrated framework aims to minimize the end-to-end computation delay while meeting the individual requirements of various applications.

Framework Design of HybridIoT

The architecture of the proposed framework, referred to as HybridIoT, is shown in Fig. Fig. 1. In the framework of HybridIoT, we choose two types of multiple access methods: NOMA and time-division multiplexing (TDM), and equip every small cell BS with one MEC server and one data classifier. Also, all MEC-enabled small cell BSs are connected through one controller and to each other via fiber and switches.

All IoT devices in the coverage of one small cell BS send their sensed data to this BS in NOMA with power domain multiplexing. The key idea of the proposed NOMA scheme is to serve simultaneous transmissions of multiple IoT devices with distinct channel conditions via successive interference cancellation (SIC) at the small cell BS, as illustrated in Fig. 2. In particular, the small cell BS sequentially decodes the signal of an individual IoT device by treating the undecoded signals of IoT devices as interference according to the decoding order of SIC. Given the number of IoT devices, the sum data rate across IoT devices is fixed; however, the date rate of individual devices is determined by the decoding order of SIC. Therefore, in order to guarantee the worst transmission delay in each cell, it is critical to optimize the decoding order of SIC for all connected IoT devices. Furthermore, multiple small cell BSs receive the data generated by IoT devices in TDM mode, in which every time frame is adaptively allocated to these small cell BSs in an orthogonal manner.

The function of a data classifier is to categorize IoT data into latency-sensitive and latency-tolerant types according to the 6-bit DSCP field in the IP header of each data packet on the side of the small cell BS. If the DSCP decimal value is in the set of {8, 10, 14, 18, 22, 24, 28, 34, 36, 38}, the data is delay-sensitive, and it is delay-tolerant otherwise. The latency-sensitive data is then forwarded to the data buffer equipped in the MEC server for the following data computing and caching. On the contrary, the latency-tolerant data is temporarily stored in the data buffer equipped at the small cell BS, and then the small cell BS forwards these data to the cloud via the Internet. Considering the imbalance of workload as well as computing and caching resource provisioning between

small cells, all MEC servers can offload workload between each other via fiber and switches under the guidance of the controller. In other words, the controller provides some computation offloading services between MEC servers according to the amount of IoT data and the capability of each MEC server.

The proposal of HybridIoT aims to improve the design of IoT networks from the physical layer and MAC layer to meet various demands on low power consumption, high data rate, low latency, and massive connectivity. To further improve the performance of HybridIoT, more efforts should be spent on the design of the SIC ordering optimization algorithm, computation offloading mechanisms, efficient network resource management, and so on.

Implementation Details of HybridIoT

We consider an IoT topology with a set $\mathcal{S} = \{1, ..., S\}$ of MEC-enabled small cell BSs, each of which can serve a set $\mathcal{N}_s = \{1, ..., N_s\}$of IoT devices. All MEC-enabled small cell BSs operate in the TDM mode, and the time frame is divided into S time slots, each of which is allocated to one MEC-enabled small cell BS. Every IoT device sn has B_{sn}-bit data to be transmitted to the nearest MEC-enabled small cell BS in the NOMA mode. When the IoT device completes its data transmission, it will be removed from the set $\mathcal{N}_s$, and the remaining IoT devices in $\mathcal{N}_s$ keep transmitting their data in the NOMA mode. Each MEC-enabled small cell BS receives the IoT data in the classify-save-then-forward mode. In particular, the data classifier first categorizes the data and stores them in two individual data buffers during the data receiving. Once the MEC-enabled small cell BS s finishes receiving the data from all IoT devices in its coverage, it forwards the latency-tolerant data to the cloud at the average data rate of R_{sc} via the Internet; meanwhile, the equipped MEC server performs the computing and caching of latency-sensitive data. Considering the imbalance of resource provisioning at every MEC server, the controller coordinates the intensive latency-sensitive workload among all MEC servers for minimizing the network-centric computation cost. Specifically, the computation workload can be offloaded between the small cell BSs s and s' at the average date rate1 of $R_{ss'}$. The details of implementing HybridIoT are shown in Fig. 3.

Performance Analysis and Simulation Verification

In this section, we present some simulation results to show the efficiency of our proposed HybridIoT. The proposed framework can reduce the end-to-end computation delay and computation cost by jointly considering hierarchical multiple access and cooperation among MECs and the cloud.

Performance Analysis

In this article, we consider the end-to-end computation delay and computation cost as two performance metrics for the proposed HybridIoT:

- **End-to-end computation delay:** the total time needed when transmitting, computing, and caching all B_{sn}-bit data is completed

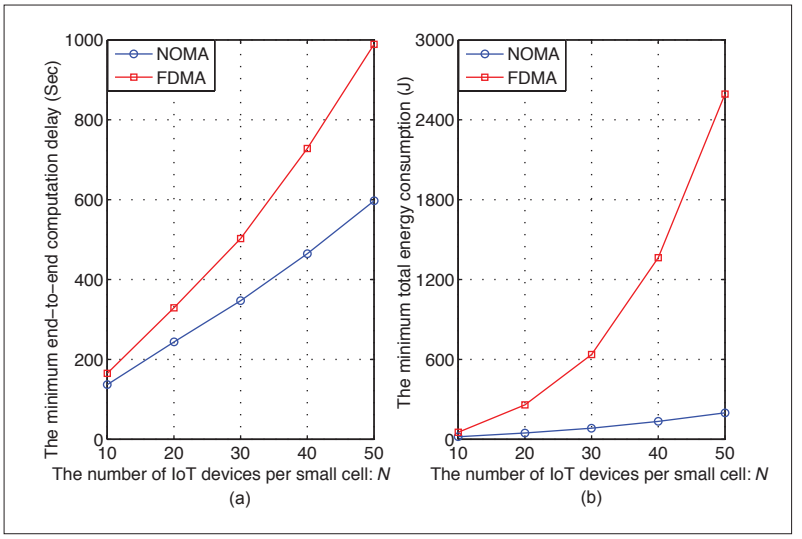

FIGURE 5. Performance evaluation of HybridIoT with different IoT device density when adopting data offloading between small cell BSs: a) the minimum end-to-end computation delay on average; b) the minimum total energy consumption on average.

- **Computation cost:** the total energy consumption during the transmitting, computing, and caching of all B_{sn}-bit data

In particular, the end-to-end computation delay can be evaluated based on the following five kinds of delay:

•The delay of transmitting the data generated by all IoT devices in the convergence of MEC-enabled small cell BS s to this BS, which is a function of the decoding order of SIC in NOMA and the time allocation in TDM

•The delay of transmitting the latency-tolerant data from small cell BS s to the cloud, which is determined by the size of latency-tolerant data and the date rate from the small cell BS s to the cloud

•The delay of computing the latency-tolerant data from small cell BS s at the cloud, which is determined by the computation resources allocated to small cell BS s

•The delay of offloading the workload between any two small cell BSs s and s', which is determined by the size of offloaded data and the transmission date rate between these two small cell BSs

•The delay of computing the latency-sensitive data at the small cell BS s, which is determined by the computation resources equipped in small cell BS s

The computation cost can be evaluated based on the following seven kinds of energy consumption:

•The energy consumption of transmitting the data generated by IoT device sn to MEC-enabled small cell BS s, which is a product of the transmit power, the data size, and the data rate emitted by IoT device sn

•The energy consumption of transmitting the latency-tolerant data from small cell BS s to the cloud, which is a product of the transmit power used by small cell BS s to transmit the latency-tolerant data to the cloud, and the corresponding transmission delay

•The energy consumption of computing the latency-tolerant data from small cell BS s at the cloud, which is a product of the latency-tolerant data size and the computation energy efficiency (Joules per bit) of the cloud allocated to computing these data

[1] The average date rate $R_{ss'}$ is dependent on the network topology connecting all small cell BSs, the communication mode between any two small cell BSs, and the radio resource allocation policy designed for all small cell BSs.

•The energy consumption of caching the latency-tolerant data from small cell BS s at the cloud, which is a product of the latency-tolerant data size, the caching energy efficiency (Joules per bit) of the cloud allocated to caching these data, and the computation ratio of the cloud

•The energy consumption of offloading the workload between any two small cell BSs s and s', which is a product of the offloading delay and the transmit power used by small cell BS s to offload the latency-tolerant data to other small cell BSs

•The energy consumption of computing the latency-sensitive data at small cell BS s, which is a product of the latency-sensitive data size and the computation energy efficiency of the MEC equipped at small cell BS s

•The energy consumption of caching the latency-sensitive data at small cell BS s, which is a product of the latency-sensitive data size, the caching energy efficiency of the MEC equipped at small-cell BS s, and the computation ratio of this MEC

In the evaluation of computation delay and computation cost, it is worth noting that they can be optimized by adjusting the factors including the decoding order of SIC in NOMA, the time allocation in TDM, the transmit power of IoT devices, and the proportion of data offloaded between any two small cell BSs.

Simulation Verification

In the simulations, we consider an LTE-A cellular network with two MEC-enabled small cell BSs, each of which serves N randomly distributed IoT devices. There is a computing cloud center connected with these two small-cell BSs via the Internet, which provides the computing and caching capabilities in the cloud. The computation energy efficiency, caching energy efficiency, and computation capacity allocated to compute the latency-tolerant data from each small cell BS are 2.5×10^{-9} W/bit, 10^{-9} W/bit, and 2 Mb/s in the cloud, respectively. Based on the LTE-A specification, 10 MHz spectrum is used in every small cell, and the noise power spectral density is set to be −140 dBm/Hz. We consider the path loss model $40\log_{10}(d)$ dB, where d means the distance between the IoT device and the connected small cell BS. The traffic volume of IoT device n associated with small cell BS s is set to be B_{sn} Mbits, where B_{sn} is uniformly distributed in [10, 20]. We set the proportion of latency-tolerant data that these two small cell BSs forward to the cloud to be 0.3 and 0.2, respectively. We set the data rate R_{sc} to be 100 Mb/ps, and the offloading data rate to be 10 Mb/s between these two small cell BSs. We set the transmit power of small cell BS and the transmit power of IoT device to be 1 W and 0.1 W, respectively. We set the computation ratio[2] of cloud and MEC servers to be 0.3 and 0.6, respectively. Also, the computation energy efficiency, caching energy efficiency, computation capacity are respectively set to be $(6, 3) \times 10^{-9}$ W/bit, $(2, 4) \times 10^{-9}$ W/bit, and(1, 0.5) Mb/s at the two MEC servers.

Figure 4 shows the performance of NOMA with different decoding orders of SIC and frequency-division multiple access (FDMA) applied to HybridIoT. We can see that the performance of NOMA is always better than that of FDMA regardless of the decoding order of SIC adopted. For NOMA, when the decoding order of SIC follows the descending order of channel gains (i.e., the IoT device with the worse channel condition suffers less co-channel interference), the best performance is achieved. Comparing Fig. 4a with Fig. 4b, we see that the optimal proportion of offloading data is totally different for minimizing the end-to-end computation delay and the total energy consumption. This implies that the method of offloading data between MEC servers must be consistent with the specific performance metric.

Figure 5 shows the performance of NOMA and FDMA with optimal data offloading between two small cell BSs under different IoT device density (i.e., with varying N IoT devices per small cell). We can see that although the performance of NOMA and FDMA are both increasing with the increase of the number of IoT devices, the performance of NOMA is always far superior to that of FDMA. This is because our proposed HybridIoT integrates the allocation of radio resource and computation resource. Since HybridIoT can provide a low-latency and low-energy-consumption experience for IoT, it can enable heterogeneous IoTs.

Conclusion and Future Work

In this article, we review recent advances in transmitting, computing, and caching oriented to the IoT. We propose a new architecture, HybridIoT, in order to efficiently transmit, compute, and cache big data generated from the massive distributed IoT devices deployed in a smart city. The proposed HybridIoT integrates UDN-based hierarchical multiple access and computation offloading between MEC and cloud, and it can substantially reduce the end-to-end computation delay and the total energy consumption for IoT compared to traditional IoT architectures.

This article provides the initial step toward multi-access mobile edge computing for IoT-based smart cities. Nevertheless, there are still many open issues that deserve in-depth investigation on the design of IoT.

Resource Allocation: There are three kinds of important resource in heterogeneous IoT: radio resource, computation resource, and caching resource. In order to meet various demands on ultra-low latency, high reliability, massive connectivity, and high data volume during transmitting, computing, and caching big IoT data, adaptive radio resource allocation solutions should be developed for IoT-based smart cities. For example, when the UDN paradigm is applied to IoT, the radio resource, such as transmit power, time slots, and frequency bands, should be properly allocated among small cells and IoT devices in accordance with channel conditions, data volume per small cell, and data volume per IoT device. From the operators' perspective, the appropriate computation-caching resource should be assigned to every MEC server and the cloud subject to the data volume to be processed and delay requirements.

Multiple Access: To serve massive IoT devices with limited radio resource and perform high-data-rate computation offloading among MEC servers, it is inevitable to explore new promising multiple access schemes. With the ultra dense deployment of small cells and IoT devices, we should keep

[2] In this article, the computation ratio means the data compression ratio after computing the data.

track of the co-channel interference mitigation when designing the multiple access scheme for the access of IoT devices to small cell BSs. On the other hand, in order to efficiently deliver data and commands among the controller and MEC servers, a practicable networking architecture as well as corresponding multiple access schemes should be carried out. After designing an efficient multiple access scheme, we need to analyze the average data rate from IoT devices to small cell BSs and that between any two small cell BSs, which will be important for evaluating the end-to-end computation delay of IoT.

Deployment of Small Cells: With the denser and denser deployment of small cells, it seems that massive IoT devices are more likely to be served. However, the radio resource limitation results in more severe co-channel interference, and thus over densely deploying small cells degrades the system performance of IoT. In practice, it is imperative to strike a balance between the deployment density of small cells and the various demands of IoT.

Cell Association: Due to the diversity of computation and caching resource among MEC-enabled small cell BSs, the existing cell association schemes that account for channel conditions, data volume, and transmit power may be highly suboptimal. New cell association schemes are therefore needed that associate IoT devices with proper small cell BSs by taking into account the computation and caching resource provisioning at each MEC server and individual service requirements in addition to data volume and channel conditions.

References

[1] X. Lyu *et al.*, "Selective Offloading in Mobile Edge Computing for the Green Internet of Things," *IEEE Network*, vol. 32, no. 1, Jan./Feb. 2018, pp. 54–60.

[2] L. Zhou *et al.*, "Greening the Smart Cities: Energy-Efficient Massive Content Delivery via D2D Communications," *IEEE Trans. Industrial Informatics*, vol. 14, no. 4, Apr. 2018, pp. 1626–34.

[3] Y. Mehmood *et al.*, "Internet-of-Things-Based Smart Cities: Recent Advances and Challenges," *IEEE Commun. Mag.*, vol. 55, no. 9, Sept. 2017, pp. 16–24.

[4] L. P. Qian *et al.*, "Joint Uplink Base Station Association and Power Control for Small-Cell Networks With Non-Orthogonal Multiple Access," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, Sept. 2017, pp. 5567–82.

[5] K. Zhang *et al.*, "Mobile Edge Computing for Vehicular Networks: A Promising Network Paradigm with Predictive Offloading," *IEEE Vehic. Tech. Mag.*, vol. 12, no. 2, June 2017, pp. 36–44.

[6] H. Guo, J. Liu, and H. Qin, "Collaborative Mobile Edge Computation Offloading for IoT over Fiber-Wireless Networks," *IEEE Network*, vol. 32, no. 1, Jan./Feb. 2018, pp. 66–71.

[7] L. Song *et al.*, "Resource Management in Non-Orthogonal Multiple Access Networks for 5G and Beyond," *IEEE Network*, vol. 31, no. 4, July/Aug. 2017, pp. 8–14.

[8] G. Pallis, "Cloud Computing: The New Frontier of Internet Computing," *IEEE Internet Computing*, vol. 14, no. 5, Sept. 2010, pp. 70–73.

[9] R. Huo *et al.*, "Software Defined Networking, Caching, and Computing for Green Wireless Networks," *IEEE Commun. Mag.*, vol. 54, no. 11, Nov. 2016, pp. 185–93.

[10] J. Choi and N. Y. Yu, "Compressive Channel Division Multiple Access for MTC Under Frequency-Selective Fading," *IEEE Trans. Commun.*, vol. 65, no. 6, June 2017, pp. 2715–25.

[11] J. Chen *et al.*, "Narrowband Internet of Things: Implementations and Applications," *IEEE Internet of Things J.*, vol. 4, no. 6, Dec. 2017, pp. 2309–14.

[12] L. Kong *et al.*, "Millimeter-Wave Wireless Communications for IoT-Cloud Supported Autonomous Vehicles: Overview, Design, and Challenges," *IEEE Commun. Mag.*, vol. 55, no. 1, Jan. 2017, pp. 62–68.

[13] L. Liu *et al.*, "Convergence Analysis and Assurance for Gaussian Message Passing Iterative Detector in Massive MU-MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, Sept. 2016, pp. 6487–501.

[14] M. Heino *et al.*, "Recent Advances in Antenna Design and Interference Cancellation Algorithms for In-Band Full-Duplex Relays," *IEEE Commun. Mag.*, vol. 53, no. 5, May 2015, pp. 91–101.

[15] Y. D. Beyener *et al.*, "NB-IoT Technology Overview and Experience from Cloud-RAN Implementation," *IEEE Wireless Commun.*, vol. 24, no. 3, June 2017, pp. 26–32.

Biographies

Li Ping Qian [SM] (lpqian@zjut.edu.cn) received her Ph.D. degree in information engineering from the Chinese University of Hong Kong in 2010. She is currently a professor with the College of Information Engineering, Zhejiang University of Technology, China. Her research interests include wireless communication and networking, resource management in wireless networks, massive IoTs, mobile edge computing, emerging multiple access techniques, and machine learning oriented toward wireless communications. She was a co-recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2011.

Yuan Wu [SM] (iewuy@zjut.edu.cn) received his Ph.D. degree in electronic and computer engineering from Hong Kong University of Science and Technology in 2010. He is currently a professor in the College of Information Engineering, Zhejiang University of Technology. His research interests focus on radio resource allocations for wireless communications and networks, and smart grid. He is the recipient of the Best Paper Award from IEEE ICC '2016.

Bo Ji [SM] (boji@temple.edu) received his Ph.D. degree in electrical and computer engineering from The Ohio State University, Columbus, in 2012. He joined the Department of Computer and Information Sciences at Temple University in July 2014, where he is currently an assistant professor. He is also a faculty member of the Center for Networked Computing (CNC) at Temple University. His research interests are in the modeling, analysis, control, optimization, and learning of computer and networking systems, such as communication networks, information-update systems, cloud/data center networks, and cyber-physical systems. He is a National Science Foundation CAREER awardee (2017) and an NSF CISE Research Initiation Initiative (CRII) awardee (2017).

Liang Huang [M] (lianghuang@zjut.edu.cn) received his Ph.D. degree in information engineering from the Chinese University of Hong Kong in 2013. He is currently an assistant professor with the College of Information Engineering, Zhejiang University of Technology. His current research focuses on queueing and scheduling in communication systems and networks.

Danny H. K. Tsang [F] (eetsang@ece.ust.hk) received his Ph.D. degree in electrical engineering from the University of Pennsylvania in 1989. He joined Hong Kong University of Science and Technology in 1992 where he is now a professor. He has been a Guest Editor of the *IEEE Journal of Selected Areas in Communications*, an Associate Editor of the *Journal of Optical Networking*, and a Guest Editor of the *IEEE Systems Journal* and *IEEE Network*. His research interests include cloud computing, cognitive radio networks, and smart grids. He is currently a Technical Editor for *IEEE Communications Magazine*. He is an HKIE Fellow.