IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 10, 2019, accepted May 31, 2019, date of publication June 12, 2019, date of current version July 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2922362

Dynamic Offloading for Energy Harvesting Mobile Edge Computing: Architecture, Case Studies, and Future Directions

BIN LI[1,2,3], ZESONG FEI[2,3], (Senior Member, IEEE), JIAN SHEN[1], (Member, IEEE), XIAO JIANG[4], AND XIAOXIONG ZHONG[5], (Member, IEEE)

[1]School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2]Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology, Nanjing 210044, China
[3]School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
[4]Shandong Honghorn Information Technology Co., Ltd., Yantai 264003, China
[5]Peng Cheng Laboratory, Shenzhen 518000, China

Corresponding author: Zesong Fei (feizesong@bit.edu.cn)

This work was supported in part by the Beijing Natural Science Foundation under Grant L182038, in part by the National Natural Science Foundation of China under Grant 61871032, Grant 61672295, and Grant 61802221, in part by the Shandong Provincial Natural Science Foundation under Grant ZR2019PF018, in part by the Chinese Ministry of Education-China Mobile Communication Corporation Research Fund under Grant MCM20170101, in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) fund, and in part by the joint Funds of the National Natural Science Foundation of China under Grant U1836115.

ABSTRACT Mobile edge computing (MEC) is envisioned as a new paradigm by integrating the mobile computing functionality into 5G wireless networks, aiming at empowering communication networks with low-latency services. In general, mobile devices have finite battery lifetime (e.g., machine-type devices) and the energy harvesting is advocated to provide perpetual energy supply for achieving sustainable operation, which is very important for facilitating sustainable computing in future applications. In this paper, we propose a wireless powered MEC network architecture that employs device-to-device (D2D) communications underlaying heterogeneous networks (HetNets) to enable the computational tasks offloading to resource-rich edge servers. A dynamic offloading decision is made to execute the computation tasks. Then, we focus on the energy-efficient offloading scheme, and joint offloading and user association scheme. From the illustrative results, we provide insights for the design of this new network architecture. Furthermore, several open research topics are discussed.

INDEX TERMS Mobile edge computing, energy harvesting, device-to-device (D2D) communication, computation offloading.

I. INTRODUCTION

With the proliferation of intelligent terminals and the emergence of modern wireless applications, the demands for mobile data traffic are dramatically increasing. According to the latest release from Cisco [1], mobile data is expected to grow 7-fold from 2016 to 2021. To alleviate traffic hot zones, deploying a large number of macro base stations (MBSs) becomes very expensive for operators. For this reason, a cost-effective network architecture is acclaimed that exploits device-to-device (D2D) communications and the deployment of small cells to coexist the macro cellular networks [2]–[4]. This improves network throughput by offloading mobile traffic from MBS.

Due to the limited processing power and storage capacity of mobile devices, running computation-intensive applications for resource-poor devices in upcoming 5G Internet of Things (IoTs) cannot be completed in real-time. Consequently, the tension between resource-hungry applications and resource-constrained devices arises a significant design challenge. It is advocated that the computation-intensive tasks can be offloaded to a centralized cloud, such a paradigm is generally known as mobile cloud computing [5], [6]. However, there exists an inherent limitation in cloud computing, namely, the long propagation distance from the end user to the remote cloud center. This may fail to catch up the requirements of delay sensitive applications (e.g., speech/face recognition, immersive gaming and augmented reality (AR) application) because of the high communication latency and possible congestions between mobile devices and the

The associate editor coordinating the review of this manuscript and approving it for publication was Zheng Chang.

cloud. [7], [8]. Nowadays, mobile edge computing (MEC) architecture could potentially overcome the drawbacks of this approach, pushing the service provisioning closer to network edge. The novel concept was introduced by the European Telecommunications Standards Institute (ETSI) as a means of extending cloud computing capabilities to the edge of network along with higher processing and storage capabilities. By leveraging the radio access networks, MEC has the advantages of achieving lower latency, proving flexible computation experience, and reducing energy cost, making it easier for both application developers and content providers to access network services. As a result, MEC is commonly agreed as a key technology to realize next-generation wireless networks [9]–[12]. At the same time, the related technical standards are being driven by prominent mobile operators (e.g., Vodafone, NTT docomo) and manufactures (e.g., IBM, Intel, Huawei). Several platforms provided by the leading companies are being employed as an immediate edge computing platform and gateway for specific types of smart domains, such as Google NEST, Intel IoT platform, Apple Homekit, IBM Node-RED Bluemix, Microsoft nitrogen.io. A laptop, and CISCO Fog Computing IOx [13], [14].

Although computation offloading is effective in exploiting the powerful computational resources at edge cloud server, for conventional battery-powered devices with finite battery capacities, the computation performance may be compromised due to insufficient battery energy for task offloading, i.e., mobile applications will be terminated and mobile devices will be out of service when the battery energy runs out. This can possibly be overcame by using larger batteries or recharging the batteries regularly. Nonetheless, using the batteries with larger capacity at mobile devices imply more hardware cost, which is not desirable. On the other hand, recharging batteries frequently may be impossible in certain application scenarios where the nodes are typically hard-to-reach [15]. To address this challenge and capture green computing, energy harvesting technology has been widely regarded as a viable and environment-friendly solution to offer power supply. It is common that renewable energy sources from solar panels and wind mills are weather-dependent, often suffering from uncertainties caused by random energy arrivals [16]. On the contrary, energy harvesting through capturing the radio-frequency (RF) electromagnetic signals is more controllable and stable, wherein the dedicated RF energy transmitters are employed to continuously recharge the battery of remote devices for realizing the capability of self-maintenance [17], [18]. Motivated by these developments, the incorporation of energy harvesting and MEC fits into a new paradigm for low-power mobile terminals to achieve sustainable and enhanced computational capabilities. Some recent works have showed the high potentials of energy harvesting MEC systems [19]–[22]. However, these works only concentrate on single BS and they fail to catch up the unique characters of 5G with D2D communications in their scheme design.

In this paper, we propose an innovative energy harvesting MEC architecture in a D2D-enabled heterogeneous network (HetNet) for solving the limited computing capability and maintaining the operation energy on mobile devices. Under this framework, mobile devices are very flexible to choose multiple patterns for task executions including local mobile execution, D2D computation offloading, small base stations (SBSs) computation offloading, and remote computation offloading at the MEC server provided by MBS. To be specific, in the following section, we review the state-of-the-art works, and then we present the proposed energy harvesting MEC architecture, along with the key technologies of D2D communications, energy harvesting and dynamic offloading. In the ensuing two sections, we focus on energy-efficient offloading scheme as well as joint offloading and user association scheme. After that, some open problems are explored, and finally we conclude this article.

II. RELATED WORKS

To date, there have been a number of research efforts to seek renewable energy sources (e.g. solar radiation, wind energy and human motion energy) to power mobile devices [23]–[26] or edge servers in outdoor scenarios [27], [28]. Nevertheless, the high intermittency and unpredictability of renewable green energy significantly exacerbates the challenge of the satisfactory computation performance.

Alternatively, some existing solutions on wireless powered MEC systems have also been proposed to exploit the ambient RF signals to supply the mobile devices. For instance, in [19], a wireless powered single-user MEC system was considered where the user harvested RF energy from a dedicated access point (AP) for computation offloading, and the CPU frequency for each required CPU cycle was optimized. In [20], the authors considered a multi-antenna AP delivering RF energy to multiple users, where the computing tasks were jointly executed by the AP and users via optimizing transmit energy beamforming, offloading decision and resource allocation with the minimum of the AP's energy consumption. The authors of [22] designed a new time frame in a binary computation offloading, in which an AP first broadcasted the RF energy in the downlink and then the energy-constrained mobile devices offloaded their tasks to the AP at their allocated time slots. With energy powered by the wireless RF signal component, the fair energy efficiency framework in a multi-user MEC system was proposed in [29] where full-duplex was employed at AP to support energy delivery and computation offloading simultaneously. The work in [30] focused on a wireless powered multi-user MEC system where a multi-antenna AP and an MEC server were separately placed. The aim was to minimize the total energy cost of all mobile devices with joint optimization of the offloading decisions, time switch, local computation/offloading powers. In order to eliminate the double-near-far effect in a two-user wireless powered MEC system, the end-user device closer to AP was selected as a relay to help offload the far-away end-user device's computation tasks to the edge cloud. In this

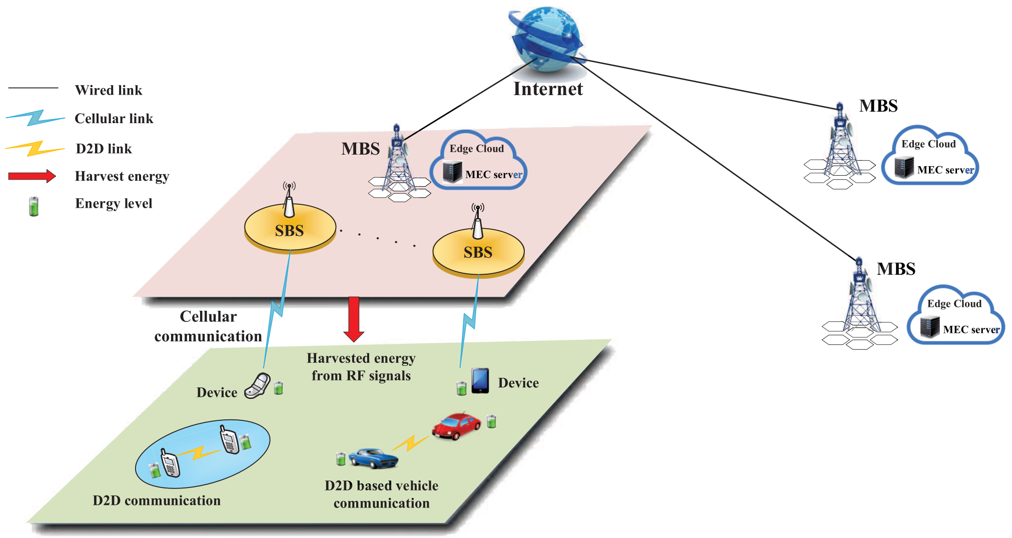


FIGURE 1. Illustration of energy harvesting MEC architecture in HetNet. On one hand, MBS and SBSs acting as RF energy sources provide energy for battery-powered mobile devices via downlink signals. On the other hand, mobile devices adopt the harvested energy to execute their computation tasks by dynamic offloading.

circumstance, [21] paid more attention to minimize the total transmit energy of the AP and [15] concentrated on the maximization of energy efficiency under the constraints of the computational tasks. Considering a hybrid MEC system consisting of a multiple-antenna cellular BS and a WiFi AP, the authors of [31] studied the problem of maximizing the total energy saving of all mobile devices by jointly optimizing the computation and radio resources along with dynamic interface selection. Recently, the study on RF-based energy harvesting was applied to an unmanned aerial vehicle-enabled multi-user MEC system [32], in which both binary and partial computation offloading modes were respectively taken into account for the computation rate maximization problems.

The aforementioned works are only view of single AP/BS system where mobile devices with energy harvesting functionalities can execute the task locally or offload it to the MEC server installed on the AP/BS. Note that the roles of heterogeneous SBSs in small cell network are not the same due to their different locations deployed and hardware equipped. Furthermore, the data transmission between mobile devices and MEC server is based on either the cellular channel or D2D communications with high efficiency and low latency. These essential characters are seldom considered, which motivates our work.

III. ENERGY HARVESTING MOBILE EDGE COMPUTING ARCHITECTURE

Owing to the physical size and battery power constraints, mobile devices with limited lifetime and low computing processor are not able to smoothly support high-performance computations applications within our expectation. This obstacle consequently drives the development of a new network architecture, i.e., wireless powered MEC, adequately integrating the MEC and the energy harvesting. An illustration of the proposed architecture is shown in Fig. 1, the deployed SBSs and the resource-constrained devices are located in the coverage range of an existing MBS. The MEC server is installed at the MBS which is provisioned by service providers to offer computation offloading services. The mobile devices in proximity can establish a direct D2D communication link and bypass the serving BS, which leverage nearby users' vacant computing resources. With the emphasis on the co-located MBS and MEC server, they not only act as energy transmitter to transfer RF power but also as information receiver to receive the offloaded tasks from mobile devices. In this system, mobile devices are equipped with energy harvesting capabilities which can be powered by downlink RF signals, and then rely on the harvested energy for local computing or uplink computation offloading.

A. COMMUNICATION MODEL

The proposed framework conducts both the cellular and D2D connections for wireless access. In the cellular communication, each good-quality device can establish a cellular link with the associated BS (which can be an MBS or an SBS) via radio access technologies. In the type of potential D2D communication, a device with low service quality (*service requester*) can establish a D2D link with peer device in proximity with better service quality (*service provider*). Here, a D2D device can either be a D2D requester or a D2D provider. Specifically, the potential D2D devices can operate in the following three resource allocation modes [33]:

- *Non-Orthogonal Sharing Mode:* The D2D devices directly transmit data to each other by allowing to reuse part of the spectrum resources occupied by cellular links, which incur co-channel interference.

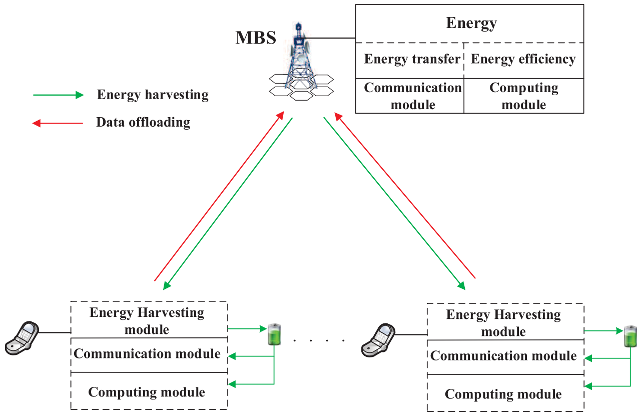


FIGURE 2. An example of energy harvesting MEC, where the energy transfer operates in the downlink and the computation offloading is in the uplink. The green color portion represents the energy flow and the red color portion represents the offloaded data flow.

- *Orthogonal Sharing Mode:* The D2D devices directly transmit data to each other by allocating a dedicated spectrum to avoid suffering interference from cellular links at the cost of reduced spectrum utilization.
- *Regular Cellular Mode:* The D2D devices relay their data through the associated BS via radio access technologies in the same way as cellular devices, which in general consume the most channel resources.

B. ENERGY HARVESTING MODEL

A detailed example of the energy harvesting from MBS is presented in Fig. 2, where the MBS comprises an energy transfer module for broadcasting RF energy to multiple user devices, a communication module for data offloading, and a computing module for remotely performing computation tasks. Each mobile device consisting of an energy harvesting module receives the energy from MBS and SBSs for replenishing the rechargeable battery, a communication module for uplink computation offloading, and a computing module for locally performing computation task. For the case of energy harvesting, assuming the battery storage of mobile device is sufficiently large such that battery-overcharging is negligible, it can store part of the newly harvested energy at each operation block. The harvested energy can not only be used for D2D communication and cellular communication, but also is very useful for low-power mobile devices to achieve self-sustainable mobile computing. To guarantee sustainable operation, the energy used for offloading data must not exceed the energy available in the battery, wherein the initial energy of each mobile device is sufficient in the very beginning for smooth communication. The energy harvesting process from SBSs is similar to that of MBS. For each device, the harvested energy is also known as its received power. By utilizing the receive power, mobile devices can perform their tasks for local computing or data offloading. In each time block with duration T, mobile device i scavenges energy from RF signals transmitted by the collection of candidate BSs Ω, the energy E_i harvested (i.e., receive power) by mobile device i from the BSs' RF signals can be given as:

$$E_i = \eta \left(P_M h_{i,M} + \sum_{j \in \Omega \backslash M} P_j^S h_{i,j} \right), \tag{1}$$

where $\eta \in (0, 1]$ is mobile device's energy conversion efficiency, $h_{i,M}$ and $h_{i,j}$ are the effective channel gain from MBS and SBS j to mobile device i, and P_M and P_j^S are the RF energy transmit power of MBS and SBS j, respectively. From the above equation, we can observe that the harvested energy (i.e., receive power) is highly related to the transmit power of BSs.

Path loss imposes a limit on the received signal power, thereby the distance between mobile device and BS (i.e., RF energy source) cannot be too long since the amount of received RF energy is required to be large enough to activate the harvesting circuit. It should be noted that mobile device i harvests energy from RF signals transmitted by the SBSs and MBS. Denote the small-scale fading channel gain from BS l ($l \in \Omega$) to the i-th mobile device as $\hat{h}_{i,l}$, thus the channel gain from BS l to the i-th mobile device can be represented by

$$h_{i,l} = \beta \hat{h}_{i,l} \left(\frac{d_0}{d_{il}} \right)^{\alpha}, \tag{2}$$

where β is the path-loss constant, α is the path-loss exponent, d_0 is the reference distance, and d_{il} is the distance from BS l to the i-th mobile device. In general, the transfer distance in small cells is short and the transmission distance in macro cell is very long. Therefore, the received power from MBS located far away from device i may be too low to be detected and unable to be harvested by the device. On the basis of this, the wireless energy harvesting from MBS may be not considered due to large transmission distance in macro cell.

C. DYNAMIC OFFLOADING

With the harvested energy, each mobile device has a computation task which can be executed either locally on itself (by the embedded micro-processor) or remotely on the *offloading destinations* via wireless access. Since a mobile device may not be able to complete a heavy computation task due to the energy- and size-constrained computing processor, it is necessary to advocate a *dynamic offloading* decision mode according to individual device's communication link condition and task requirements, such that each mobile device can flexibly choose the computation and communication services. Specifically, in this article we introduce three different offloading modes for processing the computation task at the proximity device via D2D communication, via the SBSs, or via the MEC server deployed inside the MBS, and this mode selection mainly depends on the available radio access and computing resources at these *destinations*. As shown in Fig. 3, mobile devices can perform the computation tasks through four ways, i.e., *local computing*, *D2D computation offloading*, *SBS computation offloading*, and *MBS computation offloading*. Considering mobile devices adopt a binary

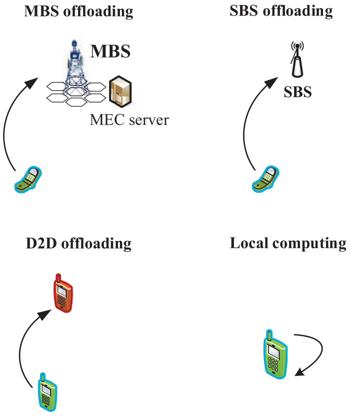


FIGURE 3. Illustration of the computation offloading, including MBS computation offloading, SBS computation offloading, D2D computation offloading and local computing. In particular, the D2D-assisted computation offloading can play a complementary role in MEC due to physical proximity, where a mobile device can directly offload its task to the nearby device with both high-computing capability and good-quality connectivity for facilitating the task execution.

computation offloading rule, they need to determine whether to be processed locally at themselves or executed remotely at one of the destinations. Note that each mobile device chooses at most one offloading mode for executing its computation task due to the delay constraint. After the computing is accomplished, the computation results will be get back to the mobile devices.

D. APPLICATIONS

There are a number of applications that can support the edge cloud service. The typical example is AR applications. Generally, AR applications rely on computationally intensive computer vision algorithms with extreme latency requirements, however, JavaScript is inadequate to provide efficient computational capability for complex matrix computations. As a result, pure front-end solutions can only use several typical algorithms to recognize some simple markers, and have weak recognition as well as tracking ability for real material objects. In such a case, the local computing operations are often restricted by the inherent hardware capabilities regarding CPU, GPU, memory, and battery. To cope with this issue, MEC can be effectively used to enhance the computing ability of mobile devices by leveraging the rich resources. In this way, mobile devices can enjoy various benefits from MEC such as latency reduction and energy savings.

In addition, we exploit a vehicular crowdsensing application to illustrate the vehicular edge cloud service. For instance, a vehicular crowdsensing application requires a crowd of vehicles to periodically sense data from immediate surroundings, in-situ process them such as data fusion and image/video processing, and transmit the processed results to the centralized application manager for post-processing. In this case, the recruited vehicles can conduct the pre-processing tasks locally and then upload the results, or can offload the sensed data to process in the edge cloud via the application manager control [34].

IV. ENERGY-EFFICIENT COMPUTATION OFFLOADING

Since sophisticated applications may require computation processing with high performance, energy-efficient computation offloading deserves much consideration for mobile devices to speedup computing and save energy. In nature, it is essential to discuss how to select an appropriate offloading service mode while achieving satisfactory user experience. The downlink energy harvesting and the uplink computation offloading can be performed over orthogonal frequency bands simultaneously, and the time for downloading the computation results from the destinations is neglected since the size of the results is much smaller than that of the input data. Adopting time division multiple access protocol to coordinate computation offloading, the mobile devices offload their individual tasks to the destinations over orthogonal time slots. Let $h_{i,M}$ and $h_{i,j}$ denote the channel gains between device i and MBS, as well as that between device i and SBS$_j$, E_i represents the amount of energy harvested by device i over specific time block T which mainly depends on the energy conversion efficiency, the RF energy signals of BSs (MBS and SBSs) and the distance from device i to BSs. It is noted that the harvested energy is proportional to the total received power.

We use a three-field notation $\mathcal{A}_i = (D_i, C_i, T_i)$ to characterize the task of device i, where the first notation stands for the input-data size of the task, the second notation is the number of CPU cycles for accomplishing this task and the last notation denotes the the maximum tolerable latency experienced by this task. In this section, we introduce the task execution model such that each device has flexible options for processing its task as follows.

- *Local Computing:* When task $\mathcal{A}_i$ is executed by local execution, the overhead of local computing contains two parts: the execution latency t_i^l and energy consumption for computing $E_{\text{loc},i}$. The execution latency is jointly determined by the total CPU cycles required for accomplishing local computing and the computing capability of device i (i.e., CPU cycles per unit time); the energy consumption for local computing is contributed by the number of CPU cycles for accomplishing this task and the energy consumption per CPU cycle of device i.
- *D2D Offloaded Execution:* If the mobile device has poor cellular connection, in this case it can choose to offload its computation task to a nearby device having large amount of idle CPU resource currently to facilitate the task execution. The short-range communications offered by D2D links reduce the energy consumption of data transmission. In such case, the processing time t_i^D of the offloaded task includes not only the transmission latency for sending data to D2D receiver but also the computation execution latency on the D2D receiver, and the corresponding energy consumption $E_{\text{off},i}^D$ is the sum of the energy used for communicating and the energy cost used for the computation.
- *SBS Offloaded Execution:* The SBSs have relatively low computing capabilities serving a small number of devices to a certain extent. If mobile device has good

TABLE 1. Qualitative comparison of offloading modes.

Offloading modes	Advantages	Disadvantages
MBS offloading	Highest computing performance and largest coverage	Highest communication cost
SBS offloading	Enhance the coverage of edge users	Relatively low computing capability over MBS
D2D offloading	Infrastructure independency and fast communication	Strict requirement on the computation duration

cellular connection, the computation task can be delivered to the associated SBS via wireless local network. Then the SBS will execute the computation task on behalf of the mobile device. In such case, the processing time t_i^S is composed of task transmission latency and the computation execution latency on SBS, and the corresponding energy consumption $E_{\mathrm{off},i}^S$ is the sum of the energy used for communicating data to the SBS and the energy cost used for the computation.

- *MBS Offloaded Execution:* The MEC server has a powerful capability to execute multiple computation tasks. For the MEC server offloading approach, the mobile devices are allowed to offload their intensive computation tasks to the MBS directly via cellular access network. When task $\mathcal{A}_i$ is offloaded to MEC server, the processing time t_i^M is composed of the transmission latency for delivering data to MEC server and the computation execution latency on the MEC server, and the corresponding energy consumption $E_{\mathrm{off},i}^M$ is the sum of the energy used for communicating data to the MBS and the energy cost used for the computation.

After successfully executing the computation tasks in the selected offloading destinations, the computation results will be sent back to the initiator devices. The details related to the mode selections for task execution are given in the following. We use an indicator variable $a_{i,m} = \{0, 1\}$ to represent the decision status of computation task. More specifically,

$$a_{i,m} = \begin{cases} 0, & \text{Local computing,} \\ 1, & \text{Computation offloading,} \end{cases} \tag{3}$$

where $a_{i,m} = 1$ represents device i's task is offloaded to destination m for executing computation; Otherwise, $a_{i,m} = 0$. Here, we denote $\{a_{i,m}\}_{m\in M=\{0,1,2,3\}}$ as the offloading decision profile, i.e., four possible offloading selections are available. More specifically, $m = 0$ represents that the mobile device executes the computation task locally on itself; $m = 1$ represents that nearby device is chosen to offload the computation task via D2D communication; $m = 2$ indicates that the computation task is offloaded via wireless local network and processed at the SBS; $m = 3$ means that the computation task is offloaded via cellular access network and processed at the MEC server co-located in MBS, which provides enough computing capability for completing the computation task. In this context, the delay-constrained computation tasks are completed with the aid of offloading destinations.

Corresponding to the above offloading methods, there are some advantages and disadvantages for each one. Generally, the MBS computation offloading offers the highest computing performance and largest coverage, but brings the highest communication cost (e.g., transmission latency). The D2D computation offloading provides a good complementary pattern for MBS computation offloading, it has the advantages of infrastructure independency and lowest cost. In particular in some extreme situations such as disaster response and military operations, a quick and flexible ad hoc D2D offloading is required. However, the disadvantage is the strict requirement on the computation duration between D2D devices to guarantee enough processing time for the offloaded computational task. The SBS computation offloading has relatively lower computing capability compared with the MEC server, and it falls somewhere in between. For the sake of clarity, some advantages and disadvantages are summarized in Table 1.

In a nutshell, in designing energy harvesting mobile devices, a basic question is whether computation can be performed locally through a microcontroller with limited processing capabilities or whether the energy-efficient computation offloading is more desirable within a given delay constraint. Three possible offloading choices are available and for each of which a different computing value is assumed. The computation offloading decisions should consider the computation abilities and communication ranges of different offloading destinations, such that the appropriate offloading destinations are selected.

A. ILLUSTRATIVE RESULTS

This subsection presents simulation result to show the performance that can be achieved by different offloading decision schemes. Regarding the simulation scenario, we consider a HetNet consisting of one macrocell (MBS), five small cells (SBSs), 20 end mobile devices coexisting 4 D2D pairs. The coverage areas of macrocell and each small cell are assumed to be circles of radius 100 m and 30 m respectively. The D2D devices use orthogonal frequency bands for simplicity and distance between D2D devices is identified as 5 m. For simplicity, the computation task is inseparable and should be offloaded as a whole. All channels are assumed to remain unchanged during a computation offloading period, while they may change over different periods. To ensure the totally consumed energy (i.e., the sum energy for both local computing and offloading) at device i cannot exceed the harvested energy at each time block, the energy harvesting constraint

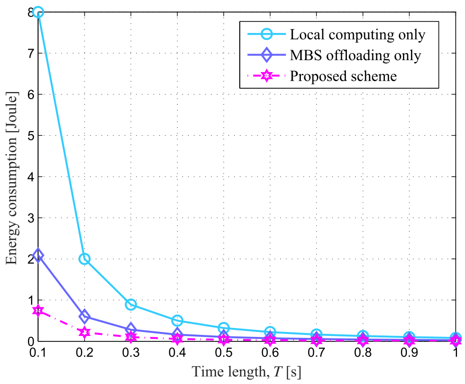


FIGURE 4. Energy consumption versus time length *T* for different decision schemes.

should be imposed. In the simulation setup, the input-data size of task D_i for different devices is randomly distributed between 100 KB and 500 KB, the total number of CPU cycles C_i randomly distributed between 200 Megacycles and 800 Megacycles. The computing ability of MBS is set as 5 GHz/sec and device *i*' computing ability is set as 0.5 GHz/sec, $\eta = 0.6$.

The curves of energy consumption (in Joules) using two decision schemes as the benchmarks versus the length of time block are depicted in Fig. 4.

- **Only offloading to MBS by all mobile devices:** All mobile devices choose to offload their tasks to the MBS with MEC server.
- **Only local computing by all mobile devices:** All mobile devices execute local computing by themselves.

We can observe that the energy consumption by all the schemes decreases as time length T increases. This is because more devices are allowed to choose the energy-efficient edge cloud computing for large T to save energy. Compared to the "local computing only" and "MBS offloading only" schemes, the proposed energy-efficient offloading scheme can decrease energy consumption, this is due to the appropriate offloading destinations can be selected. Furthermore, the "MBS offloading only" scheme is observed to achieve the lower energy consumption than the "local computing only" scheme because of the higher computing abilities of MEC server.

V. JOINT OFFLOADING AND USER ASSOCIATION

In HetNets, there are multiple BSs (i.e., MBS and SBSs) which each mobile device can choose to be associated with. Since the transmit power disparity of the BSs from different tiers varying from milliWatt to Watt, the imbalanced load distribution may be highly possible because more mobile subscribers tend to associate with the high-power MBS which is severely congested, while fewer mobile subscribers can be attracted by the low-power SBSs which are lightly loaded [35]. As a result, load balancing is a main factor influencing the network performance in HetNets. Conventionally, the signal-to-interference-plus-noise ratio (SINR) was the most prevalent one to determine whether a device should be associated with a particular BS. However, this method potentially leads to serious load imbalance in this circumstance.

It is common that D2D communications have significant potential in achieving efficient load balancing according to the real-time traffic distributions among different tiers and between the same tiers. Furthermore, the communication performance varies with user association since it depends on different channel qualities, transmission power (i.e., the harvested energy for device) and the MBS/SBS it associates to (i.e., the status of each BS). On the other hand, multiple access points are available for mobile devices, mobile devices can adaptively select the access points and transmission modes (i.e., cellular mode or D2D mode) to offload the heavy computation tasks. In order to improve the system performance of MEC-enabled HetNets (i.e., fully utilize the system resources and improve user experience), it makes sense to jointly consider user association and computation offloading. Herein, we propose a joint user association and computation offloading scheme in MEC-enabled HetNet, which is based on the instant load balancing of each BS and the overall network utility. To be specific, this scheme can be divided into the following three steps:

1) Mobile device i can be associated with either MBS or one of the SBSs by cellular mode or D2D mode for data transmission, it first chooses many nearest BSs as the candidate BSs. The collection of candidate BSs is defined as $\Omega \triangleq \{0, 1, \ldots, K\}$. The index 0 represents the MBS and the rest ones represent SBSs.
2) Then, device i will select a candidate BS from Ω. Since the computing tasks offloaded by the mobile devices could be executed either at MBS or SBSs, user association needs to be appropriately determined before evaluating the computation latency and energy consumption. We use binary control variable $x_{i,l} = \{0, 1\}$ to represent the association status between a mobile device i and a BS l. More specifically,

$$x_{i,l} = \begin{cases} 0, & \text{associated to MBS}, \; l = 0, \\ 1, & \text{associated to SBS } l, \; l \in \Omega \backslash \{0\}, \end{cases} \tag{4}$$

where $x_{i,l} = 0$ if a device i associates with MBS and D_i bits of computation task of device i can be offloaded to MBS, otherwise $x_{i,l} = 1$ and D_i bits of computation task will be offloaded to SBS l, $l \in \Omega \backslash \{0\}$.
3) When mobile device i is associated with BS j, the task offloading process begins. The mobile device i decides to choose one of the transmission modes (cellular mode or D2D mode) to offload its computation task to execution destination. Let h_{il} denote the channel gain between device i's and BS l, g_{ii} the channel gain from service requester i to service provider i, p_i the transmit power of device i, and N_0 the power of the background noise. Device i's performance is characterized by a utility function u_i (also known as uplink transmission rate in computation offloading), which is a function of

the received SINR γ_{il} that depends on its association. We denote $\mathcal{U} \triangleq \{1, \ldots, N\}$ as the set of model devices, the received SINR γ_{il} can be calculated as:

$$\gamma_{il} = \begin{cases} \dfrac{p_i h_{i,l}}{\sum_{k\in\mathcal{U}\setminus\{i\}} p_k h_{k,l} + N_0}, & \text{for cellular mode,} \\ \dfrac{p_i g_{i,i}}{\sum_{k\in\mathcal{U}\setminus\{i\}} p_k g_{k,i} + N_0}, & \text{for D2D mode,} \end{cases} \tag{5}$$

Note that the utility $u_i(\gamma_{il})$ is a popular metric, which can be defined as $u_i(\gamma_{il}) = W \log_2(1 + \gamma_{il})$ (in bps/Hz), where W denotes the bandwidth. Based on this utility function, each device builds its preference list. Within this edge cloud scenario, the association of a mobile device to a BS j depends not only on the radio channel parameters, but also the computation resources availability of the serving BS. In order to balance traffic loads and computing loads of BSs, the load balancing constraints on BSs are imposed such that the requested devices do not exceed the quota of edge cloud, i.e., $\sum_{i\in\mathcal{U}} x_{i,l} D_i \leq F_l, \forall l \in \Omega$, where F_l is the maximum computational capability that BS l can support.

We here focus on the joint design of user association and offloading selection to minimize the system latency involving mobile devices and BSs, and take into account of the constraints from both the harvested energy and the network load balance among different BSs. Note that the problem becomes more complex since it jointly captures user association which tightly couples task offloading based on the energy harvesting, and different user association may lead to different number of offloaded tasks. By examining the structure of the problem formulated, we observe that it is a mixed integer nonlinear programming problem, which can be converted into a linear optimization problem based on linear programming that simplifies the NP-hard problem. In the next subsection, simulations are conducted to demonstrate the performance of our proposed scheme.

A. ILLUSTRATIVE RESULTS

In this subsection, the simulation result is presented to illustrate the performance that can be achieved by our proposed D2D-based load balancing scheme. Here, we consider a two-tier HetNet scenario that consists of one MBS and $K = 4$ SBSs overlay the macro cell, and the neighboring SBSs are assigned orthogonal frequency bands for simplicity. The coverage areas of macrocell and each small cell are assumed to be circles of radius 200 m and 50 m, respectively. At every transmission round, each user is associated to one SBS. In the simulation setup, the input-data size of task D_i is randomly distributed between 400 KB and 800 KB, the total number of CPU cycles for completing each task C_i is randomly distributed between 900 Megacycles and 1300 Megacycles, For the purpose of simplicity, each device has the same transmission power 100 mW and the computing ability of MBS is 5 GHz/sec.

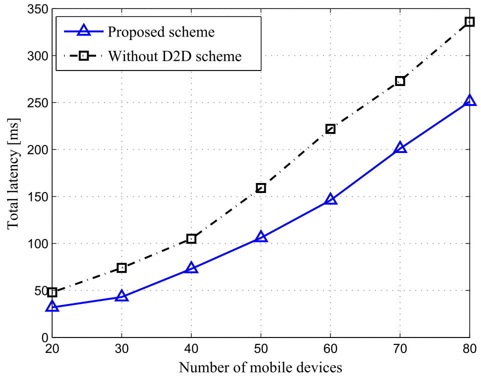


FIGURE 5. Total latency versus the number of mobile devices.

Fig. 5 compares the total latency of the proposed joint user association and offloading scheme, where the conversional SINR-based without D2D scheme is considered to be the benchmark. It can be seen that with the increase of the number of mobile devices, the increasing trend of the total latency is rapid. This is mainly because more computation tasks need to be offloaded and the interference between them becomes severe when the number of mobile devices grows. The figure also shows that the total latency of proposed joint user association and offloading scheme is lower than that of the benchmark. This demonstrates the proposed scheme can offload more data traffic from the MBS, and thus the D2D communications can achieve efficient load balancing.

VI. OPEN PROBLEMS

The research on energy harvesting MEC networks is currently in its infancy and many critical factors need to be addressed. This section will discuss the open problems in order to shed light on future research efforts.

- **System Design Requirement:** In energy harvesting MEC systems, the computation offloading and resource allocation decisions depend on the distinct amount of harvested energy. However, RF signal power decays drastically over the transmission distance due to high propagation loss. Thus, the efficiency of RF energy harvesting is relatively lower for much longer distance. To enhance the charging efficiency, novel energy delivering techniques such as energy beamforming through multi-antenna techniques and distributed multi-point wireless power transfer are urgent. On the other hand, the short range communication technique such as ultra dense networks can be exploited to improve the efficiency of energy harvesting.
- **Joint Caching and Computation Offloading:** When multiple devices aim to access the same content, the increased data traffic will put a heavy burden on the capacity-limited backhaul links. To cope with this bottleneck, caches at edge nodes (such as SBSs and D2D devices) are expected to be complementary methods for content delivery instead of core network, which can largely reduce backhaul workload and transmission latency at peak time. With the benefits of avoiding

potential network congestion and saving backhaul resource, caching popular content at MEC server is emerging as a cost-effective solution. Therefore, the joint consideration of mobile computing and caching has become a trend in the future wireless network, such that MBS has more powerful processing and storage capabilities.

- **Environment Uncertainty:** Prior network information is unpredictable ahead such as the candidate BSs, fluctuating channel conditions and random data arrival per device, especially the D2D-enabled HetNet is a very sophisticated and volatile network environment where individual devices have limited battery power and the D2D pair is not durable due to the mobility of devices. Therefore, how to dynamically obtain the future unknown information is very challenging. Machine learning is recently a potential candidate to analyze and predict the unknown behaviors by automatically learning from the environment and past experience.
- **Data Protection and Privacy:** User privacy and data security may be exposed while integrating mobile services with MEC. For example, computation-intensive tasks are offloaded to MEC server through wireless medium opening up the risk of intrusion. Moreover, different users connected to the common physical server also raise security issues. Therefore, prior to the MEC deployment, there should be an assurance that the infrastructure is well protected.

VII. CONCLUSION

This paper introduced the dynamic offloading for energy harvesting MEC networks, which become an important concept that leveraged the wireless harvested energy to achieve sustainable device operation and leveraged the MEC to enhance the computing capability on wireless devices. Furthermore, both the energy-efficient computation offloading scheme and joint user association and offloading scheme were studied. Illustrative results indicated the superior performance of the proposed computation offloading decision scheme in saving energy as well as the proposed joint user association and offloading scheme in reducing latency. In addition, we presented a number of promising research opportunities, particularly related to system design requirement, joint caching and computing, environment uncertainty, as well as data protection and privacy.

REFERENCES

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update*, Cisco, San Jose, CA, USA, Feb. 2017, pp. 2016–2021.

[2] Z. Zhou, C. Gao, C. Xu, T. Chen, D. Zhang, and S. Mumtaz, "Energy-efficient stable matching for resource allocation in energy harvesting-based device-to-device communications," *IEEE Access*, vol. 5, pp. 15184–15196, 2017.

[3] K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li, X. Peng, L. Pan, S. Maharjan, and Y. Zhang, "Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks," *IEEE Access*, vol. 4, pp. 5896–5907, 2016.

[4] X. Chen and J. Zhan, "When D2D meets cloud: Hybrid mobile task offloadings in fog computing," in *Proc. ICC*, Paris, France, May 2017, pp. 1–6.

[5] Z. Chang, S. Zhou, T. Ristaniemi, and Z. Niu, "Collaborative mobile clouds: An energy efficient paradigm for content sharing," *IEEE Wireless Commu.*, vol. 25, no. 2, pp. 186–192, Apr. 2018.

[6] L. Liu, Z. Chang, X. Guo, S. Mao, and T. Ristaniemi, "Multiobjective optimization for computation offloading in fog computing," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 283–294, Feb. 2018.

[7] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1657–1681, 3rd Quart., 2017.

[8] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 3rd Quart., 2017.

[9] A. Samanta and Z. Chang, "Adaptive service offloading for revenue maximization in mobile edge computing with delay-constraint," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3864–3872, Apr. 2019.

[10] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.

[11] Q. Hu, C. Wu, X. Zhao, X. Chen, Y. Ji, and T. Yoshinaga, "Vehicular multi-access edge computing with licensed sub-6 GHz, IEEE 802.11p and mmWave," *IEEE Access*, vol. 6, pp. 1995–2004, 2018.

[12] Y. Dai, D. Xu, S. Maharjan, and Y. Zhang, "Joint computation offloading and user association in multi-task mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12313–12325, Dec. 2018.

[13] K. Kuru and H. Yetgin, "Transformation to advanced mechatronics systems within new industrial revolution: A novel framework in automation of everything (AoE)," *IEEE Access*, vol. 7, pp. 41395–41415, 2019.

[14] Z. Chen, Q. He, Z. Mao, H.-M. Chung, and S. Maharjan, "A study on the characteristics of Douyin short videos and implications for edge caching," in *Proc. ACM Turing Celebration Conf.*, Chengdu, China, 2019, pp. 1–6.

[15] L. Ji and S. Guo, "Energy-efficient cooperative resource allocation in wireless powered mobile edge computing," *IEEE Internet Things J.*, to be published.

[16] M.-L. Ku, W. Li, Y. Chen, and K. J. R. Liu, "Advances in energy harvesting communications: Past, present, and future challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1384–1412, 2nd Quart. 2016.

[17] H. Tabassum, E. Hossain, A. Ogundipe, and D. I. Kim, "Wireless-powered cellular networks: Key challenges and solution techniques," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 63–71, Jun. 2015.

[18] D. Niyato, D. I. Kim, M. Maso, and Z. Han, "Wireless powered communication networks: Research directions and technological approaches," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 88–97, Dec. 2017.

[19] C. You, K. Huang, and H. Chae, "Energy efficient mobile cloud computing powered by wireless energy transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1757–1771, May 2016.

[20] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1784–1797, Mar. 2018.

[21] X. Hu, K.-K. Wong, and K. Yang, "Wireless powered cooperation-assisted mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2375–2388, Apr. 2018.

[22] S. Bi and Y. Zhang, "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4177–4190, Jun. 2018.

[23] Y. Mao, J. Zhang, Z. Chen, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.

[24] G. Zhang, W. Zhang, Y. Cao, D. Li, and L. Wang, "Energy-delay tradeoff for dynamic offloading in mobile-edge computing system with energy harvesting devices," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4642–4655, Oct. 2018.

[25] L. Liu, Z. Chang, and X. Guo, "Socially aware dynamic computation offloading scheme for fog computing system with energy harvesting devices," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1869–1879, Jun. 2018.

[26] M. Min, L. Xiao, Y. Chen, P. Cheng, D. Wu, and W. Zhuang, "Learning-based computation offloading for IoT devices with energy harvesting," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1930–1941, Feb. 2019.

[27] J. Xu, L. Chen, and S. Ren, "Online learning for offloading and autoscaling in energy harvesting mobile edge computing," *IEEE Trans. Cogn. Netw.*, vol. 3, no. 3, pp. 361–373, Sep. 2017.

[28] W. Chen, D. Wang, and K. Li, "Multi-user multi-task computation offloading in green mobile edge cloud computing," *IEEE Trans. Serv. Comput.*, to be published.

[29] S. Mao, S. Leng, K. Yang, X. Huang, and Q. Zhao, "Fair energy-efficient scheduling in wireless powered full-duplex mobile-edge computing systems," in *Proc. IEEE Global Commun. Conf.*, Singapore, Dec. 2017, pp. 1–6.

[30] N. Janatian, I. Stupia, and L. Vandendorpe, "Optimal offloading strategy and resource allocation in SWIPT-based mobile-edge computing networks," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Lisbon, Portugal, Aug. 2018, pp. 1–6.

[31] F. Wang and X. Zhang, "Dynamic interface-selection and resource allocation over heterogeneous mobile edge-computing wireless networks with energy harvesting," in *Proc. IEEE Conf. Comput. Commun. Workshops*, Honolulu, HI, USA, Apr. 2018, pp. 190–195.

[32] F. Zhou, Y. Wu, R. Q. Hu, and Y. Qian, "Computation rate maximization in uav-enabled wireless-powered mobile-edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1927–1941, Sep. 2018.

[33] K. Zhu and E. Hossain, "Joint mode selection and spectrum partitioning for device-to-device communication: A dynamic Stackelberg game," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1406–1420, Mar. 2015.

[34] Y. Dai, D. Xu, S. Maharjan, and Y. Zhang, "Joint load balancing and offloading in vehicular edge computing and networks," *IEEE Internet Things J.*, to be published.

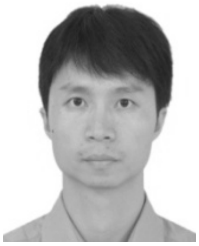
[35] M. Ali, Q. Rabbani, M. Naeem, S. Qaisar, and F. Qamar, "Joint user association, power allocation, and throughput maximization in 5G H-CRAN networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9254–9262, Oct. 2017.

BIN LI is pursuing the Ph.D. degree with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. From 2013 to 2014, he was a Research Assistant with the Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong. From 2017 to 2018, he was a Visiting Student with the Department of Informatics, University of Oslo, Norway. His research interests include physical-layer security, wireless cooperative networks, and mobile edge computing.

ZESONG FEI received the Ph.D. degree in electronic engineering from the Beijing Institute of Technology (BIT), in 2004. He is currently a Professor with the Research Institute of Communication Technology, BIT, where he is involved in the design of the next generation high-speed wireless communication. His research interests include wireless communications and multimedia signal processing. He is the Chief Investigator of the National Natural Science Foundation of China. He is the Senior Member of the Chinese Institute of Electronics and the China Institute of Communications.

JIAN SHEN received the M.E. and Ph.D. degrees in computer science from Chosun University, South Korea, in 2009 and 2012, respectively. Since late 2012, he has been a Professor with the Nanjing University of Information Science and Technology, Nanjing, China. His research interests include public key cryptography, secure data sharing, and data auditing in cloud.

XIAO JIANG received his MA.Sc degree in mechanical and electronic engineering from the Civil Aviation University of China, in 2013. From 2013 to 2016, he was a director of Emergency Rescue Information Department, Yantai Airport, China. He is currently a General Manager of Honghorn Information Technology Co. Ltd., China. His general research interests include deep learning and the internet of things.



XIAOXIONG ZHONG received the Ph.D. degree in computer science from the Harbin Institute of Technology, China, in 2015. From 2016 to 2018, he was a Postdoctoral Fellow with the Graduate School at Shenzhen, Tsinghua University, China. He is currently an Assistant Professor with Peng Cheng Laboratory, China. His general research interests include mobile edge computing and the internet of things.