Intelligent Resource Management at the Edge for Ubiquitous IoT: An SDN-Based Federated Learning Approach

Venkatraman Balasubramanian, Moayad Aloqaily, Martin Reisslein, and Anna Scaglione

Abstract

The ubiquitous nature of Internet of Things (IoT) devices has posited many challenges that need innovative solutions in the 5G era. Software defined networks (SDNs) are becoming indispensable in managing several aspects of next-generation IoT networking that arise from the need to control highly heterogeneous, geographically dispersed, mobile IoT devices. One such aspect is cache management at the edge. Recently, multiple forms of edge resources, including mobile device clouds and micro-edge data centers have emerged to provide scalable cache placement locations that reduce the costs for the mobile network operator (MNO). As all of these service locations are registered with the MNO (or links established after registration with the 5G base station, BS), content should be placed according to the user's demand and the cost the user is willing to pay to receive the desired level of QoS. To this end, it is important to understand the future popularity of the content for its optimal placement considering the highly dynamic user mobility. In this article, we address two key aspects of a mobile IoT network: security and seamless connectivity for data delivery. We rely on the federated learning (FL) architecture, which enables harnessing data and computational capabilities at end-user devices to train machine learning models. We study FL concepts in the domain of edge computing for IoT use cases, such as caching. We draw conclusions from various state-of-the-art models and posit several challenges that can be overcome via a novel proposed control algorithm.

Introduction

Motivation for Mobile Device Cloud for Ubiquitous IoT

The drastic increase in quality of service (QoS) requirements at the edge demands a practical ubiquitous system that can cover large geographical areas. In essence, the new solutions have to not only provide a seamless communication framework, but also provide a system design that can cater to the computation needs at the last mile, particularly in heterogeneous ubiquitous Internet of Things (IoT) networks. One such design is the caching of content at the edge for fast content accessibility. De-congesting the core network and reducing the delay for the last mile users have been the aims of many industrial development efforts toward high-quality network services. In particular, caching infrastructures at the edge primarily consider placing popular contents closest to the users such that frequently requested files can be retrieved faster. While considering the edge caching paradigms, two models were studied, namely the edge–data center model, known as mobile edge cloud (MEC) [1], and the low cost 5G-device-to-device (D2D) model [2], known as mobile device cloud (MDC) or mobile device edge cloud (MDEC). Both of these models have their own pros and cons while deploying services [2, 3]; however, MEC deployments are typically considered to be costly.

Due to the heterogeneous nature of ubiquitous IoT infrastructures, an integrative framework that efficiently manages communication and caching resources is essential for scalability, especially in densely crowded environments [4]. A wide variety of opportunities and innovative solutions arise when faced with the challenges from integrating multiple heterogeneous networks, such as aerial networks (e.g., unmanned aerial vehicle, UAV, networks) and ground networks such as terrestrial MDC networks. Such an integrative framework would also facilitate scalable overall coordination between heterogeneous infrastructures. In this article, we focus primarily on MDC-based applications, where mobile end devices (equipment nodes, MEs) provide scalable caching resources, and end users require prescribed levels of quality of experience (QoE).

Studies such as [2] have made MEs an integral part of caching schemes. In device-involved models, each mobile device already has a 5G subscription and a MAC-ID that is known centrally by a BS owned by a mobile network operator (MNO). A client that requests service forms a D2D resource composition using the idle device resources that are near. Such a resource-rich environment of MEs can be substituted for the expensive edge data center by forming an MDC.

Proposal for Federated-Learning-Based MDC Management

However, despite MDC's promise, privacy issues are still prevalent. Further, users' uploading contents to an edge data center results in inheriting the limitations of the existing cloud infrastructure, including the risk of the EDC being a single point of failure and security risks due to uploading sensitive information.

Digital Object Identifier: 10.1109/MNET.011.2100121

Venkatraman Balasubramanian, Martin Reisslein, and Anna Scaglione are with Arizona State University; Moayad Aloqaily is with Al Ain University.

0890-8044/21/$25.00 © 2021 IEEE

Hence, recent research has been driven toward exploring solutions that address the problem of privacy along with the allocation and redistribution of computation across multiple levels. The federated learning (FL) architecture can be a key ingredient to simultaneously overcome the challenges of privacy and to manage distributed computing resources in a scalable manner.

In conjunction with the FL modules, as most of the distribution of content over an MDC is not completely controllable due to the device movements, the software defined networking (SDN) paradigm can enable a higher level of control over the data plane. Hence, an SDN-assisted FL model may not only provide better control, but also enable seamless communication to maintain QoS.

Through this SDN-based FL model, users at the lowest layer may request services from a 5G-D2D MDC (i.e., registered with the MNO) or an EDC collocated with a 5G New Radio (NR) BS. Either of these locations can host on-demand content. Having multiple content placement locations not only forms a convenient business model, but also empowers the customers to responsibly choose the level of QoS for which they would like to pay. As most of the QoS metrics are related to where the popular content with high demand frequency is placed, we need to closely examine the content placement. Further, as both of these service locations (i.e., EDC and MDC) are registered with the MNO, the question of where the content with higher demand should be placed is directly tied to the MNO's profit margin. A controller situated at the edge could make the key decisions of content placement with the objective of maximizing the MNO's profit.

When each user downloads the FL framework from the controller for training the model with the local data, the main goal of the end user is to upload only the necessary parameter updates. The controller will then cohesively assemble the parameters received from the users via averaging algorithms and choose the popular files. To that end, the contributions of this article are:

- An SDN-controlled FL framework that provides a simple but seamless model for ensuring high QoS in a ubiquitous IoT network is proposed.
- The federated averaging paradigm is employed for ensuring local to global aggregation that enables a secure learning environment.
- A case study is simulated to examine the performance of the SDN-based FL model.

Federated Learning Models

Brief Background on Federated Learning

Federated learning is a technique that trains machine learning (ML) models harnessing data and computations on local devices (nodes) [5]. This decentralization of training provides a higher level of security to nodes that do not wish to upload all their data to a central entity. Instead, only ML model parameters are uploaded to the central entity. The central entity then merges all the information obtained from the client nodes (end devices) to provide a global update to the model. In the case of MDCs, the ME data that is used to train the FL framework model remains private. The FL framework, however, does not provide any detail on how to carry forward seamless communication of these calculated results between the ME and central entity, which can be complicated as a node may move out of its initial position.

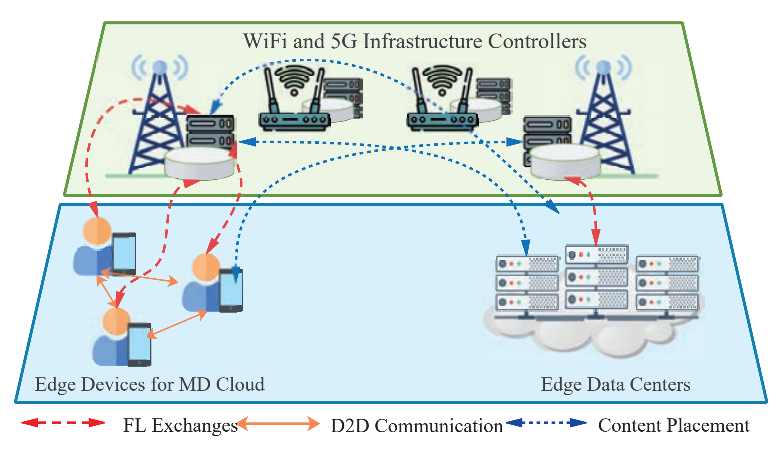
WiFi and 5G Infrastructure Controllers
Edge Devices for MD Cloud
Edge Data Centers
FL Exchanges
D2D Communication
Content Placement

FIGURE 1. Overview of the system model: A controller is collocated with each base station (BS) and WiFi access point. The individual controllers can be coordinated by an SDN control hierarchy. When a user requests a content, the available options are based on the price the user is willing to pay: either a mobile device cloud (MDC, formed via D2D communication by edge devices) or an edge data center (EDC). All the individual learning models of the edge devices are gathered (via the FL exchanges illustrated by red dashed arrows) at the control plane (central FL controller). The solid orange arrows represent the D2D communication between the MDC users. The blue dotted arrows represent the content placement on the data-plane entities.

Mobility is a key challenge while exchanging information. For instance, when a local model trained by an MDC device in one location moves to another location, the trained local model needs to be communicated to the central controller for producing a global model to achieve holistic training, or else the model update is discarded. Since mobility is an essential component for providing seamless communication, nodes have to continue to provide their locations around a particular area for completing a service that has been requested by a caching user.

Survey of the State of the Art

The study of FL training models in wireless networks is on the rise. These studies have explored various network parameters that influence the transmission of the local FL models to the BS, whereby the BS aggregates all the local FL models and maintains the global FL model. The key to our work is producing a seamless content placement/retrieval environment for cloud users, which not only benefits the MNO, but also provides a better QoE to the user considering ME mobility in FL. Figure 1 shows how the collocation of the BS controller assists in reliably placing content. Now, we compare and contrast how state-of-the-art models compare with our approach.

Edge Caching Models and Federated Learning: Recently, there have been studies showing that usage of D2D links to form clusters at the edge is a good replacement for the expensive edge caching capacity, whereby D2D clusters also provide better scalability. For instance, Zhang *et al.* [6] have shown how mobile vehicles can act as smart caching locations to enable dynamic cache storage of tasks, whereby, the vehicles communicate directly with the BS. Similarly, Wu *et al.* have designed a content distribution architecture based on the D2D-assisted caching paradigm [7]. These solutions consider dynamic caching capacity usage, but do not include demand variations. Additionally, all of these state-of-the-art models provide benefits in terms of caching, but disregard the dynamics of mobility that pose a significant challenge at the last mile. These solutions rely on probability-based estimates for judging the popularity of a content, which may be inaccurate in mobile environments.

Chen *et al.* [8] have studied FL training models in wireless networks. In particular, Chen *et al.* have studied the various network parameters that come into play while the local FL models are transmitted to the BS, which aggregates all the local FL models and maintains the global FL model at the BS. Chen *et al.* have formulated an optimization problem to capture the wireless network parameters and user choices. We take inspiration from the work by Chen *et al.* for the MDC scenario where network factors come into play while uploading the local FL models to the central controller situated at the BS. The key difference from [8] is that our interest is in producing a seamless content placement/retrieval environment for cloud users that not only benefits the MNO but also provides a better user QoE. To that end, during training, our algorithm strives to produce optimal results while minimizing losses.

Federated Learning in Wireless Networks: Yang *et al.* [9] propose an over-the-air computation model. This model exploits the super-position property of wireless channels to aggregate data. This is one of the preliminary models that provides close observation of applying FL in wireless networks. Yu *et al.* [10] propose a content caching scheme called FL-based proactive content caching (FPCC). FPCC considers a hierarchical architecture where users upload only the requisite updates to the edge server and keep all the remaining sensitive data within the devices. However, due to the complex nature of the FPCC model, a scalable deployment of the model may not be possible. On the other-hand, our proposed FL model is purely based on the probabilities of outcomes that are judged via user behavior modeling, which gives realistic outcomes for the prediction process.

Considered Use Case for Federated Learning

As illustrated in Fig. 1, the data plane comprises all the locations that can be potential content placement sites. Also, the users request content in the data plane. Every user in the data plane (whether part of the MDC or not) downloads the long short-term memory (LSTM) model from the server, trains the model with local data, and uploads the model to the server. The edge data center has a FedCo agent that is downloaded for local training from the central controller. The locally trained models are re-uploaded to the closest controller. The controller has an input graph of the network where the locations of popular contents are saved. In this way, routing the request to the placement site follows a breadth first traversal following the shortest path to the content placement site at the edge. In the case of the MDC, if the device that was trained in one location moves to another location, the closest controller communicates with the home controller and exchanges the necessary information. Therefore, placing the contents in locations where the MNO increases its revenue, and at the same time not compromising the QoS of the requesting user while moving from one place to another, are the main aims of our system model.

Federated Learning Challenges: As explained above, there are key challenges that a standalone FL approach cannot readily solve. However, challenges that are easily solvable in the purview of the FL framework, such as security and trust [11], are definitely important features. Further, the number of messages exchanged between the client nodes and the central entity should be considered. Therefore, the FL framework usage posits the following challenges in the mobile edge network setting.

- **Fast delivery:** To ensure the best performance of the task, MDCs or 5G-D2D device clouds [2] should exchange information about new requests for content with a specified deadline to maintain the QoS. However, a standalone FL framework does not necessarily perform pipe creations between geographically disparate controllers specifically for information exchanges in scenarios where the devices are moving in and out of an area, such as stadium environments.
- **Trust and privacy:** An FL framework addresses the trust and privacy issues to a large extent in 5G-D2D clouds by letting the client nodes keep their personal sensitive data within their devices. Trust and privacy are critical metrics to consider, which the FL framework naturally takes into consideration. However, a key challenge is the communication between two FL controllers that are geographically separated; SDN assistance can facilitate privacy in the communication channels between controllers.
- **Security:** The content caching service may face a security issue, such as a man-in-the-middle attack, or in the case of an EDC, a distributed denial of service attack. These risks are alleviated by the FL framework, which mainly caters to the data accountability and integrity characteristics of the secure environment by putting the end user in complete control of the data that is uploaded to the central entity.
- **QoS:** As mentioned before, at the time of mobility, nodes exchange various types of data, such as cached multimedia content or new requests for popular content. Using the FL framework as a standalone model will not be enough to maintain a seamless service.

Federated Learning for Edge Caching

Edge Caching and 5G

With the emergence of 5G, caching at the edge became a key service that could leverage the advantages of 5G technology jointly with SDN technology. Each technology can provide the other with a variety of benefits that can have a big impact on the quality of different ubiquitous IoT services:

- **Caching Policies in 5G:** A main challenge for caching policies and computing in 5G is the need for accurate recognition and control of all available resources. At the last mile, caching can be incorporated in the macrocells and small-cell clouds for which there is a need to separately analyze the data that is being processed and the resource being put into use. The resource management with an SDN controller is therefore a promising solution when multiple points of computation and caching are available.
- **Caching Policies over SDN:** As posited in [12], a caching policy is defined for an SDN controller that has an added complexity of resource partitioning and job partitioning. This approach takes into account only situations when caching happens in parts. This is one of the earliest studies that elaborates on the need of SDN control at the edge, but does not consider the privacy issues and the solutions offered by our FedCo approach.

We build on the 5G and SDN technologies to propose our FedCo concept and later provide simulations to evaluate the key performance metrics.

Concept and Proposal

Overview: In the system we consider, in combination with heavily resourced EDCs, the mobile phones communicating in an MDC form a "mini" edge cloud in their support. Such a heterogeneous system is vulnerable to many problems; hence, trust is very important. The profit margin maximization of the serving MNO entity means that its management policies are quite complex and require high-speed communication, caching, and computation; also, the MDC participants need to provide identification and go through authentication.

In this section, we present the design of a hierarchical architecture for content delivery and caching designed to serve these networks. Specifically, our proposed solution includes an FL framework for highly secure non-intrusive management of demand information, in conjunction with an SDN controller used to manage the dynamics of the user locations as they move from one BS to another. In our model, the MNO places the 5G-D2D compositions in different strategic areas along with other edge computing entities, allowing users to choose their offloading points. We employ the FL framework to predict the users' demands for a particular content in order to provide the exact location for the content placement. The SDN edge controller behaves as the global aggregation point, while the users provide the locally processed information

System Architecture: Figure 2 presents the various system architecture components and the interactions between the components of the architecture. The control plane consists of three important modules:

1. The Location Classifier module
2. The Federated Averaging Engine module
3. The Placement Decision Handler module

The Location Classifier module inside the FedCo controller keeps track of where a particular content is placed (e.g., MDC or EDC). This module makes ongoing procedural calls to the respective MDCdb and DCdb databases to understand the popularity of a content and manages the placement.

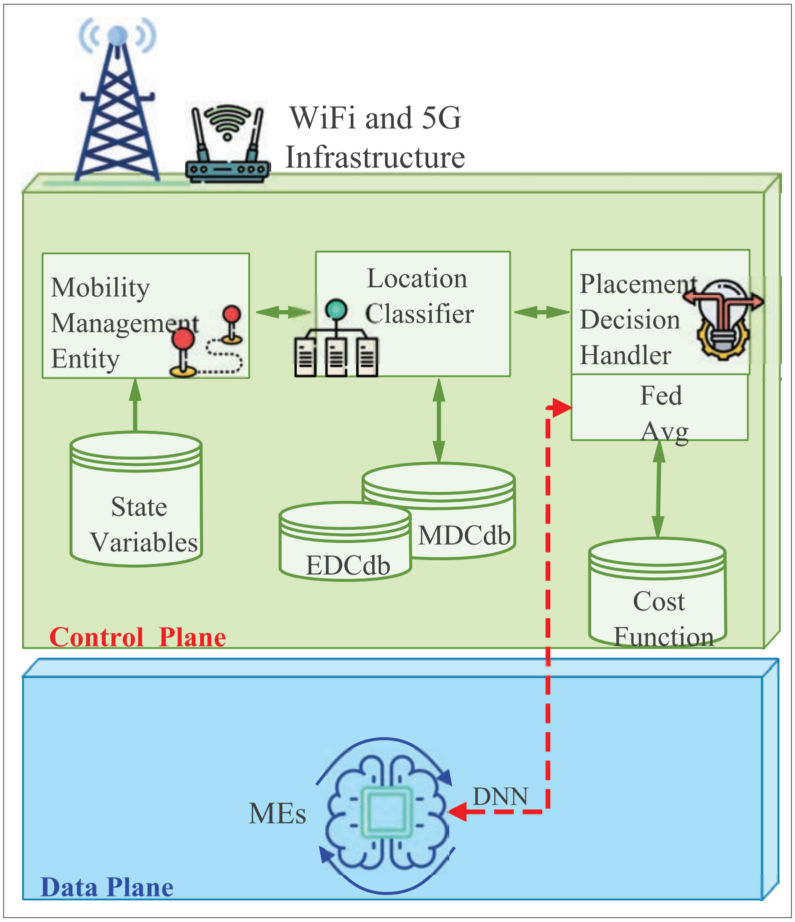

FIGURE 2. System architecture: Each MDC user device (ME) downloads the FL framework from the controller and trains the model with the local data; the locally trained models are then re-uploaded to the closest controller, specifically, the Federated Averaging Engine (Fed Avg).

The Federated Averaging Engine is tasked to train a model that predicts two important quantities, $L(\cdot)$ and m_s, in our FL framework. The quantity $L(\cdot)$ is the weighting function that captures the probabilistic outcome of popular placements. $L(\cdot)$ can be interpreted as the perception of the operator about a probabilistic outcome. For instance, a content may be placed in an MDC or an EDC. A content placed in an MDC might not always be profitable due to the mobile nature of the MDC if we do not ensure other mobility parameters. m_s is the value function that captures the subjective operation of the QoS perceived by the user. m_s can be interpreted as the probability of achieving the highest QoS when the computation is performed on a resource s. The mathematical model is elaborated in [13]. In the data plane, the users update the deep neural network (DNN) model, producing a result that will be uploaded to the control plane once the computation is complete.

The Placement Decision Handler module considers the placement popularity score $L(\cdot)$ and the QoS users' scores m_s to decide how to best place the content. These scores can be represented

Characteristics	Basic edge caching	5G-based edge caching	Federated-learning -based edge caching (FedCo)
Communication range	Mobile device clouds use limited-range based technologies (e.g., Zigbee or WiFi)	Range depends on 5G base station (typically higher than for basic edge caching)	Range depends on 5G base station
Quality of Service	Data delivery as a QoS measure; throughput is typ. limited.	Data delivery is a QoS measure, typ. average throughput and reliability	Data delivery is a QoS measure; high throughput and reliability through FL framework.
Security	↓	↓	↑ Sensitive information is stored in local devices, ensuring data security
Privacy	↓	↓	↑ Sensitive information is stored locally in the devices, ensuring data privacy
Trust	↓	↓	↑ As all participants register with their IDs, trust and privacy of information is maintained
Distribution of Control	Centralized	Centralized	Decentralized
Resource orchestration	Depends on the state-of-the-art technology (e.g., Bluetooth or Zigbee are candidates)	Follows the norms of managing 5G resources	Advanced resource management techniques combined with FL resource management
Scalability	↓	↓	↑ Improved scalability as there are multiple points where the contents can be placed
Intelligence	↓	↓	↑ FL framework provides a high intelligence to the system to self-learn and provide a high-profit operation
Autonomy	Limited	Limited	Very advanced

TABLE 1. Comparison between federated learning approaches. ↑ means relatively high, ↓ means relatively low.

through a cost function, the details of which are beyond the scope of this overview article.

Federated-Learning-Based Management: Table 1 elaborates the characteristics and subjective evaluation of the FL-based approaches. Table 1 compares and contrasts various factors, such as the control distribution, of the proposed FedCo approach compared to basic edge caching and 5G-based edge caching. The devices are subscribed to the BS, and once the D2D links are in place, the intermediate devices communicate via the BS at the time of mobility. The MEs either volunteer or lease their own idle storage resources for enhancing the MDC capacity. All mobile devices typically already have 5G subscriptions and MAC-IDs, which are known centrally by a BS (owned by an MNO). A client that requests a service forms a D2D resource composition using the idle device resources that are near (i.e., exploits the MDC). However, despite the MDC promises, privacy issues are still prevalent.

FedCo overcomes these privacy issues via local training on the devices. We use the long short-term memory (LSTM) DNN architecture, which is a type of recurrent neural network, to predict contents' popularity and users' QoS scores. More specifically, each user in the MDC downloads the initial LSTM model from the central controller at the BS. The devices train the model with the local device data and re-upload their local update to the central controller. In particular, the local device data used to train the model encompasses popularity indices that are stored on the device based on the user's daily activities (e.g., the past history of videos the user watched and related contextual information). Based on this local device data, each device trains its own model; that is, there is one model per device location. This raw local device data is never revealed. The central controller only knows the device ID, which is used to subscribe to the service and the updated learning model (from which the raw local data cannot be readily derived). Our LSTM framework receives the traffic matrix as input, whereby the traffic matrix (requests of clients, content data held by owner of data in an MDC, features) is part of the initial model download. The central controller co-located with a BS aggregates all the federated local models received from the devices in its coverage area and applies the averaging algorithm to determine the predictions of the contents' popularity and the user QoS scores. Based on the content popularity and user QoS score predictions of the LSTM DNN architecture (operating in a federated mode), the FedCo module (centrally with an ML strategy) decides on the placement of the popular contents in favorable positions for the user (QoS-wise) and for the operator (revenue-wise). The placement sites are the MDC and EDC locations.

Based on the content placement, the SDN controller adapts the routing paths to the placement sites. More specifically, the placement locations are fed as a network graph to the SDN controller. Once the prediction from the LSTM is obtained (i.e., future popularity, QoS scores) and the FedCo module has decided on the content placement, the SDN controller determines and activates the routing paths toward the placement sites. If a device serving a request has moved out of a location, the closest controller forwards the uploaded model, whereby the forwarding routing paths are determined by the SDN controller.

Through the FedCo model, users at the lowest layer may request services from a 5G-D2D mobile device cloud (registered with the MNO) or an EDC collocated with the 5G-NR BS. The multiple available content placement locations form a convenient business model, while at the same time empowering the customers to select the desired level of QoS that they can afford (in terms of cost). The QoS is strongly influenced by the placement decisions for popular content; therefore, content management is very important. Accordingly, we assume a network graph that is given as input to the LSTM at the time of training. Further, as both of the service locations (i.e., the EDC and MDC) are registered with the MNO, the question of where the content with higher demand should be placed influences the MNO's profit margin. Therefore, a controller situated at the edge makes the key decisions of content placement. The routing of the requests is based on the state of the network. That is, there are no static routes defined as is common in traditional networks. Based on the FedCo updates, new paths can be added from a source to a destination utilizing 5G-D2D edge routing [2]. As shown in Fig. 1, when each user downloads the FL framework from the controller for training the model with the local data, the main goal of the data plane entities is to upload only the necessary parameter updates.

Advantages

The proposed SDN based FL framework has several advantages:

- Improved network QoS in terms of latency, efficiency achieved and security, which are key requirements for the caching service.
- Information privacy and reliability of caching locations and content delivery sites.
- Identification of mobile devices participating in forming the device cloud is maintained as a log for new cloud formations that provides trustworthiness and anonymous presence for the mobile devices, which in turn protects them from security attacks.
- Better resource orchestration that enables better network control via the SDN controller, especially at the time of device handovers in case of MDCs and total available data center resources in case of EDCs. This essentially provides better scalability, accessibility, and power efficiency.

All of these characteristics are showcased in the next section devoted to analyzing the performance of the proposed architecture.

Performance Evaluation: A Case Study

To test the proposed FL framework and comment on its effectiveness, a simulation case study has been performed via the Mininet simulator [14], which is based on Python. We employed Pytorch, a common tool for FL simulations, with typical synthetic i.i.d. datasets. The experiments have been performed on an Intel Core i7-4510U CPU with 2.0 GHz and 8 GB of RAM. The simulator quantities are provided in Table 2. In a network region of 1500 m × 1500 m, up to 40 mobile devices are scattered. Devices can share idle memory of up to 1.0 GB per device in order to collectively store multiple parts of 100 MB test videos. Two 5G NR BSs are placed similar to the scenarios in [2] to cover the entire simulation environment with 5G-D2D. The communication between the mobile nodes and the requesting mobile node is conducted via the BS. The devices are allowed to move anywhere in the network area, and the handovers between the BSs are managed by a single POX controller. We compare our model with random caching [15] and EdgeBoost caching [2] with an SDN controller. We evaluate the delay aspects at the time of request generation from the user and the cache hit ratio (CHR).

Parameters	Numerical values
Communication medium	IEEE 802.11n (for 5G-D2D) LTE gNB with 1 Gb/s bandwidth (for edge data center access)
Area considered	1500 m × 1500 m
No. of BSs	2
No. of MEs and edge nodes	40 & 5
Edge data center placement	Uniform rand. distribution
Mobility model	Random-waypoint model
Simulation duration	100 s
Packet length	1024 bytes
Packet interval	5 ms

TABLE 2. Simulator settings.

Cache Hit Ratio

The CHR is defined as the percentage of satisfied requests out of the total number of received requests. The results depicted in Fig. 3 indicate that the proposed FedCo solution outperforms traditional techniques and the SDN-assisted state-of-the-art EdgeBoost solution. The CHRs of all the methods increase with growing caching capacity. This can be attributed to the fact that as the caching size grows, the CHR values increase as the probability of yielding the right location becomes higher. Due to the precise and timely estimate of demand for content, FedCo achieves a higher CHR than the other methods across the entire range of considered caching space (inclusive of both the EDC as well as MDCs). For example, for a 4 percent cache size, the FedCo CHR is roughly 12 percent higher than the random caching CHR. These results indicate that in dynamic mobility cases, SDN assistance is capable of adjusting locations and continuously monitoring the device locations to meet service demands. Additionally, as the node density increases, the CHR is likely to increase.

Average Delay

We define the average access latency (AAL) or average delay as the time interval between sending the request and receiving the first packet of the requested content. As the number of relay nodes increase, the packets generated increases depending on the number of relay nodes. This is the case observed in the MDC scenario. Thus, by using the MDC, the number of packets generated increases to assist in the communication between authenticated MDC devices. Thus, there is an efficiency trade-off. That is, the MDC approach incurs a higher communication overhead and in return lowers delays. We measure the average delay specifically for the MDC approach. We find that as the number of devices increases, the delay is reduced as the content is easily found through the collaboration of the devices. For example, when there were only 10 devices, the FedCo delay is roughly 50 μs; when the number of devic-

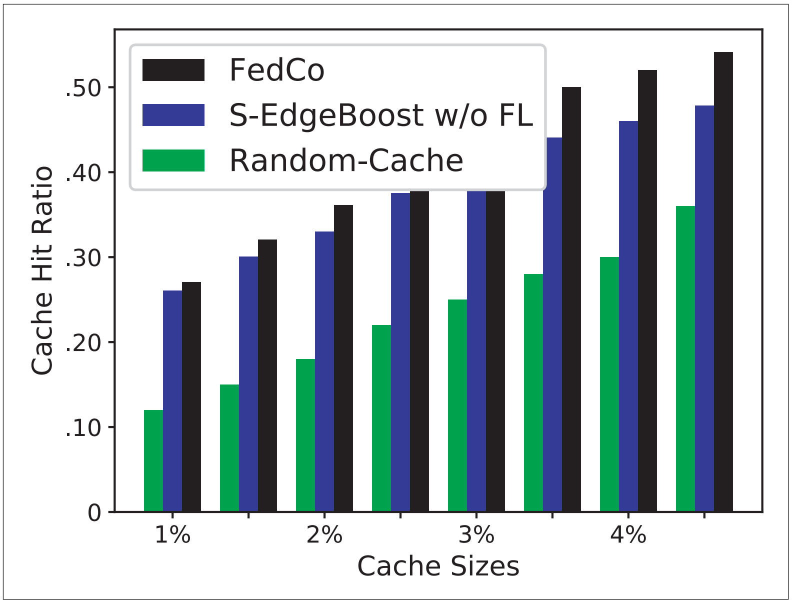

FIGURE 3. Cache hit ratio as a function of cache size.

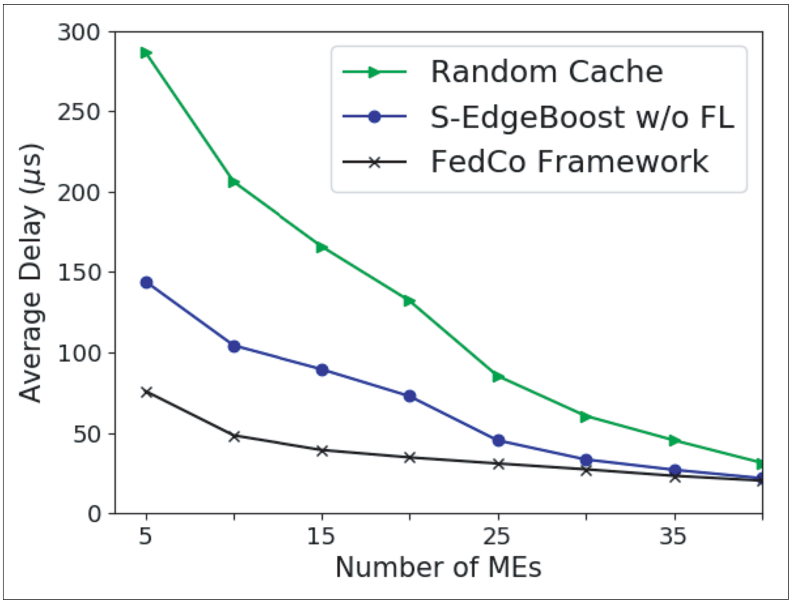

FIGURE 4. Average delay as a function of number of mobile devices.

es is increased to 40 MEs, the delay is reduced to around 20 µs. This delay reduction is achieved through the collaboration and willingness of the devices to participate in the MDC. However, a key point to note here is that FedCo still outperforms other state-of-the-art frameworks, mainly because of the relational learning via the DNN that constantly updates the requirements and the ME locations monitored by the SDN controller.

Challenges and Future Research

The SDN-controller-assisted FL framework has advantages, including but not limited to security and resource orchestration features. However, the deployment of such a framework faces challenges that need to be addressed in future research:

Scalability: MDC deployment has challenges mainly in the form of data generated between devices. Hence, the key challenge is in terms of scalability. Additionally, throughput and latency are also affected due to the rapid growth of data moved between the MDC systems. Further, the willingness of the participants is also an important consideration. This participation in essence affects scalability, and is also a critical challenge that can cause QoS reduction.

Federated Learning in MDCs: It is well known that 5G technology provides a resource-rich communication infrastructure that can run complex applications across geographically diverse regions. Forming clouds via the idle device resources of end users is one example that can leverage the 5G infrastructure. In the case of MDC, mobile devices can be utilized as intermediate nodes to forward messages. Moreover, via FL mechanisms, MDCs can assist in processing the collected data and training the FL models locally. This also raises questions about device resource constraints, such as CPU, battery, and RAM for computation. Moreover, a sophisticated resource allocation strategy needs to be in place to prolong device lifetimes and to optimize computation on the device.

Information Diversity: Mobile devices exchange information that needs to be processed differently in an MDC. Some of the control information exchanged is purely messages that are used to maintain the 5G-D2D control link between the devices. These could be different instructions received from the SDN controller or simple keep-alive messages between devices in an MDC.

5G Infrastructure: There are a handful of notable places or cities that have started deploying the 5G infrastructure. However, there is a need for 5G infrastructure to be installed across continents for better acceptance of such frameworks. Various methods as to how such a deployment can be made quicker is still an open challenge due to cost of installing the 5G infrastructures.

Automated MDC Network: The proposed framework is a self-stabilizing or self-adjusting network that allows the network to adapt to environmental changes. Such models achieve a high degree of automation, which is the key to designing a mobile device cloud service. It is also important to understand the needs of communications that are going to be separated by geographically diverse regions such as land, sea, and air.

Dense Number of Nodes: As the node density increases, it is integral to understand the dynamics of the links between the nodes. Various physical layer characteristics, such as fading losses, location ID, and channel interference management, play a crucial role in monitoring the MDCs.

Conclusion

In this article we present an SDN-assisted federated learning framework that not only provides secure and trustworthy service delivery but also ensures seamless communication in a scalable manner to end users. With the aid of SDN-assisted FL, mobile devices can act as key computation and caching resources. Via intermediate nodes, routing can be achieved that leverages all the resources available at the edge in a distributed/hierarchical manner. With the advent of ubiquitous IoT, service requests are going to be irregular and may span across geographically diverse regions. Thus, such a framework may not only be economically profitable, but also improve the service quality. Having intelligent resource management, such as an SDN-assisted FL framework, to manage the resources in the network is beneficial to overcome a static (potentially over- or under-provisioned) edge network. We provide a simulation analysis on a Mininet simulator to check the feasibility of our solution. As key metrics, we consider the cache hit ratio and average delay. We provide a comparison analysis of the proposed FL-based solution and other state-of-the-art caching techniques for a better understanding of the framework.

References

[1] B. Brik, P. A. Frangoudis, and A. Ksentini, "Service-Oriented MEC Applications Placement in a Federated Edge Cloud Architecture," *Proc. IEEE ICC*, 2020, pp. 1–6.

[2] V. Balasubramanian *et al.*, "Edge-Boost: Enhancing Multimedia Delivery with Mobile Edge Caching in 5G-D2D Networks," *Proc. IEEE Int'l. Conf. Multimedia and Expo*, 2019, pp. 1684–89.

[3] C. Long *et al.*, "Edge Computing Framework for Cooperative Video Processing in Multimedia IoT Systems," *IEEE TMM*, vol. 20, no. 5, 2018, pp. 1126–39.

[4] M. Aloqaily *et al.*, "Data and Service Management in Densely Crowded Environments: Challenges, Opportunities, and Recent Developments," *IEEE Commun. Mag.*, vol. 57, no. 4, Apr. 2019, pp. 81–87.

[5] Z. Chang *et al.*, "Learn to Cache: Machine Learning for Network Edge Caching in the Big Data Era," *IEEE Wireless Commun.*, vol. 25, no. 3, June 2018, pp. 28–35.

[6] K. Zhang *et al.*, "Cooperative Content Caching in 5G Networks with Mobile Edge Computing," *IEEE Wireless Commun.*, vol. 25, no. 3, June 2018, pp. 80–87.

[7] D. Wu *et al.*, "Collaborative Caching and Matching for D2D Content Sharing," *IEEE Wireless Commun.*, vol. 25, no. 3, June 2018, pp. 43–49.

[8] M. Chen *et al.*, "A Joint Learning and Communications Framework for Federated Learning Over Wireless Networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, 2021, pp. 269–83.

[9] K. Yang *et al.*, "Federated Learning Via Over-the-Air Computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, 2020, pp. 2022–35.

[10] Z. Yu *et al.*, "Federated Learning Based Proactive Content Caching in Edge Computing," *Proc. IEEE GLOBECOM*, 2018, pp. 1–6.

[11] J. Park *et al.*, "Wireless network Intelligence at the Edge," *Proc. IEEE*, vol. 107, no. 11, 2019, pp. 2204–39.

[12] K. Tantayakul, R. Dhaou, and B. Paillassa, "Mobility Management with Caching Policy Over SDN Architecture," *Proc. IEEE Conf. Network Function Virtualization and Software Defined Networks*, 2017, pp. 1–7.

[13] V. Balasubramanian, M. Aloqaily, and M. Reisslein, "FedCo: A Federated Learning Controller for Content Management in Multi-Party Edge Systems," *2021 30th Int'l. Conf. Computer Commun. and Networks*, 2021, pp. 1–9.

[14] R. R. Fontes *et al.*, "Mininet-WiFi: Emulating Software-Defined Wireless Networks," *2015 11th Int'l. Conf. Network and Service Management*, Nov 2015, pp. 384–89.

[15] C. Xu *et al.*, "Optimal Information Centric Caching in 5G Device-to-Device Communications," *IEEE Trans. Mob. Comp.*, vol. 17, no. 9, Sept. 2018, pp. 2114–26.

Biographies

Venkatraman Balasubramanian [S'19] (vbalasubramanian@asu.edu) received his MS degree in computer engineering from the University of Ottawa, Ontario, Canada, in 2017. He is currently working toward a Ph.D. degree in the School of Electrical and Computer Engineering, Arizona State University.

Moayad Aloqaily [S'12, M'17] (maloqaily@ieee.org) received his Ph.D. degree in electrical and computer engineering from the University of Ottawa in 2016. He is currently with the Cybersecurity Program, Al Ain University, United Arab Emirates. His current research interests include applications of AI and ML, connected and autonomous vehicles, blockchain solutions, and sustainable energy and data management. He is an ACM member and a Professional Engineer Ontario (P.Eng.).

Martin Reisslein [S'96, M'98, SM'03, F'14] (reisslein@asu.edu) is a professor in the School of Electrical, Computer, and Energy Engineering at Arizona State University, Tempe. He currently serves as Associate Editor for *IEEE Transactions on Mobile Computing*, *IEEE Transactions on Education*, and *IEEE Access*.

Anna Scaglione [M.Sc.'95, Ph.D.'99, F'11] (ascaglio@asu.edu) is currently a professor in electrical and computer engineering at Arizona State University. Her research is rooted in statistical signal processing and spans many disciplines that relate to network science, including communication, control, and energy delivery systems.