On-device Computational Caching-Enabled Augmented Reality for 5G and Beyond: A Contract Theory-Based Incentive Mechanism

Tri Nguyen Dang, Kitae Kim, Latif U. Khan, S. M. Ahsan Kazmi, Member, IEEE, Zhu Han, Fellow, IEEE, Choong Seon Hong, Senior Member, IEEE

Abstract—Recently, we have witnessed an increasing demand in augmented reality (AR) based Fifth-Generation (5G) and beyond applications, such as smart gaming, smart navigation, smart military wearable, and smart industries. These AR-based applications require on-demand computational and caching resources with low latency that can be provided via Multi-access Edge Computing (MEC) server. However, due to the massive growth of AR-enabled devices, the MEC server resources might be insufficient. To overcome this challenge, we can utilize the computational and caching resources of user equipment (UE) to serve the other UEs in its close vicinity. Successfully enabling such interaction among devices requires an attractive incentive mechanism. Therefore, we propose a contract theory-based incentive mechanism for enabling on-device caching for AR-based applications. In our approach, the MEC offers a reward to the UE for providing its resources (i.e., storage capacity, power, etc.). Furthermore, under the information asymmetry problem, we derive an optimal mechanism via the contract theory for enabling on-device caching subject to the individual rationality and incentive-compatible constraints. Finally, we perform numerical evaluations to validate the effectiveness of our proposed scheme.

Index Terms—Augmented reality, contract theory, computational caching.

I. INTRODUCTION

Recently, augmented reality (AR) has been witnessed to pave off the way towards the development of based Fifth-Generation (5G) and beyond smart applications, such as smart navigation, smart gaming, smart industries, smart military wearable, and smart infotainment [1]–[5]. The adoption of AR in various smart applications significantly increases the AR traffic. The AR-based devices communicate with the remote cloud via telecommunication networks for their operation. The massive number of AR-based devices require ultra-high transmission rates, and thus, put significant overhead on communication networks. For instance, consider Pokemon Go which uses mobile devices and AR for enabling an immense gaming experience to user equipments’ (UEs). Enabling of the Pokemon Go requires high-computational capability and high-quality connectivity. However, mobile devices have limited computational power and backup power. One way of enabling Pokemon Go on a mobile device is to offload its tasks to a remote cloud. However, this approach will suffer from the issues of high-latency and significant overhead on the back-haul network. On the other hand, edge computing can enable numerous smart Internet of Things (IoT) applications by offering on-demand computing resources with extremely low-latency [6], [7]. Therefore, edge computing seems to be a preferable solution to enable Pokemon Go by providing on-demand computational resources with low-latency. Furthermore, the processing of computationally expensive tasks at the network edge improves the network throughput by minimizing overhead on the back-haul links.

Another example of AR-based smart applications is smart remote live support. The smart remote live support consists of coping with insolvable machine error using AR and video transmission with a remote expert. The process starts with a machine operator wearing AR glasses which is followed by sending requests to remote experts for help. Additionally, a live video stream is sent by the AR glasses to the expert to enable assessment of the situation. The remote expert adds annotations in the video via AR to assist the machine operator in solving the problem. AR applications use different solutions to add annotations of the remote expert at the proper locations of the three-dimensional (3D) objects. The primary challenge in enabling correct positioning of annotations on 3D objects is the high computational power requirement by AR algorithms. However, AR is generally portable and has insufficient computational power. On the other hand, strict latency is another factor that must take into account while designing an AR-based remote live support system. Therefore, it is indispensable to use edge computing for processing computationally expensive AR algorithms with low-latency. The primary objective of using edge computing in AR-based applications is to strategically cache the contents required for their operation [8]. Caching at edge servers significantly reduces the access delays [9]. However, communication with

Authorized licensed use limited to: Kyunghee Univ. Downloaded on May 19,2021 at 10:21:22 UTC from IEEE Xplore. Restrictions apply.
the edge server might face a challenge in the unavailability of communication resources or insufficient edge server resources [10]. Additionally, transmission latency is more for communication between the device and edge server computing server than device-to-device (D2D) communication [11]. This striking feature of ultra-low latency of D2D communication opens up the opportunity of using on-device caching for AR-based applications. However, there must be some attractive incentive mechanism to motivate the participation of devices in providing on-demand caching to other devices for enabling AR-based applications.

There are various approaches for designing an incentive mechanism for on-device caching to enable AR/VR applications. For instance, contract theory, auction theory, bargaining game, and Stackelberg game, among others, can be used for incentive mechanism design. However, most of these approaches (i.e., Auction theory, Nash Bargaining game, and Stackelberg game) are based on iterative mechanism, which has limitations of long convergence time and requirement of multiple information exchanges between agents and principal [12]. To avoid these limitations, one can use contract theory to design an efficient incentive mechanism. In this work, we design an incentive mechanism based on contract theory to enable on-device caching for supporting AR applications. In the designed approach, the base station (BS) offers a reward to the UEs for providing its resources (i.e., storage capacity, power). The reward can be given as free data or any monetary benefit [13]. Depending on the reward from the BS, UEs decide to participate in caching for AR. After the participation decision by the UEs, BS offers a proportional reward to the amount of effort of UEs (i.e., power consumption for transmitting the cached content). In the contract theory-based incentive mechanism, the BS might not have all the UEs information (i.e., evaluation function) due to privacy concerns. Furthermore, UEs might not have knowledge of certain information such as content popularity and request rate. Such information is generally available at the BS. The aforementioned information at each side such as the BS and UE, are considered private and unknown to the other side. Such information asymmetry poses significant challenges to incentive mechanism design. Therefore, it is necessary to handle the information asymmetry in incentive mechanism design. Although the BS is not aware of the exact UE evaluation function, it has prior knowledge about property pertaining to the evaluation function such as strictly concave and non-decreasing nature. Therefore, one can use such type of information in incentive mechanism design.

Our key contributions can be summarized as follows:

- We present a novel system model that enables on-device computational caching in AR-based applications and serving AR service via D2D communication. Then, we formulate an optimization problem for maximizing the social welfare of the network concerning on-device caching capacities. Obtaining an optimal solution for this problem is extremely difficult due to the large size of constraints, and information asymmetry.
- Then, we propose an equivalent problem and a solution mechanism based on contract theory with information asymmetry, i.e., content popularity, UE request rate, and evaluation function.
- We derive an optimal contract scheme for on-device caching based on individual rationality and incentive-compatible constraints. Moreover, we present the numerical results to validate the superiority of our proposal under different scenarios.

The rest of the paper is organized as follows. Section II outlines and provides a critical analysis of the recent literature of edge computing-based AR, contract theory, and on-device caching. System model and problem formulation are discussed in Section III. In Section IV, a contract theory-based model is proposed. Finally, numerical results are presented in Section V and paper is concluded in Section VI.

II. RELATED WORKS

Several papers [8], [15]–[19] considered jointly cloud/edge computing and AR. A summary of the related works is given in Table I. Huang et al. in [15] proposed the CloudRidAR framework for mobile AR (MAR). The CloudRidAR uses the remote cloud for providing computing capabilities to MAR devices due to their low computational power. However, the inherent latency of the cloud might not be desirable for numerous latency-sensitive AR-based smart applications. In [16], the authors studied task offloading for AR applications in a multiple UE scenario under latency and power constraints. An optimization problem is formulated to minimize the energy required for offloading the AR task to an edge computing server. Furthermore, a successive convex approximation (SCA) based scheme is proposed to solve the optimization problem.

In [17], Schneider et al. proposed edge computing-based architecture for enabling AR through offloading of computationally expensive algorithms to high-end PC acting as an edge node. A use case of remote live support is considered to allow experts in assisting machine operators from remote locations. An acceptable end-to-end latency of 50ms considering tracking, receiving, sending, and the annotation delay for a 752 × 480 compressed video frame is computed using a real-time implementation of their architecture. The primary limitation of the authors’ work lies in the non-useage of caching with AR. In another study in [19], a novel protocol, namely dynamic adaptive AR over the edge (DARE) has been proposed to provide mobile nodes with the capability of changing their AR configurations subjected to variable channel conditions. This adaptive nature of the proposed algorithm makes it more attractive for adaptation in different smart applications. Finally, the authors tested their proposed protocol for the small-scale test-bed. In [17] and [19], the authors did not consider on-device caching for AR that can be used to improve the performance further.

In [18], Ren et al. proposed a hierarchical computation architecture for multi-access edge computing (MEC) enabled AR. The proposed architecture consists of three layers: the UE layer, the edge layer, and the cloud layer. The two layers, such as the UE layer and cloud layer are similar to traditional AR architecture. The novelty of their architecture lies in the introduction of the virtualized controller, operation...
In 

On the other hand, D2D communication studies have shown significant improvement in the network throughput and enhanced communication resource re-usability in access networks [11], [28], [29]. The works in [25]–[27] considered socially-aware D2D communication for throughput enhancement. Generally, a social group of UEs have similar interests and can trust each other. UEs located within a social group can serve other UEs of its group to simultaneously reduce traffic overhead at the base station, and increase the device’s throughput. However, there must be some attractive incentive mechanisms to ensure device participation. There are two aspects of D2D communication: social-pairs formation for D2D communication in cellular networks is proposed. Additionally, a case study for validation of handling uncertainties and the effectiveness of contract design is presented. In [22] the cooperative spectrum sharing between a single primary user and secondary users is proposed under the information asymmetry problem. Moreover, they achieved a sub-optimal solution via a decompose-and-compare approximate algorithm. In [23], the incentive mechanism is proposed for reward-based collaboration in which the total reward is sharing between multiple collaborators. The work in [24] has proposed a framework for sponsored content sharing for user/subscribers without considering the impact of D2D communication in the networks.

Numerous papers considered contract theory incentive mechanism design [13], [14], [20]–[24]. In [14], an overview of the concepts and models of contract theory is presented. A comparison between contract theory with other methods of economics was discussed. Furthermore, they presented an incentive mechanism design for spectrum trading, mobile crowdsensing, and traffic offloading. The work in [13] discussed the incentive mechanism design for D2D communication. On the other hand, a contract theory-based incentive mechanism for resource trading in a small-cell caching system has been proposed in [20]. In [21], a contract theory-based incentive platform, and communication unit based on software-defined networking in the edge layer for seamless communication with the UE and cloud layers. Apart from architecture, a mechanism for simultaneously supporting numerous AR applications is presented. The key technologies required to enable AR via edge computing are also discussed. Finally, the proposed architecture is tested using simulation results. Sun et al. in [8] considered computation, communication, and caching for mobile virtual reality (MVR). The authors proposed a framework that consists of a single cache and computational resources enabled MVR device, an edge computing server, and a remote cloud. The prime limitation of their work lies in considering a single-UE scenario. In this paper, we consider multiple UEs scenario with on-device caching and computation.

### TABLE I: Comparison of related works with proposed work.

<table>
<thead>
<tr>
<th>Reference</th>
<th>AR/VR</th>
<th>Incentive Mechanism</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun et al., [8]</td>
<td>✓</td>
<td>☑</td>
<td>Considered computation, caching, and communication for AR.</td>
</tr>
<tr>
<td>Zhang et al., [13]</td>
<td>×</td>
<td>☑</td>
<td>Proposed an incentive mechanism for D2D communication.</td>
</tr>
<tr>
<td>Zhang et al., [14]</td>
<td>×</td>
<td>☑</td>
<td>Contract theory-based modeling was discussed.</td>
</tr>
<tr>
<td>AR. Huang et al., [15]</td>
<td>✓</td>
<td>×</td>
<td>Proposed a CloudRadar framework for mobile AR.</td>
</tr>
<tr>
<td>Al-shuwaili et al., [16]</td>
<td>✓</td>
<td>×</td>
<td>Considered task offloading for AR.</td>
</tr>
<tr>
<td>Schneider et al., [17]</td>
<td>✓</td>
<td>×</td>
<td>Task offloading of AR contents to edge.</td>
</tr>
<tr>
<td>Ren et al., [18]</td>
<td>✓</td>
<td>×</td>
<td>Proposed three layers architecture for MEC-based AR.</td>
</tr>
<tr>
<td>Liu et al., [19]</td>
<td>✓</td>
<td>✓</td>
<td>Proposed a dynamic adaptive AR framework using edge computing was proposed.</td>
</tr>
<tr>
<td>Liu et al., [20]</td>
<td>×</td>
<td>✓</td>
<td>Presented an incentive mechanism for resource trading in a small-cell caching system.</td>
</tr>
<tr>
<td>Chen et al., [21]</td>
<td>×</td>
<td>✓</td>
<td>Contract theory-based incentive mechanism for D2D communication was proposed.</td>
</tr>
<tr>
<td>Duan et al., [22]</td>
<td>×</td>
<td>✓</td>
<td>Spectrum sharing between primary user and secondary users.</td>
</tr>
<tr>
<td>Duan et al., [23]</td>
<td>×</td>
<td>✓</td>
<td>Data acquisition and distributed computing applications.</td>
</tr>
<tr>
<td>Xiong et al., [24]</td>
<td>×</td>
<td>✓</td>
<td>Sponsored content management, and interactions among the mobile network operator, the sponsored content provider and end users</td>
</tr>
<tr>
<td>Li et al., [25]</td>
<td>×</td>
<td>×</td>
<td>Socially-aware D2D communication.</td>
</tr>
<tr>
<td>Zhao et al., [26]</td>
<td>×</td>
<td>×</td>
<td>Socially-aware D2D communication.</td>
</tr>
<tr>
<td>Zhang et al., [27]</td>
<td>×</td>
<td>✓</td>
<td>Socially-aware D2D communication.</td>
</tr>
<tr>
<td>Proposed</td>
<td>✓</td>
<td>✓</td>
<td>Computational results caching for AR service. AR contents sharing via D2D communication. Spectrum sharing between cellular users and D2D users.</td>
</tr>
</tbody>
</table>
communication and incentive mechanism design. Motivated by the incentive mechanism design for D2D communication in cellular networks in [13], we propose an incentive mechanism for on-device caching for serving other devices within a social group. To the best of our knowledge, we are the first to consider incentive mechanism design using contract theory for enabling on-device caching to serve different AR-based applications.

III. System Model and Problem Formulation

A. System Model

We consider a system (shown in Fig. 1) that includes a single base station (BS) equipped with a MEC server that has a fixed storage capability $S$. This BS serves a set $\mathcal{N}$ of $N$ user equipment (UE) that demand AR services under its coverage. Moreover, we assume in this model that each UE has a limited caching capability that can store only one content from a set of the AR contents $\mathcal{B} = \{b_1, b_2, \ldots, b_M\}$. The AR content of UE $i \in \mathcal{N}$ is represented by an positive tuple $b_i = \{s_i, c_i\}$, where $s_i$ is the size of content $i$, $c_i$ is the requirement in terms of CPU cycles to process a bit of AR content $i$. In AR-based applications, one typical requirement is to provide high computational resources to achieve low-latency. Latency can be reduced if a system can cache frequently requested contents which are then used to enhance the performance of AR-based applications [30]. Indeed, the edge computing paradigm can be used to enable on-demand computational resources and caching. However, the UEs might not be able to access the MEC server due to limited communication or computation resources. An alternative promising approach can enable D2D communication among AR-enabled devices and on-device caching. D2D communication can overcome this challenge by reusing the already occupied bandwidth of other UEs while protecting them by keeping the interference level below a maximum allowed limit. Therefore, in this work, we aim to use on-device caching and D2D communication to enhance the performance of AR-based applications further. Consider a typical example like Pokemon Go, in which the end-users are using their device to capture a real world video where the virtual Pokemon character appears. Then, the end-user can collect that Pokemon into their collection. The complete process can be described in three separate steps. In the first step, the end-user captures an image or video of a real world environment. In the second step, the computing server analyses the environment and location then generating 3D objects, i.e., Pokemon character and mapping them together. Finally, in the third step, the computing server feedback the required output to the end-user via a wireless link. However, we can neglect the second step if the results are computed once beforehand by adopting the idea of the Content Centric Network (CCN) in which results of a task can be cached based on the popularity itself. Moreover, the results are able to be cached at both the MEC side or UE sides. If a result is cached at the MEC side, we can reduce the power consumption for content analysis process and computational delay. On the other hand, if the result is cached on the UE side, we can improve the QoS by utilizing the high bandwidth and the low-latency link of D2D communication. This advantage of utilizing UE caching capacity and enabling D2D communication among UEs not only enhanced the QoS but also significantly increases the cache capability of the overall network.

In order to cache the content at UEs, it is indispensable first to compute the popularity of the contents. In our model, we compute the popularity of the contents via the Zipf distribution. Let $p = \{p_1, p_2, \ldots, p_M\}$ be the popularity distribution of $M$ tasks and $p$ be modeled by the Zipf distribution with parameter $\alpha$ as follows:

$$p_i = \frac{1/i^\alpha}{\sum_{m=1}^{M} 1/m^\alpha}, \quad (1)$$

where $\alpha$ is a positive value that represents the task’s popularity.
The larger the value of $\alpha$, the higher the value of $p$ will be observed and vice versa. Next, we discuss another challenge pertaining to the device availability in the BS coverage. Each AR-enabled device $i \in \mathcal{N}$ has its own independent behavior based upon its requirements for staying in a BS coverage. Such types of mobile devices’ behavior are typically unknown and unpredictable which might result in the unavailability of on-device cached content at times. Therefore, we assume the BS tackles such on-device content unavailability cases. In the next section, first, we discuss the computational caching model employed in our study.

B. Communication caching model

The performance of the D2D communication can be reflected by the achievable rate between D2D pairs. This can be controlled by the transmitter power of the D2D to achieve a required data rate. Moreover, in this paper, we assume an underlay model of D2D communication, i.e., re-use resources occupied by a cellular UE. Therefore, the signal to interference plus noise ratio (SINR) $\Gamma_i$ of D2D pair $(i)$ on a specific resource block (RB) $r$ for an underlay model is given as follows:

$$\Gamma_i^r = \frac{\delta_i^r g_i^r}{\Gamma_0^r + \Gamma_D^r + I_0},$$

(2)

where $I_0^r = \delta_{r,0}^r g_{r,0}^r$ is the interference from the cellular transmitter, $\Gamma_D^r = \sum_{k \neq i} \delta_{r,k}^\Omega g_{r,k}^r$ is the interference from the other D2D pair using the same channel as D2D pair $i$, and $I_0$ is the additive white Gaussian noise (AWGN). $\delta_{r,i}^\Omega, \delta_{r,k}^\Omega, \delta_{r,i}^r$ is the transmit power of cellular UE $i_0$, D2D UE $k$ and D2D UE $i$, respectively. Similarly, $g_{r,i}, g_{r,k}, g_{r,i_0}$ are the channel power gain between D2D pair $i$, cellular UE $i_0$, and D2D pair $k$, respectively. Moreover, we assume that the AR service requires a minimum data rate given by $\beta_{min}$, and a system bandwidth $W$ for successful service delivery. Then, all D2D transmitters have to adjust their transmit powers to meet the minimum requirement of rate. Based on (2), we can measure the transmit power of a D2D pair $i$ as follows:

$$\delta_i^r = \frac{I_0^r + \Gamma_D^r + I_0}{g_i^r} \left(2^{\beta_{min}/W} - 1 \right).$$

(3)

On the other hand, the SINR of the cellular user occupying RB $r$ which is reused by D2D UEs is given by

$$\Gamma_{cu}^r = \frac{\delta_{cu}^r g_{cu}}{\sum_{i \in \Omega_r} \delta_i^r g_{r,i} + I_0}.$$  

(4)

C. Computational Caching Model

In this work, we assume that in a fixed service area, there are $N$ AR contents, and each content is available at a single device that has cached it after its request to the BS in the previous time-slots. The specification of the AR content is dependent on its location, resolution, etc. In Fig. 2, at time $t_0$, UE A requests AR services for content 1. Initially, there is no result cached at the MEC as well as at the other devices. Therefore, the MEC needs to process the requested content and provide the result to UE A. Here, we assume that UE A has enough capability to cache the AR content of content 1. In the next time slot $t = 1$, UE B also requests the AR services for content 1. Since the content was cached in UE A, then, the BS informs that UE A can serve the request of UE B via the D2D communication link. Typically, the total number of contents in the service area is larger than the cache capability of a MEC, meanwhile utilizing UEs to cache contents is a promising solution to increase the number of cached contents in the service area. Then, we propose a system model in which the AR content can be cached either at the MEC cache storage or at the UEs. However, UE also has a limitation on its storage capacity, and therefore, in this model, we assume that a UE can cache at most one content at a given time, and the MEC has the information of these cached contents based on the previous UE’s request history. Hence, the MEC carefully chooses potential UEs to participate in this system by aiming to neglect UEs that have less power capability or storage capacity.

The performance of computational caching can be measured over the end-to-end delay, power consumption, cache hit ratio or a combination of the aforementioned factors. Intuitively, the power consumption for serving an AR content $i$ can be calculated as follows:

$$E_i = \begin{cases} \delta_i(s_i/\beta_{min}), & \text{If result is cached at UE}, \\ \delta_i(s_i/\beta_{min}), & \text{If result is cached at MEC}, \\ \delta_i(s_i/\beta_{min}) + \kappa h_i^2 s_i, & \text{If result is not cached}, \end{cases}$$

(5)

where $\kappa$ is a constant of energy consumption that depends on the CPU architecture, $h_i$ is the computational resources allocated to process task $i$. On the other hand, the cache utility of an AR task $i$ in term of latency is represented as follows:

$$L_i = \begin{cases} \varphi_i^c + \rho_i^c, & \text{Without caching}, \\ \varphi_i^c, & \text{If result is cached at the MEC}, \\ \varphi_i^d, & \text{If result is cached at UE}, \end{cases}$$

(6)

where $\varphi_i^c = s_i/\beta_{min}$ is the transmission delay of MEC, $\rho_i^c = s_i c_i/h_i$ is processing delay of MEC, and $\varphi_i^d = s_i/\beta_{min}$ is the transmission delay of D2D pair. It is evident that both (5), and (6) functions are linear with respect to $s_i$; however, they have different units, i.e., time and power consumed. Hence, we can rewrite the utility into a linear function. Without loss generality, we use a logarithm form to represent the utility of a MEC that aims to balance between power consumed and the...
time required. Therefore, let \( u(b_i) \) denote the utility associated with content \( i \). Then, \( u(b_i) \) can be calculated as follows:

\[
u(b_i) = \eta_i \log(s_i),
\]

where \( \eta_i \) is the associated coefficient to control parameters such as power, data rate, CPU cycles, etc. On the other hand, the UE request pattern within the coverage range of BS is assumed to follow the Poisson distribution with parameter \( \lambda \). Intuitively, the UE request pattern at location is uniformly distributed over \( N \) and is represented as \( \lambda_i = \lambda/N \). Hence, the probability of \( k \) requests within duration \( t \) can be calculated as follow:

\[
Pr[X = k|\lambda_i, t] = \frac{(\lambda_i t)^k \exp(-\lambda_i t)}{k!},
\]

where \( k \) is any positive integer value, \( k! \) represents the factorial of \( k \). For any given AR content, we assume that the utility function is proportional to its size. Next, we define the profit function based on two parameters, i.e., the task size, and the popularity as \( u(p_i, b_i) = p_i u(b_i) \). The utility of any cached content \( i \) associated with the number of time slot \( t \) is given as:

\[
U(\lambda_i, s_i, p_i, t) = \sum_{k=1}^{K} Pr[X = k|\lambda_i, t]k p_i u(b_i).
\]

Based on the Taylor series and the mean of the Poison distribution, we can derive the utility function associated with the number of time slot as follows:

\[
U(i, t) = tp_i \lambda_i u(b_i), \forall i \in N.
\]

Note that in a practical setting, a BS is unaware of the time a device will be activated under its coverage range, and thus, the utility of a BS becomes a conjecture. Typically, a device remains active for a limited amount of time in which it completes its required task. Thus, an approach is required to motivate devices to remain in the network to support other devices in its vicinity via D2D communication. In the next section, we propose a solution based on contract theory that aims to motivate devices to cache and remain active as long as possible such that both the utility of a BS and AR-enabled devices are maximized.

IV. CONTRACT THEORY MODEL AND PROPOSED SOLUTION

In this section, we present our approach which aims to maximize the utility of both the BS and AR-enabled devices based on contract theory. Note that, the set of UEs \( N \) in our network have heterogeneous capabilities: storage capacity, connectivity time with the BS. Moreover, this type of information is only available on the UE side. However, this type of information is needed for the design of our incentive mechanism by the BS. To motivate the UEs for offering their services (i.e., computation and caching) for AR-based applications, the BS will specify a bundle of contract \( \{U(R), R\} \). Where \( R \) is the payoff to the device for using its resources and \( U(R) \) represents the expected benefit gained by the BS. This function \( U(R) \) is strictly increasing functions with respect to \( R \), which means any device that has more expected benefit value will get more reward, and vice versa. Next, we design the contract types based on different UE types in our model.

A. User Equipment Types

We consider the UEs type depending on their availability under the coverage area of the BS. The UE availability is proportional to its mobility. A UE with high mobility has a shorter activation time and vice versa. We assume a group of UEs are active until specific time slots and are classified to be in the same type. Therefore the type of UE depends on the number of time slots it remains active in the BS coverage. Let say a UE belongs to a group type \( \theta_i \) if and only if the number of total active time slots is \( t \). The BS does not know exactly how long UEs will be activated in its coverage range. Therefore, the BS needs to design multiple contracts with the expected maximum number of time slots \( T = \{1, ..., t, ..., T\} \), where \( t \) can be represented in units of minutes, hours, etc. As we have \( N \) devices in our network caching AR content in its storage, each device is supposed to have different behaviors in terms of being active in the next \( T \) time slots. Furthermore, based on (10), we can clearly see that the cached content utility function can be rewritten as follows:

\[
U(i, t) = tp_i \lambda_i u(b_i) \equiv \theta_i p_i \lambda_i u(b_i) = U(i, \theta_i). \quad (11)
\]

Then, for each device \( i \), the BS will offer \( T \) different contract bundles \( \{U(R(i, \theta_i)), R(i, \theta_i)\} \). For simplicity, we assume that the total expected time slot that a UE be active in the coverage range of the BS is equally likely \( T_1 = T_2 = ... = T_N = T \).

**Remark 1.** For each content \( b_i \), the BS will design a contract based on the UE type represented as:

\[
\theta_i = \{\theta_{i,1}, ..., \theta_{i,t}, ..., \theta_{i,T}\}, \quad (12)
\]

and, the UE type of all contents represented by following:

\[
\theta = [\theta_1, \theta_2, ..., \theta_N]^T. \quad (13)
\]

**Definition 1.** The type of UEs are sorted in ascending order

\[
\theta_{i,1} < ... < \theta_{i,t} < ... < \theta_{i,T}, \forall t \in T, \forall i \in N. \quad (14)
\]

The type of UE depends upon the total number of time slots its remains to activate state. If the UE is active for less time slot, it will be associated with a lower type, and vice versa.

It must be noted that we consider single-dimensional UE types in this work. However, our proposal can be easily extended to multi-dimensional UEs types by performing dimensional reduction via the weighted sum method [31], [32]. Furthermore, the BS offers a set of contract bundles to UE e.g., \( T \) contract bundle. The UE can either choose one of these contract bundles to participate in or reject the contract offered. In case of UE declining to participate, we can assume that a UE has signed a contract of \( U(i, 0), R(i, 0) \). Representing the UE and BS will be awarded \( R(i, 0) \equiv U(i, 0) \equiv 0 \) utility. Next, we define the BS and UE equipment utilities for designing the contract in our model.
B. BS Utility

The utility function of content is linear with respect to the content’s size, its popularity, and the availability of a UE. Thus, the utility of the BS is given by the expected utility of cached content, e.g., $U_{i,t}(\theta_{i,t})$ minus for the cost (e.g., reward for UE) for utilizing UE’s resources. Then, the utility of the BS for a given content $i$ at the type $\theta_{i,t}$ is:

$$U_{BS}(i, \theta_{i,t}) = U(i, \theta_{i,t}) - \gamma R(i, \theta_{i,t}),$$

(15)

where $R(i, \theta_{i,t})$ is the incentive a BS has to pay to the UE $i$ for its service respected to $\theta_{i,t}$. This incentive can be in the form of any privileged access, such as an amount of free data [13]. As there are $T$-types of UEs for any content $i$, the expected utility of the BS can be calculated as follows:

$$U_{BS} = \sum_{i=1}^{N} \sum_{t=1}^{T} \xi_{i,\theta_{i,t}} U_{BS}(i, \theta_{i,t}),$$

(16)

where $\xi_{i,\theta_{i,t}}$ is the probability that UE $i$ belongs to type $\theta_{i,t}$.

C. UE Equipment Utility

The utility of UE $i$ with type-$t$ ($\theta_{i,t}$) is associated with a contract bundle $\left[ U \left( R(i, \theta_{i,t}) \right), R(i, \theta_{i,t}) \right]$ denoted as:

$$U_{UE}(i, \theta_{i,t}) = \theta_{i,t} \nu \left[ R(i, \theta_{i,t}) \right] - \rho U \left( i, \theta_{i,t} \right),$$

(17)

where $\nu[\cdot]$ is the self evaluation function [14] reflecting the rewards which are strictly concave, increasing function of $R(i, \theta_{i,t})$, where $\nu(0) = 0$, $\partial \nu[ R(i, \theta_{i,t}) ] / \partial R(i, \theta_{i,t}) > 0$ and $\partial^2 \nu[ R(i, \theta_{i,t}) ] / \partial^2 R(i, \theta_{i,t}) < 0$. And, $\rho$ is the additional effort put in by the UE according to the expected utility $U(i, \theta_{i,t})$. This can be represented as additional power consumed by the device to be active in the coverage area, or power consumption for cache maintenance. Given the above information, a UE needs to choose the bundle of contract that maximizes its utility.

D. Social Welfare

Next, we design the social welfare of the proposed approach. In this work, a UE caches at most one content, and belong to only one type of contract $\theta_{i,t}$. For simplicity, we assume that $\rho = 1$, and $\xi_{i,\theta_{i,t}}$ is uniform over $T$, then, $\sum_{t=1}^{T} \xi_{i,\theta_{i,t}} = 1, \forall i \in N, \forall t \in T$. Hence, the social welfare can be defined as the summation of both the BS and UEs utilities as follows.

$$\Pi = \sum_{i=1}^{N} \sum_{t=1}^{T} \left( U_{BS}(i, \theta_{i,t}) + U_{UE}(i, \theta_{i,t}) \right)$$

$$= \sum_{i=1}^{N} \sum_{t=1}^{T} \theta_{i,t} \nu \left[ R(i, \theta_{i,t}) \right] - \gamma R(i, \theta_{i,t}) \right) \right).$$

(18)

E. Contract Feasible Conditions

We assume that the contract bundle of AR contents is identical and independent with each other. Therefore, the BS can separate the problem into $N$ explicit sub-problems each representing a content. For simplicity, we use $\nu \left[ R(i, \theta_{i,t}) \right]$ to denote $\nu \left[ R(i, \theta_{i,t}) \right]$. A contract bundle is called feasible if and only if two conditions hold: Individual Rationality (IR), and Incentive Compatibility (IC) [33], which are defined in the following definitions.

Definition 2. Individual Rationality (IR): The utility of a UE $i$ when participated in any contract bundle $\left[ U \left( R(i, \theta_{i,t}) \right), R(i, \theta_{i,t}) \right]$ must be non-negative,

$$U_{UE}(i, \theta_{i,t}) = \theta_{i,t} \nu \left[ i, \theta_{i,t} \right] - U(i, \theta_{i,t}) \geq 0, \forall i \in N, \forall t \in T.$$

(19)

This property of contract theory aims to motivate UEs to participate in caching contents and serving the demands of other UEs via D2D communication. If the utility is negative, UE will choose not to participate in the contract. Next, we formally define the next property, i.e., incentive compatibility, as follows:

Definition 3. Incentive Compatible (IC): For any UE $i$ of type $\theta_{i,t}$, the maximum utility a UE $i$ can achieve is at the designated contract bundle of type $\theta_{i,t}$:

$$U_{i,t} \nu \left[ i, \theta_{i,t} \right] - U(i, \theta_{i,t}) > \theta_{i,t} \nu \left[ i, \theta_{i,t} \right] - U \left( i, \theta_{i,t} \right),$$

(20)

$$\forall i \in N, t \in T, \theta_{i,t} \neq \theta_{i,t}'.$'

This property of contract theory states that a UE must choose the right contract which depends on its type to maximize its utility. Next, we need to define the monotonicity condition for the reward function.

Definition 4. Monotonicity: Given any content $i$, for any feasible contract bundle $[ U \left( R(i) \right), R(i) \right]$, the reward function must satisfy the following condition,

$$0 = R(i, 0) < R(i, \theta_{i,T}) < \ldots < R(i, \theta_{i,T}) < \ldots < R(i, \theta_{i,T}).$$

(21)

Monotonicity means that the higher the type of a UE, the higher its reward compared to the lower type. It is equivalent to stating that if a UE participated longer time in the network, it would get more benefit compared to the UE that participated for a shorter time. From the aforementioned conditions, we can formulate the problem of maximizing social welfare as follows:

$$\max_{\theta_{i,t}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \theta_{i,t} \nu \left[ R(i, \theta_{i,t}) \right] - \gamma R(i, \theta_{i,t}) \right).$$

s.t.:

C1: $\theta_{i,t} \nu \left( R(i, \theta_{i,t}) \right) - U(i, \theta_{i,t}) \geq 0,$

C2: $\theta_{i,t} \nu \left[ i, \theta_{i,t} \right] - U(i, \theta_{i,t}) \geq \theta_{i,t} \nu \left[ i, \theta_{i,t} \right] - U \left( i, \theta_{i,t} \right),$,

C3: $0 \leq R(i, \theta_{i,t}) < \ldots < R(i, \theta_{i,t}) < \ldots < R(i, \theta_{i,T})$,

$$\forall i \in N, \forall t \in T, \theta_{i,t} \neq \theta_{i,t}'.$'

(22)

The constraint (C1) represents the IR constraint, and the IC constraint is represented by (C2). The constraint (C3) represents the reward function which is an increasing function of $\theta_{i,t}$, e.g., monotonicity condition. The problem (22) has a very large size due to constraints (C2) (i.e., $N \times T \times (T - 1)$).
Lemma 1. (IR constraints reduction): Given any content $i$, for any feasible contract bundle $[U (R(i, \theta_{i,t})), R(i, \theta_{i,t})]$. If the IR constraint at type $\theta_{i,t} = \hat{1}$ is held, then the number of IR constraints can be reduced from $N \times T$ to $N$ constraints.

From Definition 2, and Definition 4, we can easily see that Lemma 1 always hold at type $\theta_{i,t} = 1$. Therefore, it is straightforward proof.

Lemma 2. (Local Downward Incentive Constraints): For any feasible contract bundle $[U (R(i, \theta_{i,t})), R(i, \theta_{i,t})]$ the utility of UE is satisfied with the following:

$$\theta_{i,t}, U(i, \theta_{i,t}) - U(i, \theta_{i,t}), T \geq t_t > t_1 \geq 0, \nu_{i,T} \geq \nu_{i,t} \geq 0.$$  \hfill (23)

If and only if,

$$\theta_{i,\theta_{i,t}}(i, \theta_{i,t}) - U(i, \theta_{i,t}) \geq \theta_{i,t}U(i, \theta_{i,t-1}) - U(i, \theta_{i,t-1}), \forall i \in N, 0 < t \leq T.$$  \hfill (24)

Proof: Following the IC constraint we have:

$$\theta_{i,t}, U(i, \theta_{i,t}) - U(i, \theta_{i,t}), T \geq t_t > t_1 \geq 0, \nu_{i,T} \geq \nu_{i,t} \geq 0.$$  \hfill (25)

where $t_1, t_2 \in T$. Based on (10) and strictly concave function $\nu()$, we can easily see that the equality only occurs if and only if $t_1 = t_2$. When $t_1 \neq t_2$, after some manipulations we can achieve the following:

$$\nu(i, \theta_{i,t}) (\theta_{i,t} - \theta_{i,t+1}) > \nu(i, \theta_{i,t+1}) (\theta_{i,t} - \theta_{i,t+1})$$  \hfill (26)

Since $t_1 < t_2$ then we divide both sides of inequality with a negative term ($\theta_{i,t} - \theta_{i,t+1}$), we get $\nu(i, \theta_{i,t}) < \nu(i, \theta_{i,t+1})$. Thus, for given any $t \geq 1$, we have

$$\theta_{i,t+1} [\nu(i, \theta_{i,t}) - \nu(i, \theta_{i,t-1})] \geq \theta_{i,t} [\nu(i, \theta_{i,t}) - \nu(i, \theta_{i,t-1})] \geq U(i, \theta_{i,t}) - U(i, \theta_{i,t-1}).$$  \hfill (27)

Hence, we have completed the proof for Local Downward Incentive constraints.

Lemma 3. (Local Upward Incentive Constrains): For any feasible contract bundle $[U (R(i, \theta_{i,t})), R(i, \theta_{i,t})]$ the utility of a UE is satisfied:

$$\theta_{i,t}, U(i, \theta_{i,t}) - U(i, \theta_{i,t}), T \geq t_t > t_1 \geq 0, \nu_{i,T} \geq \nu_{i,t} \geq 0.$$  \hfill (29)

If and only if,

$$\theta_{i,t}U(i, \theta_{i,t}) \geq \theta_{i,t}U(i, \theta_{i,t-1}) - U(i, \theta_{i,t-1}), \forall i \in N, 0 < t < T.$$  \hfill (30)

Proof: Similar to the proof of Lemma 2. For any $t > 1$, we can archive:

$$\theta_{i,t-1}U(i, \theta_{i,t-1}) \geq \theta_{i,t-1}U(i, \theta_{i,t}) - U(i, \theta_{i,t}),$$

$$\theta_{i,t-1}U(i, \theta_{i,t}) \geq \theta_{i,t-1}U(i, \theta_{i,t+1}) - U(i, \theta_{i,t+1}),$$

$$\theta_{i,t-1}U(i, \theta_{i,t}) \geq \theta_{i,t-1}U(i, \theta_{i,T}) - U(i, \theta_{i,T}).$$  \hfill (31)

Hence, we have completed the proof for the discrete case of the UEs types. On the other hand, for continuum case of the UEs types $\theta_{i,t} \in [\theta_{\min}, \theta_{\max}]$. Without loss of generality, we can approximate the infinity set $[\theta_{\min}, \theta_{\max}]$ into a finite set number by choosing a number $\epsilon$ such that $\epsilon$ is strictly positive, small enough and satisfies the condition $\theta_{i,t} - \epsilon < \theta_{i,t} < \theta_{i,t} + \epsilon$ where $\theta_{i,t} - \epsilon \geq \theta_{\min}, \theta_{i,t} + \epsilon \leq \theta_{\max}$. Therefore, the analysis and proofs presented for the discrete case pertaining to IR and IC constraints also hold for the continuum type case.

F. Optimal Contract Design

Based on the aforementioned lemmas and feasible conditions, we can rewrite the reduced problem of (22) as an equivalent problem as follows:

$$\max_{\theta, R} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \theta_{i,t}U \left[R(i, \theta_{i,t})\right] - R(i, \theta_{i,t}) \right) \text{ s.t.:}$$

$$C1: \theta_{i,1}U \left[R(i, \theta_{i,1})\right] - U(i, \theta_{i,1}) \geq 0,$$

$$C2: \theta_{i,t}U \left[i, \theta_{i,t}\right] - U(i, \theta_{i,t}) \geq \theta_{i,t-1}U(i, \theta_{i,t-1}) - U(i, \theta_{i,t-1}),$$

$$C2: \theta_{i,t}U \left[i, \theta_{i,t}\right] - U(i, \theta_{i,t}) \geq \theta_{i,t-1}U(i, \theta_{i,t}) - U(i, \theta_{i,t}),$$

$$\forall i \in N, \forall t \in T.$$  \hfill (32)

We can see that the constraint C1 is reduced from $O(T)$ to $O(1)$, and the constraint C2 is reduced from $O(T \times (T - 1))$ to $O(T)$. Since the number of IR and IC constraints are now reduced. The problem in (32) can be solved by using the Lagrangian Multiplier Method by treating the monotonicity constrain at a projection function [13]. Therefore, we propose a contract-based incentive mechanism as presented in Algorithm 1. First of all, the initialization of parameters is performed (line 3) using the input of Algorithm 1 (line 1). Then, the information broadcasting step is performed (lines 4), the BS broadcasts the values of popularity ($p$) and request rate ($\lambda$) to all the UEs. Next, the BS solves the problem.
Algorithm 1 On-device Caching-contract Based Incentive Algorithm

1: Input: $S, N, T, B, \eta, \lambda, \alpha, p, \theta$
2: Output: $[U(R(\cdot)), R(\cdot)]$
3: Initialization;
4: Broadcast information: BS broadcasts the information $\lambda_i$ and $p_i$ values for UEs;
5: Optimum Contract: BS solve problem (32);
6: UE Estimate Effort: UEs can estimate the effort it required to serve individually by the information of request rate, popularity and size of content;
7: Broadcast Contract: BS broadcasts bundle contract $\{U(R(i, \theta_1, \alpha)), R(i, \theta_1, \alpha)\}$ for all UEs $i \in N$;
8: Feedback Information: UEs accept or reject the contract $\{U(R(i, \theta_1, \alpha)), R(i, \theta_1, \alpha)\}$ based on its plan to stay under the coverage of the BS.
9: Contract Execution;
10: if A request hit content $i$ with in contract time $t$ then
11: UE $i$ will serve the demand of other UE by D2D communication;
12: end if

stated in (32) to develop the optimal contracts bundle (line 5). Then, based on the information of popularity, and request rate, the UEs estimate the effort for each contract ($b_i$) using (10) (line 6). After the estimation of the utility computation by all the UEs, the BS informs all the bundles’ contracts to all UEs (line 7). The UEs then decide to accept or reject the contract based on its plan (local information) to stay under the coverage area of the BS (line 8). Once the contract is accepted by a UE it serves other UEs if required as a representative of the BS. Then, if there is a requested cached content in the UEs vicinity, the BS will inform the requesting UEs to be served by D2D communication via the representative UE (lines 10-11). Note that the complexity of the proposed scheme (Algorithm 1) for solving problem (32) reduces from $O(T^2)$ to $O(T \log(T))$. Thus, the proposed scheme can easily find an optimal solution for problem (32) by using Lagrangian Multiplier Method [35]. Moreover, if the type of UEs is not static or the environment is dynamic, it will fall into the category of the Ex-Ante Contracting [36]. In such a case, there is a need to execute the proposed Algorithm 1 according to changing states of network.

V. NUMERICAL RESULTS

A. Simulation setup

To show the feasibility of the contract subject to the IR and IC constraints, we choose a network with $N$ ranging from $[10, 50]$ UEs. In this network, we consider to have the total number of contents $M = 100$, and each UE can cache a content. Therefore, the total number of contents that can be cached at devices is $B \equiv N$. Moreover, we evaluated the proposed solution for $T = 20$ time-slots. Next, the evaluation function of a UE $\nu(\cdot)$ is considered as a quadratic functions. The main parameters used in our simulations are presented in Table III. To validate our proposal, there is a need to make sure that the contact feasibility conditions (i.e., IR and IC) are fulfilled. Furthermore, there is no real dataset to justify the practicality of our approach. Therefore, we provide comprehensive numerical results to show the validity of our proposed approach. All the results are computed by taking the average of 100 runs. Next, we present the contract feasibility conditions in the next subsection.

B. Contract Feasibility

For any contract, the contract is called feasible if and only if the two contract feasibility constraints are holding e.g., Individual Rationality, Incentive Compatibility. Moreover, our objective is to maximize social welfare which guarantees the fairness of both sides, e.g., MEC and, UEs. Firstly, in Fig. 3a, we have shown that the social welfare of our proposed model (18) is strictly positive and increasing as $t$ increases. We can also observe that the proposed solution significantly outperforms the linear pricing model in terms of social welfare and achieves performance close to no information asymmetry. For instance, the social welfare of the proposed solution can achieve an average performance benefit of up to 65% compared to the linear pricing method. Similarly, no information asymmetry solution is about 8.7% higher than the proposed approach. Next, we validate the IR constraints.

1) IR constraints: In Fig. 3b, we validate the IR constraint. Here, we present the utility of UE by varying its types from 0 to 20. Note that the type of UEs in our model depends upon the number of active time-slots. Thus, the higher type of UE, the longer time it remains active in the network. As shown in the simulation results, our proposed approach achieved better performance compared to the linear pricing model. However, in the case of no information symmetry, the BS has complete knowledge about the UEs, e.g., UE types. Thus, BS uses a selfish approach by treating the reward for each UE which are the least positive or mostly close to zeros as represented in Fig. 3c. Hence, the utility of UE is the lowest and even worse than the linear pricing approach. Therefore, in terms of fairness, our proposal is a guarantee for UE in the case of information symmetry. Furthermore, we validate the utility of BS under different settings of UEs, e.g., UE types. The utility

<table>
<thead>
<tr>
<th>TABLE III: Simulation parameters.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Total number of AK content ($M$)</td>
</tr>
<tr>
<td>On-device caching UEs ($N$)</td>
</tr>
<tr>
<td>White-noise ($\lambda_0$)</td>
</tr>
<tr>
<td>System bandwidth</td>
</tr>
<tr>
<td>Bandwidth of each RB ($W$)</td>
</tr>
<tr>
<td>Number of subcarriers per RB</td>
</tr>
<tr>
<td>Neighboring subcarrier spacing</td>
</tr>
<tr>
<td>Maximum D2D range ($d_{max}$)</td>
</tr>
<tr>
<td>Number of UE type ($T$)</td>
</tr>
<tr>
<td>Path loss (D2D links)</td>
</tr>
<tr>
<td>System bandwidth</td>
</tr>
<tr>
<td>Bandwidth of each RB ($W$)</td>
</tr>
<tr>
<td>Maximum transmit power of D2Ds ($P_{max}$)</td>
</tr>
<tr>
<td>Interference threshold ($\zeta_{max}$)</td>
</tr>
<tr>
<td>the Zipf’s parameter ($\alpha$)</td>
</tr>
<tr>
<td>the Poisson’s parameter ($\lambda$)</td>
</tr>
</tbody>
</table>
utility of BS is also proportionally with the UE types. Additionally, it can be inferred that the UE that remains activated for longer time-slots will receive more benefit from BS compared to the shorter time UE. On the other hand, all contracts are feasible because the utility for both sides e.g., BS, and UEs, are strictly positive and equal to zero only at a contract type $\theta = 0$. This represents the IR constraint which is the necessary condition for the correct contract allocation between UEs and the BS.

2) IC constraint: Fig. 4 presents the type of UE vs. the reward. In this simulation, we vary the type of UEs to calculate its expected utility. It can be seen that if a UE chooses the wrong type, it is penalized by getting less utility compared to the achievable utility by choosing the right type. For instance, if a UE of type 5 chooses a contract bundle of type 5, it will get the maximum utility. Any deviation from the contract bundle type 5 to any other contract types results in a reduction of utility. A similar trend can be observed for other types of UEs as well, i.e., type 10, 15, 20 in Fig. 4. This represents the IC constraint and validates that the proposed solution strictly abides by the IC constraint condition.

3) System performance: In this section, we present the system performance of the proposed solution. We evaluated the performance of the cached contents by observing the hit rate, average delay, and average achievable rate for enabling D2D communication in our network. We have analysed the impact of D2D caching into cellular network as well by varying maximum interference threshold $I_{\text{max}}$ constraints for protecting cellular users.

4) Impact of On-device caching to cellular networks: Fig. 5a, we analyze the performance of D2D communication in which we choose the total number of RBs to be 15. In this simulation, we can show the impact on the cellular network by sharing its resources, i.e., RBs to one or more than one D2D pair. Our aim is to capture the effect of our proposal in dense settings, therefore we only choose to have 15 RBs. Based on the works in [11], we verify the average achievable rate of D2D users. We set $I_{\text{max}} = -80, -100, -120$ dBm explicitly, as shown in the figure, we can see that the system is achieving the highest data rate when $I_{\text{max}} = -80$ dBm due to tight protection for cellular users. In such a tight protection scenario, every D2D pair have to control its power carefully to keep the interference to cellular users as low as possible. Moreover, the cellular user transmits with a fixed power level on its occupied RB. Hence, the average achievable rate is the highest compared to the remaining cases. This happens when the interference experienced by cellular is increased as a result of reuse of RBs by D2D users. In Fig. 5b, we calculate the total delay by varying the number of UEs in our network. Moreover, we evaluated our proposed model with different setting of interference threshold for protecting cellular users, e.g., $I_{\text{max}} = \{80, -100, -120\}$ dBm. It can be seen that our proposal outperforms the linear model significantly. Fig. 5c, analyzes the hit rate of the cached contents by using Binomial distribution for 1,000 samples with the request rate $\lambda$ and content popularity $p$. Moreover, we compare the proposed solution, i.e., on-device caching with an approach in which on-device caching is not enabled. Similarly, Fig. 5c shows that increasing the number of UE participating in the network increases the total number of cached contents in the network. Therefore, it can improve the hit rate and achieve a higher utility. For instance, when the number of on-device caching UEs is $N = 50$, the hit rate is significantly higher compared to the case when the number of on-device caching UEs is $N = 40$, and $N = 30$, etc.

5) Impact of user request rate on proposed system: In Fig. 6, we analyse the cumulative distribution function (CDF) of the user request rate $\lambda$ and content popularity $p$. It seems that the payoff of UE not only depends on its activation time in the coverage of the network, it also depends on how popular content is, and the user request rate. As shown in the figure,
we vary the parameter $k$ which represent the expected number of request for a content with request rate of the user at BS is $\lambda$. We observe that the longer active time slots of UEs, the higher chance of content being requested. For instance, when $\lambda = 5; 0.1 \leq \lambda_i \leq 0.5, p_i = 0.5$ at least after $t = 10$ time slot, the content is requested with probability 1.0. Similarly, there would be a very high chance of content being requested close to 50% after 3 time slots as seen in Fig. 6a. Similarly for the other two cases $k = 3$ in Fig. 6b, and $k = 5$ in Fig. 6c, the longer is the active time slots, the higher is the chance of content be requested.

6) Nash Equilibrium: In this paper, we consider each device only caches one content requested by itself beforehand, thus, there exists no any competitor in this game. Therefore, we have only two players for each content, e.g., UE and BS. Hence, in order to calculate the NE, we just simply balance the utility of both side, i.e., UE and BS. By binding the IR, and IC in (32) into optimal solution for an equivalent problem $U_{UE}(i, \theta_{i,t}) = U_{BS}(i, \theta_{i,t})$. We have a trivial solution at $\theta_{i,t} = 0$ which cannot be taken as a solution due to the
assumptions of contract feasibility. Since the self evaluation function of UE is quadratic, and the utility of BS is linear on $\theta_{i,t}$. We can simply visualize a solution for a content with size $s_i = 50$ (MB). As shown in the figure, our solution in Fig. 7 not only guarantee the feasibility of IR and IC, it also achieves a higher solution value than the no information asymmetry case in Fig. 8.

7) Cache hit performance: In order to validate the efficiency of our proposed approach, we take an average simulation in which the same user request pattern $\lambda$ and content popularity $p$, the number of UE participate in on-device caching is assumed to be $N = 20$, the total number of contents is $100$, and the cache size of the MEC is $20\%$ of the total size of contents. It can be inferred from Fig. 9 that our proposed method has improved the average cache hit performance by $20\%$ compared to the traditional caching system model.

VI. CONCLUSIONS

In this paper, we have proposed an incentive mechanism for on-device caching to enable AR-based applications. A contract theory-based incentive mechanism has been proposed to model interaction between the devices and BS for enabling on-device caching under the information asymmetry. We have derived an optimal contract design for on-device caching and guaranteed both IR and IC constraints. The more the UEs participating in the caching process, the more the benefit achieved both by the BS and other UEs in terms of average utility, serving delay, and hit rate. Moreover, we have shown that contract theory with information symmetry significantly enhances social welfare compared to the no information asymmetry, and the linear pricing mechanism via the numerical results. Our approach can improve the average hit rate of the network up to $20\%$ compared to the caching system only at the BS. Moreover, our proposal also increases the spectrum efficiency by utilizing D2D communication for serving the request of users.

REFERENCES


Tri Nguyen Dang is currently pursuing his Ph.D. in Computer Science and Engineering under fully funded scholarship at Kyung Hee University, South Korea. He received his B.S. in Information Teaching from Hue University’s College of Education, Vietnam in 2014. His professional experience includes Mobile Application Developer, middle-ware programming. He won a consolation prize of Olympiad in ACM ICPC Vietnam in 2013. His research interest includes Network Optimization, and Mobile Cloud computing, Multi-access Edge Computing, Internet of Things.

Kitae Kim received the B.S. and M.S degrees in Computer Science and Engineering at Kyung Hee University, Seoul, Korea, in 2017, 2019 respectively. He is currently pursuing a Ph.D degree in the Computer Science and Engineering in Kyung Hee University, Seoul, Korea. His research interests include Multi-Access Edge Computing, Wireless Network, Unmanned Aerial Vehicle Communications and Machine Learning.

Latif U. Khan is currently pursuing his Ph.D. degree in Computer Engineering at Kyung Hee University (KHU), South Korea. He is working as a leading researcher in the intelligent Networking Laboratory under a project jointly funded by the prestigious Brain Korea 21st Century Plus and Ministry of Science and ICT, South Korea. He received his MS (Electrical Engineering) degree with distinction from University of Engineering and Technology (UET), Peshawar, Pakistan in 2017. His research interests include analytical techniques of optimization and game theory to edge computing and end-to-end network slicing.

S. M. Ahsan Kazmi is with the Institute of Information Security and Cyber Physical System, Innopolis University, Innopolis, Tatarstan, Russia, where he is currently working as an Assistant Professor. He received his Master degree in Communication System Engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2012, and the Ph.D. degree in Computer Science and Engineering from Kyung Hee University (KHU), South Korea. His research interests include applying analytical techniques of optimization and game theory to radio resource management for future cellular networks.

Zhu Han (S’01–M’04–SM’09–F’14) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively.

From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015-2018, AAAS fellow since 2019 and ACM distinguished Member since 2019. Dr. Han is 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of 2021 IEEE Kiyo Tomiyasu Award, for outstanding early contributions to game theory and distributed management of autonomous communication networks.”
Choong Seon Hong (S’95–M’07–SM’11) received the B.S. and M.S. degrees in electronic engineering from Kyung Hee University, Seoul, South Korea, in 1983 and 1985, respectively, and the Ph.D. degree from Keio University, Japan, in 1997. In 1988, he joined KT, where he was involved in broadband networks as a Member of Technical Staff. Since 1993, he has been with Keio University. He was with the Telecommunications Network Laboratory, KT, as a Senior Member of Technical Staff and as the Director of the Networking Research Team until 1999. Since 1999, he has been a Professor with the Department of Computer Science and Engineering, Kyung Hee University. His research interests include future Internet, ad hoc networks, network management, and network security. He was an Associate Editor of the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT and the IEEE JOURNAL OF COMMUNICATIONS AND NETWORKS. He currently serves as an Associate Editor of the International Journal of Network Management, and an Associate Technical Editor of the IEEE Communications Magazine.