Stochastic Optimization for Green Multimedia Services in Dense 5G Networks

TENGFEI CAO, CHANGQIAO XU, MU WANG, ZHONGBAI JIANG, and XINGYAN CHEN, Beijing University of Posts and Telecommunications
LUJIE ZHONG, Capital Normal University
LUIGI ALFREDO GRIECO, Politecnico di Bari

The many fold capacity magnification promised by dense 5G networks will make possible the provisioning of broadband multimedia services, including virtual reality, augmented reality, mobile immersive video, to name a few. These new applications will coexist with classic ones and contribute to the exponential growth of multimedia services in mobile networks. At the same time, the different requirements of past and old services pose new challenges to the effective usage of 5G resources. In response to these challenges, a novel Stochastic Optimization framework for Green Multimedia Services (SOGMS) is proposed hereby that targets the maximization of system throughput and the minimization of energy consumption in data delivery. In particular, Lyapunov optimization is leveraged to face this optimization objective, which is formulated and decomposed into three tractable subproblems. For each subproblem, a distinct algorithm is conceived, namely Quality of Experience (QoE) based admission control, cooperative resource allocation, and multimedia services scheduling. Finally, extensive simulations are carried out to evaluate the proposed method against state-of-art solutions in dense 5G networks.

CCS Concepts: •**Information systems** →**Multimedia streaming;** •**Networks** →**Network performance analysis;**

Additional Key Words and Phrases: Stochastic optimizing, green multimedia, admission control, resource allocation, dense 5G networks

ACM Reference format:
Tengfei Cao, Changqiao Xu, Mu Wang, Zhongbai Jiang, Xingyan Chen, Lujie Zhong, and Luigi Alfredo Grieco. 2018. Stochastic Optimization for Green Multimedia Services in Dense 5G Networks. *ACM Trans. Web* 1, 1, Article 1 (September 2018), 23 pages.
DOI: 0000001.0000001

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant Nos.61871048, 61872253, 61762074; by the BUPT Excellent Ph.D. Students Foundation CX2018208; by the Chunhui Plan of Ministry of Education of China No.z2016081.

Author's addresses: T. Cao, C. Xu , M. Wang, Z. Jiang and X. Chen, the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications; emails: {caotf, cqxu, wangmu, zbjiang, chenxingyan}@bupt.edu.cn; L. Zhong, Information Engineering College, Capital Normal University; email: zhonglj@cnu.edu.com; L. A. Grieco, Department of Electrical and Information Engineering, Politecnico di Bari; email: a.grieco@poliba.it.
Corresponding author: Changqiao Xu.

 1559-1131/2018/9-ART1 $15.00
DOI: 0000001.0000001

1 INTRODUCTION

Video streaming is increasingly dominant in mobile traffic and undergoing an explosive growth [1] [2]. According to the latest visual network index by Cisco [3], the proportion of mobile video traffic grew more than 60 percent of global mobile traffic in 2016, and is expected to account for over 78 percent of the world's mobile data by 2021. This large amount of mobile traffic will bring a huge pressure to the currently available access network infrastructures.

Luckily, the fifth generation (5G) of mobile communications technologies is behind the corner and promises a 1000-fold magnification of the system capacity and millisecond latencies [4] [5]. Thanks to this renewed availability of resources, the ongoing exponential increase of multimedia traffic in mobile networks will be enforced by broadband multimedia services [6] including: mobile immersive video [7], Virtual Reality (VR) [8], and Augmented Reality (AR) [9].

At the same time, it is worth to note that with the ultra dense Base Stations (BSs) deployment expected in 5G networks, a higher degree of complexity and heterogeneity will be injected in the communication system [10] [11]. In particular, it will be necessary to coordinate the activities of the many available BSs to attain an efficient usage of the system capacity while providing a satisfactory Quality of Experience (QoE) to end users [12] [13]. On top of these peculiar facets, the introduction of broadband multimedia services will strengthen the need of green communication schemes [14], so that a complex multidimensional optimization problem will arise to accommodate the contrasting requirements on efficient usage of spectrum, low energy consumptions, and provisioning of satisfactorily QoE levels.

In this context, stochastic oriented optimization methods are receiving increasing attention, because of their inherent capability to embrace complex and not deterministic scenarios as the one considered herein [15] - [17]. However, existing stochastic optimization methods [18] [19] mainly consider the data scheduling process of a single BS in 5G networks, which is quite a restrictive hypothesis (a complete overview of the state of the art is detailed in Sec. 2). For this reason, a novel Stochastic Optimization framework for Green Multimedia Services named "SOGMS" over dense 5G networks is proposed hereby. In particular, the main contributions of this paper are summarized as follows:

1) Mathematical formulation of the multimedia stochastic optimization problem: we first formulate the scheduling of 5G multimedia services as a novel stochastic optimization problem which aims to jointly optimize the system throughput and energy consumption in data delivery under the constraints of request queue stability. Afterwards, the formulated optimization problem is decomposed into three independent and tractable subproblems.
2) SOGMS mechanism and implementation: three distinct algorithms are conceived to efficiently solve the three subproblems in the SOGMS mechanism, which target QoE based admission control, cooperative resource allocation, and multimedia services scheduling, respectively. Based on these algorithms, the SOGMS mechanism is implemented in a three-layer architecture.
3) SOGMS performance evaluation: we evaluate the proposed SOGMS-based method against the state-of-art solutions: Buffer-driven Dynamic Bitrate Adaptation (BDBA) [20] and Green Information-Centric Multimedia Streaming (GrIMS) [22]. The comparisons show that SOGMS outperforms BDBA and GrIMS, in terms of queue stability, total utility and energy consumption in data delivery.

Table 1. Comparison of Existing Works

Literatures	Stochastic	Green	Multimedia	5G	Ultra-dense
Stochastic Optimization in 5G [18] [19]	√	×	×	√	×
Stochastic Optimization for multimedia [20]	√	×	√	×	×
Multimedia in 5G [21] [23] [24]	×	×	√	√	×
Green multimedia [22] [25] [26]	×	√	√	×	×
Dense 5G networks [27] [28]	×	√	×	√	√
SOGMS proposed in this paper	√	√	√	√	√

The remainder of this paper is organized as follows. Section 2 describes the related work of our research. Then, the system model of SOGMS is presented in section 3. Afterwards, the optimization problem is formulated and decomposed into three tractable subproblems in section 4. Moreover, the optimal algorithms in SOGMS mechanism are proposed and successfully cast in a three-layer architecture in section 5. In addition, numerical evaluations are presented in section 6. Finally, the conclusion and future work are described in section 7.

2 RELATED WORK

Admission control optimization and multimedia resource allocation methods in 5G systems recently received increasing attention from the scientific community. In this section, related works on those subjects are summarized and qualitatively compared to our approach (see also Tab. 1).

2.1 Admission Control Optimization

There is a rising number of research on admission control problems in 5G networks. The representative works [18] - [22] are introduced as follows:

Liu *et al.* [18] designed a resource allocation and rate control algorithm, based on the heavy ball stochastic optimization method. The authors first propose a congestion control and scheduling framework in cellular networks. Based on this framework, they pursue the trade-off among the utility of throughput, transport latency, and convergence rate of optimization algorithm by using the second-order drift of the virtual service queues in a single BS. Unfortunately, this paper does not consider the presence of multiple and hopefully cooperating BSs, which is quite a common scenario in 5G.

Bao *et al.* [19] proposed a Lyapunov optimization approach for non-orthogonal multiple access (NOMA) system, which jointly consider data rate control and resource allocation at network and physical layers, respectively. In addition, this method transforms the optimization control problem into two subproblems of rate control in a single BS and power allocation among multiple users. Based on this method, both data transmitting rate and delivery delay are efficiently improved. However, also this optimal strategy does not account for the presence of multiple BSs, thus becoming unfitted to dense 5G scenarios.

Zhao *et al.* [20] proposed to maximize the uses' QoE utility of multimedia admission services in heterogeneous wireless networks, based on the stochastic dynamic adaptation streaming over HTTP (DASH). This method faces the dynamic bitrate selection based on the network congestion state and realtime throughput of users' requests. In addition, the

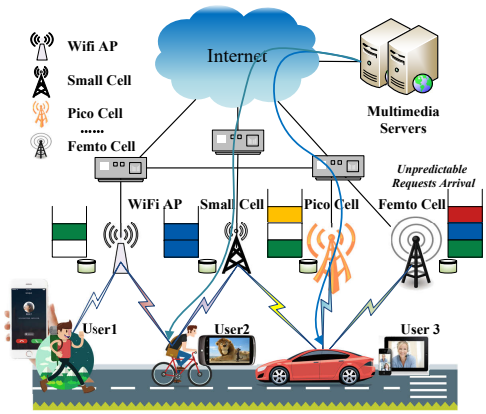

Fig. 1. The multimedia services scenario in dense 5G networks.

experiments based on a testbed validate the efficiency of this method in a real scenario. Unfortunately, this paper lacks considering the resource allocation for multimedia services in BSs, which is necessary for dense 5G networks.

Zheng *et al.* [21] proposed a mechanism for admission rate adaptation and power control in wireless networks. This method decomposes the weighted based maximum-minimum rate fairness problem into the fair sub-problems of traffic and power control. In addition, the method calculates the optimal rate of service through nonlinear Perron-Frobenius theory. However, the authors only consider the optimization of end-to-end source rates, and also this paper does not account for indispensable multimedia resource allocation of BSs in 5G access networks.

In our previous work GrIMS [22], we proposed a Markov queuing model to control multimedia service rates on resource providers (e.g BS, roadside unit, etc.) in wireless heterogeneous networks. Furthermore, a joint price optimal problem was formulated to guarantee the maximization of users' QoE and minimization of energy consumptions of the network.

Although the above studies provide important insights for multimedia request admission, most of them lack analyzing the collaborative process between the heterogeneous BSs in dense 5G networks. Therefore, in this paper, we mainly focus on multimedia admission service by integrating the Lyapunov stochastic optimization and multi-BSs collaboration in dense 5G networks.

2.2 Resource Allocation Strategy

A number of resource allocation strategies have been investigated in 5G networks. Pervez *et al.* [23] proposed a radio resource management scheme that integrates the millimeter Wave (mmWave) in vehicular networks to increase the utilization rate of spectrum resources. Khalek *et al.* [24] proposed a resource allocation policy based on the hypothesis that multiple users have different delay requirements in cellular networks. They first establish the maximal user subsets, where the users have the same delay constraint. Then, the minimization of resource allocation is formulated for each subset. In this way, each BS realizes the maximization of video quality for all scheduled users. Ho *et al.* [25] propose to effectively offload the video traffic to WiFi networks to alleviate the congestion of cellular networks, based on a game theoretic method. Moreover, a cost-effective fountain encoding technology

is utilized to achieve the multi-path transmission through cellular and WiFi, which can improve video service efficiency. Cui *et al*. [26] propose a branch and bound approach to address the user scheduling and power allocation problems in mm-Wave communications by non-orthogonal multiple access. In [27, 28], the author puts forward valuable research results for the resource allocation from the perspective of physical layer in 5G dense networks. In the case of dense cloud radio access networks (C-RAN), the authors jointly optimize the remote radio head (RRH) selection, user equipment (UE)-RRH associations and the downlink beam vector to realize the minimization of the energy consumption of the multi-channel downlink C-RAN. On this basis, a C-RAN framework centered on multiple-antenna users is further studied, in which the precoding matrices and the set of active RRHs are optimized so as to minimize the network power consumption. Based on the proposed methods, both user rate requirements and per-RRH power consumption are efficiently improved.

The aforementioned solutions introduce many strategies for resource allocation in mobile 5G networks, but almost ignore the stochastic arrival of multimedia services in dense 5G networks, which may impair the system stability and service efficiency. In comparison, our work specially focuses on stochastic oriented resource allocation in 5G networks to attain green multimedia services and satisfy users' expectations on QoE.

3 SYSTEM MODEL

In this section, the SOGMS model is presented along with its admission control, resource allocation and multimedia services scheduling functionalities. For sake of clarity, Tab. 2 briefly summarizes the key notations of this paper.

3.1 SOGMS Model

3.1.1 Network model. We assume the SOGMS Framework consisting of a platform deployed in the 5G network, serving a large number of mobile users that ask for video services. Due to the ultra dense characteristics of BSs in 5G networks [29], a large number of BSs, such as small cells, Pico cells, Femto cells, WiFi and so forth, will be deployed as roadside infrastructures (see Fig. 1).

We consider the system has continuous video requests by mobile users. Since different video types have different arrival rates, the video type is denoted by $M = \{1, 2, ..., i, ...m\}$. Furthermore, we suppose that video contents are split in chunks, and all chunks are assumed to be with equal size [30]. Let C denote the unit size of a video chunk. Each video chunk needs to be processed separately, thus the number of requests is proportional to the number of chunks processed. At the same time, we also assume that the system includes a set of BSs, denoted by $N = \{1, 2, ..., j, ...n\}$. In addition, we divide the time into equally sized slots, and each time slot t belongs to $T = \{0, 1, ..., \tau, ...\}$, where τ is a nonnegative integer number. In addition, it is assumed that our optimizing control system tries to actively satisfy the multimedia service requests of mobile users.

With SOGMS, we assume that each BS can provide all kinds of video services to mobile users. Meanwhile, various types of video requests arrive at the 5G network at every time slot t. Let $A_i(t)$ denote the number of arrived requests in the entire 5G network for video type i during time slot t, where the number of $A_i(t)$ can actually be understood as the number of chunks of requested video type i during the time slot t. To simplify our model, we assume that one user can only request one type of video service during each time slot[1] [31]. Meanwhile, one BS can serve several users in the same time slot.

[1]This hypothesis becomes more and more realistic as the slot size is made smaller.

3.1.2 Users' QoE Constraint Model. The higher quality video corresponds to the more number of the video chunks. To achieve the satisfying QoE requirements of users, more video chunks need to be requested in order to provide high quality video services [32]. However, due to the limitation of the capacity of 5G network system, it is impossible to meet the user's demands without restriction. Therefore, the QoE of the user needs to be constrained and optimized. The typical QoE utility is non-decreasing and concave. Moreover, the QoE has a diminishing property with the incremental increase of the video chunks at each time slot. Therefore, the users' QoE utility can be calculated as $U(A_i(t)) = \omega \ln(1+\gamma A_i(t))$ [33], where ω and γ are positive factors tuning. To optimize the QoE efficiency of multimedia services, the QoE constraint can be expressed as the maximization of the users' requests, which is shown in Eq (1):

$$\overline{U(A_i(t))} \geq U_i^{av}, i \in M, \tag{1}$$

where the notation of $\overline{U(A_i(t))}$ denotes the time average of QoE function, that is $\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} U(A_i(\tau))$. U_i^{av} is a constant representing the average QoE of the requests of video type i.

3.1.3 Radio Resource Model. For dense 5G communication networks in SOGMS, to further improve the spectral efficiency, the sharing and reuse of resources, Orthogonal Frequency-Division Multiple Access (OFDMA) [34] is considered as a candidate technology to transmit the data, which utilizes Signal-to-Noise Ratio (SNR) model to guarantee simultaneously wireless access of multiple users and avoid the interference among the users in data transmission. Therefore, for the generality of our system model, the transmission rate of the BS link over each time slot is calculated as follows:

$$R_{ij}(d_{ij}, t) = \frac{W_j}{h_j(t)} \log_2(1 + SNR_{ij}(d_{ij})) \tag{2}$$

where W_j denotes the link bandwidth of BS j. $h_j(t)$ denotes the number of users sharing bandwidth of BS j during time slot t. $SNR_{ij}(d_{ij})$ is mainly affected by the data transmission distance d_{ij} between the mobile user and the BS.

3.2 SOGMS Control Decisions

In every time slot, the SOGMS system model needs to undertake the following control decisions (see Fig. 2):

1) Admission Control of Video Requests: The video requests of mobile users bring a large amount of data which needs to be transmitted at each time slot. In this case, the admission control is necessary to maintain the network stability, by determining how many video requests can be admitted to the 5G system.

It is noteworthy that the SOGMS model does not enforce any special requirement on the arrival distribution of requests $A_i(t)$ [37]. And the maximum value of $A_i(t)$ is denoted as $A_i^{\max}$. To prevent network congestion, only a subset of requests for each video type can be admitted to the 5G networks. We denote the admitted number of videos of type i as $D_i(t)$ at each time slot, with $0 \leq D_i(t) \leq A_i(t)$. Then, we define the queue backlog $Q_i(t)$ as the requests awaiting to be admitted to the 5G network for video type i at the time slot t. When a request of video type i is issued in time slot t and not admitted, it will be added to the queue $Q_i(t)$ and delayed to be processed in next time slot.

2) Resource Allocation of Cooperative BSs: For each video type i, after $D_i(t)$ requests have been admitted to the 5G system, it is necessary to allocate them to available BSs.

Table 2. Table of Key Symbols

Param.	Definitions
M	Set of mobile video types
N	Set of base stations
T	Set of time slots
$A_i(t)$	Total number of requests of video type i arrived at the 5G network during time slot t
$U\left(A_i(t)\right)$	QoE utility of video type i during time slot t
$D_i(t)$	Number of requests of video type i admitted by the 5G network during time slot t
$E_{ij}(t)$	Number of requests of video type i processed by BS j during time slot t
$Q_i(t)$	Admission control queue of requests of video type i at the beginning of time slot t
$K_i(t)$	QoE constraint queue of requests of video type i at the beginning of time slot t
$H_{ij}(t)$	Resource scheduling queue of video type i processed by BS j at the beginning of time slot t
$a(t)$	Throughput utility of system during time slot t
$e_j(t)$	Energy consumption of BS j for video services during time slot t
ϕ	Utility function of the optimization objective
V	Weight of the optimization objective
μ	Weight of throughput in utility function ϕ
C	Unit size of a video chunk
W_j	Link bandwidth of BS j
$h_j(t)$	Number of users sharing bandwidth of BS j during time slot t
P_j	Transmission power of BS j
d_{ij}	Distance between the user requesting type i and BS j
$R_{ij}\left(d_{ij}, t\right)$	Data transmission rate of video type i processed by BS j

Let $D_{ij}(t)$ denote the number of ith type requests allocated to BS j. During time slot t, it yields: $D_i\left(t\right) = \sum_{j \in N} D_{ij}\left(t\right)$. Therefore, each BS maintains a queue of video requests for each type of video in which the requests are awaiting to be processed. Then, we define the queue backlog $H_{ij}(t)$ of total video requests for the ith type on the jth BS at the beginning of time slot t.

3) Multimedia Service Scheduling for Different Users: After resource allocation decisions, each BS is dynamically allocated a set of different video services D_{ij} to process. We further denote $E_{ij}(t)$ as the number of video requests of type i processed by BS j during the time slot t. $E_i(t)$ denotes the number of video services of type i provided by the 5G network system during the time slot t, with $0 \leq E_i\left(t\right) \leq D_i(t)$. During time slot t, also it yields: $E_i\left(t\right) = \sum_{j \in N} E_{ij}\left(t\right)$. Then, we need to schedule the different multimedia services $E_i(t)$ for different users at each time slot.

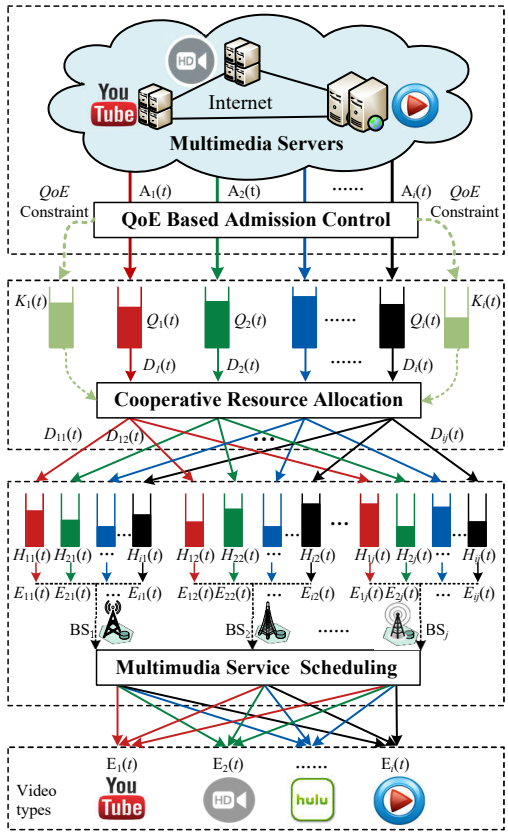

Fig. 2. The SOGMS control framework.

3.3 Dynamic Queues Setup

Based on the SOGMS model introduced previously, the dynamics of the queues for admission control, Request Allocation and QoE constraint over time can be obtained as follows:

1) Admission Control Queue: The dynamic updating queue $Q_i(t)$ for each video type i can be modeled as:

$$Q_i\left(t+1\right) = Q_i\left(t\right) - D_i\left(t\right) + A_i\left(t\right), \forall i, \forall t, \tag{3}$$

where $D_i(t) \leq A_i\left(t\right), Q_i(0) = 0, \forall i \in M$. We further define the queue stability, which is also as the constraint of admission control (see in Def. 3.1).

Definition 3.1. A queue $Q_i(t)$ is strongly stable if and only if

$$\lim_{t->\infty} \sup \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left(Q_i\left(\tau\right)\right) < \infty, \tag{4}$$

where it indicates that the upper bound of time averaged queue length should not be infinite.

2) Resource Allocation Queue: When jth BS serves the video request type i, the dynamic queues over time [38] can be modeled as:

$$H_{ij}(t+1) = H_{ij}(t) - E_{ij}(t) + D_{ij}(t), \tag{5}$$

where $H_{ij}(0) = 0, E_{ij} \leq D_{ij}, \forall i \in M, \forall j, \in N$.

3) QoE Constraint Queue: The virtual queue $K_i(t)$ over time [39] is established to transform the QoE constraint from Eq. (1), which can be modeled as:

$$K_i(t+1) = \max\left[K_i(t) - U(A_i(t)),\ 0\right] + U_i^{av}, \tag{6}$$

where $K_i(0)$ denotes a finite nonnegative value. Then, the QoE constraint requirement can be satisfied through maintaining the stability of this virtual queue $K_i(t)$.

LEMMA 3.2. *The constraint* $\lim_{t\to\infty} \frac{1}{t}\sum_{\tau=0}^{t-1} U(A_i(\tau)) \geq U_i^{av}$ *holds when the virtual queue (6) is strongly stable, i.e., when* $\lim_{t->\infty} \sup \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}(K_i(\tau)) < \infty$.

Proof of Lemma 3.2: Based on the virtual queue (6), we can derive the inequality $K_i(\tau+1) - K_i(\tau) \geq U_i^{av} - U(A_i(\tau))$. By summing this inequality over $\tau \in \{0, ..., t-1\}$, we further derive the inquality as:

$$\frac{K_i(t)}{t} - \frac{K_i(0)}{t} + \frac{1}{t}\sum_{\tau=0}^{t-1} U(A_i(t)) \geq \frac{1}{t}\sum_{\tau=0}^{t-1} U_i^{av}. \tag{7}$$

Since the initial queue length, i.e. $K_i(0)$ has finite nonnegative value, $\lim_{t\to\infty} \frac{K_i(0)}{t}$ is always zero. Only if the queue $K_i(t)$ is stable, we have $\lim_{t\to\infty} \frac{\mathbb{E}(K_i(t))}{t} = 0$. Combining with inequality (7), the constraint in Eq. (1): $\lim_{t\to\infty} \frac{1}{t}\sum_{\tau=0}^{t-1} U(A_i(\tau)) \geq U_i^{av}$ is satisfied.

Now, we prove that $\lim_{t->\infty} \sup \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}(K_i(\tau)) < \infty$ implies $\lim_{t\to\infty} \frac{\mathbb{E}(K_i(t))}{t} = 0$.

Since $K_i(t)$ satisfies independent and identically distributed over time slot t [37], $\mathbb{E}(K_i(t))$ is a constant. Then we have:

$$\lim_{t->\infty} \sup \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}(K_i(\tau)) = \sup \mathbb{E}(K_i(\tau)) < \infty. \tag{8}$$

From Eq. (8), we can always find a constant λ, that $\lambda = \sup \mathbb{E}(K_i(\tau)) < \infty$. Then, we can derive the Eq. (9):

$$\lim_{t\to\infty} \frac{\mathbb{E}(K_i(t))}{t} = \lim_{t\to\infty} \frac{\sup \mathbb{E}(K_i(t))}{t} = \lim_{t\to\infty} \frac{\lambda}{t} = 0. \tag{9}$$

Now, we proved the conclusion: $\lim_{t->\infty} \sup \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}(K_i(\tau)) < \infty$ implies $\lim_{t\to\infty} \frac{\mathbb{E}(K_i(t))}{t} = 0$.

Therefore, *Lemma 1* is proved.

4 PROBLEM FORMULATION

In this section, we formulate our SOGMS model in dense 5G network system. We first describe the utility function by jointly considering the maximization of system throughput, and the minimization of energy consumption in data delivery. Second, we propose a stochastic optimal mechanism based on Lyapunov function [33] to solve the multimedia service optimization.

4.1 Utility Function

4.1.1 Maximizing Throughput Utility. It is intuitive that the more the successful number of video services, the more the utility can be gained by the 5G network system. We consider that the throughput utility is linearly related to the total admitted number of multimedia services. Let $a(t)$ denote the sum of satisfied requests of video type i in time slot t. Therefore, we have $a(t) = \sum_{i \in M} D_i(t)$. Moreover, since the requests of multimedia services are time-varying and unpredictable, we define the utility of throughput as a time-averaged function over time slots. Then, we have the utility of throughput $\overline{a}$ in Eq. (10):

$$\overline{a} = \lim_{t->\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left(\sum_{i \in M} D_i(\tau) \right), \forall i \in M. \tag{10}$$

4.1.2 Minimizing Energy Consumption. In order to provide the mobile users with high quality services, the BSs will consume a certain amount of energy. Let $e_j(t)$ denote the energy consumption of BS j for providing services in time slot t. We have $e_j(t) = \sum_{i \in M} P_j \cdot \mathcal{T}_{ij}$ [35], where P_j represents the transmission power of BS j. $\mathcal{T}_{ij}$ indicates the time taken by the BS j to process the requests of the video type i, depended on the number of video requests to process and the data transmission distance. According to [36], we have $\mathcal{T}_{ij} = \frac{E_{ij}(t) \cdot C}{R_{ij}(d_{ij}, t)}$. Thus, the energy consumption equation can be expressed as follows:

$$e_j(t) = \sum_{i \in M} P_j \cdot \frac{E_{ij}(t) \cdot C}{R_{ij}(d_{ij}, t)}. \tag{11}$$

where $E_{ij}(t) \cdot C$ and $R_{ij}(d_{ij}, t)$ represent the total data size and data transmission rate of video type i processed by BS j during time slot t, respectively (see Eq. (2)). Similarly to time-averaged $\overline{a}$, we calculate the energy consumption utility $\overline{e_j}$ in Eq. (12):

$$\overline{e_j} = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left(\sum_{i \in M} P_j \cdot C \frac{E_{ij}(\tau)}{R_{ij}(d_{ij}, \tau)} \right), \forall i \in M, \forall j \in N. \tag{12}$$

4.1.3 Optimization Objective Function. Based on the previous definitions of throughput and energy consumption, we further define the optimization objective function by Eq. (13):

$$\begin{aligned} &\max \phi = \max \left\{ \mu \cdot \overline{a} - \sum_{j \in N} \overline{e_j} \right\} \\ &s.t. \sum_{j \in N} E_{ij}(t) = E_i(t), \forall i \in M, \forall j \in N \\ &Q_i(t),\ K_i(t),\ H_{ij}(t)\ \textit{are strongly stable.} \end{aligned} \tag{13}$$

The optimization objective function (13) is nonlinear and involves time-average problem. Thus, traditional optimization methods can hardly be applied to this problem. In the next subsection, we will exploit a stochastic optimization method to solve this objective utility function.

4.2 Problem Formulation Based on Lyapunov Optimization

To maximize the optimization objective function, a stochastic optimal mechanism based on Lyapunov optimization [33] is proposed hereby. We first consider the tradeoff between

the system stability and the maximization of objective utility, and then formulate the optimization objective based on this tradeoff.

To solve optimization objective problem (13), we define the Lyapunov function $L(t)$ as shown in Eq. (14):

$$L(t) = \frac{1}{2}\left[\sum_{i\in M} Q_i^2(t) + \sum_{i\in M} K_i^2(t) + \sum_{i\in M}\sum_{j\in N} H_{ij}^2(t)\right], \tag{14}$$

where $L(t)$ is a queue matrix, which denotes the congestion situation of the optimization objective queue. In general, the smaller value of $L(t)$, the fewer backlogs in all queues. Otherwise, it implies existing at least one queue with massive backlogs in $L(t)$. If and only if $L(t)$ maintains a small value at all time slots, the situation is named as strong stability, which is the final optimization objective.

In the following, the conditional Lyapunov drift [33] can be defined as Eq. (15):

$$\Delta L(t) = L(t+1) - L(t), \tag{15}$$

where $\Delta L(t)$ represents the changed queue backlogs based on the Lyapunov function at each time slot. In general, the smaller numerical value of $\Delta L(t)$, the more stable of all queues. Meanwhile, we define the value of $\Delta L(t)$ is non-negative.

Then, $\Delta L(t)$ can be further calculated by Eq. (3, 5 and 15) as following:

$$\begin{aligned} &Q_i^2(t+1) - Q_i^2(t) \le 2(A_i^{\max})^2 - 2Q_i(t)\cdot[D_i(t) - A_i(t)], \\ &K_i^2(t+1) - K_i^2(t) \le 2U^2(A_i(t)) - 2K_i(t)\cdot(U(A_i(t)) - U_i^{av}), \\ &H_{ij}^2(t+1) - H_{ij}^2(t) \le (E_{ij}(t))^2 + (D_{ij}(t))^2 - 2H_{ij}(t)\cdot[E_{ij}(t) - D_{ij}(t)]. \end{aligned} \tag{16}$$

$$\sum_{j\in N} E_{ij}(t) = E_i(t), \sum_{j\in N} D_{ij}(t) = D_i(t). \tag{17}$$

Next, we can combine the formulas (16 - 17) to get $\Delta L(t)$:

$$\begin{aligned} \Delta L(t) \le B &\\ &- \left\{\sum_{i\in M} Q_i(t)\cdot[D_i(t) - A_i(t)]\right\} \\ &- \left\{\sum_{i\in M} K_i(t)\cdot[U(A_i(t)) - U_i{}^{av}(t)]\right\} \\ &- \left\{\sum_{i\in M} H_{ij}(t)\cdot\left\{\sum_{j\in N}[E_{ij}(t) - D_{ij}(t)]\right\}\right\}. \end{aligned} \tag{18}$$

The first term of Eq. (18) can be proven existing an upper limit value, which is denoted by B in Eq. (19):

$$B = 2\sum_{i\in M}(A_i^{\max})^2 + \omega\sum_{i\in M}\left(\ln^{1+\gamma A_i^{\max}}\right)^2. \tag{19}$$

In order to minimize the Lyapunov drift and maximize the system optimization objective, we define the Drift-Plus-Penalty function by joining the two parts as shown in Eq. (20).

$$\Delta L\left(t\right)-V\cdot\left\{\mu\cdot\overline{a}-\sum_{j\in N}\overline{e_j}\right\}, \tag{20}$$

where V is a weight factor of constant.

With the conclusion of formula (18, 19), the upper limit of the Drift-Plus-Penalty will be drew at last. The goal of system optimization objective is to realize the admission control, resource allocation and multimedia service scheduling by minimizing the Drift-Plus-Penalty. Then, we can get the result of Drift-Plus-Penalty as:

$$\begin{aligned}\Delta L(t)-V\cdot\left\{\mu\cdot\overline{a}-\sum_{j\in N}\overline{e_j}\right\}\leq B-\\ -\sum_{i\in M}\left\{K_i\left(t\right)U\left(A_i(t)\right)-K_i\left(t\right)U_i^{av}-Q_i(t)\cdot A_i\left(t\right)\right\}\\ -\sum_{i\in M}\left\{\left(V\cdot\mu+Q_i(t)\right)\cdot D_i\left(t\right)-\sum_{j\in N}H_{ij}\left(t\right)\cdot D_{ij}\left(t\right)\right\}\\ -\sum_{i\in M}\left\{\left[H_{ij}\left(t\right)-V\cdot C\sum_{j\in N}\frac{P_j}{R_{ij}\left(d_{ij},t\right)}\right]\cdot\sum_{j\in N}E_{ij}\left(t\right)\right\}.\end{aligned} \tag{21}$$

5 SOGMS OPTIMAL MECHANISM AND IMPLEMENTATION

In this section, we first propose a novel optimal mechanism composed by three sub-algorithms to accomplish the optimization objective of SOGMS in dense 5G networks. Secondly, we implement the three sub-algorithms in a three-layer architecture, inspired by [40], [41]. The design of optimal algorithm is shown in the following subsection.

5.1 Design of The Optimal Mechanism

Based on the requirements mentioned in last section, our basic idea is to maximize the right terms of inequality (21) to realize the minimum of Drift-Plus-Penalty at each time slot. We can easily know that the variables between three right terms are not coupled. Therefore, we divide inequation (21) into three sub issues (each term denotes a issue). To solve these issues, we consider each issue exactly belongs to one of three distinct optimal algorithms, named QoE-based admission control, cooperative resource allocation, multimedia services scheduling, respectively. By executing the three algorithms, we can obtain the optimization objective result of Eq. (13). Then, we explain the three algorithms in the following:

5.1.1 QoE Based Admission Control Algorithm. Through maximizing the first term in (21), we decouple the QoE based admission control issue, as shown in Eq. (22):

$$\max_{i\in M}K_i\left(t\right)U\left(A_i(t)\right)-K_i\left(t\right)U_i^{av}-Q_i(t)\cdot A_i\left(t\right). \tag{22}$$

To obtain the maximum of (22), we transform this problem by introducing an auxiliary variable x for $A_i(t)$, where x belongs to the real number field. Then, let $F(x)$ represent Eq. (22), we can transform the problem 22 as:

$$F\left(x\right)=K_i\left(t\right)\omega\ln^{1+\gamma\cdot x}-Q_i(t)\cdot x-K_i\left(t\right)U_i^{av}. \tag{23}$$

Algorithm 1: QoE Based Admission Control

Input: Queue $Q_i(t)$ in 5G network at time slot t;
Output: $A_i(t)$ videos are admitted to 5G network system;
At the beginning of each time slot;
for $i = 1$ *to* m **do**
  if $F'(x) = 0$, *and derive* x^*;
  then
    $x^* = \frac{\omega K_i(t)}{Q_i(t)} - \frac{1}{\gamma}$;
    Let $A_i(t) = x^*$;
  end
  return $A_i(t)$;
end

The purpose of this algorithm is to calculate the value of x to maximize $F(x)$, which denotes as x^*. Since the second derivative of $F(x)$ with respect to x is calculated as in Eq. (24).

$$F''(x) = -\frac{\omega\gamma^2 K_i(t)}{(1+\gamma x)^2}. \tag{24}$$

Obviously, $F''(x) \leq 0$, therefore, the function $F(x)$ has a maximum value. According to the knowledge of derivative, x^* can be obtained as:

$$x^* = \frac{\omega K_i(t)}{Q_i(t)} - \frac{1}{\gamma}. \tag{25}$$

Then, let $A_i(t) = x^*$, we can get the maximum value of the Eq. (22), where $Q_i(t)$ represents the queue backlog of video type i requested by mobile users. If the backlog of $Q_i(t)$ is few, the 5G network system is relatively stable, thus it needs to enlarge the throughput of the 5G network system. On the contrary, it means that the 5G network system is tending to congestion, therefore it needs to control the number of admitted multimedia services. The admission control algorithm is shown in Algorithm 1.

5.1.2 Cooperative Resource Allocation Algorithm. Through maximizing the second term in (21), we decouple the cooperative resource allocation problem in Eq. (26):

$$\max_{D_i(t), D_{ij}(t)} (V \cdot \mu + Q_i(t)) \cdot D_i(t) - \sum_{j \in N} H_{ij}(t) \cdot D_{ij}(t), \tag{26}$$

$$s.t \sum_{j \in N} D_{ij}(t) = D_i(t). \tag{27}$$

To address this problem, we assume $D_i(t)$ is a fixed value. Thus, when and only when $H_{ij}(t)$ takes the minimum, the overall optimization will be the largest. The shortest queue of $H_{ij}(t)$ denotes the BS j can process more requests than others. Obviously, the resource allocation can accelerate the tradeoff of all BSs. At the same time, it can reduce the processing delay of multimedia services. Built upon the above analysis, we can redefine the resource allocation to the following Eq. (28):

$$\min \sum_{j \in N} H_{ij}(t) \cdot D_{ij}(t). \tag{28}$$

Algorithm 2: Cooperative Resource Allocation

Input: $Q_i(t)$, $H_{ij}(t)$ in 5G network at time slot t;
Output: $D_i(t)$ videos are allocated to queue $H^*_{ij}(t)$;

```
At the beginning of each time slot;
Initialize i = 0, j = 0, j* = 0;
for j = 1 to n do
    if H_{ij+1}(t) < H_{ij}(t);
    then
        j* = j;
    end
end
for i = 1 to m do
    if V · μ + Q_i(t) − H*_{ij}(t) > 0;
    then
        D_i(t) = A_i(t);
    else
        D_i(t) = 0;
    end
    return D_i(t);
end
```

In the following, we calculate the minimum of Eq. (28) to obtain the optimal solution by Eq. (29):

$$D_{ij}(t) = \begin{cases} D_i(t), if\, j = \arg\min_{j \in N}(H_{ij}(t)) \\ 0, otherwise \end{cases}. \tag{29}$$

Based on the constraints of admission control in Eq. (3), we obtain that $0 \le D_i \le A_i(t)$. Hence, the optimal solution of $D_i(t)$ can be calculated by Eq. (30):

$$D_i(t) = \begin{cases} A_i(t), (V \cdot \mu + Q_i(t)) - H^*_{ij}(t) > 0 \\ 0, otherwise \end{cases}, \tag{30}$$

where $H^*_{ij}(t)$ denotes the minimum queue backlog of video type i on BS j. From the formula 30, we can obtain the optimal cooperative scheduling of BSs. The resource allocation algorithm is shown in Algorithm 2.

5.1.3 Multimedia Service Scheduling Algorithm. Through maximizing the third term in (21), we decouple the multimedia service scheduling problem, as shown in Eq. (31):

$$\max_{E_{ij}(t)} H_{ij}(t) \cdot \sum_{j \in N} E_{ij}(t) - V \cdot C \cdot \sum_{j \in N} \frac{P_j}{R_{ij}(d_{ij},t)} \cdot \sum_{j \in N} E_{ij}(t). \tag{31}$$

It can be observed from Eq. (31) that the larger the $R_{ij}(d_{ij}, t)$, the higher the overall multimedia scheduling efficiency. In general, the data transmission rate $R_{ij}(d_{ij}, t)$ will encounter more noise as the transmission distance d_{ij} becomes larger, thus the closer the BS is, the greater the revenue will be to the transmission service. Similarly to the last problem, we calculate the maximum of Eq. (31) to obtain the optimal solution by Eq. (32):

$$\sum_{j \in N} E_{ij}(t) = \begin{cases} D_i(t), H_{ij}(t) - V \cdot C \cdot \sum_{j \in N} \frac{P_j}{R_{ij}(d_{ij},t)} > 0 \\ 0, otherwise \end{cases}, \tag{32}$$

Algorithm 3: Multimedia Services Scheduling

Input: Queue $H_{ij}(t)$ in 5G network at time slot t;
Output: $\sum_{j\in N} E_{ij}(t)$ video services are provided by 5G network system;

```
At the beginning of each time slot;
for i = 1 to m do
    if H_ij(t) − V · C · Σ_{j∈N} P_j / R_ij(d_ij, t) > 0;
    then
        Σ_{j∈N} E_ij(t) = D_i(t);
    else
        Σ_{j∈N} E_ij(t) = 0;
    end
    return Σ_{j∈N} E_ij(t);
end
```

where $\sum_{j\in N} E_{ij}(t)$ denotes the total services number of videos type i provided by the 5G network system. Then, the multimedia service scheduling algorithm is shown in Algorithm 3.

5.2 SOGMS Implementation

In this subsection, we implement the previous optimal algorithms in the SOGMS architecture (see Fig. 3), which is divided to three layers, named User Access Layer(UAL), Resource Schedule Layer(RSL) and Network Infrastructure Layer(NIL), respectively.

1) UAL: User Access Layer is at the left layer, which consists similarly to the application layer of existing networks. The mobile users from UAL can request multimedia services from 5G network system, and measure the QoE of video streaming services.

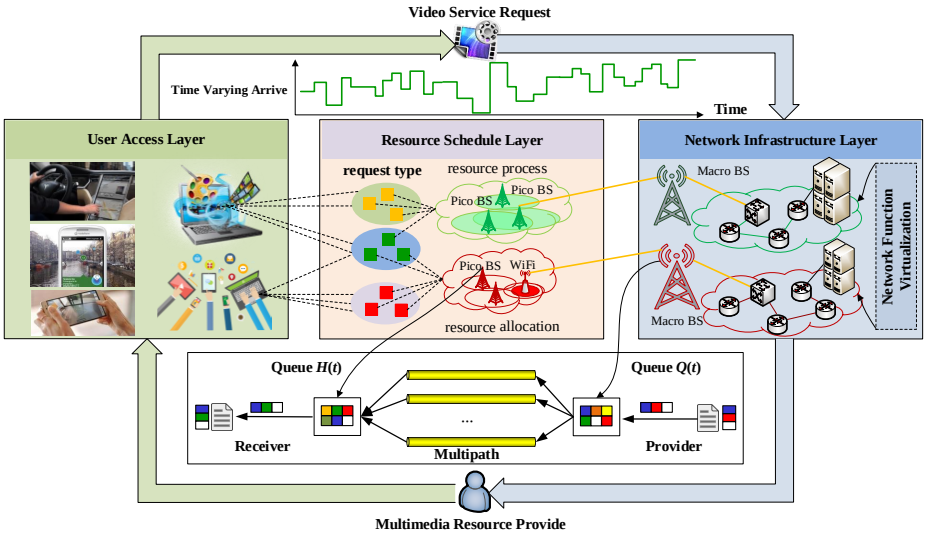

Fig. 3. The SOGMS three-layer architecture in 5G networks

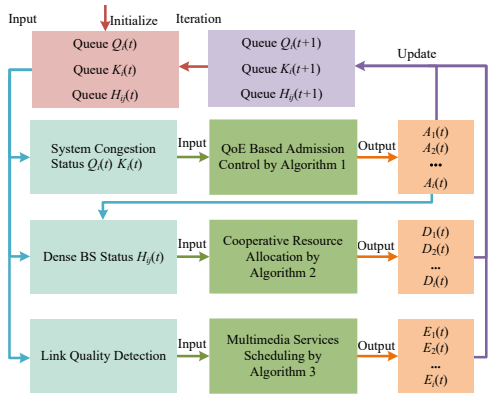

Fig. 4. The optimal control algorithm process

2) RSL: Resource Schedule Layer is at the middle layer, which plays the role like a decision controller. It's in charge of scheduling the network resource. Thanks to the RSL, each edge BS can provide the balanced and green multimedia services for mobile users.

3) NIL: Network Infrastructure Layer is at the right layer, consists of dense BSs and routers. Moreover, the network infrastructure is divided into several individual networks through virtualization and network slice [42] and provides the resources through programmable interfaces to RSL.

According to the function description of SOGMS architecture, we notice that the admission control can be addressed between UAL and NIL, the resource allocation can be addressed between NIL and RSL, meanwhile, the multimedia service scheduling can be addressed between RSL and UAL. Furthermore, we implement the optimal algorithms including QoE based admission control, cooperative resource allocation and multimedia service scheduling on the SOGMS architecture.

After accomplished the implementation of three algorithms in SOGMS architecture, we can address the system optimization problem (13) efficiently. The process of optimal algorithm in SOGMS architecture is shown in Fig. 4. According to the optimal process, the backlogs of three queues need to be updated iteratively by Eq. (3, 5 and 6), to maintain the optimization performance and stability of 5G network system.

6 NUMERICAL EVALUATION

In this section, we first introduce the simulation environment setting and the benchmark algorithms. Then we provide a detailed performance evaluation of the algorithms and demonstrate the superiority of our method.

6.1 Simulator Parameter Setting

To evaluate the performance of the approach proposed hereby, two benchmark algorithms have been selected for comparison, namely BDBA [20] and GrIMS [22].

With BDBA: The system can pursue the maximization of QoE utility by realizing the balance between the network congestion state and realtime throughput of users' requests. Since this method does not consider the resource allocation of BSs for multimedia services, we assume the resource allocation in BDBA obeys randomized strategy.

With GrIMS: The network system pursues the maximum gain of bandwidth by realizing the balance between resource allocation of service providers and video requests of mobile

users, which the resource allocation strategy is based on the greedy selection. Only when the obtained gain is positive and the service queue of providers is stable, the new requests can be admitted to the system for processing.

During the simulation, we assume there are about 1500 mobile users communicating with 10-60 BSs in 5G network system. We assume there are 5 types of video requests, which denote different quality of services. The number of multimedia categories is similar to today's playback requests, such as ultra-high resolution, high resolution, standard definition, etc. The requests of video type i arrive at the network system according to a Poisson process [43] with the distribution of intensity $\lambda_i \in \{5, ..., 10\}$. All simulations are performed over 50 iterations, which is equal to 1000 time slots. Each of the 50 iterations is repeated 10 times, and the average of 10 repeated times is taken as the final experimental result. The initial queue backlogs of all BSs are null. The mainly parameter setting is shown in Table 3.

We use the *queue backlog* and *utility* as the parameters to evaluate the performance of three methods.

Average queue backlog: the unprocessed quantity of requests of service by mobile users.

Average utility: the average gain of the 5G network system, which means the larger the system utility, the larger the system throughput and the less energy consumption of all BSs.

6.2 System Stability Analysis

We first perform the analysis of system stability on varying arrival rates, which is shown in Fig. 5 and Fig. 6.

Fig. 5 shows the system stability of different users number through five cases when the number of BSs is set to 12, the value of V is set to 50 and the number of users is set to 500-1500 respectively. The results are concluded from three phases. First of all, each curve abruptly rises up during the transient, starting from an empty queue level. Then, all the curves converge to a constant steady state value, thus demonstrating queue stability. Finally, the steady state queue level depends on the number of users: the higher the number of user the higher the steady state queue level.

Fig. 6 shows the average queue backlog on varying requests during different time slots using the three methods. In this case, the number of BSs is set to 12, the value of V is set to 50, the simulation time is enlarged to 500 iterations to better reflect the changes of the queue backlog. Fig. 6 reports that when the arrival rate of requests changes with unfixed time slots, three methods can quickly achieve stability, indicating that the three methods have superior adaptability to the time-varying requests of service. Meanwhile, the queue backlog of SOGMS is always smaller than the other two methods due to the better system stability.

Table 3. Simulator Parameter Setting

Parameter	**Value**	**Parameter**	**Value**
Simulation time	1000s	Transmission distance	100-500 m
Number of users	500-1500	Number of BSs	10-60
Number of types	5	Possion λ_i	[5,10]
Weight μ	10	V	50-150

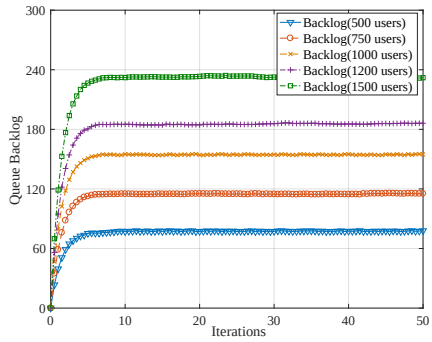

Fig. 5. Queue backlog of diff. users by SOGMS

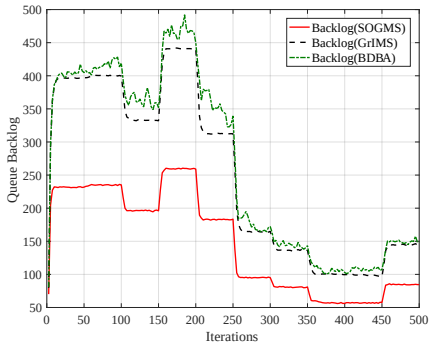

Fig. 6. Queue backlog of varying requests

6.3 Effect Under User Request Rates

We second analyze the effect of the three approaches under different user request rates. The number of users are respectively set as 500, 1000 and 1500 to guarantee the different request arriving rates. The number of BSs is set to 12. The value of V is set to 50. The results of average queue backlog and average utility are shown in Fig. 7 and Fig. 8.

Fig. 7 shows that the average queue backlog of SOGMS (red curve) is smaller than both GrIMS (black curve) and BDBA (green curve) under the three types of request rates. Notes that curves 1a, 1b and 1c represent the requests of 500 users, curves 2a, 2b, 2c represent the requests of 1000 users and curves 3a, 3b, 3c represent the requests of 1500 users, respectively. We can get the conclude that the average queue backlog of three methods significantly depends on the number of users. The larger the number of users, the higher the backlogs of three methods. And with the same number of users, the system stability of SOGMS is superior to the other two methods.

Fig. 8 shows the average utility of SOGMS is much higher than both GrIMS and BDBA under the three types of request rates. When the number of users reaches 1000, the average utility of SOGMS is nearly 1.5 times higher than GrIMS and BDBA. The reason is that SOGMS can both pay attention to the overall throughput of the system and the average processing capability of the BSs in the 5G network. Then, the average utility of three methods tends to settle when the number of users exceeds 1400.

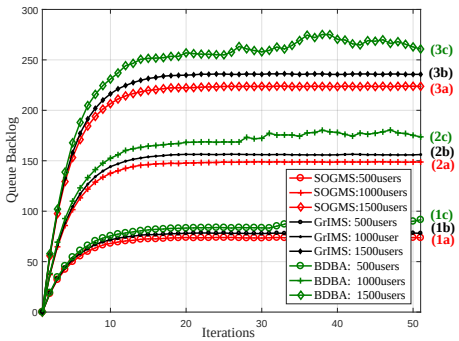

Fig. 7. Queue backlog vs. Diff. users

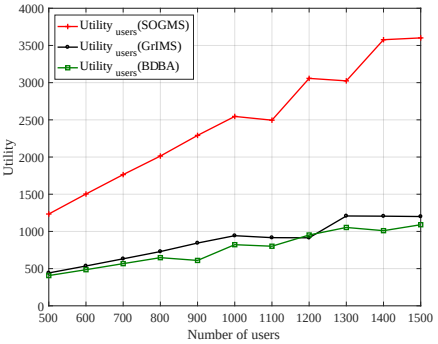

Fig. 8. Average utility vs. Diff. users

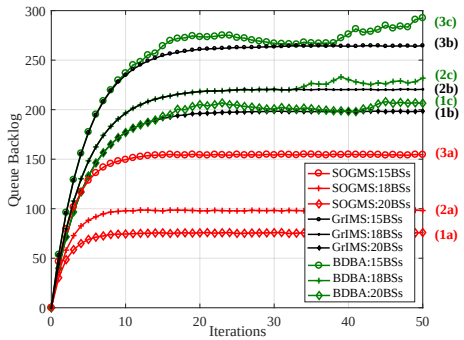

Fig. 9. Queue backlog vs. Diff. BSs

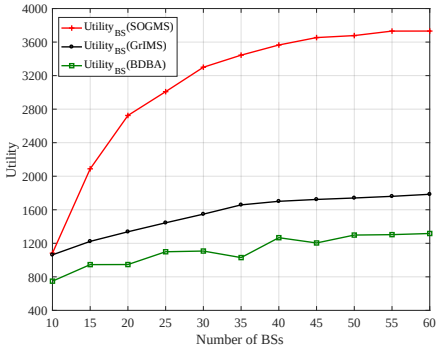

Fig. 10. Average utility vs. Diff. BSs

6.4 Effect Under Different Numbers of BSs

In order to further analyze the effect of the number of BSs, we first compare the average queue backlog achieved by the three methods. As shown in Fig. 9, the number of BSs is set to 20 (curves 1a, 1b, 1c), 18 (curves 2a, 2b, 2c) and 15 (curves 3a, 3b, 3c), respectively. The larger the number of BSs, the stronger the capacity for requests process, the smaller the queue backlog of three methods. Meanwhile, the average queue backlog of SOGMS (red curve) is much smaller than both GrIMS (black curve) and BDBA (green curve) with different number of BSs. Especially when the number of BSs is 18, the queue backlog of SOGMS is 45 percent smaller than the other two methods due to the better process capacity.

We also study the average utility with different number of BSs in Fig. 10. In order to better reflect the experimental results, we gradually expand the number of BSs from 10 to 60. Fig. 10 shows SOGMS has much higher utility than the other methods. When the number of BSs reaches 60, the utility of SOGMS is nearly 1.3 times higher than GrIMS and BDBA. Moreover, the average utility of the 5G network will gradually increases and tends to settle when more and more BSs join in the 5G network. This is because that the larger the number of BSs, the more the addressed requests of service, the more the utility of the 5G network.

6.5 Tradeoff Between Utility and Queue Backlog

Besides, we perform simulations to validate how parameter V tradeoffs the utility and queue stability. For the case of queue stability, the number of users and BSs are first set to 1000 and 15, respectively. Then, the value of V is set to 50 (curves 1a, 1b, 1c), 80 (curves 2a,

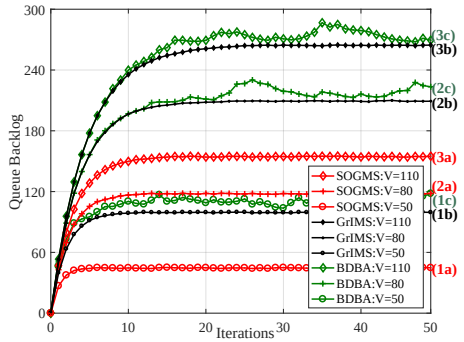

Fig. 11. Average backlog vs. Tradeoff para. V

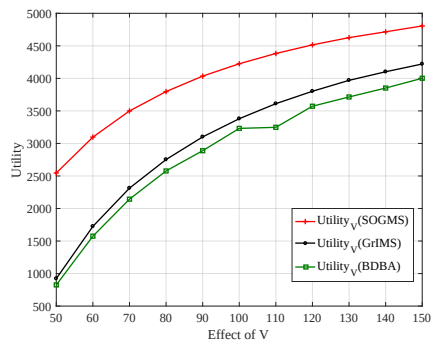

Fig. 12. Average utility vs. Tradeoff para. V

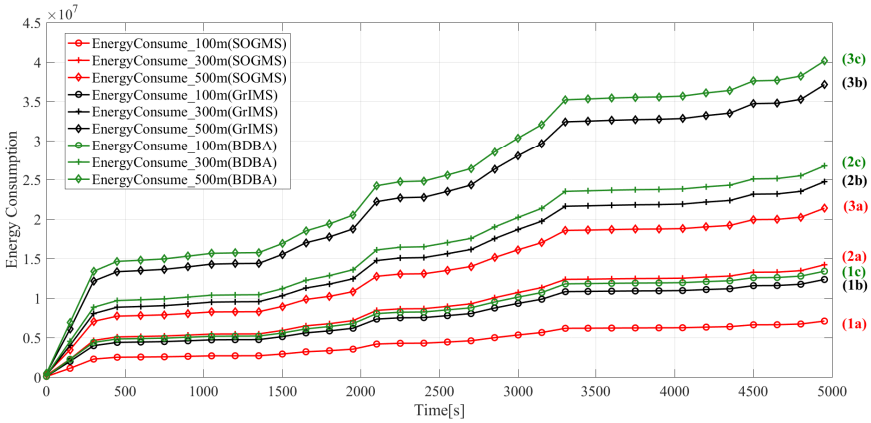

Fig. 13. Energy consumption vs. Simulation time

2b, 2c), 110 (curves 3a,3b, 3c), respectively. As shown in Fig. 11, each queue increases and gradually settles over the time. And the queue backlog becomes larger with the increase of V. This is because the greater the value of V in Lyapunov function, the more the attention payed to the benefits of the 5G network. Since the benefit is proportional to the number of processed requests, more requests will be admitted to the 5G network. Then the increased amount of requests will also lead to the increased queue backlog.

For the case of utility, Fig. 12 shows the tendency of averaged utility with different values of V. It can be observed that the utility grows continuously when the parameter V increases. Meanwhile, SOGMS has better utility than the other two methods. When V equals to 100, the utility of SOGMS is 26.5 and 37.5 percent higher than GrIMS and BDBA, respectively.

According to the above analysis, parameter V is the tradeoff between the utility and queue backlog. Since the utility increases when V enlarges. However, the improved utility would incur more queue backlog need to be processed, which may lead to the congestion of the network. Therefore, a suitable value of V should be chosen to balance the utility and queue backlog.

6.6 Effect Under Energy Consumption

Finally, we perform simulations to calculate the summation of *Energy Consumption* by three methods, which is one of the significant goals in this paper.

For the case of energy consumption, we assume that all users request multimedia services in same way. The number of BSs is set to 15, the value of V is set to 50, respectively. The baseline number of users is first set to 1000. Then, the value of *transmission distance* is set to 100 m (curves 1a, 1b, 1c), 300 (curves 2a, 2b, 2c), 500 (curves 3a, 3b, 3c), respectively.

As time goes on, the number of users fluctuates, indicating the stochastic requests of the video service. In order to better verify the experimental results, we conduct a long-term simulation. The simulation time is set to 5000 seconds. The energy consumption is defined as the unit consumption for processing a video chunk $\times$ the number of the video chunks $\times$ the data transmission distance, it yields: $e_j(t) = \sum_{i \in M} P_j \cdot \frac{E_{ij}(t) \cdot C}{R_{ij}(d_{ij}, t)}$ in Eq. (11). The total number of service includes the provided video services and current backlogs of the BSs.

As shown in Fig. 13, the ordinate represents the cumulative value of energy consumption, hence the total energy consumption continues to increase with the extension of simulation time. It is observed that the SOGMS consumes less energy than GrIMS and BDBA with

the same data transmission distance. As the transmission distance increases, the energy consumption of SOGMS will maintain a relatively low level. Namely, the advantages of SOGMS are more obvious in the case of larger data transmission distances and more video requests. Especially the consumption increment of SOGMS is 36 percent smaller than the other methods when the simulation time is at 2000 second, indicating that SOGMS has the better stability for the dynamic changes of users' request. This is because SOGMS integrates considering the energy consumption of the requests of service and the cooperative process of the BSs, which is more beneficial to the 5G network performance and the users' QoE.

7 CONCLUSIONS

In summary, according to the dynamic and complexity features of multimedia services in dense 5G networks, a novel SOGMS method is proposed for the multimedia service, aiming to face the challenges of efficient usage the system capacity and the requirement of the green communication. Then, the stochastic optimization based on Lyapunov stability theory is adopted to formulate the SOGMS mechanism, which targets the maximization of system throughput and minimization of energy consumption. Through optimizing the system performance and the queue stability, the problems can be efficiently addressed by three distinct sub-algorithms, and further implemented in SOGMS three-layer architecture. At last, the simulation results demonstrate the effectiveness of our proposed method.

In this paper, we have considered multimedia optimizing method only in dense 5G network. Future work will consider employing the proposed method in conjunction with proactive edging caching with video popularity [44] [45] to further improve the users' QoE during the video delivery in future 5G networks.

REFERENCES

[1] G. Gao, and Y. Wen. 2017. When Cloud Media Meet Network Function Virtualization: Challenges and Applications. *IEEE MultiMedia* 24, 3 (2017), 72–82. DOI:http://dx.doi.org/10.1109/MMUL.2017.3051519

[2] C. Xu, P. Zhang, S. Jia, M. Wang, and G. M. Muntean. 2017. Video Streaming in Content-Centric Mobile Networks: Challenges and Solutions. *IEEE Wireless Communications* 24, 5 (2017), 157–165. DOI:http://dx.doi.org/10.1109/MWC.2017.1600219WC

[3] Cisco 2017. *Cisco Visual Networking Index Global Mobile Data Traffic Forecast Update, 2016-2021 White Paper* (2017), San Jose, CA, USA . Retrieved from: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[4] J. G. Andrews, and et al. 2014. What Will 5G Be?. *IEEE Journal on Selected Areas in Communications* 32, 6 (2014), 1065–1082. DOI:http://dx.doi.org/10.1109/JSAC.2014.2328098

[5] X. Ge, S. Tu, G.Mao, C.X.Wang, and T. Han. 2016. 5G Ultra-Dense Cellular Networks. *IEEE Wireless Communications* 23, 1 (2016), 72–79. DOI:http://dx.doi.org/10.1109/MWC.2016.7422408

[6] S. Petrangeli, J. V. D. Hooft, T. Wauters, and F. D. Turck. 2018. Quality of Experience-Centric Management of Adaptive Video Streaming Services: Status and Challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 2s (2018), 31:1–31:29. DOI:https://doi.org/10.1145/3165266

[7] R. Mekuria, K. Blom, and P. Cesar. 2017. Design, Implementation, and Evaluation of a Point Cloud Codec for Tele-Immersive Video. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 4 (2017), 828–842. DOI:http://dx.doi.org/10.1109/TCSVT.2016.2543039

[8] K. Bahirat, C. Lai, R. P. Mcmahan, and B. Prabhakaran. 2018. Designing and Evaluating a Mesh Simplification Algorithm for Virtual Reality. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 3s (2018), 63:1–63:26. DOI:https://doi.org/10.1145/3209661

[9] R. S. Schmoll, S. Pandi, P. J. Braun, and F. H. P. Fitzek. 2018. Demonstration of VR/AR offloading to Mobile Edge Cloud for low latency 5G gaming application. *Proc. of 2018 15th IEEE Annual Consumer Communications and Networking Conference (CCNC)* , (2018), 1–3. DOI:http://dx.doi.org/10.1109/CCNC.2018.8319323

[10] A. Argyriou, K. Poularakis, G. Iosifidis, and L. Tassiulas. 2017. Video Delivery in Dense 5G Cellular Networks. *IEEE Network* 31, 4 (2017), 28–34. DOI:http://dx.doi.org/10.1109/MNET.2017.1600298

[11] B. Xie, Z. Zhang, R. Q. Hu, G. Wu, and A. Papathanassiou. 2018. Joint Spectral Efficiency and Energy Efficiency in FFR-Based Wireless Heterogeneous Networks. *IEEE Transactions on Vehicular Technology* 67, 9 (2018), 8154–8168. DOI:http://dx.doi.org/10.1109/TVT.2017.2701356

[12] L. S. Kapov, M. Varela, T. Hofeld, and K.T. Chen. 2018. A Survey of Emerging Concepts and Challenges for QoE Management of Multimedia Services. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 2s (2018), 29:1–29:29. DOI:https://doi.org/10.1145/3176648

[13] D. Bhat, A. Rizk, M. Zink, and R. Steinmetz. 2018. SABR: Network-Assisted Content Distribution for QoE-Driven ABR Video Streaming. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 2s (2018), 32:1–32:25. DOI:http://doi.acm.org/10.1145/3183516

[14] S. Buzzi, C. L. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone. 2016. A Survey of Energy-Efficient Techniques for 5G Networks and Challenges Ahead. *IEEE Journal on Selected Areas in Communications* 34, 4 (2016), 697–709. DOI:http://dx.doi.org/10.1109/JSAC.2016.2550338

[15] C. Li, H. Xiong, J. Zou, and D. O. Wu. 2018. Joint Dynamic Rate Control and Transmission Scheduling for Scalable Video Multirate Multicast Over Wireless Networks. *IEEE Transactions on Multimedia* 20, 2 (2018), 361–378. DOI:http://dx.doi.org/10.1109/TMM.2017.2745709

[16] J. Zhu, C. Xu, J. Guan, and D. O. Wu. 2018. Differentially Private Distributed Online Algorithms Over Time-Varying Directed Networks. *IEEE Transactions on Signal and Information Processing over Networks* 4, 1 (2018), 4–17. DOI:http://dx.doi.org/10.1109/TSIPN.2018.2797806

[17] J. B. Seo, B. C. Jung, and H. Jin. 2018. Nonorthogonal Random Access for 5G Mobile Communication Systems. *IEEE Transactions on Vehicular Technology* 67, 8 (2018), 7867–7871. DOI:http://dx.doi.org/10.1109/TVT.2018.2825462

[18] J. Liu , A. Eryilmaz, N. B. Shroff, and E. S. Bentley. 2016. Heavy-ball: A New Approach to Tame Delay and Convergence in Wireless Network Optimization. In *Proc. of 2016 IEEE International Conference on Computer Communications (INFOCOM)*. IEEE, 1–9. DOI:http://dx.doi.org/10.1109/INFOCOM.2016.7524474

[19] W. Bao, H. Chen, Y. Li and B. Vucetic. 2017. Joint Rate Control and Power Allocation for Non-Orthogonal Multiple Access Systems. *IEEE Journal on Selected Areas in Communications* 35, 12 (2017), 2798–2811. DOI:http://dx.doi.org/10.1109/JSAC.2017.2726357

[20] P. Zhao, W. Yu, X. Yang, D. Meng and L. Wang. 2017. Buffer Data-Driven Adaptation of Mobile Video Streaming over Heterogeneous Wireless Networks. *IEEE Internet of Things Journal* PP, 99 (2017), 1–1. DOI:http://dx.doi.org/10.1109/JIOT.2017.2763166

[21] L. Zheng, D. W. H. Cai, and C. W. Tan. 2018. Max-Min Fairness Rate Control in Wireless Networks: Optimality and Algorithms by Perron-Frobenius Theory. *IEEE Transactions on Mobile Computing* 17, 1 (2018), 127–140. DOI:http://dx.doi.org/10.1109/TMC.2017.2698469

[22] C. Xu, W. Quan, H. Zhang and L. A. Grieco. 2018. GrIMS: Green Information-Centric Multimedia Streaming Framework in Vehicular Ad Hoc Networks. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 2 (2018), 483–498. DOI:http://dx.doi.org/10.1109/TCSVT.2016.2607764

[23] F. Pervez, A. Adinoyi, and H. Yanikomeroglu. 2017. Efficient resource allocation for video streaming for 5G network-to-vehicle communications. In *Proc. of 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 1–6. DOI:http://dx.doi.org/10.1109/PIMRC.2017.8292728

[24] A. A. Khalek, C. Caramanis, and R. W. Heath. 2015. Delay-Constrained Video Transmission: Quality-Driven Resource Allocation and Scheduling. *IEEE Journal of Selected Topics in Signal Processing* 9, 1 (2015), 60–75. DOI:http://dx.doi.org/10.1109/JSTSP.2014.2332304

[25] D. Ho, G. S. Park, and H. Song. 2018. Game-Theoretic Scalable Offloading for Video Streaming Services over LTE and WiFi Networks. *IEEE Transactions on Mobile Computing* 17, 5 (2018), 1090–1104. DOI:http://dx.doi.org/10.1109/TMC.2017.2748592

[26] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan. 2018. Optimal User Scheduling and Power Allocation for Millimeter Wave NOMA Systems. *IEEE Transactions on Wireless Communications* 17, 3 (2018), 1502–1517. DOI:http://dx.doi.org/10.1109/TWC.2017.2779504

[27] C. Pan, H. Zhu, N. J. Gomes, and J. Wang. 2017. Joint User Selection and Energy Minimization for Ultra-Dense Multi-channel C-RAN With Incomplete CSI. *IEEE Journal on Selected Areas in Communications* 35, 8 (2017), 1809–1824. DOI:http://dx.doi.org/10.1109/JSAC.2017.2710858

[28] C. Pan, H. Zhu, N. J. Gomes, and J. Wang. 2017. Joint Precoding and RRH Selection for User-Centric Green MIMO C-RAN. *IEEE Transactions on Wireless Communications* 16, 5 (2017), 2891–2906. DOI: http://dx.doi.org/10.1109/TWC.2017.2671358

[29] Q. Zheng, K. Zheng, H. Zhang, and V. C. M. Leung. 2016. Delay-Optimal Virtualized Radio Resource Scheduling in Software-Defined Vehicular Networks via Stochastic Learning. *IEEE Transactions on Vehicular Technology* 65, 10 (2016), 7857–7867. DOI:http://dx.doi.org/10.1109/TVT.2016.2538461

[30] T. Cao, C. Xu, M. Wang, X. Chen, L. Zhong, and G. M. Muntean. 2018. Family-Aware Pricing Strategy for Accelerating Video Dissemination over Information-Centric Vehicular Networks. In *Proc. of 2018 IEEE International Conference on Communications (ICC)*. IEEE, 1–7. DOI:http://dx.doi.org/10.1109/ICC.2018.8422411

[31] C. Xu, S. Jia, M. Wang, L. Zhong, H. Zhang, and G. M. Muntean. 2015. Performance-Aware Mobile Community-based VoD Streaming over Vehicular Ad Hoc Networks. *IEEE Transactions on Vehicular Technology* 64, 3 (2015), 1201–1217. DOI:http://dx.doi.org/10.1109/TVT.2014.2329696

[32] C. Xu, F. Zhao, J. Guan, H. Zhang, and G. M. Muntean. 2013. QoE-Driven User-Centric VoD Services in Urban Multihomed P2P-Based Vehicular Networks. *IEEE Transactions on Vehicular Technology* 62, 5 (2013), 2273–2289. DOI:http://dx.doi.org/10.1109/TVT.2012.2228682

[33] M. J. Neely. 2010. Stochastic Network Optimization with Application to Communication and Queueing Systems. *Morgan and Claypool* , (2010). DOI:http://dx.doi.org/10.2200/S00271ED1V01Y201006CNT007

[34] F. Khoramnejad, M. Rasti, H. Pedram, and E. Hossain. 2018. On Resource Management in Load-Coupled OFDMA Networks. *IEEE Transactions on Communications* 66, 5 (2018), 2295-2311. DOI: http://dx.doi.org/10.1109/TCOMM.2018.2795603

[35] X. Ge, L. Pan, Q. Li, G. Mao, and S. Tu. 2017. Multipath Cooperative Communications Networks for Augmented and Virtual Reality Transmission. *IEEE Transactions on Multimedia* 19, 10 (2017), 2345–2358. DOI:http://dx.doi.org/10.1109/TMM.2017.2733461

[36] B. Chen, C. Yang, and A. F. Molisch. 2017. Cache-Enabled Device-to-Device Communications: Offloading Gain and Energy Cost. *IEEE Transactions on Wireless Communications* 16, 7 (2017), 4519–4536. DOI:http://dx.doi.org/10.1109/TWC.2017.2699631

[37] W. Fang, X. Yao, X. Zhao, J. Yin, and N. Xiong. 2018. A Stochastic Control Approach to Maximize Profit on Service Provisioning for Mobile Cloudlet Platforms. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 48, 4 (2018), 522–534. DOI:http://dx.doi.org/10.1109/TSMC.2016.2606400

[38] T. Liu, Y. Zhu, and A. Vasilakos. 2016. Stochastic Optimal Control for Participatory Sensing Systems with Heterogenous Requests. *IEEE Transactions on Computers* 65, 5 (2016), 1619–1631. DOI: http://dx.doi.org/10.1109/TC.2015.2452899

[39] X. Li, and J. Guan. 2017 . SORA: A stochastic optimal routing algorithm for wireless sensor networks. In *Proc. of 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 1–6. DOI:http://dx.doi.org/10.1109/PIMRC.2017.8292491

[40] H. Zhang, W. Quan, H. c. Chao, and C. Qiao. 2016. Smart Identifier Network: A Collaborative Architecture for the Future Internet. *IEEE Network* 30, 3 (2016), 46–51. DOI:http://dx.doi.org/10.1109/MNET.2016.7474343

[41] P. Dong, T. Zheng, S. Yu, H. Zhang, and X. Yan. 2017. Enhancing Vehicular Communication Using 5G-Enabled Smart Collaborative Networking. *IEEE Wireless Communications* 24, 6 (2017), 72–79. DOI:http://dx.doi.org/10.1109/MWC.2017.1600375

[42] K. Wang, F. R. Yu, H. Li, and Z. Li. 2017. Information-Centric Wireless Networks with Virtualization and D2D Communications. *IEEE Wireless Communications* 24, 3 (2017), 104–111. DOI:http://dx.doi.org/10.1109/MWC.2017.1500384WC

[43] C. Xu, S. Jia, L. Zhong, and G. M. Muntean. 2015. Socially aware mobile peer-to-peer communications for community multimedia streaming services. *IEEE Communications Magazine* 53, 10 (2015), 150–156. DOI:http://dx.doi.org/10.1109/MCOM.2015.7295477

[44] G. Gao, Y. Wen, and J. Cai. 2017. vCache: Supporting Cost-Efficient Adaptive Bitrate Streaming via NFV-Based Virtual Caching. *IEEE MultiMedia* PP, 99 (2017), 1–1. DOI:http://dx.doi.org/10.1109/MMUL.2017.265091759

[45] Y. Guo, Q. Yang, F. R. Yu, and V. C. M. Leung. 2018. Cache-Enabled Adaptive Video Streaming over Vehicular Networks: A Dynamic Approach. *IEEE Transactions on Vehicular Technology* 67, 6 (2018), 5445–5459. DOI:http://dx.doi.org/10.1109/TVT.2018.2817210

Received September 2018; revised March 2019; accepted April 2019