Dynamic Resource Allocation for Streaming Scalable Videos in SDN-Aided Dense Small-Cell Networks

Jian Yang, *Senior Member, IEEE*, Bowen Yang, Shuangwu Chen, Yongdong Zhang, *Senior Member, IEEE*, Yanyong Zhang, *Fellow, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*— Both wireless small-cell communications and software-defined networking (SDN) in wired systems continue to evolve rapidly, aiming for improving the quality of experience (QoE) of users. Against this emerging landscape, we conceive scalable video streaming over SDN-aided dense smell-cell networks by jointly optimizing the video layer selection, the wireless resource allocation, and the dynamic routing of video streams. In the light of this ambitious objective, we conceive a dense software-defined small-cell network architecture for the fine-grained manipulation of the video streams relying on the cooperation of small-cell base stations. Based on this framework, we formulate the scalable video streaming problem as maximizing the time-averaged QoE subject to a specific time-averaged rate constraint as well as to a resource constraint. By employing the classic Lyapunov optimization method, the problem is further decomposed into the twin sub-problems of video layer selection and wireless resource allocation. Via solving these sub-problems, we derive a video layer selection strategy and a wireless resource allocation algorithm. Furthermore, we propose a beneficial routing policy for scalable video streams with the aid of the so-called segment routing technique in the context of SDN, which additionally exploits the collaboration of small-cell base stations. Our results demonstrate compelling performance improvements compared with the classic PID control theory-based method.**

***Index Terms*— Dense small-cell, software-defined networking, resource allocation, scalable video streaming, segment routing.**

I. Introduction

IMMERSIVE multimedia communication is poised to pervade people's daily life. However, the enhancement of network capacity cannot keep up with the dramatically increasing demands, especially in wireless networks [1]. A report from Cisco [2] indicates that the global mobile traffic has increased exponentially over the past decade and will keep on increasing even faster, owing to the popularity of mobile devices, such as tablets, smart-phones, etc. Furthermore, the video traffic occupies more than half of all mobile data traffic due to the demand for high-quality video streaming. This imposes a challenge on the wireless network in terms of supporting high-quality video services in order to guarantee a high Quality of Experience (QoE) for the users.

The dense Small-Cell Network (DSCN) concept [3] is emerging as a promising wireless access technique for increasing the capacity of cellular networks in an economical and ecological way. It is a wide consensus that deploying dense small-cell base stations (SCBSs) having a low cost and low-power consumption is capable of improving the spectral efficiency in areas of high data traffic density [4], [5]. Additionally, in order to cater for intelligent services, the applications' prompt adaptivity to the near-instantaneous network condition is a necessary precondition of its robustness. Scalable Video Coding (SVC) [6] is eminently suitable for flexible video streaming by flexibly adapting the number of enhancement layers, both in response to time-variant network conditions and to device capabilities [7]. The spatial scalability of SVC enables the adaptation of the video services to heterogeneous devices having different resolutions, while the quality/temporal scalability can be exploited for achieving dynamic adaptation in order to cope with time-varying network conditions [8].

The high-bandwidth next-generation wireless network is expected to satisfy QoE requirements when watching videos on mobile devices. A comprehensive framework has to be conceived for adaptive video transmission. Due to the small coverage area of next-gen base stations, the users may receive packets from multiple base stations at the same time. However, the previous work has mainly been focused on the optimization of single-user or single-BS video transmission.

Although the DSCN constitutes a promising technique of providing high bandwidth for next-gen networks [9], it still faces some tough problems which are not conductive to improving the QoE for users, such as the collaboration among SCBSs, multipath, etc. As for SVC, it is still not widely exploited in practical video services. The main reason for this is that the nodes are not sufficiently controllable and

Manuscript received May 15, 2018; revised October 4, 2018; accepted November 21, 2018. Date of publication November 28, 2018; date of current version March 15, 2019. J. Yang would like to acknowledge the financial support of National Natural Science of China (No. 61573329), Equipment pre-Research Project (No. 6141B0801010a), National Defense Science and Technology Fund for Distinguished Young Scholars (2017-JCJQ-ZQ-022) and Youth Innovation Promotion Association CAS. L. Hanzo would like to acknowledge the financial support of the EPSRC projects EP/Noo4558/1, EP/PO34284/1, of the Royal Society's GRFC Grant as well as of the European Research Council's Advanced Fellow Grant QuantCom. The associate editor coordinating the review of this paper and approving it for publication was L. Huang. *(Corresponding author: Lajos Hanzo.)*

J. Yang, B. Yang, S. Chen, and Y. Zhang are with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: jianyang@ustc.edu.cn).

Y. Zhang is with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: yyzhang@winlab.rutgers.edu).

L. Hanzo is with the Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2018.2883627

transparent to the applications, which hinders their graceful in-network bitrate adaptation and stifles innovation in flawless services. In terms of these problems, Software-Defined Networking (SDN) [10] constitutes an ideal choice of eliminating those impediments. The SDN decouples the control plane from the individual network nodes and moves its functions into a centralized controller, whilst promoting collaboration between the network and the applications supported. Thus, we propose to eliminate those impediments in the new networking landscape of SDN and DSCN, which ultimately inspired us to conceive an intelligent scalable video transmission architecture, which exploits the collaboration of network access, routing and application adaptation.

Explicitly, we propose scalable video streaming over a DSCN with the aid of SDN, which allows the application to collaborate with the network for optimizing the resources allocation and hence improve the QoE for the users. To the best of our knowledge, this is the first contribution that takes the characteristics of SDN and SVC into consideration in DSCN. The main contributions of this paper include the following four aspects:

- We combine the advantages of DSCN and SDN to construct a programmable network. As a benefit of this framework, the SVC becomes capable of dynamically adjusting the number of video layers according to the prevalent network conditions estimated by the SDN controller. Furthermore, by exploiting the segment routing concept of [11] in the context of SDN, the flows of enhancement video layers can be routed to the user via multiple SCBSs by harnessing the collaboration of the SCBSs and hence improving the system's efficiency.
- Based on the prevalent network conditions estimated by the proposed system, we formulate our joint video enhancement layer selection and wireless resource allocation problem as that of maximizing the time-averaged utility in terms of the average QoE of all users subject to a specific resource constraint and rate constraint.
- By incorporating Lyapunov's stochastic optimization technique [12] into the mathematical problem proposed, the original problem is decomposed into a pair of optimization subproblems, which assists us in deriving a low-complexity video layer selection and wireless resource allocation strategy. Based on this decision strategy, we then invoke the segment routing technique for conceiving a routing strategy for each flow of the video layers to arrange for the cooperation of the SCBSs.
- The proposed algorithm relies on an online measurement-driven approach, operating without any prior statistical knowledge. The video layer selection, wireless resource allocation and transmission scheduling is carried out independently, so that our solution supports flexible scalability in deployment.

The rest of the paper is organized as follows. Section II presents the related research concerning DSCN as well as video transmission. In Section III, we elaborate on the design of SDN-aided scalable video streaming over DSCN as well as on the corresponding mathematical problem formulation. Section IV is devoted to deriving our joint video layer selection, wireless resource allocation and routing strategy conceived for dynamic video streaming over DSCN. Section V presents our experimental results for characterizing the attainable performance of the proposed approach. Finally, Section VI concludes the paper.

II. Related Work

In order to support flawless high-safe wireless connectivity, the concept of DSCN has been proposed as a new networking paradigm [13], which has received a considerable interest from both industry and academia. For instance, Kim and Cho [14] propose a joint sub-channel and power allocation scheme to maximize the system capacity. A joint admission and power control scheme was proposed by Luan *et al.* [15] for DSCN, which aims for maximizing the number of the connections admitted, while minimizing the transmission power. Improving the performance of DSCN in terms of interference management, backhauling and energy efficiency has been discussed in [16]–[22]. However, the above contributions only consider general data traffic optimization, regardless of the specific characteristics of the tele-traffic. Consequently, these approaches are not friendly for video streaming services due to ignoring the specific Quality of Service (QoS) requirements of flawless lip-synchronized video.

It should be noted that wireless video transmission has been richly characterized in the literature. Sophisticated optimization frameworks were proposed to deal with challenges of adaptive video transmission in wireless networks. Chen *et al.* [23] designed an in-network resource management framework, which considers fair resource allocation, and the stability of a user's bitrate. Abou-Zeid *et al.* [24] proposed an optimization framework for supporting energy-efficient video streaming relying on rate predictions in wireless networks. Cicaló, and Tralli [25] conceived a rate adaptation technique combined with optimized resource allocation in Orthogonal Frequency Division Multiple Access (OFDMA) for transmitting scalable video. Chen *et al.* [26] incorporated the relevant subjective quality constraints in the problem of video rate adaptation and admission control, thus improving the QoE for the video users. Similar contributions concerning cross-layer wireless video transmission can also be found in [27]–[29]. However, these valuable studies were conducted in the scenario of a single base station (BS), hence they are not applicable to the network model of dense small-cell systems.

In dense small-cell networks any user is typically within the range of multiple BSs. Hence, allowing the BSs to support cooperative transmission of video is a promising technique of improving the network's utility. Against this background, Bethanabhotal *et al.* [30] discussed the adaptive video segment downloading problem in DSCN and proposed a dynamic scheme capable of adaptive video quality reconfiguration and BS selection as well as dynamic allocation of the BS-to-user transmission rates. However, this approach has the drawback of coupling the slot duration of transmission scheduling with the size of video segments, which results in an increased start-up delay owing to the starvation of the receiver buffer. In order to rectify this deficiency, Miller *et al.* [31] invoked the classic Proportional-Integral-Derivative (PID) approach for

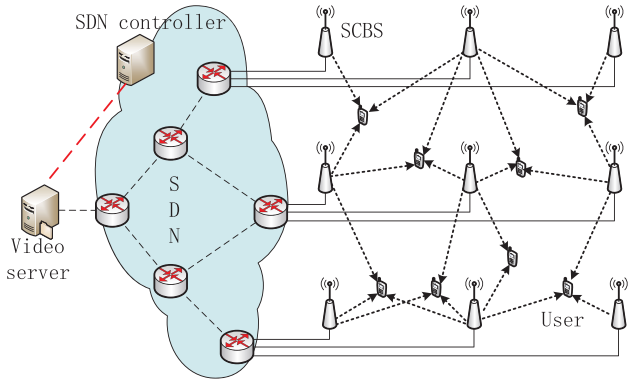


Fig. 1. Scalable video streaming over SDN-aided DSCN.

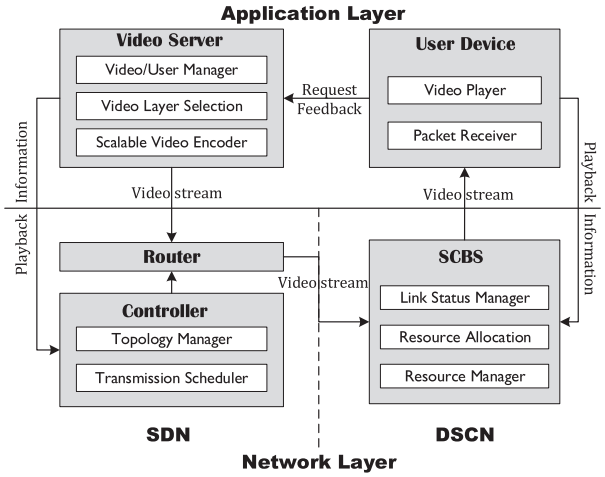


Fig. 2. Architecture of SDN-aided video streaming system.

stabilizing the fullness of the playout buffer for further increasing the QoE. Both the above mentioned treatises only focus on the wireless segment of the system under the idealized simplifying assumption of storing the video files at the BS, but neglecting the effects of the wired backhaul. By contrast, we consider the entire video transmission chain involving the wireless access, the routing and the application, rather than only the wireless access. Our proposed solution relies on the segment routing based SDN and SVC technique for achieving a fine-grained manipulation of the video traffic, which does not require caching of the video files at the BS.

III. System Model

In this section, we present the architecture of our SDN-aided scalable video streaming system operating in DSCN. We then formulate the mathematical problem of dynamically optimizing the resource allocation.

A. System Architecture

Fig.1 depicts the scalable video streaming system conceived for SDN-aided DSCN. The proposed system consists of five major components: the DSCN, the SDN network, the SDN controller, the video server and the users. In DSCN, SCBSs are deployed densely as access points, each of which connects to an edge SDN switch residing in the SDN network. The backhaul relies on an SDN network consisting of SDN switches. The centralized SDN controller has the capability of collecting the network conditions and configuring the switches' flow-control table. The video server is capable of storing and streaming videos, of managing users and of making decisions. The user devices are assumed to have multiple wireless interfaces, thus enabling them to receive video packets from multiple available SCBSs. With the aid of the SDN controller, the video server becomes capable of estimating the network conditions, including the network topology and link states for dynamically configuring the number of video layers to be streamed and transmitted, and then promptly triggers the SDN controller to appropriately schedule the different video layer flows to different SCBSs by beneficially configuring the routing paths. The SCBSs are capable of dynamically allocating the wireless resources to the associated users for delivering the appropriately selected video layers according to the prevalent channel states and to the throughput requirement of the video traffic. The above basic design assists us in combining our scalable video application, the SDN and DSCN to conceive a controllable video transmission system capable of adapting to the dynamically fluctuating network conditions.

The specific components mentioned above are illustrated in Fig. 2. The video server consists of the *Video/User Manager*, the *Video Layer Selection* and the *Scalable Video Encoder*. The *Video Layer Selection* acts as the heart of the Video Server conceived for dynamically adaptive video streaming. This component makes the optimal decisions concerning video layer selection for each users and the process is described in Section IV-B. The *Scalable Video Encoder* of Fig. 2 converts the video clip into a scalable stream having a base layer and several enhancement layers. The *Video/User Manager* of Fig. 2 is responsible for maintaining the video playback information for each users. In the SDN controller, we appropriately customize the *Topology Manager* and the *Transmission Scheduler* for assisting us in controlling the network and the selection of video streams. The *Topology Manager* of Fig. 2 estimates the network conditions, including the topology information and link states, which are used for beneficially scheduling the video transmission. The *Transmission Scheduler* of the Controller seen in Fig. 2 is employed for steering the flow of different video layers to an appropriate SCBS and this problem is presented in Section IV-D. Each SCBS of Fig. 1 has the following basic modules: the *Link Status Manager*, the *Resource Allocation* and the *Resource Manager*. The *Link Status Manager* of Fig. 2 is responsible for collecting the qualities of both wireless links of the SCBS and of its adjacent users gleaned through the users' feedback. The *Resource Allocation* of Fig. 2 makes wireless resource allocation decisions, which take both fairness and efficiency into account, and this problem is elaborated in Section IV-C. The *Resource Manager* of Fig. 2 is in charge of allocating the wireless resources to appropriate links according to the decisions made by the *Allocation Policy*. As for the user device, it contains the *Video Player* and the *Packet Receiver*. The *Packet Receiver* processes the packets received from one or several SCBSs and reassembles them into a single video

stream, so that the *Video Player* can decode and play back the video stream.

In order to derive a joint strategy of video layer selection, wireless resource allocation and transmission scheduling for our proposed framework, we will discuss the network model, wireless access model and scalable streaming model. Then in the remaining subsections we formulate a mathematical problem for characterizing the problem of streaming scalable video over the dense small-cell network.

B. Network Model

In order to exploit the diversity of SCBSs, the capability to flexibly scheduling different flows of video layers to multiple SCBSs is essential for our proposed system to attain fine-granularity manipulation of the video traffic. Here, we incorporate the advanced technique of SDN into DSCN for the sake of dynamic flow routing to different SCBSs. It should be noted that the video transmission path over this SDN-DSCN network may have to change frequently due to the dynamic channel quality fluctuation as well as owing to the dynamic activation of the video layers. Hence the SDN controller frequently reconfigures the flow-control tables of the switches in order to appropriately schedule the transmission path for each flow of video layers. The updating of the flow-control table in a SDN switch imposes a non-negligible delay, typically higher than 100msec, which results in the challenging problem of achieving fast rerouting for delivering video layers to SCBSs. By contrast, segment routing, which can be readily integrated with a controller-based SDN architecture, is capable of supporting rerouting without waiting for control messages in less than 50msec [32] which is a benefit of SDN as well as of the specific mechanism of segment routing. This motivates us to adopt the segment routing technique in constructing our proposed video delivery system.

In the context of our proposed network model, the segment routing is integrated as follows. The SDN controller relies on complete network knowledge, including the network's topology and its flows. An ingress switch triggers a request for a specific routing path to a destination with certain requirements in terms of, say, delay, bandwidth, diversity-order etc. The SDN controller computes an optimal path represented by a segment label list, in response to the requesting ingress switch. At this point, the switch becomes able to inject the traffic of the video layer by following the path specified by the segment list without any additional signaling for suitably configuring the network switches, thus achieving fast rerouting. Fig. 3 shows an example of the basic video transmission process with the aid of segment routing. The video server encodes a requested video into several layers and streams each layer into a separate flow. All flows will be forwarded to the ingress switch *A*. Then the controller calculates the forwarding path for each flow and returns the segment label lists to the ingress switch *A*. The switch *A* embeds the corresponding segment list into the head of each packet for implementing source routing. The switches of our network model relying on the assistance of segment routing no longer have to maintain a per-application and per-flow state, they simply obey the forwarding instructions encoded in the packet header in terms

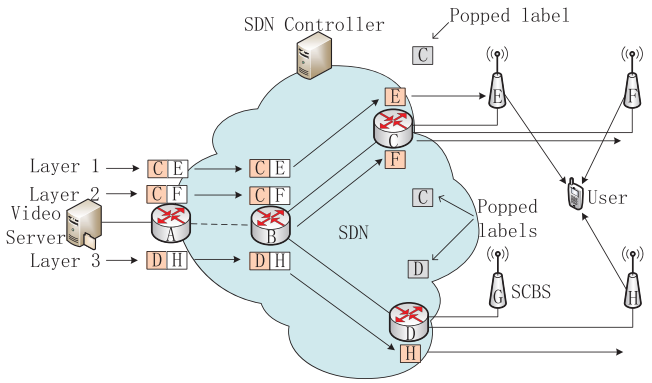


Fig. 3. Basic process of video transmission with the aid of segment routing.

of an ordered list of segments to follow. As illustrated in Fig. 3, *Layer 1* has to be transmitted from the ingress switch *A* to the SCBS *E*. When the packets of *Layer 1* arrive at switch *A*, it encodes the segment label list ("C-E") into the packet header according to the transmission path calculated by the SDN controller. Then, in the light of the top segment ID "C" in the label list, the subpath between the switch *A* and *C* can be chosen by employing the Open Shortest Path First (OSPF) protocol [33]. Once the packets arrive at the switch, it will pop the top label "C" and get its next-hop address "E" for routing the packets following the subpath from the switch *C* to the SCBS *E*. Following this procedure, we can achieve a fast rerouting for delivering the packets belonging to the same video session to diverse SCBSs.

Path/link failures are routinely encountered during the daily operation of complex heterogeneous network, both due to planned maintenance and unplanned events such as cable cuts, optical layer faults and other hardware/software bugs. Fortunately, the segment routing-based fast-rerouting solution, known as TI-LFA, provides natural network resilience against path/link failures, which has been discussed in [11]. Hence, as a benefit of segment routing, our proposed solution is capable of handling path/link failures. The wired network is typically stable due to its reliable transmission medium and has a high throughput than the wireless link. By contrast, the radio links in the wireless network have an inherently time-varying capacity and afford much lower throughput. This implies that the wireless access is likely to be the bottleneck in our system. Hence, in addition to the routing problem of the wired network side, we investigate the wireless resource allocation problem for dynamically streaming the scalable video streams.

C. Wireless Access Model

SCBSs typically operate in a time-slotted manner, where the time slot length depends on a specific communication protocol. For instance, the wireless time slot has a duration of 0.5ms in the context of Long Term Evolution (LTE) systems. Here, we consider SCBSs that apply orthogonal FDMA/TDMA and choose a different frequency for the neighboring SCBSs for achieving a sufficiently low cross-cell interference among the SCBSs.

We adopt a graph specified as $G_t = (U_t, S; E_t)$ to describe a DSCN, where t represents the time-slot index, U_t and S respectively denote the user set and the SCBS set, while the edge set E_t consists of all pairs $(s, u) \in S \times U_t$, which implies having a feasible transmission link between s and u at t. $\mathcal{N}_t(u) = \{s \in S : (s, u) \in E_t \wedge r_{su}(t) \geqslant r_{\min}\}$ is used for representing the set of the available SCBSs, which are capable of accommodating the video traffic to be conveyed to user u at t, where $r_{su}(t)$ denotes the transmission rate of the link between s and u at t, and $r_{\min}$ is the minimum acceptable transmission rate between the SCBS and the user. The user u will not be associate with the SCBS having $r_{su}(t) < r_{\min}$. Similarly, we use $\mathcal{N}_t(s) = \{u \in U_t : (s, u) \in E_t \wedge r_{su}(t) \geqslant r_{\min}\}$ to represent the users that are capable of being associated with the SCBS s, if the transmission rate satisfies $r_{su}(t) \geqslant r_{\min}$.

Here, the fading channel is assumed to be both time- and frequency-selective as in [30] and [34]. By employing OFDMA, the channel can be transformed into multiple narrow-band sub-channels in the frequency domain, where each sub-channel is time-selective and has a certain fading channel coherence time. The BSs is supposed to transmit the video data at a constant power level. The maximum achievable rate during the time slot t for the link $(s, u) \in E_t$ can be calculated according to Shannon's Continuous-Input Continuous-Output Memoryless Channel (CCMC) capacity formula given by

$$c_{s,u}(t) = W\mathbb{E}\left[\log\left(1 + \frac{Pg_{s,u}(t)}{1 + \sum\limits_{\substack{s' \neq s \\ s' \in \mathcal{N}_t(u)}} Pg_{s',u}(t)}\right)\right], \quad (1)$$

where P is the transmission power of SCBSs, and W is the bandwidth. Here, we let $g_{s,u}$ denote the path loss of the channel between s and u. The denominator of the Signal to Interference plus Noise Ratio (SINR) inside the logarithm in (1) includes the sum of the signal powers of all SCBSs $\{s' \in \mathcal{N}_t(u) : s' \neq s\}$ as well as the background noise, which models the inter-cell interference imposed on the user u, when detecting the signal received from SCBS s. It is feasible in practice to estimate the wireless channel state information. For example, in the LTE system pilot symbols are inserted in the downlink signals for the sake of estimating the channel. Hence, we assume that the user equipment is able to estimate the channel gain $g_{s,u}(t)$ and to feed it back to the associated BS. It should be noted that Eq. (1) represents the upper-bound of the transmission rate. In practice the BS adopts different digital Modulation and Coding Schemes (MCS) to control the transmission rate according to feedback as well as link layer protocol. Then the corresponding throughput is given by the Discrete-Input Continuous-Output Memoryless Channel (DCMC) capacity. Nevertheless, for simplicity we use Shannon formula to estimate transmission rate.

Since an SCBS may support multiple users, efficient resource allocation is necessary to meet bandwidth requirement of each user. In this paper, each SCBS makes wireless resource allocation decision for its associated users at the beginning of each slot. Accordingly, we use $a_{s,u}(t) \in [0, 1]$ to characterize the fraction of time/spectral resource that is allocated to user u at the time slot t. Naturally, it satisfies

$$\sum_{u \in U_t} a_{s,u}(t) \leqslant 1, \quad \forall s \in S. \quad (2)$$

Hence, the overall channel rate that user u attains at t is given by

$$c_u(t) = W \sum_{s \in \mathcal{N}_t(u)} a_{s,u}(t) \log\left(1 + \frac{Pg_{s,u}(t)}{\sigma_{s,u}}\right), \quad (3)$$

where $\sigma_{s,u} = 1 + \sum_{\substack{s' \neq s \\ s' \in \mathcal{N}_t(u)}} Pg_{s',u}(t)$.

TABLE I
MAPPING VIDEO QUALITY TO 1-5 SUBJECTIVE MOS SCALE

MOS Level	Number of Enhancement Layer	$l_u(t)$
5(Excelent)	4	5
4(Good)	3	4
3(Fair)	2	3
2(Poor)	1	2
1(Bad)	0 (base layer only)	1

D. Scalable Video Streaming Model

In this paper, we adopt the SVC technique [6], owing to its beneficial flexibility in adjusting the number of video streams, to support adaptive video transmission. The video sequences are encoded by a quality-scalable video encoder, generating a base layer and $(L-1)$ enhancement layers.

The Peak Signal to Noise Ratio (PSNR) or the Mean Opinion Score (MOS) are widely used as metrics to quantify the quality of video service. Without loss of generality, we employ the MOS level as the QoE metric in our problem formulation. Let $q_u(t)$ and $l_u(t) \in \{1, \cdots, L\}$ denote the MOS level and the number of video layers corresponding to user u at time t, respectively, where $q_u(t)$ is a function of $l_u(t)$, which is given by:

$$q_u(t) = f\left[l_u(t)\right], \quad f : l_u(t) \rightarrow MOS. \quad (4)$$

Generally, the MOS is proportional to the number of enhancement video layers. For instance, for a video with four enhancement layers, namely $l_u(t) \in \{1, 2, 3, 4, 5\}$, a commonly expressed 5-point MOS level (*i.e.*, 1-bad, 2-poor, 3-fair, 4-good and 5-excellent) is shown in Table I [35].

The *Layer Selection Policy* in the video server makes the relevant decisions to adjust the number of enhancement layers according to the prevalent network conditions estimated by the SDN controller. It should be noted that the interval of video layer selection is much longer than the length of the wireless time slot. The reason for this is that the layer switching is usually conducted at the Groups of Pictures (GOPs) boundary having a duration on the order of hundreds of milliseconds. For simplicity, we assume that the duration of each GOP is constant and equals T times the wireless time slot length. Thus, the video layer selection is made every T wireless time slots.

Once the number of video layers $l_u(t)$ is decided, the video layer configuration remains unchanged within the next T wireless time slots. Let $d_u^t(l)$ denote data packet size of the GOP containing the first $l_u(t)$ video layers, while $r_u(t)$ is defined

as the playback bit rate corresponding to the user u, which is kept constant during the time interval $t \in [(k-1)T, kT]$. Then, we have

$$r_u(t) \triangleq d_u^t(l)/T_{\mathrm{GOP}}, \quad t \in [(k-1)T, kT]. \tag{5}$$

Playback interruption is another challenge in video services, which significantly reduces the QoE perceived by the viewers. In order to achieve a smooth video playback, the transmission rate has to be kept higher than the video bitrate. This means that the following rate constraint should be satisfied for avoiding playback interruptions:

$$\overline{r_u} \leqslant \overline{c_u}, \tag{6}$$

where $\overline{r_u}$ and $\overline{c_u}$, respectively, represent the time-averaged video playback bitrate and the time-averaged transmission rate of user u. Their specific definitions are given by

$$\overline{r_u} \triangleq \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[r_u(\tau)], \tag{7}$$

and

$$\overline{c_u} \triangleq \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[c_u(\tau)]. \tag{8}$$

It should be noted that the time-averaged constraint (6) also ensures the stability of the queuing system in SCBSs. Hence, the average queue length is bounded, which implies that (6) implicitly guarantees the delay constraint as a QoE metric.

E. Problem Formulation

The instantaneous utility in terms of averaged QoE of all users is defined as

$$Q(t) = \frac{1}{N_t} \sum_{u\in U_t} q_u(t), \tag{9}$$

where $N_t = |U_t|$ denotes the number of users in the system. Then, the time-averaged utility is further defined as

$$\overline{Q} \triangleq \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[Q(\tau)]. \tag{10}$$

Since our aim is to maximize the system's utility, while satisfying the resource allocation constraint (2) and the transmission constraint (6), the Dense Small-Cell Utility Maximization (DSCUM) of our scalable video streaming problem is formulated as:

$$\underset{a_{s,u}, l_u}{\text{Maximize}} \ \overline{Q} \tag{11}$$

$$\text{Subject to } \overline{r_u} \leqslant \overline{c_u}, u \in U_t \tag{12}$$

$$a_{s,u}(t) \in [0,1], \quad \forall t \tag{13}$$

$$\sum_{u\in U_t} a_{s,u}(t) \leqslant 1, \quad \forall s \in S \tag{14}$$

$$l_u(t) \in \{1, \cdots, L\}, \quad \forall t. \tag{15}$$

The problem (11)-(15) is a classical constrained stochastic optimization problem. However, it is difficult to obtain *a priori* knowledge of the stochastic process concerning the overall available bandwidth $c_u(t)$. This motivates us to employ the powerful Lyapunov optimization framework of [12] to develop an online measurement based solution in the next section, instead of relying on any *a priori* statistical knowledge of the system.

IV. Joint Video Layer Selection, Wireless Resource Allocation and Transmission Scheduling

In this section, we commence by invoking Lyapunov drift and the associated optimization theory for converting the time-averaged optimization problem into a minimization problem. Furthermore, we decompose the minimization problem into two independent sub-problems corresponding to video layer selection and to wireless resource allocation presented in Section IV-B and Section IV-C, respectively. Then we discuss the transmission path scheduling and its implementation in Section IV-D according to the results of optimal video layer selection and wireless resource allocation. Finally, we present the intricate cooperation of the algorithms in Section IV-E.

A. Lyapunov Stochastic Optimization Formulation

Since the constraint represented by Eq. (12) is based upon time-averaged values, we construct a virtual queue $H_u(t)$ in the light of the Lyapunov optimization method, which can be employed for satisfying the time-averaged rate constraint. The update process for $H_u(t)$ is given by:

$$H_u(t+1) = [H_u(t) + r_u(t) - c_u(t)]^+, \tag{16}$$

where we have $(x)^+ \triangleq \max(x, 0)$. As expected, the virtual queue $H_u(t)$ is mean-rate-stable, hence we have $\lim_{t\to\infty} \frac{\mathbb{E}[|H_u(t)|]}{t} = 0$, because $H_u(t)$ is finite at any time. Thus, the time-averaged rate constraint (12) is satisfied according to [12]. Then, we define a quadratic Lyapunov function as

$$L(t) = \frac{1}{2} \sum_{u\in U_t} H_u(t)^2. \tag{17}$$

Let $\mathbf{H}(t) \triangleq \{H_u(t), u \in U_t\}$ denote a concatenated vector of the virtual queue length. Furthermore, we define the corresponding conditional Lyapunov drift from the time slot t to the time slot $(t+T)$ as

$$\Delta_T(t) \triangleq \mathbb{E}[L(t+T) - L(t)|\mathbf{H}(t)]. \tag{18}$$

According to Lyapunov's drift theory, minimizing the Lyapunov drift $\Delta_T(t)$ can achieve mean-rate-stability of the virtual queues, and thus enforces the constraints to be satisfied by the time-averaged video download bitrate and by the time-averaged video playback bitrate.

Since the objective of the joint layer selection, wireless resource allocation and transmission scheduling is to maximize the system's utility subject to the constraints, we define a T-slot "drift-plus-penalty" in the context of our Lyapunov optimization framework, which integrates the Lyapunov drift and the instantaneous system utility in T time slots $\sum_{\tau=t}^{t+T-1} Q(\tau)$ as follows:

$$\Delta_T(t) - V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1} Q(\tau)|\mathbf{H}(t)\right], \tag{19}$$

where $\mathbb{E}\left[\sum_{\tau=t}^{t+T-1} Q(\tau)|\mathbf{H}(t)\right]$ is the expected system utility over T time slots, and the parameter V is a non-negative value invoked for striking a trade-off between the system's utility and the queue stability in our control strategy. Hence, the optimization problem (11)-(15) is further reformulated as making a decision on the wireless resource allocation and on the video layer selection to minimize the "draft-plus-penalty" of (19).

Below, we present a bound on (19) for deriving the online optimization algorithm. It should be noted that the video layer selection is conducted at the time-granularity of every T wireless time slots. Hence, the future queue length $H(\tau), \tau \in [t, t+T-1]$ has to be known for video layer selection at the time slot t, when we employ (19) for optimizing the video layer selection. However, it is difficult to predict the future queue length $H(\tau)$, since it depends not only on the video layer selection at the time slot t, but also on the wireless resource allocation during $[t, \tau]$. In order to overcome this difficulty, following the approximation technique proposed in [36], we use the current queue length $H(t)$ as an approximate value of the near-future queue length $H(\tau), t \leqslant \tau \leqslant t+T-1$ for deriving a "loose" upper bound of the drift-plus-penalty, which has been summarized in **Lemma 1**.

Lemma 1: For any feasible decision concerning the video layer selection and the wireless resource allocation at the instant $t = kT$, where k can be any non-negative integer, we have

$$\begin{aligned}\Delta_T(t) - V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1} Q(\tau)|\mathbf{H}(t)\right] \\ \leqslant B + T\sum_{u\in U_t}\mathbb{E}\left[H_u(t)r_u(t) - \frac{V}{N_t}q_u(t)|\mathbf{H}(t)\right] \\ - \sum_{\tau=t}^{t+T-1}\mathbb{E}\left[\sum_{u\in U_t} H_u(t)c_u(\tau)|\mathbf{H}(t)\right],\end{aligned} \tag{20}$$

where the positive constant B is defined as

$$B \triangleq \frac{1}{2}T^2 N_t(r_{\max}^2 + c_{\max}^2).$$

Proof: Please see the Appendix A. ∎

According to the Lyapunov drift and Lyapunov optimization, instead of directly minimizing the "drift-plus-penalty" (19), we solve the DSCUM problem (11)-(15) by minimizing the "loose" bound (20) given in **Lemma 1**, which is formulated as

$$\begin{aligned}\underset{a_{u,s},l_u}{\text{Minimize}}\ & T\sum_{u\in U_t}\mathbb{E}\left[H_u(t)r_u(t) - \frac{V}{N_t}q_u(t)|\mathbf{H}(t)\right] \\ & - \sum_{\tau=t}^{t+T-1}\mathbb{E}\left[\sum_{u\in U_t} H_u(t)c_u(\tau)|\mathbf{H}(t)\right] + B \\ \text{Subject to}\ & a_{s,u}(t) \in [0,1], \quad \forall t \\ & \sum_{u\in U_t} a_{s,u}(t) \leqslant 1, \quad \forall s \in S \\ & l_u(t) \in \{1, \cdots, L\}, \quad \forall t.\end{aligned} \tag{21}$$

However, the solution of (21) is a sub-optimal solution for the DSCUM problem. Thus, we present **Lemma 2** for theoretically quantifying the performance bound of our sub-optimal solution.

Lemma 2: Suppose that the virtual queue length is initially zero, *i.e.*, $\mathbf{H}(t) = \mathbf{0}$. For any positive value V, the average system utility $\overline{Q_{subopt}}$ obtained by solving the problem (21) satisfies

$$\overline{Q_{subopt}} \geqslant \overline{Q^*} - \frac{B}{TV}, \tag{22}$$

where $\overline{Q^*}$ is the optimal utility of the problem (11)-(15).

Proof: Please see Appendix B. ∎

Lemma 2 implies that the sub-optimal average utility asymptotically approaches the optimal value $\overline{U^*}$ as V increases. While, the parameter V represents the weight of the penalty and a large value of V may degrade the stability of the virtual queue $H_u(t)$. Hence, in Section V we have conducted an experiment to analyze the performance sensitivity wrt the trade-off factor V.

Besides, it should be noted that for a given $H(t)$, the second term of the objective function in (21) only depends on the video layer selection variables $l_u(t)$, while the third term of the objective function only depends on the wireless resource allocation variables $a_{s,u}(t)$. This beneficial structure of the bound (20) allows us to decompose the minimization problem into two individual sub-problems, namely, that of minimizing the second term for optimizing the video layer selection and maximizing the third term for determining the wireless resource allocation.

B. Video Layer Selection

Since the video layer selection is decided at the beginning of $t = kT (k = 0, 1, 2, \ldots)$, the queue length vector $\mathbf{H}(t)$ can be observed and the number of video layers is kept unchanged in the future $(T-1)$ time slots. In order to deduce the video layer selection policy, we can minimize the expectation of the second term of the objective function in (21), which can be rewritten as:

$$\underset{l_u}{\text{Minimize}}\ \sum_{u\in U_t}\left[H_u(t)r_u(t) - \frac{V}{N_t}q_u(t)\right] \tag{23}$$

$$\text{Subject to}\ l_u(t) \in \{1, \cdots, L\}, \quad \forall u \in U_t. \tag{24}$$

The component $[H_u(t)r_u(t) - \frac{V}{N_t}q_u(t)]$ in (23) is mutually decoupled, and only relies on the video layer selection variable $l_u(t)$. This decoupling property of (23) makes it possible to further decompose the problem (23)-(24) into N_t independent sub-problems corresponding to each user as

$$\underset{l_u}{\text{Minimize}}\ \left[H_u(t)r_u(t) - \frac{V}{N_t}q_u(t)\right] \tag{25}$$

$$\text{Subject to}\ l_u(t) \in \{1, \cdots, L\}, \quad u \in U_t, \tag{26}$$

noting that we maintain the virtual queue length $H_u(t) = [H_u(t-1) + r_u(t-1) - c_u(t-1)]^+$ independently for each video session corresponding to the user u. According to (25), the queue length $H_u(t)$ can be treated as a weight factor imposed on the video bitrate. This implies that for a larger value of $H(t)$, a higher cost has to be paid for increasing the number of video layer. When $H(t)$ becomes too large, an

adjustment would be triggered to reduce the number of video layers, thus resulting in a reduction of $H(t)$. The parameter V is a positive constant invoked for striking a tradeoff between the system's utility and drift. A large value of V contributes to improving the system's utility, while a small value of V is beneficial for stabilizing the virtual queues.

In practice, the video sequence is encoded into a very limited number of layers, for example to 2-6 enhancement layers. Hence, the optimal number of video layers for user u can be determined by enumerating all possible layers $l \in \{1, \cdots, L\}$ to minimize (25). Thus, the computational complexity of the video layer selection is on the order of $O(N_t)$, which is determined by the number of users and is particularly low owing to its fugally designed limited search-space and low update frequency taking place at intervals of every T time slots.

C. Wireless Resource Allocation

In this subsection, we maximize the third term of the objective function in (21) to optimize the wireless resource allocation for scalable video streaming. Specifically, for any $t \in [kT, (k+1)T-1], k = 0, 1, 2, \ldots$, by substituting (3) into the third term of the objective function in (21), we rewrite the wireless resource allocation problem under the constraints (13) and (14) as follows:

$$\underset{a_{s,u}}{\text{Maximize}} \ W \sum_{s \in S} \sum_{u \in U_t} H_u(kT) a_{s,u}(t) G_{s,u}(t) \tag{27}$$

$$\text{Subject to } a_{s,u}(t) \in [0,1], \quad \forall t \in [kT, (k+1)T-1] \tag{28}$$

$$\sum_{u \in U_t} a_{s,u}(t) \leqslant 1, \quad \forall t \in [kT, (k+1)T-1], \tag{29}$$

where the new notation of $G_{s,u} = \log\left(1 + {}^{Pg_{s,u}(t)}/_{\sigma^2_{s,u}}\right)$ is introduced to simplify the exposition.

In (27), the parameters $H_u(kT)$ and $G_{s,u}(t)$ can be estimated *a priori* before optimizing the resource allocation variables. Hence, the optimization problem (27)-(29) is a classical discrete linear programming problem. Furthermore, (27) also satisfies a decoupling property at the granularity of BSs, *i.e.* the term of the sum $\sum_{u \in U_t} H_u(kT) a_{s,u}(t) G_{s,u}(t)$ are independent of each other. This lays the foundation of achieving distributed resource allocation by solving the following $|S|$ sub-problems:

$$\underset{a_{s,u}}{\text{Maximize}} \ \sum_{u \in U_t} H_u(kT) a_{s,u}(t) G_{s,u}(t) \tag{30}$$

$$\text{Subject to } a_{s,u}(t) \in [0,1], \quad \forall t \in [kT, (k+1)T-1] \tag{31}$$

$$\sum_{u \in U_t} a_{s,u}(t) \leqslant 1, \quad \forall t \in [kT, (k+1)T-1]. \tag{32}$$

Note that for a certain BS s, it is easy to solve the classical linear programming problem (30)-(32) by allocating all wireless resources to the specific user u, who has the highest product of $H_u(kT)G_{s,u}(t)$. In this way, we can obtain the optimal wireless resources for the entire system. However, each station carries our its the wireless allocation decision at every wireless slot (around tens of milliseconds), which leads to a change of link connection from time to time. This imposes a challenge on the transmission schedule, since the SDN has to frequently configure forwarding paths for each flow to adapt to frequent variations of the link connections.

In order to reduce the frequency of flow-scheduling, we reconfigure the wireless resource allocation on a per GOP basis. Thus, the sub-problem for each BS is given as follows:

$$\underset{a_{s,u}}{\text{Maximize}} \ \sum_{u \in U_t} H_u(kT) a_{s,u}(kT) G_{s,u}(kT) \tag{33}$$

$$\text{Subject to } a_{s,u}(kT) \in [0,1] \tag{34}$$

$$\sum_{u \in U_t} a_{s,u}(kT) \leqslant 1. \tag{35}$$

It is plausible that if we allocate wireless resources at the beginning of each GOP, a user will occupy all wireless resources until the next slot according to (33)-(35). If the GOP duration is too long, any inflexible wireless resource allocation strategy may lead to unfairness for users and would result in the problem of either wasting or lacking bandwidth, which may lead to video playback interruptions and imposes significant negative impacts on the QoE of users. Motivated by this, we proceed to design a sub-optimal solution for the problem (33)-(35) to achieve fairness among of the users as a trade-off.

In order to solve the above problem, the fair resource allocation algorithms of [37] and [38] can be employed. For simplicity, we use the classical weighted fair queue (WFQ) technique of [39], which is one of the most well-known queue-scheduling algorithms, in order to solve the problem (33)-(35). Each BS allocates wireless resources for the available users according to the virtual queues $H_u(kT), u \in \mathcal{N}_t(s)$. Based on (33), $G_{s,u}(kT)$ can be regarded as the weight of the corresponding virtual queue. Thus, the resource allocation solution of each BS relying on WFQ can be expressed as follows:

$$a_{s,u}(kT) = \frac{H_u(kT) G_{s,u}(kT)}{\sum_{i \in \mathcal{N}_t(s)} H_i(kT) G_{s,i}(kT)}. \tag{36}$$

Specifically, $\sum_{i \in \mathcal{N}_t(s)} H_i(kT) G_{s,i}(kT) = 0$ means that the wireless resources are equitably allocated to all users. The computational complexity of the wireless resource allocation algorithm is on the order of $O[\mathcal{N}_t(s)]$ for a certain BS, which is related to the number of users supported by the BS, which is low.

In this paper, we aim for designing a comprehensive video transmission system in DSCN to improve the average QoE of users. Instead of applying the fine-grained per-time-slot control, we use coarse-grained grained control for tackling the bottleneck of SCBSs. Furthermore, we adopt WFQ to achieve fairness for the users, which may slightly reduce wireless resource utilization. This allows each user to achieve fairness of wireless resource allocation at the time granularity of a GOP duration, as well as allows the SDN controller to have sufficient time to appropriately schedule the transmission paths.

D. Transmission Path Scheduling and Implementation

The video layer selection and wireless resource allocation have been figured in Sections IV-B and IV-C. The remaining problem is that of scheduling the transmission path for the

specific video flows having limited wireless resources. In this part, we proposed a heuristic algorithm, which is executed by the SDN controller, for scheduling the transmission paths for all video flows. The switches forward those flows to the appropriate BSs with the aid of segment routing. In general, the BSs store the video packets in their own buffer until forwarding them to the corresponding users, given the potential bandwidth mismatch between the wired and wireless network. It should be noted that the length of video buffered in BSs will not occupy too much storage space, since we have guaranteed that the virtual queue $H_u(t)$ is mean-rate-stable by employing sophisticated Lyapunov optimization techniques.

As described above, both the video layer selection and the wireless resource allocation are executed at $t = kT, k = 0, 1, 2, \ldots$. Let $\tilde{r}_{u,l}(t)$ denote the bitrate of the lth video layer at the time instant t, while $\tilde{c}_{s,u}(t)$ represents the achievable channel rate between user u and the SCBS s according to the wireless resource allocation. Let $\psi_u(t) = \sum_{s \in \mathcal{N}_t(u)} \tilde{c}_{s,u}(t)$ denote the total wireless channel bit rate allocated to user u, while let $\rho_u(t) = \sum_{l=0}^{l^*} \tilde{r}_{u,l}(t)$ represent the total throughput demand of user u, where l^* denotes the optimal video layer. In order to ensure fairness between the flows of different video layers, we use a proportionally fair resource allocation scheme to determine the channel bit rate $\omega_{u,l}(t)$ assigned to the lth video layer, which is given by:

$$\omega_{u,l}(t) = \frac{\tilde{r}_{u,l}(t)}{\rho_u(t)} \psi_u(t). \tag{37}$$

Considering the features of SVC, the base layer is the most important layer and the enhancement layers cannot be decoded unless the base layer as well as all the lower enhancement layers are available. Naturally, it would be better to forward video layers through a single path and use multiple paths as infrequently as possible for avoiding the extra decoding complexity imposed by the out-of-order packets. Hence, we assign a higher priority to the lower video layers, when scheduling the routing path. First we calculate the bitrate required for each video layer according to Eq. (37). Then, we assign bandwidth for each video layers in the order of their priority. Here we assign the video layer associated with the highest priority to the SCBS having the maximum available bandwidth. We summarize our proposed transmission scheduling algorithm for a certain user u in **Algorithm 1**, where we omit the index u as well as t for simplicity. Here, we use $\upsilon(s, l)$ to denote the channel bitrate of s allocated to user u for transmitting the lth video layer as the output in the proposed algorithm. This algorithm can be readily carried out in the video server at a computational complexity of $O(s)$.

Algorithm 1 Transmission Scheduling Algorithm

Input: The achievable rate for the user offered by SCBS s, $\tilde{c}_s$; The bitrate of the lth video layer,$\tilde{r}_l$; The optimal video layer,l^*.

Output: The channel bitrate of s allocated to the user for transmitting the lth video layer.

Calculate the bitrate for each video layer ω_l according to Eq. (37)
for $l = 1$ to l^* **do**
  $temp \leftarrow 0$
  repeat
    $j^* \leftarrow \arg\max_j \tilde{c}_j$
    $\upsilon(j^*, l) \leftarrow \tilde{c}_{j^*}$
    $temp \leftarrow temp + \tilde{c}_{j^*}$
    $\tilde{c}_{j^*} \leftarrow 0$
  until $temp \geqslant \omega_l$
  $\tilde{c}_{j^*} \leftarrow temp - \omega_l$
  $\upsilon(j^*, l) \leftarrow \upsilon(j^*, l) - \tilde{c}_{j^*}$
end for

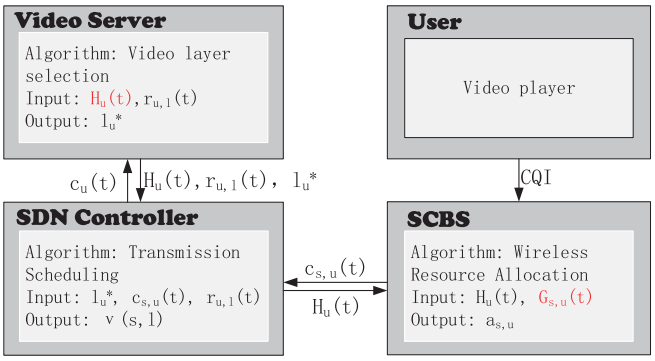


Fig. 4. Message exchanges of our proposed algorithms.

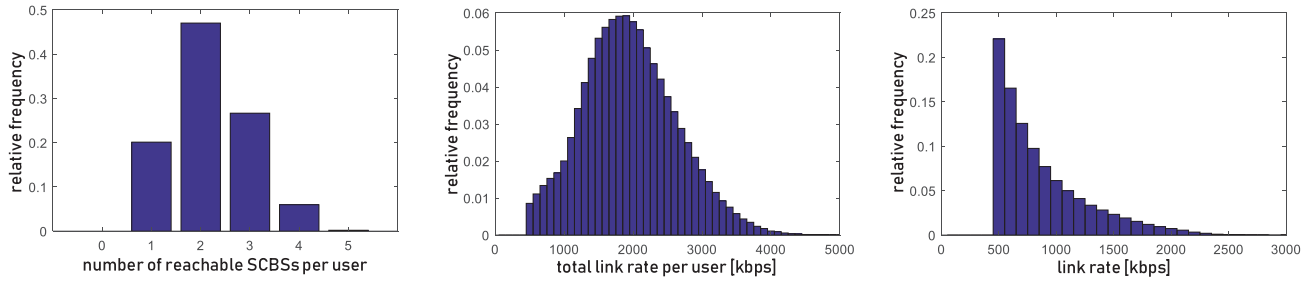
E. Interaction and Cooperation Among of the Algorithms

As described above, we proposed three algorithms for solving the video layer selection, wireless resource allocation and transmission scheduling. Below, we discuss the interaction and cooperation of these algorithms to demonstrate the integrated operation of scalable video streaming over DSCN.

Fig. 4 characterizes the message exchanges of our proposed algorithms. The video server is in charge of making decision on the video layer selection, which depends both on the virtual queue $H_u(t)$ and on the bitrate of each layer $r_{u,l}(t)$ as its input parameter. Furthermore, $H_u(t)$ is calculated by the video server according to (16), where $c_u(t)$ is obtained from the SDN controller and $r_u(t)$ is available for the video server. After determining the optimal number of video layers l_u^* for each user, the video server sends $H_u(t)$ to SCBS through the SDN controller. Then, the wireless resource allocation algorithm relies on $H_u(t)$ and $G_{s,u}(t)$ to assign the resources to user u, where $G_{s,u}(t)$ can be calculated by itself according to the user's Channel Quality Indicator (CQI) report. The SDN controller acquires $r_{u,l}(t)$ as well as l_u^* from the video server and $c_{s,u}(t)$ from the SCBSs to schedule the transmission path for each video flow.

V. Performance Evaluation

In this section, we present our experimental results for characterizing the performance of the proposed dynamic scalable video transmission (DSVT) algorithm. For the sake of performance comparison, we implemented the PID algorithm proposed in [31] as a benchmarker of our experiments. Furthermore, in order to analyze the optimality of



(a) Available SCBSs of a user (b) Achievable throughput for a user (c) Bit rate for a link

Fig. 5. Connectivity statistics.

our proposed algorithm, we also employed the non-fairness DSVT (NFDSVT) algorithm, which maximizes the system's utility without considering the fairness of users, as another benchmarker algorithm.

A. Evaluation Setup

In our experiments, we considered the scalable video streaming system described in Section III and employed in area of 100m $\times$ 100m square, which contains 25 SCBSs. We divided the square into 5×5 small square cells having the side length of 20m and each SCBS was located at the center of a small square-shaped cell. Those SCBSs chose different frequency bands around 5 GHz and the adjacent SCBSs selected orthogonal channels to reduce interference among them. We set the bandwidth of each SCBS to 330 kHz and set the transmission power to 43 dBm. The fading gain $g_{s,u}(t)$ between user u and the SCBS s was characterized by the WINNER II channel model [40] as $g_{s,u}(t) = 10^{-\frac{PL[d_{s,u}(t)]}{10}}$, where $d_{s,u}(t)$ was the distance between the BS s and the user u at t, and

$$PL(d) = A\log(d) + B + C\log(0.25 f_0) + \chi_{dB}. \quad (38)$$

In (38), the units of d and the carrier frequency f_0 are meter and GHz, respectively, while χ_{dB} denotes log-normal shadowing variable having a variance of σ^2_{dB}. The parameters A, B, C and σ^2_{dB} are scenario-dependent constants. According to the A1 model [40] listed in WINNER II, we set $A = 18.7, B = 46.8, C = 20, \sigma^2_{dB} = 9$ in the condition of $3 \leqslant d \leqslant 100$. In order to adapt this model to our experiments, we referred to [31] and extended the model by defining $PL(d) = PL(3), 0 \leqslant d \leqslant 3$. Due to the video transmission requirements, the wireless links having a bitrate below 0.5Mbps are deemed inadequate.

In order to obtain an intuitive understanding of the network conditions, we conducted an experiment, where a user is randomly positioned and collected the link states. We repeated the experiment ten thousand times and the corresponding results were shown in Fig. 5. Observe from Fig. 5(a) that the probability distribution of the number of SCBSs that are capable to serve the same user at the same time is consistent with our previous declarations. Explicitly, most users have a chance of receiving packets from multiple SCBSs. Fig. 5(b) characterizes the probability distribution of maximum achievable throughput for a user, which supports that the throughput approximately obeys the normal distribution with a mean of 2000 kbps. Fig. 5(c) shows the probability distribution of the bit rate corresponding to a link, which reflects that most links have a transmission rate of around 500 to 1000 kbps.

TABLE II
STATISTICS OF THE VIDEO SEQUENCES

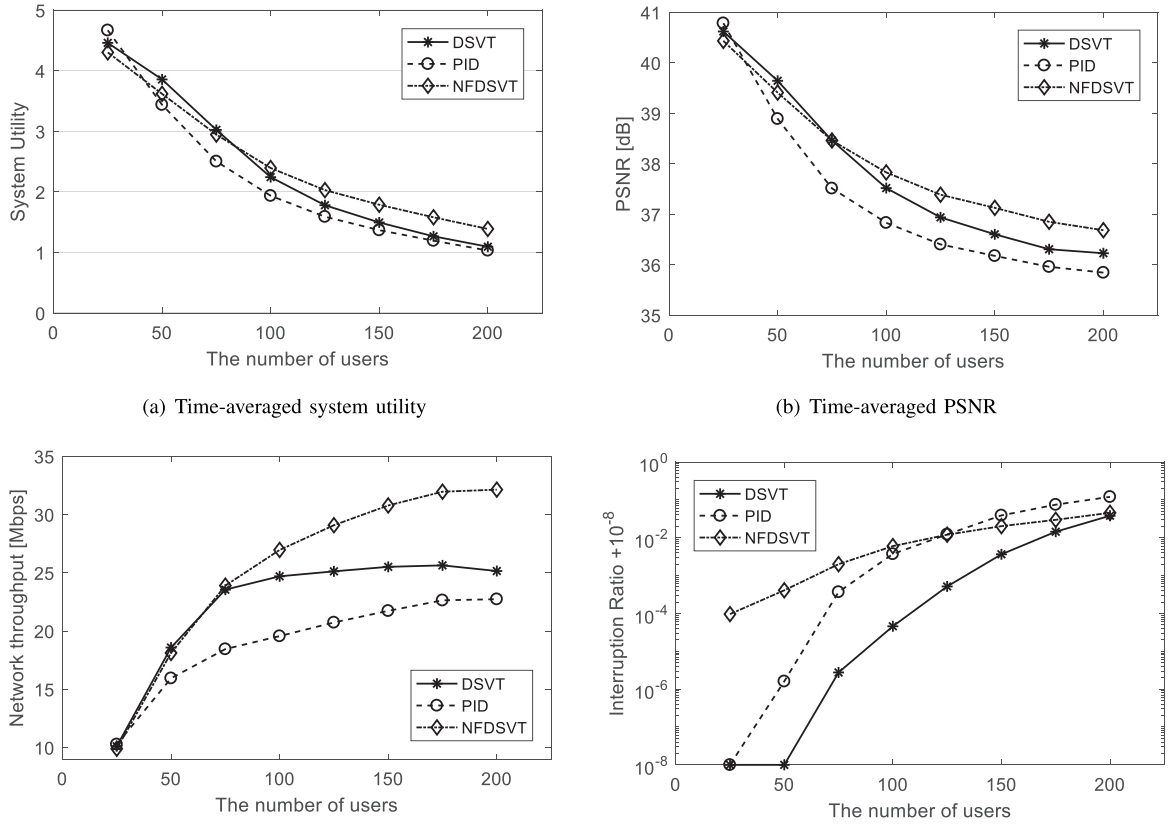
SVC Trace	Gandhi		The Terminator		Die Hard	
	Bitrate (kbps)	PSNR (dB)	Bitrate (kbps)	PSNR (dB)	Bitrate (kbps)	PSNR (dB)
Layer 1	114.83	35.10	145.34	34.89	79.28	37.30
Layer 2	139.66	36.31	128.76	36.18	125.04	39.02
Layer 3	80.07	37.77	115.27	37.80	56.62	40.77
Layer 4	54.82	38.78	72.94	38.86	36.09	41.81
Layer 5	65.96	40.27	77.93	40.15	36.22	43.01

The video sequences used in our experiments are "Gandhi", "The Terminator" and "Die Hard" [41]. Those videos are encoded by using the H.264/SVC reference software JSVM [42] at 30 Frame Per Second (FPS) and each video has 1 base layer and 4 enhancement layers. The GoP contains an I frame and 15 B frames, *i.e.*, we have a GOP size of 16 [43]. We listed the statistics of the video traces in Table II. The users in the experiment below randomly request one of the above videos to play.

B. Performance Comparison

In this subsection, we evaluate the performance of our algorithm by comparing it to both the PID and NFDSVT. The PID algorithm employs the PID controller to stabilize the length of playout buffer at the appropriate target value for users by adjusting the video quality as well as the wireless resource allocation, so as to improve the QoE of users. Here, we adopted the PID parameters recommended in [31]. As for NFDSVT, it adopts the same control policy as DSVT, apart from applying WFQ for user fairness. Since we aim for improving the QoE, we design a series of experiments to measure the performance metrics in terms of system utility, PSNR of the received video, the total network throughput and interruption ratio. In order to acquire comprehensive results concerning the system's performance, we varied the number of users from 25 to 200. The experimental results below were averaged over 50 independent runs.

Fig. 6(a) shows the time-averaged system utility defined in (10), which indicates the mean QoE of the users. Naturally,



(a) Time-averaged system utility

(b) Time-averaged PSNR

(c) Time-averaged network throughput

(d) Time-averaged playback interruption ratio

Fig. 6. Performance comparison in terms of time-averaged system utility, PSNR, network throughput and interruption ratio.

when a user has no data in the buffer to play back in the next frame, the QoE of this user is substantially reduced, and we defined that by assigning $MOS = 0$ in that time slot. We also used the PSNR to characterize the quality of the received video, as shown in Fig. 6(b). It can be observed from Fig. 6(a) and Fig. 6(b) that our algorithm achieves a higher system utility and average PSNR than PID. In combination with Table II, we can observe that DSVT increases the number of video layer by an extra layer. The basic reason for this is that PID aims for stabilizing the fullness of the playback buffer, which is not consistent with the goal of maximizing the utility in terms of the averaged QoE. By contrast, our proposed algorithm explicitly considers the utility maximization as formulated in the problem (11)-(15). Another reason for the performance erosion of PID is that the classical PID method does not perform well for complex, nonlinear and time-varying systems. As for NFDSVT, it achieves a higher system utility and higher average PSNR, when the number of system users is high. The reason for this trend is that the NFDSVT allocates all wireless resources to the particular user who can achieve the maximum system utility. However, NFDSVT achieves a lower system utility and lower PSNR, when the system has a low workload. Furthermore, NFDSVT suffers from a higher interruption ration than DSVT, because all resources are occupied by a single user and the other users cannot receive packets after the next T slots. Fig. 6(a) and Fig. 6(b) show that MOS and PSNR decrease upon increasing the number of users. This is caused by the limited wireless resources shared by the users. As the number of users increases, the average resource per user is reduced correspondingly, which results in a reduced average bit rate of the video session. These results illustrate that both our algorithm and PID have the capability of adapting themselves to different system loads.

In order to evaluate the network's efficiency, we recorded the network throughput of all SCBSs for different number of users. The time-averaged network throughput was plotted in Fig. 6(c). This result illustrates that when the system has a low workload, for instance the number of users is below 100 as shown in Fig. 6(c), the network throughput rises sharply in line with the increased the number of users, and both DSVT as well as NFDSVT achieve a more steep network throughput increase than PID. When the system's workload is relatively high, for example the number of users is above 100 as shown in Fig. 6(c), NFDSVT achieves the highest network throughput, owing to allocating all resources to the specific user having the maximum transmission rate. By contrast, the throughput of DSVT remains relatively stable in order to maintain the QoE fairness among the users. As for PID, it exhibits a slow growth, but overall its throughput remains lower than that of the other algorithms. There may be two basic reasons for these phenomena. The first one is that the PID controller is driven only by the feedback of the control error of the client playback buffer without utilizing any other information concerning the physical system. This also implies

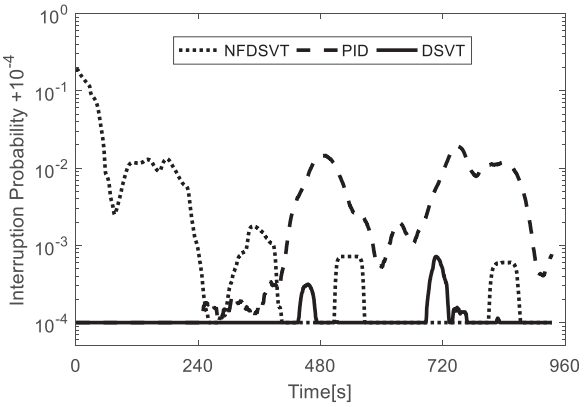
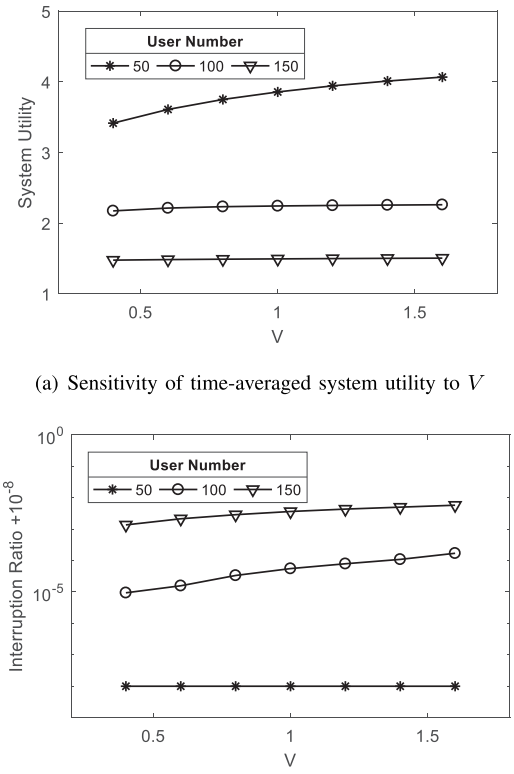


Fig. 7. Interruption probability of the video streaming system having 100 system users.

that the PID treats the transmission system as a "Blackbox", and less information is not beneficial for improving the system performance. By contrast, DSVT uses a system model that characterizes the state of each wireless channel during video transmission. The second reason is that our algorithm employs the Lyapunov optimization method to make joint decisions for video layer selection, wireless resource allocation and dynamic routing. However, PID manually and heuristically sets its parameters to control the client buffer fullness, and invokes a heuristic algorithm to allocate wireless resources. Both of them may impose a performance degradation in terms of the network's throughput.

Video playback interruption is another important factor degrading the users' QoE. Here, we define the playback interruption ratio as the proportion of the interruption time to the total playback time. We evaluated the playback interruption ratio of our system under different system workloads, and the corresponding experimental results were plotted in Fig. 6(d). This result demonstrates that upon increasing the number of users, the likelihood of playback interruptions is increased owing to the limited overall wireless resources and reduced average available resources of each user. Fig. 6(d) also shows that our proposed algorithm achieves the lowest playback interruption ratio. The main reason for this is that our proposed algorithm applies an explicit rate constraint (6) to the time-averaged video playback bitrate, which is lower than the time-averaged transmission rate, hence resulting in a reduced playback interruption ratio upon increasing of the number of users. However, PID does not apply any explicit rate constraint, it rather relies on the PID aiming for ensuring the system's stability. In addition, we employed the WFQ to fairly allocate wireless resource to users so as to avoid the starvation of service to users, which substantially improves the QoE of the users in contrast to NFDSVT.

A further experiment was conducted to verify the adaptivity of the proposed solution to the time-varying video bitrate exhibited in the time domain. Here, the number of users was set to 100 and the playback interruption probability in the time domain was characterized, as shown in Fig. 7. This illustrates that our algorithm maintains lowest interruption probability. This performance improvement is an explicit benefit of the



(a) Sensitivity of time-averaged system utility to V

(b) Sensitivity of interruption probability to V

Fig. 8. Sensitivity of system utility and interruption ratio to the parameter V.

fact that our algorithm makes joint fair decisions according to the length of the virtual queues which can reflect the current mismatch between the video bitrate and the channel bitrate at the BS. Once the bitrate of video changes, the virtual queues change promptly to trigger a corresponding action concerning the video layer selection and wireless resource allocation. By contrast, the PID simply depends on controlling the error feedback of the receiver's buffer fullness to control the number of video layers and to allocate the wireless resources. This feedback is unable to accurately characterize the real system conditions. Consequently, the heuristic PID cannot proactively respond to the burstiness of the video bitrate variations, and hence suffers from a higher video playback interruption probability. As for NFDSVT, it has a higher playback interruption probability than DSVT at the beginning of the video playback, because no data is available in the users' playback buffer at this stage, and simply allocating all resource to a single user is likely to inflict playback interruption on the other users.

C. Sensitivity Analysis

In this subsection, we conducted an experiment to investigate the performance sensitivity of the proposed algorithm to the trade-off factor V in Eq. (23). The value of V was varied from 0.5 to 1.5 and the rest of the parameters were kept the same as those in Section V-B. The experimental results were plotted in Fig. 8. It can be observed from Fig. 8(a) that the time-averaged system utility increases upon increasing V. According to Lyapunov's optimization theory, the parameter V represents the weight of penalty and a large value of V is

beneficial for improving the system's utility by making the layer selection decision to minimize (23). However, the fact that the video bit rate also increases whilst the wireless resources are limited leads to a longer virtual queue $H_u(t)$. A longer virtual queue $H_u(t)$ implies a higher mismatch between the video bit rate and the available transmission rate, hence the users suffer from a higher playback interruption probability, as illustrated by Fig. 8(b).

Fig. 8(b) further demonstrates that for a low system workload, the interruption ratio remains at a low level and it is less sensitive to any increase of V, while for a high system workload, the system's utility is susceptible to the increased value of V. This fact inspires us to choose a larger value of V for achieving a higher system utility for the scenario of low system workload, while at a high system workload we set V to a low value for achieving an acceptable interruption ratio. Hence, controlling V provides us with an efficient and flexible technique of striking a trade-off between the time-averaged system utility and the playback smoothness.

VI. Conclusions

In this paper, we investigated the video streaming problems of dense small-cell networks. We designed an SDN aided scalable video delivery system for dense small-cells with the objective of providing improved video services to users. We formulated a joint decision problem for maximizing the time-averaged system utility under a specific QoS/QoE constraint. By applying Lyapunov's optimization technique, we decomposed the problem into video layer selection and wireless resource allocation, which can be optimized independently by appropriately configuring the different system components. The solution derived for our video delivery system is an online measurement-driven approach operating without any prior statistical knowledge. By adopting the advanced technique of segment routing, the system can flexibly schedule its transmission path for each flow to achieve improved transmission efficiency. We evaluated the performance of the proposed approach using real video traces. The results verified that the proposed method is capable of providing a prompt response to dynamic environmental fluctuations, hence achieving a better performance than the benchmark algorithm.

Appendix A
Proof of Lemma 1

Squaring the virtual queue update Eq. (16) yields

$$H_u^2(t+1) \leqslant H_u^2(t) + r_u^2(t) + c_u^2(t) + 2H_u(t)[r_u(t) - c_u(t)].$$

Then, we have

$$\mathbb{E}\left[\frac{1}{2}H_u^2(t+1) - \frac{1}{2}H_u^2(t)|\mathbf{H}(t)\right] \leqslant \frac{1}{2}(r_{\max}^2 + c_{\max}^2) + \mathbb{E}[H_u(t)[r_u(t) - c_u(t)]|\mathbf{H}(t)]. \quad (39)$$

Summing (39) over $u \in U_t$, we obtain

$$\mathbb{E}\left[\frac{1}{2}\sum_{u\in U_t} H_u^2(t+1) - \frac{1}{2}\sum_{u\in U_t} H_u^2(t)|\mathbf{H}(t)\right] \leqslant \frac{1}{2}N_t(r_{\max}^2+c_{\max}^2)+\mathbb{E}\left[\sum_{u\in U_t} H_u(t)[r_u(t)-c_u(t)]|\mathbf{H}(t)\right]. \quad (40)$$

Summing (40) over $[t, t+T-1]$, we get the upper bound of the Lyapunov drift $\Delta_T(t)$ given by

$$\begin{aligned}\Delta_T(t) &= \mathbb{E}\left[\frac{1}{2}\sum_{u\in U_t} H_u^2(t+T) - \frac{1}{2}\sum_{u\in U_t} H_u^2(t)|\mathbf{H}(t)\right] \\ &\leqslant B_1 + \sum_{\tau=t}^{t+T-1}\sum_{u\in U_t}\mathbb{E}\left[H_u(\tau)r_u(\tau)|\mathbf{H}(t)\right] \\ &\quad - \sum_{\tau=t}^{t+T-1}\sum_{u\in U_t}\mathbb{E}\left[H_u(\tau)c_u(\tau)|\mathbf{H}(t)\right], \end{aligned} \quad (41)$$

where the positive constant B_1 is defined as $B_1 \triangleq \frac{1}{2}TN_t(r_{\max}^2 + c_{\max}^2)$. By substituting (9) and (41) into the "drift-plus-penalty" Equation (19), we have

$$\begin{aligned}&\Delta_T(t) - V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1} Q(\tau)|\mathbf{H}(t)\right] \\ &\leqslant B_1 - \sum_{\tau=t}^{t+T-1}\sum_{u\in U_t}\mathbb{E}\left[H_u(\tau)c_u(\tau)|\mathbf{H}(t)\right] \\ &\quad + \sum_{\tau=t}^{t+T-1}\sum_{u\in U_t}\mathbb{E}\left[H_u(\tau)r_u(\tau) - \frac{V}{N_t}q_u(\tau)|\mathbf{H}(t)\right]. \end{aligned} \quad (42)$$

On the one hand, $r_u(\tau)$ is the playback bit rate of user u and $q_u(\tau)$ is the MOS of user u in the Inequality (42). It should be noted that both of them depend on the decision of video layer selection, which is taken every T time units. Thus, for all $\tau \in [t, t+T-1]$ we have $r_u(\tau) = r_u(t)$ and $q_u(\tau) = q_u(t)$. On the other hand, we approximate $H_u(\tau)$ by $H_u(t)$ in the right-hand side of the Inequality (42) to derive a loose upper bound. By exploiting the fact that for any $\tau \in [t, t+T-1]$

$$H_u(t) - (\tau - t)c_{\max} \leqslant H_u(\tau) \leqslant H_u(t) + (\tau - t)r_{\max},$$

we have

$$\begin{aligned}&\sum_{\tau=t}^{t+T-1}\sum_{u\in U_t} H_u(\tau)r_u(\tau) \\ &= \sum_{\tau=t}^{t+T-1}\sum_{u\in U_t} H_u(\tau)r_u(t) \\ &\leqslant \sum_{\tau=t}^{t+T-1}\sum_{u\in U_t}[H_u(t)r_u(t) + (\tau-t)r_{\max}^2] \\ &= T\sum_{u\in U_t} H_u(t)r_u(t) + \frac{1}{2}T(T-1)N_t r_{\max}^2 \end{aligned} \quad (43)$$

and

$$
\begin{aligned}
&-\sum_{\tau=t}^{t+T-1}\sum_{u\in U_t} H_u(\tau)c_u(\tau)\\
&\leqslant \sum_{\tau=t}^{t+T-1}\sum_{u\in U_t}[-H_u(t)c_u(t)+(\tau-t)c_{\max}^2]\\
&=-\sum_{\tau=t}^{t+T-1}\sum_{u\in U_t} H_u(t)c_u(t)+\frac{1}{2}T(T-1)N_tc_{\max}^2. \quad (44)
\end{aligned}
$$

Substituting (43) and (44) into (42) we have

$$
\begin{aligned}
&\Delta_T(t)-V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1}Q(\tau)|\mathbf{H}(t)\right]\\
&\quad\leqslant B+T\sum_{u\in U_t}\mathbb{E}\left[H_u(t)r_u(t)-\frac{V}{N_t}q_u(t)|\mathbf{H}(t)\right]\\
&\qquad-\sum_{\tau=t}^{t+T-1}\mathbb{E}\left[\sum_{u\in U_t}H_u(t)c_u(\tau)|\mathbf{H}(t)\right], \quad (45)
\end{aligned}
$$

where the positive constant B is defined as $B\triangleq B_1+\frac{1}{2}T(T-1)N_tr_{\max}^2+\frac{1}{2}T(T-1)N_tc_{\max}^2=\frac{1}{2}T^2N_t(r_{\max}^2+c_{\max}^2)$.

APPENDIX B
PROOF OF LEMMA 2

According to [12, Th. 4.5], there exists a stationary optimal ω-only policy that achieves the optimal system utility $\overline{U^*}$ while satisfying the constraint (12)-(15). Furthermore, the optimal ω-only policy satisfies the inequality (20). Thus, we have

$$
\begin{aligned}
&\Delta_T(t)-V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1}Q^*(\tau)|\mathbf{H}(t)\right]\\
&\quad\leqslant B+\sum_{u\in U_t}\mathbb{E}\left[\sum_{\tau=t}^{t+T-1}H_u(t)[r_u^*(\tau)-c_u^*(\tau)]|\mathbf{H}(t)\right]\\
&\qquad-V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1}Q^*(\tau)|\mathbf{H}(t)\right], \quad (46)
\end{aligned}
$$

where $r_u^*(t)$ and $c_u^*(t)$ represent the playback rate and transmission rate of user u under the optimal ω-only policy, respectively. The second term at the right-side of (46) is non-positive because of the constraint (12). Since $\overline{Q_{subopt}}$ is a sub-optimal system utility, thus $\overline{Q_{subopt}}\leqslant\overline{Q^*}$ holds. Then, we have

$$
\Delta_T(t)-V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1}Q_{subopt}(\tau)\right]\leqslant B-V\mathbb{E}\left[\sum_{\tau=t}^{t+T-1}Q^*\right].
$$

Taking the expectations of the above inequality and summing it over $t=kT, k=0,1,\ldots,$ yields:

$$
\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[\sum_{u\in U_t}H_u^2(KT)\right]-\frac{1}{2}\mathbb{E}\left[\sum_{u\in U_t}H_u^2(0)\right]\\
&\qquad-V\mathbb{E}\left[\sum_{\tau=0}^{KT-1}Q_{subopt}(\tau)\right]\\
&\quad\leqslant KB-V\mathbb{E}\left[\sum_{\tau=0}^{KT-1}Q^*(\tau)\right].
\end{aligned}
$$

By dividing both sides with VKT, setting $K\to\infty$ and rearranging the terms, we have

$$
\begin{aligned}
&\lim_{K\to\infty}\frac{1}{KT}\mathbb{E}\left[\sum_{\tau=0}^{KT-1}Q_{subopt}(\tau)\right]\\
&\qquad\geqslant\lim_{K\to\infty}\frac{1}{KT}\mathbb{E}\left[\sum_{\tau=0}^{KT-1}Q^*(\tau)\right]-\frac{B}{VT},
\end{aligned}
$$

hence, $\overline{U_{subopt}}\geqslant\overline{U^*}-\frac{B}{VT}$.

REFERENCES

[1] L. Hanzo, H. Haas, S. Imre, D. O'Brien, M. Rupp, and L. Gyongyosi, "Wireless myths, realities, and futures: From 3G/4G to optical and quantum wireless," *Proc. IEEE*, (Special Centennial Issue), vol. 100, pp. 1853–1888, May 2012.

[2] *Global Mobile Data Traffic Forecast Update 2016–2021 White Paper*. Accessed: Apr. 15, 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[3] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *IEEE Veh. Technol. Mag.*, vol. 6, no. 1, pp. 37–43, Mar. 2011.

[4] N. Bhushan *et al.*, "Network densification: The dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[5] M. Kamel, W. Hamouda, and A. Youssef, "Performance analysis of multiple association in ultra-dense networks," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3818–3831, Sep. 2017.

[6] Y. Huo, C. Hellge, T. Wiegand, and L. Hanzo, "A tutorial and review on inter-layer FEC coded layered video streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 1166–1207, 2nd Quart., 2015.

[7] L. Sun, H. Shan, A. Huang, L. Cai, and H. He, "Channel allocation for adaptive video streaming in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 734–747, Jan. 2017.

[8] J. Yang, W. Cai, Y. Ran, H. Xi, and L. Hanzo, "Online measurement-based adaptive scalable video transmission in energy harvesting aided wireless systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6231–6245, Jul. 2017.

[9] X. Zhang, M. Yang, Y. Zhao, J. Zhang, and J. Ge, "An SDN-based video multicast orchestration scheme for 5G ultra-dense networks," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 77–83, Dec. 2017.

[10] S. Shenker, "The future of networking and the past of protocols (keynote presentation)," Open Netw. Summit, Stanford, CA, USA, Oct. 2011.

[11] C. Filsfils, N. K. Nainar, C. Pignataro, J. C. Cardona, and P. Francois, "The segment routing architecture," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.

[12] M. J. Neely, *Stochastic Network Optimization With Application to Communication and Queueing Systems*. San Rafael, CA, USA: Morgan & Claypool, 2010.

[13] I. Hwang, B. Song, and S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 20–27, Jun. 2013.

[14] J. Kim and D.-H. Cho, "A joint power and subchannel allocation scheme maximizing system capacity in indoor dense mobile communication systems," *IEEE Trans. Veh. Technol.*, vol. 59, no. 9, pp. 4340–4353, Nov. 2010.

[15] Z. Luan, H. Qu, J. Zhao, B. Chen, and J. C. Principe, "Correntropy induced joint power and admission control algorithm for dense small cell network," *IET Commun.*, vol. 10, no. 16, pp. 2154–2161, Nov. 2016.

[16] J. Liu, M. Sheng, L. Liu, and J. Li, "Interference management in ultra-dense networks: Challenges and approaches," *IEEE Netw.*, vol. 31, no. 6, pp. 70–77, Nov. 2017.

[17] C. Yang, J. Li, Q. Ni, A. Anpalagan, and M. Guizani, "Interference-aware energy efficiency maximization in 5G ultra-dense networks," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 728–739, Feb. 2017.

[18] B. Li, D. Zhu, and P. Liang, "Small cell in-band wireless backhaul in massive MIMO systems: A cooperation of next-generation techniques," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 7057–7069, Dec. 2015.

[19] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung, and H. V. Poor, "Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1936–1947, Sep. 2017.

[20] S. Wu, Z. Zeng, and H. Xia, "Load-aware energy efficiency optimization in dense small cell networks," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 366–369, Feb. 2017.
[21] C. Li, J. Zhang, and K. B. Letaief, "Throughput and energy efficiency analysis of small cell networks with multi-antenna base stations," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2505–2517, May 2014.
[22] A. R. Elsherif, W.-P. Chen, A. Ito, and Z. Ding, "Adaptive resource allocation for interference management in small cell networks," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2107–2125, Jun. 2015.
[23] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A scheduling framework for adaptive video delivery over cellular networks," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw.*, 2013, pp. 389–400.
[24] H. Abou-zeid, H. S. Hassanein, and S. Valentin, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2013–2026, Jun. 2014.
[25] S. Cicalò, and V. Tralli, "Distortion-fair cross-layer resource allocation for scalable video transmission in OFDMA wireless networks," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 848–863, Apr. 2014.
[26] C. Chen, X. Zhu, G. de Veciana, A. C. Bovik, and R. W. Heath, Jr., "Rate adaptation and admission control for video transmission with subjective quality constraints," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 22–36, Feb. 2015.
[27] I. Djama, T. Ahmed, A. Nafaa, and R. Boutaba, "Meet in the middle cross-layer adaptation for audiovisual content delivery," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 105–120, Jan. 2008.
[28] M. Li, Z. Chen, and Y.-P. Tan, "Scalable resource allocation for SVC video streaming over multiuser MIMO-OFDM networks," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1519–1531, Nov. 2013.
[29] H. Hu, X. Zhu, Y. Wang, R. Pan, J. Zhu, and F. Bonomi, "Proxy-based multi-stream scalable video adaptation over wireless networks using subjective quality and rate models," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1638–1652, Nov. 2013.
[30] D. Bethanabhotla, G. Caire, and M. J. Neely, "Adaptive video streaming for wireless networks with multiple users and helpers," *IEEE Trans. Commun.*, vol. 63, no. 1, pp. 268–285, Jan. 2015.
[31] K. Miller, D. Bethanabhotla, G. Caire, and A. Wolisz, "A control-theoretic approach to adaptive video streaming in dense wireless networks," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1309–1322, Aug. 2015.
[32] A. Giorgetti, A. Sgambelluri, F. Paolucci, F. Cugini, and P. Castoldi, "Segment routing for effective recovery and multi-domain traffic engineering," *J. Opt. Commun. Netw.*, vol. 9, no. 2, pp. 223–232, Feb. 2017.
[33] A. Cianfrani, V. Eramo, M. Listanti, M. Polverini, and A. V. Vasilakos, "An OSPF-integrated routing strategy for QoS-aware energy saving in IP backbone networks," *IEEE Trans. Netw. Service Manag.*, vol. 9, no. 3, pp. 254–267, Sep. 2012.
[34] D. N. C. Tse *et al.*, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[35] A.-N. Moldovan, I. Ghergulescu, and C. H. Muntean, "VQAMap: A novel mechanism for mapping objective video quality metrics to subjective MOS scale," *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 610–627, Sep. 2016.
[36] Y. Yao, L. Huang, A. B. Sharma, L. Golubchik, and M. J. Neely, "Power cost reduction in distributed data centers: A two-time-scale approach for delay tolerant workloads," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 200–211, Jan. 2014.
[37] A. Banchs and X. Perez, "Distributed weighted fair queuing in 802.11 wireless LAN," in *Proc. IEEE ICC*, vol. 5, Apr. 2002, pp. 3121–3127.
[38] A. Eryilmaz and R. Srikant, "Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1333–1344, Dec. 2007.
[39] M.-F. Homg, W.-T. Lee, K.-R. Lee, and Y.-H. Kuo, "An adaptive approach to weighted fair queue with QoS enhanced on IP network," in *Proc. IEEE TENCN*, vol. 1, Aug. 2001, pp. 181–186.
[40] J. Meinilä, P. Kyösti, T. Jämsä, and L. Hentilä, "WINNER II channel models," in *Radio Technologies and Concepts for IMT-Advanced*. Chichester, U.K.: Wiley, 2009.
[41] *Video Trace Library*. Accessed: May 2013. [Online]. Available: http://trace.eas.asu.edu
[42] P. Seeling and M. Reisslein, "Video transport evaluation with H.264 video traces," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 1142–1165, 4th Quart., 2012.
[43] R. Gupta, A. Pulipaka, P. Seeling, L. J. Karam, and M. Reisslein, "H.264 coarse grain scalable (CGS) and medium grain scalable (MGS) encoded video: A trace based traffic and quality evaluation," *IEEE Trans. Broadcast.*, vol. 58, no. 3, pp. 428–439, Sep. 2012.

Jian Yang received the B.S. and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2001 and 2006, respectively. From 2006 to 2008, he was a Post-Doctoral Scholar with the Department of Electronic Engineering and Information Science, USTC. Since 2008, he has been an Associate Professor with the Department of Automation, USTC.

He is currently a Professor with the School of Information Science and Technology, USTC. His research interests include future network, distributed system design, modeling and optimization, multimedia over wired and wireless networks, and stochastic optimization. He received the Lu Jia-Xi Young Talent Award from the Chinese Academy of Sciences in 2009.

Bowen Yang received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2014, where he is currently pursuing the Ph.D. degree with the School of Information Science and Technology.

His research interests include multimedia streaming, wireless networks, software-defined networks, and stochastic optimization.

Shuangwu Chen received the B.S. and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2011 and 2016, respectively. He is currently a Post-Doctoral Researcher at USTC.

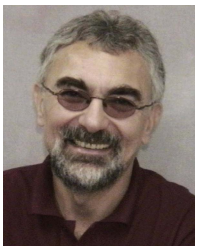
His research interests include multimedia over wired and wireless networks, future network, and stochastic optimization.

Yongdong Zhang (M'08–SM'13) received the Ph.D. degree in electronic engineering from Tianjin University, Tianjin, China, in 2002. He is currently a Professor with the School of Information Science and Technology, University of Science and Technology of China. His current research interests are in the fields of multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology.

He has authored over 100 refereed journal and conference papers. He was a recipient of the Best Paper Awards in PCM 2013, ICIMCS 2013, and ICME 2010 and the Best Paper Candidate in ICME 2011. He serves as an Editorial Board Member of the *Multimedia Systems Journal* and the IEEE TRANSACTIONS ON MULTIMEDIA.

Yanyong Zhang (F'17) received the B.S. degree from the University of Science and Technology of China (USTC) in 1997 and the Ph.D. degree from Penn State University in 2002. From 2002 to 2018, she was on the faculty of the Electrical and Computer Engineering Department, Rutgers University. She was also a member of the Wireless Information Networks Laboratory. Since 2018, she has been with the School of Computer Science and Technology, USTC.

She has 21 years of research experience in the areas of sensor networks, ubiquitous computing, and high-performance computing and has published more than 110 technical papers in these fields. She received the NSF CAREER Award in 2006. She currently serves as an Associate Editor for several journals, including the IEEE/ACM TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON SERVICE COMPUTING, and *Smart Health* (Elsevier).



Lajos Hanzo (F'04) received the five-year degree in electronics and the Ph.D. degree from the Technical University of Budapest in 1976 and 1983, respectively. In 2016, he was admitted to the Hungarian Academy of Science. During his 40-year career in telecommunications, he has held various research and academic posts in Hungary, Germany, and U.K. Since 1986, he has been with the School of Electronics and Computer Science, University of Southampton, U.K., where he is currently the Chair in telecommunications. From 2008 to 2012, he was a Chaired Professor at Tsinghua University, Beijing. He has successfully supervised 112 Ph.D. students, co-authored 18 John Wiley/IEEE Press books on mobile radio communications totaling in excess of 10 000 pages, published 1790 research contributions at the IEEE Xplore, acted both as the TPC and General Chair of IEEE conferences, presented keynote lectures, and has been received a number of distinctions. He is currently directing an Academic Research Team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council, U.K., the European Research Council's Advanced Fellow Grant, and the Royal Society's Wolfson Research Merit Award. He is an Enthusiastic Supporter of Industrial and Academic Liaison, and he offers a range of industrial courses. He is a fellow of the Royal Academy of Engineering, IET, and EURASIP. In 2009, he received an Honorary Doctorate from the Technical University of Budapest and in 2015 from The University of Edinburgh. He is also a Governor of the IEEE ComSoc and VTS. From 2008 to 2012, he was the Editor-in-Chief of the IEEE Press. For further information on research in progress and associated publications, please refer to http://www-mobile.ecs.soton.ac.uk.