Proactive Caching for Mobile Video Streaming in Millimeter Wave 5G Networks

Jian Qiao, Yejun He, *Senior Member, IEEE*, and Xuemin (Sherman) Shen, *Fellow, IEEE*

***Abstract*— Mobile video streaming is fundamental to advanced applications in the fifth generation (5G) networks. Millimeter wave (mmWave) communication represents a leading 5G technology, which provides rich bandwidth and, therefore, great potentials for high-quality mobile video streaming. However, mobile video streaming in mmWave 5G networks faces fundamental challenges due to mmWave antenna directivity and high user mobility. As such, users typically have short connection durations and frequent handoffs, making video streaming suffer from long handoff delays and connection latency. In this paper, we tackle the issues by developing a caching-based mmWave framework, which precaches video contents at the base station for handoff users and thus significantly reduces the connection and retrieval delays. As a result, high-mobility users with frequent handoffs can enjoy continuous high-quality video streaming. Specifically, we model the proposed system as a cache management problem and attain optimal video streaming quality by using Markov decision process to dynamically allocate proper cache memory space of each base station to mobile users. A cell-by-cell decomposition method is proposed to solve the dynamic programming problem with significantly reduced computational complexity. Using extensive simulations, we demonstrate that the proposed solution can effectively maintain high-quality mobile video streaming for high-mobility 5G users moving among mmWave small cells with directional antenna.**

***Index Terms*— Mobile video streaming, proactive caching, 5G, millimeter wave, directional antenna.**

I. Introduction

MOBILE data traffic increases dramatically with the exponential growth in the number of mobile devices and the emerging high-rate multimedia applications (such as video streaming for mobile gaming and social networks) [1]–[3]. Future fifth generation (5G) networks need to be designed to accommodate the overwhelming mobile traffic demands [4]. Millimeter wave (mmWave) spectrum with huge available bandwidth is a promising technology for 5G networks to satisfy the high capacity demands and overcome the bandwidth shortage at saturated microwave spectrum [5], [6]. mmWave communication has severe propagation loss because of the high frequency band [7]. Directional antenna is adopted for both mobile users and mmWave base stations to combat the severe propagation loss and achieve high data rate. Connectivity maintaining by directing antenna beams towards each other is challenging, especially for high-mobility scenarios (such as highway and rail environments). mmWave base stations are expected to be deployed densely in small cell (e.g., picocells with range under 100 meters [8]) to provide high data rates and aggregate capacity [4]. Users with high mobility would suffer frequent handoffs due to smaller cell size.

Mobile video streaming is one of the main applications for 5G networks [9]. Enabling mobile video streaming in mmWave 5G networks has several challenges in high-mobility environments. First, due to directional antenna, the connectivity from mmWave base station to mobile user can be available to deliver video content only if both antennas direct towards each other. Each mobile user has very short connection time to the base station considering high user mobility, antenna directivity, smaller cell size, and the fact that a larger number of users share the mmWave channel in time division. Second, with smaller cell size, users with high-mobility have frequent handoffs and need to frequently re-build the route to remote video server through different base stations connected to the core network, which involves heavy communication overheads and long latency. The delay for re-building the route to deliver video content has significant impact on the video streaming quality, such as video frozenness.

In this paper, to provide satisfactory quality for mobile video streaming requiring high data rate (e.g., uncompressed video streaming requires mandatory data rate of 1.78/3.56 Gbps in 5G networks), a proactive caching system is enabled by pre-loading each user's video content from the remote video server and storing the video content in the cache memory of each base station. A mobile user entering a new cell can immediately have the available video content, which mitigates the delay on frequently re-building the route to remote video server. Larger size of pre-loaded video content for a specific user in the cache memory can provide better video streaming quality while including more video delivery cost and occupying more cache memory space which can be used by other users. Smaller size of pre-loaded video content might not be sufficient for the

Manuscript received January 18, 2016; revised June 13, 2016; accepted August 1, 2016. Date of publication August 10, 2016; date of current version October 7, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61372077, in part by the Natural Science and Engineering Research Council of Canada, and in part by the Science and Technology Innovation Commission of Shenzhen under Grant ZDSYS 201507031550105. The associate editor coordinating the review of this paper and approving it for publication was B. Hamdaoui. *(Corresponding author: Yejun He.)*

J. Qiao and Y. He are with the Shenzhen Key Laboratory of Antennas and Propagation, College of Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: heyejun@126.com). J. Qiao was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: jqiao@bbcr.uwaterloo.ca).

X. Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: xshen@bbcr.uwaterloo.ca).

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/TWC.2016.2598748

user to maintain video streaming quality until the next video content is received. Therefore, how to optimize the size of pre-loaded video content in each user's subsequent base station is an important issue to improve the quality of mobile video streaming.

The main contributions of this paper are three-fold. First, a proactive caching system is proposed to allocate proper amount of cache memory space in each base station to its associated users to store the pre-loaded video content from remote server to eliminate the delay effect on mobile video streaming quality. Second, we formulate the dynamic cache memory allocation problem in mmWave 5G networks as a Markov decision process to maximize the video streaming quality for a number of users moving on the highway covered by a series of mmWave base stations. Third, an approximated cell-by-cell decomposition method is proposed to provide a general and applicable solution to the formulated Markov decision process.

The remainder of the paper is organized as follows. Section II describes the related works. Mobile video streaming system architecture and directional communication model are presented in Section III. The problem is defined in Section IV. In Section V, the problem is formulated as a Markov decision process. A cell-by-cell decomposition method is proposed in Section VI. Section VII presents the numerical results. Finally, conclusions are given in Section VIII.

II. Related Works

Quality provisioning for video streaming has been intensively studied in wireless networks, subject to wireless bandwidth limitation and wireless channel fluctuation [10]–[21]. One category of research focuses on efficient bandwidth utilization to provide the required quality of service (QoS) of video streaming [10]–[13]. In [10], an economic model is used to allocate radio resource between cellular network and wireless local area network (WLAN), aiming to optimize the total welfare of the heterogeneous networks. A practical resource allocation scheme [11] is proposed in a distributed manner for mobile users with multi-homing capability, to support video streaming requiring constant bit rate. In [12], a joint adaptation, resource allocation, and scheduling algorithm is proposed to allocate network resources and make transmission schedule, based on the required QoS of video streaming for each user. The proposed algorithm can achieve near full utilization of network resources while satisfying the delay requirement of all video frames. A delay-aware resource allocation (DARA) approach in [13] can compute optimal time slot allocation policy by maximizing the deadline-abiding delivery of all senders. The proposed DARA approach yields a non-stationary slot allocation policy depending on the allocation of previous slots.

The second category of research mitigates the wireless channel fluctuation effect on video streaming quality by buffering a certain amount of video data at end user [14]–[17]. In [14], the impacts of wireless channel dynamics on video streaming quality is analytically investigated by modeling the receiver buffer with G/G/1/M queue and G/G/1/N queue. In [15], mobile video applications (e.g., Youtube) periodically download video data aggressively into end user's buffer without considering buffer status and network resource availability, in order to reduce video frozenness. An intelligent cost-aware buffer management strategy for mobile video streaming applications is proposed in [16] to minimize the cost resulting from un-consumed video data while respecting certain quality of experience (QoE) requirements. Literature [17] jointly considers the bandwidth allocation and buffer management at the mobile user to dynamically charge/discharge the buffer to optimize video streaming quality.

Proactive caching is another category of approaches to achieve quality provisioning on mobile video streaming, by fetching and storing video content in the cache memory of base stations proactively [18]–[21]. Proactive caching can effectively improve network performance in terms of peak hour capacity and content delivery delay [18], [20]. In [19], a proactive paradigm for 5G is proposed to track and build users' demand profiles which are used to predict future transmission requests for proactive caching at the base stations. In [21], a proactive seeding technique is proposed to minimize the network peak load by proactively pushing content to selected users before they actually request it in online social networks.

Most of the existing works on proactive caching concentrate on how to predict the most popular contents of the users based on the traffic history, without considering the efficiency of proactive caching and how to allocate the cache memory spaces for different users to optimize mobile video streaming quality. In this paper, proactive caching is adopted to distribute video content from remote video server to the candidate base station ahead of the user association for high-mobility environments, so that the video content required by a mobile user is immediately available when it moves into a new cell and the handover delay can be significantly reduced. Compared with existing works, we consider the characteristics of video streaming traffic and obtain the exact amount of video content stored in the candidate base station for each user under the capacity constraint of local memory, to achieve optimal video streaming quality.

III. System Model

This section describes the system architecture for mobile video streaming in mmWave 5G networks and the directional communication model for high-mobility environments.

A. Video Streaming Architecture for Highway Scenario

In this paper, the hybrid architecture of 4G+mmWave for 5G wireless networks is adopted to provide ubiquitous coverage and high data rate in most coverage areas. Mobile video streaming with high-rate requirements can be supported with mmWave networks, which are deployed with cellular topology in small cell as shown in Fig. 1. We consider a set of mobile users $\mathcal{U} = \{U_1, U_2, \dots, U_n, \dots U_{\mathcal{N}}\}$ moving on a segment of highway which is covered by a set of mmWave base stations $\mathcal{B} = \{B_1, B_2, \dots, B_k, \dots B_{\mathcal{K}}\}$. Both mmWave base stations and mobile users are equipped with electronically steerable directional antennas. The connection between the

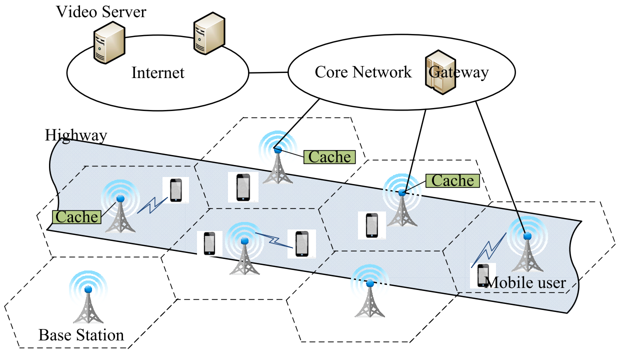

Fig. 1. Video Streaming Architecture with Cache Memory.

mobile user and the base station can be built when both the transmission and reception antennas direct their beams towards each other with beamforming technologies [22].

Traditionally, if user U_n enters the cell covered by base station B_k, user U_n sends the transmission request to B_k which is responsible for forwarding the request to the remote video server, downloading the video content to B_k, and transmitting the video content to user U_n through mmWave channel. Since the user's movement can be tracked when it moves on the highway, the next associated base station for user U_n can be predicted. With proactive caching approach, when user U_n is currently associated with base station B_k, the system can predetermine the amount of video content to be cached in the local memory of base station B_{k+1} in order to reduce the latency and improve video streaming quality.

B. Directional Communication With Mobility

Directional antenna radiates power to all the directions while having focus on specific directions. The radiation model of directional antenna can be described as

$$g(\alpha) = \frac{G(\alpha)}{G_{max}} \tag{1}$$

where $G_{max} = \max_{\alpha} G(\alpha)$. α is the horizontal angle towards different directions. In order to combat the high propagation loss over distance, directional antenna with high directivity gain is used. The flat-top antenna model [6] is adopted for simplification purpose. Specifically,

$$g(\alpha) = \begin{cases} 1, & \mid \alpha \mid \leq \dfrac{\Delta\alpha}{2} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\Delta\alpha$ is the beamwidth with $\Delta\alpha = \frac{2\pi}{N}$ (N is the number of beams for each antenna).

With the flat-top antenna model, mobile users and base station can communicate with each other if and only if both of them are within each other's beamwidth, i.e., each of them directs its beam towards each other. Fig. 2 shows the directional communication between mobile user U_n and the associated base station B_k. When mobile user U_n enters the cell shown in Fig. 2(a), it can not receive video content from B_k until it moves into the beamwidth of B_k as shown in Fig. 2(b). The connection is discontinued after U_n moves out of the beamwidth of B_k as shown in Fig. 2(c). It can be seen that the communication time between mobile user and its associated base station is quite short with directional communication, high user mobility, and smaller cell size. The video content delivery delay plays a significant role on video streaming quality of the mobile user.

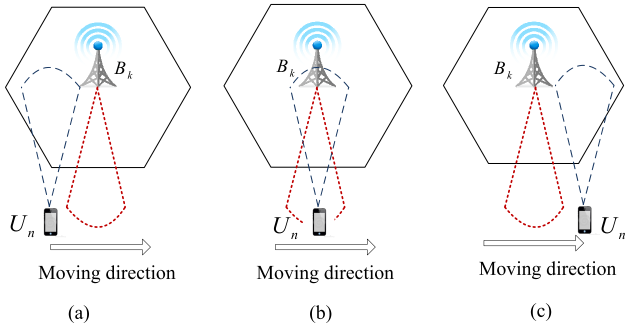

Fig. 2. Directional mmWave Connection with High Mobility.

Remark: Since the highway is usually constructed in rural area, most of the traffic demands in each cell come from the highway area rather than other areas within the cell. It is assumed that the directional antenna of each base station is always directed to the highway to provide connections for mobile users. In addition, it is possible that there are multiple mobile users connected to the same base station for video streaming. The medium access control (MAC) schemes (such as TDMA) for these mobile users with directional communication are beyond the scope of this paper.

IV. Problem Definition

A set of adjacent mobile users $\mathcal{U} = \{U_1, U_2, \ldots, U_n, \ldots U_{\mathcal{N}}\}$ requiring video streaming services move towards a certain direction on the highway covered by a set of adjacent mmWave base stations $\mathcal{B} = \{B_1, B_2, \ldots, B_k, \ldots B_{\mathcal{K}}\}$ corresponding to a set of cells $\mathcal{C} = \{C_1, C_2, \ldots, C_k, \ldots C_{\mathcal{K}}\}$. Each mmWave base station B_k covering a cell C_k has a local cache memory with size M_k ($k = 1, 2, \ldots, \mathcal{K}$). We consider the number of mmWave base stations $\mathcal{K}$ is large enough, such that the mobile users do not exceed the area on the highway covered by the base station set $\mathcal{B}$ within the considered time duration. mmWave base stations are connected to the gateway through optical fiber with high capacity. It is assumed that the connection between mmWave base station and gateway is large enough to deliver the pre-loaded video content. If the required amount of video content can not be delivered to the base station due to limited capacity of optical fiber, the actual amount of pre-loaded video content should be bounded by the integration of optical fiber capacity over video content delivery duration. Mobile users switch connections with different mmWave base stations through hard handoffs. The video content is segmented into blocks, and the size of local memory M_k is in the unit of block. Different users' video blocks pre-loaded in the local memory of each base station are stored based on the order in which the users become associated with the base station. Video blocks of the same user associated with a specific base station are stored in local memory from lower address to higher address depending on the availability

TABLE I
NOTATIONS FOR MODELING

U_n	$\triangleq$	Mobile user n
C_k	$\triangleq$	mmWave cell k
B_k	$\triangleq$	mmWave base station k
M_k	$\triangleq$	Cache memory size in B_k
$\mathcal{M}_{n,k}$	$\triangleq$	Cache memory size in B_k allocated to U_n
$\mathcal{U}$	$\triangleq$	Set of mobile users
$\mathcal{B}$	$\triangleq$	Set of mmWave base stations
$g(\alpha)$	$\triangleq$	Normalized antenna radiation pattern over angle α
$\mathbb{M}$	$\triangleq$	Matrix describing memory allocation state for $\mathcal{K}$ base stations and $\mathcal{N}$ users
$\mathbb{R}$	$\triangleq$	Matrix describing a user changes its associated base station
$\Psi_{n,k}$	$\triangleq$	Time duration for user U_n staying in the cell C_k
$\psi_{n,k}$	$\triangleq$	Arrival rate of the event that the user U_n enters a new cell from cell C_k
$\mathcal{P}$	$\triangleq$	State space of the system
$\mathcal{M}$	$\triangleq$	Set of all local memory allocation solutions
$\zeta^{(i)}$	$\triangleq$	Set of possible actions for state i
$\mathcal{C}$	$\triangleq$	Control space
$a(i)$	$\triangleq$	Action for state i
$p_{i,j}(a(i))$	$\triangleq$	State transition probability from state i to state j with action $a(i)$
Υ_i	$\triangleq$	Overall state transition rates from state i
μ_n	$\triangleq$	Required data rate for video streaming of user U_n
Ω_n	$\triangleq$	Buffer size of user U_n
t_χ	$\triangleq$	Starting time of the χ^{th} stage in the Markov decision process
$\mathbb{Y}(t)$	$\triangleq$	State of the system at any instant t
τ_χ	$\triangleq$	Stage duration
$W_n^{(\chi)}$	$\triangleq$	Cost for user U_n in the χ^{th} stage
$\Omega_n(t_\chi)$	$\triangleq$	U_n's buffer status at the beginning of the χ^{th} stage
$W_{n,loss}^{(\chi)}$	$\triangleq$	Stage cost in terms of video quality loss duration for the user U_n in the χ^{th} stage
$W_{n,froz}^{(\chi)}$	$\triangleq$	Stage cost in terms of video frozenness duration for the user U_n in the χ^{th} stage
$W_n^{(\chi)}$	$\triangleq$	Total stage cost for the user U_n
$\widetilde{\mathcal{W}}$	$\triangleq$	The average cost per unit time per user
W_{cons}	$\triangleq$	Average video streaming quality consistency per user in the system
$d(i)$	$\triangleq$	Differential cost for state i
ϕ	$\triangleq$	Policy of a state in Bellman's Equation
$\gamma_n^{(\chi-1)}$	$\triangleq$	Coefficient obtained based on the memory space allocation records in the previous $(\chi-1)$ stages for the user U_n

of the memory. When user U_n is currently associated with base station B_k, the next associated base station B_{k+1} can be predicted based on its moving direction and the coverage area of each base station. Before a user U_n moving into a new cell C_{k+1}, the system can determine a specific amount of cached video content and can pre-load the video content from remote video sever to the local memory of base station B_{k+1}. Thus, it is assumed that each base station knows the video content of each user and the video content of each user cached in different base stations is continuous without content overlap. Since the video content is pre-loaded from remote video sever to mmWave base station through cable with perfect channel quality, lost blocks are not considered in this paper. The parameters for system model are shown in Table I.

This paper investigates the impact of proactive caching on video streaming quality in 5G networks and addresses the problem of how to allocate the proper cache size to pre-load the video content in the next associated base station for each mobile user in order to achieve global optimal video streaming quality, under the space size constraint of each local memory. The video content proactive caching problem in this paper is a decision making problem, i.e., based on the current state of the system (each mobile user U_n with corresponding associated base station B_k and its allocated memory size $\mathcal{M}_{n,k}$ in that base station), the system determines the newly allocated memory size $\mathcal{M}_{n,k+1}$ for the user moving into a cell covered by the next base station B_{k+1}. The overall mobile video streaming quality for all the users in the system can be significantly affected by the amount of pre-loaded video content, which is the action of each state. Note that the next state of the system (which describes the local memory allocation for each user when there is a user entering a new cell) only depends on the current state and the allocated cache memory space to the new user, which satisfies Markov property.

Markov Decision Process is a stochastic control process for decision making. At each time step, the process is in a specific state, and the decision maker chooses an action available in that state. Then, the process responds at the next time step by moving into a new state with a reward corresponding to the selected decision. The problem addressed in this paper exactly fits the Markov decision process, and we formulate the proactive caching problem as a Markov decision process, composed of *the state space*, *the control space*, *the state transition probabilities* and *the cost function*.

V. FORMULATE THE PROBLEM AS MARKOV DECISION PROCESS

As discussed in Sec. IV, to formulate the problem of proactive caching for mobile video streaming as a discrete-state finite horizon Markov decision process, we first define the state as a combination of a cell-based memory allocation configuration and a movement event. Then, we provide the control space, the state transition probability, and the cost function.

A. The Space of the Combined State

At any particular instant, given the set of mobile users $\mathcal{U} = \{U_1, U_2, \ldots, U_n, \ldots U_{\mathcal{N}}\}$ with video streaming applications and the set of base stations $\mathcal{B} = \{B_1, B_2, \ldots, B_k, \ldots B_{\mathcal{K}}\}$ covering a segment of highway, the status of local memory allocation for all the users associated with the corresponding base stations can be described by a $\mathcal{N} \times \mathcal{K}$ matrix:

$$\mathbb{M} = \| \widetilde{\mathcal{M}}_{n,k} \|_{\mathcal{N}\times\mathcal{K}} = \left\| \begin{matrix} \widetilde{\mathcal{M}}_{1,1} & \widetilde{\mathcal{M}}_{1,2} & \cdots & \widetilde{\mathcal{M}}_{1,\mathcal{K}} \\ \widetilde{\mathcal{M}}_{2,1} & \widetilde{\mathcal{M}}_{2,2} & \cdots & \widetilde{\mathcal{M}}_{2,\mathcal{K}} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\mathcal{M}}_{\mathcal{N},1} & \widetilde{\mathcal{M}}_{\mathcal{N},2} & \cdots & \widetilde{\mathcal{M}}_{\mathcal{N},\mathcal{K}} \end{matrix} \right\|, \tag{3}$$

where $\widetilde{\mathcal{M}}_{n,k}$ is the allocated local memory size in base station B_k to user U_n. $\widetilde{\mathcal{M}}_{n,k}$ is given as

$$\widetilde{\mathcal{M}}_{n,k} = \begin{cases} \mathcal{M}_{n,k}, & (U_n \text{ associated with } B_k) \\ 0, & (\text{otherwise}) \end{cases} \tag{4}$$

where $\mathcal{M}_{n,k}$ is the allocated local memory size of mobile user U_n when it is associated with B_k. $\widetilde{\mathcal{M}}_{n,k} = 0$ can occur in two cases. First, $\widetilde{\mathcal{M}}_{n,k}$ can be zero if the user U_n is not associated with B_k in the current state. Second, even if the user U_n is currently associated with B_k, the local memory in B_k is fully occupied by other users and there is no available memory space to be allocated to the user U_n. Since each base station B_k has a local memory with space M_k, we have

$$\sum_{n=1}^{\mathcal{N}} \widetilde{\mathcal{M}}_{n,k} \leq M_k, \text{ for } k = 1, 2, \ldots, \mathcal{K} \tag{5}$$

The time duration $\Psi_{n,k}$ for the user U_n staying in the cell C_k until it moves into the next cell C_{k+1} is independent of the previous visited cells and the previous length of video streaming duration. The random variable $\Psi_{n,k}$ can be affected by the length of the highway located in the cell C_k and the speed of the mobile user. It is assumed that $\Psi_{n,k}$ is exponentially distributed with rate $\psi_{n,k}$ which can be statistically obtained through measurements or other means. Therefore, the arrival of the event that the user U_n enters a new cell is a Poisson process with the rate of $\psi_{n,k}$. The process of all the users' movement within the considered highway segment is a combination of several independent Poisson processes. A random event of the system that a user changes its associated base station can be described by the following $\mathcal{N} \times \mathcal{K}$ matrix:

$$\mathbb{R} = \| r_{n,k} \|_{\mathcal{N}\times\mathcal{K}} = \begin{Vmatrix} r_{1,1} & \cdots & r_{1,\mathcal{K}} \\ \vdots & r_{n,k} & \vdots \\ r_{\mathcal{N},1} & \cdots & r_{\mathcal{N},\mathcal{K}} \end{Vmatrix} \tag{6}$$

where $r_{n,k}$ indicates a specific user moving from the current cell into the next cell and can be given as

$$r_{n,k} = \begin{cases} -1, & (\text{User } n \text{ intends to leave cell } C_k) \\ 1, & (\text{User } n \text{ intends to enter } C_k) \\ 0, & (\text{otherwise}). \end{cases} \tag{7}$$

The matrix $\mathbb{R}$ has two non-zero adjacent elements since each $\mathbb{R}$ describes the movement of one user.

The state of the system at any particular instant can be described by the combination of the two matrices:

- The matrix $\mathbb{M}$ describes the local memory allocation to $\mathcal{N}$ mobile users moving towards a certain direction among $\mathcal{K}$ cells;
- The matrix $\mathbb{R}$ describes the next random event that a particular mobile user U_n changes its associated base station.

Let $\mathcal{P}$ denote the state space of the system. The space $\mathcal{P}$ is composed of all possible combinations of the matrix $\mathbb{M}$ and $\mathbb{R}$, given that the set of users $\mathcal{U}$ move towards a certain direction and each user can only move from the current cell C_k to the next cell C_{k+1} with $k \in \{1, 2, \ldots, \mathcal{K} - 1\}$. $\mathcal{P}$ can be given as

$$\mathcal{P} = \{i = (\mathbb{M}^{(i)}, \mathbb{R}^{(i)}) \mid \mathbb{M}^{(i)} \in \mathcal{M}, \mathbb{R}^{(i)} \in \mathcal{R}\} \tag{8}$$

where $\mathcal{M}$ is the set of all local memory allocation solutions, and $\mathcal{R}$ is the set of random events indicating which user is the next moving user and the corresponding moving-out and moving-into cells. Note that the set of $\mathcal{R}$ depends on the primary state component $\mathbb{M}^{(i)}$.

Based on the definition of the system state $i = (\mathbb{M}^{(i)}, \mathbb{R}^{(i)})$, there are several observations: (1) The system enters a new state $j = (\mathbb{M}^{(j)}, \mathbb{R}^{(j)})$ as soon as a particular user U_n moves into a new cell determined by the matrix $\mathbb{R}^{(i)}$; (2) The system allocates a specific local memory space in the next associated base station to the user $U_{n'}$ in order to pre-load the content for video streaming if the user $U_{n'}$ in state $j = (\mathbb{M}^{(j)}, \mathbb{R}^{(j)})$ initiates the coming random event of user movement; (3) Both $\mathbb{M}$ and $\mathbb{R}$ are sparse matrixes. The non-zero elements of the matrix $\mathbb{M}$ move from the left side towards the right side of $\mathbb{M}$ as the system undergoing different stages.

Initially, the set of users are located in a number of cells at the beginning of the considered highway segment. At the end of the considered duration, the set of mobile users would be located in a number of cells at the end of the considered highway segment. If it is assumed that the mobile users have similar speed on the highway, the system state transition stops and the system enters the final stage when a user moves into the last cell $C_{\mathcal{K}}$.

Remark: We consider the local memory allocation problem for proactive caching among a fixed number of active users (users with video streaming) in the system. It is possible that there are user arrivals and user departures in the considered system, which leads to a dynamic number of active users. The whole process with a dynamic number of active users can be viewed as a series of sub-processes, each of which has a fixed number of active users contending for the local memory spaces. The admission control problem for new user arrivals and user departures in the system is beyond the scope of this paper. Although we consider single-direction scenario on the highway, the model in this paper can also be used for two-direction scenario by considering users from the other direction without changes on problem formulation. The matrix $\mathbb{M}$ would have more columns as more users are involved. For single-direction scenario, users always move from cell C_k to cell C_{k+1} while users can move from cell C_{k+1} to C_k and from cell C_k to C_{k+1} for two-direction scenario which affects matrix $\mathbb{R}$ indicating user's movement.

B. The Control Space

The local memory allocation problem for proactive caching is a decision-making problem. Specifically, when the system is under a state $i = (\mathbb{M}^{(i)}, \mathbb{R}^{(i)})$, the system knows that the user U_n will enter a new cell C_{k+1} by reading matrix $\mathbb{R}^{(i)}$. Then the system selects an action indicating the allocated local memory for the user U_n in the next associated base station B_{k+1} covering the new cell C_{k+1}. The action taken in each state of the whole process has a significant impact on the overall video streaming quality of all the users. The possible actions of all the states compose the *Control Space*.

For each state $i = (\mathbb{M}^{(i)}, \mathbb{R}^{(i)}) \in \mathcal{P}$, the set of possible actions $\zeta^{(i)}$ is the set of available memory spaces for the user U_n in the next associated base station B_{k+1}. In specific, if $r_{n,k} = -1$ and $r_{n,k+1} = 1$ in matrix $\mathbb{R}^{(i)}$, $\zeta^{(i)} = \{0, 1, 2, \ldots, S(\mathbb{M}^{(i)}, \mathbb{R}^{(i)})\}$ where $S(\mathbb{M}^{(i)}, \mathbb{R}^{(i)})$ is the number of available blocks in base station B_{k+1} based on the local memory allocation matrix $\mathbb{M}^{(i)}$. $S(\mathbb{M}^{(i)}, \mathbb{R}^{(i)})$ can be given as

$$S(\mathbb{M}^{(i)}, \mathbb{R}^{(i)}) = M_{k+1} - \sum_{n=1}^{\mathcal{N}} \widetilde{\mathcal{M}}_{n,k+1} \tag{9}$$

Therefore, the control space can be defined as $\mathbb{C} = \bigcup_{i \in \mathcal{P}} \zeta^{(i)}$.

The determined action $a(i) \in \zeta^{(i)}$ in each state i has a great effect on the video streaming quality. For example, if the pre-loaded video content in the allocated memory space is not large enough, it is possible that the mobile video streaming would be frozen. The larger allocated local memory space in the next associated base station would store more video content which can be delivered to mobile user to support video streaming for relatively longer duration. However, this can lead to the smaller allocated local memory space for other users associated with the same base station. Therefore, how to determine the proper allocated local memory size to users entering different cells to optimize the video streaming quality is an important and challenging issue.

C. The State Transition Probabilities

Given two states $i = (\mathbb{M}^{(i)}, \mathbb{R}^{(i)})$ and $j = (\mathbb{M}^{(j)}, \mathbb{R}^{(j)})$, the state transition probability from state i to state j can be given as the state transition rate from state i to state j divided by the overall state transition rates starting from state i. For any network configuration $\mathbb{M}^{(i)}$, the duration until the next random event $\mathbb{R}^{(i)}$ occurs depends on the movements of all the users in the system. Specifically, the overall state transition rates Υ_i are the sum occurring rates of all possible events $\mathbb{R}^{(i)}$ and can be obtained as

$$\Upsilon_i = \sum_{n=1}^{\mathcal{N}} \sum_{k=1}^{\mathcal{K}} \psi_{n,k} \Gamma_{n,k} \tag{10}$$

where $\Gamma_{n,k}$ indicates the user U_n is currently in the cell C_k if $\Gamma_{n,k} = 1$ while $\Gamma_{n,k} = 0$ if the user U_n is not located in the cell C_k. Given the state $i = (\mathbb{M}^{(i)}, \mathbb{R}^{(i)})$, the state transition probability from state i to state j is given as

$$p_{i,j}(a(i)) = \frac{1}{S(\mathbb{M}^{(i)}, \mathbb{R}^{(i)})} \frac{\psi_{n,k}\Gamma_{n,k}}{\sum_{n=1}^{\mathcal{N}} \sum_{k=1}^{\mathcal{K}} \psi_{n,k}\Gamma_{n,k}} \tag{11}$$

where $1/S(\mathbb{M}^{(i)}, \mathbb{R}^{(i)})$ is the probability to select one possible local memory space in the next associated base station among $S(\mathbb{M}^{(i)}, \mathbb{R}^{(i)})$ available blocks.

D. The Cost Function

The Markov decision process is composed of a series of stages, each of which is characterized by a specific state. The objective is to determine the proper action for each stage in order to achieve optimal video streaming quality for all the users over a time duration. In order to evaluate video streaming quality, we define a metric, *video streaming quality consistency* as the percentage of time when there is no video frozenness or/and no video packet dropping out of the total time period of video streaming.

Each user U_n has a specific required data rate μ_n for video streaming. There is a buffer with the size of Ω_n associated with each mobile device U_n to store the video content for playback. For each stage, the local memory size $\mathcal{M}_{n,k+1}$ in base station B_{k+1} for the user U_n currently associated with base station B_k need to be determined so that the video content can be pre-loaded in base station B_{k+1} before U_n entering B_{k+1} to reduce the delay of delivering video content from remote video server to the base station. Due to the high user mobility and directional communication, mobile user and the base station has extremely short time to transmit video content. The proactive caching method can effectively deliver video content to mobile users with high data rate requirement, by reducing the transmission delay and handoff delay. In order to simplify the formulation of video streaming quality consistency, the following assumptions are made: a) The transmission data rate of mmWave channel from base station to mobile user is large enough. Thus, the video streaming quality is mainly affected by the amount of pre-loaded video content. b) The video content can be delivered to mobile device at the beginning of each stage. c) We consider fully uncompressed video streaming, and the video data packets are not correlated. d) User Datagram Protocol (UDP) is adopted for video packet transmission, in which there is no packet re-transmission if video packet is dropped.

Let t_χ denote the starting time of the χ^{th} stage in the Markov decision process. Thus, the time set $T = \{t_1, t_2, \ldots, t_\chi, \ldots, t_{\mathbf{X}}\}$ describes the starting time of a Markov decision process with total number of $\mathbf{X}$ stages. Let $\mathbb{Y}(t)$ represent the state of the system at any instant t with $\mathbb{Y}(t) = (\mathbb{M}^{(t)}, \mathbb{R}^{(t)}) \in \mathcal{P}$. $a(t_\chi) \in \mathcal{C}$ denotes the corresponding control for the χ^{th} stage. The duration of the χ^{th} stage can be given as

$$\tau_\chi = t_{\chi+1} - t_\chi \tag{12}$$

with the expected stage duration $\overline{\tau}_\chi = \frac{1}{\psi_{n,k}}$ if $r_{n,k}^{(\chi-1)} = -1$ and $r_{n,k+1}^{(\chi-1)} = 1$ in the matrix $\mathbb{R}^{(\chi-1)}$.

The cost of the user U_n in the χ^{th} stage is denoted by $W_n^{(\chi)}$ and is defined as the time duration when either there is video frozenness or there is packet dropping. If mobile user U_n changes its associated base station from B_k to B_{k+1} at time t_χ (characterized by $r_{n,k}^{(\chi-1)} = -1$ and $r_{n,k+1}^{(\chi-1)} = 1$), there will be video content with data size of $\mathcal{Z}\mathcal{M}_{n,k+1}$ transmitted to U_n where $\mathcal{Z}$ is the data size of each video block. Considering U_n's buffer size Ω_n and the buffer status $\Omega_n(t_\chi)$ at the beginning of the χ^{th} stage, if $\mathcal{Z}\mathcal{M}_{n,k+1} + \Omega_n(t_\chi) > \Omega_n$, packet dropping can occur and the stage cost $W_{n,loss}^{(\chi)}$ in terms of video quality loss duration resulting from video packet dropping for the user U_n in the χ^{th} stage is given as

$$W_{n,loss}^{(\chi)} = \frac{\mathcal{Z}\mathcal{M}_{n,k} + \Omega_n(t_\chi) - \Omega_n}{\mu_n}. \tag{13}$$

With the condition of $\mathcal{Z}\mathcal{M}_{n,k+1} + \Omega_n(t_\chi) > \Omega_n$, the duration it takes to fully empty the buffer by video streaming for the user U_n is

$$t_{em}^{(\chi)} = \Omega_n/\mu_n. \tag{14}$$

If the stage duration $\tau_\chi > \Omega_n/\mu_n$, the stage cost in terms of video frozenness duration is

$$W_{n,froz}^{(\chi)} = \tau_\chi - \frac{\Omega_n}{\mu_n}. \tag{15}$$

Otherwise, there is no frozenness duration in the χ^{th} stage with $\tau_\chi \le \Omega_n/\mu_n$.

If $\mathcal{Z}\mathcal{M}_{n,k+1} + \Omega_n(t_\chi) \le \Omega_n$, there is no packet dropping and the stage cost in terms of video quality loss duration $W_{n,loss}^{(\chi)}$ is zero. With the condition of $\mathcal{Z}\mathcal{M}_{n,k+1} + \Omega_n(t_\chi) \le \Omega_n$ and $\tau_\chi > [\mathcal{Z}\mathcal{M}_{n,k+1} + \Omega_n(t_\chi)]/\mu_n$, the stage cost in terms of video frozenness duration is

$$W_{n,froz}^{(\chi)} = \tau_\chi - \frac{\mathcal{Z}\mathcal{M}_{n,k+1} + \Omega_n(t_\chi)}{\mu_n}. \tag{16}$$

If $\tau_\chi \le [\mathcal{Z}\mathcal{M}_{n,k+1} + \Omega_n(t_\chi)]/\mu_n$, there is no video frozenness duration in the stage cost.

The stage cost includes video quality loss duration and video frozenness duration, i.e.,

$$W_n^{(\chi)} = W_{n,loss}^{(\chi)} + W_{n,froz}^{(\chi)}. \tag{17}$$

Therefore, in the χ^{th} stage, the stage cost of user U_n with $r_{n,k}^{(\chi-1)} = -1$ and $r_{n,k+1}^{(\chi-1)} = 1$ can be summarized in (18).

For the user U_n without changing the associated base station in the χ^{th} stage (i.e., $r_{n,k}^{(\chi-1)} = 0$ and $r_{n,k+1}^{(\chi-1)} = 0$), there is no video quality loss duration in the stage cost. The video frozenness duration could be included in the stage cost if the stage duration τ_χ is less than $\frac{\Omega_n(t_\chi)}{\mu_n}$. Therefore, the stage cost for the user U_n can be given as

$$W_n^{(\chi)} = \begin{cases} \tau_\chi - \dfrac{\Omega_n(t_\chi)}{\mu_n}, (\tau_\chi \le \dfrac{\Omega_n(t_\chi)}{\mu_n}), \\ 0, \quad (\tau_\chi > \dfrac{\Omega_n(t_\chi)}{\mu_n}). \end{cases} \tag{19}$$

Given the stage cost for each user defined either in (18), as shown at the bottom of this page, or in (19), the average cost per unit time (denoted by $\widetilde{\mathcal{W}}$) throughout the whole process is

$$\widetilde{\mathcal{W}} = \frac{1}{E(t_{\mathbf{X}})} \sum_{\chi=1}^{\mathbf{X}} \sum_{n=1}^{\mathcal{N}} W_n^{(\chi)}. \tag{20}$$

Then, the average video streaming quality consistency of all the users in the system (denoted by W_{cons}) can be given as

$$W_{cons} = 1 - \widetilde{\mathcal{W}} = 1 - \frac{1}{E(t_{\mathbf{X}})} \sum_{\chi=1}^{\mathbf{X}} \sum_{n=1}^{\mathcal{N}} W_n^{(\chi)} \tag{21}$$

The objective is to select a proper control at each stage in order to maximize the video streaming quality consistency (W_{cons}), which is equivalent to minimizing the average cost per unit time ($\widetilde{\mathcal{W}}$). Given the initial buffer state for all the users $\{\Omega_1(t_1), \Omega_2(t_1), \ldots, \Omega_{\mathcal{N}}(t_1)\}$ and the required video streaming data rates for all the users $\{\mu_1, \mu_2, \mu_3, \ldots, \mu_{\mathcal{N}}\}$, the objective function can be described by the following optimization problem (P1)

$$\text{P1:} \quad \min \frac{1}{E(t_{\mathbf{X}})} \sum_{\chi=1}^{\mathbf{X}} \sum_{n=1}^{\mathcal{N}} W_n^{(\chi)} \tag{22}$$

For mmWave networks with dense population, the local memory spaces would be not sufficient for all the users with multi-Gbps data rate requirements for the applications of uncompressed video streaming [23]. With the large enough buffer size in the mobile device, the dominant cost in each stage is the video frozenness duration. The optimal policy ϕ^* (indicating the action of each stage) obtained by the optimization problem P1 can lead to a biased solution, i.e., the memory space would incline towards those users with lower required data rates while some unlucky users will suffer from more video frozenness. To achieve fairness in video streaming quality among the users, a coefficient γ is used in a new objective function P2 as shown in (23), based on the weighted fair queuing [24].

$$\text{P2:} \quad \min \frac{1}{E(t_{\mathbf{X}})} \sum_{\chi=1}^{\mathbf{X}} \sum_{n=1}^{\mathcal{N}} \gamma_n^{(\chi-1)} W_n^{(\chi)} \tag{23}$$

where $\gamma_n^{(\chi-1)}$ can be obtained based on the memory space allocation records in the previous $(\chi - 1)$ stages for the user U_n. Specifically, $\gamma_n^{(\chi-1)}$ is defined as

$$\gamma_n^{(\chi-1)} = \frac{\delta_n}{[\omega + (1 - \widetilde{\mathcal{W}}_n^{(\chi-1)})]^\lambda} \tag{24}$$

where ω is a small positive scalar to prevent zero denominator, δ_n is a factor for user U_n to provide differentiated services according to the required data rate μ_n of video streaming, and λ is a parameter to make a tradeoff between fairness and total cost. $\widetilde{\mathcal{W}}_n^{(\chi-1)}$ is the average cost per unit time during the previous $(\chi - 1)$ stages of the user U_n and can be defined as

$$\widetilde{\mathcal{W}}_n^{(\chi-1)} = \frac{\Sigma_{m=1}^{\chi-1} W_n^{(m)}}{\Sigma_{m=1}^{\chi-1} \tau_m} \tag{25}$$

The user U_n with less accumulated video smooth duration (calculated by $1 - \widetilde{\mathcal{W}}_n^{(\chi-1)}$) has a larger value of $\gamma_n^{(\chi-1)}$. Therefore, the user U_n is likely to be allocated more memory space in order to obtain less stage cost $\gamma_n^{(\chi-1)} W_n^{(\chi)}$.

$$W_n^{(\chi)} = \begin{cases} \dfrac{\mathcal{Z}\mathcal{M}_{n,k} + \Omega_n(t_\chi) - \Omega_n}{\mu_n} + \tau_\chi - \dfrac{\Omega_n}{\mu_n}, & (\mathcal{M}_{n,k+1} > \dfrac{\Omega_n - \Omega_n(t_\chi)}{\mathcal{Z}}, \tau_\chi > \dfrac{\Omega_n}{\mu_n}), \\ \dfrac{\mathcal{Z}\mathcal{M}_{n,k} + \Omega_n(t_\chi) - \Omega_n}{\mu_n}, & (\mathcal{M}_{n,k+1} > \dfrac{\Omega_n - \Omega_n(t_\chi)}{\mathcal{Z}}, \tau_\chi \le \dfrac{\Omega_n}{\mu_n}), \\ \tau_\chi - \dfrac{\mathcal{Z}\mathcal{M}_{n,k+1} + \Omega_n(t_\chi)}{\mu_n}, & (\mathcal{M}_{n,k+1} \le \dfrac{\Omega_n - \Omega_n(t_\chi)}{\mathcal{Z}}, \tau_\chi > \dfrac{\mathcal{Z}\mathcal{M}_{n,k+1} + \Omega_n(t_\chi)}{\mu_n}), \\ 0, & (\mathcal{M}_{n,k+1} \le \dfrac{\Omega_n - \Omega_n(t_\chi)}{\mathcal{Z}}, \tau_\chi \le \dfrac{\mathcal{Z}\mathcal{M}_{n,k+1} + \Omega_n(t_\chi)}{\mu_n}). \end{cases} \tag{18}$$

When $\lambda = 0$, the memory space is allocated to users without taking into account the history of the obtained cost.

Remark: Although the cost function of the MDP model is obtained based on the scenario of uncompressed video streaming, it is easily extended to the scenario of compressed video streaming with the knowledge of packet correlation to calculate the stage cost of each user in (18).

VI. Solve Markov Decision Process Problem

The problem of proactive caching for mobile video streaming is formulated as an average cost Markov decision problem in Sec. V. This section discusses how to solve the formulated Markov decision problem. First, a typical dynamic programming method (i.e., Bellman's equation), is described to solve the Markov decision problem. Based on Bellman's equation, we propose a cell-to-cell decomposition method to achieve sub-optimal solution. Let ϕ be any stationary policy which gives the control $\phi(i) \in \zeta^{(i)} \subset \mathcal{C}$ for each state i. The average cost per unit time under the policy ϕ is denoted by θ^{ϕ} while the optimal policy ϕ^* minimizes the average cost per unit time θ^*.

A. Bellman's Equation

Markov decision process can be solved by Bellman's Equation to achieve optimality [25]. For a stationary system, the average cost θ^{ϕ} under the policy ϕ is independent of the initial state. Consequently, "differential cost" is used to compare different states. The set $\mathcal{D} = \{d(i),\ i \in \mathcal{P}\}$ defines the differential costs for all the possible states. Let $d^*(i)$ be the differential cost of state i under an optimal policy ϕ^*. Based on Bellman's Equation, the optimal average cost θ^* and the set $\{d^*(i),\ i \in \mathcal{P}\}$ satisfy the following equations:

$$d^*(i) = \min_{a \in \zeta^{(i)}} \{W^{(i)}(a) - \theta^* \overline{\tau}_i(a) + \sum_{j \in \mathcal{P}} p_{i,j}(a) d^*(j)\}, \quad (26)$$

where $W^{(i)}(a)$ is the cost of the system under state i with the control of a and is defined as

$$W^{(i)}(a) = \sum_{n=1}^{\mathcal{N}} W_n^{(i)}(a). \quad (27)$$

Given the differential cost set $\{d^*(i),\ i \in \mathcal{P}\}$ and the optimal average cost θ^*, for all $i \in \mathcal{P}$, the optimal decision $\phi^*(i)$ at state i can be obtained by

$$\phi^*(i) = \arg \min_{a \in \zeta^{(i)}} \{W^{(i)}(a) - \theta^* \overline{\tau}_i(a) + \sum_{j \in \mathcal{P}} p_{i,j}(a) d^*(j)\}. \quad (28)$$

Several value iteration and policy iteration algorithms were developed to compute the optimal average cost θ^* and the differential cost set $\{d^*(i),\ i \in \mathcal{P}\}$ in order to solve the Markov decision problem [25], [26]. Most of these iteration algorithms determining $\{d^*(i),\ i \in \mathcal{P}\}$ and θ^* have the computational complexity of $O(\|\mathcal{C}\|\|\mathcal{P}\|^2)$ [26], where $\|.\|$ is the cardinality of a set. Overwhelming computation on iteration is required to determine these values due to the huge number of system states in such a high dynamic system with a large number of mobile users moving on the highway covered with mmWave small cells. Since the system needs real-time information to determine the memory allocation for each stage, suboptimal methods for the solution must be used to obtain real-time solution.

B. Cell-by-Cell Decomposition Method

As discussed in Sec. VI-A, the size of the state space is too large for exact dynamic programming to be practical. In this section, an approximate cell-by-cell decomposition method based on Bellman's Equation is proposed to solve the formulated Markov decision problem.

For convenient usage in the following of the paper, for all $i \in \mathcal{P}$ and $a \in \zeta^{(i)}$, we define

$$D^*(i,a) = W^{(i)}(a) - \theta^* \overline{\tau}_i(a) + \sum_{j \in \mathcal{P}} p_{i,j}(a) d^*(j). \quad (29)$$

Then, the corresponding optimal control $\phi^*(i)$ for state i under the optimal policy ϕ^* is

$$\phi^*(i) = \arg \min_{a \in \zeta^{(i)}} D^*(i,a), \qquad \text{for all } i \in \mathcal{P} \quad (30)$$

and the optimal differential cost $d^*(i)$ is

$$d^*(i) = \min_{a \in \zeta^{(i)}} D^*(i,a), \qquad \text{for all } i \in \mathcal{P}. \quad (31)$$

In this paper, an approximation of $D^*(i,a)$ which is referred as $\widetilde{D}(i,a)$ is constructed to determine the control of each state, instead of calculating the actual differential cost set $\{d^*(i),\ i \in \mathcal{P}\}$ and the optimal average cost θ^*. $\widetilde{D}(i,a)$ can be given as

$$\widetilde{D}(i,a) = W^{(i)}(a) - \widetilde{\theta} \overline{\tau}_i(a) + \sum_{j \in \mathcal{P}} p_{i,j}(a) \widetilde{d}(j) \quad (32)$$

where $\widetilde{\theta}$ is the approximated optimal average cost and $\widetilde{d}(i)$ is the approximated optimal differential cost for state i. We consider cell-by-cell decomposition combined with feature extraction to construct $\widetilde{D}(i,a)$ in order to achieve a much smaller size of state space to apply dynamic programming. For each cell k, we obtain the estimation of the average cost $\widetilde{\theta}_k$ and the estimation of the differential cost $\widetilde{d}_k(i)$. The approximated optimal average cost $\widetilde{\theta}$ and approximated optimal differential cost $\widetilde{d}(i)$ for state i can be given by the summation of the corresponding values of all the cells:

$$\widetilde{\theta} = \sum_{k=1}^{\mathcal{K}} \widetilde{\theta}_k, \quad \text{and} \quad \widetilde{d}(i) = \sum_{k=1}^{\mathcal{K}} \widetilde{d}_k(i), \quad i \in \mathcal{P}. \quad (33)$$

With the approximation value of $\widetilde{\theta}$ and $\widetilde{d}(i)$, the control of each state can be determined by

$$\begin{aligned} \phi(i) &= \arg \min_{a \in \zeta^{(i)}} \widetilde{D}(i,a) \\ &= \arg \min_{a \in \zeta^{(i)}} \{W^{(i)}(a) - \overline{\tau}_i(a) \sum_{k=1}^{\mathcal{K}} \widetilde{\theta}_k + \sum_{j \in \mathcal{P}} p_{i,j}(a) \sum_{k=1}^{\mathcal{K}} \widetilde{d}_k(i)\}. \end{aligned} \quad (34)$$

A few features of cell k are extracted to form a Markov decision process for cell k to obtain the average cost $\widetilde{\theta}_k$ and the differential cost $\widetilde{d}_k(i)$. The featured Markov decision process for cell k has a limited number of states so that traditional iteration method is practical to calculate $\widetilde{\theta}_k$ and $\widetilde{d}_k(i)$. Let $\mathcal{Y}_k^{(i)} = \{y_{k,1}^{(i)}, \ldots, y_{k,\mathcal{J}}^{(i)}\}$ be the set of features to describe the state of cell k when the system is in state i, where $\mathcal{J}$ is the number of features for each cell. The approximated differential cost function $\widetilde{d}(i)$ can be given as:

$$\widetilde{d}(i) = \sum_{k=1}^{\mathcal{K}} \widetilde{d}_k(\mathcal{Y}_k^{(i)}) \tag{35}$$

In summary, the steps of the cell-by-cell decomposition method can be given as follows:

1) Determine the features $\mathcal{Y}_k^{(i)} = \{y_{k,1}^{(i)}, \ldots, y_{k,\mathcal{J}}^{(i)}\}$ for each cell k and form the feature-based Markov decision process for each cell with the state composed of the extracted features;
2) Calculate the average cost $\widetilde{\theta}_k$ and the differential cost $\widetilde{d}_k(i)$ for each cell k, and determine the approximated $\widetilde{\theta}$ and $\{\widetilde{d}(i),\ i \in \mathcal{P}\}$ based on (33);
3) Make a decision for each state i according to $\phi(i) = \arg\min_{a \in \zeta^{(i)}}[W^{(i)}(a) - \widetilde{\theta}\overline{\tau}_i(a) + \sum_{j \in \mathcal{P}} p_{i,j}(a)\widetilde{d}(j)]$, for all $i \in \mathcal{P}$.

The extracted features to represent the state for each cell is the local memory allocation for a number of users located in the cell, considering the Markov decision process for the system formulated in Sec. V. Specifically, we use a set $m_k = \{m_{k,1}, \ldots, m_{k,\mathcal{V}}\}$ to indicate the allocated memory space to each user to describe the state of cell k, where $\mathcal{V}$ is the upper bound of the number of users in each cell. Since we consider mobile users moving toward a certain direction in the highway covered by a series of mmWave cells, there should be a limitation for the number of mobile users in each cell (e.g., bounded by the number of passengers in the vehicles). Therefore, suppose there are V_k users currently located in cell k, thus the elements from m_{k,V_k+1} to $m_{k,\mathcal{V}}$ in set $m_k = \{m_{k,1}, \ldots, m_{k,\mathcal{V}}\}$ are zero and these elements are for future potential handoff users entering cell k. The state of cell k can be characterized by two sets: $m_k = \{m_{k,1}, \ldots, m_{k,\mathcal{V}}\}$ and $r_k = \{r_{k,1}, \ldots, r_{k,\mathcal{V}}\}$, where $m_{k,\upsilon}$ indicates the amount of memory space (in the unit of block) allocated to the υ^{th} user in cell k and $r_{k,\upsilon}$ indicates the next random movement event of the υ^{th} user with

$$r_{k,\upsilon} = \begin{cases} -1, & \text{(User } \upsilon \text{ intends to leave cell } C_k) \\ 1, & \text{(User } \upsilon \text{ intends to move into cell } C_k) \\ 0, & \text{(User } \upsilon \text{ intends to remain in cell } C_k). \end{cases} \tag{36}$$

Then, the state of the Markov decision process for cell k can be given as the combination of the two sets, i.e., the state space can be defined as $\mathcal{P}_k = \{i \mid i = (m_k, r_k)\}$. Following similar procedures in Sec. V, the memory allocation problem in each cell can be formulated as a Markov decision process with state space, control space, state transition probability, and the cost function. The cost function calculates the video frozenness period and video quality loss period. Based on the formulated Markov decision process, we can adopt iteration method [26] to calculate the average cost $\widetilde{\theta}_k$ and the differential cost $\widetilde{d}_k(i)$ for each cell k in order to determine the approximated $\widetilde{\theta}$ and $\widetilde{d}(i)$. The corresponding control for each state of the whole system (indicating the allocated memory space in next associated base station B_{k+1} for the user U_n entering in cell k) can be determined by (34).

Remark: To implement the proposed proactive caching method based on traditional Bellman's equation, a central controller is necessary and connected to all the mmWave base stations to collect the update information (such as buffer status at the beginning of each stage, the movement status of the users, and the required video content of each each user) to make the decision when system entering a new stage. For the cell-by-cell decomposition method, the total amount of calculation grows linearly with the number of cells. Moreover, the cell-by-cell decomposition method depends on local information and can be easily implemented in a distributive manner.

VII. Numerical Results

In this section, we conduct numerical results to show the performance of the cache memory allocation solution achieved from the Markov decision process with cell-by-cell decomposition method. There are another two solutions used in this paper for comparison. Specifically, random memory allocation solution determines a random memory space for a mobile user to pre-load the video content, under the cache memory space constraint of each base station. Another solution for comparison is the caching-free solution where there is no local memory space in each base station to pre-load and store the video content and the mobile user need to build the connection through base station and Internet to remote video server. Video quality consistency is the main metric to demonstrate the performance of the memory allocation solution on mobile video streaming. Additionally, we also show the performances in terms of the impact of cache memory size on video quality, the distribution of video smooth duration, and impact of required video data rate on system performance.

The numerical results are conducted in a segment of highway, covered by a series of 150 mmWave base stations. The speed of mobile user and the length of the highway in each cell determine the duration that the mobile user stays within each cell. The time duration $\Psi_{n,k}$ for the user U_n staying in the cell C_k until it moves into the next cell C_{k+1} is exponentially distributed with average duration of 1 s. The directional communication time between mobile user and mmWave base station is randomly selected from the range of [0.1s, 0.3s]. Initially, a number of users are located in the cells at one end of the highway, and move towards another end of the highway passing the cells sequentially. At the beginning of the considered period, the buffer installed in each mobile device is empty. The memory space size in each mmWave base station is in the unit of video block. The required data rate of the multimedia application in each user is randomly selected from the range of [0.8 Gbps, 2 Gbps]. The system parameter settings are summarized in Table II.

A. Video Streaming Quality Consistency

The results of video streaming quality consistency for the three solutions on video streaming over various numbers

TABLE II
SYSTEM RELATED PARAMETERS

Parameters	Value
Block size ($\mathcal{Z}$)	1 Mbits
Buffer size (Ω)	3 Gbits
Number of base stations ($\mathcal{K}$)	150
Number of users ($\mathcal{N}$)	50
Total considered duration	10 mins
Antenna beam width	60°
Delay for remote connection	5 ms

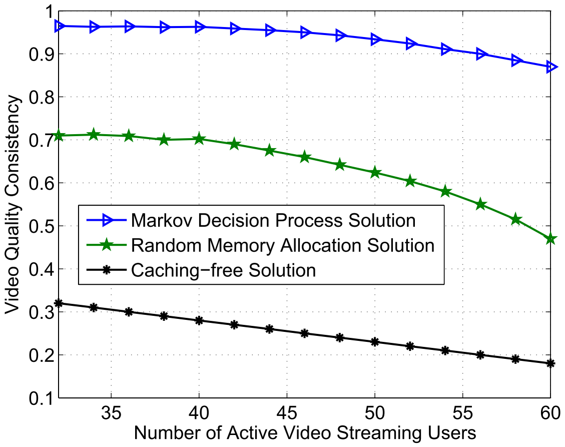

Fig. 3. Video Quality Consistency for Various Numbers of Active Users.

of active users in the system are shown in Fig. 3. Video quality consistency decreases as more number of active users included in the system to compete for the local memory spaces. Markov decision process solution and random memory allocation solution achieve significant performance improvement on video quality consistency compared with caching-free solution, because the proactive caching system can pre-load the video content to base stations to eliminate delay effect and reduce video frozenness. Markov decision process solution outperforms random memory allocation solution since Markov decision process solution can determine proper amount of video content stored in the base station to achieve optimal video streaming quality for all the users in the system.

B. Impact of Local Memory Space

Fig. 4 shows the impact of local memory space on maintaining the video streaming quality for the Markov decision process solution and the random memory allocation solution, for a fixed number of active video streaming users. Demonstrated in Fig. 4, as the local memory space in each base station increases, video quality consistency can be improved. With more local memory space in each base station, mobile users can be allocated more memory space to store video content and maintain video playout quality. If the directional mmWave connection has large enough transmission rate and the buffer size at mobile user is large enough to mitigate video packet overflow, 100 percent video quality consistency can be achieved with large enough memory space in each base station.

Fig. 5 shows video quality consistency of the Markov decision process solution with different numbers of mmWave base stations deployed in the considered section of highway.

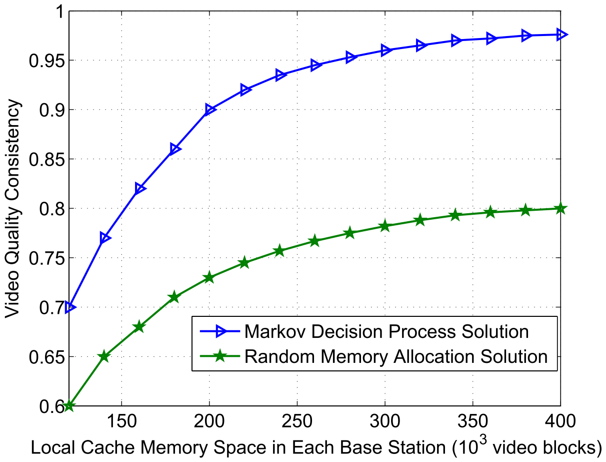

Fig. 4. Impact of Local Memory Space on Video Streaming Quality.

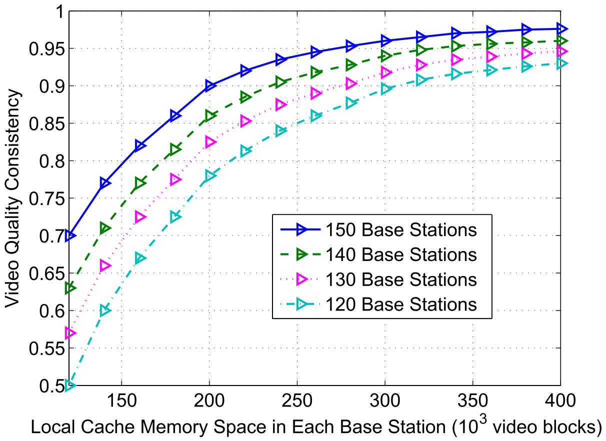

Fig. 5. Tradeoff Between mmWave Base Station Density and Cache Memory Space.

As fewer mmWave base stations are deployed on the highway, video quality consistency decreases since the total available cache memory space to pre-load video content becomes less. To achieve the same video quality consistency (e.g., 0.9), each base station requires larger cache memory space as fewer mmWave base stations are deployed. It is a tradeoff between mmWave base station density and cache memory space in each mmWave base station.

C. Probability Distribution of Video Smoothness Duration

Video streaming would be smooth if there is no frozenness and/or quality loss. In this paper, the quality loss resulting from packet dropping is because of user's buffer overflow since it is assumed that the impact of mmWave channel variation on packet reception is beyond the scope of this paper. If a relatively large buffer is installed in user's device, there would be not buffer overflow (i.e., no quality loss). Thus, the main impact on video streaming quality considered in this paper is video frozenness which happens when the pre-loaded video content in the base station is not large enough to maintain the video playout quality for a specific duration. Fig. 6 shows the probability distribution of the normalized video smoothness duration of all the active users. The Markov decision process solution can extend the video smoothness duration significantly (i.e., video smoothness durations in Markov decision

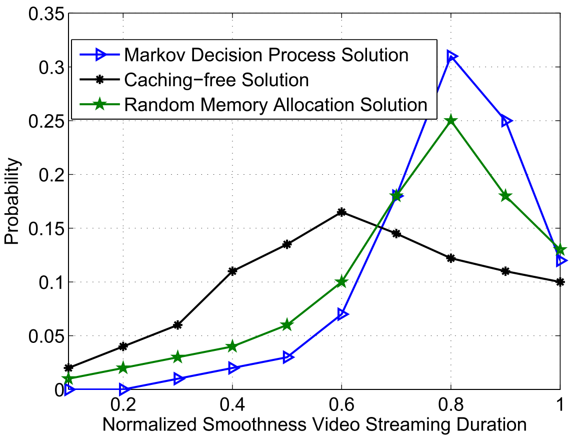

Fig. 6. Probability Distribution of Video Smoothness Duration.

process solution are located in the long duration range with higher probability) by allocating proper cache memory space to each mobile user. The proposed memory allocation method is effective to maintain video streaming quality when the mobile user and base station have very short connection time for data transmission. The random memory allocation solution can reduce the frozenness compared with caching-free solution.

VIII. Conclusions

In this paper, we have investigated the problem of cache memory allocation in each base station to 5G mobile users with video streaming on the highway covered by mmWave small cells. Markov decision process is used to formulate the dynamic cache memory allocation problem for users moving among different cells in order to achieve optimal video streaming quality. We use cell-by-cell decomposition method to reduce the size of the state space in order to practically solve Markov decision problem with dynamic programming. Demonstrated by the numerical results, the obtained solution indicating a proper cache memory space in the subsequent base station for each user entering a new cell can effectively improve the video streaming quality by pre-loading video content stored in the allocated cache memory. The proposed proactive caching method is effective on quality provisioning for non-real time video streaming and can be applied for the scenarios that the users with high mobility have frequent handoff among small cells.

References

[1] T. S. Rappaport *et al.*, “Millimeter wave mobile communications for 5G cellular: It will work!” *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[2] K. Zheng, L. Zhao, J. Mei, M. Dohler, W. Xiang, and Y. Peng, “10 Gb/s HetSNet with millimeter-wave communications: Access and networking—Challenges and protocols,” *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 222–231, Jan. 2015.

[3] J. Qiao, X. Shen, J. W. Mark, Q. Shen, Y. He, and L. Lei, “Enabling device-to-device communications in millimeter-wave 5G cellular networks,” *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 209–215, Jan. 2015.

[4] J. G. Andrews *et al.*, “What will 5G be?” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[5] M. R. Akdeniz *et al.*, “Millimeter wave channel modeling and cellular capacity evaluation,” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[6] J. Qiao, L. X. Cai, X. Shen, and J. W. Mark, “Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3824–3833, Nov. 2011.

[7] Z. Pi and F. Khan, “An introduction to millimeter-wave mobile broadband systems,” *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[8] J. G. Andrews, “Seven ways that HetNets are a cellular paradigm shift,” *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 136–144, Mar. 2013.

[9] A. Osseiran *et al.*, “Scenarios for 5G mobile and wireless communications: The vision of the METIS project,” *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[10] X. Pei, T. Jiang, D. Qu, G. Zhu, and J. Liu, “Radio-resource management and access-control mechanism based on a novel economic model in heterogeneous wireless networks,” *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 3047–3056, Jul. 2010.

[11] M. Ismail and W. Zhuang, “A distributed multi-service resource allocation algorithm in heterogeneous wireless access medium,” *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 425–432, Feb. 2012.

[12] J. Huang, Z. Li, M. Chiang, and A. K. Katsaggelos, “Joint source adaptation and resource allocation for multi-user wireless video streaming,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 5, pp. 582–595, May 2008.

[13] J. Xu, Y. Andrepoulos, Y. Xiao, and M. van der Schaar, “Non-stationary resource allocation policies for delay-constrained video streaming: Application to video over Internet-of-Things-enabled networks,” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 4, pp. 782–794, Apr. 2014.

[14] T. H. Luan, L. X. Cai, and X. Shen, “Impact of network dynamics on user’s video quality: Analytical framework and QoS provision,” *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 64–78, Jan. 2010.

[15] Y. Liu, L. Guo, F. Li, and S. Chen, “An empirical evaluation of battery power consumption for streaming data transmission to mobile devices,” in *Proc. ACM Int. Conf. Multimedia*, Nov. 2011, pp. 473–482.

[16] J. He, Z. Xue, D. Wu, D. O. Wu, and Y. Wen, “CBM: Online strategies on cost-aware buffer management for mobile video streaming,” *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 242–252, Jan. 2014.

[17] J. Qiao, X. Shen, J. W. Mark, and L. Lei, “Video quality provisioning for millimeter wave 5G cellular networks with link outage,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5692–5703, Oct. 2015.

[18] H. Ahlehagh and S. Dey, “Video caching in radio access network: Impact on delay and capacity,” in *Proc. IEEE WCNC*, Apr. 2012, pp. 2276–2281.

[19] E. Bastug, M. Bennis, and M. Debbah, “Living on the edge: The role of proactive caching in 5G wireless networks,” *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[20] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, “GreenDelivery: Proactive content caching and push with energy-harvesting-based small cells,” *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 142–149, Apr. 2015.

[21] A. Mishra, M. Shin, and W. A. Arbaush, “Context caching using neighbor graphs for fast handoffs in a wireless network,” in *Proc. IEEE INFOCOM*, Mar. 2004, pp. 351–361.

[22] J. Qiao, X. Shen, J. W. Mark, and Y. He, “MAC-layer concurrent beamforming protocol for indoor millimeter-wave networks,” *IEEE Trans. Veh. Technol.*, vol. 64, no. 1, pp. 327–338, Jan. 2015.

[23] D. Niyato and E. Hossain, “WLC04-5: Bandwidth allocation in 4G heterogeneous wireless access networks: A noncooperative game theoretical approach,” in *Proc. IEEE GLOBECOM*, Nov./Dec. 2006, pp. 1–5.

[24] A. Demers, S. Keshav, and S. Shenker, “Analysis and simulation of a fair queueing algorithm,” *ACM SIGCOMM Comput. Commun. Rev.*, vol. 19, no. 4, pp. 1–12, Sep. 1989.

[25] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley, 1994.

[26] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed., vol. I and II. Belmont, MA, USA: Athena Scientific, 2011.

Jian Qiao received the B.E. degree from the Beijing University of Posts and Telecommunications, China, in 2006, and the M.A.Sc. and Ph.D. degrees from the Department of Electrical and Computer Engineering in the University of Waterloo, Canada, in 2010 and 2015, respectively. He is currently a Post-doctoral Fellow with the College of Information Engineering, Shenzhen University, Shenzhen, China. His research interests include 5G mobile networks, millimeter-wave networks, haptic communications, medium access control, resource management, and mobile video streaming.

Yejun He (SM'09) received the Ph.D. degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2005. From 2005 to 2006, he was a Research Associate with the Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong. From 2006 to 2007, he was a Research Associate with the Department of Electronic Engineering, Faculty of Engineering, Chinese University of Hong Kong, Hong Kong. In 2012, he was a Visiting Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. From 2013 to 2015, he was an Advanced Visiting Scholar (Visiting Professor) with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Since 2011, he has been a Full Professor with the College of Information Engineering, Shenzhen University, Shenzhen, China, where he is also the Director of Shenzhen Key Laboratory of Antennas and Propagation, Shenzhen. He has authored or co-authored over 100 research papers, books (chapters), and holds 13 patents. His research interests include channel coding and modulation, 4G/5G wireless mobile communication, space-time processing, antennas and RF.

Dr. He is a Senior Member of the China Institute of Communications and a Senior Member of the China Institute of Electronics. He has also served as a Technical Program Committee Member or a Session Chair for various conferences, including the IEEE Global Telecommunications Conference, the IEEE International Conference on Communications, the IEEE Wireless Communication Networking Conference, and the IEEE Vehicular Technology Conference. He has been an Associate Editor of *Security and Communication Networks* since 2012. He also serves as a Guest editor for *Mobile Information Systems*. He was the TPC Co-Chair of WOCC 2015. He has served as a Reviewer for various journals such as the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, the IEEE WIRELESS COMMUNICATIONS, the IEEE ACCESS, the IEEE COMMUNICATIONS LETTERS, *International Journal of Communication Systems*, *Wireless Communications and Mobile Computing*, and *Wireless Personal Communications*. He is the Principal Investigator for over 20 current or finished research projects including NSFC of China, the Integration Project of Production Teaching and Research by Guangdong Province and Ministry of Education as well as the Science and Technology Program of Shenzhen City.

Xuemin (Sherman) Shen (M'97–SM'02–F'09) received the B.Sc. degree from Dalian Maritime University, China, in 1982, and the M.Sc. and Ph.D. degrees from Rutgers University, NJ, USA, in 1987 and 1990, respectively, all in electrical engineering. He is currently a Professor and the University Research Chair with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is also the Associate Chair for Graduate Studies. His research focuses on resource management in interconnected wireless/wired networks, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He is an elected member of IEEE ComSoc Board of Governors, and Chair of the Distinguished Lecturers Selection Committee. He served as the Technical Program Committee Chair/Co-Chair for IEEE Globecom'16, Infocom'14, IEEE VTC'10 Fall, and Globecom'07, the Symposia Chair for IEEE ICC'10, the Tutorial Chair for IEEE VTC'11 Spring and IEEE ICC'08, the General Co-Chair for ACM Mobihoc'15, Chinacom'07 and QShine'06, the Chair for IEEE Communications Society Technical Committee on Wireless Communications, and P2P Communications and Networking. He also serves/served as the Editor-in-Chief of the IEEE NETWORK, *Peer-to-Peer Networking and Application*, and *IET Communications*, a Founding Area Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Associate Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *Computer Networks*, and ACM/*Wireless Networks*; and the Guest Editor of the IEEE JSAC, the IEEE WIRELESS COMMUNICATIONS, the *IEEE Communications Magazine*, and *ACM Mobile Networks and Applications*. He is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Distinguished Lecturer of IEEE Vehicular Technology Society and Communications Society. He received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004, 2007, 2010, and 2014 from the University of Waterloo, the Premier's Research Excellence Award in 2003 from the Province of Ontario, Canada, and the Distinguished Performance Award in 2002 and 2007 from the Faculty of Engineering, University of Waterloo.