Caching Meets Millimeter Wave Communications for Enhanced Mobility Management in 5G Networks

Omid Semiari[†], Walid Saad[†], Mehdi Bennis[‡], and Behrouz Maham[*]

[†]Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, USA,
Emails: {osemiari,walids}@vt.edu

[‡] Centre for Wireless Communications, University of Oulu, Finland, Email: bennis@ee.oulu.fi

[*]Department of Electrical and Electronic Engineering, Nazarbayev University, Astana, Kazakhstan, Email:
behrouz.maham@nu.edu.kz

Abstract

One of the most promising approaches to overcome the uncertainty and dynamic channel variations of millimeter wave (mmW) communications is to deploy dual-mode base stations that integrate both mmW and microwave (μW) frequencies. If properly designed, such dual-mode base stations can enhance mobility and handover in highly mobile wireless environments. In this paper, a novel approach for analyzing and managing mobility in joint μW-mmW networks is proposed. The proposed approach leverages device-level caching along with the capabilities of dual-mode small base stations (SBSs) to minimize handover failures, reduce inter-frequency measurement energy consumption, and provide seamless mobility in emerging dense heterogeneous networks. First, fundamental results on the caching capabilities, including caching probability and cache duration are derived for the proposed dual-mode network scenario. Second, the average achievable rate of caching is derived for mobile users. Moreover, the impact of caching on the number of handovers (HOs), energy consumption, and the average handover failure (HOF) is analyzed. Then, the proposed cache-enabled mobility management problem is formulated as a *dynamic matching game* between mobile user equipments (MUEs) and SBSs. The goal of this game is to find a distributed handover mechanism that, under network constraints on HOFs and limited cache sizes, allows each MUE to choose between: a) executing an HO to a target SBS, b) being connected to the macrocell base station (MBS), or c) perform a transparent HO by using the cached content. The formulated matching game inherently captures the dynamics of the mobility management problem caused by HOFs. To solve this dynamic matching problem, a novel algorithm is proposed and its convergence to a two-sided dynamically stable HO policy for MUEs and target SBSs is proved. Numerical results corroborate the analytical derivations and show that the proposed solution will provides significant reductions in both the HOF and energy consumption of MUEs, resulting in an enhanced mobility management for heterogeneous wireless networks with mmW capabilities.

I. Introduction

The proliferation of bandwidth-intensive wireless applications such as social networking, high definition video streaming, and mobile TV have drastically strained the capacity of wireless cellular networks. To cope with this traffic increase, several new technologies are anticipated for 5G cellular systems: 1) dense deployment of small cell base stations (SBSs), 2) exploitation of the large amount of available bandwidth at *millimeter wave (mmW)* frequencies, and 3) enabling of *content caching* directly at the user equipments (UEs) to reduce delay and improve quality-of-service (QoS). The dense deployment of SBSs with reduced cell-sizes will boost the capacity of wireless networks by decreasing UE-SBS distance, removing coverage holes, and improving spectral efficiency. Meanwhile, mmW communications will provide high data rates by leveraging directional antennas and transmitting over a large bandwidth that can reach up to 5 GHz. In addition, exploiting the high storage capacity of the modern smart handhelds to cache the data at the UE increases the flexibility and robustness of resource management, in particular, for *mobile UEs (MUEs)*. In fact, caching allows the network to store the data content in advance, while enabling MUEs to use the cached content when sufficient wireless resources are not available.

However, dense heterogeneous networks (HetNets), composed of macrocell base stations (MBSs) and SBSs with various cell sizes, will introduce three practical challenges for mobility management. First, MUEs will experience frequent handovers (HOs), while passing SBSs with relatively small cell sizes, which naturally increases the overhead and delay in HetNets. Such frequent HOs will also increase handover failure (HOF), particularly for MUEs that are moving at high speeds [1]. In fact, due to the small and disparate cell sizes in HetNets, MUEs will not be able to successfully finish the HO process by the time they trigger HO and pass a target SBS. Second, the inter-frequency measurements that are needed to discover target SBSs can be excessively power consuming and detrimental for the battery life of MUEs, especially in dense HetNets with frequent HOs. Third, microwave (μW) frequencies are stringently congested, and thus, frequent HOs may introduce unacceptable overhead and limit the available frequency resources for the static users. In this regard, offloading MUEs from heavily utilized μW frequencies to mmW frequencies can substantially improve the spectral efficiency at the μW network.

To address these challenges and enhance mobility management in HetNets, an extensive body of work has appeared in the literature [2]–[17]. In [2], the authors provide a comprehensive overview on mobility management in IP networks. The authors in [3] present different distributed mobility management protocols at the transport layer for future dense HetNets. In [4], an energy-efficient SBS discovery method is proposed for HetNets. The work in [5] investigates HO decision algorithms that focus on improving HO between femtocells and LTE-Advanced systems. The work presented in [6] overviews existing approaches for vertical handover decisions in HetNets. In [7], the authors study the impact of channel fading on mobility management in HetNets. In addition, the work in [7] shows that increasing the sampling period for HO decision decreases the fading impact, while increasing the ping-pong effect. In [8], the authors propose an HO scheme that takes into account the speed of MUEs to decrease frequent HOs in HetNets. The authors in [9] propose an HO scheme that supports soft HO by allowing MUEs to connect with both a macrocell base station (MBS) and SBSs. Furthermore, a distributed mobility management framework is proposed in [10] which uses multiple frequency bands to decouple the data and control planes.

Although interesting, the body of work in [2]–[10] does not consider mmW communications and caching capabilities for mobility management and solely focuses on HetNets operating over μW frequencies. In addition, it does not study the opportunities that caching techniques can provide for mobility management. In [11], an HO scheme for mmW networks is proposed in which the MBS acts as an anchor for mmW SBSs to manage control signals. However, [11] assumes that line-of-sight (LoS) mmW links are always available and provides no analytical results to capture the directional nature of mmW communications. In [12], the authors propose a resource allocation scheme for hybrid mmW-μW networks that enhances video streaming by buffering content over mmW links. However, [12] does not address any mobility management challenge, such as frequent HOs or HOF. Our early work in [13] provided some of the basic insights on mobility management in μW-mmW networks. However, in contrast to this work, [13] solely focuses on an average performance analysis, does not consider dynamic HO problem for multi-MUE scenarios, and does not propose any energy management mechanisms for handling inter-frequency measurements.

Proactive caching for enhancing mobility management has been motivated by the works in [14]–[17]. In [14], the authors discuss the potential of content caching at either evolved packet core network or radio access network to minimize the traffic overhead at the core network. Moreover, the authors in [15] propose a proactive caching framework in which an ongoing IP service can be cached in advance and continuously transferred among different data centers as MUEs move across different cells. In [16], the authors propose a caching framework that stores different parts of a content at different base stations, allowing MUEs to randomly move across different cells and download different cached parts of the original content whenever possible. In addition, in [17], a proactive caching solution is proposed for mobility management by exploiting MUEs' trajectory information. Although interesting, the body of work in [14]–[17] focuses on adopting protocols that are designed for higher network layers. Moreover, these solutions do not consider caching directly at the MUEs and focus on mobility management at the core network. However, we will show how leveraging high capacity mmW communication complements the notion of caching at MUEs. In addition, caching at MUEs will provide opportunities to perform *transparent* HOs in HetNets, without requiring any data session with a target SBS.

The main contribution of this paper is a novel mobility management framework that addresses critical handover issues, including frequent HOs, HOF, and excessive energy consumption for seamless HO in emerging dense wireless cellular networks with mmW capabilities. In fact, we propose a model that allows MUEs to cache their requested content by exploiting high capacity mmW connectivity whenever available. As such, the MUEs will use the cached content and avoid performing any HO, while passing SBSs with relatively small cell sizes. First, we propose a geometric model to derive tractable, closed-form expressions for key performance metrics, including the probability of caching, cumulative distribution function of caching duration, and the average data rate for caching at an MUE over a mmW link. Moreover, we provide insight on the achievable gains for reducing the number of HOs and the average HOF, by leveraging caching in mmW-μW networks. Then, we formulate the proposed cache-enabled mobility management framework as a dynamic matching game, so as to provide a distributed solution for mobility management in HetNets, while taking the dynamics of the system into account. To solve the formulated dynamic matching problem, we first show that conventional algorithms such as the deferred acceptance algorithm adopted in [18] and [19], fail to guarantee a dynamically stable HO between MUEs and SBSs. Therefore, we propose a novel distributed algorithm that is guaranteed to converge to a dynamically stable HO policy in dense HetNets. Subsequently, the complexity of the proposed algorithm in terms of signaling overhead is analyzed. Under practical settings, we show that the proposed cache-enabled HO framework can decrease the average HOF rate by up to 45%, even for MUEs with high speeds. In addition, simulation results provide insights on the

Table I: Variables and notations

Notation	Description	Notation	Description
K	Number of SBSs	$\mathcal{K}$	Set of SBSs
U	Number of MUEs	$\mathcal{U}$	Set of MUEs
θ_u	Moving angle of MUEs	v_u	Speed of MUEs
p_k	Transmit power of SBS k	B	Segment size of video (bits)
Ω_u	Cache size of MUE u	$\Omega_u^{\max}$	Maximum cache size
t_u^c	Caching duration of MUE u	Q	Video play rate
$R^c(u,k)$	Average achievable caching rate	d^c	Traversed distance using cached content
ΔT	Time-to-trigger (TTT)	r^c	Traversed distance in caching duration
T_s	Inter-frequency cell scanning interval	t_{MTS}	Minimum time-of-stay (ToS)
θ_k	Beamwidth for SBS k	E^s	Consumed energy per cell search
$t_{u,k}$	Time-of-stay for MUE u at SBS k	t_{MTS}	Minimum required time-of-stay

achievable gains by the proposed distributed algorithm, in terms of reducing energy consumption for cell search, as well as increasing traffic offloads from the μW frequencies.

The rest of this paper is organized as follows. Section II presents the system model. Section III presents the analysis for caching in mobility management. Performance analysis of the cache-enabled mobility management is provided in Section IV. Section V formulates the mobility management as a dynamic matching and presents the proposed algorithm. Simulation results are presented in Section VI and conclusions are drawn in Section VII.

II. System Model

Consider a HetNet composed of an MBS and K SBSs within a set $\mathcal{K}$ distributed uniformly across an area. Each SBS $k \in \mathcal{K}$ can be viewed as a picocell or a femtocell, depending on its transmit power p_k. Picocells are typically deployed in outdoor venues while femtocells are relatively low-power and suitable for indoor deployments. The SBSs operate at μW frequencies that are different than those used by the MBS and, thus, there is no interference between SBSs and the MBS [4], [20]. The SBSs are also equipped with mmW front-ends to serve MUEs over either mmW or μW frequency bands [21]. The dual-mode capability allows to integrate mmW and μW radio access technologies (RATs) at the medium access control (MAC) layer of the air interface and reduce the delay and overhead for fast vertical handovers between both RATs [21]. Within this network, we consider a set $\mathcal{U}$ of U MUEs that are distributed randomly and that move across the considered geographical area during a time frame T. Each user $u \in \mathcal{U}$ moves in a random direction $\theta_u \in [0, 2\pi]$, with respect to the $\theta = 0$ horizontal angle, which is assumed fixed for each MUE over a considered time frame T. In addition, we consider that an MUE u moves with an average speed $v_u \in [v_{\text{min}}, v_{\text{max}}]$. The MUEs can receive their requested traffic

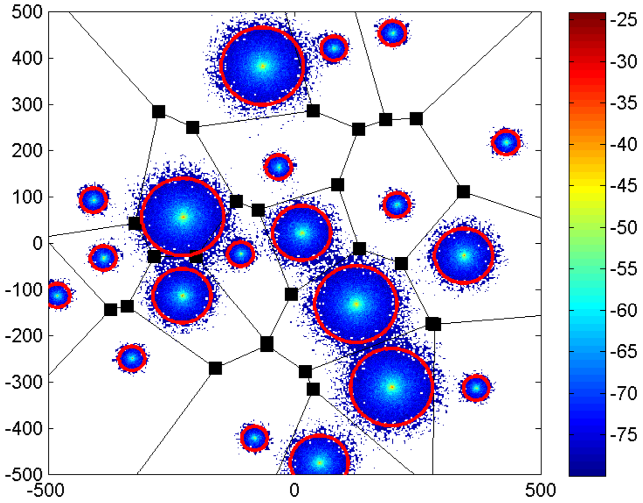

Figure 1: SBSs coverage with RSS threshold of -80 dB. Red circles show the simplified cell boundaries.

over either the mmW or the μW band.

A. Channel model

The large-scale channel effect over mmW frequencies for a link between an SBS k and an MUE $u \in \mathcal{U}$, in dB, is given by[1]:

$$L(u,k) = 20\log_{10}\left(\frac{4\pi r_0}{\lambda}\right)+10\alpha\log_{10}\left(\frac{r_{u,k}}{r_0}\right)+\chi, \tag{1}$$

where (1) holds for $r_{u,k} \geq r_{\text{ref}}$, with r_{ref} and $r_{u,k}$ denoting, respectively, the reference distance and distance between the MUE u and SBS k. In addition, α is the path loss exponent, λ is the wavelength at carrier frequency $f_c = 73$ GHz over the E-band, due to the low oxygen absorption, and χ is a Gaussian random variable with zero mean and variance ξ^2. The path loss parameters α and ξ will have different values, depending on whether the mmW link is line-of-sight (LoS) or non-LoS (NLoS). Over the μW frequency band, the path loss model follows (1), however, with parameters that are specific to sub-6 GHz frequencies.

An illustration of the considered HetNet is shown in Fig. 1. The coverage for each SBS at the μW frequency is shown based on the maximum received signal strength (max-RSS) criteria with a threshold of -80 dB. White spaces in Fig. 1 delineate the areas that are covered solely by the MBS. Here, we observe that shadowing effect can adversely increase the ping-pong effect

[1] The free space path loss model in (1) has been adopted in many existing works, such as in [22], that carry out real-world measurements to characterize mmW large scale channel effects.

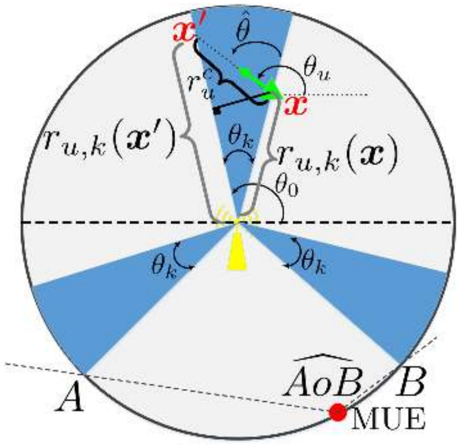

Figure 2: Antenna beam configuration of a dual-mode SBS with $N_k = 3$. Shaded areas show the mmW beams.

for MUEs. To cope with this issue, the 3GPP standard suggests L1/L3 filtering which basically applies averaging to RSS samples, as explained in [7].

B. Antenna model and configuration

To overcome the excessive path loss at the mmW frequency band, the MUEs will be equipped with electronically steerable antennas which allow them to achieve beamforming gains at a desired direction. The antenna gain pattern for MUEs follows the simple and widely-adopted sectorized pattern which is given by [23]:

$$G(\theta) = \begin{cases} G_{\text{max}}, & \text{if} \quad \theta < |\theta_m|, \\ G_{\text{min}}, & \text{otherwise}, \end{cases} \tag{2}$$

where θ and θ_m denote, respectively, the azimuth angle and the antennas' main lobe beamwidth. G_{max} and G_{min} denote, respectively, the antenna gain of the main lobe and side lobes. For SBSs, we use a model similar to the sectorized pattern in (2), however, we allow each SBS k to form N_k beams, either by using N_k antenna arrays or forming multi-beam beamforming. The beam patten configuration of an SBS k is shown in Fig. 2, where $N_k = 3$ equidistant beams in $\theta \in [0, 2\pi]$ are formed. To avoid the complexity and overhead of beam-tracking for mobile users, the direction of the SBSs' beams in azimuth is fixed. In fact, an MUE can connect to an SBS k over a mmW link, if the MUE traverses the area covered by the k's mmW beams. It is assumed that for a desired link between an SBS k and an MUE u, the overall transmit-receive gain is $\psi_{u,k} = G_{\text{max}}^2$.

C. Traffic model

Video streaming is one of the wireless services with most stringent quality-of-service (QoS) requirement. Meeting the QoS demands of such services is prone to the delay caused by frequent handovers in HetNets. In addition, HOFs can significantly degrade the performance by making frequent service interruptions. Therefore, our goal is to enhance mobility management for MUEs that request video or streaming traffic. Each video content is partitioned into small segments, each of size B bits. The network incorporates caching to transmit incoming video segments to an MUE, whenever a high capacity mmW connection is available. In fact, high capacity mmW connection, if available, allows to cache a large portion or even the entire video in a very short period of time. We define the cache size of $\Omega_u(k)$ for an arbitrary MUE u, associated with an SBS k, as the number of video segments that can be cached at MUE u as follows:

$$\Omega_u(k) = \min\left\{\left\lfloor \frac{\bar{R}^c(u,k)t_u^c}{B}\right\rfloor, \Omega_u^{\max}\right\}, \tag{3}$$

where $\lfloor . \rfloor$ and $\min\{.,.\}$ denote, respectively, the floor and minimum operands and $\Omega_u^{\max}$ is the maximum cache size. In addition, t_u^c is the *caching duration* which is equal to the time needed for an MUE u to traverse the mmW beam of its serving SBS. Considering the small green triangle in Fig. 2 as the current location of an MUE crossing a mmW beam, the caching duration will be $t_u^c = r_u^c/v_u$ where r_u^c is the distance traversed across the mmW beam. Moreover, $\bar{R}^c(u,k)$ is the *average achievable rate* for the MUE u during t_u^c. Given $\Omega_u(k)$ and the video play rate of Q, specified for each video content, the distance an MUE u can traverse with speed v_u, while playing the cached video content will be

$$d^c(u,k) = \frac{\Omega_u(k)}{Q} v_u. \tag{4}$$

In fact, the MUE can traverse a distance $d^c(u,k)$ by using the cached video content after leaving its serving cell k, without requiring an HO to any of the target cells. Meanwhile, the location information and control signals, such as paging, can be handled by the MBS during this time. As we discuss in details, such caching mechanism will help MUEs to avoid redundant cell search and HOs, resulting in an efficient mobility management in dense HetNets.

D. Handover procedure and performance metrics

The HO process in the 3GPP standard proceeds as follows: 1) Each MUE will do a cell search every T_s seconds, which can be configured by the network or directly by the MUEs, 2) If any target cell offers an RSS plus a hysteresis that is higher than the serving cell, even after L1/L3 filtering of input RSS samples, the MUE will wait for a time-to-trigger (TTT) of ΔT seconds to measure the average RSS from the target cell, 3) If the average RSS is higher than that of the serving SBS during TTT, the MUE triggers HO and sends the measurement report to its serving cell. The averaging over the TTT duration will reduce the ping-pong effect resulting from instantaneous CSI variations, and 4) HO will be executed after the serving SBS sends the HO information to the target SBS.

In our model, we modify the above HO procedure to leverage the caching capabilities of MUEs during mobility. Here, we let each MUE u dynamically determine T_s, depending on the cache size Ω_u, the video play rate Q, and the MUE's speed v_u. That is, an MUE u is capable of muting the cell search while Ω_u/Q is greater than ΔT, which enables it to have ΔT seconds to search for a target SBS before the cached content runs out.

Next, we consider the HOF as one of the key performance metrics for any HO procedure. One of the main reasons for the potential increase in HOF in HetNets is due to the relatively small cell sizes, compared to MBS coverage. In fact, HOF is typical if the time-of-stay (ToS) for an MUE is less than the minimum ToS (MTS) required for performing a successful HO. That is,

$$\gamma_{\text{HOF}}(u,k) = \begin{cases} 1, & \text{if } t_{u,k} < t_{\text{MTS}}, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where $t_{u,k}$ is the ToS for MUE u to pass across SBS k coverage. Although a short ToS may not be the only cause for HOFs, it becomes very critical within an ultra dense small cell network that encompasses MUEs moving at high speeds [24].

To search the μW carrier for synchronization signals and decode the broadcast channel (system information) of the detected SBSs, the MUEs have to spend an energy E^s per each cell search [4]. Hence, the total energy consumed by an MUE for cell search during time T will be

$$E^s_{\text{total}} = E^s \frac{T}{T_s}. \tag{6}$$

Note that increasing T_s reduces E^s_{total} which is desirable. However, less frequent scans will be equivalent to less HOs to SBSs. Therefore, there is a tradeoff between reducing the consumed power for cell search and maximizing traffic offloads from the MBS to SBSs. Content caching will allow increasing T_s, while maintaining traffic offloads from the MBS.

Next, we propose a geometric framework to analyze the caching opportunities, in terms of the caching duration t^c, and the average achievable rate $\bar{R}^c$, for MUEs moving at random directions in joint mmW-μW HetNets.

III. Analysis of Mobility Management with Caching Capabilities

In this section, we first investigate the probability of serving an arbitrary MUE over mmW frequencies by a dual-mode SBS.

A. Probability of mmW coverage

In Fig. 2, the small circle represents the intersection of an MUE u's trajectory with the coverage area of an SBS k. In this regard, $\mathbb{P}_k^c(N_k, \theta_k)$ represents the probability that MUE u with a random direction θ_u and speed v_u crosses the mmW coverage areas of SBS k. From Fig. 2, we observe that the MUE will pass through the area within mmW coverage only if the MUE's direction is inside the angle $\widehat{AoB}$. Hence, we can state the following.

Theorem 1. If an SBS k has formed a mmW beam pattern with $N_k \geq 2$ main lobes, each with a beamwidth $\theta_k > 0$, the probability of content caching will be given by:

$$\mathbb{P}_k^c(N_k, \theta_k) = \left[\frac{N_k \theta_k}{2\pi}\right] + \left[1 - \frac{N_k \theta_k}{2\pi}\right] \left[\frac{1}{2}\left(1 - \frac{1}{N_k}\right) + \frac{\theta_k}{4\pi}\right]. \tag{7}$$

Proof. Due to the equidistant beams, we have

$$\widehat{AoB} = \frac{1}{2}\widehat{AB} = \frac{1}{2}\left[2\pi - \widehat{AoB}\right] = \frac{1}{2}\left[2\pi - \left(\frac{2\pi}{N_k} - \theta_k\right)\right] = \left(1 - \frac{1}{N_k}\right)\pi + \frac{\theta_k}{2}. \tag{8}$$

Given that an arbitrary MUE can enter the circle in Fig. 2 at any direction, this MUE will be instantly covered by mmW with probability $\mathbb{P}(\boldsymbol{x}_u \in \mathcal{A}) = \frac{N_k \theta_k}{2\pi}$, where $\mathcal{A} \subset \mathbb{R}^2$ denotes the part of circle's perimeter that overlaps with mmW beams. Therefore,

$$\mathbb{P}_k^c(N_k, \theta_k) = \mathbb{P}(\boldsymbol{x}_u \in \mathcal{A}) + \left[1 - \mathbb{P}(\boldsymbol{x}_u \in \mathcal{A})\right] \frac{1}{2\pi}\widehat{AoB}, \tag{9}$$

where (9) results from the fact that $\theta_u \sim U\left[0, 2\pi\right]$. Therefore, from (8) and (9), the probability of crossing a mmW beam follows (7). ■

We can verify (7) by considering an example scenario with $N_k = 3$ and $\theta_k = \frac{2\pi}{3}$. For this example, (7) results in $\mathbb{P}_k^c(N_k, \theta_k) = 1$ which correctly captures the fact that the entire cell is covered by mmW beams.

B. Cumulative distribution function of the caching duration

To enable an MUE to use the cached content while not being associated to an SBS, it is critical to analyze the distribution of caching duration t^c for an arbitrary MUE with a random direction and speed. In this regard, consider the small green triangle in Fig. 2, which represents the location of an arbitrary MUE u, $\boldsymbol{x_u} = (x_u, y_u) \in \mathbb{R}^2$, crossing a mmW beam. First, we note that the geometry of the mmW beam of any given SBS can be defined by the location of the SBS, as well as the sides of the beam angle. Without loss of generality, we assume that the SBS of interest is located at the center, such that $\boldsymbol{x}_k = (0, 0)$. Therefore, the two sides of the beam angle will be given by

$$y = x \tan(\theta_0 - \theta_k), y = x \tan(\theta_0), \quad x > 0. \tag{10}$$

Assuming that the MUE u is currently located on the angle side $x = y \cos(\theta_0 - \theta_k)$, as shown by the small triangle in Fig. 2, then θ_0 in (10) will be $\theta_0 = \arccos\left(\frac{x_u}{r_{u,k}(\boldsymbol{x_u})}\right) + \theta_k$, where $r_{u,k}(\boldsymbol{x}) = \sqrt{x_u^2 + y_u^2}$. Hereinafter, we will use the parameter θ_0 to simplify our analysis. Let $F_{t^c}(.)$ be the cumulative distribution function (CDF) of the caching duration t^c. Thus,

$$F_{t_u^c}(t_0) = \mathbb{P}(t_u^c \leq t_0) = \mathbb{P}(r_u^c \leq v_u t_0), \tag{11}$$

where r_u^c is the distance that MUE u will traverse across the mmW beam, as shown in Fig. 2. Given the location of MUE $\boldsymbol{x}_u$, the minimum possible distance to traverse, $r_u^{\min}$, is

$$r_u^{\min} = \frac{|x_u \tan\theta_0 - y_u|}{\sqrt{1 + \tan^2\theta_0}}. \tag{12}$$

In fact, (12) gives the distance of the point $\boldsymbol{x}_u$ from the beam angle side $y = x\tan(\theta_0)$. If $r_u^{\min} > v_u t_0$, then $F_{t_u^c}(t_0) = 0$. Therefore, for the remainder of this analysis we consider $r_u^{\min} \leq v_u t_0$. Next, let $\boldsymbol{x}'_u$ denote the intersection of the MUE's path with line $y = x\tan(\theta_0)$. It is easy to see that $\boldsymbol{x}'_u = (x_u + r_u^c \cos\theta_u, y_u + r_u^c \sin\theta_u)$. Hence, $y_u + r_u^c \sin\theta_u = [x_u + r_u^c \cos\theta_u] \tan\theta_0$, and r_u^c, i.e., the distance that MUE u traverses during the caching duration t^c, is given by:

$$r_u^c = v_u t_u^c = \frac{y_u - x_u \tan\theta_0}{\tan\theta_0 \cos\theta_u - \sin\theta_u}. \tag{13}$$

Next, from (11) and (13), the CDF can be written as

$$F_{t_u^c}(t_0) = \mathbb{P}\left(\frac{y_u - x_u \tan\theta_0}{\tan\theta_0 \cos\theta_u - \sin\theta_u} \leq v_u t_0\right). \tag{14}$$

Using the geometry shown in Fig. 2, we find the CDF of the caching duration as follows:

Lemma 1. The CDF of the caching duration, t^c, for an arbitrary MUE u with speed v_u is given

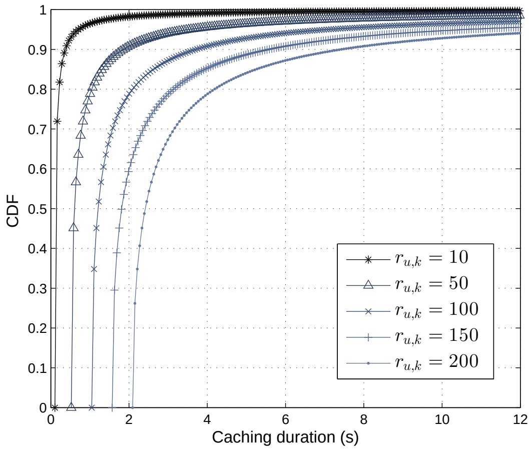

Figure 3: CDF of caching duration t^c.

by

$$F_{t^c}(t_0) = \frac{1}{\pi - \theta_k}\Bigg(\arccos\left(\frac{r_u^{\min}}{v_u t_0}\right) + \min\left\{ \arccos\left(\frac{r_u^{\min}}{r_{u,k}(\boldsymbol{x})}\right), \arccos\left(\frac{r_u^{\min}}{v_u t_0}\right) \right\}\Bigg). \tag{15}$$

Proof. From (11), $F_{t^c}(t_0) = \mathbb{P}(r_u^c \leq v_u t_0)$. To find this probability, we note that $r_u^c \leq v_u t_0$ if MUE moves between two line segments of length $v_u t_0$ that connect MUE to line $y = x \cos\theta_0$. Depending on $r_{u,k}(\boldsymbol{x})$, the intersection of line segment with $y = x\cos\theta_0$ may have one or two solutions. In case of two intersection points, the two line segments will make two equal angles with the perpendicular line from $\boldsymbol{x}_u$, to $y = x\cos\theta_0$, which each is obviously equal to $\pi - (\pi/2 - \theta_k) - \hat{\theta} = \pi/2 + \theta_k - \hat{\theta} = \arccos\left(\frac{r_u^{\min}}{v_u t_0}\right)$. Therefore,

$$F_{t^c}(t_0) = \frac{2}{\pi - \theta_k} \arccos\left(\frac{r_u^{\min}}{v_u t_0}\right). \tag{16}$$

In fact, θ_u must be within a range of $\pi - \theta_k$ for $r_u^c \leq v_u t_0$ to be valid. Now, if this angle is greater than $\pi/2 - \theta_k$, only one intersection point exists. Equivalently,

$$F_{t^c}(t_0) = \frac{1}{\pi - \theta_k}\left(\arccos\left(\frac{r_u^{\min}}{v_u t_0}\right) + \arccos\left(\frac{r_u^{\min}}{r_{u,k}(\boldsymbol{x})}\right)\right). \tag{17}$$

Integrating (16) and (17), the CDF for caching duration can be written as (15). ■

The CDF of t^c is shown in Fig. 3 for different MUE distances from the serving SBS. Fig. 3 shows that as the MUE is closer to the SBS, t^c takes smaller values with higher probability which is expected, since the MUE will traverse a shorter distance to cross the mmW beam.

IV. Performance Analysis of the Proposed Cache-enabled Mobility Management Scheme

Next, we analyze the average achievable rate for content caching, for an MUE with speed v_u, direction θ_u, and initial distance $r_{u,k}(\boldsymbol{x})$ from the serving dual-mode SBS. In addition, we evaluate the impact of caching on mobility management. For this analysis, we ignore the shadowing effect and only consider distance path loss.

A. Average Achievable Rate for Caching

The achievable rate of caching is given by:

$$R^c(u,k) = \frac{1}{v_u t_u^c}\int_{r_{u,k}(\boldsymbol{x})}^{r_{u,k}(\boldsymbol{x}')} w \log\left(1+\frac{\beta P_t \psi r_{u,k}^{-\alpha}}{wN_0}\right) dr_{u,k}, \tag{18}$$

where $\beta = (\frac{\lambda}{4\pi r_0})^2 r_0^{\alpha}$. The integral in (18) is taken over the line with length r_u^c that connects the MUE location $\boldsymbol{x}$ to $\boldsymbol{x}'$, as shown in Fig. 2. With this in mind, we can find the average achievable rate of caching $\bar{R}^c$ as follows.

Theorem 2. The average achievable rate for an MUE u served by an SBS k, $\bar{R}^c(u,k)$, is:

$$\bar{R}^c(u,k) = \mathbb{P}_k^c(N_k,\theta_k) R^c(u,k), \tag{19}$$

$$= \delta_2 \int_{f(\theta_k)}^{f(0)} \frac{1}{f^2(\theta)} \log\left(1+\delta_1 f^{\alpha}(\theta)\right) df(\theta), \tag{20}$$

$$\overset{\text{(a)}}{=} \frac{\delta_2}{\ln(2)}\bigg[2\sqrt{\delta_1}\arctan(\sqrt{\delta_1}f(\theta_k)) - \frac{\ln(\delta_1 f^2(\theta_k)+1)}{f(\theta_k)},$$

$$-2\sqrt{\delta_1}\arctan(\sqrt{\delta_1}f(0)) + \frac{\ln(\delta_1 f^2(0)+1)}{f(0)}\bigg], \tag{21}$$

where $\delta_1 = \frac{\beta P_t \psi}{wN_0}\left[r_{u,k}(\boldsymbol{x})\sin\hat{\theta}\right]^{-\alpha}$. Moreover, $\delta_2 = w r_{u,k}(\boldsymbol{x})\sin\hat{\theta}\mathbb{P}_k^c(N_k,\theta_k)/v_u t^c$, and $\hat{\theta} = \theta_u - \theta_0 + \theta_k$. For (a) to hold, we set $\alpha = 2$ which is a typical value for the path loss exponent of LoS mmW links [22].

Proof. Theorem 1 implies that with probability $1-\mathbb{P}_k^c(N_k,\theta_k)$, only μW coverage is available for an MUE. Therefore, the average achievable rate for caching over the mmW frequencies is given by (19). To simplify (19), we have

$$r_{u,k}\cos\theta = r_{u,k}(\boldsymbol{x}) + r_u\cos\hat{\theta},\ \ r_{u,k}\sin\theta = r_u\sin\hat{\theta}, \tag{22}$$

where $\hat{\theta} = \theta_u - \theta_0 + \theta_k$ and θ is an angle between the line connecting MUE to SBS, ranging from 0 to θ_k. Moreover, r_u is the current traversed distance, with $r_u = r_u^c$ once the MUE reaches $\boldsymbol{x}'$ by the end of caching duration, as shown in Fig. 2. From (22), we find $r_{u,k} = r_{u,k}(\boldsymbol{x}) \sin\hat{\theta} / \sin(\hat{\theta} - \theta)$. By changing the integral variable r_u to θ, we can write (19) as

$$\bar{R}^c(u,k) = \delta_2 \int_0^{\theta_k} \log\left(1 + \delta_1 \sin^\alpha(\hat{\theta} - \theta)\right) \frac{\cos(\hat{\theta} - \theta)}{\sin^2(\hat{\theta} - \theta)} d\theta, \tag{23}$$

where $\delta_1 = \beta P_t \psi (r_{u,k}(\boldsymbol{x}) \sin\hat{\theta})^{-\alpha} / w N_0$ and $\delta_2 = w r_{u,k}(\boldsymbol{x}) \sin\hat{\theta} \mathbb{P}_k^c(N_k, \theta_k) / v_u t^c$. Next, we can directly conclude (20) from (23) by substituting $f(\theta) = \sin(\hat{\theta} - \theta)$ in (23). For $\alpha = 2$, which is a typical value for the path loss exponent for LoS mmW links, (20) can be simplified into (21) by taking the integration by parts in (20). ■

B. Achievable gains of caching for mobility management

From (3), (4), and (21), we can find $d^c(u,k)$ which is the distance that MUE u can traverse, while using the cached video content. On the other hand, by having the average inter-cell distances in a HetNet, we can approximate the number of SBSs that an MUE can pass over distance $d^c(u,k)$. Hence, the average number of SBSs that MUE is able to traverse without performing cell search for HO is

$$\eta \approx \left\lfloor \frac{\mathbb{E}\left[d^c(u,k)\right]}{l} \right\rfloor, \tag{24}$$

where the expected value is used, since $d^c(u,k)$ is a random variable that depends on θ_u. Moreover, l denotes the average inter-cell distance. Here, we note that

$$\mathbb{E}\left[d^c(u,k)\right] = \int_0^\infty \left(1 - F_{t_u^c}(v_u t)\right) dt, \tag{25}$$

where $F_{t^c}(.)$ is derived in Lemma 1. We note that (25) is the direct result of writing an expected value in terms of CDF. Based on the definition of η in (24) and considering that the inter-frequency energy consumption linearly scales with the number of scans, we can make the following observation.

Remark 1. The proposed caching scheme will reduce the average energy consumption E^s for inter-frequency cell search by a factor of $1/\eta$ with η being defined in (24).

Furthermore, from the definition of γ_{HOF} in (5), we can define the probability of HOF as $\mathbb{P}(D_{u,k} < v_u t_{\text{MTS}})$ [25], where $D_{u,k} = t_{u,k}/v_u$, and $t_{u,k}$ is the ToS. To compute the HOF probability, we use the probability density function (PDF) of a random chord length within a circle with radius a, as follows:

$$f_D(D) = \frac{2}{\pi\sqrt{4a^2 - D^2}}, \tag{26}$$

where (26) relies on the assumption that one side of the chord is fixed and the other side is determined by choosing a random $\theta \in [0, \pi]$. This assumption is in line with our analysis as shown in Fig. 2. Using (26), we can find the probability of HOF as follows:

$$\mathbb{P}(D_{u,k} < v_u t_{\text{MTS}}) = \int_0^{v_u t_{\text{MTS}}} \frac{2}{\pi\sqrt{4a_k^2 - D^2}} dD = \frac{2}{\pi} \arcsin\left(\frac{v_u t_{\text{MTS}}}{2a_k}\right). \tag{27}$$

In fact, γ_{HOF} is a binomial random variable whose probability of success depends on the MUE's speed, cell radius, and t_{MTS}. Hence, by reducing the number of HOs by a factor of $1/\eta$, the proposed scheme will reduce the expected value of the sum $\sum \gamma_{\text{HOF}}$, taken over all SBSs that an MUE visits during the considered time T.

Thus far, the provided analysis are focused on studying the caching opportunities for the mobility management in single-MUE scenarios. However, in practice, the SBSs can only serve a limited number of MUEs simultaneously. Therefore, an HO decision for an MUE is affected by the decision of the other MUEs. In this regard, we propose a cache-enabled mobility management framework to capture the inter-dependency of HO decisions in dynamic multi-MUE scenarios.

V. Dynamic Matching for Cache-enabled Mobility Management

Within the proposed mobility management scenarios, the MUEs have a flexibility to perform either a vertical or horizontal HO, while moving to their chosen target cell. Additionally, as elaborated in Section IV, caching enables MUEs to skip a certain HO, depending on the cache size Ω. In fact, there are three HO actions possible for an arbitrary MUE that is being served by an SBS: 1) Execute an HO for a new assignment with a target SBS, 2) Use the cached content and mute HO, 3) Perform an HO to the MBS. Similarly, an MUE assigned to the MBS can decide whether to handover to an SBS, use cached content, or stay connected to the MBS.

Our next goal is to maximize possible handovers to the SBSs in order to increase the traffic offload from the MBS, subject to constraints on the HOF, SBSs' quota, and limited cache sizes. With this in mind, our goal is to find an HO policy $\boldsymbol{\zeta}$ for MUEs and target BSs[2] that

[2] For brevity, if not specified, we refer to a base station (BS) as either an SBS $k \in \mathcal{K}$ or the MBS k_0.

satisfies:

$$\underset{\boldsymbol{\zeta}}{\operatorname{argmin}} \sum_{u\in\mathcal{U}} \zeta(u,k_0), \tag{28a}$$

$$\text{s.t.}\quad \mathbb{P}\left(\sum_{k\in\mathcal{K}} \zeta(u,k) D_{u,k} < v_u t_{\text{MTS}}\right) \leq P_u^{\text{th}}, \tag{28b}$$

$$\left[1 - \sum_{k\in\mathcal{K}'} \zeta(u,k)\right] T_s \leq \frac{\Omega_u}{Q}, \tag{28c}$$

$$\sum_{k\in\mathcal{K}'} \zeta(u,k) \leq 1, \tag{28d}$$

$$\sum_{u\in\mathcal{U}} \zeta(u,k) \leq U_k^{\text{th}}, \qquad \forall k \in \mathcal{K}, \tag{28e}$$

$$\zeta(u,k) \in \{0,1\}, \tag{28f}$$

where $\mathcal{K}' = \mathcal{K} \cup \{k_0\}$ and $\boldsymbol{\zeta}$ is a vector of binary elements $\zeta(u,k) \in \{0,1\}$. In fact, a variable $\zeta(u,k) = 1$, if MUE u is chosen to execute an HO to the target cell k, otherwise, $\zeta(u,k) = 0$. Constraints (28b)-(28f) must hold for all $u \in \mathcal{U}$. In fact, the objective in (28a) is to minimize the number of MUEs associated with MBS k_0. (28b) ensures that once an MUE u is assigned to an SBS, i.e. $\sum_{k\in\mathcal{K}} \zeta(u,k) = 1$, the probability of HOF must be less than a threshold P_u^{th}, determined based on the QoS requirement of the MUE u's service. Constraint (28c) ensures that if an MUE u is not assigned to any SBS nor the MBS, there will be enough cached video segments for the next T_s time duration. Moreover, constraints (28d) and (28e) indicate, respectively, that each MUE can be assigned to at most one BS and each SBS can serve maximum U_k^{th} MUEs simultaneously.

We note that using (27), we can rewrite (28b) as $\sum_{k\in\mathcal{K}} \frac{2}{\pi} \arcsin\left(\frac{v_u t_{\text{MTS}}}{2a_k}\right) \zeta(u,k) \leq P_u^{\text{th}}$, which is a linear constraint. Hence, the posed problem in (28a)-(28f) is an integer linear programming (ILP), and thus, it is NP-hard. Although an approximation algorithm can be employed to solve (28a)-(28f), centralized algorithms are not scalable and typically introduce latency which is not desired for real-time applications such as streaming for mobile users. Moreover, these solutions will typically rely on the current network instances, such as the location, speed and cache size of the MUEs, and, hence, they fail to capture the dynamics of the system. To show this, we consider two critical scenarios, shown in Fig. 4, as follows:

Illustrative Scenario 1: Consider a feasible solution for (28a)-(28f), where an MUE u is not

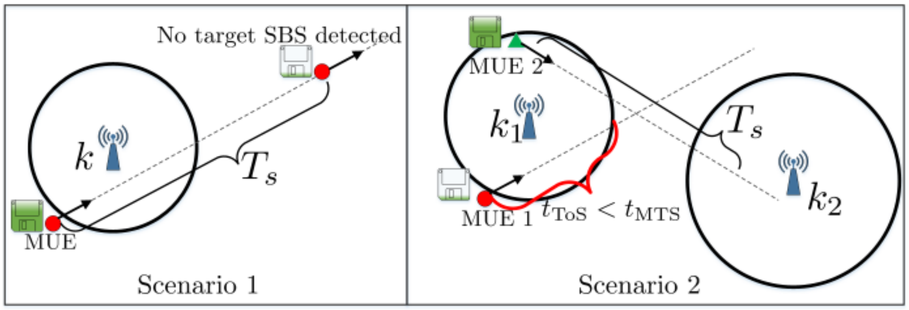

Figure 4: Two dynamic HO scenarios for cache-enabled mobile users.

assigned to the target SBS k and will use the cached content for the next T_s time duration, as shown in scenario 1 of Fig. 4. However, the MUE has to be assigned to the MBS after T_s, since eventually no target SBS is detected. Alternatively, the MUE could be assigned to k initially and fill up the cache, while later, it could use the saved cached content to reach the next target cell without requiring to be assigned to the MBS.

Illustrative Scenario 2: Consider a feasible solution for (28a)-(28f) which assigns an arbitrary MUE 1, in Fig. 4, to a target SBS k_1. If there are not enough cached contents for the MUE 1 to move to the next SBS and at the same time HO fails, the MUE has to be assigned to the MBS as shown in scenario 2 of Fig. 4. Alternatively, we could assign MUE 2 with a large cache size to the SBS k_1 such that in case of an HOF, the MUE 2 can reach the next target SBS k_2 by using its available cached contents.

These examples show that taking into account the future network information, such the estimated distance from the next target SBS, is imperative to effectively maximize the traffic offloads from the MBS. Therefore, an efficient HO policy must take into account post-handover scenarios that may occur due to the HOFs. In this regard, we propose a framework based on *dynamic matching theory* [26] which allows effective mobility management, as presented in (28a)-(28f), while capturing the future network instances, such as the cache size, MUEs' trajectory, and the topology of the network. Next, we present the fundamentals of matching theory and explain how the proposed problem can be formulated as a dynamic matching problem.

A. Handover as a matching game: Preliminaries

Matching theory is a mathematical framework that provides polynomial time solutions for combinatorial assignment problems such as (28a)-(28f) [18]. In a static form, a matching game is defined as a two-sided assignment problem between two disjoint sets of players in which the players of each set are interested to be matched to the players of the other set, according to their preference profiles. A *preference profile* for player i, denoted by $\succ_i$, is defined as a complete, reflexive, and transitive binary relation between the elements of a given set. Within the context of our proposed cache-enabled HO problem, we define the matching problem as follow:

Definition 1. Given the two disjoint sets of MUEs and BSs, respectively, in $\mathcal{U}$ and $\mathcal{K}' = \mathcal{K} \cup \{k_0\}$, a single-period *HO matching* is defined as a many-to-one mapping $\mu : \mathcal{U} \cup \mathcal{K}' \to \mathcal{U} \cup \mathcal{K}'$ that satisfies:

1) $\forall u \in \mathcal{U}$, $\mu(u) \in \mathcal{K}' \cup \{u\}$. In fact, $\mu(u) = k$ means u is assigned to k, and $\mu(u) = u$ indicates that the MUE u is not matched to any BS, and thus, will use the cached content.
2) $\forall k \in \mathcal{K}'$, $\mu(k) \subseteq \mathcal{U} \cup \{k\}$, and $\forall k \in \mathcal{K}$, $|\mu(k)| \leq U_k^{\text{th}}$. In fact, $\mu(k) = k$ implies that no MUE is assigned to the BS k.
3) $\mu(u) = k$, if and only if $u \in \mu(k)$.

Note that, by definition, the matching game satisfies constraints (28d)-(28f). More interestingly, the matching framework allows defining relevant utility functions per MUE and SBSs, which can capture the preferences of MUEs and SBSs. In this regard, the utility that an arbitrary MUE $u \in \mathcal{U}$ assigns to an SBS $k \in \mathcal{K}$ will be:

$$\Phi(u,k) = P_u^{\text{th}} - \mathbb{P}\left(\sum_{k \in \mathcal{K}} \zeta(u,k) D_{u,k} < v_u t_{\text{MTS}}\right) = P_u^{\text{th}} - \frac{2}{\pi} \arcsin\left(\frac{v_u t_{\text{MTS}}}{2a_k}\right). \tag{29}$$

Here, we observe that the utility in (29) is larger for SBSs having a larger cell radius a_k. In addition, as the speed of the MUEs increases, the utility generated from those MUEs being assigned to an SBS decreases. Meanwhile, the utility that an SBS k assigns to an MUE u is given by

$$\Gamma(u,k) = T_s - \frac{\Omega_u}{Q}. \tag{30}$$

In fact, an SBS assigns higher utility to MUEs that are not capable of using caching for the next time duration T_s. Based on the defined utility functions, the preference profile of an arbitrary MUE u, $\succ_u$, will be:

$$k \succ_u k' \;\Leftrightarrow\; \Phi(u,k) > \Phi(u,k'), \tag{31a}$$

$$u \succ_u k \;\Leftrightarrow\; \Phi(u,k) < 0, \tag{31b}$$

where $k \succ_u k'$ implies that SBS k is strictly more preferred than SBS k' by MUE u. Moreover, $u \succ_u k$ means that an SBS k is not acceptable to an MUE u, if and only if the assigned utility is negative. In fact, (31b) known as *an individual rationality constraint* and is in line with satisfying

Algorithm 1 DA Algorithm for Single-period Association Between MUEs and SBSs

```
Inputs: $\Pi \triangleq (\mathcal{U} \cup \mathcal{K}, \succ_{\boldsymbol{u}}, \succ_{\boldsymbol{k}})$.
Outputs: Stable matching $\mu^*$.
1: If not already accepted by an SBS, each unmatched MUE $u \in \mathcal{U}$ applies for its most preferred SBS $k \succ_u u$.
   Remove $k$ from $u$'s preference profile $\succ_u$.
2: Each SBS $k \in \mathcal{K}$ receives the proposals from the applicants in Step 1, tentatively accepts $U_k^{\text{th}}$ of most preferred
   MUEs from new applicants and the MUEs that are so far accepted in $\mu(k)$, and rejects the rest.
3: repeat Steps 1 to 2
4: until Each MUE $u$ is accepted by an SBS, or $u$ is applied for all SBSs $k \succ_u u$.
5: if $\exists u \in \mathcal{U}, \mu(u) \notin \mathcal{K}$ and $\Omega_u/Q < T_s$, then
6:     $\mu(u) = u$,
7: else
8:     Assign $u$ to the MBS.
9: end if
```

the feasibility condition in (28b). Similarly, we can define the preference profile of an SBS k, $\succ_k$, as follows

$$u \succ_k u' \;\Leftrightarrow\; \Gamma(u,k) > \Gamma(u',k), \tag{32a}$$

$$k \succ_k u \;\Leftrightarrow\; \Gamma(u,k) < 0, \tag{32b}$$

where (32b) is the individual rationality requirement for SBSs which is equivalent to satisfying the feasibility constraint in (28c). With this in mind, the proposed matching game is formally defined as a tuple $\Pi \triangleq (\mathcal{U} \cup \mathcal{K}, \succ_{\boldsymbol{u}}, \succ_{\boldsymbol{k}})$, where $\succ_{\boldsymbol{u}} = \{\succ_u\}_{u \in \mathcal{U}}$ and $\succ_{\boldsymbol{k}} = \{\succ_k\}_{k \in \mathcal{K}}$.

To solve this game, one desirable solution concept is to find a *two-sided stable matching* between the MUEs and SBSs, μ^*, which is defined as follow [27]:

Definition 2. An MUE-SBS pair $(u,k) \notin \mu$ is said to be a *blocking pair* of the matching μ, if and only if $k \succ_u \{\mu(u), u\}$ and $u \succ_k \{\mu(k), k\}$. Matching μ is *stable*, $\mu \equiv \mu^*$, if there is no blocking pair.

A two-sided stable association between MUEs and SBSs ensures fairness for the MUEs. That is, if an MUE u envies the association of another MUE u', then u' must be preferred by the SBS $\mu^*(u')$ to u, i.e., the envy of MUE u is not justified.

Remark 2. For a given single-period HO matching game Π, the *deferred acceptance (DA)* algorithm [18], presented in Algorithm 1, is guaranteed to find a two-sided stable association μ^* between MUEs and SBSs.

Unfortunately, the DA algorithm is not suitable to capture the dynamics of the system which arise from the mobility of the MUEs. In fact, the preference profiles of the MUEs and SBSs only depend on the current state of the system, such as the location of the MUEs, and the cache sizes. In addition, the DA algorithm cannot guarantee stability, if the preference of the MUEs change after HOFs. Thus, to be able to achieve stability for dynamic settings, such as in Scenarios 1 and 2, we need to incorporate the post-HO scenarios into the matching game, such that no MUE can block the stability even after experiencing an HOF. To this end, we extend the notion of one-stage stability in Algorithm 1 into a *dynamic stability* concept that is suitable for the problem at hand.

B. Dynamic matching for mobility management in heterogeneous networks

To account for possible scenarios that may occur after HO, we consider a two-stage dynamic matching game that incorporates within the preference profiles, some of the possible scenarios that may face the MUEs and base stations after handover execution. Such a dynamic matching will allow the MUEs to build preference profiles over different *association plans* rather than SBSs. An association plan is defined as a sequence of two matchings for a given MUE or SBS. For example, kk' is an association plan that indicates an MUE will be assigned to the SBS k followed by another HO to SBS k'. In this regard, $k_1k_2 \succ_u k_1'k_2'$ means that MUE u prefers plan k_1k_2 to $k_1'k_2'$. With this in mind, we can modify the one-period matching in Definition 1 to a relation $\mu^\dagger : \mathcal{U} \cup \mathcal{K}' \rightarrow (\mathcal{U} \cup \mathcal{K}')^2$, such that $\mu^\dagger(u) = (\mu_1(u), \mu_2(u))$, where μ_1 and μ_2 are one-period matchings. For example, $\mu^\dagger(u) = (k, u)$ indicates that MUE u will first perform an HO to SBS k, $\mu_1(u) = k$, followed by using the content of the cache after exiting the coverage of SBS k, $\mu_2(u) = u$. Next, we use the following definitions to formally define the stability in dynamic matchings [26]:

Definition 3. An MUE-BS pair (u, k) can *period-1 block* the matching, if any of the following conditions is satisfied: 1) $kk \succ_u \mu^\dagger(u)$ and $uu \succ_k \mu^\dagger(k)$; 2) $ku \succ_u \mu^\dagger(u)$ and $uk \succ_k \mu^\dagger(k)$; 3) $uk \succ_u \mu^\dagger(u)$ and $ku \succ_k \mu^\dagger(k)$; or 4) $uu \succ_u \mu^\dagger(u)$ and $kk \succ_k \mu^\dagger(k)$. A matching is *ex ante stable*, if it cannot be period-1 blocked by any MUE/BS or MUE-BS pair.

In a dynamic matching problem, either the MUEs or the BSs may block the matching, after knowing the outcome of the first matching μ_1. In this regard, we define the notion of period-2 blocking and dynamic stability as follows:

Definition 4. An MUE u can *period-2 block* a matching $\mu^\dagger$ if $(\mu_1(u), u) \succ_u \mu^\dagger(u)$. Similarly, an MUE-BS pair (u, k) can *period-2 block* if any of the following conditions is satisfied: 1) $(\mu_1(u), k) \succ_u \mu^\dagger(u)$ and $(\mu_1(k), u) \succ_k \mu^\dagger(k)$, or 2) $(\mu_1(u), u) \succ_u \mu^\dagger(u)$ and $(\mu_1(k), k) \succ_k \mu^\dagger(k)$.

A matching is said to be *dynamically stable*, if it cannot be period-1 or period-2 blocked by any MUE or MUE-BS pair[3].

From Definitions 3 and 4, we can see that, any dynamically stable matching is also an ex ante stable matching. However, ex ante stability does not guarantee dynamic stability. For example, if $\mu^{\dagger}(u) = (k, u)$ for an MUE u, ex ante stability does not guarantee that the MUE commits to use the cache, if the first handover to SBS k fails. In other words, the MUE may block an ex ante stable matching after the actual outcome of the first matching is known. To help better understand the stability for dynamic matchings, we consider the following simple example.

Example 1. Consider a dynamic matching game $\Pi^{\dagger}$, composed of MUEs $\mathcal{U} = \{u_1, u_2\}$, MBS k_0, and SBSs $\mathcal{K} = \{k_1, k_2\}$, with $U_k^{\text{th}} = 1$ for $k = k_1, k_2$, as shown in scenario 2 of Fig. 4. The preference plans of MUEs, MBS k_0, and SBSs are as follows:

$\succ_{u_1}$: $k_1k_0, \underline{k_1u_1}, u_1k_0, u_1u_1$; $\succ_{u_2}$: $k_1u_2, \underline{u_2k_2}, u_2u_2$;

$\succ_{k_1}$: $\underline{u_1k_1}, u_2k_1, k_1k_1$; $\succ_{k_2}$: $\underline{k_2u_2}, k_2k_2$; $\succ_{k_0}$: $k_0u_1, \underline{k_0k_0}$;

where the preference profiles are sorted in descending order and association plans that are not included do not meet the individual rationality constraint. Here, the underlined matching is one of the possible ex ante stable matchings. However, this matching is not dynamically stable. That is because conditioned to $\mu_1(u) = k_1$, the MUE-MBS pair (u_1, k_0) will period-2 block the matching, since $k_1k_0 \succ_{u_1} k_1u_1$ and $k_0u_1 \succ_{k_0} k_0k_0$. In practice, such a blocking occurs if the MUE experiences an HOF with its first matching to k_1.

Next, we propose an algorithm that finds a dynamically stable solution for the proposed mobility management problem.

C. Dynamically stable matching algorithm for mobility management

To find the dynamically stable solution, we note that the solution must first admit the ex ante stability. Therefore, we propose an algorithm, inspired from [26] that yields an ex ante stable association in the first stage, followed by a simple modification to resolve any possible period-2 blocking cases. For each MUE u, let $\mathcal{P}_u = \cup_{k \in \mathcal{K}}\{kk, uk, ku\}$ be the set of all plans considered by u. The algorithm proceeds as follows:

Stage-1 (Finding an ex ante stable matching):

[3] In general, a matching is dynamically stable for any time t, if it cannot be period-t blocked by any MUE or MUE-BS pair. Extending the dynamic matching to more than two periods depends on how much information is available for MUEs about the network. In this work, we focus on a two-period matching problem, since it is more tractable and practical.

1) For each MUE $u \in \mathcal{U}$, if $uu \succ_u \kappa$, for all $\kappa \in \mathcal{P}_u$, then u does not send any plan proposal to the BSs. Otherwise, MUE u sends a plan proposal to a BS, according to the most preferred plan κ_u^* as follows. If $\kappa_u^* = kk$, MUE u sends a request for a two-period association to the BS k. If $\kappa_u^* = ku$, the MUE sends an association request to BS k, only for period-1. Similarly, if $\kappa_u^* = uk$, the MUE sends an association request to k only for period-2. The MUE removes κ_u^* from its preference profile for the rest of the procedure.
2) Each SBS $k \in \mathcal{K}$ receives the plan proposals and tentatively accepts the most preferred plans, such that the quota U_k^{th} is not violated at each period. Clearly, any accepted plan κ by SBS k satisfies $\kappa \succ_k kk$.
3) MUEs with rejected plans apply in the next round, based on their next most preferred plan. The first stage of the algorithm converges, once no plan is rejected.

Proposition 1. Stage-1 of the proposed algorithm in Algorithm 2 converges to an ex ante stable association between MUEs and BSs.

Proof. Assume an MUE-BS pair (u, k) period-1 blocks the matching $\mu^\dagger$. In consequence, there is a plan $\kappa \in \{ku, uk, kk\}$ for u and a corresponding plan for k that both prefer to their current matching in $\mu^\dagger$. If $\kappa \succ_u \mu^\dagger(u)$, then the MUE u must have sent a proposal for κ to k prior to its associated plan in $\mu^\dagger$. Since κ is not eventually accepted, that means at some point, the SBS k has rejected κ in favor of another plan. Since the matching for SBSs improves at each round, we conclude that κ is less preferred by k compared to $\mu^\dagger(k)$. This contradicts the first assumption, thus, such a period-1 blocking pair does not exist and $\mu^\dagger$ is ex ante stable. ■

To avoid period-2 blockage, we introduce a certain structure to the preference profile of the SBSs as follows. For any SBS for whom the maximum quota of U_k^{th} MUEs are assigned, i.e. $|\mu_2^\dagger(k)| = U_k^{\text{th}}$,

$$\mu^\dagger \succ_k \left(\mu_1^\dagger(k), \tilde{\mu_2}^\dagger(k) \cup \{u\}\right), \tag{33}$$

where $\tilde{\mu_2}^\dagger(k)$ is ${\mu_2}^\dagger(k)$ with one associated MUE removed to accommodate a new matching with MUE u. In fact, (33) implies that an MUE cannot period-2 block the matching with any SBS k that is associated to U_k^{th} MUEs. In addition,

$$(\mu_1(k_0), u) \succ_{k_0} \mu^\dagger \iff P^{\text{th}}_{\mu_1^\dagger(u)} - \frac{2}{\pi} \arcsin\left(\frac{v_u t_{\text{MTS}}}{2a_{\mu_1^\dagger(u)}}\right) < \epsilon, \tag{34}$$

where ϵ is a non-negative scalar. In fact, (34) allow MUEs that are assigned to SBSs in period 1, with not small enough HOF probability, to be assigned to the MBS in period 2. Another

Algorithm 2 Proposed Algorithm for Dynamic Matching Between MUEs and BSs

```
Inputs: Preference plans $\kappa$ for all MUEs, MBS, and SBSs.
Outputs: Dynamically stable matching $\mu^*$.
   Phase 1:
1: For each MUE $u \in \mathcal{U}$, if $uu \succ_u \kappa$, for all $\kappa \in \mathcal{P}_u$, then $u$ does not send any plan proposal to the BSs.
   Otherwise, MUE $u$ sends a plan proposal to a BS, according to the most preferred plan $\kappa_u^*$.
2: Each SBS $k \in \mathcal{K}$ receives the plan proposals and tentatively accepts most preferred plans (also compared to
   plans that are previously accepted), such that the quota $U_k^{\text{th}}$ is not violated at each period. Clearly, any accepted
   plan $\kappa$ by SBS $k$ satisfies $\kappa \succ_k kk$.
3: repeat Steps 1 to 2
4: until No plan is rejected. The yielded ex ante stable matching is denoted by $\mu^\dagger = (\mu_1^\dagger, \mu_2^\dagger)$.
   Phase 2:
5: if $\exists u \in \mathcal{U}, \mu_2^\dagger(u) = u$, then apply DA algorithm in Algorithm 1 to the subset of MUEs with $\mu_2^\dagger(u) = u$ and
   the subset of BSs with $|\mu_2^\dagger(k)| < U_k^{\text{th}}$, considering the constraints in (33) and (34). Return yielded matching.
6: else
7:     return $\mu^\dagger$.
8: end if
```

alternative was to set P^{th} a small value from the start. However, this policy will discourage MUEs to be assigned to SBSs and could increase the load on the MBS. With this in mind, we construct the second stage of the algorithm as follows:

Stage-2 (Remove period-2 blocking pairs): Apply the deferred acceptance algorithm shown in Algorithm 1 to a subset of MUEs with $\mu_2^\dagger(u) = u$, and subset of BSs with $|\mu_2^\dagger(k)| < U_k^{\text{th}}$, while considering the constraints in (33) and (34). The proposed two-stage algorithm is summarized in Algorithm 2. Reconsidering Example 1, it is easy to follow that Algorithm 2 yields the following solution which is dynamically stable[4]:

$$\succ_{u_1}: \underline{k_1k_0}, k_1u_1, u_1k_0, u_1u_1; \qquad \succ_{u_2}: k_1u_2, \underline{u_2k_2}, u_2u_2;$$

$$\succ_{k_1}: \underline{u_1k_1}, u_2k_1, k_1k_1; \qquad \succ_{k_2}: \underline{k_2u_2}, k_2k_2; \qquad \succ_{k_0}: \underline{k_0u_1}, k_0k_0.$$

For the proposed algorithm, we can state the following results:

Theorem 3. The proposed two-stage algorithm in Algorithm 2 is guaranteed to converge to a dynamically stable association between MUEs and BSs.

Proof. From Proposition 1, the solution is guaranteed to be ex ante stable. Therefore, MUEs and BSs will not period-1 block the matching. The rest of the proof easily follows the fact that the BSs will not make a period-2 blocking pair with any MUE, due to the constraints in (33) and (34). In fact, if there is any period-2 blocking pair (u, k), there are four possible cases to consider: 1) $kk \succ_u \mu^*(u)$ and $uu \succ_k \mu^*(k)$, 2) $uk \succ_u \mu^*(u)$ and $ku \succ_k \mu^*(k)$, 3) $uk \succ_u \mu^*(u)$

[4] Here, we assume that (34) holds for u_1. Otherwise, the ex ante stable solution in Example 1 is also dynamically stable, since k_0 will not make a period-2 block pair with u_1.

Table II: Simulation parameters

Notation	Parameter	Value
f_c	Carrier frequency	73 GHz
$P_{t,k}$	Total transmit power of SBSs	$[20, 27, 30]$ dBm
K	Total number of SBSs	50
w	Available Bandwidth	5 GHz
$(\alpha_{\text{LoS}},\alpha_{\text{NLoS}})$	Path loss exponent	$(2, 3.5)$ [22]
d_0	Path loss reference distance	1 m [22]
G_{max}	Antenna main lobe gain	18 dB [23]
G_{min}	Antenna side lobe gain	-2 dB [23]
N_k	Number of mmW beams	3
θ_m, θ_k	beam width	10° [23]
N_0	Noise power spectral density	-174 dBm/Hz
t_{MTS}	Minimum time-of-stay	1s [24]
Q	Play rate	1k segments per second
B	Size of video segments	1 Mbits
$(v_{\text{min}}, v_{\text{max}})$	Minimum and maximum MUE speeds	$(1, 16)$ m/s
E^s	Energy per inter-frequency scan	3 mJ [20]

and $u'u \succ_k \mu^*(k)$, where $u' \neq u$, or 4) $k'k \succ_u \mu^*(u)$ and $u'u \succ_k \mu^*(k)$, where $k' \neq k$ and $u' \neq u$. The first two cases are not possible, since they indicate that (u, k) can period-1 block μ^* which contradicts ex ante stability. Considering the last two cases, since MUE u is not associated with SBS k in period-2, that implies that k has already been assigned to U_k^{th} MUEs. Otherwise, u would be assigned to k during the second stage of Algorithm 2. Hence, due to the constraint (33), k will not make a period-2 blocking pair with u. Similarly, the MBS will not make a period-2 blocking pair with any MUE u that is not assigned to the MBS during the second stage. Thus, no period-2 blocking pair exists and the solution μ^* satisfies dynamic stability. ■

To analyze the signaling overhead of the proposed algorithm, we consider the total number of HO requests sent to a target SBS by the MUEs. Additional control signals from the SBSs to MUEs can be managed by using a broadcast channel and do not significantly contribute to the overhead of the proposed scheme. In this regard, consider the worst-case scenario in which the initial cache size is $\Omega_u = 0$ for all $u \in \mathcal{U}$. Therefore, all MUEs seek to perform an HO to the target SBS k by sending a request for plan $\kappa = ku$ during Stage-1 of the proposed algorithm. The SBS k accepts up to U_k^{th} association plans and rejects the rest. Clearly, if there is one target SBS for the MUEs, the signaling overhead will be $\mathcal{O}(U)$. Otherwise, rejected MUEs will send an HO request to the next target SBS, based on their preference profiles. The maximum signaling overhead occurs for a case when all MUEs have the same preference profile as it introduces the highest competition among MUEs. In this case, the signaling overhead of the proposed algorithm will be $\mathcal{O}(UK)$. In addition, in Section VI, we will discuss how caching capabilities will reduce the overhead of the proposed algorithm.

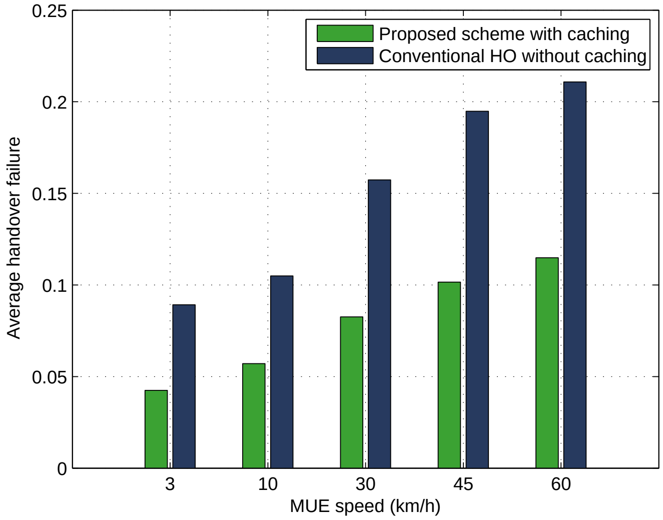

Figure 5: HOF vs different MUE speeds.

VI. Simulation Results

For simulations, we consider a HetNet composed of $K = 50$ SBSs distributed uniformly across a circular area with radius 500 meters with a single MBS located at the center and a minimum inter-cell distance of 30 meters. Moreover, the transmit power of SBSs are chosen randomly from the set of powers in $[20, 27, 30]$ dBm. The main parameters are summarized in Table II. In our simulations, we consider the overall transmit-receive antenna gain from an interference link to be random. All statistical results are averaged over a large number of independent runs. Next, we first investigate the gains achievable by the proposed cache-enabled scheme for a single user scenario. Then, we will evaluate the performance of the proposed dynamic matching approach by extending the results for scenarios with multiple MUEs in which SBSs can only serve a limited number of MUEs.

A. Analysis of the proposed cache-enabled mobility management for single user scenarios

Fig. 5 compares the average HOF of the proposed scheme with a conventional HO mechanism without caching. The results clearly demonstrate that caching capabilities, as proposed here, will significantly improve the HO process for dense HetNets. In fact, the results in Fig. 5 show that caching over mmW frequencies will reduce HOF for all speeds, reaching up to 45% for MUEs with $v_u = 60$ km/h.

Fig. 6 shows the achievable rate of caching for an MUE with $v_u = 60$ km/h, as a function of different initial distances $r_{u,k}(\boldsymbol{x})$ for various θ_u. The results in Fig. 6 show that even for MUEs with high speeds, the achievable rate of caching is significant, exceeding 10 Gbps, for all θ_u

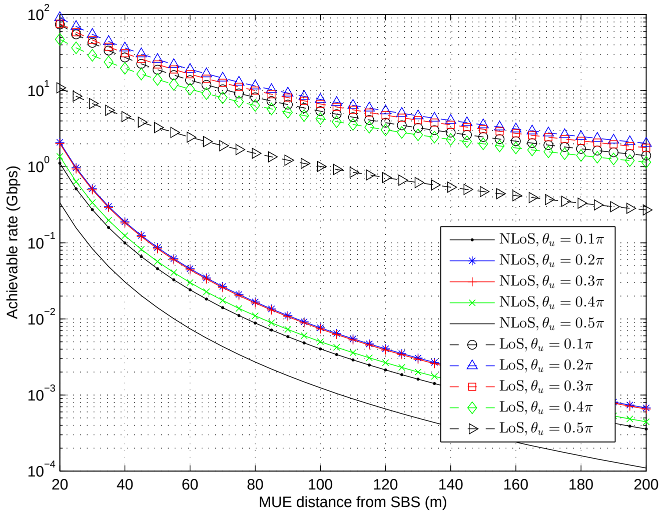

Figure 6: Achievable rate of caching vs $r_{u,k}(\boldsymbol{x})$ for different θ_u.

values and inital distance of 20 meters from the SBS. However, we can observe that the blockage can noticeably degrade the performance. In fact, for NLoS scenarios, the maximum achievable rate at a distance of 20 meters decreases to 2 Gbps.

B. Performance of the proposed dynamically stable mobility management algorithm

Here, we consider the set of MUEs entering a target cell coverage region with random directions and speeds. Moreover, the cache sizes of the MUEs are initially $\Omega_u = 10^4$ segments for all MUEs. In addition, each SBS can serve up to $U_k^{\text{th}} = 10$ MUEs. Depending on the speed of the MUE, its direction, and the location of the next target SBS, MUEs form their preferences over different plans as elaborated in Section V.

In Fig. 7, the average HOF probability of the proposed algorithm is compared with a conventional scheme that does not incorporate caching, versus the speed of the MUEs. The HOF probability is defined as the ratio of the MUEs with HOF to the total number of MUEs, for $U = 20$ and $U_k^{\text{th}} = 10$. The results in Fig. 7 show that the HOF probability increases with the speed of the MUEs, since the ToS will decrease for higher MUE speeds. In addition, we observe that the proposed algorithm can significantly reduce the HOF probability by leveraging the information on the MUE's trajectory and the network's topology. Fig. 7 also shows that for a non-zero initial cache sizes of $\Omega_u = 10^4$ segments, the algorithm is considerably robust against HOF. In fact, the HOF probability declines for speeds beyond $v_u = 8$ m/s, since higher speed allows the MUE to traverse larger distance before the cached video segments run out. Therefore,

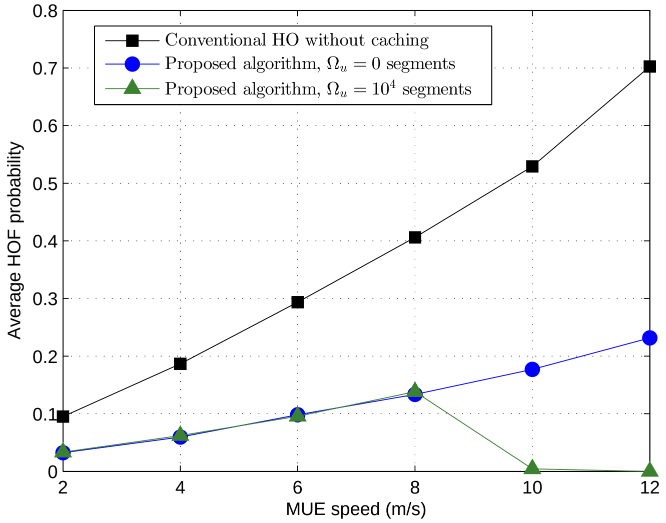

Figure 7: Average HOF probability versus MUEs' speeds.

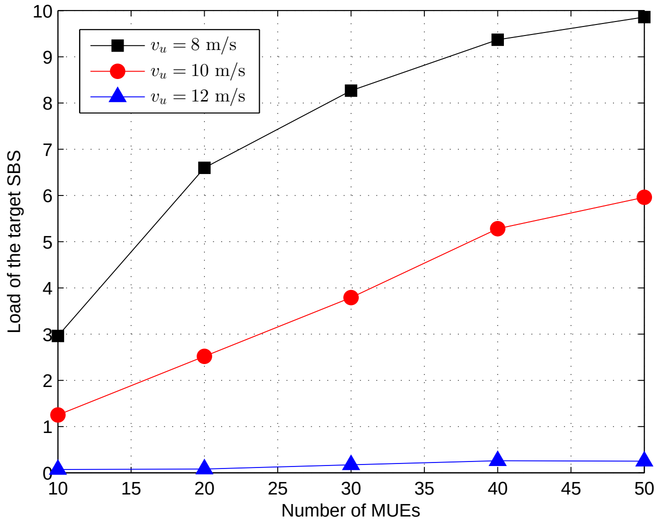

Figure 8: Load of the target SBS vs the number of MUEs.

more MUEs will be able to skip an HO to the target cell and use the cached content to move to the next available SBS.

Fig. 8 shows the load of the target cell versus the number of MUEs for different MUE speeds $v_u = 8, 10$, and 12 m/s, SBS quota $U_k^{\text{th}} = 10$, and initial cache size $\Omega_u = 10^4$ segments. Here, we observe that the proposed algorithm associates less MUEs to the target cell as the speed increases. That is due to two reasons: 1) higher speeds decrease the ToS and increase the chances of HOFs, and 2) with higher speeds, MUEs can traverse longer distances by using Ω_u cached segments and it is more likely that they can reach to the next target SBS. Fig. 8 shows that the load of the target cell reduces up to 45% when v_u increases from 8 to 10 m/s for $U = 40$ MUEs.

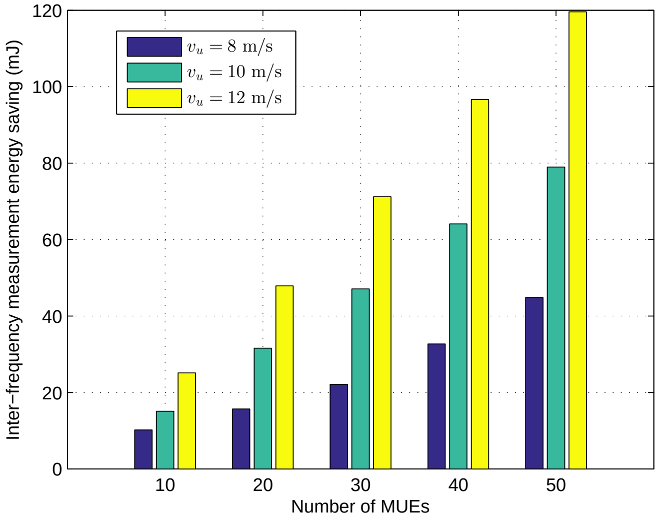

Figure 9: Energy savings for inter-frequency measurements vs number of MUEs.

In Fig. 9, the inter-frequency measurement energy savings yielded by the proposed algorithm are shown as a function of the number of MUEs. Fig. 9 shows the total saved energy for MUEs that will use the cached content and do not perform any inter-frequency measurements for handover to an SBS for an initial cache size of $\Omega_u = 10^4$ segments and different MUE speeds. For $U = 50$, MUEs that perform conventional handover without caching will require $UE^s = 150$ mJ total energy for performing inter-frequency measurements. However, the results in Fig. 9 show that the proposed scheme achieves up to 80%, 52%, and 29% gains in saving energy, respectively, for MUE speeds $v_u = 8, 10,$ and 12 m/s by leveraging cached segments and muting unnecessary cell search. Given that the required energy for measurements linearly scales with the number of MUEs, the results in Fig. 9 can also be interpreted as the offloading gains of the proposed approach, compared with conventional HO with no caching. Moreover, these results are consistent with those shown in Fig. 8. In fact, as the speed of MUEs increases, the HOF probability increases, and thus, MUEs tend to be assigned to the MBS or use their cached content. In addition, fast moving MUEs are more likely to reach the next target cell before the cached content runs out.

In Fig. 10, we show the signaling overhead resulting from the proposed algorithm versus the number of MUEs, for $\Omega_u = 10^4$ initial cache size and different MUE speeds. Here, we refer to the signaling overhead as the number of HO requests sent to the target SBS by the MUEs. Fig. 10 shows that for low speeds $v_u = 2$ m/s, almost all MUEs will attempt to hand over to

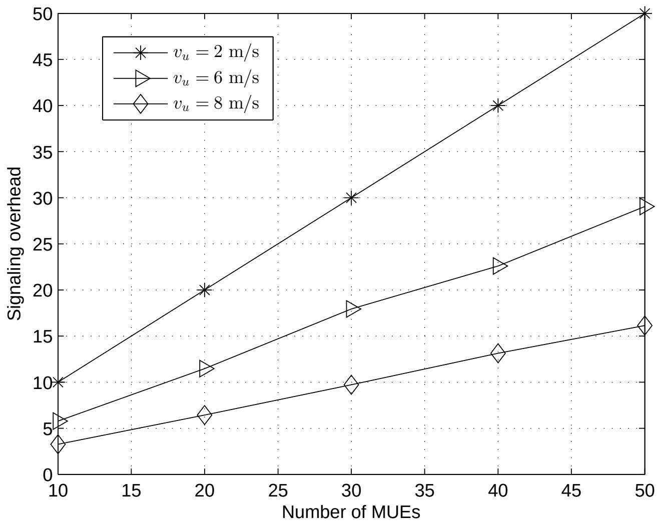

Figure 10: Signaling overhead vs number of MUEs.

the target SBS, since the time needed for traversing the SBS coverage is longer than the time available by using the cached content. Nonetheless, the results in Fig. 10 clearly demonstrate that the proposed algorithm has a manageable overhead, not exceeding 17 requesting signals for a network size of $U = 50$ with $v_u = 8$ m/s. In fact, it is interesting to note that although mobility management is, in general, more challenging for high speed MUEs, the overhead of the proposed algorithm decreases for high speed scenarios. This is due to the fact that high speed MUEs use the cached content more effectively than slow-moving MUEs.

VII. Conclusions

In this paper, we have proposed a comprehensive framework for mobility management in integrated microwave-millimeter wave cellular networks. In particular, we have shown that by smartly caching video contents while exploiting the dual-mode nature of the network's base stations, one can provide seamless mobility to the users. We have derived various fundamental results on the probability and the achievable rate for caching video contents by leveraging millimeter wave high capacity transmissions. In addition, to capture the dynamics of the mobility management, we have formulated the multi-user handover problem as a dynamic matching game between the mobile users and small base stations. To solve this game, we have proposed a novel algorithm that is guaranteed to converge to a dynamically stable handover mechanism. Moreover, we have shown that the proposed cache-enabled mobility management framework provides significant gains in reducing the number of handovers, energy consumption for inter-frequency scanning, as well as mitigating the handover failure. Numerical results have corroborated our analytical results and showed that the significant rates for caching can be achieved over the mmW frequencies, even for fast mobile users. In addition, the results have shown that the proposed approach substantially decreases the handover failures and provides significant energy savings in heterogeneous networks.

References

[1] D. Lopez-Perez, I. Guvenc, and X. Chu, "Mobility management challenges in 3GPP heterogeneous networks," *IEEE Communications Magazine*, vol. 50, pp. 70–78, December 2012.

[2] I. F. Akyildiz, J. Xie, and S. Mohanty, "A survey of mobility management in next-generation all-IP-based wireless systems," *IEEE Wireless Communications*, vol. 11, pp. 16–28, August 2004.

[3] F. Giust, L. Cominardi, and C. J. Bernardos, "Distributed mobility management for future 5G networks: overview and analysis of existing approaches," *IEEE Communications Magazine*, vol. 53, pp. 142–149, January 2015.

[4] A. Prasad, O. Tirkkonen, P. Lundn, O. N. C. Yilmaz, L. Dalsgaard, and C. Wijting, "Energy-efficient inter-frequency small cell discovery techniques for LTE-advanced heterogeneous network deployments," *IEEE Communications Magazine*, vol. 51, pp. 72–81, May 2013.

[5] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility management for femtocells in LTE-advanced: Key aspects and survey of handover decision algorithms," *IEEE Communications Surveys Tutorials*, vol. 16, pp. 64–91, First 2014.

[6] A. Ahmed, L. M. Boulahia, and D. Gaiti, "Enabling vertical handover decisions in heterogeneous wireless networks: A state-of-the-art and a classification," *IEEE Communications Surveys Tutorials*, vol. 16, pp. 776–811, Second 2014.

[7] K. Vasudeva, M. Simsek, D. Lopez-Perez, and I. Guvenc, "Impact of channel fading on mobility management in heterogeneous networks," in *2015 IEEE International Conference on Communication Workshop*, June 2015.

[8] M. Khan and K. Han, "An optimized network selection and handover triggering scheme for heterogeneous self-organized wireless networks," *Mathematical Problems in Engineering*, vol. 16, pp. 1–11, 2014.

[9] H. Zhang, N. Meng, Y. Liu, and X. Zhang, "Performance evaluation for local anchor-based dual connectivity in 5G user-centric network," *IEEE Access*, vol. 4, pp. 5721–5729, September 2016.

[10] I. Elgendi, K. S. Munasinghe, and A. Jamalipour, "Mobility management in three-tier sdn architecture for densenets," in *2016 IEEE Wireless Communications and Networking Conference*, pp. 1–6, April 2016.

[11] S. G. Park and Y. S. Choi, "Mobility enhancement in centralized mmwave-based multi-spot beam cellular system," in *2015 International Conference on Information and Communication Technology Convergence*, pp. 200–205, October 2015.

[12] J. Qiao, X. S. Shen, J. W. Mark, and L. Lei, "Video quality provisioning for millimeter wave 5G cellular networks with link outage," *IEEE Transactions on Wireless Communications*, vol. 14, pp. 5692–5703, October 2015.

[13] O. Semiari, W. Saad, M. Bennis, and B. Maham, "Mobility management for heterogeneous networks: Caching meets millimeter wave to provide seamless handover," in *submitted to the 2017 IEEE International Symposium on Information Theory*, June 2017.

[14] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *IEEE Communications Magazine*, vol. 52, pp. 131–139, February 2014.

[15] Y. Rao, H. Zhou, D. Gao, H. Luo, and Y. Liu, "Proactive caching for enhancing user-side mobility support in named data networking," in *2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pp. 37–42, July 2013.

[16] K. Poularakis and L. Tassiulas, "Exploiting user mobility for wireless content delivery," in *2013 IEEE International Symposium on Information Theory*, pp. 1017–1021, July 2013.

[17] A. S. Gomes, B. Sousa, D. Palma, V. Fonseca, Z. Zhao, E. Monteiro, T. Braun, P. Simoes, and L. Cordeiro, "Edge caching with mobility prediction in virtualized LTE mobile networks," *Future Generation Computer Systems*, 2016.

[18] D. Gale and L. Shapley, "College admissions and the stability of marriage," *Amer. Math. Monthly*, vol. 69, 1962.

[19] E. Jorswieck, "Stable matchings for resource allocation in wireless networks," in *Proc. of 17th International Conference on Digital Signal Processing (DSP)*, (Corfu, Greece), July 2011.

[20] A. Prasad, O. Tirkkonen, P. Lunden, O. N. Yilmaz, L. Dalsgaard, and C. Wijting, "Energy-efficient inter-frequency small cell discovery techniques for LTE-Advanced heterogeneous network deployments," in *IEEE Communications Magazine*, pp. 72–81, May 2013.

[21] A. Ravanshid, P. Rost, D. S. Michalopoulos, V. V. Phan, H. Bakker, D. Aziz, S. Tayade, H. D. Schotten, S. Wong, and O. Holland, "Multi-connectivity functional architectures in 5g," in *2016 IEEE International Conference on Communications Workshops*, pp. 187–192, May 2016.

[22] A. Ghosh, R. Ratasuk, P. Moorut, T. S. Rappaport, and S. Sun, "Millimeter-Wave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1152 –1163, June 2014.

[23] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Tractable model for rate in self-backhauled millimeter wave cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 2196–2211, October 2015.

[24] 3GPP, "E-UTRA: Mobility enhancements in heterogeneous networks," *3rd Generation Partnership Project*, vol. Rel 11, September 2012.

[25] C. H. M. de Lima, M. Bennis, and M. Latva-aho, "Modeling and analysis of handover failure probability in small cell networks," in *Proc. of IEEE Conference on Computer Communications Workshops*, pp. 736–741, April 2014.

[26] S. Kadam and M. H. Kotowski, "Multi-period matching," *Available online at http://scholar.harvard.edu/kadam/publications/multi-period-matching*, April 2016.

[27] A. E. Roth and M. A. O. Sotomayor, *Two-sided matching: A study in game-theoretic modeling and analysis*. Cambridge University Press, 1992.