Performance Analysis of Cache-Enabled Handover Management for Vehicular Networks

Neetu R.R, Gourab Ghatak, Vivek Ashok Bohara and Anand Srivastava

***Abstract*—Emerging vehicular applications and high data demands have led to an increasing need for storage and computational resources in vehicular networks (VNs). Due to the limited resource availability, it is challenging to provide high-quality of service (QoS) to the end users. In our work, we propose a cache-enabled handover management scheme leveraging proactive caching at mobile terminals (MTs) to minimize handovers (HOs) and improve the average rate of the network. We employ a stochastic geometry approach to analyze the network performance, where the small base stations (SBSs) are distributed as 1-D Poisson point process (PPP). Considering the time overhead for each handover, we derive the analytical expression for the effective average rate experienced by the typical MT. We derive the distribution of total interference power experienced by the MT at a given time instant. Further, we calculate the minimum cache size required to obtain the maximum average rate in the network for a noise-limited and interference-dependent network. Finally, we study the impact of frequency reuse on the network performance and obtain an optimal value of frequency reuse factor for which the interference-dependent network offers the same performance as the noise-limited network for a given play rate.**

***Index Terms*—Vehicular networks, local caching, handover management, stochastic geometry**

I. Introduction

The fifth-generation (5G)-enabled VNs have led to a wide variety of applications, which include autonomous driving, auto navigation, and delay-sensitive applications such as Voice over Internet Protocol (VoIP) and infotainment applications [1]. These require high computation and storage resources [2]. The resource constraints in VNs inhibit the quality of service (QoS) guarantees to the user, which is a major bottleneck in the evolution of VNs [3]. In order to address this challenge, proactive caching is considered a key solution that overcomes resource constraints by bringing data closer to the vehicle terminals (VTs) [4]. In 5G-enabled VNs, because of their faster download speed and reduced latency, caching capabilities of the networks have increased and have empowered real-time video streaming. Local caching reduces the number of handovers by utilizing the cache data for a seamless handover process in connected networks ensuring high quality of service to the user. The dominant form of communications between the connected autonomous VNs are vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications. In the case of the former, essential safety-related messages like speed, location, braking, the direction of travel, etc., are exchanged between the vehicles [5]. In V2I communications, vehicles receive traffic-related contents such as radio-frequency indication (RFID) information, congestion, road conditions, and infotainment contents directly from the content service provider (CSP) with minimal delay [6]. In a local caching-enabled framework, the user requesting the content is served by caching the data at the VTs without accessing the core network. This leads to significant improvement in spectral efficiency and a decrease in the end-to-end latency of the network. The ultra-densification approach is considered an effective solution to satisfy the increasing demands in the network and improve coverage and capacity. However, due to the high mobility of vehicles, frequent switching occurs among multiple base stations (BSs), specifically when the BSs are densely located. This leads to an increase in handover rates and handover failures, thus degrading the service continuity and quality of experience (QoE) at the user end. In this paper, we focus on reducing the handover rates in an ultra-dense network, leveraging caching capabilities at the MT, and improving the network performance.

Neetu R.R, Vivek Ashok Bohara and Anand Srivastava are with the Department of Electronics and Communication Engineering, IIIT-Delhi, India. (email: {neetur, vivek.b, anand} @iiitd.ac.in) and Gourab Ghatak is with the Department of Electrical Engineering, IIT Delhi, India (email: gghatak@ee.iitd.ac.in).

A. Related Work

1) Caching in V2X Networks: The authors in [4] have proposed a novel architecture for data dissemination in V2X networks, considering the proactive caching and file-sharing strategy. They have considered a unmanned-aerial vehicle (UAV)-enabled scheduling protocol and showed that the adoption of proactive caching in VTs can overcome the limited endurance issue in UAVs. The authors proposed a dynamic trajectory scheduling for UAVs so that caching duration is reduced, which increases the network throughput. In [7], the authors discuss various schemes for caching data at the vehicular level and BS level for integrated VLC/RF technologies. They developed an optimal caching positioning, observing various parameters, such as rate and latency, for minimizing the communication latency of the vehicles. It was also observed that caching at vehicles and BSs decreases the expected retrieval time of the contents, thus reducing the latency in the network. The authors in [8] developed an architecture to enable autonomous driving in vehicular networks. They leveraged caching to improve vehicular content delivery to meet real-time constraints in autonomous vehicles. The authors also observed that using caching at the edge reduces the elapsed time, which consists of request time, process time, and response time, thus ensuring a seamless driving experience for the user.

2) Handover Management: The study of HOs using different schemes, e.g., velocity-aware scheme [9], route-aware

scheme [10] and alternate HO scheme [11] and other schemes have been discussed in the literature [12] - [15]. These works analyzed the impact of user equipment (UE) velocity and BS intensity on various metrics like HO cost, average throughput, and HO failure rate. The authors in [16] have developed a time-based handover skipping scheme in which they control the frequency of HOs for a moving UE. Specifically, the authors have set a threshold time, referred to as skipping time, and any HOs attempted before this threshold time are skipped. Considering a random walk model of user mobility, the authors have constructed an evaluation function of transmission performance, considering the HO rate and expected downlink data rate for this scheme. Further, they have derived expressions for optimal skipping time, which maximizes the evaluation function and other system parameters in their scheme. However, they haven't evaluated the optimal skipping time based on the variation of expected data rate with respect to other parameters like the intensity of BSs, interference model, transmit power of the BSs etc., other than the velocity of UE.

In [17], the authors have proposed a user-centric handover skipping scheme, which is cooperation based, to reduce the number of HOs. They also attempt to reduce the handover cost for frequent cluster reformations in ultra-dense networks (UDNs). The proposed group-cell handover skipping scheme (GCHO-S) is modeled to support the MT along the UE trajectory in skipping the best-connected group-cell, which reduces frequent group-cell reformations. The authors observed a significant decrease in HO rate using the proposed scheme. However, they have studied the model only from HO rate perspective, which is not desirable, because many other factors are affected by skipping the HOs, such as the expected throughput and energy efficiency of the system. Furthermore, the authors haven't considered the impact of interference between the BSs in each group-cell region, which significantly deteriorates the system's performance. The authors in [18] have developed a feasible heterogeneous networks (HetNets) model for traffic hotspots area. The locations of users and clustered SBSs are coupled in hotspots to capture the correlation between them. A modified random waypoint (MRWP) model is proposed to eliminate the density wave phenomenon, i.e., the distribution of nodes in a finite area tends to concentrate on the center and thus increase the accuracy of the handover decision. The authors have analyzed the effect of intensity of BSs, the velocity of the user, and transmit power on HO rate, HO failure rate, and ping-pong rate. However, they haven't analyzed the effect of the above parameters, such as the intensity of BSs and the user's velocity on the expected data rate experienced by the user in the network. The authors in [19] analyzed the performance of the vehicular network with platooned vehicular traffic modeled ad 1-D Matern cluster process (MCP) on the road. The BSs or roadside units (RSUs) are modeled as 1-D PPP. They study the coverage and reliability performance of 1-D vehicular network with platooned traffic. They also observed that interchanging platoon radius and density causes similar performance. In [20], the author has proposed a two-phase transmission policy for the propagation of critical data in a 1-D V2X network within a fixed deadline. The two phases are broadcast and relay. They employed resource partitioning between the two phases and observed a reduction in information outage for a 1-D V2X network.

B. Motivation and Contribution

In the above literature, many have studied the handovers due to user mobility and their effect on various performance metrics. The work in [21] is closely related to our work, where the authors have studied the advantages of caching to address critical handover issues such as frequent handovers, handover failures, the load of target BSs etc. They exploited the high-capacity millimeter wave (mmWave) connections of a dual-mode BS, allowing the UE to cache their requested content and avoid unnecessary HOs. They evaluated the expected rate of caching the data at the UE and studied the effect of the user's velocity on the caching rate and the average HO failure. They designed an optimization problem to reduce the load on the macro base station (MBS) by maximizing the possible HOs to SBS, thereby increasing the traffic offload from MBS.

However, in our work, we perform a downlink analysis from the perspective of a mobile terminal (MT) streaming a video at a fixed play rate. We propose a cache-based handover management scheme by leveraging the caching capabilities at the MT for skipping HOs and improving the network performance in terms of average rate and energy efficiency. We perform temporal averaging followed by spatial averaging over realizations to present the efficiency of our analysis. Generally, stochastic geometry papers deal with static scenarios considering the typical user at the origin, and averaging is performed over different spatial realizations. However, the dynamic nature of the user will influence the realizations of the network in a different manner. Hence, performing this dual expectation is challenging and is not usually present in literature.

In [22], the authors adopt a comprehensive strategy for maximizing the effective area spectral efficiency (ASE) of downlink transmissions to a mobile user. They characterize the average spectral efficiency per unit time by spatially averaging across network realizations, by considering the typical user at the origin at a given time. However, they did not perform the dual-expectation analysis that spans both temporal dynamics and network realizations, which differs from our work. Similarly, in [9], [11] and [16], the authors discussed different HO skipping strategies. They characterized the average throughput experienced by the MT by averaging over the network realizations without considering the temporal dynamics, thus making our analysis distinct. We observe that leveraging local caching to skip HOs, thereby improving the network performance, is not a straightforward solution. In this regard, we observe a trade-off between the cache size and the energy efficiency of the network. In [21], the authors did not consider this trade-off and have not presented the network performance in terms of the QoS experienced by the MT.

Motivated by this, in this paper, we study the limitations in [21], and present a handover management scheme, leveraging local caching capabilities at MT, and characterize the average rate experienced by the MT and energy efficiency of the network, using stochastic geometry.

The main contributions of this paper are summarized as follows:

1) We propose a novel stochastic geometry-based framework to characterize the user performance in terms of the average rate and energy efficiency for a cache-enabled vehicular network. We develop a HO management scheme for the vehicular network by integrating caching at the MTs and thus reducing the frequency of HOs in the network while maintaining the QoE at the user end.
2) We derive the analytical expressions for the cache distance, i.e., distance traveled by the MT using the cached data, considering the impact of interference power, velocity of MT, and cache size at the MT. The HO rate experienced by the MT for conventional and cache-based HO management scheme conditioned on the interference power at time t is derived.
   We derive the analytical expressions for the effective average rate experienced by the MT considering the HO time overhead and the impact of interference power experienced by the MT. The time-averaged rate is derived for a given realization which is spatially averaged over all the realizations. This spatio-temporal expectation is challenging and has not previously been attempted in any other stochastic geometry literature. Our results recommend selecting the optimal deployment density for maximizing the effective average rate of the MT for a given play rate.
3) We consider a power consumption model considering the energy consumed for caching and inter-frequency measurements for each HO. We observe a trade-off between cache size and energy efficiency in the network, i.e., as the cache size increases, the effective average rate experienced by the MT degrades because of the increase in energy consumption to cache data.
4) We obtain a range of play rates for the MT, such that local caching at the MT provides better performance than the conventional HO schemes. We provide recommendations for choosing minimum cache memory size at the MT in order to maximize the effective average rate experienced by the MT.
5) Finally, we employ frequency reuse in the network and compare the coordinated and uncoordinated transmission schemes. We observe coordinated scheme performs better than the uncoordinated scheme. We obtain an optimal frequency reuse factor for both schemes. The interference-dependent network offers the same performance as a noise-limited network in terms of the average rate the user experiences for a given set of parameters.

The rest of the paper is organized as follows. In Section II, we introduce our network model and outline the study objectives. In Section III, we discuss the distribution of interference power experienced by the user at a certain time. In Section IV, we derive the analytical expressions for the average rate for conventional and cache-based HO schemes. In Section V, we discuss the numerical results, and finally, the paper concludes in Section VI and future directions in Section VII.

II. System Model

Table I provides the notation used in the paper.

A. Network Model

We consider a street consisting of several SBSs, whose locations are distributed according to a homogeneous PPP Φ on $\mathbb{R}$ with intensity λ. Each SBS has a reliable backhaul connection to the core network. We assume an MT whose movement is uni-directional, i.e., an MT moves on a randomly oriented straight line with velocity v [23]. Due to the stationarity of the PPP [24], we assume the MT moves only forward in a straight line along the x-axis. Without loss of generality, we perform a downlink analysis from the perspective of an MT moving in a straight line along the x-axis, passing through the origin. We consider at a given time $t = 0$, the MT is served by an SBS located at x_0, where $x_0 \in \Phi$. This SBS is referred to as reference SBS. Without loss of generality, we assume the MT is located at the origin at that given time [24]. Averaging over the PPP Φ, this MT is referred to as typical MT.

The typical MT is associated with a SBS based on the maximum received signal strength indication (RSSI) association scheme, i.e., the MT will get associated with the SBS from which it receives maximum received power. Assuming that the typical MT is located at the origin o at time $t = 0$, therefore, at $t = t_1$, the typical MT is located at $u(t_1) = (vt_1, 0)$, where v is the velocity of MT. We have considered the large-scale path loss model. The small-scale fading is not considered in our work because of the spatial and temporal dynamics in the network and is analytically intractable. We assume when at time t, the received power experienced by the user is represented as $PK|x_0 - u(t)|^{-\alpha}$, where x_0 is the location of the reference SBS the MT is associated to, P is the downlink transmit power, and α is the path loss exponent. K is the path-loss coefficient given by $K=(\frac{\lambda_c}{4\pi})^2$, where λ_c is the carrier wavelength.

B. Cache-enabled Handover Management Scheme

Video streaming is the continuous transmission of video files from a server to a client, which requires a stringent QoS requirement. Frequent handovers will affect the QoS demands of such services. In a conventional HO strategy, the MT does a cell search periodically after every t_{S} seconds [21], measures the highest reference signal received power (RSRP) from the SBSs and initiates a HO to the best serving SBS. As a result, there is a time overhead corresponding for each HO procedure, represented as t_{H} [22]. In order to mitigate this delay, we describe a cache-based HO management scheme, which enhances the handover decision at the typical MT that requests infotainment-related contents. Leveraging the caching capabilities of MT, unnecessary HOs are skipped, which assists in maintaining a high-end QoE in terms of the effective average rate.

A detailed illustration of the management architecture of the cache-based HO is given in Fig. 1. From Fig. 1, we have illustrated a directional MT moving in a straight line with a constant velocity v [23]. We have assumed that the MT is

TABLE I
NOTATION

Symbol	Definition	Symbol	Definition
λ	SBS Intensity	f_c	Carrier frequency
B	Bandwidth	N_0	Noise Power
P	Transmit Power	α	Path-loss exponent
v	MT velocity	p	Play rate
K	Path loss coefficient	I	Total Interference power
r	Caching Radius	t_{C}	Time for which HOs are skipped
d_{C}	Distance for which HOs are skipped	T	Total time frame
D	Distance between t=0 and next SBS associated	R_{co}	Average rate (conventional)
R_{ca}	Average rate (proposed)	d_{L}	Distance travelled with low-play rate
N_{co}	No.of HOs skipped (conventional)	Δ	Frequency Reuse factor
G	Cache Size	μ_{co}	HO rate (conventional)
μ_{ca}	HO rate (proposed)	R_{eco}	Effective Average rate (conventional)
R_{eca}	Effective Average rate (proposed)	t_H	HO Time overhead
N_{ca}	No.of HOs skipped (proposed)	t_{F}	Time required to fill the cache
t_{R}	Time required to cross caching region	D_{C}	Total data cached
P_{C}	Power for caching	w_c	Power efficiency of caching
P_{I}	Power for inter-frequency measurements	P_{TC}	Total power consumed

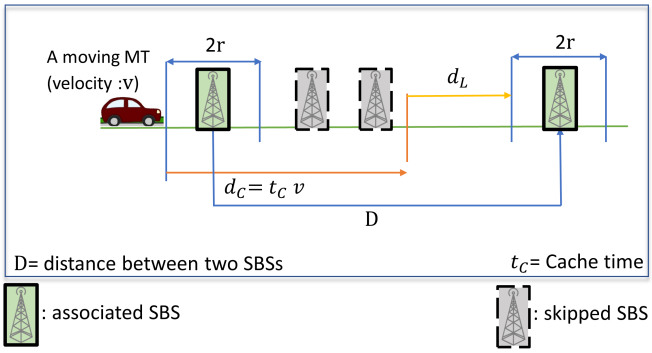


Fig. 1. Illustration of typical MT moving with a velocity v

playing the video with a fixed play rate p. On that premise, we define a caching region of $2r$, referred to as C_{R}, across each SBS, allowing the MT to cache a portion of the video segments. C_{R} is twice the distance of the MT from the SBS, where it experiences a download rate equal to the play rate p. Hence, C_{R} distance depends on transmit power of SBS, noise power, path loss coefficient, the sum of interference powers from the other SBSs, bandwidth, and path loss exponent. We are referring to r as the caching radius, which is evaluated as

$$r = \left(\frac{(N_0 + I)(2^{(p/B)} - 1)}{PK} \right)^{-1/\alpha}, \tag{1}$$

p is the play rate. I is the total interference power which is the sum of all the powers received by the MT from the SBS other than the reference SBS when the MT is located at a distance of r from the reference SBS. Without loss of generality, we assume at time $t = 0$, the MT is at a distance of r from the reference SBS at x_0.

The time travelled by typical MT by skipping the HOs and playing with a play rate p is referred to as cache time, t_{C}. When MT is outside C_{R} where the download rate is less than the play rate, the MT makes use of the data cached inside C_{R}. The MT moves forward, playing with play rate p utilizing the cached data, maintaining the connection with the associated SBS, and skips the consecutive HOs and makes the next HO when the cache data is fully exhausted. The distance traveled by the MT without initiating a HO until exhausting the cached data is referred to as cache distance as given in Fig. 1, expressed as $d_{\mathrm{C}} = t_{\mathrm{C}} v$.

From Fig. 1, we can observe the distance d_{C} starts from the caching region of associated SBS, which is at a distance of r from x_0 to the point where the MT exhausts all the cached data. The MT checks at the caching region of every SBS it passes through, that if there is enough cached data left at the MT to cover the caching region of that SBS. If there is no enough cached data left, it makes the HO. We assume the next associated SBS is at a distance of D from the associated SBS at x_0. Therefore, d_{C} takes values less than or equal to D, where D depends on the no.of HOs skipped. If N HOs are skipped for a distance of d_{C}, D is the distance of the $(N+1)^{th}$ SBS from the associated one. The number of HOs skipped for a distance of d_{C} is Poisson distributed, with parameter λ, and its average value is given by λd_{C}. After moving for a distance of d_{C}, the MT moves for a distance of d_{L} before entering the caching region of the next associated SBS at a distance of D, if $d_{\mathrm{C}} < D$. Throughout this distance of d_{L}, MT plays the video at a rate equal to the download rate experienced by the MT from the associated SBS. So, the distance d_{L} is referred to as low play rate region or L_{R}. Fig. 2 depicts the flowchart for the proposed cache based HO management scheme.

In this regard, we characterize the average rate, which is the expected rate experienced by the MT throughout its journey. We consider the distribution of total interference power to characterize the effect of interference on the average rate. The averaging of the rate is performed at a temporal and spatial level. First, as the MT moves, a temporal averaging of rate is done for a given spatial realization to cover the two neighboring distances. Second, spatial averaging is performed to obtain the expected rate experienced by the MT throughout.

Mathematically, let R' be the rate experienced by the MT at time t taking into account the sum of the powers from all the interfering SBSs in PPP at time t, represented as I. The pdf of the total interference power experienced by the MT at time t is $f_I(y,t)$. The distance between two associated SBSs is L. Therefore, the average rate experienced by the MT to move for a distance of L is given as

$$\bar{R} = \int_0^{\infty} \int_0^{L/v} R'(t,y) f_I(y,t) \mathrm{d}t \; \mathrm{d}y \tag{2}$$

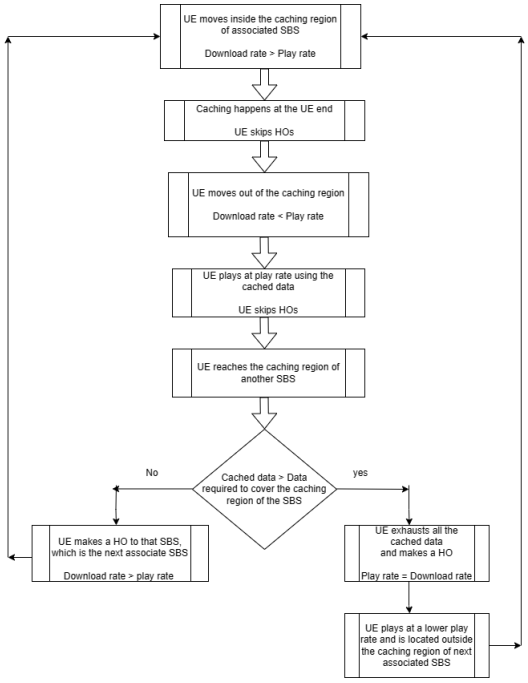


Fig. 2. Flowchart for the proposed cache-based HO management scheme

C. Frequency Reuse

We study the impact of frequency reuse on the average rate in our work. Frequency reuse is an inter-cell interference coordination technique. Each cell is allocated with resources so that interference in the network is minimized, which improves the network's performance. In our model, we consider a decentralized operation of the SBS, i.e., the frequency band of operation for each SBS is chosen probabilistically [25]. We assume that each SBS transmits in a frequency sub-band randomly selected from the range of frequency indices $\{1,2,...\Delta\}$ [25]. As the value of Δ increases, the interference in the network decreases. We assume the SBS's operations are not coordinated, and each SBS independently and randomly chooses the frequency band with a probability of $\frac{1}{\Delta}$. We have compared the probabilistic scheme with the coordinated scheme, where each SBS shares the channel with its $(\Delta+1)^{th}$ neighboring SBS and operates on the corresponding frequency. However, analytical expressions are intractable as there is dependent thinning of SBSs, i.e., the $(\Delta+1)^{th}$ SBSs are thinned.

D. Power consumption model

In our work, the total power consumption in the network consists of power consumption for caching, power consumption for frequent inter-frequency handovers, and transmit power required at the SBSs [26].

1) Power consumption for caching: We consider a power-proportional model [27]. The total power consumed for caching depends on the amount of data being cached at the MT and the caching power efficiency w_{c}. Caching power efficiency is the power consumed for caching one bit of data. It depends on the caching technology employed at MT. In this work, we consider a high-speed solid state disk (SSD) will be employed for storing the cached data, whose power efficiency is given as $6.25\text{x}10^{-12}$ watt/bit [28]. Therefore, the total power consumed for caching is given as $P_{\mathrm{C}} = Gw_{\mathrm{c}}$, where G is the cache size and w_c is the power efficiency of caching.

2) Power consumption for Inter-frequency Measurements: For every HO happening in the network, a certain amount of energy is consumed by the MT for inter-frequency measurements. The amount of energy consumed for one inter-frequency measurement is, E_{I}=3mJ [29]. Therefore, the total power consumed for inter-frequency measurements is given as $P_{\mathrm{I}} = HE_{\mathrm{I}}$, where H is the expected number of HOs initiated by the MT per unit time.

Hence, the total power consumed in the network is given as $P_{\mathrm{TC}} = P + P_{\mathrm{I}} + P_{\mathrm{C}}$.

III. Distribution of total interference power

Lemma 1. *The probability density function (pdf) of the total interference power experienced by the MT at time t, in a network across realizations, where the locations of SBSs are modeled as a 1-D PPP is given as*

$$f_I(y,t) = \frac{1}{2\pi}\int_0^{\infty} \exp\big(-j\omega y\big)\psi_I(\omega,t)\mathrm{d}\omega, \tag{3}$$

where,

$$\psi_I(\omega,t) = \exp\left(-\frac{\lambda}{\Delta}\int_r^{\infty} 1-\exp\Big(j\omega PKg^{-\alpha}\Big)\mathrm{d}g\right). \tag{4}$$

where $\psi_I(\omega,t)$ is the characteristic function of the interference power at time t. The distance of the interfering SBSs from the MT at time t is represented as g.

Proof. We assume that all the SBSs are operating in the same frequency. Therefore, the typical MT experiences interference from all the other SBSs in the network, except from which it is associated. As we assume the MT moves in the positive x-axis, the location of the MT at time t is denoted as $\mathbf{u}(t)$= $(vt,0)$.

The total interference power experienced by the MT at time t is given as

$$I(t) = \sum_{j=1}^{\infty} PKa_j(t)^{-\alpha}, \tag{5}$$

where

$$a_j(t) = |x_j - u(t)| \quad \forall x_j \in \Phi\backslash x_0, \tag{6}$$

where $\mathbf{x}_j$ is the location of the j^{th} interfering SBS and a_j is the distance between the j^{th} interfering SBS and the MT. We assume all the SBSs are operating with the same transmit power P.

To find the pdf of total interference power experienced by the MT at time t, we derive the characteristic function, which is given as

$$\psi_I(\omega,t) = \mathbb{E}\Big[\exp\Big(j\omega I(t)\Big)\Big]. \tag{7}$$

Substituting (5) in (7),

$$\psi_I(\omega,t) = \mathbb{E}\Big[\exp\Big(j\omega\sum_{j=1}^{\infty} PKa_j(t)^{-\alpha}\Big)\Big] \quad (8)$$

$$= \mathbb{E}\Big[\prod_{j=1}^{\infty}\exp\Big(j\omega PKa_j(t)^{-\alpha}\Big)\Big] \quad (9)$$

Calculating the probability generating functional (PGFL) over the PPP ϕ in (9), we obtain the characteristic function in (4), where

$$g = a_j(t) = |x_j - u(t)| \quad \forall x_j \in \Phi\backslash x_0 \quad (10)$$

which is given in (6).

Finally, applying the Gil-Pelaez inversion theorem [30], the pdf of total interference power I experienced by the MT at time t is given in (3). □

1) Assumption: Here we can see that the distance of MT to the first interfering SBS can take values $r \leq d_1 < \infty$. Therefore, for the calculation of the pdf of interference power, there is an impact of the caching radius r, which itself is a function of interference power I as seen in (1). Hence, for the analytical calculations, the impact of interference power on the caching radius r is neglected. This assumption is accurate for low interference scenarios, i.e., when the intensity of base stations is low. Also, for a lower play rate, the assumption is accurate because, for higher play rates, the caching radius across the SBS will be too large. Therefore, the impact of interference will reduce this caching radius but not as much to impact the HO rate and average experienced by the MT. Therefore, the caching radius r is written as

$$r = \left(\frac{N_0(2^{(p/B)}-1)}{PK}\right)^{-1/\alpha}, \quad (11)$$

Fig. 3 shows the CDF of the total interference power versus the power values in watts at a time instant for different values of Δ. When Δ=1, the MT experiences interference from all the SBSs in the point process Φ with intensity λ. When Δ increases, the interference power experienced by the MT decreases because the MT experiences interference from the SBSs in the point process Φ with the intensity of $\frac{\lambda}{\Delta}$. The intensity of SBSs for the point process Φ is thinned by a factor of $\frac{1}{\Delta}$.

IV. Characterization of Average Rate

In the following sections, we derive the expressions for the average rate for a vehicular network with conventional HO strategy and cache-enabled HO strategy.

A. Conventional HO scheme

In a conventional HO scheme, the HO happens every time MT crosses a cell boundary. The MT associates with the SBS from which it receives maximum power. We derive the expression of the average rate for a vehicular network for a cell-boundary HO scheme.

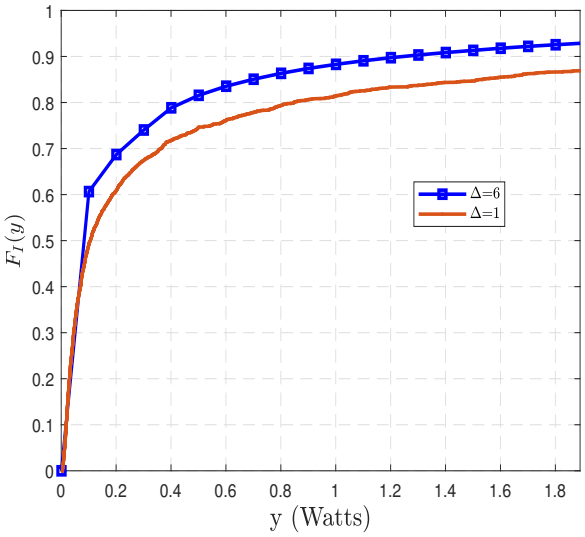


Fig. 3. CDF of total interference power

Lemma 2. *The average rate R_{co} experienced by the typical MT in a conventional HO scheme is given as*

$$R_{\text{co}} = \int_0^{\infty}\int_0^{\infty} R_0(r_1,r_2) f_{l_1}(r_1) f_{l_2}(r_2)\mathrm{d}r_1\mathrm{d}r_2, \quad (12)$$

where

$$R_0(l_1,l_2) = \int_0^{\infty}\frac{2v}{(l_1+l_2)} \int_0^{\frac{(l_1+l_2)}{2v}} B\log_2\left(1+\frac{PK/(N_0+L(t,l_1,l_2)+b)}{|(\frac{l_1}{2}-vt)^{\alpha}|}\right) f_S(t,l_1,l_2)\ \mathrm{d}t\ \mathrm{d}b. \quad (13)$$

where $b = S(t,l_1,l_2)$ i.e., the total interference power experienced by the MT from the SBSs, conditioned on l_1 and l_2, at time t.

The pdf of the total interference power $S(t,l_1,l_2)$, represented as $f_S(t,l_1,l_2)$, is given as

$$f_S(t,l_1,l_2) = \frac{1}{2\pi}\int_0^{\infty}\exp\big(-j\omega b\big)C_S(\omega,t,l_1,l_2)\mathrm{d}\omega \quad (14)$$

where $C_S(\omega,t,l_1,l_2)$ is the characteristic function given as

$$C_S(\omega,t,l_1,l_2) = \exp\left(-\frac{\lambda}{\Delta}\int_0^{\infty} 1-\exp\Big(j\omega PK(a+l_1/2+vt)^{-\alpha}\Big)\mathrm{d}a\right)\cdot \exp\left(-\frac{\lambda}{\Delta}\int_0^{\infty} 1-\exp\Big(j\omega PK(a+l_2+l_1/2-vt)^{-\alpha}\Big)\mathrm{d}a\right)$$

N_0 is the noise power, B is the total bandwidth and v is the velocity.

Proof. See Appendix A. □

B. Cache-enabled HO scheme

In this section, we derive the analytical expressions for cache time t_{C} and average rate R_{ca} for cache-enabled vehicular network, with the cache size of G bits.

When the MT with limited local cache enters C_{R} and starts caching the data, the time at which the MT reaches its maximum caching capacity is denoted by t_{F}. As referred before, t_{C} is the time MT travels without initiating the HOs, and making use of this cached data outside C_{R}. Let t_{R} be the time MT takes to move through the entire caching region C_{R} with a velocity v. Therefore, t_{R}=$\frac{2r}{v}$.

The time at which the cache is full, i.e., the amount of bits cached is equal to G bits, is referred to as t_{F}. The amount of data bits cached is the difference between downloaded and played data bits. Therefore, it can be derived numerically as a solution to the following equation.

$$\int_0^{t_{\mathrm{F}}} \left[B\log_2\left(1+\frac{PK}{(N_0+I)(|r-vt|)^{\alpha}}\right) - p \right] dt = G. \quad (15)$$

Hence, the cache time t_{C} for the cache-enabled vehicular network with limited cache is evaluated for two conditions such as:

1) $t_{\mathrm{F}} < t_{\mathrm{R}}$
   The cache memory reached its maximum caching capacity before the MT crosses the entire caching region C_{R}.
2) $t_{\mathrm{F}} \geq t_{\mathrm{R}}$.
   There is enough cache memory left even after the MT crosses the entire caching region C_{R}.

Therefore, the cache time t_{C} for a cache-enabled vehicular network with a limited cache size of G bits at MT is given as the solution to the given equation.

$$\begin{array}{ll} G + \int_{t_{\mathrm{R}}}^{t_{\mathrm{C}}} B\log_2\left(1+\frac{PK(|r-vt|)^{-\alpha}}{(N_0+I(t))}\right)\mathrm{d}t & \\ \quad = p(t_{\mathrm{C}}-t_{\mathrm{R}}) & t_{\mathrm{F}} < t_{\mathrm{R}}. \\ \int_0^{t_{\mathrm{C}}} B\log_2\left(1+\frac{PK(|r-vt|)^{-\alpha}}{(N_0+I(t))}\right)\mathrm{d}t & \\ \quad = pt_{\mathrm{C}} & t_{\mathrm{F}} \geq t_{\mathrm{R}}. \end{array} \quad (16)$$

$I(t)$ is the sum of the received powers from the interferes experienced by the MT at time t for a given spatial realization. It is challenging to solve a closed-form expression for t_{C}. Hence, using numerical integration, we solve for the cache time t_{C} in (16).

For the first condition $t_{\mathrm{F}} < t_{\mathrm{R}}$, we find the time at which the difference between played data bits and downloaded data bits is equal to G. And for $t_{\mathrm{F}} \geq t_{\mathrm{R}}$, we find the time at which the difference between downloaded data bits and played data bits is equal to zero. We know, the next associated SBS is at a distance of D from the associated SBS. Then, cache time t_{C} is given as

$$t_{\mathrm{C}} = \min\left\{t_{\mathrm{C}}, \frac{D}{v}\right\} \quad (17)$$

From (16), we can see that t_{C} is a function of interference, which is represented as $t_{\mathrm{C}}(I)$ and is spatially averaged later. The cache distance, $d_{\mathrm{C}}(I) = t_{\mathrm{C}}(I)v$ is the distance at which the typical MT skips the unnecessary HOs. Therefore, $d_{\mathrm{C}}(I)$ can be less than or equal to D.

Special Case: For unlimited cache size, i.e G= ∞, the time at which MT reaches its maximum capacity G is t_{F}, which in this case is infinite. Therefore, t_{F} is always greater than t_{R}. Therefore, using second condition, t_{C} is given as

$$\int_0^{t_{\mathrm{C}}} B\log_2\left(1+\frac{PK(|r-vt|)^{-\alpha}}{(N_0+I(t))}\right)\mathrm{d}t = pt_{\mathrm{C}}. \quad (18)$$

To derive the average rate experienced by the MT for a cache-enabled HO scheme, we divide the distance traveled by the MT into two: caching distance $d_{\mathrm{C}} = vt_{\mathrm{C}}$ and the distance traveled by the MT after the caching distance, before entering the caching region of the next associated SBS. Throughout the caching distance d_{C}, the MT plays at the required play rate. Therefore, the average rate experienced by the MT throughout the caching distance is the required play rate. After the caching distance before entering the caching region of the next associated SBS, the MT travels for a distance of d_{L}. The value of d_{L} depends on the location of the next associated SBS, D, which is random. Therefore, d_{L} is also random, which is given as

$$d_{\mathrm{L}} = \begin{cases} D - d_{\mathrm{C}}(I) & d_{\mathrm{C}} < D \\ 0 & d_{\mathrm{C}} = D. \end{cases} \quad (19)$$

The CCDF of D is given in (37). Hence, we derive the pdf of d_{L}, where the MT plays at play rate lower than p. The distance d_{L} is conditioned on interference power I for a given realization represented as $d_{\mathrm{L}}(I)$.

Lemma 3. *The pdf of d_{L} conditioned on the interference power I is given by*

$$f_{d_{\mathrm{L}}|I}(z|y) = \sum_{n=0}^{\infty} \exp\left(-2\frac{\lambda}{\Delta}(z+d_{\mathrm{C}}(y))\right) \frac{(2\frac{\lambda}{\Delta}(z+d_{\mathrm{C}}(y)))^{(n+1)}}{(z+d_{\mathrm{C}}(y))\Gamma(n+1)} \frac{\exp\left(-\frac{\lambda}{\Delta}d_C(y)\right)(\frac{\lambda}{\Delta}d_C(y))^n}{n!}, \quad (20)$$

where n is assumed to be the number of HOs skipped for a distance of d_{C}, and it follows the distribution of the number of points in a PPP [13].

Proof: See Appendix B. ∎

Using the pdf of d_{L}, we derive the expression for average rate of a cache-enabled cellular network.

Lemma 4. *The average rate of a cache-enabled vehicular network is given as*

$$R_{\mathrm{ca}} = \int_0^{\infty} R_2(y) f_I(y) \mathrm{d}y, \quad (21)$$

where

$$R_2(y) \approx \int_0^{\infty} \frac{v}{(d_{\mathrm{C}}(y)+Z)} \left[\frac{p d_{\mathrm{C}}(y)}{v} + \int_0^{Z/v} B\log_2(1+\frac{PK/(N_0+y)}{(Z+r-vt)^{\alpha}} f_I(y,t)\mathrm{d}t \right] f_{d_{\mathrm{L}}}(Z)\mathrm{d}Z. \quad (22)$$

where $Z = d_{\mathrm{L}}(y)$ is a function of the interference power y, experienced by the user at time t, $f_I(y,t)$ is the pdf of the total interference power at a given time t, p is the play rate and $f_{d_{\mathrm{L}}}(z)$ is derived in (20).

TABLE II
SIMULATION PARAMETERS

Notation	Parameter	Value
P	Transmit power	35 dBm [31]
λ	Intensity of SBSs	1 Km^{-1}
B	Bandwidth of the system	100 MHz
α	Path-loss Exponent	2
f_c	Carrier frequency	3.5 Ghz
σ^2	Noise density	-174 dBm/Hz [22].
p	Play rate	2 Gbps.
t_H	Handover time overhead	43 ms [22].
t_S	Search time	20 ms [22].

Proof. See Appendix C. □

(21) is solved using numerical integration by taking the spatial average over the interference term.

For the noise-limited case, the interference power is zero. Hence, the interference term in (22) is equated to zero to obtain the average rate for the noise-limited cache-based HO scheme.

In this section, we discuss the number of HOs initiated by the MT for conventional and cache-enabled HO schemes.

Lemma 5. *For a conventional HO scheme, expected number of HOs initiated by MT for 1D network is given as*

$$N_{\text{co}} = \max\{\lambda v T, 1\} - 1, \tag{23}$$

where T is the total time travelled by the MT and v is velocity.

Lemma 6. *For a cache-enabled HO scheme, the expected number of HOs initiated by the MT is given as*

$$N_{\text{ca}} = N_{\text{co}} - \lambda(\max\{d_{\text{C}}, r\} - r), \tag{24}$$

where r is half of caching region distance of the SBS which MT is associated to.

Proof: See Appendix D. ■

Next, we derive the effective average rate considering the effect of HO rate and the time overhead for each HO on the average rate experienced by the MT.

The time overhead t_H for a SBS handover is 43ms [22].

Lemma 7. *The effective average rate of a conventional HO scheme is given as [22].*

$$R_{\text{eco}} = R_{\text{co}}(1 - \mu_{\text{co}} t_H)^+, \tag{25}$$

where $(1 - \mu_{\text{co}} t_H)^+ = \max(0, (1 - \mu_{\text{co}} t_H))$, μ_{co} is the HO rate for conventional HO scheme given as $\mu_{\text{co}} = \frac{N_{\text{co}}}{T}$ and R_{co} is given in (12).

Lemma 8. *The effective average rate of a cache-enabled HO scheme is given as*

$$R_{\text{eca}} = R_{\text{ca}}(1 - \mu_{\text{ca}} t_H)^+. \tag{26}$$

μ_{ca} is the HO rate for cache-based HO scheme given as $\mu_{\text{ca}} = \frac{N_{\text{ca}}}{T}$ and R_{ca} is given in (21).

V. NUMERICAL RESULTS AND DISCUSSION

In this section, we validate our analytical framework using Monte-Carlo simulations, providing a precise analysis equivalent to executing experiments, and present some numerical results to discuss the salient features of the network. The conventional scheme is taken as the benchmark here. The simulation parameters are shown in Table II.

We have evaluated the effective average throughput and energy efficiency and their variations with the intensity of SBSs, the velocity of MT, and the play rate. We assume the conventional scheme as the benchmark for comparing our proposed model.

A. Handover rate

Fig. 4 shows that the analytical result on the HO rate closely matches the Monte-Carlo simulations. We observe that as the intensity of SBSs increases, HO rate increases. For lower intensity values, the expected number of handovers or the handover rate for the cache-enabled scheme and the conventional scheme is the same. However, for higher values of intensity, the cache-enabled scheme experiences less HOs compared to the benchmark scheme because unnecessary HOs are skipped in the proposed model.

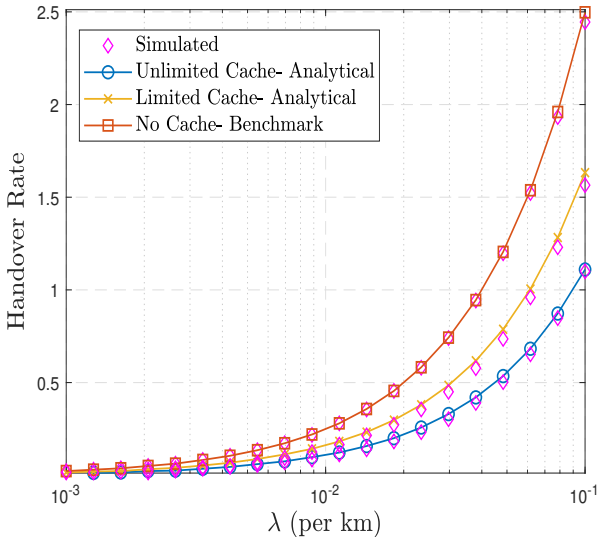


Fig. 4. Handover rate versus Intensity of SBSs

B. Effective Average Rate

In Fig. 5 and Fig. 6, we plot the effective average rate versus intensity of SBSs for interference-dependent and noise-limited scenarios, respectively. In Fig. 5, we plot the derived analytical expressions for the average rate considering the distance distribution of the total interference power at a given time instant. It is considered the upper bound, as the effect of interference on the caching region r is not taken into account. Here, we show that the analytical result on the effective average rate closely matches the Monte-Carlo simulations.

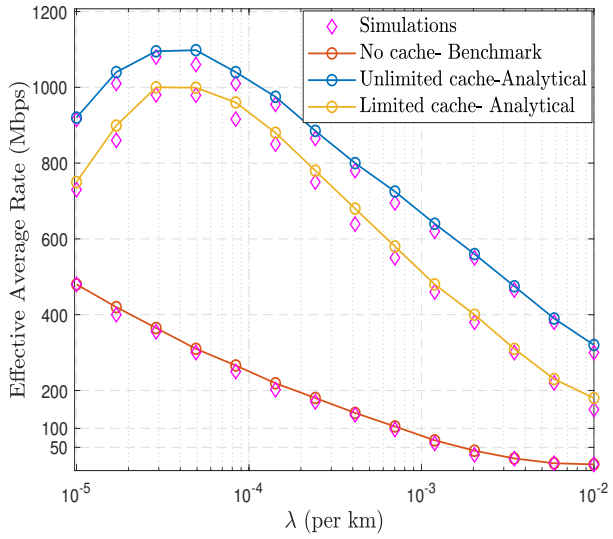


Fig. 5. Effective average rate versus Intensity of SBSs for interference-dependent network for v=25 m/s

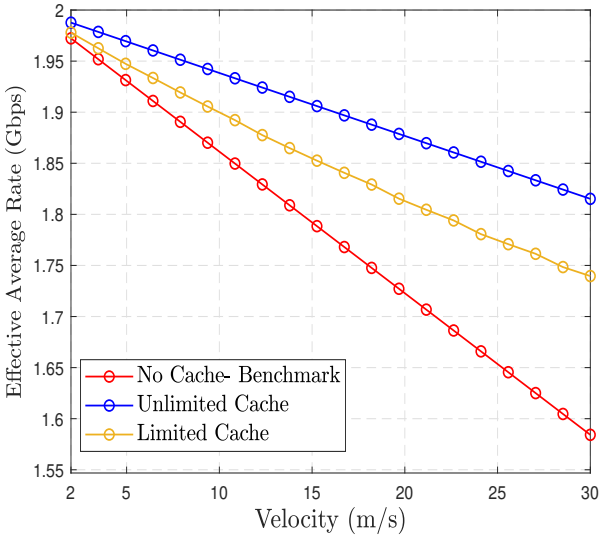


Fig. 7. Effective average rate versus velocity of MT for $\lambda = 10^{-3}$

In both Fig. 5 and Fig. 6, we observe for lower values of intensity, the effective average rate increases, and the intensity of SBSs increases. This is because there is less number of HOs for lower intensity values. After a particular value of intensity (defined as the optimal value), the effective average rate starts to decrease because of the increase in the number of HOs, leading to an increase in HO time overhead. This would help the network operator decide on the optimal deployment densities to obtain the maximum average rate for different cache sizes of the proposed scheme.

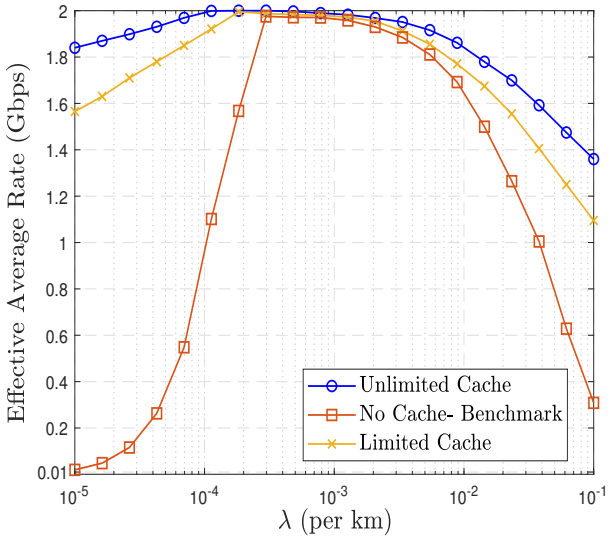


Fig. 6. Effective average rate versus intensity of SBS for noise-limited network for v=25 m/s

In Fig. 7, we plot the effective average rate versus velocity of typical MT for the noise-limited network. As velocity increases, the effective average rate decreases because of the increase in HO rate with velocity. We observe that, for pedestrian users, e.g., the velocity of 5m/s, we obtain a high value of average rate compared to vehicular users, e.g., the velocity of 25 m/s, for a specific set of parameters. We also observe that skipping unnecessary HOs through the cache-enabled scheme is expendable for pedestrian users. This is because for a particular value of intensity and the lower values of velocity, the number of HOs skipped will be significantly less, indicating the performance is the same as that of the conventional scheme. However, for higher velocity values, the cache-enabled scheme performs better than the conventional scheme because of skipping more HOs.

C. Maximum Effective Average Rate for Minimum Cache Size

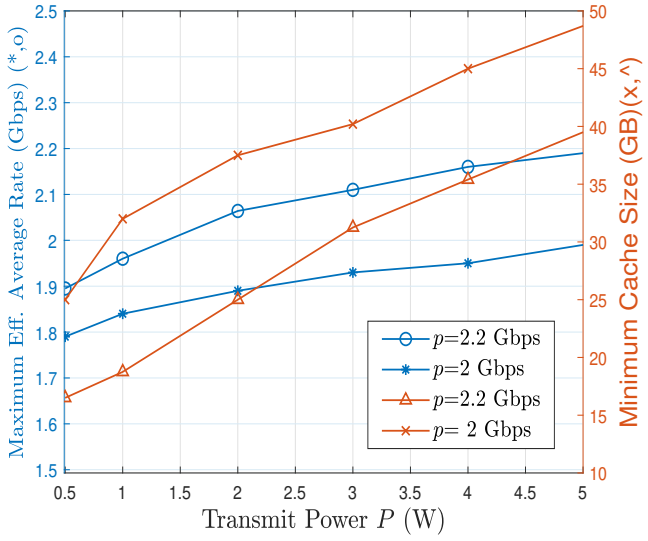


Fig. 8. Maximum average rate for minimum cache size versus transmit power P for the noise-limited network.

Fig. 8 and Fig. 9 show the minimum cache size required for obtaining the maximum effective average rate for different values of play rate p for a noise-limited network and interference-dependent network, respectively. We observe that as the play rate increases, the cache size needed to obtain the maximum average rate decreases. This is because if the play rate is high, the amount of data stored in the cache reduces even though it experiences a high average rate. For the network operator, this reveals how our framework helps to decide the cache limit and transmit power of the SBSs for the MT to play the videos at a particular play rate. For instance, if the MT has a cache size of 32GB and plays the video at the rate of 2.2 Gbps, the maximum average rate experienced by the MT will be 2.1 Gbps for a downlink transmit power of 3 W.

In an interference-dependent network, we observe an increase in the cache size in order to include the effect of interference given in Fig. 9. So, the minimum cache size for the network has increased to obtain the maximum average rate. This is because of a decreased downloaded data at the MT, which requires more cache to obtain the maximum rate.

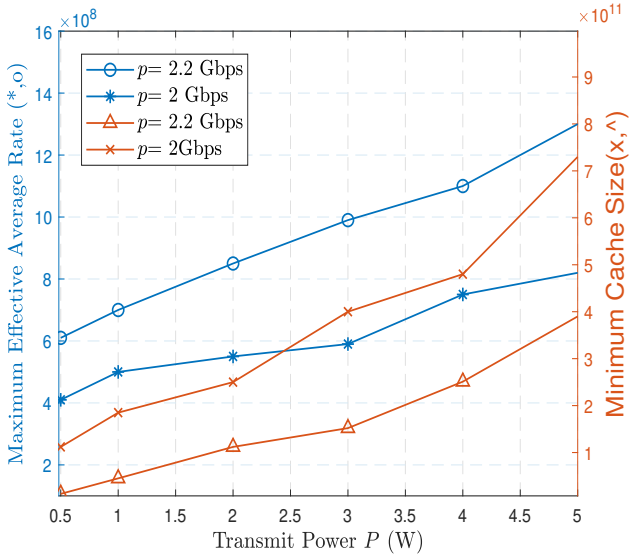


Fig. 9. Maximum effective average rate for minimum cache size versus transmit power P for λ=10^{-3} and v=25 for interference-dependent network.

D. Frequency Reuse

In Fig. 10, we plot the effective average rate versus the frequency reuse factor Δ for the probabilistic frequency reuse scheme. Here, we observe that as Δ increases, the effective average rate increases. For an unlimited cache scenario, any values of Δ greater than 14 give the same performance as the noise-limited network for a fixed play rate. While for limited cache, the Δ value is 22, and for the benchmark scheme, it is 38. We obtain an optimal Δ for different cache sizes at the MT.

In Fig. 11, we plot the effective average rate versus Δ for a coordinated operation between the SBS. We observe that there is a significant decrease in the optimal value of Δ because of the reduction in the interference power experienced by the MT in the network. We conclude that coordinated transmission reduces interference in the network. However, because of analytical intractability, we implement the probabilistic frequency reuse scheme. This provides valuable insights to the network operator to determine the minimum number of frequency bands to be allocated to the SBSs.

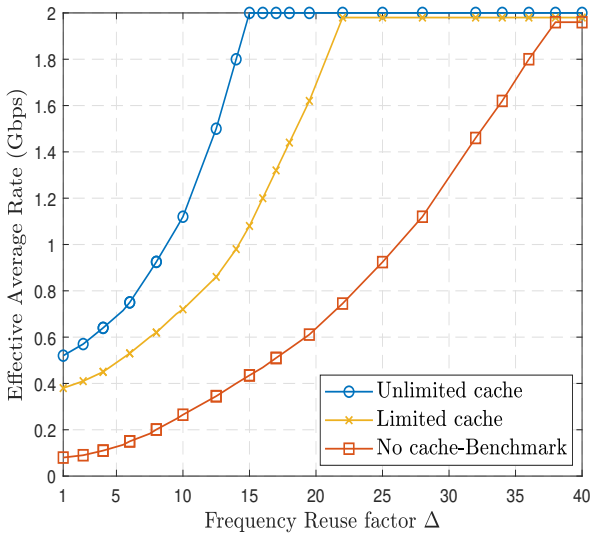


Fig. 10. Effective average rate versus Frequency reuse factor Δ for λ=10^{-3} and v=25 for probabilistic scheme

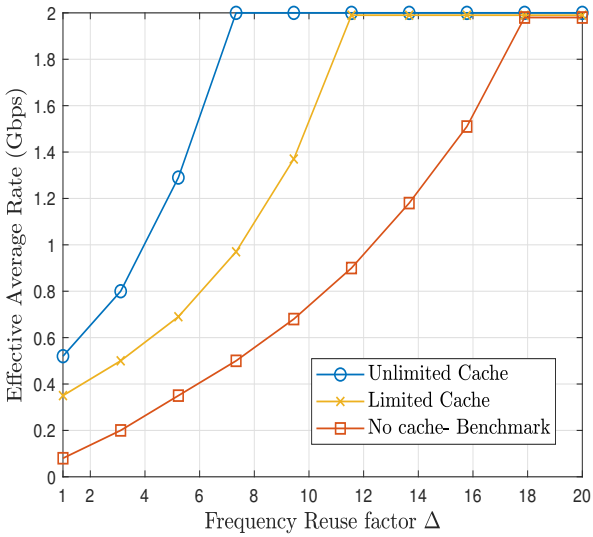


Fig. 11. Effective Average Rate versus Frequency reuse factor Δ for coordinated scheme

E. Energy Efficiency

For Fig. 12, we plot the energy efficiency of the network versus the play rate for the noise-limited network. We notice that until a particular value of play rate, the cache-enabled scheme outperforms the conventional scheme. For higher values of play rate, the cached data tends to be negligible,

leading the cache-enabled scheme to perform the same as the conventional scheme. However, an extra amount of energy is consumed for caching in the cache-enabled scheme, causing the energy efficiency of that scheme to perform poorly. We conclude from the plot that the cache-enabled scheme achieves better than the benchmark scheme for a play rate of less than 3 Gbps. For a network operator, any play rate of more than this can lead to performance degradation.

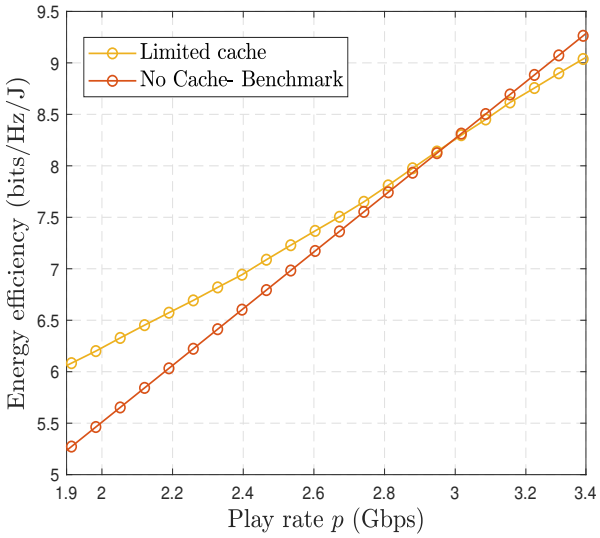


Fig. 12. Energy efficiency versus play rate for P=35 dBm,v=25 and λ=10^{-3} for noise-limited network

From Fig. 13, we observe a preferred play rate for an improvement in the performance of the proposed network in an interference-dependent scenario. Also, we observe that for lower values of cache size, there is a maximum range of play rate the MT can experience in the proposed method, giving an improved performance compared to the benchmark scheme. As cache size increases, the maximum range of play rate reduces. In fact, for $G = 400$ GB, the MT can play at the rate from 1.8 Gbps to 3.3 Gbps for improved performance in terms of energy efficiency by a factor of almost 20%. But, for $G = 800$ GB, the play rate range is reduced to 1.8 Gbps-3 Gbps for this improved performance.

F. System Design Insights

We outline the system design insights based on our results:

- Our analysis reveals to the network operator that beyond a certain limit, network densification is not a solution for maximizing the average rate experienced by the mobile terminal (MT). Increasing the intensity of SBS leads to increased handovers and interference in the network, degrading system performance.
- A linear relationship between the cache size and transmit power at the SBS is displayed, enabling network operators to tune their network configurations to meet specific QoS demands efficiently. In case, the device manufacturers have the flexibility to increase the cache limit at the MT, the network operators should consider an increase in transmit power at the SBSs so as to maintain the required quality of service (QoS) demanded by the MT and vice versa.
- By implementing the frequency reuse in the network, it is possible to determine the minimum number of shared frequency channels issued to the SBSs in order to maintain the required QoS for MTs. If the cache limit at the MT is provided, the network operator can determine the minimum required frequency bands to be allocated to the SBSs.
- When the device manufacturers have the flexibility to expand cache size, network operators can restrict the service demand at the MT, so that power consumption in the network is controlled. Moreover, beyond a specific play rate, the cache memory at the MT can be turned off in order to reduce the energy consumption due to caching while maintaining the required QoS for users.

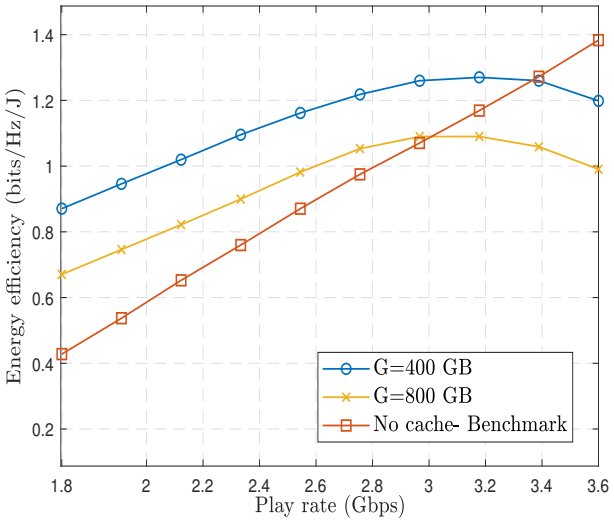


Fig. 13. Energy Efficiency versus play rate for $\lambda = 10^{-3}$, v=25 for interference-dependent network

VI. Conclusion

In this paper, we studied the performance variations experienced by the MT in a uni-directional street by leveraging cache capabilities for vehicular and pedestrian users. Streaming a video for a moving MT has severe QoS requirements and undergoes many handovers. Therefore, we propose a cache-enabled handover management scheme in 5G-enabled VNs, leveraging the caching capabilities at the MT, to improve the network performances, such as average rate and energy efficiency. We observe the proposed scheme significantly reduced the HO rate and improved the average rate of the network. The cache size at the MT terminal significantly impacts the overall performance of the proposed system, and the rate at which the MT plays the video affects the minimum cache size required at the MT. Accordingly, we prescribe an optimal deployment solution, such that the deployment density, transmit power, and minimum cache size can be

decided to obtain the maximum average rate for a fixed play rate. We studied the power consumption at the MT due to caching and inter-frequency measurements and prescribed a preferred range of play rate at which MT can play the data. We observe that the proposed scheme performs almost 20% better than the benchmark scheme in terms of the expected rate experienced by the MT. We studied the impact of interference on the proposed scheme and prescribed an optimal value for the frequency reuse factor so that the interference-dependent network performs the same as the noise-limited network for a given value of play rate of the MT.

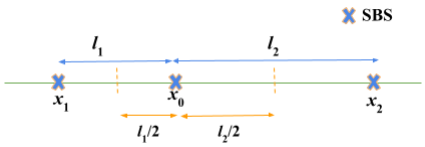
VII. Future Work

The study of HOs and mobility management in multi-tier UAV-enabled heterogeneous networks are interesting directions of research that we will address in the future. Also, using aerial re-configurable intelligent surfaces (RIS) mounted on UAVs and RIS-assisted edge computing to develop an efficient mobility management scheme is an exciting research direction that we will address in the future.

VIII. Acknowledgement

This research is supported by the IIT Palakkad Technology IHub Foundation Doctoral Fellowship IPTIF/HRD/DF/026.

Appendix A
Proof of Lemma 2



In the above figure, the first SBS located at x_1 is at a distance of l_1 to the left of MT, and the second SBS located at x_2 is at a distance of l_2 to the right of MT. At a given time, the location of the SBS to which the user is associated, i.e., reference SBS, is x_0. At $t = 0$, the MT is at a distance of $l_1/2$ to the left of the reference SBS. At that point, the MT makes a HO from first to reference SBS.

All the SBSs are distributed as a 1-D PPP, and we are conditioning on the distances of first and second SBS from the reference SBS at x_0, i.e., l_1 and l_2, where l_1 and l_2 takes random values.

The pdf of l_1, i.e., the distance between two SBSs in a 1D PPP, is given as

$$f_{l_1}(r) = \lambda \exp(-\lambda r). \tag{27}$$

Similarly, we can obtain the pdf of l_2.

In this regard, the total interference power experienced by the MT at time t is given as

$$M(t,l_1,l_2) = \underbrace{PK(l_1/2+vt)^{-\alpha} + PK(l_2+l_1/2-vt)^{-\alpha}}_{I} + \underbrace{\sum_{j=3}^{\infty} PKa_j(t,l_2,l_2)^{-\alpha}}_{II} \tag{28}$$

where a_j is the distance of the j^{th} interfering SBS other than x_1 and x_2 from the MT at time t, which is conditioned on l_1 and l_2. I consists of the interference powers from the first and second SBSs, represented as

$$L(t,l_1,l_2) = PK(l_1/2+vt)^{-\alpha} + PK(l_2+l_1/2-vt)^{-\alpha} \tag{29}$$

The total interference power from all the SBSs other than the SBSs at x_1 and x_2. It is represented as

$$S(t,l_1,l_2) = \sum_{j=3}^{\infty} PKa_j(t,l_2,l_2)^{-\alpha} \tag{30}$$

We evaluate the characteristic function of $S(t,l_1,l_2)$ which is given as

$$C_S(\omega,t,l_1,l_2) = \mathbb{E}\Big[\exp\big(j\omega \sum_{j=3}^{\infty} PKa_j(t,l_1,l_2)^{-\alpha}\big)\Big] \tag{31}$$

$$C_S(\omega,t,l_1,l_2) = \mathbb{E}\Big[\prod_{j=3}^{\infty} \exp\big(j\omega PKa_j(t,l_1,l_2)^{-\alpha}\big)\Big] \tag{32}$$

Applying PGFL,

$$C_S(\omega,t,l_1,l_2) = \exp\left(-\frac{\lambda}{\Delta}\int_0^{\infty} 1-\exp\Big(j\omega PK(a+l_1/2+vt)^{-\alpha}\Big)\mathrm{d}a\right)\cdot \exp\left(-\frac{\lambda}{\Delta}\int_0^{\infty} 1-\exp\Big(j\omega PK(a+l_2+l_1/2-vt)^{-\alpha}\Big)\mathrm{d}a\right)$$

where t varies from $0 \le t \le \frac{(l_1+l_2)}{2v}$

From the above equation, we can obtain the characteristic function $C_S(\omega,t,l_1,l_2)$.

Therefore, the pdf of the total interference power from the SBSs other than the first and the second SBSs is given as

$$f_S(t,l_1,l_2) = \frac{1}{2\pi}\int_0^{\infty} \exp\big(-j\omega b\big) C_S(\omega,t,l_1,l_2)\mathrm{d}\omega \tag{33}$$

The average rate experienced by the MT moving from the cell region of first SBS to second SBS, conditioned over l_1 and l_2, is given as

$$R_0(l_1,l_2) = \int_0^{\infty} \frac{2v}{(l_1+l_2)} \int_0^{\frac{(l_1+l_2)}{2v}} B\log_2\left(1+\frac{PK/(N_0+L(t,l_1,l_2)+b)}{|(\frac{l_1}{2}-vt)^{\alpha}|}\right) f_S(t,l_1,l_2)\,\mathrm{d}t\,\mathrm{d}b. \tag{34}$$

where $b = S(t, l_1, l_2)$, given in (30) i.e., the total interference power experienced by the MT from the SBSs, conditioned on l_1 and l_2, at time t.

Taking the expectation over l_1 and l_2, by multiplying the pdf given in (27), we obtain the average rate for conventional HO scheme in (12).

APPENDIX B
PROOF OF LEMMA 3

From Fig. 1, we observe that

$$d_L(I) = \begin{cases} D - d_{\mathrm{C}}(I) & d_{\mathrm{C}}(I) < D \\ 0 & d_{\mathrm{C}}(I) = D. \end{cases} \tag{35}$$

where D is the distance between the SBS to which the MT is associated, i.e., zeroth SBS and the next associated SBS, i.e., the SBS with which the MT will get to associate after exhausting the cached data by playing at the rate p.

The distance traveled by the MT by skipping the HOs until the cached data is exhausted referred to as d_{C}. Therefore,

$$d_{\mathrm{C}} \leq D. \tag{36}$$

Let N be the number of HOs skipped for a distance of d_{C}, where N can take random values 0,1,..∞ which is Poisson distributed. Then the MT gets associated to $(N+1)^{th}$ SBS, which is at a distance of D from the zeroth SBS. Therefore, D is the distance to the first SBS from the zeroth SBS when no HOs are skipped. Also, D can be the distance to the second SBS from the zeroth SBS when one HO is skipped, and so on.

The CCDF of D is the probability that there are less than $(N+1)$ nodes closer than b [32]

$$P_N := \mathbb{P}(0....N \text{ nodes within } b) = \sum_{k=0}^{N} \frac{(\frac{\lambda}{\Delta} 2b)^k}{k!} \exp\left(-\frac{\lambda}{\Delta} 2b\right). \tag{37}$$

where $2b$ is the standard Lesbegue measure or m-dimensional volume of the ball of radius b, where $b > 0$. We assume $m = 1$ since we consider a 1-D PPP.

Finding the pdf of d_{L},

$$\begin{aligned} F_{d_{\mathrm{L}}}(z) &= \mathbb{P}(d_{\mathrm{L}} \leq z) \\ &= \mathbb{P}(D \leq z + d_{\mathrm{C}}). \end{aligned}$$

The pdf of d_L, $f_{d_{\mathrm{L}}}(z)$= - $\frac{dP_N}{db}$ [32]

$$f_{d_{\mathrm{L}}}(z) = \frac{(N+1)}{(z+d_{\mathrm{C}})} \frac{(2\frac{\lambda}{\Delta}(z+d_{\mathrm{C}}))^{N+1}}{(N+1)!} \exp\left(-2\frac{\lambda}{\Delta}(z+d_{\mathrm{C}})\right). \tag{38}$$

Conditioning on the interference power, the above equation can be represented as (20).

APPENDIX C
PROOF OF LEMMA 6

We know that for a distance of d_C, the MT plays at the play rate. Further, for the distance d_L, it plays with a rate equal to the download rate, i.e., a lower play rate.

The rate experienced by the MT for the proposed scheme is given as

$$R_2(y) = \frac{v}{(d_{\mathrm{C}}(y) + d_{\mathrm{L}}(y))} \left[\frac{p d_{\mathrm{C}}(y)}{v} + \int_0^{d_{\mathrm{L}}/v} B \log_2\left(1 + \frac{PK/(N_0+y)}{(d_L(y) - vt)^{\alpha}} f_I(y,t) \mathrm{dt}\right) \right]. \tag{39}$$

Here, d_{L} is random. Therefore, taking the expectation over d_{L}, to obtain the average rate experienced by the MT in the network.

$$R_2(y) \approx \int_0^{\infty} \frac{v}{(d_{\mathrm{C}}(y) + Z)} \left[\frac{p d_{\mathrm{C}}(y)}{v} + \int_0^{Z/v} B \log_2\left(1 + \frac{PK/(N_0+y)}{(Z - vt)^{\alpha}} f_I(y,t) \mathrm{dt}\right] \mathrm{f}_{\mathrm{d_L}}(\mathrm{Z}) \mathrm{dZ}. \tag{40}$$

where $Z = d_{\mathrm{L}}(y)$, which is a function of interference I.

APPENDIX D
PROOF OF LEMMA 6

The distance traveled by MT by utilizing the cached data is cache distance d_{C}. From Fig.1, we observe after associating to a SBS, all the HOs after that are skipped. Therefore, the distance at which the HOs are skipped is $(d_{\mathrm{C}} - r)$.

The number of HOs skipped for a distance of $d_{\mathrm{C}} - r$ is given as $\lambda(d_{\mathrm{C}} - r)$. Therefore, the number HOs initiated by the MT for cache-based HO scheme is given in (24).

REFERENCES

[1] S. Garg, K. Kuljeet, K. Georges, A. Syed Hassan, and N. K. J. Dushantha, "SDN-based secure and privacy-preserving scheme for vehicular networks: A 5G perspective," *IEEE Transactions on Vehicular Technology 68*, pp. 8421–8434, 05 2019.

[2] J. Zhao, Q. Li, Y. Gong, and K. Zhang, "Computation offloading and resource allocation for cloud assisted mobile edge computing in vehicular networks," *IEEE Transactions on Vehicular Technology*, pp. 7944–7956, Aug 2019.

[3] J. Zhang, H. Guo, J. Liu, and Y. Zhang, "Task offloading in vehicular edge computing networks: A load-balancing solution," *IEEE Transactions on Vehicular Technology*, pp. 2092–2104, Feb 2020.

[4] R. Zhang, R. Lu, X. Cheng, N. Wang, and L. Yang, "A uav-enabled data dissemination protocol with proactive caching and file sharing in V2X networks," *IEEE Transactions on Communications*, pp. 3930–3942, June 2021.

[5] J. B. Kenney, "Dedicated short-range communications (dsrc) standards in the united states," *Proc. IEEE, vol. 99, no. 7*, 2011.

[6] Wang, Siming, and et al, "Low-latency caching with auction game in vehicular edge computing." *2017 IEEE/CIC International Conference on Communications in China (ICCC)*, 10 2017.

[7] T. Ismail, M. E.Gad, and B. Mokhtar, "Integrated VLC/RF wireless technologies for reliable content caching system in vehicular networks," *IEEE Access, vol. 9*, pp. 51 855–51 864, 2021.

[8] Lee, Sanghoon, and et al, "Design of V2X-based vehicular contents centric networks for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems (2021)*, 2021.

[9] Arshad, Rabe, and et al, "Velocity-aware handover management in two-tier cellular networks," *IEEE Transactions on Wireless Communications 16.3*, 2017.

[10] Jingwen, Bai, and et al, "Route-aware handover enhancement for drones in cellular networks," *2019 IEEE Global Communications Conference (GLOBECOM). IEEE*, pp. 1–6, Dec 2019.

[11] Arshad, Rabe, and et al, "Handover management in dense cellular networks: A stochastic geometry approach," *2016 ieee international conference on communications (icc). IEEE*, pp. 1–7, May 2016.

[12] Semiari, Omid, and et al, "Caching meets millimeter wave communications for enhanced mobility management in 5g networks," *IEEE Transactions on Wireless Communications 17.2 (2017)*, pp. 779–93, Nov 2017.

[13] Arshad, Rabe, and et al, "Integrating UAVs into existing wireless networks: A stochastic geometry approach," *2018 IEEE Globecom Workshops (GC Wkshps). IEEE*, pp. 1–6, Dec 2018.

[14] Wei, Bao, and et al, "Stochastic geometric analysis of user mobility in heterogeneous wireless networks," *IEEE Journal on Selected Areas in Communications 33.10 (2015)*, pp. 2212–25, May 2015.

[15] Ahmad, A. A., and et al, "A comprehensive survey on handover management for vehicular ad hoc network based on 5g mobile networks technology," *Transactions on Emerging Telecommunications Technologies 30.3 (2019)*, 2019.

[16] K. Tokuyama, T. Kimura, and N. Miyoshi, "Time-based handover skipping in cellular networks: Spatially stochastic modeling and analysis," *arXiv preprint arXiv:2008.10535 (2020)*, Aug 2020.

[17] Kibinda, M. Nyaura, and G. Xiaohu, "User-centric cooperative transmissions-enabled handover for ultra-dense networks," *IEEE Transactions on Vehicular Technology 71.4 (2022)*, pp. 4184–97, Jan 2022.

[18] Zhou, He, and et al, "Heterogeneous ultra-dense networks with traffic hotspots: A unified handover analysis," *arXiv preprint arXiv:2204.03294 (2022)*, Apr 2022.

[19] P. Rangesh, K. Pandey, and A. Gupta, "On the performance of communication in a vehicular network with platooned traffic," *2023 National Conference on Communications (NCC)*, pp. 1–6, 2023.

[20] G. Ghatak, "Cooperative relaying for URLLC in V2X networks," *IEEE Wireless Communications Letters, vol. 10*, pp. 97–101, Jan 2021.

[21] Semiari, Omid, and et al, "Caching meets millimeter wave communications for enhanced mobility management in 5G networks," *IEEE Transactions on Wireless Communications 17.2 (2017)*, pp. 779–93, Nov 2017.

[22] Kalamkar, S. S, and et al., "Beam management in 5G: A stochastic geometry analysis," *IEEE Transactions on Wireless Communications 21.4 (2021),pp. 2275-2290*, Sep 2021.

[23] M. Salehi and E. Hossain, "Stochastic geometry analysis of sojourn time in multi-tier cellular networks," *IEEE Transactions on Wireless Communications*, pp. 1816–30, Nov 2020.

[24] M. Haenggi, "Stochastic geometry for wireless networks," *Cambridge, U.K.: Cambridge Univ. Press, 2012*, 2012.

[25] T. D. Novlan, R. K. Ganti, A. Ghosh, and J. G. Andrews, "Analytical evaluation of fractional frequency reuse for ofdma cellular networks," *IEEE Transactions on wireless communications 10.12 (2011)*, pp. 4294–305, Oct 2011.

[26] Soh, Y. Sheng, and et al., "Energy efficient heterogeneous cellular networks," *IEEE Journal on selected areas in communications 31.5 (2013)*, pp. 840–50, April 2013.

[27] N. Choi, K. Guan, D. C. Kilper, and G. Atkinson, "In-network caching effect on optimal energy consumption in content-centric networking," *2012 IEEE international conference on communications (ICC)*, pp. 2889–2894, Jun 2012.

[28] D. Perino and M. Varvello, "A reality check for content centric networking," *Proceedings of the ACM SIGCOMM workshop on Information-centric networking*, 2011.

[29] Prasad, Athul, and et al, "Energy-efficient inter-frequency small cell discovery techniques for lte-advanced heterogeneous network deployments," *IEEE Communications Magazine 51.5 (2013)*, pp. 72–81, May 2013.

[30] M. Di Renzo and P. Guan, "Stochastic geometry modeling of coverage and rate of cellular networks using the gil-pelaez inversion theorem," *IEEE Communications Letters, vol. 18*, pp. 1575–1578, Sept 2014.

[31] G. Ghatak, A. D. Domenico, and M. Coupechoux, "Performance analysis of two-tier networks with closed access small-cells," *2016 14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 1–8, May 2016.

[32] M. Haenggi, "On distances in uniformly random networks," *IEEE Transactions on Information Theory, 51(10)*, pp. 3584–3586, Sep 2005.

Neetu R.R received her BTech and Mtech in communication engineering and signal processing from the University of Calicut, India. Currently, she is pursuing her Ph.D. from Indraprastha Institute of Information Technology (IIIT), Delhi. During her MTech, she worked as a Research Assistant in the Defence Research and Development Organization (DRDO), Bangalore, India. Her research interests are stochastic geometry for heterogeneous networks (HetNets) and wireless communications

Gourab Ghatak received his Ph.D. from Telecom ParisTech (University of Paris Saclay), France, during which he was also employed in CEA Grenoble, France. Currently, he is an Assistant Professor at the Department of Electrical Engineering at the Indian Institute of Technology Delhi (IIT Delhi). His research interests are stochastic geometry and machine learning for wireless communications.

Vivek Ashok Bohara a (Senior Member, IEEE) received his Ph.D. degree from Nanyang Technological University, Singapore, in 2011. From 2011 to 2013, he was a Postdoctoral Researcher (Marie Curie fellowship), ESIEE Paris, University Paris-East, Créteil, France. In 2013, he joined Indraprastha Institute of Information Technology-Delhi, India, where he is currently Head and Professor (ECE). He has authored and co-authored more than 50 publications in major IEEE/IET journals, refereed international conferences, two book chapters, and one patent. His research interests include next-generation communication technologies, such as device-to-device communication, carrier aggregation, and visible light communications. Dr. Bohara was the recipient of the First Prize in National Instruments ASEAN Virtual Instrumentation Applications Contest in 2007 and 2010, Best Poster Award at the IEEE ANTS 2014 and the IEEE Comsnets 2015 and 2016 conferences.

Anand Srivastava did his M.Tech. and Ph.D. from IIT Delhi and is currently working at IIIT Delhi as a Professor in ECE department since Nov. 2014 and also Director at IIIT Delhi Incubation Center (a Section 8 company). He is also Adjunct faculty in Bharti School of Telecom Technology at IIT Delhi. He was Dean & Professor in the School of Computing and Electrical Engineering at Indian Institute of Technology Mandi, HP, India for 2 years. In his initial stint of 20 years, he was with the Center for Development of Telematics (CDOT), a telecom research center of Govt. of India where he was Director and member of the CDOT Board. Currently, he is driving VLC/LiFi standardization activities under the aegis of TSDSI. His research work is in the area of optical core & access networks, Vehicle-to-vehicle communications, Fiber-Wireless (FiWi) architectures, and Visible light communications.