Cache-Enabled Millimeter Wave Cellular Networks With Clusters
Yu Ye, Student Member, IEEE, Shaocheng Huang, Student Member, IEEE, Ming Xiao, Senior Member, IEEE, Zheng Ma, Senior Member, IEEE, and Mikael Skoglund, Fellow, IEEE

Abstract—Wireless content caching in cellular networks is an efficient way to reduce the service delay and alleviate backhaul pressure. For the benefits of sharing spectral and storage resources, clustering in cached networks has recently attracted significant research interests. Meanwhile, since the multimedia content (e.g., video) of caching networks may require a huge transmission rates, millimeter wave (mmWave) communication is considered to be an efficient transmission scheme for cache-enabled networks. We investigate the ergodic rate and average service delay for typical user terminal (UT) in the clustered cache-enabled small cell networks (SCN) and ultra dense networks (UDN) with mmWave channels. In SCN, each cluster consists of cache-enabled UTs, and in the UDN a cluster is formed by cache-enabled UTs and small base stations (SBSs) with non-uniform caching capacity. The clusters are assumed to be discs and content sharing is only possible within clusters through mmWave device-to-device (D2D) tier and SBS tier communications. With stochastic geometry methods, the distributions of content sharing distance and signal-to-interference-noise-ratio (SINR) of typical UT in a cluster are derived for both SCN and UDN scenarios. To minimize the average service delay in high SINR region, we provide an algorithm to jointly optimize caching scheme for SBSs and UTs. By simulations, we validate our theoretical analysis and the performance of proposed caching scheme. The numerical results also show that there exists best radius in the design of cluster for UDNs.

Index Terms—Caching, D2D communication, heterogeneous networks, ultra dense networks, mmWave, clustering.

I. INTRODUCTION

Wireless data traffic has experienced tremendous growth in recent years. According to [1], [2], the data traffic is increasingly concentrated to hotspots and is expected to increase almost tenfold by 2020 compared with 2016. The heavy traffic is normally connected to the core network (CN) through backhauling, which may have limited capacity. This makes the backhaul a bottleneck of wireless networks. One promising technology to reduce the expensive usage of the backhaul and to shorten service delay is wireless caching [3]–[11]. In a cache-enabled system, storage memory is utilized to prefetch popular contents in small cells during off-peak hours before being requested by user terminals (UTs). The requests can be served directly through wireless links if the requested contents are already stored in the local small base stations (SBSs) instead of fetching from the remote CN. Thus the service delay can be largely reduced. Since the UTs, e.g., smart phones and laptops, are normally equipped with storage units, popular contents can also be stored at the local cache of UTs. Then the requests can be served locally if the requested contents are already available at local storage. Moreover, the saved contents can also be shared via device-to-device (D2D) communication among neighboring UTs [12]–[18], through which the backhaul bandwidth and service delay can be further reduced. However, due to the co-channel interference caused by D2D communications, and the limited energy resources for UTs, the resource management for D2D communications should be well designed. In [15], a joint D2D link scheduling and power allocation problem is formulated in order to maximize the system throughput. The trade-off between the energy efficiency (EE) and spectral efficiency (SE) is investigated in [19]. The cluster-based D2D communication is studied in [20]–[24]. Reference [20] presents a system model where the cell is divided into clusters. Each cluster is formed by UTs with caching, and the UTs in a same cluster can share stored contents via D2D communication. With clustering, the same time-frequency resource can be reused in each cluster, through which the intra-cluster interference can be avoided. A cluster-based multicast transmission method for D2D communication with the objective of decreasing the data distribution time is proposed in [21]. In [22], a probabilistic caching scheme is designed to minimize the energy consumption of a clustered cache-enabled D2D network. Reference [23]
investigates and maximizes the cache offloading gain and rate coverage probability for a clustered D2D caching network, where a sub-optimal caching solution is obtained. Besides, a joint caching and spectrum partitioning scheme is proposed to reduce the average service delay for clustered D2D networks in [24].

Meanwhile, as one promising technology for the fifth generation (5G) and beyond mobile networks, millimeter wave (mmWave) communication can support higher data rate than microwave communications [25]–[27]. According to references [28]–[30], mmWave communications can facilitate popular content exchange in cache-enabled small cell networks (SCNs), especially for video streaming, which may require very high transmission rates [31]. The offloading gain of D2D-aware device caching with mmWave D2D communication is revealed in [28]. While the successful content delivery probability is studied in [29]. In [30], the success probabilities and area spectral efficiencies are analyzed in a novel cache-enabled heterogeneous networks (HetNets), where macro base stations (BSs) with traditional sub-6 GHz are overlaid by dense mmWave pico BSs. Besides, it is shown that combining caching and mmWave at SBSs can significantly reduce the connection and retrieval delays [31]. Thus, it is expected that mmWave will be widely used in mobile networks and it is valuable to investigate wireless caching networks with mmWave communications, which is significantly different from traditional microwave communications in terms of channel models, performance and design principles etc. In the above-mentioned works, none of them investigates the clustering scheme for cache-enabled mmWave cellular networks.

Another capacity-increasing approach in 5G systems is deploying ultra dense networks (UDN), where the network is closer to UTs than traditional SCNs. Different from SCN, the density of access points (APs), including macro BSs, micro BSs (pico BSs and femto BSs), and relays, are much higher than that in SCN. The distance between the APs and UTs is further decreased. Thus, the network capacity is increased with UDN [32]. The cache-enabled SCN is investigated in [33]–[37], while the cache-enabled UDN is investigated in [38]–[43]. Reference [35] considers a cluster-centric SCN with the combined design of cooperative caching and transmission policy. In [36], UTs with similar content popularity are grouped into a cluster, and served by a same SBS. As for UDN, a clustering approach is considered as a promising way to simplify the topology structures of UDN and to reduce complexity [38]. In [39], the user-centric cooperative interference mitigation strategy is proposed, where dense SBSs are able to store contents. In [42], the small cells are grouped into disjoint clusters. However, there are no results investigating the clustering scheme for the cache-enabled mmWave SCN and UDN, to our knowledge.

We focus on studying the effect of cluster size, caching policies and capacity of UTs and SBSs by characterizing the ergodic rate and average service delay for typical UT in clustered cache-enabled SCN and UDN. The main contributions of this paper are listed as follows.

- We investigate the service delay for typical UT located at the center of a cluster in cache-enabled SCN and UDN, where both mmWave D2D and SBS tiers are involved.
- We derive the exact distribution of serving distance for typical UT in cache-enabled SCN and UDN with clusters, based on which the closed-form expression of SINRs with mmWave communications are calculated through Gaussian-Chebyshev quadrature approximation.
- We derive the exact ergodic rate and average service delay for the typical UT in SCN and UDN scenarios. To gain further insights, we assume having a high SINR and the closed-form approximation for ergodic rates is derived.
- To minimize the average service delay of typical UT, we propose an Alternate Optimization based algorithm to jointly optimize the caching policy for SBSs and UTs in both SCN and UDN scenarios.
- Numerical results show that: 1) in mmWave SCN and UDN, the approximations of SINRs and ergodic rates in high SINR region coincide with theoretical results; 2) the proposed caching scheme can achieve smaller service delay compared with other policies; 3) there exists a best size of clusters in cache-enabled SCN upon the service delay of typical UT.

The rest of the paper is organized as follows. We present the model of clustered cache-enabled SCN and UDN with mmWave communication in Section II. Then for both scenarios, the distributions of content transmission and SINR for typical UT are provided in Section III. Furthermore, in Section IV, ergodic rate and average service delay for typical UT are derived. Numerical experiments are presented to evaluate the performance of the clustering scheme in Section V. Finally we make conclusions in Section VI.

II. SYSTEM MODEL

The system model of clustered cache-enabled networks with mmWave channel is illustrated in Fig. 1. We assume that the locations of UTs follow a two-dimensional homogeneous Poisson Point Process (PPP) \( \Phi_1 \) with density \( \lambda_1 \). The locations of SBSs are spatially distributed following a two-dimensional homogeneous PPP \( \Phi_2(\Phi_2') \) with density \( \lambda_2(\lambda_2') \) in SCN(UDN) [44]. The popular contents are denoted by \( \mathcal{F} = \{1, \ldots, F\} (F \in \mathbb{N}) \). Without loss of generality, we assume that each content has the same length \( l \), and they are stored as a whole at each UT and SBS. The storage capacity for UT and SBS is \( \Theta_1 \) and \( \Theta_2 \) respectively. Notation is summarized in Table I.

A. Network Architecture

We consider clustering schemes for two scenarios, where clusters are disjoint discs with common radius \( d \). Fig. 1(a) shows the clustered cache-enabled SCN, where the density of SBSs is much lower than that of UTs, i.e. \( \lambda_1 \gg \lambda_2 \). For this scenario, cells are divided into virtual clusters as demonstrated in [20]. Within each cluster, UTs can share stored contents with each other through D2D links, where the content sharing is completed in one hop to reduce the delay and complex communication protocols for multi-hop transmission. When
a content is requested by a UT, the request will be served immediately if the content is stored at the cache. Otherwise, the desired content can be obtained from the nearest UT where it is available. When the requested content is not available in the cluster, it needs to be retrieved from the cache of the nearest SBS or the CN.

While for the cache-enabled UDN shown in Fig. 1 (b), the considered area is divided into virtual clusters that consist of SBSs and UTs. In UDN, the density of the SBS is comparable with that of UT. Hence in the design of the clustering scheme, we need to take both SBSs and UTs into account.

According to resource sharing schemes in [45]–[48], we assume that in each cluster, local D2D links reuse the same radio resource. That is, concurrent transmissions of D2D links are available. References [45], [46] have shown that the use of directional antenna and high propagation loss in mmWave bands can result in relatively lower mutual interference or even no interference by properly selecting the concurrent links formed by geographically distributed wireless devices. Thus, we ignore the interference for concurrent transmissions in each cluster, which may provide an upper bound of system performance. To make the interference of D2D links between clusters negligible, according to [20], we allow the maximum number of D2D links in a disc of radius \(d\) as \(N_d = \left\lfloor d/r_d \right\rfloor\), where \(r_d\) is D2D collaboration distance [20]. With multiple D2D links in a cluster, we use orthogonal radio resource (frequency or time) to avoid mutual intra-cluster interference among D2D links [47]–[50]. We consider equal bandwidth partition among D2D links in each cluster. There is only one SBS link is allowed to be active at one time in each cluster. For tractable analysis, the concurrent transmissions of D2D links and D2B link in a cluster is not allowed. Our models (SCN and UDN) are applicable to any radio allocation schemes and non-orthogonal resource allocation schemes (e.g., non-orthogonal multiplexing access [51]).

### B. Transmission Model With mmWave Channel

We assume the connections between SBS and UT, UT and UT are modeled as mmWave channels, where the D2D tier and SBS tier share the same frequency resource. Then we consider directional beamforming and use a sector model to analyze effective beam pattern [25], [52]. Let \(G_{\varphi_m} = g_M g_m(\varphi)\) denote the sectored antenna pattern, where \(\varphi\) is the angle off the boresight direction, \(\varphi_m \in [0, 2\pi]\) is main-lobe beamwidth, \(g_M\) (dB) is main-lobe gain, and \(g_m\) (dB) is side-lobe gain. Then the antenna gain between a transmitter and a receiver is a discrete random variable, which is described by

\[
G = \begin{cases}
  g_M^2, & P_{TM_M} = \frac{\varphi_m^2}{4\pi^2}; \\
  g_M g_m, & P_{TM_m} = \frac{(2\pi - \varphi_m)\varphi_m}{4\pi^2}; \\
  g_m^2, & P_{TM_m} = \frac{(2\pi - \varphi_m)^2}{4\pi^2}; \\
  g_m^2, & P_{TM_m} = \frac{(2\pi - \varphi_m^2)}{4\pi^2}.
\end{cases}
\]

where \(P_{Tlk} (l, k \in \mathcal{M} = \{m, M\})\) is the probability for antenna gain \(g_{lk}\). We further assume that perfect beam alignment can be achieved at the desired link with perfect
Thus, the directivity gain for the desired signal link is $G_0 = g_M^2$.

In both SCN and UDN scenarios, we assume that the total bandwidth for each cluster is $W$. Without loss of generality, we suppose that each user connects only one UT(SBS) when it is served. In addition we consider both path loss and small-scale fading. For small-scale fading of each link, the Nakagami-$M$ fading channel is considered as in [52], [53]. Besides, in SCN and UDN, we consider line-of-sight (LOS) for links within the cluster and non-line-of-sight (NLOS) for links outside the cluster according to [25]. The path loss exponents for the LOS and NLOS links are denoted as $\alpha_L$ and $\alpha_N$, respectively [27].

### C. Caching Model

Similar to [16], we consider homogeneous content preference in SCN and UDN scenarios, where all UTs have the same content preference distribution. In other words, each UT independently and randomly requests contents according to the same popularity distribution $p \in (0, 1)^{1 \times F}$. According to results in [54], $p$ follows a Zipf-like distribution, while the element $p_f$ represents the probability of UT requesting content $f$. Defining the content popularity index as vector $\omega = \{\omega_1, \ldots, \omega_F\}$, the request probability of the $i$-th most popular content $\omega_i$ can be calculated as

$$p_{\omega_i} = \frac{i^{-\eta}}{\sum_{j=1}^{F} j^{-\eta}}, \quad i \in F,$$

where $\eta$ is the shape parameter. The vector $\omega$ is permutation of $F$. All UTs(SBSs) follow the same policy $\Pi_1(\Pi_2)$ (e.g., random caching (RC)) to store contents. Denote the index of cached contents for UT as $\omega_{\Pi_1} \subset \omega$ with length $\Theta_1$, while the index of cached contents for SBS as $\omega_{\Pi_2} \subset \omega$ with length $\Theta_2$. The caching capacity of SBSs and UTs is non-uniform, i.e., $\Theta_2 > \Theta_1$. We define $\text{Pr}_{1, f} \triangleq \text{Pr}\{f \in \omega_{\Pi_1}\}$ and $\text{Pr}_{2, f} \triangleq \text{Pr}\{f \in \omega_{\Pi_2}\}$. When $\Pi_1(\Pi_2)$ is fixed, the probability $\text{Pr}_{1, f}(\text{Pr}_{2, f})$ is identical among UTs(SBSs).

### III. SINR Analysis in Clustered Cache-Enabled Networks

We first present the distribution of transmission distance for a typical UT in clustered SCN and UDN. Then considering mmWave channel in D2D and SBS tiers, the SINR distribution is derived for both scenarios.

#### A. Distance Distribution in Clustered SCN

Since the density of SBS is much smaller than that of UT, i.e., $\lambda_1 \gg \lambda_2$, SBSs in $\Phi_2$ are located only in a few clusters. Hence we consider a cluster $A$ with radius $d$, which does not include a SBS. Moreover, to simplify the analysis, we focus on the typical UT $o$, which is inserted at the center of $A$. As illustrated in Section II, the typical UT can retrieve desired content from neighboring UTs within $A$ with distance $X_1$, or from the nearest SBS with $X_2(> d)$. The probability density functions (PDFs) of $X_1$ and $X_2$ are given in Lemma 1.

![Figure 2. Content sharing within a cluster in cache-enabled mmWave SCN.](image)

**Lemma 1:** In clustered cache-enabled SCN, for the typical UT $o$ requesting content $f$, which is located at the center of $A$ with radius $d$, the PDF of $X_1$ is

$$f_{X_1}(x) = \begin{cases} \frac{2\lambda_1 f^{-\lambda_1} \pi x e^{-\lambda_1 x \pi^2}}{1 - e^{-\lambda_1 x \pi^2}}, & 0 \leq x \leq d; \\ 0, & \text{otherwise}, \end{cases}$$

where $\lambda_1 = \text{Pr}_{1, f} \lambda_1$. The PDF of SBS tier distance $X_2$ is

$$f_{X_2}(x) = \begin{cases} 2\lambda_2 \pi x e^{-\lambda_2 \pi^2 (x^2 - d^2)}, & d \leq x; \\ 0, & \text{otherwise}. \end{cases}$$

**Proof:** See Appendix A the Proof of Lemma 1. For the derivation of $f_{X_1}(x)$, we thin the PPP $\Phi_1$ to $\Phi_{1, f}$ by only considering the UTs which have desired content $f$. The derivation of $f_{X_2}(x)$ is based on the condition that content $f$ is not available within $A$.

#### B. SINR Distribution in mmWave SCN

Based on the distance distribution of content sharing in Lemma 1, we then investigate the SINR distribution for the typical UT when mmWave channel is applied. At first the SINR distribution for UT $o$ is derived with mmWave D2D transmission. Let $D_1 \triangleq \Phi_1 \setminus A$ and denote $s \in D_1$ the interference SBS outside of cluster $A$. Thus according to [25], the SINR between the typical UT and the serving UT can be expressed as

$$\Gamma_1 = \frac{P_1 G_0 e^{-\alpha_L} |h_0|^2}{I_1 + \sigma^2},$$

where $I_1 = \sum_{s \in D_1} P_0 G_r s^\alpha |h_s|^2$ is the inter-cluster interference with mmWave channel and $r_s$ is the distance between the interference SBS and typical UT. $P_1$ is the transmission power of D2D tier, while $P_0$ is the average transmission power for SBSs within $D_1$. Then we derive the cumulative distribution function (CDF) of $\Gamma_1$ in the following lemma.

**Lemma 2:** In cache-enabled mmWave SCN, the CDF of $\Gamma_1$ for typical UT requesting $f$ at cluster $A$ can be approximated as

$$F_{\Gamma_1}(\gamma) \approx \sum_{m=0}^{M} \left( \begin{array}{c} M \\ m \end{array} \right) (-1)^m \sum_{i=1}^{I} w_i \Psi_{X_1, f, \lambda_2, \psi_{1, f}}(u_i),$$

where $t$ is a predefined integer and

$$w_i = \frac{\pi}{t+1} \sin^2 \left( \frac{i}{t+1} \pi \right), \quad u_i = \cos \left( \frac{i}{t+1} \pi \right),$$

When $m \geq 4$, we have

$$\left( \begin{array}{c} M \\ m \end{array} \right) (-1)^m \sum_{i=1}^{I} w_i \Psi_{X_1, f, \lambda_2, \psi_{1, f}}(u_i) \approx \sum_{i=1}^{I} w_i \Psi_{X_1, f, \lambda_2, \psi_{1, f}}(u_i),$$

where $\Psi_{X_1, f, \lambda_2, \psi_{1, f}}(u)$ is the cumulative distribution function (CDF) of $X_1$. The approximation is justified by the fact that $\sum_{j=0}^{m} \binom{M}{j} (-1)^j \Psi_{X_1, f, \lambda_2, \psi_{1, f}}(u_j)$ is the cumulative distribution function of $X_1$, and $\sum_{j=0}^{m} \binom{M}{j} (-1)^j \Psi_{X_1, f, \lambda_2, \psi_{1, f}}(u_j) \approx \sum_{j=0}^{m} \binom{M}{j} (-1)^j \Psi_{X_1, f, \lambda_2, \psi_{1, f}}(u_j)$ as $m \to \infty$.
where \( \Psi_{X_{1}, f, \lambda_{2}, \psi_{1}, l, k} (u_i) \) is given in (8), shown at the bottom of this page, \( \psi_{1, l, k} = \frac{G_{0, 0} \pi}{\pi r_{G}}, \kappa = M(M!)^{-\frac{1}{2}}, \) and \( 2F_{1} (\cdot; \cdot; \cdot) \) is the Gaussian Hypergeometric function.

Proof: See Appendix B the Proof of Lemma 2.

To obtain the closed-form expression for \( F_{\Gamma_{2}} (\gamma) \), we adopt Gaussian-Chebyshev quadrature in the derivation, which approximates the definite integral of a function by weighted sum of function values at specified points within the domain of integration. \( t \) balances the trade-off between the complexity and accuracy of Gaussian-Chebyshev quadrature approximation. The detailed derivation of \( \Psi_{X_{1}, f, \lambda_{2}, \psi_{1}, l, k} (\cdot) \) can be found in Appendix B.

When \( j \) is not available for the typical UT in \( A \), the content will be retrieved from the nearest SBS. Let \( X_2 \) denote the distance between the typical UT and the nearest SBS, and \( A_2 \) denote the disc with radius \( X_2 \). Thus, when the nearest SBS provides service, we obtain the SINR for the UT o as

\[
\Gamma_{2} = \frac{P_{2} G_{0} r^{-\alpha_{N}} |h_{0}|^{2}}{I_{2} + \sigma^{2}},
\]

where \( I_{2} = \sum_{s \in D_{2}} P_{0} G_{s} r^{-\alpha_{N}} |h_{si}|^{2} \) is the inter-cluster interference, and \( r_{s} \) is the distance between the interference SBS in \( D_{2} \triangleq \Phi_{2} (b(o, X_{2})) \) and UT o, where \( b(o, X_{2}) \) is the disc of radius \( X_2 \) centered at \( o \). \( P_{2} \) is the transmission power of \( \text{mmWave SBS tier} \). Following similar approaches in Lemma 2, we can obtain the CDF of \( \Gamma_{2} \).

**Lemma 3:** In cache-enabled \( \text{mmWave SCN} \), the CDF of \( \Gamma_{2} \) for typical UT at cluster \( A \) can be approximated as

\[
F_{\Gamma_{2}} (\gamma) \approx \sum_{m=0}^{M} \binom{M}{m} (-1)^{m} \sum_{i=1}^{4} w_{i} \Lambda_{\lambda_{2}, \psi_{2}, l, k} (u_{i}),
\]

where \( w_{i} \) and \( u_{i} \) are given in (7), and

\[
\Lambda_{\lambda_{2}, \psi_{2}, l, k} (u_{i}) = 8 \lambda_{2} \pi d^{2} e^{2\pi d^{2}/(u_{i}+1)^{2}} e^{-4\lambda_{2} \pi d^{2}/(u_{i}+1)^{2}} \frac{\text{erfc}(2\pi d^{2}/(u_{i}+1)^{2})}{\sqrt{1-u_{i}^{2}}^{}},
\]

where \( \psi_{2, l, k} = \frac{G_{0, 0} \pi}{\pi r_{G}} \).

Proof: From the proof of Lemma 2, the CDF of \( \Gamma_{2} \) can be obtained as

\[
F_{\Gamma_{2}} (\gamma) \approx \sum_{m=0}^{M} \binom{M}{m} (-1)^{m} E_{X_{2}} \left[ \sum_{l, k} \varepsilon_{l, k} e^{-\frac{r_{l, k}^{2}}{r_{G}^{2}}} \right],
\]

where

\[
E_{X_{2}} \left[ e^{-\frac{r_{l, k}^{2}}{r_{G}^{2}}} \right] = \frac{1}{2} \left( 1 - e^{-\frac{\lambda_{f} \pi d^{2}}{u_{i}+1}} \right) \sqrt{1-u_{i}^{2}} e^{-\frac{\lambda_{f} \pi d^{2}}{u_{i}+1}} \sqrt{1-u_{i}^{2}},
\]

Then substituting (13) into (12) and calculating the expectation over \( X_{2}, \) the final result (10) is derived.

Similar to \( \Psi_{X_{1}, f, \lambda_{2}, \psi_{1}, l, k} (\cdot) \), the derivation of \( \Lambda_{\lambda_{2}, \psi_{2}, l, k} (\cdot) \) is also obtained from the Gaussian-Chebyshev quadrature approximation \( \int_{1}^{\infty} \sqrt{1-x^{2}} g(x) dx \approx \sum_{i=1}^{4} w_{i} g(u_{i}) \). But note that \( \Lambda_{\lambda_{2}, \psi_{2}, l, k} (\cdot) \) is only determined by the density of SBSs, i.e. \( \lambda_{2} \).

**C. Distance Distribution in Clustered UDN**

For the cache-enabled mmWave UDN with clusters shown as Fig. 1 (b), the density of SBSs is much higher than that of \( \text{SCN} \). The clusters in UDN are formed by UTs and SBSs, where two independent PPPs, \( \Phi_{1} \) and \( \Phi_{2} \), are included. In cluster \( A \), a UT has access to the contents stored at the caches of neighboring UTs and SBSs. Specifically, we consider the typical UT o, which is inserted at the center of cluster A. This is shown in Fig. 3, where \( A \) is a disc with radius \( d \). To differentiate UDN from the \( \text{SCN} \) scenario, we denote the density of SBSs PPP \( \Phi_{2} \) as \( \lambda_{2} \), where \( \lambda_{2} \gg \lambda_{2} [32] \). Besides, without loss of generality, we assume that there exists at least one SBS within cluster \( A \). In what follows, we will first investigate the distribution of transmission distance for the typical UT. Then the distribution of SINR for UT o with mmWave channel will be given.

Considering UT o requesting content \( f \) in cluster \( A \), we denote the distance for the D2D tier between typical UT o and nearest UT which has content \( f \) as \( X_{3} \). The distance that o retrieves \( f \) from nearest SBS which stores \( f \) is denoted as \( X_{4} \). When \( f \) is not available in \( A \), the typical UT is served by the nearest SBS with distance \( X_{5} \). Then denoting \( \Lambda_{2, f} = \text{Pr}_{2, f} \lambda_{2} \), we have the following results.

\[
\Psi_{X_{1}, f, \lambda_{2}, \psi_{1}, l, k} (u_{i}) = \frac{\bar{\Lambda}_{1, f} \pi d^{2} (u_{i}+1)}{2 (1 - e^{-\bar{\Lambda}_{1, f} \pi d^{2}}) \sqrt{1-u_{i}^{2}}} e^{-\frac{\bar{\Lambda}_{1, f} \pi d^{2}}{u_{i}+1}^{2}} \times e^{-\frac{\lambda_{2} \sum_{l, k} \varepsilon_{l, k} d^{2}}{r_{G}^{2}}} \varepsilon_{l, k}^{2} \left( 2F_{1} \left( M, \frac{2}{\pi d^{2}}; \frac{2}{\pi d^{2}}; \frac{2}{\pi d^{2}}; \frac{2}{\pi d^{2}}; -\gamma m_{\psi_{1}, l, k}(u_{i}+1)\right) \right).
\]
Lemma 4: In clustered cache-enabled UDN, for typical UT requesting content \( f \), which is located in the center of \( A \) with radius \( d \), the PDF \( f_{X_3}(x) \) is the same with \( f_{X_1}(x) \) given in (3), while the PDFs \( f_{X_4}(x) \) and \( f_{X_5}(x) \) are obtained by substituting \( X_1 \) with \( X_2 \) and \( X_2 \) respectively.

Proof: Recalling the proof of Lemma 1, the only difference in UDN is that the density of SBSs becomes \( \lambda_2 \). Based on the assumption \( N_2 > 0 \), the distributions can be derived through calculating \( \Pr \{ X_3 \leq x | N_2 > 1 \} \) and \( \Pr \{ X_4 \leq x | N_2 > 1 \} \) respectively as given in (33), where \( N_2 \) denotes the number of SBSs in \( A \) which have content \( f \).

\[ \Xi = \sum_{j \in F} p_j \left( 1 - \Pr_{1,j} \right) e^{-\lambda_1 \pi d^2}. \]

The \( \Xi \) represents the probability that a SBS is active. Let point \( s \in \Phi_2 \setminus \mathcal{A} \) denote the interference SBS outside of disc \( \mathcal{A} \). Thus the SINR experienced by the typical UT for three scenarios shown in Fig. 3 can be defined as

\[ \Gamma_3 = \frac{P_0 G_0 \lambda_2}{I + \sigma^2} ; \quad \Gamma_{4,5} = \frac{P_0 G_0 \lambda_2}{I + \sigma^2}, \]

where \( I = \sum_{s \in \mathcal{P}} \frac{P_0 G_0 \lambda_2 |h_0|^2}{I + \sigma^2} \) is the inter-cluster interference with mmWave channel. \( P_0 \) is the average transmission power for SBSs within \( D \). The CDFs of \( \Gamma_3(j = 3, 4, 5) \) are given in Lemma 5.

Lemma 5: In cache-enabled mmWave UDN, the CDF of \( \Gamma_3,4,5 \) for typical UT requesting \( f \) at cluster \( \mathcal{A} \) can be approximated as

\[ F_{\Gamma_1}(\gamma) \approx \sum_{m=0}^{M} \binom{M}{m} (-1)^m \sum_{i=1}^{t} w_i \psi_{a_3, a_4, b_3, b_4, b_5} (u_i), \]

where \( a_3 = \lambda_1, f \), \( b_3 = \psi_{1, ik} \); \( a_4 = \lambda_2, f \), \( b_4 = \psi_{2, ik} \); \( a_5 = \lambda_2 \), \( b_5 = \psi_{2, ik} \).

\[ \triangleq \]

D. SINR Distribution in mmWave UDN

Based on the distance distribution of content sharing given in Lemma 4, we then investigate the SINR distribution for the typical UT in cache-enabled mmWave UDN. Since only one SBS link is active in each cluster at a time, there is at most one active SBS in neighboring clusters. Hence to ensure that the expected amount of SBSs in \( D = \Phi_2 \setminus \mathcal{A} \) is equal to the case where one SBS is placed at each cluster in average sense, we let the density of SBSs in \( D \) be \( \lambda_2 = \frac{1}{\pi d^2} \). Besides, not all the SBSs in \( D \) are active for serving due to the presence of D2D tier content exchanging. In order to characterize the SBSs which interfere UT \( o \), we further thin the density of SBSs in area \( D \) to \( \lambda_2, f = \Xi \cdot \lambda_2 \), where

\[ \Xi = \sum_{f \in F} p_f \left( 1 - \Pr_{1,f} \right) e^{-\lambda_1 \pi d^2}. \]

The CDFs of \( \Gamma_{j} \) for typical UT requesting \( f \) at cluster \( \mathcal{A} \) can be obtained through calculating

\[ \frac{\ln 2}{\gamma_{max}} \int_{0}^{\gamma_{max}} \frac{1}{1 + \gamma} d\gamma \]

where \( \gamma_{max} = 2R_{max} - 1 \) is the SINR threshold determined by practical constraints for the radio frequency circuit, and \( R_{max} \) is the maximum achievable rate.

Theorem 1: The ergodic rate of a typical UT in cluster \( \mathcal{A} \) for cache-enabled mmWave SCN and UDN is given by

\[ R_{i} = \frac{W}{\ln 2} \int_{0}^{\gamma_{max}} \frac{1}{1 + \gamma} (1 - F_{\Gamma_i} (\gamma)) d\gamma, \quad i = 1, 2, 3, 4, 5, \]

where \( W_{1,3} = \frac{W}{\lambda_2} \) and \( W_{2,4,5} = W \).

Proof: Following the proof of Lemma 2, we can similarly obtain the results in (16) but with different density of interference SBSs and the densities given by (5).

The calculation of the interference in Lemma 5 by using the thinned density \( \lambda_2, f \) is only an approximation to the true case. Comparing to \( F_{\Gamma_1} \) in (6), the CDFs \( F_{\Gamma_{3,4,5}} \) have similar form. This is because the densities of SBSs with \( f \) and interference SBSs in \( D \) in UDN are different from those in SCN.

IV. PERFORMANCE ANALYSIS AND OPTIMIZATION

In this section, we will derive the ergodic rate of a typical UT at a cluster for both cache-enabled mmWave SCN and UDN. Then the average service delay for typical UT will be studied.

A. Ergodic Rate Analysis

The ergodic rate refers to the mean data rate when adaptive modulation/coding is used over many fading realization for a given average SINR [44], which is given by

\[ R = \frac{W}{\ln 2} \int_{0}^{\gamma_{max}} \frac{1}{1 + \gamma} (1 - F_{\Gamma_i} (\gamma)) d\gamma, \]

where \( \gamma_{max} = 2R_{max} - 1 \) is the SINR threshold determined by practical constraints for the radio frequency circuit, and \( R_{max} \) is the maximum achievable rate.

When typical UT is served by the SBS, the distribution for transmission distance should be obtained similar as \( X_4 \) in Lemma 4 but on the condition \( N_2 = 1 \).
Corollary 1: In high SINR region, the ergodic rate of D2D tier for typical UT at cluster $\mathcal{A}$ in cache-enabled mmWave SCN is approximated by

$$R_1^+ = \frac{W_1}{\ln 2} \left( \left( 1 + Y_{\lambda_1,f,\lambda_2,\psi,1,ik} \right) \ln \left( 1 + \gamma_{\max} \right) - Y_{\lambda_1,f,\lambda_2,\psi,1,ik} \gamma_{\max} \right),$$

(21)

where

$$Y_{\lambda_1,f,\lambda_2,\psi,1,ik} = \frac{2}{\alpha_N - 2} \left( \lambda_2 \pi d^2 + 1 \right) \sum_{k,l \in \mathcal{M}} M^2 Pr_{ik} \psi_{2,ik}$$

$$+ \frac{M \lambda_2 \sigma^2 \theta_{\max} d^2}{P_2 G_0 (\lambda_2 \pi d^2)^{\frac{1}{2}} \Gamma \left( 1 + \alpha_N \frac{1}{2}, \lambda_2 \pi d^2 \right)}.$$

(22)

Proof: See Appendix C the Proof of Corollary 1. □

Under high SINR approximation, the closed-form expression of $Y_{\lambda_1,f,\lambda_2,\psi,1,ik}$ can be obtained directly through integration without using Gaussian-Chebyshev quadrature approximation and Gaussian Hypergeometric function.

Corollary 2: In high SINR region, the ergodic rate of SBS tier for typical UT at cluster $\mathcal{A}$ in cache-enabled mmWave SCN is approximated by

$$R_2^+ = \frac{W_2}{\ln 2} \left( \left( 1 + \Omega_{\lambda_2,\psi,2,ik} \right) \ln \left( 1 + \gamma_{\max} \right) - \Omega_{\lambda_2,\psi,2,ik} \gamma_{\max} \right),$$

(23)

where

$$\Omega_{\lambda_2,\psi,2,ik} = \frac{2}{\alpha_N - 2} \left( \lambda_2 \pi d^2 + 1 \right) \sum_{k,l \in \mathcal{M}} M^2 Pr_{ik} \psi_{2,ik}$$

$$+ \frac{M \lambda_2 \sigma^2 \theta_{\max} d^2}{P_2 G_0 (\lambda_2 \pi d^2)^{\frac{1}{2}} \Gamma \left( 1 + \alpha_N \frac{1}{2}, \lambda_2 \pi d^2 \right)}.$$

(24)

Proof: This can be similarly proved by following the process of Appendix C but with substituting $X_1,I_1$ with $X_2,I_2$ and the corresponding distributions. □

In clustered cache-enabled UDN, the ergodic rate of typical UT in high SINR region can be also derived with Lemma 5.

Corollary 3: In high SINR region, the ergodic rates $R_j^+$ ($j = 3, 4, 5$) of typical UT at cluster $\mathcal{A}$ in cache-enabled mmWave UDN are derived as

$$R_j^+ = \frac{W_j}{\ln 2} \left( \left( 1 + Y_{a_j,\lambda_j,\psi,2,ik, bj} \right) \ln \left( 1 + \gamma_{\max} \right) - Y_{a_j,\lambda_j,\psi,2,ik, bj} \gamma_{\max} \right),$$

(25)

where ($a_j, b_j$) are given by (16).

Proof: Following the proof of Corollary 2, we can derive the final results by changing the densities of SBSs ($\lambda_{2,j},\lambda_j$) within $\mathcal{A}$ and interference SBSs ($\lambda_{2,j}$) in $\mathcal{D}$. □

It is worth noting that except for $R_2$, the other ergodic rates given in Theorem 1 and Corollaries 1, 2 and 3 are derived with respect to (w.r.t.) the requested content $f$. Besides, the feasible domain for transmission power $P_1$ and $P_2$ can be obtained by letting $R_i^+ \geq 0 (i = 1, 2, 3, 4, 5)$.

B. Service Delay Analysis

With the derived ergodic rates, the delay for two-nodes transmission is described as $T = l/R$. We then evaluate the average service delay for the typical UT$^2$ which is inserted at the center $o$ of cluster $\mathcal{A}$.

Proposition 1: In cache-enabled mmWave SCN, the average service delay $\bar{T}_1$ for typical UT $o$ at cluster $\mathcal{A}$ is evaluated by

$$\bar{T}_1 = \sum_{f \in \mathcal{F}} p_{n,f} (1 - Pr_{1,f}) \left( \left( 1 - e^{-Pr_{1,f} \lambda_1 \pi d^2} \right) \frac{l}{R_1} + e^{-Pr_{1,f} \lambda_1 \pi d^2} \left( \frac{l}{R_2} + (1 - Pr_{2,f}) T_0 \right) \right).$$

(26)

In cache-enabled mmWave UDN, the average service delay $\bar{T}_2$ for typical UT at cluster $\mathcal{A}$ is given by

$$\bar{T}_2 = \sum_{f \in \mathcal{F}} p_{n,f} (1 - Pr_{1,f}) \left( \left( 1 - e^{-Pr_{1,f} \lambda_1 \pi d^2} \right) \frac{l}{R_3} + e^{-Pr_{2,f} \lambda_2 \pi d^2} \left( \frac{l}{R_4} + (1 - Pr_{2,f}) T_0 \right) \right).$$

(27)

Proof: In cache-enabled mmWave SCN, considering the request for content $f$ from typical UT at center of cluster $\mathcal{A}$, the delay $T_1$ for fetching content from local cache, the UTs within $\mathcal{A}$, the cache of nearest SBS and the CN is $0, l/R_1, l/R_2$ and $l/R_2 + T_0$, respectively. Accordingly, the probability for delay $T_1$ is given by

$$\begin{align*}
&\left\{ Pr\{T_1 = 0\} = Pr_{1,f}; \\
&Pr\{T_1 = l/R_1\} = (1 - Pr_{1,f}) Pr\{N_{1,f} > 0\}; \\
&Pr\{T_1 = l/R_2\} = (1 - Pr_{1,f}) Pr\{N_{1,f} = 0\} Pr_{2,f}; \\
&Pr\{T_1 = l/R_2 + T_0\} = (1 - Pr_{1,f}) Pr\{N_{1,f} = 0\} (1 - Pr_{2,f});
\end{align*}$$

(28)

where $T_0$ is the consumed time for fetching content from CN. From the proof of Lemma 1, it has $Pr\{N_{1,f} = 0\} = e^{-\lambda_1 \pi d^2}$. Hence with the probability $p_f$ for requesting content $f$, the average service delay $\bar{T}_1$ for typical UT in SCN is deduced in (26). Similarly, we can derive the expression for $\bar{T}_2$. □

Since all the rates $R_i (i = 1, 3, 4, 5)$ except $R_2$ are determined by $\Theta_1$ and $\Theta_2$, it is necessary to investigate how the performance of average service delay behaves with adjusting caching capacity.

Proposition 2: For fixed caching policy $\Pi_1(\Pi_2)$, the probability $Pr_{1,f}(Pr_{2,f})$ is non-decreasing with $\Theta_1(\Theta_2)$, while $\bar{T}_{1,2}$ are non-increasing with $\Theta_2$ when $\frac{\lambda_1}{\lambda_2} \leq \frac{\alpha_N}{\alpha_N} < \frac{\lambda_2}{\lambda_2} + T_0$.

Proof: When $\Theta_1(\Theta_2)$ increases, more contents can be stored at UTs(SBSs). Hence $Pr_{1,f}(Pr_{2,f})$ will grow when $\Pi_1(\Pi_2)$ is fixed. This can also be shown by looking at the derivative of $Pr_{1,f}$ for the policies described in Section V. Considering the average service delay $\bar{T}_{1,2}$, the derivative satisfies $\frac{\partial \bar{T}_{1,2}}{\partial Pr_{2,f}} < 0$. Since $Pr_{2,f}$ is non-decreasing with $\Theta_2$

$^2$ The average service delay for UTs within cluster $\mathcal{A}$ can also be analyzed in the similar way to the typical UT. Due to space limitation, we omit the details here.
from Proposition 1, it can be concluded that $T_1$ is non-increasing with $\Theta_2$. As for the UDN scenario, it can be shown that when $\frac{1}{T_i} \leq \frac{1}{T_0} + T_0$, the average service delay $T_2$ decreases with $Pr_{2,f}$ (i.e., $\frac{\partial T_2}{\partial Pr_{2,f}} \leq 0$), while $\frac{\partial T_2}{\partial Pr_{1,f}} \leq 0$ under the condition $\frac{1}{T_i} \leq \frac{1}{T_0} \leq \frac{1}{T_i} + T_0$. □

The condition presented in Proposition 2 demonstrates that for a typical UT in cache-enabled UDN, the delay of fetching content from the cache of SBS is smaller than D2D content share, while that of fetching content from CN is the largest. This condition can be easily satisfied when the density of SBSs is comparable with UTs in UDN. From the expressions of $T_1$ and $T_2$, it is easy to find out that the advantage of enabling caching at SBSs is to offload backhauling to CN. Meanwhile the caching capability of UTs can determine the rates of UT both in clustered mmWave SCN and UDN. In what follows, we study the caching policy which can minimize the service delay for typical UT.

Proposition 3: We can obtain the optimal caching scheme that minimizes average service delays $T_{1,2}$ by solving

\begin{align}
    (P1): & \quad \min_{Pr_1, Pr_2} T_{1,2}, \quad \text{(29a)} \\
    & \text{s.t.} \quad \sum_{f \in \mathcal{F}} Pr_{1,f} = \Theta_1, \quad \sum_{f \in \mathcal{F}} Pr_{2,f} = \Theta_2; \quad \text{(29b)} \\
    & \quad 0 \leq Pr_{1,f} \leq 1, \quad 0 \leq Pr_{2,f} \leq 1, \quad f \in \mathcal{F}. \quad \text{(29c)}
\end{align}

Proof: The constraint (29c) is due to that $Pr_{1,f}$ and $Pr_{2,f}$ are probabilities. Suppose the constraints (29b) are not active, i.e., $\sum_{f \in \mathcal{F}} Pr_{1,f} < \Theta_1, \sum_{f \in \mathcal{F}} Pr_{2,f} < \Theta_2$ where $Pr_{1,2}^*$ are optimal solution for (P1). Then we can always find $i$ with $0 < Pr_{1,i} \leq 1$, and $\epsilon \in [0, \min(1 - Pr_{1,i}, \Theta_1 - \sum_{f \in \mathcal{F}} Pr_{1,f})]$. Since $T_1$ increases over $Pr_{1,f}$ with fixed $Pr_2$, we can substitute $Pr_{1,i}^*$ with $Pr_{1,i}^* + \epsilon$ to further reduce $T_1$. Similarly, with fixed $Pr_{1,i}^*$, we can also introduce a small increment on some $Pr_{2,f}$ to reduce $T_2$. Hence the constraint (29b) must be active for the optimum. Following the same clue, it can be proved that the constraint (29b) should also be active in minimizing $T_2$. □

Since $Pr_{1}$ and $Pr_{2}$ are coupled variables in objective (29a), the caching strategies of UTs and SBSs should be optimized jointly. We adopt an Alternate Optimization (AO) approach to solve (P1), which is concluded in Alg. 1. At the first step of each iteration, we optimize the caching policy of UTs with fixed caching scheme of SBSs. Then the caching policy of SBSs is optimized with updated $Pr_{1}$. At the second step, it can be verified that the average service delay reduces at each iteration of Alg. 1. In high SINR region, where $P_1$ and $P_2$ are assumed to be large enough to make $R_i(f) \rightarrow R_{\text{max}}(i \in \{1, 2, 3, 4, 5\}, \forall f \in \mathcal{F})$, we can obtain the approximation $T_{1,2}$ by applying $R_i \rightarrow R_{\text{max}}$ in (26) and (27).

Remark 1: By substituting $T_{1,2}$ with $T_{1,2}$ in (29a), the iterates $(Pr_{1}^k, Pr_{2}^k)$ generated in Algorithm 1 converges monotonically.

Proof: Since the objectives $T_{1,2}$ are convex over $Pr_{1}$ when $Pr_{2}$ is fixed, and linearly decrease over $Pr_{2}$ with fixed $Pr_{1}$, in each sub-step of one iteration in Algorithm 1, the objectives $T_{1,2}$ are monotonically non-increasing due to the minimization. Since the service delays $T_{1,2}$ are bounded below, the sequence $\{T_{1,2}(Pr_{1}^k, Pr_{2}^k)\} \in \mathbb{R}$ converges to a limit value [55]. □

It should be noted that Alg. 1 is a Probabilistic Caching (PC) based policy. Different from the PC scheme provided in [56], we jointly optimize the caching policy for SBSs and UTs to minimize the service delay for typical UT. With the output $Pr_{1}(Pr_{2})$ from Alg. 1, UTs(SBSs) randomly build a list of up to $\Theta_1(\Theta_2)$ contents to be cached as [56].

V. NUMERICAL RESULTS

In what follows, we will provide some numerical results to validate the analysis of distance and SINR distributions, and the ergodic rate for typical UT. Moreover, the effects of caching capacity and policies of SBSs and UTs, as well as the radius of clusters, on the average service delay for both cache-enabled mmWave SCN and UDN with clusters will be provided. The parameters used for simulations are given in Table I.

We consider homogeneous content preference among UTs with $\omega = \mathcal{F}$. Besides, we apply the proposed caching scheme Alg. 1, and the following caching policies for SBSs and UTs:

- **RC**: UTs(SBSs) choose to store contents randomly until $\Theta_1(\Theta_2)$ is occupied, where

\[ Pr_{i,f} = \frac{\Theta_i}{\mathcal{F}}, \quad f \in \mathcal{F}. \]  \hspace{1cm} (30)

- **Most popular caching (MPC)**: UTs(SBSs) choose to store contents according to the order of $\omega$ until $\Theta_1(\Theta_2)$ is occupied, where

\[ Pr_{i,f} = \begin{cases} 1, & f \in \mathcal{F}_{\omega_i} = \{\omega_1, \ldots, \omega_{\Theta_i}\}; \\
0, & f \in \mathcal{F}\setminus\mathcal{F}_{\omega_i}. \end{cases} \]  \hspace{1cm} (31)

- **PC**: UTs(SBSs) store contents up to $\Theta_1(\Theta_2)$ according to the probabilistic content caching method as [56], but with the following probability

\[ Pr_{i,f} = \min\left(\frac{\mu f(\omega_i + \mu^* + 1), f \in \mathcal{F}}{\mathcal{F}}, \Theta_i \right), \]  \hspace{1cm} (32)

where $\mu^*(\geq 0)$ is found to make $\sum_{f \in \mathcal{F}} Pr_{i,f} = \Theta_i$ satisfied.

Since we assume homogeneous content preference and caching capacity for UTs, the contents that are stored at UTs will be unique if the MPC scheme is adopted. Besides, to exploit the diversity of caching content at UTs, we adopt PC scheme for UTs, i.e., $\Pi_1 = PC$. 

\[ \text{Algorithm 1: AO Algorithm for (P1)} \]

1. Initialize: $\{Pr_{1,f} = \frac{\Theta_1}{\mathcal{F}}, f \in \mathcal{F}\}$
2. for Iteration $k = 1, 2, 3, \ldots$ do
3. Step 1: solve (P1) over $Pr_1$ with fixed $Pr_2$
4. Step 2: solve (P1) over $Pr_2$ with fixed $Pr_1$
5. end for
It has content $f$ determined by the density $\lambda$. For the transmission distance is within interval $A$, Distributions of Distance and SINR. Fig. 4 (a). Because the transmission distance is within interval $A$, Distributions of Distance and SINR. Fig. 4 (a). CDFs of transmission distance $F_{X_1,4,5}(x)$. (b) CDF of transmission distance $F_{X_2}(x)$. Fig. 5. CDFs of SINR $F_{1,2,3,4,5}(\gamma)$ for typical UT for both SCN and UDN scenarios.

A. Distributions of Distance and SINR

Fig. 4 presents the CDFs $F_{X_1}(x)$ with radius $d = 20$ m, $\eta = 1.2$, $\Theta_2 = 2$ Gb and $\Pi_2 = \text{PC}$. Since the distribution for $X_1$ and $X_2$ is the same, we show the CDFs of $X_{1,4,5}$ Fig. 4 (a). Because the transmission distance is within interval $(0, d]$, both CDFs have $F_{X_{1,4,5}}(d) = 1$. As for content $f = 10$, the probability $P_{T_1,f=10}$ is less than 1. The UTs which have content $f = 10$ follows the PPP $\Phi_{1,f=10}$ with density $P_{T_1,f=10} \lambda_1 < \lambda_1$. Hence the CDF $F_{X_1}(x)$ is not larger than $F_{X_1}(x)$ with fixed $x$. With enlarging caching capacity $\Theta_1$, the probability $P_{T_1,f=10}$ is non-decreasing according to Proposition 1. This densities the PPP $\Phi_{1,f=10}$, which hence increases the CDFs $F_{X_1}(x)$. The CDF $F_{X_2}(x)$ is only determined by the density of SBS $\lambda_2$. For the CDF $F_{X_2}(x)$, it is always smaller than $F_{X_2}(x)$ for fixed $x$ since $\lambda_2 f \leq \lambda_2$. The CDF $F_{X_2}(x)$ is shown in Fig. 4 (b), which is only determined by the density $\lambda_2$. As the result shown, it has almost probability 1 for the typical UT to connect a SBS within distance 200 m in SCN.

Fig. 5 presents the CDFs $F_{1,2,3,4,5}(\gamma)$ for SINRs of typical UT in clustered SCN and UDN. We set $P_1 = 0$ dBm, $P_2 = 20$ dBm, $\eta = 1.2$, $f = 1$, $\Theta_1 = 0.5$ Gb, $\Theta_2 = 2$ Gb and $d = 10$ m. The CDFs $F_{1,2,3,4,5}(\gamma)$ in theory are obtained by applying integration on (35), (12) and variants of (35) directly. While $F_{1,2,3,4,5}(\gamma)$ in approximation are given in Lemmas 2, 3 and 5. From Fig. 5, it can be concluded that Gaussian-Chebyshev quadrature with $t = 30$ can provide accurate approximation for $F_{1,2,3,4,5}(\gamma)$. For $\alpha_L \in \{2.5, 3\}$ and $\Pi_2 = \{\text{RC}, \text{PC}\}$, the SINR of D2D tier for typical UT in mmWave SCN is expected to be larger than that of SBS tier and the SINR of typical UT in mmWave UDN. This is due to the truth that the interference experienced by the typical UT in mmWave UDN is stronger than that in SCN. When the pathloss exponent of LOS $\alpha_L$ increases to 3, the performance of SINR for all scenarios become worse. While the CDF $F_{1,2,3,4,5}(\gamma)$ keeps the same since it is only determined by the pathloss exponent of NLOS $\alpha_N$. For the considered caching setup, the SINRs $F_{1,2,3,4,5}(\gamma)$ in both SCN and UDN cenarios with caching policy $\Pi_2 = \text{PC}$ are larger than those achieved by $\Pi_2 = \text{RC}$ policy. This is because with $P_{T_2,f=1,\Pi_2=\text{PC}} \geq P_{T_2,f=1,\Pi_2=\text{RC}}$, more SBSs will store content $f = 1$ when policy PC is adopted rather than RC. Hence the typical UT is likely to retrieve content $f = 1$ from a closer SBS with PC than RC, i.e., $E_{X_2,f=1,\Pi_2=\text{PC}}[x] \leq E_{X_2,f=1,\Pi_2=\text{RC}}[x]$ and $E_{X_4,f=1,\Pi_2=\text{PC}}[x] \leq E_{X_4,f=1,\Pi_2=\text{RC}}[x]$.

B. Ergodic Rates

We evaluate ergodic rates $R_1$ and approximations $R_1^+$ with $\Pi_2 = \text{PC}$, $d = 20$ m, $\eta = 1.2$, $\alpha_L = 3$ and $\alpha_N = 3.5$. From the results presented in Fig. 5, we show that adopting PC policy at SBSs can achieve better SINR performance than RC for popular contents (e.g., $f = 1$). Hence we only consider the $\Pi_2 = \text{PC}$ policy here. The ergodic rates $R_1,3$ and $R_{1,3}^+$ are shown in Fig. 6 (a). As $P_1$ increases from $-40$ dBm to $-10$ dBm, the ergodic rate $R_1(f = 2)$ and $R_3(f = 2)$ will grow to the maximum achievable rate $R_{\text{max}} = 2$ Gbps. Moreover, when $P_1$ is higher than $-30$ dBm, the approximation
Tier transmission power UT in clustered mmWave SCN and UDN w.r.t. the D2D channel.

Average Service Delay

For fixed radius \(d\), different caching policies satisfy the gap from theory \(R_1 = 10\) dBm, \(R_2 = 10\) dBm, respectively.

In Fig. 7, we evaluate the average service delay for typical UT in mmWave UDN is smaller than that in mmWave SCN. Hence the advantage brought by caching unit becomes weaken, which leads to a larger service delay.

Finally, we present the impact of radius \(d\) of clusters.

### C. Average Service Delay

In Fig. 7 we evaluate the average service delay for typical UT in clustered mmWave SCN and UDN w.r.t. the D2D tier transmission power \(P_1\). In the simulation, we set \(P_2 = 10\) dBm, \(\eta = 1.2\), \(\Theta_1 = 2\) Gb, \(\Theta_2 = 2\) Gb and \(\alpha_L = 2.5\).

For fixed radius \(d = 10\) m, the average service delay for different caching policies satisfy \(T_{1,\text{MPC}} < T_{1,\text{PC}} < T_{1,\text{RC}}\) and \(T_{2,\text{MPC}} < T_{2,\text{PC}} < T_{2,\text{RC}}\). When the caching policies PC and RC are adopted by SBSs, the average service delay of typical UT in mmWave UDN is smaller than that in mmWave SCN, i.e., \(T_2 < T_1\), when \(P_1\) is larger than \(-15\) dBm with \(d = 10\) m. But with \(\Pi_2 = \text{MPC}\) at SBSs and \(P_1 \in [-40, -10]\) dBm, the average delay of typical UT in clustered UDN is smaller than that in SCN with \(d = 10\) m. It is because that with \(\Pi_2 = \text{MPC}\), more requests from the typical UT are served by the D2D tier both in clustered SCN and UDN than cases \(\Pi_2 \in \{\text{RC, PC}\}\). Furthermore, the interference for the typical UT in mmWave UDN is stronger than that in SCN. Hence the average delay presents the trend \(T_{2,\text{MPC}} < T_{2,\text{MPC}}\) when \(P_1\) is small enough, which can also explain the case \(d = 15\) m.

By fixing \(\Pi_2 = \text{PC}\), \(P_1 = 0\) dBm and \(P_2 = 20\) dBm, we present the average service delay over the caching capacity \(\Theta_2\) in Fig. 8. As shown in this figure, both delay \(T_1\) and \(T_2\) decrease with enlarging \(\Theta_2\). This is because larger \(\Theta_2\) can offload more backhauling to CN. Besides, the delay can be reduced by enlarging \(\Theta_1\). This phenomenon can be explained from two aspects. With a larger \(\Theta_1\), more requests from typical UT can be served by its local cache, which consumes 0 s delay. Besides, the expected distance for D2D tier transmission will be shorter since the requested contents have higher probability to be cached by the neighboring UTs within the same cluster. When the shape parameter of Zipf-like distribution reduces from 1.2 to 0.9, the content preference of UTs becomes less concentrated. Hence the advantage brought by caching unit becomes weaken, which leads to a larger service delay.

Finally, we present the impact of radius \(d\) of clusters on the service delay for typical UT in SCN and UDN in Fig. 9, with \(P_1 = 0\) dBm, \(P_2 = 20\) dBm, \(\alpha_L = 2.5\), \(\Theta_1 = 0.5\) Gb and \(\Theta_2 = 2\) Gb. When caching policies PC, RC and MPC are adopted at SBSs, the delays \(T_{1,2}\) decrease with \(d\) in range \([5, 15]\) m. The reason is that the increased amount of D2D tier content sharing can reduce the delay. Besides, the interference for typical UT is non-increasing with radius \(d\) for both mmWave SCN and UDN scenarios. However, service delay \(T_{2,\text{RC}}, T_{2,\text{PC}}\) and \(T_{2,\text{MPC}}\) degrade by enlarging cluster size to \(d > r_d = 15\) m. This is because multiple D2D links are allowed when \(d > r_d\), and the D2D transmission bandwidth for the typical UT reduces due to spectrum partition. When the cluster size gets larger, we have...
For simplicity, we refer to $b_{in}$ as the number of UEs falling in $b(o,x)$, where $b(o,x)$ is the disc of radius $x$ centered at $o$. For simplicity, we refer to $N_{1,2}(b(o,d))$ as $N_{1,2}$. The D2D tier transmission happens only when the requested content exists in cluster $A$. Considering $P_{1,1}$, which represents the probability that $f$ exists at any UEs, we then the $P_{1}$ by deleting each UT with probability $P_{1,1}$ to obtain a new PPP denoted as $F_{1,1}$, which has density $\lambda_{1,1} = P_{1,1} \lambda_{1}$ and consists of the UEs with $f$. Denote $N_{1,1}(b(o,x))$ as the number of UT in $F_{1,1}$ located within disc $b(o,x)$ while $N_{1,1,1}(b(o,d))$ is as the number of UT in $F_{1,1}$ located within disc $b(o,d)$ while $N_{1,1,1}(b(o,d))$ is. The probability $\Pr\{X_{1} \leq x|Y_{1,1} \geq 1\}$ conditioned on $Y_{1,1} \geq 1$ implies that there exists at least one UT with content $f$ in area $A$ with radius $d$, which is expressed as

$$\Pr\{X_{1} \leq x|N_{1,1,1} \geq 1\} = 1 - \Pr\{X_{1} > x|\exists Y_{1,1} \geq 1\}$$

VI. CONCLUSIONS

We studied clustering schemes for cache-enabled mmWave SCN and UDN through investigating the ergodic rate and service delay for typical UT in a cluster. By defining the clusters as disjoint discs with equal size, the closed-form expressions of SINR distribution and ergodic rates in high SINR region are derived for both scenarios, which are characterized by the radius $d$ of clusters, caching capacity of UEs and SBSs, and caching policies. Besides, we propose a PC based caching policy which jointly optimize the caching scheme for SBSs and UEs to minimize the service delay. Through analysis and simulations, we show that the average service delay of typical UT can be reduced by enlarging the caching capacity of UEs and SBSs. Moreover, we reveal that the service delay in cache-enabled mmWave UDN with clusters can lead a conditional performance gain compared with those in mmWave SCN due to deploying dense SBSs. It shows that the proposed caching scheme can further reduce the service delay of typical UT compared with benchmark policies. Moreover, simulation results show that there exists a best radius $d$ on average service delay in mmWave SCN, which can guide the design of the clustering scheme.

APPENDIX A

PROOF OF LEMMA 1

Denote $N_{1,2}(b(o,x))$ as the number of UT or SBS falling in $b(o,x)$, which is the disc of radius $x$ centered at $o$. For simplicity, we refer to $N_{1,2}(b(o,d))$ as $N_{1,2}$. The D2D tier transmission happens only when the requested content exists in cluster $A$. Considering $P_{1,1}$, which represents the probability that $f$ exists at any UEs, we then the $P_{1}$ by deleting each UT with probability $P_{1,1}$ to obtain a new PPP denoted as $F_{1,1}$, which has density $\lambda_{1,1} = P_{1,1} \lambda_{1}$ and consists of the UEs with $f$. Denote $N_{1,1,1}(b(o,x))$ as the number of UT in $F_{1,1}$ located within disc $b(o,x)$ while $N_{1,1,1}(b(o,d))$ is as the number of UT in $F_{1,1}$ located within disc $b(o,d)$ while $N_{1,1,1}(b(o,d))$ is. The probability $\Pr\{X_{1} \leq x|Y_{1,1} \geq 1\}$ conditioned on $Y_{1,1} \geq 1$ implies that there exists at least one UT with content $f$ in area $A$ with radius $d$, which is expressed as

$$\Pr\{X_{1} \leq x|N_{1,1,1} \geq 1\} = 1 - \Pr\{X_{1} > x|\exists Y_{1,1} \geq 1\}$$

APPENDIX B

PROOF OF LEMMA 2

The CDF of $G_{1}$ can be expressed as

$$F_{G_{1}}(\gamma) = 1 - \Pr \left\{ \frac{P_{1} G_{0} x^{-\alpha L} |h|^{2}}{I_{1} + \sigma^{2} x^{2}} > \gamma \right\}$$

$$= 1 - \Pr \left\{ |h|^{2} > \frac{(I_{1} + \sigma^{2} x^{2}) \gamma}{P_{1} G_{0} x^{-\alpha L}} \right\}$$

$$\approx 1 - \mathbb{E}_{X_{1,1}} \left\{ 1 - e^{-\frac{\mu_{1} x^{2}}{P_{1} G_{0} x^{-\alpha L}}} \right\}$$

$$\approx \sum_{m=0}^{M} \left( \begin{array}{c} M \\ m \end{array} \right) (-1)^{m} \mathbb{E}_{X_{1}} \left\{ e^{-\frac{\mu_{1} x^{2}}{P_{1} G_{0} x^{-\alpha L}}} \right\}$$

where (a) follows from the CDF approximation of normalized gamma random variable $|h|^{2} \sim \text{Gamma}(M, \frac{1}{\sigma^{2}})$ [52], and (b) follows from the Binomial Theorem. The term $B_{1}$ in (35) is given by

$$B_{1} = e^{-\frac{\mu_{1} x^{2}}{P_{1} G_{0} x^{-\alpha L}}}$$

$$= e^{-\frac{\mu_{1} x^{2}}{P_{1} G_{0} x^{-\alpha L}}}$$

$$= e^{-\frac{\mu_{1} x^{2}}{P_{1} G_{0} x^{-\alpha L}}}$$

$$= e^{-\frac{\mu_{1} x^{2}}{P_{1} G_{0} x^{-\alpha L}}}$$

$$= e^{-\frac{\mu_{1} x^{2}}{P_{1} G_{0} x^{-\alpha L}}}$$

$$= e^{-\frac{\mu_{1} x^{2}}{P_{1} G_{0} x^{-\alpha L}}}$$
\[
(F) = e^{-\pi \lambda_2 \sum_{i,k \in M} P_{tk} d^2 \left( x F_i \left( M, -\frac{1}{N} - \frac{1}{N}; -\gamma \psi_{1,ik} x^\alpha L d^{-\alpha N} \right) \right)},
\]

(36)

where (c) follows from the expectation of the antenna gain \( G \), (d) follows from the moment-generating function of gamma distribution, (e) follows from the probability generating functional of a PPP \([58]\), and (f) is from the property of Gaussian Hypergeometric function and holds for \( \alpha_1, \alpha_N > 2 \).

Then plugging (36) into (6) and taking expectation over \( X_1 \), we obtain the CDF as

\[
F_{11}(\gamma) \approx \sum_{m=0}^{\infty} \frac{M^m}{m!} (-1)^m \int_0^d f_{X_1}(x) e^{-\frac{mk_g x^{2-\alpha}}{P_1 G_0}} \times e^{-\pi \lambda_2 \sum_{i,k \in M} P_{tk} d^2 \left( x F_i \left( M, -\frac{1}{N} - \frac{1}{N}; -\gamma \psi_{1,ik} x^\alpha L d^{-\alpha N} \right) - 1 \right) dx}
\]

\[
(37)
\]

where (g) is obtained by invoking approximation of Gaussian-Chebyshev quadrature, which follows \( \int_1^\infty \sqrt{1-x^2} g(x) dx \approx \sum_{i=1}^t w_i g(u_i) \) and \( \Psi_{1,1,1,1} \) are given in (7) and (8) respectively.

**APPENDIX C**

**PROOF OF COROLLARY 1**

From (35) and (36), the CDF of \( I_1 \) in high SINR region can be derived as

\[
F_{11}(\gamma) \approx \mathbb{E}_{X_1, I_1} \left[ \left( 1 - e^{-\frac{k_g x^{2-\alpha}}{P_1 G_0}} \right)^M \right]
\]

(a)

\[
= \mathbb{E}_{X_1} \left[ \mathbb{E}_{x \in D_1} \left[ \sum_{s \in D_1} G_{s,i} h_{i,s} \frac{M \gamma P_{0} G_{s} |h_{i,s}|^2}{P_{1} G_{0x} x^\alpha L r_s^\alpha} \right] \right]
\]

+ \frac{M \gamma^2}{P_{1} G_{0x} x^\alpha L}

\[
= \mathbb{E}_{X_1} \left[ \mathbb{E}_{x \in D_1} \left[ \sum_{s \in D_1} \sum_{k \in M} M^2 P_{tk} |\psi_{1,ik}| \frac{x^{\alpha L} r_s^{\alpha}}{x^\alpha L r_s^{\alpha}} \right] \right]
\]

+ \frac{M \gamma^2}{P_{1} G_{0x} x^\alpha L}

\[
= \mathbb{E}_{X_1} \left[ \frac{2\pi \lambda_2 d^{2-\alpha N} \sum_{k \in M} M^2 P_{tk}}{\alpha_N - 2} \right]
\]

(b)

where (a) follows from the approximation of \( \lambda \exp(-x)^M \approx \lambda x \) when \( x \to 0 \), (35) is the gamma function, and (b) is since the truth that \( \mathbb{E} \left[ \sum_{x \in D} f(x) \right] = \lambda \int_{x=2} f(x) dx \), where \( \lambda \) is the density of PPP \( \Phi \). Then substituting (38) into (19) the desired result for \( R_1^* \) can be obtained. Similarly integrating Lemmas 3 and 5, the expressions for \( R_2^* \) can be similarly derived but with different constant \( C_2 \).

**REFERENCES**

[1] Cisco Visual Networking Index, “Global mobile data traffic forecast update, 2016-2021,” Cisco, San Jose, CA, USA, White Paper, Mar. 2017.

[2] D. Mavarakis and M. Kamal-Saad, “Mobile broadband access at home: The business case for femtocells, UMA and IMS/VC dual mode solutions,” in Informa Telecoms Media. Sep. 2008.

[3] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” IEEE Trans. Inf. Theory, vol. 60, no. 5, pp. 2856–2867, May 2014.

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

[5] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, “FemtoCaching: Wireless content delivery through distributed caching helpers,” IEEE Trans. Inf. Theory, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[6] D. Liu, B. Chen, C. Yang, and A. F. Molisch, “Caching at the wireless edge: Design aspects, challenges, and future directions,” IEEE Commun. Mag., vol. 54, no. 9, pp. 22–28, Sep. 2016.

[7] N. Zhao, X. Liu, F. R. Yu, M. Li, and V. C. M. Leung, “Communications, caching, and computing oriented small cell networks with interference alignment,” IEEE Commun. Mag., vol. 54, no. 9, pp. 29–35, Sep. 2016.

[8] E. Zeydan et al., “Big data caching for networking: Moving from cloud to edge,” IEEE Commun. Mag., vol. 54, no. 9, pp. 36–42, Sep. 2016.

[9] Y. Ye, Z. Zhang, G. Yang, and M. Xiao, “Minimum cost based clustering scheme for cooperative wireless caching network with heterogeneous file preference,” in Proc. IEEE Int. Conf. Commun. (ICC), May 2017, pp. 1–6.

[10] N. Zhao et al., “Caching UAV assisted secure transmission in hyperdense networks based on interference alignment,” IEEE Trans. Commun., vol. 66, no. 5, pp. 2281–2294, May 2018.

[11] F. Cheng, G. Gui, N. Zhao, Y. Chen, J. Tang, and H. Sari, “UAV-relaying-assisted secure transmission with caching,” IEEE Trans. Commun., vol. 67, no. 5, pp. 3140–3153, May 2019.

[12] M. Gregori, I. Gómez-Vilarelle, J. Matamoros, and D. Glündüz, “Wireless content caching for small cell and D2D networks,” IEEE J. Sel. Areas Commun., vol. 34, no. 5, pp. 1222–1234, May 2016.

[13] M. Ji, G. Caire, and A. F. Molisch, “Fundamental limits of caching in wireless D2D networks,” IEEE Trans. Inf. Theory, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[14] Y. Guo, L. Yuan, and R. Zhang, “Cooperative local caching under heterogeneous file preferences,” IEEE Trans. Commun., vol. 65, no. 1, pp. 444–457, Jan. 2017.

[15] L. Zhang, M. Xiao, G. Wu, and S. Li, “Efficient scheduling and power allocation for D2D-assisted wireless caching networks,” IEEE Trans. Commun., vol. 64, no. 6, pp. 2438–2452, Jun. 2016.

[16] Z. Chen, N. Pappus, and K. Kontoulis, “Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal,” IEEE Commun. Lett., vol. 21, no. 3, pp. 584–587, Mar. 2017.

[17] R. Amer, M. M. Butt, M. Bennis, and N. Marchetti, “Inter-cluster cooperation for wireless D2D caching networks,” IEEE Trans. Wireless Commun., vol. 17, no. 9, pp. 6108–6121, Sep. 2018.

[18] N. Deng and M. Haenggi, “The benefits of hybrid caching in Gauss-Poisson D2D networks,” IEEE J. Sel. Areas Commun., vol. 36, no. 6, pp. 1217–1230, Jun. 2018.
[19] Z. Zhou, M. Dong, K. Otu, J. Wu, and T. Sato, “Energy efficiency and spectral efficiency tradeoff in device-to-device (D2D) communications,” *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 485–488, Oct. 2014.

[20] N. Golrezaee, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, “Base-station assisted device-to-device communications for high-throughput wireless video networks,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.

[21] B. Peng, T. Peng, Z. Liu, Y. Yang, and C. Hu, “Cluster-based multicast transmission for device-to-device (D2D) communication,” in *Proc. IEEE 7th Veh. Technol. Conf. (VTC Fall)*, Sep. 2013, pp. 1–5.

[22] R. Amer, M. M. Butt, H. Elsawy, M. Bennis, J. Kibiłda, and N. Marchetti, “On minimizing energy consumption for D2D clustered caching networks,” in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[23] B. Peng, H. Elsawy, J. Kibiłda, M. M. Butt, and N. Marchetti, “Cooperative transmission and probabilistic caching for clustered D2D networks,” in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–6.

[24] R. Amer, H. Elsawy, M. M. Butt, E. A. Jorswiec, M. Bennis, and N. Marchetti, “Optimized caching and spectrum partitioning for D2D enabled cellular systems with clustered devices,” *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3558–3574, Jul. 2017.

[25] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, Jr., “Modeling and analyzing millimeter wave systems,” *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 403–430, Jan. 2017.

[26] M. Cheng, J.-B. Wang, Y. Wu, X.-G. Xia, K.-K. Wong, and M. Lin, “Coverage analysis for millimeter wave cellular networks with imperfect beam alignment,” *IEEE Veh. Technol. Conf.*, vol. 67, no. 9, Sep. 2018.

[27] Y. Zhou, V. W. S. Wong, and R. Schober, “Coverage and rate analysis of millimeter wave NOMA networks with beam misalignment,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8211–8227, Dec. 2018.

[28] N. Giatsoglou, K. Ntontin, E. Kartalaki, A. Antonopoulos, and C. Verikoukis, “D2D-aware device caching in mmWave-cellular networks,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2025–2037, Sep. 2017.

[29] Y. Zhu, G. Zheng, L. Wang, K. Wong, and L. Zhao, “Content placement in cache-enabled sub-6 GHz and millimeter-wave multi-antenna dense small cell networks,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2843–2858, May 2018.

[30] W. Yi, Y. Liu, and A. Nallanathan, “Cache-enabled HetNets with millimeter wave small cells,” *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5497–5511, Nov. 2018.

[31] J. Qiao, Y. He, and X. S. Shen, “Proactive caching for mobile video streaming in millimeter wave 5G networks,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 7187–7198, Oct. 2016.

[32] M. Kamel, W. Hamouda, and A. Youssef, “Ultra-dense networks: A survey,” *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2522–2545, 4th Quart., 2016.

[33] C. Yang, Y. Yao, Z. Chen, and B. Xia, “Analysis on cache-enabled wireless heterogeneous networks,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.

[34] J. Hachem, N. Karamchandani, and S. Diggavi, “Content caching and delivery over heterogeneous wireless networks,” in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2015, pp. 756–764.

[35] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, “Cooperative caching and transmission design in cluster-centric small cell networks,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.

[36] M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-Aho, “Content-aware user clustering and caching in wireless small cell networks,” in *Proc. 11th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2014, pp. 945–949.

[37] F. Song et al., “Probabilistic caching for small-cell networks with terrestrial and aerial users,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 9162–9177, Sep. 2019.

[38] J. Zhang, X. Zhang, Z. Yan, Y. Li, W. Wang, and Y. Zhang, “Social-aware cache information processing for 5G ultra-dense networks,” in *Proc. 8th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2016, pp. 1–5.

[39] J. Liu and S. Sun, “Energy efficiency analysis of dense small cell networks with caching at base stations,” in *Proc. 2nd IEEE Int. Conf. Comput. Commun. (ICCC)*, Oct. 2016, pp. 2944–2948.

[40] L. Sun, W. Wang, Y. Jia, and S. Fu, “A cluster-based energy-efficient resource management scheme for ultra-dense networks,” *IEEE Access*, vol. 4, pp. 6823–6832, 2016.
Shaocheng Huang (Student Member, IEEE) received the bachelor's degree in electronic and information engineering from the Changsha University of Science and Technology, Changsha, China, in 2013. He is currently pursuing the Ph.D. degree with the Division of Information Science and Engineering, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. His current research interests include massive MIMO, millimeter-wave communications, machine learning, and physical layer security.

Ming Xiao (Senior Member, IEEE) received the bachelor's and master's degrees in engineering from the University of Electronic Science and Technology of China, Chengdu, in 1997 and 2002, respectively, and the Ph.D. degree from the Chalmers University of Technology, Sweden, in November 2007. From 1997 to 1999, he worked as a Network and Software Engineer at ChinaTelecom. From 2000 to 2002, he held a position at the Sichuan Communications administration. Since November 2007, he has been with the Department of Information Science and Engineering, School of Electrical Engineering and Computer Science, Royal Institute of Technology, Sweden, where he is currently an Associate Professor. He was an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS (2012–2017), IEEE COMMUNICATIONS LETTERS (Senior Editor since January 2015), and IEEE WIRELESS COMMUNICATIONS LETTERS (2012–2016), and has been an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since 2018. He was the Lead Guest Editor of IEEE JSAC Special issue on “Millimeter Wave Communications for future mobile networks” in 2017. He has been an Area Editor of IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY since 2019.

Zheng Ma (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees in communications and information systems from Southwest Jiaotong University, in 2000 and 2006, respectively. He was a Visiting Scholar with the University of Leeds, U.K., in 2003. In 2003 and 2005, he was a Visiting Scholar with The Hong Kong University of Science and Technology. From 2008 to 2009, he was a Visiting Research Fellow with the Department of Communication Systems, Lancaster University, U.K. He is currently a Professor with Southwest Jiaotong University, and also serves as the Deputy Dean for the School of Information Science and Technology. His research interests include information theory and coding, signal design and applications, FPGA/DSP implementation, and professional mobile radio. He has authored over 120 research papers in high-quality journals and conferences. He is also the Vice-Chairman of the Information Theory Chapter in IEEE Chengdu Section. He received the Marie Curie Individual Fellowship in 2018. He is currently the Area Editor of IEEE COMMUNICATIONS LETTERS.

Mikael Skoglund (Fellow, IEEE) received the Ph.D. degree from the Chalmers University of Technology, Sweden, in 1997.

In 1997, he joined the Royal Institute of Technology (KTH), Stockholm, Sweden, where he was appointed to the Chair in Communication Theory in 2003. At KTH, he heads the Division of Information Science and Engineering, and the Department of Intelligent Systems. He has worked on problems in source-channel coding, coding and transmission for wireless communications, Shannon theory, information and control, and statistical signal processing. He has authored and coauthored over 160 journal and 380 conference papers.

Dr. Skoglund was the General Co-Chair for IEEE ITW 2019, and will serve as the TPC Co-Chair for IEEE ISIT 2022. He has served on numerous technical program committees for IEEE sponsored conferences. He was an Associate Editor of IEEE TRANSACTIONS ON COMMUNICATIONS (2003–2008) and was on the editorial board of IEEE TRANSACTIONS ON INFORMATION THEORY (2008–2012).