This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

Radio Resource Sharing and Edge Caching with Latency Constraint for Local 5G Operator: Geometric Programming Meets Stackelberg Game

Tachporn Sanguanpuak, Dusit Niyato, Nandana Rajatheva, Matti Latva-aho

***Abstract*—The rapidly increasing demand in indoor small cell networks has given rise to the concept of local 5G operator (OP) for local service delivery. In this regard, we develop a novel game-theoretic framework with geometric programming to model and analyze cache-enabled small cell base stations (SBSs) with infrastructure sharing for local 5G OP networks. In such a network, the local 5G OP provides wireless network in indoor area and rent out the infrastructure which are RAN and cache storage to multiple mobile network operators (MNOs) while guarantee the quality-of-experience (QoE) at the users (UEs) of MNOs. We formulate a Stackelberg game model where the local 5G OP is the leader and the MNOs are the followers. The local 5G OP aims to maximize its profit by optimizing its infrastructure rental fee, and the MNOs aim to minimize their renting cost of infrastructure by minimizing the "cache intensity" subject to latency constraint at each UE. Here, the cache intensity is defined as the product of the number of SBSs per unit area and the number of popular files stored in each SBS. The optimization problems of the local 5G OP and the MNOs are transformed into geometric programming. Accordingly, the subgame perfect equilibrium of Stackelberg game is obtained through the succesive geometric programming (SGP) method. Since the MNOs share their rented infrastructure, for cost sharing, we apply the concept of Shapley value to divide the cost among the MNOs. We show that the cost sharing problem can be mapped into a simplified "airport runway cost sharing problem", in which the Shapley value can be computed efficiently. Finally, we present an extensive performance evaluation that reveals interesting insights into designing resource sharing with edge caching in local 5G OP networks.**

***Index Terms*—5G, beyond 5G (B5G), edge caching, latency constraint, local 5G operator, micro-operator, stochastic geometry, geometric programming, Stackelberg game, Shapley value.**

I. Introduction

A. Motivation

The 5G and beyond 5G (B5G) technologies will need to support extremely diverse use-cases for example, (i) Extreme mobile broadband (xMBB) with data rates up to several Gbps, more videos, more live streaming and reliable broadband access over large coverage areas. (ii) Massive machine type communications (mMTC) which is a service category consisting of sensing, tagging, and monitoring require high connection density. (iii) Ultra reliable low latency (uRLLC) which is a service category to support the latency-sensitive services such as, remote control, autonomous driving car and tactile Internet [2]. Industrial, manufacturing companies, sport arenas and smart hospitals cannot completely rely on unlicensed wireless band. Also, traditional macro cellular networks deployed by the mobile network operators (MNOs) is insufficient to rapidly serve the UEs in an indoor area with the quality of experience (QoE)/quality of service (QoS) guaranteed.

Since 80% of traffic of the above applications is generated from indoor areas [2], the new business model of the MNOs needs to be developed for local service delivery with specific requirements in indoor [3]–[5]. In this regard, the most prominent and efficient solution is the deployment of local 5G operator (OP) as to offer services in specific indoor with locally licensed spectrum. The local 5G OP business model becomes very promising in ultra dense networks deployment as (i) reduce the network total cost of the MNOs and (ii) the UEs can be served with the QoE/QoS satisfied. The facility owner with the capability of deploying small cell base stations (SBSs) in an indoor with licensed band can become local 5G OP [6]–[7]. Recently, Qualcomm has proposed the concept of local 5G services for industrial automation with reliability/latency satisfied [8]. The overview of local infrastructure provided by the local 5G OP (micro-OP) to serve IoT devices was studied in [9].

The local 5G OP can serve the MNO's UEs with licensed bands while renting out the SBSs to the MNOs. In this regard, the resource/network virtualization will be implemented for the local 5G OP to reduce capital expenditure (CAPEX) and operational expenses (OPEX) and improve network resource utilization. Furthermore, the deployment of local 5G OP can support the possible use cases of B5G, for example, using short-range communication in high frequencies of licensed band to provide extreme data rate in indoor areas, i.e., terabit/sec for downlink transmission [10]. Under the IEEE 802.15 standard, the data rate 100 gigabit/sec is enough for virtual/augmented reality and kiosk for a single UE, however, it is not enough for multiple simultaneous UEs and instantaneous large file download. Therefore, local 5G OP with proper infrastructure is essential to serve UEs with such a very high data rate.

However, none of the existing work has formally formulated the business model of local 5G OP and multiple MNOs using

T. Sanguanpuak, N. Rajatheva, and M. Latva-aho are with 6G Flagship, Centre for Wireless Communications (CWC), University of Oulu, Finland (emails: {tachporn.sanguanpuak, nandana.rajatheva, matti.latva-aho}@oulu.fi). D. Niyato is with School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore (email: dniyato@ntu.edu.sg).

A part of the paper appeared in IEEE Globecom 2018 [1].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

game theoretical approaches while taking latency constraint at each UE into account. In [11], the authors studied resource sharing through simulations in which spectrum including mmWave band and/or network infrastrucure resources can be shared by MNOs. The stochastic geometry modeling for BS placement with single/multiple sellers and multiple buyer MNOs and Cournot oligopoly game was proposed in [12]–[13]. Multiple-MNO spectrum sharing using matching game for small cell networks was explored in [14]. Apart from the efficient resource utilization, achieving low latency for xMBB with partial/full virtual reality is also a critical challenge for the MNOs. To tackle this challenge, the concept of proactive caching was introduced [16]–[19] in which popular contents are stored at the edge/radio access network (RAN), e.g., cache-enabled BSs to reduce the wireless access delay. In [15], the virtualization technique in the downlink transmission of limited fronthaul capacity cloud-radio access networks was considered. The authors formulated an optimization problem to maximize network energy efficiency by a joint design of virtual computing resources, transmit beamforming, remote radio head (RRH) selection, and RRH-UE association.

In the context of economic modeling of caching, the work in [17] used stochastic geometry method to characterize the probability that a UE finds a video file from a content provider in the cache of a BS located closest to it. Using this probability, a Stackelberg game was formulated and solved to maximize the average profit of the network service providers, which act as the leaders, and the content providers, which act as the followers. In [18], the authors considered a Stackelberg game with a single MNO and multiple content providers. The MNO, as the leader, decides on the price to charge to content providers such that the revenue is maximized. The content providers, as the followers, compete with each other to obtain sufficient cache space to improve the QoS to its UEs. In [19], the cache is partitioned into slices and each partition is allocated to the content providers. The utility based approach is used to formulated the problem of content providers. In [20], the authors considered that the network operator (NO) leases the resources of a high-tier central cloudlet for task offloading. The NO tries to minimize its computational cost and devices' energy consumption in a multi-tier mobile edge computing (MEC) system by optimizing the offloading decision, transmit power and radio resources in uplink channel.

The majority of the above work consider neither infrastructure sharing (i.e., virtualization of network resources) nor the latency constraint at the UE in the context of local 5G OP with edge caching. Also, only deterministic channel models with caching are considered, e.g., in [16]–[19]. We therefore significantly extend the existing work by developing a framework to model and analyze latency-constrained for radio resource sharing with caching-enabled SBSs in local 5G OP networks with the SBS modeling using stochastic geometry.

Fig. 1 illustrates the business model between the local 5G OP, MNOs, application provider (APV), and infrastructure provider (InP) considered in this paper. The local 5G OP leases available licensed subbands from multiple MNOs and obtains the videos/contents from the APV. The core networks can be

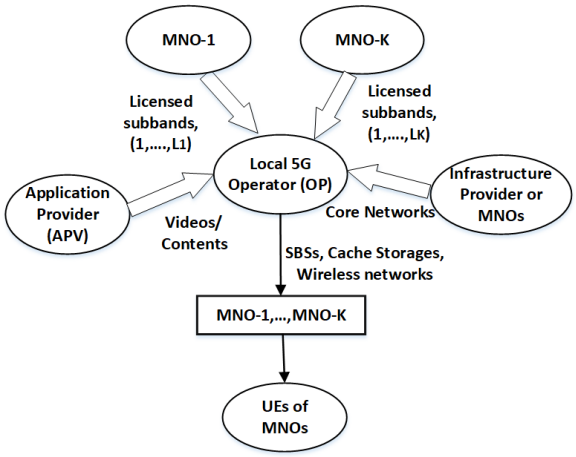

Fig. 1. Business model between the local 5G OP, MNOs, application provider (APV), and infrastructure provider.

provided by either InP or MNOs. Then, the local 5G OP will assign the licensed subbands to the SBSs in order to give the wireless service to the UEs of MNOs. We consider only a part of local 5G OP leasing out the SBSs and cache storage to the MNOs and allowing the MNOs to rent and share the same resources. Some part of the business model in Fig. 1, includes one MNO and one local 5G OP, has been initially implemented in real-world scenario, i.e., in our 5G test network (5G TN), 6G Flagship, university of Oulu, Finland [21]. The 5G TN has received the licensed spectrum subbands, frequency band 7 which is from 2110 – 2170 MHz for downlink transmission. The core network has been provided by nokia, Oulu, Finland, to connect to each SBS. In this case, our 5G TN becomes the local 5G operator (local 5G OP) who deploys SBSs with licensed subbands and each SBS connects to core networks.

For the theoretical aspects, our work is the first work that formulates the optimization problems of each MNO and the local 5G OP as geometric programming and formulate an entire problem as a Stackelberg game where the local 5G OP is the leader and the MNOs are the followers. Also, the MNOs cooperate with each other to share the cost of leasing. The motivation behind our proposed game theoretical frameworks are (i) The local 5G OP installs fixed amount of cache-enable SBSs per unit area, so it knows the available cache intensity to provide MNOs for renting. (ii) The local 5G OP and the MNOs can exchange information between each other. Thus, the local 5G OP will have sufficient foresight to anticipate the strategy of each MNO and the local 5G OP is able to estimate the amount of infrastructure that all MNOs need. Therefore, we formulate the business model of local 5G OP and the MNOs by using Stackelberg game. Then, the local 5G OP declares the total rent to MNOs, and since the MNOs are able to communicate with other. Accordingly, we propose the coalitional formation game where the MNOs cooperate with each other to share the total rent.

B. The contributions of the paper

The main contributions of the paper are as follows

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

- From the perspective of cache-enabled SBSs, we derive expressions for two important caching performance metrics, namely, cache hit probability and wireless access delay.
- Using the derived caching performance metrics, we model the virtualization problem of a cache-enabled large-scale cellular networks as a Stackelberg game, where the local 5G OP is the leader and the MNOs are the followers.
- For radio resource sharing with edge caching, using a geometric program, the optimal strategies of each MNO are obtained analytically in terms of SBS intensity and cache size where the MNO aims at minimizing the cost of renting cache intensity from the local 5G OP subject to the latency constraint at the UE.
- To use the storage resources efficiently, from the perspective of resource sharing of local 5G OP, it handles only the largest cache intensity required by the MNO. The pricing problem of the local 5G OP is formulated as a Stackelberg game where the objective is to maximize its revenue while minimizing the power consumption at the SBSs. We obtain the subgame perfect equilibrium of the Stackelberg game analytically via successive geometric programming.
- We develop a method based on the Shapley value to share the cost of leasing among the MNOs. We show that our cost sharing problem can be mapped into a simplified form, namely, an airport runway cost sharing problem, in which the Shapley value can be computed efficiently.

C. Organization

The rest of the paper is organized as follows. Section II describes the system model. In Section III-A, the cache hit probability and wireless access delay are derived. Section IV presents the optimization of each MNO as to minimize the *cache intensity* subject to the latency constraint at a UE. This corresponds to the follower subgame in the Stackelberg game formulation. Section V presents the problem of the local 5G OP as to maximize the revenue which corresponds to the leader subgame in the Stackelberg game formulation. Also, the cooperation among MNOs for sharing the infrastructure using Shapley value to divide the rent among the MNOs is proposed. The numerical results are presented in Section VI before the paper is concluded in Section VII.

II. Local 5G Operator Virtualized Cache-enable SBSs to MNOs

A. System Model

We consider a heterogeneous network with a local 5G OP and a set $\mathcal{K}$ of MNOs such that $|\mathcal{K}| = K$. The MNOs are assumed to be co-located and serve their UEs in the same area. The local 5G OP is assumed to provide the licensed spectrum band while installing a set of cache-enabled SBSs, Φ_b, which are spatially distributed according to a homogeneous Poisson point process (PPP) with spatial intensity λ, i.e., the number of SBSs per unit area. All SBSs are identical in terms of edge caching capabilities.

TABLE I
List of Common Notations

Notation	Description
$\|\mathcal{K}\|$	A set of MNOs where $\|\mathcal{K}\| = K$.
Φ_b	A set of cached-enabled SBSs installed by the local 5G OP.
Φ_k	A set of cache-enabled SBSs rented by MNO-k.
Φ_{u_k}	The set of UEs that subscribe to MNO-k.
λ	The intensity of SBSs installed by the local 5G OP.
λ_k	The intensity of SBSs rented by MNO-k.
W_k	Licensed banwidth leased out by the MNO-k.
L_k	Number of subchannels of the MNO-k.
λ_I	The intensity of interfering SBSs which causes intra-MNO interference of MNO-k.
λ_A	The intensity of SBSs that a typical UE of MNO-k can associate itself with.
ξ_k	The intensity of UEs subscribe to MNO-k.
p_k	The transmit power of the rented SBS of MNO-k.
I_k	The intra-MNO interference of MNO-k.
g_j	The channel gain between the tagged UE and interfering SBS-j
r_j	The distance between the tagged UE and the interfering SBS-j,
P_c	The coverage probability.
G_k	The throughput of the tagged UE served by the nearest SBS.
$\|\mathcal{K}\|$	The set of files available for caching in the cloud, where $\|\mathcal{K}\| = K$.
S_k	The number of files store in cache-enable SBS.
ν	Zipf exponent.
P_{hit}	Cache hit probability.
x_f	Size of the file which is assumed to be fixed.
D_{Tx}	Transmission delay.
D_{bh}	Backhaul delay.
D_k	Total delay.
ω	The price of cache per unit area.

Each MNO-k, $k \in \mathcal{K}$, wants to rent a fraction of the set Φ_b. The intensity of the SBSs λ_k rented and utilized by the MNO-k, is given by the thinning of Φ_b of the SBSs owned by the local OP, which yields another homogeneous PPP Φ_k. We can express the property of thinning and sharing of SBSs by $\Phi_k \subseteq \Phi_b$ such that $\bigcup_{k\in\mathcal{K}} \Phi_k = \Phi_b$, where $k, l \in \mathcal{K}$ and $k \neq l$. First of all, the local 5G OP leases some portions of available licensed spectrum from multiple MNOs then, attach the licensed subbands to its SBSs while allowing multiple MNOs to rent its SBSs and cache storage.

Fig. 2 gives an example for the general case of the average number of cache-enabled SBSs required by MNO-1, MNO-2 and MNO-3. Since the local 5G OP deploys resource virtualization, when MNOs rent the cache-enabled SBSs from the local 5G OP, some of the SBSs including cache storage can be utilized by all three MNOs simultaneously while some are used by one or two MNOs. However, in our scenario, we assume that the each cache-enable SBS of local 5G OP are shared by all K MNOs simultaneously. Each MNO-k operates over orthogonal spectrum, and thus there is no inter-MNO interference. The MNO-k leases out the licensed bandwidth of W_k Hz, which is divided into L_k subchannels to the local 5G OP. Each SBS operates in one of the L_k available subchannels assigned to it by the local 5G OP. The subchannel of the same MNO can be accessed by more than one SBS and therefore,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

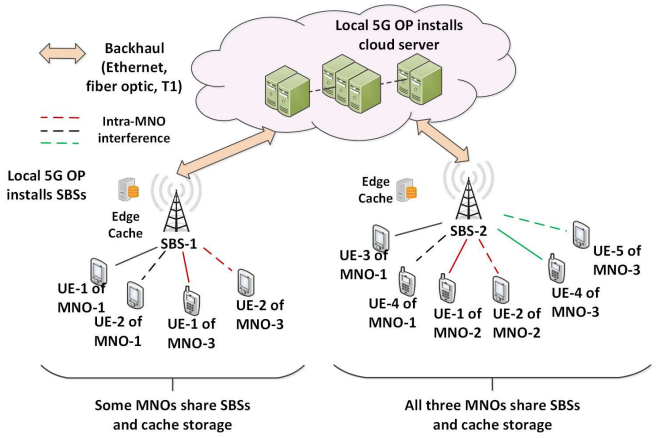

Fig. 2. Virtualized Cache-enabled SBSs shared between MNO-1, MNO-2, and MNO-3.

the intensity of interfering SBSs which causes intra-MNO interference of the MNO-k is given by $\lambda_I = \frac{\lambda_k}{L_k}$. Every SBS and UE are assumed to be equipped with a single antenna.

For each MNO-k, the SBS serves one UE in a given time slot by using the assigned subchannel with the maximum transmit power p_k. A UE subscribed to an MNO-k associates with the nearest available SBS that the MNO-k rents from the local 5G OP. We assume that each MNO-k is free to use any fraction of the total available SBSs. The net intensity of the SBSs that a typical UE of the MNO-k can associate itself with is $\lambda_A = \lambda_k$. The set of UEs that subscribe to MNO-k is denoted by Φ_{u_k}. The UEs are assumed to be spatially distributed according to a homogeneous PPP with spatial intensity ξ_k.

In this part, we consider the analysis of a single MNO-k downlink SINR coverage probability and throughput. We consider a tagged UE of MNO-k to be located at the origin, which associates with the nearest SBS. Let us label the nearest SBS as SBS-0. We assume that the signal undergoes Rayleigh fading with the channel gain, g_0. Let $\alpha_k > 2$ denote the path-loss exponent for the path-loss model $r_0^{-\alpha_k}$, where r_0 is the distance between the tagged UE and the nearest SBS-0, $0 \in \Phi_b$. Let σ_k^2 denote the noise variance, and again p_k denote the transmit power of all the SBSs of the MNO-k.[1] The downlink $\mathrm{SINR_k}$ at the tagged UE is given by $\mathrm{SINR_k} = \frac{g_0 r_0^{-\alpha_k} p_k}{I_k + \sigma_k^2}$. The interference experienced by the tagged UE associated with SBS-0 comes from the transmitted signal from other SBSs of the same MNO-k to the UEs in the same time slot. Thus, $I_k = \sum_{j \in \Phi_b \backslash \{0\}} g_j r_j^{-\alpha} p_k$. Here g_j is the channel gain between the tagged UE and interfering SBS-j, and r_j is the distance between the tagged UE and the interfering SBS-j, where $j \in \Phi_b \backslash \{0\}$.

For a given threshold $\overline{T}$, the SINR coverage probability for the tagged UE is defined as $P_c = \mathbb{P}(\mathrm{SINR_k} > \overline{T})$.

[1]The UEs are assumed to associate to the SBSs based on their average received signal strength and the average received signal power at each UE from its corresponding SBS will be strictly higher than the average interference power from the interfering SBS. As such, power control is not crucial for the network operation due to interference protection introduced by the UE association criterion.

Following the approach given in [22, Theorem 1], we first condition on the nearest SBS at the distance r_0 from the tagged UE. Since the Rayleigh fading channel gain follows an exponential distribution, P_c can be expressed by taking expectation with respect to the interference power as in (4)–(6) in [12]. The P_c under our system model is given as [12, Prop. 1]: $P_c = \pi\lambda_A \int_0^\infty \exp\{-(\overline{A}z + \overline{B}z^{\alpha/2})\}dz$, with $\lambda_A = \lambda_k$, $\lambda_I = \lambda_k/L_k$. The coefficients $\overline{A}$ and $\overline{B}$ are given by $\overline{A} = \pi[\lambda_I(\beta - 1) + \lambda_A]$ and $\overline{B} = \frac{\overline{T}\sigma_k^2}{p_k}$, where $\beta = \frac{2(\overline{T}/p_k)^{2/\alpha_k}}{\alpha_k}\mathbb{E}_g[g^{2/\alpha}(\Gamma(-2/\alpha_k, \overline{T}g/p_k)) - \Gamma(-2/\alpha_k)]$. We can evaluate P_c using equation (13) in [12] as $P_c \simeq \pi\lambda_A \left[\overline{A} + \frac{\alpha_k}{2}\frac{\overline{B}^{2/\alpha_k}}{\Gamma\left(\frac{2}{\alpha_k}\right)}\right]^{-1}$. Therefore, the coverage probability can be expressed as,

$$P_c = \left[1 + \frac{\beta - 1}{L_k} + \frac{\alpha_k}{2\pi\lambda_k\Gamma\left(\frac{2}{\alpha_k}\right)}\left(\frac{\overline{T}\sigma^2}{p_k}\right)^{2/\alpha_k}\right]^{-1}, \quad (1)$$

where $\Gamma(z)$ is the Gamma function. For the interference-limited case, when $\sigma^2 \to 0$, or when $\lambda_k \to \infty$, the last term in (1) will become 0. The expression in (1) simplifies to

$$P_c \simeq \frac{L_k}{\beta + L_k - 1}. \quad (2)$$

Note from (2) that as $L_k \to \infty$, $P_c \to 1$.

Next, we define the *throughput* of the tagged UE served by the nearest SBS as $G_k = \frac{P_c W_k}{L_k}\log_2(1 + \theta)$, where P_c is the downlink coverage probability, and W_k/L_k is the channel bandwidth. We can approximate the throughput by using P_c in (1) for the general case, or P_c in (2) for the interference-limited case. In this regard, we use P_c of the interference-limited case from (2) to express G_k as follows:

$$G_k = \frac{W_k}{\beta + L_k - 1}\log_2(1 + \theta). \quad (3)$$

Note that G_k is independent of λ_k. Also, increasing the number of subchannels L_k improves the coverage probability but reduces the throughput.

III. Modeling of Content Caching and Wireless Access Delay

The idea of caching at the SBSs allows us to reduce the latency of data delivery to the UEs. The local 5G OP can store the most popular files in cache storage of all the SBSs to serve the UEs of MNOs. In this section, we present the analysis of cache hit probability and the delay modeling.

A. Content Popularity-Based Edge Caching: Cache Hit Probability

Let $\mathcal{F} = \{f_1, \ldots, f_F\}$ be the set of files available for caching in the cloud, where $F = |\mathcal{F}|$. Considering edge caching based on the file popularity, let $\mathcal{S}_k \subseteq \mathcal{F}$ be the set of files can be stored at each SBS of each MNO-k. For simplicity, we assume that all the files are of equal size. If a random file $f \in \mathcal{F}$ is requested by a UE, let $P_{\mathrm{hit}}(\mathcal{S}_k) = \mathbb{P}(f \in \mathcal{S}_k)$ denote the probability that the file f is available at the SBS cache, which we refer to as the "*cache hit probability*."

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

We assume that the cache policy is to store the S_k most popular files from $\mathcal{F}$. We can model the popularity of the files by the Zipf distribution given by $p_d = \frac{1/d^\nu}{\sum_{j=1}^{F} 1/j^\nu}$, where p_d is the probability of d-th most popular file being requested and the exponent $\nu > 0$ reflects the *skewness* of the file popularity distribution. Larger values of ν lead to fewer popular files in the content requests. The probability that the requested file $f \in \mathcal{F}$ is stored in the cache is $P_{\text{hit}}(S_k) = \mathbb{P}(d \leq S_k)$, where d is the random popularity rank of file f. Since $\mathbb{P}(d \leq S_k)$ is the cumulative distribution function (CDF) of the Zipf distribution, we can express $P_{\text{hit}}(S_k)$ as follows:

$$P_{\text{hit}}(S_k) = \frac{\sum_{d=1}^{S_k} 1/d^\nu}{\sum_{j=1}^{F} 1/j^\nu} = \frac{H_{S_k,\nu}}{H_{F,\nu}}. \tag{4}$$

In (4), we concisely express P_{hit} using generalized harmonic numbers, i.e., $H_{S_k,\nu}$ and $H_{F,\nu}$, where

$$H_{S_k,\nu} = \sum_{n=0}^{S_k-1} \frac{1}{(n+1)^\nu}, \tag{5}$$

and $H_{F,\nu}$ is defined similarly.

B. Asymptotic Approximation for the Cache Hit Probability

To facilitate subsequent analysis, we now derive the following lemma on the asymptotic approximation (i.e., when $S_k \to \infty$) on the hit probability:

Lemma 1. *When the cache size S_k is large and $\nu \neq 1$, the probability that a requested file $f \in \mathcal{F}$ is in the cache is asymptotically given by*

$$P_{hit}(S_k) \simeq \frac{1}{H_{F,\nu}} \left[\zeta(\nu) - \frac{(S_k+1)^{1-\nu}}{\nu - 1} \right], \tag{6}$$

where $\zeta(\nu)$ is the Riemann zeta function.

Proof: The generalized harmonic number $H_{S_k,\nu}$ does not have a closed-form expression. Nevertheless, for analytical tractability, we can make an asymptotic approximation in terms of S_k and ν. To do so, we relate the generalized harmonic number to the Hurwitz zeta function and then use the properties of Hurwitz zeta function. The Hurwitz zeta function, $\zeta(s, a)$, is defined as follows [23, Eqn 25.11.1]:

$$\zeta(s, a) = \sum_{n=0}^{\infty} \frac{1}{(n+a)^s}, \tag{7}$$

where $\Re(s) > 1$ and $a \neq 0, -1, -2, \ldots$. The Hurwitz zeta function reduces to the Reimann zeta function when $a = 1$, i.e., $\zeta(s, 1) = \zeta(s)$, where $\zeta(s)$ is the Riemann zeta function. Also, harmonic sums can be expressed in terms of Hurwitz zeta function as follows [23, Eqn 25.11.4]:

$$\sum_{n=0}^{m-1} \frac{1}{(n+a)^s} = \zeta(s, a) - \zeta(s, a+m). \tag{8}$$

In our case, comparing (5) and (8), we can express the generalized harmonic sum $H_{S_k,\nu}$ in terms of the Hurwitz zeta function as follows:

$$H_{S_k,\nu} = \sum_{n=0}^{S_k-1} \frac{1}{(n+1)^\nu} = \zeta(\nu) - \zeta(\nu, S_k + 1). \tag{9}$$

Now, as $S_k \to \infty$, the asymptotic expansion of the Hurwitz zeta function is given by [23, Eqn 25.11.43]

$$\zeta(\nu, S_k+1) \sim \frac{(S_k+1)^{1-\nu}}{\nu-1} + \frac{1}{2}(S_k+1)^{-\nu} + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} (\nu)_{2k-1} (s_k+1)^{1-\nu-2k}, \tag{10}$$

where B_{2k} are Bernoulli numbers and $(\nu)_{2k-1} = \nu(\nu+1)\cdots(\nu+2k-2)$ are Pochammer's symbol for rising factorial. Taking only the first dominant term from (10) and substituting it in (9), we obtain the asymptotic approximation for the generalized harmonic number as follows:

$$H_{S_k,\nu} \sim \zeta(\nu) - \frac{(S_k+1)^{1-\nu}}{\nu-1}. \tag{11}$$

Thus, from the above arguments, using (4) and (11), we prove the lemma. ∎

Here (6) is asymptotic in the sense that a larger value of S results in a greater accuracy of the approximation. Note that although it is required that $\Re(\nu) > 1$ in the definition of the Hurwitz zeta function in (7), the Riemann zeta function $\zeta(\nu)$ has a unique analytic continuation to the entire complex plane, excluding $\nu = 1$, which corresponds to a simple pole [24]. Similar analytic continuation holds for the Hurwitz zeta function as well [25]. Thus, so long as $\nu \neq 1$, the approximation in (6) is applicable for any Zipf's exponent $\nu > 0$.

In reality, Zipf's exponent is found to be close to, but never exactly equal to, 1. There is no consensus on the actual setting of ν value [26], [27], with the considered value varying widely, i.e., $\nu \in [0.5, 2.5]$. Also, since we expect the cache size to be $1 < S_k \ll F$, the above approximation holds with very small margin of error. In Fig. 3, we compare the hit probability versus cache size using the exact values of $H_{s,\nu}$ from (9) and the asymptotic approximation of $H_{s_k,\nu}$ from (11) when $F = 10^3$. The relative error is shown in Fig. 4. We observe that the relative error decreases with an increase of cache size. The error tends to decrease more rapidly for larger values of ν. For $\nu \geq 0.5$, the relative error is less than 1% for $S_k \geq 30$, while for $\nu \geq 1.5$ the relative error is less than 1% for $S_k \geq 10$. Finally, we also note that the formula is applicable even when $\nu < 0$.

Remark 1. *Similar to $H_{S_k,\nu}$, applying the asymptotic approximation for the generalized harmonic number to $H_{F,\nu}$, we have*

$$P_{hit}(S_k) \sim \frac{\zeta(\nu) - \frac{(S_k+1)^{1-\nu}}{\nu-1}}{\zeta(\nu) - \frac{(F+1)^{1-\nu}}{\nu-1}}. \tag{12}$$

Thus, for a fixed value of S_k, P_{hit} decreases with increasing F.

Remark 2. *Let $s_k = \frac{S_k+1}{F+1}$ be the fixed fraction of files cached at the SBS. Dividing the numerator and denominator of the right hand side of (12) by $(F+1)^{1-\nu}/(\nu-1)$, we see that as $F \to \infty$ and as S_k changes such that the fraction s is fixed, we obtain*

$$P_{hit} \sim s_k^{1-\nu}, \tag{13}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

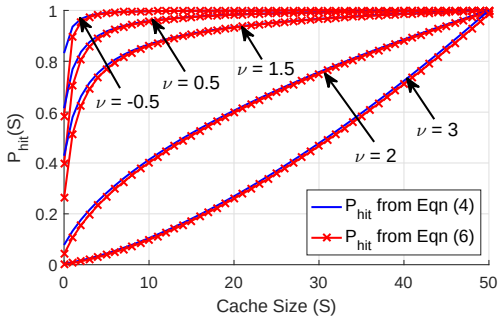

Fig. 3. Cache hit probability versus cache size S.

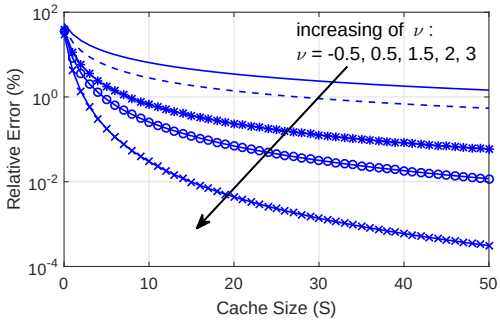

Fig. 4. Relative error versus cache size S.

for $0 < \nu < 1$. *This result has an important implication in that, to achieve a desirable hit probability, the fraction of total files that needs to be cached at the SBS is*

$$s_k \sim P_{hit}^{\frac{1}{1-\nu}}, \tag{14}$$

so long as $0 < \nu < 1$. *If* $\nu > 1$, *then* $P_{hit} \approx 1$ *as* $F \to \infty$.

C. Modeling Wireless Access Delay in a Cache-Enabled Cellular Network

With caching at the SBSs, by *wireless access delay* we refer to the delay between a request is made by a UE to download a file and the time the file is downloaded at the UE. Considering that the downlink is the bottleneck, we consider only expected downlink transmission delay plus expected backhaul delay along with the cache hit probability[2].

1) Expected Downlink Transmission Delay: The delay in the transmission of a file between the cached SBS and UE is referred to as the *downlink transmission delay*. If a file requested by a UE is available in the cache of the serving SBS, then there is only transmission delay. This delay is attributed to a number of factors, including finite channel capacity, size of the file, and the number of UEs in a cell. For each MNO-k, each rented SBS serves N_k UEs in the cell in a given time slot based on round-robin scheduling, the throughput per UE becomes G_k/N_k, and the delay at each UE is N_k/G_k. In order to transfer a file of fixed size x_f, the transmission delay is $D_{\text{Tx}} = \frac{N_k x_f}{G_k}$. Here, N_k is a random variable. The expected number of UEs inside an average Voronoi cell formed by the PPP SBSs is given by $\mathbb{E}[N_k] = \frac{\xi_k}{\lambda_k}$, where ξ_k is the intensity of the UEs.

$$\mathbb{E}[D_{\text{Tx}}] = \frac{\mathbb{E}[N_k] x_f}{G_k} = \frac{\xi_k x_f}{\lambda_k G_k}. \tag{15}$$

Since we are using G_k for interference limited case in (3), the throughput of each UE becomes a constant. In (15) the SBS intensity λ_k is a variable that the MNO will need to decide when renting the infrastructure from the local 5G OP. We can oSBServe from (15) that increasing the SBS intensity lowers the transmission delay. This also verifies a peculiarity of modeling the SBSs as a homogeneous PPP which is a well-known result in stochastic geometry.

2) Expected Backhaul Delay: If the requested file from a UE is *not* available in the cache of the serving SBS, the SBS needs to fetch the file from a cloud server through optical fiber. Here, the delay of the file transfer between the SBS and the cloud server is referred to as the *backhaul delay*. We model the process of a SBS fetching files from the cloud server using an $M/M/m$ queue [28][3]. When the number of cloud servers is $m = 1$, the expected backhaul delay is given as follows [29]:

$$\mathbb{E}[D_{\text{bh}}] = \left(\frac{c_a^2 + c_s^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\tau + \tau, \tag{16}$$

where τ is the average time taken for the server to deliver x_f bits of each file to the SBS. The arrival rate of file requests to the server is ϕ, the service rate of the server is $\mu = 1/\tau$, and $\rho = \phi/\mu$ is the server utilization. Assuming $\rho < 1$ for the steady-state system. Here c_a and c_s are coefficients of variations of the inter-arrival time and the service time, respectively.

3) Expected Total Delay: The delay experienced by a UE while downloading a file is only due to the downlink transmission delay, D_{Tx}, if the requested file is already cached at its serving SBS. If this is not the case, then the wireless access delay experienced by the UE is the sum of the downlink transmission delay and the backhaul delay, $D_{\text{Tx}} + D_{\text{bh}}$. Since the availability of a file in the SBS cache is given by the hit probability, P_{hit}, the expected total delay is given by

$$\begin{aligned}\mathbb{E}[D_k] &= \mathbb{E}[D_{\text{Tx}}]\, P_{\text{hit}} + \mathbb{E}[D_{\text{Tx}} + D_{\text{bh}}]\,(1 - P_{\text{hit}}) \\ &= \mathbb{E}[D_{\text{Tx}}] + \mathbb{E}[D_{\text{bh}}](1 - P_{\text{hit}}).\end{aligned} \tag{17}$$

Since $0 \le P_{\text{hit}} \le 1$, note that the average total delay is bounded by $\mathbb{E}[D_{\text{Tx}}] \le \mathbb{E}[D_k] \le \mathbb{E}[D_{\text{Tx}}] + \mathbb{E}[D_{\text{bh}}]$. Since P_{hit} depends on the cache size S, this implies that the minimum expected total delay that we can achieve by changing only the cache size is $\mathbb{E}[D_{\text{Tx}}]$.

In the next section, we propose the MNO strategy subject to the latency constraint, i.e., the total transmission delay from the SBS to UE. Then, we formulate the business model of the local 5G OP and multiple MNOs by using Stackelberg game, where the local 5G OP the leader and MNOs are the followers.

[2]The uplink delay (i.e., delay between the request is sent by the UE and it is received by the SBS) is not considered in our case.

[3]In our framework, any other queueing model can be also used to characterize the backhaul delay.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

IV. The MNO strategy : Minimization of Cache Intensity with Latency Constraint

In this section, we deal with the minimization of *cache intensity* for each MNO-k in a large cellular network. The cache intensity is defined as the product of SBS intensity and cache size. For the interference-limited transmission scenario, we then transform the problem into a geometric program and provide an exact solution.

A. Optimization Problem Formulation

The optimization problem for an MNO-k, where $k \in \mathcal{K}$, so as to minimize the cost of renting the amount of cache per unit area (cache intensity) while satisfying the latency of a tagged UE is as follows:

$$(P0) \quad \min_{\lambda_k, S_k} \quad \omega \lambda_k S_k \tag{18}$$
$$\text{s.t.} \quad \mathbb{P}(D_k \geq D_{\text{th}}) \leq \epsilon, \tag{19}$$
$$S_k \leq F, \tag{20}$$

where $\lambda_k \geq 0$ and $S_k \geq 0$. Here, the latency constraint is given in (19). It is a probabilistic constraint that limits the probability of end-to-end delay to be above a certain threshold D_{th} to a small value ϵ, $\epsilon \in (0, 1)$. The ω is the price per unit of cache intensity, which is set by the local 5G OP.

Since D_k is a random variable whose distribution is not known, in order to make the optimization problem more tractable, we can use the Markov's inequality to linearize the probabilistic constraint in (19). Using Markov's inequality, we have $\mathbb{P}(D_k \geq D_{\text{th}}) \leq \frac{\mathbb{E}[D_k]}{D_{\text{th}}}$. If we ensure that $\frac{\mathbb{E}[D_k]}{D_{\text{th}}} \leq \epsilon$, then the Markov inequality implies that constraint (19) is also satisfied. Accordingly, the probabilistic constraint in (19) can be replaced by the constraint $\mathbb{E}[D_k] \leq \epsilon D_{\text{th}}$. Thus, we have a more tractable problem:

$$(P0') \quad \min_{\lambda_k, S_k} \quad \omega \lambda_k S_k \tag{21}$$
$$\text{s.t.} \quad \mathbb{E}[D_k] \leq \epsilon D_{\text{th}}, \tag{22}$$
$$S_k \leq F. \tag{23}$$

SuSBStituting the expression for $\mathbb{E}[D_k]$ from (17) into (22), we obtain after some algebra,

$$1 - \frac{\epsilon D_{\text{th}} - \mathbb{E}[D_{\text{Tx}}]}{\mathbb{E}[D_{\text{bh}}]} \leq P_{\text{hit}}(S_k). \tag{24}$$

Since $P_{\text{hit}}(S_k) \leq 1$ for any $S_k \leq F$, the left-hand-side of (24) must be less than or equal to unity. As such, it must be the case that

$$\mathbb{E}[D_{\text{Tx}}] \leq \epsilon D_{\text{th}}. \tag{25}$$

Thus, we have **Lemma 2** in the following subsection to guarantee the feasibility of latency constraint.

B. Trade-off Between Cache Storage and SBS Intensity

Lemma 2. *The constraint in (22) is feasible for $S_k \leq F$ if and only if $\mathbb{E}[D_{Tk}] \leq \epsilon D_{th}$ is satisfied.*

Proof: The proof of statement in the forward direction is as given above. For the reverse direction, we are given $\mathbb{E}[D_{\text{Tk}}] \leq \epsilon D_{\text{th}}$. To check if there exists some S_k such that $S_k \leq F$ which satisfies $\mathbb{E}[D_k] \leq \epsilon D_{\text{th}}$, we have from (17), $P_{\text{hit}}(S) = 1 - \frac{\mathbb{E}[D_k] - \mathbb{E}[D_{\text{Tk}}]}{\mathbb{E}[D_{\text{bh}}]} \leq 1 - \frac{\epsilon D_{\text{th}} - \mathbb{E}[D_{\text{Tk}}]}{\mathbb{E}[D_{\text{bh}}]}$. Since $\mathbb{E}[D_{\text{Tk}}] \leq \epsilon D_{\text{th}}$, this means that $P_{\text{hit}}(S) \leq 1$. Hence, there must exist some S_k such that $S_k \leq F$ which satisfies (22). ■

Lemma 2 gives us the necessary and sufficient condition under which both the constraints in (22) and (23) can be satisfied. We see that although the latency constraint in (22) is over the total delay, since the backhaul delay is a constant, from (25), the transmission delay is the most significant component. However, if we have $\mathbb{E}[D_{\text{Tk}}] = \epsilon D_{\text{th}}$, then the cache size is $S_k = F$. That is, each SBS has to cache all the available files in $\mathcal{F}$. This is certainly unrealistic in practice. Hence, practically, it should be the case that $\mathbb{E}[D_{\text{Tk}}] < \epsilon D_{\text{th}}$ so that $S_k < F$. We can also express the constraint in (25) in terms of the average number of UEs served per SBS, $\frac{\xi_k}{\lambda_k}$ which lead to the following proposition:

Proposition 1. *For the latency constraint in (25) to be satisfied for $S_k \leq F$, the average number of UEs per SBS, $\frac{\xi_k}{\lambda_k}$, must satisfy $\frac{\xi_k}{\lambda_k} \leq \frac{\epsilon D_{th} G_k}{x_f}$.*

Proof: We substitute the expression for $\mathbb{E}[D_{\text{Tk}}]$ from (15) in (25), we obtain a lower bound for the SBS intensity required to satisfy the constraint in (25) and hence the constraint in (22) as follows :

$$\lambda_k \geq \frac{\xi_k x_f}{\epsilon D_{\text{th}} G_k}. \tag{26}$$

This gives us the relationship between the SBS intensity λ_k and the UE intensity ξ_k for the latency constraint to be satisfied for $S_k < F$. If this condition is violated, then the required S_k will be greater than F, leading to a contradiction. ■

Therefore, the above proposition gives us a simple condition under which caching at the SBS will satisfy the required latency constraint. Indeed, if λ_k is held fixed, we cannot change the transmission delay and can change only the total delay by varying the cache size, S_k. Hence, we need to vary both λ_k and S_k that results in problem (P1) given in the next subsection.

C. Optimal Strategy of each MNO

We will now solve the general problem when both λ_k and S_k are jointly optimized. Given the popularity-based caching, for large S_k the optimization problem (P0') is a geometric program [30]. In the following, we first express the primal problem in (22) in the standard form of a geometric program, after which, we give the solution to the problem via its dual problem.

First, we expand $\mathbb{E}[D_k]$ in (17) using (15) and (6) in terms of λ_k and S_k as follows:

$$\mathbb{E}[D_k] = \frac{\xi_k x_f}{G_k \lambda_k} + \mathbb{E}[D_{\text{bh}}] \left[1 - \frac{1}{H_{F,\nu}} \left(\zeta(\nu) - \frac{(S_k + 1)^{1-\nu}}{\nu - 1}\right)\right]$$
$$= C_1 + \frac{C_2}{\lambda_k} + C_3 (S_k + 1)^{1-\nu}, \tag{27}$$

where $C_1 = \mathbb{E}[D_{\text{bh}}] \left(1 - \frac{\zeta(\nu)}{H_{F,\nu}}\right)$, $C_2 = \frac{\xi_k x_f}{G_k}$, and $C_3 = \frac{\mathbb{E}[D_{\text{bh}}]}{(\nu - 1) H_{F,\nu}}$. Since we consider the case that the cache size, S_k, is large, without loss of generality we can assume $S_k + 1 \approx S_k$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

in the right most term of (27). Then, substituting (27) in the constraint in (22), we obtain, $C_1 + \frac{C_2}{\lambda_k} + C_3 S^{1-\nu} \leq \epsilon D_{\text{th}} \iff \left(\frac{C_2}{\epsilon D_{\text{th}} - C_1}\right) \frac{1}{\lambda} + \left(\frac{C_3}{\epsilon D_{\text{th}} - C_1}\right) S_k^{1-\nu} \leq 1$. Therefore,

$$A\lambda_k^{-1} + V S_k^{1-\nu} \leq 1, \tag{28}$$

where $A = \frac{C_2}{\epsilon D_{\text{th}} - C_1}$ and $V = \frac{C_3}{\epsilon D_{\text{th}} - C_1}$. Similarly, substituting (26) to the constraint in (23), we can express the primal problem in (P0') as a geometric program.

Proposition 2. *For content popularity-based edge caching, assuming the cache size, S_k, to be sufficiently large and the constants $A > 0$, $V > 0$, $\nu \neq 1$, we can transform the problem (P0') into the following geometric program:*

$$(P1) \quad \min_{\lambda, S} \quad g = \omega \lambda_k S_k \tag{29}$$

$$\text{s.t.} \quad A\lambda_k^{-1} + V S_k^{1-\nu} \leq 1, \tag{30}$$

$$R\lambda_k^{-1} \leq 1, \tag{31}$$

where $R = \frac{C_2}{\epsilon D_{th}}$.

The optimization problem in (P1) can be solved analytically. In geometric programming, when the orthogonality and normality conditions with dual variables δ_i are satisfied, the maximum of dual function is equal to the minimum of primal function g [30]. As such, we can express the dual maximization problem as follows:

$$\max_{\boldsymbol{\delta}} q = \left(\frac{\omega}{\delta_1}\right)^{\delta_1} \left(\frac{A}{\delta_2}\right)^{\delta_2} \left(\frac{V}{\delta_3}\right)^{\delta_3} \left(\frac{R}{\delta_4}\right)^{\delta_4} (\delta_2 + \delta_3)^{\delta_2 + \delta_3} \delta_4^{\delta_4}, \tag{32}$$

$$\text{s.t.} \quad \delta_1 = 1, \tag{33}$$

$$\begin{pmatrix} 1 & -1 & 0 & -1 \\ 1 & 0 & 1-\nu & 0 \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{pmatrix} = 0, \tag{34}$$

where $\delta_i \geq 0$ for $i = 1, \ldots, 4$. The degree of difficulty of this geometric program is 1. In our case, (33) gives the normality condition while (34) gives the orthogonality condition. In geometric programming, we focus on finding the optimal point of the dual variables $\boldsymbol{\delta}^* = (\delta_1^*, \delta_2^*, \delta_3^*, \delta_4^*)$ that maximizes the dual function q subject to the orthogonality and normality conditions. Note that this dual problem is a convex program with a concave objective function and linear constraints.

Using (33) and (34), we can directly solve for $\boldsymbol{\delta}^*$. Here, matrix multiplication from (34) yields

$$\delta_1 - \delta_2 - \delta_4 = 0, \quad \text{and} \quad \delta_1 + (1-\nu)\delta_3 = 0.$$

Since $\delta_1 = 1$, we have $\delta_3 = \frac{1}{\nu - 1}$ and $\delta_2 + \delta_4 = 1$. Let $\delta_2 = r$, so that $\delta_4 = 1 - r$. Since $\delta_2 \geq 0$ and $\delta_4 \geq 0$, we then have a bound over r as $0 \leq r \leq 1$. Substituting the values of $\boldsymbol{\delta}$ in the dual problem, we obtain a simpler problem constrained over a single variable r as given in (37)-(38). To find the optimal r, we first take the logarithm of q in (37) and differentiate it with respect to r. Since A, R, and ν are all positive, we obtain

$$\frac{\partial \log(q)}{\partial r} = \log\left(\frac{A}{rR}\right) + \log\left(\frac{1 + r(\nu - 1)}{\nu - 1}\right). \tag{35}$$

Solving the optimality condition $\frac{\mathrm{d}\log(q)}{\mathrm{d}r} = 0$ for r, we obtain the maxima at $r = \frac{1}{(\nu-1)\left(\frac{R}{A} - 1\right)}$. Since r is bounded between $0 \leq r \leq 1$, we have the optima of the modified dual problem at

$$r^* = \max\left(0, \min\left(1, \left[(\nu - 1)\left(\frac{R}{A} - 1\right)\right]^{-1}\right)\right). \tag{36}$$

Let q^* be the optimal value of the modified dual problem (37)-(38). For the optimal primal variables λ_k^* and S_k^*, we have

$$\omega \lambda_k^* S_k^* = \delta_1^* q^* = q^*, \qquad A(\lambda^*)^{-1} = \delta_2^* q^* = r^* q^*,$$

$$V(S_k^*)^{1-\nu} = \delta_3^* q^* = \frac{q^*}{\nu - 1}, \quad R(\lambda_k^*)^{-1} = \delta_4^* q^* = (1 - r^*) q^*.$$

By adding the expressions for $A(\lambda_k^*)^{-1}$ and $R(\lambda_k^*)^{-1}$, we obtain $A(\lambda_k^*)^{-1} + R(\lambda_k^*)^{-1} = q^*$, which we can solve to obtain $\lambda_k^* = \frac{A+R}{q^*}$. Also, we have $S_k^* = \left(\frac{V(\nu-1)}{q^*}\right)^{1/(\nu-1)}$. Note that for S_k^* to be positive, we must have $\nu > 1$. Hence, we have the following proposition:

Proposition 3. *The optimal solution to problem (P1) for each MNO-k, which is the local optimal solution of the problem (P0), for $A > 0$, $V > 0$ and $\nu > 1$ is given by*

$$\lambda_k^* = \frac{A + R}{q^*}, \tag{39}$$

$$S_k^* = \left(\frac{V(\nu - 1)}{q^*}\right)^{1/(\nu-1)}, \tag{40}$$

where q^* is the optima of the one-dimensional problem (37) – (38) evaluated at r^* in (36).

Note that the value of q is indeterminate at $r = 0$ and $r = 1$. We see that $\lim_{r \to 1}(1-r)^{1-r} = 1$ and $\lim_{r \to 1}(\frac{R}{1-r})^{1-r} = 1$, and the limit of q^* as $r \to 1$ is

$$\lim_{r^* \to 1} q^* = \omega A \left((\nu - 1)V\right)^{1/(\nu-1)} \left(\frac{\nu}{\nu - 1}\right)^{\frac{\nu}{\nu-1}}.$$

Likewise, since we have $\lim_{r \to 0}(A/r)^r = 1$, the limit of q^* as $r \to 0$ is

$$\lim_{r^* \to 0} q^* = \omega R \left((\nu - 1)V\right)^{1/(\nu-1)} \left(\frac{1}{\nu - 1}\right)^{\frac{1}{\nu-1}}.$$

Therefore, given the price of infrastructure ω, the optimal strategy of each MNO-k, which is computed by **Proposition 3**, gives the minimum amount of cache per unit area, $\lambda_k^* S_k^*$, while satisfying the latency constraint for each UE.

V. The Local 5G OP: Stackelberg Game Model-Based Cache Intensity Pricing for Multiple MNOs

In this section, we develop a novel strategy of the local 5G OP for renting out its infrastructure to K MNOs using

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

$$\max_r\ q = \omega\left(\frac{A}{r}\right)^r\left((\nu-1)V\right)^{\frac{1}{\nu-1}}\left(\frac{R}{1-r}\right)^{1-r}\left(r+\frac{1}{\nu-1}\right)^{r+\frac{1}{\nu-1}}(1-r)^{1-r} \tag{37}$$

$$\text{s.t.}\quad 0 \le r \le 1. \tag{38}$$

Stackelberg game and then propose the coalitional game to for the MNOs to share the rent among each other. Recall that the SBSs and cache storage can be shared by multiple MNOs. Given the infrastructure sharing scheme and the popularity based caching at the SBSs by each MNO, a question emerges: *how many files the Local 5G OP should be cache to serve all MNOs*? Clearly, it is sufficient for the local 5G OP to cache only the largest number of files requested by the MNOs, rather than the aggregate amount of files requested. This is due to the fact that when the MNOs request files based on popularity, there is an overlapping of the most popular files.

For example, let MNO-1 requests $S_1 = 10$ most popular files $\{f_1, \ldots, f_{10}\}$ to be cached and MNO-2 requests $S_2 = 15$ most popular files $\{f_1, \ldots, f_{15}\}$, assuming that the popularity rank of the file corresponds to the file's index, with file f_1 being the most popular and file f_{15} being the least popular. The local 5G OP can satisfy the requests of both MNOs by caching $S_I^* = \max(S_1, S_2) = 15$ most popular files $\{f_1, \ldots, f_{15}\}$, since $\{f_1, \ldots, f_{10}\} \subset \{f_1, \ldots, f_{15}\}$. This is much less than the aggregate amount $S_1 + S_2 = 25$ required to be stored when the cache is not shared among the MNOs.

In general, since $\mathcal{S}_k$ is the set of most popular files requested to be cached by the MNO-k, we can order the sets $\mathcal{S}_k$ as $\mathcal{S}_{\pi(1)} \subseteq \cdots \subseteq \mathcal{S}_{\pi(K)}$, where π represents the permutation of set $\mathcal{K}$. Thus, it is sufficient for local 5G OP to cache the largest set $\mathcal{S}_{\pi(K)}$ of a certain MNO that also meets the demands of all other MNOs. Since the largest set of most popular files also contains the smaller sets of most popular files and the local 5G OP allows cache-enable SBSs to be shared among multiple MNOs, the local 5G OP needs to handle only the largest cache intensity required by the MNO with the largest set of file demand denoted as $\lambda_I^* S_I^* = \max_k\{\lambda_k S_k\}$. In this regard, we model the pricing problem of the infrastructure as a Stackelberg game, where the local 5G OP is the leader and the MNOs are the followers. The local 5G OP, as the leader, will then compute the optimal price of cache intensity, ω^*, accordingly. Since the local 5G OP only needs to handle the largest cache intensity, the game is essentially simplified to a one-leader one-follower game, where the single follower is the MNO with the largest demand. The leader subgame problem is shown in (Q0) in (41).

After the local 5G OP computes the price of cache intensity of the infrastructure, the local 5G OP will declare the total rent to all MNOs. Since the MNOs are able to communicate with each other, we propose the coalitional formation game where the MNOs cooperate with each other to share the total rent. The relationship among the MNOs and the local 5G OP, and rent sharing among the MNOs are illustrated in Fig. 5.

In order to obtain the Stackelberg equibrium, the backward induction method is used. Therefore, each MNO-k, $k \in \mathcal{K}$, will send its best response in terms of its demand for cache intensity $\lambda_k^* S_k^*$ to the local 5G OP. The local 5G OP will handle the largest required cache intensity as $\lambda_I^* S_I^* = \max_k(\lambda_k^* S_k^*)$. Also, the local 5G OP will declare the price, ω^*, to all K MNOs that maximizes its revenue. Finally, using Shapley value, the MNOs can share the rental cost of cache intensity in a fair manner.

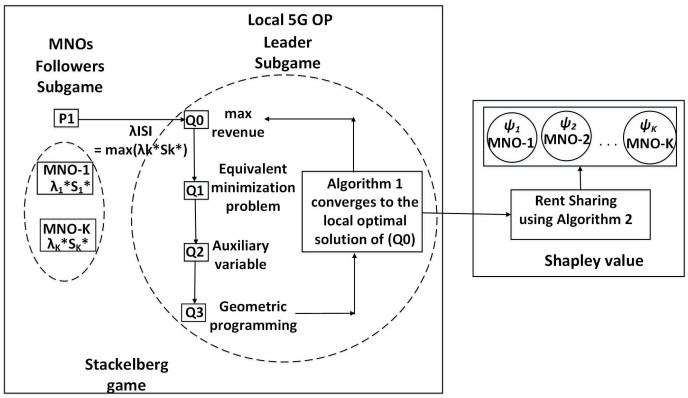

Fig. 5. Basic Idea of Hierarchical Relationship between the MNOs and the Local 5G OP.

A. Optimization Problem of the local 5G OP

In a Stackelberg game, the leader is assumed to have sufficient foresight to be able to anticipate the strategy of the follower. The local 5G OP aims at maximizing its revenue obtained by renting out the cache SBSs to the MNOs, while minimizing the operational expenses in terms of power consumption. Since the SBSs can be randomly selected by the MNOs, we do not know the number of MNOs that will utilize the SBSs in advance. In the worst-case scenario, all K MNOs will use the same SBSs simultaneously and the transmit power at each SBS will be $p_t = Kp_k$. Since the local 5G OP is renting out λ_I^* SBSs per unit area and, since we assume the worst-case scenario, the power consumption per unit area is then given by $Y(\lambda_I^*) = \lambda_I^*(Kp_k + p_c)$, where p_c denotes a fixed amount of circuit power.

When the cache-enable SBSs are shared by K MNOs, we can formulate the optimization problem of the local 5G OP as follows:

$$(Q0)\quad \max_\omega\quad \omega\lambda_I^* S_I^* - \gamma Y(\lambda_I^*), \tag{41}$$

where ω is the price of cache per unit area, and γ is the price of areal power consumption, where $\omega, \gamma > 0$. Note that we are not dealing with how other resources, e.g., computing, server, or transmission capacity are shared. We only consider the case where the cache storage in a unit area is shared among the MNOs.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

To obtain the solution of the Stackelberg game, we use backward induction method. Accordingly, we first solve the follower subgame problem. This essentially is to solve the optimal strategy of the MNO with the largest required cache intensity. The follower's solution is then used in the leader subgame problem, after which the leader problem is solved. The solution to the leader subgame gives the subgame perfect equilibrium of Stackelberg game.

Accordingly, we compute the largest cache intensity that the local 5G OP needs to provide to MNOs as $\lambda_I^* S_I^* = \max(\lambda_k^* S_k^*)$ by using λ_k^* and S_k^* of each follower MNO-k from **Proposition 3**. From **Proposition 3**, we can express λ_k^* and S_k^* in terms of ω as $\lambda_k^* = \frac{T_k}{\omega}$ and $S_k^* = U_k \omega^{-\frac{1}{(\nu-1)}}$, where $T_k = (A+R)/(q_k^*/\omega)$ and $U_k = [V(\nu-1)/(q_k^*/\omega)]^{1/(\nu-1)}$. Note that, in these expressions for T_k and U_k, from (37), the term q_k^*/ω is independent of ω, making T_k and U_k independent of ω as well. This transforms the first term of (41), which is $\omega \max_k\{\lambda_k^* S_k^*\} = \omega \lambda_I^* S_I^*$, into

$$\max_k\{U_k T_k\}\ \omega^{-1/(\nu-1)} = UT\omega^{-1/(\nu-1)}.$$

That is, $UT = \max_k\{U_k T_k\}$. Also, the second term in (41) is transformed into

$$\gamma Y(\lambda_I^*) = \lambda_I^* \gamma (K p_k + p_c) = T\overline{p}\omega^{-1},$$

where $\overline{p} = \gamma(K p_k + p_c)$. Therefore, we can rewrite the maximization problem in (41) as an equivalent minimization problem:

$$(Q1) \quad \min_{\omega>0} \quad T\overline{p}\omega^{-1} - UT\omega^{-1/(\nu-1)}. \tag{42}$$

The problem (Q1) is a signomial optimization problem over the price variable ω. In general, the problem (Q1) is a non-convex problem. However, this problem becomes convex at some values of ν. Nevertheless, we can obtain the solution to the problem in (Q1) via successive geometric programming (SGP). In order to solve (Q1), let us introduce an auxiliary variable $z \geq 0$ such that it upper bounds the objective function in (42) as follows:

$$z \geq T\overline{p}\omega^{-1} - UT\omega^{-1/(\nu-1)}. \tag{43}$$

Since minimizing the upper bound z minimizes the objective function in (42) as well, the problem (Q1) can be equivalently re-written in terms of this auxiliary variable as follows:

$$(Q2) \quad \chi = \min_{\omega>0} z \tag{44}$$

$$\text{s.t.} \quad \frac{T\overline{p}\omega^{-1}}{z + UT\omega^{-1/(\nu-1)}} \leq 1. \tag{45}$$

Here, the constraint in (45) is obtained after some algebraic manipulations of the bound in (43). To see the equivalence, for fixed ω, the optimal value of z is $z = T\overline{p}\omega^{-1} - UT\omega^{-1/(\nu-1)}$.

Since the constraint in (45) is a ratio of two posynomials, the problem (Q2) is also referred to as a complementary geometric program. Following the approach outlined in [31], [32], the problem (Q2) can be rewritten in the form of a geometric program by substituting the posynomial in the denominator of (45) by a monomial. We can transform a posynomial into a monomial using the arithmetic-geometric mean inequality $\sum_i x_i \geq \prod_i \left(\frac{x_i}{w_i}\right)^{w_i}$ for non-negative numbers $x_i \geq 0$, where $\sum_i w_i = 1$ and equality if and only if all x_i/w_i are the same. Accordingly, let us denote the posynomial in the denominator of (45) as $Q(z,\omega) = z + UT\omega^{-1/(\nu-1)}$, and its evaluation at point $(\overline{z},\overline{\omega})$ as $\overline{Q} = Q(\overline{z},\overline{\omega}) = \overline{z} + UT\overline{\omega}^{-\frac{1}{\nu-1}}$. Substituting each term of $Q(z,\omega)$ into x_i and since $(\overline{z} + UT\overline{\omega}^{-\frac{1}{\nu-1}})/\overline{Q} = 1$, we have the lower bound from the arithmetic-geometric mean inequality as $Q(z,\omega) \geq Q(z,\omega,\overline{z},\overline{\omega})$, where $Q(z,\omega,\overline{z},\overline{\omega})$ is a monomial given by

$$Q(z,\omega,\overline{z},\overline{\omega}) = \left(\frac{z\overline{Q}}{\overline{z}}\right)^{\frac{\overline{z}}{\overline{Q}}} \left(\frac{\omega^{\frac{-1}{\nu-1}}\overline{Q}}{\overline{\omega}^{\frac{-1}{\nu-1}}}\right)^{\frac{UT\overline{\omega}^{\frac{-1}{\nu-1}}}{\overline{Q}}} = E z^{\overline{\alpha}} \omega^{\overline{\beta}}.$$

The parameters are given by

$$\overline{\alpha} = \frac{\overline{z}}{\overline{Q}}, \qquad \overline{\beta} = \frac{-UT\overline{\omega}^{\frac{-1}{\nu-1}}}{(\nu-1)\overline{Q}}, \qquad E = \frac{\overline{Q}}{\overline{z}^{\overline{\alpha}}\overline{\omega}^{\overline{\beta}}}. \tag{46}$$

Therefore, the problem (Q2) can be transformed into an approximate geometric program by approximating the denominator of (45) by the lower bound $Q(z,\omega,\overline{z},\overline{\omega})$ as follows:

$$(Q3) \quad \chi = \min_{\omega} z \tag{47}$$

$$\text{s.t.} \quad \frac{T\overline{p}\omega^{-1}}{E z^{\overline{\alpha}}\omega^{\overline{\beta}}} \leq 1. \tag{48}$$

The problem (Q3) is now a geometric program which can be solved analytically. Note that the degree of difficulty of this problem is zero. The dual maximization problem of (Q3) is given by

$$\max_{\overline{\boldsymbol{\delta}}} \chi_d = \left(\frac{1}{\overline{\delta}_1}\right)^{\overline{\delta}_1} \left(\frac{T\overline{p}}{\overline{\delta}_2 E}\right)^{\overline{\delta}_2} \overline{\delta}_2^{\overline{\delta}_2}, \tag{49}$$

$$\text{s.t.} \quad \overline{\delta}_1 = 1, \tag{50}$$

$$\begin{pmatrix} 0 & -1-\overline{\beta} \\ 1 & -\overline{\alpha} \end{pmatrix} \begin{pmatrix} \overline{\delta}_1 \\ \overline{\delta}_2 \end{pmatrix} = 0. \tag{51}$$

We can solve for the optimal values of dual variables $\overline{\boldsymbol{\delta}}$ directly from (50) and (51). The matrix multiplication from (51) gives $(-1-\overline{\beta})\overline{\delta}_2 = 0$ and $\overline{\delta}_1 - \overline{\alpha}\overline{\delta}_2 = 0$. Summing the two equations gives $(-1-\overline{\beta})\overline{\delta}_2 + \overline{\delta}_1 - \overline{\alpha}\overline{\delta}_2 = 0$. Since $\overline{\delta}_1 = 1$, we obtain $\overline{\delta}_2 = 1/(1+\overline{\alpha}+\overline{\beta})$. By substituting the values of $\overline{\delta}_1$ and $\overline{\delta}_2$ in the dual problem χ_d, the optimal dual function χ_d^* yields

$$\chi_d^* = \left(\frac{T\overline{p}}{E}\right)^{\frac{1}{1+\overline{\alpha}+\overline{\beta}}}. \tag{52}$$

Since χ_d^* is the optimal solution of the dual problem, we can find the optimal variables z^* and ω^* of the problem (Q3) from χ_d^* as follows:

$$z^* = \overline{\delta}_1^* \chi_d^* = \left(\frac{T\overline{p}}{E}\right)^{\frac{1}{1+\overline{\alpha}+\overline{\beta}}} \quad \text{and}$$

$$\left(\frac{T\overline{p}(\omega^*)^{-1}}{E(z^*)^{\overline{\alpha}}(\omega^*)^{\overline{\beta}}}\right) = \delta_2^* \chi_d^* = \left(\frac{1}{1+\overline{\alpha}+\overline{\beta}}\right)\left(\frac{T\overline{p}}{E}\right)^{\frac{1}{1+\overline{\alpha}+\overline{\beta}}}.$$

Accordingly, we can compute $\omega^* = \left(\frac{T\overline{p}(1+\overline{\alpha}+\overline{\beta})}{E(z^*)^{\overline{\alpha}}}\left(\frac{E}{T\overline{p}}\right)^{1/(1+\overline{\alpha}+\overline{\beta})}\right)^{1/(1+\overline{\beta})}$. Since this is the first approximate values of z^* and ω^*, we can substitute

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

Algorithm 1 Successive Geometric Programming (SGP)

1: Set $z(0)$ and $\omega(0)$ to any arbitrary feasible values
2: **repeat**
3: Calculate $\overline{\alpha}(t), \overline{\beta}(t), E(t)$ using (46)
4: Calculate $\chi_d^*(t)$ using (52)
5: $\omega^*(t+1) \leftarrow \left(\frac{T\overline{p}(1+\overline{\alpha}+\overline{\beta})}{E(z^*)^{\overline{\alpha}}} \left(\frac{E}{T\overline{p}} \right)^{1/(1+\overline{\alpha}+\overline{\beta})} \right)^{1/(1+\overline{\beta})}$
6: $z^*(t+1) \leftarrow \chi_d^*(t)$
7: **until** convergence

these values back into $Q(z, \omega, z^*, \omega^*)$ and repeat the process, leading to new values of to z^* and ω^*. Thus, we have a SGP algorithm as given in **Algorithm 1**. This algorithm converges to a local optima of (Q2) [33] which is the local optimal solution of (Q0). This leads to the following proposition.

Proposition 4. ***Algorithm 1*** *converges to a locally optimal solution of the primal problem (Q0) for the local 5G OP. This, along with the largest cache intensity from the MNOs, gives the subgame perfect equilibrium of the one-leader and one-follower Stackelberg game.*

B. Rent Sharing Among MNOs using Shapley Value

After the local 5G OP computes the price per unit of cache intensity of the infrastructure from **Algorithm 1**, the local 5G OP will declare the total rent $\omega^* \lambda_I^* S_I^*$ to all MNOs. The MNOs will cooperate with each other and fairly divide the infrastructure rental fee among each other. Therefore, we model this situation as a cooperative game for rent sharing using the Shapley value. The coalition form of the K-person game is given by $(\mathcal{K}, v)$, where $\mathcal{K}$ is the set of K MNOs. The characteristic function of the game is denoted by v, where $v : 2^K \rightarrow \mathbb{R}$. The characteristic function maps every 2^K possible coalitions to a real number, referred to as the value of a coalition. When all K MNOs cooperate and form a single coalition, it is called the "grand coalition". The value of the grand coalition is $v(\mathcal{K})$. For the case when a subset of MNOs cooperate with each other and form a coalition $\mathcal{C} \subseteq \mathcal{K}$, we define the characteristic function of $\mathcal{C}$ as follows:

$$v(\emptyset) = 0 \quad \text{and} \quad v(\mathcal{C}) = \max_{k \in \mathcal{C}} \omega^* \lambda_k^* S_k^*, \tag{53}$$

where ω^* is the optimal price set by the local 5G OP following **Proposition 4**.

Proposition 5. *The characteristic function v given in (53) satisfies the sub-additivity property, $v(\mathcal{C}_1 \cup \mathcal{C}_2) \leq v(\mathcal{C}_1) + v(\mathcal{C}_2)$ for any two coalitions $\mathcal{C}_1$ and $\mathcal{C}_2$, where $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$.*

Proof: Let $k^* = \text{argmax}_{k \in \mathcal{C}_1 \cup \mathcal{C}_2} \omega^* \lambda_k^* S_k^*$. Since $\mathcal{C}_1$ and $\mathcal{C}_2$ are disjoint, k^* must belong to either $\mathcal{C}_1$ or $\mathcal{C}_2$, but not both. If $k^* \in \mathcal{C}_1$, we have $v(\mathcal{C}_1 \cup \mathcal{C}_2) = v(\mathcal{C}_1)$. Therefore, $v(\mathcal{C}_1 \cup \mathcal{C}_2) \leq v(\mathcal{C}_1) + v(\mathcal{C}_2)$. Likewise, if $k^* \in \mathcal{C}_2$, then $v(\mathcal{C}_1 \cup \mathcal{C}_2) = v(\mathcal{C}_2)$. Again, $v(\mathcal{C}_1 \cup \mathcal{C}_2) \leq v(\mathcal{C}_1) + v(\mathcal{C}_2)$. Thus, v given in (53) is sub-additive. ∎

Note that the sub-additivity implies that the MNOs forming a bigger coalition will have smaller cost. Let the cost allocated to MNO-k be ψ_k. Then, the value of coalition $\mathcal{C}$ should be divided among each MNO in the coalition $\mathcal{C}$ such that $v(\mathcal{C}) = \sum_{k \in \mathcal{C}} \psi_k$. The Shapley value function, ψ_k, is a function that assigns to each possible characteristic function of a K-MNO game, v with a K-tuple, $\psi(v) = (\psi_1(v), \psi_2(v), \ldots, \psi_K(v))$. Here $\psi_k(v)$ represents the worth or value of MNO-k in the game with characteristic function v and is defined by the following axioms of fairness:

1) Efficiency: $\sum_{k \in \mathcal{K}} \psi_k(v) = v(\mathcal{K})$.
2) Symmetry: If k and l are such that $v(\mathcal{C} \cup \{k\}) = v(\mathcal{C} \cup \{l\})$ for every coalition $\mathcal{C}$ not containing k and l, then $\psi_k(v) = \psi_l(v)$.
3) Dummy: If i is such that $v(\mathcal{C}) = v(\mathcal{C} \cup \{i\})$ for every coalition $\mathcal{C}$ not containing i, then $\psi_i(v) = 0$.
4) Additivity: If u and v are characteristic functions, then $\psi(u + v) = \psi(u) + \psi(v)$.

There exists a unique function that satisfies all these fairness axioms which is given by:

$$\psi_k(v) = \sum_{\substack{\mathcal{C} \subseteq \mathcal{K} \\ k \in \mathcal{C}}} \frac{(|\mathcal{C}| - 1)!(n - |\mathcal{C}|)!}{k!} [v(\mathcal{C}) - v(\mathcal{C} - \{k\})]. \tag{54}$$

This gives the average marginal contribution made by MNO-k when it joins a random coalition $\mathcal{C}$. We take this value as the fair payoff allocation among the MNOs inside the coalition.

The direct computation of the Shapely value using the analytical formula given in (54) quickly becomes computationally infeasible as the number of MNOs increases. However, due to the special structure of the characteristic function for our rent sharing problem, as given in (53), we can apply a simple algorithm to allocate the cost among the MNOs by recognizing that our problem is equivalent to the airport runway cost sharing problem studied by Littlechild and Owens [34]. In our case, the local 5G OP sets the total rent, $\omega^* \lambda_I^* S_I^*$ using **Algorithm 1**, and the MNOs divide the rent among each other according to their required cache intensities $\lambda_k^* S_k^*$ obtained from **Proposition 3**.

Algorithm 2 presents the rent sharing algorithm among MNOs. Depending on the requirement of the MNOs, the MNOs are first sorted in an ascending order of their cache intensities, i.e., demands, $\lambda_{\pi(1)}^* S_{\pi(1)}^* \leq \cdots \leq \lambda_{\pi(K)}^* S_{\pi(K)}^*$, where π denotes the permutation of set $\mathcal{K}$. The cost of meeting the smallest demand, $\omega^* \lambda_{\pi(1)}^* S_{\pi(1)}^*$, is divided equally among all K MNOs. Then, the incremental cost $\Delta = \omega^* \lambda_{\pi(2)}^* S_{\pi(2)}^* - \omega^* \lambda_{\pi(1)}^* S_{\pi(1)}^*$ of meeting the second smallest demand is shared equally among all the MNOs, except the MNO with the smallest demand. The process is repeated until the incremental cost $\Delta = \omega^* \lambda_{\pi(K)}^* S_{\pi(K)}^* - \omega^* \lambda_{\pi(K-1)}^* S_{\pi(K-1)}^*$ is allocated only to the MNO with the largest demand. The cost allocated among the MNOs by **Algorithm 2** is equivalent to the Shapley value of the coalition game $(\mathcal{K}, v)$ [34]. Note that **Algorithm 2** has the worst-case computational complexity of $O(K^2)$.

VI. Numerical Results

Unless otherwise stated, the transmit power of SBS is $p = 1$ Watt, noise power is $\sigma^2 = -150$ dBm, the number of video files in the cloud is $F = 10^5$, the size of the file requested by each UE is $x_f = 10^5$ bits, path-loss exponent is $\alpha = 5$, i.e.,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

Algorithm 2 Cost sharing among MNOs

1: Initialize $\lambda_0 S_0 = 0$ and $\psi_k = 0$ for all $k \in \mathcal{K}$
2: Arrange the MNOs in an ascending order of their cache intensity $\lambda^*_{\pi(1)} S^*_{\pi(1)} \leq \cdots \leq \lambda^*_{\pi(K)} S^*_{\pi(K)}$, where π is the permutation of set $\mathcal{K}$
3: **for** $k = 1$ **to** K **do**
4: $\quad \Delta = \omega^* \lambda^*_{\pi(k)} S^*_{\pi(k)} - \omega^* \lambda^*_{\pi(k-1)} S^*_{\pi(k-1)}$.
5: $\quad$ **for** $i = k$ **to** K **do**
6: $\quad\quad \psi_{\pi(i)} \leftarrow \psi_{\pi(i)} + \Delta/(K - k + 1)$
7: $\quad$ **end for**
8: **end for**

suburban area without line of sight [35]. The SINR threshold is $\theta = 10$ dB. Each MNO is assumed to have the same number of subchannels as $L = 4$. Based on 5G requirements in [36], the latency of data transmission should be less than 10^{-3} sec. Therefore, we limit the total delay to be $\mathbb{P}(D \geq 10^{-3}) \leq 0.01$ in (19). We assume that there is a single server, $m = 1$, in the cloud, where the mean arrival rate of file requests is $\phi = 0.8$, mean service time $\tau = 5 \times 10^{-3}$, the coefficients of variation of inter-arrival time and service time are $c_a = 2$ and $c_s = 1$, respectively, as such, $\mathbb{E}[D_{\text{bh}}] = 0.0051$ sec.

A. Optimal Strategy of an MNO-k from **Proposition 3**

We first show the optimal strategy of a single MNO-k from **Proposition 3** by varying the bandwidth as $W_k = 1$ gigaherz (GHz), 2GHz, 3GHz, and 4GHz while the price of cache intensity is ω is assumed to be unity. The UE intensity is $\xi_k = 20/(\pi \times 100^2)$.

In Fig. 6, we see that λ^*_k increases when ν is increased. However, λ^*_k does not change with W_k for a given ν. On the other hand, in Fig. 7, S^*_k increases when W_k is increased when ν is fixed. The reason is that a higher bandwidth yields a larger throughput. When the SBS can transmit a file faster, the MNO renting the SBS can benefit more from serving its UEs and hence the MNO should cache more files. Hence, for $\nu > 1$, increasing W_k leads to a decrease in the cost q^*, which in turn leads to an increased S^*_k. On the other hand, since $R \propto 1/G_u$, from (39), the G_u term cancels out in the expression for λ^*_k, making λ^*_k independent of W_k. We can see from Fig. 7 that for $\nu \in (1.8, 2.3)$ most of the files are equally popular. Therefore, the cache size is made small to reduce the cost for the MNO. By contrast, as ν increases, some files become more popular than the other files. Therefore, it is worth for the MNO-k to request for increased cache size. However, when ν is large, i.e., $\nu > 2.3$, only few files are very popular, and hence the cache size decreases. We see that while S^*_k decreases with an increasing of ν, λ^*_k also increases. Fig. 6 and Fig. 7 show the tradeoff between λ^*_k and S^*_k for a given ν.

In Fig. 8, we set $W_k = 2$ GHz. We can observe that increasing the UE intensity, ξ_k, decreases the cache size, S^*_k. Also, increasing the number of subchannels L_k reduces the cache size S^*_k. This is due to the fact that increasing UE intensity or the number of subchannels reduces the throughput of the UE, hence increases the downlink transmission delay. The MNO will cache fewer files to keep the total delay less than a given threshold. We can see that the UE intensity has a significant impact on the cache size.

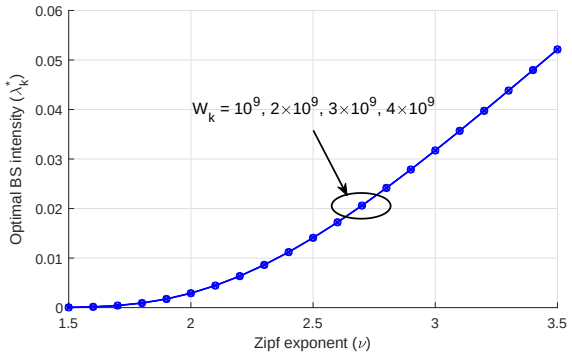

Fig. 6. Optimal SBS intensity (λ^*_k) versus Zipf exponent (ν).

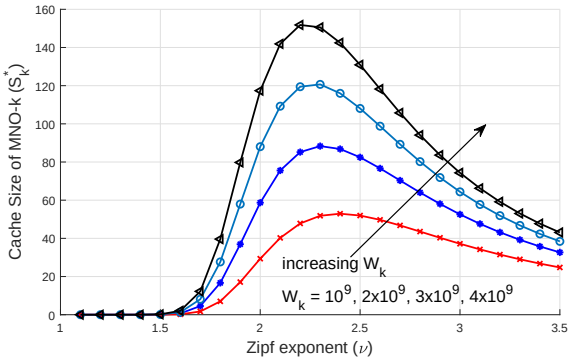

Fig. 7. Optimal cache size (S^*_k) versus Zipf exponent (ν).

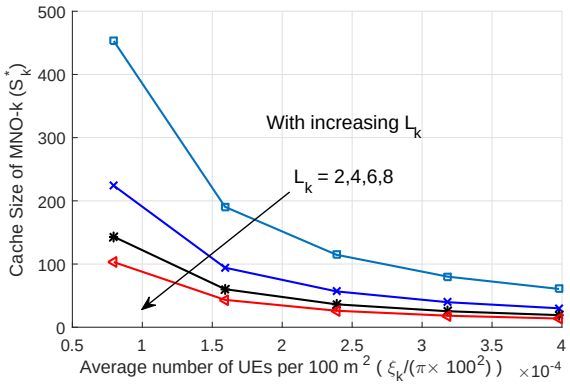

Fig. 8. Cache size of MNO-k (S^*_k) versus average number of UEs per $100m^2$ with varying number of subchannels L_k.

B. Optimal Price (ω^*) and Maximum Profit (z^*) of the Local 5G OP at Subgame Perfect Equilibrium of Stackelberg Game

After obtaining the best response of each follower MNO from **Proposition 3**, with the infrastructure sharing deployment, the local 5G OP supports the largest demand required by the MNOs. The local 5G OP will then compute the optimal price ω^* of the infrastructure so as to maximize its profit z^* by using the SGP in **Algorithm 1**. In Figs. 9–12, we demonstrate the optimal strategy of the leader local 5G OP and the optimal strategy of three follower MNOs at the subgame

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

perfect equilibrium of the Stackelberg game. We assume the bandwidth and the UE intensity to be $[W_1, W_2, W_3]$= [1GHz, 1.5GHz,2GHz] and $[\xi_1, \xi_2, \xi_3] = [10, 15, 20]/(\pi \times 100^2)$, respectively.

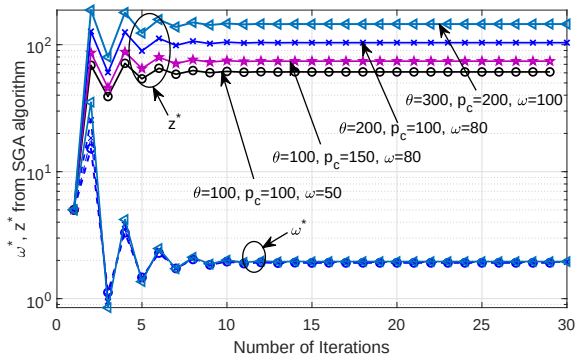

Fig. 9. Convergence of the optimal price of infrastructure (ω^*) and maximum profit (z^*) from **Algorithm 1**.

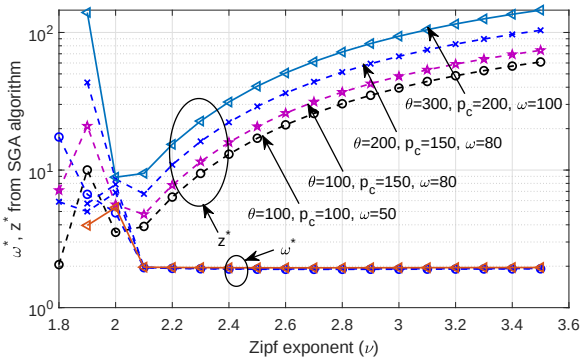

Fig. 10. Optimal price of infrastructure (ω^*) and maximum profit (z^*) from **Algorithm 1** versus Zipf exponent (ν).

The best responses of an MNO (in terms of optimal SBS intensity and optimal cache size to be leased) at the subgame perfect equilibrium of Stackelberg game are shown in Fig. 11 and Fig. 12, respectively. In Fig. 9, we show the convergence of the optimal price of infrastructure, ω^*, and the maximum profit, z^*. When the price of areal power consumption, γ, and the circuit power, p_c, increase, z^* also increases, while both γ and p_c have very small effect on ω^*. The SGP algorithm is effective since both ω^* and z^* converge within a few iterations. We see that with increasing ν, ω^* remains constant. The reason is that the price of infrastructure depends only on the maximum demand $\lambda_I^* S_I^*$ required by the MNOs. However, when ν increases, z^* also increases. By increasing ν, the amount of infrastructure required by the MNO increases. This gives higher profit to the InP. The best response λ_k^* and S_k^* of each MNO-k at the equilibrium are shown in Fig. 11 and Fig. 12, respectively.

In Fig. 11, the curves for λ_k^* of all K MNOs are identical. The trends in Fig. 11 are very similar to those in Fig. 6. Also, varying of price of infrastructure ω and circuit power p_c does not have any impact on λ_k^*. In Fig. 12, we plot S_k^* of different MNOs versus ν. When the value of ν changes, the cache size S_1^* of MNO-1 is the largest while S_3^* of MNO-3 becomes the smallest. Although the MNO-3 has the largest

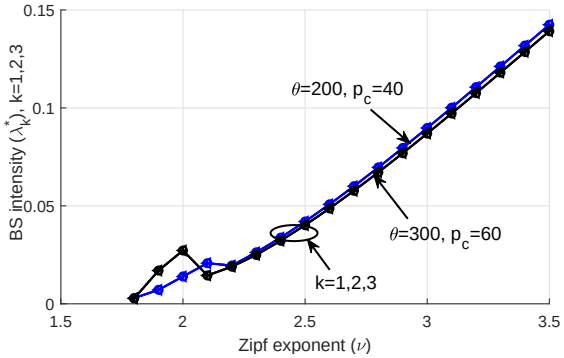

Fig. 11. Optimal SBS intensity (λ_k^*) versus Zipf exponent (ν) at subgame perfect equilibrium of Stackelberg game.

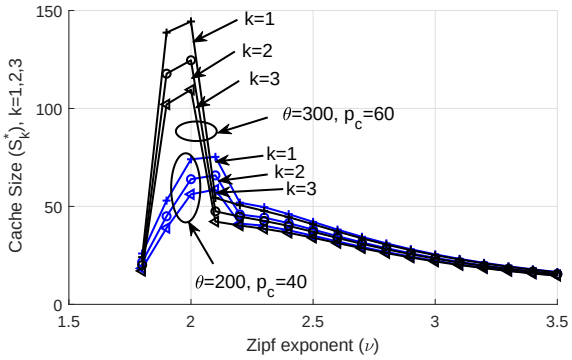

Fig. 12. Optimal cache size (S_k^*) Zipf exponent (ν) with (ω^*) from **Algorithm 1**.

bandwidth, the UE intensity has a more significant effect on S_k^*. Accordingly, the UE intensity of MNO-1, ξ_1, is the lowest and the S_1^* becomes the highest. Varying the price of areal power consumption γ and circuit power p_c does not change optimal S_k^*. We see that the initial curves of S_k^* in Fig. 12 fluctuates, which is due to the random initial values of Z^* and S^* of SGP in **Algorithm 1**.

C. Maximum Profit of the local 5G OP at Subgame Perfect Equilibrium of Stackelberg Game and Shapley Value

Fig. 13 presents the total rent $\omega^* \lambda_I^* S_I^*$, which is the first term of problem (Q0), and the Shapley value of each MNO-k, ψ_k, versus the Zipf exponent, ν. For a given value of ν, we observe that ψ_1 is the highest while ψ_3 is the lowest. The reason is that the MNO-1 has the lowest number of UEs per unit area while MNO-3 has the highest UE intensity. As discussed in Section VI-A, a lower ξ_k tends to increase the cache size S_k^*. Thus, the MNO-1 will try to buy the highest amount of infrastructure when compared with those for MNO-2 and MNO-3. It can be seen that all three MNOs can divide the rent in a fair manner by using **Algorithm 2**, which is the Shapley value of their cooperative game. The maximum profit of local 5G OP versus ν is plotted in Fig. 14, which is obtained by solving the problem (Q0) using **Algorithm 1**. We observe that when ν increases, the maximum profit of the local 5G OP also increases. Also, when γ and p_c increase, the maximum profit of the InP enhances significantly.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2019.2948630, IEEE Transactions on Mobile Computing

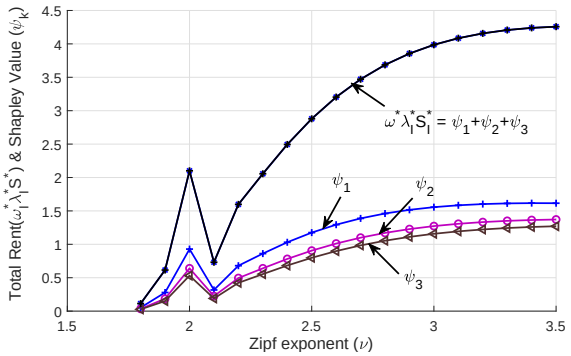

Fig. 13. Total rent $\omega^*\lambda^*S_k^*$ and ψ_k versus Zipf exponent (ν) at Stackelberg equilibrium.

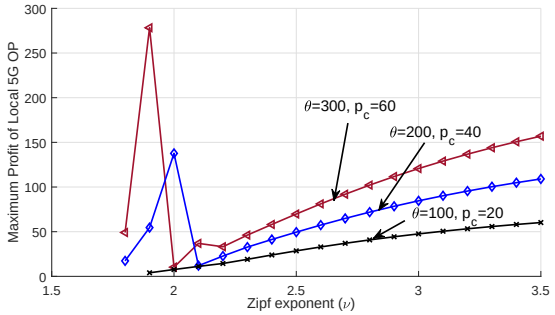

Fig. 14. The maximum profit versus Zipf exponent (ν) at Stackelberg equilibrium.

The major observations from these numerical results are: (i) The Zipf exponent, ν, has a significant impact on the optimal strategy of each MNO-k. (ii) When the bandwidth increases, the cache size is also increased for a given ν. However, changing the bandwidth does not affect the SBS intensity. (iii) The UE intensity and subchannels have high influence on the cache size. (iv) The price of power consumption per unit area, γ, and the circuit power, p_c, affect the profit of the local 5G OP significantly.

VII. Conclusion

We have proposed a novel deployment of indoor wireless networks for local 5G OP with virtualized cache-enabled SBSs. The local 5G OP provides the infrastructure, consisting of RAN and and edge caching, to multiple MNOs. With infrastructure sharing deployment, multiple MNOs are able to use the common infrastructure simultaneously. The throughput of videos/contents transmission from the SBS to each UE for interference limited case has been derived. Each MNO aims to minimize the cost of rented cache intensity subject to latency constraint at each UE while SBSs transmit contents/videos to the UEs. The problem of each MNO has been transformed into geometric program and a closed-form solution has been obtained. Likewise, we have modeled the pricing problem for sharing the cache-enabled SBSs infrastructure between the local 5G OP and the MNOs as a Stackelberg game where the local 5G OP is the leader and the MNOs are the followers. With infrastructure sharing deployment, we have shown that the single leader multi-followers Stackelberg game has become a single leader single follower Stackelberg game. Then, we have obtained the optimal strategy of the local 5G OP at the subgame perfect equilibrium of Stackelberg game via successive geometric programming. Lastly, sharing of the rent of infrastructure among the MNOs has been done via Shapely value. However, the proposed framework can be enhanced by considering spectrum sharing in addition to infrastructure sharing among the MNOs as well as the SBSs and cache storage can be implemented by multiple local 5G OPs.

Acknowledgment

This work has been financially supported by the Academy of Finland 6Genesis Flagship (grant no. 318927).

References

[1] T. Sanguanpuak, S. Guruacharya, N. Rajatheva, and M. Latva-aho, "Edge Caching for Cache Intensity under Probabilistic Delay Constraint," *Proc. IEEE Global Commun. Conf.*, Dec. 2018.

[2] Cisco, "Cisco visual networking index: forecast and methodology, 2016-2021," *White Paper*, Jun., 2017.

[3] M. Latva-aho, "Micro Operators for Vertical Speific Service Deliver in 5G," [Online]. Available: http://5g.ieee.org/images/files/pdf/Workshop8Dec16/D1-Latva-aho-Micro-Operator.pdf, Jan., 2017.

[4] T. Sanguanpuak, *et al.*, "On spectrum sharing among micro-operators in 5G," *Proc. European Conference on Networks and Communications (EuCNC)*, pp.1–6, 2017.

[5] M.G. Kibria, *et al.*, "Shared Spectrum Access Communications: A Neutral Host Micro Operator Approach," ," *IEEE J. on Sel. Areas in Commun.*, vol. 35, no, 8, Aug. 2017.

[6] T. Sanguanpuak, D. Niyato, N. Rajatheva, and M. Latva-aho, "Network Slicing with Mobile Edge Computing for Micro-Operator Networks in Beyond 5G," *to be appeared in Proc. IEEE WPMC*, 2018.

[7] M. Matinmikko, M. Latva-aho, M. Ahokangas, *et al.*,"Micro Operators to Boost Local Service Delivery in 5G," *Wireless Personal Communications* Springer, pp.69-82, Jul. 2017.

[8] Qualcomm, "Ultra-Reliable Low Latency 5G for Industrial Automation," *Heavy Reading White Paper*, 2018.

[9] K. Ishizu, "R&D status on micro cell operator and spectrum sharing toward 5G and beyond," [Online]. Available: https://www.nict.go.jp/en/asean_ivo/4otfsk000040lmol-att/a1513219427659.pdf, Nov. 2017.

[10] P. Nikolich, *et al.*, "Standards for 5G and Beyond: Their Use Cases and Applications," *IEEE 5G Tech Focus* Vol. 1, No. 2, Jun., 2017.

[11] M. Rebato, M. Mezzavilla, S. Rangan, and M. Zorzi, "Resource Sharing in 5G mmWave Cellular Networks," *IEEE Conf. on Computer Commun. Workshops (INFOCOM WKSHPS)*, pp. 271-276, Apr. 2016.

[12] T. Sanguanpuak, *et al.*, "Infrastructure sharing for mobile network operators: analysis of trade-offs and market," *IEEE Trans. on Mobile Computing*, 2018.

[13] T. Sanguanpuak, *et al.*, "Inter-operator infrastructure sharing: trade-offs and market," *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, pp.73–78, 2017.

[14] T. Sanguanpuak, *et al.*, "Multi-Operator Spectrum Sharing for Small Cell Networks: A Matching Game Perspective," *IEEE Trans. on Wireless Commun.*, vol. 16, no. 6, pp. 3761-3774, 2017.

[15] P. Luong, F. Gagnon, C. Despins, and L. Tran, "Joint Virtual Computing and Radio Resource Allocation in Limited Fronthaul Green C-RANs," *IEEE Trans. on Wireless Commun.*, vol. 17, no. 4, pp. 2602-2617, 2018.

[16] X. Wang, *et al.*, "Cache in the air: exploiting content Caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, Feb. 2014.

[17] J. Li, *et al.*, "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE J. on Sel. Areas in Commun.*, vol. 34, no. 8, Aug. 2016.

[18] F. Shen, *et al.*, "A Stackelberg game for incentive proactive caching mechanisms in wireless networks," *Proc. IEEE Conf. on Global Commun. (GLOBECOM)*, Dec. 2016.

[19] M. Dehghan, W. Chu, P. Nain, and D. Towsley, "Sharing LRU Cache Resources among Content Providers: A Utility-Based Approach,"*arXiv:1702.01823v1 [cs.NI] 6 Feb. 2017.*

[20] E. El Haber, T. M. Nguyen, and C. Assi, "Joint Optimization of Computational Cost and Devices Energy for Task Offloading in Multi-Tier Edge-Clouds," *IEEE Trans. on Commun.*, vol. 67, no. 5, pp. 3407-3421, 2019.
[21] https://www.oulu.fi/6gflagship/.
[22] J.G. Andrews, F. Baccelli, and R.K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. on Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.
[23] F.W.J. Olver, *et al.*, eds., "Hurwitz Zeta Function," *NIST Digital Library of Mathematical Functions*, [Online]. Available: http://dlmf.nist.gov/25.11 Release 1.0.16 of 2017-09-18.
[24] J. Sondow and E.W. Weisstein, "Riemann Zeta Function," *MathWorld – A Wolfram Web Resource*, [Online]. Available: http://mathworld.wolfram.com/RiemannZetaFunction.html
[25] J. Sondow and E.W. Weisstein, "Hurwitz Zeta Function." *MathWorld – A Wolfram Web Resource*, [Online]. Available: http://mathworld.wolfram.com/HurwitzZetaFunction.html
[26] M. Cha, *et al.*, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Trans. on Networking*, vol. 17, no. 5, pp. 1357–1370, Oct. 2009.
[27] Christine Fricker, *et al.* "Impact of traffic mix on caching performance in a content-centric network," *Proc. IEEE INFOCOM Workshops*, pp. 310–315, 2012.
[28] J. Vilaplana, *et al.*, "A queuing theory model for cloud computing," *The Journal of Supercomputing*, pp.492–507, Jul. 2014.
[29] L. Kleinrock, *Queueing Systems, Vol.I: Theory. Wiler, New York*, 1975.
[30] R.J. Duffin, *Geometric Programming Theory and Application*. John Wiley & Sons, 1967.
[31] M. Avriel and A.C. Williams, "An extension of geometric programming with applications in engineering optimization," *Journal of Engineering Mathematics*, vol. 5, no.3, pp 187–194, 1971.
[32] A. J. Morris, "Approximation and complementary geometric programming," *SIAM Journal on Applied Mathematics*, vol. 23, no. 4, pp. 527–531, Dec. 1972.
[33] M. Avriel and A.C. Williams, "Complementary geometric programming," *SIAM Journal on Applied Mathematics*, vol. 19, no. 1, pp. 125–141, 1970.
[34] S.C. Littlechild and G. Owen, "A simple expression for the Shapley value in a special case," *Management Science*, vol. 20, no. 3, Theory Series, 370–372, 1973.
[35] 3GPP TR.36.814 v9.0.0 (2010-03), "3rd Generation partnership project; technical specification group radio access network; evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRA physical layer aspects (Release 9); *http://www.qtc.jp/3GPP/Specs/36814-900.pdf*
[36] 5G PPP Architecture Working Group, "View on 5G architecture (Version 2.0)" *www.5G-ppp.eu/wp-content/uploads/2017.*

Tachporn Sanguanpuak is a posdoctoral research fellow at Centre for Wireless Communications (CWC) at the University of Oulu, Finland. She received the B.Eng. degree in Telecommunication Engineering from King Mongkuts Institute of Technology Ladkrabang (KMITL), Thailand; M.Eng. in Telecommunications Engineering from Asian Institute of Technology (AIT), Thailand; and Ph.D. in Communication Engineering from university of Oulu, Finland. She currently works in an Academy of Finland 6Genesis Flagship project. Her research interests are in the areas of future wireless radio networks (beyond LTE, and beyond 5G), game theory, distributed optimization, drone networks, vehicular networks, positioning, mmWave, machine learning, deep learning, and applications of AI in wireless communications.

Dusit Niyato is currently a professor in the School of Computer Engineering, at the Nanyang Technological University, Singapore. He obtained his Bachelor of Engineering in Computer Engineering from King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok, Thailand. He received Ph.D. in Electrical and Computer Engineering from the University of Manitoba, Canada. His research interests are in the area of radio resource management in cognitive radio networks and broadband wireless access networks.

Nandana Rajatheva (SM'01) received the B.Sc. (Hons.) degree in Electronics and Telecommunication Engineering from the University of Moratuwa, Sri Lanka, in 1987 ranking first in the graduating class. He obtained M.Sc. and Ph.D. degrees from the University of Manitoba, Winnipeg, MB, Canada, in 1991 and 1995, respectively. He was a Canadian Commonwealth Scholar during the graduate studies in Manitoba. He held Professor / Associate Professor positions at University of Moratuwa and Asian Institute of Technology (AIT), Thailand from 1995-2010. He is currently a Professor with the Centre for Wireless Communications (CWC), University of Oulu, Finland. His research interests include waveforms, channel coding for 5G, small cells, smart grid communications, and application of machine learning in 5G use cases.

Matti Latva-aho received the M.Sc., Lic.Tech. and Dr. Tech (Hons.) degrees in Electrical Engineering from the University of Oulu, Finland in 1992, 1996 and 1998, respectively. From 1992 to 1993, he was a Research Engineer at Nokia Mobile Phones, Oulu, Finland after which he joined Centre for Wireless Communications (CWC) at the University of Oulu. Prof. Latva-aho was Director of CWC during the years 1998-2006 and Head of Department for Communication Engineering until August 2014. Currently he is Professor of Digital Transmission Techniques at the University of Oulu. He serves as Academy of Finland Professor in 2017 – 2022. His research interests are related to mobile broadband communication systems and currently his group focuses on 5G systems research. Prof. Latva-aho has published 300+ conference or journal papers in the field of wireless communications. He received Nokia Foundation Award in 2015 for his achievements in mobile communications research.