IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON NEW WAVEFORM DESIGN AND AIR-INTERFACE FOR FUTURE HETEROGENEOUS NETWORK TOWARDS 5G

Received December 31, 2018, accepted January 21, 2019, date of publication February 19, 2019, date of current version March 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2900195

INVITED PAPER

Integrated Use of Licensed- and Unlicensed-Band mmWave Radio Technology in 5G and Beyond

XI LU[1], EDUARD SOPIN[2], VITALY PETROV[3], OLGA GALININA[2,3], DMITRI MOLTCHANOV[3], KIRILL AGEEV[2], SERGEY ANDREEV[3], YEVGENI KOUCHERYAVY[3], KONSTANTIN SAMOUYLOV[2], AND MISCHA DOHLER[1]

[1]King's College London, London WC2R 2LS, U.K.
[2]Department of Applied Mathematics and Probability Theory, Peoples' Friendship University of Russia (RUDN University), 117198 Moscow, Russia
[3]Laboratory of Electronics and Communications Engineering, Tampere University of Technology, 33720 Tampere, Finland

Corresponding author: Xi Lu (xi.lu@kcl.ac.uk)

This work was supported by the RUDN University Program 5-100 and funded by RFBR according to the research projects under Grant 18-00-01555 (18-00-01685) and Grant 19-07-00933.

ABSTRACT The 3GPP standardization rapidly moves forward with studies of a wide-bandwidth waveform as well as an adaptation of the emerging 5G new radio (NR)-based access to the unlicensed spectrum (NR-U). One of the basic architectures for NR-U involves carrier aggregation of an anchor—licensed—NR carrier and a secondary carrier in unlicensed spectrum, which altogether allows for seamless traffic offloading in scenarios where multi-gigabit data rates are required. While today's research on NR-U addresses mostly physical- and protocol-layer aspects, a system-level performance of the NR-U offloading mechanisms has not been investigated thoroughly. In this paper, we develop a mathematical queuing-theoretic framework that is mindful of the specifics of millimeter-wave (mmWave) session dynamics and may serve as a flexible tool for the analysis of various strategies for the integrated use of licensed and unlicensed mmWave bands in terms of the session drop probability and system utilization. To illustrate this, we select three distinct strategies (based on sequential service, probabilistic offloading, or proportional splitting), and complement our mathematical models with a detailed performance evaluation in a representative massive augmented/virtual reality scenario. Based on this quantitative analysis of the selected schemes, we conclude that proportional splitting of traffic between the two mmWave bands leads to a better performance. We believe that the contributed mathematical analysis can become an important building block in further system development and service optimization across many usage scenarios.

INDEX TERMS
5G and beyond, 5G-U, integration of licensed and unlicensed bands, mathematical analysis, mmWave-based access, massive AR/VR, NR-U, radio resource allocation.

I. INTRODUCTION

Information and communications technology continues to develop at a rapid pace by already surpassing the stage of fifth-generation (5G) pre-commercial trials. According to CCS Insight [1], some countries announced their plans to deploy 5G by the end of 2019, and over the following 6 years the number of 5G subscriptions is expected to count to 2.6 billion. The mass adoption of 5G will likely coexist with an increased popularity of emerging services, such as autonomous driving, massive drone surveillance, Extended Reality (XR) – that unifies Augmented Reality (AR), Virtual Reality (VR), Mixed Reality (MR) [2], and Synchronized Reality (SR) [3] – and many others. To support these advanced and *bandwidth-hungry* applications, major vendors and operators embark on the path of complementing their conventional microwave network deployments with an emerging radio technology of significantly higher capacity [4] operating in the millimeter-wave (mmWave, up to 100 GHz) spectrum.

Particularly, the 3GPP is currently exploring several licensed mmWave sub-bands (e.g., 24.25 – 27.5, 27.5 – 29.5,

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

37 − 40, 64 − 71 GHz) for the 5G mmWave cellular networks [5]. The bandwidths that such systems can potentially utilize are between 500 MHz and 2 GHz, which results in the cell capacities of several gigabits per second [6], [7]. Meanwhile, IEEE is actively exploring the unlicensed band at 60 GHz for the next-generation wireless local area networks (WLANs) and direct device-to-device (D2D) interactions [8]. The recent specifications of IEEE 802.11ay technology offer the channel bandwidths from 2.16 GHz to 8.64 GHz, which leads to the theoretic capacity of up to few tens (optimistically, 100) of gigabits per second [9].

While the envisioned performance of these mmWave systems may be beyond the typical demands of today's bandwidth-hungry 5G services, the consumers are increasingly involved in utilizing more and more advanced data-driven applications. As a result, the shortage of available wireless spectrum may arise soon even for more capable mmWave network layouts. This is expected for beyond-5G massive and dense deployments of intelligent devices, each requiring reliable high-rate connectivity [10]: e.g., a football stadium where spectators are equipped with AR/VR glasses [11], dense urban traffic of autonomous vehicles engaged in collective driving [12], large swarms of flying drones [13], etc.

A standalone mmWave cellular technology may be insufficient to support these *extremely bandwidth-hungry* scenarios, which calls for either exploring even higher frequency bands (can be challenging in terms of the radio equipment capabilities [14]) or integrating licensed and unlicensed mmWave frequencies in extreme operating environments [11]. Exploring the latter option, the 3GPP standardization rapidly moves forward with evolving the legacy licensed assisted access (LAA) to 5G New Radio (NR) based access to the unlicensed spectrum (NR-U), which is also known as 5G-U in [11]. One of the basic architectures for NR-U assumes carrier aggregation of an anchor – licensed – NR carrier and a secondary carrier in unlicensed spectrum, which altogether allows for seamless traffic offloading in extreme scenarios where multi-gigabit data rates are required.

Given the fact that today's studies on NR-U address primarily physical- and protocol-layer features, a system-level performance characterization of NR-U offloading mechanisms has not been investigated thoroughly. This work is a systematic account of our recent research efforts in this space, which proposes a mathematical methodology to understand effective mmWave spectrum integration and leverage it for improved beyond-5G system performance. In particular, based on our rationale as well as a supportive review of technology background and requirements, we contribute the following:

- A mathematical queuing-theoretic framework that allows for analyzing the integrated use of licensed- and unlicensed-band mmWave access on the system level by evaluating a number of key performance indicators (i.e., system resource utilization and session drop probability) for different offloading strategies.
- A detailed numerical analysis of three example offloading strategies with quantitative conclusions on their expected performance.

The rest of this text is organized as follows. In Section II, we review the current trends in the discussed mmWave radio technology integration as well as outline the attractive scenarios therein. Our system model is then specified in Section III, which is followed by the proposed mathematical framework in Section IV. The key results of our numerical study are introduced and explained in Section V. The paper terminates with some concluding remarks.

II. CURRENT TRENDS AND PROSPECTIVE APPLICATIONS

A. INTEGRATION OF LICENSED AND UNLICENSED mmWAVE BANDS

Driven by a lack of spectrum and its expensive licensing, wireless industry has historically been interested in traffic offloading. This includes approaches to operate cellular signal on unlicensed frequencies, which have been further accommodated by 3GPP LTE radio technology [15], [16]. Among them, there are LTE on Unlicensed bands (LTE-U) and LAA technologies [17], which constitute industry-grade solutions to aggregate licensed and unlicensed microwave bands with the aim to meet the ever-growing user traffic demands.

The Carrier Aggregation (CA) [18] capabilities that enable LTE-U/LAA also make it possible to extend these concepts from ultra-high and super-high bands to extremely-high frequencies, namely, mmWave spectrum around and above 30 GHz. Hence, an attractive goal is to integrate, e.g., the 28 GHz licensed band employed by 3GPP NR and the 60 GHz unlicensed band, which is utilized for IEEE 802.11ad solution by Wireless Gigabit Alliance (WiGig) [11] and its recent successor, IEEE 802.11ay [8], as well as by the legacy Wireless HD, ECMA-387, and 802.15.3c radios.

Theoretically, for the 28 GHz licensed band, there is approximately 2 GHz of free bandwidth, while unlicensed mmWave spectrum offers around 14 GHz of spare bandwidth in the US (7 GHz in most other countries) [8]. Naturally, 28 GHz frequency may become the primary carrier – to reliably provide secure and robust connectivity – while 60 GHz unlicensed frequency might be dedicated to offloading cellular traffic as a secondary carrier. Consequently, the connectivity therein is more opportunistic and can only offer best-effort quality.

B. EMERGING APPLICATIONS FOR INTEGRATED mmWAVE SYSTEMS

The envisioned integration of mmWave bands can support multiple scenarios (see Fig. 1). One example conveying extensive amounts of information between a vehicle and its nearby entities is Vehicle-to-Everything (V2X) communications (see Fig. 1(a)) [19], which is also known as an umbrella term for a range of dedicated connectivity options, such as V2I (Vehicle-to-Infrastructure), V2V (Vehicle-to-Vehicle), V2P (Vehicle-to-Pedestrian), V2D (Vehicle-to-Device), and V2G (Vehicle-to-Grid). As an integral part of 5G-grade

FIGURE 1. Prospective applications for 5G and beyond. (a) V2X communications in urban environments. (b) Extreme augmented and virtual reality. (c) Airborne communications systems.

Internet of Things (IoT) [20], V2X demands data rates on the order of up to tens of Gbit/s and latencies at the level of milliseconds, which are only possible with abundant mmWave bandwidths [21].

Further, Extreme Virtual Reality (X-VR) and Extreme Augmented Reality (X-AR) become of interest (see Fig. 1(b)) [11]. The conventional AR and VR services are typically enabled with wired transmissions, pre-cached data, dedicated users, and heavy audio/visual content, or cloud-assisted AR/VR [22]. In contrast, X-VR and X-AR are truly wireless, real-time, and can support multiple perceptions: beyond audio/visual content and including touch, heartbeat, tactile, and other types of experience.[1] As follows from examples of live broadcasts in stadiums and operas, consistently high data rates become a key demand, while any significant delay immediately leads to a service failure [23].

Last but not least, a new class of scenarios that the envisioned mmWave integration might support relatively soon is Unmanned Aerial Vehicles (UAVs, see Fig. 1(c)) [24], which are highly mobile and have stringent safety requirements as well as demand reliable and high-rate wireless connections. Notably, the licensed component of the integrated mmWave radio system may supply UAVs with robust and secure connections; on the other hand, the unlicensed system component can enable timely offloading of traffic in critical and/or massive UAV applications [11]. Thereby, an important direction of further research is to evaluate the performance promise of the integrated mmWave system across these emerging scenarios.

C. TOWARDS PERFORMANCE ANALYSIS OF INTEGRATION OPTIONS

To analyze the above use cases on the system level, it is essential to develop a mathematical abstraction of, e.g., X-AR/X-VR applications in a representative scenario as displayed in Fig. 1. This should reflect various formats of multimedia content in a highly dense area as well as more stringent service requirements as compared to enhanced mobile broadband. For instance, while the data rate for a flat 4K video of 4096×2048 px may be estimated as approximately 15.6 Mbps [25], the support of 4K 360^o video streaming of minimum 12288×6144 px (which is a 9-fold increase) in a football stadium with the consumer density of $4/m^2$ will incur the data rate of 561.6 Mbps/m^2, or even 748.8 Mbps/m^2 in the case of a 12-fold expansion.

Moreover, the link capacity requirements of, e.g., an Oculus VR set[2] producing VR images of 75 fps with the quantization parameter of 15 is around 40 Mbps [26]; therefore, assuming 200 VR users one may expect that the total throughput may reach an unprecedented value of 8 Gbps. Similarly, X-AR also imposes extreme demands on ultra-reliable and low-latency communications: no extra delay is allowed in live streaming; and this is along with high data rate demands for a large number of users. Arguably, with the growing numbers and densities of subscribers, any current radio technology alone – be it 3GPP LTE, 5G NR, or WiGig – will become insufficient to support the emerging X-AR/X-VR services.

To this end, current research is revisiting the conventional approaches to spectrum sharing and licensed–unlicensed system management for higher frequencies. In particular, Cognitive Radio (CR) that enables opportunistic spectrum utilization in the context of 3GPP LTE femtocells has been analyzed in [27] and [28]. Further, LTE-U based on 3GPP Release 10 technology and first defined in Non-Listen Before Talk (NLBT) regions [18], and LAA in Listen Before Talk (LBT) regions also ratified by 3GPP [29] have been comprehensively studied in [30]–[32]. Finally, a substantial

[1] Another recent term for a similar immersive experience is hyper-reality.

[2] https://www.oculus.com/

effort has been devoted in [33] and [34] to evaluating LTE-WLAN Aggregation (LWA) [35], which originates from LTE Dual Connectivity (DC) in 3GPP Release 12, as well as a spectrum sharing framework named Licensed Shared Access (LSA) in Europe and its alternative solution known as Spectrum Access Systems (SAS) in the US.

At the same time, an extension of protocols of 5G-U or NR-U has been discussed in both academic and industrial communities. In September 2017, 3GPP standardization has introduced a study item on NR-based access to unlicensed spectrum, which includes NR-based physical layer design for unlicensed operation, unlicensed bands below and above 6 GHz (up to 52.6 GHz), MAC-layer considerations (channel access, scheduling, radio-link monitoring/failure), as well as coexistence within and between NR-based operation in unlicensed and LTE-based LAA and with other radios [36]. Subsequent 3GPP studies address different architectural scenarios, such as NR-based LAA cells connected to NR anchor cells operating in licensed spectrum or NR-based cell operating standalone in unlicensed spectrum, as well as include waveform considerations for the 60 GHz band (at the moment, scheduled for Release 17 and 18) [37], while Release 16 aims to define the notion of NR-U already in 2019 [38].

On the academic side, spectrum sharing and interference mitigation as well as dynamic frequency selection (DFS) are being studied at the early stages of NR-U development [39]; similarly, a Listen-Before-Receive (LBR) technique for unlicensed shared spectrum under the coexistence with NR is considered in [40]. Moreover, Qualcomm as a driver behind LTE-U has also suggested the potential use cases and the underlying spectrum sharing properties of NR-U, which has attracted focused attention to this initiative [41].

However, most of today's research efforts in spectrum integration concentrate on microwave bands; therefore, a timely target is to conduct a first-order assessment of the forthcoming mmWave-based integration technologies by offering a suitable modeling framework, which is capable of accounting for mmWave-specific radio propagation properties. Particularly, rapid fluctuations in the amounts of demanded radio resources for a given session caused by inherent dynamics of the mmWave wireless links have to be incorporated into the framework [42]–[44]. Along these lines, the currently available microwave mechanisms listed above may become the benchmark solutions for the subsequent system-level performance comparison. This will deliver substantiated conclusions on the expected operating potential on top of the existing legacy schemes in typical usage scenarios.

III. CONSIDERED SYSTEM MODEL

In this section, we formulate a representative scenario for analyzing the envisioned system operation and summarize the main assumptions of the corresponding system model. We aim at evaluating a crowded outdoor scenario (e.g., associated with a mass sports event or a fair) featuring extremely high data rate and latency-sensitive demand generated by e.g., X-AR/X-VR users therein.[3] These engage in immersive participation through their personal devices, such as smartphones, tablets, or headgear. The proposed methodology is specifically tailored to the session-level analysis of mmWave communications, where one has to ensure that the traffic load does not exceed the system capacity in the long run. Local surges in the user demand are assumed to be handled at lower layers by taking advantage of buffering, scheduling, as well as radio-level mechanisms, such as beamforming, beamsteering, and power control [21], [45], [46]. We specifically focus on the ability of the mmWave system to handle its offered load with a selected strategy for the integrated use of licensed and unlicensed bands. Below, our key assumptions and system parameters are detailed, while Table 1 collects the notation utilized throughout this paper.

TABLE 1. System modeling notation.

Symbol	Description
N_k	Maximum number of sessions in RP $k, k = 1, 2$
R_k	Number of resource units per RP $k, k = 1, 2$
λ	Arrival rate of session requests
μ^{-1}	Average session duration
$F(x)$	CDF of requested resource shares
γ	Arrival rate of triggers
ζ	Probability to be routed to RP 1
ϵ	Share of resources requested from RP 1
p_r	Probability that a session requires r resource units
$\xi(t)$	Number of sessions in the system at time t
$\theta_i(t)$	Resource occupied by session i at time t
$p_r^{(n)}$	Probability that n sessions occupy r resources
$\tilde{\lambda}$	Session arrival rate to RP 2 (sequential strategy)
$q_n(r)$	Stationary probability of the aggregated process
δ_k	Resource utilization coefficient for RP k
δ	System resource utilization coefficient
$p_{B,k}$	Arriving session drop probability by RP k
$p_{T,k}$	Ongoing session drop probability by RP k
p_B	Total probability of dropping a session upon arrival
p_T	Total probability of dropping an accepted session

The RP index: 1 for unlicensed, 2 for licensed.

A. SYSTEM MODELING ASSUMPTIONS

Wireless users running their resource-hungry applications are assumed to be by default connected to the 5G mmWave cellular network operating in *licensed* frequency bands (e.g., 28 GHz) as well as managing session admissions and offloading procedures. To augment the capacity while maintaining the required levels of service quality, the network provider may employ radio resources available in *unlicensed* spectrum (e.g., 60 GHz). Hence, any data connection of an end user (i.e., a session) may be transparently offloaded onto this additional pool of resources. A decision whether to exploit one pool or another (or both at the same time) is made

[3] Intel®, "Get Courtside with Intel True VR", available at: https://www.intel.com/content/www/us/en/sports/nba/overview.html

by a centralized scheduling entity once and immediately before a session starts.

1) RADIO RESOURCE STRUCTURE

By abstracting a particular implementation of serving mmWave technologies, we consider a system of *two* independent radio resource pools, which differ in terms of their size and structure. Each resource pool (RP) i, $i = 1, 2$, incorporates a total of C_i discrete resource units and may accommodate up to N_i simultaneous sessions from different users. Importantly, these two pools are assumed to be operating on orthogonal frequencies, and their respective data transmissions do not interfere with each other. Assuming that all of the users are initially registered on the cellular network, we may omit the connection establishment phase and instead concentrate on the dynamics of user requests to serve their high-rate data sessions.

2) SESSION REQUEST ARRIVALS AND REQUIREMENTS

We assume an infinite population of users and let their session requests arrive according to a homogeneous Poisson process with the intensity of λ. Each new session has a random duration, which is distributed exponentially with the mean of μ^{-1} and requires a random number of radio resource units. The number of requested resource units is independent and identically distributed (i.i.d.) for all users and is drawn from a general discrete distribution with the cumulative distribution function (CDF) of $F(x)$. We also assume that the particularities of the network infrastructure deployment, geometry information, channel quality, and application rate requirements are all incorporated into $F(x)$, similarly to, e.g., [44].

3) CONNECTION ADMISSION CONTROL

Upon arrival of a session request, the cellular system initiates an admission control procedure. If the amount of the currently unoccupied resources is sufficient (as determined by the *offloading strategies* below), this session is accepted and the system reserves the requested radio resources. Whenever sufficient resources cannot be guaranteed at the moment of its arrival, a session is considered *dropped*.

4) SESSION TRIGGERS AND REALLOCATION

In our system, each active session may be interrupted by a *trigger*, which indicates that the corresponding user modifies its data rate demands. That may happen as a result of, e.g., suddenly degraded channel quality in cases of the radio link blockage [44], which alters the video quality regime [47], or changes incurred by the ongoing user application. We emphasize that incorporating these demand fluctuations into the system model is imperative for a comprehensive analysis of high-rate mmWave scenarios. To the best of our knowledge, this feature has not been captured analytically in the existing literature.

For a certain tagged session, the arrivals of triggers follow a Poisson process with the intensity of γ. Each new trigger induces changes in the current resource allocation i.e., the respective ongoing session immediately releases all of its occupied resources and re-applies for another resource grant. Without the loss of generality, we assume that this change occurs instantly, which is equivalent to ''shifting'' the trigger by a fixed signaling delay. The size of a newly generated resource request follows the same CDF $F(x)$ as the initial resource requirement does, and remains independent of the previous history of this particular demand.

We note that if the reallocation request produced by a trigger cannot be supported under the below offloading strategies, then the entire session is terminated and regarded as dropped during service. If throughout the session service period the system is always able to allocate the required number of radio resource units, the session in question is considered to be served successfully.

B. HEURISTIC OFFLOADING STRATEGIES

Addressing the operation of our target scenario, we aim at comparing the following three offloading strategies (illustrated in Fig. 2):

- *Baseline Sequential Strategy* We assume that upon a new arrival, the system first attempts to offload the entire session to the licensed-band resource pool (RP 1 in our notation). If the licensed spectrum cannot offer a sufficient amount of resources, the system moves this session to the unlicensed-band resource pool (RP 2). If the latter also lacks the required number of resource units, the session is dropped. If either of the two attempts succeeds, the session is accepted to the corresponding pool.
- *Joint Probabilistic Strategy* Here, we assume that an arriving flow of session requests is divided in two. The probability that a session is routed to the first resource pool is ζ. With the complementary probability, $1-\zeta$, this session requests resources from the second pool. A session is dropped if no sufficient resources are available in the selected resource pool.
- *Joint Proportional Strategy* Here, each session is served by both resource pools simultaneously. That is, upon its arrival, a share ϵ of the requested resources is granted from the licensed-band pool, while the other part, $1-\epsilon$, of resources comes from the unlicensed-band pool. A session is dropped if there are no sufficient resources at either of these resource pools.

C. MAIN METRICS OF INTEREST

In this work, we consider both user- and system-centric metrics of interest. These include (i) arriving session drop probability, p_B; (ii) ongoing session drop probability, p_T; and (iii) system resource utilization coefficient, δ. Let $U_i(t)$, $t > 0$, $i = 1, 2$, be the amount of resources occupied at time t in the RPs 1 and 2, respectively. The resource utilization coefficients for the individual pools and the joint utilization

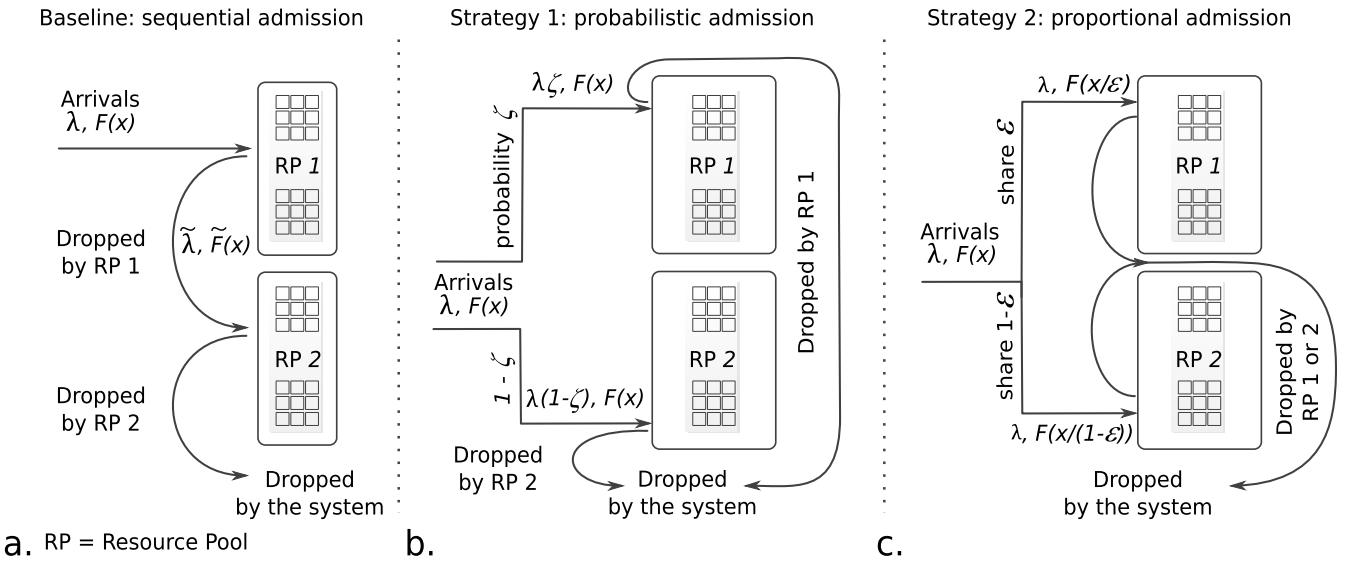


FIGURE 2. An illustration of three considered resource allocation strategies.

coefficient are then defined by,

$$\delta_i = \lim_{t\to\infty} \frac{1}{tR_i} \int_0^t U_i(t)dt, \quad i = 1, 2,$$
$$\delta = \lim_{t\to\infty} \frac{1}{t(R_1 + R_2)} \int_0^t [U_1(t) + U_2(t)]\, dt. \qquad (1)$$

IV. PERFORMANCE EVALUATION FRAMEWORK

In this section, we develop a performance evaluation framework that captures the essentials of the previously discussed resource allocation strategies. Below, we begin by describing the basic modeling principles and then proceed with specifying and assessing the queuing models associated with the considered strategies. Finally, the sought performance metrics are derived.

A. SOLUTION AT A GLANCE

Our proposed analysis is based on the tools of queuing theory, which allow us to evaluate the performance of the three offloading strategies as defined above: (i) baseline sequential strategy, (ii) joint probabilistic strategy, and (iii) joint proportional strategy.

1) BASELINE SEQUENTIAL STRATEGY

In case of the sequential strategy, we model the service process in licensed and unlicensed bands separately as two independent but successive queuing systems with random resource requirements and trigger-induced demand fluctuations. The session drop probability in the first (''licensed'') queuing system, $p_{B,1}$, corresponds to the case where a newly arriving session observes all of the licensed-band resources occupied and is then routed to unlicensed bands. The second (''unlicensed'') queuing system with demand fluctuations captures offloading onto unlicensed bands and receives a thinned flow of sessions with the request arrival rate of $p_{B,1}\lambda$. The total arriving session drop probability for this strategy may be established as $p_B = p_{B,1}\, p_{B,2}$.

The system utilization coefficient, δ, can be calculated as

$$\delta = \frac{\delta_1 C_1 + \delta_2 C_2}{C_1 + C_2}, \qquad (2)$$

where δ_1 and δ_2 are the resource utilization coefficients for licensed and unlicensed bands, respectively, while C_1 and C_2 denote the corresponding available resources.

2) JOINT PROBABILISTIC STRATEGY

According to the joint probabilistic strategy, the arrival flow is divided probabilistically between licensed (probability ζ) and unlicensed (probability $1 - \zeta$) bands. In this case, we again rely upon a queuing system with random resource requirements and model the service process in the two bands separately. The ongoing session drop probability is defined as $p_B = \zeta p_{B,1} + (1 - \zeta)p_{B,2}$, while the overall system resource utilization may be obtained similarly to (2).

3) JOINT PROPORTIONAL STRATEGY

A key feature of the joint proportional strategy is that upon a session arrival its resource demand is split into two parts. The CDFs of the resource requests on licensed and unlicensed pools, $F_1(x)$ and $F_2(x)$, are given as,

$$F_1(x) = F(x/\epsilon), \quad F_2(x) = F(x/(1-\epsilon)), \qquad (3)$$

where $F(x)$ is the CDF of the required number of resource units.

If both pools are able to allocate the corresponding shares of the requested demand, the session in question is accepted by the system. Otherwise, if the amount of resources in either of the pools is insufficient, this session is dropped permanently. Based on these assumptions, we represent our system

$$\sum_{s=0}^{C_1}\sum_{i=0}^{\min(s,r)}\frac{p_{r-i}p_i^{(n-1)}}{p_r^{(n)}}p_{s-i}+\sum_{i=0}^{r}\frac{p_{r-i}p_i^{(n-1)}}{p_r^{(n)}}\left(1-\sum_{k=0}^{C_1-i}p_k\right)=1+\sum_{i=0}^{r}\frac{p_{r-i}p_i^{(n-1)}}{p_r^{(n)}}\sum_{s=i}^{C_1}p_{s-i}-\sum_{i=0}^{r}\frac{p_{r-i}p_i^{(n-1)}}{p_r^{(n)}}\sum_{k=0}^{C_1-i}p_k=1. \quad (4)$$

as a queuing model with resource requests that are random and variable over the session duration, as well as having *two* resource pools of size C_1 and C_2.

In contrast to the previous two strategies, where one may reduce the model to considering only one resource pool, this case requires the characterization of joint dynamics across both interdependent queues. The queuing model with two resource pools, random resource requirements, and trigger-induced fluctuation is an extension of the system with a single resource pool. Hence, we subsequently focus on the queuing model with *two resource pools* as a more complex and general case.

B. QUEUING SYSTEM WITH TWO RESOURCE POOLS

1) DECOMPOSITION INTO TWO QUEUING SYSTEMS

Here, we consider the system with two resource pools. We decompose the model into two independent queues that correspond to the available resource pools and assume that the respective arrival flows are independent (in general they are not, but this assumption results in a tight approximation as shown in Section V). Recall that the sessions arriving into the first queuing system are characterized by the rate of λ and the CDF $F(x)$ of the requested resource units. Hence, the arrival rate into the second queue is $\tilde{\lambda} = \lambda p_{B,1}$ and the associated number of the requested resource units follows the CDF $\tilde{F}(x)$, see Fig. 2.

We further refer to the distributions of the amounts of the resource units by employing the probability mass functions $\{p_r\}$ and $\{\tilde{p}_r\}$, respectively, where $p_r, r \geq 0$, is the probability that an arriving session requires r discrete resource units. The behavior of the first queuing system may be described by a stochastic process $X_1(t) = (\xi(t), \theta_1(t), \ldots, \theta_{\xi(t)}(t))$, where $\xi(t)$ is the number of sessions in the system at time t and $\theta_i(t)$ is the number of resource units occupied by i-th customer. The set of states of $X_1(t)$ is defined as follows,

$$S_1 = \bigcup_{n=0}^{N_1} S_{1,n}, \quad (5)$$

where for n active sessions the subset $S_{1,n}$ is given by,

$$S_{1,n} = \{(n, r_1, \ldots, r_n) : \sum_{i=1}^{n} r_i \leq C_1, p_{r_i} > 0, i \geq 1\}. \quad (6)$$

The actual size of S_1 depends on the distribution $\{p_r\}$. For example, if $p_r > 0$ for any discrete r that satisfies $0 < r \leq C_1$, then the number of states corresponding to n sessions in the system and occupying j resource units ($\sum_{i=1}^{n} r_i = j$, $r_i > 0$) equals the number of n-combinations with repetitions produced by a set of $j - n$ elements (that is, $\binom{j-1}{n-1}$). Summing up all of the binomial coefficients $\binom{j-1}{n-1}$ from n to C_1 yields that the number of states in $S_{1,n}$ is given by $\binom{C_1}{n}$. Then, if $N_1 = C_1$, the total number of states in S_1 is 2^{C_1}, which is impossible to handle in practical calculations. Below, we develop an efficient state aggregation approach that allows decreasing the number of states in the considered system.

2) STATE AGGREGATION FOR THE PROCESS $X_1(T)$

Since the complexity of direct analysis of the process $X_1(t)$ is extremely high, we exploit the *state aggregation* technique [48]. The core idea is to keep track of the number of sessions in the system and only the aggregated amount of the occupied resources. The principal challenge of this approach is that the number of resources released upon a service completion does not coincide with the session resource request distribution. We address this issue by utilizing a Bayesian estimate of the CDF of the number of released resource units.

It has been demonstrated in [49] that for the system without triggers, where the resources are released according to the proposed Bayesian approach, the steady-state distributions in the original and the modified system with the state aggregation are identical. Let A denote the event that n sessions occupy r resource units in total, while event B corresponds to the case where a session releases j resource units upon its departure. Then, the conditional probability $P(B|A)$, which indicates that j resource units are released when n sessions occupy a total of r resource units, may be obtained by using Bayes' formula,

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)} = \frac{p_j p_{r-j}^{(n-1)}}{p_r^{(n)}}, \quad (7)$$

where $p_r^{(n)}$ is the probability that n sessions occupy r resource units, which may be calculated from the distribution $\{p_r\}$ by utilizing a convolution i.e.,

$$p_r^{(n)} = \sum_{i=0}^{r} p_i p_{r-i}^{(n-1)}, \quad n \geq 2, \quad (8)$$

where $p_{r-j}^{(n-1)}$ is the probability that $r - j$ resource units are occupied by $n-1$ sessions and p_j is the probability that exactly j resource units are occupied by a single session. Note that $p_r^{(1)} = p_r, r \geq 0$, and $p_r^{(0)} = \delta_{0,r}$, where $\delta_{i,j}$ is the Kronecker's delta function.

3) MARKOV MODEL WITH AGGREGATED STATES

The behavior of the system with aggregated states can be described by a Markov process $X_2(t) = (\xi(t), \delta(t))$, where $\xi(t)$ is the number of sessions at time t and $\delta(t)$ is the total

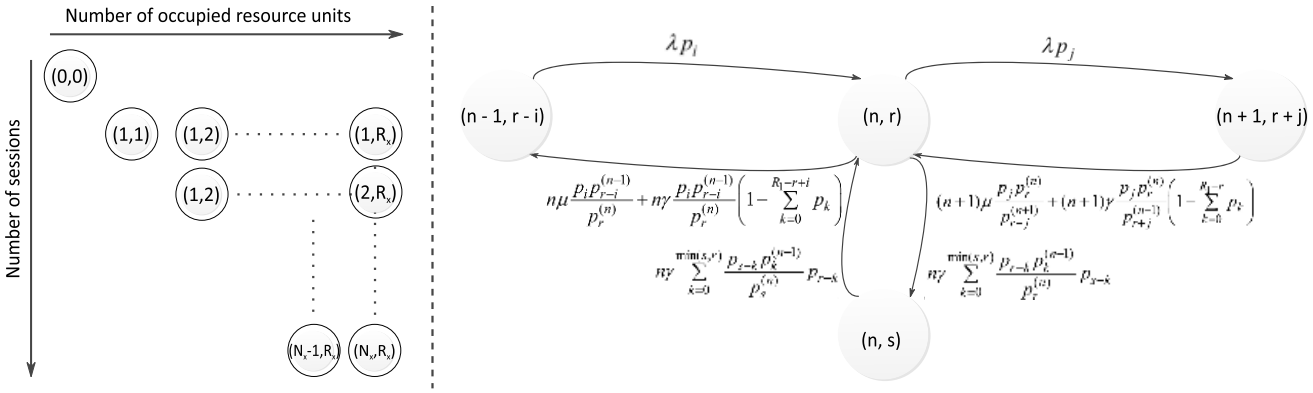


FIGURE 3. An illustration of state transitions in the Markov model with aggregated states.

amount of the occupied resources. The set of states is then given by,

$$S_2 = \bigcup_{0 \le n \le N_1} S_{2,n}, \tag{9}$$

where the subset $S_{2,n}$ for n active sessions is defined as,

$$S_{2,n} = \left\{ (n, r) : 0 \le r \le C_1, p_r^{(n)} > 0 \right\}. \tag{10}$$

Here, Fig. 3 illustrates a fragment of the state transition diagram for the state (n, r), $n \in (0, N)$ and $r \in (0, C_1)$. Upon a new session arrival, the system shifts from state (n, r) to state $(n + 1, j)$, $j \in [r, C_1]$, with the probability p_{j-r}, and the corresponding transition rate equals λp_{j-r}. The session drop rate in the state (n, r) equals $\lambda(1 - \sum_{j=0}^{C_1-r} p_j)$. Upon a session departure from the system, the probability that $r - i$ resource units are released and the system shifts to the state $(n - 1, i)$ is given by (7). Therefore, the transition rate from state (n, r) to state $(n - 1, i)$ is $n\mu p_{r-i} p_i^{(n-1)} / p_r^{(n)}$.

The same Bayesian estimate is employed in case of a trigger arrival. Particularly, in state (n, r), a session releases $r - k$ resource units upon the trigger arrival with the probability $p_{r-k} p_k^{(n-1)} / p_r^{(n)}$. If the new resource requirement is $s - k$ with the probability p_{s-k}, then the transition rate from (n, r) to (n, s) is given by $n\gamma \sum_{k=0}^{\min(s,r)} p_{s-k} p_{r-k} p_k^{(n-1)} / p_r^{(n)}$. Otherwise, if this new resource requirement exceeds $C_1 - k$, the session is dropped. Further, under the law of total probability we observe that

$$\sum_{i=0}^{r} \frac{p_{r-i} p_i^{(n-1)}}{p_r^{(n)}} = 1,$$

and the service completion rate in state (n, r) is $n\mu$.

Similarly, the trigger arrival rate in state (n, r) is given by $n\gamma$. This may be confirmed by observing that (4), shown at the top of the previous page, holds true.

4) STATIONARY PROBABILITIES FOR THE AGGREGATED PROCESS

Further, we introduce the stationary probabilities of the process $X_2(t)$ as

$$q_n(r) = \lim_{t \to \infty} P\{\xi(t) = n, \delta(t) = r\}, \quad (n, r) \in S_{2,n}. \tag{11}$$

Observing Fig. 3, we may derive the balance equations for states (n, r), $n \in (0, N_1)$, $(n, r) \in S_{2,n}$ as given by (12), as shown at the top of the next page. The equations corresponding to the boundary states are obtained similarly, see (13) and (14), as shown at the top of the next page, for detail. The system of equilibrium equations (12)-(14) (along with the normalization condition) has a unique solution, which produces the stationary distribution (11). We continue by characterizing the metrics of interest.

The arriving session drop probability $p_{B,1}$ may be calculated as

$$p_{B,1} = 1 - \sum_{(n,r) \in S_2, n < N_1} q_n(r) \sum_{j=0}^{C_1 - r} p_j, \tag{15}$$

while the ongoing session drop probability $p_{T,1}$ is given by

$$p_{T,1} = \sum_{\substack{(n,r) \in S_2, \\ n > 0}} q_n(r) \sum_{j=0}^{r} \frac{p_j p_{r-j}^{(n-1)}}{p_r^{(n)}} \left(1 - \sum_{i=0}^{C_1 - r + j} p_i \right). \tag{16}$$

Finally, the utilization coefficient δ_1 for the RP 1 can be derived as

$$\delta_1 = \frac{1}{C_1} \sum_{(n,r) \in S_2} r q_n(r). \tag{17}$$

5) SECOND RESOURCE POOL

As established above, the arrival rate into the second queuing system is given by $\tilde{\lambda} = \lambda p_{B,1}$, and the probability distribution $\{\tilde{p}_r\}$ can be produced by the conditional probability approach (7). The probability that a session dropped at the

$$\left(\lambda\sum_{j=0}^{C_1-r}p_j+n\mu+n\gamma\right)q_n(r)=\lambda\sum_{j:(n-1,r-j)\in S_{2,n-1}}q_{n-1}(r-j)p_j+(n+1)\mu\sum_{j:(n+1,r+j)\in S_{2,n+1}}q_{n+1}(r+j)\frac{p_jp_r^{(n)}}{p_{r+j}^{(n+1)}}$$
$$+(n+1)\gamma\left(1-\sum_{j=0}^{C_1-r}p_j\right)\sum_{j:(n+1,r+j)\in S_{2,n+1}}q_{n+1}(r+j)\frac{p_jp_r^{(n)}}{p_{r+j}^{(n+1)}}$$
$$+n\gamma\sum_{j:(n,j)\in S_{2,n}}q_n(j)\sum_{i=0}^{\min(j,r)}\frac{p_{j-i}p_i^{(n-1)}}{p_j^{(n)}}p_{r-i},\ 0<n<N,\ (n,r)\in S_{2,n},\tag{12}$$

$$\lambda\sum_{j=0}^{C_1}p_jq_0(0)=\mu\sum_{j:(1,j)\in S_{2,1}}q_1(j)+\gamma\left(1-\sum_{j=0}^{C_1}p_j\right)\sum_{j:(1,j)\in S_{2,1}}q_1(j),\tag{13}$$

$$(N_1\mu+N_1\gamma)\,q_{N_1}(r)=\lambda\sum_{j:(N_1-1,r-j)\in S_{2,N_1-1}}q_{N_1-1}(r-j)p_j+N_1\gamma\sum_{j:(N_1,j)\in S_{2,N_1}}q_{N_1}(j)$$
$$\times\sum_{i=0}^{\min(j,r)}\frac{p_{j-i}p_i^{(N_1-1)}}{p_j^{(N_1)}}p_{r-i},\ (N_1,r)\in S_{2,N_1}.\tag{14}$$

first queue requires r resource units is determined as

$$\tilde{p}_r=\frac{1}{p_{B,1}}p_r\sum_{(n,j)\in S_2,j\geq C_1-r+1}q_n(j),\quad r\geq 0.\tag{18}$$

The equilibrium equations for the second queuing system can be obtained similarly to (12)-(14). Further, the session drop probabilities $p_{B,2}$, $p_{T,2}$, and the utilization coefficient δ_2 are calculated based on the stationary distribution for the second system in (15)-(17), respectively. Finally, for the sequential admission strategy, the arriving session drop probability is established as

$$p_B=p_{B,1}p_{B,2},\tag{19}$$

while the ongoing session drop probability is given by

$$p_T=(1-p_{B,1})p_{T,1}+p_{B,1}p_{T,2}.\tag{20}$$

C. ANALYSIS OF NON-SEQUENTIAL STRATEGIES

1) JOINT PROBABILISTIC ADMISSION STRATEGY

In case of the probabilistic admission strategy, both resource pools operate independently. Hence, the balance equations for the two resource pools are derived in (12)-(14) correspondingly. The only difference is that the arrival rate for the RP 1 equals $\zeta\lambda$ and the arrival rate for the RP 2 is $(1-\zeta)\lambda$, while the distribution of the required number of resource units remains the same and is given by $\{p_r\}$, $r\geq 0$.

The arriving session drop probability may thus be calculated as

$$p_B=\zeta p_{B,1}+(1-\zeta)p_{B,2},\tag{21}$$

while the ongoing session drop probability equals

$$p_T=\zeta p_{T,1}+(1-\zeta)p_{T,2}.\tag{22}$$

2) JOINT PROPORTIONAL ADMISSION STRATEGY

For the proportional admission strategy, the two resource pools may not be modeled as independent service entities. In this case, the required resource request distribution is represented by a *two-dimensional* probability mass function,

$$\hat{p}_{r,j}=p_k,\quad r=\lfloor\epsilon k\rfloor,\quad j=k-r,\quad k\geq 0,\tag{23}$$

where $\lfloor\epsilon x\rfloor$ is a floor function.

The total number of servers in the system is $N=N_1+N_2$ and the resource vector is $\mathbf{C}=(C_1,C_2)$. The equilibrium equations (12)-(14) should thus be extended to the vector-type form. The arriving session drop probability, the ongoing session drop probability, and the system utilization coefficient are then evaluated similarly to (15), (16), and (17), respectively.

D. SOLUTION ALGORITHM

To obtain the stationary probabilities, one needs to consider a system of linear equations (12)-(14) together with the normalization condition. The coefficients corresponding to $q_n(r)$ constitute an infinitesimal matrix denoted as $\mathbf{A}$. Therefore, the system (12)-(14) can be represented in the matrix form as

$$\mathbf{qA}=\mathbf{0},\tag{24}$$

where $\mathbf{q}$ is the vector of stationary probabilities $q_n(r)$, $(n,r)\in S_2$, and $\mathbf{0}$ is a vector of zeros. We may observe that $\mathbf{A}$ is the matrix of an irreducible aperiodic Markov chain and thus (24) has a unique solution [50]. Since any state from the subset $S_{2,n}$ has non-zero transition probabilities only to the states from the same subset $S_{2,n}$ and the adjacent subsets $S_{2,n-1}$ and $S_{2,n+1}$, $\mathbf{A}$ can be represented in a block tri-diagonal form. Hence, one may apply UL-decomposition [51] to simplify the solution of (24).

TABLE 2. Default numerical parameters.

Par.	Value	Description
N_1	60	Maximum number of sessions in RP 1
N_2	60	Maximum number of sessions in RP 2
R_1	160	Number of resource units per RP 1
R_2	130	Number of resource units per RP 2
C_0	50	Data rate per one resource unit, Mbps
λ	0.1	Arrival rate of sessions
μ^{-1}	420	Average session duration, s
γ	0.01	Arrival rate of triggers
ζ_O	0.55	Probability to be routed to RP 1 (opt.)
ϵ_O	0.55	Share requested from RP 1 (opt.)

V. SELECTED NUMERICAL RESULTS

In this section, we apply our developed mathematical framework to the representative crowded setup and assess the performance of the considered offloading strategies by exploring (i) the arriving session drop probability, p_T, (ii) the ongoing session drop probability, p_T, and (iii) the system resource utilization, δ. We begin by defining the scenario of interest as well as its important technical parameters. The variable parameters ζ and ϵ of the introduced resource allocation strategies are optimized in what follows, where we also compare our analytical results to those produced by system-level simulations. We also illustrate the implications of preferring one resource allocation strategy over another.

A. SCENARIO OF INTEREST

We model a futuristic scenario where spectators in a large and crowded stadium engage in X-VR applications through their head-mounted displays while watching the game. User terminals are served by beyond-5G wireless access points (B5G-APs) equipped with IEEE 802.11ad[4] and 5G NR chipsets,[5] which are able to operate concurrently in 28 and 60 GHz, respectively.

We consider 5G mmWave cellular to be the primary radio access technology that operates at 28 GHz and exploits 1 GHz of bandwidth with the maximum spectral efficiency of 8 bps/Hz. Hence, the overall cell capacity constitutes approximately 8 Gbit/s. As the secondary radio access technology, we adopt IEEE 802.11ad at 60 GHz and assume 2.16 GHz of channel bandwidth. Estimating the IEEE 802.11ad spectral efficiency to be around 3 bps/Hz, we arrive at the theoretic cell capacity of about 6.5 Gbit/s [52].

The users are assumed to initiate new sessions with the rate of $\lambda = 0.1\,\mathrm{s}^{-1}$, while the trigger arrival rate is set to $\gamma = 0.01\,\mathrm{s}^{-1}$. The probability mass function of the number of requested resource units is given by a geometric distribution $p_r = (1-p)p^{r-1}, r \geq 1$, with the mean value $\frac{1}{1-p}$. The average session data is calculated as $E[R] = C_0\frac{1}{1-p}$, where C_0 is the average data rate achieved by using one resource unit. By default, it is assumed that rate $C_0 = 50$ Mbps and $p = 0.2$. Other numerical parameters are summarized in Table 2.

For tractability, we model an idealistic cell operation with no extra costs due to dynamic resource reallocation between the active sessions. In addition, since the use of directional antennas at mmWave frequencies allows for reducing the levels of interference between the neighboring access points [53], we assume no harmful inter-cell interference. We compare the analytical results with those produced by our event-driven simulation tool written in Java, which specifically captures: (i) the arrivals of new sessions; (ii) the departures of sessions after a successful service; and (iii) the trigger events.

As sessions are arriving into an empty system, each of the simulation rounds (replications) models the arrivals of 5,000 sessions. Our system reaches its stationary state after approximately 2,000 arrivals. The statistical data are not collected during this "warm-up" period to avoid any bias in the output results. The interval between the 2,000-th and the 5,000-th arrival is then used to collect the statistics in each round, which is later averaged across multiple rounds to eliminate the residual dependence by ensuring the accuracy of the output results.

B. OPTIMIZING RESOURCE ALLOCATION STRATEGIES

We begin by optimizing the parameter ζ of the *joint probabilistic strategy*, which determines the optimal value of the probability to route an arriving session to the first resource pool. To this aim, Fig. 4 illustrates the arriving session drop probability, p_B, the ongoing session drop probability, p_T, and the system resource utilization coefficient, δ; all as functions of ζ. As one may observe in Fig. 4(a) and Fig. 4(b), both drop probabilities reach the minimum point at around $\zeta = 0.55$. At the same time, the resource utilization coefficient shown in Fig. 4(c) has the highest value of about $0.4 - 0.7$ (depends on μ and $E[R]$), also at $\zeta \approx 0.55$.

Therefore, we may conclude that 0.55 is the optimal value of ζ for the given set of input parameters, which is in line with an intuitive assumption that the optimal ζ_O is proportional to the total amount of resources in the first pool i.e., $\zeta_O = C_1/(C_1 + C_2)$. We now proceed with the analysis of the *joint proportional strategy* and optimize the share of the resources requested from the first resource pool, ϵ. Accordingly, Fig. 5 presents the same metrics of interest – p_B, p_T, and δ – as functions of ϵ.

We observe that the optimal value of ϵ, ϵ_O, remains the same across all of the considered performance metrics and can be evaluated as $C_1/(C_1 + C_2)$ (0.55 in our scenario). Comparing the analytical results in Fig. 4 and Fig. 5 with the simulation data, we may confirm the accuracy of our approximation, which also verifies our system modeling assumptions. A similar match maintains for the baseline strategy as well as across other sets of input parameters. Building on

[4]RF modules are available on the market today.

[5]First mmWave QTM052 antenna module is announced by Qualcomm in 2018; first commercial device is expected by the end of 2019.

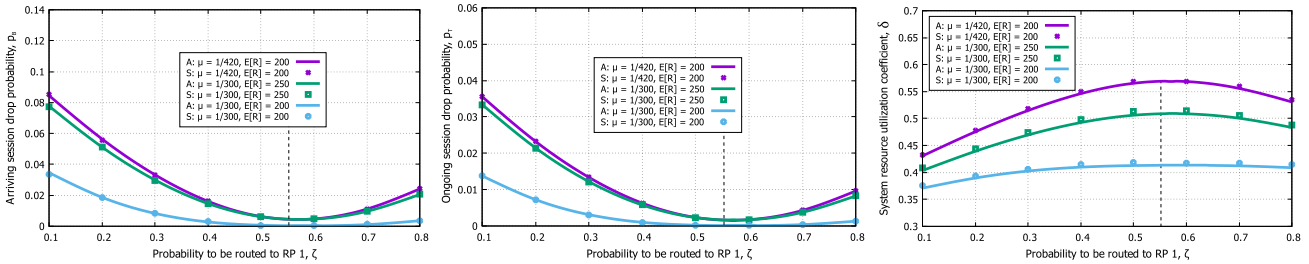


FIGURE 4. Joint Probabilistic Strategy: optimizing probability of an arriving session to be routed to the first resource pool, ζ. (a) Arriving session drop probability p_B vs. ζ. (b) Ongoing session drop probability p_T vs. ζ. (c) Resource utilization coefficient δ vs. ζ.

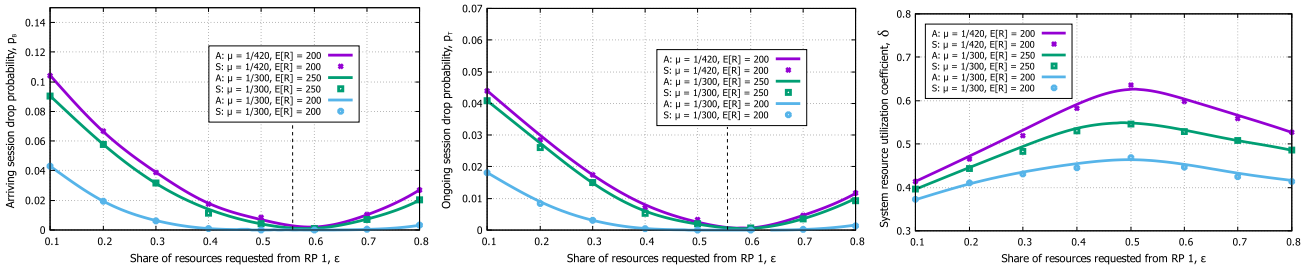


FIGURE 5. Joint Proportional Strategy: optimizing share of resources requested from the first resource pool, ϵ. (a) Arriving session drop probability p_B vs. ϵ. (b) Ongoing session drop probability p_T vs. ϵ. (c) Resource utilization coefficient δ vs. ϵ.

this observation, the following subsections only refer to the analytical results.

C. COMPARISON OF RESOURCE ALLOCATION STRATEGIES

In this subsection, we compare the performance of the three considered offloading strategies. For the *joint probabilistic* and *joint proportional* strategies, we adopt the optimal values of ζ and ϵ as derived in our previous analysis i.e., $\zeta_O = \epsilon_O = 0.55$.

1) EFFECT OF AVERAGE SESSION DURATION

Here, we study the impact of the average session duration, $1/\mu$, shown in Fig. 6 and Fig. 7. As follows from the ongoing session drop probability analysis in Fig. 6(a), in case of shorter sessions (under approximately 170 s), all three strategies handle the incoming traffic effectively and avoid any notable session drops. After the point of 170 s, the probability p_B increases for the baseline strategy, while the other two strategies demonstrate relatively low values up until approximately 250 s.

Meanwhile, according to Fig. 6(a), the value of p_B for these strategies grows faster and exceeds the corresponding value for the baseline strategy already after 500 s of the average session duration. Consequently, we conclude that more advanced strategies operate better under low and medium loads, while the baseline strategy remains preferable at higher loads. Continuing with the analysis of the two drop probabilities, we focus on Fig. 6(b) that presents the ongoing session drop probability, p_T, as a function of the session duration $1/\mu$.

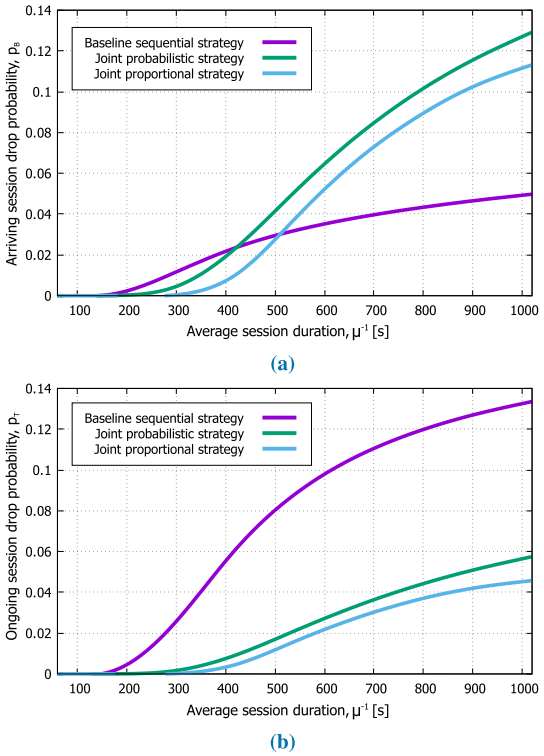


FIGURE 6. Comparing user-centric metrics for varying μ^{-1}. (a) Arriving session drop probability p_B vs. μ^{-1}. (b) Ongoing session drop probability p_T vs. μ^{-1}.

Here, we may notice that the baseline strategy demonstrates worse results across the entire range of input parameters. Combining the trade-offs from Fig. 6(a) and Fig. 6(b),

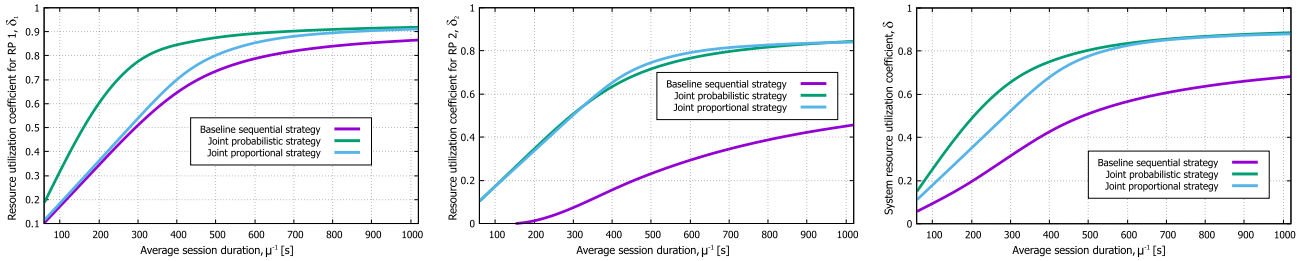


FIGURE 7. Comparing network-centric metrics for varying μ^{-1}. (a) Resource utilization coefficient for RP 1 δ_1. (b) Resource utilization coefficient for RP 2 δ_2. (c) Resource utilization for the entire system δ.

we conclude that the *joint proportional strategy* is the most desirable with respect to the user-centric metrics. We also note that despite the fact that the baseline strategy demonstrates slightly lower chances to drop an arriving session, this positive effect is outweighed by the considerably higher values of the ongoing session drop probability.

We finally study the network-centric performance in terms of the resource utilization coefficients for the first resource pool, the second resource pool, and the entire system: δ_1, δ_2, and δ, respectively (see Fig. 7(a), Fig. 7(b), and Fig. 7(c)). We observe that the *joint probabilistic strategy* yields the most efficient utilization of the network resources, while the *joint proportional strategy* results in slightly lower utilization values: the relative difference is between 5% and 40% depending on the average session duration.

2) EFFECT OF AVERAGE SESSION DATA RATE

We continue by studying the impact of the average session data rate, $E[R]$, on both user- and network-centric performance. To this end, Fig. 8 and Fig. 9 illustrate the dependencies similar to those discussed previously but constructed for varying $E[R]$. First, we analyze the user-centric metrics of interest, p_B and p_T, outlined in Fig. 8(a) and Fig. 8(b), respectively.

Similarly to the effects noted in Fig. 6, we observe that the use of the baseline strategy leads to slightly lower p_B values in high-load regimes as well as considerably greater p_T over the entire range of the considered $E[R]$ values. Therefore, we conclude that the *joint proportional strategy* results in the preferred user-centric performance across a wide range of input parameters. Finally, we study the network-centric performance indicators, namely, δ_1, δ_2, and δ as functions of $E[R]$ (illustrated in Fig. 9).

Here, the baseline strategy heavily overloads the first resource pool, while keeping the second one severely underloaded: e.g., for $E[R] = 250$, δ_1 exceeds 0.8, whereas $\delta_2 \approx 0.2$. This imbalance yields the worst performance of the baseline strategy in terms of the overall resource utilization coefficient δ in the considered range of $E[R]$, as confirmed by Fig. 9(c). The highest resource utilization is achieved by the *joint proportional strategy*. We may conclude that the *joint proportional strategy* achieves the best performance almost

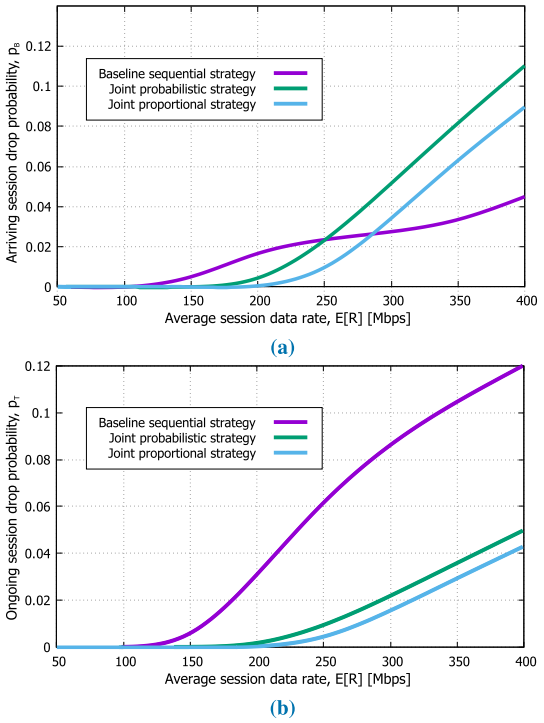


FIGURE 8. Comparing user-centric metrics for varying $E[R]$. (a) Arriving session drop probability p_B vs. $E[R]$. (b) Ongoing session drop probability p_T vs. $E[R]$.

in all cases, while the other two outperform it only over the very narrow ranges of μ and $E[R]$.

3) EFFECT OF mmWAVE CHANNEL DYNAMICS

We finally investigate the impact of dynamics brought by the mmWave wireless channel with respect to the selected metrics of interest. As the time instants when mmWave channel quality fluctuates can be approximated with a memoryless process [54], we study the said impact by varying the rate of trigger events, δ. Fig. 10a illustrates the arrival session drop probability as a function of δ, and we make three important qualitative conclusions. First, we notice that p_B changes significantly under the growth of δ: from $p_B \approx 0.03$ for $\delta = 0.004$ to $p_B \approx 0.02$ for $\delta = 0.02$ for the baseline strategy. This observation accentuates the importance of accounting

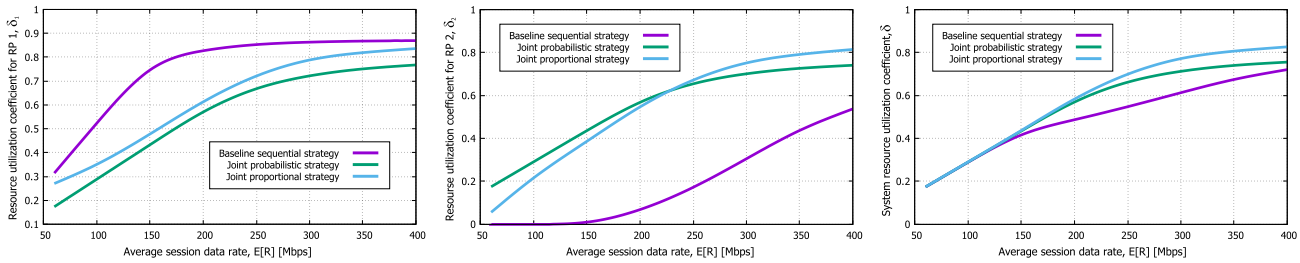


FIGURE 9. Comparing network-centric metrics for varying $E[R]$. (a) Resource utilization coefficient for RP 1 δ_1. (b) Resource utilization coefficient for PR 2 δ_2. (c) Resource utilization for the entire system δ.

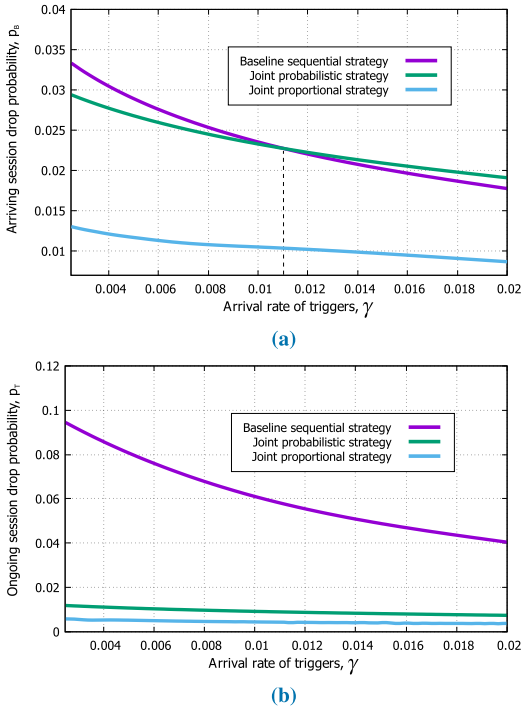


FIGURE 10. Comparing user-centric metrics of interest for γ. (a) Arriving session drop probability p_B vs. γ. (b) Ongoing session drop probability p_T vs. γ.

for the mmWave channel dynamics when modeling such a system.

Second, the *joint probabilistic strategy* slightly outperforms the baseline case starting from $\delta = 0.01$, thus highlighting the fact that the former is preferable in more dynamic channel conditions: $p_B \approx 0.017$ for the *joint probabilistic strategy* vs. $p_B \approx 0.019$ for the baseline case at $\delta = 0.02$. This fact is mainly explained by a more efficient utilization of the system resources with the *joint probabilistic strategy* in dynamic conditions. Finally, the *joint proportional strategy* performs considerably better than the other two counterparts across the entire range of δ, which confirms our previous consideration that this option should be preferred for future system design. The corresponding dependencies for the ongoing session drop probability, p_T as a function of δ outlined in Fig. 10b support this observation as well.

VI. CONCLUSIONS

Harnessing multiple mmWave-based radio access technologies is the next step towards terabit-per-second wireless systems [55]. Our work makes a decisive contribution in this direction by developing a novel mathematical framework that is capable of modeling an integrated extremely high frequency system, which aggregates licensed and unlicensed mmWave radio access technologies. The developed queuing-theoretic formulation makes it possible to evaluate both user- and network-centric performance metrics as well as compare potential radio resource allocation strategies that employ the two radio access technologies, which helps improve system operation and service reliability.

With our numerical study, we demonstrate that the *joint proportional strategy* – based on a proportional splitting of traffic between the two mmWave radio technologies – leads to better performance (both user- and network-centric) as compared to other considered solutions. At the same time, implementation of this attractive strategy in practical systems calls for further research into combining data across multiple streams handled by different access technologies. Therefore, simpler strategies may serve as feasible alternatives for the initial deployments of the envisaged multi-radio mmWave systems. We also believe that the contributed mathematical framework can be employed as an important building block in future system development and service optimization across many usage scenarios.

REFERENCES

[1] H. Bridgland. (2017). *CCS Insight Predicts 1 Billion Users of 5G by 2023, With More Than Half in China*. Accessed: Dec. 2018. [Online]. Available: https://www.ccsinsight.com/press/company-news/3240-ccs-insight-predicts-1-billion-users-of-5g-by-2023-with-more-than-half-in-china

[2] Y. Ohta and H. Tamura, *Mixed Reality: Merging Real and Virtual Worlds*. Berlin, Germany: Springer, 2014.

[3] S. S. Kim, M. Dohler, and P. Dasgupta, "The Internet of skills: Use of fifth-generation telecommunications, haptics and artificial intelligence in robotic surgery," *BJU Int.*, vol. 122, no. 3, pp. 356–358, 2018.

[4] M. Xiao *et al.*, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.

[5] Qualcomm Technologies. (2016). *Making 5G NR a Reality: Leading the Technology Innovations for a Unified, More Capable 5G Air Interface*. Accessed: Dec. 2018. [Online]. Available: https://www.qualcomm.com/media/documents/files/making-5g-nr-a-reality.pdf

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

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

[8] Y. Ghasempour, C. R. C. M. da Silva, C. Cordeiro, and E. W. Knightly, "IEEE 802.11ay: Next-generation 60 GHz communication for 100 Gb/s Wi-Fi," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 186–192, Dec. 2017.

[9] P. Zhou *et al.*, "IEEE 802.11ay-based mmWave WLANs: Design challenges and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1654–1681, 3rd Quart., 2018.

[10] M. Cheffena, "Industrial wireless communications over the millimeter wave spectrum: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 66–72, Sep. 2016.

[11] X. Lu, M. Lema, T. Mahmoodi, and M. Dohler, "Downlink data rate analysis of 5G-U (5G on unlicensed band): Coexistence for 3GPP 5G and IEEE802.11ad WiGig," in *Proc. 23th Eur. Wireless Conf.*, May 2017, pp. 1–6.

[12] J. Choi, V. Va, N. G.-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, Jr., "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 160–167, Dec. 2016.

[13] Z. Xiao, P. Xia, and X.-G. Xia, "Enabling UAV cellular with millimeter-wave communication: Potentials and approaches," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 66–73, May 2016.

[14] V. Petrov, D. Moltchanov, and Y. Koucheryavy, "Applicability assessment of terahertz information showers for next-generation wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–7.

[15] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, 1st Quart., 2014.

[16] F. Ghavimi and H.-H. Chen, "M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 525–549, 2nd Quart., 2015.

[17] Y. Mekonnen, M. Haque, I. Parvez, A. Moghadasi, and A. Sarwat, "LTE and Wi-Fi coexistence in unlicensed spectrum with application to smart grid: A review," in *Proc. IEEE/PES Transmiss. Distrib. Conf. Expo. (T&D)*, Apr. 2018, pp. 1–5.

[18] A. Al-Dulaimi, S. Al-Rubaye, Q. Ni, and E. Sousa, "5G communications race: Pursuit of more capacity triggers LTE in unlicensed band," *IEEE Veh. Technol. Mag.*, vol. 10, no. 1, pp. 43–51, Mar. 2015.

[19] R. Trivisonno, M. Condoluci, X. An, and T. Mahmoodi, "mIoT slice for 5G systems: Design and performance evaluation," *Sensors*, vol. 18, p. 635, Feb. 2018.

[20] V. Petrov *et al.*, "When IoT keeps people in the loop: A path towards a new global utility," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 114–121, Jan. 2019.

[21] V. Petrov *et al.*, "Achieving end-to-end reliability of mission-critical traffic in softwarized 5G networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 485–501, Mar. 2018.

[22] V. Chang, "An overview, examples, and impacts offered by emerging services and analytics in cloud computing virtual reality," *Neural Comput. Appl.*, vol. 29, no. 5, pp. 1243–1256, 2018.

[23] S. F. Kuliga, T. Thrash, R. C. Dalton, and C. Hölscher, "Virtual reality as an empirical research tool—Exploring user experience in a real building and a corresponding virtual model," *Comput., Environ. Urban Syst.*, vol. 54, pp. 363–375, Nov. 2015.

[24] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[25] P. Dempsey, "VR in. . . sport," *Eng. Technol.*, vol. 11, no. 3, p. 51, 2016.

[26] X. Hou, Y. Lu, and S. Dey, "Wireless VR/AR with edge/cloud computing," in *Proc. 26th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2017, pp. 1–8.

[27] J. Xiang, Y. Zhang, T. Skeie, and L. Xie, "Downlink spectrum sharing for cognitive radio femtocell networks," *IEEE Syst. J.*, vol. 4, no. 4, pp. 524–534, Dec. 2010.

[28] M. Matinmikko, H. Okkonen, M. Palola, S. Yrjola, P. Ahokangas, and M. Mustonen, "Spectrum sharing using licensed shared access: The concept and its workflow for LTE-advanced networks," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 72–79, Apr. 2014.

[29] A. Mukherjee *et al.*, "System architecture and coexistence evaluation of licensed-assisted access LTE with IEEE 802.11," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 2350–2355.

[30] S. Dama, A. Kumar, and K. Kuchi, "Performance evaluation of LAA-LBT based LTE and WLAN's co-existence in unlicensed spectrum," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–6.

[31] H. Lee, H. Kim, H. J. Yang, J. T. Kim, and S. Baek, "Performance analysis of license assisted access LTE with asymmetric hidden terminals," *IEEE Trans. Mobile Comput.*, vol. 17, no. 9, pp. 2141–2154, Sep. 2018.

[32] Y. Gao, Q. Huang, S. Xu, H. Li, Z. Li, and W. Tang, "Experimental performance evaluation and analysis of LAA and Wi-Fi coexistence in the unlicensed spectrum," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[33] D. López-Pérez *et al.*, "Long term evolution-wireless local area network aggregation flow control," *IEEE Access*, vol. 4, pp. 9860–9869, 2016.

[34] P. Fotiadis, P. Soldati, and P. Legg, "Design and evaluation of LTE/WLAN traffic steering and link aggregation algorithms," in *Proc. IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2016, pp. 1–6.

[35] N. Zhang, S. Zhang, S. Wu, J. Ren, J. W. Mark, and X. Shen, "Beyond coexistence: Traffic steering in LTE networks with unlicensed bands," *IEEE Wireless Commun.*, vol. 23, no. 6, pp. 40–46, Dec. 2016.

[36] *Revised SID: Study on NR-Based Access to Unlicensed Spectrum*, document 3GPP TSG RAN RP-172021, Sep. 2017.

[37] *Potential Solutions and Techniques for NR Unlicensed*, document 3GPP TSG RAN, R1-1802865, Mar. 2018.

[38] *Study on NR-Based Access to Unlicensed Spectrum*, document 3GPP TSG RAN Meeting 75, RP-170828, Qualcomm CDMA Technologies, 2017.

[39] E. Semaan, J. Ansari, G. Li, E. Tejedor, and H. Wiemann, "An outlook on the unlicensed operation aspects of NR," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[40] S. Lagen and L. Giupponi, "Listen before receive for coexistence in unlicensed mmWave bands," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–6.

[41] Qualcomm Technologies. (2017). *What Can We do With 5G NR Spectrum Sharing that is not Possible Today?* Accessed: Dec. 2018. [Online]. Available: https://www.qualcomm.com/media/documents/files/new-3gpp-effort-on-nr-in-unlicensed-spectrum-expands-5g-to-new-areas.pdf

[42] K. Venugopal, M. C. Valenti, and R. W. Heath, Jr., "Device-to-device millimeter wave communications: Interference, coverage, rate, and finite topologies," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6175–6188, Sep. 2016.

[43] M. Gapeyenko *et al.*, "On the degree of multi-connectivity in 5G millimeter-wave cellular urban deployments," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1973–1978, Feb. 2018.

[44] V. Petrov *et al.*, "Dynamic multi-connectivity performance in ultra-dense urban mmWave deployments," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2038–2055, Sep. 2017.

[45] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 102–108, Jun. 2017.

[46] S. Al-Rubaye, A. Al-Dulaimi, J. Cosmas, and A. Anpalagan, "Call admission control for non-standalone 5G ultra-dense networks," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 1058–1061, May 2018.

[47] J. Kua, G. Armitage, and P. Branch, "A survey of rate adaptation techniques for dynamic adaptive streaming over HTTP," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1842–1866, 3rd Quart., 2017.

[48] V. Naumov, K. Samouylov, E. Sopin, and S. Andreev, "Two approaches to analyzing dynamic cellular networks with limited resources," in *Proc. 6th Int. Congr. Ultra Mod. Telecommun. Control Syst. Workshops (ICUMT)*, 2014, pp. 485–488.

[49] V. A. Naumov, K. E. Samuilov, and A. K. Samuilov, "On the total amount of resources occupied by serviced customers," *Autom. Remote Control*, vol. 77, no. 8, pp. 1419–1427, 2016.

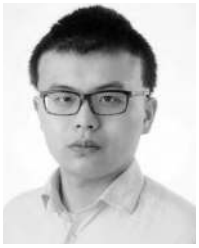
[50] J. G. Kemeny and J. L. Snell, *Finite Markov Chains: With a New Appendix 'Generalization of a Fundamental Matrix'*. New York, NY, USA: Springer, 1983.

[51] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, vol. 71. Philadelphia, PA, USA: SIAM, 2000.

[52] *IEEE Draft Standard for Information Technology–Telecommunications and information exchange between systems local and metropolitan area networks–Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Standard P802.11bb, 2017.

[53] V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference analysis of EHF/THF communications systems with blocking and directional antennas," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–7.

[54] M. Gapeyenko *et al.*, "On the temporal effects of mobile blockers in urban Millimeter-wave cellular scenarios," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10124–10138, Nov. 2017.
[55] F. Khan, "Multi-comm-core architecture for terabit-per-second wireless," *IEEE Commun. Mag.*, vol. 54, no. 4, pp. 124–129, Apr. 2016.



XI LU received the B.Eng. degree in information engineering from the Wuhan University of Technology, Wuhan, China, in 2013, and the M.Sc. degree in telecommunications and internet technology from the King's College London, London, U.K., in 2015, where he is currently pursuing the Ph.D. degree in telecommunications research with the Centre for Telecommunications Research. He was a Trainee with China Telecom, Wuhan, in 2012, and a Visiting Scholar with the Laboratory of Electronics and Communications Engineering, Tampere University, Tampere, Finland. His current research interests include heterogeneous 5G (and beyond) networks, unlicensed spectrum sharing, mmWave communications, the Internet of Things, and resource sharing.

EDUARD SOPIN received the B.Sc. and M.Sc. degrees in applied mathematics from the Peoples' Friendship University of Russia (RUDN University), in 2008 and 2010, respectively, and the Ph.D. degree in applied mathematics and computer science, in 2013. Since 2009, he has been with the Telecommunication Systems Department, RUDN University, where he is currently an Associate Professor with the Department of Applied Probability and Informatics. His current research interests include the performance analysis of modern wireless networks and cloud/fog computing.

VITALY PETROV received the Specialist degree from SUAI University, St. Petersburg, Russia, in 2011, and the M.Sc. degree from the Tampere University of Technology, Finland, in 2014, where he is currently pursuing the Ph.D. degree with the Laboratory of Electronics and Communications Engineering. He has co-authored more than 30 published research works on terahertz band/mmWave communications, the Internet-of-Things, nanonetworks, cryptology, and network security. He was a recipient of the Best Student Paper Award at the IEEE VTC-Fall, Boston, USA, in 2015, and the Best Student Poster Award at the IEEE WCNC, San Francisco, USA, in 2017.

OLGA GALININA received the B.Sc. and M.Sc. degrees in applied mathematics from the Department of Applied Mathematics, Faculty of Mechanics and Physics, St. Petersburg State Polytechnical University, Russia, and the Ph.D. degree from the Tampere University of Technology, Finland, where she is currently a Research Scientist with the Laboratory of Electronics and Communications Engineering. Her research interests include applied mathematics and statistics, queueing theory and its applications, wireless networking and energy efficient systems, and machine-to-machine and device-to-device communication.

DMITRI MOLTCHANOV received the M.Sc. and Cand.Sc. degrees from the St. Petersburg State University of Telecommunications, Russia, in 2000 and 2002, respectively, and the Ph.D. degree from the Tampere University of Technology, Finland, in 2006, where he is currently a Senior Research Scientist with the Laboratory of Electronics and Communications Engineering. He has authored over 80 publications. His research interests include performance evaluation and optimization issues of wired and wireless IP networks, the Internet traffic dynamics, quality of user experience of real-time applications, and traffic localization P2P networks. He serves as a TPC member in a number of international conferences.

KIRILL AGEEV received the M.Sc. degree in computer science from the Peoples' Friendship University of Russia (RUDN University), in 2017, where he is currently pursuing the Ph.D. degree with the Applied Probability and Informatics Department. His current research interests include teletraffic theory, queueing systems, and advanced telecommunications.

SERGEY ANDREEV received the Specialist and Cand.Sc. degrees from the Saint Petersburg State University of Aerospace Instrumentation, St. Petersburg, Russia, in 2006 and 2009, respectively, and the Ph.D. degree from the Tampere University of Technology, Finland, in 2012, where he is currently an Assistant Professor with the Laboratory of Electronics and Communications Engineering. He has co-authored over 150 published research works on wireless communications, energy efficiency, heterogeneous networking, cooperative communications, and machine-to-machine applications.

YEVGENI KOUCHERYAVY received the Ph.D. degree from the Tampere University of Technology, Finland, in 2004, where he is currently a Full Professor with the Laboratory of Electronics and Communications Engineering. He has authored numerous publications in the field of advanced wired and wireless networking and communications. His current research interests include various aspects in heterogeneous wireless communication networks and systems, the Internet of Things and its standardization, and nanocommunications. He is an Associate Technical Editor of the *IEEE Communications Magazine* and an Editor of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.

KONSTANTIN SAMOUYLOV received the Ph.D. degree from Moscow State University and the D.Sc. degree from the Moscow Technical University of Communications and Informatics. From 1985 to 1996, he held several positions with the Faculty of Sciences, Peoples' Friendship University of Russia (RUDN University), where he became the Head of the Telecommunication Systems Department, in 1996. Since 2014, he has been the Head of the Department of Applied Informatics and Probability Theory, RUDN University. During last two decades, he has been conducting research projects for the Helsinki University of Technology and Lappeenranta University of Technology, Moscow Central Science Research Telecommunication Institute, several Institutes of Russian Academy of Sciences, and a number of Russian network operators. He has authored over 150 scientific and technical papers and three books. His current research interests include performance analysis of 4G networks (LTE and WiMAX), teletraffic of triple play networks, signaling network planning, and cloud computing.

MISCHA DOHLER is currently a Full Professor of wireless communications with the King's College London, London, U.K., and a Co-Founder of the pioneering smart city company Worldsensing. He is a Fellow of the Royal Society of Arts and the Distinguished Member of Harvard Square Leaders Excellence. He is a Frequent Keynote, Panel, and Tutorial Speaker. He has pioneered several research fields, contributed to numerous wireless broadbands, the IoT/M2M, and cyber security standards, holds a dozen patents, organized, and chaired numerous conferences, has over 200 publications, and authored several books. He acts as a Policy, Technology, and Entrepreneurship Adviser. He is also an Entrepreneur, a Composer, and a Pianist with four albums on iTunes, and fluent in six languages. He has talked at TEDx. He had coverage by the national and international TV and radio, and his contributions have featured on the BBC and *The Wall Street Journal*.