Achieving End-to-End Reliability of Mission-Critical Traffic in Softwarized 5G Networks

Vitaly Petrov, Maria A. Lema, Margarita Gapeyenko, Konstantinos Antonakoglou, Dmitri Moltchanov, Fragkiskos Sardis, Andrey Samuylov, Sergey Andreev, *Senior Member, IEEE*, Yevgeni Koucheryavy, *Senior Member, IEEE*, and Mischa Dohler, *Fellow, IEEE*

***Abstract*— Network softwarization is a major paradigm shift, which enables programmable and flexible system operation in challenging use cases. In the fifth-generation (5G) mobile networks, the more advanced scenarios envision transfer of high-rate mission-critical traffic. Achieving end-to-end reliability of these stringent sessions requires support from multiple radio access technologies and calls for dynamic orchestration of resources across both radio access and core network segments. Emerging 5G systems can already offer network slicing, multi-connectivity, and end-to-end quality provisioning mechanisms for critical data transfers within a single software-controlled network. Whereas these individual enablers are already in active development, a holistic perspective on how to construct a unified, service-ready system as well as understand the implications of critical traffic on serving other user sessions is not yet available. Against this background, this paper first introduces a softwarized 5G architecture for end-to-end reliability of the mission-critical traffic. Then, a mathematical framework is contributed to model the process of critical session transfers in a softwarized 5G access network, and the corresponding impact on other user sessions is quantified. Finally, a prototype hardware implementation is completed to investigate the practical effects of supporting mission-critical data in a softwarized 5G core network, as well as substantiate the key system design choices.**

***Index Terms*— Software-defined networking, fifth-generation (5G) mobile systems, millimeter-wave communications, heterogeneous networks, 5G network function virtualization, mission-critical traffic.**

Manuscript received October 4, 2017; revised February 5, 2018; accepted February 27, 2018. Date of publication March 12, 2018; date of current version May 21, 2018. This work was supported in part by the 5G UK Testbeds and Trials program through the Department of Digital Culture Media and Sports and in part by the Ericsson 5G Tactile Internet Industry Grant to the King's College London, and in part by the Academy of Finland (Projects WiFiUS and PRISMA). The work of V. Petrov was supported in part by the Nokia Foundation and in part by the HPY Research Foundation through Elisa, and has been partly completed during the research visit to King's College London, U.K. *(Corresponding author: Vitaly Petrov.)*

V. Petrov, M. Gapeyenko, D. Moltchanov, A. Samuylov, and Y. Koucheryavy are with the Tampere University of Technology, 33720 Tampere, Finland. (email: vit.petrov@gmail.com).

M. A. Lema, K. Antonakoglou, F. Sardis, and M. Dohler are with King's College London, London WC2R 2LS, U.K.

S. Andreev is with the Tampere University of Technology, 33720 Tampere, Finland, and also with King's College London, London WC2R 2LS, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2018.2815419

I. Introduction

THE next generation of mobile networks brings along the unprecedented levels of heterogeneity, further aggravated by the adoption of the New Radio (NR) technology in the millimeter-wave (mmWave) spectrum [1]. The inherent intermittency of NR transmissions, as well as new types of highly-mobile users, such as connected vehicles and drones, pose significant challenges in terms of network management. They also require prompt decision-making, thus calling for novel means to enable efficient orchestration of network resources [2]. Since the applicability of existing solutions is limited to hard-coding all possible events and corresponding actions [3], the community demands increased degrees of freedom and improved intelligence for the underlying network elements [4], [5].

Known as *Network Softwarization*, this trend becomes a major new wave in communications engineering and is based on several paradigm shifts [6]. First, Software-Defined Networking (SDN) primarily focuses on decoupling the software-based control plane from the hardware-based data plane within the network infrastructure, thus allowing to adjust the behavior of network nodes in an automated and dynamic manner [7]. Intelligent implementation of the SDN concept leads to a number of benefits, which include more efficient resource allocation [8] and facilitated integration of multiple Radio Access Technologies (RATs) [9]. In addition, SDN offers flexibility in terms of traffic management and network control [10], [11].

The second enabler is Network Function Virtualization (NFV), which creates another level of abstraction by deploying network functions as software components, named virtual network functions (VNFs) [12]. This approach simplifies network control by providing a distinct separation between a certain functionality and its actual deployment [13]. The use of NFV enables lower-cost and seamless network upgrades. NFV also improves energy-efficiency, since a physical server that combines several virtual roles consumes less energy [14]. The third major shift is ubiquitous edge caching and computing, which aims to merge the communications infrastructure with storage and computing capabilities [4], [15]. Network softwarization develops synergy between these three planes across a number

of scenarios. This leads to emergence of fundamentally novel network architectures with enhanced flexibility [16].

Many discussions are ongoing on the desired degree of integration that software-driven communications will have with the 5G systems [10], and there are multiple research concerns to be resolved [17], [18]. The consensus is that network softwarization is identified as a key component to improve performance and reliability, as well as reduce expenditures and operating costs of next-generation wireless deployments [19]. Consequently, intelligent softwarized 5G networks are envisioned to become enablers for advanced network architectures [16] by supporting emerging applications, such as Tactile Internet, Internet of Skills, and many more [20].

A. Critical Traffic Management in 5G and Beyond

A key emerging use case that may benefit substantially from network softwarization is handling mission-critical traffic in 5G networks [21]. Ranging from control of Unmanned Aerial Vehicles (UAVs) to remote surgery, these applications generate traffic that is of critical nature and has to be supported by all means [22]. Considering the properties of these critical data, not only accurate and timely information delivery is necessary, but also other stringent key performance indicators (KPIs) need to be maintained, such as guaranteed bitrate, latency, and reliability [23]. Since static reservation of radio resources to accommodate the envisioned volumes of mission-critical traffic may lead to substantial over-provisioning [24], on-demand and instant resource allocation is desired [25].

Responding to these needs, new technologies that prioritize mission-critical traffic over regular (e.g., "best-effort") data are demanded on both radio access network (RAN) and core network (CN) levels. Particular enablers that are capable of handling mission-critical traffic include utilization of higher frequencies with the mmWave RAT, multi-RAT, and multi-connectivity schemes with dynamic fallback to a backup serving station or RAT, and dynamic radio resource re-allocation in both radio and core network [26]–[28]. Specifically, softwarization mechanisms to efficiently manage the system with the above features are in prompt need, which will give priority to mission-critical traffic at all times [8]. Here, the concept of network slicing supported by SDN and NFV technologies promises to achieve virtualization of the infrastructure components and thus help manage QoS as well as satisfy the target KPIs for the critical data [29]–[31].

At the same time, the cornerstone question is that of how the changes in the network behavior to support these high-rate and high-priority sessions will affect the performance of regular user data. More specifically, the concern is related to how much the end-to-end prioritization of mission-critical traffic in multi-RAT softwarized 5G networks will affect the QoS of other sessions, which may occupy the network resources at the moment when a mission-critical session arrives.

The research problem at hand has recently gained increased attention in the community. The NFV-based QoS provisioning for software-defined optical networks has been delivered by [27]. A constructive approach to achieve close to real-time guarantees by using NFV has been presented in [30]. The work in [32] discussed the growth of reliability with the use of multi-connectivity features within the 5G architecture, while a roadmap towards achieving more reliable critical communications has been outlined in [33]. While careful initial steps to address these crucial research questions have been made recently, we maintain that a holistic methodology in this important area has not been proposed as of yet.

B. Our Motivation and Contribution

Achieving end-to-end reliability for mission-critical communications via softwarization of the network components and their dynamic (re)configuration to guarantee the required QoS for the mission-critical sessions is an important research area. There are multiple recent contributions illustrating the *feasibility* of this feature, including the underlying architectures, algorithms, and protocols to enable it in 5G. At the same time, the question of *impact* of mission-critical traffic in softwarized 5G – particularly, its effects on the regular user traffic as well as the network operation at large, when such mission-critical sessions will massively emerge and become prioritized at all levels – has been insufficiently studied so far. The existing results on this topic are fragmented and a holistic evaluation methodology to address this problem both at the access and the core network levels has not been proposed yet.

Motivated by this gap, we contribute a novel integrated methodology to account for the implications of handling mission-critical traffic in softwarized 5G networks. We first outline the developments in 5G architecture to incorporate the benefits of network softwarization by focusing specifically on management of mission-critical sessions. We then develop a comprehensive mathematical methodology to model the softwarized 5G RAN that is handling mission-critical data from a moving high-priority user on top of the regular data transmissions from other users. We finally summarize a measurement-based campaign to assess the softwarized 5G CN that accommodates jointly the mission-critical and the best-effort data flows. The contributions of this work are thus summarized as follows:

- *Softwarized 5G architecture*: An NFV-based architecture is proposed where all 5G network functions run as pieces of software and the data plane is controlled by SDN-like features end-to-end. This architecture enables support of mission-critical services as well as ensures co-existence with other, less critical services, such as mobile broadband. To achieve this, two key and complementary technologies, namely, SDN and NFV, are employed. On the one hand, ETSI's NFV allows to deploy network services, such as 3GPP's 5G system-level architecture, as VNFs controlled by the centralized orchestrator. On the other hand, SDN provides the necessary QoS monitoring and path management capabilities.
- *Enabling mathematical methodology*: An elaborate mathematical methodology is developed that takes into account performance dynamics of multi-RAT softwarized 5G RAN together with mobility of mission-critical users in urban deployments, mindful of the unique features of the mmWave RAT. The 5G-centric technology enhancements to support session continuity of mission-critical transmissions are also modeled, including

multi-RAT, multi-connectivity, and dynamic on-demand radio resource re-allocation for mission-critical applications. The model is specifically tailored to the softwarized 5G architecture and targets to assess the RAN-level performance as well as quantify the effects of high-rate mission-critical sessions on the QoS of regular user traffic.

- *Advanced experimental study*: A detailed practical framework is constructed for the end-to-end performance evaluation of software-assisted 5G CN that handles a mixture of best-effort traffic from regular users and mission-critical traffic from the selected prioritized user. It incorporates a comprehensive prototype comprising major network elements based on the real-world hardware as part of the 5G test network deployment in the UK.[1] The field measurements are conducted to assess the CN behavior in the softwarized 5G system. The proposed approach targets to answer the questions related to the coexistence of stationary best-effort and spontaneous mission-critical traffic at the softwarized 5G network core level.

We detail our considered softwarized 5G architecture in the following section.

II. Envisaged Softwarized 3GPP 5G Architecture

A. Core Design Principles

The architecture outlined in this work considers 3GPP's 5G System Architecture described at length in [34], but introduces it as part of the global ETSI's NFV architecture for virtualization.[2] In particular, all network functions are considered to be pieces of software, which can run in standard hardware and may in principle be moved around across different locations subject to the requirements of the communications provider.

We also address the role of softwarization from the end-to-end perspective, including (i) the software-defined traffic steering decisions in the access part, (ii) the role of SDN in the core part, and (iii) how it should be interfaced with the 3GPP architecture. In general, software-defined rules have to be enforced in the physical infrastructure to satisfy a certain level of QoS, starting from accurate selection of access nodes and up to managing traffic in the transport network where the CN resides. In this work, we consider mission-critical services; hence, the use of softwarization in policy and QoS enforcement is crucial to ensure appropriate traffic isolation as well as correct delivery of user-plane data, which allows for prioritization in the transport network if required.

Network virtualization and softwarization are actively studied in the context of the H2020 EU projects, providing architectural solutions that can satisfy the requirements of different use cases (e.g., 5G Car); contributing to softwarization of radio access and core networks (e.g., 5G COHERENT); and focusing on the integration of network services in an orchestration platform (e.g., SONATA). While there are numerous existing approaches to construct a flexible 5G network [5], [9], [35], our outlined architecture combines the components, which are key in delivering QoS-enabled services in well-established solutions (such as 3GPP and ETSI NFV). We combine both RAN- and CN-related aspects in the context of QoS and policy management, with the objective of providing the overall end-to-end reliability. We also ensure consistency between the enforcement of policies across the entire chain of the provided network service.

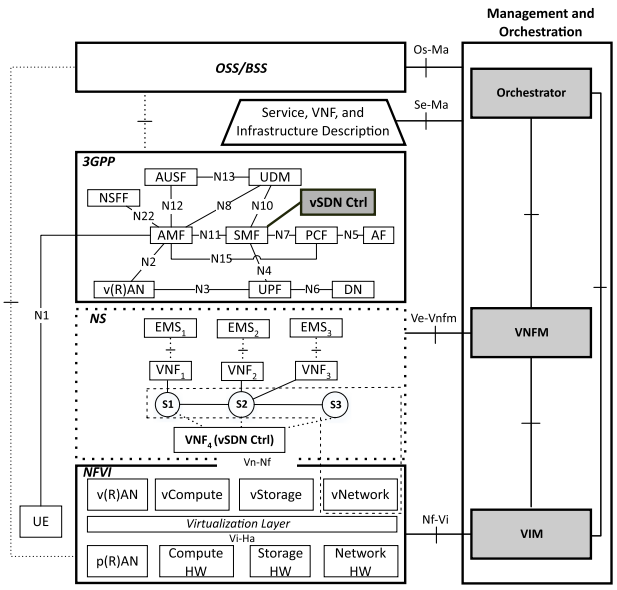

Fig. 1. Considered architecture of a softwarized 5G network.

B. NFV and SDN Architecture Description

In more detail, Fig. 1 outlines the standard ETSI NFV architecture, where the bottom layer comprises a set of virtualized resources running on top of the physical infrastructure. Network functions are then introduced as Virtual Machines (VMs) or Containers running inside the virtual infrastructure. These network functions can perform various operations, such as load balancing, firewalling, switching, 3GPP-defined network functions, as well as act as a virtualized SDN controller responsible for managing VNF instances of OpenFlow switches.

The envisaged system employs SDN virtualization as a means of offering control of the virtualized network infrastructure to network services, such as 3GPP architecture, on top of the physical network infrastructure, which is managed by an SDN controller. An example of SDN virtualization can be found in [35], where the authors present an architecture that allows virtualized SDN infrastructures to be controlled by virtualized SDN controllers. The latter reside on top of the physical SDN infrastructure, which is managed by an operator's SDN controller.

C. 3GPP System Architecture Description

A distinct feature of 5G system-level architecture defined by 3GPP is the modular principle in the network function design, which brings the needed flexibility to run network functions as well as enables the concept of network slicing.

[1]Jack Loughran, "UK government grants £16m for 5G test networks at three UK universities," *Engineering and Technology*, July 2017.

[2]ETSI, "Network Functions Virtualisation (NFV); Architectural Framework," ETSI GS NFV 002 V1.1.1, October 2013.

The modularity in the 5G architecture is designed to support deployments using both NFV and SDN technologies. In this work, we consider that all the network functions are in their turn running as VNFs within the virtualized infrastructure i.e., cloud infrastructure. While there are multiple ways to encapsulate VNFs, including containers [36] and unikernels [37], in this work we discuss VMs as an illustrative example, since paravirtualization is an established technology and OpenStack is widely used in telco provider clouds [38]. The main network functions as defined in 3GPP TR 23.501 [34] are:

- *User Plane Function (UPF)* is in charge of handling the user plane path of a data session and will provide an interface to the data network (outside the 3GPP domain).
- *Session Management Function (SMF)* is in charge of the establishment, modification, and release of the session. Amongst the multiple tasks of the SMF, some of the most relevant to our analysis are: policy control, QoS enforcement, data plane routing information to the UPF.
- *Access and Mobility Management Function (AMF)* controls the access decisions as well as handles all related mobility procedures. Based on multiple inputs and metrics, software-defined rule enforcement allows the AMF to decide on the best access point for the user.
- *Policy Control Function (PCF)* is in charge of providing the policy rules to the relevant control plane functions and supports a unified policy framework across the network.

The methodology presented in this paper considers the 5G system-level architecture as its baseline, where distributed software-defined decisions are made independently for both the access and the core network. The end-to-end QoS/Policy framework is thus a combination of distributed functions that monitor network performance and enforce QoS policies based on the session's context-related information as well as a centralized PCF that ensures an efficient coordination and alignment among all these distributed functions.

Within the 3GPP community, there has been a lot of discussion on how to ensure adequate alignment between the QoS enforcement rules in the User Plane network functions (such as interactions between the PCF, SMF, and AMF) and the QoS enforcement rules in the transport network. To this end, we propose a mechanism of the interface from the SMF to the virtual SDN controller, which will allow the 3GPP system to control the forwarding rules in the network infrastructure and thus satisfy the service requirements. In this sense, once the UE/eNB initiates a service with a particular QoS requirement, the SMF will be in charge of enforcing those rules in the physical network as well as ensuring that there is consistency between radio access, core, and transport networks.

D. SDN and Virtualization for End-to-End Service

In our outlined architecture, all 3GPP network functions are running as VNFs; hence, a virtualized SDN controller becomes an important network element that can be contacted by any of the 3GPP functions [39]. More specifically, the SMF can instruct the SDN controller to adjust flow rules within the virtualized 3GPP network for the QoS management purposes. This enables full control of the virtualized network resources by the 3GPP system as opposed to having a static configuration that is dictated by the NFV operator's SDN controller. The existing solutions developed in the context of intent APIs, such as the ONOS Intent Framework and the OpenDaylight Network Intent Composition, offer comfortable configurability of the network interfaces but, at the same time, are featured by the limited functionality, which is insufficient to support the intended compound solution. Therefore, our proposed architecture suggests that the physical SDN controller manages the physical infrastructure that connects the physical nodes of the Virtual Infrastructure Manager between them as well as to external networks. In its turn, the virtualized SDN controller manages virtualized switches created as VNFs to support the virtualized 3GPP functions deployed within the tenant's slice.

At the higher layer of the architecture, network service descriptors as per the NFV specifications are utilized to describe the network services and network functions, while the OSS/BSS is used by the NFV operator for management and billing purposes. On the right-hand side of Fig. 1, the MANO layer spans across all the layers of the NFV architecture and is responsible for managing the infrastructure, monitoring the status of the network functions, coordinating their life-cycle, and finally maintaining a catalog of descriptors for the network services and functions that can be deployed in the system.

To implement a 3GPP system, the NFV operator needs to provide VNF descriptors to the Orchestrator. These descriptors contain configuration and deployment information, such as the number of CPU cores, RAM, storage, and network interfaces for each VM that will host a VNF. The descriptors can include additional information that is specific to the VNF, such as 3GPP configuration options for a Packet Gateway (PGW). The Operator will need to on-board the VNFs to the VM by adding the images of these VNFs into the virtual storage pool. At this point, a Network Service descriptor can be used to create a complete network service by including all of its component VNFs and the virtual network infrastructure that will connect them together to deliver a functional 3GPP system.

In the next sections, we follow the major design choices of the outlined architecture and proceed with the mathematical analysis of the softwarized 5G operation in the RAN domain.

III. Mathematical Framework for Softwarized 5G RAN: System Model and Methodology Overview

In this section, we specify an illustrative use case for high-rate mission-critical traffic in a softwarized 5G network. We begin by outlining the corresponding system model, then detail our mathematical framework, and finally proceed with summarizing the proposed methodology.

A. Proposed System Model

We analyze a characteristic use case, where mission-critical traffic needs to be served by the operator's 5G network. Particularly, we focus on an ambulance vehicle that is transporting a patient to the hospital, while the paramedics in the vehicle and doctors in the hospital are jointly assessing the patient's condition. In this specific scenario, softwarized 5G infrastructure enables bulky data transfer (from cameras,

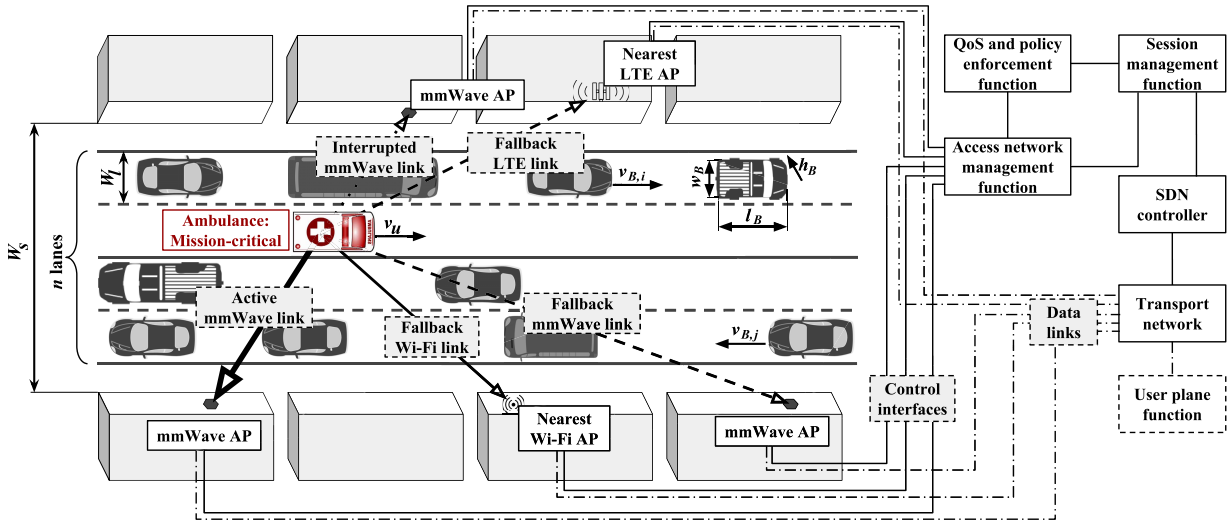

Fig. 2. Considered system model for the target use case of mission-critical traffic support in softwarized beyond-5G networks.

sensors, and robotic systems) to a remote clinician, who could then guide the paramedics on more advanced pre-hospital treatment [22]. Since the requirements for serving mission-critical traffic in terms of its data rate and transfer delay are more stringent than what microwave systems can typically support, the use of 5G mmWave cellular is considered here as a primary option for the ambulance vehicle, while other radio technologies are only left as a backup [40].

1) Environment and Target User: We assume an urban deployment typical for e.g., central London. Specifically, we model a straight segment of a street with left-side traffic, see Fig. 2. The total street width is W_S and there are n lanes in each direction, each having the width of W_L. The ambulance vehicle is driving on the right (the fastest) lane with the constant speed of v_U. Since the ambulance flashing lights are on, we assume an optimistic case where no other vehicles drive on the same lane. The height of the vehicle is h_U, while all the antennas are placed on its roof for better channel conditions.

2) Moving Vehicles on Side Lanes: Other moving vehicles are assumed to be deployed randomly in the street. The vehicles on the lane i are located according to a Poisson process with the intensity of λ_i. The length and height of other vehicles are l_B and h_B, respectively, having the probability density function (pdf) of $f_{l_B}(x)$ and $f_{h_B}(x)$. For the sake of analytical tractability, all the vehicles are assumed to have a constant width, w_B, and drive along the center of their lanes.

The inter-vehicle distance, d_I, is also random with the pdf of $f_{d_I}(x)$ and subject to the selected lane. The speed of the vehicles in the lane i, $v_{B,i}$, is random and follows the pdf of $f_{v_{B,i}}(x)$. For simplicity, no parked vehicles are considered.

3) Access Network and Propagation: The wireless network comprises three segments: (i) cellular mmWave network, (ii) cellular microwave network (LTE), and (iii) non-3GPP microwave network (Wi-Fi). The mmWave access points (APs) are assumed to be deployed on the building sides at a constant height of h_A. The APs are deployed regularly having the distance of d_A between each other, alternately on the left and the right sides of the street. For microwave access, we assume an extreme case where both LTE and Wi-Fi connectivity is always available. Particularly, the LTE network covers the entire city and there is always at least one Wi-Fi AP within the coverage area around the ambulance vehicle.

For mmWave, there are no buildings in the way from the ambulance to the AP in the modeled scenario; hence, the links are always in the line-of-sight (LoS) conditions. However, vehicles on the side lanes act as blockers. As a result, two conditions are distinguished by our model: non-blocked LoS and blocked LoS. Each of these is featured by its own path loss formula: $P_{nb}(x)$ and $P_b(x)$, where x is the link length. We follow the standard 3GPP mmWave propagation models when delivering our numerical study in Section VII. For LTE, we adopt [41] by assuming that SNR is maintained at a constant level via adequate power control. For Wi-Fi, we consider fixed-power transmissions, where the effective spectral efficiency is determined by true distance between the node and the AP [42].

4) Mission-Critical Traffic: The target data rate of the considered mission-critical traffic is constant and equals r_c. Aiming for higher reliability when transmitting critical data, the vehicle always keeps a number of simultaneous data sessions to the nearest mmWave APs, whereas their number is named the *degree of multi-connectivity*. Hence, the ambulance is in outage on the mmWave network only when it has outage on all the currently serving mmWave APs. In this case, the ongoing critical data transmission falls back from the mmWave to the microwave segment, where the choice of which RAT to use (LTE or Wi-Fi) is specified by the network configuration. The network is configured to always grant the requested data rate for the mission-critical traffic, even if it leads to deteriorating the QoS of the best-effort sessions for regular users.

5) Best-Effort Traffic: The best-effort traffic coming from the regular users is not expected to compete for radio resources with the mission-critical data in well-provisioned mmWave deployments due to more abundant spectral resources [43].

Therefore, we do not model the best-effort traffic in the mmWave network segment. In contrast, the target data rate of a mission-critical session may be comparable to the overall capacity of an individual LTE or Wi-Fi AP. To understand the effects of mission-critical traffic on the best-effort transmissions, we model random best-effort traffic according to a homogeneous Poisson process with the intensity of θ [44]. The duration of each session is exponential with the parameter μ. In case where there are insufficient free resources to handle the fallback of the mission-critical session, a certain number of best-effort sessions may be dropped. The resources utilized by the mission-critical data are released once the ambulance is connected back to the mmWave network.

Below, we introduce our mathematical framework to capture the core dynamics of the considered system.

B. Our Framework at a Glance

The mathematical framework developed in this paper can be decomposed into two phases. First, we formalize the mmWave LoS blockage process with multi-connectivity of degree M and mobility of both the ambulance as well as the vehicles in other lanes. The **first phase** is divided into three steps.

- *Step 1:* We specify the dynamics of the blockage process in the presence of mobility of vehicles in other lanes. We show that when the position of the ambulance is known, the blockage adheres to a stationary alternating renewal process.
- *Step 2:* We extend the LoS blockage model to the case of multi-connectivity of degree M by assuming that the ambulance can be connected simultaneously to at most M nearest mmWave APs. The blockage process in the presence of multi-connectivity is then formulated as a superposition of individual LoS blockage processes.
- *Step 3:* We extend the LoS blockage model by considering mobility. This extension makes the LoS blockage process non-stationary, since the distance from the ambulance to its ith neighboring AP changes in time. The resulting formulation is a non-homogeneous Markov chain, whose time-dependent intensities are functions of the system parameters.

The **second phase** of our framework characterizes the amounts of available radio resources at LTE and Wi-Fi APs, as detailed in Section V. We demonstrate that the multi-dimensional Markov chain capturing system dynamics can be reduced to its one-dimensional projection by applying stochastic averaging techniques. Having the dynamic LoS blockage model and the models for quantifying the amounts of available resources at Wi-Fi and LTE APs, we proceed with characterizing the system-level performance.

IV. Mathematical Framework for Softwarized 5G RAN: Studying Dynamics of mmWave Links

In this section, we present the first phase of our proposed mathematical framework for the mission-critical use case introduced above. The major notation is summarized in Table I.

TABLE I
NOTATION USED BY THE MATHEMATICAL FRAMEWORK

Parameter	Definition
Environmental parameters	
h_U, l_U, w_U, v_U	Height, length, width, and speed of ambulance vehicle
$h_B, f_{h_B}(x)$	Height of vehicles and its pdf
$l_B, f_{l_B}(x)$	Length of vehicles and its pdf
W_S, w_B	Width of the street and the regular vehicle
h_A	Height of mmWave APs
$d_I, f_{d_I}(x)$	Inter-vehicle distance and its pdf
$v_{B,i}, f_{v_{B,i}}(x)$	Speed of vehicles in lane i and its pdf
M	Degree of multi-connectivity
n	Number of lanes (n is even)
λ_i	Intensity of vehicles in lane i
$\lambda_{E,i}$	Intensity of vehicles in lane i blocking the LoS path
Geometric parameters	
d_A	Distance between mmWave APs
d_W	Distance from ambulance to Wi-Fi AP
$d_{0,1}$	Interval, ambulance is associated with "left" APs
$d_{1,1}$	Interval, ambulance may fall in outage to "left" APs
$d_{2,1}$	Interval, ambulance is associated with "right" APs
AP service process	
θ, μ	Arrival and service rates of sessions
$\phi(t)$	Total number of sessions at time t
$\vec{\psi}(t)$	Vector of resources required by sessions at time t
$\delta(t)$	Amount of resources occupied at time t
L, W	RVs denoting LTE/Wi-Fi session requirements
$P_k(x)$	Stationary resource distribution of i sessions
$F_V(y\|x)$	Amount of resources occupied by a session

A. Dynamics of mmWave Blockage Process

The geometry of the target scenario is shown in Fig. 3, while the distances $d_{0,1}$, $d_{1,1}$, $d_{2,1}$, $d_{C,1}$, and the minimal height of the vehicle in lane i to occlude the LoS path, $h_{B,i}$, are derived with the use of trigonometry tools. Below, we consider the mmWave blockage process for the odd number of APs. The process for the even number is the special case of our model.

1) Case of a Single Lane: Consider the non-blocked interval caused by the vehicles in an arbitrarily chosen lane. Depending on the height $h_{B,i}$, not all of the vehicles occlude the LoS path [45]. The probabilities that a randomly chosen vehicle blocks the LoS path of the target mission-critical link is

$$p_{B,i} = \Pr\{h_B > h_{B,i}\} = \int_{h_{B,i}}^{\infty} f_{h_B}(x)dx, \quad i = 1, 2, \quad (1)$$

where $f_{h_B}(x)$, $x > 0$, is the pdf of the vehicle height.

Assuming the independence of vehicle heights, we observe that the number of vehicles between two consecutive vehicles that block the LoS path follow a geometric distribution with the parameter $q_i = 1/\lambda_{E,i}$, $i = 0, 1, 2$, where $\lambda_{E,i} = p_{B,i}\lambda_i$ is the effective intensity of vehicles in lane i blocking the LoS path. Recalling that the vehicle length and inter-vehicle distance follow the pdfs of $f_{l_B}(x)$ and $f_{d_I}(x)$, $x > 0$, the pdf of the distance between two vehicles occluding the LoS is

$$f_{d_{L,i}}(x) = q_0 f_{d_I}(x) + \sum_{i=1}^{\infty} q_i [f_{l_B}(x) * f_{d_I}(x)]^{(i)}, \quad (2)$$

where the superscript (i) denotes i-fold convolution.

To determine the pdf of time that LoS is not blocked, we need to take into account the random speed of vehicles in a lane. Let $T_{L,i}$ be the random variable (RV) denoting the LoS time and $f_{T_{L,i}}(x)$ be its pdf. Observing that $T_{L,i} = d_{L,i}/v_{B,i}$, where both $d_{L,i}$ and v_i are the RVs, we need to determine the

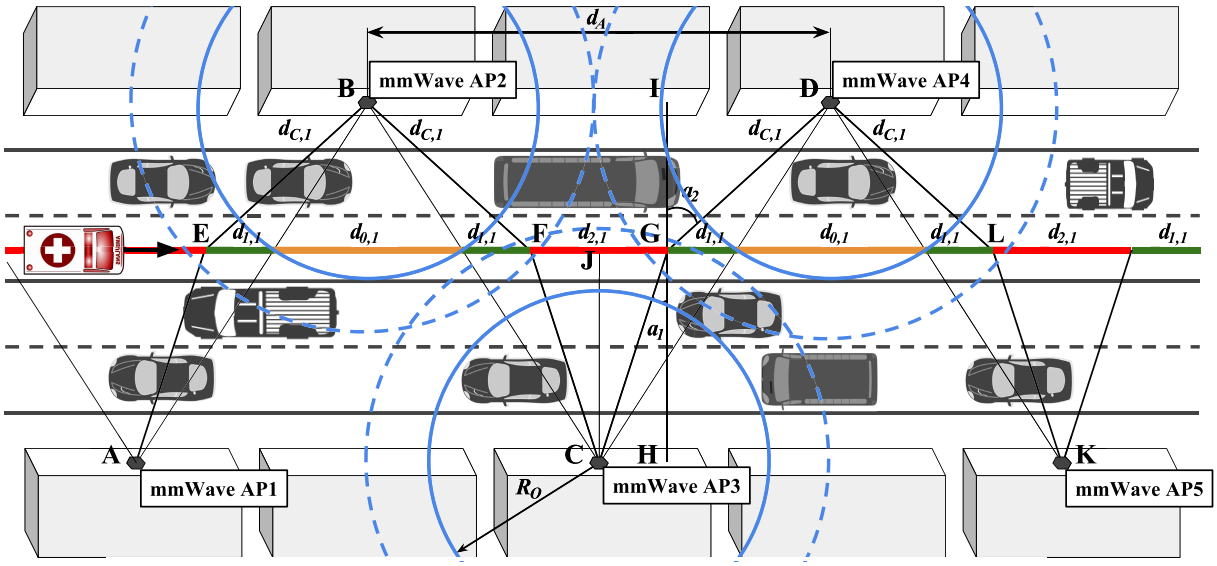

Fig. 3. Periodic connectivity pattern with a single mmWave link. Multi-connectivity of degree M is a superposition of M individual processes and is too complex to be displayed (yellow – no outage, red – possible outage with odd APs, green – possible outage with even APs).

RV that is defined as a ratio of two RVs. The joint pdf of $T_{L,i}$ and $v_{B,i}$ can be obtained as

$$f_{T_{L,i},v_i}(y_1, y_2) = f_{d_{L,i},v_{B,i}}|J(x_1, x_2)|^{-1}, \tag{3}$$

where $J(x_1, x_2)$ is the Jacobian of the transformation.

Observing the structure of the LoS interval in a single lane, we learn that the contribution of the components decreases exponentially as the number of vehicles between two vehicles occluding the LoS increases. In fact, (2) resembles the kernel density approximation of the exponential distribution. Hence, the LoS blockage distance in one lane can be approximated by the exponential distribution with the parameter

$$\lambda_{L,i} = \frac{1}{\lambda_{E,i}} \int_0^\infty f_{w_B}(x)dx \int_0^\infty f_{v_{B,i}}(x)dx, \tag{4}$$

where $\lambda_{L,i}$ is the effective density of vehicles in lane i.

2) Case of Multiple Lanes: Now consider the vehicles in multiple lanes blocking the LoS. Assuming the independence of vehicles deployed across the lanes, the CDF of the LoS interval duration with i lanes is given by the minimum of the LoS intervals in i lanes as

$$F_{T_L}(x) = 1 - \prod_{i=1}^{n}[1 - F_{T_{L,i}}(x)], \tag{5}$$

where $F_{T_{L,i}}(x)$ is the CDF of the LoS duration in lane i.

The blockage interval duration caused by a single vehicle, $T_{B,i}$, is defined by the length of vehicles, l_B, as well as their speed, $v_{B,i}$, as $T_{B,i} = l_B/v_{B,i}$, and can be determined by using (3). To establish the blockage interval duration caused by the vehicles in multiple lanes, observe that the service time distribution is the weighted blockage time caused by a single vehicle in all the lanes, $T_B^\star$, where the pdf of the latter is

$$f_{T_B^\star}(x) = \sum_{j=1}^{i} \frac{E[v_{B,i}]}{\sum_{k=1}^{i} E[v_{B,i}]} f_{T_{B,j}}(x). \tag{6}$$

The blockage time distribution can be approximated by employing the means $E[T_L]$ and $E[T_B^\star]$ for the corresponding queuing systems and then utilizing the Laplace transform (LT) that obeys the Kendall functional equation in the form

$$L_{T_B}(s) = L_{T_B^\star}\left(s + \frac{1}{E[T_L]} + \frac{1}{E[T_B^\star]} - L_{T_B}(s)\right). \tag{7}$$

The approximation of the mean is readily given by [46] as

$$E[T_B] = \frac{E[T_B^\star]E[T_L]}{E[T_L] - E[T_B^\star]}. \tag{8}$$

This specified blockage model can be further extended to the case of e.g., three-sided mobility where all of the entities affecting the channel conditions – namely, the UE, the blockers, and the AP – are potentially mobile. A major extension needed to capture the mobility of the APs is to utilize the movement trajectories of the AP and the UE to characterize the dynamically changing distance between these entities as well as employ it to estimate the Markov blockage model for any fixed time t. Such an extension will enable the analysis of UAV and vehicle-to-vehicle (V2V) communications [47] together with even more flexible network setups including moving access points deployed on vehicles and drones [16].

B. Single- and Multi-Connectivity in mmWave

Let us further incorporate the effects of multi-connectivity. According to our assumptions, at each point of the ambulance trajectory it is allowed to be connected to up to M nearest mmWave APs. For any value of M, the trajectory of the ambulance is divided into periodically repeating segments. Segment lengths as well as distances to the APs can be computed similarly to the single-connectivity case.

First, consider the ambulance vehicle in a certain position at the red segment of the trajectory (see Fig. 3). Observing the positions of APs, we see that for the realistic inter-AP distances, d_A, and any $M \leq 3$, the blockage processes to the

APs are mutually independent. To describe the dynamics of the blockage process for $M = 1$, we model the process of outage at the mmWave AP by using a time-homogeneous continuous-time Markov chain (CTMC) with two states, which has its infinitesimal generator in the following form

$$\Lambda_1 = \begin{pmatrix} -\alpha_1 & \alpha_1 \\ \beta_1 & -\beta_1 \end{pmatrix}, \quad (9)$$

where $\alpha_1 = 1/E[T_{L,1}]$ and $\beta_1 = 1/E[T_{B,1}]$ are the mean durations of the LoS and the blocked intervals.

Consider now the process of outage with the two nearest APs. The associated CTMC model, $\{S_2(t), t > 0\}$, $S_2(t) \in \{1, 2, \ldots, 2^M\}$, is determined by a superposition of the outage processes to the first two APs with the infinitesimal generator

$$\Lambda_2 = \begin{bmatrix} -\alpha_2 - \alpha_1 & \alpha_2 & \alpha_1 & 0 \\ \beta_2 & -\beta_2 - \alpha_1 & 0 & \alpha_1 \\ \beta_1 & 0 & -\beta_1 - \alpha_2 & \alpha_2 \\ 0 & \beta_1 & \beta_2 & -\beta_1 - \beta_2 \end{bmatrix}, \quad (10)$$

where state 1 corresponds to the outage state.

The CTMCs capturing the outage processes from M APs are irreducible and aperiodic, thus implying that there are final state probabilities coinciding with the steady-state probabilities given by $\vec{\pi}_M$. One can obtain $\vec{\pi}_M$ by solving the linear system $\vec{\pi}_M \Lambda_M = 0$, $\vec{\pi}_M \vec{e}^T = 1$, where $\vec{e}$ is the unit vector of size 2^M.

After approximating the LoS blockage process for M nearest APs by using the time-homogeneous CTMC, we can proceed with addressing the ambulance vehicle itself. Consider the case of $M = 1$ and concentrate on the odd APs as illustrated in Fig. 3. The distance from the ambulance to the mmWave AP as the ambulance travels along its trajectory is

$$g(x) = \sqrt{||d_{C,1}||^2 + x^2}, \quad -||CH|| < x < ||CH||, \quad (11)$$

where $||CH||$ is the length of the CH line segment in Fig. 3.

Observe that for a stationary ambulance at the 2D distance x from an AP, the fraction of time when the AP is blocked coincides with the ergodic blockage probability. Hence,

$$f_B(x) = \frac{E[T_B(x)]}{E[T_B(x)] + E[T_L(x)]}, \quad (12)$$

where $E[T_B(x)]$ and $E[T_L(x)]$ are the mean blockage and LoS periods that were derived previously, and x indicates that these quantities are functions of the current distance x to the AP.

Using (11) and (12), the fraction of the blockage time during the interval that the ambulance spends in the red zone in Fig. 3 (where a blockage leads to outage) is given by

$$f_B = \int_{-||CH||}^{||CH||} \frac{\sqrt{||d_{C,1}||^2 + x^2} E[T_B(x)]}{v_U (E[T_B(x)] + E[T_L(x)])} dx. \quad (13)$$

Now observe that the fraction of the blockage time when connected to the even APs can be established similarly. An extension to the multi-connectivity case is also straightforward here. The only difference is that the ergodic blockage probability in (12) needs to be calculated by taking into account M active APs. The time-dependent blockage process can be further obtained from the time-homogeneous CTMC that models the blockage at a certain 2D distance from the AP. Considering the odd APs and $M = 1$ as an example, the infinitesimal generator of the CTMC that captures the time-dependent blockage process is

$$\Lambda_1(x) = \begin{pmatrix} -\alpha_1(x) & \alpha_1(x) \\ \beta_1(x) & -\beta_1(x) \end{pmatrix}, \quad (14)$$

where $\alpha_1(x) = v_U/E[T_{L,1}]$ and $\beta_1(x) = v_U/E[T_{B,1}]$ are the mean durations of the LoS and the blocked intervals with the nearest mmWave AP. The infinitesimal generator can be found by using the Kronecker product of the generators, which represent CTMCs that model the LoS blockage process.

We further apply the above results for the mmWave blockage process in the presence of multi-connectivity.

V. Mathematical Framework for Softwarized 5G: RAN Performance With Mission-Critical Traffic

We proceed with the second phase of our mathematical framework for the ambulance use case specified in Section III.

A. AP Service Process

Consider a multi-server queue with N servers, which corresponds to the maximum number of supported sessions at a single microwave AP. Note that assuming randomness in radio resource requirements allows to represent not only the random session rates required from the network, but also the use of multiple modulation and coding schemes (MCSs) in Wi-Fi and LTE technologies. A translation of random session rates into Hz/s is described in detail in [44]. Assuming that the system can accommodate an unbounded number of sessions, a session is considered to be lost when the amount of available radio resources at the moment of its arrival is insufficient.

A complete description of the system at hand requires the use of a multi-dimensional Markov chain, $\{\vec{S}(t), t > 0\}$, which is defined over $\vec{S}(t) = (\phi(t), \vec{\psi}(t))$, where $\phi(t)$ is the total number of sessions at the state changing time t and $\vec{\psi}(t) = (\psi_1, \psi_2, \ldots, \psi_{\phi(t)})$ is the amount of resources allocated to the ith session. The reason for this complex description is that one needs to keep track of the amount of resources allocated to each active session. This information is to be used at the session departure time to release resources. When the number of sessions allowed per AP is not limited, $N = \infty$, the number of dimensions is infinite. Relying on the stochastic averaging techniques, the model in question can be simplified.

Let $\delta(t) = \sum_{i=1}^{\phi(t)} \psi_i(t)$ be the total amount of occupied resources at time t and consider the two-dimensional process $\{\phi(t), \delta(t), t > 0\}$. As one may observe, this process is non-Markov, since we do not keep track of the amount of resources occupied by each session in service. However, if we assume that at the moment of a session departure, τ, it frees the session-"average" CDF of resources conditioned on the current amount of occupied resources i.e.,

$$Pr\{\nu \le x | \phi(\tau), \delta(\tau) = y\} = F_k(x|y), \quad k = 1, 2, \ldots, \quad (16)$$

where $F_k(x|y)$ is defined as

$$F_k(x|y) = Pr\{r_k \le x | r_1 + r_2 + \cdots + r_k = y\}, \quad (17)$$

$$\theta F(B)p_0 = \mu P_1(B),$$
$$\theta \int\limits_{0\le y\le x} F(B-y)P_1(dy) + \mu P_1(x) = \theta F(x)p_0 + 2\mu \int\limits_{\substack{0\le x\le y\le B\\ y-z\le x}} F_2(dz|y)P_2(dy), \quad 0\le x\le B,$$
$$\theta \int\limits_{0\le y\le x} F(B-y)P_k(dy) + k\mu P_k(x) = \theta \int\limits_{0\le y\le x} F(x-y)P_{k-1}(dy) + (k+1)\mu \int\limits_{\substack{0\le x\le y\le B\\ y-z\le x}} F_{k+1}(dz|y)P_{k+1}(dy), \quad 0\le x\le B. \quad (15)$$

where r_i, $i = 1, ..k$, is the amount of resources occupied by the ith session, then the process $\{\phi(t), \delta(t), t > 0\}$ is Markov. The steady-state distribution of the amount of the occupied resources in the original and the modified systems coincides as proven in [48].

Note that the acceptance of a session implies that the CDF of the amount of requested resources and the CDF of the resource requests by the sessions accepted into the system are generally different. Note that $F_v(x|y)$ is such that $\phi(t_+) = 0$, then all the resources are freed by implying that $F_k(x|y)$ can be calculated recursively from $F_1(x|y)$. Since the following holds

$$Pr\left\{y - r_k \le x \Big| \sum_{i=1}^{k} r_i = y\right\} = Pr\left\{\sum_{i=1}^{k-1} r_i \le x \Big| \sum_{i=1}^{k} r_i = y\right\}, \quad (18)$$

the conditional CDF $F_v(x|y)$ satisfies the following

$$\int\limits_{\substack{0<z<y\le B,\\ y-z\le x}} F_k(dz|y)F^{(k)}(dy) = \int\limits_{0\le y\le x} F(B-y)F^{(k-1)}(dy), \quad (19)$$

as both sides are equal to the joint CDF

$$Pr\{r_1 + \cdots + r_{k-1} \le x, r_1 + \cdots + r_k \le B\}, \quad (20)$$

where $F(y)$ is the CDF of the amount of resources requested by a single session, while superscript $(k-1)$ denotes the k-fold convolution of $F(y)$, and B is the total volume of resources.

As B is limited, there always exists a bivariate mixed stationary distribution of the Markov chain $\{\phi(t), \delta(t), t > 0\}$,

$$p_0 = \lim_{t\to\infty} Pr\{\phi(t) = 0\},$$
$$P_k(x) = \lim_{t\to\infty} Pr\{\phi(t) = k, \delta(t) \le x\}, \quad (21)$$

which satisfies the set of Kolmogorov equations in (15), as shown at the top of this page, that can be solved by applying the conventional method for mixed discrete-continuous Markov chains.

B. Implications for Best-Effort Traffic

After defining the dynamic blockage process as well as the stationary distribution of radio resources at Wi-Fi and LTE APs, we can proceed with analyzing the system performance by concentrating on the upper bound of the fallback time and the intensity of dropped best-effort sessions.

1) Upper Bound on Fallback Time: Observe that the ambulance vehicle may only experience outage in mmWave access, which leads to a fallback onto Wi-Fi or LTE when passing the $d_{2,1}$ segment. The outage process at this segment, given the degree of multi-connectivity M, is governed by a non-homogeneous CTMC. As the exact analysis of the outage time is extremely convoluted, here we derive an upper bound on the outage time. Particularly, we determine the pdf and the mean value of the outage time when upon entering $d_{2,1}$ the ambulance is in the blocked state.

The sought distribution is of the phase-type nature $(\vec{\alpha}, S)$, where α is the initial state distribution upon entering $d_{2,1}$, which is defined over $\{2, 3, \ldots, 2^i\}$. The pdf is then given by

$$f_P(t) = \vec{\alpha}e^{St}\vec{s}_0, \quad t > 0, \quad (22)$$

where $\vec{s}_0 = -S\vec{1}$, $\vec{1}$ is the vector of ones with the size of $1 \times 2^i - 1$, and e^{St} is the matrix exponential defined as

$$e^{St} = \sum_{k=0}^{\infty} \frac{1}{k!}(St)^k. \quad (23)$$

The initial state probability vector $\vec{\alpha}$ can be established from the steady-state distribution, $\vec{\pi}$, of the time-homogeneous CTMC blockage model for the ambulance located at the entrance point to the segment $d_{2,1}$, which is a solution to the system $\vec{\rho}\Lambda_i = \vec{\rho}$, $\rho\vec{e} = 1$, where $\vec{e}$ is the unit vector. Accordingly, we have

$$\alpha_i = \begin{cases} 1, & i = 1, \\ 0, & i = 2, 3, \ldots. \end{cases} \quad (24)$$

2) Intensity of Dropped Sessions With LTE Fallback: Let L be the amount of resources that are requested from the LTE by a session transferred from the mmWave segment and define

$$P(x) = \sum_{i=0}^{\infty} P_i(x), \quad 0 < x < B, \quad (25)$$

to be the CDF of the amount of resources occupied by an LTE session at the LTE AP, where $P_i(x)$ is the joint distribution that there are i sessions in the LTE system and they occupy x amount of resources, see subsection V-A. Note that taking into account the presence of the LTE power control function, we may assume that L is constant.

Assume that the session that is transferred onto LTE is dropped when there are insufficient resources at the LTE AP. It can only happen at some segment $d_{2,1}$, when the following two events occur simultaneously. First, the ambulance vehicle

has to experience outage at some point of $d_{2,1}$ and then there must be insufficient resources at the LTE AP. Since these two events are independent, we have session drop probability, p_D, as $p_D = p_{D,M}p_{D,L}$, where $p_{D,M}$ is the session drop probability due to mmWave outage and $p_{D,L}$ is the probability that there are insufficient resources at the LTE AP. The latter can be easily established as the probability of the event $\{L-(B-\delta) > 0\}$, where δ is the RV denoting the steady-state distribution of the amount of occupied resources. We thus have

$$p_{D,L} = Pr\{\delta - B + L > 0\} = \int_0^B P(x + B - L)dx. \quad (26)$$

It is important to note that there might be multiple outages occurring inside a single segment $d_{2,1}$. However, for the realistic values of the offered traffic load on LTE, the system does not return to the steady-state between two consecutive outage events. This implies that if the transferred session sees sufficient resources at its first outage, this will also happen at further outage(s) within the same $d_{2,1}$. Assuming the independence of the blockage processes at successive segments $d_{2,1}$, we conclude that the number of segments of length $d_{2,1} + 2d_{1,1} + d_{0,1}$ until the segment where the outage occurs follow a geometric distribution with the parameter p_D. Letting T_2 denote the time to pass the segment $2d_{1,1} + d_{0,1}$, the probability mass function of time until the outage is

$$f_{T_O,i} = (1 - p_D)p_D^i, \quad i = 0, T_1 + T_2, \ldots, \quad (27)$$

with the mean delivered by the Wald identity

$$E[T_O] = \frac{1 - p_D}{p_D}(T_1 + T_2). \quad (28)$$

Define the intensity of session drops at LTE, $\gamma_{D,L}$, as

$$\gamma_{D,L} = \lim_{t\to\infty} \frac{E[N(t)]}{t}, \quad (29)$$

where $E[N(t)]$ is the number of dropped sessions in $(0, t)$.

Observe that the durations T_1 and T_2 are deterministic; hence, we may employ a geometric distribution with the mean $(T_1 + T_2)(1 - p_{D,M}p_{D,L})/(p_{D,M}p_{D,L})$, where $p_{D,L}$ is provided in (26). Therefore, one can write

$$\gamma_{D,L} = \lim_{t\to\infty} \frac{E[N(t)]}{t} = \frac{p_{D,M}p_{D,L}E[N]}{(T_1 + T_2)(1 - p_{D,M}p_{D,L})}, \quad (30)$$

where $E[N]$ is the only unknown, which corresponds to the number of sessions dropped during a single session drop event.

Recall that the amount of resources occupied by an LTE session when there are x resources occupied in the system is different from the amount of resources requested by an LTE session upon its arrival. The CDF of the former is given by $F_k(y|x)$ and has been computed in subsection V-A. The conditional mean is then given by

$$E[S_L|\delta] = \sum_{k=1}^{\infty} \int_0^B k f_k(ky|x)dy, \quad (31)$$

where $f_k(ky|x)$ is the pdf of δ/k and δ is the RV denoting the amount of resources occupied in the steady-state.

Observe that the RV $L-(B-\delta)$ is conditioned on the event that there is an overflow at the LTE AP, $L - (B+\delta) > 0$.

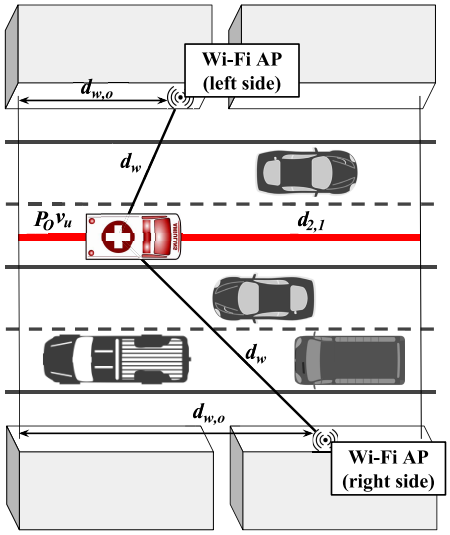

Fig. 4. Illustration of the random distance to Wi-Fi AP.

If a session has to be dropped, the corresponding amount of resources that needs to be vacated to accept it should be characterized. Conditioning on $\delta = x$, we account for different mean sizes of the resources occupied by a single connection, which makes $L - (B - \delta)$ a constant. Hence, for $E[N]$ we have

$$E[N] = \int_{B-L}^{B} \frac{y - B + x}{E[S_L|\delta]} p(x)dx, \quad (32)$$

where $E[S_L|\delta]$ has been obtained in (31).

3) Intensity of Dropped Sessions With Wi-Fi Fallback: Addressing the Wi-Fi technology that is also used as a back-up connectivity option, we note that the absence of power control mechanism requires to explicitly track the distance between the Wi-Fi AP and the ambulance vehicle at the first session outage. Assuming that the position of the Wi-Fi AP is distributed along the street segment uniformly over $d_{2,1}$, the distance between the ambulance and the Wi-Fi AP at the time instant of the first outage is given by (see Fig. 4)

$$d_W = \begin{cases} \sqrt{\left(\frac{n+1}{2}d_L\right)^2 + (P_O v_U - d_{W,O})^2}, & \text{“left”-side AP,} \\ \sqrt{\left(\frac{n-1}{2}d_L\right)^2 + (P_O v_U - d_{W,O})^2}, & \text{“right”-side AP,} \end{cases} \quad (33)$$

where the former case assumes that the Wi-Fi AP is on the “left” side of the street, while the latter one models the situation where the Wi-Fi AP is located on the “right” side, see Fig. 4. Further, P_O is the RV describing the time until the first outage (given that it happens), $d_{W,O}$ is the position offset of the Wi-Fi AP. Therefore, d_W is an RV, whose pdf for each case can be established by utilizing the RV transformation techniques similarly to (3). Observing that the probability of having the Wi-Fi AP at any side of the street is 0.5, the resulting pdf $f_{d_W}(x)$ is defined as a weighted sum of the components.

The time from the starting point of the segment $d_{2,1}$ to the outage event (if it happens at this segment) is

calculated by

$$f_{P_O}(t) = f_P(t) \Big/ \int_{d_{2,1}/v_u}^{\infty} f_P(t)dt, \quad 0 < t < d_{2,1}/v_u, \quad (34)$$

where $f_P(t)$ is the pdf of the first-passage time from the set of the non-blocking states, $\{2, 3, \ldots, 2^i\}$, to the blocking state, i is the degree of multi-connectivity obtained similarly to (22).

Let W be an RV denoting the amount of resources requested from the Wi-Fi AP. Once $f_{d_W}(x)$ is established, the said amount is derived similarly to [44] by explicitly taking into account the set of MCSs. The fundamental difference between the LTE and the Wi-Fi modeling here is that in the latter case the amount of requested resources during the outage is an RV. In this case, the mean of the RV describing the amount of resources taken from the ongoing LTE sessions $W - (B - \delta)$, conditioned on a certain amount of the occupied resources and the positiveness of $W - (B - \delta)$, can be expressed as

$$E[\delta + W - B|\delta] = \int_0^B \frac{f_L(z + B - x)}{\int_{-B}^0 f_L(z + B - x)dz} z dz, \quad (35)$$

thus leading to the following expression for $E[N]$

$$E[N] = \int_{B-L}^{B} \frac{E[x + W - B|\delta]}{E[S_W|\delta]} p(x)dx, \quad (36)$$

where $E[S_W|\delta]$ is the mean amount of resources occupied by a single Wi-Fi session, given that the total amount of the occupied resources is $\delta = x$, calculated similarly to (31). The rest of this analysis is similar to the LTE fallback case.

In the following section, we augment the proposed framework with a measurement-based campaign to study the practical effects of the mission-critical traffic. We thus complement the RAN-specific performance indicators that can be directly derived from our mathematical framework with the CN-specific numerology obtained via field measurements.

VI. Experimental Study of Softwarized 5G: Effects of Mission-Critical Traffic on CN

In this section, we present the rationale behind our trial implementation of a softwarized 5G network. We utilize this implementation later on to characterize the mission-critical data flows in the softwarized 5G CN. The contributed study effectively complements those related to softwarized 5G RAN and conducted with the aid of the mathematical framework developed in Sections IV and V. It therefore allows to comprehensively characterize the impact of mission-critical traffic on the softwarized 5G network from an end-to-end perspective.

A. Features of Softwarized 5G Architecture

Reliable and timely delivery of mission-critical data over wireless alone is not sufficient to guarantee the required end-to-end reliability, since the CN part has to also be considered. Particularly, the transport network plays an important role in the successful delivery of KPIs, such as reliability. Both mission-critical and best-effort traffic will have to travel through a complex network while competing for resources during transmission, buffering, and computing. In order to achieve the desired levels of end-to-end reliability, the mission-critical traffic has to be properly marked, identified, and prioritized at all the network elements.

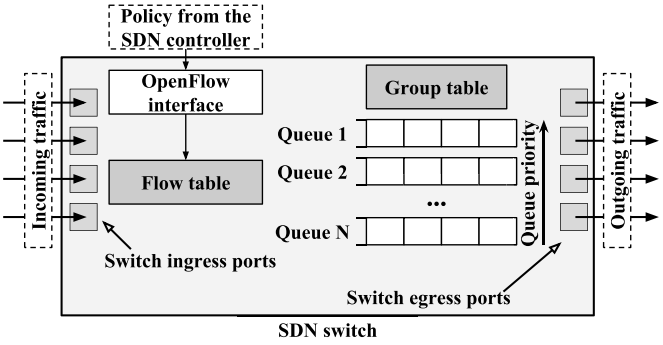

Fig. 5. Components of OpenFlow switch in softwarized 5G CN.

To this aim, in our considered softwarized 5G architecture (see Fig. 1 in Section II), the QoS management and policy enforcement functions within the user plane are controlled by the SMF, which is also in charge of the traffic steering at the UPF to route data to its intended destination. To ensure appropriate mapping of QoS onto adequate forwarding rules in the transport network, the SMF talks to the vSDN controller during the session initiation. The vSDN controller configures the transport network elements with the forwarding rules.

Therefore, SDN is introduced to manage the different flows arriving from the access network, create a complete isolation between them, and apply the necessary forwarding rules to ensure that the QoS is maintained throughout the network. In a network with multiple flows that require different levels of QoS, SDN allows to dynamically control their respective KPIs depending on the effective demands by each of these flows.

The SDN control plane uses *OpenFlow*, which is a communication protocol between the control and the infrastructure. The SDN controller can modify the forwarding rules for each flow that are kept at the infrastructure layer in the form of a flow table. In order to maintain the priority of the critical flows with respect to other flows (e.g., best-effort traffic), the SDN controller maps a flow onto a priority queue. To this end, Fig. 5 presents a diagram of the main components of an OpenFlow switch utilized for the purposes of this study.

B. Representing Softwarized 5G Data Networks

Here, we consider a network comprising SDN-capable switches connected to a number of servers, which are running various 5G VNFs in both the user and the control plane. The physical network topology is illustrated in Fig. 6, where the network elements are also mapped onto the elements of the system model described in Section III.

To make sure the QoS is aligned across the entire user plane path, the SDN controller has to receive the traffic shaping configuration from the control plane function in charge of the traffic steering and policy/QoS control (according to the considered architecture, the SMF). The SDN controller will then configure all the switches, so that these are able to differentiate between the various types of traffic and isolate the

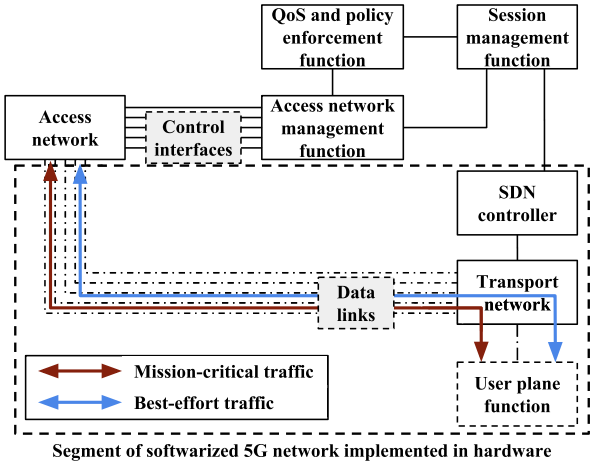

Fig. 6. Softwarized 5G network and its segment used for the study on the data flows coexistence that was implemented in hardware.

mission-critical data accordingly. In other words, SDN controller reserves a slice of the transport link capacity, such that the network can satisfy the minimum guaranteed bitrate at all times as well as ensure no packet losses.

While the OpenFlow controls the application of traffic forwarding rules, the OVSDB (Open vSwitch Database) protocol is used for the configuration and management of the OpenFlow switches. OVSDB provides an interface for the centralized controller to configure the underlying hardware of switches and thus meet the network requirements. The parameters can be configured from the SDN controller, which includes configuring data-paths and assigning ports to them, as well as managing link speeds, queue configuration, and policing.

C. Implementing Softwarized 5G Data Networks

In our test implementation, we employ equipment similar to that utilized in a practical 5G CN. Particularly, we reuse the real-world segment of the UK 5G test network, which is responsible for the data network part. To demonstrate the end-to-end approach advocated in this paper, we construct a simple but representative proof-of-concept implementation. We thus use a Pica8 SDN-capable switch and five Linux servers connected to it, whereas the rack of the relevant equipment utilized in our Lab for testing is displayed in Fig. 7.

In our test implementation, one of the Linux machines acts as the SDN controller, while others are hosts used to either generate or receive data traffic. Thus generated data flows of UDP packets represent either mission-critical or best-effort traffic, which is created with iperf. The latter tool also allows to measure bandwidth, jitter, and packet loss. In order to assess the round-trip time (RTT), we utilize ping, using which requires that an L2-switch module be additionally installed inside the controller. This makes the switch capable of handling the ICMP packets from ping.

Further, we employ the OpenDaylight (ODL) SDN controller to manage the SDN-capable switch. The communication between the ODL controller and the switch is based on the OpenFlow protocol, and in our case has been implemented via the REST API in order to apply the appropriate flow configuration to each switch. Relying on this controller, we assign

Fig. 7. Network equipment for experimental testing of SDN features.

TABLE II
OUR EXPERIMENTAL SETTINGS

Nodes	Traffic type	Queue	Queue min and max rates
APa to UPFa	Critical	Q1	35Mpbs and 50Mbps
APb to UPFb	Best-effort	Q0	20Mbps and 55Mbps

the data flows initiated by the Linux hosts to the queues. To distinguish the critical and the best-effort traffic, we identify the destination IP as encapsulated in the packets. The queues prioritize the data flows at the egress interface of the switch as displayed in Table II. The minimum (guaranteed) and maximum service rate need to be maintained in each of the queues to avoid disturbance from the best-effort traffic toward the critical traffic, since the sum of both data flows tends to reach the full capacity of the link.

We further utilize our developed platform to obtain the following numerical results and thus complement the approach of the proposed mathematical framework.

VII. MAIN NUMERICAL RESULTS AND DISCUSSION

In this section, we summarize our numerical results on the performance of the described system. We address a realistic urban scenario, which may serve as a representative setup against a variety of possible deployment configurations. It helps illustrate the typical dependencies and understand the performance insights of softwarized 5G networking, as it also completes our holistic evaluation approach.

A. Parametrization of Target Scenario

We model a 4-lane bidirectional street[3] having the width of 20 m in total, whereas the lane width is set to 3.65 m.

[3]UK Government Standard "Design manual for roads and bridges," vol. 6, February 2005.

We assume the Mercedes Benz Sprinter form factor with the dimensions of 6025 mm, 2380 mm, and 2630 mm as the ambulance vehicle.[4] Following the data from [49], we allow the length of other vehicles, l_B, to follow a Gamma distribution with the mean of 4.5 m and the variance of 0.5 m^2. The height of other vehicles, h_B, is assumed to adhere to a Gamma distribution with the mean of 1.7 m and the variance of 0.3 m^2.

The speed of the ambulance vehicle varies from 50 km/h to 110 km/h, while the speed of vehicles in other three lanes is assumed to be the same and follow a uniform distribution from 15 km/h to 50 km/h, which is representative of dense urban traffic in large cities. The target data rate for a mission-critical session varies from 10 Mbit/s to 50 Mbit/s, whereas the data rate demanded by the best-effort sessions on Wi-Fi is assumed to be 1 Mbit/s with the average duration of 120 s (corresponding to a short audio call). The data rate demanded by the best-effort sessions on LTE is assumed to be 5 Mbit/s with the average duration of 60 s (modeling a short video call).

Radio propagation over mmWave follows the default 3GPP model [50], where the case of blockage is pessimistically modeled as nLoS conditions. This is because in case of blockage the mmWave RAT is capable of establishing an alternative nLoS path that is currently non-blocked. We also adopt the numerology typical of mmWave cellular. Particularly, the transmit power is set to 25 dBm and the center frequency is 28 GHz with 1 GHz of bandwidth. The antenna gains at the AP and the ambulance vehicle are assumed to be 15 dB and 10 dB, respectively. The noise floor is equal to −80 dB.

Assuming the best case of perfect end-to-end reliability of mission-critical traffic – so that none of the components handling the mission-critical transmissions are failing over an infinitely-long time period,[5] we now focus on what are the costs to achieve it in the softwarized 5G network. In other words, the question is what would be the performance indicators associated with the network load and the best-effort sessions from the regular users. In the sequel, we focus on the following five metrics of interest:

- *Fallback time fraction*. Assuming that the mission-critical session is sufficiently long, this is the fraction of time that it is served by the microwave radio (either Wi-Fi or LTE).
- *Fallback time interval*. This is the worst-case uninterrupted time interval, where the ambulance is in the outage on the mmWave network and has to transfer its mission-critical session onto the microwave radio.
- *Drop rate of best-effort sessions*. This is the average number of best-effort sessions dropped per a time unit in order to vacate sufficient radio resources for the mission-critical traffic.
- *Packet delay*. This is the average delay of the best-effort traffic, when a mission-critical session falls back onto the microwave radio.
- *Jitter*. This is the variation of the best-effort data packet delay, when a mission-critical session falls back onto the microwave radio.

[4]NSW Ambulance, "Vehicle and stretcher dimensions," April 2015.

[5]ETSI, "Network Functions Virtualisation (NFV); Report on Models and Features for End-to-End Reliability," GS NFV-REL 003 V1.1.1, April 2016.

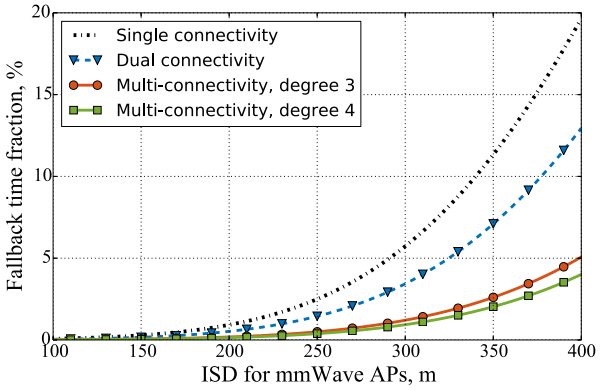

Fig. 8. Fallback time fraction vs. NR ISD.

The first three KPIs are obtained within our mathematical framework, whereas the latter two are derived with the help of our practical implementation.

B. Interpretation of Performance Results

In what follows, the obtained modeling and measurement results are summarized and explained.

1) Degree of Multi-Connectivity in mmWave: We begin with Fig. 8, which illustrates the effect of inter-site distance (ISD) between mmWave APs on the fallback time fraction. First, we observe that the considered system does not consume much microwave resources in case of ultra-dense NR deployment (ISD $\leq$ 150 m). We also observe that dual-connectivity and multi-connectivity with the degree of 3 lead to considerable gains over single connections in the modeled scenario. A decrease in the fallback time fraction is from 20% down to 13% and then down to 5%, respectively. We finally notice that the degree of multi-connectivity higher than 3 does not produce any advantages for the street deployment of mmWave APs.

2) Inter-Site Distance Between mmWave APs: We continue with analyzing the duration of the fallback time interval. Since precise derivation of this parameter is not feasible in the considered model (see Section IV for details), here we provide its worst-case estimate by assuming that the mmWave outage begins at the point where it would statistically last for the longest duration. Fig. 9 reports on the pdf of this interval subject to the speed of the ambulance vehicle and the ISD. We clearly observe that the higher speed of the ambulance results in smaller mean and variance of this KPI. We also notice that the ISD has a more pronounced impact on the fallback time interval: the latter for 200 m ISD (65 km/h speed) is around 4 times shorter than the corresponding value for 400 m ISD. Similar observations hold for other values of the speed.

3) Speed of Ambulance Vehicle: Further, we study the effects of node mobility on the performance of the softwarized network. We thus investigate the relation between the impact of the ambulance speed on the drop rate of the best-effort sessions and the fallback time interval, which is illustrated in Fig. 10. The data rate of the mission-critical traffic is set to 30 Mbit/s. Here, we notice the negative effect the ambulance speed has on the drop rate for both LTE and Wi-Fi fallback.

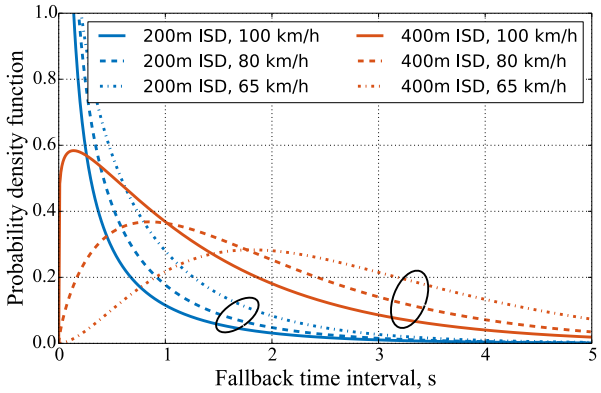

Fig. 9. Fallback time interval vs. ISD and ambulance speed.

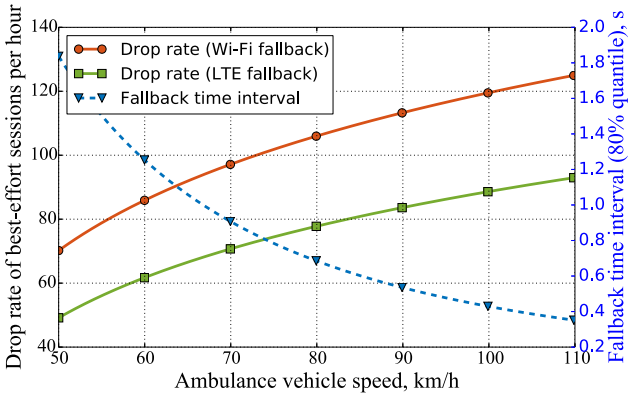

Fig. 10. Drop rate of best-effort sessions vs. ambulance speed.

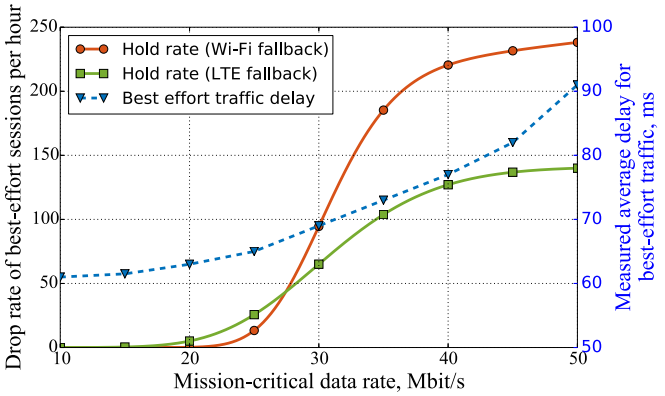

Fig. 11. Drop rate and measured delay of best-effort sessions vs. mission-critical data rate.

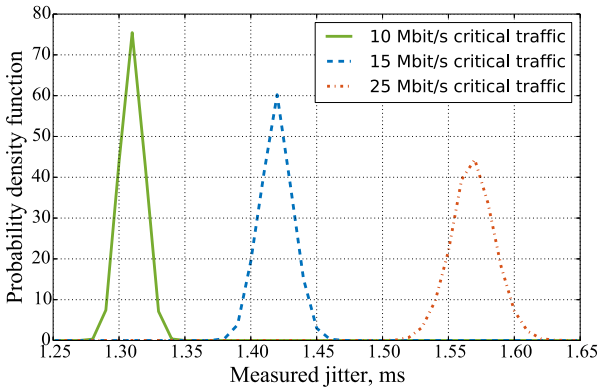

Fig. 12. Measured jitter when best-effort and mission-critical traffic coexist in CN.

Particularly, the average rate of dropped sessions increases by about 90% and 100% for Wi-Fi and LTE fallback, respectively, as the ambulance speed grows from 50 km/h to 110 km/h.

4) Volume of Mission-Critical Traffic: We proceed with Fig. 11, where both the modeling and the measurement results are displayed, since the target data rate of the mission-critical traffic influences the operation of both the RAN and the CN. First, we confirm that the drop rate grows with the increasing data rate. However, the real challenges emerge only after around 20 Mbit/s. This is because for the considered set of parameters lower data rates can be accommodated by spare radio resources. Then, both curves for the drop rate clearly face saturation at about 50 Mbit/s, as most of the best-effort sessions are dropped during a fallback when the mission-critical data rate is comparable with the system data rate (50 Mbit/s for LTE and 72 Mbit/s for Wi-Fi).

Finally, the growth of the Wi-Fi fallback curve is slower than that of the LTE fallback curve for lower data rates, whereas it becomes exactly the opposite for higher data rates of the mission-critical traffic. In contrast, the Wi-Fi sessions having the same data rate may occupy considerably different amounts of radio resources, which depend on the relative locations of the user and the AP. Hence, a softwarized Wi-Fi network will first drop larger sessions and only then discard smaller ones.

We consider Fig. 11 to be of particular importance for network engineers when provisioning for mission-critical traffic, since it offers insights into crucial design choices. Particularly, Fig. 11 allows to evaluate the implications of splitting a mission-critical session in case of its fallback onto several microwave RATs. Accordingly, the data flow may be divided into "virtual" data streams with the corresponding drop rates. For instance, if a 50 Mbit/s mission-critical session (equivalent to either 240 Wi-Fi or 130 LTE sessions) is split into a 20 Mbit/s flow over LTE and a 30 Mbit/s flow over Wi-Fi, the negative impact on the best-effort traffic decreases down to 93 Wi-Fi and 7 LTE sessions dropped per hour.

5) Bandwidth Reservation in CN: Continuing with Fig. 11, we now focus on the results for the CN, which are obtained according to the methodology described in Section VI. We confirm the trend for the average delay to grow with the increasing mission-critical data rate. This trend is super-linear: the gap between 10 Mbit/s and 20 Mbit/s is only 2 ms, whereas 40 Mbit/s and 50 Mbit/s differ by 14 ms. Hence, with higher mission-critical traffic load the service of the best-effort traffic degrades dramatically.

Finally, Fig. 12 presents empirical pdf for the jitter of the best-effort traffic. Comparing the results for 10 Mbit/s, 15 Mbit/s, and 25 Mbit/s data rate of the mission-critical traffic, we note that both the average value and the variance of jitter become considerably worse at higher loads. We also notice that the absolute values of the averages for 10 Mbit/s and 25 Mbit/s are on the order of 0.2 ms, with the relative difference of over 20%. Since only a segment of the real network has been used for our measurements, the ultimate performance

degradation caused by serving mission-critical traffic for the jitter of the best-effort data can become even more severe.

VIII. Conclusions

The ongoing softwarization of network elements is expected to become the next big wave in communications, thus enabling a number of advanced applications one could have considered infeasible only a decade ago. One of such services is provisioning end-to-end reliability for high-rate mission-critical traffic between a server in the data network and the highly-mobile user. A solution to this challenge is extremely convoluted and requires efficient orchestration of network functionality at different levels. It should account for spontaneous changes of the serving access point according to the user mobility pattern as well as highly dynamic channel conditions. In this work, we first introduced a softwarized 5G architecture tailored to this challenging use case. We then presented our mathematical framework, which is capable of capturing the system dynamics at the access network level. We finally developed a hardware implementation of the 5G core network segment and studied the coexistence of the mission-critical and the best-effort traffic at the core network level.

Our performance predictions reveal that even in the idealistic cases of uninterrupted microwave coverage, there is a notable cost of supporting mission-critical traffic in 5G systems. Particularly, both high data rate of the critical sessions and high velocity of the target user bring considerable degradation to the service of other user sessions. At the same time, we show that intelligent use of multi-connectivity in mmWave access as well as splitting the fallback traffic across multiple microwave technologies can mitigate these negative effects for other users. Hence, network configuration has to be carefully adapted in order to balance the impact on different user classes and the complexity of the overall system, mindful of the operator's cost function. Our proposed evaluation methodology can be further employed to analyze particular (beyond-)5G scenarios, which involve different deployments, various user mobility models, desired programmable connectivity policies, and, consequently, optimize the network deployment configurations.

References

[1] M. Mirahsan, R. Schoenen, and H. Yanikomeroglu, "HetHetNets: Heterogeneous traffic distribution in heterogeneous wireless cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2252–2265, Oct. 2015.

[2] Y. Hirota, K. Takahashi, H. Tode, and K. Murakami, "P2P-based ultra high definition multi-view video distribution system with best-effort and bandwidth guaranteed networks," in *Proc. 13th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2016, pp. 774–775.

[3] 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.

[4] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1657–1681, 3rd Quart., 2017.

[5] I. Farris, T. Taleb, H. Flinck, and A. Iera, "Providing ultra-short latency to user-centric 5G applications at the mobile network edge," *Trans. Emerg. Telecommun. Technol.*, to be published. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ett.3169

[6] H. Freeman and R. Boutaba, "Networking industry transformation through softwarization [the president's page]," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 4–6, Aug. 2016.

[7] S. Fichera, M. Gharbaoui, P. Castoldi, B. Martini, and A. Manzalini, "On experimenting 5G: Testbed set-up for SDN orchestration across network cloud and IoT domains," in *Proc. IEEE Conf. Netw. Softw. (NetSoft)*, Jul. 2017, pp. 1–6.

[8] M. Mu *et al.*, "A scalable user fairness model for adaptive video streaming over SDN-assisted future networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2168–2184, Aug. 2016.

[9] M. Amani, T. Mahmoodi, M. Tatipamula, and H. Aghvami, "SDN-based data offloading for 5G mobile networks," *ZTE Commun.*, vol. 12, pp. 34–40, Jun. 2014.

[10] Z. Zaidi, V. Friderikos, Z. Yousaf, S. Fletcher, M. Dohler, and H. Aghvami. (Aug. 2017). "Will SDN be part of 5G?" [Online]. Available: https://arxiv.org/abs/1708.05096

[11] L. Mamatas, S. Clayman, and A. Galis, "A service-aware virtualized software-defined infrastructure," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 166–174, Apr. 2015.

[12] S. Abdelwahab, B. Hamdaoui, M. Guizani, and T. Znati, "Network function virtualization in 5G," *IEEE Commun. Mag.*, vol. 54, no. 4, pp. 84–91, Apr. 2016.

[13] A. Nakao, "Network virtualization as foundation for enabling new network architectures and applications," *IEICE Trans. Commun.*, vol. E93-B, no. 3, pp. 454–457, Mar. 2010.

[14] H. Farhady, L. HyunYong, and N. Akihiro, "Software-defined networking: A survey," *Comput. Netw.*, vol. 81, pp. 79–95, Apr. 2015.

[15] S. Andreev *et al.*, "Exploring synergy between communications, caching, and computing in 5G-grade deployments," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 60–69, Aug. 2016.

[16] S. Andreev, V. Petrov, M. Dohler, and H. Yanikomeroglu. (2017). "Future of ultra-dense networks beyond 5G: Harnessing heterogeneous moving cells." [Online]. Available: https://arxiv.org/abs/1706.05197

[17] M. Casado, N. Foster, and A. Guha, "Abstractions for software-defined networks," *Commun. ACM*, vol. 57, pp. 86–95, Sep. 2014.

[18] O. Rottenstreich, I. Keslassy, Y. Revah, and A. Kadosh, "Minimizing delay in network function virtualization with shared pipelines," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 1, pp. 156–169, Jan. 2017.

[19] M. Karimzadeh, L. Valtulina, H. V. D. Berg, A. Pras, M. Liebsch, and T. Taleb, "Software defined networking to support IP address mobility in future LTE network," in *Proc. Wireless Days*, Mar. 2017, pp. 46–53.

[20] M. Dohler *et al.*, "Internet of Skills, where robotics meets AI, 5G and the Tactile Internet," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2017, pp. 1–5.

[21] "5G for mission critical communication," Nokia, Espoo, Finland, White Paper, 2016. [Online]. Available: http://www.hit.bme.hu/~jakab/edu/litr/5G/Nokia_5G_for_Mission_Critical_Communication_White_Paper.pdf

[22] V. Oleshchuk and R. Fensli, "Remote patient monitoring within a future 5G infrastructure," *Wireless Pers. Commun.*, vol. 57, pp. 431–439, Apr. 2011.

[23] A. Aijaz, M. Dohler, A. H. Aghvami, V. Friderikos, and M. Frodigh, "Realizing the Tactile Internet: Haptic communications over next generation 5G cellular networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 82–89, Apr. 2017.

[24] X. Ge, S. Tu, G. Mao, and C. X. Wang, "5G ultra-dense cellular networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.

[25] V. Petrov *et al.*, "Vehicle-based relay assistance for opportunistic crowdsensing over narrowband IoT (NB-IoT)," *IEEE Internet Things J.*, to be published. [Online]. Available: http://ieeexplore.ieee.org/document/7857676/

[26] "Leading the world to 5G," Qualcomm, San Diego, CA, USA, White Paper, 2016. [Online]. Available: https://www.qualcomm.com/media/documents/files/qualcomm-5g-vision-presentation.pdf

[27] R. F. Moyano, D. Fernández, L. Bellido, N. Merayo, J. C. Aguado, and I. de Miguel, "NFV-based QoS provision for software defined optical access and residential networks," in *Proc. IEEE/ACM 25th Int. Symp. Quality Service (IWQoS)*, Jun. 2017, pp. 1–5.

[28] V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference and SINR in millimeter wave and terahertz communication systems with blocking and directional antennas," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1791–1808, Mar. 2017.

[29] N. Zhang, Y.-F. Liu, H. Farmanbar, T.-H. Chang, M. Hong, and Z.-Q. Luo, "Network slicing for service-oriented networks under resource constraints," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2512–2521, Nov. 2017.

[30] Y. Li, L. T. X. Phan, and B. T. Loo, "Network functions virtualization with soft real-time guarantees," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2016, pp. 1–9.

[31] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, May 2017.
[32] A. Ravanshid *et al.*, "Multi-connectivity functional architectures in 5G," in *Proc. IEEE ICC Workshops*, May 2016, pp. 187–192.
[33] R. Ford, M. Zhang, M. Mezzavilla, S. Dutta, S. Rangan, and M. Zorzi, "Achieving ultra-low latency in 5G millimeter wave cellular networks," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 196–203, Mar. 2017.
[34] *System Architecture for the 5G System*, document TR 23.501, v1.4.0, 3GPP, 2017.
[35] A. Blenk, A. Basta, and W. Kellerer, "HyperFlex: An SDN virtualization architecture with flexible hypervisor function allocation," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage. (IM)*, May 2015, pp. 397–405.
[36] A. Sheoran, X. Bu, L. Cao, P. Sharma, and S. Fahmy, "An empirical case for container-driven fine-grained VNF resource flexing," in *Proc. IEEE Conf. Netw. Function Virtualization Softw. Defined Netw. (NFV-SDN)*, Nov. 2016, pp. 121–127.
[37] A. Ledjiar, E. Sampin, C. Talhi, and M. Cheriet, "Network function virtualization as a service for multi-tenant software defined networks," in *Proc. 4th Int. Conf. Softw. Defined Syst. (SDS)*, May 2017, pp. 168–173.
[38] S. Oechsner and A. Ripke, "Flexible support of VNF placement functions in OpenStack," in *Proc. 1st IEEE Conf. Netw. Softw. (NetSoft)*, Apr. 2015, pp. 1–6.
[39] F. Z. Yousaf, M. Bredel, S. Schaller, and F. Schneider, "NFV and SDN—Key technology enablers for 5G networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2468–2478, Nov. 2017.
[40] "5G and e-Health," 5G-PPP, White Paper, Feb. 2015.
[41] S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, and Y. Koucheryavy, "Analyzing assisted offloading of cellular user sessions onto D2D links in unlicensed bands," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 67–80, Jan. 2015.
[42] O. Galinina, A. Pyattaev, S. Andreev, M. Dohler, and Y. Koucheryavy, "5G multi-RAT LTE-WiFi ultra-dense small cells: Performance dynamics, architecture, and trends," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1224–1240, Jun. 2015.
[43] T. Bai, A. Alkhateeb, and R. W. Heath, Jr., "Coverage and capacity of millimeter-wave cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 70–77, Sep. 2014.
[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] 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.
[46] S. M. Ross and S. Seshadri, "Hitting time in an $M/G/1$ queue," *J. Appl. Probab.*, vol. 36, no. 3, pp. 934–940, 1999.
[47] V. Petrov, J. Kokkoniemi, D. Moltchanov, J. Lehtomaki, M. Juntti, and Y. Koucheryavy, "The impact of interference from the side lanes on mmWave/THz band V2V communication systems with directional antennas," *IEEE Trans. Veh. Technol.*, to be published. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/8272491/
[48] 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.
[49] *Car Dimensions of any Make and Model in the European Market*. Accessed: Sep. 24, 2017. [Online]. Available: http://www.automobiledimension.com/
[50] *Channel Model for Frequency Spectrum Above 6 GHz (Release 14)*, document 3GPP TR 38.900 V14.3.1, 3GPP, 2017.

Vitaly Petrov received the M.Sc. degree in information systems security from the Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia, in 2011, and the M.Sc. degree in communications engineering from the Tampere University of Technology, Tampere, Finland, in 2014, where he is currently pursuing the Ph.D. degree. He was a Visiting Scholar with the Georgia Institute of Technology, Atlanta, GA, USA, in 2014, and a Strategic Intern with the Nokia Research Center, Helsinki, Finland, in 2012. His current research interests include Internet-of-Things, mmWave/THz band communications, nanonetworks, cryptology, and network security. He was a recipient of the Best Student Paper Award at the IEEE VTC-Fall'15 and the Best Student Poster Award at the IEEE WCNC'17.

Maria A. Lema is currently the Technical Project Manager for the 5G UK Testbeds and Trials project and the lead Researcher for the Ericsson-sponsored 5G Tactile Internet project with the Centre for Telecommunications Research, King's College London. She has been heavily involved in the definition of use cases for 5G, working together with various industry verticals and other telecom players to identify the main requirements and challenges to successfully bring 5G to market. She has been driving different demonstrations with key industrial partners that show the advantages of vertical integration and partnerships in 5G. Most of her research in this area has been focused on designing end-to-end utra-reliable low latency 5G networks in the context of the Internet of Skills.

Margarita Gapeyenko received the B.Sc. degree in radio-engineering, electronics, and telecommunications from Karaganda State Technical University, Kazakhstan, in 2012, and the M.Sc. degree in telecommunication engineering from the University of Vaasa, Finland, in 2014. She is currently pursuing the Ph.D. degree with the Laboratory of Electronics and Communications Engineering, Tampere University of Technology, Finland. Her research interests include mathematical analysis, performance evaluation, and optimization methods of future wireless networks, device-to-device communication, and 5G-grade heterogeneous networks.

Konstantinos Antonakoglou received the B.Sc. degree in electronics engineering from the Technological Educational Institute of Piraeus and the M.Sc. degree in electronic automation from the National and Kapodistrian University of Athens. He is currently pursuing the Ph.D. degree with the Department of Informatics, King's College London, funded by the EPSRC Grant. His current research interests include haptic communication over 5G networks, and the quality of service and the quality of experience in mobile networks for multi-modal media.

Dmitri Moltchanov received the M.Sc. and Cand.Sc. degrees from the Saint Petersburg State University of Telecommunications, Russia, in 2000 and 2002, respectively, and the Ph.D. degree from the Tampere University of Technology in 2006. He is currently a Senior Research Scientist with the Laboratory of Electronics and Communications Engineering, Tampere University of Technology, Finland. He authored over 50 publications. His research interests include the performance evaluation and optimization issues of wired and wireless IP networks, 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.

Fragkiskos Sardis received the M.Sc. degree in computer networks in 2010 and the Ph.D. degree in mobile cloud computing and edge caching. He is currently a Research Associate with King's College London, where he is involved in 5G infrastructures in the context of the EU H2020 Project VirtuWind and the 5GUK Project. His areas of interest include cloud computing, edge caching, software defined networks, radio access and core networks and haptic applications. In 2011, he received the Academic Excellence Scholarship for his Ph.D. degree.

Andrey Samuylov received the M.Sc. degree in applied mathematics and the Cand.Sc. degree in physics and mathematics from RUDN University, Russia, in 2012 and 2015, respectively. Since 2015, he has been with the Tampere University of Technology as a Researcher, where he is involved in the analytical performance analysis of various 5G wireless networks technologies. His research interests include P2P networks performance analysis, performance evaluation of wireless networks with enabled D2D communications, and mmWave band communications.

Sergey Andreev received the Specialist and Cand.Sc. degrees from the Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia, in 2006 and 2009, respectively, and the Ph.D. degree from the Tampere University of Technology, Finland, in 2012. He is currently a Senior Research Scientist with the Laboratory of Electronics and Communications Engineering, Tampere University of Technology. Since 2018, he has also been a Visiting Senior Research Fellow with the Centre for Telecommunications Research, King's College London. He has authored or 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, in 2004. He is currently a Professor with the Laboratory of Electronics and Communications Engineering, Tampere University of Technology, Finland. He is the author of 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 *IEEE Communications Magazine* and an Editor of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.

Mischa Dohler (F'14) is currently a Full Professor in wireless communications with King's College London, driving cross-disciplinary research and innovation in technology, sciences and arts. He is also the Director of the Centre for Telecommunications Research, a Co-Founder of the pioneering smart city company WorldSensing, a fellow of the Royal Academy of Engineering and the Royal Society of Arts, and a Distinguished Member of Harvard Square Leaders Excellence. He acts as a Policy, Technology, and Entrepreneurship Adviser. He has over 200 highly-cited publications and authored several books. He has pioneered several research fields, contributed to numerous wireless broadband, IoT/M2M and cyber security standards. He holds a dozen patents. He has received numerous awards. He organized and chaired numerous conferences. He was the Editor-in-Chief of two journals. He has talked twice at TEDx. He had coverage by national and international TV and radio, and his contributions have featured on the BBC, The Wall Street Journal, and many others. He is a frequent keynote, panel, and tutorial speaker.