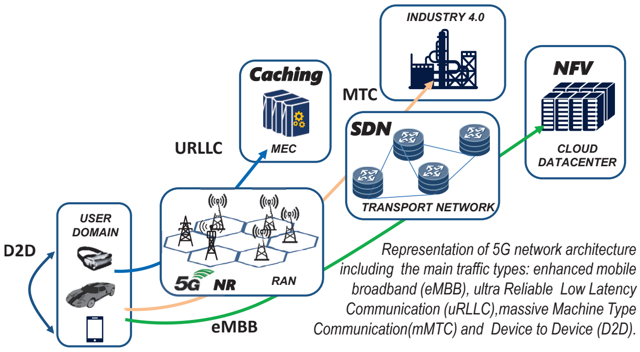
On Guaranteeing End-to-End Network Slice Latency Constraints in 5G Networks

Lanfranco Zanzi, Vincenzo Sciancalepore
NEC Laboratories Europe GmbH, Germany
{name.surname}@neclab.eu

***Abstract*—The upcoming 5th generation (5G) of mobile networks is being designed to significantly improve the performance of the current network deployments by introducing more flexibility and scalability while, at the same time, optimizing the spectrum utilization and energy efficiency of radio communications. Among such novelties, *Network Slicing* is emerging as the key-concept in the 5G landscape, able to provide the means for the concurrent deployment of heterogeneous services over a common physical network. In this paper, we investigate current technologies, open issues and possible solutions while addressing the most critical requirement envisioned with the advent of advanced services, i.e., the provisioning of stringent end-to-end delay guarantees as a pillar of the novel Ultra Reliable and Low Latency Communication (URLLC) service type.**

***Index terms*—** 5G, Network Slicing, URLLC, End-to-end delay, Low latency, SDN, NFV.



Representation of 5G network architecture including the main traffic types: enhanced mobile broadband (eMBB), ultra Reliable Low Latency Communication (uRLLC),massive Machine Type Communication(mMTC) and Device to Device (D2D).

Fig. 1. Schematic view of the 5G network architecture

I. Introduction

The new generation of mobile networks (5G) brings together novel advanced services to offer a solid user experience, such as tactile internet, high resolution (4K) video streaming, advanced sensing and monitoring, autonomous driving. However, the need to provide seamless integration of such services on a common infrastructure hardly fits with the current capabilities of mobile network deployments, both in terms of throughput and latency guarantees.

In this context, the *network slicing* paradigm is getting ahead as the main key-enabler for the coexistence, over the same physical premises, of heterogeneous services with a wide set of different requirements. In particular, the telecom operator may decide to properly "slice" its own infrastructure so as to offer service-tailored (and isolated) network slice to independent network tenants, e.g., Mobile Virtual Network Operators (MVNOs), vertical industries, Over-The-Top service providers (OTTs), in a flexible and dynamic manner. The network slice includes a set of radio and transport resources depending on the specific requirements as well as fine tuning of computing capabilities so that the ad-hoc applications can run as virtualized functions over cloud premises providing the service functionalities [1].

The ever-increasing throughput requirements, e.g., those envisioned for enhanced/extreme mobile broadband (e/xMBB) services targeting peak data rates up to 10 or 20 Gbps, can be properly handled by means of an efficient exploitation of both licensed and non-licensed spectrum [2]. In addition, novel technologies such as carrier aggregation (CA), mm-Wave, massive multiple-input multiple-output (mMIMO) and advanced channel coding schemes are necessary to provide enhancements in the spatial and spectrum efficiency of the system. While these enhancements appear to be technologically addressed, the need of low latency communications exacerbates the design of 5G networks.

Latency-sensitive services such as autonomous driving and tactile internet applications demand for extremely high reliability, availability and security that can not be ensured in today's networks. The packet-based mobile networks are able to guarantee scalability, but once data is encapsulated and sent through the transport network, failures, traffic congestions or different routing policies outside the operator domain may introduce unexpected delay that, finally, impairs the overall performance.

The main enhancements are required across the Radio Access Network (RAN) and transport network in order to deal with such unpredictable behaviors while guaranteeing user plane latency of few milliseconds in both uplink and downlink directions [3]. This turns into a novel service class, namely Ultra Reliable Low Latency Communications (URLLC) that requires novel concepts such as Software Define Network (SDN), Network Virtualized Function (NFV), Multi-access Edge Computing (MEC) and caching solutions, as shown in Fig. 1. The main advantages of these technologies are represented by a better utilization of physical resources as well as a dynamic displacement of software functionalities in standalone hardware. To support an easy function-deployment, network programmability and softwarization paradigms come to help while the radio access part still represents the bottleneck of the new communication characteristics. In this paper, we

978-1-5386-5005-9/18/$31.00 ©2018 IEEE

shed the light on the main challenges and potential solutions that aim at reducing this gap while providing reasonable and realistic implementation details for the current standardization roadmap.

The remainder of this paper is structured as follows: Section II introduces the main improvements in the management of radio resources necessary to cope with 5G requirements. Section III discusses and analyzes the achievable benefits in terms of lower latency and reliability by introducing SDN and NFV paradigms into the mobile network framework. Section IV provides an overview of caching and Mobile-Edge-Computing (MEC) solutions aimed at improving the Quality of Service (QoS) and Quality of Experience (QoE) of 5G-services end-users. Finally, Section V concludes the paper.

II. Radio Access Network Enhancements and 5G New Radio

In the Long-Term Evolution (LTE or 4G) systems, the radio frame structure has been designed to find a trade-off between latency and reliability so as to improve bandwidth performance with respect to older mobile network generations. This has been achieved with the introduction of Orthogonal Frequency Division Multiplexing (OFDM) techniques in the wireless medium access.

The radio structure is summarized in the following. Each frame has duration of 10 ms and is divided into smaller time transmission intervals (TTIs) of 1 ms, respectively. Each TTI is composed of 14 orthogonal frequency-division multiplexing (OFDM) symbols carrying data and control plane information spread over 12 consecutive subcarriers spaced of $\Delta f = 15$KHz [4]. Control plane activities represent a significant overhead in LTE systems. The achievable bit rate during the communication depends on modulation and coding scheme, which adapts to the current physical channel conditions. This is done by means of a constant monitoring of pilot symbols sent through dedicated frequency carriers over the control channels. For example, in a 5 MHz bandwidth system with 2x2 Multiple-Input Multiple-Output (MIMO) setup, downlink signal uses 4 reference symbols in every third subcarrier, resulting in about 5% of transmission overhead.

Also, the Physical Downlink Control Channel (PDCCH) takes from 1 to 3 symbols out of 14 in each subframe, resulting (on average) in 18% of overhead. The total count including all the other control channel activities (e.g., synchronization signalling, broadcasting, etc.) sums up to about 25% of radio resources utilization just for control purposes [5]. Moreover, while scheduling of radio resources has 1 ms granularity the re-transmission of packets in case of failures during the wireless communication typically occurs in 8 ms with common Hybrid Automatic-Repeat-Request (HARQ) protocol setup [6]. The definition of novel mechanisms for more reliable transmission with less communication overhead are fundamental to avoid service interruption and assure a major resource utilization.

It is safe to assert that from the radio communication point of view, these performance metrics do not meet the requirements of delay-sensitive services, as those envisioned for the 5G era and summarized in Table I. Moreover, from the intra-slice perspective, this framework lacks flexibility and solutions to provide prioritized dispatching of critical traffic over the non-critical one.

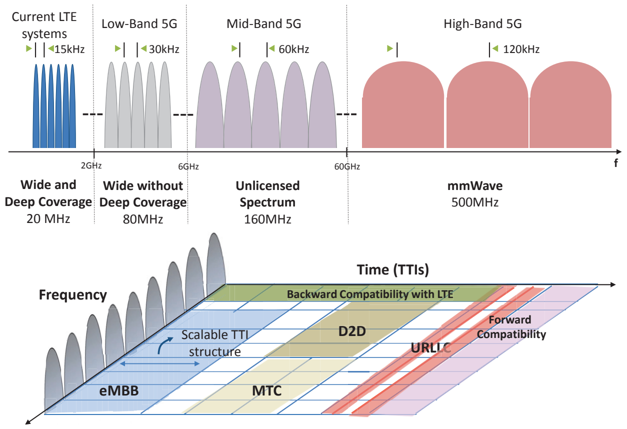


Fig. 2. Representation of the main novelties in the 5G New Radio (NR) spectrum usage.

To overcome these limitations a new physical air interface, namely *5G New Radio* (NR), is currently under definition. Fig. 2 shows the 5G NR radio structure and the main novelties introduced. First, it should be noticed that with respect to current mobile systems the new 5G radio interface will use a wider set of frequencies (up to 60 GHz), and a wider channel bandwidth (up to 1GHz). The overall resource availability greatly increases, and different portion of the time-spectrum grid can be simultaneously allocated to heterogeneous traffic kinds. In general, shorter TTIs enable shorter processing time, i.e., lower latency during the communication [7]. One approach suggests expanding the subcarrier spacing Δf from 15KHz to 30KHz, reducing $T_{OFDM} = 1/\Delta f$ from $66.67\mu s$ to $33.33\mu s$. For backward compatibility with the LTE-A systems, the frame period $T = 10$ ms can be divided into 40 subframes, each one with duration $250\mu s$, keeping almost unchanged the sampling frequency [8]. With the same calculations, in case of $\Delta f = 60$KHz the symbol duration can be further reduced to $16.67\mu s$, with the current LTE frame duration possibly splitted into 80 *mini-slots* of duration $125\mu s$ each.

In this way, the fixed frame structure of LTE can be upgraded in favor of more a flexible one, which supports scheduling of resources with a smaller time granularity, i.e., allocation of a variable number of TTIs or even a fractional part of them [9], [10]. With smaller and scalable time transmission intervals, new and optimized OFDM-based waveforms and advanced modulation and coding schemes, the 5G NR is filling up the gap of provisioning of throughput and low-latency requirements.

Flexible resource allocation and isolation among multiple slices are additional challenges that must be tackled to enable network slicing [11]. Unlike wired networks where the environment is hard to change, the resource allocation mechanism in wireless scenario is more challenging due to interference,

TABLE I
DIFFERENT 5G USE CASES WITH VERY STRINGENT LATENCY REQUIREMENTS

Use Case	Latency (ms)	Use Case Description
Virtual Reality (VR)	1	The users interact with an artificial environment provided by a computer. The environment can be experienced through visual and haptic sensory stimuli as if the user belongs to the same scenario. VR represents an attractive solution for the commercial entertainment sector.
Augmented Reality (AR)	$\leqslant 10$	The users experience an enhanced version of reality created by the use of this technology. It represents a promising solution in safety scenarios, where additional information into the users field of view allows to take decision and act in a safer way.
eHealth	1-10	Health-care practice supported by electronic processes and remote communication schemes based on real-time control and feedback loops.
Smart Factory	5	Factory automation with real-time control of production machine and system with very limited human involvement.
Automotive	10	Autonomous driving, road safety, traffic control and optimization for future intelligent transport system.
Remote Control	$\leqslant$10	Remote controlled robots with haptic feedback for diverse applications in wide set of sectors such as building sector and safety scenarios.

user mobility, and radio channel variability. These two aspects are actually coupled. If resource allocation is not performed correctly, changes in one slice may influence the others, thus leading to an overall degradation of the quality of service in the system, obviously with greater impact on URLLC traffic.

A network slice instance could be isolated from another network slice instance in several ways, e.g., full or partial isolation and logical or physical isolation [12], and each setup has proper advantages or disadvantages when compared with the others. It is very hard to find a general scheme suitable in all the scenarios and for every network deployment. Each physical network infrastructure has operator, vendor and technological peculiarities that must be addressed and optimized case by case according to the necessary level of isolation.

III. TRANSPORT NETWORK ENHANCEMENTS

We previously discussed lower layers improvements as demanded by novel 5G use cases. However, the deployment of multiple end-to-end network slices also requires additional enhancements in the network management capabilities. Each network slice can be internally orchestrated by tenants as an independent and isolated environment, so that the underlying topology must be flexible enough to accommodate different technical and business exigencies [13].

Current delay-sensitive services, e.g., those relative to security applications, often have dedicated network (hardware and software) components able to guarantee reliability and assure delay upper bounds during the communication among different nodes. Despite the positive resulting performance, this scenario does not scale in wide networks and results in a significant increase of costs from both the deployment and management viewpoints.

With the advent of Software Defined Networking (SDN) this is no longer an issue. The possibility to have a global view of the network exploiting the well-known centralized management system provides significant advantages when compared to traditional networks. The SDN Controller has almost complete control of the data paths and thus does not have to compete with other control plane elements, which simplifies scheduling and resource allocation over the network. Appropriate rules can be simply *pushed down* to the different nodes of the system when new network slices are demanded, while enforcing bandwidth and resource allocation setups to meet the traffic requirements.

Unfortunately, a number of issues arises when theory becomes practice. For example, the orchestration of a new network slice implies the definition a feasible path into the network with an adequate resource availability. In case of URLLC traffic, such a route must also provide delay guarantees without affecting other instances already deployed. As highlighted in [14], existing SDN systems can reason only about bandwidth and/or the number of hops in the network, without the possibility to build routing strategies based on delay parameters.

To solve these issues, the same authors proposed a heuristic approach aimed at optimizing a multi-constraint problem based on the reservation for each delay-critical flow of one queue per switch port. Their approach allows to almost avoid the introduction of queue and buffering delays, but with the expenses of over-provisioning the resource reservation. Despite modern hardware routers and switches can provide significant number of ports and queues, with this solution only one queue per port can be used to admit a slice request, thus binding the maximum number of slices to the hardware capabilities. To overcome this problem, multiple physical ports can be logically combined to form a single virtual port increasing the number of available queues. Another approach includes the multiplexing of more than one traffic flow, i.e., slice, per queue. For this last option to be feasible, advanced mechanisms for monitoring and validation of end-to-end latency requirements must be in place to avoid the introduction of queue delay and, in turn, the resulting service degradation.

A. NFV and Reliability aspects

To offer cost-efficient, scalable and flexible provisioning of network slices services, Network Functions Virtualization (NFV) paradigm may be leveraged to deliver added-value services as a chain of Virtual Network Functions (VNFs).

Traditional network functions are executed and installed in middleboxes, usually deployed within the data center premises. In the 5G era, these hardware modules are substituted by flexible and easy-to-maintain software instances hosted in commodity physical machines such as servers, storage nodes or router/switches, providing operational cost reduction and paving the road to automatic network reconfiguration.

NFV architecture allows to execute network functions on commodity servers regardless the real location of the physical machines, relying on SDN capabilities to manage the network flows. It is clear that virtual functions placement has an impact on reliability and performance of the network. A failure (hardware or software) of any VNF of a service chain leads to break down the entire chain, causing significant delays and resource wastage as well as possible interruption for other co-hosted applications. To achieve reliability targets, the VNF placement process usually deals with redundant deployments over different physical host in a static or dynamic manner.

As highlighted in [15], once selected the subset of VNFs necessary to support the network service and defined the placement strategy for the backups, traffic flows will also pass through those duplicates, practically resulting in a longer service's chain deployment. The additional routing across different copies of the same VNF will impact of the experienced end-to-end delay leading to a fundamental trade-off. Where lower latency is expected, shorter chains should be deployed while for service reliability higher degree of VNF redundancy is demanded.

The same authors proposed a joint optimization framework to solve this issue based on an iterative backup selection procedure with routing implementation. Their solution investigates the feasibility of the service deployment and incrementally provisions the primary VNFs with an adequate number of backup VNFs accounting for delay requirements. Differently from other existing reliability-aware schemes, this solution distributes backup VNFs over multiple paths to avoid the introduction of additional delays, but at the expenses of an increase in the overall bandwidth consumption.

IV. MEC and Caching Solutions

It is widely accepted the inability of providing very demanding end-to-end delay guarantees in legacy mobile network systems composed by RAN, transport and core domains. Although advanced optimization schemes for radio resource scheduling [16] and transport resource allocation [17] of slices have been proposed in the literature to cope with delay guarantees, the main drawback lies in the fact that in today's networks most of the services run outside the mobile operator domain, where advanced solutions for the dispatching of delay-sensitive traffic can not be applied.

The need to move services closer to end-users where higher bandwidth and lower latency are available, led to the birth of the Mobile Edge Computing (MEC) concept. MEC is envisioned as one of the most promising solution in the 5G landscape, able to bring for the first time cloud computing capabilities at the edge of the network [18]. The European Telecommunications Standards Institute (ETSI) started the specification of MEC platforms as an operator-owned system, where MEC *providers* allow third party entities, namely MEC tenants, to install their applications on a shared platform. Each MEC host is associated with a set of Base Stations (BSs). Such BSs build the so-called *MEC coverage area*. Mobile users in the proximity of this area might access the services through a prioritized path, which are affected almost uniquely by the delay introduced in the wireless channel and the one introduced by the fronthaul links between the MEC host and the base stations.

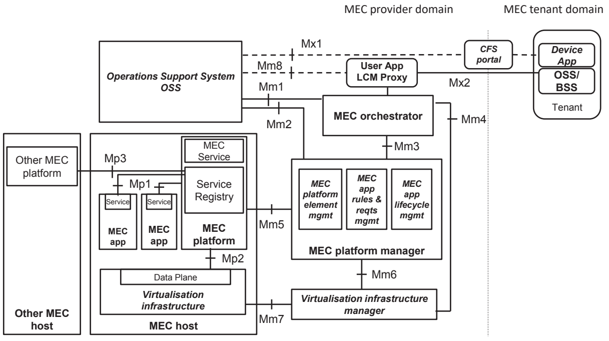


Fig. 3. Simplified ETSI Mobile Edge Computing (MEC) architecture [19].

A simplified description of the ETSI MEC logical architecture is depicted in Fig. 3. It can be noticed the clear separation of provider and tenant domains, with the Customer Facing Service (CFS) only point of interaction among the two entities. In this *Platform-as-a-Service* kind of business, the MEC provider has all the administrative privileges. He receives from the tenant the descriptor files of the service to be deployed and manages the enforcement of traffic policies in order provide adequate quality of service to the tenants [19].

The MEC Orchestrator entity deals with requests of instantiation or termination of software applications and instruct the MEC platform manager to fulfill them. The authors of [20] enhanced this framework with the introduction of a management entity aimed at extending the network slicing concept to MEC premises. The presented orchestration solution is corroborated by simulation and the impact of stringent delay requirements on the consolidation capabilities of the system is also investigated. The outcomes show that provide high bandwidth and real-time access capabilities decreases the possibility to implement profitable functional placement and load-balancing issues may arise in the system. These aspects must be considered during the admission and control phase in order to avoid Service Level Agreements (SLAs) degradation.

In the MEC and network slicing context end-user mobility raises a set of issues to be addressed and the definition of appropriate mechanisms for the handover management is still an open challenge. During the handover procedure, user equipment can not exchange data with any of the base stations till the end of the cell re-selection and the resource release process at the source eNodeB. In this period of time, no service or application can be provided. The soft-handover procedure in LTE requires from 10 ms to 15 ms in both TDD and FDD operational mode as well as a significant level of synchronization among the base station [21]. These numbers are not deterministic and greatly influenced by traffic congestions in the network. The access to the shared medium may be delayed in case of exponential back-off periods due to multiple and concurrent trials coming from other mobile users. In case users move from/to a different MEC coverage area a hard-handover process must relocate the application and user context to another MEC host, trying to avoid service interruption and quality of service degradation. This task is particularly challenging if the target application is not yet available in the destination host and migration or re-instantiation of virtual machines are involved.

On the one hand, migration techniques assure application states conservation and service continuity, but requires additional time to move and rebuild the instance impacting on the user experience. On the other hand, re-instantiation techniques provide much faster but stateless application handover [22], thus breaking the service provisioning.

A. Smart caching

The MEC platform provides not only the possibility to directly host application and services but also to store contents. In this way, most of the issues with unpredictable network behaviours are solved. Delay-sensitive flows can be offloaded and processed at the edge of the network, avoiding time consuming routing activities towards the end of the network. These peculiar characteristics are the reason why, with the advent of Mobile Edge Computing, a new wave of caching solutions started. The idea is to take advantage of the unique location of the MEC entity to provide added-value services to the end-users.

Proactive caching of contents to provide faster access at commonly used information is a well-known concept that has been exhaustively addressed in the literature under the name of Content Delivery Networks (CDNs). Benefits of CDNs include reducing backbone bandwidth utilization and costs, improving time to load and access contents, and increasing global availability of multimedia files. CDN nodes are usually deployed in correspondence of strategic points of the network, for example where multiple backbones links are available or in highly populated areas where many users can be reached and CDN advantages exalted.

Advanced caching mechanisms have been proposed in the context of 5G with the objective of reduce the traffic load of macro-BSs in case of unavailability of high-speed backhaul connection toward the cloud.

The cache placement problem starts with the definition of which content store in the BSs premises based on user requests. Typical solutions include centralize edge caching wherein a coordinator with access to all the information about storage capacity, user connectivity, etc. schedules the content delivery, and distributed edge caching in which small-cells BSs and macro-BSs coordinate their transmissions such that multimedia content at the small-cells can be delivered with lower latency.

As caching of contents becomes more and more important for next generation wireless networks, it increases the need to understand and solve the trade-off between the storage capacity requirements and lower-latency achievements. Different metrics can be defined to investigate this issue such as normalized delivery time (NDT) and delivery time per bit (DTB). NDT represents the worst-case file delivery normalized to a reference interference-free system with unlimited caching capabilities. NDT has been defined in [23] to investigate the trade-off between latency and the cache storage capacity of the edge nodes. The same authors in [24] and [25] investigated the total delivery latency over fronthaul and wireless link in a Fog RAN environment under different settings, using NDT to derive upper and lower bounds of delivery latency as a function of cache and fronthaul resources. DTB has been defined in [26] and represents the ratio between time of transmission and the file size in bits. The authors investigated and characterized the minimal delivery latency for a reference system architecture with an information-theoretic model, also analyzing the potential degradation of service delivery due to multiple small-cells in a binary fading one-sided interference downlink channel.

V. Conclusion

A compelling increase of throughput and connectivity demand with novel vertical use-cases give raise to unprecedented ultra reliable and low-latency traffic characteristics. To handle their design challenges, the fifth generation of mobile networks (5G) aims at re-designing the mobile network concept following novel paradigms such as network softwarization and programmability. In this context, network slicing has been elected as the key- enhancement to provide guaranteed traffic performance for a heterogeneous set of services over virtual networks sharing the same physical infrastructure.

In this paper, we have gathered and organized the main solutions in the literature according to different network domains shedding the light on the stringent end-to-end delay requirements. We have provided a comprehensive analysis of the network slice enablers and identified the pillars of the novel radio frame structure that might allow an easy-to-deploy solution focusing on low-latency services. In addition, we have showcase the SDN/NFV paradigm and the MEC concept to further support limited (and reasonable) delay characteristics across the transport and core/cloud domains.

Acknowledgments

The research leading to these results has been supported by the H2020-MSCA-ITN-2015 research and innovation program under grant agreement number 675806 (5G AURA).

References

[1] Lingen, F. V. *et al.*, “The Unavoidable Convergence of NFV, 5G, and Fog: A Model-Driven Approach to Bridge Cloud and Edge,” *IEEE Communications Magazine*, vol. 55, pp. 28–35, Aug. 2017.

[2] Third Generation Partnership Project (3GPP), "Study on Scenarios and Requirements for Next Generation Access Technologies (Rel.14) ," 3GPP TR 38.913, May 2017.

[3] Joachim Sachs, T. D. R. B. and Kittichokechai, K., "5G Radio Network Design for Ultra-Reliable Low-Latency Communication," *IEEE Network*, vol. 32, pp. 24–31, Aug. 2017.

[4] Third Generation Partnership Project (3GPP), "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation (Rel. 15)," 3GPP TS 36.211, Mar. 2018.

[5] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Rel. 15)," 3GPP TS 36.213, Apr. 2018.

[6] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification," 3GPP TS 36.321, Apr. 2015.

[7] Takeda, K., Wang, L. H., and Nagata, S., "Latency Reduction toward 5G ," *IEEE Wireless Communications Magazine*, vol. 52, no. 11, pp. 65–75, Jun. 2017.

[8] Parvez, I., Rahmati, A., Guvenc, I., Sarwat, A. I., and Dai, H., "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," 2017. [Online]. Available: {https://arxiv.org/abs/1708.02562}

[9] Pedersen, K. I., Berardinelli, G., Frederiksen, F., Mogensen, P., and Szufarska, A., "A flexible 5G frame structure design for frequency division duplex cases," in *IEEE Communications Magazine*, vol. 54, no. 3, Mar. 2016, pp. 53–59.

[10] Fountoulakis, E., Pappas, N., Liao, Q., Suryaprakash, V., and Yuan, D., "An Examination of the Benefits of Scalable TTI for Heterogeneous Traffic Management in 5G Networks ," in *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2017 15th International Symposium on*, Apr. 2017.

[11] Sciancalepore, V., Zanzi, L., Costa-Perez, X., and Capone, A., "ONETS: Online Network Slice Broker From Theory to Practice," Jan. 2018. [Online]. Available: https://arxiv.org/abs/1801.03484

[12] Kotulski, Z., Nowak, T., Sepczuk, M. *et al.*, "On end-to-end approach for slice isolation in 5G networks. Fundamental challenges," in *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*, Sep. 2017.

[13] Zanzi, L., Sciancalepore, V., Garcia-Saavedra, A., and Costa-Perez, X., "OVNES: Demonstrating 5G Network Slicing Overbooking on Real Deployments," in *The 37th Annual IEEE International Conference on Computer Communications (INFOCOM 2018)*, Apr. 2018.

[14] Kumary, R., Hasany, M., Padhyy, S., Evchenkoy, K., Piramanayagamk, L., Mohany, S., and Bobbax, R. B., "End-to-End Network Delay Guarantees for Real-Time Systems using SDN," in *IEEE Real-Time Systems Symposium*, Dec. 2017.

[15] Qu, L., Assi, C., Shaban, K., and Khabbaz, M. J., "A Reliability-Aware Network Service Chain Provisioning With Delay Guarantees in NFV-Enabled Enterprise Datacenter," *IEEE Transaction on Network and Service Management*, vol. 14, no. 3, pp. 554–568, Sep. 2017.

[16] Foukas, X., Nikaein, N., Kassem, M. M., Marina, M. K., and Kontovasilis, K., "FlexRAN: A Flexible and Programmable Platform for Software-Defined Radio Access Networks ," in *Proceedings of the 12th ACM CoNEXT*. ACM, 2016.

[17] Leconte, M., Paschos, G., Mertikopoulos, P., and Kozat, U., "A Resource Allocation Framework for Network Slicing," in *IEEE International Conference on Computer Communications (INFOCOM 2018)*, Apr. 2018.

[18] Giust, F., Sciancelpore, V. *et al.*, "Multi-access Edge Computing: The driver behind the wheel of 5G-connected cars," *IEEE Communications Standards Magazine*, Mar. 2018.

[19] ETSI MEC ISG, "Mobile Edge Computing (MEC); Framework and reference architecture," ETSI, DGS MEC 003, Apr. 2016.

[20] Zanzi, L., Giust, F., and Sciancalepore, V., "M2EC: A Multi-tenant Resource Orchestration in Multi-access Edge Computing Systems," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2018.

[21] Singhal, D., Kunapareddy, M., Chetlapalli, V., James, V. B., and Akhtar, N., "LTE-Advanced: Handover Interruption Time Analysis for IMT-A Evaluation," in *Proc. of 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN)*, Jul. 2011.

[22] Hassan Hawilo, M. J. and Shami, A., "Orchestrating Network Function Virtualization Platform: Migration or Re-Instantiation? ," in *Cloud Networking (CloudNet), 2017 IEEE 6th International Conference on*, Sep. 2017.

[23] Sengupta, A., Tandon, R., and Simeone, O., "Cache aided wireless networks: Tradeoffs between storage and latency," in *Proc. Annual Conference on Information Science and Systems (CISS)*, Mar. 2016, pp. 320–325.

[24] Koh, J., Simeone, O., Tandon, R., and Kang, J., "Cloud-aided edge caching with wireless multicast fronthauling in fog radio access networks," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, Mar. 2017, pp. 1–6.

[25] Sengupta, A., Tandon, R., and Simeone, O., "Pipelined Fronthaul-Edge Content Delivery in Fog Radio Access Networks," in *Proc. IEEE Globecom Workshop (GC Wkshps)*, Dec. 2016, pp. 1–6.

[26] S. M. Azimi, R. T., "Fundamental Limits on Latency in Small-Cell Caching Systems: An Information-Theoretic Analysis," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016.