©2020 Elsevier. This manuscript version is made available under the CC-BY-NC-ND 4.0 license
http://creativecommons.org/licenses/by-nc-nd/4.0/

Are Mobility Management Solutions Ready for 5G and Beyond?☆

Akshay Jain[1,*], Elena Lopez-Aguilera[1], Ilker Demirkol[1]

[1] *Department of Network Engineering, Universitat Politècnica de Catalunya BarcelonaTECH, Barcelona 08034, Spain.*

Abstract

Enabling users to move to different geographical locations within a network and still be able to maintain their connectivity and most essentially, continuity of service, is what makes any wireless network ubiquitous. Whilst challenging, modern day wireless networks, such as 3GPP-LTE, provision satisfactory mobility management (MM) performance. However, it is estimated that the number of mobile subscriptions will approximately touch 9 billion and the amount of data traffic will expand by 5 times in 2024 as compared to 2018. Further, it is expected that this trend of exponential growth will be maintained well into the future. To cope with such an exponential increase in cellular traffic and users alongside a burgeoning demand for higher Quality of Service (QoS), the future networks are expected to be highly dense and heterogeneous. This will severely challenge the existing MM solutions and ultimately render them ineffective as they will not be able to provide the required reliability, flexibility, and scalability. Consequently, to serve the 5G and beyond 5G networks, a new perspective to MM is required. Hence, in this article we present a novel discussion of the functional requirements from MM strategies for these networks. We then provide a detailed discussion on whether the existing mechanisms conceived by standardization bodies such as IEEE, IETF, 3GPP (including the newly defined 5G standards) and ITU, and other academic and industrial research efforts meet these requirements. We accomplish this via a novel qualitative assessment, wherein we evaluate each of the discussed mechanisms on their ability to satisfy the reliability, flexibility and scalability criteria for future MM strategies. We then present a study detailing the research challenges that exist in the design and implementation of MM strategies for 5G and beyond networks. Further, we chart out the potential MM solutions and the associated capabilities they offer to tackle the persistent challenges. We conclude this paper with a vision for the 5G and beyond MM mechanisms.

Keywords: 5G, Beyond 5G, 6G, Mobility Management, SDN, Meta-Surfaces.

1. Introduction

Future wireless networks define a very challenging environment for mobility management (MM) solutions, due to the significant increase in density (in terms of both users and deployed access points), in heterogeneity (given the various radio access technologies (RATs) supported), as well as in programmability (the network as well as the environment can be programmable). To achieve an ubiquitous network service in such challenging environments, it is critical to devise effective MM strategies that facilitate seamless mobility by allowing users to traverse through the network without losing connectivity and service continuity.

One of the traditional approaches for allowing applications to serve a user in mobile scenarios has been to maintain network connectivity through handovers based on criteria such as Radio Signal Strength Indicator (RSSI), Signal to Interference and Noise Ratio (SINR), Reference Signal Received Quality (RSRQ), Reference Signal Received Power (RSRP), etc. However, in addition to the signal quality parameter centric handovers, modern day applications necessitate that other parameters such as available core network bandwidth, End-to-End (E2E) latency, backhaul bandwidth and backhaul reliability [1] are also taken into consideration. Moreover, maintaining Quality of Service (QoS), e.g., provisioning service continuity, link continuity, required bit-rate and latency, during mobility scenarios has been one of the primary objectives for novel MM mechanisms. Multiple strategies to satisfy such QoS criteria such as service migration [2], service replication [3], path reconfiguration [4], etc., have been proposed by the research community. MM solutions for 5G and beyond networks are also expected to ensure E2E connectivity and session continuity through the maintenance/preservation of IP address of the user towards the core network entity that provisions the service for the corresponding user.

To motivate further, we consider an illustrative example of the future mobility scenario as presented in Figure 1. It shows the extraordinary nature of complexity that the future networks will present for MM. As shown in

☆This work has been supported in part by the EU Horizon 2020 research and innovation programme under grant agreement No. 675806 (5GAuRA), and by the ERDF and the Spanish Government through project RYC-2013-13029.

*Corresponding author
Email address: akshay.jain@upc.edu (Akshay Jain)

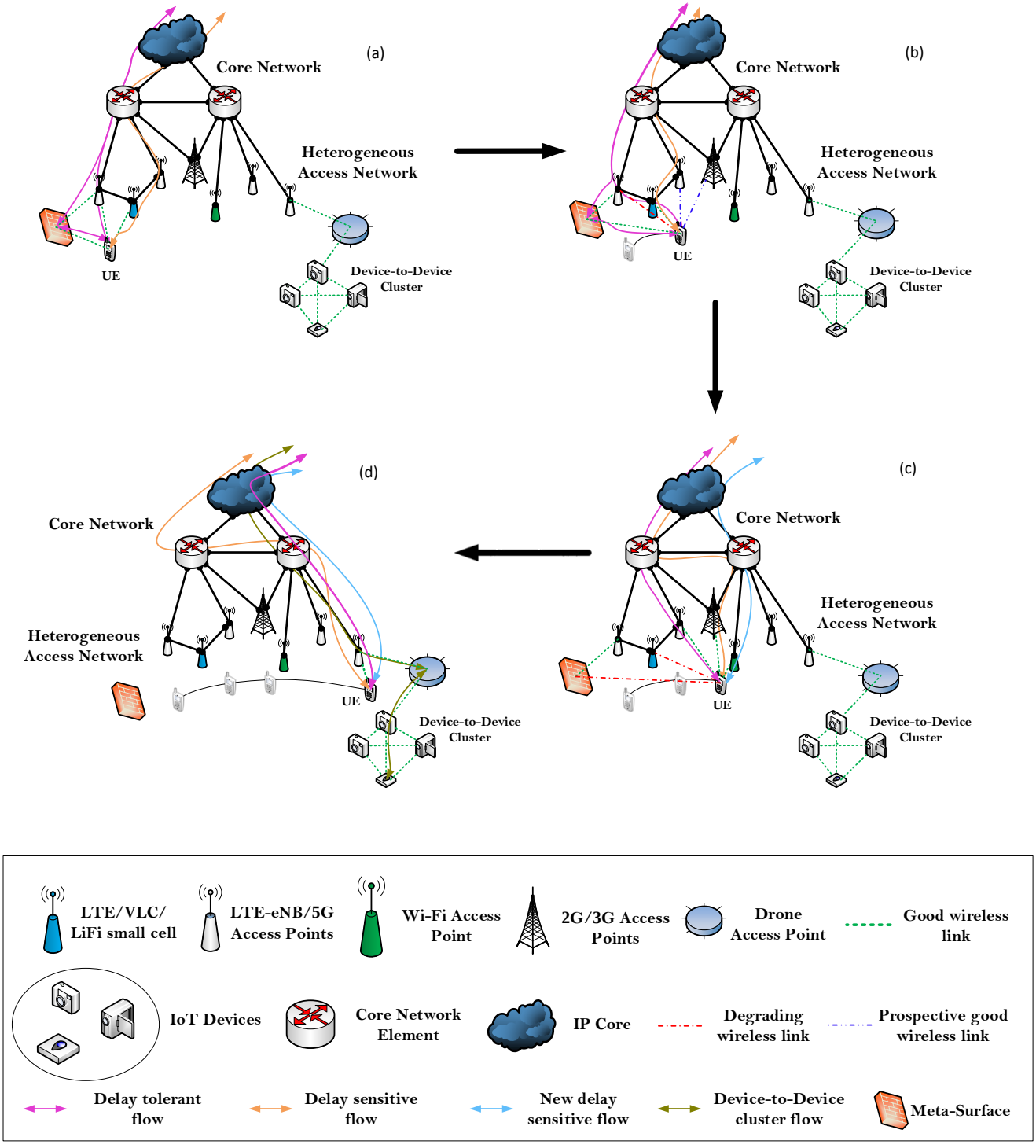

Figure 1: An illustrative 5G and beyond network mobility scenario.

Figure 1(a), a mobile user equipment (UE) is connected to multiple RATs (5G Access Point (AP)/ Long Term Evolution (LTE) eNode B (eNB)/visible light communications (VLC) and Light Fidelity (LiFi) small cells [5–7], etc.), while having a delay tolerant and a delay sensitive application datastream (flows) with distinct QoS profiles. Also, the AP through which the delay tolerant flow is being served to the user has a good wireless link with a meta-surface in the vicinity. While, traditionally the environment between the user and an AP is considered as an adversary in all the generations of mobile communications, including 5G, in beyond 5G (B5G) networks the environment will be programmable and hence, will be an ally by provisioning favorable transmission channels [8–10]. These favorable channels will essentially consist of reflected signals, the phases and polarizations of which will be adjusted by thin (but electrically significant) surfaces, also known as meta-surfaces, so that they interfere constructively at the receiver [8, 9]. In addition to the meta-surfaces, future networks will also consist of mobile APs such as drones, as shown in Figure 1(a). Note that, the density of meta-surfaces and drone APs will also be extremely high in future networks. Further, in the scenario illustrated, we consider the use case wherein the drone AP is servicing a device-to-device (D2D) cluster, and connecting it to the core network through one of the ground based APs. The D2D cluster over the course of its existence does not generate packets as frequently as the other users, since the cluster devices mainly host Internet of Things (IoT) applications.

Next, in Figure 1(b), as the user moves, it starts to register wireless links with better signal quality from other APs as compared to those it is already associated to. It is imperative to state here that, the APs can be from the same or different network operators. Henceforth, a careful and efficient RAT and AP selection for each flow will be necessary as part of the future MM mechanisms. It is interesting to observe that while the AP used for serving the delay tolerant flow in Figure 1(a) no longer has a good link quality, through the meta-surfaces and their programmable nature it still has a good wireless link to the user and hence is able to serve it.

Following the new RAT/AP association, flows pertaining to the user are redirected through the most optimal path. Novel MM mechanisms that aim to service the 5G and B5G networks will require efficient route optimization methods to perform the same. Additionally, the MM mechanisms will also need to implement IP forwarding so as to ensure E2E link continuity. In Figure 1(c) we then observe that as the user moves further, the RAT/AP selection and optimal routing methods are continually implemented. Further, when a new application request is generated, as seen in Figure 1(c), an appropriate RAT and AP for the given flow is selected alongside the route that satisfies the requested QoS. Lastly, in Figure 1(d), it can be seen that alongside the user's flows, the D2D cluster's flows are also being serviced by network. However, the D2D cluster is firstly serviced by a drone AP, which then relays information to/from the ground based APs. These ground based APs assist in serving the data flows generated from the devices in the D2D cluster by relaying the data to the relevant servers in the core network.

Given the complexity of the scenario presented in Figure 1, it is evident that no single MM mechanism will form the solution to all the possible situations and scenarios that will be prevalent. And, although current MM mechanisms propose methods for careful RAT and AP selection, IP packet forwarding, route optimization, and session management, a more than 10-fold increase in user density coupled with the heterogeneity in flow types and network will extremely limit their capabilities, as explained in the subsequent sections in detail. New user applications such as Augmented Reality, Virtual Reality, Vehicle-to-Everything (V2X), etc., will present very restrictive delay requirements, exceptionally high reliability and bandwidth requirements [11], that will consequently severely challenge the capabilities of current MM strategies. Further, the radio access network (RAN) technologies themselves are expected to undergo important transformation in the future networks given the significant interest in VLC, LiFi, etc., [5, 6]. Whilst both LiFi and VLC, being TeraHertz (THz) bandwidth technologies, enable near Terabits per second (Tbps) speeds, they are significantly impaired by the environment. This consequently has significantly more detrimental effects on the user QoS during mobility scenarios, which we will discuss in further detail in the later sections.

Also, owing to the telecom operators' desire to serve more industry verticals, a new set of mobility patterns will emerge. For example, a platoon of vehicles moving coherently together, vehicles disbanding from one platoon to join another, ultra-fast moving users (in excess of 500 km/h), moving access points (such as those on drones [12]), etc., thus introducing another dimension to the MM problem. Henceforth, the ability to serve devices with mobility patterns that will be more diverse and challenging as compared to current day network scenarios, will be a significant challenge towards the design, development and deployment of 5G and beyond MM mechanisms. An additional yet significant challenge will be to manage and potentially reduce the control plane (CP) signaling load [13] due to mobility events.

Thus, a fresh perspective, wherein MM solutions are decentralized and flexible, can support multiple use cases simultaneously and account for the various other radical changes in 5G and B5G networks with reliability, is required. Note that, decentralization will permit MM mechanisms to service the exponentially increasing number of users coupled with different mobility profiles (e.g., static IoT devices and users in high-speed trains). On the other hand, flexibility will allow them to adapt to the user, network and/or environment context (e.g., QoS, user mobility profile, network load, flow types, meta-surfaces, etc.). Additionally, reliability will aid in provisioning seamless mobility as well as in satisfying the ultra-reliable criterion

for future wireless network applications.

References [14] and [15] aim to provide new MM strategies via Software Defined Networking (SDN) based MM and multi-RAT mobility. However, they do not elaborate on the myriad challenges that future MM mechanisms will encounter, such as time complexity, signaling overhead, etc. Similarly, while in [16] MM strategies, such as advanced cell association, group handovers, etc., have been discussed to address the heterogeneity in the mobility patterns and profiles that will arise in 5G, they fall short in addressing the challenges such as core network signaling, complexity, etc., that 5G and beyond MM solutions will face. Further, surveys such as [17] and [18] are restricted to the current network architecture, and hence, fail to provide a MM perspective for 5G and beyond networks. In addition, while [7] aims to provide insights into the requirements, architecture and key technologies for B5G networks, it does not address the critical issue of MM in B5G networks. Hence, to the best of our knowledge, no study has ever provided a comprehensive view of the functional requirements, challenges and potential solutions with regards to the future MM strategies, essential to realizing the future networks. We now list the contributions of this paper, which aim to address these aforementioned gaps, as follows:

1. We present a novel discussion on the functional requirements and design criteria for 5G and Beyond MM mechanisms.

2. We develop a novel qualitative analysis for the legacy mechanisms as well as the current state of the art MM mechanisms on the basis of reliability, flexibility and scalability, towards their utility for 5G and beyond wireless networks.

3. We provide a novel classification of the current state-of-the-art mechanisms based on where they are implemented or create an impact within the network, i.e., core network (CN), access network (AN) and extreme edge network. Additionally we also provision a mapping of these classifications onto the 5G service based architecture (SBA) defined by 3GPP [19], which will consequently assist to indicate explicitly the gaps that exist currently.

4. We then provide the first discussion in literature with regards to how the current state-of-the-art strategies will fare towards MM for potential B5G solutions envisioned.

5. Following the discussions and qualitative analysis we have elucidated the various challenges that the design and development of future MM mechanisms will face.

6. We then provide a discussion on the potential strategies that will help them overcome these persistent challenges. We accompany these discussions with a novel mapping between the potential strategies and the aforementioned challenges that they will help resolve.

7. Lastly, we develop and provision a novel and unified vision for the 5G and beyond MM solution.

The rest of this paper is organized as follows: Section 2 presents the functional requirements and design criteria for the 5G and beyond MM mechanisms. Section 3 defines the criteria for the qualitative analysis as well as the parameters that govern the fulfillment of these criteria. Section 4 presents the novel qualitative analysis for the legacy mechanisms and establishes their pros and cons for 5G and beyond MM. Section 5 introduces a similar analysis for the current state of the art mechanisms as well as their utility towards the MM solutions fr future networks. Section 6 then presents the persistent challenges, the potential strategies that will assist in resolving these challenges whilst aiming to satisfy the requirements defined in Section 2, and the proposed framework for 5G and beyond MM. We then conclude this paper in Section 7.

2. 5G and Beyond MM: Functional Requirements and Design Criteria

Future wireless networks, in addition to being dense, heterogeneous and extensively programmable, will serve multiple industry verticals as well as accommodate multiple tenants on the same network infrastructure [9, 20]. These transformations, some of which are being discussed by the research community [8, 21], represent a paradigm shift from the current network architecture design. As a consequence, MM mechanisms need to be re-evaluated and/or re-designed. For this, we first present the functional requirements of MM mechanisms for future wireless networks in Table 1, based on the characteristics we derive from the current and future network scenarios.

From Table 1, it can be observed that the MM solutions for 5G and beyond networks will have to adapt and evolve, so as to be able to serve the future wireless networks efficiently. As seen from the table, MM solutions will need to be redesigned so that they are flexible, scalable and reliable to ensure the requested QoS and seamless mobility. Apart from these requirements, there are certain criteria that will impact the design and development of future MM solutions. Consequently, in the following text we present an insight into these myriad design criteria and their impact on 5G and beyond MM.

2.1. Centralized vs. Hierarchical vs. Distributed Solution

While a centralized solution might offer optimality given its global view, a distributed approach can offer more reliability by eliminating the Single Point of Failure (SPoF) problem as well as avoiding congestion at a specific network node. Instead, a hierarchical approach can incorporate the benefits of both aforesaid techniques. For example, in LTE, MME is the mobility management entity

Table 1: Functional Requirements from 5G and beyond MM

Req #	Current Scenario	5G and Beyond Scenario	Resulting MM Functional Requirement
R1	Single RAT connectivity	UE connected to multiple RATs	Provision support for multi-RAT MM as well as efficient RAT selection methods.
R2	UEs with predominantly mobile broadband applications request MM support	UEs with enhanced broadband (eMBB), massive machine type communications (mMTC) and Ultra-reliable low latency communication (URLLC) applications will request MM support. These applications will have different QoS requirements [22]. For example: minimum data rate, latency, reliability, etc.	Provide MM support based on context, i.e., based on application requirements, user mobility, network conditions, etc.
R3	Density of UEs in the current scenario is $10^5 devices/km^2$ [23]	Density of UEs in 5G and beyond will be $\geq 10^6 devices/km^2$ [23]	MM mechansims should be able to scale and provision support for the increasing user density
R4	Network is vendor driven [24]	Network is softwarized [24]	MM solutions should evolve to utilize the benefits provided by softwarized 5G and beyond enablers such as SDN, Network Function Virtualization (NFV), etc.
R5	Network is predominantly ground based with static radio towers	APs and relay stations may be carried on drones in 5G and beyond networks [25, 26]	MM solutions for 5G and beyond networks should support mobility of both UEs and APs
R6	4G, 3G and 2G are standardized and the MM protocols provision support for all these devices	5G and beyond networks and devices will be gradually rolled out. They are fundamentally different from 4G, 3G and 2G networks	Backwards compatibility to support legacy devices will be needed from MM solutions for 5G and beyond.
R7	Sub 6 GHz is the frequency range for data transfer	Sub 6 GHz, millimeter Wave (mmWave) [21], Terahertz communication [5, 6] will be utilized in 5G and beyond networks	Increased robustness, given that VLC and mmWave will be significantly impacted by the environment, thus challenging seamless mobility in 5G and beyond networks.
R8	Finest granularity of tracking and localization is $< 50m$ [27]	Finest granularity of tracking and localization is a beam ($< 10cm$) [27]	MM solutions should evolve to utilize the advanced level of granularity to provision better mobility and tracking performance in dense urban or high speed scenarios
R9	The complexity is driven mainly via user requirements in a homogeneous network	The complexity in 5G and beyond networks is a combination of different user types, different QoS requirements, heterogeneous RAT scenarios, heterogeneous backhaul scenarios [28] and ultra dense nature of the network	MM mechanisms need to ensure adequate flexibility (they should accommodate for the increased heterogeneity) and tractable solutions (fast and low computational complexity) with well managed power consumption for the increased network complexity
R10	Requested services and data is always hosted in the IP Multimedia Subsystem (IMS) core	Requested services and data in 5G and beyond networks can now be hosted at the network edge, through MECs [24]	MM mechanisms should provide adequate Service Migration [2]/ Service Replication [3] support to ensure the required QoS from the applications
R11	Support for mobility up to 350 km/h	Support for mobility up to and beyond 500 km/h proposed [23]	MM solutions need to ensure the required flexibility to accommodate multiple demanding mobility profiles, avoiding the *one size fits all* approach

with the Serving Gateway (S-GW) being the mobility anchor, and hence, it is a centralized solution. However, Distributed Mobility Management (DMM) [29] assists in decentralization of the traditional MM mechanisms, wherein instead of having a single MM anchor for all the flows on a UE, the anchors are now distributed. By distribution of MM anchors here we mean that, when a flow is initiated to/from a UE, the anchor may be chosen dependent on the flow requirements. For example, given a new flow originating to/from a UE, a MM anchor is chosen which might be very close to the UE to assist in network offloading purposes, whereas pre-existing flows might still be served from the MM anchors to which they were first assigned, so as to avoid service disruptions. Hence, it would provide more reliability. The hierarchical method on the other hand, will combine the centralized and distributed approaches to offer the reliability of the distributed approach (through decentralization of mobility anchors) and the optimality of the centralized approach (e.g. through master and slave network management entities). An example of such a distributed/hierarchical approach can be found in the upcoming 5G networks, wherein through SDN and NFV there is a separation between the CP, i.e., Access and Mobility Function (AMF)- Session Management Function (SMF) for mobility management, and the data plane (DP), i.e., OpenFlow (OF) switches, etc., [4, 30].

2.2. Computational Resources

The computational resource locations and their corresponding computational power will determine the degree to which the mobility management mechanisms can be distributed. For example, edge clouds can aid not only in MM related computation (e.g., RAT and AP selection) but can also enable faster access to content through caching. In addition, for 5G and beyond networks, it will also be critical for the MM mechanisms to determine whether services need to be migrated or replicated [3, 31, 32], so as to maintain service continuity and hence ensure the required QoS. Note that, by service replication we mean that the services being requested by a user undergoing a mobility event are replicated to other edge servers. Further, by service migration [2, 31] we imply that the services being accessed by a user undergoing a mobility event are migrated to the next edge cloud server where the user is expected to move to.

2.3. Backhaul Considerations

Network densification and the prohibitively expensive nature of installing optical fibre as backhaul [28] will render the backhaul scenario in 5G and B5G wireless networks to be extremely heterogeneous, i.e., they will be composed of both wired and wireless links. Further, the backhaul wireless links will consist of multiple radio access techniques such as microwave, mmWave, VLC or LiFi, co-existing together [6]. These transformatory trends will need to be taken into consideration while developing future MM mechanisms, as:

- Congestion or multiple-hops in the backhaul can impact the E2E latency, and consequently, the perceived QoS [33].
- Backhaul reliability will be critical given the relatively poor penetration capability of mmWave [34] and additionally, strong atmospheric absorption features for VLC [5]. Thus, during mobility, attaching to an AP with a poor backhaul link quality can correspondingly lead to degradation in QoS since, there can be increased packet loss or even an outage altogether.

2.4. Context

A multitude of parameters, such as user mobility profiles, type of flows, network and user policies, AP signal quality, network load, backhaul-fronthaul options, etc., constitute the context. Additionally, MM mechanisms for 5G and B5G networks will have to service users with different mobility profiles, accessing different services. Hence, the available contextual information will be valuable for any future MM mechanism. For example, in [15], network load aware MM methods present an improvement of 75% in throughput at the cell edge as compared to the context agnostic methods, thus reinforcing the aforesaid criteria.

2.5. Granularity of Service

Granularity in MM services (e.g., based on flow, subscriber or mobility profile) will be an important component for MM methods to provision optimal solutions for 5G and B5G networks. Further, the type of granularity offered, i.e., per-flow based, mobility based, etc., will depend on the user context as well as the network conditions. Hence, innovative mechanisms like the Mobility Management-as-a-Service (MMaaS) paradigm [35] will be required. In MMaaS, on-demand MM solutions can be employed by or assigned to UEs. For example, if a device is moving at a high speed ($\sim$ 300km/h) and there is another device, say an IoT device, that is stationary, then a mobility based granularity of service can be adopted. Based on this service granularity provision, the high mobility device can be allocated resources on macro-cells whilst the stationary device can be served by small cells. Another important example being that of network slices. Network slicing, the concept, typically refers to a resource based logical slicing of the existing network infrastructure to support multiple verticals and corresponding operators that serve them [36]. In such scenarios, on-demand MM will be necessitated by the network slices, as they will cater to services with differing mobility requirements and patterns, such as the URLLC and eMBB services.

2.6. D2D Service Availability

The availability of D2D services will determine how the mobility management mechanism is executed, as D2D can assist in providing seamless mobility through CP information and/or DP data forwarding. This will be specially

relevant in scenarios involving V2X [37], wherein for example, the vehicles, that are outside the coverage area of the infrastructure network (IN) or are experiencing a deep fade with the IN, can exchange data with it by relaying their information through other vehicles, over the PC5 interface [37], that might be nearby and within the coverage area of the IN or are experiencing better channel conditions with it.

2.7. Physical Layer Considerations

The introduction of massive MIMO and mmWave technology will certainly impact current MM methods. Concretely, in urban environments the mmWave links will face extensive blockage alongside their limited range due to the propagation characteristics. Hence, this will require densification, which introduces the possibilities of frequent handovers (FHOs). Here by FHOs, we refer to the fact that in a dense network environment, such as those in 5G, the users will be subjected to handover (HO) scenarios more frequently as compared to that in the current networks. On the other hand, beamforming through massive MIMO antennas can be utilized to track moving users and hence, provide them with high QoS through higher throughput and better localization services.

Further, for B5G networks, VLC and meta-surfaces have emerged as the main enablers. Note that, VLC will be challenged extremely by the existing environment. This is so because, it operates in the Terahertz range of frequencies, thus making most objects in the environment as blocking agents. Also, meta-surfaces will lead to programmable environments, which will create the issue of dimensionality for an optimal solution.

Henceforth, the physical (PHY) layer techniques require consideration in any MM mechanism development for 5G and beyond networks.

2.8. Control Plane Signaling

An important target of future MM mechanisms will be to reduce the CP signaling induced during handovers. Studies such as [38], have proposed enhanced handover signaling mechanisms for an SDN-based core network architecture, such that the transmission and processing cost as well as the overall latency during a handover process is reduced whilst ensuring the Capital Expenditure (CAPEX) does not rise significantly. Such a procedure will enhance the QoS for the user while switching access points and hence, will be critical to the future MM suite.

Although, a complete overhaul of MM mechanisms for future wireless networks might result in optimal solutions, the time to develop and market them will be correspondingly longer. Hence, in the following sections, we perform a novel qualitative analysis for the various legacy as well as current state-of-the-art mechanisms and standardization efforts, and evaluate their suitability as *enablers for MM* in 5G and beyond wireless networks.

3. Qualitative Analysis Criteria

Present day MM mechanisms and standards are extremely stable and also readily implementable. Given the challenging nature of 5G and B5G network scenarios, it is of significant interest that these mechanisms and standards be explored for their potential inclusion – whole or in part – as enablers for future MM solutions. Hence, we perform a novel qualitative analysis of these mechanisms on the basis of reliability, flexibility and scalability: the three pillars of any future MM strategy. As part of this qualitative analysis, we firstly present a detailed description of these three criteria, as follows:

1. *Reliability* will help to determine whether the MM mechanisms employed will be able to ensure guaranteed and continuous service in any given network topology. Such reliability requirements entail not only continuous connectivity whilst traversing a geographic area, they also include reliability in delivery of packets for critical and delay sensitive services. Further, reliability from a MM mechanism also envelops factors such as tolerance to congestion (through for example, Distributed MM), ensuring faster yet trustworthy reconnection and authentication whilst mobile, ensuring appropriate levels of redundancy in the number of flows, connections, and hosts, and also ensuring appropriate resource allocation for users with myriad mobility and application profiles at the edge, access and core network.

2. *Flexibility* as a qualitative analysis tool helps to determine the adaptability that MM mechanisms will provide to the network, which as discussed will be heterogeneous and dense in all perceivable aspects. The flexibility provisioned by MM mechanisms for future networks hence envelops factors such as the ability to formulate and deploy MM policies depending on individual user profiles, flow profiles or based on a slice profile. Further, ensuring the possibility of multi-connectivity through various layers such as transport layer (Stream Control Transmission Protocol (SCTP)/ Multi-Path Transmission Control Protocol (MPTCP)), IP layer (Multi-homing), Medium Access Control (MAC)-PHY layer (Dual Connectivity), will be an important factor for ensuring a flexible MM policy. Additionally, factors such as multi-objective access point selection/user association taking into account factors such as congestion, QoS requirements, backhaul reliability, etc., will be critical to a flexible MM mechanism.

3. *Scalability* aspect allows one to determine if the future MM mechanisms can serve the increasing number of user devices with a corresponding increase in requested QoS with heterogeneous mobility profiles. A measure of scalability of MM mechanisms can be

Table 2: Governing Parameters for the Reliability, Scalability and Flexibility of a MM mechanism/standard

#	**Reliability**	**Contribution to Reqs.**	#	**Flexibility**	**Contribution to Reqs.**	#	**Scalability**	**Contribution to Reqs.**
RL1.	Redundancy in the number of flows, connections, etc.	R7	FL1.	Granularity of service. E.g. per flow, per connection, per user, etc.	R9, R11	SL1.	Manageable number of connections with increasing number of users	R3, R9
RL2.	Seamless handover capability[†]	R1, R7, R8	FL2.	Capability to enable connectivity to multiple APs	R1, R9	SL2.	Manageable signaling load with increasing number of users	R3, R9
RL3.	Decentralization	R3, R4	FL3.	Handover service support at multiple network levels. E.g. Core network, Access network, etc.	R4, R9	SL3.	Manageable processing load with increasing number of users/devices	R3, R9
RL4.	Fast path re-routing at CN	R5, R10	FL4.	Handover decision making utilizing multiple parameters. E.g. network load, requested QoS, etc.	R1, R9	SL4.	Decentralization	R4
RL5.	Congestion aware	R2	FL5.	Context awareness	R2, R9, R10	SL5.	Ease of implementation and integration	R6

[†]Here seamless handover capability refers to the ability of a MM mechanism to permit vertical (inter-RAT) as well as horizontal (intra-RAT) handover.

gained by analyzing factors such as number of connections that can be managed given an increasing number of user devices, management of the signaling load generated due to mobility events, management of the increasing load due to processing the many CP messages generated in mobility events, as well as the ability to permit decentralization (which in essence would ensure scalability) and being easily deployable on a large scale given a new MM mechanism.

We summarize the aforesaid criteria into a list of parameters for each criteria and present them in Table 2. Additionally, we also indicate the requirements (Table 1) for whose fulfilment each of these parameters contribute towards. Note that, compliance with each of the stated parameters in Table 2 for the reliability, flexibility and scalability criteria will be essential towards ensuring that the MM mechanism under consideration satisfies the requirements (Table 1) defined for the upcoming 5G and beyond networks. We now elaborate upon the parameter-requirement relationships that have been illustrated in Table 2, with the objective of enhancing the comprehensiveness of the evaluation criteria.

3.1. Reliability: Parameter to Requirement mapping

The provision of redundancy in the number of flows and connections, i.e., by satisfying parameter *RL1*, can help fulfil requirement *R7* presented in Table 1. This is so because, redundancy in connections will help overcome the fragile nature of wireless channels in the frequency bands that constitute VLC and mmWave communications. Next, satisfying the parameter *RL2* will contribute towards fulfilling the requirements *R1*, *R7*, and *R8* (Table 1). Here, the ability to provision seamless handover assists in supporting mobility amongst multiple RAT(s) (*R1*), supporting multi-connectivity and thus reliability (*R7*), and utilize enhanced localization capabilities to accomplish the same in dense urban scenarios (*R8*). Additionally, the *RL3* parameter for the reliability criteria, when satisfied, will help to fulfill the *R3* and *R4* requirements (Table 1). The reason being, decentralization will allow for efficient handling of the number of devices (*R3*). Moreover, to establish an effective level of decentralization, such as for accessing cached data at the edge and in the IMS core, enablers such as NFV and Mobile Edge Computing (MEC) will be utilized (*R4*). Furthermore, the *RL4* parameter holds significant relevance towards fulfilling the requirements *R5* and *R10* (Table 1). Specifically, fast path rerouting in the CN ensures that the increased dynamism, due to the mobility of both the UE and APs (*R5*), is catered to in the CN. In addition, data path modifications, due to service migration and service replications, do not lead to extensive delays is also ensured through parameter *RL4*. Lastly, satisfying the *RL5* parameter will help towards fulfilling the *R2* requirement (Table 1), since guaranteeing congestion awareness helps service the different QoS requirements of the applications, such as virtual reality and emergency services, with better reliability.

3.2. Flexibility: Parameter to Requirement mapping

When a MM mechanism under study satisfies the flexibility parameter *FL1*, it correspondingly helps to fulfil the *R9* and *R11* requirements (Table 1). This is so because, *FL1* states that a MM mechanism should support granularity of service. This will correspondingly assist in accommodating the multitude of service requirements independently (*R9*) as well as avoid the *one size fits all* approach (*R11*). Next, *FL2* parameter will help in satisfying the *R1* and *R9* requirements (Table 1). Essentially, the capability to be able to connect with multiple APs will assist in multi-RAT MM (*R1*) as well as in provisioning enhanced agility for MM mechanisms in a dense and heterogeneous network (*R9*). Further, when the *FL3* parameter is satisfied, it helps to fulfil the *R4* and *R9* requirements. The reason being, to enable handover support at multiple levels of the network, usage of SDN, NFV and MEC platform will be necessitated for efficient implementation (*R4*). Moreover, such multi-level handover support will also provision flexibility for the network (*R9*). Additionally, satisfying parameter *FL4* enables the MM mechanism under study to contribute towards satisfying the *R1* and *R9* requirements (Table 1). Specifically, having a handover decision mechanism that utilizes multiple parameters aids in handling MM amongst multiple RAT(s) more flexibly and hence, efficiently (*R1*). Also, such strategies will ensure that alongside flexibility, solutions are computationally tractable and energy efficient (*R9*). Finally, parameter *FL5*, when satisfied, will be relevant for the fulfilment of requirements *R2*, *R9* and *R10* (Table 1). To elaborate, the context awareness feature of a MM mechanism will assist in provisioning MM support dependent on application, user and network context (*R2*), flexibility to handle the increased heterogeneity in the network (*R9*), and ensure QoS whilst performing complex tasks such as migrating or relocating services based on user mobility events (*R10*) through appropriate path and resource management.

3.3. Scalability: Parameter to Requirement mapping

For the scalability criteria, when parameter *SL1*, *SL2* and *SL3* are satisfied by a MM mechanism, they correspondingly also assist in fulfilling the *R3* and *R9* requirements (Table 1). Concretely, the ability to be able to manage increasing number of connections, signaling load and processing load with the number of increasing users will correspondingly assist in handling a user density of more than 10^6 devices per km^2 in 5G and beyond networks (*R3*). Also, they will help in ensuring the required scalability to accommodate the increasing heterogeneity in the network as well as the corresponding tractability of the MM solution (*R9*). Next, when parameter *SL4* for the scalability criterion is met, it helps to fulfil the *R4* requirement (Table 1). Specifically, to accomplish decentralization objective the MM mechanism under study will

need to utilize enablers such as NFV and MEC. Lastly, satisfying parameter *SL5* will help to meet the requirement *R6* (Table 1). The reason being that, ease of implementation usually arises from the fact that a MM mechanism has been used/deployed before, as well as is suitable to accommodate legacy devices whilst catering to a new set of service and devices. Hence, satisfying the *SL5* parameter will assist in ensuring that backwards compatibility requirements (*R6*) are adhered to.

And so, from the aforementioned elaborate understanding of the mapping, it can be deduced that the criteria chosen for our qualitative analysis are comprehensive in nature and approach. Moreover, and considering only the 5G networks since their KPIs have been defined [22], provisioning beyond 99.999% reliability will be ensured through the reliability metric during mobility scenarios. Further, latency less than 5 ms for connected cars and 10 ms for virtual reality and broadband applications, will be guaranteed through the reliability and flexibility metric. Specifically, the reliability metric will help provision congestion awareness, reliable link selection, etc., while flexibility will allow multiple type and number of connections during mobility scenarios. In addition, support for nearly 1 million devices per km^2 with different application and mobility profiles will be ensured through the scalability criterion. Consequently, this further reinforces the comprehensiveness of the criteria chosen for the qualitative analysis that follows.

4. Legacy mechanisms and standards: 5G and Beyond MM enablers?

We evaluate certain widely employed/studied legacy standards and mechanisms based on the criteria (reliability, flexibility and scalability) listed in Table 2. It is important to state here that, the goal of the following analysis is not to compare the considered standards and mechanisms against each other but rather to highlight the extent of their suitability for 5G and beyond networks.

4.1. IETF MPTCP-SCTP

4.1.1. Discussion

Being transport layer protocols, MPTCP (through multiple TCP connections) [39, 40] and SCTP (through its multi-homing capabilities) [41] can provide multiple TCP paths for flows originating at the host. Generally utilized for increasing data rates [42] and improving the QoS, the provision of multipath redundancy [43–46] and congestion awareness (at the transport layer level) [41, 47–49] will facilitate reliability for 5G and beyond MM mechanisms. Additionally, MPTCP and SCTP satisfy the granularity of service criterion (by provision of per-flow level granularity of service), which will be essential for the future MM mechanisms. Further, according to [39, 40, 50], for MPTCP to be implemented without altering the legacy systems, proxy servers supporting MPTCP will need to be installed in front of the legacy devices, such as the middleboxes installed by service providers. The legacy systems can then communicate with the proxies using the legacy TCP protocol, while the proxies utilize MPTCP for communicating with the destination MPTCP capable device. However, it is the requirement of these additional proxies that will impact the scalability of the MPTCP solution for 5G and beyond MM mechanisms. Moreover, for SCTP, both the user and server protocol stacks need to be updated [41]. Given the number of users in future networks, it will pose a scalability challenge for the deployment of SCTP as part of the 5G and beyond MM mechanisms.

4.1.2. Analysis

Given our objective of determining the suitability of MPTCP and SCTP for 5G and beyond MM mechanisms, we enlist their *pros* and *cons* as follows:

- MPTCP Pros
  - Allows for multiple data flows at the transport layer level [39, 40, 45], and hence, provisions for resiliency against connection failures, given the multipath feature [43–45]
  - Provisions congestion awareness, with studies such as [47] proposing specific congestion control methods for MPTCP
  - Through its ability to divide a connection into multiple sub-flows, MPTCP provisions the capability to handle each flow independently [45, 49]
- MPTCP Cons
  - The middleboxes installed by service providers are not optimized to support MPTCP [39, 40]
  - MPTCP requires proxies to allow MPTCP enabled devices to take its full benefits [50]
- SCTP Pros
  - Allows for multiple data flows at the transport layer level [41, 46], and hence, provisions for resiliency against connection failures, given the multipath feature.
  - Provisions congestion awareness, wherein reference [41] establishes the presence of congestion avoidance methods within the SCTP suite
  - Assists in network level fault tolerance through support for multi-homing [41, 46]
- SCTP Cons
  - Requires both host and destination device protocols stacks to be updated with the SCTP protocol [41]

From the *pros* and *cons* of both MPTCP and SCTP, as listed above, it can be concretely stated that the IETF MPTCP-SCTP methods satisfy parameters *RL1* (allowing for multiple flows over the network for any given user) and *RL5* (provisioning congestion awareness as part of the transport layer characteristic for MM) for the reliability criterion. Further, for flexibility, from our discussion above, it is clear that IETF MPTCP-SCTP only satisfies parameter *FL1* (by allowing for multiple flows, flow level granularity can be induced).

4.2. IEEE 802.21

4.2.1. Discussion

Network layer protocols will play a critical role in ensuring seamless mobility during inter-RAT mobility events, given the fact that a change in IP anchors/addresses invariably leads to a dropped session. A significant effort in this direction is provided by IEEE 802.21, which is an inter-RAT handover protocol allowing devices to move seamlessly between the various IEEE 802.x technologies [17, 51–54]. Sitting just above the MAC layer, it provides information and command service to higher layers thus permitting the users to perform a media independent handover. 3GPP technologies can also utilize this information and hence, allow devices to handover from 3GPP to IEEE 802.x RATs and vice versa. Consequently, IEEE 802.21 can provision certain degree of reliability and flexibility for 5G and beyond MM mechanisms. However, note that the protocol stack of all the users would have to be modified to implement the IEEE 802.21 mechanism.

4.2.2. Analysis

For the purpose of analysis, we list the *pros* and *cons* of the IEEE 802.21 mechanism towards 5G and beyond MM strategies, as follows:

- IEEE 802.21 Pros
  - Provisions seamless handover capability, as it allows users to switch between multiple RATs [17, 51, 54]
  - Provisions the possibility for a UE to connect to multiple APs [17, 52]
- IEEE 802.21 Cons
  - Requires the protocol stacks of both the host and destination devices to be modified, so as to enable the IEEE 802.21 functionality [51, 53]

And so, given the aforesaid *pros* and *cons* with regards to IEEE 802.21, it can be deduced that it satisfies parameter *RL2* for reliability (allowing for seamless movement between different RATs) and *FL2* for flexibility (allowing for the possibility to connect with multiple RATs) criteria.

4.3. IETF PMIPv6

4.3.1. Discussion

Proxy Mobile IPv6 (PMIPv6) is a layer-3 MM protocol that allows a network based MM solution by utilizing gateways and anchors, i.e., Mobile Access Gateway (MAG) and Local Mobility Anchor (LMA), respectively [55, 56]. An LMA manages multiple MAGs, and is responsible for the assigment of the IP prefix which the UE retains during its entire duration within an LMA, i.e., a PMIPv6, domain [55, 56]. Concretely, it is the topological anchor for the UE. On the other hand, MAG is responsible for performing mobility related signaling, on behalf of the UE, with the LMA. Furthermore, it mains the assigned IPv6 prefix as the UE roams around the MAGs within an LMA domain[55, 56]. It is noteworthy that PMIPv6 has also been adopted by 3GPP networks [57], thus reflecting the maturity and reliability of the solution with regards to its utility for future MM solutions. However, being centralized in nature, it can impact the network scalability and reliability in dense and heterogeneous future network environments, as a large volume of the traffic will pass through a single anchor. This can consequently lead to SPoF and congestion [58], thus making it less favorable for 5G and beyond MM mechanisms. And so, certain studies such as [58, 59] provide discussions on scalable methods for PMIPv6. Specifically, in [59] a PMIPv6 based DMM approach has been proposed. The DMM approach essentially aids in improving the reliability and scalability aspects, as it would provide a decentralized method (without any mobility anchors) and eliminate SPoFs. Furthermore, in [58], a cluster based approach was proposed to enhance the scalability of the existing PMIPv6 protocol.

4.3.2. Analysis

Based on the discussions carried out in Section 3.3.1, we now enlist the *pros* and *cons* of the PMIPv6 strategy with regards to its utility for 5G and beyond MM mechanisms, as follows:

- PMIPv6 Pros
  - Given that PMIPv6 is adopted by 3GPP and it forms a relatively agnostic setup for an UE towards its mobility signaling, it can thus provision seamless mobility [55–57]
  - Through the DMM based PMIPv6 approach, decentralization can be introduced [59]. Furthermore, other approaches, such as the clustering based approach in [58], can grant enhanced scalability and reliability to the PMIPv6 approach
  - Given that it has already been adopted by 3GPP for LTE, the available implementational expertise will enhance the ease with which it can be adopted in future networks

- PMIPv6 Cons
  - In its original flavor, PMIPv6 suffers from scalability and reliability issues due to the SPoF formed by the LMA in its architecture [58]
  - An explicit treatment of PMIPv6 with regards to the parameters for flexibility criterion is missing in [55–59]

And so, it can be deduced that the IETF PMIPv6 in its original flavor, given its maturity in development and deployment, satisfies the seamless handover parameter *RL2* in the reliability criteria. However, with enhancements from the use of DMM and cluster based methods, PMIPv6 can be decentralized and scaled thus satisfying parameters *RL3* and *SL4* in reliability and scalability, respectively. Furthermore, since it has already been explored and implemented in the LTE networks, it satisfies parameter *SL5* owing to its relative ease of implementation as against any other new protocol.

4.4. LTE MM mechanisms

4.4.1. Discussion

A. *Handover:* Whilst LTE mobility management derives its characteristics from the PMIPv6 MM strategy [60], LTE-X2 offers a method to decentralize it. In the presence of an X2 interface between two LTE eNodeB's (eNBs), instead of involving the core network for resource negotiation and data forwarding tasks, the eNBs communicate amongst themselves. This allows for a fast handover and also reduces signaling in the core network [61]. And so, due to the ability of LTE-X2 to provision seamless handover alongside decentralization, it can grant reliability and scalability to 5G and beyond MM mechanisms. Further, since it provisions decentralization and reduces CN signaling, it also reduces the processing load for the CN. Hence, LTE-X2 can facilitate scalability for 5G and beyond MM. Lastly, since LTE-X2 only enables multi-level HO service support, i.e., HO can be executed either at the access (through X2 HO) or core network level (through S1 HO), it offers limited flexibility.

   However, note that, LTE-S1 handover involves resource negotiation and routing decisions through the MME [61]. Due to this centralized approach, there will be extensive CN signaling, which will lead to congestion and a SPoF. Thus, in its own capacity, LTE-S1 handover is not foreseen as an enabler for future MM strategies.

B. *Traffic Offloading:* 3GPP, through Release-10, introduced Local IP Access (LIPA) and Selected IP Traffic Offloading (SIPTO) [62] protocols. Concretely, LIPA allows for a local breakout, wherein a mobile device can communicate with another device through a private network, i.e., the data flow does not pass through the 3GPP CN, or to a public network, if the private network connects to it [62]. An important challenge of LIPA with regards to MM is that, session continuity for LIPA connections during mobility events is not supported.

   On the other hand, SIPTO is an orthodox traffic offloading mechanism, wherein the goal is to offload the IP traffic to an eNB or a gateway that is closer to the UE. Next, during 3GPP Release-10, the concept of IP Flow Mobility (IFOM) was also introduced. IFOM allows a UE to offload, if possible, data sessions to the Wi-Fi network from the 3GPP network. Consequently, through IFOM, a UE can maintain data flows belonging to the same packet data network (PDN) connection simultaneously on both the 3GPP and the Wi-Fi network [62].

   Given these aforesaid traffic offloading strategies, they can consequently aid in managing any increase in traffic load within the network, as well as the processing load on specific network nodes, due to the increase in the number of users/devices. Thus, these mechanisms can provision scalability for 5G and beyond MM.

C. *Dual Connectivity and LTE-WiFi Aggregation:* The Dual Connectivity (DC) concept allows a user to camp on two APs simultaneously. Concretely, a UE can be connected to a Small-cell (SC) and a Macrocell (MC) at the same time, wherein the MC and SC are connected to each other via the X2 interface, and the MC is the master eNB. According to 3GPP, all control plane communications, including resource allocation on SC, are performed via the corresponding MC, i.e., the master eNB, to which the UE is associated to. Note that, DC was introduced by 3GPP for LTE during Release-12. But, it is in Release-13 that this concept matured, wherein multiple usage scenarios, architecture and the operational characteristics were defined. A detailed description of the same has been presented in [63]. Furthermore, during Release-13, the concept of LTE-WiFi aggregation (LWA) was standardized [64]. Through LWA, a UE can simultaneously receive packets over both the LTE and the Wi-Fi interfaces, wherein the aggregation of these two physically distinct data streams takes place at the Packet Data Convergence Protocol (PDCP) layer in the protocol stack. However, note that the LWA functionality is defined only for the downlink [65]. Henceforth, given that the DC and LWA strategies provision the ability to connect to multiple APs at the same time, they can provision reliability and flexibility for 5G and beyond MM mechanisms.

4.4.2. Analysis

For the 3GPP based MM mechanisms, we firstly highlight the *pros* and *cons* for the handover, traffic offloading and DC and LWA strategies, as follows:

- LTE Handover Pros
  - The LTE-X2 and S1 mechanisms together offer handover support at the access and core network level [61]
  - Through LTE-X2 handover mechanism, CN signaling can be avoided [61]
  - LTE-X2 permits decision making for a handover to be taken at the access network level. Hence, it reduces the processing load on the CN entities as well and also permits fast handover capabilities [61, 66]
- LTE Handover Cons
  - The S1 based handover mechanism involves signaling through the CN, which creates increased load on the CN [66] as well as introduces SPoFs
- LTE Traffic Offloading Pros
  - Provision a method for managing the traffic load given that the number of users/devices will increase significantly [62]
  - Provision a method for managing the processing load in the network nodes [62]
- LTE Traffic Offloading Cons
  - LIPA does not support session continuity during mobility events, as well as it requires an additional gateway [62]
  - SIPTO is not helpful in mitigating radio congestion [62]
  - IFOM is significantly harder to implement as it necessitates coordination with the non-3GPP networks [62]
- LTE DC and LWA Pros
  - Provisions the ability to connect to multiple 3GPP as well as Non-3GPP RATs [63–65, 67]
  - Provisions the capability to have multiple physical paths for data transmission, and thus better fault tolerance [63–65, 67]
- LTE DC and LWA Cons
  - 3GPP LWA is only applicable for downlink

From the *pros* and *cons* for the LTE MM mechanism, it is clear that they provision redundancy in data paths (through DC and LWA), decentralization (through X2 and traffic offloading) and seamless handover (through X2 and S1 handover), thus satisfying *RL1*, *RL2* and *RL3* parameters for the reliability criterion. Further, for the flexibility criterion, LTE MM mechanisms offer the possibility of a multi-level HO support (through X2 and S1 handover) as well as the ability to connect to multiple APs/RATs at the same time (through DC and LWA), thus satisfying parameters *FL2* and *FL3* for flexibility. Lastly, LTE MM mechanisms offer enhanced support with regards to the scalability criterion for 5G and beyond MM, as they satisfy parameters *SL2* to *SL5*, given their decentralization, ease of integration, multi-level handover mechanisms (X2 and S1 handover), and traffic offloading characteristics.

4.5. Non-3GPP Multi-Connectivity Solutions

4.5.1. Discussion

Multi-connectivity enables the users to establish and maintain physical and logical connections to multiple access points (possibly belonging to different RAT(s)) at the same time. Certain standards and mechanisms, apart from those developed by 3GPP (Section 4.4), that utilize this concept are ITU-Vertical multihoming (ITU-VMH) and the Co-ordinated Multipoint (CoMP) strategy.

Specifically, ITU-VMH permits the user to camp on more than one RAT, via multiple physical channels, at any given moment [68]. Through such provision, ITU-VMH ensures path redundancy. Further, through interactions between the various Open Systems Interconnection (OSI) layers, techniques such as MPTCP/SCTP in combination with ITU-VMH can also aid in the provision of path redundancy [68]. And so, ITU-VMH via its redundancy and seamless handover capabilities ensures reliability for 5G and beyond MM mechanisms. Note that, the seamless handover capability is facilitated by the ability of ITU-VMH to allow the user to connect to a multitude of APs, thus reducing the possibility of outage (as compared to standard HO process) during mobility events. Further, via the provision of multi-connectivity, ITU-VMH also permits per-channel granularity of service. Hence, it also provisions flexibility for future MM mechanisms. However, ITU-VMH, like the IEEE 802.21 standard, would require a transformation in the protocol stack to permit efficient resource allocation at all the protocol layers [68]. Such a transformation might be difficult to scale to all the user devices, and hence, ITU-VMH is not a very scalable solution for 5G and B5G networks.

Lastly, the Co-ordinated Multipoint (CoMP) strategy involves multiple access points co-ordinating with each other to serve a given user [69]. Similar to ITU-VMH, CoMP can also provision path redundancy as well as seamless handover capability, owing to its coordinated feature. And hence, it is also a reliable strategy for future MM strategies. Further, similar to ITU-VMH, CoMP can configure multi-connectivity alongside per-channel granularity (multiple APs permit multiple channels for transmission of data and hence, per-channel granularity of service can be provisioned) [69]. Consequently, it is qualitatively a

flexible mechanism for 5G and B5G networks. However, since CoMP will involve centralized scheduling operations, it will lead to SPoF as well as challenge the scalability of backhaul networks. Consequently, this also renders CoMP as not being a very scalable proposition towards the objective of developing 5G and beyond MM solutions.

4.5.2. Analysis

We now present the *pros* and *cons* for ITU-VMH and CoMP strategies, as follows:

- ITU-VMH Pros
  - Provisions path redundancy through multi-homing [68]
  - Provisions the capability to connect to multiple RAT(s) at any given time [68]
  - Per-channel granularity of service is possible
- ITU-VMH Cons
  - It will require the transformation of the entire protocol stack [68]
- CoMP Pros
  - Provisions path redundancy through its ability to coordinate data transmission from multiple APs, which may also belong to different RATs [69, 70]
  - Provisions the capability to connect to multiple RAT(s) at any given time [69, 70]
  - Through the use of multiple APs for transmission, per-channel granularity of service is made possible
- CoMP Cons
  - Centralized processing introduces the possibility of SPoF [69, 71]
  - Backhaul networks will need to have extremely high capacity and extremely low latency characteristics, so as to support CoMP whilst maintaining QoS [71]

Concretely, ITU-VMH and CoMP satisfy parameters *RL1* (allowing for the possibility of redundant physical connections) and *RL2* (allowing for seamless mobility) for the reliability criterion, and parameters *FL1* (provisioning the possibility of per-channel granularity for MM) and *FL2* (allowing for the possibility of connecting to multiple RATs/APs) for the flexibility criterion.

4.6. RSS based AP selection methods

4.6.1. Discussion

The erstwhile Received Signal Strength (RSS) based methods employ a very simplistic approach to AP selection, i.e., comparing the detected AP link quality (RSSI/RSRP/RSRQ) levels [57, 72, 73]. The aforesaid simplistic nature can hence permit scalability for the future MM mechanisms as it is easy to implement, and does not entail a high processing and signaling load either. However, such an approach can be plagued by multiple issues. For example, APs with a good RSS might be overloaded (as more users will be assigned to them) whilst others maybe under-utilized. Such a scenario also implies that RSS based methods are not reliable as a better RSS does not always guarantee better QoS, since, congestion will lead to degraded QoS. Moreover, in dense scenarios, even with the implementation of a hysteresis, UEs will be subject to FHOs due to the fluctuating RSS and availability of multiple candidate APs. This further exemplifies the unreliability of RSS based methods. Additionally, these methods are one-dimensional, given that they consider only RSS as a decision parameter. The RSS methods also do not provision any granularity of service, context awareness, multiple levels of HO support, etc. Hence, they do not offer any flexibility to the MM mechanisms for 5G and B5G networks.

4.6.2. Analysis

Based on the discussion, we present here the *pros* and *cons* of the RSS based AP selection methods as follows:

- RSS based methods Pros
  - Easy to implement, given that it has already been adopted by 3GPP [57, 72, 75]
  - Relatively low processing and signaling load, owing to its simplicity [75]
- RSS based methods Cons
  - FHOs in ultra dense scenarios is a pertinent issue [74]
  - It is agnostic of other parameters related to the UE and the network, such as the load, UE context, etc., thus making it unreliable and one-dimensional [57, 72, 74]

Given these *pros* and *cons*, the erstwhile RSSI based method due to its existence and maturity can ensure mobility between multiple RAT(s), hence, satisfying parameter *RL2* for reliability criteria. Furthermore, owing to the aforementioned simplicity and maturity in development and deployment it also satisfies parameters *SL2*, *SL3* and *SL5* for the scalability criterion.

To summarize, we introduce Table 3 wherein we indicate the parameters that each of the explored methods satisfies for the reliability, scalability and flexibility criteria.

Table 3: Compliance with the Reliability, Scalability and Flexibility criteria for the legacy MM mechanism/standard

		Mechanisms											
		IETF MPTCP-SCTP		**IEEE 802.21**		**IETF PMIPv6**		**LTE MM mechanisms**		**Non-3GPP Multi-connectivity solutions**		**RSS based handover methods**	
		Cnf.[†]	Refs.[δ]	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.
Reliability	**RL1**	✓	[42–48]	×	[17, 51] [54]	×	[55, 56] [59]	✓	[61, 67] [63–65]	✓	[68–70]	×	[57, 72] [73]
	RL2	×		✓		✓		✓		✓		✓	
	RL3	×		×		✓		✓		×		×	
	RL4	×		×		×		×		×		×	
	RL5	✓		×		×		×		×		×	
Flexibility	**FL1**	✓	[43, 49] [45, 46]	×	[17, 52] [54]	×	[55–59]	×	[61, 67] [63–65]	✓	[68–70]	×	[57, 72] [73, 74]
	FL2	×		✓		×		✓		✓		×	
	FL3	×		×		×		✓		×		×	
	FL4	×		×		×		×		×		×	
	FL5	×		×		×		×		×		×	
Scalability	**SL1**	×	[39, 40] [50]	×	[51, 53]	×	[59] [55, 56]	×	[61, 62]	×	[69, 71]	×	[57, 72] [73, 75]
	SL2	×		×		×		✓		×		✓	
	SL3	×		×		×		✓		×		✓	
	SL4	×		×		✓		✓		×		×	
	SL5	×		×		✓		✓		×		✓	

[†]The conformance (Cnf.) of a given mechanism for a given criterion.

[δ]The corroborating references (Refs.), if any, for the specified conformance of a mechanism for a given criterion

We also enlist the important references that have lead us to the development of Table 3, as presented in this article. From the discussions, analysis and Table 3 it can be deduced that none of the legacy mechanisms that have been studied achieve the requirements as necessitated by 5G and B5G networks. Concretely, none of the studied mechanisms satisfy all the parameters of the criteria utilized for the qualitative analysis. Notably, the 3GPP based LTE MM mechanisms provision the best basis and support for 5G and beyond MM mechanism, given that they collectively satisfy the most parameters amongst other strategies explored.

Additionally, through this qualitative analysis, whilst we have presented the offered capabilities from legacy mechanisms towards 5G and B5G MM, we have also exposed the gaps that exist. This reinforces the fact that a holistic MM strategy for future wireless networks still remains elusive. Hence, in the following section, we explore the current state-of-the-art in MM solutions for 5G and beyond networks.

5. 5G and Beyond MM: Current State of the Art

Global efforts have spinned up consortiums that have provided impetus to the development of 5G, including that of MM strategies. Further, for B5G networks, such as 6G, certain collaborative efforts have already started. References [5, 6, 8, 9, 76] highlight the advances that have been made with regards to identifying the enablers and core principles of B5G networks. Hence, in this section we first detail the current state of the art in MM mechanisms and the parameters they satisfy from Table 2. We then follow this with a first discussion in literature that elaborates upon the utility of the current state of the art in MM for B5G networks.

As a prologue to the aforementioned discussion, we introduce Figure 2, wherein the 5G architecture standardized by 3GPP has been presented [19]. Correspondingly, we have also presented the classification of the various mechanisms that we explore in Sections 5.1 and 5.2 with respect to the 5G architecture in Figure 2. This classification is dependent on the portion of the network that is impacted (directly or indirectly) the most by a particular MM scheme. Furthermore, we have illustrated whether the studied mechanisms are either control plane or data plane solutions. Concretely, a CP solution would primarily impact MM via either CP signaling or decisions, while a DP solution would entail provisioning alternate and more efficient data paths. A detailed discussion with regards to these classifications has been provided in Sections 5.1 and 5.2.

Concretely, the 5G architecture, as shown in Figure 2, consists of two main core network functions, i.e., the Session Management Function (SMF) and the Access and Mobility Management Function (AMF). The SMF communicates with the User Plane Function (UPF) over the N4 interface, while the AMF is responsible for communicating with the RAN side over the N2 interface. Furthermore, the AMF and SMF communicate with other network functions, such as the Policy Control Function (PCF), Authentication Server Function (AUSF), etc., to execute their defined functionalities within the ambit of the policies and existing user and network context. For the sake of brevity, in Figure 2 we club all of these functions into a single entity box called *Network Functions*. Moreover, the AMF also has an N26 interface that connects to the Evolved Packet Core (EPC) to facilitate Inter-RAT mobility, while an N32 interface exists in the event of a change in Public Land Mobile Network (PLMN) with 5G Core (5GC) as the CN for both the visted and home networks. Note that, the interfaces *N2*, *N4*, *N26* and *N32* are all control plane paths, with the AMF, SMF and other network functions forming the control plane entities.

In addition, the AMF in 5G networks is the equivalent of the MME in LTE-4G networks. It focuses on handling mobility at the access network level (such as AP selection, resource allocation, etc.). The SMF on the other hand handles the CN related tasks during mobility events (such as path re-routing, etc.). Next, in Figure 2, it can be seen that the RAN interacts with the UPF through interface N3, and the UPFs use the N9 interface to communicate amongst themselves. Also, the 5G networks provision a local breakout through the N6 interface from an UPF. The interfaces *N3*, *N9* and *N6* constitute the data plane paths, with the RAN and UPF forming the data plane entities. Lastly, the UE, which is also a data plane entity, interacts with the AMF through the N1 interface. However, to maintain clarity, we have omitted the illustration of this interface from Figure 2. Thus, with this background, we now explore the 3GPP 5G MM mechanisms as well as other research efforts with regards to MM for 5G networks.

5.1. 3GPP 5G MM Mechanisms

3GPP, through TS 23.501 [19], TS 23.502 [77] and TS 38.300 [78], has provided significant insights into the design and development of 5G MM strategies. New session management methods, service continuity states, UE mobility monitoring, provisioning for multi-homing, load balancing strategies, provision of on-demand MM, resource allocation due to mobility events, the new MM module, i.e., AMF, inter- and intra- next generation core (NGC) handovers, and LTE-EPC 5G-NGC interworking have been introduced in the aforesaid 3GPP specifications. These techniques through the provision of a softwarized solution and a global view of the network scenario alongside user context appear to facilitate the efficient operation of 5G and beyond MM mechanisms. Consequently, in the discussions that follow, we investigate these newly defined 3GPP MM mechanisms and elaborate their *pros* and *cons* for future MM mechanisms.

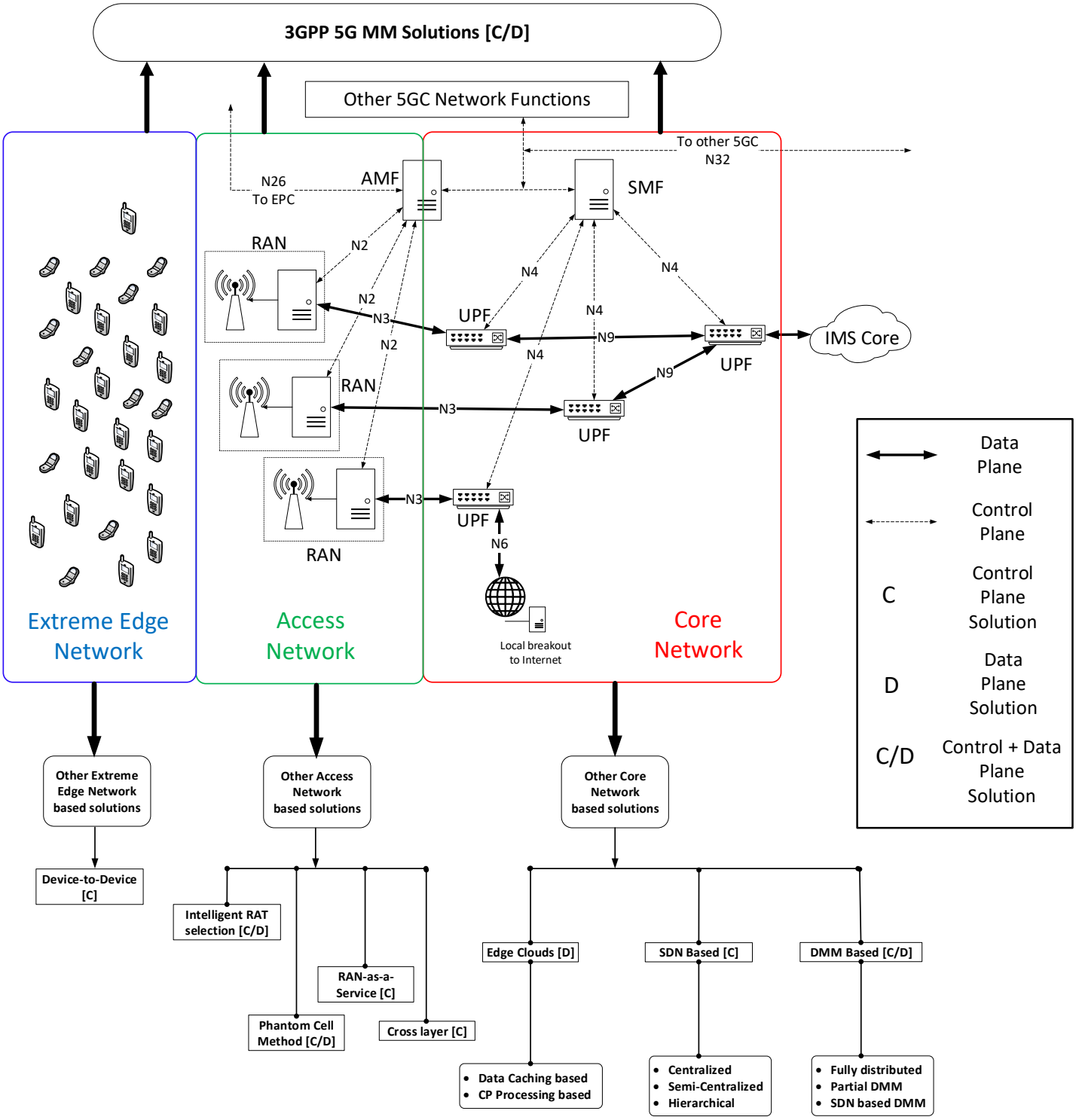

Figure 2: Classification of the state of the art in MM strategies on the 5G architecture.

5.1.1. Discussion

A. *UE Mobility monitoring:* In TS 23.501 [19], details with regards to how the UE mobility is monitored and the corresponding actions with regards to resource allocation and context updates have been specified. Concretely, when a UE is mobile, the 5G standards define that the AMF will be responsible for monitoring its movement and hence, its mobility pattern. Furthermore, during a UE mobility event, new resources on the destination AP are managed by the AMF through the RAT and Frequency Selection Parameter (RFSP). Such a process simplifies the identification of the required resources, as well as migration of these resources to the destination network. Moreover, the AMF manages the UE mobility event notification, i.e., it provisions details with regards to the mobility event as well as the areas of interest (Tracking areas, Cells, RANs, etc., to which a UE might migrate to). The other Network Functions (NF), such as the SMF, can subscribe to these notifications so as to employ their decisions and policies.

B. *Session Management:* Through TS 23.501 [19], the various modes that can be utilized to manage the multiple heterogeneous sessions for a given user has been defined. Notably, if a UE is connected to multiple RATs then, for a given Protocol Data Unit (PDU) session, the UE has the choice to select the access network over which this PDU session will be served. In addition, the UE, in the event of mobility or congestion, can request a PDU session to be transferred from 3GPP to non-3GPP RAT(s). Furthermore, in roaming scenarios, PDU sessions can either avail a local breakout or be routed through the home network. Specifically, each PDU session can be granted, independently, different routing modes. To do so, the SMF in the 5G CN controls and monitors the status of the data paths. Moreover, the SMF also provisions the capability of performing selective traffic routing by the application of Uplink Classifier (UL CL) on certain data plane entities, i.e., UPFs. A UPF essentially performs the function of a router in the 5G network.

C. *IPv6 multihoming:* The new 5G standards, as specified in TS 23.501 [19], have formalized the use of IPv6 multi-homing so as to reap the benefits from the multiple physical channels that will be available for use through multi-connectivity. Specifically, according to TS 23.501, more than one session anchor can be specified for a PDU session. Note that, a PDU session anchor's primary role is to assign the IPv6 prefixes that are used by the UE for a given PDU session to communicate with the public network. However, all these PDU session anchors will have a single UPF as a branching point. Next, during a mobility event, a make-before-break approach for a PDU session is adopted to provision service continuity. It must be stated here that, service continuity is ensured through the Session and Service Continuity (SSC) modes, which we will discuss next.

D. *Session and Service Continuity Modes:* 3GPP, through TS 23.501 defines the SSC modes, which are critical for the network in determining the level of service continuity offered to a PDU session [19]. Concretely, three modes are defined, i.e., *SSC mode 1*, *SSC mode 2* and *SSC mode 3*. We briefly describe them as follows:

   - *SSC mode 1:* This mode ensures that the IP address is preserved. Specifically, the PDU session anchor is maintained regardless of the access technology being used by the PDU session after the mobility event. Furthermore, the IP address is maintained throughout the lifetime of the PDU session. Additionally, more PDU session anchors might be allocated for additional IP addresses, however, it is not necessary that they be maintained just like the initial IP address and PDU session anchor.
   - *SSC mode 2:* In this mode, if needed, the network can release a PDU session and request the UE to immediately establish a new PDU session with the same network. Moreover, if the UE has multiple PDU session anchors, the additional anchors can be released or allocated (for new IP prefixes/addresses).
   - *SSC mode 3:* In this mode, IP address is not preserved. This consequently makes any changes in the user plane visible to the UE. However, to ensure that an acceptable level of QoS, and hence, service continuity is maintained, a *make-before-break* approach is followed. This essentially determines the destination PDU session anchor before relieving the resources the PDU session occupies at its current anchor.

   It must be stated here that the SSC mode for a UE is selected by the SMF depending on the UE subscription details as well as the PDU session type.

E. *User Plane aspects:* In 5G networks, UPFs will be utilized to handle the data plane traffic. Concretely, they can be thought of as routers, on whom the routing rules are programmed by the SMF. In TS 23.501 [19], the aforesaid specifics have been defined. However, note that the methodology to establish these paths still involves exchanging Tunnel Endpoint Identifiers (TEIDs) between CN entities. This, as we will state in the analysis, can be a cause of increased network load. Additionally, traffic re-routing, in the event of mobility or load balancing, is handled by the SMF, wherein it sends the necessary information, such as the forwarding target information, to the UPFs. Lastly, in

the event of mobility of a UE, packet buffering is also provisioned so as to minimize the loss of packets and hence, QoS.

F. *Dual Connectivity:* Through TS 23.501 [19] and TS 37.340 [79], 3GPP has also concretized and standardized the integration of Multi-RAT Dual Connectivity (MR-DC) into 5G. Concretely, the UEs will now have the capability and possibility to connect to two APs belonging to the same RAT (LTE-LTE, 5G New Radio (NR) - 5G NR) or to different RAT(s) (LTE - 5G NR). As in LTE-DC, this can be configured to allow fast-switching (fast HO), since control plane is not changed unless the Master Node is changed. Also, the UP is terminated at MC, so, no CN signalling is necessary for intra-MC HO.

G. *Edge Computing:* TS 23.501 [19] defines the support for edge computing platforms in 5G networks. Concretely, these are utilized in the non-roaming or local breakout roaming modes. By local breakout, we mean that a UE can access public network without traversing the core network via additional gateways that are placed within the network. Note that, the 5G CN is responsible for selecting a UPF that is close to the UE and also has access to an edge compute node. Consequently, traffic steering is performed at this UPF towards the edge compute node.

H. *Network Slicing:* The concept of enabling a telecom operator to be able to slice its infrastructure network into logically separated networks and consequently service multiple tenants, e.g., virtual network operators, services (eMBB, URLLC, mMTC), etc., using the same, wherein the logical separation involves dynamic allocation of network resources, is termed as network slicing [36]. 3GPP, in TS 23.501 [19], has discussed network slicing in detail, wherein its support for roaming as well as its involvement in the inter-working process between 5G CN and LTE EPC has been elaborated. Specifically, support for migrating and translating the Single Network Slice Selection Assistance Information (S-NSSAI), which consists of the necessary information with regards to an assigned network slice for a UE, between the Home PLMN (H-PLMN) and the Visited PLMN (V-PLMN) has been detailed. Similarly, for the inter-working process, 3GPP charts out the principles for migration, translation and creation of S-NSSAIs whenever a UE undergoes mobility and changes from a 5G network to an LTE network, and vice versa. Moreover, the support has been defined for scenarios where the N26 interface, which is the standard 5G CN and LTE EPC inter-working interface, may or may not be present [19].

On the other hand, and importantly, the concept of network slicing also assists in provisioning tailor-made MM solutions for the tenants that each network slice will cater to. This consequently helps to deploy on-demand MM strategies.

I. *Load Balancing and Congestion Awareness:* In TS 23.501 [19], 3GPP has defined procedures for load balancing at the AMF and SMF, as well as congestion awareness within the core network. Concretely, two specific strategies, i.e., load balancing and load re-balancing, have been provisioned. Within the load balancing paradigm, new users incoming into an AMF region, if necessary, are directed to an appropriate AMF in order to manage the load of the AMFs. To do this, appropriate weights, indicative of the load on each AMF, are assigned and updated at appropriate intervals (typically on a monthly basis). On the other hand, if an AMF becomes overloaded, then load re-balancing is performed. Here, already registered users are migrated to other AMFs that are not overloaded while ensuring minimum service disruption [19]. Note that, the new AMF chosen should belong to the same AMF set. An AMF set is defined as the AMFs which belong to the same PLMN, have the same AMF region ID and the same AMF set ID value [19]. These parameters are pre-configured by the network operator. Lastly, 3GPP also provisions extensive details with regards to handling congestion control for the Non Access Stratum (NAS) messages. This is important from the perspective of MM, as MM messages are carried over NAS to the CN nodes. For further details with regards to the specifics of the congestion control procedures, the reader is referred to TS 23.501 [19].

J. *Cell, Beam and Network Selection:* Through TS 23.501 [19] and in particular through TS 38.300 [78] details with regards to cell, beam and network selection have been specified. For *cell selection* these standards documents, developed by 3GPP, specify support for cell selection procedures given that the UE is in either Radio Resource Control (RRC) idle, or RRC inactive or RRC connected state. Note that, RRC idle state refers to a UE that can listen to paging channels, broadcasts and multicasts, as well as perform cell quality measurements. The RRC inactive state refers to a UE that can roam within the RAN-based notification area (RNA) without informing the NG-RAN. The RRC connected state for a UE implies that it has an active connection and data flow. Most notably, for the RRC connected state, cell mobility and beam mobility have been specified. As the name suggests, a UE can either undergo a cell handover or it can switch between the beams that a given AP uses. To perform this, procedures for beam quality and cell quality measurements have also been defined in [78]. The beam quality measurements are performed in the physical layer for multiple beams being transmitted by a given cell. These measurements are filtered and aggregated

at the RRC layer to obtain the cell quality measurements. Note that, these quality measurements are still performed using the RSSI/RSRP/RSRQ/SINR metrics. Furthermore, in [78] procedures for cell selection and handover involving intra- and inter-frequency handover in 5G NR, Inter-RAT handover within 5G CN, Inter-RAT handover from 5GC to EPC and vice versa, have been specified. For the sake of brevity, we do not detail these procedure and refer the reader to TS 38.300 [78]. Moreover for Inter-RAT handovers, procedures for packet buffering and forwarding as well as data path switching, to ensure the requested QoS, have also been defined. Lastly, roaming and access restrictions are also appropriately defined based on the user subscription to both the SMF and AMF. This facilitates the selection of the right AP and PLMN for a given user [19, 78].

K. *Inter-Working, Migration and Handover signaling:* While TS 38.300 [78] specified certain handover procedures for both the CP and DP, a detailed description of the handover signaling, inter-working between 5G CN and EPC, and migration of PDU sessions has been provided in TS 23.502 [77] and TS 23.501 [19]. Concretely, through [77] the CN signaling process for the various stages in a handover, i.e., handover request, handover preparation, handover complete/cancel/reject, have been presented in detail. These handover signaling strategies have been detailed for Intra-RAT HO (N2 and Xn handovers) as well as for Inter-RAT handovers (involving 5GC and EPC). Moreover, the handover signaling procedures have also been defined for the scenarios wherein the EPC-5GC inter-working interface, i.e., N26, may or may not be present. Also note that, the 5G-N2 handover is similar to the LTE-S1 handover (specified in Section 4.4.1) and the 5G-Xn handover is similar to the LTE-X2 handover (specified in Section 4.4.1). Next, for the 5GC and EPC inter-working, in TS 23.501 [19] the principles for maintaining IP address continuity in the event of UE mobility from 5GC to EPC or vice versa have been provisioned. However, it is also specified that in the event a UE transitions from 5G to 3G or 2G and vice versa, the IP address continuity might not be maintained. Furthermore, procedures for transferring the PDN/PDU sessions established by a UE over a 4G/5G network, when it transitions to the 5GC/EPC, over the N26 interface have been provisioned in [19]. Also, traffic steering and forwarding procedures have also been elaborated. Lastly, procedures for migrating PDU sessions from non-3GPP access to the 3GPP access, when a UE undergoes a mobility event from 5GC to EPC, is also supported [19].

L. *D2D mobility support:* With the standardization of Proximity Services (ProSe) in 3GPP Release-12 and 13 [64], 5G networks can utilize the capability to orchestrate data forwarding/relaying in both DP and CP. This can consequently enhance the ability of the network to provide a proactive and seamless handover procedure [80].

5.1.2. Analysis

Given the extensive overview with regards to the MM solutions that have been provisioned by the 5G standards [19, 77, 78], we now, as part of our qualitative analysis, present the *pros* and *cons* for the same, as follows:

- 3GPP 5G MM Pros
  - Provisions monitoring of UE mobility, mobility event notifications and resource negotiation mechanisms at destination networks [19]
  - Employs flexible session management strategies, wherein provision of per-PDU session granularity, through path selection, roaming support and traffic steering, has been detailed [19]
  - Support for IPv6 multi-homing [19]
  - Provision for multiple sessions and service continuity modes [19]
  - Support for Multi-RAT DC [19]
  - Support for Edge Computing [19]
  - Network slicing information migration support in the event of inter-/intra- RAT mobility [19]
  - Network slicing support for provisioning on-demand MM
  - Ability to provision context awareness via network slicing
  - Provision for managing core network load by introducing load balancing and re-balancing principles on the AMF [19]
  - Provision of congestion awareness on the CP handling MM messages, i.e., NAS [19]
  - Introduction of beam level MM support [78]
  - Intra-RAT (5GC to 5GC) and Inter-RAT (5GC to EPC and vice versa) HO support [19, 77]
  - Well defined EPC and 5GC inter-working interface, i.e., N26 [19, 77]
  - Mobility support at the D2D level [64, 80]
- 3GPP 5G MM Cons
  - Handover signaling in the CN is extremely suboptimal [13]
  - RAT selection still relies on received signal quality fundamentals only [78]
  - A unified framework for cross-layer mechanisms, such as MPTCP-SCTP (transport layer), IPv6 multi-homing (network layer) and MR-DC (Physical and MAC layer) working together, has not been provisioned

– In IPv6 multi-homing, a single point of failure (SPoF) still exists, as the multiple PDU session anchors are still connected to a single UPF from where the paths branch out [19]
– Co-ordination between D2D peers for enacting an efficient MM strategy is not explored explicitly in the standards

From the *pros* and *cons*, it can be deduced that the 3GPP 5G MM mechanisms will be able to support reliability parameters *RL1* (owing to the support for MR-DC and IPv6 multi-homing, and hence, redundancy in the number of connections and flows), *RL2* (owing to the support for MR-DC and handover procedures defined, thus ensuring seamless handover capability), *RL3* (owing to managing mobility at the access, core and extreme edge network levels as well as local breakouts, thus introducing decentralization) and *RL5* (owing to the congestion awareness feature in NAS). Next, for the flexibility parameters, 3GPP 5G MM mechanisms satisfy *FL1* (owing to the granularity of service support per PDU session as well as per mobility level, and the ability to support on-demand MM through network slicing support), *FL2* (owing to the ability to connect to multiple APs through MR-DC and IPv6 multi-homing support), *FL3* (owing to the handover support at the access, core and extreme edge network levels via the Xn handover, N2 handover and 3GPP ProSe, respectively) and *FL5* (owing to the ability to take into account the context of the tenant via network slicing). Lastly, 3GPP 5G MM mechanisms, for the scalability criterion, satisfy parameters *SL1* (owing to the AMF load balancing strategies, local breakout strategies, multi-level handover support as well as the granularity in service per mobility levels), *SL4* (owing to local breakout and support for edge computing, thus leading to decentralization) and *SL5* (since these are standards, implementation and integration is not a bottleneck).

Note that, scalability parameters *SL2* and *SL3* are not supported owing to the sub-optimality in CN handover signaling as well as the presence of SPoFs, as stated in the *cons* for the 3GPP 5G MM mechanisms. Also, given that the 3GPP 5G MM mechanisms provision both CP and DP related strategies as well as the core, access and extreme edge network related mechanisms, in Figure 2 they have been classified as illustrated.

5.2. Other Research Efforts: Core, Access and Extreme Edge Network Solutions

From the perspective of MM strategies in 5G networks, the main objective of the ongoing academic and industrial research efforts has been to provision mechanisms that cater to the myriad user mobility and application profiles, as well as to ensure context/on-demand based service provision and continuity [81]. For example, in [82], a wide swathe of avenues that exist in the 5G MM design have been explored. It discusses an SDN based framework that can encompass strategies and techniques which grant certain level of adaptability (feedback based), flexibility (in terms of granularity provisions) and reliability (through availability of multiple paths) for 5G MM solutions. Notably, and apart from the aforementioned broad study, specific areas of MM have also been tackled through research efforts such as [13] wherein optimal handover signaling strategies for 4G-5G networks have been proposed.

Hence, given that we will be analyzing a wide range of mechanisms and strategies, we have broadly classified them as being *Core Network*, *Access Network* and *Extreme Edge Network* based solutions, as shown in Figure 2. These classifications reflect the regions in the network where the respective mechanisms generate the most impact. Concretely, *Core network* based solutions will invoke solutions that primarily assist in the provision of MM services through the core network. Similarly, the *Access network* and *Extreme Edge network* solutions assist in provision of MM services through the access and extreme edge portion of the wireless network. And so, we now present a detailed discussion of these solutions alongside their efficacy in satisfying the criteria listed in Table 2.

5.2.1. Core Network Solutions

5.2.1.1. Discussion

Core network solutions have been categorized further as either being *SDN*, *DMM* or *Edge clouds* based. Solutions that utilize SDN to implement MM can be equipped with a global or locally-global network view. This top-view of the network enables MM solutions to offer a high degree of optimality. However, as a result of the convoluted 5G network scenario, the design of SDN CP also becomes increasingly crucial. Hence, the placement of SDN controller(s) (SDN-C) in the overall network topology is an important factor to consider [83]. Consequently, we present a brief discussion on the SDN based solutions, which might be Centralized, Semi-Centralized or Hierarchical [84–86].

A centralized MM solution will consist of a single global SDN-C which monitors and manages the entire network. With the global view, it enables the formulation of optimal MM solutions. However, the centralized nature might not offer the scalability and reliability (SDN-C can be a SPoF) [84, 87] needed by 5G MM solutions. Note that, even though SDN-Cs might appear as SPoFs, corresponding clustering for load sharing and redundancy can help alleviate this issue. Specifically, and similar to the method proposed by 3GPP to pool the Mobility Management entities (MMEs) to avoid SPoF problem and to share the workload between MME instances, SDN-Cs can be clustered together to provision redundancy (and hence no SPoF) and workload sharing. Next, semi-centralized approaches divide the entire geographical region into smaller domains, each managed by a separate SDN-C. This SDN-C, responsible for handling MM in its domain, helps to enhance the network scalability. However, since each domain still has a single SDN-C managing it, SPoF issue might become a limiting factor. Further, for inter-domain HO, extensive signaling would be required between two SDN-Cs whilst

the same would be non-existent in a centralized approach [84]. On account of this trade-off, a semi-centralized approach can be successful if an appropriate number of SDN domains are created, which do not increase the signaling burden while reinforcing the network reliability and scalability characteristics [87]. A combination of the aforementioned approaches, i.e., hierarchical approach, consists of SDN-Cs at multiple levels [84]. Whilst the global SDN-C behaves as a master (tuning HO parameters, manage inter domain HOs, etc.), the SDN-Cs in the lower hierarchical levels manage MM within their domains and function as slaves. Such an approach can hence provide the scalability and reliability required by 5G MM solutions.

Next, similar to the SDN based solutions, DMM based approaches will contribute significantly to the design and functioning of 5G networks. With the ability to provide a distributed DP in conjunction with a distributed/centralized CP [4, 29, 88–90], DMM can enhance the scalability (by removing anchors prevalent in current MM solutions, i.e., decentralization) and flexibility (by allowing the most optimum access router for each flow independently) of the 5G networks. These approaches can be classified as being fully distributed, partially distributed and SDN based.

The fully distributed approach whilst ensuring reliability and scalability by distributing both DP and CP, will encounter extensive amount of handover signaling between access routers (ARs) during a mobility event [4, 88]. Note that the DP functionalities and location of ARs are the same as that of the UPFs. However, depending on the type of DMM approach, the CP is fully or partially located on the ARs themselves, instead of being located in a centralized controller. And so, while the fully distributed approach is challenged by the signaling between ARs, the partially distributed (P-DMM) approach centralizes the CP, hence, alleviating this concern [4, 88]. The P-DMM approach also maintains the benefit of avoiding a single mobility anchor. However, an enhancement of this approach is the SDN based approach. Similar to the P-DMM approach, the CP is still at a central controller, i.e., SDN-C, however, the signaling between the controller and the DP devices is far more simplified as compared to the partially distributed approach. The reason being, in an SDN based approach, the ARs are converted to mere forwarding devices and it is the SDN-C that orchestrates the forwarding rules (routing table) on them to realize the data paths for the existing sessions in the network. Concretely, in the SDN based approach the DP devices no longer need to perform a handshake, like in the P-DMM approach, with the central controller to establish a route, instead the routing information is now fed to the DP devices by the SDN controller [88, 89]. These enhancements are further quantified in [88] by the fact that the mean HO latency for SDN based DMM is reduced by 3.94% as compared to P-DMM, while the E2E delay is reduced by 39.55%.

Subsequent to these discussions, and given that the current standardization in 5G [19, 77] stipulates the functionality for mobility management to be split up between the AMF and the SMF NFs, it is noteworthy that the decoupling of the CP and DP and subsequent utilization of the aforesaid NFs via an SDN-C can provision the capability to implement fast and efficient MM solutions for 5G and beyond networks. Such solutions, on the basis of the discussions thus far, will be reliable, flexible and –to an extent– scalable. Since, CN signaling during mobility events will still be a challenge, given the future network scenario, there remains a possibility for the SDN and DMM based 3GPP 5G MM solutions to be rendered sub-optimal.

Lastly, edge clouds, which essentially refer to data clouds/processing centers close to the RAN within a given network infrastructure, can have a profound impact on the user QoS during mobility scenarios (through fast access to data and compute resources) [91]. Henceforth, several studies such as [2, 31, 32, 92–94] alongside 3GPP and ETSI [95], have studied the fundamental concepts of utilizing the edge clouds for fast data access (via data caching) as well as for processing capabilities (i.e., performing certain MM operations without the messages having to traverse the entire CN). Note that, we classify the edge clouds to be a CN solution, even though we state that they are most likely to be closer to the RAN, because, certain topology designs might entail a hierarchical setup. In this hierarchical setup, there will be some edge clouds that are placed close to the RAN and some of them being placed further away from the RAN, say close to the S-GW and Packet Gateway (P-GW) in an LTE network [91]. Such an approach can help in caching data according to their level of popularity, taking into account CN traffic as well as the latency to retrieve the requested content [91].

5.2.1.2. Analysis

For analyzing the core network solutions we utilize the generic classifications, i.e., SDN based, DMM based and Edge Cloud solutions, and firstly list their *pros* and *cons*.

- SDN based mechanism Pros
  - Provisions global view of the network [84, 86]
  - Provisions hierarchical solutions, thus enabling decentralization [84]
  - Provisions the ability to manage CN signaling, and hence, DP paths during mobility events [84–86]
  - Provisions a single point of collection for network statistics thus enabling the design and development of context based MM mechanisms [87]
- SDN based mechanism Cons
  - Extensive CN signaling for managing handovers in a centralized/semi-centralized approach [84]
  - It does not alleviate the issue of mobility anchors which can lead to SPoFs in the DP

- DMM based mechanism Pros
  - Provision decentralization of the mobility management anchors [4, 88–90]
  - Assist the CN in implementing efficient data paths for UEs undergoing mobility [4, 29, 88]
- DMM based mechanism Cons
  - Fully decentralized solution introduces extensive CN signaling in order to manage the changes in data paths and mobility anchors, and hence, handovers [88]
  - Partially distributed solution, while solving the extensive CN signaling, introduces a central controller, and hence, an SPoF [88]
  - Co-existence and integration with already deployed networks and devices will be a significant challenge [29]
- Edge clouds Pros
  - Ensure data offloading opportunities, and hence, reduction in CN traffic load [31, 95]
  - Facilitate processing of MM related tasks without the messages having to traverse the CN [93]
  - Context awareness [93, 94]
- Edge clouds Cons
  - Require dedicated infrastructure and appropriate placement [2, 31, 95]
  - Require fast service migration strategies to ensure seamless mobility [32]

From these *pros* and *cons* as well as the preceding discussions, it is evident that the SDN based solutions satisfies parameter *RL2* (allowing for seamless mobility), *RL3* (through the provision of decentralized solutions), *RL4* (through the ability to re-program paths in CN via orchestration of OF rules) and *RL5* (through the ability to utilize network statistics for traffic steering with the CN) for the reliability criterion. For the flexibility criteria, the SDN based mechanisms satisfy the parameters *FL1* (through the capability of orchestrating policies dependent on flow type, slice, etc.), *FL3* (by allowing for CN based MM solutions that will work in synergy with the access network based solutions) and *FL4* (through the global view of the network wherein a variety of parameters such as network load, QoS requirements, etc., are considered). In terms of scalability, SDN based solutions satisfy parameters *SL1* to *SL3* (given the ability to manage and steer traffic flows with the ability of having a distributed, hierarchical or centralized implementation) and *SL4* (due to the possibility of having a decentralized configuration).

The DMM based solutions, however only satisfy parameters *RL2* (allowing for seamless handovers) and *RL3* (due to the decentralized nature) in the reliability criterion. Further, for the flexibility criterion, DMM based solutions only satisfy parameter *FL1*, i.e., they only offer granularity of service by preventing any mobility anchor. It is noteworthy though that, from the scalability aspect DMM based solutions, like SDN based solutions, satisfy parameters *SL1* to *SL4*, and for the same reasons.

Lastly, for the edge cloud based solutions, parameters *RL2* (allowing for seamless mobility through fast access to data/processing capabilities upon migration to the target network) and *RL3* (allowing decentralization of MM based services) are satisfied for the reliability criterion. For the flexibility criteria, parameters *FL1* (due to the ability to provision services based on mobility and application profiles), *FL3* (by allowing for MM methods at the edge network level in addition to the access and core network based solutions), *FL4* (by provisioning processing capabilities for user association/AP selection services) and *FL5* (by allowing for context awareness in data caching according to user mobility) are satisfied. Additionally, for the scalability criteria, parameters *SL1* to *SL4* are satisfied by the edge cloud solutions. The reason being, they allow for decentralization which can consequently permit better capability to manage connections and control messages due to increasing number of users.

It is important to state here that, given the SDN based mechanisms assist in MM through CP procedures, DMM based solutions assist through CP procedures as well as provision alternate and effective DP paths, and Edge clouds provision alternate and effective data paths, they have been classified as being CP, CP/DP and DP procedures, respectively, in Figure 2.

5.2.2. Access Network Solutions

5.2.2.1. Discussion

As part of access network strategies, one of the key approaches that has been proposed, and similar to LTE dual connectivity, is the concept of phantom cell [96]. It allows the UE to camp its CP on a MC, while its DP is being handled at the small cells that lie within the coverage of the earlier mentioned MC. This, in essence offers a low signaling cost regime to perform the intra-MC HOs as the UE does not need to access the CN for radio resource management operations during HO. Concretely, the MC handles the radio resource allocation operations for the phantom cells, and hence, during HOs between the phantom cells the CN signaling is avoided [63].

Moreover, owing to the softwarization of the complete network, the process of exchanging information between the various OSI layers, i.e., implementation of the cross layer strategy, is eased. This in turn allows the network to formulate solutions that are optimal, taking into cognizance the impact and benefits that the solution will produce at various levels of the network [97–99]. However, to realize cross-layer techniques, significant modifications to the software architecture of the protocol stack will be necessary [97–99]. Another consequence of the softwarization

process is the RAN-as-a-Service (RANaaS), also known as Cloud-RAN (C-RAN), which allows on-demand allocation of access network resources (e.g., Baseband unit (BBU) pool, BBU- Remote Radiohead (RRH) functional splitting) depending on the network and user context [100–102]. Additionally, the BBU pool, through close interaction of various RATs at a single location, can orchestrate fast handovers on-demand [103].

However, in order to choose the best APs to connect to in a multi-RAT scenario, computationally tractable RAT selection mechanisms need to be adopted. The multi-RAT solutions are a broad classification for the myriad RAT selection processes (Optimization based, Fuzzy logic and Genetic Algorithm based, RSSI based, etc. [18, 104–106]) that have been proposed. From our earlier discussions it is evident that RSSI based methods, although simple, do not weigh in other parameters such as network load, backhaul conditions, or user/network policies, for a RAT selection decision. This will most certainly result in sub-optimal solutions. But, optimized mechanisms, that can facilitate closed form solutions and are computationally tractable, will be able to capture more features from the network. Consequently, context aware mechanisms, such as [107, 108], will lead to optimal solutions that can be implemented for real-time scenarios.

It must be stated here that, the aforesaid HO decision may be executed either at the UE (user-centric) [107], at the network, or as a joint effort between the UE and the network (hybrid decision process).

5.2.2.2. Analysis

As part of the analysis for the access network solutions, we firstly present the *pros* and *cons* for each mechanism discussed above, as follows:

- Phantom Cell method Pros
  - Grants the ability to a UE to connect to multiple APs simultaneously, thus also granting redundancy in physical layer connections [96]
  - Provisions the ability to allow per-flow and per-user granularity of service [35, 96]
  - Handover support at access network level [96]
  - Ease of implementation due to existing standards on MR-DC [19, 96]
- Phantom Cell method Cons
  - Handovers between different MC domains will still entail service disruption [19, 96]
  - Inter-MC domain handover signaling will still be a significant burden on the CN [13, 96]
- RANaaS Pros
  - Provisions on-demand allocation of network resources at the RAN level [100–102]
  - Provisions the ability to execute on-demand handovers, through close interaction between the various RATs that are integrated at a BBU pool [103]
  - Assists in allowing UEs to camp on more than one AP
  - Introduces support for executing handovers at the access network level [103]
  - Introduces the ability to utilize per-flow/channel granularity of service by being able to manage the physical connections more centrally [100–103]
- RANaaS Cons
  - Requires a complete architectural overhaul at the RAN side of the network [100–102]
- Cross layer Pros
  - Allows for the sharing of network statistics between the various OSI layers [97–99]
  - Allowing for interaction between multiple OSI layers, thus facilitating the possibility of efficient utilization of multi-homing [68, 97–99]
- Cross layer Cons
  - Requires significant software modifications to the existing modular nature of the protocol structure [97–99]
- Intelligent RAT selection Pros
  - Optimized RAT Selection strategies [18, 104–107]
  - Utilization of parameters such as AP load, UE context, etc., for RAT selection [18, 104–107]
  - Provisioning the ability to select RATs per-slice/user/flow [107]
  - Provisioning the ability to select multiple APs (possibly belonging to multiple RATs) [108]
- Intelligent RAT selection Cons
  - Requires rapid collection of network statistics to perform well informed selection
  - Computational complexity and convergence time of RAT selection algorithms will be critical, given the QoS requirements in 5G [108]

Given the discussions in Section 5.2.2.1 and the *pros* and *cons* listed above, we now determine the parameters, listed in Table 2, satisfied by each of the mechanisms explored. Concretely, for the phantom cell method, parameters *RL1* (redundancy in physical layer connections) and *RL2* (seamless mobility) are satisfied for the reliability criterion. For the flexibility criterion, parameters *FL1* (by

permitting the possibility of per-flow and per-user based MM), *FL2* (allowing for connectivity to multiple APs potentially belonging to different RATs) and *FL3* (provisioning handover support at the access network level that will work in synergy with CN based mechanism) are satisfied. In terms of scalability, the phantom cell method satisfies parameters *SL1* to *SL3* (owing to the handling of handover related computation and decision at the access network) and *SL5* (owing to the existing standards on MR-DC, as discussed in Section 3).

Next, the RAN-as-a-service concept satisfies parameters *RL2* (allowing for seamless handovers) and *RL5* (the softwarized nature enables dynamic initiation for RAN functionality such as BBU resources, functional splits, etc., depending on the network and user context) for reliability, parameters *FL1* (allowing for per-flow, per-user, per-slice, etc., service granularity through its softwarized nature), *FL2* (allowing the possibility for connecting a user to multiple APs through its softwarized nature), *FL3* (provisioning handover support at the access network which will work in synergy with the CN and edge network based methods) and *FL4* (enabling the possibility of collection and utilization of RAN based information and generating intelligent AP selection/user association decisions) for flexibility, and parameters *SL1* to *SL3* (by offloading handover decision making and signaling to the access network) for scalability.

On the other hand, the cross-layer method only satisfies parameters *RL2* (allowing for seamless handover) and *RL5* (allows for congestion aware method by sharing statistics about queue lengths, buffer sizes, etc., amongst the various layers) for the reliability criteria. Further, for the flexibility criteria it satisfies only parameters *FL2* (by allowing for the possibility of multi-homing, etc.) and *FL4* (allowing for the possibility of sharing statistics and other information amongst the various OSI layers and enabling joint optimization for AP selection, path re-routing, etc.).

Lastly, for the intelligent RAT selection methods parameter *RL2* (allowing for seamless handover through optimized decisions on RAT selection) is satisfied for the reliability criterion. For the flexibility criterion, parameters *FL1* (allowing for the possibility of flow/user/slice based RAT selection), *FL2* (allowing for the possibility to select multiple RATs for a given user) and *FL5* (via the ability to utilize user and network context for RAT selection) are satisfied, while for scalability only parameter *SL5* (owing to the extensive body of research for optimal RAT selection strategies) is satisfied.

It is important to state here that, given the intelligent RAT selection mechanism assists in MM through RAT selection (which is a CP task) and provision of effective and alternate DP paths, the phantom cell method provisions support for MM by handling the CP signaling for SC selection as well as provision alternate and effective DP paths via SCs, and RANaaS and Cross layer strategies assist through efficient resource allocation decisions (which is a CP task), thus they have been classified as being CP/DP, CP/DP, CP and CP procedures, respectively, in Figure 2.

5.2.3. Extreme Edge Network Solutions

5.2.3.1. Discussion

Contrasting to the design and implementation of access and core network based methods, the extreme edge network based solutions consider the potential of utilizing D2D techniques for facilitating seamless HO. Multiple research efforts, such as [109–113], have provisioned methodologies to handle mobility of D2D pairs. Concretely, in [109] two types of handovers for D2D pairs have been provisioned. These are either *D2D aware* and *D2D triggered* handovers. They take into account the fact that the control of the D2D pair can be handed over independently of the actual cellular handover. And so, for the *D2D aware* handover, the D2D pair control (and if possible the cellular control) is handed over from the source eNB to the target eNB only after both the devices in the D2D pair satisfy the conditions to handover to the target eNB. On the other hand, the *D2D triggered* handover mechanism aims at clustering the devices of a D2D group in minimum number of cells. Hence, during mobility events the algorithm tries to determine the cell to which the majority of devices within the D2D group belong too.

Similarly, in [110] two handover management mechanisms have been proposed. While the joint handover strategy aims at migrating both the devices in a D2D pair simultaneously to the target eNB, the half handover stipulates that such a migration can be asynchronous. Furthermore, the D2D handover decision has also been specified in [110]. The Channel Quality Information (CQI) criteria has been utilized for the same. Next, in [111], a markov chain based model has been proposed for D2D mobility.

Lastly, the work done in references [112, 113] develops a model and simulation framework analyzing D2D mobility. Specifically, it considers a D2D pair with one of them being a transmitter (TX) and the other being just a receiver (RX). Thus, a handover procedure is defined for the scenario when the TX moves to the target eNB. In this procedure, the control of the D2D pair is transferred to the target eNB as soon as the TX migrates to it.

5.2.3.2. Analysis

We firstly present the *pros* and *cons* for the D2D strategies as follows:

- D2D strategy Pros
  - Provisions D2D handover management strategies [109, 110, 113]
  - Provisions MM support at the extreme edge network level [109–113]
  - Provisions the ability to decentralize MM functionality

- D2D Strategy Cons
  - Control signaling overhead will be a challenge [109, 110]
  - The viability with regards to energy efficiency of D2D peers as well as latency incurred in conveying the decisions with regards to MM are unexplored questions

Based on the discussions and the aforesaid *pros* and *cons*, the device-to-device methods satisfy parameter *RL2* (through the provision of various seamless handover management studies) for reliability, parameter *FL3* (provisioning mobility support at the edge network level which will work in synergy with access and core network based methods) for flexibility, and parameter *SL4* (allowing for the decentralization of MM functionality) for scalability.

Note that, given the D2D mechanism assists in MM through provision of CP assistance, thus they have been classified as being CP procedure in Figure 2.

5.3. B5G Networks

In addition to the discussions in Sections 5.1 and 5.2, in this section we present a short study detailing the challenges that current state-of-the-art mechanisms will continue to face for B5G networks. Furthermore, given the special characteristics that B5G networks will pose, as shown in Figure 1, we also list potential research areas for MM in B5G networks. Note that, these are then utilized in the subsequent section wherein we define challenges and potential solutions for 5G and beyond MM.

Concretely, while *SDN* and *NFV* will provide the tools for the B5G networks to provision rapid programmability of the meta-surfaces, during mobility scenarios they will be challenged critically. The reason being that, while current networking paradigms permit anywhere between 1 ms–10 ms time interval for performing any programmability task (latency restrictions, as specified in current 5G networks [11], on most services), in B5G networks this will be constrained even further as additional surfaces need to be programmed and orchestrated. Specifically, an increased number of surfaces/network nodes leads to more data required to be processed for generating appropriate programmability decisions. These decisions then need to be sent out (orchestrated) to the relatively large number of network nodes (including meta-surfaces), to execute the given task. Hence, this leads to an increased latency constraint on the network programmability aspect. Further, while the meta-surfaces provide a higher degree of freedom to the operator, they need to be programmed, as mentioned above. This introduces the challenging aspect of managing the SDN domains, NFV orchestration and the related signaling. As a consequence, the compactness as well as the efficiency of the current state-of-the-art SDN and NFV procedures will be challenged.

Next, with techniques such as DC, the challenge will be multi-fold as B5G networks will not just comprise of meta-surfaces, which can also act as a MIMO array, but they will also be equipped with Terahertz and mobile AP based multi-tier networks. And while, DC and multi-RAT procedures, as stated in Sections 5.1 and 5.2, will aid in ensuring a context-aware network selection procedure, the complexity for the access network techniques will be compounded by the fact that not only will they need to ensure QoS requirements, but they will have to also ensure sufficient available access bandwidth as well as backhaul bandwidth. Note that with the backhaul bandwidth there will be a significant design challenge since VLC technology is capable of carrying data rates of up to 1 Tbps. Current backhaul technologies cannot provision such high bandwidths [28]. Further, it is important to reiterate that the network will be composed of not only 4G-LTE and mmWave APs, but there will also be VLC and drone based APs, which essentially are the main reason for the increased complexity as discussed above.

Moreover, for the edge clouds, while they aid in allowing low latency access to cached content as well as the compute resources, the deployment strategies will need to be rethought given the ongoing growth pattern for data usage as well as the number of served devices coupled with more resource hungry services. Certain important recent studies in this direction have been provisioned via references [114, 115].

Given these significant shortcomings in the current state-of-the-art mechanisms towards B5G networks as well as taking into account the seminal works in the area of B5G techniques [5, 6, 8, 9] [12], the potential areas of research in MM for these networks are as follows:

- Characterization of the channel between meta-surface and the users, and meta-surface and the AP, in the event of user/AP being mobile, for the purpose of MM decisions
- Consideration of reliability and coverage of VLC link for MM decisions
- Characterization of the computational complexity for re-calibrating the meta-surfaces alongside the network, during mobility events
- Impact of mobility upon the programmable environment[1] concept, drone based communication and VLC
- Optimal RAT and AP selection with a programmable environment
- Optimal RAT and AP selection in scenarios where both the UE and AP (drone based) are mobile
- Characterizing the computational complexity of optimization methodologies for user association

[1] By environment, we refer to the physical environment that lies between the transmitter and receiver.

- Methods to handle possible increase in handover signaling/messaging during other network processes, such as reprogramming meta-surfaces to serve mobile users
- Formulation of a sound heterogenous RAT strategy, just like the 4G-5G concept, given mmWave and Terahertz technologies and their associated challenges related to coverage.

Note that, the aforementioned research areas do not form an exhaustive list, but are broadly indicative of what aspects remain to be explored with regards to MM in B5G networks.

To summarize, in this section we firstly introduced the 5G service based architecture and the classification of the various mechanisms that we analyzed, through Figure 2. Following this, we qualitatively analyzed the 3GPP 5G MM mechanisms as well as other research efforts with regards to their efficacy towards 5G and beyond MM solutions. Consequently, we introduce Table 4 wherein we indicate the parameters that each of the explored methods satisfies for the reliability, scalability and flexibility criteria (Table 2). We also enlist the important references that have lead us to the development of Table 4, as presented in this article. And so, from the capability profiles of each mechanism, as illustrated in Table 4, it is evident that even after significant efforts none of them completely meet the specified requirements as expected for the 5G and beyond MM mechanisms. Concretely, neither the 3GPP 5G MM mechanisms nor the other academic and industrial research efforts satisfy all the criteria completely. Subsequently, it is deduced that none of the analyzed mechanisms satisfy the requirements for the future MM mechanisms, as listed in Table 1. Hence, through the aforesaid qualitative analysis we have further exposed the gaps in the design and development for 5G and beyond MM mechanisms.

6. Challenges, Potential Solutions and Future framework

From our discussions in Sections 2 to 5, we have highlighted the requirements from MM mechanisms as well as the criteria that future MM mechanisms should satisfy to meet these requirements in Tables 1 and 2, respectively. Further, we have analyzed the legacy mechanisms and the current state of the art towards their utility for 5G and B5G networks in Tables 3 and 4, respectively. However, we have observed that gaps in fulfilling the requirements still persist. Concretely, we have demonstrated that none of the strategies evaluated satisfy the reliability, flexibility and scalability criteria in their entirety. Hence, to be able to design and develop a holistic MM mechanism, it is of substance to our study to understand the challenges/questions that persist. We consolidate, from earlier works in literature and the discussion in Sections 2-5, these key challenges/questions in the text that follows.

6.1. Challenges

6.1.1. Handover Signaling

Even after the release of 3GPP specifications for 5G [116], HO signaling is still a challenge. Hence, reducing HO signaling to ensure system scalability and reliability will be one of the key challenges. Certain studies such as [13] have provided methods to help overcome this challenge, and hence, can be actively pursued by the research and industrial community.

6.1.2. Network Slicing

Network slices have been defined to ensure different service types are served according to their own resource demands. Hence, it will be a key challenge to design MM strategies that either jointly take into account the requirements of multiple network slices or provide individual solutions for each network slice.

6.1.3. Integration framework for MM solutions

The state of the art and 3GPP specifications ensure to some extent the provision of flexibility, reliability and scalability for 5G MM solutions, as discussed earlier. However, since these solutions function at different sections of the network (Figure 2), the challenge will be to design them such that collectively they ensure the appropriate levels of flexibility, scalability and reliability in MM mechanisms to cope with the diversity in mobility profiles and applications the devices will access. Also, a part of this challenge will be to ensure that the CAPEX and Operating Expenditure (OPEX), owing to the architectural (software or hardware) transformations stemming from these redesigned MM mechanisms, are manageable.

6.1.4. Ensuring Context Awareness

Context based MM solutions accounting for factors such as network load, user preference, network policy, mobility profiles, etc., to ensure best possible provision of requested QoS will be important. The criticality of this challenge is enhanced by the fact that, low computational complexity whilst executing these solutions will be of the essence to meet the strict latency constraint requirements.

6.1.5. Architectural Evolution Costs

SDN and edge cloud capabilities will be important for enhancing the user experience during mobility, as discussed in Section 5. However, a key challenge will be to ensure appropriate scalability while maintaining a manageable CAPEX and OPEX.

Table 4: Compliance with the Reliability, Scalability and Flexibility criteria of the Current state-of-the-art MM mechanism/standard

		3GPP 5G MM mechanism		**Other Research Efforts**															
				Core Network Solutions						**Access Network Solutions**								**Extreme Edge Network Solutions**	
				SDN based		**DMM based**		**Edge Clouds**		**Phantom Cell Method**		**RAN-as-a-Service**		**Cross layer**		**Intelligent RAT selection**		**Device-to-Device**	
		Cnf.†	Refs.δ	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.
Reliability	**RL1**	✓	[19, 77]	×	[84–87]	×	[4, 29]	×	[31]	✓	[96]	×	[100]	×	[68, 97]	×	[18, 104]	×	[109]
	RL2	✓		✓		✓		✓		✓		✓		✓		✓		✓	
	RL3	✓		✓		✓		✓		×		×		×		×		×	[110]
	RL4	×	[64, 80]	✓		×	[88–90]	×	[93–95]	×		×	[101, 102]	×	[98, 99]	×	[105–107]	×	[112, 113]
	RL5	✓		✓		×		×		×		✓		✓		×		×	
Flexibility	**FL1**	✓	[19, 64]	✓	[84, 85]	✓	[29, 88]	✓	[31]	✓	[35, 96]	✓	[100]	×	[68, 97]	✓	[104]	×	[109]
	FL2	✓		×		×		×		✓		✓		✓		✓		×	
	FL3	✓	[77, 78]	✓		×		✓		✓		✓		×		×		✓	[110, 111]
	FL4	×	[80]	✓		×		✓	[93–95]	×		✓	[101–103]	✓	[98, 99]	×	[105–107]	×	[112, 113]
	FL5	✓		×		×		✓		×		×		×		✓		×	
Scalability	**SL1**	✓	[19, 64]	✓	[86, 87]	✓	[29]	✓	[31]	✓	[19, 96]	✓	[100]	×	[97]	×	[18, 104]	×	[109]
	SL2	×		✓		✓		✓		✓		✓		×		×		×	
	SL3	×		✓		✓		✓		✓		✓		×		×		×	[110, 111]
	SL4	✓	[78]	✓		✓	[88–90]	✓	[93–95]	×		×	[101–103]	×	[98, 99]	×	[105–107]	✓	[112, 113]
	SL5	✓		×		×		×		✓		×		×		✓		×	

†The conformance (Cnf.) of a given mechanism for a given criterion.

δThe corroborating references (Refs.), if any, for the specified conformance of a mechanism for a given criterion

6.1.6. Frequent Handovers

Reducing frequent handovers, ping-pong effects and devising an optimized HO strategy will still be a key challenge, given the dense and heterogeneous future network environment. This is further exacerbated by the fact that current methods, such as IEEE 802.21 and 3GPP specifications, fail to integrate cellular and non-3GPP networks effectively for seamless HO between them. For example, while methods such as LWA have been explored extensively [117, 118], an effective handover methodology between 3GPP and non-3GPP networks still remains elusive.

6.1.7. Security

An important challenge for ensuring service continuity and seamless mobility in an extremely dense and heterogeneous network environment, such as 5G and beyond networks, will be to ensure that security related tasks, such as authenticating the user as well as the network, be completed as efficiently as possible. By efficiently here we mean that the authentication should guarantee a required level of security whilst provisioning low computational complexity [119] as well as latency [120]. Again this task will become even more critical in scenarios where mobility occurs between 3GPP to non-3GPP networks.

6.1.8. Energy Efficiency

Given that one of the goals of 5G is to ensure enhanced battery lives for the devices, it will be a critical component for 5G MM services to ensure that the mobility of the devices is handled in an energy efficient way [121]. Additionally, 5G MM services will also need to ensure that the energy footprint goal for 5G networks is achieved via techniques such as smart AP selection methodologies [122] and reduced CN signaling [13]. By smart AP selection methodologies we refer to being able to not only account for the user energy consumption over the course of its mobility, but also accounting for the energy consumed whilst performing such selections.

6.1.9. Meta-surface Reconfiguration for mobility support

For the B5G networks, finding the optimal configuration of meta-surfaces during mobility related scenarios will be challenging. This is because, the physical characteristics of the surfaces will have to be altered rapidly so as to have the signals arriving at the user in a constructive manner.

6.1.10. Beyond 5G Network: Handovers

A fundamental question that will be posed in B5G is – how frequently and when will the handovers be needed? The reason this question is a challenge because, up until now the rate of power loss in an urban environment is characterized by a R^4 factor (where R is distance between the transmitter and receiver) given the destructive interference encountered. However, with programmable environments, according to [9], this decay will now be similar to the free space scenario, i.e., R^2, since all signals can be modulated in phase and polarization to interfere at the receiver in a constructive manner. And so, in mobile environments, the power decay will not be significant even at distances further away. Hence, the handover triggering methods and their execution procedure need to be revisited as currently they do not expect such a reliable behavior from the channel.

6.1.11. Beyond 5G Network: Protocol stack

A next fundamental question posed in B5G, with reference to meta-surfaces, is: What is the impact on the existing layers? The reason this question is a challenge because, the MAC, Radio Link Control (RLC), PDCP and TCP layers, they all have error control, packet reordering, transmission repeat request and other reliability control mechanisms in-built. These were designed keeping in mind that the environment is unreliable and randomly varying. However, with programmable surfaces the environment will be much more deterministic and reliable. Thus, there arises a case for either eliminating/modifying some of these layers (for example, a lightweight version of TCP may be utilized , as the channel is deterministic and the probability of having lost packets due to error or timeout is significantly lower since the multipaths can be redirected to interfere constructively at the receiver by the meta-surfaces, or the User Datagram Protocol (UDP) can be utilized with much more reliability), which play a critical part in MM procedures, or revisiting their original implementation to adapt to these programmable environments.

6.1.12. Dynamic Network Topology

In terms of user association for B5G networks, the challenge will now not be to just choose an AP with the best SINR/RSSI/RSRP/RSRQ, but it will rather be to choose or program an AP/programmable surface configuration/drone, depending on the user mobility, location and coverage from these sources. While it still reduces to the problem presented for 5G networks, the increased dimensionality and heterogeneity of the problem will provide formidable challenges to existing methods.

6.1.13. Edge Node configuration in B5G networks

Edge nodes' placement for supporting user mobility will also be challenged. This is so because the possibility of supporting better QoS over longer distances can reduce the requirements for service replication/service migration. This is a consequence of the fact that the handovers would be impacted given the programmability of the environment and the squared decay instead of a fourth power decay in the received signal power.

6.1.14. IP address continuity

The vision for near zero latency by 3GPP [123] necessitates that E2E link continuity is ensured given any network

and mobility scenario. Hence, maintaining IP address continuity during mobility events will remain a critical challenge as the complexity of the networks increases in 5G and B5G.

The aforementioned key challenges define the technology gap towards fulfilling the MM governing parameters listed in Table 2. In the following subsection we list the potential solutions that can fill this technology gap.

6.2. Potential Solutions

6.2.1. Smart CN signaling

Utilizing the properties of SDN, the signaling performed within the CN for handover and re-routing purposes can be optimized further. This will enable more scalability and better support to users with high mobility. Concretely, techniques such as graph theory, Machine Learning [124] as well as the recently established intelligent Information Elements (IE) mapping methods [13], etc., can enable faster and efficient CN signaling, as mentioned above. Here by efficiency we imply that the transmission cost, processing cost and other CN signaling related metrics [13] are reduced/optimized.

6.2.2. On demand MM

Given the functional requirements (Section 2), legacy methods (Section 4) and the state of the art (Section 5), on demand MM strategies (such as [35]) will allow future MM mechanisms to serve users with different mobility profiles, accessing different services and accessing networks with differing loads, more effectively. As an example, slice based MM strategies can enable independent strategies for the various network slices that the 5G networks will serve. This will help cater to the different network slices according to their mobility demands, and avoid the sub-optimal *one size fits all* approach.

6.2.3. Deep learning

Learning network parameters such as network load, congestion statistics at access and core network, user mobility trends, etc., enable the network to devise effective and optimal MM strategies for a highly dynamic network environment such as that in 5G and B5G networks. Hence, deep learning methods such as reinforcement learning can assist in such tasks.

6.2.4. SDN-NFV integrated DMM

DMM facilitates the distribution of MM functionality throughout the network and avoiding single MM anchors, which consequently assists in alleviating issues such as SPoF and congestion. Note that, SDN and NFV will assist in DMM as network programmability facilitates fast switching while the user/device transits through the network.

6.2.5. D2D CP-DP extension

D2D clustering and support for communication with devices in such clusters has been formalized since 3GPP Release-13. Thus, through an extension of CP-DP capabilities of the current D2D framework, i.e., by utilizing the relaying strategies for CP/DP information, handover performance for devices migrating within the network and in such clusters can be enhanced. Further, policy based methods, which take into account the presence of D2D communications between vehicles and other V2X scenarios, will also enable future MM mechanisms to serve the complex scenarios that will prevail in 5G and B5G networks better.

6.2.6. Service Continuity through Edge Computing

For serving fast moving users, such as vehicles, and satisfying their latency and bandwidth requirements, edge computing solutions for MM will play a major role in 5G and B5G networks [125]. And while service migration strategies will play a critical role in ensuring seamless connectivity, a fine balance between service replication and service migration will help mitigate the multitude of challenges that arise for such strategies. Further, given that users might crossover to other PLMNs during the duration of mobility [126], which can lead to a change in the edge cloud that serves them, effective service migration strategies will greatly enhance the QoS during mobility.

6.2.7. Clean Slate Methods

Current networks rely on resolving the IP addresses of the hosts for the applications requested by the users. However, such a resolution can lead to delays [127]. And so, Information Centric Networking (ICN), and specifically Named Data Networking (NDN) paradigm, avoid this process thus making the network more flexible and faster. Additionally, with the proposition of having in-network caching, ICN and NDN paradigms enable caching capabilities near the users.

Another class of such clean slate methods is MobilityFirst [128]. In MobilityFirst, a new paradigm to networking, like in ICN and NDN, has been proposed. In this paradigm, IP based resolution of nodes has been deprecated and name based resolution is proposed. Further, concepts similar to ICN and NDN, such as in-network caching etc., have also been proposed. Additionally, and different to the ICN-NDN paradigm, ensuring security in a fully dynamic scenario has been considered as one of the guiding principles of MobilityFirst. Further, MobilityFirst also introduces support for migration of entire networks and not just the end nodes.

Consequently, such methods together can provision more scalable, flexible and reliable MM strategies.

And so, up until now in this section, we have highlighted the multiple challenges that the 5G and beyond MM mechanism will face, given our qualitative evaluation for legacy

Table 5: Mapping potential solutions to MM challenges

Challenges	**Recommended Potential Solutions**	**Comments**	**Param. Satisfied***
Handover Signaling	Smart CN Sig. & SDN-NFV integ. DMM	In addition to the existing strategies, a smart CN signaling method, such as that in [13] will assist in relieving the handover signaling load significantly. DMM strategies will assist in decentralization of MM anchors and hence, more reliability in mobile environments	RL3, RL5, SL1 – SL4
Network Slicing	On demand MM	An on demand strategy will assist the network slices to assist in provisioning tailor made mobility solutions for the corresponding tenants	FL1, FL5
Integration framework for MM solutions	*Design*	This is a design challenge and hence, should collectively take into account all the other non-design challenges as well as other necessary factors, such as efficacy and delays	SL5
Ensuring Context Awareness	On demand MM	It will ensure that the user, network and application context is taken into account and appropriate MM solution is provisioned as and when needed	FL5
Architectural Evolution Costs	*Design*	This is a design challenge and hence, should collectively take into account all the other non-design challenges as well as other necessary factors, such as cost of infrastructure	SL5
Frequent Handovers	Deep learning	Learning the network conditions, mobility profiles and the corresponding impact on the handovers is a complex task. Deep learning can help predict/estimate valuable system parameters, such as SINR, to avoid the frequent handover condition via appropriate AP-user association	RL1, RL2, FL2, FL3, FL4
Security	Smart CN Signaling	Effective CN signaling will assist in maintaining/migrating security context when required, thus reducing the latency as well as complexity to ensure the same	RL2, SL3
Energy Efficiency	Deep learning and Smart CN Signaling	Whilst deep learning methodologies can in general provision an optimal solution for handling user mobility whilst adhering to the energy constraints, smart CN signaling, via reduction in signaling messages during mobility, can enhance energy efficiency of the MM strategy	SL1
Meta-surface Reconfiguration for mobility support	Deep learning	Based on the user mobility deep learning algorithms can assist in understanding how the meta-surface configurations have to be adjusted so as to ensure the requested QoS for the users	RL1,RL2 and FL3
B5G: Handovers	Smart CN Sig., Serv. Cont. through Edge Comp. & D2D CP-DP Ext.	Edge compute platforms can assist in faster and effective handover decisions, given their capability to provision compute power closer to the access network. Smart CN signaling can assist in efficient and low latency handover signaling in the CN. D2D networks can assist in extended coverage and hence, smoother handovers	RL2, RL4, FL3, SL1 – SL4
B5G: Protocol Stack	*Design*	This is a design challenge and hence, should collectively take into account all the other non-design challenges as well as other necessary factors, such as efficacy and delays	SL5
Dynamic Network Topology	Deep learning	The ability to understand complex associations will make deep learning methodologies essential in determining the optimal user-AP association in an increasingly dynamic and multi-dimensional network, such as the B5G networks	RL1, RL2, FL3, FL4
Edge Node Configuration in B5G Networks	*Design*	This is a design challenge and hence, should collectively take into account all the other non-design challenges as well as other necessary factors, such as efficacy and infrastructure cost	SL5
IP address continuity	Clean Slate Methods	Given their ability to resolve destinations based on names and not the IP address, clean slate methods can assist in maintaining a single IP address throughout with respect to the destination server	RL2, RL4

* Details regarding the parameters and the requirements that they help satisfy are provided in Table 2.

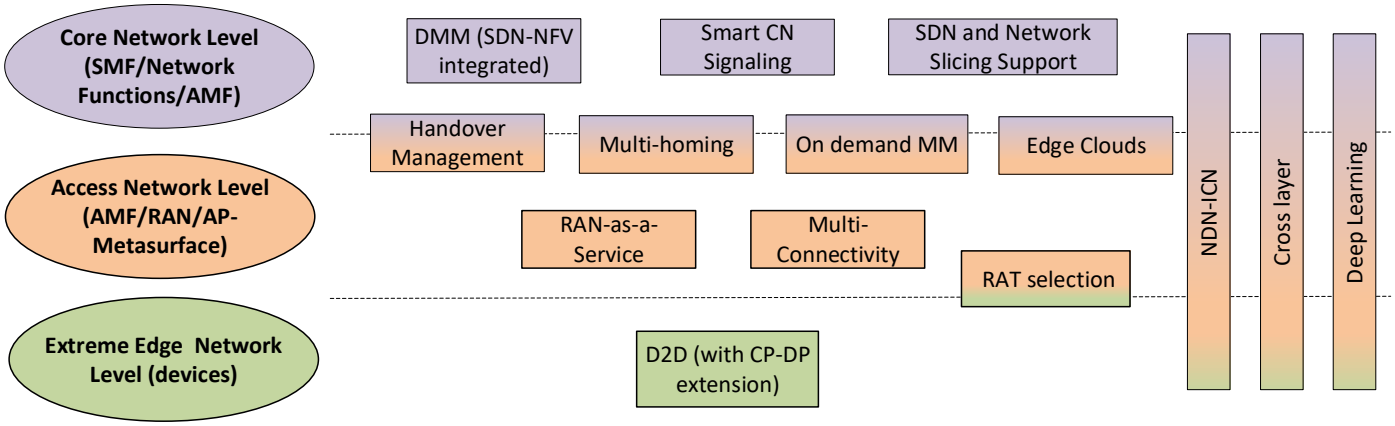

Figure 3: Proposed 5G and beyond MM framework.

and current state-of-the art methods in Sections 2-5. We have then provisioned a brief discussion on the potential solutions that can assist in addressing these challenges. We illustrate a novel mapping between these challenges and potential solutions in Table 5. Additionally, we have also listed the parameters for the qualitative analysis (and hence the requirements specified in Table 1) that they satisfy. This, as a result, reinforces the completeness of our current study. Hence, in the next subsection, utilizing the inferences from Sections 2-5 and Table 5, we propose a framework for 5G and beyond MM.

6.3. Proposed 5G and beyond MM framework

We utilize the earlier established classification process for the current state-of-the-art strategies to define our vision for 5G and beyond MM in Figure 3. Concretely, we have categorized the MM mechanisms as *Core Network level*, *Access Network level* and *Extreme Edge Network level*, depending on where they will be creating an impact on/from. The specific entities (based on the 5G architecture illustrated in Figure 2), to which these aforesaid levels correspond to, have also been mentioned in Figure 3.

To elaborate, the core network strategies encompass the DMM, SDN and Network slicing paradigms to provision the necessary reliability, flexibility and scalability from a more global perspective. Additionally, the aforesaid core network strategies need to be well complemented with an efficient CN signaling strategy. Next, handover management, on-demand MM, IPv6 multi-homing and Edge cloud related MM strategies will be enacted not only in the core network or the access network level, but jointly at both levels thus provisioning the necessary flexibility and reliability. Further, RAN-as-a-Service and Multi-connectivity provisions at the access network level will assist in utilizing the multiple RATs and APs effectively. Moreover, it is envisioned that the RAT selection process maybe either at the access network or at the device level. The D2D techniques, on the other hand, are expected to provide added assistance for mobility at the device level through DP as well CP functionality.

Complementing these mechanisms, NDN-ICN support will be provisioned at all levels, thus assisting in maintaining IP addresses/prefixes during mobility whilst resolving destinations via names. Note that, traditional IP address/prefix allocation strategies are not intended to be changed. Instead, the NDN-ICN concept provisions an over-the-top assistance. Further, the cross layer strategies, as the name suggests, will spawn across the multiple levels and enact policies, utilizing the available information at each of these levels, which assist in optimal MM related decisions across the network. Lastly, the deep learning strategies will again assist across the multiple levels by learning the complex features about the network context, user mobility and overall QoS requirements, and formulating effective MM related decisions.

Hence, given that we utilize the potential solutions for overcoming the technology gap, specified in Section 6.2, alongside certain strategies from the state of the art and legacy MM mechanisms, specified in Sections 4 and 5, it can be inferred from Tables 2-5 that our proposed framework will satisfy all the parameters for the reliability, flexibility and scalability criteria. Consequently, it can be stated that the proposed framework in Figure 3 will also satisfy all the requirements as defined in Table 1, thus provisioning a holistic solution. With this vision, in the following section we summarize the main findings of this article and conclude this paper.

7. Conclusions

Given the complexity of future network scenarios, i.e., 5G and B5G, a full view of the MM strategies, their capabilities, the persistent challenges and the possible solutions to them, will enable the research community to design better MM strategies.

In this paper, through Section 2 and Table 1, we firstly presented the important functional requirements and design criteria to be considered when devising 5G and B5G MM solution. We then presented the multiple parameters that the future MM mechanisms needs to satisfy for

each of the evalutation criteria, i.e., scalability, flexibility and reliability, in Section 3 and Table 2. Next, from our discussions in Section 4 it is clear that the legacy MM solutions fail in provisioning scalability, flexibility and reliability simultaneously. Nevertheless, the current standards and research efforts explored in Section 5 are promising as they provide enhanced capabilities towards future MM solutions. We have summarized these conclusions effectively in Tables 3 and 4. And as a consequence, through this qualitative analysis the various benefits and shortcomings of the legacy and the current state of the art mechanisms, studied in this paper, can be understood easily by the research community. Subsequently, we established that none of the mechanisms fulfill the complete 5G and beyond MM mechanism requirements.

And so, it is evident that a holistic MM mechanism for 5G and B5G networks remains elusive. Thus, certain challenges that will still persist for the design, development and deployment of future MM mechanisms have been detailed in this paper in Section 6.1. Furthermore, we have provided a concise discussion on the potential MM strategies that the research community can explore so as to solve these persistent challenges and the technological gaps they present, in Section 6.2. Following this, we have also provisioned a novel mapping between the potential strategies and the persistent challenges in Table 5, thus highlighting the efficacy of our current study. Based on the inferences drawn, we have concluded our study by provisioning a novel framework for the 5G and beyond MM strategies through Section 6.3 and Figure 3.

References

References

[1] A. Sutton, et al., Wireless Backhaul: Performance Modeling and Impact on User Association for 5G, IEEE Transactions on Wireless Communications 17 (5) (2018) 3095–3110.

[2] A. Machen, et al., Live Service Migration in Mobile Edge Clouds, IEEE Wireless Communications 25 (1) (2018) 140–147.

[3] P. A. Frangoudis, A. Ksentini, Service migration versus service replication in Multi-access Edge Computing, 14th International Wireless Communications and Mobile Computing Conference , IWCMC (2018) 124–129.

[4] H. Yang, Y. Kim, SDN-based distributed mobility management, in: Int. Conf. Inf. Netw., 2016, pp. 337–342. doi:10.1109/ICOIN.2016.7427127.
URL http://ieeexplore.ieee.org/document/7427127/

[5] A. Boulogeorgos, et al., Terahertz Technologies to Deliver Optical Network Quality of Experience in Wireless Systems beyond 5G, IEEE Communications Magazine 56 (6) (2018) 144–151.

[6] M. Z. Chowdhury, et al., Optical Wireless Hybrid Networks for 5G and beyond Communications, 9th International Conference on Information and Communication Technology Convergence, ICTC (2018) 709–712.

[7] Z. Zhang, et al., 6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies, IEEE Vehicular Technology Magazine.

[8] E. Basar, et al., Wireless Communications Through Reconfigurable Intelligent Surfaces (June) (2019) 1–20. arXiv:1906.09490.
URL http://arxiv.org/abs/1906.09490

[9] M. D. Renzo, et al., Smart radio environments empowered by reconfigurable AI meta-surfaces: an idea whose time has come, Eurasip Journal on Wireless Communications and Networking (1). arXiv:arXiv:1903.08925v1.

[10] Y. L. Chung, L. J. Jang, Z. Tsai, An efficient downlink packet scheduling algorithm in LTE-Advanced systems with Carrier Aggregation, 2011 IEEE Consumer Communications and Networking Conference (2011) 632–636.

[11] I. Parvez, et al., A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions, IEEE Communications Surveys and Tutorials.

[12] S. Sekander, H. Tabassum, E. Hossain, Multi-Tier Drone Architecture for 5G/B5G Cellular Networks: Challenges, Trends, and Prospects, IEEE Communications Magazine 56 (3) (2018) 104–111.

[13] A. Jain, et al., Evolutionary 4G/5G Network Architecture Assisted Efficient Handover Signaling, IEEE Access 7 (2019) 256–283.

[14] I. F. Akyildiz, et al., SoftAir: A software defined networking architecture for 5G wireless systems, Comput. Networks 85 (2015) 1–18. doi:10.1016/j.comnet.2015.05.007.

[15] S. Andreev, et al., Intelligent access network selection in converged multi-radio heterogeneous networks, IEEE Wirel. Commun. 21 (6) (2014) 86–96.

[16] P. Fan, J. Zhao, C.-L. I, 5G high mobility wireless communications: Challenges and solutions, China Communications 13 (2) (2016) 1–13. doi:10.1109/cc.2016.7405718.

[17] S. Ferretti, et al., A survey on handover management in mobility architectures, Comput. Networks 94 (2016) 390–413.

[18] M. Zekri, et al., A review on mobility management and vertical handover solutions over heterogeneous wireless networks, Comput. Commun. 35 (17) (2012) 2055–2068.

[19] 3GPP, 5G; System architecture for the 5G System (5GS) (3GPP TS 23.501 version 15.8.0 Release 15) (2020) 1–251.
URL https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[20] P. Rost, et al., Mobile network architecture evolution toward 5G, IEEE Commun. Mag. 54 (5) (2016) 84–91.

[21] I. F. Akyildiz, et al., 5G roadmap: 10 key enabling technologies, Comput. Networks 106 (2016) 17–48.

[22] S. E. Elayoubi, et al., 5G service requirements and operational use cases: Analysis and METIS II vision, EUCNC 2016 - Eur. Conf. Networks Commun. (2016) 158–162.

[23] ITU, IMT Vision Framework and overall objectives of the future development of IMT for 2020 and beyond, M Series, Recommendation ITU-R M.2083-0 (09/2015).

[24] M. A. Habibi, et al., A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System, IEEE Access 7 (2019) 70371–70421.

[25] W. Khawaja, et al., A Survey of Air-to-Ground Propagation Channel Modeling for Unmanned Aerial Vehicles, IEEE Communications Surveys & Tutorials 21 (3) (2019) 2361 – 2391.

[26] B. Li, Z. Fei, Y. Zhang, UAV communications for 5G and beyond: Recent advances and future trends, IEEE Internet of Things Journal 6 (2) (2019) 2241–2263.

[27] H. Wymeersch, et al., 5G mmwave positioning for vehicular networks, IEEE Wireless Communications 24 (6) (2017) 80–86.

[28] M. Jaber, et al., 5G Backhaul Challenges and Emerging Research Directions: A Survey, IEEE Access 4 (2016) 1743–1766.

[29] D. Liu, H. Chan, RFC 7429 Distributed Mobility Management: Current Practices and Gap Analysis (2015) 1–34.

[30] C. Chen, et al., Mobility management for low-latency handover in SDN-based enterprise networks, in: IEEE WCNC, 2016, pp. 1–6.

[31] R. Urgaonkar, et al., Dynamic service migration and workload scheduling in edge-clouds, Performance Evaluation 91 (2015) 205–228.

[32] S. Wang, J. Xu, N. Zhang, Y. Liu, A Survey on Service Migra-

tion in Mobile Edge Computing, IEEE Access 6 (2018) 23511–23528.

[33] R. I. Rony, et al., Joint access-backhaul perspective on mobility management in 5G networks, in: 2017 IEEE Conf. Stand. Commun. Netw., 2017, pp. 115–120.

[34] T. Bai, R. W. Heath, Coverage analysis for millimeter wave cellular networks with blockage effects, 2013 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013 - Proceedings (2013) 727–730.

[35] A. Jain, et al., Mobility Management as a Service for 5G Networks, in: IEEE ISWCS 2017 Work., 2017, pp. 1–6. arXiv:1705.09101.
URL http://arxiv.org/abs/1705.09101

[36] L. Zanzi, V. Sciancalepore, On Guaranteeing End-to-End Network Slice Latency Constraints in 5G Networks, Proceedings of the International Symposium on Wireless Communication Systems (2018) 1–6doi:10.1109/ISWCS.2018.8491249.

[37] R. Molina-Masegosa, J. Gozalvez, LTE-V for Sidelink 5G V2X Vehicular Communications, IEEE Veh. Technol. Mag. (2017) 30–39.

[38] A. Jain, et al., Enhanced Handover Signaling through Integrated MME-SDN Controller Solution, in: 2018 IEEE 87th Veh. Technol. Conf. (VTC Spring), 2018, pp. 1–7.

[39] A. Ford, et al., Architectural Guidelines for Multipath TCP Development, RFC 6182 (2011) 1–28.

[40] A. Ford, et al., TCP Extensions for Multipath Operation with Multiple Addresses, RFC 6824 (2013) 1–64.

[41] R. Stewart, Stream Control Transmission Protocol, RFC 4960 (2007) 1–152arXiv:arXiv:1011.1669v3, doi:10.1017/CBO9781107415324.004.

[42] G. Hampel, T. Klein, Enhancements to Improve the Applicability of Multipath TCP to Wireless Access Networks, IETF (2011) 1–25.

[43] S. Zannettou, M. Sirivianos, F. Papadopoulos, Exploiting path diversity in datacenters using MPTCP-aware SDN, Proceedings - IEEE Symposium on Computers and Communications (2016) 539–546arXiv:1511.09295, doi:10.1109/ISCC.2016.7543794.

[44] C. D. Phung, et al., MPTCP robustness against large-scale man-in-the-middle attacks, Computer Networks 164 (2019) 106896. doi:10.1016/j.comnet.2019.106896.
URL https://doi.org/10.1016/j.comnet.2019.106896

[45] Y. Liu, A. Neri, A. Ruggeri, A. M. Vegni, A MPTCP-Based Network Architecture for Intelligent Train Control and Traffic Management Operations, IEEE Trans. Intell. Transp. Syst. 18 (9) (2017) 2290–2302. doi:10.1109/TITS.2016.2633531.

[46] P. Natarajan, F. Baker, C. Systems, P. D. Amer, J. T. Leighton, SCTP : What , Why , and How, IEEE Internet Comput. 13 (5) (2009) 81–85.

[47] C. Raiciu, M. Handly, D. Wischik, Coupled Congestion Controol for Multipath Transport Protocols, IETF RFC6356 (2011) 1–12arXiv:arXiv:1011.1669v3, doi:10.1017/CBO9781107415324.004.

[48] D. Wischik, et al., Design, implementation and evaluation of congestion control for multipath TCP, Proceedings of NSDI 2011: 8th USENIX Symposium on Networked Systems Design and Implementation (2011) 99–112.

[49] P. Ignaciuk, M. Morawski, Discrete-time MPTCP flow control for channels with diverse delays and uncertain capacity, 2018 22nd International Conference on System Theory, Control and Computing, ICSTCC 2018 - Proceedings (2018) 722–727doi:10.1109/ICSTCC.2018.8540742.

[50] X. Wei, C. Xiong, E. Lopez, MPTCP proxy mechanisms, Internet-Draft draft-wei-mptcp-proxy-mechanism-02, Internet Engineering Task Force, work in Progress (Jun. 2015).
URL https://datatracker.ietf.org/doc/html/draft-wei-mptcp-proxy-mechanism-02

[51] A. de la Oliva, et al., An Overview of IEEE 802.21: Media-Independent Handover Services, IEEE Wireless Communications (Aug.) (2008) 96–103.

[52] L. Eastwood, et al., Mobility Using IEEE 802.21 in a Heterogeneous IEEE 802.16/802.11-based, IMT-Advanced (4G) Network, IEEE Wireless Communications (Apr.) (2008) 26–34.

[53] IEEE, IEEE 802.21c-2014: IEEE Standard for Local and metropolitan area networks Part 21 : Media Independent Handover Services Amendment 3: Optimized Single Radio Handovers, no. June, 2014. doi:10.1109/IEEESTD.2012.6198737.

[54] J.-S. Wu, S.-F. Yang, B.-J. Hwang, A terminal-controlled vertical handover decision scheme in IEEE 802.21-enabled heterogeneous wireless networks Jung-Shyr, Int. J. Commun. Syst. 22 (2009) 819–834. doi:10.1002/dac.

[55] S. Gundavelli, et al., Proxy Mobile IPv6, RFC 5213 (2008) 1–92arXiv:arXiv:1011.1669v3, doi:10.1017/CBO9781107415324.004.

[56] C. Bernados, Proxy Mobile IPv6 Extensions to Support Flow Mobility, RFC 7864 (2016) 1–19arXiv:arXiv:1011.1669v3, doi:10.1017/CBO9781107415324.004.

[57] 3GPP, Universal Mobile Telecommunications System (UMTS); LTE; Proxy Mobile IPv6 (PMIPv6) based Mobility and Tunnelling protocols; Stage 3 (3GPP TS 29.275 version 8.6.0 Release 8) (2010) 1–73.

[58] H. N. Nguyen, C. Bonnet, Scalable proxy mobile IPv6 For heterogeneous wireless networks, Proceedings of the International Conference on Mobile Technology, Applications, and Systems, Mobility'08doi:10.1145/1506270.1506352.

[59] F. Giust, et al., Distributed mobility management for future 5G networks: overview and analysis of existing approaches, IEEE Commun. Mag. 53 (1) (2015) 142–149.

[60] M. I. Sanchez, et al., Experimental evaluation of an SDN-based distributed mobility management solution, in: Proc. Work. Mobil. Evol. Internet Archit. - MobiArch '16, 2016, pp. 31–36.

[61] 3GPP, LTE; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (3GPP TS 23.401 version 13.9.0 Release 13), Tech. rep. (2017).

[62] C. B. Sankaran, Data offloading techniques in 3GPP Rel-10 networks: A tutorial, IEEE Communications Magazine 50 (6) (2012) 46–53. doi:10.1109/MCOM.2012.6211485.

[63] E. Dahlman, et al., 4G: LTE Advanced Pro and the Road to 5G, 3rd Edition, Academic Press, 2016.

[64] 3GPP, ETSI, 3GPP TS 36.300 version 13.2.0 Release 13 LTE; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 Terrestrial Radio Access (E-UTRA) Overall description, Tech. rep. (2013).

[65] D. P. Ibarra Barreno, LTE / WIFI AGGREGATION IMPLEMENTATION AND EVALUATION, Ph.D. thesis, UPC (2017).

[66] S. Oh, B. Ryu, Y. Shin, EPC signaling load impact over S1 and X2 handover on LTE-Advanced system, 2013 3rd World Congress on Information and Communication Technologies, WICT 2013 (2014) 183–188doi:10.1109/WICT.2013.7113132.

[67] D. Astely, E. Dahlman, G. Fodor, S. Parkvall, J. Sachs, LTE Release 12 and Beyond David, Ieee Wirel. Commun. Mag. (July) (2013) 154–160. doi:10.1109/MWC.2009.5361183.

[68] ITU-T, Framework of vertical multihoming in IPv6-based next generation networks.

[69] R. Irmer, et al., Coordinated multipoint: Concepts, performance, and field trial results, Communications Magazine (2011) 102–112doi:10.1109/MCOM.2011.5706317.
URL http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=5706317

[70] W. Sun, J. Liu, Coordinated multipoint-based uplink transmission in internet of things powered by energy harvesting, IEEE Internet Things J. 5 (4) (2018) 2585–2595. doi:10.1109/JIOT.2017.2782745.

[71] J. Lee, Y. Kim, H. Lee, B. Ng, D. Mazzarese, J. Liu, W. Xiao, Y. Zhou, Coordinated multipoint transmission and reception in LTE-advanced systems, IEEE Commun. Mag. 50 (11) (2012) 44–50. doi:10.1109/MCOM.2012.6353681.

[72] 3GPP, 3gpp TS 36.331 – 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved

Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 10) The (June).

[73] D. Xenakis, et al., Mobility management for femtocells in LTE-advanced: Key aspects and survey of handover decision algorithms, IEEE Communications Surveys and Tutorials 16 (1) (2014) 64–91. doi:10.1109/SURV.2013.060313.00152.

[74] C. Shen, M. Van Der Schaar, A learning approach to frequent handover mitigations in 3GPP mobility protocols, IEEE Wirel. Commun. Netw. Conf. WCNC (2017) 1–6doi:10.1109/WCNC.2017.7925950.

[75] A. Ahmed, L. M. Boulahia, D. Gaïti, Enabling vertical handover decisions in heterogeneous wireless networks: A state-of-the-art and a classification, IEEE Commun. Surv. Tutorials 16 (2) (2014) 776–811. doi:10.1109/SURV.2013.082713.00141.

[76] J. Zhao, A Survey of Reconfigurable Intelligent Surfaces: Towards 6G Wireless Communication Networks with Massive MIMO 2.0 (2019) 1–7arXiv:1907.04789.
URL http://arxiv.org/abs/1907.04789

[77] 3GPP, 5G; Procedures for the 5G System (5GS) (3GPP TS 23.502 version 15.8.0 Release 15) (2020) 1–362.

[78] 3GPP, 5G; NR; Overall description; Stage-2 (3GPP TS 38.300 version 15.8.0 Release 15) 1 (2020) 1–102.
URL https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[79] 3GPP, Universal Mobile Telecommunications System (UMTS); LTE; 5G; NR; Multi-connectivity; Overall description; Stage-2 (3GPP TS 37.340 version 15.5.0 Release 15), Tech. rep. (2019). doi:10.1002/9781118256114.ch46.

[80] S. Jung, J. Kim, A new way of extending network coverage: Relay-assisted D2D communications in 3GPP, ICT Express 2 (3) (2016) 117–121. doi:10.1016/j.icte.2016.08.001.

[81] M. Kantor, et al., A policy-based per-flow mobility management system design, in: Proc. Princ. Syst. Appl. IP Telecommun. - IPTComm '15, 2015, pp. 35–42.

[82] M. Gramaglia, et al., Definition of connectivity and QoE / QoS management mechanisms 5G Norma deliverable D5.1.

[83] G. Schütz, A k-Cover Model for Reliability-Aware Controller Placement in Software-Defined Networks, in: Computational Science – ICCS 2019, Springer International Publishing, 2019, pp. 604–613.

[84] S. Kuklinski, et al., Handover management in SDN-based mobile networks, in: 2014 IEEE Globecom Work. (GC Wkshps), 2014, pp. 194–200.

[85] F. Meneses, C. Guimares, D. Corujo, R. L. Aguiar, SDN-based Mobility Management: Handover Performance Impact in Constrained Devices, in: 2018 9th IFIP Int. Conf. New Technol. Mobil. Secur., IEEE, 2018, pp. 1–5. doi:10.1109/NTMS.2018.8328716.
URL http://ieeexplore.ieee.org/document/8328716/

[86] T. D. Assefa, et al., SDN-based local mobility management with X2-interface in femtocell networks, IEEE Int. Work. Comput. Aided Model. Des. Commun. Links Networks, CAMAD (2017) 3–8doi:10.1109/CAMAD.2017.8031628.

[87] S. Basloom, N. Akkari, Mobility Management in SDN and NFV-based Next- Generation Wireless Networks : An Overview and Qualitative Evaluation, 2018 1st International Conference on Advanced Research in Engineering Sciences (ARES) 1–8.

[88] T.-T. Nguyen, et al., SDN-based distributed mobility management for 5G networks, in: IEEE WCNC., 2016, pp. 1–7.

[89] I. Elgendi, K. S. Munasinghe, A. Jamalipour, A Three-Tier SDN based distributed mobility management architecture for DenseNets, 2016 IEEE Int. Conf. Commun. ICC 2016doi:10.1109/ICC.2016.7511020.

[90] D. Battulga, et al., Handover Management for Distributed Mobility Management in SDN-based Mobile Networks, 27th Int. Telecommun. Networks Appl. Conf.

[91] Q. Li, et al., Edge Cloud and Underlay Networks: Empowering 5G Cell-Less Wireless Architecture, in: 20th Eur. Wirel. Conf., 2014, pp. 1–6.

[92] R. Architecture, Mobile Edge Computing (MEC): End to End Mobility Aspects, Etsi 1 (2016) 1–18.

[93] A. Mtibaa, et al., Towards edge computing over named data networking, Proceedings - 2018 IEEE International Conference on Edge Computing (2018) 117–120.

[94] P. Mach, Z. Becvar, Mobile Edge Computing: A Survey on Architecture and Computation OffloadingarXiv:1702.05309, doi:10.1109/COMST.2017.2682318.
URL http://arxiv.org/abs/1702.05309{%}0Ahttp://dx.doi.org/10.1109/COMST.2017.2682318

[95] ETSI, MEC in 5G networks, White Paper (28) (2018) 1–28.

[96] T. Nakamura, et al., Trends in small cell enhancements in LTE advanced, IEEE Commun. Mag. 51 (2) (2013) 98–105.

[97] F. A. A. Emam, M. E. Nasr, S. E. Kishk, Coordinated Handover Signaling and Cross-Layer Adaptation in Heterogeneous Wireless Networking, Mob. Networks Appl. 25 (2020) 285–299.

[98] F. A. A. Emam, M. E. Nasr, S. E. Kishk, Context-aware parallel handover optimization in heterogeneous wireless networks, Ann. Telecommun. 75 (2020) 43–57.

[99] A. Al-rubaye, J. Seitz, A Cross-Layer Mobility Management with Multi-Criteria Decision Making, 2016 Eighth Int. Conf. Ubiquitous Futur. Networks (2016) 821–826doi:10.1109/ICUFN.2016.7537152.

[100] N. Nikaein, et al., Demo Closer to Cloud-RAN : RAN as a Service, ACM Mobicom (2015) 193–195.

[101] A. Outtagarts, et al., When IT meets Telco : RAN as a Service, 2015 IEEE/ACM 8th Int. Conf. Util. Cloud Comput. (2015) 422–423doi:10.1109/UCC.2015.75.

[102] D. Sabella, et al., RAN as a Service: Challenges of Designing a Flexible RAN Architecture in a Cloud-based Heterogeneous Mobile Network, 2013 Futur. Netw. Mob. Summit (2013) 1–8.

[103] L. Liu, et al., Analysis of Handover Performance Improvement in Cloud-RAN Architecture, 7th Int. Conf. Commun. Netw. China (2012) 850–855doi:10.1109/ChinaCom.2012.6417603.

[104] V. Passast, et al., Dynamic RAT Selection and Pricing for Efficient Traffic Allocation in 5G HetNets, IEEE ICC (2019) 1–6doi:10.1109/ICC.2019.8761831.

[105] S. Goudarzi, et al., A hybrid intelligent model for network selection in the industrial Internet of Things, Appl. Soft Comput. J. 74 (2019) 529–546. doi:10.1016/j.asoc.2018.10.030.

[106] X. Wang, et al., Intelligent User-Centric Network Selection : A Model-Driven Reinforcement Learning Framework, IEEE Access 7. doi:10.1109/ACCESS.2019.2898205.

[107] D. Calabuig, et al., Resource and Mobility Management in the Network Layer of 5G Cellular Ultra-Dense Networks, IEEE Commun. Mag. 55 (6) (2017) 162–169.

[108] A. Jain, E. Lopez-Aguilera, I. Demirkol, User association and resource allocation in 5g (aura-5g): A joint optimization framework (2020). arXiv:2003.10605.

[109] O. N. C. Yilmaz, et al., Smart mobility management for D2D communications in 5G networks, 2014 IEEE Wirel. Commun. Netw. Conf. Work. WCNCW 2014 (2014) 219–223doi:10.1109/WCNCW.2014.6934889.

[110] K. Ouali, B. Kervella, An Efficient D2D Handover Management Scheme for SDN-based 5G networks, 2020 IEEE 17th Annu. Consum. Commun. Netw. Conf. (2020) 1–6.

[111] R. Klempous, J. Nikodem, Smart Innovations in Engineering and Technology, Vol. 15, 2020. doi:10.1007/978-3-030-32861-0.
URL http://link.springer.com/10.1007/978-3-030-32861-0

[112] S. Barua, R. Braun, Mobility management of D2D communication for the 5G cellular network system: A study and result, 2017 17th Int. Symp. Commun. Inf. Technol. Isc. 2017 (2017) 1–6doi:10.1109/ISCIT.2017.8261187.

[113] S. Barua, R. Braun, A novel approach of mobility management for the D2D communications in 5G mobile cellular network system, in: 2016 18th Asia-Pacific Netw. Oper. Manag. Symp., 2016, pp. 1–4. doi:10.1109/APNOMS.2016.7737272.
URL http://ieeexplore.ieee.org/document/7737272/

[114] A. Santoyo-González, C. Cervelló-Pastor, Latency-aware cost optimization of the service infrastructure placement in 5G networks, J. Netw. Comput. Appl. 114 (2018) 29–37. doi:10.1016/j.jnca.2018.04.007.
URL https://doi.org/10.1016/j.jnca.2018.04.007

[115] I. Leyva-Pupo, A. Santoyo-González, C. Cervelló-Pastor, A framework for the joint placement of edge service infrastructure and user plane functions for 5G, Sensors (Switzerland) 19 (18). doi:10.3390/s19183975.

[116] 3GPP, TS 23.502: Procedures for the 5G System (Stage 2), Tech. Rep. Release-15 (2017).

[117] R. Ratasuk, N. Mangalvedhe, A. Ghosh, LTE in unlicensed spectrum using licensed-assisted access, 2014 IEEE Globecom Workshops, GC Wkshps 2014 (2014) 746–751.

[118] R. Alkhansa, H. Artail, D. M. Gutierrez-Estevez, LTE-WiFi carrier aggregation for future 5G systems: A feasibility study and research challenges, Procedia Computer Science 34 (2014) 133–140.

[119] M. A. Ferrag, et al., Security for 4G and 5G Cellular Networks: A Survey of Existing Authentication and Privacy-preserving Schemes (2017) 1–24.
URL http://arxiv.org/abs/1708.04027

[120] M. Jawad Alam, M. Ma, DC and CoMP Authentication in LTE-Advanced 5G HetNet, IEEE Global Communications Conference (GLOBECOM) 2017 (2018) 1–6. .

[121] G. Qiao, et al., Joint Deployment and Mobility Management of Energy Harvesting Small Cells in Heterogeneous Networks, IEEE Access 5 (2017) 183–196.

[122] A. Habbal, et al., Context-aware Radio Access Technology Selection Approach in 5G Ultra Dense Networks, IEEE Access 5 (2017) 6636 – 6648.

[123] 3GPP, TS22.261: Service requirements for the 5G system (Stage 1) (2018).

[124] A. Sadeghian, et al., Semantic Edge Labeling over Legal Citation Graphs, in: LTDCA, 2018.

[125] M. Boban, et al., Use Cases, Requirements, and Design Considerations for 5G V2X (2017) 1–10.
URL http://arxiv.org/abs/1712.01754

[126] ETSI, GR MEC 022 V2.1.1: Study on MEC Support for V2X Use Cases (2018) 1–19.

[127] L. Zhang, et al., Named Data Networking, ACM SIGCOMM Computer Communication Review 44 (3) (2014) 66–73.

[128] D. Raychaudhuri, et al., MobilityFirst : A Robust and Trustworthy Mobility- Centric Architecture for the Future Internet, ACM SIGMobile Mob. Comput. Commun. Rev. (2012) 1–12. .

View publication stats