Distributed cloud and de-centralized control plane
A proposal for scalable control plane for 5G

Amir Roozbeh
Ericsson Research, Cloud Technology
Stockholm, Sweden
Email: amir.roozbeh@ericsson.com

Abstract—5G is the next generation of mobile network. The aim is to launch service starting in 2020. There are many requirements on 5G, such as high capacity, low latency, flexibility, and support for any-to-any communication. Cloud technology, in the form of a distributed cloud (also known as a network embedded cloud), is an enabler technology for 5G by providing flexible networks to meet different user application requirements. On the other hand, Machine Type Communication (MTC) is a primary application for 5G, but it can add a high volume of control signaling. To manage the expected high volume of control signaling introduced by MTC, we identified the main control events that generate signaling messages in the network, i.e., session management, hand over management, and tracking area update. Then, we proposed a decentralized core network architecture optimized for the identified control events. The proposed control plane functions are independent of each other in the sense that each function can be executed separately. The control functions can utilize a distributed cloud (embedded in the 5G core network) to manage the enormous amount of control signaling by handling this signaling locally, i.e., close to the end user. Additionally, we present an analysis of the control signaling performance for each proposed control function. The result shows that it is beneficial to move session management to data centers collocated with the base station in the 5G network when there is high user density.

Keywords: 5G, decentralized control function, core network, network enabled cloud

I. INTRODUCTION

Mobile communications have had a significant impact on the way that people and devices interconnect. Historically there has been an evolution of wireless technology in terms of radio access technology, data rates, operational bandwidth, and mobile network architecture every ten years [1].

First generation mobile communication aimed at connecting people by combining communications and mobility in one package utilizing a circuit switch network in 1980. The (r)evolution in mobile communication began with adapting the mobile architecture to support packet switch networking. This lead to the second generation 2G (i.e., introduced in 1990s), such as Global System for Mobil Communications (GSM), and later 3rd generation (3G) mobile networks. 2G introduced increased capacity and coverage, enabling additional services beyond voice, such as SMS and Email. Later, 3G introduced mobile broadband (i.e., up to 2 Mbps) by combining high-speed mobile access with IP services. In recent years, wireless technologies (r)evolution resulted in an all-IP architecture, i.e., fourth generation (4G) together with Long Term Evolution Advanced (LTE Advanced). The term all-IP refers to the fact that all data and signaling within the mobile network utilize the Internet Protocol (IP) on the network layer.

Today there are wide ranges of requirements on the next generation of mobile communication, i.e., fifth generation (5G). These requirements include: higher data rates and increased traffic capacity (i.e., 1000 time more capacity than 4G), improved reliability, support for enormous numbers of devices, and lower latency. As a result, 5G has to be more efficient and scalable than 4G [2]. 5G is concerned not only with providing a service for people but also serving any devices that may benefit from being connected. In another word, the next generation of mobile communication should enable anyone or anything to access information and share data anywhere, anytime [3].

IOT is one of the important applications for 5G. In the transition to a networked society and connected planet with 5G, a massive number of devices, which can be embedded into the environment, will be connected to networks. The concept of IOT combines different Machine-Type Communications (MTC) (i.e., Machine-to-Machine and Machine-to-Human) with human communication (i.e., Human-to-Human and Human-to-Machine). Wearable devices, smart homes, smart parking places, smart and self-driving cars, and smart sensors are some example applications of MTC. IOT devices can be categorized based on many factors such as their portability (i.e., fixed or mobile) or communication direction (i.e., one-way or two-way communication). Most IOT devices are characterized by their simplicity, the fact that they usually transmit small amounts of data, and whether they transmit frequent or only from time to time.

5G has a requirement to support hundreds of billions of low-power connections in order to bring IOT to peoples daily lives. The enormous number of IOT devices and IOT application characteristics in a future communication ecosystem will have an impact on future mobile communication architectures. IOT and MTC will effect both the user plane and control plane traffic that need to be handled by future 5G networks. Management of billions of IOT devices is expected to add considerable signaling loads on the control plane of the network, but this may lead to higher response times and increased congestion at network control plane entities. The future 5G control plane should be designed to avoid congestion and reduce latency of the control plane, while providing a flexible framework which can adapt to user requirements in a wide variety of different scenarios.
In this paper, we present a decentralized control plane function for 5G to provide the scalability needed for IOT control signaling traffic. In this proposal, the mobile network control plane is divided into different functions based on the primary control events expected in a mobile network. Each of the proposed functions can execute in a decentralized or centralized way, depending upon the network’s demands to handle signaling traffic.

The reminder of this paper is organized as follows: Section II presents an overview of LTE network and 5G characteristics and concept. Section III describes the concept of IOT and how this can effect signaling load. Section IV gives a description of the proposed control plane architecture for 5G. Section V gives an analytic model used to calculate signaling load on different control plane functions in 5G. Additionally, a numerical analysis is made based on the analytic model and metrics from an LTE network as a baseline of 5G. Finally, we conclude the paper in Section VII.

II. FIFTH GENERATION (5G)

The 5th generation of mobile network (5G) is in its early stage and lack a standard architecture and protocol. As a result, this paper uses Long Term Evolution (LTE) technology as a baseline to identify the main control plane events that produce signaling load. The assumption is that a future 5G architecture will evolve from today's All-IP LTE. In this section we provide overview about the LTE network followed by a description of the expected 5G characteristics and underlying concepts.

A. LTE NETWORK

The LTE network architecture separates the control plane and user plane (or data plane) [4]. The control plane is responsible for control and transmission of signaling information, while the user plane is responsible for forwarding user traffic. This flat architecture has two parts: the Evolved-Universal Terrestrial Radio Access Network (E-UTRAN) and Evolved Packet Core (EPC). E-UTRAN consists of evolved node-Bs (eNodeBs) that are responsible for providing wireless connectivity to the user entities (UEs). EPC consists of four primary entities: Mobility Management Entity (MME), Home Subscriber Server (HSS), Packet Data Network Gateway (P-GW), and serving gateway (S-GW). EPC provides functions for user plane and control signaling management. Fig. 1 shows the LTE network architecture. The EPC also includes standard component of IP networks, such as switches, DNS, and NTP servers (although these are not shown in this figure).

An MME is a critical control node in the EPC and supports the most important control plane functions, specifically mobility management managing security when device attaches to the access network, and tracking and paging to support devices in idle mode. MME uses information provided by HSS to authenticate devices and updates UE’s location information in HSS as part of its mobility management.

HSS is a database that contains subscriber-related information. HSS provides support functions related to call and session setup, user authentication, access authorization, and mobility management.

The P-GW and S-GW of the core network act as a gateway for user plane. These two elements provide connectivity between the UE and external IP networks. The S-GW is logically connected to a P-GW and provides the interconnection between the radio access network and EPC. The P-GW interconnects the EPC with external IP networks. The P-GW offers several functions, such as IP address management, policy control, and charging.

LTE network elements communicate with each other by using standard interfaces defined by the 3rd Generation Partnership Project (3GPP). For example, eNodeBs are interconnected with each other by means of the X2 interface [5], while S10 is the standard interface between two MMEs. There are four main events that initiate signaling control in LTE network: a UE-originated session, a UE-terminated session, Handover (HO), and Tracking Area Update (TAU).

A UE-originated session occurs when a UE attempts to establish a connection to the EPC in order to receive or send data. The signaling messages involved in this procedure are shown in Fig. 2. To initiate the communication link, UE starts the Radio Resource Control (RRC) connection procedure with an eNodeB to ask for resources from access network, (in step 1). In step 2, the UE sends an ATTACH and PDN CONNECTIVITY request to MME to set up an internet connection (this is referred to as a Non-Access Stratum (NAS) service). If the MME is unable to identify the user (for example because the UE has just at powered on or at the time initial access to the network), then the MME starts identification and authorization as NAS common procedure (in step 3). This process involves six messages between the UE and MME. The MME updates the HSS with the UE’s location while asking for this subscriber’s profile. Also, there is a negotiation between MME and S-GW and P-GW to establish an initial bearer for this UE (in steps 4-7). Next the MME sends a context setup request to the eNodeB that results in a signaling message being sent between the eNodeB and UE to establish a bearer between these two entities (in steps 8-10). When the initial context configuration is satisfactory, the MME configure the user plane of S-GW and sends an ‘attach accept” to UE (in steps 11-15). At this point, the data flow between the UE and external network is established. Finally, when the session is complete the MME can release all bearers allocated for this session (in steps 16-21).

A UE-terminated session occurs when the network has data for an idle UE. In this situation, the MME starts by paging all
eNodeBs (within a tracking area size that UE registered with it) in order to notify the UE that there is pending data for it. The remainder of the procedure is same as for UE-originated session, as shown in Fig. 3.

Handover (HO) can occur when the UE is idle or in connected mode. We ignore those cases when HO occurs when UE is in idle mode or when a HO occurs within a single eNodeB (i.e., when a UE changes its cell sector), as in both of these cases the MME is not involved in the signaling procedure [6]. During a HO the UE may stay with a given MME (i.e., an intra-MME HO) or need to change its MME (i.e., an inter-MME HO). Intra-MME HO occurs when an MME serves multiple eNodeBs. In this scenario, a UE can continue to use its S-GW (i.e., intra-MME/SGW HO) or changes its S-GW (i.e., intra-MME/inter-SGW HO). The source eNodeB (SeNB) triggers a HO based on measurements reported by a UE. A summary of intra-MME/SGW HO signaling flows and a description of this HO procedure is depicted in Fig. 4. These scenarios assume that the a X2 interface between source and destination eNodeB exists\(^*\).

The intra-MME/inter-SGW HO signaling flows contain two additional pairs of signaling messages. The first pair between MME and source S-GW (SSGW) release the bearer, while the second pairs between MME and target S-GW (TSGW) establish the bearer.

The inter-MME HO happens when more than one MME is available in the network. Upon HO the UE can stay with a given S-GW (i.e., inter-MME/intra-SGW HO) or changes its S-GW (i.e., inter-MME/SGW HO). The source MME (S-MME) controls the source eNodeB (SeNB) and source S-GW (S-SGW) [if applicable], while the target MME (T-MME) controls target eNodeB (T-eNB) and target S-GW (T-SGW) [if applicable]. The message flow and description of the procedure for inter-MME HO is depicted in Fig. 5. These scenarios assume that the HO initiation needs to use the S2 interface.

Tracking Area Update (TAU) happens when a UE detects that it has entered a new tracking area. If so, the UE updates the network with this new tracking area information. If this new tracking area is served by same MME, then this MME accepts the TAU and register a new location for this UE without interaction with HSS. However, if the serving MME changes (i.e., inter-MME TAU), then the network forwards the TAU to target MME. Next, the target MME updates the HSS database with this new location information and the identity of the target MME. The HSS cancels the UE’s location in source MME and send the UE’s subscription data to the target MME. The Inter-MME TAU message flow is depicted in Fig. 6.

B. 5G

The 5th generation (5G) is the next generation of wireless technology and is expected to be deployed in the post 2020 timeframe. Currently, there is a global discussion on the

---

*If this interface does not exist, then the HO procedure is very similar to the inter-MME HO except for the involvement of an MME coordinating the HO signaling loads.
definition of 5G [7]–[9]. NO recognized standards body has yet defined 5G. There are several proposals for 5G network architectures in literature [3], [14], [15]. These proposals generally assume a network upgrade from an All-IP network architecture, and utilize cloud technology and variety of different radio access technologies.

There are many factors that are driving the development of 5G such as: multi-gigabit per second (Multi-Gbps), Internet of Things (IOT), ubiquitous access to mobile services, cloud technology, software defined networks (SDNs), and increasing service complexity (e.g., 3D video and gaming, augmented reality, and self-driving cars). Study shows that telecommunication services are increasingly to be virtualized and will migrate to the cloud [10], [11]. While previous generations of mobile networks were communication centric, 5G is about both communication and computing. As a result 5G is expected to extend the continuing revolution of moving the cloud to the network in the form of a network embedded cloud.

To achieve higher data capacity, cell deployment will be denser, i.e., gradually increasing up to ten times the density of deployment in today’s networks [2]. Additionally, cooperative communication and joint processing techniques, such as coordinated multi-point (CoMP), massive multiple-input-multiple-output (MIMO), and joint radio resource scheduling can be utilized to provide higher data rates. These methods introduce additional cost due to interfering links and intensive data exchange and computation. Additionally, the 5G core network must be able to handle all of the control and user plane traffic generated by the increasing large number of network connected devices.

Based upon the different demands for communication in the future it is unlikely that one network deployment model be able to serve all use cases and scenarios in 2020. As a result, flexibility to adapt to different scenarios with different requirements will be an inherent part of future 5G networks. One way to provide flexibility for future mobile network deployment is leveraging cloud technology, specifically telecommunication cloud or network embedded cloud as a distributed cloud architecture in which network providers add data centers within their network to offer cloud services [12]. The distribution of cloud deployment in operator networks can range from medium size data centers at a back office to Nano data centers coexisting with a Base Station (BS). Operators can exploit this cloud to offload the computing requirement of mobile networks (e.g., cloud-radio access network (C-RAN) or to virtualize mobile core network functions) [13].

III. INTERNET OF THINGS (IOT)

In a future networked society, every thing that benefits from a connection will be connected. As a result, IOT communication such as MTC, vehicle-to-vehicle (V2V), Machine-to-Machine (M2M), smart objects and sensors can be embedded in almost everything everywhere. IOT can be described as an application that communicates with other IOT devices or servers within the cloud (either cloud inside or outside an operator’s network) using a communication network and without human interaction. IOT devices can use any type of network for their communication, such as wireless local area networks, mobile networks including UMTS, LTE, and future 5G depending on the required QoS, cost, and network support. As mobile networks already have significant coverage in order to services to humans, they might also be used to carry IOT traffics [16]. However, current mobile networks were not designed for MTC, hence 5G aims to support the IOT by meeting the requirements for this communication.

Monitoring information from embedded sensors, remote device configuration, and triggering alarms based on data received from IOT devices are examples of IOT applications. IOT devices can have a fixed location (e.g., sensors measuring humidity in environment or other sensors inside a smart home) or be mobile (e.g., sensors embedded within a car, a human body, or other moving objects).

Data generated by IOT devices differs from data generates by humans. Additionally, not all MTC applications have the
same characteristics [17], IOT data can be small in size but frequently transmitted. IOT may deploy billions of devices (i.e., many more than the number of humans). The increasing numbers of devices utilizing a mobile network is a primary concern and may lead to problems such as congestion and overhead both on the data and control planes. Congestion can occur both in the radio access and core network. Although each IOT device may send and receive a small amount of data, the cumulative traffic from all of these devices can lead to congestion. Additionally, IOT devices also generate control and signaling traffic (even while they only transfer a small amount of data) due to their need to attach or re-attach to the network, and this control traffic can negatively affect core network control entities.

There is an ongoing discussion regarding the prerequisite for the network to meet MTC demands [17], [18]. Additionally, there efforts have been made to address congestion control for IOT and MTC, especially for LTE network. J. Wang, et al. [19] proposed TCP-FT as an enhancement for TCP to reduce congestion. Another suggestion was group based traffic management [20], where in each group only one IOT device communicated directly to the mobile network, hence reducing the control signaling overhead in mobile network. The time controlled policy introduced by 3GPP [21] is another suggestion, in which a device only can communicate with a network if it is not in a blocked period. The intention of most of these proposals is to reduce congestion caused by IOT devices in the radio access network. This paper proposes a solution to reduce the load and probability of congestion in the core network control entities for 5G.

### IV. Proposed Solution

This proposal assumes that the 5G architecture will not be limited by current flat LTE architecture. Fig. 7 describes the proposed architecture for future 5G. In this proposal, the core network consists of two categories of nodes: control nodes and gateway nodes. Control nodes are responsible for processing and handling control signaling, while gateway nodes are responsible for handling user plane traffic. The proposed 5G control node characteristics are

- **Control functions are independent and can run independently, in the same or separate physical locations.**
- **Each of control function can be executed in a different physical location while maintaining a parent-child relationship.**
- **Each control functions can act as a parent, a child, or both. The child acts the same as its parent and is able to update its parent if required.**
- **Control functions communicate with each other when needed.**

One suggestion to manage control signaling traffic in order to avoid congestion in the 5G core network is placing those control functions that make high demands upon the network due to users, as a child function close to UE’s current radio access point. Combining the telecommunication cloud and the proposed 5G control node characteristics are the primary enablers for this proposal. In this way, control functions can be launched as a child in data centers within the core network or data centers co-located with radio base stations. An SDN can be employed to provide a communication between parent and child control functions, i.e., enabling the child to migrate to an other location.

To achieve this goal, the design and implementation of the control node should be done such that it can satisfies the proposed characteristic of the 5G core network. The benefit of this approach is that by distributing control functions as children running close to UEs, the signaling load can be processed and handled locally. Additionally, this reduces the response time to control function requests. Moreover, the parent control functions are unaffected by a amount of signaling traffic that might be generated by large numbers of UE, such as IOT devices. Finally, this approach reduces the control plane’s bandwidth consumption, while adding additional capacity to the core network to handle user plane traffic.

![Proposed 5G core architecture](image)

Fig. 7. Proposed 5G core architecture.

1) A control node is virtualized and realize all required control functions, such as (re)attachment handling, mobility management, authorization, and authentication.

### V. Signaling Load Analysis

This section gives a mathematical model that can be used to calculate the control node(s) signaling load for a future
In CONNECT mode in a UE-terminated session, there is no signaling overhead to set up a connection for originating a session to send data. Also, if a UE is in CONNECT mode in a UE originated session, there is no signaling overhead to set up a connection for receiving data. As a result, for a UE in idle state number of messages processed by the control node (entering or leaving) due to UE-terminated session is given by

\[ L_t = M_t \times \lambda_t \times \rho \times A \times C \times (1 - P_{AO}) / K \]

where \( M_t \) denotes the number of control messages processed by control function(s) to set up a connection for UE-terminated session and \( P_{AO} \) denotes the probability that an IOT device is connected in a UE-originated session when data arrives for an IOT device. From Fig.3 \( M_t = M_o + C_o + 1 \) for a LTE network where \( C_o \) represents the number of eNodeBs per tracking area. Also, based on the independent assumption of \( \lambda_o \) and \( \lambda_t \), \( P_{AO} \) can be calculated as

\[ P_{AO} = \lambda_o \times \gamma_o \]

where \( \gamma_{out} \) is the average duration of originated session (in seconds). The load due to session management can be calculated as

\[ L_{sm} = L_o + L_t \]

To calculate control node load due to hand over classical Fluid-Flow Mobility [23] is employed to estimate a UE HO rate. Based on this model for a circular region with a population density of \( \rho \), an average velocity \( \vec{v} \), and region diameter of \( D \), the average number of site crossings per unit time is \( N_{avg} = \rho \pi D \vec{v} \). The centralized approach of HO management function (i.e., \( K = 1 \)) presents a scenario where a central HO function in the network manages all HO's within a network and no changing of HO management node occurs during HO. In this case, the total number of message per hour at the control function due to HO is given by:

\[ L^h_{K=1} = [M_{rec}^sg \times P_{csg} + M_{rec}^ngc \times (1 - P_{csg})] \times N_{avg} P_A \times C \]

where \( P_{csg} \) is S-GW relocation probability, \( M_{rec}^sg \) denotes the number of HO signaling messages processed by the control function when there is no change in control node and no change in S-GW, \( M_{rec}^ngc \) denotes number of HO signaling messages processed by the control function when there is no change in control node but S-GW relocation is required, and \( P_A \) denotes probability of device being in CONNECTED mode. \( P_A \) can be well approximated by \( P_{AT} + P_{AO} \).

When \( K = C \) we have full distribution of HO management entity. Also, we assume that co-located with each BS is an embedded data center to host a control function for HO management. We also assume that the rate of UEs leaving a BS equals to the rate of UEs joining a BS as a result of uniform user density. In this procedure there is possibility that

\*We assumes the paging message will not be lost on the back-haul or air interface and there is no need for paging re-transmission.
a UE stays with same S-GW or S-GW relocation also occurs during the HO procedure. As a result, the load on the HO control function is sum of load on HO management function as a source control function ($L_{isc}$) (i.e., due to UEs leave the cell) and the loads on HO management function as a target control function ($L_{tsc}$) (i.e., due to UEs joining a cell). These are given by

$$L_{h}^{b} = [(M_{isc} + M_{tsc}) × P_{isc} + (M_{isc} + M_{tsc}) × (1 - P_{isc})] × N_{avg} × P_{h}$$

(7)

where $M_{isc}$ denotes the number of HO signaling messages of the source HO function when S-GW relocation is required and $M_{tsc}$ denotes the number of HO signaling messages of the source HO control function when the UE maintains the same S-GW. Additionally, $M_{isc}$ denotes the number of HO signaling messages on the target HO control function when S-GW relocation is required and $M_{tsc}$ denotes the number of HO signaling messages on the target HO management function when the UE maintains the same S-GW.

When $1 < k < C$ the coverage region divides to K areas* where each area has its own HO management entity. In this case, the signaling load on the HO management entity in each area is the sum of the load due to HO inside that area (i.e., HO without changing HO management function) and the load due to HO between two areas (i.e., HO cases when the UE changes HO management function). The signaling load on each HO management function can be approximated as

$$L_{1<k<C}^{b} = L_{K=1}^{b} × (\frac{1}{K} - \sqrt{\frac{C}{K}}) + L_{K=C}^{b} × \sqrt{\frac{C}{K}}$$

(8)

Another event that produces signaling load is TAU. If $K'$ represents the distribution factor of tracking area, then $C_{a} = \frac{C}{K'}$. In this case, the total number of signaling messages generated on the network due to tracking area update is approximated as

$$L_{TAU}^{b} = [(M_{isc,TAU} × P_{isc} + M_{isc,TAU} × (1 - P_{isc})) × N_{avg} × \sqrt{C × K'}$$

(9)

where $M_{isc,TAU}$ denotes the number of signaling messages for intra-session management TAU, $M_{isc,TAU}$ denotes the number of messages for inter-session management TAU, and $P_{isc}$ represents the probability of changing session management function when changing tracking area region. $P_{isc}$ can be approximated as $\sqrt{\frac{K}{C}}$ when $K \leq K'$ and to 1 for other cases.

VI. NUMERICAL ANALYSIS

This section presents the numerical analysis of the control signaling load for both centralized and de-centralized control approach. The assumption in future 5G is that the number of signaling message processed at control functions for each controlling event is equal to messages proceed at an LTE MME due to same events (as discussed in section II-A). In this analysis we assumed a region size of 1000 $Km^{2}$, and $\lambda_{o} = 3\lambda_{t} = 1$ per minute. Additionally, based on 5G UE terminal target capacity (i.e., 10 Gbps) and the assumption of only a very small amount of data being transmitted by IOT device (i.e., assumed 500 Kb for up-link and 1500 Kb per down-link) session duration can be calculated as $\gamma_{t} = 3 × \gamma_{o} = 1.5e^{-6}$. Finally, we assumed the average speed of devices $\bar{v} = 20 \text{ km/h}$.

Equations 2 and equation 3 shows that user density and the number of BSs connected to one session management function are two factors that effect session management load due to UE originated and UE terminated scenarios. As a result, in future 5G deployments with denser cell deployment and high user density, the load on the session management function is expected to increase. One approach for reducing this load would be to decrease the area covered by a session management function (i.e., to reduce the number of BSs attached to a single session management function). This reduction can be a factor of co-locating session management functions at each BS, i.e., session management distribution factor $K = C$ (as we discussed in section IV). Fig. 9 shows how user density and the distribution factor of the session management function effect the session management load. In this figure we assumed that the size of a tracking area was $C_{a} = 10$.

Table I gives examples of the number of messages processed at each session management function due to UE originated and UE terminated scenario in units of millions of messages per hour. It is worth mentioning that increasing distribution factor of session management function will not effect the total number of signaling message in the network. The total number of signaling message generated in the network for different values of $\rho$ is shown in the first column of Table I when the $K = 1$.

<table>
<thead>
<tr>
<th>Distribution factor K</th>
<th>100</th>
<th>200</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 1000$</td>
<td>556</td>
<td>5.56</td>
<td>2.15</td>
</tr>
<tr>
<td>$\rho = 2000$</td>
<td>2224</td>
<td>22.24</td>
<td>11.12</td>
</tr>
<tr>
<td>$\rho = 8000$</td>
<td>4448</td>
<td>44.48</td>
<td>22.24</td>
</tr>
</tbody>
</table>

*‘Inter HO management function’ describes a scenario

![Fig. 9. Session management function load per hours based on session management distribution factor K and user density ρ. The "Z" axis uses logarithm scale.](image-url)
wherein after HO the UE changes HO management function, while “intra HO management function” is an HO within a single HO management function coverage area. As discussed previously, inter HO management function scenarios produce greater load than intra HO management function scenarios. On the other hand, increasing UE density or dense deployment of BSs will increase HO rate in the mobile network resulting in higher load. In a fully distributed HO management function approach, all HOs are inter HO management function, but each HO management function handles fewer HO. Fig. 10 shows how increasing user density and the distribution factor of the HO function affects the HO management function loads.

Table II gives an example of the number of messages processed at each HO management function and the total HO signaling load generated on the network when the user density $\rho = 8000$ UE/Km$^2$. From this table, we can observe that increasing the distribution factor of HO management function will reduce the load on each HO management function but, at the cost of increasing the total signaling load generated on the network. However, the total number of messages per hour is negligible. The high capacity of future 5G networks and low amount of data transmitted in each communication results in low value of $P_A$ (i.e., $8.340e-09$ based on our assumptions). The low value for $P_A$ indicates a low probability of a device being in a connected state when HO occurs, hence the signaling between UE and HO management function are unneeded.

<table>
<thead>
<tr>
<th>Distribution factor $K$ =</th>
<th>1</th>
<th>100</th>
<th>200</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load at each HO management function</td>
<td>1.318</td>
<td>0.031</td>
<td>0.019</td>
<td>0.015</td>
</tr>
<tr>
<td>Total load generated due to HO in network</td>
<td>1.318</td>
<td>3.993</td>
<td>3.910</td>
<td>4.538</td>
</tr>
</tbody>
</table>

The tracking area information will be utilized by session management entities when there is incoming data for an idle UE. Tracking area size affects the load generated because of the need to send TAU (i.e., when a UE finds itself in a new tracking area it must send a TAU) and also the signaling load due to the UE terminated scenario. In this part of the analysis, we assume that the agent for handling TAU is co-located with the session management functions. From equation 9, it is clear that increasing user density results in linear growth of load due to TAU. There are two more factors that can effect on TAU: distribution factor of session management functions ($K$) and the distribution factor of the tracking area ($K'$). These two factors determine the number of BSs in one tracking area and the number of tracking areas supported by one session management. Fig. 11 shows how these two factors affect the number of messages processed because of TAU.

Table III gives an example of the number of messages processed at each session management function due to TAU.
in units of millions messages per hour. It is worth mentioning that increasing $K'$ result in more tracking area but with smaller size thus increasing TAU load as the user more frequent needs to send a TAU. $K < K'$ means that one TA will be covered by numbers of session management. As a result, the load of TAU will split between those session management functions that result in slow growth in TA management function when $K'$ increase. $K > K'$ means that each session management function cover several TAs. As a result, the load on each session management function will be the sum of the load generated by each TA, thus leading to a rapid growth in session management load when $K'$ increase. Fig. 12 depicts this behavior when the session management distribution factor $K = 150$.

**TABLE III. NUMBER OF MESSAGES PROCEED AT SESSION MANAGEMENT FUNCTION DUE TO TAU BASED ON SAMPLE VALUE OF SESSION MANAGEMENT DISTRIBUTION FACTOR $K$ AND TA DISTRIBUTION FACTOR $K'$ IN UNITS OF MILLIONS MESSAGES PER HOUR.**

<table>
<thead>
<tr>
<th>$K'$ =</th>
<th>1</th>
<th>100</th>
<th>200</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1506.622</td>
<td>5.381</td>
<td>2.690</td>
<td>1.794</td>
</tr>
<tr>
<td>200</td>
<td>2175.261</td>
<td>12.067</td>
<td>3.805</td>
<td>2.557</td>
</tr>
<tr>
<td>300</td>
<td>2688.325</td>
<td>17.198</td>
<td>6.370</td>
<td>5.107</td>
</tr>
</tbody>
</table>

On the other hand, increasing the distribution factor of TAU (i.e., results in smaller $C_u$) decreases the load on each session management functions due to UE terminated sessions. As a result from a session management function point of view the maximum value possible for the TA distribution factor results in lower load due to UE terminated scenarios. Based on the assumption of Co-locating the session management function and TA management function to find appropriate value for TA distribution factor the aggregation of UE termination and TAU signaling load should be considered. Table IV shows the number of messages processed at the session management function due to UE originated and UE terminated scenario based on sample values of $K$ and $K'$.

**TABLE IV. AGGREGATION OF MESSAGES PROCEED AT SESSION MANAGEMENT FUNCTION DUE TO TAU AND UE OriginATED AND UE TERMINATED BASED ON SAMPLE VALUE OF SESSION MANAGEMENT DISTRIBUTION FACTOR $K$ AND TA DISTRIBUTION FACTOR $K'$ IN UNITS OF MILLION MESSAGES PER HOUR.**

<table>
<thead>
<tr>
<th>$K'$ =</th>
<th>1</th>
<th>100</th>
<th>200</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5.61474</td>
<td>0.615</td>
<td>0.308</td>
<td>0.205</td>
</tr>
<tr>
<td>100</td>
<td>1507.985</td>
<td>9.394</td>
<td>2.698</td>
<td>1.799</td>
</tr>
<tr>
<td>200</td>
<td>2178.392</td>
<td>12.051</td>
<td>3.811</td>
<td>2.541</td>
</tr>
<tr>
<td>300</td>
<td>2689.647</td>
<td>17.211</td>
<td>6.377</td>
<td>5.111</td>
</tr>
</tbody>
</table>

Fig. 13 shows how changing $K$ and $K'$ effects on the total number of message processed on the network due to handling TAU. It worth mentioning that increasing the distribution factor of TAU increase the total load generated due to TAU on network. The pattern of increase is the same as shown Fig. 12. On the other hand, increasing the session management distribution factor decreases the total TAU load on the network. When $K = K'$, the total load generated on the network due to TAU will remain constant. Table V gives examples of the total number of messages generated on the network due to TAU.

**TABLE V. TOTAL NUMBER OF SIGNALING MESSAGE GENERATED WITHIN NETWORK BASED ON SAMPLE VALUE OF SESSION MANAGEMENT DISTRIBUTION FACTOR $K$ AND TA DISTRIBUTION FACTOR $K'$ IN UNITS OF MEGA MESSAGE PER HOUR.**

<table>
<thead>
<tr>
<th>$K$</th>
<th>1</th>
<th>100</th>
<th>200</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K'$ =</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>53.801</td>
<td>53.808</td>
<td>53.808</td>
<td>53.808</td>
</tr>
<tr>
<td>200</td>
<td>53.808</td>
<td>53.801</td>
<td>53.808</td>
<td>53.808</td>
</tr>
<tr>
<td>300</td>
<td>53.808</td>
<td>53.808</td>
<td>53.808</td>
<td>53.808</td>
</tr>
</tbody>
</table>

Increasing the “session management function” distribution factor results in a division of session management signaling messages to more session management functions without adjusting the total number of message generated on the network. A more highly loaded network indicates a need for more compute resources and capacity both of which will increase operational costs.

The high capacity of future 5G networks and limited data transmitted by IOT devices affects most of the HO scenarios as the UE is likely to stay in idle mode. In this case, a low amount of HO signaling is expected (i.e., even in a network with very high UE density) as for the idle UEs no signaling messages are required between the UE and the HO management function. As a result, the distribution of HO management function does not add considerable value for the network operators. Future IOT device that are not concerned with battery consumption can choose “always-on”* approach that raise signaling load because of HO. On the other hand, for devices with “always-on” approach session management function signaling load will decrease as there is no paging and session management required. Further study is needed to analyze the side effect of the always-on approach, but this is out of the scope of this paper.

The TA distribution factor determines the size and number of TAs covered by each session management. A larger value for the TA distribution factor will result in lower load for UE terminated scenarios, while increasing the aggregated load due to the TAU on the network.

**VII. CONCLUSION**

This paper discusses a novel control plane architecture for future 5G to provide a flexible and scalable network. The proposed network architecture consists of independent control function responsible for different control events in the network. The control functions can leverage cloud technologies (e.g., network embedded cloud) and run in a virtualized environment to implement a flexible and scalable control plane function distributed within 5G mobile networks. This approach aims to enable 5G networks to manage the enormous amount of control signaling load on each function and total load generated within the network for each control events. Reducing the load at each control function by distributing those functions enables network operators to better scale their network in order to manage the huge number of signaling messages generated by different types of users in a future 5G network. The second factor that we should consider is the total number of message generated on the network. A more highly loaded network indicates a need for more compute resources and capacity both of which will increase operational costs.

*This approach keeps continuously UEs in a connected state.
signaling produce by IOT and MTC. We utilized 3GPP LTE to identify the required control plane events, and we assumed that 5G takes evolutionary path from LTE. We have shown that moving the control plane session management function close to UEs at a data center co-located with the BS is beneficial in terms of control signaling load management.

REFERENCES


