International Journal for Innovative
Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

Title: Efficiient Network Modelling Analysis for 5G Systems.

Volume 06, Issue 05, Page No: 2210 - 2224.

Paper Authors

*** RAMA KRISHNA DOKKA, K. SRIKANTH.**

* Sree Dattha Group of Institutions – Integrated Campus.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

EFFICIIENT NETWORK MODELLING ANALYSIS FOR 5G SYSTEMS

***RAMA KRISHNA DOKKA, **K. SRIKANTH**
*PG Scholar, DEPT OF wireless mobile communications, Sree Dattha Group of Institutions – Integrated Campus. Sheriguda Village, Nagarjuna Sagar Road, Ibrahimpatnam, Hyderabad - 505450
**Assistant professor, Dept of VLSI & Embedded systems, Sree Dattha Group of Institutions . Sheriguda Village, Nagarjuna Sagar Road, Ibrahimpatnam, Hyderabad - 505450

ABSTRACT:

One of the main challenges of the upcoming 5G net-works is to accommodate the high demand of data raised from the increasing number of devices. In this vein, deploying small cells should be considered with high interest to overcome this issue. 5G networks would deploy densely self-organizing low-cost and low power small base-stations. However, deploying high number of small cells would increase the signalling overhead caused by the tracking and paging of User Equipment (UE). Combined with the high number of UEs and Machine Type Communication (MTC) devices [1], [2], the use of small cells will introduce a major challenge in term of signalling overhead for 5G networks. In order to tackle the increased data rate expected from the usage of the envisioned 5G network, the signalling overhead should be minimized as much as possible.

INTRODUCTION:

Usually, the Radio Access Network (RAN) of a mobile operator is organized into a set of cells (including small cells) that covers several geographical areas. UEs in a specific area are attached to a base station (eNodeB), which manages their access to the mobile core network. UEs are usually in idle mode and have no call activity for some duration. When a connection request comes for a UE in idle mode, the Mobility Management Entity (MME) sends a signalling message, namely paging, to all eNodeBs to find the UE's location (i.e., cell) in the network. Accordingly, in case a high number of UEs need to be paged, a massive number of downlink signal-ing messages have to be transmitted, resulting in high signalling overhead and wasting scarce resources of the mobile network. To overcome this issue, the Tracking Area (TA) concept has been introduced in Release 8 of the 3GPP mobile network specifications (i.e., replacing the Routing Area concept in previous releases). The key idea beneath the TA principle consists in grouping several cells or sites into one TA. MME keeps record of the location of UEs in idle mode at the TA granularity. Thus, when a connection setup request comes for a UE in idle mode, the UE in question is paged only within its current TA, which would mitigate the overhead of paging in the network. Each time a UE moves to a new location and connects to a new cell not belonging to its current TA, the UE sends an uplink message, namely Tracking Area Update (TAU), to MME, which subsequently updates the TA of the UE. In this vein, it is worth noting that a TA is also defined as an area where the UE can move without transmitting TAU messages to MME. Despite the advantages of the TA concept in minimizing the paging overhead, it has the following limitations on the TAU signalling: (i) many TAU signalling messages might be generated due to ping-pong effect, i.e, a UE keeps hopping between two adjacent cells belonging to different TAs, which could be exacerbated in case of densely deployed small cells; (ii) the mobility signalling congestion due to a large number of UEs having a similar behaviour, e.g. massive number of UEs simultaneously moving from one TA to another TA (train scenario); (iii) the use of TA strategy has the symmetry limitation: If two cells are in the same TA, then neither of them can be in any other TA. To overcome this limitation, introduces the Tracking Area List (TAL) concept in order to simplify the TA configuration. The TAL concept aims for reducing the TAU signalling messages by grouping several TAs in one TAL and allowing the overlapping of TAs. Each time a UE visits a new TA that does not belong to its TAL, a TAU message is sent to the MME. Upon receiving the TAU message, MME assigns a new TAL to the UE. The new TAL should include the visited TA. Furthermore, Release 12 allows network operators to include up to 15 TAs in each TAL and the MME always adds the last visited TA to the list to overcome the problem of frequent updates due to ping-pong situations. Given that TALs are overlapped, the above-mentioned limitations of conventional TAs, defined in Release 8, can be accordingly mitigated. However, the current LTE specifications do not provide any details on how to define TALs and allocate them to UEs. Each time a UE moves to a new location and connects to a new TA not belonging to its current TAL, the UE sends a TAU message to MME. On the other hand, when a connection request comes for a UE, the MME sends a paging message to all TAs (i.e., TAL) where the UE is registered. An increase in TALs size leads to a rise in paging signalling messages and a decrease in TAU signalling messages. Fig. 1 shows the trade off between TAU and paging overheads when forming TALs. In the figure, we assume that the network contains four TAs along a railway path, in which each TA has two other neighbouring TAs on the left and the right sides. From Fig. 1(a), we observe that the organization of each TA in a separate TAL causes many TAU signalling messages in the network, which are generated and forwarded from the RAN to the evolved packet core (EPC). Whereas Fig. 1(b) and Fig. 1(c) show that increasing TAL size reduces TAU overhead and increases paging overhead. Fig. 1(c) shows that the TAU overhead can be ignored if all TAs are organized in the same TAL. Several research works have been conducted to solve the TAL problem, whereby the aim is to capture the trade off that mitigates the overhead of TAU and paging messages when constructing and assigning TALs to UEs. Most of these solutions formulate the problem using a multi-objectives optimization technique to achieve a fair trade off between signalling messages overhead of TAL and paging, i.e. minimize both signalling messages due to TAU and paging. In this paper, we devise an efficient tracking area list management (ETAM) framework for 5G cloud-based mobile networks [3], [4]. The proposed framework consists of two independent parts. The first part is executed offline and is responsible of assigning TAs to TALs, whereas the second one is executed online and is responsible of the distribution of TALs on UEs during their movements across TAs. For the first part, we propose three solutions, which are: (a) F-PAGING favouring the paging overhead over TAU, (b) F-TAU favouring TAU over paging, and (c) FOTA (i.e., Fair and Optimal Assignment of TALs to TAs) for a solution that uses bargaining game to ensure a fair trade off between TAU and paging overhead. For the second part, two solutions are proposed to assign TALs to UEs. The com-potation load is kept lightweight in both solutions not to downgrade the network performance. Furthermore, both solutions do not require any additional new messages when assigning TALs to UEs. The first solution takes into account only the priority between TALs. As for the second one, in addition to the priority between TALs, it takes into account the UEs activities (i.e., in terms of in-coming communication frequency and mobility patterns) to enhance further the network performance. The remainder of this paper is organized as follows. Section 2 introduces some related research work. Section 3 presents the envisioned network model and formulates the target problem. It also presents an overview of the ETAM framework. Section 4 presents the online part of the ETAM framework for assigning TALs to UEs. The three solutions proposed for the offline part of the ETAM framework are described in Section 5. Section 6 details a Markov-based analytical model for the three offline solutions. Besides the numerical results obtained by solving the Markov model, Section 7 presents the simulation setup to evaluate the performance of ETAM and discusses the obtained results. Finally, the paper is concluded in Section 8.

2 LITERATURE SURVEY:

Mitigating signalling overhead, due to UE mobility in cellular mobile networks, has attracted high attention during the last years. As stated earlier, in the Evolved Packet System (EPS), MMEs keep records of UEs' positions in order to adequately forward their relevant in-coming connections. For this purpose, 3GPP introduced two types of signalling messages to support UE mobility: (i) paging messages from the network, namely MME, in order to find the locations of UEs in idle mode; (ii) TAU messages from UEs to MME to update their positions. A TAU message is sent each time a UE enters into a new location (cell) that does not belong to its current TA. Conventional TA assignment procedures whereby the network assigns only one TA for different UEs is not sufficient when UEs are highly mobile. Indeed, high number of TAU messages could be sent by UEs as they frequently cross their corresponding TA borders. An enhancement to the conventional procedure was envisioned to reduce TAU overhead by i) grouping several cells (i.e., eNodeBs) in one TA or ii) introducing delays between TAU messages sent by UEs. Another solution to reduce the impact of TAU messages on the network was proposed in [5] whereby queuing models and buffer information at eNodeBs are used to delay the TAU frequency.

To further alleviate the effect of TAU messages on the network performance, 3GPP has introduced the concept of TAL in Long Term Evolution (LTE), wherein each cell (eNodeB) assigns different TALs to UEs [6], [7]. Since TALs are overlapped, the number of UEs performing TAU when crossing TA border drastically decreases. Besides reducing the number of TAU messages, TAL pre-vents the ping-pong effect, i.e., frequent TAU messages when a UE keeps hopping between adjacent TAs. Nev-ertheless, the current LTE specifications do not provide any details on how to define TALs and allocate them to UEs. To address this open issue, several solutions have been proposed. In [8], Chung et. al. proposed a solution that organizes cells into rings, where UEs in each ring use the same TAL. Solutions, proposed in [9] and [10], use the same concept as in [8] by assigning the same TAL to different UEs when visiting a cell in the network. However, all these solutions [8]–[10] have not fully explored the advantage of TAL against the conventional TA approach. In [7] and [11], Razavi et. al. overcome this limitation by allowing UEs residing in the same cell to register with different TALs. Indeed, in [7] they proposed a solution for congestion mitigation along a railway path. On the other hand, in [11] an extension of the former work is proposed with two new aspects: i) the solution is generalized for any arbitrary network instead of only train scenario; ii) a new solution that handles the extenuation of paging signalling messages via TAL management is proposed.

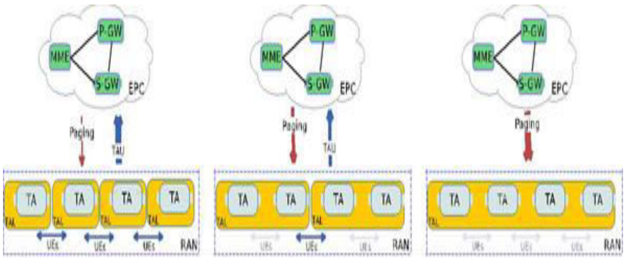


Fig. 1: The trade off between TAU and paging overhead in 4G and beyond mobile networks.

Generally speaking, assigning TALs to UEs shall de-pend on the mobility patterns of UEs as well as on their geographical distribution and density. MME may group, under the same TAL, a large number of TAs in an area that has low density to reduce the impact of TAU overhead on the network performance. Similarly, MME may group under the same TAL a small number of TAs serving a highly densed area. Indeed, to alle-viate the impact of paging messages on the network performance, it is worth assigning more than one TAL to the same TA.

To the best knowledge of the authors, most existing solutions focus only on the offline part for assigning the TAs to TALs. Moreover, they consider only the TAU overhead and ignore the paging overhead. The only research work that addressed both constraints is presented in [11], wherein Razavi et al. proposed two separate solutions, addressing the impact of TAU and paging overhead, respectively. Both solutions are based on multi-objectives optimization techniques for assigning the TAs to TALs. The first one tries to minimize the TAU overhead while setting paging as a constraint, and the second one minimizes the paging overhead while fixing the TAU overhead as a constraint. In contrast to the existing works, in this paper, we propose a framework optimizing the management of TALs and consisting in: (i) an offline part that assigns TAs to TALs; (ii) an online part that assigns TALs to UEs. Two solutions are proposed to achieve the aim of the online part. The first one takes into account only the priority between TALs, whereas the second one, in addition to the priority between TALs, takes into account the UE behaviour in terms of mobility and connection frequency. Regarding the offline part, we have devised three solutions, which differ from the existing ones on their way to cope with the problem. Indeed, most existing solutions assign the same TAL:

i) to the same TAs in a static manner [8]–[10]; or ii) with the same probability [7], [11]. In contrast, the devised solutions dynamically assign the same TAL to different TAs with different probabilities. The first one, dubbed F-PAGING, is proposed for a network known with a high rate of paging (i.e., for voice call as well as for IP-based web applications) in comparing to the mobility rate. This solution may be designated for small cities with high-density populations. The second one, dubbed F-TAU, is proposed for a network which is known with a high mobility rate compared to the paging rate. Such kind of solution may be useful for a network known with low-density populations and/or high mobility. The last one, dubbed FOTA, is proposed to be generic for any kind of networks. It takes advantage of both previous solutions, jointly addressing the overhead due to both TAU and paging messages. FOTA uses Nash bargaining game to ensure a fair trade off between both conflicting overhead, i.e., TAU and paging signalling messages.

3 ENVISIONED NETWORK MODEL AND FRAMEWORK OVERVIEW

3.1 ETAM framework overview

Fig. 2 depicts a general overview of the ETAM frame-work. We assume that the network is subdivided into N TAs, named N = f1; 2; Ng. Each TA consists of a set of cells, whereby a cell is managed by an eNodeB (i.e., base station). As depicted in the figure, the geo-graphically close eNodeBs can be grouped in the same TA, using any existing algorithm [12], [13], to optimize the network performance in terms of paging overhead.

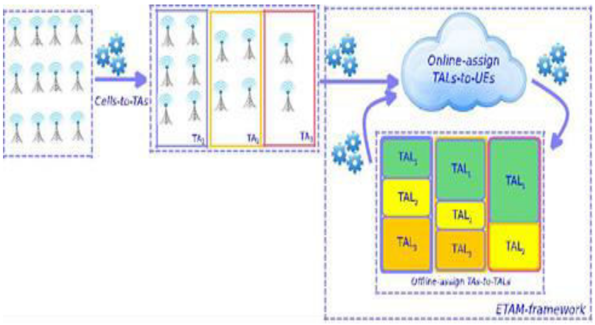


Fig. 2: The proposed framework for tackling TAU and paging overhead in 4G and beyond mobile networks.

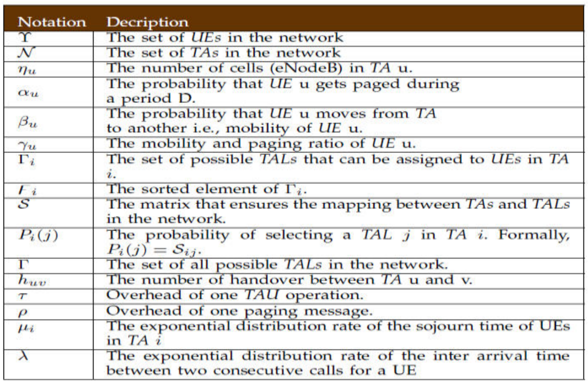
Initially, the ETAM framework starts by an inefficient solution and then converges, through iterations, to the optimal one. As depicted in Fig. 2, ETAM framework starts by considering each TA as a separated TAL. Then it executes, repetitively, two steps to converge to the optimal solution. The first step is the offline-assignment of TAs-to-TALs, whereas the second one is the online-assignment of TALs-to-UEs. To efficiently map between TAs and TALs, the information about TAU and paging signalling messages are transferred from the online step to the offline one. The latter enhances the mapping between TALs and TAs and then provides the former with the new mapping to optimize further the network performance. The online step is executed during a specified period D, where all the information about the TAU and paging overhead are gathered from the network to be transferred to the offline step. The duration D may be fixed by the network operator, but it can be changed when there is a noticeable update in the network.

Since there is no exact indication on the trajectory of UEs, during the online-assignment of TALs-to-UEs, we use a probability strategy to assign TALs to UEs. In each visited TA, TALs are assigned to visiting UEs with different probabilities. Indeed, the TAL that reduces more the TAU and paging signalling messages would have more priority to be assigned to a UE. There is a trade off between TAU and paging signalling messages. Clearly, the smaller the size of TALs is, the higher the TAU overhead is, but the smaller the paging overhead becomes. For the online-assignment of TALs-to-UEs, we consider two solutions. The first one takes into account only the priority between TALs that was learned from the offline step. Whereas, the second one, in addition to the priority between TALs, takes into account the UEs behaviour, in terms of incoming communication frequency and mobility patterns. For the offline-assignment of TAs-to-TALs, we consider three different solutions, which define the core of our ETAM framework. It is worth recalling that (i) the first solution favours the paging overhead when forming TALs; (ii) the second one favours the TAU overhead; and (iii) the third solution uses the bargaining game theory to distribute TALs among TAs by capturing a fair trade off between TAU and paging overhead. The TAL that exhibits the highest fairness in the TAU and paging overhead has the highest probability to be as-signed to a UE.

3.2 Network model and notations

Let denote the set of all possible TALs in a mobile network, and let A denote the set of possible TALs that can be assigned to UEs in TA A. As mentioned earlier, each time a UE visits a new TA that does not belong to its TAL, a TAU message is sent to the MME. Upon receiving the TAU message, MME computes and sends a new TAL to the UE. The new TAL should include the visited TA. From Release 12 of the 3GPP specifications, the operator can specify for each TAL a list of up to 15 TAs and the MME always adds the last visited TA to the list to prevent the risk of ping-pong updates. For this reason, is formed by considering the different possible combinations of TAs, such that the length of each element in should be higher or equal to one and less than 16, i.e. each TAL i 2 should contain at least 1 TA and at most 15 TAs to allow the MME to add the last visited TA. Throughout the paper, we will refer to the example depicted in Fig. 3 in order to show how should be constructed. In this example, we assume that the network consists of five TAs, named A, B, C, D and E. The blue arrows between TAs denote the movement of different UEs in the network. The movement of UEs can be deduced from the handover statistics of different eNodeBs or from the handover command messages sent by MME. To form , we begin by forming the neighboring graphs G from the network as depicted in Fig. 3(b). An edge between two vertices (i.e., TA) A and B exists, if there is a TAU possibility between them. In Fig. 3(b), an edge is generated between the vertices A and B, if there is a blue arrow between TAs A and B in Fig. 3(a), which means the possibility of UEs movement between these TAs. In Fig. 3(b), we do not construct an edge between vertices A and E since a direct blue arrow does not exist between them; UEs cannot move from A to E without passing by another TA (i.e., B or D). Finally, A is formed from the neighboring graph G. Indeed, the different elements of A are those having all vertices of all sub-graphs of G that contain the vertex A and their length do not exceed 15. Thus, the vertices of a sub-graph of G that contain the vertex A are considered as one element in A.

We assume that each UE has a specific probability to be called/paged (i.e., for voice call as well as for IP-based web applications). Further, each UE follows a different mobility pattern, hence the number of sites (cells) visited by each UE is different. In the online-assignment of TALs-to-UEs step, the network is monitored in order to track the number of signalling messages (i.e., TAU and paging) sent and received by different UEs. We denote by $\alpha = \{\alpha_1, \alpha_2 \cdots\}$ and $\beta = \{\beta_1, \beta_2 \cdots\}$the probability of paging and TAU of UEs in the network, respectively. In other words, in the offline-assignment step, we have the information about different existing UEs in the network. We denote by the different UEs. For each UE belongs to ϒ2 , we have its probability u to send a TAU message and its probability u to be called (i.e., cause a paging). We denote by = $\gamma = \{\gamma_1, \gamma_2, \cdots\}$ 1; 2; g the overhead of mobility and paging ratio of different UEs. u denotes the overhead of mobility and paging ratio of UEu, i.e. the ratio between the paging and the TAU of a UEu. Formally, u is computed as follows



Notation	Decription
Υ	The set of *UEs* in the network
$\mathcal{N}$	The set of *TAs* in the network
η_u	The number of cells (eNodeB) in *TA* u.
α_u	The probability that *UE* u gets paged during a period D.
β_u	The probability that *UE* u moves from *TA* to another i.e., mobility of *UE* u.
γ_u	The mobility and paging ratio of *UE* u.
Γ_i	The set of possible *TALs* that can be assigned to *UEs* in *TA* i.
$\digamma_i$	The sorted element of Γ_i.
$\mathcal{S}$	The matrix that ensures the mapping between *TAs* and *TALs* in the network.
$P_i(j)$	The probability of selecting a *TAL* j in *TA* i. Formally, $P_i(j) = \mathcal{S}_{ij}$.
Γ	The set of all possible *TALs* in the network.
h_{uv}	The number of handover between *TA* u and v.
τ	Overhead of one *TAU* operation.
ρ	Overhead of one paging message.
μ_i	The exponential distribution rate of the sojourn time of UEs in *TA* i
λ	The exponential distribution rate of the inter arrival time between two consecutive calls for a UE

TABLE 1: Notations used in the paper

5G TECHNOLOGY REQUIREMENTS

As a result of this blending of requirements, many of the industry initiatives that have

progressed with work on 5G (see Appendix A) identify a set of eight requirements:

- 1-10Gbps connections to end points in the field (i.e. not theoretical maximum)
- 1 millisecond end-to-end round trip delay (latency)
- 1000x bandwidth per unit area
- 10-100x number of connected devices
- (Perception of) 99.999% availability
- (Perception of) 100% coverage
- 90% reduction in network energy usage
- Up to ten year battery life for low power, machine-type devices

Because these requirements are specified from different perspectives, they do not make an entirely coherent list – it is difficult to conceive of a new technology that could meet all of these conditions simultaneously. Equally, whilst these eight requirements are often presented as a single list, no use case, service or application has been identifed that requires all eight performance attributes across an entire network simultaneously. Indeed some of the requirements are not linked to use cases or services, but are instead aspirational statements of how networks should be built, independent of service or technology – no use case needs a network to be significantly cheaper, but every operator would like to pay less to build and run their network. It is more likely that various combinations of a subset of the overall list of requirements will be supported 'when and where it matters'.

GSMA Intelligence Understanding 5G, Finally, while important in their own right, six of these requirements are not generation defining attributes. These are considered below:

Perceived 99.999% availability and 100% geographical coverage:

These are not use case drivers, nor technical issues, but economic and business case decisions. 99.999% availability and 100% coverage are achievable using any existing technology, and could be achieved by any network operator. Operators decide where to place cells based on the cost to prepare the site to establish a cell to cover a specific area balanced against the benefit of the cell providing coverage for a specific geographic area.

This in turn makes certain cell sites and coverage areas - such as rural areas and indoor coverage - the subject of difficult business decisions. Whilst a new generation of mobile network technology may shift the values that go in to the business model that determines cell viability, achieving 100% coverage and 99.999% availability will remain a business decision rather than a technical objective. Conversely, if 100% coverage and 99.999% availability were to be a 5G 'qualifying criteria', no network would achieve 5G status until such time as 100% coverage and 99.999% availability were achieved.

Connection density (1000x bandwidth per unit area, 10-100x number of connections):

These essentially amount to 'cumulative' requirements i.e. requirements to be met by networks that include 5G as an incremental technology, but also require continued support of pre-existing generations of network technology. The support of 10-100 times the number of connections is dependent upon a range of technologies working together, including 2G, 3G, 4G, Wi-fi, Bluetooth and other complementary technologies. The addition of 5G on top of this ecosystem should not be seen as an end solution, but just one additional piece of a wider evolution to enable connectivity of machines. The Internet of Things (IoT) has already begun to gain significant momentum, independent of the arrival of 5G. Similarly, the requirement for 1,000 times bandwidth per unit area is not dependent upon 5G, but is the cumulative effect of more devices connecting with higher bandwidths for longer durations. Whilst a 5G network may well add a new impetus to progression in this area, the rollout of LTE is already having a transformational effect on the amount of bandwidth being consumed within any specific area, and this will increase over the period until the advent of 5G. The expansion of Wi-fi and integration of Wi-fi networks withcellular will also be key in supporting greater data density rates. Meeting both of these requirements will have significant implications for OPEX on backhaul and power, since each cell or hotspot must be powered and all of the additional traffic being generated must be backhauled.

Reduction in network energy usage and improving battery life:

The reduction of power consumption by networks and devices is fundamentally important to the economic and ecological sustainability of the industry. A general industry principle for minimising power usage in network and terminal equipment should pervade all generations of technology, and is recognised as an ecological goal as well as having a GSMA Intelligence Understanding 5G significant positive impact on the OPEX associated with running a network. At present it is not clear how a new generation of technology with higher bandwidths being deployed as an overlay (rather than a replacement) on top of all pre-existing network equipment could result in a net reduction in power consumption. Some use cases for M2M require the connected device in the field to lie dormant for extended periods of time. It is important that innovation in how these devices are powered and the leanness of the signalling they use when becoming active and connected is pursued. However, this requirement is juxtaposed with 5G headline requirements on data rate – what is required for mass sensor networks is very occasional connectivity with minimal throughput and signalling load. Work to develop such technology predates the current 5G requirements and is already being pursued in Standards bodies. These six requirements should be and are being pursued by the industry today using a range of techniques (some of which are covered later in the paper) but these amount to evolutions of existing network technology and topology or opportunities enabled by changing hardware characteristics and capabilities. These will in turn open business opportunities for operators and third parties. However, none of these business opportunities exist today – they are constrained by limitations greatly governed by economics, and much of these six requirements are motivated by improving the economic viability of those opportunities, rather than filling technological gaps that explicitly prohibit these opportunities, regardless of the amount they might cost to enable. Thus in the strictest terms of measurable network deliverables which could enable revolutionary new use case scenarios, the potential attributes that would be unique to 5G are limited to sub-1ms latency and >1 Gbps downlink speed.

PRPOSED DESIGN:

Along with the publication of 5G service and performance requirements by International Telecommunication Union-Radio communication (ITU-R) [1], many research groups today are conducting research and standardization activities of 5G mobile communication system actively [2]. 3GPP has specified new use cases that cannot be met with 4G Evolved Packet System (EPS) [3]. The proposed use cases can be classified into five categories according to their examples as follows.

•Enhanced mobile broadband: Ultra High Definition (UHD), virtual presence.

•Critical communication: Robot/drone, emergency.

•Massive machine type communication: eHealth.

•Network operation: Network slicing, interworking.

•Enhancement of vehicle-to-everything: Autonomous driving.

It is necessary to reduce latency and connect many devices to the network while increasing data rate in order to support the above services. 5G mobile communication system will enable not only existing services, but also new services in various fields. Through realization of the new services, Information and Communication Technology (ICT) convergence will come true in real life.

3GPP has been working on the standardization of 5G mobile communication system for the commercialization of 5G in 2020. Radio Access Network (RAN) and Service and System Aspects (SA) are the representative Technical Specification Groups (TSGs) within 3GPP. The TSG RAN is developing documents covering radio access architecture and radio interface protocol aspects of new Radio Access Technology (RAT) [4,5]. SA Working Group 2 (SA2) within the TSG SA studies architecture and main functions of the 5G network system under the study item of Next Generation system (NextGen). The SA2 finalized the NextGen Phase 1 study in December 2016 and published the 3GPP TR 23.799 specification as an outcome of the study [6]. NextGen Phase 2 study is also expected to begin mid 2017. Based on the agreements within the NextGen study, SA2 group is currently conducting normative standardization work for 5G mobile core network architecture by aiming to finalize the initial architecture by the end of 2017. In this paper, we present the architecture and functions of 5G mobile communication system by referring to the 3GPP SA2 NextGen Phase 1 study.

2. Architecture for next generation core network

In 4G core network, called Evolved Packet Core (EPC), protocol and reference points are defined for each entity such as Mobility Management Entity (MME), Serving Gateway (S-GW), and Packet Data Network Gateway (P-GW). On the other hand, in NextGen, protocol and reference points are defined for each Network Function (NF). We now present the NextGen architecture and the agreements of overall architecture.

2.1. NextGen architecture overview

Overview: Fig. 1 shows the NextGen architecture composed of NFs and reference points connecting NFs. User Equipment (UE) is connected to either RAN or Access Network (AN) as well as Access and Mobility Function (AMF). RAN represents a base station using new RAT and evolved LTE while AN is a general base station including non-3GPP access, e.g., Wi-Fi. The NextGen core network consists of various NFs. In Fig. 1, there are seven NextGen core NFs, namely, (1) AMF, (2) Session Management Function (SMF), (3) Policy Control Function (PCF), (4) Application Function (AF), (5) Authentication Server Function (AUSF), (6) User Plane Function (UPF), and (7) User Data Management (UDM).

Architecture and reference points for NextGen

Download high-res image (120KB)Download full-size image

Fig. 1. Architecture and reference points for NextGen.

Network functions: NF, 3GPP-adopted processing function in NextGen, has both functional behavior and interface. An NF can be implemented either as a network element on a dedicated hardware, as a software instance running on a dedicated hardware, or as a virtualized function instantiated on an appropriate platform, e.g., a cloud infrastructure [6].

Functional description: AMF provides UE-based authentication, authorization, mobility management, etc. A UE even using multiple access technologies is basically connected to a single AMF because the AMF is independent of the access technologies. SMF is responsible for session management and allocates IP addresses to UEs. It also selects and controls the UPF for data transfer. If a UE has multiple sessions, different SMFs may be allocated to each session to manage them individually and possibly provide different functionalities per session. AF provides information on the packet flow to PCF responsible for policy control in order to support Quality of Service (QoS). Based on the information, PCF determines policies about mobility and session management to make AMF and SMF operate properly. AUSF stores data for authentication of UE while UDM stores subscription data of UE. Data network, not part of NextGen core network, provides Internet access or operator services.

Reference points: Reference point representation of the architecture can be used to develop detailed call flows in the normative standardization. Next Generation (NG)1 is defined to carry signalling between UE and AMF. The reference points for connecting between AN and AMF and between AN and UPF are defined as NG2 and NG3, respectively. There is no reference point between AN and SMF, but there is a reference point, NG11, between AMF and SMF. Therefore, we can confirm that SMF is controlled by AMF. NG4 is used by SMF and UPF so that the UPF can be set using the control signal generated by the SMF, and the UPF can report its state to the SMF. NG9 is the reference point for the connection between different UPFs, and NG14 is the reference point connecting between different AMFs, respectively. NG15 and NG7 are defined since PCF applies policy to AMF and SMP, respectively. NG12 is required for the AMF to perform authentication of the UE. NG8 and NG10 are defined because the subscription data of UE is required for AMF and SMF.

2.2. Agreements on overall architecture

Separation of control and user planes: NextGen aims at separating user plane and control plane. The user plane carries user traffic while the control plane carries signalling in the network. In Fig. 1, the UPF is in the user plane and all other NFs, i.e., AMF, SMF, PCF, AF, AUSF, and UDM, are in the control plane. Separating the user and control planes guarantees each plane resource to be scaled independently. It also allows UPFs to be deployed separately from control plane functions in a distributed fashion. In this architecture, UPFs may be deployed very close to UEs to shorten the Round Trip Time (RTT) between UEs and data network for some applications requiring low latency.

Modularization: NG architecture is composed of modularized functions. For example, the AMF and SMF are independent functions in the control plane. Separated AMF and SMF allow independent evolution and scaling. Other control plane functions like PCF and AUSF can be separated as shown in Fig. 1. Modularized function design also enables NextGen to support various services flexibly.

Interaction: Each NF interacts with another NF directly. It is not impossible to use an intermediate function to route messages from one NF to another NF. In the control plane, a set of interactions between two NFs is defined as service so that its reuse is possible. This service enables support for modularity. The user plane supports interactions such as forwarding operations between different UPFs.

Roaming: NextGen considers the architecture for supporting roaming in a manner similar to how EPS does. There are two kinds of deployment scenarios, i.e., Home Routed (HR) and Local Break Out (LBO).

Agreements on NextGen functions

In the NextGen core network, existing 4G core network technologies are required to be improved in order to efficiently support various services and requirements. In this section, we introduce the agreements on the NextGen functions discussed in 3GPP, namely, (1) mobility management, (2) session management and service continuity, (3) QoS framework, and (4) network slicing.

3.1. Mobility management

State model: Fig. 2 shows the state models of EPS and NextGen. Three types of states are shown in the EPS model, i.e., EPS Mobility Management (EMM) state, EPS Connection Management (ECM) state, and Radio Resource Control (RRC) state [7]. EMM and ECM states are managed by the core network, where the EMM state represents whether a UE is registered in the EPC, and the ECM state shows whether Non Access Stratum (NAS) signalling connection between UE and MME is established. On the other hand, RRC state is managed by RAN, and it represents whether a connection between UE and RAN, i.e., evolved Node B (eNB), exists or not. A UE in the ECM-CONNECTED state needs to be in the RRC-CONNECTED state, because radio link connection is required to establish NAS signalling connection.

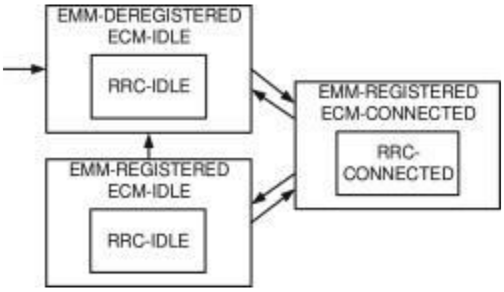
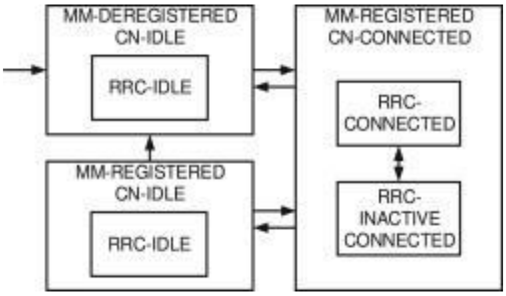


Fig. 2(a). (a) EPS state model.



In NextGen, the MM state of a UE can be MM-REGISTERED or MM-DEREGISTERED state depending on whether the UE is registered in NextGen core network, which is very similar to EMM-REGISTERED and EMM-DETERSITERD states [6]. When the UE is registered, the UE is in either Core Network (CN)-IDLE state or CN-CONNECTED state according to the existence of NAS layer connection. The definitions of MM and CN states are almost the same as those of EMM and ECM states in the EPS. On the other hand, RRC-INACTIVE CONNECTED state is newly introduced as a state of the RRC state model. The new state is proposed to be used as a primary sleeping state prior to RRC-IDLE state [8]. When a UE moves to the new state, both the UE and RAN keep the context information of the

UE's RRC connection, such as UE capabilities and security context, that have been obtained during the RRC connection setup. Therefore, the new state enables a lightweight transition from inactive to active data transmission.

Handover and cell reselection: In the EPS, when a UE is in the RRC-CONNECTED state, the serving eNB evaluates the reported signal strength between the UE and the eNB, and performs a handover procedure when the signal strength is weakened. However, in the RRC-IDLE state, where the eNB is not aware of the existence of the UE, the UE decides whether to camp on the current cell or to reselect a neighboring target cell based on signal strength measurements. This procedure is referred to as cell reselection. In the EPS, the utilized mobility procedures are fixed as handover in the RRC-CONNECTED state and cell reselection in the RRC-IDLE state, respectively. On the other hand, in NextGen, it is expected that the core network is able to flexibly control whether to perform handover or cell reselection for a UE in CN-CONNECTED state.

Location tracking: In the EPC, the location of a UE is managed by the MME. The level of a UE's location is different according to the RRC state of the UE. In the RRC-CONNECTED state, the UE's location is tracked in the cell level while in the RRC-IDLE state, its location is tracked in the Tracking Area (TA) level, which is a set of cells. Similarly, in NextGen, the core network can track the location of UE at the CN location area level in the CN-IDLE state, and the UE location is known at the level of the serving RAN to the core network in the CN-CONNECTED state. A CN location area is an area allocated by the network registering the UE, i.e., TA list. The NextGen core network can allocate TA list composed of cells using new RAT and evolved LTE. On the other hand, in NextGen, RAN also needs to support the location tracking for the UE in RRC-INACTIVE CONNECTED state. In that state, the core network understands that the UE is located within the RAN area, but the RAN needs a new location tracking functionality to determine the exact location of the UE because the connection between the UE and the RAN is not active.

Paging: In the EPS, when Down Link (DL) traffic for a UE in the RRC-IDLE state arrives at the S-GW, the MME performs a paging procedure based on the detected location of the UE. On the other hand, it was agreed to support the following two kinds of paging, i.e., CN paging and RAN paging, in the NextGen system. CN paging, the default paging procedure, is requested by the core network when the UE is in the CN-IDLE state. A newly introduced RAN paging is needed for UE in the RRC-INACTIVE CONNECTED state. Since a UE in the RRC-INACTIVE CONNECTED state is in the CN-CONNECTED state in the core network's viewpoint, the core network simply forwards the data or the singling message to the corresponding RAN when the data or signalling message arrives. Therefore, RAN itself generates the paging message and performs paging to find the exact location of the UE, and then to send the data or signalling message to the UE. The NextGen core network can transmit additional assistance information to RAN for RAN paging.

Mobility on demand: Mobility on demand is a concept to support mobility not to all devices but only to devices that need it [9]. It also includes supporting UE's mobility at its appropriate level. There had been many discussions on mobility on demand, and it is divided into aspects of mobility restriction and mobility pattern (or mobility level). The mobility restriction is addressed in terms of area, which is divided into allowed area, non-allowed area, and forbidden area. The granularity of the area is at least TA level. In the allowed area, UE can communicate through the control plane or the user plane. UE cannot send service request and session management signalling in the non-allowed area. However, periodic registration update is possible. It can also respond to the paging of the NextGen core network. Moreover, emergency calls or multimedia priority service are allowed. In the forbidden area, UE is not allowed to have any communication with the network except for the emergency services.

The mobility pattern is used as a concept to describe the expected mobility of UE in the NextGen core network, not a parameter delivered on the interface defined in the standard. The mobility pattern is determined by considering subscription, location, capabilities, and mobility information statistics of UE, network policies, etc. The NextGen core network can use the mobility pattern to optimize the mobility management procedure and related parameters of UE.

Mobile Originated (MO) only mode: Internet of Things (IoT) service is an important 5G service. IoT devices, e.g., sensor devices, mostly send MO data. For this kind of devices, MO only mode is defined in NextGen, and the NextGen core network determines whether to apply the MO only mode to a UE during the registration procedure based on the subscription data of the UE and network policy. The MO only mode is allocated to a UE, which does not require Mobile Terminated (MT) traffic. Therefore, the UE in MO only mode does not listen to the paging message. The NextGen core network does not need to manage the UE's location while it is registered in the NextGen core network. For optimization, the NextGen core network may decide to deregister after the MO data communication is finished, without transferring the UE's state into the CN-IDLE state in the MM-REGISTERED state, because most functions supported in the CN-IDLE state is not meaningful for the UE in MO only mode, e.g., UE location tracking and reachability management. In such cases, the UE needs to perform attach procedure whenever the MO data transmission is necessary to communicate with the core network.

3.2. Session management and service continuity

PDU session: UE receives services through a Protocol Data Unit (PDU) session, which is a logical connection between the UE and data network. In NextGen, various PDU session types are supported, e.g., IPv4, IPv6, Ethernet, etc. Unlike the EPS, where at least one default session is always created while the UE attaches to the network, NextGen can establish a session when service is needed independently of the attachment procedure of UE, i.e., attachment without any PDU session is possible. NextGen also supports UE establishing multiple PDU sessions to the same data network or to different data networks over a single or multiple access networks including 3GPP and non-3GPP accesses.

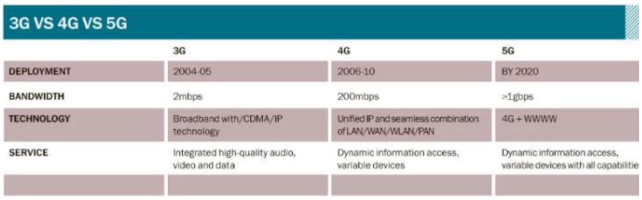
The number of UPFs for a PDU session is not specified. At least, deployment with one UPF is essential to serve a given PDU session. For a UE with multiple PDU sessions, there is no need for convergence point like S-GW in the EPC. In other words, the user plane paths of different PDU sessions are completely disjoint. This implies that there is a distinct buffering node per PDU session for the UE in the RRC-IDLE state.

3.2 Comparison 3g VS 4g VS 5G:

London will be ready to host the fifth generation of mobile telecoms technology by 2020, according to Boris Johnson, the capital's mayor. 5G technology promises mobile data speeds that far outstrip the fastest home broadband network currently available in the UK.

With speeds of up to 100 gigabits per second, 5G will be as much as 1,000 times faster than 4G, the latest iteration of mobile data technology. The gains brought about by 4G are already being felt by businesses whose employees are often on the move.

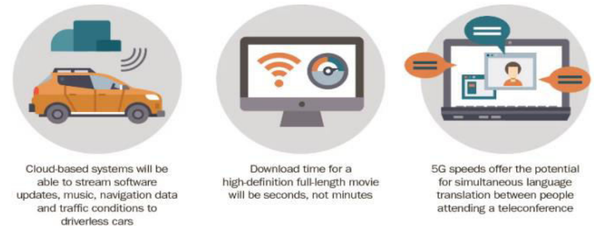
3G does not have the capacity to cope with modern mobile working demands. Workers in urban centres often feel the effects the most, with slow and sometimes non-existent mobile connections common in peak hours. Because 4G networks have higher base speeds, they experience less of this peak-hour strain. There is more than enough capacity to share for core services, such as e-mail and web browsing.



3G VS 4G VS 5G

	3G	4G	5G
DEPLOYMENT	2004-05	2006-10	BY 2020
BANDWIDTH	2mbps	200mbps	>1gbps
TECHNOLOGY	Broadband with/CDMA/IP technology	Unified IP and seamless combination of LAN/WAN/WLAN/PAN	4G + WWWW
SERVICE	Integrated high-quality audio, video and data	Dynamic information access, variable devices	Dynamic information access, variable devices with all capabilitie

POTENTIAL USES OF 5G

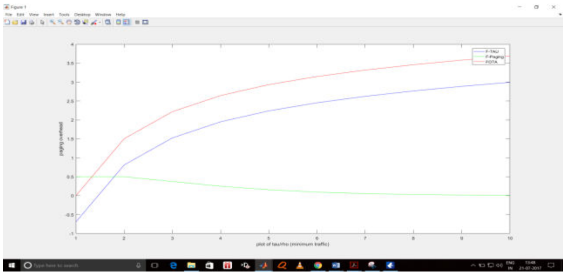


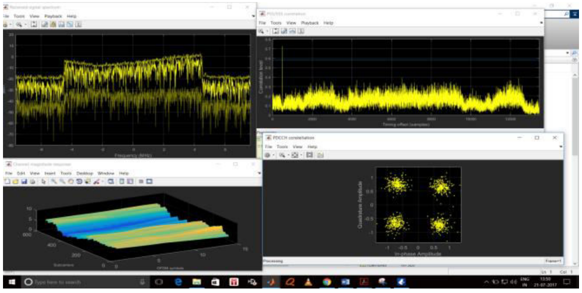
Cloud-based systems will be able to stream software updates, music, navigation data and traffic conditions to driverless cars

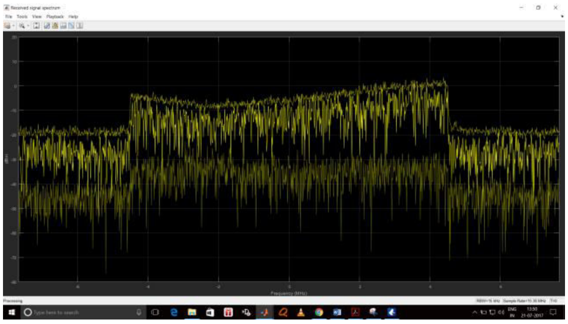
Download time for a high-definition full-length movie will be seconds, not minutes

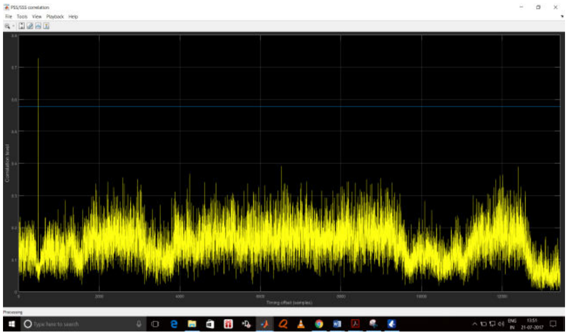
5G speeds offer the potential for simultaneous language translation between people attending a teleconference

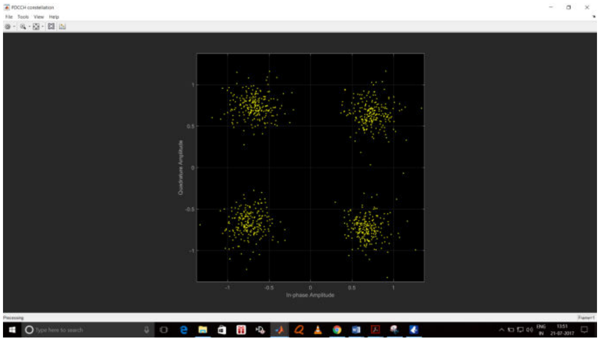
RESULTS AND COCLUSION:

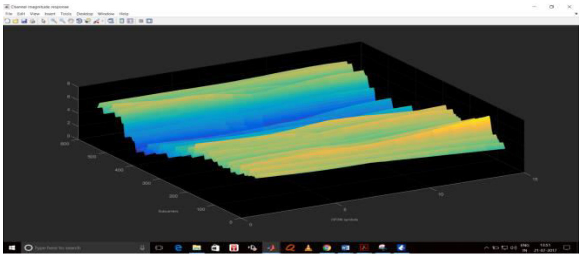












In this paper, we presented the agreements on the architecture and functions of NextGen that aims to create a 5G architecture supporting a variety of services, e.g., virtual presence, eHealth, connected car, etc. In the NextGen architecture, NFs are modularized in order to make the network flexible and scalable. New technology such as network slicing is introduced, and functions such as mobility management, session management, and QoS framework are also developed in different ways from 4G EPS. Detailed procedures about functions as well as normative standardization will be defined in the NextGen Phase 2 study based on the agreements made in Phase 1.

REFERENCES

[1] ITU-R M.2083-0, IMT-Vision-Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond (Sep. 2015).

[2] R.N. Mitra, D.P. Agrawal 5G mobile technology: A survey ICT Express, 1 (3) (2016), pp. 132-137

[3] 3GPP TR 22.891, Feasibility Study on New Services and Markets Technology Enablers; Stage 1, V.14.0.0 (Mar. 2016).

[4] 3GPP TR 38.801, Study on New Radio Access Technology; Radio Access Architecture and Interfaces, V.1.0.0 (Dec. 2016).

[5] 3GPP TR 38.804, Study on New Radio Access Technology; Radio Interface Protocol Aspects, V.0.4.0 (Nov. 2016).

[6] 3GPP TR 23.799, Study on Architecture for Next Generation System, V.1.2.1 (Nov. 2016).

[7] 3GPP TS 23.401, General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access, V.14.2.0 (Dec. 2016).

[8] I.L. Da Silva, G. Mildh, M. Säily, S. Hailu, A novel state model for 5G radio access networks, in: Proc. IEEE ICC, 2016, pp. 632–637.

[9] NGMN Alliance, 5G White paper, V.1.0 (Feb. 2015).