

On the Scalability of Access and Mobility Management Function: The Localization Management Function Use Case

Domenico Scotece^{1b}, *Member, IEEE*, Giuseppe Santaromita^{1b}, *Member, IEEE*,
 Claudio Fiandrino^{1b}, *Member, IEEE*, Luca Foschini^{1b}, *Senior Member, IEEE*,
 and Domenico Giustiniano^{1b}, *Senior Member, IEEE*

Abstract—The adoption of Service-Based Architecture (SBA) in 5G Core Networks (5GC) has significantly transformed the design and operation of the control plane, enabling greater flexibility and agility for cloud-native deployments. While the infrastructure has initially evolved by implementing key functions, there remains significant potential for additional services, such as localization, paving the way for the integration of the Location Management Function (LMF). However, the extensive functional decomposition within SBA leads to consequences, such as the increase of control plane operations. Specifically, we observe that the additional signaling traffic introduced by the presence of the LMF overwhelms the Access and Mobility Management Function (AMF) which is responsible for authentication and mobility. In fact, in mobile positioning, each connected mobile device requires a significant amount of control traffic to support location algorithms in the 5GC. To address this scalability challenge, we analyze the impact of three well-known optimization techniques on location procedures to reduce control message traffic in the specific context of the 5GC, namely a caching system, a request aggregation system, and a service scalability system. Our solutions are evaluated in an OpenAirInterface (OAI) emulated environment with real hardware. After the analysis in the emulated environment, we select the caching system—due to its feasibility—for being analyzed in a real 5G testbed. Our results demonstrate a significant reduction in the additional overhead introduced by the LMF, improving scalability by minimizing the impact on AMF processing time up to a 50% reduction.

Index Terms—5G localization, 5G core, SBA, AMF, localization management function (LMF).

Received 24 February 2025; revised 15 October 2025; accepted 9 February 2026. Date of publication 13 February 2026; date of current version 24 February 2026. This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on “Telecommunications of the Future” (PE00000001 - Program “RESTART”) CUP: J33C22002880001; by 6G-ELSA Project PID2022-136769NB-I00 funded by MCIN/AEI /10.13039/501100011033/ and the European Union ERDF “A way of making Europe.” Claudio Fiandrino is a Ramón y Cajal awardee (RYC2022-036375-I), funded by MCIU/AEI/10.13039/501100011033 and the ESF. The associate editor coordinating the review of this article and approving it for publication was B. Martini. (*Corresponding author: Domenico Scotece.*)

Domenico Scotece and Luca Foschini are with the Department of Engineering and Computer Science, University of Bologna, 40136 Bologna, Italy (e-mail: domenico.scotece@unibo.it; luca.foschini@unibo.it).

Giuseppe Santaromita, Claudio Fiandrino, and Domenico Giustiniano are with the IMDEA Networks Institute, Leganés, 28918 Madrid, Spain (e-mail: giuseppe.santaromita@imdea.org; claudio.fiandrino@imdea.org; domenico.giustiniano@imdea.org).

Digital Object Identifier 10.1109/TNSM.2026.3664546

I. INTRODUCTION

UNLIKE the previous generations of mobile networks, the 5th Generation (5G) of mobile communication networks accommodates a diverse array of compelling use cases and a continuously growing multitude of interconnected mobile devices [1]. In addition to innovations on the radio access, a key enabler of this paradigm shift is the introduction of the Service-Based Architecture (SBA) within the 5G Core Network (5GC). The SBA allows for a flexible, cloud-native core with modular network functions, service-based interfaces, and event-driven communication, improving scalability, efficiency, and interoperability for advanced 5G use cases. Its standardization activity has occurred since the Release 15 of the 3rd Generation Partnership Project (3GPP) [2].

Among the various use cases, Location Based Service (LBS) leveraging 5G networks have been experiencing a growing interest, being one of the most active areas of standardization since the Release 16 of the 3GPP [3]. Commercial applications such as Industry 4.0, autonomous vehicles, emergency services, augmented reality, and Internet of Thing (IoT) for mobile health and precision agriculture exemplify instances where localization accuracy and reliability are imperative, aligning with the specifications set by 5G [4], [5]. The increase in adoption of LBS is fueled by significant technological advancements on the 5G radio access, including network densification, expanded network bandwidth, and scalable support for both antennas and users. Localization also plays a crucial role in optimizing specific network mechanisms such as power control, scheduling and handover management, as well as network planning [6], particularly in the context of millimeter-wave networks [7].

The location techniques for User Equipment (UE)’s as defined by the 3GPP are categorized into three main groups: standalone, user-based, and network-based [8]. Earlier versions of 3GPP prior to 5G primarily emphasized user-based and standalone approaches, wherein users were responsible for self-localization with or without network assistance, often relying on Global Navigation Satellite Systems (GNSS) technology integrated into their mobile devices. On the contrary, network-based methods depend on the cellular network to conduct location measurements for the UE. These measurements rely on Reference Signals defined in 5G for

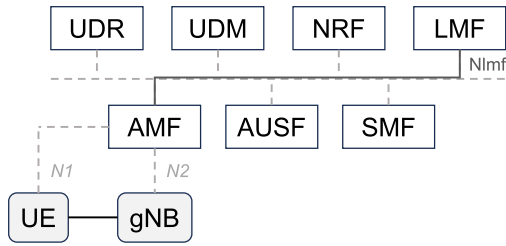


Fig. 1. UE positioning applicable to 5G core service-based architecture.

the purpose of positioning. In situations where GNSS alone may be ineffective, it has been shown that their usage can enhance localization [9], thus providing significant potential for telecom operators and other stakeholders to deliver precise and scalable positioning services.

Since Release 16, the 3GPP has put forth network functions within the 5GC for managing suitable position measurements between the UE and the 5GC. These modules facilitate the implementation of various network location-based applications. In Fig 1, the pivotal component for localization purposes, as suggested by the 3GPP, is the Location Management Function (LMF) [10]. As an integral part of the 5GC, the LMF is capable of executing diverse localization algorithms. It collects location data from the Next Generation Radio Access Network (NG-RAN) associated with each Next Generation Node Base station (gNB) involved in the process, ultimately providing an estimate of the UE's location.

However, with the presence of the LMF and the growing number of connected devices and services, the surge in control traffic raises concerns about scalability and the risk of signaling storms [11]. The elevated number of Network Function (NF) handling incoming traffic, along with the increased volume of messages exchanged between them, has resulted in substantial signaling overhead [12]. In particular, LMF-triggered measurement campaigns (e.g., Sounding Reference Signal (SRS)-based uplink sounding for positioning) can create significant short-burst signaling. For example, we observe that, if each UE's SRS spans 48 Physical Resource Blocks (PRBs) on a 100 MHz bandwidth, practical NR configurations allow multiplexing on the order of hundreds of UEs within just one frame of 10 ms.

This problem is timely: on the practical side, the deployment of 5G networks strictly follows the 3GPP standards and it is occurring at a fast pace. The first commercial 5G networks have first upgraded the Radio Access Network (RAN) to support new commercial 5G smartphones, with ongoing SBA deployments. More recently, 5G standalone networks have started to be deployed as well, with the cloudification of the 5GC through the SBA. In particular, location support such as LMF and corresponding network protocols have started to be supported by major vendors, and will be soon part of future 5G standalone deployments. Therefore, the problems addressed in this work should be promptly addressed for scalable operation of the network while supporting location functions.

Accordingly, 3GPP establishes Quality of Service (QoS) policies for location management functions [13], specifically defining various classes of QoS for location services, including

Best Effort Class, Multiple QoS Class, and Assured Class. Each class is defined based on the parameters of Horizontal Accuracy, Vertical Accuracy, and Response Time. When combined with the performance requirements for positioning service levels [14], this provides a comprehensive view of the specific requirements for each positioning service level. For instance, positioning service level 1 requires adhering to a maximum latency of 1 second. Essentially, with a specific service layer, we can adjust between different classes of positioning QoS to meet service requirements even during network congestion.

The main research question that we address in this work is how to efficiently use scalability techniques to lower the overhead at the Access and Mobility Function (AMF) side upon including the LMF into the 5G core and, at the same time, maintaining the protocol compliant with the 3GPP standard. While our past work has introduced a preliminary evaluation of the LMF [15], this paper aims at solving practical issues with the current SBA deployments. In particular, the primary challenge is to develop scalability techniques tailored for location-based services in the 5GC that can address the following issues: *i)* significant signaling overhead between network functions, in particular in the AMF during the execution of location-based service procedures; *ii)* increased latency on the location procedure times according to the 3GPP location service QoS specifications [13]; and *iii)* limited location procedures defined by the 3GPP protocol. Therefore, to address the aforementioned issues, in this work we propose the following:

- 1) We propose to use three well-known scalability techniques such as a cache system, a request aggregation system, and a network service scalability that support location service QoS. In particular, we implement into the LMF, gNB, and the AMF the location function and the efficient procedures to handle location service.
- 2) We conduct extensive evaluations of the proposed scalability techniques in the OpenAirInterface (OAI) simulation. Moreover, we evaluate the cache system in a real 5G testbed.
- 3) The results demonstrate that the proposed scalability techniques can meet service requirements in challenging situations like signal storms according to the defined QoS and service requirements. Compared to the case of using standard LMF procedures, our proposed scalability techniques reduce the positioning service latency.

The remainder of this paper is organized as follows. In Section II, we provide detailed background information and outline the motivation behind this work. We introduce the proposed scalability techniques in Section III. Implementation details are proposed in Sections IV. We evaluate the performances in Section V. After reviewing the related works in Section VI, we report a thorough retrospective discussion of our solution in Section VII. The conclusion of this paper is presented in Section VIII.

II. BACKGROUND AND MOTIVATION

In this section, we delve into the structure of 5GC, focusing on its LMF and examining how communication between its

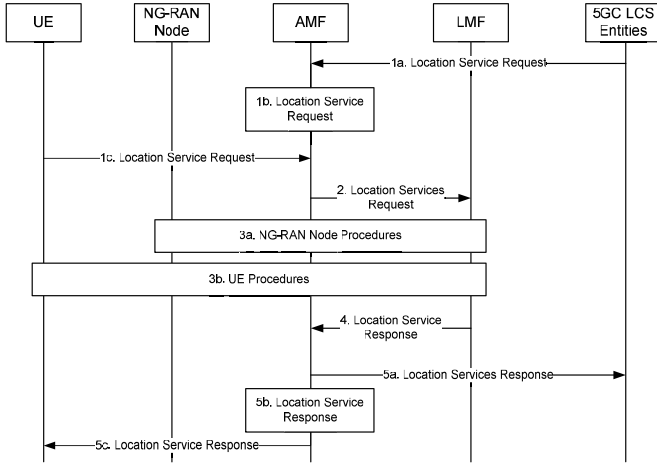


Fig. 2. Location service support by NG-RAN.

entities is established in the control plane. Moreover, we provide motivation for this work.

A. Background

The 3GPP TS 38.305 [16] defines the methodologies in 5G System (5GS) for UE positioning, and it also introduces the supporting architecture that enables localization. When another NF sends a request to the AMF for a specific location service related to a target UE, the AMF proceeds to forward a location services request to an LMF. The LMF handles the location services request, which may involve providing assistance data to the target UE for UE-based and/or UE-assisted positioning, or directly performing positioning of the target UE. Subsequently, the LMF sends the location service result back to the AMF, which then returns it to the requesting entity.

Figure 2 shows the overall sequence of events applicable to the UE, NG-RAN and LMF for any location service. The protocol defines two different location procedures such as NR Positioning Protocol A (NRPPa) and LPP LTE Positioning Protocol (LPP), steps 3a and 3b respectively in Fig. 2. On the one hand, the NRPPa protocol carries information between the NG-RAN Node and the LMF. On the other hand, the LPP directly involves the UE. Since this work is mainly focused on the NRPPa protocol, we investigate the NRPPa protocol that is transparent to the AMF. In particular, the AMF transparently routes the NRPPa from the LMF to the NG-RAN. To ensure this, the NG Application Protocol (NGAP) is involved in the communication between AMF and NG-RAN. The NGAP protocol is specified in the 3GPP TS 38.412 [17].

As stated before, the NRPPa is completely transparent to the AMF. The AMF efficiently directs the NRPPa PDUs without being aware of the specific NRPPa transaction, using a Routing ID that corresponds to the relevant LMF node over the NG interface, ensuring transparent routing. Figure 3 shows NRPPa PDU transfer between an LMF and NG-RAN Node to support positioning of a specific UE. The procedure starts when the LMF sends a NRPPa message to the serving

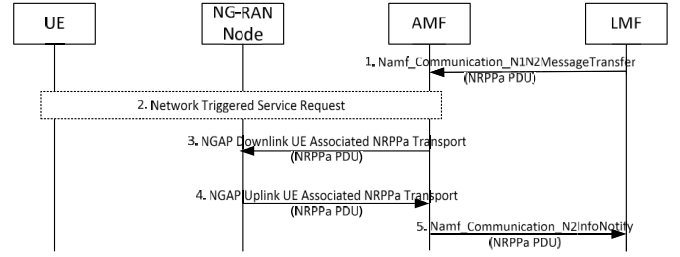


Fig. 3. NRPPa PDU transfer between an LMF and NG-RAN node for UE positioning.

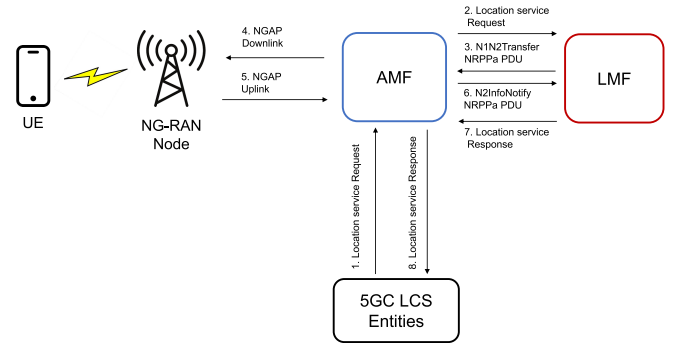


Fig. 4. 3rd-party service localization message flow.

NG-RAN Node to locate a UE. With Step 1, the LMF then invokes the *Namf_Communication_NIN2MessageTransfer* service operation towards the AMF to request the transfer of a NRPPa PDU to the serving NG-RAN Node for the UE. The *Namf_Communication_NIN2MessageTransfer* is a REST request defined in the 3GPP TS 29.518 [18] and includes the NRPPa PDU together with the LCS Correlation ID in the N2 Message Container. Then, step 3, the AMF forwards the NRPPa PDU to the serving NG-RAN Node in an **NGAP Downlink UE Associated NRPPa Transport** message including the ROUTING ID related to the LMF. After that, the NG-RAN Node sends an NRPPa PDU to the AMF in an **NGAP Uplink UE Associated NRPPa Transport** message, step 4. Finally, the AMF sends the *Namf_Communication_N2InfoNotify* service operation towards the LMF indicated by the Routing ID. Overall, the location manager protocol designed by 3GPP involves at least 8 network communications between entities in the 5GS.

B. Motivation

The issues that we address here are how to efficiently handle multiple UEs location requests and the support for the last OAI implementation. In this section, we will explain the problem, the assumptions that we have considered for this problem, and directions to solve this problem.

1) *Problem Description*: We assume a 5GC network composed of a single AMF module connected with multiple gNB. Moreover, we assume there are third-party services that request UEs location information. As stated before, the LMF module is in charge of handling location service requests by leveraging the AMF module to complete the UE localization. In particular, for each UE location request, at least 8 different

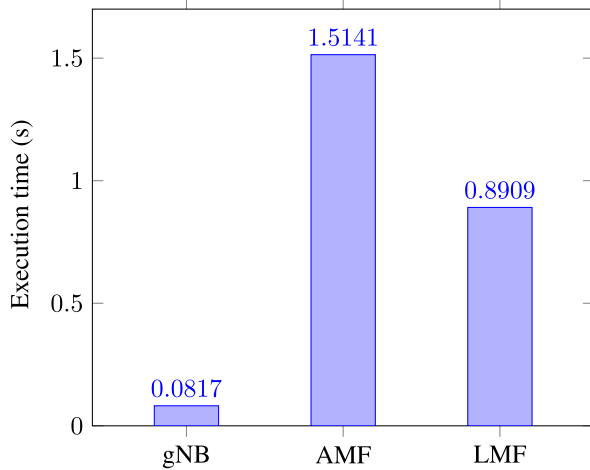


Fig. 5. Total execution time of gNB and key NFs over 500 location requests.

messages are involved in the communication, as is represented in Fig. 4. The communication flow it can be divided into two stages: the AMF and LMF stage, and the AMF and the NG-RAN stage. As discussed before, the first stage involves the REST protocol, while the second stage involves the NGAP protocol. This results in an increased number of messages that AMF has to handle to support location-based services.

The total execution time for NFs involved in the location protocol, shown in Fig 5, is calculated as the cumulative time required to handle 500 location requests from a network service following a Poisson distribution. As you can see, the load demanded by the AMF is 100% higher than that required by the LMF. However, in a scenario where the AMF must process multiple location requests in addition to standard system signals, it may sometimes be unable to handle everything.

Finally, the 3GPP protocol TS 38.305 [16] provides the mechanisms to support or assist the calculation of the geographical position of a UE. At this stage, this specification limits its purpose to outline the NG-RAN UE positioning architecture, its functional entities, and the related operations required to support positioning methods. However, it does not include how to exploit the obtained positioning results within the 5GC or the NG-RAN, nor does it introduce mechanisms aimed at optimizing or accelerating the positioning procedures.

2) *Assumptions*: In this paper, we address the issue of signaling storms, noting that compliance with 3GPP standards has already been shown to be inefficient [19]. The widespread deployment of smartphones, tablets, and wearable devices, each equipped with a diverse array of sensors, positions them as valuable sources of information. Requesting location information of multiple devices might be a very challenging task according to the 3GPP location management protocol. The signaling storm at the control plane can effectively create congestion at the AMF module that, standing at protocol, has to manage every control signal at the 5GC [20].

3) *Directions*: Evidence suggests that the scalability of the 3GPP location management protocol is insufficient to handle a large number of location requests. Consequently, the location

management protocol should align with contemporary design principles found in the existing literature. For instance, the work presented in [21], supports the low latency communication in the 5GC between NFV-based NFs. This shall help NFs to communicate with lower latency. On the contrary, the work presented in [22] proposes another direction such as the study of stateless mobile core network functions. Lastly, the work presented in [23], evaluates the performance of the state-of-the-art network stacks for CPU-intensive VNFs of the 5G mobile packet core.

On the contrary, while these approaches highlight promising directions for improving the performance and scalability of mobile core networks, there is still insufficient discussion on how such strategies could be applied to the specific limitations of 3GPP-defined location procedures. In particular, mechanisms to reduce control-plane overhead, mitigate AMF congestion, and accelerate positioning protocols remain largely unexplored within the standard specifications. We next discuss three potential scalability techniques to mitigate AMF congestion when handling multiple location requests. In particular, we also highlight possible strategies to navigate the limitations of location procedures as defined by the 3GPP protocol, ensuring compliance while improving efficiency.

III. SCALABILITY TECHNIQUES

As we have motivated in Section II-B, the introduction of the location management introduces scalability problems. To alleviate such issues, this section provides details on implementing different scalability techniques in the field of 5GC. Specifically, we discuss and analyze the pros and cons of the various techniques that we considered for the Location Management Services protocol.

A. Location Information Caching

A localization request must align with the existing Key Performance Indicator (KPI), such as latency, accuracy, and inter-packet gap (IPG) specified in the 3GPP requirements [10]. However, when a 5GC Location Service (LCS) entity starts a new localization request, it terminates at the LMF and leverages a recent, accurate localization result. This requires the definition of QoS for requests, enabling the dismissal and termination of those with lenient requirements that allow for flexibility. By establishing these objectives, requests with more relaxed criteria can be ignored and effectively terminated, ensuring the utilization of a pre-existing, current localization outcome. Nevertheless, it is important to delve into the trade-offs regarding the extent to which termination occurs—greater termination yields greater benefits, while potentially leading to lower localization accuracy. This depends on the specific KPIs and QoS for UEs location requests. In conclusion, the 3GPP protocol defines periodic location request messages (such as a pub-sub subscription schema). Still, it does not provide a flexible schema to re-use previous location information according to the KPIs and QoS.

B. Location Requests Aggregation

The main idea is to aggregate location requests from the LMF to the AMF, reducing excessive inter-NF

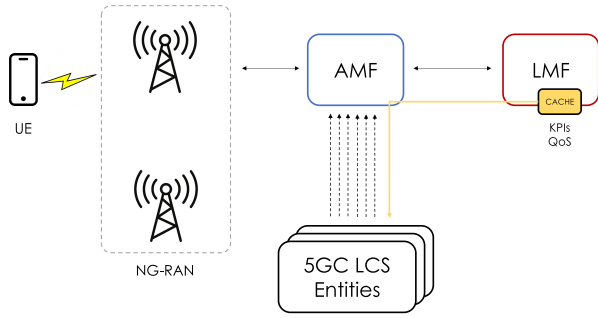


Fig. 6. Efficient schema for location information caching.

communications. This approach significantly decreases message overhead, as each location service operation typically involves multiple signaling exchanges. In scenarios involving multiple location requests, these can be efficiently aggregated into a single location request. However, there exists a tradeoff depending on how often and how many messages are aggregated. The higher number of aggregated requests, the higher is the overhead reduction at the expense of latency (some messages need to wait longer than others to be served). The higher the frequency of sending aggregated requests, the lower the impact on latency variability at the expense of a lower overhead reduction. Data aggregation techniques as such have been studied in the literature in recent years and have proven to be effective in several contexts [24], [25]. However, the 3GPP protocol, as default, does not provide a mechanism to aggregate different location requests.

C. Amf Scaling

A potential strategy to reduce inter-NF communications between the LMF and the AMF for location service operations is to scale the AMF, as it is responsible for handling these requests. There exists preliminary results that aim to understand how the AMF function scales with load [26]. However, the utilization of conventional orchestrators like ETSI Management and Orchestration (MANO) and Kubernetes for the automated scaling of the NFs results in a heightened level of intricacy and elevated management expenses [27]. The balancing act lies in determining an equitable quantity of AMF instances while simultaneously minimizing the volume of control messages exchanged among the AMFs, ensuring an efficient and streamlined operational process.

IV. SYSTEM DESIGN AND IMPLEMENTATION

This section provides details on the implementation of the previously discussed techniques for avoiding the congestion of the AMF system in the LMF use case. In this work, we utilize the implementation based on the OAI version¹ both for the 5GC and the NG-RAN part.

A. Location Information Caching

According to the 3GPP protocol, location requests start and terminate at the LMF. As previously mentioned, each location

request passes through the AMF at least twice. Therefore, the algorithm proposed in this work for location information caching is based on the schema shown in Fig. 6. The cache system is provided at the LMF stage and it works based on the KPIs and QoS for UEs location requests.

Algorithm 1 aclmf Caching Procedure

Require: ΔT : cache validity threshold (in seconds)

Ensure: Return the most recent valid UE position if valid

- 1: **Input:** Location request for a specific UE received from the AMF
- 2: Retrieve UE_ID (e.g., SUPI/IMSI) from the incoming request
- 3: **if** UE_ID not found in the cache **then**
- 4: Execute the positioning procedure for the given UE
- 5: Obtain the estimated position (x, y, z) and current timestamp t_{now}
- 6: Store $\{UE_ID, \text{position}, t_{now}\}$ in the cache
- 7: Return the new position to the AMF
- 8: **else if** the difference $(t_{now} - t_{cached}) < \Delta T$ **then**
- 9: Retrieve the cached position from the database
- 10: Return the cached position to the AMF
- 11: **else**
- 12: Execute the positioning procedure for the given UE
- 13: Obtain the estimated position (x, y, z) and current timestamp t_{now}
- 14: Update $\{UE_ID, \text{position}, t_{now}\}$ in the cache
- 15: Return the new position to the AMF
- 16: **end if**

The caching mechanism saves at least half of the messages that pass through the AMF module. The proposed algorithm's procedure is shown in Algorithm 1. The core idea behind this algorithm is to reuse the last known position of the UE, if it is still valid, that is, the stored position has not yet expired. However, it only operates at the LMF level, avoiding any intervention at the AMF level. Furthermore, the important point is that the caching mechanism does not impact in any way the 3GPP location protocol. This makes caching an easy and fast mechanism to implement even in cases where a 5G system is already in operation, as it would only require deploying the new LMF module with a minimal impact to the entire system.

We implemented the caching system using the open-source 5GC implementation provided by OAI. In particular, we leveraged and modified the LMF module.² We added a dedicated function that intercepts location requests from external applications. This function utilizes the UE ID injected as the IMSI value in the SUPI field, as specified in the 3GPP 5G APIs [28], to check whether a valid location value already exists in the cache.

Note that, if the cache does not contain the UE's location information or the location data is expired, the location procedure follows the standard process as described in Section II and adheres to the 3GPP location procedure.

¹<https://gitlab.eurecom.fr/oai>

²We build upon the version corresponding to commit hash 62870658.

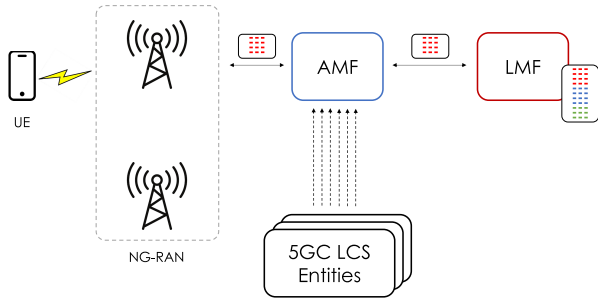


Fig. 7. Efficient schema for Location requests aggregation.

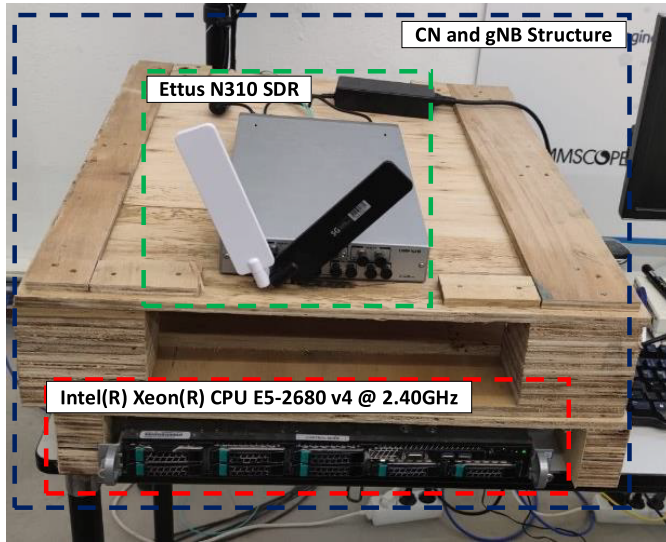


Fig. 8. 5GC (CN in the figure) and gNB of the real 5G testbed.

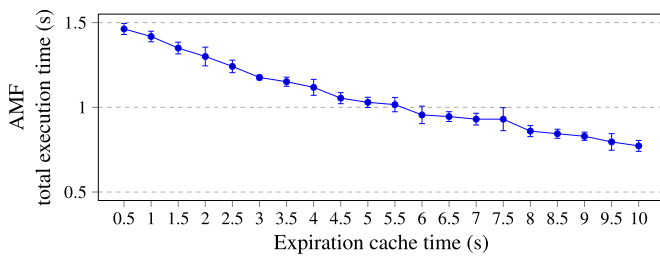


Fig. 9. Total AMF execution time over different cache expiration time.

B. Location Requests Aggregation

Since the 3GPP defines the 5GC as a network built on microservices, it is important to address the potential increase in the number of requests to prevent possible issues. Especially, if these microservices are hosted in cloud environments where service auto-scaling has a high cost. However, request aggregation patterns for services hosted in cloud environments have been studied in recent works [29]. Here, we propose a solution for multiple location requests aggregation based on the schema shown in Fig. 7. Conceptually, multiple location requests from different UEs can be aggregated, for example, by grouping UEs belonging to the same user class or sharing the same priority level.

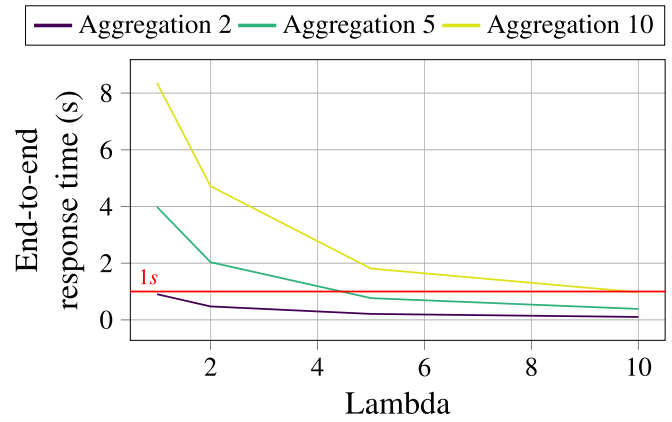


Fig. 10. End-to-end time for a request over different lambda values and aggregation level.

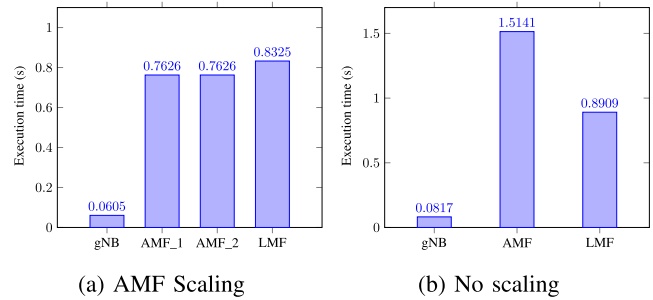
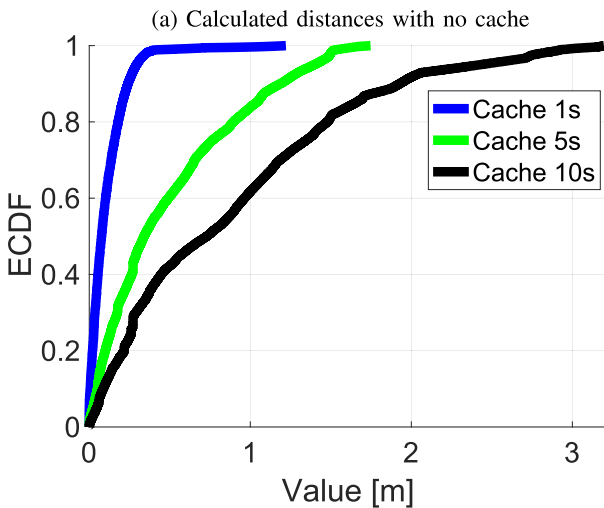
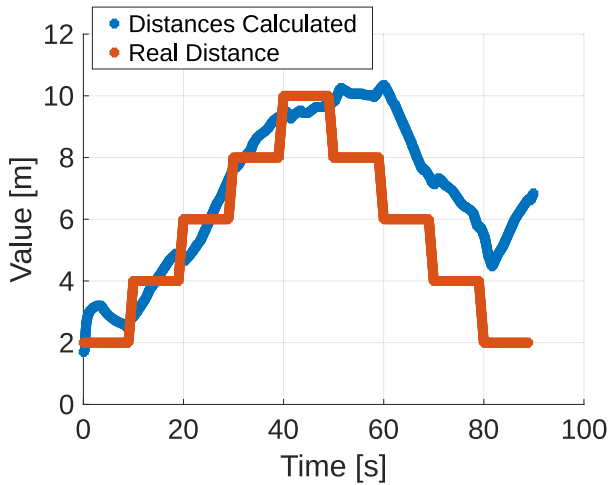


Fig. 11. Total execution time for the gNB and key NFs over 500 location requests.

Contrary to the location caching mechanism, it is necessary to slightly modify the 3GPP protocol for location request aggregation. In particular, based on the last specification [30], the NRPPa protocol specifies several functions for gathering UE location information from the NG-RAN node, including, for instance, the **E-CID Location Information Transfer** and the **OTDOA Information Transfer**. Conversely, the NRPPa Transport Messages include the **DOWNLINK UE ASSOCIATED NRPPa TRANSPORT** and the **UPLINK UE ASSOCIATED NRPPa TRANSPORT** [31]. The first message, sent by the AMF, carries the NRPPa message over the NG interface. Meanwhile, the second message, sent by the NG-RAN node, serves the same purpose of transporting the NRPPa message over the NG interface. Both messages include the NRPPa-PDU field, defined as a generic OCTET STRING that should contain the related location procedure function. This allowed us to leverage the field for aggregating multiple requests. We remark that we do not bypass or decrypt control-plane encryption: NRPPa PDUs are transported as opaque application payloads inside NGAP messages and the AMF only transparently routes these OCTET-STRING PDUs between the LMF and NG-RAN; our aggregation uses the NRPPa_PDU OCTET_STRING for bundling but does not inspect or decrypt the payload, therefore the 3GPP control-plane security model is preserved. An example of the code modifications we implemented is presented in Listings 1 and 2.



(b) ECDF comparing the deviation of localization results of various cache values with respect to the case without cache

Fig. 12. Evaluation of the information caching technique on the Real 5G Testbed evaluation.

```
void DownLinkUeAssociatedNRPPaTransportMsg
::setNRPPa_PDU(char* s) {
    OCTET_STRING_fromBuf(&NRPPa_PDU,
        s, strlen(s));
}
```

Listing 1. DownLinkUEAssociatedNRPPaTransport.cpp.

```
...
DownLinkUeAssociatedNRPPaTransportMsg
nrppamessage = {};

nrppamessage.setNRPPa_PDU("PROCEDURE -
    ALL UE IDs");
...
```

Listing 2. amf_n2.cpp

As shown, the fundamental approach, which does not completely alter the 3GPP protocol, is to utilize the NRPPa_PDU field within the *DownLinkUEAssociatedNRPPaTransport* class. This approach allowed requests to be

aggregated in the LMF, enabling the AMF to initiate a single location procedure for multiple requests. Finally, the NG-RAN utilizes the same field to populate the required parameters for the location process.

C. Amf Scaling

In this section, we present a possible implementation of the scaling algorithm for the AMF. Specifically, the proposed solution operates in two distinct areas: one focuses on scaling the AMF module up and down, while the other addresses the load balancing algorithm used to manage UE location requests. Therefore, for scaling the AMF up and down the proposed solution leverages the OAI Docker Compose deployment script to provide multiple instances of the AMF.

Algorithm 2 Round Robin Load Balancing

```
1: procedure RoundRobinLoadBalancing(AMF_List,
    Requests)
2:    $num\_AMFs \leftarrow length(AMF\_List)$ 
3:    $current\_index \leftarrow 0$ 
4:   for each req in Requests do
5:     assign_request(AMF_List[current_index], req)
6:      $current\_index \leftarrow (current\_index + 1) \bmod$ 
        $num\_AMFs$ 
7:   end for
8: end procedure
```

On the contrary, we implemented a round robin load balancing algorithm at the LMF level to distribute UE location requests across a group of AMF servers. A snippet of the straightforward round robin algorithm is shown in Algorithm 2. The Round Robin Load Balancing algorithm evenly distributes incoming location requests among the available AMFs. Each request is sequentially assigned to an AMF in a cyclic order, ensuring that all instances receive an approximately equal share of the workload. Once the last AMF in the list is reached, the allocation restarts from the first one. While this approach provides a simple yet effective mechanism for balancing processing loads without requiring complex state tracking or performance estimation.

Finally, note that it is crucial to have a dedicated framework, such as Kubernetes, to automatically scale the AMF up or down based on the control plane flow. While this proposed solution is a straightforward approach to scalability, it is highly resource-intensive.

D. Implications of Virtualization on the 5G SBA

The 5G SBA is by design highly virtualized and this virtualization has practical consequences for control-plane functions such as the LMF. Virtualized deployments introduce performance variability (due to multi-tenant resource sharing and noisy neighbors), cold-start overheads for newly spawned instances, and non-negligible state-transfer costs when sessions are migrated across instances. These factors affect latency, freshness of location estimates, and cost-efficiency; they therefore motivate our design choices for caching, token-aware aggregation, and the use of state-affinity or consistent-hash routing to reduce state migration. Prior experimental and

system work documents these effects and mitigation strategies: virtualization and network-stack variability can materially change control-plane latency and throughput [23], autoscaling systems that use warm pools and proactive scaling reduce cold-start penalties and improve resilience [32], and high-availability/session-replication schemes increase state-transfer overheads that routing affinity can mitigate [33]. Additionally, decentralized authorization and slice isolation in modern core frameworks demand careful token/scope design when exposing SBA functions to external consumers [34], and slice-level scalability tradeoffs further motivate per-slice freshness and admission-control policies [35]. Our previous work has observed and quantified some of these effects in LMF and edge/emulation settings, specifically caching accuracy, age and aggregation-driven AMF load reduction [15], [36].

V. PERFORMANCE EVALUATION

We now evaluate the proposed scalability techniques. First, we explain the methodology and the evaluation settings that we used for the experiments. Second, we evaluate the proposed scalability techniques in terms of the total AMF execution time. Finally, we show the results of experiments executed on a real-testbed to test the cache system with a location procedure implemented in the core of the network.

A. Evaluation Methodology

The 5GC, UEs, and the gNB operate on an Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz, 64 GB of RAM and Linux Ubuntu 18.04 LTS. All components are implemented using the dockerized version of OAI. The UEs and the gNB are simulated, and the 5GC communicates with each other through local interfaces. The total number of simulated UEs is limited to 10 due to OAI simulation constraints. Additionally, we increase the number of threads used by the OAI web server from 1 to 10, as we encountered crashes when the system was heavily stressed during a surge of location requests.

To assess the performance of the proposed scalability techniques, we implement our version of the LMF, AMF, and gNB based on the OAI codebase, as outlined in the previous section. In particular, we rely on these commit hashes^{3,4,5}. Specifically, we implement a generic Application Function (AF) that sends location requests based on varying values of a lambda distribution. Location requests start from the latter module and are processed in the AMF before being forwarded to the LMF. Finally, we develop a lightweight Python application that generates and manages location requests. In our evaluation analysis, we focus on the total time taken for requests to be processed within the AMF.

To evaluate our methodology in a real 5G testbed scenario, we use the same configuration described above for the 5GC and the gNB, adding a radio interface to the gNB in order to communicate with the UE over the air. Specifically, the gNB is equipped by an Ettus N310 [37], a Software-Defined Radio (SDR) which supports a comprehensive range of 5G NR

TABLE I

TOTAL AMF EXECUTION TIME OVER DIFFERENT AGGREGATION LEVELS

Aggregation	Average AMF execution time (s)	Std-dev (s)
Aggregation 2	0.7291	0.0272
Aggregation 5	0.3036	0.0061
Aggregation 10	0.1522	0.0026

bands. Finally, a Google Pixel 7 Pro with a 5G USIM card is used as UE. This smartphone is powered by the Google Tensor G2 chip and runs on Android 13 software [38]. Figure 8 shows the structure of 5GC and gNB described. The radio interface of the gNB (Ettus N310 SDR) is heightened with a pallet in order to have the same height as the smartphone.

B. Experimental Results

1) *Caching Results:* Using the LMF caching Algorithm 1 outlined in the previous section, this testbed evaluates the total execution time for processing location requests at the AMF module. To achieve this, we measure the processing time at the AMF for 500 location requests across 10 simulated users while varying the cache expiration time. As previously explained, the location information for each user is stored locally in the LMF for a specific duration. Based on this timing, if the LMF receives a location request for a specific user and a valid value already exists in the cache, that value is returned directly, bypassing the execution of the location protocol. For this specific experiment, the cache expiration time for each user varies from 0.5s to 10s and the location requests follow a Poisson distribution with a lambda value set to 1. This specific range has been chosen because 3GPP mandates that the 5G system must support positioning services with a Time To First Fix (TTFF) of less than 10s [14]. The results, presented in Fig. 9, show the average and standard deviation of the total execution time at the AMF across 30 different complete experiments. We observe a linear decrease in the total execution time at the AMF as the cache expiration time increases. Therefore, the cache expiration duration should be adjusted based on the user's mobility. As thoroughly analyzed in our real testbed evaluation, using a high expiration time for the cache system results in a significant loss of information.

2) *Aggregation Results:* We extensively evaluate and confirm the feasibility of the location request aggregation solution. This section showcases a detailed selection of experimental results obtained from our simulated deployment scenario. The simulation measures the end-to-end response time for a single location request by analyzing various Lambda values and aggregation levels.

Initially, we evaluate the total AMF execution time by generating 500 location requests across 10 different simulated UEs with varying lambda values. Specifically, we conduct 500 location requests for lambda 1, lambda 2, lambda 5, and lambda 10. The distribution of the requests is based on the analysis of user inter-arrival at production LTE base stations that was also used in our previous work [36]. Table I shows the average AMF execution time across 30 trials. Notably, the total execution time at the AMF decreases as the aggregation level increases and it is independent of the lambda value. Therefore,

³commit hash 62870658 for the LMF.

⁴commit hash c21d74c4 for the AMF.

⁵commit hash 2b717d49 for the gNB.

a more relevant test focuses on the end-to-end response time for each request. As discussed in Section I, 3GPP establishes QoS policies for location management functions [14]. Most of the service levels defined in the standard specify that the positioning service latency is set to 1 second. This ensures that a service consumer must receive the location information in less than 1 second. Figure 10 illustrates the end-to-end response time for various lambda values and aggregation levels. As shown, location requests with low lambda values require a low aggregation level to ensure responsiveness remains under 1 second. On the contrary, the system can support high aggregation levels when the frequency of location requests increases. Therefore, the aggregation level can be tailored to ensure the end-to-end response time fulfilling QoS policies for location management.

3) *Amf Scaling Results*: For AMF scaling, we use a basic round robin mechanisms because the AMF instances are homogeneous, with same Docker image and CPU/RAM allocation. Further, the distribution of location requests is drawn from real-world traces [36], which makes the methodology for this evaluation sound. We report in Fig. 11 the comparison in total execution times among the 5GC modules involved in the location procedure between a system with two AMFs and a system with one AMF. The drop in AMFs execution time is calculated considering the round-robin algorithm specified in Section IV-C. We utilize the Docker Compose file from the OAI GitLab project to scale the AMF across two separate network units. Requests received by the LMF were seamlessly forwarded to a single AMF. In this specific experiment, we generate 500 location requests with a fixed lambda value set to 1. The results indicate that the reduction in a single AMF execution time depends on the number of AMF instances present in the system. However, generally speaking, this scalability technique proves to be the least efficient due to the high cost of instantiating multiple AMF instances.

C. Real 5G Testbed Evaluation

For the real 5G testbed evaluation, we use location information caching. The limited number of UEs available, i.e one Google Pixel 7 Pro smartphone, makes the use of location request aggregation and the AMF scaling meaningless, as those are designed for a large volume of UEs. Although the caching system is also designed for a large amount of UEs and localization requests, the benefits can also be appreciated for a single UE.

Since the total execution time at the AMF for processing location requests is virtually the same as we showed in Fig. 9, scaled for one UE, we focus the study on the location information loss as the selected cache expiration time varies. The localization algorithm used leverages the cross-correlation between a known transmitted signal and its reception, so it is possible to determine the time, and thus the distance, between the transmitter and the receiver. The known signal is called Sounding Reference Signal (SRS).

Fig. 12a shows the estimated distance tracking (blue line) of the UE in an experiment where the UE stands stationary for 9 seconds and then moves in the approximately one-second time lapse to the next position (real distance in red

line). Fig. 12b shows the location information loss as the selected cache expiration time varies in terms of Empirical Cumulative Distribution Function (ECDF), Indicating the error due to using location information caching method versus not using it.

We observe that we have a median error of around $0.0764m$, $0.3407m$, and $0.7302m$, for a cache expiration time of $1s$, $5s$, and $10s$, respectively, and an 80-th percentile around $0.1795m$, $0.8944m$ and $1.4775m$. We can infer that, for systems that require sub-meter accuracy, a few-second cache is reliable and does not have high location information loss, while a high cache is still a good trade-off for systems that do not require high localization accuracy.

VI. RELATED WORK

We present related works in the areas of 5G control plane network functions scalability and an overview of the research on location procedures in 5G networks.

A. Control Plane Scalability

The design and implementation of mobile core systems has been an active area of research in recent years, with the majority of studies, including ours, focusing on the control plane. The study in [32] introduced CoreKube, a cloud-native mobile core design that achieves truly stateless and efficient processing by employing a generic worker capable of handling any control plane message. The solution leverages Kubernetes for orchestrating control plane modules. While we proposed a similar approach for managing AMF scaling, the alternative scalability techniques, specifically tailored for the LMF module, require less effort to implement. Solutions like [21] achieve low latency in the 5GC control plane by leveraging software-based 5GC NFs and well-known technologies like shared memory communication and zero-cost state update. However, at this stage, this solution supports a limited number of user sessions. In [39], the authors have proposed a stateless and procedure-based system for 5GC specifically for four different control plane procedures. In this way, they significantly reduce the control signal traffic. Overall, both solutions require the implementation and use of specific protocols to enable rapid communication between 5GC NFs.

Following the same research line, the work proposed by Ashwin et al. [23] evaluates the performance of several state-of-the-art network stacks in the context of the VNFs of the 5GC. The results reveal that while modern stacks surpass the Linux kernel stack in handling I/O-intensive VNFs, the performance difference is less pronounced for CPU-intensive VNFs in the 5G core. Another approach for designing 5GC is presented in [40] where the authors present an approach to run the 5GC NFs as Cloud-Native applications that can be easily scaled on demand to match the constraints of different verticals.

The 5GC is designed as a set of VNFs hosted on Commercial-Off-the-Shelf (COTS) hardware. According to this, some works have been started evaluating the performances of the Control Plane on Public Cloud. In [41], the authors evaluated the performance overhead for various 5G

use cases using different core deployment strategies in the Amazon Web Services (AWS) infrastructure. Finally, the work proposed in [42] examines the evolution of 5G SBA to enable flexible service routing, discovery, and a service dataflow layer, streamlining Service Function development and optimizing resources.

B. 5G Localization

Existing work on localization has mainly focused on the study of suitable methods using 5G NR Reference Signals measured in the NG-RAN, addressing positioning and its applications [43], [44]. However, it is not usual for localization context work to address resource scalability aspects.

The authors in [45] proposed the positioning quantification as an analytics function in the 5GC. They developed a virtualized system based on a simulated LMF module and a machine learning-based approach for predicting and updating the level of localization uncertainty in a monitored environment. The authors in [46] analyze a latency reduction approach based on preconfigured assistance data. In this approach, a UE receives assistance data needed to process Positioning Reference Signal (PRS) at its current and anticipated future locations together with information on when to use the assistance data. Results show that this approach reduces average latency by 50% compared with traditional approaches and achieves a 20% reduction in power consumption compared to Release 16. Finally, the authors in [8] proposed a high-level analysis of cutting-edge applications in 5G and beyond, focusing on network-centric location-based analytics, and a proposal to send positioning data from LMF to a new module called LDAF to extract on-demand analytics that serves third-party applications or optimize the network performance.

Our recent work has introduced a preliminary evaluation of the LMF [15], [36] and these have been a starting point for this work. We introduced a 3GPP standard compliant 5G LMF design, implementation and evaluation. The works were focused on the evaluation of LMF performance and user QoS, with parameters such as CPU utilization, throughput and latency, proving that the performance satisfies the 5G KPIs required by 3GPP for localization.

VII. DISCUSSION

This paper primarily focuses on minimizing the overhead of control plane messages processed by the AMF for each location service request. For this purpose, we presented three different scalability techniques and we showed results about AMF execution times and end-to-end response time according to our implementation and realistic simulation. Consider a scenario involving remote sensing for agricultural development controlled by a private 5G Network. In this context, advanced technologies like drones, and IoT sensors are deployed to monitor and manage agricultural activities. Specifically, multiple sensors are deployed to cover a large area, each serving different mobility purposes. In the case of *Location Information Caching*, two main trade-offs arise. The first is between location accuracy and the cache expiration value. As illustrated in Fig. 12, for low-mobility sensors, a longer cache

expiration value can be advantageous, whereas for drones or other high-speed devices, a shorter and carefully configured expiration value is preferable. In both cases, properly tuning the cache expiration at the LMF substantially reduces location information errors. The second trade-off concerns the AMF execution time and the cache expiration value. Under similar conditions, selecting an appropriate cache expiration value considerably decreases the overall execution time at the AMF when processing multiple location requests, as shown in Fig. 9. It is worth noting that cache expiration is adjustable and should always be configured according to the dynamism of the scenario.

Conversely, we can consider different parameters for *Location Requests Aggregation*. Here, the primary variable of interest is the end-to-end response time, which represents the time a client spends waiting for location information. According to Fig. 10, increasing the number of aggregated requests directly impacts the end-to-end response time, highlighting the trade-off. The challenge lies in the system waiting until a certain number of location requests accumulate before processing them. Thus, it is independent of the scenario's dynamism but depends on the request frequency (Lambda value in the experiment). As a result, scenarios with low request frequency require a low level of aggregation to ensure a response time of less than one second, in accordance with QoS policies. On the contrary, scenarios with high request frequencies may allow for a high level of aggregation to meet QoS requirements. This leads to a significant reduction in AMF processing time. Finally, we present a preliminary evaluation of AMF scaling; however, due to hardware and software constraints, we cannot provide precise results. Instead, we can only achieve an approximate level of truthfulness and individual rationality.

In summary, these solutions represent an initial approach to reducing the flow of AMF control plane messages. Specifically, this method could be applied to various service applications where the AMF serves as the central service component. The microservice architecture of the 5G core offers several advantages; however, challenges related to enhancing communication still persist.

VIII. CONCLUSION

In this paper, we studied the problem of optimizing location services in the 5GC, with particular emphasis on the control plane traffic handled by the AMF. To tackle this issue, we proposed three different scalability techniques applicable to both the AMF and LMF sides. Specifically, the first proposed method, such as the LMF caching system, does not alter the 3GPP protocol, and its modifications affect only the LMF module. The second proposed technique, location request aggregation, required slight modifications to the 3GPP protocol and impacts both the AMF and LMF sides. The last proposed technique concerned the physical scaling of the AMF. To evaluate these techniques, we implemented them in the OAI open-source code and demonstrated their performance through simulations. The results have shown significant gains in reducing the number of control plane messages handled by the AMF and a substantial decrease in the AMF processing

time, while ensuring that the end-to-end response time aligns with QoS policies for location management.

REFERENCES

- [1] (2024). *Ericsson Mobility Report*. [Online]. Available: <https://www.ericsson.com/4adb7e/assets/local/reports-papers/mobility-report/documents/2024/ericsson-mobility-report-november-2024.pdf>
- [2] *System Architecture for the 5G System (5GS)*, document TS 23.501, 2023. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144>
- [3] *Telecommunication Management; Charging Management; Location Services (LCS) Charging*, document TS 23.271, 3GPP, 2024. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1902>
- [4] S. Bartoletti, A. Conti, D. Dardari, and A. Giorgetti, "5G localization and context-awareness," in *5G Italy White Book: From Research To Market*, 2018, pp. 167–187.
- [5] S. Bartoletti et al., "Positioning and sensing for vehicular safety applications in 5G and beyond," *IEEE Commun. Mag.*, vol. 59, no. 11, pp. 15–21, Nov. 2021.
- [6] R. Di Taranto, S. Muppirisetty, R. Raulefs, D. Slock, T. Svensson, and H. Wymeersch, "Location-aware communications for 5G networks: How location information can improve scalability, latency, and robustness of 5G," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 102–112, Nov. 2014.
- [7] C. Fiandrino, H. Assasa, P. Casari, and J. Widmer, "Scaling millimeter-wave networks to dense deployments and dynamic environments," *Proc. IEEE*, vol. 107, no. 4, pp. 732–745, Apr. 2019.
- [8] S. Bartoletti et al., "Location-based analytics in 5G and beyond," *IEEE Commun. Mag.*, vol. 59, no. 7, pp. 38–43, Jul. 2021.
- [9] D. Giustiniano, G. Bianchi, A. Conti, S. Bartoletti, and N. B. Melazzi, "5G and beyond for contact tracing," *IEEE Commun. Mag.*, vol. 59, no. 9, pp. 36–41, Sep. 2021.
- [10] *5G System; Location Management Services*, document TS 29.572, 3GPP, 2023. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3407>
- [11] A. Banerjee, R. Mahindra, K. Sundaresan, S. Kaser, K. Van der Merwe, and S. Rangarajan, "Scaling the LTE control-plane for future mobile access," in *Proc. 11th ACM Conf. Emerg. Netw. Exp. Technol.*, Dec. 2015, pp. 1–13.
- [12] A. Mohammadkhan, K. K. Ramakrishnan, and V. A. Jain, "CleanG—Improving the architecture and protocols for future cellular networks with NFV," *IEEE/ACM Trans. Netw.*, vol. 28, no. 6, pp. 2559–2572, Dec. 2020.
- [13] *5G System (5GS) Location Services (LCS); Stage 2*, document TS 23.273, 3GPP, 2024. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3577>
- [14] *Service Requirements for the 5G System*, document TS 22.261, 3GPP, 2024. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3577>
- [15] A. Pinto, G. Santaromita, C. Fiandrino, D. Giustiniano, and F. Esposito, "Characterizing location management function performance in 5G core networks," in *Proc. IEEE Conf. Netw. Function Virtualization Softw. Defined Netw. (NFV-SDN)*, Nov. 2022, pp. 66–71.
- [16] *NG Radio Access Network (NG-RAN); Stage 2 Functional Specification of User Equipment (UE) Positioning in NG-RAN;*, document TS 38.305, 3GPP, 2023. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3310>
- [17] *NG-RAN; NG Signalling Transport;*, document TS 38.412, 3GPP, 2023. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3222>
- [18] *5G System; Access and Mobility Management Services; Stage 3*, document TS 38.412, 3GPP, 2023. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3339>
- [19] M. T. Raza, D. Kim, K.-H. Kim, S. Lu, and M. Gerla, "Rethinking LTE network functions virtualization," in *Proc. IEEE 25th Int. Conf. Netw. Protocols (ICNP)*, Oct. 2017, pp. 1–10.
- [20] M. Hoffmann and P. Kryszkiewicz, "Signaling storm detection in IIoT network based on the open RAN architecture," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2023, pp. 1–2.
- [21] V. Jain et al., "L25GC: A low latency 5G core network based on high-performance NFV platforms," in *Proc. ACM SIGCOMM 2022 Conf.*, 2022, pp. 143–157, doi: [10.1145/3544216.3544267](https://doi.org/10.1145/3544216.3544267).
- [22] Y. Li et al., "A case for stateless mobile core network functions in space," in *Proc. Conf. ACM Special Interest Group Data Commun.*, 2022, pp. 298–313.
- [23] A. Kumar, P. Naik, S. Patki, P. Chaudhary, and M. Vutukuru, "Evaluating network stacks for the virtualized mobile packet core," in *Proc. 5th Asia-Pacific Workshop Netw. (APNet)*, Jun. 2021, pp. 72–79, doi: [10.1145/3469393.3469402](https://doi.org/10.1145/3469393.3469402).
- [24] Z. Qin, D. Wu, Z. Xiao, B. Fu, and Z. Qin, "Modeling and analysis of data aggregation from convergecast in mobile sensor networks for industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4457–4467, Oct. 2018.
- [25] H. Harb, A. Makhoul, S. Tawbi, and R. Couturier, "Comparison of different data aggregation techniques in distributed sensor networks," *IEEE Access*, vol. 5, pp. 4250–4263, 2017.
- [26] I. Alawe, Y. Hadjadj-Aoul, A. Ksentini, P. Bertin, and D. Darche, "On the scalability of 5G core network: The AMF case," in *Proc. 15th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2018, pp. 1–6.
- [27] D. Scotece, A. Noor, L. Foschini, and A. Corradi, "5G-kube: Complex Telco core infrastructure deployment made low-cost," *IEEE Commun. Mag.*, vol. 61, no. 7, pp. 26–30, Jul. 2023.
- [28] *Lmf Location Service*. Accessed: Dec.2025. [Online]. Available: https://forge.3gpp.org/swagger/ui/?url=https://forge.3gpp.org/rep/all/5G_APIs/raw/REL-18/TS29_572_Nlmf_Location.yaml#/Determine%20Location/DetermineLocation
- [29] G. Kousiouris, "A self-adaptive batch request aggregation pattern for improving resource management, response time and costs in microservice and serverless environments," in *Proc. IEEE Int. Perform., Comput., Commun. Conf. (IPCCC)*, Oct. 2021, pp. 1–10.
- [30] *NG-RAN; NR Positioning Protocol A (NRPPa)*, document TS 38.455, 3GPP, 2024. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3256>
- [31] *NG-RAN; NG Application Protocol (NGAP)*, document TS 38.413, 3GPP, 2024. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3223>
- [32] J. Larrea, A. E. Ferguson, and M. K. Marina, "CoreKube: An efficient, autoscaling and resilient mobile core system," in *Proc. 29th Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2023, pp. 1–15, doi: [10.1145/3570361.3592522](https://doi.org/10.1145/3570361.3592522).
- [33] S. Vittal, S. Sarkar, and A. A. Franklin, "Revamping the resilience and high availability of 5G core for 6G ready network slices," *IEEE Trans. Netw. Service Manage.*, vol. 21, no. 2, pp. 2287–2302, Apr. 2024.
- [34] P. Sharma, T. Atalay, H. A. Gibbs, D. Stojadinovic, A. Stavrou, and H. Wang, "5G-WAVE: A core network framework with decentralized authorization for network slices," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2024, pp. 1–10.
- [35] C. H. T. Arteaga, A. Ordoñez, and O. M. C. Rendon, "Scalability and performance analysis in 5G core network slicing," *IEEE Access*, vol. 8, pp. 142086–142100, 2020.
- [36] A. Pinto, G. Santaromita, C. Fiandrino, D. Giustiniano, and F. Esposito, "Experimenting with localization management functions in 5G core networks," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2022, pp. 806–807.
- [37] (2023). *Usrp N310*. [Online]. Available: <https://www.ettus.com/all-products/usrp-n310/>
- [38] *Google Pixel 7 Pro*. Accessed: Jan. 20, 2024. [Online]. Available: https://www.gsmarena.com/google_pixel_7_pro-11908.php
- [39] E. Goshi, R. Stahl, H. Harkous, M. He, R. Pries, and W. Kellerer, "PP5GS—An efficient procedure-based and stateless architecture for next-generation core networks," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 3, pp. 3318–3333, Sep. 2023.
- [40] O.-M. Ungureanu and C. Vlădeanu, "Leveraging the cloud-native approach for the design of 5G NextGen core functions," in *Proc. 14th Int. Conf. Commun. (COMM)*, Jun. 2022, pp. 1–7.
- [41] T. O. Atalay, D. Stojadinovic, A. Famili, A. Stavrou, and H. Wang, "A first look at 5G core deployments on public cloud: Performance evaluation of control and user planes," 2023, *arXiv:2312.04833*.
- [42] G. Baldoni, J. Quevedo, C. Guimarães, A. de la Oliva, and A. Corsaro, "Data-centric service-based architecture for edge-native 6G network," *IEEE Commun. Mag.*, vol. 62, no. 4, pp. 32–38, Apr. 2024.
- [43] H. Ryden, S. M. Razavi, F. Gunnarsson, and I. Olofsson, "Cellular network positioning performance improvements by richer device reporting," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Jun. 2018, pp. 1–5.

- [44] S. Dwivedi et al., "Positioning in 5G networks," *IEEE Commun. Mag.*, vol. 59, no. 11, pp. 38–44, Nov. 2021.
- [45] S. Bartoletti et al., "Uncertainty quantification of 5G positioning as a location data analytics function," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, Jun. 2022, pp. 255–260.
- [46] B. Ghimire, R. Shreevastav, and X. Jiang, "Preconfigured assistance data for reduction in latency and power consumption," in *Proc. IEEE 97th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2023, pp. 1–6.



Claudio Fiandrino (Member, IEEE) is currently a Research Assistant Professor with the IMDEA Networks Institute. His research focuses on explainable and robust AI for next-generation mobile networks. He has received numerous awards for his research, including a Fulbright scholarship, several Spanish national grants, and several best paper awards. He is a member of ACM and Editorial Board of *IEEE TRANSACTIONS ON MOBILE COMPUTING*, *IEEE NETWORKING LETTERS*, and *Computer Networks* (Elsevier).



Domenico Scotece (Member, IEEE) received the Ph.D. degree from the University of Bologna, Italy, in April 2020. He is currently a Junior Assistant Professor with the University of Bologna. His research interests include pervasive computing, middleware for fog and edge computing, the software-defined networking, the Internet of Things, and 5G network planning and design.



Luca Foschini (Senior Member, IEEE) received the Ph.D. degree in computer science engineering from the University of Bologna, Italy. He is currently a Full Professor in distributed systems with the University of Bologna. His research interests span from integrated management of distributed systems and services to mobile crowd-sourcing/-sensing, from infrastructures for industry 4.0 to fog/edge cloud systems.



Giuseppe Santaromita (Member, IEEE) is currently a Post-Doctoral Researcher with the IMDEA Networks Institute, Pervasive Wireless Systems Group. His research activities are focused on wireless networks. In particular on the programmable PHY layer for optimizing wireless networks performance and on the low-latency and high-accuracy localization methods, mainly on 5G new radio networks, including 5G non-terrestrial networks.



Domenico Giustiniano (Senior Member, IEEE) received the Ph.D. degree in telecommunication engineering from the University of Rome Tor Vergata in 2008. He is currently a Research Professor (tenured) with IMDEA Networks, Madrid, Spain. Before joining IMDEA, he was a Senior Researcher and a Lecturer at ETH Zurich. He also worked for four years as a Post-Doctoral Researcher with Disney Research Zurich and Telefonica Research Barcelona. His current research interests cover battery-free IoT, large-scale spectrum analytics, autonomous aerial networks, and 6G localization systems.