RESEARCH-ARTICLE

Demystifying the Mobile Control Plane Characteristics for Ubiquitous Connectivity

SHIYI LIU, University of Chinese Academy of Sciences, Beijing, China

YANBIAO LI, University of Chinese Academy of Sciences, Beijing, China

XIN WANG, Stony Brook University, Stony Brook, NY, United States

XINYI ZHANG, University of Chinese Academy of Sciences, Beijing, China

ZHUORAN MA, Hunan University, Changsha, Hunan, China

HAITAO LIU, University of Chinese Academy of Sciences, Beijing, China

View all

Published: 28 October 2025

IMC '25:ACM Internet Measurement Conference
October 31, 2025
WI, Madison, USA

Conference Sponsors:
SIGMETRICS
SIGCOMM

https://doi.org/10.1145/3730567.3764463
ISBN: 9798400718601

Demystifying the Mobile Control Plane Characteristics for Ubiquitous Connectivity

Shiyi Liu
Computer Network Information Center, CAS
Beijing, China
University of Chinese Academy of Sciences
Beijing, China
syliu@cnic.cn

Yanbiao Li*
Computer Network Information Center, CAS
Beijing, China
Hangzhou Institute for Advanced Study, UCAS
Hangzhou, China
lybmath@cnic.cn

Xin Wang
Stony Brook University
New York, United States
x.wang@stonybrook.edu

Xinyi Zhang
Computer Network Information Center, CAS
Beijing, China
University of Chinese Academy of Sciences
Beijing, China
xyzhang@cnic.cn

Zhuoran Ma
Hunan University
Changsha, China
Computer Network Information Center, CAS
Beijing, China
mazhuoran@hnu.edu.cn

Haitao Liu
Computer Network Information Center, CAS
Beijing, China
University of Chinese Academy of Sciences
Beijing, China
liuhaitao@cnic.cn

Gaogang Xie*
Computer Network Information Center, CAS
Beijing, China
University of Chinese Academy of Sciences
Beijing, China
xie@cnic.cn

Abstract

The evolution of mobile networks toward ubiquitous connectivity envisioned by International Mobile Telecommunications-2030 has caused a surge in control plane traffic. A deep understanding of the control plane's internal characteristics and mechanisms is crucial for delivering optimal services. However, existing measurements often neglect the control plane or treat it as an opaque box, focusing on overall performance instead of its intrinsic characteristics.

In this paper, we introduce a 3GPP-compliant control plane evaluation framework and conduct the first in-depth analysis of the characteristics and overheads exhibited by various network functions (NFs) under large-scale connectivity conditions, based on empirical measurements. We selected three core network systems and conducted performance measurements on 500,000 User Equipment during UE registration and PDU session establishment procedures. We reveal the substantial resource demands and limited scalability of the Access and Mobility Management Function (AMF) and the Network Repository Function (NRF). Furthermore, our analysis identifies a significant need for an enhanced state management mechanism. The insights derived from our measurements underscore the immense potential for optimization within the core network. Key optimization pathways include enhancing protocol stack processing, mitigating potential leverage-based attacks, and implementing an integrated state management framework.

CCS Concepts

• **Networks** → **Network measurement**.

Keywords

Mobile Networks; Control Plane

ACM Reference Format:
Shiyi Liu, Yanbiao Li, Xin Wang, Xinyi Zhang, Zhuoran Ma, Haitao Liu, and Gaogang Xie. 2025. Demystifying the Mobile Control Plane Characteristics for Ubiquitous Connectivity. In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25), October 28–31, 2025, Madison, WI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3730567.3764463

*Corresponding authors: lybmath@cnic.cn and xie@cnic.cn.

IMC '25, Madison, WI, USA.

ACM ISBN 979-8-4007-1860-1/25/10
https://doi.org/10.1145/3730567.3764463

1 Introduction

Ubiquitous connectivity—the seamless interconnection of users across diverse geographical locations and access technologies—is a central vision of mobile networks[33, 75]. It helps bridge the digital gap by connecting unconnected populations[34], and enables the development of emerging applications, such as drone-based delivery, satellite services, and large-scale IoT deployment[1, 57]. To realize this vision, the mobile core requires a robust and scalable control plane that can simultaneously achieve low latency and high scalability.

However, achieving ubiquitous connectivity inevitably introduces additional design constraints and exacerbates the limitations of existing 5G control plane architectures. First, the joining of unconnected elements into the network increases the connection density from $10^6/km^2$ to $10^8/km^2$[18], subsequently leading to a significant rise in control plane signaling. Second, the new control plane introduces more NFs (from 6 to more than 30[59, 60]), which adds even more control plane signaling. Third, ubiquitous connectivity enables users to switch between different networks[61], further increasing the control plane signaling. Collectively, these factors elevate the challenges to an unprecedented level and may significantly degrade the performance of the control plane.

To foster ubiquitous connectivity and address the aforementioned challenges, numerous studies have explored optimization strategies for the mobile control plane. For latency reduction, L^{2}5GC [35], SoftBox[45], and Neutrino[3] optimize control plane responsiveness through mechanisms such as fast path processing and lightweight signaling. For enhanced scalability, PEPC[66], MMLite[49], and CoreKube[37] aim to enhance scalability by improving resource elasticity, load balancing, and autoscaling strategies. Despite the importance and large amount of effort in increasing the efficiency of the control plane, there lacks a deep understanding of its characteristics. This would make the designs from various sources ad hoc and incoherent.

Existing measurements have explored the performance of Access Network (AN) - another essential component of mobile networks [7, 15, 21, 42, 51, 52, 79], cross-layer assessments [53, 77], and Key Performance Indicators[20, 27, 39, 63, 78]. However, they either entirely exclude the control plane or treat it as a monolithic, opaque box, which is unable to serve as guidance for other deployments and limits their effectiveness in delivering profound insights for the optimization of the overall network services.

This paper delves into the performance analysis of the control plane through the measurement of control plane prototypes. Our principles and contributions are as follows:

- Generalizability: We designed a measurement framework that integrates 3GPP standard analyses, studies on cutting-edge academic research, and industrial configurations. We address conclusions related to 3GPP standards, which are beneficial to both academia and industry, rather than being limited to a specific implementation.
- Practicality: We conducted tests using real-world systems. Although we were unable to access commercial core networks, we performed measurements on three open-source, 3GPP-compliant 5G core network implementations, some of which have been adopted by consortium-based 5G frameworks such as Aether[2] and SD-CORE[70].
- Novelty: To the best of our knowledge, this work introduces the first systematic measurement framework and performs the first comprehensive measurements that open up the control plane opaque box. Our measurements have uncovered a range of previously undiscovered and overlooked issues, offering valuable insights for the standardization and implementation of next-generation networks.

Our key findings are as follows:

AMF exhibits limited scalability. In the control plane, AMF handles essential tasks such as mobility management and connection management[60]. As expected, AMF consumes substantial resources: 29.3% of control plane CPU usage, 39.1% of memory usage, and generates over 40% of the network traffic. However, AMF shows poor scalability as the service rate rises, with the CPU demand growing by 20.6% and memory demand increasing by 50%.

- *Implications.* Ubiquitous connectivity increases complexity in mobility management. The traditional over-provisioning strategy, aimed at scalability, is ineffective as it causes low resource use and poor adaptation to more fluctuating demands, underscoring the need to focus on system mechanisms[10]. Our work identifies inefficient signaling processing as one of the major causes of these issues.

Overlooked NRF poses risks to latency. NRF is a novel NF within the 5G system and is crucial for enabling seamless and automated NF discovery, a key requirement of ubiquitous connectivity. [59, 60]. However, this frequently overlooked NF profoundly impacts system performance. We find it accounts for 20.1% of the control plane CPU usage, over 20% of network traffic and incurs significant temporal overhead.

- *Implications.* With the expansion of ubiquitous connectivity, NRF will extend its role from local NF discovery to cross-network NF discovery, further increasing both complexity and significance. Consequently, NRF will face more significant challenges, necessitating a more refined design and more efficient NF discovery algorithms.

Limitations of Existing State Management. State management forms the backbone of advanced NF functionalities. However, our measurements reveal a hidden bottleneck: as system load increases, state management triggers a superlinear surge in memory consumption across critical NFs—a behavior that remains invisible under conventional low-load evaluations.

- *Implications.* As ubiquitous connectivity drives rapid user growth, state management must evolve from fragmented and disjointed practices toward coordinated collaboration between local memory and persistent databases. Current flat, coarse-grained, and single-medium approaches are no longer effective. A new framework is essential to unify user state lifecycles, correlate signaling events, and accommodate diverse access patterns.

The structure of this paper is organized as follows. We provide the necessary background information in §2 and elaborate on the principles of our measurement framework in §3. We then present the measurement results and analyses for control plane NFs, network traffic, and database in §4, §5, and §6, respectively. Finally, we present a discussion in §7, review related work in §8, and conclude

Table 1: Our findings

Section	Category	Description	Root cause	Figure/Table/Chapter
Control Plane NFs §4	General observations	Stable NFs and bursting NFs	3GPP standards[a]	**Fig. 6**
		Heavy NFs and light NFs	3GPP standards	**Fig. 7**
	Potential bottlenecks	The scalability issue in AMF and NRF	3GPP standards	**Fig. 8**
		Inefficient protocol processing	3GPP standards	**Fig. 9 and 11**
		Inefficient state management	Implementation[b]	**Fig. 10 and 12**
Network §5	General observations	Leading NFs - AMF and SMF	3GPP standards	**Fig. 13**
		Promising growth in NRF traffic	3GPP standards	**Fig. 13**
		Booming demands for differentiated data retrieval	3GPP standards	**Fig. 14**
	Potential bottlenecks	Leverage effect on AN-CN traffic poses security risks	3GPP standards	**Fig. 15**
Database §6	General observations	Core network data access characteristics	3GPP standards	**Fig. 16 and 17**
		New demands for database	3GPP standards	**Tables 6 and 7**
	Potential bottlenecks	Non-scalable data organization	Implementation	**§6.1**
		Non-scalable access frequency	3GPP standards	**Fig. 16**

[a]This indicates the corresponding finding is implementation-agnostic and applies to all 3GPP-compliant systems, including commercial ones.
[b]This indicates the corresponding finding is due to the specific implementation and the lack of standardization in the specifications.

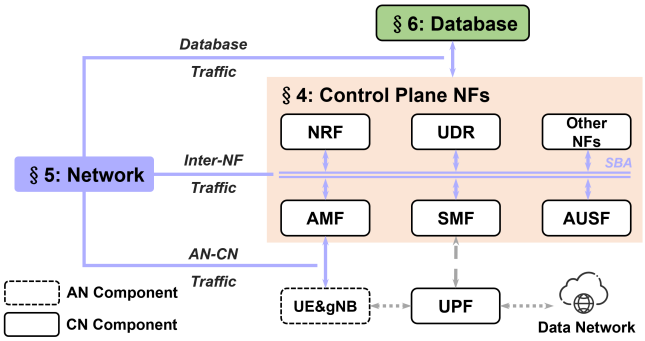


Figure 1: 5G architecture

in §9. Additionally, we summarize the findings of this paper in Table 1, and the abbreviations related to 5G are provided in Section D to facilitate reader comprehension.

2 Background

5G Architecture. The 5G standards are developed by the 3GPP group. 5G networks are composed of the AN and the core network (CN), as shown in Fig. 1. The 5G AN consists of User Equipment (UE) (*e.g.*, smartphone, laptop) and gNodeB (gNB) (commonly known as the base station). The 5G core network adopts a control-/user-plane separation design, in which the control plane processes all user mobile signaling, while the data plane transfers data between the AN and the Data Network (*e.g.*, the Internet). The control plane consists of AMF, NRF, Session Management Function (SMF), Authentication Server Function (AUSF), Unified Data Management (UDM), Unified Data Repository (UDR), Policy Control Function (PCF), and other NFs. These NFs utilize the HTTP/2 protocol for communication. The control plane also utilizes databases for data storage. However, the selection of databases is delegated to individual implementations. The data plane consists of one or more User Plane Functions (UPFs), which receive session management instructions from SMF and transfer data traffic accordingly.

UE-Data Network connection. Connections between UE and the Data Network form the foundation for users to access external networks. It starts from UE and goes through gNB and several UPFs before finally reaching the Data Network. This connection is essential for providing fundamental network services, regardless of variations in the deployment environment and access method. It requires two indispensable control plane procedures: UE Registration (UER) and Protocol Data Unit Session Establishment (PDUSE).

- The UER procedure serves as the foundation for all other procedures between UE and the control plane. It handles UE identification, authentication, and UE context configuration. Afterward, UE can initiate subsequent advanced procedures.
- The PDUSE procedure establishes a PDU session, which provides end-to-end user plane connection management and enables communication between UE and the Data Network[60].

Ubiquitous connectivity will undoubtedly enhance the importance of the UER procedure and the PDUSE procedure. When transitioning between different access networks, the UER procedure can be used to re-establish user context, while the PDUSE procedure can be triggered to complete the PDU session handover.

Ubiquitous connectivity requirements for 5G. Ubiquitous connectivity—among everything, by which every means, and everywhere [18]—presents challenges to 5G networks in two aspects: the dramatic increase in the number of UEs and the interworking of multiple access methods.

For the 5G control plane, the former emphasizes that the control plane should maximize scalability while maintaining the current latency. The latter, on the other hand, highlights the importance of

procedures and signaling responsible for interworking, such as the UER procedure and PDUSE procedure.

3 Methodology

3.1 Principles

Our measurement is based on the following principles:

- Generalizability: Our framework adheres to 3GPP standards, incorporates academic and industrial insights, and focuses on common patterns across diverse scenarios and deployments, ensuring the relevance of our findings beyond a specific implementation.
- Practicality: We base our study on measurements from real-world systems, complemented by 3GPP standard analysis, to ensure that observed phenomena and issues are genuine and practical while remaining agnostic to specific technical implementations.
- Novelty: Our goal is to identify critical issues that have been overlooked or remain undiscovered but are essential bottlenecks either now or in the future.

3.2 Aspects

As shown in Fig. 1, we partition the core network into three aspects and evaluate each one separately:

Control plane NFs aspect. In 5G, NFs require considerable CPU and memory resources to execute computational tasks and store associated states. Therefore, in §4, we primarily focus on the CPU load and memory load from a holistic perspective, transition to an NF perspective, and then conduct a detailed analysis of representative NFs.

Network aspect. The new 5G Service-Based Architecture brings innovation and flexibility. However, compared to 4G, this architecture significantly increases the number of NFs and signaling. We aim to unveil the characteristics of traffic volume and interactions between NF pairs within the 5G core network in §5. We partition the network part of the 5G core network into three distinct segments:

- *Inter-NF Traffic.* This segment is known as the Service-Based Architecture, which enables intercommunication among different NFs. It predominantly relies on the HTTP/2 protocol for communication.
- *Database Traffic.* This segment involves specific NFs communicating with external databases. The 3GPP standards delegate this choice to implementations.
- *AN-CN Traffic.* This segment is known as the N1/N2 reference point, which facilitates control plane procedures between the core network and the UE/gNB. It primarily utilizes the Non-Access Stratum (NAS) protocol carried over the New Generation Application Protocol (NGAP) to communicate.

We investigate the characteristics of traffic volume and interactions between NF pairs of the three segments.

Database aspect. The 3GPP standards leave the selection of databases to implementations. However, databases can profoundly influence performance. We would like to understand how open-source implementations utilize databases in §6. We choose to evaluate their performance using metrics such as throughput, query time, and the most frequently accessed data.

3.3 Experimental Parameters

In this part, we discuss how we configure the control plane procedure, UE number, and service rate to adapt to the demands of ubiquitous connectivity. A summary is provided in Table 5.

Control plane procedure. In the core network control plane, the most common procedures are UER, PDUSE, *handover*, and *service request*. Based on our principles, we consider UER and PDUSE to be the most representative procedures. As shown in Table 2, we present the number of NFs involved in different procedures and the corresponding number of signaling associated with each procedure.

The UER and PDUSE procedures involve interactions with most NFs and more signaling messages. They are most likely to reveal system scalability bottlenecks and best reflect the NFs and their associated capabilities (*e.g.*, NF discovery, data retrieval). In contrast, the others engage with far fewer NFs and signaling [64]. Furthermore, a significant number of studies and optimizations[22, 23, 35, 37, 38, 67, 72] target the UER and PDUSE procedures, indicating that they are of paramount importance in research on the 5G core network.

In most of our experiments, we initiate the PDUSE procedure 1 second after the UER procedure is finished for each UE, which is the smallest interval that our emulator can set. When conducting independent PDUSE procedure tests, we first wait for all UEs to complete the UER procedure. This ensures that any influence from the UER procedure is excluded.

UE number. In our experiment, the number of UEs is the most critical parameter for reflecting workload under ubiquitous connectivity. Traditional measurements often select values that are too low or too high, resulting in either the failure to reveal the true bottleneck or the exaggeration of issues that are not relevant to the current stage of development.

As shown in Table 3, we utilized 5G standards, academic research, and real-world case studies to derive an appropriate UE number. They indicate that a UE number on the order of 100,000 per core network is appropriate. Specifically, we present the details of our specific decisions:

- Standards: The 3GPP standards propose requirements for UE density in different scenarios. The requirements for UE density can reach up to $500{,}000/km^2$ (*broadband access in a crowd* case)[58].
- Academia: Research [44] in the academic community indicates that a capacity of 250,000 UEs per core network is recommended.
- Use case - smart factory: Typically, only a single core network is deployed in a smart factory. An Ericsson report[16] shows that a typical smart factory needs 0.5 connected devices per square meter. Therefore, this means that a factory with an area of 1 square kilometer would need 500,000 UEs.
- Use case - connected vehicle: Emerging applications like connected vehicle have set stringent performance demands (*e.g.*, latency < 20 ms[5]), necessitating the core network be deployed in the same city. Given that the average population of the top 150 most populous cities in the United States is 496,328[76], the core network workload in future scenarios can be a scale of this population figure.

Therefore, we set the UE number to 500,000, a suitable benchmark that is aligned with standards and provides excellent generalizability for diverse use cases under ubiquitous connectivity.

Table 2: The relationship between procedures and NFs

Procedure	AMF	SMF	AUSF	PCF	NRF	NSSF	UDM	UDR	The number of exchanged signalings[64]
UER[61]	✓		✓	✓	✓		✓	✓	67
PDUSE[61]	✓	✓		✓	✓	✓	✓	✓	53
Service Request[61]	✓	✓		✓					11
Handover[61]	✓	✓							10

Table 3: Real-world guidance on the UE number

Category	Conclusion	The order of magnitude of UE
3GPP standards[58]	UE density can reach up to 500,000/km^2	100,000
an academic study[44]	250,000 UEs are associated with each core network	100,000
use case: smart factory[16]	typical UE density is 0.5/m^2	100,000
use case: connected vehicle[76]	average population of the top 150 US cities is 496,328 (assuming one vehicle per person)	100,000

Table 4: Real-world guidance on the service rate

Category	Conclusion
An ITU report[32]	A region in Tunisia with 527,947 users experiences an average of **1** attach event per user per hour.
An academic research (Esa and Juha[17])	Attach events occur at a rate of **0.5** per user per hour.
An academic research (Archibald *et al.* [4])	It distinguishes between busy-hour and common-hour load, with a ratio of **1.5 to 2**.

Table 5: Experimental Parameters

Parameter	Value	Unit
Procedures	UER, PDUSE	-
Service Rate	150 (if unspecified)	procedures/s
UE Number	500,000	-

Furthermore, we also conducted tests with 250,000, 1 million, and 2 million UEs, all of which exhibited similar results to the 500,000 workload.

Service rate. We define *service rate* as the number of UER procedures and PDUSE procedures executed per second. Each UER procedure is followed by the execution of a PDUSE procedure, aligning with the 4G experience. We set its value based on existing standards and the workload of 500,000 UEs.

Since 3GPP standards do not specify the ratio between the UE number and the corresponding service rate, we have compiled the information from available sources to the best of our ability, as shown in Table 4. We also compile a list of literature sources that are not utilized and provide the reasons for their exclusion in Section B. The two key points from the above information are: first, we need to distinguish between busy hours and common hours to reflect the temporal characteristics of the real-world workload accurately, and second, the service rate is approximately 0.5 to 1 per hour per UE, which translates to about 75 to 150 per second for 500,000 UEs.

Therefore, we configured two workloads, busy hours and common hours, with a ratio of 2. We set their absolute quantities to 150 for busy hours and 75 for common hours based on our specified 500,000 UEs. Furthermore, we put the default service rate at 150 and set 75 as a baseline unit to observe the trend as the service rate increases.

3.4 Testbed Setup

Our experiment aims to measure the control plane prototypes adhering to 3GPP standards. However, due to the unavailability of a commercial core network, we focus our analysis on three open-source, 3GPP-compliant 5G core network systems: Free5GC [24], Open5GS[25], and OpenAirInterface[26] to ensure generalizability. Our goal is not to compare the strengths and weaknesses of open-source core network implementations but rather to present the common characteristics of the control plane prototypes.

Since we conducted our analysis from different perspectives, we have selected measurement approaches based on their respective characteristics:

- For the control plane NFs aspect, 3GPP standards provide detailed specifications for the internal functions of each NF to enforce uniformity, and studies have shown their similarities in terms of control plane NFs[9, 36, 47, 68]. Finally, we consider Free5GC as the exemplary core network to present the common characteristics, as it supports 80 operations—45 by OpenAirInterface and 40 by Open5GS—and offers the best implementation of the selected procedures[47].
- For the network aspect, the 3GPP standards define mandatory communication interfaces and protocols between NFs, with detailed

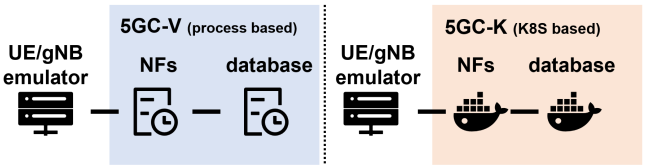


Figure 2: Testbed architecture

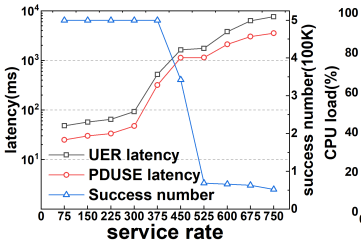


Figure 3: General performance

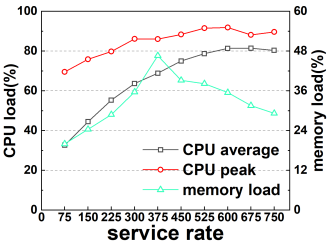


Figure 4: The variations in CPU & memory load

specifications for each packet. Therefore, the network aspect is entirely implementation-agnostic[47, 73]. For brevity and consistency with the previous sections, we use results from Free5GC as a representative example..

- For the database aspect, since the 3GPP standards leave the choice of database to implementations, each implementation may differ. Therefore, we analyze all three open-source implementations.

We provided more detail more detailed similarities and differences among the three core networks in Section E.

We deploy the 5G system separately in a virtual machine (5GC-V) and Kubernetes[6] (5GC-K), as shown in Fig. 2.

5GC-V is suitable for measuring bare core network characteristics under controlled resources because process-based NFs eliminate the overhead introduced by containerized NFs in 5GC-K and are often used by 5G service providers. Each type of NF is instantiated as a process. This virtual machine has 32 CPU cores and 197GB memory, running on Ubuntu 20.04.4 LTS and Linux 5.4.0-125-generic kernel. It is deployed on a DELL R7525 server with two AMD EPYC 7742 64-Core Processor clocked at 3.4GHz and 640GB 3200MHz DRAM, running CentOS 7.4.1708. The resources are sufficient for our chosen workload, ensuring no performance impact, as shown in Fig. 3.

5GC-K is suitable for monitoring network traffic due to the container-to-container traffic monitoring tools available in Kubernetes. Each type of NF is instantiated as a container. We use the Kubernetes client v1.27.0 and server v1.27.1, deployed on a DELL R7525 server with 2 AMD EPYC 7763 64-Core Processor and 504GB memory to ensure sufficient resources.

We deployed a high-performance commercial UE/gNB emulator on another server and connected it directly to the core network via a 10 GB Ethernet link (this bandwidth is demonstrated to be adequate, as detailed in §5). The emulator emulates a specified number of UEs and gNBs per second based on the configured service rate and UE number. Each gNB then establishes a connection with AMF in the core network using the NGAP protocol. We performed each experiment three times and used the mean value. Unless mentioned otherwise, the coefficient of variation was below 5% in all cases.

For the control plane NFs aspect, we use 5GC-V to conduct this experiment. We use Linux tools from *sysstat* package[74] and *procps* package[65] to record and analyze CPU load and memory load. For the network aspect, we use *sysstat* tools to monitor coarse-grained overall traffic patterns under 5GC-V and use in-built tools of Kubernetes to monitor finer-grained traffic patterns between NF pairs under 5GC-K. For the database aspect, we use 5GC-V and the built-in monitoring tool for the databases.

4 Control Plane NF Characteristics

We first check the general performance trend of the core network. As shown in Fig. 3, with 500,000 UEs and a gradually increasing service rate, system latency significantly rises, and success numbers sharply drop once the service rate exceeds 375. Fig. 4 reveals inadequate CPU resources, causing failures in handling upcoming UE requests and a decrease in memory load due to the drop in success numbers. Moreover, latency remains relatively high even at lower service rates, indicating the need for deeper investigation.

In this section, we begin by examining the core network from a holistic view (§4.1). Next, we delve into the NF level (§4.2). Then, we focus on representative NFs (§4.3). Finally, we provide a summary of the findings (§4.4).

4.1 Holistic View

Ascending and fluctuating CPU. As shown in Fig. 5, we present the CPU load of the core network using three metrics: the average, maximum, and minimum values within a 60-second window. The minimum value remained stable at approximately 30%. The maximum value had high fluctuations and exhibited a gradual ascent, ultimately reaching around twice the magnitude of the minimum value. The average value, following an almost linear trend, steadily increased from 34% to 54%.

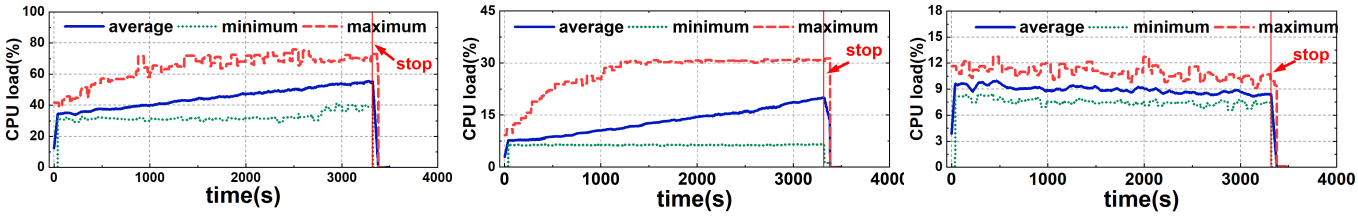
This significant gap between the maximum and minimum values, along with their increasing deviation from the average value, implies the high fluctuations of the CPU load. High fluctuations necessitate additional redundant CPU resources and introduce unpredictability, while the increasing average value suggests a growing demand for CPU resources over time under a fixed service rate. This reveals a significant scalability issue. We next explore this matter at the NF level.

4.2 NF Level View

We analyze from two perspectives: the temporal trend of resource demand during a single test and the total resource demand across different tests.

Temporal trend. We observe two types of NFs based on the temporal trends of CPU load: "stable NFs" and "bursting NFs". Fig. 6 shows two representative NFs from each type.

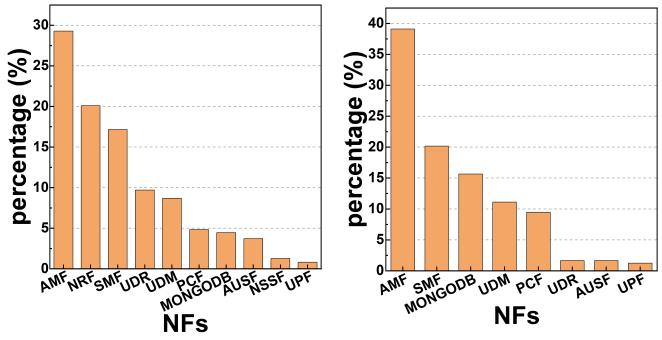
Stable NFs include NRF, Network Slice Selection Function (NSSF), and MongoDB database. Fig. 6b displays the stable CPU trend of



(a) A bursting NF (AMF) (b) A stable NF (NRF)

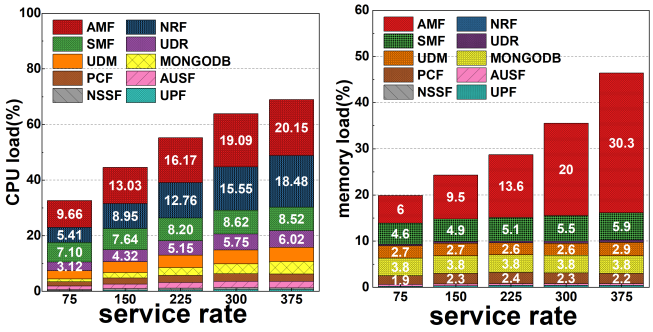
Figure 5: CPU load of the entire 5G core network

Figure 6: CPU load of two NF types



(a) CPU (b) Memory

Figure 7: The proportion of resources used by NFs



(a) CPU breakdown (b) Memory breakdown

Figure 8: The variations in resource load with service rate

NRF, with maximum, minimum, and average values around 11%, 7%, and 9%, respectively.

Bursting NFs include AMF, SMF, AUSF, PCF, UDM, and UDR. Fig. 6a shows that the maximum CPU values of AMF exhibits a near-linear increase toward our predetermined maximum, rising from 7% to 30%. Upon reaching this maximum, it remains consistently at the peak level, while the average value continues to grow linearly. It is noteworthy that the gap between the minimum and maximum values progressively widens. Consequently, the overall CPU performance is likely to become increasingly disordered and unpredictable.

Percentage and Growth rate. Fig. 7 depicts the proportions of the average CPU load and memory load of NFs over the entire duration of the experiment, respectively. AMF, NRF, and SMF together use 66% of CPU, and AMF and SMF together use 60% of memory, while the remaining NFs collectively utilize the rest of the resources. Fig. 8a shows the average CPU load under different service rates while serving the same number of UEs. AMF and NRF demand most of the CPU resources, both displaying nearly linear increases in CPU load. Their average growth rates are 20.6% and 37.2%, respectively. The remaining seven NFs demand modest CPU resources. Fig. 8b depicts the final memory load under different service rates while serving the same number of UEs. The primary contributor to memory growth is AMF, which has an average growth rate of 50%. The remaining NFs exhibit much lower growth rates.

To sum up, **AMF** is noteworthy from the perspectives of CPU and memory due to its significant proportion and scalability issues, which align with 4G experiences[8, 49]. **NRF** is particularly noteworthy from the CPU perspective due to its high proportion and scalability issues because there is no direct counterpart to NRF in the 4G system; instead, its functionalities are primarily fulfilled by DNS, a shared network infrastructure rather than an NF within the mobile core network. This represents both a technical gap, requiring a new mechanism to adapt to 5G characteristics, and a performance gap, with significant scalability issues.

4.3 Analyses on Representative NFs

Based on the previous discussion, we focus on AMF and NRF as representative NFs for analyses in this part. We profile the starting and ending 5 minutes of an experiment to analyze CPU and memory usage. We focus on two aspects: dynamic change, which examines the difference between periods, and static proportion, which looks at the proportion at a specific period.

AMF CPU analysis. Fig. 9 depicts the profiled CPU time of AMF. We further divide it into 4 logical types: *runtime*, *NGAP*, *HTTP/2*, and *HTTP*. *NGAP*, *HTTP/2*, and *HTTP* represent the corresponding protocol handling modules. *Runtime* refers to the underlying supporting environment provided by the programming language. Additionally, we further elaborate on the functionality of these modules for each type in Section C. The unit *minute/core* refers to the amount of computational work executed by a single processor core in one minute.

- Dynamic changes: It reveals a rapid increase in the *gcBgMarkWorker* module, which rose from 10.3% to 61.1%, accounting for

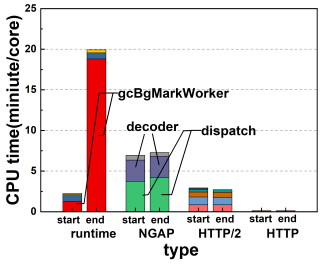


Figure 9: CPU breakdown of AMF

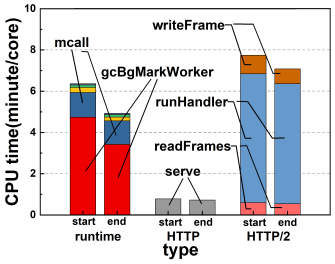


Figure 10: CPU breakdown of NRF

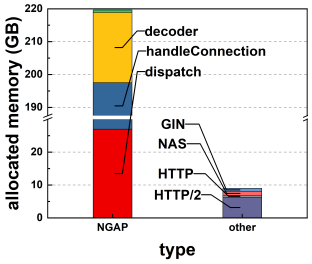


Figure 11: Allocated memory breakdown of AMF

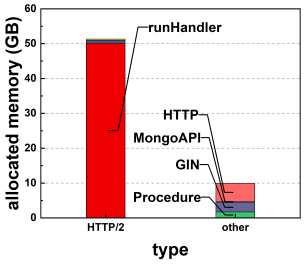


Figure 12: Allocated memory breakdown of NRF

most of the overall increase. This surge indicates that garbage collection becomes a major bottleneck as more UEs connect. Garbage collection is required for the data structures involved in UE context retrieval, incoming NGAP packet processing, and other related operations, as their numbers vary with the change in the number of UEs.

• Static proportion: Excluding the surge in *gcBgMarkWorker*, the growth of profiled CPU time in other types is relatively minor. Their proportion ratio fluctuates between 0.9 and 1.1, indicating an acceptable variation. In this context, *NGAP* type and *HTTP/2* type consume most of the CPU time. Their ratio is around 7:3. Since *NGAP* type mainly includes the *dispatch* and *decoder* modules, optimizing these two would be highly beneficial.

AMF Memory analysis. Fig. 11 depicts the profiled allocated memory of AMF during the starting 5 minutes. We exclude the ending 5 minutes as the allocated memory remains stable. We aim to emphasize the memory consumption of *NGAP* type. It consumes the most memory, around 220GB, while other components use only around 10GB. In the *NGAP* type, the *handleConnection* module, which maintains NGAP connections, uses around 170GB. The *dispatch* module processes AMF logic and occupies around 27GB, and the *decoder* module consumes around 21GB.

NRF CPU analysis. Fig. 10 depicts the profiled NRF CPU time, where we show the time of three major logical types: *runtime*, *HTTP/2*, and *HTTP*.

• Dynamic changes: The disparity in profiled CPU time between the starting and ending periods is minimal.

• Static proportion: The majority of CPU time is consumed by *HTTP/2* type and *runtime* type, accounting for 51.4% and 42.8%, respectively. The primary module within *HTTP/2* type is *runHandler*, constituting 41.5% of the total profiled CPU time. The *runtime* type, similar to that in AMF, is mainly composed of *gcBgMarkWorker* module and *mcall* module, with *gcBgMarkWorker* module consuming 31.6% of profiled CPU time.

NRF Memory analysis. Fig. 12 shows the profiled allocated memory in the starting 5 minutes. *HTTP/2* accounts for the majority of the memory, approximately 51GB, with other types accounting for a significantly lower portion, approximately 10GB.

In summary, **protocol processing** consumes substantial CPU and memory resources, such as NGAP in AMF and HTTP/2 in NRF. Specifically, protocol encoding and decoding use a significant portion of CPU and memory resources, accounting for an average of 30% of the resources in AMF and NRF. This feature is also found in other NFs. Therefore, enhancing protocol processing can significantly improve scalability. Besides, inefficient **state management** is the root cause of CPU spikes in NFs and contributes to the lack of scalability. Stable NFs, such as NRF and NSSF, manage fewer states proportional to the number of NFs, while bursting NFs manage states proportional to the number of UEs, resulting in high memory usage.

Bursting NFs and stable NFs differ fundamentally in the target of their state management. Bursting NFs primarily manage state on a per-UE basis. These NFs handle a large volume of bursty access requests from numerous users and maintain substantial per-UE information. In contrast, stable NFs serve other NFs. On one hand, NFs have significantly less state information; on the other hand, compared to the number of UEs, the number of NFs is much smaller, and their access patterns are more static and predictable.

Therefore, we recommend shifting the design of bursting NFs from a user-centric paradigm to a data-centric paradigm. Currently, all data of a user is organized as a large monolithic structure, which leads to substantial data redundancy. In contrast, a data-centric paradigm leverages similarities within the data, as well as their access frequency and importance, to achieve effective information consolidation.

4.4 Summary

The main findings are as follows:

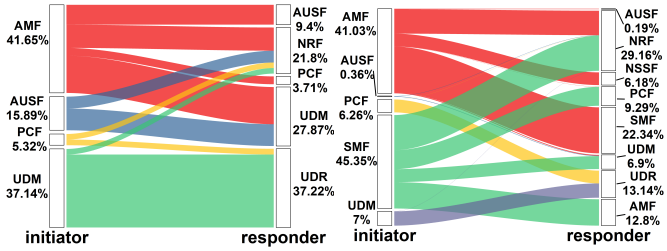
General observations.

• Stable NFs and bursting NFs: Based on the temporal trend of CPU load, we distinguish between stable NFs and bursting NFs. Bursting NFs exhibit CPU spikes, while stable NFs maintain a steady CPU load. The former can result in inefficient resource utilization.

• Heavy NFs and light NFs: We identified resource-heavy NFs (AMF, SMF, and NRF) that consume the majority of resources, while the remaining NFs share significantly fewer resources. It is essential to prioritize resource-heavy NFs.

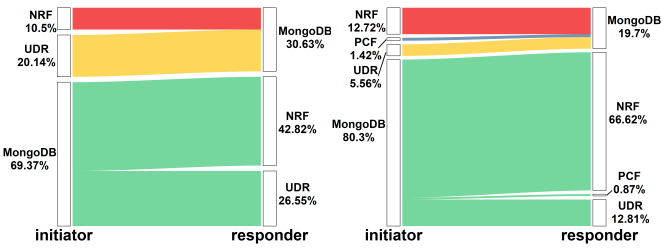
Potential bottlenecks.

• The scalability issue in AMF and NRF: The resource allocation for AMF and NRF is significant, with demand increasing rapidly. Some resource demands display superlinear growth, posing a challenge for efficient utilization. This highlights the need for optimization of



(a) The UER procedure case (b) The PDUSE procedure case

Figure 13: Inter-NF traffic breakdown



(a) The UER procedure case (b) The PDUSE procedure case

Figure 14: Database traffic breakdown

AMF and NRF in response to the growing demand for ubiquitous connectivity.

- Inefficient protocol processing: Both NGAP and HTTP/2 protocols consume significant resources. Methods such as L^{2}5GC[35] and Neutrino[3] address this issue by using shared memory and space-for-time techniques. However, these solutions have significant limitations, sacrificing scalability and compatibility, and there remains considerable potential for further exploration of this issue.
- Inefficient state management: Inefficient state management has caused CPU spikes and heavy memory load and will be exacerbated by frequent state transitions and surging UE numbers under ubiquitous connectivity. However, there is a limited exploration of state management strategies at present. A stateless core represents the opposite extreme, decoupling computation from the state entirely, trading communication overhead for more efficient state management.

5 Network Characteristics

The 5G core network utilizes the HTTP/2 protocol for NF interconnection, enabling NF deployment across multiple hosts and regions. However, recent research highlights the high latency[35] and computational burden[46] of the HTTP/2 protocol. Therefore, a deeper understanding of the HTTP protocol is essential. This hinges on understanding the traffic patterns between NFs. This section begins with traffic between NFs (§5.1), followed by a discussion of database traffic (§5.2), then traffic between the AN and the CN (§5.3), and concludes with a summary (§5.4).

5.1 Inter-NF Traffic

Fig. 13a and Fig. 13b show the traffic volumes across different NF pairs during the UER procedure and the PDUSE procedure, respectively. We have identified the following findings:

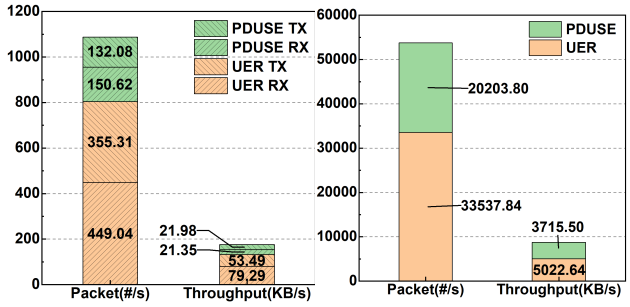
Key NFs driving the main traffic. In the UER and PDUSE procedures, AMF and SMF dominate the traffic volumes and initiate the main transactions in logic, with the traffic of AMF and SMF accounting for 41.65% and 67.69% of the total traffic in the two procedures, respectively. AMF and SMF also lead the traffic in logic. The UER procedure comprises four AMF-centric sequential transactions: (1) AMF-AUSF pair for authentication; (2) AMF-PCF pair for AM policy association; (3) AMF-UDM pair for UE context retrieval; (4) AMF-NRF pair for NF discovery. The PDUSE procedure comprises four SMF-centric sequential transactions: (1) SMF-AMF pair for message transfer; (2) SMF-PCF pair for SM policy association; (3) SMF-UDM pair for UE context retrieval; (4) SMF-NRF pair for NF discovery. The remaining transactions can be regarded as consequences of these main transactions. AMF and SMF are subject to a high percentage of traffic load, making it essential to focus on optimizing their performance.

Promising demand for NF discovery in NRF. As shown in Fig. 13, NRF accounts for 21.8% and 29.16% of the traffic in the UER procedure and the PDUSE procedure, respectively. The high percentage of NRF traffic volumes indicates a significant demand for NF discovery. With numerous NF instances (*e.g.*, cloud), the importance of NRF is undeniable. NRF prevents each NF from storing information about thousands of other NFs, reducing significant lookup and storage overhead. In future trends, the role of NRF will become increasingly critical. NRF must handle a larger number of NFs and more diverse NF queries, positioning itself as an enabler of distributed intelligent networking and automatic NF discovery and management. Moreover, since NRF requires regular communication with each NF, it also holds the potential for efficient fault detection. Therefore, a significant increase in NRF traffic in the future is inevitable.

5.2 Database Traffic

Fig. 14a and Fig. 14b show the traffic volumes between different NFs and the MongoDB database in the UER procedure and the PDUSE procedure, respectively. The MongoDB database primarily communicates with three NFs: NRF, PCF, and UDR. The NRF-MongoDB database pair accounts for the majority of traffic, at 53.31% and 79.34%, respectively, while the UDR-MongoDB database pair accounts for 46.68% and 18.37%, respectively. The high percentage of traffic volume between NRF and the MongoDB database supports the point we made earlier about NRF.

3GPP standards specify three common data retrieval approaches. The first approach is the UDM-UDR approach, which is adopted by most NFs (*e.g.*, AMF, SMF, and AUSF). In this method, UDM acts as a proxy to store subscription data. The second approach is the direct UDR approach. UDR is responsible for storing policy data, structured data for exposure, and application data. For instance, PCF communicates directly with UDR through the N36 reference point and retrieves policy data[60]. The third approach bypasses UDR and utilizes the Unstructured Data Storage Function (UDSF) or a database for data storage. This strategy is employed by NRF.



(a) External traffic (b) Internal traffic

Figure 15: The comparison of external and internal traffic

It is worth noting that the third, differentiated method accounts for a significant proportion, and we anticipate its future share to increase. Firstly, as analyzed in §4, state management will become a bottleneck limiting the scalability of NFs. Unstructured data, such as UE context, will necessitate specialized databases for management. Secondly, future networks will require more Operations, Administration, and Maintenance functionalities, which will demand substantial data storage (*e.g.*, performance log, fault log). These unstructured data are extensive in volume and generated at high velocity, thereby placing substantial performance demands on network transmission. This necessitates the development of novel architectural and protocol designs.

5.3 AN-CN Traffic

The leverage effect can lead to an attack. Fig. 15 compares the internal and external traffic in terms of packet number and throughput. The pronounced difference in packet count and throughput between internal and external traffic demonstrates a notable leverage effect. Based on logical attribution, we classify *Inter-NF Traffic* and *Database Traffic* as *Internal Traffic*, while *AN-CN Traffic* is designated as *External Traffic*. We observe that the ratio of internal to external traffic is approximately 50:1 in both packet number and throughput. Moreover, different procedures have different ratios. The UER procedure has a ratio of 42:1, while the PDUSE procedure has a higher ratio of 72:1.

This traffic pattern resembles the DNS system, where a single request can generate numerous processing packets. Consequently, the core network is vulnerable to attacks akin to DNS reflection attacks. Fortunately, the authentication mechanism can halt further processing if authentication fails, reducing the internal packet volume, but it can be bypassed[61] in special cases (*e.g.*, emergency cases), leaving room for exploitation.

5.4 Summary

The main findings are as follows:

General observations.

- Leading NFs - AMF and SMF: We find that AMF and SMF serve as the leading NFs regarding traffic volume and traffic logic, while other NFs act as ancillary NFs to them. They should be prioritized in network optimization.
- Promising growth in NRF traffic: Currently, NRF handles a high percentage of traffic. In the future, it will be a key enabler for distributed intelligent networking, automated NF discovery, management, and efficient fault detection. Its increased traffic demands more efficient mechanisms.
- Booming demands for differentiated data retrieval: Differentiated data access demands currently account for a significant portion of data traffic, with expectations for further growth. Operations and Maintenance and NF context management demand improved transmission rates and quality, necessitating new architectures and protocols.

Potential bottlenecks.

- Leverage effect on AN-CN traffic poses security risks: We find that external traffic has a leverage effect on internal traffic, potentially resulting in attacks, with the rise of external traffic under ubiquitous connectivity further exacerbating the risk. Traffic monitoring and analysis mechanisms will be crucial in addressing threats, particularly for monitoring complex procedures such as UER and PDUSE.

6 Database Characteristics

In the 5G core network, critical data are stored in databases and require retrieval during runtime. Databases play a pivotal role in overall system performance, but the 3GPP standards leave database selection to implementations. In this section, we examine the current database usage in implementations (§6.1) and the characteristics of data access in 5G systems (§6.2) to provide valuable insights for the standardization and implementation of future mobile networks. We further explore the new requirements for databases under future mobile networks and conclude with a summary (§6.3).

6.1 Data Types and Granularity

In general, 5G NFs can persistently store provisioned data and temporarily store runtime data.

Provisioned data is typically managed by UDR. All mainstream open-source implementations primarily use UDR to manage provisioned data. As shown in Table 7, UDR always employs modern databases for data management, such as MongoDB and MySQL.

Runtime data, such as NF profiles (by NRF), is managed by the corresponding NF. 5G NFs may require database support to store large runtime data, as exclusive reliance on in-memory storage could lead to inefficient data lookup and resource wastage. As shown in Table 7, open-source implementations have already recognized the necessity of using databases. For instance, Open5GS and Free5GC have integrated databases with PCF and NRF for data management.

However, as demonstrated in §4, other NFs (*e.g.*, AMF) have an even greater need for database support in managing runtime data, as in-memory storage has already led to significant scalability issues. With the development of ubiquitous connectivity, it is foreseeable that more NFs will require assistance from databases to manage runtime data effectively. However, there is currently a lack of de jure or de facto specifications.

Table 6 further elaborates on the types of provisioned and runtime data and the extent of usage in open-source implementations. We observe that different implementations store varying amounts of data: Free5GC stores both provisioned data and runtime data,

Table 6: The classification of the data stored in databases

	Functionality	Free5GC	Open5GS	OpenAirInterface
Provisioned Data	Identification (IMSI, MSISDN) [a]	yes	yes	yes
	Authentication (subscriber key, authentication method)	yes	yes	yes
	Slice and Session (SST, SD, session type, PCC rule)[b]	yes	yes	no
	The Others (subscribed AMBR, PLMN id)[c]	yes	yes	no
Runtime Data	Authentication Status	yes	no	no
	3GPP Access Information	yes	no	no
	NF profiles	yes	no	no

[a]IMSI: International Mobile Subscriber Identity; MSISDN: Mobile Subscriber ISDN Number;
[b]SST: Slice and Service Type; SD: Slice Differentiator; PCC: Policy and Charging Control;
[c]AMBR: Aggregate Maximum Bit Rate; PLMN: Public Land Mobile Network;

Table 7: NF data handling in open-source implementations

	PCF	NRF	UDR
Free5GC	UDR+MongoDB	MongoDB	MongoDB
Open5GS	UDR+MongoDB	Local Memory	MongoDB
OpenAirInterface	Local Memory	Local Memory	MySQL

followed by Open5GS, which stores provisioned data, while OpenAirInterface stores only a subset of provisioned data. We believe that with the realization of ubiquitous connectivity, the increasing reliance on databases for managing runtime data will become inevitable in the future.

Additionally, the granularity of data organization also differs among the implementations. OpenAirInterface and Open5GS organize data by user, with all data for a single user stored together in one document or a row in a table. In contrast, Free5GC distinguishes between different types of data. Simply consolidating data can lead to several limitations: Storing all user data in a single row or document may hinder scalability, as rows or documents can become excessively large, leading to performance degradation. It may also encounter concurrency and synchronization issues, where extended locks on large data structures can reduce system efficiency and increase contention.

6.2 Data Access Frequency

We further examine the access frequencies of different types of data. Fig. 16 illustrates the relationship between service rate and the time spent on different collections during a 1-second window. We observe that the time increases in proportion to the service rate. At loads of 75, 150, and 375, the cumulative time for accessing all collections is approximately 150ms, 350ms, and 900ms, respectively. When the load exceeds 375, most of the time within the 1-second window is consumed, increasing database query latency. This aligns with the performance in Fig. 3, where latency worsens beyond a load of 375.

Fig. 17 presents the time spent on different collections, procedures, and read/write instructions in detail. We observed that different collections exhibit varying read/write characteristics, which can be leveraged to guide data partitioning strategies. By understanding

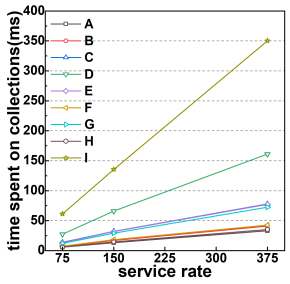


Figure 16: The variations in query time of collection with service rate

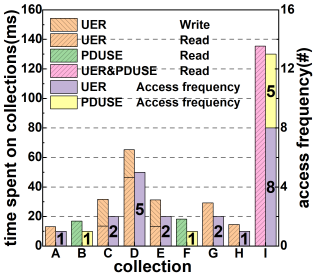


Figure 17: The correlation of query time and access frequency in different collections

A:policyData.amData; B:policyData.smData; C:authenticationStatus; D:authenticationSubscription; E:amf3gppAccess; F:smData; G:amData; H:smfSelectionSubscriptionData; I:NfProfile

these characteristics, we can optimize data distribution to minimize read/write contention and improve overall system performance.

Besides, we observed that the access frequencies of different collections are closely correlated with latency, which suggests potential scalability issues, especially when dealing with the *NfProfile* collection. As shown in Fig. 18, for each UE, *NfProfile* is accessed 8 times in the UER procedure and 5 times in the PDUSE procedure through NRF. More information about *NfProfile* is provided in Section F.

To alleviate the scalability challenges of NRF querying, caching strategies have been adopted in some commercial core networks[11]. These approaches can reduce redundant queries; however, their effectiveness is constrained by the sensitivity to query parameters. Any variation in the query parameters often leads to cache misses, thereby limiting the long-term benefits of caching.

Reducing the overall number of queries is therefore considered a promising direction. One effective method is procedure-aware batch querying, which leverages the logical dependencies among signaling procedures. By anticipating the information needed for upcoming procedures, it is possible to prefetch and consolidate

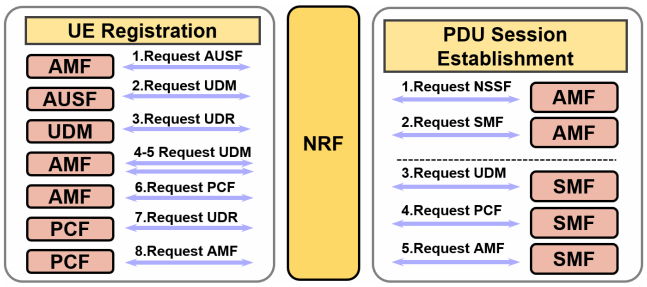


Figure 18: NF Requests to NRF for NF Profiles

multiple data items in a single query, significantly reducing the querying frequency.

Nevertheless, these optimizations face inherent limitations under ubiquitous connectivity. The exponential growth in the number of UEs, coupled with increased heterogeneity and mobility, leads to more frequent cache replacements and higher cache miss rates. Moreover, due to frequent cross-operator roaming and geographic dispersion of NF instances, a single UE may be served by NFs registered in different NRFs. This fragmentation further complicates caching and batching, as the queries may need to be resolved across multiple network domains. Hence, while caching and batch querying offer partial mitigation, standalone optimization techniques are insufficient. Targeted architectural support must be incorporated into the design of next-generation mobile networks.

6.3 Summary

The main findings are as follows:

General observations.

- Core network data access characteristics: We detail the correspondence between procedures and collections, and the read/write frequencies on collections. This information helps refine data model design and improve retrieval mechanisms.
- New demands for databases: NFs will face stronger and more refined demands for managing runtime data via databases due to increased user volume and insufficient local memory, limited flexibility of UDR, and the trend towards distributed deployment. The current utilization of databases is both inadequate and imprecise, leaving significant room for adaptation to NF and future distributed deployments.

Potential bottlenecks.

- Non-scalable data organization: We found that the granularity of data decomposition in implementations is insufficient. Simply aggregating a user's data results in overly large data structures, which potentially leads to concurrency and synchronization issues. Currently, there is a lack of a data partition framework that can ensure efficient management based on read/write frequency, lifecycle, and signaling logic.
- Non-scalable access frequency: High access frequency to the database can lead to query timeouts. Minimizing the frequency of database accesses can alleviate the problem, such as by employing caching mechanisms and the procedure-aware batch querying we suggested.

7 Discussion

Enabling ubiquitous connectivity. Ubiquitous connectivity of 6G introduces higher connection density, more diverse access methods, and tighter cross-domain collaboration. This trend is driving the evolution of core networks from centralized cloud-based deployments toward edge-oriented and distributed deployments, thereby imposing significantly higher performance demands on NF coordination and protocols.

On one hand, the importance of AMF and NRF has substantially elevated in response to massive device connectivity and the need for dynamic inter-NF coordination. A major challenge herein lies in state management, and we considered it necessary to transition from a user-centric to a data-centric paradigm. The current paradigm aggregates per-user data into a monolithic structure and incurs substantial redundancy and inefficiency. In contrast, the data-centric paradigm can enhance efficiency by utilizing data similarities, access frequency, and criticality.

On the other hand, mobile protocols have become increasingly critical for state synchronization, backup, and recovery among NFs in distributed deployments. Conventional TCP-based HTTP suffers from head-of-line blocking and high handshake latency, failing to meet demands for low latency and frequent small transfers. We propose QUIC-based HTTP/3 as a superior alternative. It overcomes TCP's transport-layer limitations, significantly improving efficiency and robustness for distributed deployments.

Generalizability of our work. Although commercial core networks are proprietary, confidential, and optimized for better capital expenditure, our findings can still contribute to their improvement. First, commercial systems strictly adhere to 3GPP standards, making their architecture consistent with our measured open-source core network. Regarding NF deployment, although mainstream commercial systems have integrated certain NFs, they still maintain separation of critical NFs such as the AMF, SMF, and NRF[14, 28, 54, 80]. Regarding communication protocols, these systems also employ standard HTTP to ensure interoperability with other commercial core networks[13, 30, 55, 81]. Second, our measurement framework aligns with the settings of commercial systems. Key parameters used in our analysis—such as the number of UEs[29, 31] and service rate[12]—are consistent with those deployed in operational commercial core networks.

8 Related Work

Measurements on the 5G core network. Since the deployment of 5G systems, numerous measurements have emerged. However, research on measuring and analyzing the 5G core network remains few and far between.

Some research focuses on the analysis of signaling and standards. Silveira *et al.* [73] presented a tutorial on the protocols and evaluated three open-source 5G core network implementations for compliance with 3GPP standards. Meng *et al.* [43] collected signaling data from real-world systems and analyzed its distribution. However, these studies did not conduct tests on real systems, making it impossible to identify bottlenecks in the 5G core network.

Some research focuses on testing actual systems. Lando *et al.* [36], Chen *et al.* [9], and Reddy *et al.* [68] conducted preliminary analyses on the performance of open-source core networks. In their tests,

the UE numbers were limited (≤ 100), and the specific service rates were not disclosed. Besides, their focus was more on comparing the strengths and weaknesses of different implementations rather than addressing common issues. Mukute *et al.* [47] measured the system call overhead in open-source core networks. Instead of focusing on common issues arising from 3GPP standards, this work emphasizes the differences in system call usage across various open-source implementations.

Compared to these studies, we combine standard analysis with actual system testing and establish a reliable measurement framework. We focus on the common issues arising from 3GPP standards, aiming to provide universally applicable conclusions and suggestions.

Additional 5G Measurements. Most research focuses on the AN or the whole 5G system. A significant amount of work [15, 21, 42, 50–52] has measured the AN from various aspects. Besides, considerable research has also measured various key performance indicators of 5G as a whole, such as reliability[39], bandwidth[41, 78], mobility[19, 20, 27, 63], and round-trip time[69]. Additionally, some studies[53, 77] measured cross-layer metrics.

These studies have all made outstanding contributions to the development of 5G. However, they treat the control plane as an opaque box, resulting in a lack of in-depth analysis of the control plane. Our work fills this gap by conducting pioneering measurements of the control plane and providing a measurement framework that lays the foundation for future measurement efforts.

9 Conclusion

This paper presents a comprehensive measurement of the 5G core network control plane in the context of ubiquitous connectivity, focusing on three aspects: NFs, network, and database. Our in-depth analyses of these aspects identify AMF and NRF as key performance bottlenecks and show that state management significantly impacts the performance of the core network. Based on these findings, we propose recommendations for future NF development and discuss the potential optimizations from these findings for both the academic and open-source communities.

10 Acknowledgments

We sincerely thank our shepherd and the anonymous reviewers for their insightful and constructive suggestions. This work was supported by the National Natural Science Foundation of China under Grant No. 62072430.

References

[1] 3GPP. 2024. NR Support for UAVs. https://www.3gpp.org/technologies/nr-uav Accessed: 2024-05-01.
[2] Aether. 2024. Cloud Native 5G Connectivity Service. https://aetherproject.org/ Accessed: 2024-05-01.
[3] Mukhtiar Ahmad, Syed Muhammad Nawazish Ali, Muhammad Taimoor Tariq, Syed Usman Jafri, Adnan Abbas, Syeda Mashal Abbas Zaidi, Muhammad Basit Iqbal Awan, Zartash Afzal Uzmi, and Zafar Ayyub Qazi. 2023. Neutrino: A Fast and Consistent Edge-Based Cellular Control Plane. *IEEE/ACM Transactions on Networking* 31, 2 (2023), 754–769.
[4] Rennie Archibald, Dhruv Gupta, Rittwik Jana, Vijay Gopalakrishnan, Ashok Sunder Rajan, Kannan Babu Ramia, Dan Dahle, Jacob Cooper, George Kennedy, Nikhil Rao, et al. 2016. An IoT control plane model and its impact analysis on a virtualized MME for connected cars. In *2016 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*. IEEE, 1–6.
[5] Tolga O. Atalay, Dragoslav Stojadinovic, Alireza Famili, Angelos Stavrou, and Haining Wang. 2023. A First Look at 5G Core Deployments on Public Cloud: Performance Evaluation of Control and User Planes. arXiv:2312.04833 [cs]
[6] The Kubernetes Authors. 2024. Kubernetes website. https://kubernetes.io/ Accessed: 2024-05-01.
[7] Stefanos Bakirtzis, André Felipe Zanella, Stefania Rubrichi, Cezary Ziemlicki, Zbigniew Smoreda, Ian Wassell, Jie Zhang, and Marco Fiore. 2023. Characterizing Mobile Service Demands at Indoor Cellular Networks. In *Proceedings of the 2023 ACM on Internet Measurement Conference*. Association for Computing Machinery, 645–659.
[8] Arijit Banerjee, Rajesh Mahindra, Karthik Sundaresan, Sneha Kasera, Kobus Van Der Merwe, and Sampath Rangarajan. 2015. Scaling the LTE Control-Plane for Future Mobile Access. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*. ACM, 1–13.
[9] Peng-Yu Chen and Chin-Ya Huang. 2024. Evaluation of UE Registration Performance in Open-Source 5G Core Networks. In *2024 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS)*. IEEE, 1–5.
[10] cisco. 2024. AMF overview. https://www.cisco.com/c/en/us/td/docs/wireless/ucc/amf/2022-01/config-and-admin/b_ucc-5g-amf-config-and-admin-guide_2022-01/m_amf-overview.html Accessed: 2024-05-01.
[11] cisco. 2024. NRF cache. https://www.cisco.com/c/en/us/td/docs/wireless/ucc/smf/b_SMF/b_SMF_chapter_011100.html Accessed: 2024-05-01.
[12] CISCO. 2024. Ultra Cloud Core 5G Access and Mobility Management Function Configuration and Administration Guide. https://www.cisco.com/c/en/us/td/docs/wireless/ucc/amf/2025-02/config-and-admin/b_ucc-5g-amf-config-and-admin-guide_2025-02/m_sample_configuration.html Accessed: 2024-05-01.
[13] CISCO. 2024. Ultra Cloud Core 5G Policy Control Function, Release 2024.01 - Configuration and Administration Guide. https://www.cisco.com/c/en/us/td/docs/wireless/ucc/pcf/2024-01/configuration-admin/b_ucc-5g-config-and-admin-guide_2024-01/m_chapter-http-tls.html Accessed: 2024-05-01.
[14] CISCO. 2024. Wireless Products. https://www.cisco.com/c/en/us/support/wireless/index.html Accessed: 2024-05-01.
[15] Phuc Dinh, Moinak Ghoshal, Dimitrios Koutsonikolas, and Joerg Widmer. 2022. Demystifying Resource Allocation Policies in Operational 5G mmWave Networks. In *2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. 1–10.
[16] Ericsson. 2024. Realizing smart manufacturing through IoT. https://www.ericsson.com/en/reports-and-papers/mobility-report/articles/realizing-smart-manufact-iot Accessed: 2024-05-01.
[17] Juha TT Salmelin Esa Markus Metsälä. 2015. *LTE Backhaul*. 3–43. doi: 10.1002/9781118924655.ch2
[18] FG-NET-2030. 2024. A Blueprint of Technology, Applications and Market Drivers Towards the Year 2030 and Beyond . https://www.itu.int/en/ITU-T/focusgroups/net2030/Documents/White_Paper.pdf Accessed: 2024-05-01.
[19] Claudio Fiandrino, David Juárez Martínez-Villanueva, and Joerg Widmer. 2023. A Study on 5G Performance and Fast Conditional Handover for Public Transit Systems. *Computer Communications* 209 (2023), 499–512.
[20] Moinak Ghoshal, Imran Khan, Z. Jonny Kong, Phuc Dinh, Jiayi Meng, Y. Charlie Hu, and Dimitrios Koutsonikolas. 2023. Performance of Cellular Networks on the Wheels. In *Proceedings of the 2023 ACM on Internet Measurement Conference*. Association for Computing Machinery, 678–695.
[21] Moinak Ghoshal, Z. Jonny Kong, Qiang Xu, Zixiao Lu, Shivang Aggarwal, Imran Khan, Yuanjie Li, Y. Charlie Hu, and Dimitrios Koutsonikolas. 2022. An In-Depth Study of Uplink Performance of 5G mmWave Networks. In *Proceedings of the ACM SIGCOMM Workshop on 5G and Beyond Network Measurements, Modeling, and Use Cases*. Association for Computing Machinery, 29–35.
[22] Endri Goshi, Vignesh Karunakaran, Hasanin Harkous, Rastin Pries, and Wolfgang Kellerer. 2023. Procedure-Aware Stateless Systems for 5G & Beyond Core Networks. In *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*. 5403–5408.
[23] Endri Goshi, Raffael Stahl, Hasanin Harkous, Mu He, Rastin Pries, and Wolfgang Kellerer. 2023. PP5GS—An Efficient Procedure-Based and Stateless Architecture for Next-Generation Core Networks. *IEEE Transactions on Network and Service Management* 20, 3 (2023), 3318–3333.
[24] Free5gc group. 2024. free5gc implementation. https://github.com/free5gc/free5gc Accessed: 2024-05-01.
[25] Open5GS group. 2024. open5GS implementation. https://github.com/open5gs/open5gs Accessed: 2024-05-01.
[26] OpenAirInterface group. 2024. openAirInterface core network implementation. https://gitlab.eurecom.fr/oai/cn5g Accessed: 2024-05-01.
[27] Ahmad Hassan, Arvind Narayanan, Anlan Zhang, Wei Ye, Ruiyang Zhu, Shuowei Jin, Jason Carpenter, Z. Morley Mao, Feng Qian, and Zhi-Li Zhang. 2022. Vivisecting Mobility Management in 5G Cellular Networks. In *Proceedings of the ACM SIGCOMM 2022 Conference*. ACM, 86–100.
[28] HUAWEI. 2024. 5G Network Architecture. https://carrier.huawei.com/~/media/CNBG/Downloads/Program/5g_nework_architecture_whitepaper_en.pdf Accessed: 2024-05-01.

[29] HUAWEI. 2024. HUAWEI UEN. https://e.huawei.com/en/products/wireless/cloud-core/uen Accessed: 2024-05-01.

[30] HUAWEI. 2024. STRATEGY ANALYTICS. https://carrier.huawei.com/~/media/cnbgv2/download/products/core/strategy-analytics-5g-signaling-en.pdf Accessed: 2024-05-01.

[31] INTEL. 2024. ZTE's High Performance 5G Core Network UPF Implementation. https://builders.intel.com/docs/networkbuilders/zte-s-high-performance-5g-core-network-upf-implementation-based-on-3rd-generation-intel-xeon-scalable-processors-1618205183.pdf Accessed: 2024-05-01.

[32] ITU. 2024. core network dimensioning. https://www.itu.int/en/ITU-D/Regional-Presence/AsiaPacific/SiteAssets/Pages/Events/2016/Aug-WBB-Iran/Wirelessbroadband/core%20network%20dimensioning.pdf Accessed: 2024-05-01.

[33] ITU. 2024. Framework and overall objectives of the future development of IMT for 2030 and beyond . https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2160-0-202311-I!!PDF-E.pdf Accessed: 2024-05-01.

[34] ITU. 2024. ITU advances the development of IMT-2030 for 6G mobile technologies. https://www.itu.int/en/mediacentre/Pages/PR-2023-12-01-IMT-2030-for-6G-mobile-technologies.aspx Accessed: 2024-05-01.

[35] Vivek Jain, Hao-Tse Chu, Shixiong Qi, Chia-An Lee, Hung-Cheng Chang, Cheng-Ying Hsieh, K. K. Ramakrishnan, and Jyh-Cheng Chen. 2022. L ^{2} 5GC: A Low Latency 5G Core Network Based on High-Performance NFV Platforms. In *Proceedings of the ACM SIGCOMM 2022 Conference*. ACM, 143–157.

[36] Gabriel Lando, Lucas Augusto Fonseca Schierholt, Mateus Paludo Milesi, and Juliano Araujo Wickboldt. 2023. Evaluating the Performance of Open Source Software Implementations of the 5G Network Core. In *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 1–7.

[37] Jon Larrea, Andrew E. Ferguson, and Mahesh K. Marina. 2023. CoreKube: An Efficient, Autoscaling and Resilient Mobile Core System. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, 1–15.

[38] Yuanjie Li, Hewu Li, Wei Liu, Lixin Liu, Yimei Chen, Jianping Wu, Qian Wu, Jun Liu, and Zeqi Lai. 2022. A Case for Stateless Mobile Core Network Functions in Space. In *Proceedings of the ACM SIGCOMM 2022 Conference*. Association for Computing Machinery, 298–313.

[39] Yang Li, Hao Lin, Zhenhua Li, Yunhao Liu, Feng Qian, Liangyi Gong, Xianlong Xin, and Tianyin Xu. 2021. A Nationwide Study on Cellular Reliability: Measurement, Analysis, and Enhancements. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*. ACM, 597–609.

[40] Yuanjie Li, Zengwen Yuan, and Chunyi Peng. 2017. A Control-Plane Perspective on Reducing Data Access Latency in LTE Networks. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 56–69.

[41] Tong Liu, Jiangyu Pan, and Ye Tian. 2020. Detect the Bottleneck of Commercial 5G in China. In *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*. 941–945.

[42] Yanbing Liu and Chunyi Peng. 2023. A Close Look at 5G in the Wild: Unrealized Potentials and Implications. In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*. 1–10.

[43] Jiayi Meng, Jingqi Huang, Y. Charlie Hu, Yaron Koral, Xiaojun Lin, Muhammad Shahbaz, and Abhigyan Sharma. 2023. Modeling and Generating Control-Plane Traffic for Cellular Networks. In *Proceedings of the 2023 ACM on Internet Measurement Conference*. Association for Computing Machinery, 660–677.

[44] Ali Mohammadkhan, KK Ramakrishnan, Ashok Sunder Rajan, and Christian Maciocco. 2016. Considerations for re-designing the cellular infrastructure exploiting software-based networks. In *2016 IEEE 24th International Conference on Network Protocols (ICNP)*. IEEE, 1–6.

[45] Mehrdad Moradi, Yikai Lin, Z. Morley Mao, Subhabrata Sen, and Oliver Spatscheck. 2018. SoftBox: A Customizable, Low-Latency, and Scalable 5G Core Network Architecture. *IEEE Journal on Selected Areas in Communications* 36, 3 (2018), 438–456.

[46] Mehrdad Moradi, Karthikeyan Sundaresan, Eugene Chai, Sampath Rangarajan, and Z. Morley Mao. 2018. SkyCore: Moving Core to the Edge for Untethered and Reliable UAV-based LTE Networks. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 35–49.

[47] Tariro Mukute, Lusani Mamushiane, Albert A. Lysko, Elena-Ramona Modroiu, Thomas Magedanz, and Joyce Mwangama. 2024. Control Plane Performance Benchmarking and Feature Analysis of Popular Open-Source 5G Core Networks: OpenAirInterface, Open5GS, and free5GC. *IEEE Access* 12 (2024), 113336–113360.

[48] Tariro Mukute, Lusani Mamushiane, Albert A Lysko, Elena-Ramona Modroiu, Thomas Magedanz, and Joyce Mwangama. 2024. Control plane performance benchmarking and feature analysis of popular open-source 5g core networks: Openairinterface, open5gs, and free5gc. *IEEE Access* 12, 113336–113360.

[49] Vasudevan Nagendra, Arani Bhattacharya, Anshul Gandhi, and Samir R. Das. 2019. MMLite: A Scalable and Resource Efficient Control Plane for Next Generation Cellular Packet Core. In *Proceedings of the 2019 ACM Symposium on SDN Research*. Association for Computing Machinery, 69–83.

[50] Arvind Narayanan, Eman Ramadan, Jason Carpenter, Qingxu Liu, Yu Liu, Feng Qian, and Zhi-Li Zhang. 2020. A First Look at Commercial 5G Performance on Smartphones. In *Proceedings of The Web Conference 2020*. Association for Computing Machinery, 894–905.

[51] Arvind Narayanan, Eman Ramadan, Rishabh Mehta, Xinyue Hu, Qingxu Liu, Rostand A. K. Fezeu, Udhaya Kumar Dayalan, Saurabh Verma, Peiqi Ji, Tao Li, Feng Qian, and Zhi-Li Zhang. 2020. Lumos5G: Mapping and Predicting Commercial mmWave 5G Throughput. In *Proceedings of the ACM Internet Measurement Conference*. ACM, 176–193.

[52] Arvind Narayanan, Muhammad Iqbal Rochman, Ahmad Hassan, Bariq S. Firmansyah, Vanlin Sathya, Monisha Ghosh, Feng Qian, and Zhi-Li Zhang. 2022. A Comparative Measurement Study of Commercial 5G mmWave Deployments. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*. 800–809.

[53] Arvind Narayanan, Xumiao Zhang, Ruiyang Zhu, Ahmad Hassan, Shuowei Jin, Xiao Zhu, Xiaoxuan Zhang, Denis Rybkin, Zhengxuan Yang, Zhuoqing Morley Mao, Feng Qian, and Zhi-Li Zhang. 2021. A Variegated Look at 5G in the Wild: Performance, Power, and QoE Implications. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*. ACM, 610–625.

[54] NOKIA. 2024. Cloud-native packet core network functions. https://www.nokia.com/asset/f/207513/ Accessed: 2024-05-01.

[55] NOKIA. 2024. Cloud signaling director. https://www.nokia.com/core-networks/cloud-signaling-director/ Accessed: 2024-05-01.

[56] OpenAirInterface. 2024. OpenAirInterface 5G Core Network Basic Deployment using Docker-Compose. https://gitlab.eurecom.fr/oai/cn5g/oai-cn5g-fed/-/blob/master/docs/DEPLOY_SA5G_BASIC_DEPLOYMENT.md Accessed: 2024-05-01.

[57] 3GPP organization. 2024. 5g system overview. https://www.3gpp.org/technologies/5g-system-overview. Accessed: 2024-05-01.

[58] 3GPP organization. 2024. TS 22.261. https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3107 Accessed: 2024-05-01.

[59] 3GPP organization. 2024. TS 23.401. https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=849 Accessed: 2024-05-01.

[60] 3GPP organization. 2024. TS 23.501. https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144 Accessed: 2024-05-01.

[61] 3GPP organization. 2024. TS 23.502. https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3145 Accessed: 2024-05-01.

[62] 3GPP organization. 2024. TS 24.501. https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3370 Accessed: 2024-05-01.

[63] Yueyang Pan, Ruihan Li, and Chenren Xu. 2022. The First 5G-LTE Comparative Study in Extreme Mobility. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 6, 1 (2022), 20:1–20:22.

[64] Taeho Park, Hochan Lee, Heewon Kim, Subin Han, Taeyun Kim, and Sangheon Pack. 2023. Divide and Cache: A Novel Control Plane Framework for Private 5G Networks. In *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*. 417–422.

[65] procps. 2024. Linux From Scratch. https://www.linux.co.cr/ldp/lfs/appendixa/procps.html Accessed: 2024-05-01.

[66] Zafar Ayyub Qazi, Melvin Walls, Aurojit Panda, Vyas Sekar, Sylvia Ratnasamy, and Scott Shenker. 2017. PEPC: A High Performance Packet Core for Next Generation Cellular Networks. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*. ACM, 348–361.

[67] Shixiong Qi, K. K. Ramakrishnan, and Jyh-Cheng Chen. 2024. L26GC: Evolving the Low Latency Core for Future Cellular Networks. *IEEE Internet Computing* (2024), 1–7.

[68] Rekha Reddy, Michael Gundall, Christoph Lipps, and Hans Dieter Schotten. 2023. Open Source 5G Core Network Implementations: A Qualitative and Quantitative Analysis. In *2023 BlackSeaCom*. 253–258.

[69] Justus Rischke, Christian Vielhaus, Peter Sossalla, Sebastian Itting, Giang T. Nguyen, and Frank H. P. Fitzek. 2022. Empirical Study of 5G Downlink & Uplink Scheduling and Its Effects on Latency. In *WoWMoM 22*. IEEE, 11–19.

[70] SD-CORE. 2024. SD-CORE website. https://opennetworking.org/sd-core/ Accessed: 2024-05-01.

[71] Rinku Shah, Vikas Kumar, Mythili Vutukuru, and Purushottam Kulkarni. 2020. TurboEPC: Leveraging Dataplane Programmability to Accelerate the Mobile Packet Core. In *Proceedings of the Symposium on SDN Research*. ACM, 83–95.

[72] Bhavishya Sharma, Shwetha Vittal, and A Antony Franklin. 2023. FlexCore: Leveraging XDP-SCTP for Scalable and Resilient Network Slice Service in Future 5G Core. In *Proceedings of the 7th Asia-Pacific Workshop on Networking*. ACM, 61–66.

[73] Lucas BD Silveira, Henrique C de Resende, Cristiano B Both, Johann M Marquez-Barja, Bruno Silvestre, and Kleber V Cardoso. 2022. Tutorial on communication between access networks and the 5G core. *Computer Networks* 216 (2022), 109301.

[74] sysstat. 2024. Performance monitoring tools for Linux. https://github.com/sysstat/sysstat Accessed: 2024-05-01.

[75] WP 5D Management Team. 2024. The ITU-R Framework for IMT-2030 . https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2030/Document

s/IMT-2030%20Framework_WP%205D%20Management%20Team.pdf Accessed: 2024-05-01.
[76] Wikipedia. 2024. List of United States cities by population. https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population Accessed: 2024-05-01.
[77] Dongzhu Xu, Anfu Zhou, Xinyu Zhang, Guixian Wang, Xi Liu, Congkai An, Yiming Shi, Liang Liu, and Huadong Ma. 2020. Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*. ACM, 479–494.
[78] Xinlei Yang, Hao Lin, Zhenhua Li, Feng Qian, Xingyao Li, Zhiming He, Xudong Wu, Xianlong Wang, Yunhao Liu, Zhi Liao, Daqiang Hu, and Tianyin Xu. 2022. Mobile Access Bandwidth in Practice: Measurement, Analysis, and Implications. In *Proceedings of the ACM SIGCOMM 2022 Conference*. ACM, 114–128.
[79] André Felipe Zanella, Antonio Bazco-Nogueras, Cezary Ziemlicki, and Marco Fiore. 2023. Characterizing and Modeling Session-Level Mobile Traffic Demands from Large-Scale Measurements. In *Proceedings of the 2023 ACM on Internet Measurement Conference*. Association for Computing Machinery, 696–709.
[80] ZTE. 2024. ZTE Cloud Core. https://www.zte.com.cn/china/solutions_latest/cloud_core.html Accessed: 2024-05-01.
[81] ZTE. 2024. ZTE Launches the 2/3/4/5G Fully-Integrated Common Core Solution. https://www.zte.com.cn/global/topics/mwc2018/newsdetail.aspx-id=122688892.html Accessed: 2024-05-01.

Appendix

A Ethics

This work does not raise any ethical issues.

B Literature Sources Unused in Service Rate Formulation

- In reference to[40], it is observed that the session establishment occurs every 106.9s for each device on average. The term "session establishment" refers to the service request procedure[62] in their work and does not apply to our experiment.
- In reference to[43], they do not provide a specific service rate for attach or registration procedures.
- In reference to[71], although it claims a service rate of 9,000 per second for 1 million subscribers, the cited source does not substantiate this service rate.
- The memory growth of the database is minimal, as it pre-allocates a sufficient amount of memory during the startup phase.

C Functionalities of Main Modules for Each Type in Representative NF Breakdown

- In *runtime* type, the *gcBgMarkWorker* module is a background marking worker function in the Go language's garbage collector. Its primary function is to mark active objects during the garbage collection process, enabling the safe reclamation of unmarked objects afterward; the *mcall* module is a critical function in the Go runtime, primarily responsible for switching from Go code to runtime system code to perform various low-level system operations. This function is invoked in various scenarios, such as scheduling, garbage collection, and system call handling.
- In *NGAP* type, the *handleConnection* module is responsible for maintaining the SCTP connection used for transmitting NGAP messages, as NGAP messages are encapsulated within SCTP; the *decode* module is used for decoding NGAP messages and extracting the control plane signaling contained within them; *dispatch* module is used to distribute the decoded control plane signaling to specific processing functions.
- In *HTTP/2* type, the *serve* module is the main module for handling HTTP packets.
- In *HTTP* type, the *readFrames* module and the *writeFrame* module are responsible for performing read and write operations on HTTP messages; the *runHandler* module is specifically responsible for handling the logic within HTTP/2.

D Abbreviations

Please refer to Table 8.

E Comparison of Different Open-source Core Networks

We compare the similarities and differences among open-source core networks from three perspectives—NF, network, and database—as summarized in Table 9.

From the NF perspective, their architecture remains uniform across core networks, although their realization strategies are divergent. Regarding architecture, all core networks strictly adhere to the standard Service-Based Architecture, and all communication between NFs are conducted using the HTTP protocol. Regarding realization strategies, there are several points of difference. First, differences in the underlying libraries and programming languages lead to variations in performance. Table 10 presents the proportion of CPU usage and memory increase of each NF in three core networks under normal operational conditions [1]. Second, the user scaling methods are implemented differently. Free5GC and OpenAirInterface do not require a pre-defined total user number and instead allocate resources dynamically as users access the system. Open5GS, however, depends on a pre-set limit and pre-initialized resources, which does not fit well with the highly dynamic nature of user numbers under ubiquitous connectivity. Third, the completeness of implementing individual NFs varies significantly. For instance, NRF exhibits notable differences in implementation depth and feature support. Specifically, Open5GS and OpenAirInterface implement NRF using local memory, as shown in Table 7, which eliminates the overhead of accessing a remote database. More importantly, both Open5GS and OpenAirInterface employ an oversimplified NF profile search mechanism, which only supports searching for NFs using a limited set of static fields. Although these mechanism greatly reduces the computational load on NRF, it is inadequate for supporting ubiquitous connectivity, as such environments require more sophisticated, dynamic, and flexible retrieval capabilities. Therefore, we selected Free5GC in order to demonstrate the performance of NRF under ubiquitous connectivity conditions.

[1]The service rate is 75 for Free5GC and Open5GS, whereas it is 25 for OpenAirInterface. When the service rate of OpenAirInterface exceeds 25, the success rates of UER and PDUSE decrease significantly. To ensure full functionality, we choose a service rate of 25 for OpenAirInterface.

Table 8: Abbreviations used in this paper

Abbreviation	Full form
3GPP	3rd Generation Partnership Project
AMF	Access and Mobility Management Function
AN	Access Network
AUSF	Authentication Server Function
CN	Core Network
gNB	gNodeB
NAS	Non Access Stratum
NF	Network Function
NGAP	New Generation Application Protocol
NRF	Network Repository Function
NSSF	Network Slice Selection Function
PCF	Policy Control Function
PDUSE	Protocol Data Unit Session Establishment
SMF	Session Management Function
UDM	Unified Data Management
UDSF	Unstructured Data Storage Function
UDR	Unified Data Repository
UE	User Equipment
UER	UE Registration
UPF	User Plane Function

Table 9: Comparison of capabilities on three open-source core networks

Capabilities	Free5GC	Open5GS	OAI
Architecture & Standards Compliance			
Service-Based Architecture	Supported	Supported	Supported
HTTP/2 Compliance	Supported	Supported	Supported
NGAP Protocol Compliance	Supported	Supported	Supported
Selected 5G Procedures	Supported	Supported	Supported
Deployment			
Cloud-Native Deployment Support	Supported	Supported	Supported
UE Scaling Method	Dynamic	Static	Dynamic
Functional Completeness & Performance			
NF Completeness	High	Low	Low
Supported Service Rate	High	High	Low
Signaling Protocol Overhead	High	High	High
Database Utilization Efficiency	High	Medium	Low
NF Discovery Completeness	High	Low	Low

Table 10: Comparison of control plane NFs on three open-source core networks

	Resources	AMF	SMF	NRF	UDM	UDR	PCF	AUSF	Database
Free5GC	CPU	29.59%	21.73%	16.56%	9.05%	9.56%	4.52%	4.11%	3.09%
	Memory	37.27%	28.57%	<1%	16.77%	1.86%	11.80%	2.48%	<1%
Open5GS	CPU	54.77%	6.77%	<1%	8.92%	6.37%	4.88%	2.67%	9.67%
	Memory	20.53%	21.05%	<1%	19.65%	5.96%	12.98%	5.96%	<1%
OAI	CPU	56.04%	4.21%	1.44%	6.26%	3.05%	\[a]	3.36%	25.23%
	Memory	44.48%	37.55%	<1%	7.54%	<1%	\	1.59%	<1%

[a]OAI basic deployment [56] does not contain a PCF.

Beyond the noted differences, the systems share universal challenges. First, the overall ranking of NFs is consistent, with the AMF confirmed as the performance bottleneck. As shown in Table 10, although the CPU and memory usage proportions of the individual NFs differ slightly, the overall ranking of resource consumption is similar. Critical NFs consume a significant amount of resources, with AMF, in particular, occupying the majority of them. Second, state management presents a significant challenge. Although the three core networks differ in their scaling methods (dynamic access vs. pre-configuration), state management consistently proves to be a time-consuming process that consumes substantial computational and storage resources. Third, protocol processing introduces a significant overhead. Our flame graph analysis of all three core networks reveals that the HTTP and NGAP protocols account for a high proportion of the processing resources.

From the network perspective, the strict definition of the standards results in a similar overall traffic profile across the core networks. For NGAP signaling, all exhibit a high degree of compliance with the 3GPP specifications. For HTTP signaling, the overall traffic volume is similar across the core networks; however, the specific signaling messages are not identical. This discrepancy is attributable to varying degrees of implementation fidelity to the standards among the different core networks. A more complete implementation inherently generates a larger volume of HTTP signaling. Further details on the implementation completeness can be found in [48].

From the database perspective, the differences far outweigh the similarities. This is a consequence of the lack of standard definitions in this domain. Nonetheless, based on the requirements of ubiquitous connectivity, we can infer potential directions for future standardization and identify mechanisms for more efficient

implementation. Potential standardization directions could include specifying cache strategies and cache invalidation mechanisms, defining data lifecycle policies, and establishing requirements for data synchronization timeliness. Recommended implementation mechanisms may involve more optimized data models and improved data compression strategies to enhance performance.

F Details of NFProfile

In 5G Networks, NRF maintains crucial metadata for each registered NF instance. This metadata is essential for enabling service discovery, lifecycle management, and ensuring seamless communication within the network. The following list delineates the fundamental attributes associated with an NF instance record stored within the NRF:

- **NF Type&ID**: Denotes the functional type (e.g., AMF, SMF, UDM) or a unique identifier assigned to the specific NF instance. This is the primary key for distinguishing different NFs.
- **Services Provided**: Enumerates the standardized service-based interfaces (e.g., namf-communication, nsmf-pdusess-ion) that this particular NF instance is capable and authorized to provide. Consumers query the NRF to discover instances offering a specific service.
- **sNSSAI**: Represents a unique identifier for a Network Slice. This attribute is critical for associating the NF instance with a specific network slice, ensuring that service discovery and requests are confined within the intended slice boundaries.
- **IPv4 Address**: Specifies the IPv4 endpoint address of this NF instance. This address is utilized by other NFs for establishing HTTP/2 communication to consume the services provided by this instance.
- **NF Status**: Indicates the current registration status (e.g., *REGISTERED*) of the NF in the NRF. This status reflects the operational availability and reachability of the NF instance.
- **PLMN**: Identifies the mobile network operator to which this NF instance belongs. This is a key parameter for facilitating inter-PLMN operations and ensuring network isolation and security.