Evaluating CAP Theorem on a Stateless 5G Core

Siddhesh Pratim Sovitkar, Shwetha Vittal, Antony Franklin A
Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad
Email: {cs21mtech12005, cs19resch01001, antony.franklin}@iith.ac.in

2024 IEEE 21st Consumer Communications & Networking Conference (CCNC) | 979-8-3503-0457-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CCNC51664.2024.10454811

***Abstract*—With multiple Network Functions (NFs) participating, the current Service Based Architecture (SBA) of 5G Core (5GC) is crafted stateless, scalable, robust, and reliable to cater to various service requirements. While Beyond 5G (B5G) and 6G networks aim to achieve higher resilience, studying the Consistency (C), Availability (A), and Partition (P) tolerance (CAP) theorem becomes a critical need for SBA-based 5GC to evaluate its impact. In this paper, we evaluate the CAP theorem on the 3GPP-compliant non-roaming stateless 5GC and show the trade-off between Consistency and Partition tolerance (CP) mode and Availability and Partition tolerance (AP) mode of functioning. Experiments prove that the 5GC in CP mode serves more PDU session requests while remaining less available as compared to the 5GC in AP mode.**

I. Introduction

Built in, two important parts (Radio Access Network and Core Network), the current 5G system is serving heterogeneous users for a variety of services. Specifically, on the 5G Core (5GC) part, this has been made possible with key principles of Service Based Architecture (SBA) and control plane and user plane separation framework. Distributed by nature, the SBA of stateless 5GC involves multiple Network Functions (NFs) to provide the ultimate services to the 5G users (User Equipments (UEs)). Access and Mobility Management Function (AMF), Unified Data Management (UDM), Authentication Server Function (AuSF), Network Repository Function (NRF), and Session Management Function (SMF), to name the primary ones on the control plane. These NFs serve different requests arriving from 5G users by communicating with each other on a service-based interface.

However, the overwhelming arrival of a variety of users connecting to the 5GC has urged the Mobile Network Operator (MNO) to dive deep into studying the key factors that challenge the robustness, scalability, consistency, resilience, and high availability of the service. In this regard, the Consistency, Availability, and Partition tolerance (CAP) theorem [1] sets forth a challenging principle. It states that "A distributed system can deliver only two of three desired characteristics: Consistency (C), Availability (A), and Partition tolerance (P) at any given time." Fig. 1 depicts this flow. As shown here, in the case of a network partition, the distributed system must choose between consistency and availability. However, the ultimate service satisfaction depends on the High Availability (HA) with consistency and resilience to the best possible level. Therefore, in this work, we study the CAP theorem on the stateless 5GC to identify the possible impacts on the resilience and HA of service of the 5GC.

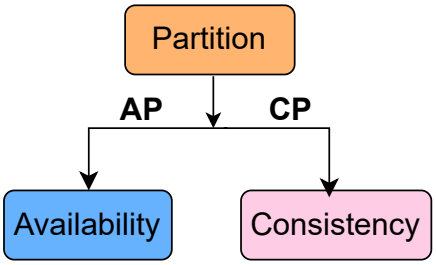


Fig. 1: View of CAP Theorem.

II. Motivation & Related Work

A. Need for Studying CAP Theorem on 5G Core

To justify the need for studying the CAP theorem on the 5GC, we briefly list a set of factors highlighting its need.

- 5GC is distributed by nature with multiple NFs involved to handle different types of user requests. However, as proven in the work [2], there is a tight inter-dependency between these NFs. That makes it more crucial to evaluate the consistency and availability challenges on the 5GC.
- 5GC is supposed to serve a variety of users for a variety of services. While most users are already registered, some users connect to the 5GC in emergencies such as disaster management and users traveling across regions during handovers. In such a situation, HA and resilience are of utmost importance, as otherwise, these users cannot avail the service.

B. Related Work

While the 5G was evolving, an initial study of the CAP theorem has happened for 4G networks in [3]. Here, the authors perform an empirical study of the CAP theorem in a 4G mobility situation and confirm that the incurred data service unavailability time is comparable to that induced by wireless connectivity setup. They also highlight that the desirable mobility control for the 5G networks should take a paradigm shift. While no work so far has specifically evaluated the CAP theorem on 5GC, recently, multiple works have emphasized on catering to resilience and HA service of the 5GC. In this context, authors in [4] highlight the significant need for handling short-time scale traffic variation in the 5GC and its impact on resilience. In a similar context, the authors in [2] show the impact on the procedure completion time, due to tight inter-NF dependency in the 5GC, for which the same authors have proposed a Per Procedure 5G system

Authorized licensed use limited to: Indian Institute of Technology Hyderabad. Downloaded on October 10,2026 at 15:42:09 UTC from IEEE Xplore. Restrictions apply.

(PP5GS) [5]. Here in [5], they have developed Per-Procedure NFs in the 5GC, such as an NF for serving UE registration, another NF for serving PDU session establishment, and so on. With this, the authors remove the dependency on multiple NFs for every UE procedure and illustrate resilience by remaining highly available to the best possible level. However, even in this work, Unstructured Data Storage Function (UDSF), UDM, and NRF are stateful and hence remain as external NFs without getting absorbed into any of the PP5GS NFs.

Overall, these works emphasize the significance of maintaining HA, reliability, and resilience of the ultimate service while preserving the architectural and scalability needs. However, the CAP theorem brings in the trade-off between consistency and availability and their influence on the overall resilience of the ultimate service provided. Hence, in this work, we dive deep into evaluating the CAP theorem on the 5GC. To the best of our knowledge, this is the first work to evaluate the CAP theorem on the stateless 5GC and identify the problems arising from its properties. Note that, as we specifically focus on the stateless 5GC, in the rest of the paper, we refer to it as 5GC only for brevity.

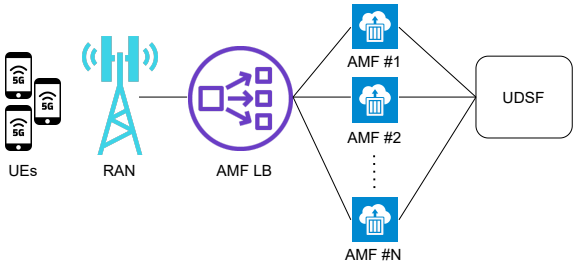


Fig. 2: View of scalable and stateless working of AMF with a load balancer, multiple serving instances, and a UDSF.

III. Study of CAP Theorem on the 5G Core

We study the CAP theorem on the 5GC in two stages. In the first stage, we identify the critical path(s) that could affect the HA service of the 5GC while serving different user requests. Then in the second stage, we apply the CAP theorem to the identified critical path(s), to eventually realize the outcomes of it.

A. Stage 1: Identification of Critical Path

As such by the statement of CAP, we realize that when there is no partition in the 5GC, both consistency and availability can be catered to the full extent. Therefore, in the first stage, we check the possible situations where a partition could happen. In this regard, we list the various NFs involved in serving different user requests. As specified by 3GPP [6] and listed in works [5, 7], AMF, AuSF, and UDM are the main NFs in serving UE registration, while SMF and User Plane Function (UPF) are involved along with AMF to serve the PDU session related messages. However, in the stateless architecture [5, 8], UDSF [9] has an important role to play. It (UDSF) is a data store for storing UE context. Mainly the UE state, the current PDU sessions the UE has established, security information like the Non Access Stratum (NAS)-related keys, and so on. This UE context is being written and read to and from the UDSF [8] regularly at the end and start of respective UE procedures like UE registration, and PDU session establishment, until the UE de-registers, upon which the corresponding UE context is removed from the UDSF. Fig. 2 depicts the scalable and stateless working of the AMF with a load balancer, multiple serving instances working in the stateless mode, and a UDSF.

Additionally, the UDSF can be global or local by scope. Specifically, when the UDSF is global, every other NFs like AMF and SMF requiring it can update and read the required UE context from it. But when it is local in scope to the respective NF, like a UDSF for AMF (AMF-UDSF), a UDSF for SMF (SMF-UDSF), and so on, multiple instances of the corresponding NF leverage their respective UDSF for UE context state synchronization to improve the needs of scalability, resilience, and HA of the service [8].

Nevertheless, we realize that in the stateless architecture, UDSF (global or local) should be reachable all the time, especially at the beginning and the end of handling the different UE procedures (user service requests), to ensure that the UE context remains up-to-date and consistent for access by multiple NFs or instances to serve a variety of requests (UE procedures) from the corresponding user. Hence, we identify the UDSF as an important NF in the 5GC and the path involving UDSF as a critical path for catering to resilient and HA service.

B. Stage 2: Studying CAP on the Critical Path

From Stage 1 (see Section III-A), we identify that the UDSF is situated on the critical path, and in order to cater to the consistency and HA of the service, it should be reachable all the time for serving different types of user requests. Taking this observation, we now inject the partition into this critical path for CAP theorem analysis.

So, we consider the partition as a situation of UDSF being unreachable from other NFs (in the case of global UDSF) or from the instances of NF (in the case of local UDSF) due to sudden link failures or the UDSF NF failure. While the type and amount of information stored in global and local UDSF can be different, for simplicity, here onwards, we consider the local UDSF, as that is sufficient to study the CAP theorem.

Now, that the partition is introduced in this form, we evaluate the CAP theorem (see Section I for the definition of CAP theorem), checking if consistency and availability can be maintained at the same time or not. Note that, while our study on CAP theorem is applicable for other NFs too, for illustration and brevity, henceforth we focus on AMF, as AMF is the uplink entry point for all the control plane interactions from UE and Radio Access Network (RAN).

- **Availability with Partition tolerance (AP):** Availability in the 5GC [8] is evaluated by the presence of the required NFs (along with a load balancer in the scalable architecture) to serve a variety of arriving user requests.

Authorized licensed use limited to: Indian Institute of Technology Hyderabad. Downloaded on October 10,2026 at 15:42:09 UTC from IEEE Xplore. Restrictions apply.

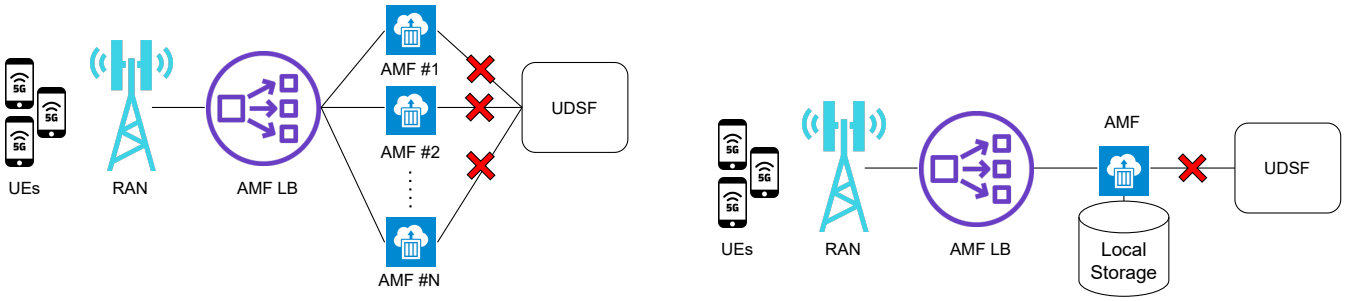


Fig. 3: View of AMF functioning with stateless AP mode in a) and stateful CP mode in b).

Before the partition occurs, the situation is normal and hence the new users' requests for registration can be handled to the successful completion. However, when the partition occurs such that UDSF is not reachable, the available AMF instances can still serve the new registration requests but not the PDU session establishment requests from the already registered users. This is because they (AMF instances) work in stateless mode and cannot reach the UDSF to read the UE state information when these requests arrive. Hence, in this situation, though multiple AMF instances are available, they cannot serve the already registered users who may be in need of data services. We term this situation as *AP*, representing Availability and Partition tolerance, as here the consistency is sacrificed to Availability when the Partition occurs.

- **Consistency with Partition tolerance (CP):** Consistency in the 5GC is evaluated with respect to serving a huge number of users for a variety of arriving user requests with up-to-date access to the UE state information. However, when the partition occurs such that UDSF is not reachable, the available AMF instances cannot serve the PDU session requests, as they cannot reach the UDSF for reading the UE state information. But to maintain consistency, as typically done in the distributed database-enabled systems, only a single AMF instance is kept active to serve these UE requests (PDU session requests), while the rest are made inactive or unavailable. So, here, the only alive instance can serve these requests without any consistency issues on the UE state. Essentially, the single alive AMF instance handles both registration and PDU session establishment requests, by maintaining the UE state information locally. However, in this case, since a single instance is present, it can handle only a limited number of such user requests to its capacity limit. Additionally, as this single instance could be overloaded, served user requests' procedure completion time [10] gets affected too, compared to multiple such instances actively serving more such user requests in parallel [8]. Hence, overall service availability gets disturbed bringing down the overall performance of the system. We term this situation as *CP*, representing Consistency and Partition tolerance, as here the availability is sacrificed to Consistency when the Partition occurs.

IV. Performance Evaluation

A. Experimental Setup

For the performance evaluation of 5GC in the AP and 5GC in the CP, we picked the in-house 5GC prototype. Similar to openly available 5G cores like free5GC and OAI-5G Core, the in-house 5GC prototype is 3GPP-compliant with control plane NFs including AMF, AuSF, UDM, NRF, SMF, AMF-UDSF and UPF for the data plane services. All the control plane NFs support service-based interactions with each other on HTTP/2. In addition, for end-to-end testing, a RAN+UE emulator is also developed as part of this prototype. The emulator generates the UE messages for registration and PDU session establishment towards the 5GC provisioned by the RAN portion inside it. So, using this prototype, we set up the respective frameworks shown in Fig. 3a and Fig 3b.

For the end-to-end testing, we bring up the 5GC first with required NFs, including AMF with 2 instances, AuSF, UDM, NRF, SMF, and UPF. Then we start the RAN+UE emulator in both modes. Upon start, the RAN+UE emulator feeds a different number of UEs which register and try to establish the PDU session in sequence to the 5GC. Note that the emulator sends different UE requests to both the AMF instances explicitly, without any AMF Load balancer in between the emulator and the AMF instances. The average time taken to complete a single PDU session establishment procedure on this test setup is '20 ms'. Using this '20 ms' value, we considered the ground truth capacity of processing 50 such requests per second for every AMF instance. During the testing, we inject a partition into the test setup, by shutting down the AMF-UDSF momentarily. When this partitioning occurs, for 5GC in CP mode, only a single AMF instance is made available as per Fig. 3b. At this point, we measure the different metrics from individual setups.

Authorized licensed use limited to: Indian Institute of Technology Hyderabad. Downloaded on October 10,2026 at 15:42:09 UTC from IEEE Xplore. Restrictions apply.

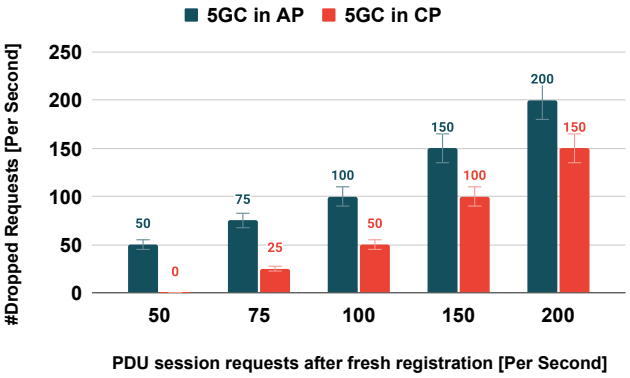


Fig. 4: Number of dropped PDU session requests.

B. Evaluating Resilience with Consistency

To measure the resilience with consistency during the partition of UDSF being unreachable, we capture the 5GC's UE request tolerance in the form of the number of impacted UEs (users) by referring [11, 4]. These UEs first register and then trigger PDU session establishment towards the 5GC. Fig. 4 compares the number of PDU session establishment requests dropped by the two modes. We can observe that 5GC in AP mode drops all these requests, as none of the AMF instances can access the UE state, due to UDSF's unreachability. On the other hand, with a single AMF serving instance working in stateful mode, the 5GC in CP mode drops only the excess requests beyond its capacity.

C. Evaluating High Availability

When a partition occurs (with UDSF going unreachable), and different user requests continue to arrive at the 5GC, we measure the total availability of the AMF. We refer to the total availability calculations of AMF from [8], shown below (the equations). Availability of individual serving instances of AMF is given by

$$A_{AMFi} = \frac{Uptime\ of\ AMFi}{Uptime\ of\ AMFi\ +\ Downtime\ of\ AMFi} \quad (1)$$

The total availability of AMF without AMF-UDSF is given by

$$A_{AMF} = A_{AMF-LB} * (1 - (1 - A_{AMFi})^n)) \quad (2)$$

where 'n' is the total number of serving AMF instances.

TABLE I: High availability of AMF

Availability of 5GC in AP	Availability of 5GC in CP
99.972%	98.33%

As shown in (1) and (2), the total availability of AMF (when AMF-UDSF is not reachable), depends on the availability of AMF load balancer and the availability of individual serving instances. Note that, we assume 100% availability for the load balancer in both modes, and handling its failure is outside the scope of this work. Here, first, we compute the downtime of every AMF instance. i.e., for a total time of 5 minutes of testing, we restart every available AMF instance once. We observe that it takes 5 seconds for a new AMF instance to come up, resulting in 98.33% availability for it, as per (1). Using (2)), Table I lists the total availability calculated for both the modes. Here, we can realize that with the presence of 2 instances, 5GC in AP mode proves higher availability as compared to the 5GC in CP mode.

V. Conclusion and Future Work

This paper studied the CAP theorem on the 5GC and identified the trade-off between 5GC in AP and CP modes of functioning when the partition occurs with UDSF going unreachable. Our experiments on the 3GPP-compliant 5GC testbed show that the 5GC in CP mode serves more PDU session requests while remaining less available as compared to the 5GC in AP mode. In the future, we plan to propose an effective solution that could possibly address the trade-off identified in this work.

Acknowledgment

This work was supported by the Department Of Science & Technology, India as part of the "Next Generation Wireless Research and Standardization on 5G and Beyond" project.

References

[1] Seth Gilbert and Nancy Lynch, "Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services", *SIGACT News*, vol. 33, no. 2, pp. 51–59, jun 2002.

[2] Endri Goshi, Michael Jarschel, Rastin Pries, Mu He, and Wolfgang Kellerer, "Investigating inter-nf dependencies in cloud-native 5g core networks", in *2021 17th International Conference on Network and Service Management (CNSM)*, 2021, pp. 370–374.

[3] Yuanjie Li, Zengwen Yuan, Chunyi Peng, and Songwu Lu, "CAP on Mobility Control for 4G LTE Networks", in *Proceedings of the 3rd Workshop on Hot Topics in Wireless*, New York, NY, USA, 2016, HotWireless '16, p. 33–37, Association for Computing Machinery.

[4] Rui Li, Bertrand Decocq, Anne Barros, Yi-Ping Fang, and Zhiguo Zeng, "Estimating 5G Network Service Resilience Against Short Timescale Traffic Variation", *IEEE Transactions on Network and Service Management*, vol. 20, no. 3, pp. 2230–2243, 2023.

[5] Endri Goshi, Raffael Stahl, Hasanin Harkous, Mu He, Rastin Pries, and Wolfgang Kellerer, "PP5GS—An Efficient Procedure-Based and Stateless Architecture for Next-Generation Core Networks", *IEEE Transactions on Network and Service Management*, vol. 20, no. 3, pp. 3318–3333, 2023.

[6] 3GPP, "Procedures for the 5G System", Tech. Rep. TS 23.502.

[7] Shwetha Vittal, Aditya Chilukuri, Sourav Sarkar, Akshitha Shinde, and Antony Franklin A, "Performance study of large scale network slice deployment in a 5g core testbed", in *2021 IEEE 4th 5G World Forum (5GWF)*, 2021, pp. 311–316.

[8] Shwetha Vittal and A. Antony Franklin, "HARNESS: High Availability Supportive Self Reliant Network Slicing in 5G Networks", *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, pp. 1951–1964, 2022.

[9] 3GPP, "Unstructured Data Storage Services", Tech. Rep. TS 29.598, 2023.

[10] Bhavishya Sharma, Shwetha Vittal, and A Antony Franklin, "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*. 2023, APNET '23, p. 61–66, Association for Computing Machinery.

[11] ENISA, "Resilience Metrics and Measurements: Technical Report", https://www.enisa.europa.eu/publications/metrics-tech-report.

Authorized licensed use limited to: Indian Institute of Technology Hyderabad. Downloaded on October 10,2026 at 15:42:09 UTC from IEEE Xplore. Restrictions apply.