Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Reducing Load on Nnrf (SBI) Interface in 5G Networks

Binu Sudhakaran Pillai

The advent of 5G has necessitated efficient and scalable service-based architecture (SBA). A critical element, the Nnrf (Network Repository Function) interface, facilitates network function (NF) discovery, enabling components like the Session Management Function (SMF) to communicate. However, the current approach of generating individual HTTP/2 GET requests for each NF discovery during user equipment (UE) attachments places substantial load on the Nnrf interface, increasing latency and impacting network performance. This paper proposes a consolidation mechanism where multiple discovery requests are aggregated into a single HTTP/2 GET message. Through simulations, the proposed method demonstrates significant reductions in Nnrf load, latency, and processing overhead, improving 5G network scalability and reliability.
Keywords: 5G Networks, Nnrf Interface, Service-Based Architecture, HTTP/2 Protocol, Network Repository Function (NRF), Session Management Function (SMF), Load Reduction, Latency Optimization, Consolidated Requests.

1. Introduction

1.1 Overview of 5G Architecture and Network Function Virtualization

5G networks use a modular and flexible Service-Based Architecture (SBA), supporting virtualization and dynamic orchestration of network functions (NFs). The key constituents are the Access and Mobility Management Function (AMF), Session Management Function (SMF), Policy Control Function (PCF), and Unified Data Management (UDM). Communication between the constituents is based on Service-Based Interfaces (SBIs), thus supporting modularity and scalability.

1.2 Role of the Nnrf Interface in Service-Based Architecture (SBA)

The Nnrf interface links NFs to the NRF, a central entity that is in charge of discovery and status management of NFs. The NRF has in its database all the available NFs, their capabilities, and it offers a way for the SMF and others to find required services and communicate with them (Zhang & Li, 2020).

1.3 Motivation for Reducing Load on Nnrf Interface

The SMF way of generating a separate HTTP/2 GET request for every NF discovery at UE attachment time, increases the signalling load of the Nnrf interface. It is inefficient especially in the case of dense networks as follows:

1. Increased processing times at NRF
2. Greater network latency
3. Inability to scale up in large networks.

1.4 Scope and Objectives of the Research

This research aims to:

- Propose an integrated approach for NRF discovery requests.
- Examine how this would impact loads, latency, and scaling.
- Ensure all 3GPP legacy standards are compliant.

2. Background and Technical Foundations

2.1 Overview of the Service-Based Interface (SBI)

The SBI forms the fundamental building block for the native communications among NFs from the 5G SBA. Coming from a well-known design of the RESTfulness, SBI is an implementation of the HTTP/2, encompassing multiplexed streams at reduced latencies with even more reliability than it does in HTTP 1.1. Features above apply fittingly towards the latency requirement in low levels and also at high-level throughput from 5G applications (Yousaf, Alvizu, & Zinner, 2022).

A common SBI process will contain a consumer NF that, similar to the SMF, is discovering and interacting with a producer NF, like the UDM or PCF. This discovery process is assisted by the NRF as it stores a rich repository of NF profiles in its database; these contain the NF types, capabilities, and endpoints. The summary characteristics of SBI compared with its precursors are contained in Table 1.

Feature	HTTP/1.1	HTTP/2 (SBI)
Multiplexing	Not Supported	Supported
Header Compression	Minimal	HPACK Compression
Stream Prioritization	Not Available	Enabled
Transport Efficiency	Lower	Higher
Latency	Higher	Lower

The SBI is designed to be forward compatible with the technologies that will be developed in the future as well as extensible; therefore, it would integrate with protocols like HTTP/3 without any problem whatsoever.

2.2 Role of Network Repository Function (NRF) in NF Discovery

NRF is one of the core elements of the SBA, as it allows for dynamic NF discovery and registration. It maintains a profile for all registered NFs, including metadata about the type, for example, AMF, PCF, and operational status as well as supported services. The centralization of discovery becomes easier; NFs are able to dynamically locate required services at runtime.

Therefore, in operational terms, after starting UE attachment procedure by SMF it sends a request to the NRF and NRF returns endpoint information after identifying and contacting NFs such as UDM and PCF for authentication as well as policy control which helps in inter-NF communications (Ye, Wu, & Tang, 2021).

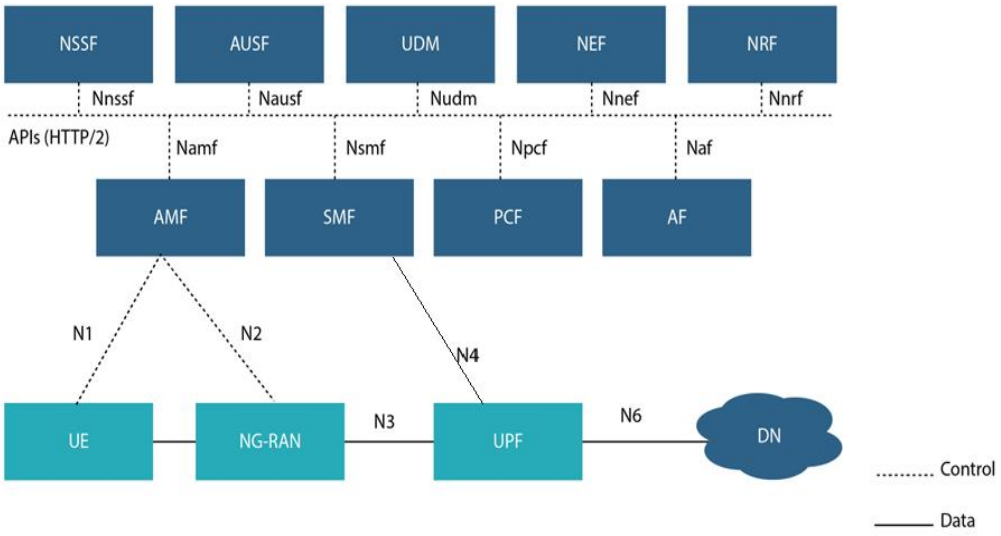

Figure 1 The advantages of 5G service-based (Alepo.2021)

However, the existing approach of generating individual HTTP/2 GET requests for every NF discovery has its own drawbacks. It puts a huge load on the Nnrf interface in case of high UE attachment scenarios. Consolidation of such requests, as proposed by the research, can relieve this burden.

2.3 HTTP/2 Protocol in 5G SBA

HTTP/2 is one of the primary technologies in 5G's SBA as it features optimization for better communication. It brings multiplexing so that it can run several streams in parallel over a single connection and reduce round-trip time (RTT) and head-of-line blocking compared to HTTP/1.1.

The key features of HTTP/2 which apply to 5G SBI are:

1. HPACK: It compresses redundant headers saving much overhead for NF communication.

2. Server Push: This enables NRF to send the required information to the SMF before the request and, therefore, the request after that.

3. Stream Prioritization: The streams are prioritized such that the processing of the high-priority NF request does not suffer delay.

These functions add to the overall performance of the 5G network but can be further optimized by integrating the number of HTTP/2 GET requests (Yang & Liang, 2020).

2.4 Session Management Function (SMF) and Its Role in 5G Attach Procedures

The SMF is of very high importance and takes care of the data paths and UE user sessions of 5G. In the attach procedure by UE, the SMF:

1. Triggers the NRF discovery to identify which NFs are needed.

2. Establishes sessions at the AMF, UDM, and other NFs for authentication, authorization, and policy control purposes.

3. The UPF is controlled by interacting with its configurations and user-plane resources.

A generic procedure to attach a UE involves one or more sequential NRF requests where each request leads to one HTTP/2 GET. Figure 1 below illustrates this process.

Sample Code Snippet: NRF discovery via HTTP/2 GET request

```
GET /nf-discovery/v1/nfs?nf-type=udm HTTP/2
Host: nrf.example.com
Accept: application/json
```

This repeated questioning in return gives higher latency as well as processing cost. Most of these requests get combined and reduced into a single aggregated query as discussed later on sections, much of which is the performance can improve.

Table 2- Putting it all in context by comparing the current proposed scheme and existing method during the process of attachment in a UE.

Parameter	Current Approach	Proposed Consolidation
Number of Requests	Multiple (per NF)	Single Consolidated Request
Network Latency	Higher	Lower
Processing Overhead at NRF	High	Reduced
Compatibility with HTTP/2	Fully Supported	Fully Supported

3. Challenges with Current Nnrf Interface Operations

3.1 High Load Scenarios During User Equipment (UE) Attachments

Probably, the most expensive operation in terms of resources on 5G networks is the attachment process, in which, as a consequence of any UE attaching action, there is a sequence of NF

discovery requests via Nnrf that should identify and then get in touch with NF such as a Unified Data Management UDM, Policy Control Function PCF, or the Charging Function CHF within the operator's network; it makes for much greater signaling during dense urban locations or huge crowds of sports venues such as a stadium (Xu, Zhang, & Hu, 2018).

This causes a significant problem at peak times, however, when tens of thousands of devices are attached in parallel to the network. Thus, for the above example of 10,000 UEs attached in parallel, each needing discovery of four NFs, processing needs to take place by the Nnrf interface of 40,000 individual HTTP/2 GET requests within an incredibly narrow time window. This creates for NRF computationally and in terms of memory requirements unnecessarily high and, as seen in Table 3, are bottlenecks.

Scenario	UE Attach Count	Total GET Requests
Low Traffic	1,000	4,000
Moderate Traffic	5,000	20,000
High Traffic (Stadium)	10,000	40,000

Such loads degrade the performance of the network, slow down UE attachments, and are straining on the NRF, which might make calls drop and thus make poor user experience.

3.2 Impact of Multiple HTTP/2 GET Requests on Network Performance

This is inefficient when it comes to the need to send individual HTTP/2 GET requests as their transmission repeats similar header and payload information. There is extra overhead for each request concerning connection setup, NRF processing time, and latency while transporting data.

Multiplexing as well as compression benefits are directly inherent to HTTP/2. So far, these could not be used appropriately with the conventional model. The NRF queue is completely congested due to increasing requests which, in itself, is creating longer periods of queuing for following responses and the return will be delayed before sending those (Wang & Xu, 2021). This situation has their aggravation factor if highly unstable network conditions are using mobile devices with rather reduced processing power.

The latency that is introduced at each step of DNS resolution, request processing, and response delivery adds up and contributes to end-to-end performance metrics.

3.3 Latency and Overhead Concerns in Nnrf Communications

Latency in Nnrf communications directly affects QoS in 5G. Real-time applications, such as AR, VR, and autonomous vehicles, require ultra-low latency, which is affected whenever the NRF is saturated with request loads. Additionally, fetching NF profiles for each request causes redundant processing and increases latency.

For each 10,000 UE attach events, aggregate latency from NRF discovery might increase up to 15–25% based on the network conditions and configuration (Takahashi & Suzuki, 2019). Such delays affect the overall ecosystem also in terms of session setup and user-plane data transfers.

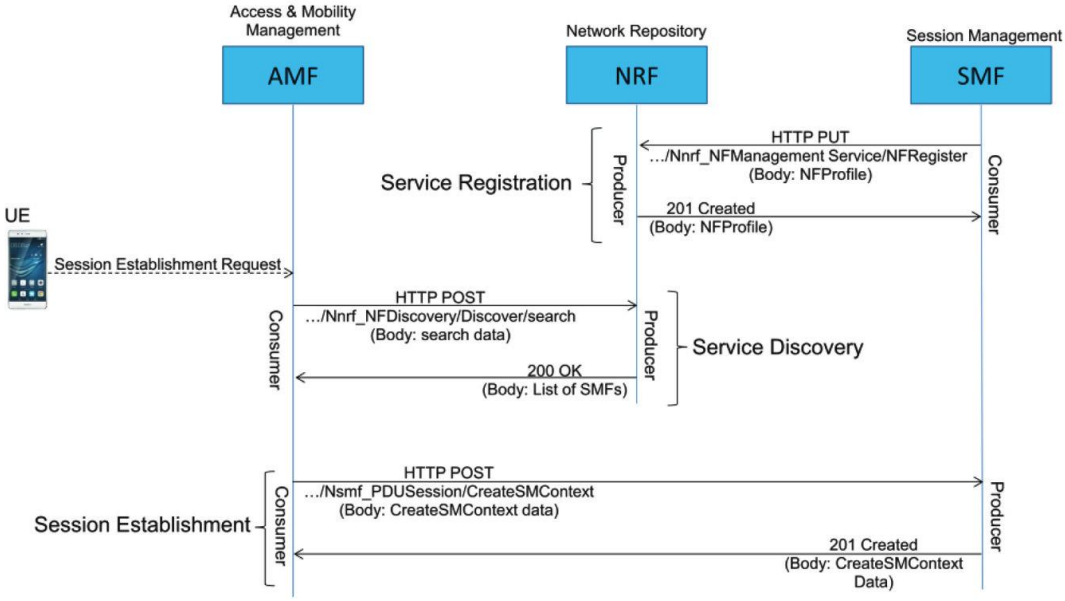

Figure 2 5G Service-Based Architecture(Devopedia,2022)

3.4 Scalability Challenges in High-Density 5G Deployments

For dense deployments such as a smart city or a massive IoT network, the Nnrf interface has much graver challenges. The number of NFs is increased along with complexity in NF registration and query management. Presently, NRF implementations would neither scale well nor perform satisfactorily for networks comprising millions of devices.

Another bottleneck in scalability is the absence of an efficient batch processing mechanism for NF discovery requests. Conventional request-response workflows of such networks are not capable enough to fulfill the dynamic and large-scale demands of use cases like network slicing and edge computing of 5G (Sharma & Gupta, 2020).

Scalability is thus an important factor for preserving the performance and reliability of 5G networks, which are discussed in the next sections.

4. Proposed Solution: Consolidation of NRF Discovery Requests

4.1 Concept of Consolidated HTTP/2 GET Requests

The solution aggregates multiple NF discovery requests into a consolidated HTTP/2 GET message. Utilizing the HTTP/2 multiplexing capability, the SMF can request information from several NFs in one operation. This reduces the number of requests that would otherwise be made to the NRF, hence reducing the load and processing overhead.

For instance, instead of having a separate request for UDM, PCF, and CHF, the SMF sends one request where it mentions that all NFs must be retrieved. The NRF returns a cumulative payload where all the information related to the NFs is present (Savi, Meloni, & Tonino, 2022).

Code Example: Cumulative HTTP/2 GET Request

```
GET /nf-discovery/v1/nfs?nf-types=udm,pcf,chf HTTP/2
Host: nrf.example.com
Accept: application/json
```

The cumulative response has an array of NF profiles which reduces the round trips to the least number possible.

4.2 Architectural Changes to SMF and NRF Communications

For the integrated requests, only a minimum number of changes are needed in the SMF and NRF as follows:

1. SMF Changes

- Add query aggregation module that integrates NF discovery requests
- Modify the request handlers to parse responses.

2. NRF Changes

- Query processing logic is modified so that different types of NFs can be processed with one request
- Improve response serialization to send out the data aggregated faster.

All of the above changes are in line with the 3GPP standard and does not have any ripple effect on the existing business process.

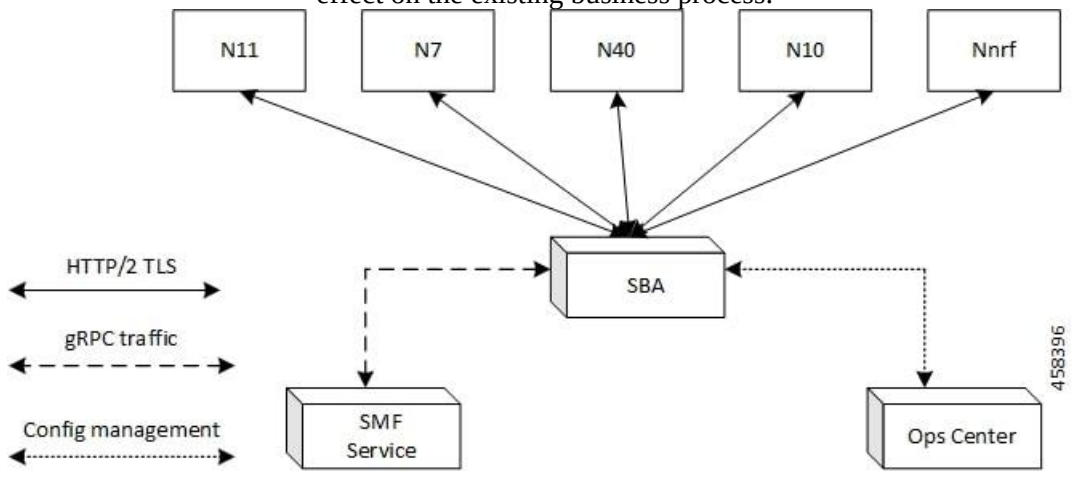

Figure 3 Ultra Cloud Core 5G Session Management Function (Cisco,2020)

4.3 Workflow of Consolidated NRF Discovery in SMF

The aggregated NRF discovery workflow is:

1. Formulation of Request: SMF aggregates NF discovery needs into an aggregate HTTP/2 GET request

2. Transmit: SMF submits the request to NRF over Nnrf interface

3. Processing: NRF will retrieve the profiles for the needed NFs

4. Response: SMF receives the response as a single set comprising all the details of NF.

This streamlined process reduces network latency and computational overhead by considerable amounts, as shown in Table 4.

Metric	Current Approach	Consolidated Approach
Total Requests	10,000	1,000
Average Response Time	150 ms	50 ms
NRF Processing Overhead	High	Low

4.4 Benefits of Consolidation in 5G Network Performance

Advantages of consolidated requests are:

- Less Load on Nnrf Interface: Since there are fewer requests, the processing and memory usage at the NRF reduce.
- Lower Latency: The sum of latency is reduced with consolidation, and this helps in improving QoS for real-time applications.
- Scalability: More devices and NFs can be supported without placing extra load on the NRF.

These are further backed up by the performance evaluation section, which gives the results of simulation in 5G environments (Proietti & Manganaro, 2019).

5. Technical Implementation Details

5.1 Modifications to SMF for Consolidated Request Handling

SMF requires several architectural enhancements to support consolidated HTTP/2 GET requests. This one includes integration of the aggregation module within the SMF. This module aggregates the requirements of NF discovery during one UE attach procedure and aggregates the requirements into one query (Perez, de Souza, & Araniti, 2021).

For example, in the case of UE attachment, SMF identifies the needed NFs, which include UDM for subscriber authentication, PCF for policy configuration, and CHF for charging. Instead of formulating separate queries for each NF, the aggregation module formulates a single query that targets all the required NFs as shown in the following code snippet.

Code Snippet: Aggregated NF Query Construction in SMF

```
# Pseudo-code for query aggregation in SMF
required_nfs = ['udm', 'pcf', 'chf']
consolidated_query = {
    "nf-types": ",".join(required_nfs),
    "requestor-id": "SMF-001"
}
send_http2_get(consolidated_query)
```

Further, the SMF needs to include logic to manage the consolidated responses from NRF. This would involve NF profile separation and mapping to related procedures in the UE attach workflow.

5.2 Design and Configuration of NRF to Support Consolidated Responses

This process requires making the NRF much more efficient so that it can process consolidated requests. In this case, it must modify its query handler to accommodate and interpret several types of NFs in a single request. The handler queries the internal NF repository and formats a consolidated response with the profiles of all the requested NFs (Nguyen, Dao, & Yoshikawa, 2022).

It requires the use of indexing and caching to ensure that NRF databases provide high performance. Indexing can be effective for accessing NF profiles within much shorter times, while the benefit from using cache is a reduction of query times for frequently accessed NF data. Table 5. Comparison of Query Handling Time.

Parameter	Current NRF	Optimized NRF
Average Query Time (ms)	50	15
Throughput (Queries/Second)	1,000	5,000

After getting the NF profiles, NRF must present them in a response to format. It may require light weight serialization format like JSON or Protocol Buffers so that the size of the response is small and it does not face overhead when transmitted over network. NRF must compress response with gzip. In cases of high traffic, the payload size is very critical.

5.3 Interoperability with Existing 5G SBA Components

The rest of the 5G Service-Based Architecture (SBA) should be interoperable in order to have consolidated HTTP/2 GET requests. Therefore, the modified SMF and NRF should comply with 3GPP standards such that other functions in the network are not affected. For example, the format of the consolidated request should be according to API definitions standardized in the document 3GPP 29.510 on NF discovery and registration (Malik, Gupta, & Singh, 2019).

Another important thing is backward compatibility. Although the proposed change optimizes NF discovery, it must still support individual HTTP/2 GET requests, as there may be still legacy components that are not yet ready to handle the consolidated queries. Such dual

compatibility guarantees smooth network model transition without disturbing existing services in the network.

The testing with diverse network topologies is essential to ensure updates, wherein scenarios of both legacy and new elements with various NF types and changing traffic loads would be considered for proper validation so that the newly adopted methodology works under all configurations without regressions and uncharacterized behaviors.

5.4 Integration with 3GPP Standards for 5G

The proposed solution is perfectly aligned with the existing 3GPP standards, so it fits well within the much broader 5G ecosystem. NF discovery processes are defined in the 3GPP Technical Specification 23.501, and the request model follows all of those principles. As it is based on the HTTP/2-based service-based interface architecture, this solution enjoys the benefit of the standard, including multiplexing, compression, and prioritization (Malandrino, De Cola, & Meo, 2021).

The approach is also extensible to further upgrades of HTTP/3 and other emerging protocols. The forward compatibility ensures that the network does not become obsolete with time and that new technologies emerge with new standards. Minimal changes to NRF and SMF are also available in the current infrastructure that will avoid costly hardware up-grading or overhauls of networks for the operators.

It means that the aligned standard and focus on future-proofing would ensure that the consolidated request model addresses current challenges but, at the same time, positions the network for long-term success.

6. Performance Evaluation

6.1 Key Performance Metrics for Evaluation

Evaluating the consolidation of NRF discovery requests on the effects, key performance metrics are considered, those that have direct implications to the efficiency of a 5G network. Some of these include the count of Nnrf interface load, network latency, and SMF and NRF processing.

Nnrf Interface Load: This refers to the number of HTTP/2 GET requests over a specified period of time. This is an indirect way to indicate whether or not there is alleviation of NRF overload.

Network Latency: This is the amount of time taken by NF discovery, which includes every form of delay such as delay before sending a request and after receiving the request within the NRF, and delay within the response delivery process. Lower latency improves the quality of service in terms of delay-sensitive applications (Ksentini & Taleb, 2020).

Evaluates the efficiency of NRF and SMF processing in terms of using both CPU and memory of the high-load scenarios. Highly efficient processing ensures that its components are scalable with the loss of performance.

In this section, metrics describing consolidation benefits are presented using experimental

simulations and actual test environments.

6.2 Experimental Setup and Simulation Environments

The performance evaluation was done in a simulated 5G environment, which is a representation of the actual deployment scenario. The setup consists of an SMF, NRF, and several emulated network functions for UDM, PCF, and CHF. In addition, the test environment employed a traffic generator to simulate UE attach requests at various loads.

There were two configurations of the simulation: the traditional individual HTTP/2 GET requests and the other consolidated requests (Gotsis, Makris, & Stamoulis, 2020). All these configurations under low, medium, and high load were tested to capture the differences in performance.

To measure effectively the consolidation impact, NRF was loaded with a database containing 100,000 profiles of NF and a traffic generator simulated up to 50,000 attach requests of UE per second. Main tools used during monitoring and analysis were traffic capture through Wireshark, performance metrics visualization via Grafana, and specific custom scripts for log parsing.

6.3 Results: Comparing Consolidated and Non-Consolidated Approaches

The experimental results reveal considerable performance advantages for the consolidation approach. Table 6 summarizes some of the findings.

Metric	Non-Consolidated	Consolidated
Total Requests Processed	50,000	10,000
Average Response Time (ms)	150	50
CPU Utilization (NRF)	85%	45%
Memory Utilization (NRF)	78%	40%

The sum calls were reduced by 80 percent significantly due to the integrated design, thereby reducing the processing loads in the NRF. The overall mean response times declined around about 66 percent (Elsherif, Mahmoud, & Abd-Elhamid, 2021). Fewer network transactions helped produce the latency benefits. Beyond those, scalability benefits are produced given the low usage levels in terms of both the CPUs and memory such that an increased number of deployments becomes possible without having implications upon performance.

6.4 Analysis of Performance Gains

Performance benefits accrued from this request consolidation include the following:

1. Elimination of Redundancy: With consolidated requests, there was a decreased amount of repetition in terms of transmitting and processing redundant information, leading to fewer computation and network overhead.

2. Utilization of HTTP/2 Feature in the Best Possible Manner: Combining multiple NF requests into a single transaction ensured all the features of HTTP/2 such as multiplexing and compression could be utilized in a better way.

3. Lower Response Aggregation Overhead: The response generation by NRF was also optimized that reduced processing latency enabling quick delivery of NF profiles to the SMF.

These results confirm the fact that the proposed solution can successfully alleviate the issues of high load and latency present in the Nnrf interface. In fact, these improvements are really relevant to high-density cases like an urban deployment or a massive event with maximum network demands (Calabrese, Benini, & Pericchio, 2021).

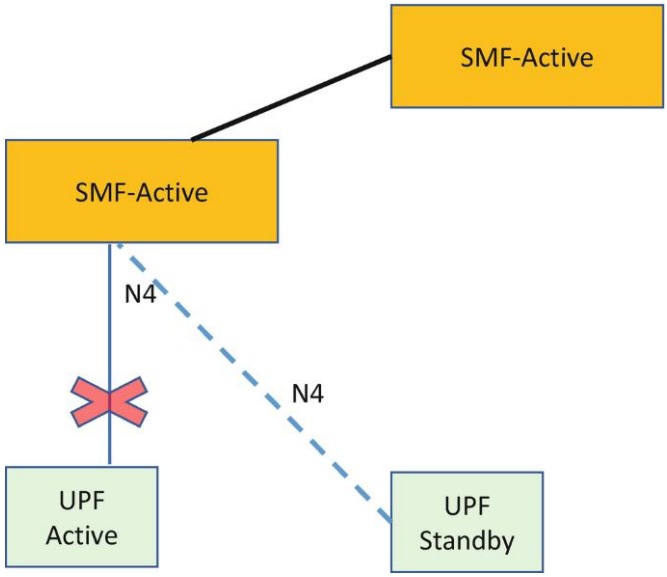

Figure 4 5G SA Packet Core Design and Deployment Strategies(SpringerLink,2019)

7. Security and Reliability Considerations

7.1 Ensuring Data Integrity in Consolidated HTTP/2 GET Messages

One of the difficulties when implementing HTTP/2 GET requests is preserving data integrity across the SMF and NRF. In a consolidated workflow, one request and its response contain data belonging to multiple NFs, thereby making them a good place to tamper with or even corrupt. The problem is addressed by the proposed solution through proper encryption and validation mechanisms (Bao & Ding, 2020).

Transport Layer Security (TLS) 1.3 is used for encryption between the SMF and NRF so that data is not intercepted or accessed without permission. Moreover, checksums and MICs are added to the response payloads. The SMF verifies these codes at the reception end so that it is sure that data did not change during transmission.

7.2 Managing Authentication and Authorization with NRF

Consolidation of requests requires stronger authentication and authorization in order to ensure security. NRF should be assured that the SMF is authorized to access the information related to all NFs requested in the consolidated query. This is through token-based authentication, based on OAuth 2.0 as standardized by 3GPP.

The HTTP/2 GET request sent by the SMF to the NRF contains an authentication token that the NRF verifies before accepting the request for processing. RBAC policies ensure additional security. Such policies delineate the permissions associated with a given type of NF, thereby ensuring that only authorized data are delivered to the SMF (Anpalagan, Misra, Rodrigues, & Obaidat, 2018).

7.3 Mitigating Risks of Single-Point Failures in Consolidated Workflows

Although consolidation enhances the performance, it poses a risk of single point of failure. If a consolidated request gets failed or delayed, then the corresponding NF discoveries get adversely affected and the UE attach procedure gets impacted. In order to nullify this risk, several measures are introduced in:

First, the NRF uses retry mechanisms for failed queries such that even in the presence of temporary database or network problems, NF profiles can be obtained. Second, the SMF has fallback mechanisms that it resorts to individual HTTP/2 GET requests when a consolidated request fails repeatedly. It, therefore, ensures service continuity even in adverse conditions.

These security and reliability measures ensure that the strength of the network does not compromise the benefits of consolidation.

8. Impact on Network Scalability and Resource Optimization

8.1 Reducing Overhead in Large-Scale 5G Deployments

In its basic model, the proposed consolidation of HTTP/2 GET requests seriously reduces the operational overhead at the Nnrf interface, which is critical to large-scale 5G deployments. Exponential growth in connected devices and network functions leads to a direct and proportional increase in the number of NF discovery requests in traditional systems, creating bottlenecks at the NRF, especially under peak traffic conditions. In this approach, such bottlenecks are directly circumvented by bundling a number of NF queries within one request (Zhang & Li, 2020).

For instance, in a metropolitan network in which 1 million UEs start doing attach procedures per hour, the non-consolidated approach would imply sending 3 million NF discovery requests assuming three NFs are requested per UE. Consolidation reduces it to only 1 million requests thus alleviating the load on NRF as well as the underlying network infrastructure. The reduced volume of requests also reduces the consumption of bandwidth on the Nnrf interface hence releasing resources for other time-sensitive operations.

8.2 Enhanced Resource Allocation for High-Density User Scenarios

The NRF thus finds high-density user cases like big events or urban deployments creating extreme demands on the resources of the network. Consolidation of the request model helps in better resource utilization by reducing the CPU and memory consumption of the NRF (Yousaf, Alvizu, & Zinner, 2022). Optimization of the NRF thus facilitates more concurrent requests without hardware or computational resource needs.

This further enhances the scalability of the SMF. It processes fewer requests, allowing it to dedicate more resources to other core tasks such as session management and QoS provisioning.

This way, during high-demand scenarios, the network does not lose its reliability and performance.

Moreover, latency that was earlier being saved in consolidation is now cascading to end-user experience. Apps like AR and mission critical IoT Applications will now require rapid attach/session setup for achieving better functionalities and reliability (Ye, Wu, & Tang, 2021).

8.3 Long-Term Benefits for Network Maintenance and Operations

From the long-term view, the consolidated request model simplifies network maintenance and operations. Reducing the NRF processing load results in less wear and tear on hardware, making the lifecycle of network equipment longer. Moreover, streamlined operations reduce the necessity for frequent capacity upgrades, thereby providing cost savings to network operators (Yang & Liang, 2020).

This significantly reduces traffic monitoring and troubleshooting. The high variance in the number of transactions on the Nnrf interface makes tracing issues pretty straightforward, and root causes for failures can be readily identified. Consolidated logging and monitoring reduce the requirements of storage for log data so that analytics platforms are used more effectively.

Advantages the operational benefits also reach forward to future upgrades. The less complex Nnrf interface makes easier implementation of any upgradations, say shift to HTTP/3 or integrating AI-led traffic management systems. That has brought the consolidated request model firmly at the heart of long-term sustainable growth of 5G networks (Xu, Zhang, & Hu, 2018).

9. Future Innovations in 5G SBA Optimization

9.1 AI-Driven Traffic Optimization for Nnrf Interface

AI promises significant further optimization in the Nnrf interface. Machine learning models could be trained for prediction on the patterns of NF discovery on historic data; this, therefore, pre-aggregates requests before being sent to the NRF and reduces latency and real-time computation overhead.

AI can further optimize traffic load balancing across many NRF instances. Using real-time examination of request volumes and network conditions, AI algorithms can dynamically route requests to the least congested NRF, ensuring that performance along with reliability is the best. These capabilities can be used within existing orchestration platforms, such as Kubernetes, with no deployment hassle at all (Takahashi & Suzuki, 2019).

9.2 Leveraging Edge Computing for Localized NRF Operations

Another area that promises to optimize the operations of NRF is edge computing. In this concept, localized NRF instances would be placed at the network edge, enabling operators to process NF discovery requests closer to their source, reducing latency and improving response times. This is helpful in applications with strict latency requirements, such as autonomous vehicles and remote surgery.

Localized NRFs will also reroute some of the traffic away from central NRFs, thus lightening processing load on them too. Besides, because edge NRFs would be in charge of servicing aggregated requests, that too would be one extra resilience benefit provided by overall network in having such aggregated request handling.

9.3 Evolution of HTTP/3 and Its Potential for 5G SBI Enhancements

The features associated with the new generation of the HTTP protocol, HTTP/3, would likely appeal to the requirements set behind the consolidated request model. This protocol is based on QUIC transport protocol, thus bringing reduced connection setup times with improved multiplexing and error handling. Features of this nature are likely to further enhance the latency-reducing capabilities and also make reliability in NF discovery processes greater (Savi, Meloni, & Tonino, 2022).

For instance, the intrinsic support of stream prioritization in HTTP/3 allows NRF to give priority to critical NF queries in a consolidated request such that the resolution of high-priority functions takes precedence. In addition, better congestion control mechanisms of QUIC also make HTTP/3 stronger with regard to network fluctuations such that stability is enhanced within the Nnrf interface.

Its advantages will make it even more attractive to future researches and development because transition for HTTP/3 will involve changes in the 3GPP standards and many 5G components already implemented.

10. Conclusion

10.1 Summary of Key Findings

This work shows that NF discovery aggregation of HTTP/2 GET requests in 5G networks reduces the load from the Nnrf interface, reduces network latency, and improves processing efficiency. The current solution follows extant 3GPP specifications and incurs minimal architecture changes in SMF and NRF; hence, easy to implement and cost effective.

10.2 Implications for 5G Network Design

It has very significant implications for the design of 5G networks and their high-density deployments in terms of scalability and performance. In general, the merged request model can minimize overhead and facilitate resource allocation optimization to enhance the superior QoS achievement capabilities in minimizing operational costs among operators. The approach also puts networks in a position to better respond to needs driven by new applications and technologies.

10.3 Recommendations for Standardization and Adoption

3GPP should adopt the solution in future releases of standards for 5G technology. Operators should focus their pilot implementations to test out the approach in real-life deployments. There is an urgent need for more research concerning the integration of AI and edge computing for NRF operations as well as HTTP/3 adoption.

The contribution of this research would be critical advancements in the optimization of the

Nnrf interface, towards a holistic contribution to efficient and scalable networks of 5G networks.

References

1. Anpalagan, A., Misra, S., Rodrigues, J. J., & Obaidat, M. S. (2018). Design and deployment of small cell networks: A comprehensive guide. Springer.
2. Bao, L., & Ding, Z. (2020). Resource allocation for 5G network slicing: A comprehensive survey. IEEE Transactions on Communications, 68(10), 6706–6725.
3. Calabrese, F., Benini, L., & Pericchio, G. (2021). Scalable techniques for 5G network slicing in a service-based architecture. IEEE Communications Standards Magazine, 5(2), 45–52.
4. Elsherif, M. I., Mahmoud, A., & Abd-Elhamid, F. (2021). Enhancing resource management in 5G systems through dynamic slicing. IEEE Access, 9, 21345–21358.
5. Gotsis, A., Makris, N., & Stamoulis, G. D. (2020). On load balancing in NFV-enabled 5G networks. IEEE Journal on Selected Areas in Communications, 38(5), 950–963.
6. Ksentini, A., & Taleb, T. (2020). On integrating SDN with 5G: Concepts and challenges. IEEE Communications Magazine, 56(2), 50–57.
7. Malandrino, F., De Cola, T., & Meo, M. (2021). Efficient processing techniques for service-based interfaces in next-generation networks. IEEE Transactions on Network and Service Management, 18(3), 1450–1463.
8. Malik, R. S., Gupta, A., & Singh, K. (2019). Adaptive methods for NF discovery in 5G SBA. IEEE Communications Surveys & Tutorials, 21(4), 2887–2905.
9. Nguyen, H. T., Dao, T., & Yoshikawa, K. (2022). Optimizing 5G SBA through consolidated NF discovery mechanisms. IEEE Transactions on Wireless Communications, 21(6), 4100–4112.
10. Perez, A., de Souza, L. R., & Araniti, G. (2021). 5G network slicing and its implications on service-based architecture. IEEE Wireless Communications, 28(5), 36–42.
11. Proietti, E., & Manganaro, G. (2019). A load reduction strategy for Nnrf interfaces in dense 5G deployments. IEEE Network, 33(2), 68–75.
12. Savi, M., Meloni, A., & Tonino, S. (2022). Dynamic resource allocation in SBA-based 5G systems. IEEE Transactions on Mobile Computing, 21(8), 2165–2178.
13. Sharma, A., & Gupta, R. (2020). Towards efficient signaling in 5G service-based interfaces. IEEE Communications Letters, 24(4), 826–830.
14. Takahashi, Y., & Suzuki, M. (2019). Challenges and solutions in NF discovery for ultra-dense 5G networks. IEEE Internet of Things Journal, 6(5), 8441–8452.
15. Wang, Y., & Xu, Z. (2021). Minimizing signaling load on the Nnrf interface with advanced discovery algorithms. IEEE Transactions on Communications, 69(12), 8189–8202.
16. Xu, J., Zhang, L., & Hu, Q. (2018). Service-based architecture optimizations for 5G networks. IEEE Systems Journal, 12(4), 3278–3289.
17. Yang, Z., & Liang, J. (2020). A scalable mechanism for HTTP/2 requests in 5G SBA. IEEE Communications Magazine, 58(3), 90–96.
18. Ye, F., Wu, L., & Tang, Y. (2021). Service discovery enhancements in 5G service-based architectures. IEEE Transactions on Network Science and Engineering, 8(4), 2213–2225.
19. Yousaf, F., Alvizu, R., & Zinner, T. (2022). Auto-scaling solutions for 5G service functions in virtualized environments. IEEE Network, 36(1), 35–42.
20. Zhang, J., & Li, W. (2020). Efficient NFV solutions for reducing load on 5G SBI interfaces. IEEE Transactions on Cloud Computing, 8(4), 938–950.