3GPP TR 29.843 V16.0.0 (2019-09)

Technical Report

3rd Generation Partnership Project;

Technical Specification Group Core Network and Terminals;

Study on Load and Overload Control of 5GC Service Based Interfaces;

(Release 16)

** 

The present document has been developed within the 3rd Generation Partnership Project (3GPP TM) and may be further elaborated for the purposes of 3GPP.  
The present document has not been subject to any approval process by the 3GPPOrganizational Partners and shall not be implemented.  
This Report is provided for future development work within 3GPPonly. The Organizational Partners accept no liability for any use of this Specification.  
Specifications and Reports for implementation of the 3GPP TM system should be obtained via the 3GPP Organizational Partners' Publications Offices.

Keywords

<keyword[, keyword]>

***3GPP***

Postal address

3GPP support office address

650 Route des Lucioles - Sophia Antipolis

Valbonne - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Internet

http://www.3gpp.org

***Copyright Notification***

No part may be reproduced except as authorized by written permission.  
The copyright and the foregoing restriction extend to reproduction in all media.

© 2019, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).

All rights reserved.

UMTS™ is a Trade Mark of ETSI registered for the benefit of its members

3GPP™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners  
LTE™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners

GSM® and the GSM logo are registered and owned by the GSM Association

Contents

Foreword 5

Introduction 5

1 Scope 6

2 References 6

3 Definitions, symbols and abbreviations 6

3.1 Definitions 6

3.2 Abbreviations 6

4 Architectural Baseline 7

5 Service Based Architecture Deployment Topologies 7

5.1 Introduction 7

5.2 Direct Interaction Models without Intermediaries 7

5.2.1 Direct NF to NF Interaction 7

5.2.2 NF Services as Distributed Collection 7

5.2.3 Interpretation of Overload 8

5.3 Indirect Interaction Models through Intermediaries 8

5.3.1 NF Service Instances behind a Service Communication Proxy without delegated NF Service discovery 8

5.3.2 NF Service Instances behind a Service Communication Proxy with delegated NF Service discovery 9

5.3.3 Interpretation of Overload 9

6 Overload Scenarios 10

6.1 Introduction 10

6.2 Overload Caused due to UE Initiated Signalling 10

6.3 Overload Caused due to Failure and Restart of Network Functions 11

6.4 Overload Caused due to Notifications 11

7 Key Requirements for Load Control 11

7.1 Introduction 11

7.2 Key Requirement #1: Load Information Conveyance 11

7.3 Key Requirement #2: Co-existence with Rel-15 Load Control 12

7.4 Key Requirement #3: Regulation of Load Information Reception and Conveyance in the NRF 12

8 Key Requirements for Overload Control 12

8.1 Introduction 12

8.2 Key Requirement #1: Overload Conveyance 12

8.3 Key Requirement #2: Mechanisms for avoiding and mitigating overload 13

8.4 Key Requirement #3: Overload Control 13

9 Evaluation of Release 15 Load and Overload Control 13

9.1 Introduction 13

9.2 R15 Overload Control 14

9.2.1 Behaviour 14

9.2.2 Evaluation and Conclusion 15

9.3 R15 Load Control Analysis 16

9.3.1 Analysis 16

9.3.2 Evaluation and Conclusion 16

10 Solutions for Load Control 17

10.1 Introduction 17

10.2 Solution #1: Direct Load Information Conveyance 17

10.3 Solution #2: Load Control Co-Existence 17

10.4 Solution #3: Regulation of Load Information Change Notifications Conveyance by the NRF 18

10.5 Solution #4: Regulation of Load Information Updates Reception at the NRF 18

10.6 Evaluation and Conclusion 19

11 Solutions for Overload Control 19

11.1 Introduction 19

11.2 Principles of Overload Control 19

11.3 Solution #1: Conveying Overload Information via Signalling 20

11.4 Solution #2: Explicit Overload Control Information Conveyance 22

11.5 Solution #3: Overload Control Information 22

11.6 Solution #4: Overload Avoidance and Mitigation 24

11.7 Solution #5: Frequency of Inclusion - Piggyback in every HTTP signalling message 26

11.8 Solution #6: Conveying Overload Information via NRF 26

11.9 Solution #7: Converying OCI via Overload Control Service 28

11.10 Solution #8: Overload Control enforcement behaviour 29

11.11 Solution #9: Location of OCI 30

11.12 Solution#10: Overload control by redirecting requests to another NF within a NF set 31

11.13 Solution #11: Overload Control Enforcement by SCPs 32

11.14 Conveyance of Overload Information with hyperlink 32

11.15 Evaluation and Conclusion 34

11.15.1 Evaluation of the Solutions 34

12 Selected solutions 36

12.1 Selected solution for Overload Control 36

12.2 Solution outline and implications on SBI specifications 36

12.2.1 Overload Conveyance - General (key requirement #1.1) 36

12.2.2 Overload Conveyance - Semantics (key requirement #1.2) 36

12.2.3 Overload Conveyance - Frequency (key requirement #1.3) 37

12.2.4 Overload Conveyance - Where to include OCI (key requirement #1.4) 37

12.2.5 Overload Conveyance - HATEOAS (key requirements #1.1 and #1.4) 37

12.2.6 Overload - Avoiding and mitigating (key requirement #2) 37

12.2.7 Overload Control (key requirement #3) 38

Annex: Change history 39

# Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

# Introduction

This clause is optional. If it exists, it is always the second unnumbered clause.

# 1 Scope

The present document identifies key requirements for the load and overload control on the service based interfaces, identifies the deployment topologies to consider for analysing the solutions for the load and overload control, analyses the 3GPP Release 15 overload control mechanism for the service based interfaces and identifies solutions for addressing the requirements for load and overload control on the service based interfaces.

# 2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non‑specific.

- For a specific reference, subsequent revisions do not apply.

- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] 3GPP TS 23.501: "System Architecture for the 5G System; Stage 2".

[3] 3GPP TS 29.500: "5G System; Technical Realization of Service Based Architecture; Stage 3".

[4] 3GPP TR 23.742: "Study on Enhancements to the Service-Based Architecture".

[5] 3GPP TS 29.501: "5G System; Principles and Guidelines for Services Definition; Stage 3".

[6] 3GPP TS 29.502: "5G System; Session Management Services; Stage 3".

[7] 3GPP TR 23.843: "Study on Core Network Overload Solutions".

[8] 3GPP TS 23.527: "5G System; Restoration Procedures".

[9] 3GPP TS 29.510: "5G System; Network Function Repository Services; Stage 3".

# 3 Definitions, symbols and abbreviations

## 3.1 Definitions

For the purposes of the present document, the terms and definitions given in TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in TR 21.905 [1].

## 3.2 Abbreviations

For the purposes of the present document, the abbreviations given in TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in TR 21.905 [1].

OCI Overload Control Information

SCP Service Communication Proxy

# 4 Architectural Baseline

3GPP Release 15 Service Based Architecture as specified in TS 23.501 [2] and the Technical Realization of the Service Based Architecture as specified in TS 29.500 [3] shall be taken as the baseline for the study on load and overload control in service based interfaces. Further the conclusions from 3GPP release 16 FS\_eSBA study as specified in TR 23.742 [4] shall be taken into account.

# 5 Service Based Architecture Deployment Topologies

## 5.1 Introduction

This clause identifies the various deployment models possible with Release 15 service based architecture. These deployment models are used to evaluate the various load and overload control mechanisms.

## 5.2 Direct Interaction Models without Intermediaries

### 5.2.1 Direct NF to NF Interaction

In this topology, each NF and correspondingly the NF services interact directly with its peer NF and the NF services. The following are the key characteristics of this topology:

- TCP connection is directly between the interacting NF services.

- For https scheme URIs, the TLS connection is direct between the interacting NF services.

- Detection of overload of the peer NF/NF service can be done based on the direct interaction with the peer.

- Any new overload conveyance mechanism need not be concerned about the issues related to presence of intermediaries (i.e proxies).

### 5.2.2 NF Services as Distributed Collection

In this topology, an NF / NF service instance can provide an API URI root which is used by the NF service consumers for initiating the creation of resources. The initial HTTP/2 messages from the NF service consumers for the creation of resources are handled by an origin server that is authoritative for this API URI, while the resource is created at a backend processing instance of the NF service instance. The authority part of the API may be changed due to this and the exact resource URI is returned in the Location header of the resource creation response. Subsequent interactions with the resource happen towards the resource URI returned in the Location header. This is illustrated in figure 5.3-1.



Figure 5.2.2-1: NF Services as Distributed Collection

This topology is specified in release 15 in TS 29.501 [5] and is currently used by the SMF services as specified in TS 29.502 [6]. The following are the key characteristics of this topology:

- HTTP/2 connection for the initial create request terminates at the initial origin server that handles the resource create requests.

- HTTP/2 connection for the subsequent interaction (e.g update resource methods) with the created resource terminates at the origin server where the resource was created.

- The authority part of the resource URI is changed during resource creation, and hence the initial create request and subsequent requests are handled by two different origin servers.

### 5.2.3 Interpretation of Overload

For the direct interaction topologies specified above, the following points apply for the interpretation of the detected overload condition.

- For the Release 15 overload detection mechanism, whether the detected overload is associated with the NF service instance or with the origin server is upto HTTP/2 client implementations.

- If the HTTP/2 client implementations associate the overload with an origin server, then for the model specified in clause 5.2.2, overload detection will happen independently for the initial origin server that handles resource create requests and the origin servers where the resources are actually created.

- Any new overload conveyance mechanism may consider associating the overload with a NF service instance instead of per origin server, if required.

## 5.3 Indirect Interaction Models through Intermediaries

### 5.3.1 NF Service Instances behind a Service Communication Proxy without delegated NF Service discovery

In this deployment model, a NF Service Consumer first discovers the API endpoint (service URI) of the NF Service Producer by querying the NRF. The NF Service Consumer then routes the request to the NF Service Producer through a Service Communication Proxy (SCP, e.g a HTTP forward and reverse proxy). SCPs may be deployed in a distributed manner in which case, a SCP on the side of the NF Service Consumer acts as a HTTP forward proxy while a SCP on the side of the NF Service Producer acts as a HTTP reverse proxy. The SCP routes the request to appropriate instance of the NF Service Producer. This topology is specified in Annex C of TR 23.742 [4] as model C. In this deployment model, the "authority" part of the URI used by the NF service consumer to access the NF service producer may be:

- the FQDN / IP address representing the NF service producer, if the NF service consumer does the NF service producer discovery and selection but is not aware of the number of backend instances behind the SCP in which case the SCP (acting as a HTTP reverse proxy) selects an instance; or

- the FQDN / IP address representing the NF service producer instance, if the NF service consumer does NF service producer instance discovery and selection.

### 5.3.2 NF Service Instances behind a Service Communication Proxy with delegated NF Service discovery

In this deployment model, a NF Service Consumer does not discover the API endpoint (service URI) of the NF Service Producer by itself. Instead, the NF Service Consumer routes the request to a SCP and the SCP based on the contents of the request URI and/or HTTP headers performs a NF service discovery either by configuration or optionally by querying the NRF that has been assigned for the set of slices or the slice instance corresponding to NF services involved in the indirect communication. SCPs may be deployed in a distributed manner in which case, a SCP on the side of the NF Service Consumer acts as a HTTP forward proxy while a SCP on the side of the NF Service Producer acts as a HTTP reverse proxy. An SCP in the communication path does the NF service producer instance selection. The SCP(s) then routes the request to a selected instance of the NF Service Producer. This topology is specified in Annex C of TR 23.742 [4] as model D. In this deployment model, the "authority" part of the URI used by the NF service consumer to access the NF service producer may be:

- for initial requests, the FQDN representing the NF type if a deployment delegates the discovery of NF service producer instance or NF service producer set and corresponding instance within the set to the SCP (see clause 6.3.1.x of 3GPP 23.501 [2]); or

- for subsequent requests, the FQDN representing the NF service producer set if a deployment delegates NF service provider instance selection within a set to the SCP (i.e. concept of a set is visible to the consumer and the NF service consumer delegates the NF service producer instance selection within the set to the SCP and there is no binding between a NF service consumer and a particular NF service producer instance); or

- the FQDN / IP address representing the NF service producer if a deployment delegates NF service producer instance selection to the SCP (i.e. NF service consumer is not aware of the number of backend instances behind the SCP in which case the SCP acting as a HTTP reverse proxy selects an instance, and there is no binding required between a NF service consumer and a particular NF service producer instance).

### 5.3.3 Interpretation of Overload

For the indirect interaction topologies specified above, the following points apply for the interpretation of the detected overload condition.

- For the Release 15 overload detection mechanism, whether the detected overload is associated with the NF service instance or the NF service instance set or the SCP depends on what the authority part of the request URI contains.

- Any new overload conveyance mechanism may consider associating the overload with a specific scope (e.g. at NF service instance level or NF service set level).

# 6 Overload Scenarios

## 6.1 Introduction

3GPP has done a detailed study of scenarios leading to overload in the core network elements in 3GPP Release 12, available in TR 23.843 [7]. Many of the scenarios are applicable to 5GC as well, while some scenarios like overload due to frequent IDLE-CONNECTED state transition can be mitigated in 5GC by using solutions like RRC-INACTIVE state. The following clauses provide a list of scenarios that may lead to overload situation in the 5GC service based interfaces.

## 6.2 Overload Caused due to UE Initiated Signalling

The following different UE initiated and/or UE induced signalling events can cause overload in the 5GC service based interfaces.

1. Large number of UEs performing mobility registration procedure. In densely populated areas, mass rapid transport systems transport large number of people from one location to another at high speeds. This will lead to a large number of UEs doing mobility registrations simultaneously after they move from one registration area to another. In the case of 5G this problem can be further aggravated since it is not only the smartphones that get transported along with the humans that travel, but there will be a lot of IoT devices like wearables, V2X UEs and sensors monitoring and transmitting active vehicle diagnostic information that will be generating signalling towards the core network due to rapid mobility.

2. Large number of UEs generating application signalling that induces signalling at the 3GPP 5GC elements

- An example scenario is large number of people watching a live sports event at a stadium and trying to catch the action replay videos, videos of player profiles, use of Augmented Reality (AR) glasses or headsets to get augmented information at real time related to the event and players. These scenarios can create or update or delete QoS flows which can lead to massive signalling in the 5GC service based interfaces.

- Another example is the application level heartbeat messages from large number of UEs in a region, causing frequent IDLE-CONNECTED state transitions leading to NG-RAN to AMF to SMF to UPF signalling. Though this can be mitigated in 5GC by use of RRC-INACTIVE state, such scenarios may still cause a core network overload if RRC-INACTIVE is not used.

3. Large number of UEs generating application layer broadcasts which results in signalling flood towards each member of that application layer broadcast domain. For example, when a large number of UEs is part of an Ethernet DN and they run IP above the Ethernet PDU session, ARP or IPv6 Neighbour Discovery can generate huge signalling if ARP proxying / IPv6 ND proxying are not implemented. As per TS 23.501 [2], clause 5.6.10.2

*NOTE 4: Entities on the LAN connected to the 5GS by the UE may have an IP address allocated by the DN but the IP layer is considered as an application layer which is not part of the Ethernet PDU Session.*

and

*Neither a MAC nor an IP address is allocated by the 5GC to the UE for a PDU Session.*

Since IP address allocation to entities on the LAN (including the UE which is part of the LAN) is considered as an application layer functionality, it is possible that 3GPP 5GC NFs like UPF and SMF have no visibility into the MAC address to IP address mapping and hence ARP / IPv6 ND proxying cannot be effectively implemented. In such cases, an ARP / IPv6 Neighbour Solicitation request may result in being broadcasted to every Ethernet entity (including UEs) that are part of the DN. If many UEs are in CM-IDLE state, this will result in N11 signalling for triggering N3 tunnel setup which would further lead to massive paging and subsequent service request leading to N11 communication again.

TS 23.501 [2], clause 5.6.10.2 specifies mechanisms to avoid overloads due to such broadcast / multicast traffic over Ethernet PDU sessions.

## 6.3 Overload Caused due to Failure and Restart of Network Functions

The following different scenarios related to failure and restart of network functions can cause overload in the 5GC service based interfaces.

1. Restart of a network function, indicated by its recovery timestamp, can lead to:

- re-creation of resources at the restarted NF, if the peer NFs decide to restore the resources; or

- it can lead to clean-up of resources, which would likely lead to the deregistration or release of PDU sessions towards a large number of UEs.

See TS 23.527 [8], clause 6. In all cases, restart of a network function can result in large number of signalling messages on the 3GPP 5GC service based interfaces within a short interval of time.

2. When the FTEID allocation is done by the UPF, upon restart of a V-UPF, then restoration of the PFCP sessions as specified in clause 4.3.4 of TS 23.527 [8], can result in large number of signalling messages on the N16 interface for Home Routed PDU sessions, if the V-UPF allocated different FTEIDs.

## 6.4 Overload Caused due to Notifications

The following different scenarios related to notifications can cause overload in the 5GC service based interfaces. Notifications may be at per UE level or per PDU session level or per UE group level or per NF level.

1. Some of the 5GC NFs support EventExposure service. This allows any authorized NF to subscribe for event notifications. Depending on the type of event, some events may occur very frequently for a large number of UEs. Examples are: Location change, UE presence in an area of interest report. This can lead to a lot of event notifications on the service based interfaces leading to overload.

2. NRF supports subscribing for NF status reports at an NF type level. Any change of NF profile of NFs of the subscribed type can lead to a notification. 3GPP Release 15 has specified mechanisms to filter such notifications based on monitored and unmonitored attributes of NF profile. When such filtering mechanisms are not used, then, if there are large number of instances of a particular NF type, this may result in frequent notifications leading to overload in the service based interfaces.

# 7 Key Requirements for Load Control

## 7.1 Introduction

This clause contains the key requirements to be addressed for load control of service based interfaces.

## 7.2 Key Requirement #1: Load Information Conveyance

As identified in clause 9.3.2, the release 15 load control mechanism has some limitations. Consequently solutions for conveying the load information of NF service producer to NF service consumers need to be identified.

This key issue shall study solutions for:

- Sub-Issue#1: Alternative mechanisms for conveying the load information of NF service producers to NF service consumers apart from the Rel-15 mechanism.

- Sub-Issue#2: Content of the conveyed load information.

- Sub-Issue#3: Frequency of inclusion / conveyance.

- Sub-Issue#4: Where to include the conveyed load information (e.g. HTTP headers, payload)

## 7.3 Key Requirement #2: Co-existence with Rel-15 Load Control

Even if new mechanisms for conveying load information is agreed in Rel-16 it is expected that it has to co-exist with existing Rel-15 deployments using the Rel-15 load conveyance and control mechanism.

This key issue shall study solutions for:

- Sub-issue#1: How an NF service producer discovers an NF service consumer's support for new Rel-16 load conveyance mechanism.

- Sub-issue#2: Recommendations for NF service producers on the use of Rel-15 mechanism to update the load information in NRF when there are NF service consumers supporting Rel-16 mechanism.

- Sub-issue#3: How an NF service consumer discovers an NF service producer's support for new Rel-16 load conveyance mechanism.

## 7.4 Key Requirement #3: Regulation of Load Information Reception and Conveyance in the NRF

As highlighted in clause 9.3.2, Rel-15 load control mechanism has some limitations with regards to the frequency of load information change notifications sent by the NRF to the subscribed NFs, which might lead in some scenarios to overload situations in these NFs. In a similar way, defining a highly frequent load information update by NFs in the NRF might also lead in some cases to overloading the NRF itself, e.g. due to the associated processing load. Therefore, this key requirement aims at addressing these two identified issues and opens for potential solutions.

This key issue shall hence study solutions for:

- Sub-Issue#1: Defining a mechanism that enables to control and manage the frequency of load information change from the NRF to the subscribing NFs.

- Sub-Issue#2: Defining a mechanism that enables to control and manage the frequency of load information updates by NFs in the NRF.

# 8 Key Requirements for Overload Control

## 8.1 Introduction

This clause contains the key requirements to be addressed for overload control of messages sent over the service based interfaces.

## 8.2 Key Requirement #1: Overload Conveyance

As identified in clause 9, the time to detect overload condition and the ease of overload situation at a NF service producer by a NF service consumer depends on the value of W and K. It is identified that low values of W would lead to false positives. Hence with larger values of W, the number of rejects taken to identify overload condition and/or the ease of overload condition increases. This requires alternate solutions to be studied for conveying the current overload situation at a NF service producer to a NF service consumer so that the NF service consumer can react immediately.

This key issue shall study solutions for the following:

- Sub-Issue#1: Solutions for conveying current overload situation at a NF service producer to a NF service consumer

- Sub-Issue#2: Information to be included in the conveyed overload information. Information used to associate the conveyed overload information with a particular scope (e.g. NF service level, NF service instance level, specific DNN and/or S-NSSAI level, upstream NF services like PCF, H-SMF etc.)

- Sub-Issue#3: Frequency of inclusion.

- Sub-Issue#4: Where to include the conveyed overload information (e.g. HTTP headers, payload)

## 8.3 Key Requirement #2: Mechanisms for avoiding and mitigating overload

Since the architecture of 5GC allows virtualized deployment of NF services, it is possible to avoid and/or mitigate overload situations by scaling the virtualized instances of the NF service in some deployments. In addition, the Release 16 eSBA architecture allows NF services to be deployed as a set allowing incoming requests to be routed to any NF service instance within a set. This mechanism also allows for horizontally scaling the number of instances within a set and avoid overload situations.

This key issue shall study the following:

- Mechanisms for avoiding overload scenarios at NF services;

- Mechanisms for mitigating overload situation when they arise at NF services, without impacting ongoing signalling and without any reactive behaviour from the NF Service Consumers (HTTP Clients);

- Deployment scenarios where such mechanisms are applicable.

## 8.4 Key Requirement #3: Overload Control

In Release-15, once a HTTP client learns that a HTTP Service is overloaded based on the "503 Service Unavailable" or "429 Too Many Requests" status codes as specified in clause 6.4 of TS 29.500 [3], the client applies an adaptive throttling mechanism as specified in Annex A of TS 29.500 [3].

Assuming that solutions for feeding back exact overload status of HTTP server (NF Service Producer) are identified in Release-16, this key issue shall study solutions for:

- How an NF Service Consumer (HTTP Client) reacts based on the information received in Overload Control Information.

- How intermediaries (e.g. V-SMF, SCP) react

# 9 Evaluation of Release 15 Load and Overload Control

## 9.1 Introduction

The release 15 overload control behaviour is based on the HTTP/2 client side doing adaptive throttling of requests based on the number of requests sent and the number requests accepted. This is specified in Annex A of TS 29.500 [3]. The following clauses analyze the behavior of release 15 mechanism.

Release 15 also supports NF / NF service selection based on the load information provided by NRF during NF / NF service discovery procedure. This clause also analyses the Release 15 load control mechanism.

## 9.2 R15 Overload Control

### 9.2.1 Behaviour

The Release 15 HTTP client determines that the HTTP server is overloaded by keeping track of the number of requests sent vs the number of requests accepted. In order to be meaningful in determining the overload condition without false positives, it is suggested to keep a window for tracking the number of requests sent vs number of requests accepted and apply the congestion detection algorithm only after the window is full.

NOTE: If the client side overload detection is applied as soon as the first reject is received from the server even before the window is full at the client side, it will result in showing up as number of accepts vs number of requests sent as minimal triggering a false state of congestion at server even though the reject received could be just one off.

The detection is based on the following formula as specified in Annex A of TS 29.500 [3].

https://landing.google.com/sre/book/images/equation/eqn-3.png

where K. determines how agressively the HTTP client detects and applies the adaptive throttling. Assuming the following:

- Client side signalling rate is X requests per T time units;

- Window size, W, as a multiple "N" of X (where N can be integer or fractional).

Table 9.2-1 provides a calculation of the number of rejects up to which the client can tolerate before detecting an overload condition at the server, for various values of K and N. The reject tolerance calculation formula is derived as follows:

- (requests - K \* accepts) > 0

- requests > (K \* accepts)

- requests > (K \* (requests - rejects))

- (requests/K) > (requests - rejects)

Therefore, rejects > (requests - (requests / K)

Using a window size of W, the reject tolerance point at which overload is detected can be written as:

- **CT** = Ceiling (W - W/K)

Table 9.2-1: Reject Tolerance before Detecting Overload by HTTP Client using Release 15 Overload Control Algorithm

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl.No** | **Value of K** | **Number of requests sent within a time interval "t" - X** | **Client side window size multiplier N.**  **Window size W = N times X** | **Number of rejects client can tolerate until detecting overload at the server, CT = Ceiling (W-W/K).** |
| 1 | 1.2 | 1500 | 1 (= 1500) | 250 |
| 2 | 1.2 | 1500 | 2 (=3000) | 500 |
| 3 | 1.3 | 1500 | 1 (= 1500) | 347 |
| 4 | 1.3 | 1500 | 2 (=3000) | 693 |
| 5 | 1.4 | 1500 | 1 (= 1500) | 429 |
| 6 | 1.4 | 1500 | 2 (=3000) | 858 |
| 7 | 1.5 | 1500 | 1 (= 1500) | 500 |
| 8 | 1.5 | 1500 | 2 (=3000) | 1000 |
| 9 | 1.6 | 1500 | 1 (= 1500) | 563 |
| 10 | 1.6 | 1500 | 2 (=3000) | 1125 |
| 11 | 1.8 | 1500 | 1 (= 1500) | 667 |
| 12 | 1.8 | 1500 | 2 (=3000) | 1334 |
| 13 | 2 | 1500 | 1 (= 1500) | 750 |
| 14 | 2 | 1500 | 2 (=3000) | 1500 |
| 15 | 1.5 | 1500 | 0.01 (=15) | 5 |

The following observations can be made from the table:

- The larger the window size W, the greater is the reject tolerance and hence greater is the time taken to detect overload;

- The larger the value of K, the greater is the reject tolerance and hence greater is the time taken to detect overload;

- Very small window sizes could lead to false positives;

- Once the overload is detected, the number of rejects should fall below the tolerance point **CT,** in order for the client to detect that overload condition has eased and not apply any client side throttling;

- Hence the time required to detect that overload condition has eased at the server depends on W and K.

Once a HTTP client detects that the number of rejects is below the tolerance point **CT,** the client stops throttling and allows all the request messages be sent to the server. However the server may still be in overload, albeit to a lesser degree, and will continue rejecting messages at a lesser rate.

### 9.2.2 Evaluation and Conclusion

The Rel-15 Overload Control Mechanism (Adaptive Throttling at the client side) as specified in clause 6.4 and Annex A of TS 29.500 [3] has the following deficiencies:

a) Latency for converging to optimal output traffic the sever can handle, when overload increases or alleviates

Clients have to send a certain number of requests before detecting that the server is overloaded, which causes additional load on the server.

With a bigger K value throttling is more permissive resulting in increased latency to reduce traffic when overload increases but faster convergence when overload alleviates, and vice versa.

When the client throttling rate reaches a high value, latency may happen for client's traffic to increase when overload alleviates.

b) Client parameters may result in too aggressive or too permissive behaviour towards the overloaded server.

c) The mechanism is less optimal for bursty traffic and rapid increase of traffic. When highly overloaded the server cannot request clients to stop sending requests. The mechanism does not enable server feedback for rapid adaption when traffic level increase.

d) Clients have to keep track of traffic history (requests accepted/rejected) per API.

e) Different clients can behave very differently, depending on the algorithm implemented, parameters K and window size, and frequency of requests sent. This can result in unfair server resource usage by clients.

f) Adoptive throttling parameters (e.g. K, window size) need to be configured in all clients.

## 9.3 R15 Load Control Analysis

### 9.3.1 Analysis

Release 15 supports the following mechanisms for controlling the load on an NF / NF service:

- Register an NF or NF service load percentage (0-100) and capacity with NRF (see clause 6.1.6.2.2 and 6.1.6.2.3 of TS 29.510 [9]).

- Update the dynamic load and capacity of an NF or NF service with NRF (see clause 5.2.2.3.2 of TS 29.510 [9]) using heartbeat messages or dedicated update requests.

- Notification of change of dynamic load of an NF or NF service to the consumers that are subscribed for that notification (see clause 6.1.6.2.43 of TS 29.510 [9]).

- NF / NF service instance selection based on the load information provided by NRF during NF / NF service instance discovery (see clause 6.3.2 of TS 23.501 [2]).

The following are the characteristics of the Release 15 load control mechanism:

- Dynamic load information of NFs / NF services are discovered via NRF.

- Any subsequent updates to the already discovered load information requires subscription to NRF for notification about NF status.

- For home routed roaming scenarios, NF service consumers (e.g. AMF) may need to subscribe for notifications about NF / NF service status from HPLMN (e.g. H-SMF).

- The NRF can take dynamic load and capacity of NFs / NF services into account when processing NF discovery requests.

### 9.3.2 Evaluation and Conclusion

The following are the issues identified with Rel-15 load control mechanism:

- Learning of the current load of a NF service producer is based on notifications from NRF. This could potentially lead to a massive number of notifications in a network if all the NF service consumers subscribe for notifications about the current load of NF service producers as the load will keep changing dynamically.

- The periodicity of notifications depends on the periodicity of the heartbeat mechanism between NF service consumer and the NRF. In order to avoid stale load information in the NRF, an NF service producer has to update its NF status in the NRF more frequently using this heartbeat mechanism, thus leading to frequent large number of notifications from the NRF to all NF service consumers that have subscribed to this service producer's NF status change notifications.

- If NF service consumers wish to avoid frequent notifications by not subscribing to the NF status notifications from the NRF, then they will not be aware of the NF service producer's actual load at a given point of time. This would lead to NF service consumers incorrectly selecting an NF service producer instance that is loaded more than another lightly loaded instance.

- NRF becomes the single point of contact for learning load information of NF service producers and hence has the risk of becoming a single point of failure. High availability of the NRF is hence required.

- depends on the NRF supporting this Rel-15 load control mechanism.

# 10 Solutions for Load Control

## 10.1 Introduction

## 10.2 Solution #1: Direct Load Information Conveyance

This solution addresses key requirement#1, sub-issues #1 and #2 specified in clause 7.2.

While in the existing Rel-15 solution load and capacity information is conveyed between NFs indirectly via the NRF, it is proposed for Rel-16 as an option (making use of the supported feature mechanism) to allow in addition to the existing Rel-15 solution also direct signalling of Load Control Information (LCI) between NFs piggybacked on existing messages.

Information conveyed directly between NFs shall contain the identity of the LCI reporting NF/ NF service (i.e. the scope of the load information), a static capacity value, a load percentage value indicating the percentage of the static capacity that is actually not available for new requests, a timestamp or sequence number used to identify stale information.

The conveyed LCI shall be used for load balancing among a set of candidate NF / NF service instances.

Direct LCI conveyance piggybacked on existing messages allows clients to detect a server's current load

- without the need for potentially massive notification traffic from the NRF; clients can avoid massive notification traffic by unsubscribing from notifications, but then they are not aware of the server's actual load.

- in a timely manner independent from heartbeat frequency;

- at a finer granularity;

- independent from NRF deployments;

- frequently, as the reporting frequency naturally increases with the NF load;

- without relying on the NRF (single point of failure) that may fail and lose information after restart.

## 10.3 Solution #2: Load Control Co-Existence

This solution addresses key requirement#2 specified in clause 7.3.

It is proposed that the Supported Features mechanism specified in Rel-15 be used to discover the support of Rel-16 load control mechanism. A NF service producer may convey the load information to a NF service consumer, using a Rel-16 mechanism only if the NF service consumer has included in the Supported Features that it supports Rel-16 load conveyance / control mechanism.

A Rel-16 NF service producer is still required to update its load information periodically to the NRF since the load information is returned in NF discovery response and based on that a NF service consumer performs the selection of NF service producer instances.

However to avoid frequent notifications of the load from NRF to NF service consumers, it shall be possible, based on operator policy, to configure the NRF to not notify changes in the "load" parameter and subscriptions for NF status notifications be restricted only to changes in the other parameters (e.g. not allow subscriptions for "/load" attribute "monitoredAttributes" parameter, see clause 6.1.6.2.43 of TS 29.510 [9]) of the NF / NF service profile.

Editor's Note: It is FFS how and when to restrict the sending of "load" parameter (e.g. not allow subscriptions for "/load" attribute "monitoredAttributes" parameter).

It is recommended to apply such restrictions in the NRF as given below:

- For intra PLMN cases, apply only if it is known that all NF service consumers in the PLMN support Rel-16 load conveyance / control mechanism.

- For inter PLMN cases, apply based on operator policy.

## 10.4 Solution #3: Regulation of Load Information Change Notifications Conveyance by the NRF

This solution addresses key requirement#3, sub-issue#1 specified in clause 7.4.

As clearly highlighted in clause 9.3.2 and clause 7.4, it might be possible that the frequency of load information change notifications increases for some reason (e.g. a NF producer load increases rapidly in a short timeframe) in such a way that leads to overload situations in the subscribed NFs.

This solution proposes thus to define a threshold-based mechanism to enable a NF (e.g. NF consumer) to regulate the frequency of change notifications of load information of another NF (NF consumer) it has subscribed to in order to protect itself from getting to overload situations due to a potential rapid increase of the traffic associated to these notifications. The operation of this mechanism is described hereinafter:

- When subscribing with the NRF to notifications of changes in load information of another NF, a NF shall be able to specify a set of conditions on the frequency of reception of these notifications;

- This set of condition can consist on:

- The definition of one or several thresholds, e.g. the subscribing NF requests the NRF to send notifications only when certain specific load level(s) is/are reached and/or when the load level has increased/decreased by a certain amount, and/or;

- The specification of some periodicity parameters and/or timers, e.g. the subscribing NF requests the NRF to send notifications with respect to a certain periodicity and not systematically each time there is a change in the load information of the concerned NF, and/or;

- Any other relevant parameter, e.g. other conditions defined by the operator;

- The mechanism used to manage the logical combinations of the above mentioned conditions that control the triggering of notifications is operator specific, e.g. notifications are sent by the NRF if one of the conditions is satisfied;

- NFs shall be able to modify these conditions settings anytime in the NRF;

- If no conditions are present in the subscription to notifications request, the NRF shall behave as in Rel-15 and hence send notifications each time there is a change of load information in the concerned NF.

## 10.5 Solution #4: Regulation of Load Information Updates Reception at the NRF

This solution addresses key requirement#3, sub-issues#2 specified in clause 7.4.

As clearly highlighted in clause 9.3.2 and clause 7.4, it might be possible that the frequency by which NFs update their load information in the NRF becomes high for some reason (e.g. a configuration error or NFs load increases rapidly in a short timeframe) in such a way that leads to overload situations in the NRF, e.g. due to the necessary associated processing load.

This solution proposes thus to define a threshold-based mechanism to enable a NRF to regulate the frequency of load information updates it receives from a NF so as to protect itself from getting to overload situations due to a potential rapid increase of the traffic related to these updates. The operation of this mechanism is described hereinafter:

- The NRF shall be able to configure and indicate to a NF a set of conditions that have to be respected for the communication of load information updates;

- This set of condition can consist on:

- The definition of one or several thresholds, e.g. the NRF requests the NF to send load information updates only when certain specific load level(s) is/are reached and/or when the load level has increased/decreased by a certain amount, and/or;

- The specification of some periodicity parameters and/or timers, e.g. the NRF requests the NF to send load information updates with respect to a certain periodicity and not systematically each time there is a change in the load information of the concerned NF, and/or;

- Any other relevant parameter, e.g. other conditions defined by the operator;

- The mechanism used to manage the logical combinations of the above mentioned conditions that control the triggering of load information updates is operator specific, e.g. load information updates are sent if one of the conditions is satisfied;

- If no conditions are set by the NRF, the NF shall behave as it is currently described in Rel-15.

## 10.6 Evaluation and Conclusion

As documented in clause 9.3.2, the release 15 load control mechanism has some limitations. This clause proposes an evaluation of the proposed solutions for Rel-16 load control improvements. The following recommendations are hence made as a conclusion for this study:

- It is recommended that the solution for key issue #1 described in clause 10.2, i.e. to enable direct signalling of Load Control Information (LCI) piggybacked on existing signalling messages between NFs, will be used as a basis for further normative work.

- To avoid frequent notifications of the load from NRF to NF service consumers (key Issue #2), the following will be addressed in the normative work:

- the NF consumer supporting Rel-16 load control mechanism should unsubscribe from notifications of load information updates of a NF producer with the NRF as soon as it discovers the support of the same Rel-16 load control mechanism by this NF producer.

- the NRF and 5G NFs may also be configured, e.g. based on exchanged conditions for regulating notifications and/or on operator policy, and/or may consider the service consumer requested regulations of Load Information Change Notifications e.g.as documented in solution 10.x.1, in order to reduce regulate the frequency of traffic related to notifications of changes in the "load" parameter.

Editor's note: The above bullet for reducing/regulating load change updates and notifications is ffs.

# 11 Solutions for Overload Control

## 11.1 Introduction

This clause identifies the solutions for the key requirements identified in clause 8 for overload control. Each solution specifies the key issues and the specific aspects of those key issues it addresses.

## 11.2 Principles of Overload Control

Reactive overload control based on HTTP reject status codes by HTTP clients is already specified in 3GPP release 15 in TS 29.500 [3]. The high level principles of new overload control mechanisms to be adopted in Release 16 are summarized below:

a) Any new overload conveyance and overload control mechanism shall co-exist with release 15 mechanism;

b) Conveyance of overload control information by any means is optional;

c) If a receiving NF / NF service consumer instance does not understand the conveyed overload control information it shall ignore it. The overload control adopted by such receiving NF / NF service instances shall be the release 15 mechanism;

d) An NF / NF service producer (instance) compliant to R16 overload control mechanism, shall also return the specific HTTP status codes specified in TS 29.500 [3], clause 6.4 when the overload condition has not eased even after conveying overload control information (i.e the sending NF / NF service (instance) assumes that the receiving NF / NF service instance only supports release 15 overload control mechanism);

e) It shall be possible to associate the conveyed overload control information with a specific scope. The scopes shall be identified as part of the study;

f) The overload control mechanism shall allow preferential treatment of priority users (eMPS) and emergency services;

g) It shall be possible for an NF / NF service consumer to know the overload status of an NF / NF service producer even if they are not directly interacting (i.e interacting via an intermediary). Example use cases are AMF sending a PDU session establishment request to H-SMF via a V-SMF and SMF communicating with AMF via a I-SMF.

Editor's Note: Other principles, if any, are FFS.

## 11.3 Solution #1: Conveying Overload Information via Signalling

This solution addresses the subissue#1 in Key Issue#1 specified in clause 8.2

- Solutions for conveying current overload situation at a NF service producer to a NF service consumer

In this solution, the following principles are followed

- The NF Service Producer shall be able to provide the current overload situation via an Overload Control Information (OCI) IE in existing signaling messages;

- Generation and inclusion of Overload Control Information is implementation specific.

- The contents of the Overload Control Information are not addressed in this solution. They shall be addressed through solutions for key issue#1 subissue#2, specified in clause 8.2;

- The frequency of inclusion of the Overload Control Information is not addressed in this solution. It shall be addressed through solutions for key issue#1 subissue#3, specified in clause 8.2.

- How a NF Service Consumer that receives the OCI IE reacts is not addressed in this solution. It shall be addressed through solutions for key issue#3. Consequently aspects related to how intermediaries like HTTP proxies or SCP react to OCI shall also be addressed through solutions for key issue#3.

The overload control information of the NF service producer is conveyed to the NF service consumer as follows:



Figure 11.3-1 Conveying Overload Information via Signalling

1. The NF Service Producer determines that it is overloaded.

2. The NF Service Producer receives a HTTP request from the NF Service Consumer for any NF service that the NF Service Producer exposes.

3. The NF Service Producer includes an Overload Control Information (OCI) in the HTTP response. For deployments where the NF Service Consumer Instance and NF Service Producer Instance interact directly, the OCI information is provided by the NF Service Producer Instance. For deployments where the NF Service Producer instance is deployed behind a SCP and the SCP exposes the API endpoints on behalf of the NF Service Producer, the SCP shall include the OCI information after considering the overall overload situation of all the instances of the NF Service Producer that the SCP is acting on behalf.

4. The NF Service Consumer receives the OCI information and starts applying the overload control procedures specified as solutions for keyissue#3 in clause 8.x.

5. The NF Service Producer may piggyback OCI information in the notification messages (i.e., existing notification messages).

6. The NF Service Consumer responds to the notification.

7. If the NF Service Consumer received OCI information in a notification, it shall apply the overload control mechanisms for the subsequent request message it is going to generate towards the NF Service Producer.

## 11.4 Solution #2: Explicit Overload Control Information Conveyance

This solution addresses Key-Requirement #1 and proposes to convey Overload Control Information (OCI) explicitly between NFs (addressing Sub-Issue#1) within 3GPP custom headers (addressing Sub-Issue#4). The OCI may be piggybacked on existing HTTP response signalling in order not to add additional signalling load to the involved NFs.

Editor's note: OCI handling at the SCP is ffs.

Addressing Sub-Issue#2, the information conveyed within the OCI shall contain the identity of the overloaded NF/ NF service, a requested traffic reduction or limitation, a validity period and a sequence number used to identify stale OCIs.

This solution addresses Key-Requirement #3 as follows: NF service consumers shall act on received OCIs by reducing the number of API invocations according to the requested reduction or limitation metric value during the validity period, taking into account message priority considerations.

The traffic reduction (throttling) shall rely on the Loss Algorithm with similar principles as have been defined for Diameter and GTP-C overload control.

Support of solution #2 is proposed to be optional in Rel-16 and its use may be based on deployment/operator policy. The feature negotiation mechanism shall be used to detect/advertise support of solution #2 by NFs/NF services.

The proposed overload control mechanism based on explicit OCI conveyance achieves the following advantages:

a) Clients have the accurate server overload information to determine how to throttle traffic resulting in immediate convergence of output traffic to the amount of traffic the overloaded server allows.

b) Clients do not need to send requests that are rejected to detect the server's overload.

c) Clients do not need to keep track of traffic history.

d) Burst traffic is well addressed with low latency.

e) Homogenous client behaviour results in fair usage of the overloaded server's resources.

f) In high overload situations the overloaded server can request clients to stop traffic.

## 11.5 Solution #3: Overload Control Information

The following information shall be considered for inclusion in the Overload Control Information

Table 11.5-1: Overload Control Information

|  |  |  |  |
| --- | --- | --- | --- |
| Sl.No | Information | Type | Purpose |
| 1 | Overload-Reduction-Metric | enumerated | Contains the reduction metric to be applied for the experienced overload situation. This Overload-Reduction-Metric can either be a percentage of traffic reduction (e.g. loss algorithm), a specific message traffic rate (e.g. rate algorithm), or any other relevant reduction or limitation metric type that the sender of the overload control information requests the receiver to apply.  The computation of the exact value for this parameter is left as an implementation choice at the sending node.  This IE is mandatory within the Overload-Control-Information. |
| 2 | Period-of-Validity | DateTime | The Period-Of-Validity indicates the time upto which, the overload condition specified by the Overload Control Information IE is to be considered valid (unless overridden by a subsequent Overload Control Information IE for the same scope).  This IE is mandatory within the Overload-Control-Information. |
| 3 | Overload-Sequence-Number | integer | Due to availability of multiple paths / multiple TCP connections to reach a API endpoint or to deliver notifications, it is possible that an overload control information received at once instance may be less recent than an information received at a previous instance in another path. The Overload-Sequence-Number aids in sequencing the overload information received from an overloaded NF service.  The Overload-Sequence-Number contains a value that indicates the sequence number associated with the Overload Control Information IE. This sequence number is used to differentiate two Overload Control Information IEs generated at two different instants by the same NF service for the same scope.  The sender of this information shall increment the Overload-Sequence-Number associated to a particular overload scope whenever modifying some information in the Overload-Control-Information IE. The Overload-Sequence-Number shall not be incremented otherwise.  This IE is mandatory within the Overload-Control-Information. |
| 4 | Scope | OCIScope | The conveyed OCI information is associated with the HTTP signalling that is in the same scope. |

The high level content of OCIScope is shown in table 11.5-2 below.

Table 11.5-2: OCIScope content

|  |  |  |
| --- | --- | --- |
| Number | Information | Description |
| 1 | Uri | This identifies the API URI. The NF service consumer uses it to identify whether the OCI is associated with the NF it is directly communicating with (e.g. V-SMF) or with the NF it is communicating with via an intermediary (e.g. H-SMF). |
| 2 | NfInstanceId | The OCI is associated with a specific NF instance. |
| 3 | nfServiceInstanceIdList | The OCI is associated with a specific list of NF service instances. |
| 4 | Dnn list) | The OCI is associated with a specific list of DNNs. |
| 5 | sNssaiSmfInfoList | List of S-NSSAIs See 1 29.510 [9] clause 6.1.6.2.29 |
| 6 | NF Set ID | To be updated during normative phase depending on the data type chosen for NF Set ID in eSBA work. |
| NOTE: The more precise and detailed the information provided in the OCI Scope, the more precise the parts of the concerned NF for which overload control should be applied. | | |

When scope information is not included, the overload control information conveyed shall be associated with the "authority" part of the NF service API.

## 11.6 Solution #4: Overload Avoidance and Mitigation

This solution addresses keyissue#2. In deployments where NF services are deployed as instances (virtual) behind an API endpoint that terminates the API URI (e.g. SCP instance), the following mechanisms can be used to avoid and mitigate overload situations:

NOTE: For ease of understanding the entity that terminates the NF Service API endpoint is called a "Reverse Proxy" here onwards.



Figure 11.6-1: Overload Avoidance and Mitigation via OAM and Service Instance Scaling

1. OAM or a monitoring entity can monitor the resource usage (e.g. CPU, memory, link utilization etc.,.) of the NF Service instances.

2. If the monitored resource usage is nearing a configured threshold, the OAM system can spawn off new instances of the NF service, with the same API endpoint as the rest of the instances (e.g. API endpoint terminating at the Reverse Proxy).

3. The Reverse Proxy is made aware of the availability of new NF Service instance(s).

4. The Reverse Proxy routes subsequent new resource creation requests to lightly loaded NF Service instances.

5. When the overload situation eases, the OAM may shutdown some of the NF Service Instances, if the resource state in those NF Service instances are shared with other NF Service instances through UDSF or if the session contexts can be transferred by R16 eSBA solutions.

6. In order to avoid signaling overload of the Reverse Proxy itself, the following deployment options may be considered:

- The Reverse Proxy is deployed with multiple interfaces / IP links all having the same anycast IP address;

- The Reverse Proxy has a large CPU pool and memory to handle signalling surges, to handle those signalling requests in parallel and to distribute them to backend NF Service instances

7. If the frequency and the quantum of signalling messages crosses beyond the capacity upto which the Reverse Proxy can handle, scaling of service instances cant help avoid / mitigate the overload. Either Release-15 based Overload control or new overload control mechanisms are required.

## 11.7 Solution #5: Frequency of Inclusion - Piggyback in every HTTP signalling message

This solution addresses key requirement#1, sub-issue#3 specified in clause 8.2.

The following are the key principles of this solution:

- NF service producer monitors its overload situation and identifies when to start communicating the overload control information to NF service consumers, depending on configured thresholds or implementation specific triggers;

- Once the NF service producer decides to include OCI towards NF service consumer, it shall piggyback it in every HTTP response message and may also piggyback it in every HTTP notification message that the NF service producer would normally send towards the NF service consumer (for the already subscribed notifications).

- The OCI information contains a sequence number. The sequence number shall be incremented whenever the overload reduction percentage value changes from a previously communicated value.

- OCI IEs carrying the same sequence number shall have the same content.

- If an NF service consumer has already received an OCI with sequence number#n it shall ignore the subsequent OCI IEs that it receives with the same or lower sequence number#n (except when there is rollover of sequence number).

NOTE: How to avoid that a receiver uses only actual and not out of date OCI information is left to the normative work,

**Advantages**:

- NF service producer implementation is simple as it can include the OCI information in every message without taking any additional considerations.

- Resiliency against message losses. The NF service consumer will know the current overload situation of the NF service producer even if some of the messages to NF service consumer are lost in the network.

- Since the OCI IE is included in HTTP response messages and in the notification messages for which the NF service consumer subscribed, an NF service consumer that has indicated support for Rel-16 overload control will be able to process the OCI IE.

**Disadvantages:**

- Since OCI IE is included in every message, an unnecessary increase in the size of every message, especially considering that except the first occurrence of OCI IE with the sequence number#n, all other occurrences of the same or lower sequence number are ignored by the NF service consumer.

## 11.8 Solution #6: Conveying Overload Information via NRF

This solution addresses the subissue#1 in Key Issue#1 specified in clause 8.2

- Solutions for conveying current overload situation at a NF service producer to a NF service consumer

In this solution, the following principles are followed

- The NFProfile in NRF shall be able to store the Overload Control Information related to an NF Service Producer;

- The contents of the Overload Control Information are not addressed in this solution. They shall be addressed through solutions for key issue#1 subissue#1, specified in clause 8.2;

- It shall be possible to get the Overload Control Information of an NF Service Producer through NF discovery and NF status notification procedures;

The overload control information of the NF service producer is conveyed to the NF service consumer as follows:



Figure 11.8-1 Conveying Overload Information via NRF

1. NF Service Consumers may use the Nnrf\_NFManagement\_NFStatusSubscribe service operation to subscribe for notifications on the change of overload control information of a NF Service Producer. The NotifCondition attribute as specified in clause 6.1.6.2.43 may be used to set the monitoredAttributes IE to the Overload Control Information IE.

2. The NRF creates the resource for the NF status subscription and responds to the NF Service Consumer.

3. When the NF Service Producer considers itself overloaded, it updates the Overload Control Information in the NFProfile in NRF using the Nnrf\_NFManagement\_NFUpdate service operation. The frequency of inclusion is based on operator policy.

4. NRF responds to the NF Service Producer.

5. [Optional] If the NF Service Consumer has subscribed for NF Status notification in step 1, the NRF notifies the NF Service Consumer providing the Overload Control Information of the NF Service Producer.

6. [Optional] The NF Service Consumer responds to the notification.

7. [Conditional] If the NF Service Consumer has received the notification in step 6 and if the NF Service Consumer has active HTTP/2 signalling interaction with the NF Service Producer, it shall use the Overload Control Information received to apply overload control behaviors.

For scenarios, where the NF service consumer has been already instructed to use a specific instance of an NF service producer but the NF service consumer has to discover the API endpoint of the AMF service:

- If the NF Service Consumer gets NF profile of an NF Service Producer (e.g. AMF) before it communicates with the NF Service Producer and if the NF profile has Overload Control Information, then the NF Service Consumer uses the Overload Control Information to apply overload control behaviors.

8. [Conditional] If a NF Service Consumer is discovering the NF Service Producer, it requests the NRF with Nnrf\_NFDiscovery\_NFDiscover service operation.

9. [Conditional] The NRF responds to the NF Service Consumer providing the NFProfile of the NF being discovered. The NFProfile shall contain the Overload Control Information, with the information updated in step 3.

10. The NF Service Consumer shall consider the Overload Control Information for the NF Service Producer instance selection. If the NF Service Producer is experiencing signalling overload, it is beneficial if new resource creations, which may lead to further increased signalling later, are not sent to the overloaded NF Service Producer instance.

In deployments using the Release-16 eSBA mechanism with delegated discovery, the role of NF Service Consumer in this solution is played by the SCP.

## 11.9 Solution #7: Converying OCI via Overload Control Service

This solution addresses Key Issue#1 as documented in clause 8.2.

This solution includes the following aspects:

- Introduce a new NF service, Overload Control Service, which shall compliant to the requirement as specified in clause 7.1 "Network Function Service Framework" in 3GPP TS 23.501 [2], i.e. the Overload Control service shall be registered to the NRF, and discoverable and the service shall only be authorized to be used by the eligible NFs.

- The Overload Control Service may be supported by any existing 3GPP NFs. e.g. for AMF, the Overload Control Service may be supported in addition to the existing NF services, i.e. Namf\_Communication, Namf\_Eventexposure, Namf\_MT and Namf\_Location services.

- The Overload Control Service may provision an OCI on NF level and optionally on the NF Service level.

- The Service Operations supported by the Overload Control Service includes Subscribe, Unsubscribe and Notify.

- The Overload Control Service may support Subscribe-Notify communication pattern.

- Any 3GPP NF entity (e.g. an AMF, a SCP), and/or any network entity (e.g. OAM or a 3rd party service), regardless whether it is a service consumer of another service offered by an overloaded NF service producer, as long as being authorized, may subscribe to receive Overload Control Information notifications from the overloaded NF service producer.

- A NF A, if it is consuming an NF service (other than Overload Control Service) offered by a NF B, shall subscribe to the Overload Control Service if it is offered by the NF B, to receive OCI notification for NF B and/or that NF service. The NF A shall be considered as the Overload Enforcement node, i.e. shall take actions, e.g. throttling signaling traffic sent from NF A to NF B, to mitigate the Overload situation in the NF B if the NF B is overloaded.

- When subscribing to receive OCI notification, the consumer may request to be notified with certain conditions, e.g. frequency of notifications, though the Overload Control Service producer takes final decision.

- The OCI information is sent via Notifications towards all subscribed entities and included in the message body.

The following figure illustrates the population of OCI via a new NF service - Overload Control Service.



Figure 11.9-1 Conveying OCI via Overload Control Service

1. The NF B (as Service Producer) register offered services including the Overload Control Service on the NRF.

2. The NF A (as Service Consumer) perform a Service Discovery procedure, and the NF B is selected.

3. The NF A consume a service (other than the Overload Control service) offered by the NF B.

4. The NF A consume the Overload Control Service offered by NF B by subscribing to get OCI notifications.

5. If the NF B or a service on NF B is overloaded, and according to the subscription.

6. The NF B sends a notification to NF A includes its OCI.

7. The NF B sends further notification if the OCI information is changed.

## 11.10 Solution #8: Overload Control enforcement behaviour

This solution addresses key requirement#3 specified in clause 8.4.

The following are the key aspects of this solution to depict an NF service consumer behavior during overload situations:

- The NF service consumer shall reduce the number of messages it would normally generate towards the NF service producer, in accordance with the overload reduction percentage requested by the NF service producer in the OCI IE. For example, if the NF service producer has requested an overload reduction percentage of 20%, then out of 100 messages that the NF service consumer generates towards the NF service producer it shall drop 20 messages. This mechanism is called message throttling.

- Since the OCI IE is communicated end to end, the enforcement of the overload control based on overload reduction percentage shall be applied by the NF service consumer. SCPs are not expected to react on OCI IE in the scope of this solution.

- However SCPs can help mitigate overload as specified in solution#4 in clause 11.6.

- When applying message throttling, the NF service consumer shall give preference to messages related to priority users (eMPS) and emergency services and shall avoid throttling the messages related to such users / services.

- A prioritization for not throttling messages based on the system procedure the message is related to and parameters in the message (e.g. DNN, S-NSSAI) shall be employed.

- The following table provides a suggested prioritization of system procedures. Messages related to system procedures of higher priority (lower number indicates higher priority) shall be given lower preference for throttling. This priority mechanism should be used in combination with the existing 3gpp-Sbi-Message-Priority (SMP) mechanism, see TS 29.500 [3], clause 6.8.

Table 11.10-1: Priority of System Procedures

|  |  |  |  |
| --- | --- | --- | --- |
| Priority | System Procedure | Related Service Operations | Description |
| 1 | Service Request and Handover | Nsmf\_PDUSession\_UpdateSMContext,request / response  Nsmf\_PDUSession\_Update request request / response  Nsmf\_PDUSession\_CreateSMContext request / response and Nsmf\_PDUSession\_Create request/response for EPS to 5GS Handover Procedure | Service request and handover procedures are expected to finish under tight time deadlines (less than 20 ms) and hence messages related to those |
| 2 | PDU session release, Any procedure causing a DELETE on a HTTP resource | Any service operation using DELETE method | Resource deletions would lead to lesser number of resource related signalling and hence should be treated on priority to ease the overload situation. |
| 3 | Other procedures that don't fall under category 1, 2 and 4 | Any service operation not falling under category 1,2 and 4 (e.g. Namf\_Communication service operations, Namf\_EventExposure service operations, Nudm\_UECM service operations etc,.,) |  |
| 4 | Any procedure causing new resource creation (e.g. PDU session establishment) | Any service operation using POST or PUT method to create a resource | Resource creations will result in more signalling in future on those resources leading to overload. Hence these messages should be throttled more. |

## 11.11 Solution #9: Location of OCI

This solution addresses key requirement#1, sub-issue#4 specified in clause 8.2.

There are two possible locations within a HTTP message to include the Overload Control Information (OCI).

1. Include OCI as a JSON object within the payload of HTTP messages.

- Provides flexibility to expand OCI in future releases as payloads can be larger when compared to HTTP headers.

2. Include OCI encoded within a custom HTTP header.

- Can be used on all messages (including responses carrying 204 No Content status code).

- Does not require OpenAPI updates every time a new API / method to an existing API is added.

## 11.12 Solution#10: Overload control by redirecting requests to another NF within a NF set

This solution addresses the Key Issue#3 specified in clause 8.4

In this solution, the following principles are followed

- The OCI includes an indication on whether the NF service Consumer can select another NF service producer from the same NF Set, if the scope of the OCI is at NF service producer level (and not at NF set level)

- If it is allowed, the NF service Consumer selects a new NF service producer from the same NF set, and send the request message to the new NF service producer instance.

- The newly selected NF service producer instance is expected to have the resource context information related to the request message. How the selected NF service producer instance gets it from the previous NF service producer instance will be addressed as part of eSBA solutions for context transfers within a NF set.

- This solution does not address SCP with delegated discovery deployment scenario.

The overload control is handled as follows:



Figure 11.12-1 Overload control by redirecting requests to another NF within a NF set

1a. [Optional] NF Service Consumers has received a response message which includes OCI reflecting the overload of the NF Service Producer.

1b. [Optional] NF Service Consumer has received Nnrf\_NFDiscovery Response or Nnrf\_NFManagement\_NFStatusNotify, in which OCI is included, refelecting the overload of the NF Service Producer.

2. [Optional] If the OCI reflects the scope at NF instance Id level or NF service instance Id level and also includes an indication saying that a new NF Service Producer from the same NF Set can be selected to replace the original NF Service Producer, the NF Service Consumer triggers Nnrf\_NFDiscovery\_NFDiscover Request (NF Set ID) to discover a new NF Service Producer from the NF Set.

3. [Optional] The NRF responds with Nnrf\_NFDiscovery\_NFDiscover Response. The Message includes a list of NF profiles of NF Instance that have no OCI in their profile from the requested NF Set.

4. The NF Service Consumer then selects one NF Service Producer from the list of NF Instances provided by NRF, and sends the request to this new NF Service Producer Instance.

The NF Service Consumer replaces the authority part of the resource URI of the original request message with the FQDN or IP address of the new NF Service Producer Instance.

## 11.13 Solution #11: Overload Control Enforcement by SCPs

This solution addresses key requirement#3 specified in clause 8.4.

The main characteristic of this solution is that the SCP is considered here as the overload control enforcement entity. SCP is considered in this solution as deployed in either option C (Indirect communication without delegated discovery) or option D (Indirect communication with delegated discovery) communication models defined in annex E of TS 23.501 [2]. In this sense, the SCP is terminating the TLS connection on both sides of the communication path.

- At reception of OCI information from a NF producer, the SCP shall reduce the number of messages it would normally send towards the NF service producer, in accordance with the overload reduction percentage requested by the NF service producer in the OCI IE;

- When performing this throttling mechanism, the behaviour of the SCP with regards to the concerned NF service consumers is to be specified by this solution, e.g. an appropriate error message could be sent to the concerned NF consumer(s) with an error code and error details that are in accordance with the overload situation that is experienced;

- The SCP shall also be able to select another NF service producer, e.g. from the same NF Set, if it is allowed, and send the request message to the new NF service producer instance. This solution shall also specify how and when the SCP performs this reselection;

- Depending on the ongoing overload situation and the efficiency of the overload control mechanisms that are applied by the SCP and not involving upstream NFs, the SCP shall also be able to send (e.g. derived or simply relayed) upstream OCI information to the concerned NF consumer(s) if it considers it beneficial for the overall management of the overload situation;

- For example:

- If the NF service producer has requested an overload reduction percentage of 20%, then out of 100 messages that the SCP has to route to this NF service producer, only 80 messages would be maintained and the remaining 20 messages would if possible rather be routed to another NF service Producer from the same NF Set. If no NF service producer could be identified / is available e.g. from the same NF set, then the SCP shall drop the extra messages in accordance with the overload reduction percentage requested by the NF service producer in the OCI IE;

- The SCP shall apply a proper message prioritization mechanism when applying throttling, e.g. the same message prioritization mechanisms described in solution#8 in clause 11.10.

## 11.14 Conveyance of Overload Information with hyperlink

This solution addresses the subissue#1 and subissue#4 in Key Issue#1 specified in clause 8.2

- Solutions for conveying current overload situation at a NF service producer to a NF service consumer.

- Where to include the conveyed overload information (e.g. HTTP headers, payload).

In this solution, the following principles are followed

- The NF service producer stores the overload control information (when necessary) in a dedicated storage, the storage may either collocate with the NF service producer or be standalone storage. The NF service producer may update the overload control information in the storage periodically or when the load status exceeds a threshold;

- The interface between NF service producer and overload information storage is not covered in this solution;

- When overload information is to conveyed to the service consumer, the service producer includes a hyperlink in the response message and the hyperlink points the relevant overload information stored in the overload control information storage;

- Upon receiving the response containing the hyperlink described as above, the service consumer retrieves the overload control information according to the hyperlink, this mechanism is based on HATEOAS as described in clause 4.7 of TS 29.501 [5];

The overload control information of the NF service producer is conveyed to the NF service consumer as follows:



Figure 11.14-1 Conveying Overload Control Information via hyperlink

1. NF Service Producer stores its overload control information on the dedicated OCI storage.

2. The NF Service Consumer invokes service towards the NF Service Producer, the NF service Consumer shall indicate support of HATEOAS related features as described in clause 4.7.6 of TS 29.501 [5].

3. When needed, the NF Service Producer responds to the service invoking request, in addition it includes in the response message the hyperlink pointing to its overload control information on the OCI storage. Based on overload control strategy the NF Service Producer may:

- always include the hyperlink in response messages if possible; or

- include the hyperlink in response messages when it is under overload condition.

4. If the NF Service Consumer receives a hyperlink pointing to the overload control information in step 3, then it retrieves the overload control information based on the hyperlink.

NOTE: The NF Service Consumer retrieves the overload control information only once for each time it receives a hyperlink pointing to the overload control information.

5. The OCI storage sends the overload control information which previously stored by the NF Service Producer to the NF Service Consumer.

## 11.15 Evaluation and Conclusion

### 11.15.1 Evaluation of the Solutions

In the following, an evaluation of all the proposed solutions is made with regards to each key requirement and associated sub-issues defined in clause 8. For addressing Key Requirement #1, Sub-Issue#1:

- A combination of Solution #1 and Solution #2 should be used. Overload control information is signaled from the concerned NF service producer to the NF service consumer either in a response or in a notification. This enables to mitigate most of the limitations identified for the Rel-15 overload control mechanism and depicted in clause 9.2.2.

- Solution #7 can also be considered as a complementary solution. Indeed, assuming that the overload control service is supported by a different service instance than the one providing the overloaded service of the NF producer, this service can be useful in case the concerned service producer instance is overloaded to a point that it cannot react to received requests (construct and send responses) or send notifications that can be used to include OCI towards NF consumers. This solution enables the overloaded service producer to signal OCI to all concerned network entities (who have subscribed) intermediately and at the same time without waiting until signalling exchanges (to consume the overloaded services) takes place. Other network entities (other than the concerned NF consumers), e.g. O&M, SCP and/or a 3rd party service can also make use of this new service to retrieve OCI information if required. However, it also implies a tight coupling in the implementation between this overload service and the other services supported by the NF service producer.

- Also alternatively, Solution #6 can be considered. However, it is to be noted that this solution can have the same limitations as the Rel-15 load control mechanism (see clause 9.3.2) in the sense that it can lead to additional overload situations in the NRF and NF consumers if there is no possibility to regulate the frequency of overload control updates (from NFs to the NRF) and notifications (from the NRF to subscribing consumer NFs).

For addressing Key Requirement #1, Sub-Issue#2 on the content of OCI:

- Solution #3 is advised. The information conveyed within the OCI shall contain the identity of the overloaded NF/ NF service, a requested reduction or limitation metric, a validity period, a sequence number used to identify stale OCIs and the scope of the conveyed OCI.

For addressing Key Requirement #1, Sub-Issue#3 on the frequency of OCI conveyance:

- Solution #5 ought to be used. OCI is included in every signalling message, HTTP response message and also notifications, sent by the NF producer to the NF consumer.

- If Solution #7 is used as described above, frequency of notifications of OCI towards subscribed NF consumers is suggested by the service consumer, and the concerned service producer takes that into account, together with its local policy when setting up this frequency.

For addressing Key Requirement #1, Sub-Issue#4 on where to include the conveyed OCI.- 3 variants are described in the different solution on how to signal OCI information in this TR:

- OCI is signaled by Subscription/Notification via a new service offered by the NF service producer (see clause 11.9 Solution #7)

- OCI is signaled by being piggybacked in the existing signalling messages sent between the NF producer and the NF consumer (see clause 11.11 Solution #1)

- OCI is retrieved by the service consumer using a HATEOAS links provided by the overloaded NF service producer. This solution has the following advantages:

a) this solution is inline with the service based architecture, with this solution the overload control component can be decoupled from the service logic component in the NF, thus more flexibility is provided for network deployments;

b) the burden of conveying OCI can be offloaded from the server;

c) the client is able to make the decision whether to retrieve the OCI (the client may not continue to contact that server thus the OCI may be useless), this can save the unnecessary OCI conveyance and provide more flexibility to the overall overload control solution.

Table 11.15-1: Comparison of the solutions on location of signalling OCI information

|  |  |  |
| --- | --- | --- |
| Method | Pros | Cons |
| OCI included in HTTP messages payload | 1. Provides flexibility to expand OCI in future releases as payloads can be larger when compared to HTTP headers. | 1. Most off the shelf HTTP proxies do not have readymade solutions for inserting content into HTTP payload.  2. Some response messages (204 No Content) do not have any payload.  3. Every time a new API is added and/or a HTTP method is added to an existing API, the OCI IE needs to be included in the payload of the messages in OpenAPI. |
| OCI included in HTTP custom headers | 1. Can be used on all messages (including responses carrying 204 No Content status code).  2. Does not require OpenAPI updates every time a new API / method to an existing API is added. | 1. Headers should be generally short for better performances. Including OCI in headers would restrict the size available for OCI IE. |

- OCI included in HTTP custom headers has the least impact on the openAPI, as it is expected that OCI information will not increase much and thus the size restriction in the header may not be an issue, given that the contents of OCI as identified in clause 11.5 are limited.

- Alternatively and in case Solution #7 is used as described above, it should hence not be precluded to use the associated notification mechanism to convey OCI.

For addressing Key Requirement #2:

- Solution #4 can be considered depending on the deployment scenarios.

For addressing Key Requirement #3 on Overload Control enforcement:

- A combination of Solution #8 and Solution #10 for Overload Control enforcement should be followed. The NF consumer is expected to take actions (throttling, reselection, etc.) taking into account the information provided by the overloaded NF producer in the OCI IE.

- Alternatively and in case an SCP is used with communication options C or D (refer to annex E of TS 23.501 [2]), Solution #y (Overload Control Enforcement by SCPs), should be used as it provides a centralized overload control management for all 5G NFs and enables to have the least impact on NF consumers and service delivery in general. Indeed, the SCP is in charge of managing the overload situation of a NF producer for all concerned NF consumers by taking into consideration all the available parameters that it has (e.g. OCI, other available NF instances, overall load distribution among NF instances, etc.).

The following considerations should also be taken into account:

- When applying throttling, the message prioritization scheme should be defined as described in Solution #8 and in accordance with the mechanisms already defined in Rel-15.

- Also when applying throttling, the error message type and content to be sent upstream should be defined and is hence left for the normative phase.

# 12 Selected solutions

## 12.1 Selected solution for Overload Control

Based on the provisions in clause 11.15 "Evaluation and Conclusion", CT shall study how to combine the following solutions to address all requirements with one unified solution:

- Key requirement #1.1 (meaning requirement #1, sub-issue #1): Solutions 1 and 2.

- Key requirement #1.2: Solution 3.

- Key requirement #1.3: Solution 5.

- Key requirement #1.4: Solutions 2 and 9 (solution #9 offers two alternatives, but only the second one, i.e.OCI conveyed via an HTTP custom header should be adopted.

- Key requirement #2: Solution 4.

- Key requirement #3: Solutions 8, 10 and 11.

## 12.2 Solution outline and implications on SBI specifications

Below is a higher level summary of the overload control solution. Potential impacts on normative SBI specs are also identified.

### 12.2.1 Overload Conveyance - General (key requirement #1.1)

How to meet key requirement #1.1 is defined in solution #1:

- The overloaded NF Service Producer includes an Overload Control Information (OCI) in the HTTP response messages.

- For deployments where the NF Service Consumer Instance and the NF Service Producer Instance interact directly, the OCI information is provided by the NF Service Producer Instance.

- For deployments where the NF Service Consumer Instance and the NF Service Producer Instance do not interact directly, i.e. the NF Service Producer instance is deployed behind a SCP and the SCP exposes the API endpoints on behalf of the NF Service Producer, the SCP shall include the OCI information after considering the overall overload situation of all the instances of the NF Service Producer that the SCP is acting on behalf.

- The NF Service Producer may also piggyback OCI information in the notification messages (i.e., existing notification messages).

Editor's note: This will likely impact the following 3GPP specs under CT4 responsibility: TS 29.500, 29.502-505, 29.509-511, 29.518, 29.531and 29.572 (use this info in the new WID, when specifying the list of impacted normative specs).

See more details in clause 11.3.

### 12.2.2 Overload Conveyance - Semantics (key requirement #1.2)

How to meet key requirement #1.2 is defined in solution #3. The following information shall hence be included in the Overload Control Information (OCI):

- Overload-Reduction-Metric

- Period-of-Validity

- Overload-Sequence-Number

- Scope

Editor's note: This will likely impact at least TS 29.571 (use this info in the new WID, when specifying the list of impacted normative specs).

See more details in clause 11.5.

### 12.2.3 Overload Conveyance - Frequency (key requirement #1.3)

How to meet key requirement #1.3 is defined in solution #5:

- The NF service producer monitors its overload situation and identifies when to start communicating the overload control information to NF service consumers, depending on configured thresholds or implementation specific triggers;

- Once the NF service producer decides to include OCI towards NF service consumers, it may piggyback it in every HTTP response message (the decision is implementation dependent) and may also do the same in every HTTP notification message that the NF service producer would normally send towards upstream NF service consumers (for the already subscribed notifications).

Editor's note: How to handle the absence of the OCI in subsequent messages will be addressed during the normative work.

Editor's note: This will likely impact the following 3GPP specs under CT4 responsibility: TS 29.500, 29.502-505, 29.509-511, 29.518, 29.531and 29.572 (use this info in the new WID, when specifying the list of impacted normative specs).

See more details in clause 11.7.

### 12.2.4 Overload Conveyance - Where to include OCI (key requirement #1.4)

How to meet key requirement #1.4 is defined in solutions #2 and #9, which propose to convey Overload Control Information (OCI) explicitly between NFs within 3GPP custom headers (solution #9 offers two alternatives, but only the header based alternative should be adopted).

Editor's note: This will likely impact the following 3GPP specs under CT4 responsibility: TS 29.500, 29.502-505, 29.509-511, 29.518, 29.531and 29.572 (use this info in the new WID, when specifying the list of impacted normative specs).

See more details clauses 11.4 and 11.11.

### 12.2.5 Overload Conveyance - HATEOAS (key requirements #1.1 and #1.4)

For deployments where HATEOAS is used, solution #12 may be used.

Editor's note: Summary of the solution for HATEOAS option will be addressed by a separate pCR.

See more details in clause 11.14.

### 12.2.6 Overload - Avoiding and mitigating (key requirement #2)

How to meet key requirement #2 is defined in solution #4. If NF services are deployed as virtual instances behind an API endpoint that terminates the API URI (e.g. SCP instance), the following mechanisms can be used to avoid and mitigate overload situations:

- OAM or any other monitoring entity monitors the resource usage (e.g. CPU, memory, link utilization etc.) of the NF Service instances.

- If the monitored resource usage is nearing a configured threshold, the monitoring system initiates new instances of the NF service, with the same API endpoint, which terminates at the Reverse Proxy.

- The Reverse Proxy routes subsequent new resource creation requests to lightly loaded NF Service instances.

- When the overload situation ceases, the OAM may shutdown some of the NF Service Instances, if the resource state in those NF Service instances are shared with other NF Service instances through UDSF or if the session contexts can be transferred by 3GPP Rel-16 eSBA solutions.

Editor's note 1: This solution should be aligned with the outcome of the 5G\_eSBA work, which is still ongoing at SA2.

Editor's note 2: This will likely impact the following 3GPP specs under CT4 responsibility: TS 29.500, 29.502-505, 29.509-511, 29.518, 29.531and 29.572 (use this info in the new WID, when specifying the list of impacted normative specs).

See more details in clause 11.6.

### 12.2.7 Overload Control (key requirement #3)

How to meet key requirement #3 is defined in solutions #8, 10 and 11.

- For deployments where SCPs are not used and direct communication models are implemented:

- The NF service consumer shall reduce the number of messages it would normally generate towards the NF service producer (message throttling), in accordance with the overload reduction or limitation metric requested by the NF service producer in the OCI IE (solution #8).

- When applying message throttling, the NF service consumer shall give preference to messages related to priority users (eMPS) and emergency services and shall avoid throttling the messages related to such users / services (solution #8). This priority mechanism should be used in combination with the existing 3gpp-Sbi-Message-Priority (SMP) mechanism, see TS 29.500 [3], clause 6.8.

- The OCI may include an indication on whether the NF service consumer can select another NF service producer from the same NF Set, if the scope of the OCI is at NF service producer level (solution #10).

- For deployments where SCPs are used to implement indirect communications between NF consumers and NF producers, the SCP is the overload control enforcement entity and shall perform the following functions (solution #11):

- Message throttling;- Message prioritization;

- If allowed by the service producer, selection of another NF service producer, e.g. from the same NF Set.

Editor's note: This will likely impact the following 3GPP specs under CT4 responsibility: TS 29.500, 29.502-505, 29.509-511, 29.518, 29.531and 29.572 (use this info in the new WID, when specifying the list of impacted normative specs).

See more details in clauses 11.10, 11.12, 11.13.

Annex:  
Change history

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Change history** | | | | | | | |
| **Date** | **Meeting** | **TDoc** | **CR** | **Rev** | **Cat** | **Subject/Comment** | **New version** |
| 2018-07 | CT4#85bis | C4-185418 |  |  |  | Skeleton of the TR | 0.1.0 |
| 2018-10 | CT4#86bis | C4-187647 |  |  |  | Implementation of PCRs agreed in CT4#86bis - C4-187163, C4-187527, C4-187528, C4-187601 | 0.2.0 |
| 2018-12 | CT4#87 | C4-188697 |  |  |  | Implementation of PCRs agreed in CT#87 - C4-188587 | 0.3.0 |
| 2019-03 | CT4#89 | C4-190634 |  |  |  | Implementation of PCRs agreed in CT4#89 - C4-190209, C4-190518, C4-190528, C4-190615, C4-190616, C4-190521, C4-190192, C4-190522, C4-190523, C4-190524, C4-190525 and C4-190527 | 0.4.0 |
| 2019-03 | CT#83 | CP-190049 |  |  |  | Presented for information | 1.0.0 |
| 2019-04 | CT4#90 | C4-191025 |  |  |  | General Cleanup | 1.1.0 |
| 2019-04 | CT4#90 | C4-191460 |  |  |  | Key Issues on Load Control | 1.1.0 |
| 2019-04 | CT4#90 | C4-191462 |  |  |  | Scope of OCI | 1.1.0 |
| 2019-04 | CT4#90 | C4-191464 |  |  |  | Evaluation of Rel-15 Load Control | 1.1.0 |
| 2019-04 | CT4#90 | C4-191466 |  |  |  | Solution for Frequency of OCI Inclusion | 1.1.0 |
| 2019-04 | CT4#90 | C4-191490 |  |  |  | Key Issue: Co-existence of load control mechanism with Rel-15 mechanism | 1.1.0 |
| 2019-04 | CT4#90 | C4-191491 |  |  |  | Solution for Load Control Co-Existence | 1.1.0 |
| 2019-04 | CT4#90 | C4-191492 |  |  |  | Solution for Conveying Overload Information via NRF | 1.1.0 |
| 2019-04 | CT4#90 | C4-191493 |  |  |  | Solution for populating Overload Control Information via new service | 1.1.0 |
| 2019-04 | CT4#90 | C4-191517 |  |  |  | Solution for Overload Control | 1.1.0 |
| 2019-04 | CT4#90 | C4-191518 |  |  |  | Solution for OCI Location | 1.1.0 |
| 2019-04 | CT4#90 | C4-191538 |  |  |  | Reselect producer instance when OCI indicates overload of original prod | 1.1.0 |
| 2019-06 | CT4#91 | C4-192034 |  |  |  | Key Requirement on the regulation of load information conveyance via NRF | 1.2.0 |
| 2019-06 | CT4#91 | C4-192138 |  |  |  | Solution for the Regulation of Load Information Change Notifications Conveyance by the NRF | 1.2.0 |
| 2019-06 | CT4#91 | C4-192139 |  |  |  | Solution for the Regulation of Load Information Updates Reception at the NRF | 1.2.0 |
| 2019-06 | CT4#91 | C4-192421 |  |  |  | Solutions for Overload Control enforcement by SCPs | 1.2.0 |
| 2019-06 | CT4#91 | C4-192422 |  |  |  | Clarification of the Overload Reduction Metric in the OCI IE | 1.2.0 |
| 2019-06 | CT4#91 | C4-192528 |  |  |  | Conveyance of OCI | 1.2.0 |
| 2019-06 | CT4#91 | C4-192531 |  |  |  | Load Control Evaluation and Conclusion | 1.2.0 |
| 2019-06 | CT4#91 | C4-192532 |  |  |  | OverLoad Control Evaluation and Conclusion | 1.2.0 |
| 2019-09 | CT4#93 | C4-193520 |  |  |  | Overload Control proposal | 1.3.0 |
| 2019-09 | CT4#93 | C4-193521 |  |  |  | Evaluation and Conclusion on hypermedia based OCI conveyance solution | 1.3.0 |
| 2019-09 | CT#85 | CP-192204 |  |  |  | Presented for approval | 2.0.0 |
| 2019-09 | CT#85 | CP-192241 |  |  |  | Presented for approval | 2.1.0 |
| 2019-09 | CT#85 |  |  |  |  | Approved at CT#85 | 16.0.0 |