Performance Assurance for 5G Networks Including Network Slicing

Yizhi Yao[1] and Xiaowen Sun[2]

[1] *3GPP SA5 rapporteur, Intel, USA*
[2] *3GPP SA5 co-rapporteur, China Mobile, China*
E-mail: yizhi.yao@INTEL.COM; sunxiaowen@chinamobile.com

Received 04 March 2019;
Accepted 20 May 2019

Abstract

As part of the 5G (5th Generation) family in 3GPP (3rd Generation Partnership Project), the standardization of performance assurance for 5G networks including network slicing has been started in 3GPP SA5 (Service and System Aspects Working Group 5) since release 15, and being continued in release 16 with further enhancements.

The network performance is directly associated to the experience of the end users, and concerns the user's satisfaction and loyalty. The 3GPP management system is designed to provide the capabilities for performance assurance, including:

- Near real-time and non-real time performance data collection and reporting;
- Instant performance threshold monitoring;
- MDA (Management Data Analytics) to support automated, preventative and predictive operations in network management and orchestration.

The performance assurance for 5G networks including network slicing targets at the entire 3GPP system across the NF (Network Function), NSSI (Network Slice Subnet Instance) and NSI (Network Slice Instance) perspectives.

Journal of ICT, Vol. 7_2, 93–108. *River Publishers*
doi: 10.13052/jicts2245-800X.722

Keywords: 5G, network slicing, performance assurance, measurement job, performance data file, real-time, performance data streaming, threshold monitoring, MDA, NF, NSSI, and NSI.

1 Introduction

The 5G network featured with network slicing capability is betaken to support diverse services, such as IoT (Internet of Things), cloud-based services, industrial control, autonomous driving, mission critical communications, etc. Each kind of service has its specific performance requirements, for instance massive connectivity, super-high bandwidth, ultra-low latency and ultra-high reliability.

The performance of 5G networks including network slicing needs to be assured in order to meet the performance requirements of the services. The performance of the NFs, NSSIs and NSIs needs to be monitored and analysed, to figure out the present and potential issues that are dragging down (or may drag down) the performance.

The performance assurance of 5G networks including network slicing relies on a set of management services with the relevant management data, e.g., performance measurements, KPIs (Key Performance Indicators) and management analytical data.

The performance assurance related management services, performance measurements and KPIs described in this paper are based on the approved 3GPP specifications TS 28.550-16.1.0 [1], 28.552-16.1.0 [2], 28.554-16.0.0 [3], 32.425-16.3.0 [4] and 28.532-16.0.0 [5] respectively.

2 Performance Assurance Services

2.1 General

The management services in terms of performance assurance include the measurement job control service, performance data file reporting service, performance data streaming service, performance threshold monitoring service and management data analytics service. The performance data includes performance measurements and KPIs for NFs, NSSIs and NSIs. The performance data of NSSI is generated based on the aggregation and calculation of performance data of NFs, and the performance data of NSI is produced based on the aggregation and calculation of performance data of NSSIs and NFs.

2.2 Measurement Job Control Service

The measurement job control service allows the consumer to create, stop and list the measurement jobs for collecting the performance measurements of NFs, NSSIs and NSIs. The consumer can choose to get the measurement results by file or by streaming. Reporting the performance data by file allows the GP (granularity period) of the measurement job no less than 5 mins, which is backward compatible with the legacy networks; while the performance data streaming supports the GP down to second level which is near real-time.

2.2.1 Measurement job creation for NF(s)

The procedure for creating a measurement job for NF(s) is illustrated in the Figure 1.

If the NF measurement job is successfully created, the NF measurement job control service producer will collect the performance data from the NF(s) accordingly.

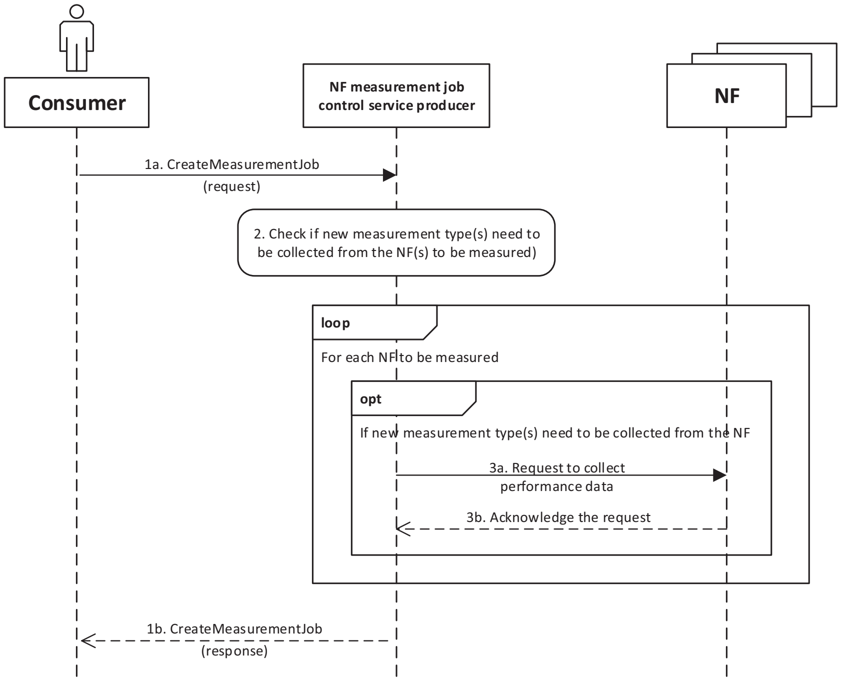

Consumer
NF measurement job control service producer
NF
1a. CreateMeasurementJob (request)
2. Check if new measurement type(s) need to be collected from the NF(s) to be measured)
loop
For each NF to be measured
opt
If new measurement type(s) need to be collected from the NF
3a. Request to collect performance data
3b. Acknowledge the request
1b. CreateMeasurementJob (response)

Figure 1 Procedure for NF measurement job creation.

2.2.2 Measurement job creation for NSSI(s)

The NSSI measurement job has a dependency on the NF measurement job(s). For the case that the NSSI measurement type(s) can be decomposed into the measurement data type(s) of the constituent NSSI(s) and NF(s), the procedure for creating a measurement job for NSSI(s) is illustrated in the Figure 2.

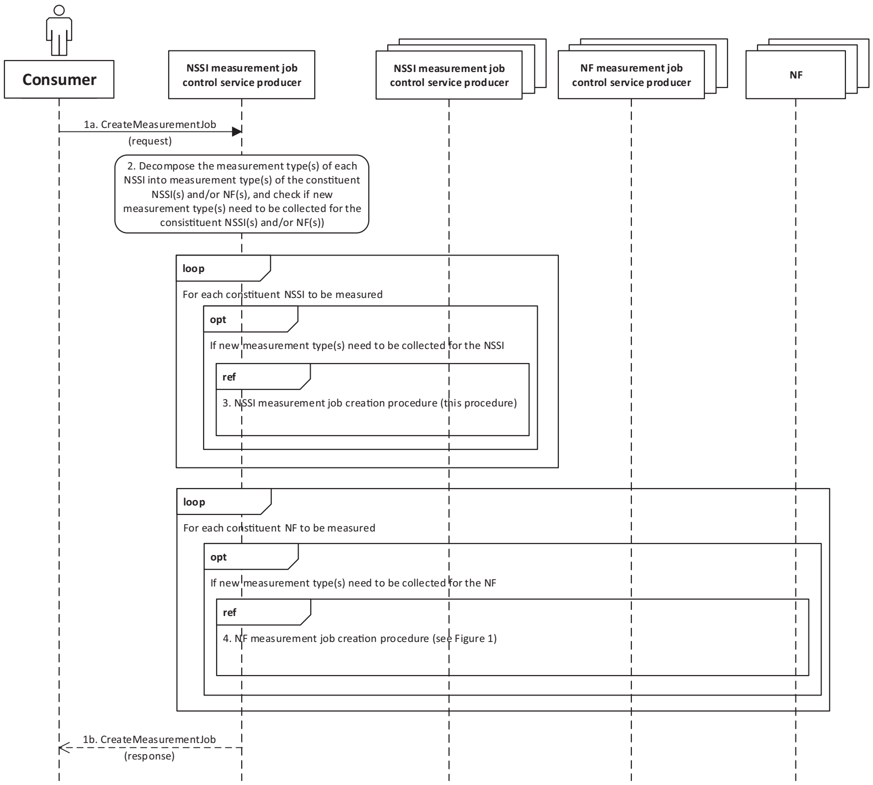

Consumer
NSSI measurement job control service producer
NSSI measurement job control service producer
NF measurement job control service producer
NF
1a. CreateMeasurementJob (request)
2. Decompose the measurement type(s) of each NSSI into measurement type(s) of the constituent NSSI(s) and/or NF(s), and check if new measurement type(s) need to be collected for the consistituent NSSI(s) and/or NF(s))
loop
For each constituent NSSI to be measured
opt
If new measurement type(s) need to be collected for the NSSI
ref
3. NSSI measurement job creation procedure (this procedure)
loop
For each constituent NF to be measured
opt
If new measurement type(s) need to be collected for the NF
ref
4. NF measurement job creation procedure (see Figure 1)
1b. CreateMeasurementJob (response)

Figure 2 Procedure for NSSI measurement job creation.

If the NSSI measurement job is successfully created, the NSSI measurement job control service producer will collect the performance data for the constituent NSSI(s) and/or NF(s) accordingly and generate the measurement results for the measured NSSI(s).

2.2.3 Measurement job creation for NSI(s)

The NSI measurement job has a dependency on the NSSI measurement job(s) and/or NF measurement job(s). For the case that the NSI measurement type(s) can be decomposed into the measurement data type(s) of the constituent

NSSI(s) and/or NF(s), the Figure 3 illustrates the procedure for creating a measurement job for NSI(s).

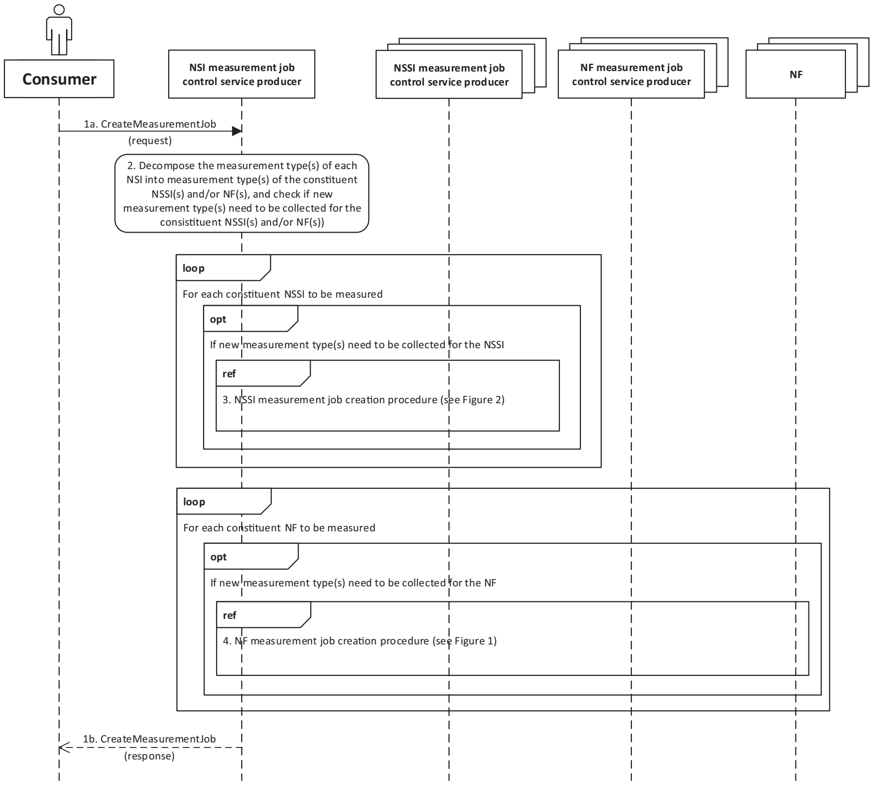

Consumer
NSI measurement job control service producer
NSSI measurement job control service producer
NF measurement job control service producer
NF
1a. CreateMeasurementJob (request)
2. Decompose the measurement type(s) of each NSI into measurement type(s) of the constituent NSSI(s) and/or NF(s), and check if new measurement type(s) need to be collected for the consistituent NSSI(s) and/or NF(s))
loop
For each constituent NSSI to be measured
opt
If new measurement type(s) need to be collected for the NSSI
ref
3. NSSI measurement job creation procedure (see Figure 2)
loop
For each constituent NF to be measured
opt
If new measurement type(s) need to be collected for the NF
ref
4. NF measurement job creation procedure (see Figure 1)
1b. CreateMeasurementJob (response)

Figure 3 Procedure for NSI measurement job creation.

If the NSI measurement job is successfully created, the NSI measurement job control service producer will collect the performance data for the constituent NSSI(s) and/or NF(s) accordingly and generate the measurement results for the measured NSI(s).

2.2.4 CreateMeasurementJob operation

The *createMeasurementJob* operation is used by the authorized consumer to request the measurement job control related service producer to create the measurement job.

One measurement job can collect the values of one or multiple measurement types. The measurement types are the performance measurements defined in TS 28.552 [2].

When a measurement type is collected by one measurement job for a given instance (e.g., an NF instance), another measurement job creation request to collect the same measurement type for the same instance with different GP may be rejected. This behaviour shall be consistent for a given implementation by a specific management service producer.

There are two different methods for the performance data to be reported:

- Performance data file method: In this method the performance data is accumulated for a certain period of time before it is reported; the data will be delivered as a file.
- Performance data streaming method: In this method, the performance data streaming producer, when the performance data is ready, sends it to the consumer (i.e., stream target). The volume of the performance data reported by streaming is expected to be small, and the GP of the performance data stream needs to be configurable and is expected to be short (in seconds).

In the request of *createMeasurementJob* operation, the consumer provides the following inputs for creating the job:

- object class name and object instance(s), whose measurement type(s) are to be collected;
- measurement type(s) to be collected;
- reporting method of the collected performance data, i.e., performance data file or performance data streaming;
- GP, i.e., the period between two successive measurements;
- reporting period, i.e., the period between two successive performance data reporting;
- start time, stop time and schedule of the measurement job;
- target of the performance data streaming (when the selected reporting method is performance data streaming);
- optionally the priority and reliability of the measurement job.

In the measurement job is successfully created, the service producer responds to the consumer with the identifier of the measurement job, otherwise the service producer provides the detailed information about the reason that the job creation failure.

2.3 Performance Data File Reporting Service

The performance data file reporting service reports the performance data by the file to the consumer. This is consistent with the performance data reporting method of the legacy networks.

The measurement job control related service producer (e.g., NF measurement job control service producer, NSSI measurement job control service producer, NSI measurement job control service producer) provides the measurement results (i.e. the value of the measurement type(s)) to the performance data reporting related service producer, and the performance data reporting related service producer generates the performance date file(s) for the consumer(s) and emits the *notifyFileReady* or *notifyFilePreparationError* notifications to the subject consumer(s) that have subscribed to these notifications.

The performance data reporting related service producer shall be able to allow the consumer to access the file using the following FTP (File Transfer Protocol) or SFTP (Secure File Transfer Protocol), and the performance data reporting related service producer shall always act as server while the consumer shall always act as the initiator (client) of file transfer actions.

2.4 Performance Data Streaming Service

The performance data streaming service reports the performance data to the consumer in near real-time, with the interval at second level. This advanced reporting approach provides fundamental support to the features such as MDA, SON (Self-Organizing Network) and other applications (e.g., Edge Computing applications) that require near real-time performance data.

The holistic sequence of performance data streaming is illustrated in the Figure 4.

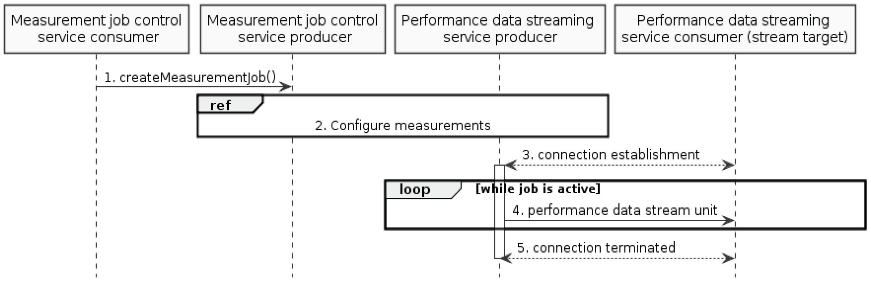

Measurement job control service consumer
Measurement job control service producer
Performance data streaming service producer
Performance data streaming service consumer (stream target)
1. createMeasurementJob()
ref
2. Configure measurements
3. connection establishment
loop
[while job is active]
4. performance data stream unit
5. connection terminated

Figure 4 Holistic sequence of performance data streaming.

In order to send the performance data via streams to the consumer (i.e., stream target), the performance data streaming service producer initiates the streaming connection establishment (using *establishStreamingConnection* operation) with the consumer. One streaming connection supports one or more streams. Each stream is designated by the producer to report a specific list

Table 1 Performance data stream unit content description

Performance Data Stream Unit Content	Description
streamId	The streamId of the performance data stream.
granularityPeriodEndTime	Time stamp referring to the end of the granularity period.
measResults	This parameter contains the sequence of result values for the observed measurement types. The "measResults" sequence shall have the same number of elements, following the same order, as the measurement types presented in `"measTypes"` for the subject stream in the input parameter `streamInfoList` of the *establishStreamingConnection* operation.

of measurements for one specific measured object, and the information about each stream is sent to the consumer during the establishment of the connection.

Once the streaming connection is successfully established, the producer sends the Performance Data Stream Units (using *reportStreamData* operation) to the consumer on this connection according to the information of the allocated streams when the performance data is ready for each GP.

When the performance data is no longer to be reported on the connection, the streaming service producer terminates the streaming connection (using *terminateStreamingConnection* operation) with the consumer.
The content of a Performance Data Stream Unit is described in the Table 1.

2.5 Performance Threshold Monitoring Service

It is a fairly big number of performance measurements that have been defined, and the amount of 5G network nodes is expected to be large, thus reporting the values for all of the measurements for all of the network nodes requires huge network bandwidth, and processing the values of all of the measurements needs many processing resources. To save the network bandwidth and processing resources, for some cases the operator/consumer may only need to be alerted when the performance value reaches or exceeds a certain threshold, without collecting all of the measurements.

The performance threshold monitoring service supports instant alert (by *notifyThresholdCrossing* notification) when the monitored performance value reaches or exceeds the pre-defined threshold. The consumer can create one or more threshold monitors with one or more threshold levels for a managed entity (which could be a sub-network, a manged/network element, a managed/network function, or a network slice subnet as modelled in Figure 5), and get notified when any threshold level is reached or crossed.

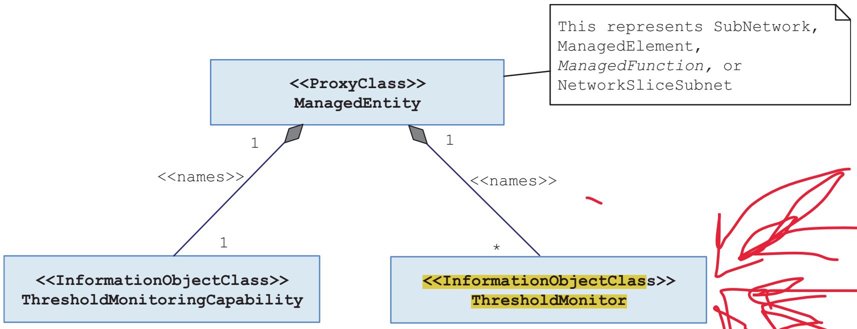


Figure 5 Threshold monitor models.

For the threshold monitoring of performance measurements that are cumulative counters, the *notifyThresholdCrossing* notification is emitted immediately when the cumulative counter of measured events reaches the threshold, without waiting to the end of the monitoring GP, as illustrated in Figure 6.

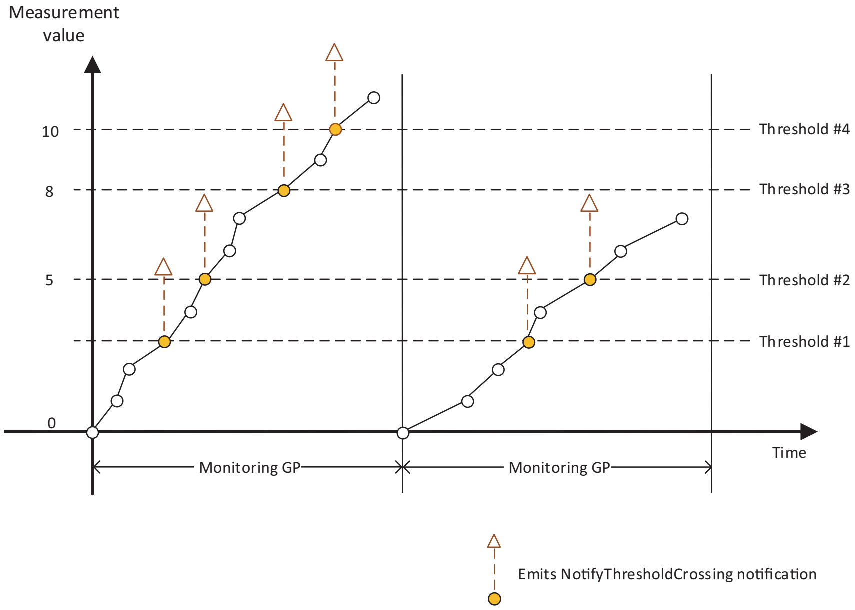


Figure 6 Threshold crossing notifications triggering for cumulative counters.

For the threshold monitoring of performance measurements that are not cumulative counters, the *notifyThresholdCrossing* notification is emitted at the end of the monitoring GP if the measurement value reached or crossed the threshold, as illustrated in Figure 7.

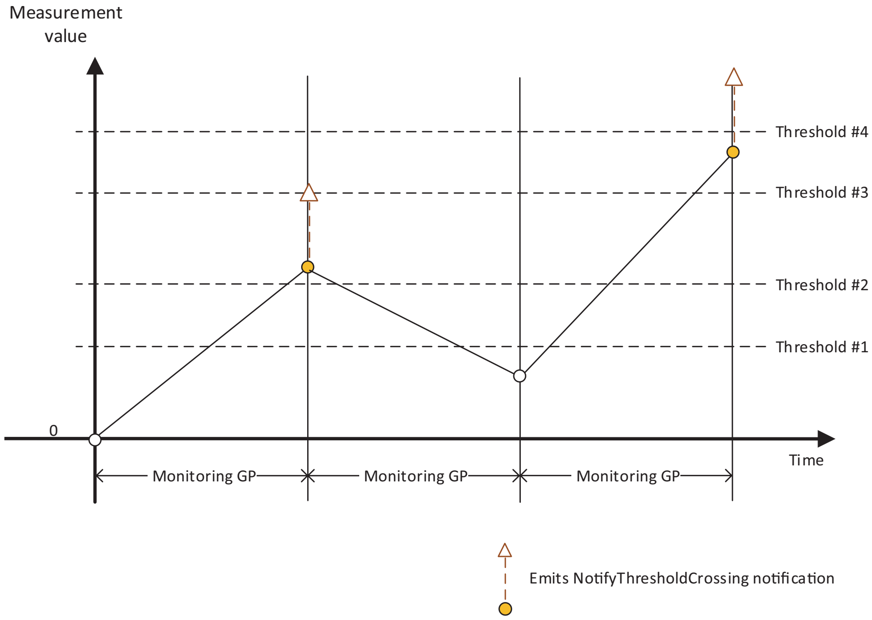


Figure 7 Threshold crossing notifications triggering for measurements that are not cumulative counters.

2.6 Management Data Analytics Service

The raw performance measurements, together with other management data (e.g., alarm information, configuration data), can be further analysed and processed by the MDAF (management data analytics function) to form management analytical data (e.g., analytical KPI). The management analytical data, derived from the current and historical performance measurements, can be used to diagnose ongoing issues and also to predict any potential issues. The consumer can use the management analytical data to locate and solve the issues, or prevent the potential failures. The MDAS (management data analytics service) can also provide the analytical input to SON functions for the network management automation, as shown in Figure 8.

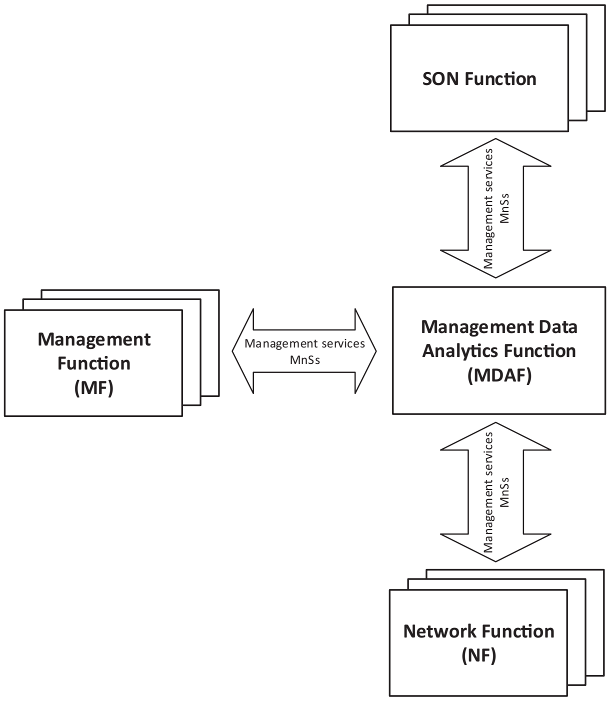


Figure 8 Management Data Analytics Service.

3 Performance Measurements and KPIs

The performance measurements are used to monitor the performance of NF functionalities, Virtualized Resource utilization and the E2E (End to End) QoS (Quality of Service) of the 5G networks including network slicing.

The KPIs are high-level performance indicators of key areas of the network from end to end point of view. The KPIs are generated by post processing (e.g., aggregation, calculation) of the performance measurements.

3.1 Performance Measurements

The performance measurements are defined for NFs including NG-RAN (the new Radio Access Network for 5G) and 5GC (5G Core Network), and for NSSIs and NSIs in terms of E2E QoS.

The performance measurements for NG-RAN are defined for both the so-called functional split deployment scenarios (e.g., Central Unit and Distributed

Unit split) and the non-split deployment scenario, so far related to the following aspects:

- Packet Delay
- Packet Loss Rate
- Packet Drop Rate
- IP Latency
- Radio resource utilization
- UE (User Equipment) IP (Internet Protocol) throughput (down link and uplink link)
- RRC (Radio Resource Control) connection
- PDU (Protocol Data Unit) Session Management
- Handovers
- Transport Blocks (TB)
- DRB (Data Radio Bearer) Setup Management
- QoS flow management
- UE Context management
- PDCP (Packet Data Convergence Protocol) data volume

The performance measurements for 5GC are defined for AMF (Access and Mobility Management Function), SMF (Session Management Function), UPF (User Plane Function), PCF (Policy Control Function) and UDM (Unified Data Management), etc.

For AMF, the measurements are related to the procedures such as registration and service requests.

For SMF, the measurements cover the aspects of PDU session management and QoS flow management.

For UPF, the measurements are to monitor the performance of the N3, N6 and N4 interfaces.

For PCF, the measurements are about AM (Access Management) policy association and SM (Session Management) policy association, etc.

For UDM, the numbers (mean and maximum) of registered subscribers are measured.

The measurements for NFs are split into subcounters per Network Slice Instance identifier (e.g., S-NSSAI – Network Slice Selection Assistance Information) and/or 5QI (5G QoS Identifier) or mapped 5QI if possible, in order to reflect the performance pertinent to the NSI and the QoS requirements.

There are also some measurements common to NG-RAN and 5GC NFs, such as the VR (Virtualized Resource) usage generated based on correlation and mapping of the measurements received from the MANO (Management and Orchestration) system defined by the ETSI NFV group (see ETSI GS NFV 002 [6]).

The performance measurements related to E2E network and network slicing are defined for monitoring the resource utilization and E2E QoS (e.g., E2E latency).

The performance measurements for EN-DC (Evolved-UMTS Terrestrial Radio Access Network New Radio – Dual Connectivity) are also defined (in TS 32.425 [4]), one example is the measurements related to secondary node additions.

Note that the coverage of the performance measurements is expanding, as the work is still ongoing and more and more performance measurements are being defined to support various use cases.

3.2 KPIs

The KPIs are categorized into the following aspects:

- Accessibility
- Integrity
- Utilization
- Retainability
- Availability
- Mobility

And, the concrete KPIs are defined for each category from end to end point of view, based on the processing of the performance measurements (see subclause 3.1).

The accessibility KPIs include, but are not limited to, the number of registered subscribers and registration success rate on an NSI or a network.

The integrity KPIs cover the E2E latency of the 5G network, upstream and downstream throughput of an NSI or a network utilization.

The retainability KPIs reflect the ability that the resource is retained when it is in use.

One example of the retainability KPI is the QoS flow retainability which shows how often an end-user abnormally loses a QoS flow during the time the QoS flow is used. The QoS flow retainability can be measured for a single QoS level (R1), and it is fairly straight forward:

$$R1_{Q\mathrm{o}S\_x} = \frac{QosFlow.RelActNbr.Q\mathrm{o}S_{Q\mathrm{o}S\_x}}{QoSFlow.SessionTimeQ\mathrm{o}S.Q\mathrm{oS}_{Q\mathrm{o}S\_x}}$$

The QoS flow retainability can be also measured from UE perspective (R2), and it is not as straightforward as R1, as for a UE there might be multiple QoS flows active at the same time, hence aggregating the QoS level measurements for session time will give a larger session time than the total UE session time, see the Figure 9 for example.

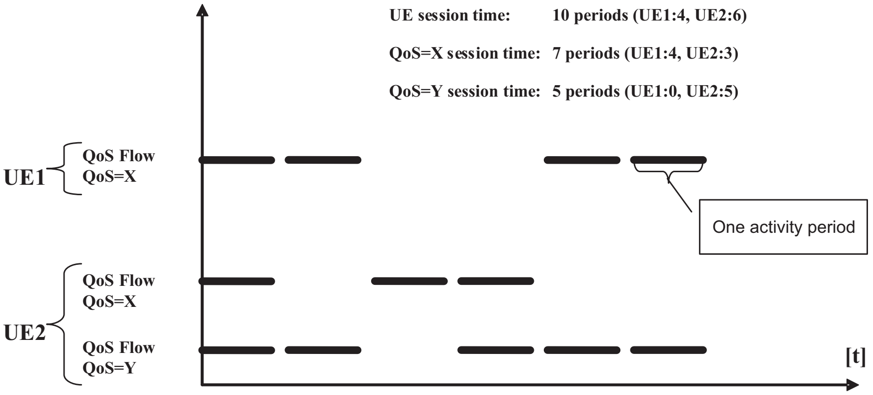

UE session time: 10 periods (UE1:4, UE2:6)
QoS=X session time: 7 periods (UE1:4, UE2:3)
QoS=Y session time: 5 periods (UE1:0, UE2:5)
UE1
QoS Flow QoS=X
UE2
QoS Flow QoS=X
QoS Flow QoS=Y
One activity period
[t]

Figure 9 QoS flow retainability for UEs.

$$R2 = \frac{\sum_{QoS} QosFlow.RelActNbr.[QoS]}{QosFlow.SessionTimeUE}$$

The utilization KPIs comprise the mean number of PDU sessions of an NSI or a network and the VR utilization of an NSI.

Given that the work is still progressing, more KPIs are being defined.

4 Conclusion

The performance assurance services defined by 3GPP SA5 provide the capabilities of near real-time and non-real-time performance data collection and reporting, instant performance threshold monitoring and intelligent analytics of management data for 5G networks including network slicing.

The performance data streaming service and performance data file reporting service give the consumers flexibility for selection or combination of real-time and non-real-time performance data collection in appropriate scenarios.

The instant performance threshold monitoring service keeps consumers alerted immediately when the performance reaches or exceeds a specific threshold, while saving the bandwidth and resources without collecting all of the performance data.

The management data analytics service brings the intelligence to the network operations and SON.

The performance assurance feature is the foundation of automated, preventative and predictive network management and orchestration for pursuing satisfactory performance of 5G networks including network slicing.

Acknowledgements

The authors wish to thank Thomas Tovinger, Chairman of 3GPP SA5, for his invitation, coordination and reviews for this paper.

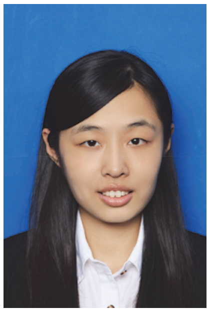
References

[1] 3GPP TS 28.550 V16.1.0 “Management and orchestration; Performance assurance”.
[2] 3GPP TS 28.552 V16.1.0 “Management and orchestration; 5G performance measurements”.
[3] 3GPP TS 28.554 V16.0.0 “Management and orchestration; 5G end to end Key Performance Indicators (KPI)”.
[4] 3GPP TS 32.425 V16.3.0 “Performance Management (PM); Performance measurements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN)”.
[5] 3GPP TS 28.532 V16.0.0 “Management and orchestration; Generic management services”.
[6] ETSI GS NFV 002 V2.4.1 “Network Functions Virtualisation (NFV), Architectural Framework”.

Biographies

Yizhi Yao is a system architect at Intel Corporation, United States, which he joined in 2016. He worked for Motorola and Nokia from 2005 to 2016. Yizhi has been actively participating 3GPP SA5 since 2005, and has served as Chairman of the OAM SWG from 2011 to 2015. He has also been involved in some other standardization organizations, such as ETSI ISG NFV, ETSI ISG MEC, etc. Yizhi has taken numerous rapporteurships in the standardization organizations, and has been endeavoring to promote the standards and solutions of mobile network management and orchestration towards cloudification, intelligentization (e.g., making the system more intelligent with AI) and automation.

Yizhi graduated from Beijing University of Posts and Telecommunications (BUPT) in 2003.



Xiaowen Sun is a Core Network Researcher employed by China Mobile Research Institute. She has much experience in international standardization activities and collaborations, and has 4 years working experience in network slicing technology areas. Xiaowen has been participating in 3GPP SA5 since 2017, and NGMN NWMO until its completion in 2019. She also takes part in ITU-T SG13, ETSI ISG MEC and GSMA NEST. Now she is vice chair of GSMA NEST group. Xiaowen graduated from Shanghai Jiao Tong University (SJTU) in 2015.