Oracle® Communications

Cloud Native Core, Network Exposure Function Benchmarking Guide

Release 24.2.0
G11758-01
July 2024

ORACLE

Oracle Communications Cloud Native Core, Network Exposure Function Benchmarking Guide, Release 24.2.0

G11758-01

Contents

My Oracle Support

My Oracle Support (https://support.oracle.com) is your initial point of contact for all product support and training needs. A representative at Customer Access Support can assist you with My Oracle Support registration.

Call the Customer Access Support main number at 1-800-223-1711 (toll-free in the US), or call the Oracle Support hotline for your local country from the list at http://www.oracle.com/us/support/contact/index.html. When calling, make the selections in the sequence shown below on the Support telephone menu:

- For Technical issues such as creating a new Service Request (SR), select **1**.
- For Non-technical issues such as registration or assistance with My Oracle Support, select **2**.
- For Hardware, Networking and Solaris Operating System Support, select **3**.

You are connected to a live agent who can assist you with My Oracle Support registration and opening a support ticket.

My Oracle Support is available 24 hours a day, 7 days a week, 365 days a year.

What's New in This Guide

This section introduces the documentation updates for release 24.2.x.

Release 24.2.0 - G11758-01, July 2024

- Updated the following sections to provide the latest performance testing values:
  - 5G Callflow Testcases
  - Monitoring Event Test Scenario 1
  - Monitoring Event Test Scenario 2
  - Quality of Service Test Scenario 1
  - Quality of Service Test Scenario 2
  - Traffic Influence Test Scenario 1
  - Traffic Influence Test Scenario 2
  - Device Trigger Test Scenario 1
  - Device Trigger Test Scenario 2
  - MSISDN-Less-MO-SMS Test Scenario 1
  - MSISDN-Less-MO-SMS Test Scenario 2

Acronyms

The following table provides information about the acronyms used in the document:

Table Acronyms

Acronym	Description
AMF	Access and Mobility Management Function
AUSF	Authentication Server Function
CPU	Central Processing Unit
cnDBTier	Cloud Native Database Tier
CNE	Oracle Communications Cloud Native Core, Cloud Native Environment
DB	Database
GW	Gateway
GPSI	Generic Public Subscription Identifier
HTTP	Hypertext Transfer Protocol
HPA	Horizontal Pod Autoscaling
HTTP	Hypertext Transfer Protocol
MPS	Messages Per Second
ms	Microsecond
NEF	Oracle Communications Cloud Native Core, Network Exposure Function
NF	Network Function
NRF	Network Repository Function
NSSF	Network Slice Selection Function
NS or Ns	Network Slice
PVC	Persistent Volume Claim
RAM	Random Access Memory
SMF	Session Management Function
TPS	Transactions Per Second
UDM	Unified Data Management

Terminologies

The following table provides information about the terminologies used in the document:

Table Acronyms

Terminology	Description
Message Per Second (MPS)	The rate of data transfer between connections is measured by message per second or MPS.
Transaction Per Second (TPS)	Transaction per second (TPS) is the number of transactions executed per second. The throughput in performance testing is expressed as TPS.
Ingress Gateway	An ingress gateway describes a load balancer operating at the edge of the mesh that receives incoming HTTP/TCP connections. It is an entry point for accessing OCNSSF supported service operations and provides the functionality of an OAuth validator.
Egress Gateway	An egress gateway describes a load balancer operating at the edge of the mesh that defines outgoing HTTP/TCP connections. It is responsible to route OCNSSF initiated egress messages to other NFs.
cnDBTier	cnDBTier is the geodiverse database layer provided as part of Oracle Communications Cloud Native Environment (OCCNE).

1 Introduction

This document provides information about the performance benchmarking of Oracle Communications Cloud Native Core, Network Exposure Function (NEF) and its microservices.

NEF is a key component of the 5G Service Based Architecture (SBA). NEF provides a platform to securely expose the network services and capabilities offered by the 5G Network Functions (NFs) to either external third-party applications or the internal Application Functions (AFs). As an interface between 5G core network and different application functions, NEF enables the operators to customize their network and provide innovative services to the end users.

NEF communicates with different application functions and the 5G core network to support the above mentioned functions. The application consists of the following components running on separate namespaces in the cloud native environment:

- **Common API Framework (CAPIF):** The service interfaces between NEF and the external third-party applications or internal Application Functions (AFs).
  CAPIF is a 3GPP defined secured framework to expose network service interfaces. It enables the API invokers (external applications) to discover and communicate with service APIs of the API provider (NEF). This framework manages API security, logging of events, auditing capability, multiple service exposure, policy based routing, dynamic routing of information, and so on.
- **Network Exposure Function (NEF):** The core component that runs the business logic of NEF. It consists of various services that interact with the CAPIF and perform the core functionality of NEF.

> **Note**
>
> CAPIF and NEF can be installed in different cluster also, but the cluster has to be reachable.

For more information about the services, see *Oracle Communications Cloud Native Core, Network Exposure Function User Guide*.

1.1 Purpose and Scope

This document is designed to help operators measure the capacity and performance of NEF, NEF microservices, and deployment environment setup and software details.

The document provides the following information:

- Benchmarking data of NEF performance and capacity
- Benchmarking done for NEF Model B deployment
- Benchmarking done for each service (ME/QoS/TI) separately
- Performance numbers provided were calculated with pre-loaded database with one million subscription records for ME/QoS/TI microservices
- The logging levels in all the involved microservices and stubs were set to ERROR

- Transactions per second for ME/QOS/TI services, when CPU average utilization reaches near about 70 percent
- Benchmarking data from the Oracle labs
- Key metrics used to manage NEF performance and capacity
- Recommendations on how to use the data obtained from the metrics

It is recommended that NEF is run through a benchmark on the target cloud native infrastructure to determine the capacity and performance in the target infrastructure. This information can be used to adjust the initial deployment resources and to help predict resource requirements when NEF is scaled up.

1.2 References

You can refer to the following documents for more information:

- *Oracle Communications Cloud Native Core, Network Exposure Function User Guide*
- *Oracle Communications Cloud Native Core, Network Exposure Function Installation, Upgrade, and Fault Recovery Guide*
- *Oracle Communications Cloud Native Core, Cloud Native Environment Installation, Upgrade, and Fault Recovery Guide*
- *Oracle Communications Cloud Native Core, DBTier User Guide*

2
Deployment Environment

This section provides information about the NF deployment platform, such as CNE, the services used for fetching counters or metrics, and the software requirements for NEF benchmark testing.

2.1 Deployed Components

Deployment Platform

Oracle Communications Cloud Native Core, Cloud Native Environment (CNE) 24.2.0 and Bare Metal CNE 24.2.0 can be used for performing benchmark tests for NEF deployment.

Observability Services

The following table lists services that are part of CNE and used for fetching NEF metrics.

Table 2-1 Observability Services

Service Name	Version
OpenSearch	2.11.0
OpenSearch Dashboard	2.11.0
logs	3.1.0
Kyverno	1.9
Fluentd	3.1.0
Prometheus	2.51.1
prometheus-kube-state-metrics	2.9.2
prometheus-node-exporter	1.5.0
Grafana	9.5.3
Jaeger	1.52.0
MetalLb	0.14.4
metrics-server	0.6.1
tracer	1.21.0

Cloud Native Orchestrator

Kubernetes 1.29.x is used to manage application pods across the cluster.

cnDBTier

cnDBTier 24.2.0 is used to perform benchmark tests.

For more information about above mentioned software, see *Oracle Communications Cloud Native Core, Network Exposure Function Installation, Upgrade, and Fault Recovery Guide*.

2.2 Deployment Resources

The performance and capacity of NEF can vary based on the chosen environment and how NEF is deployed. This section provides information about CNE and cnDBTier resources used to perform benchmark tests.

2.2.1 CNE Cluster Details

The following table provides information about the types of servers and the number of servers used in Bare Metal CNE clusters.

Table 2-2 Bare Metal CNE

Nodes	Type	Count
Worker Nodes	HP ProLiant BL460c Gen8	12
Primary Nodes	HP ProLiant BL460c Gen8	3

2.2.1.1 CNE Common Services Observability Resources

The following table provides information about the number of pods required by each CNE service.

Table 2-3 CNE Common Services Observability Resources

Service Name	No. of Pods	RAM Request/Limit	vCpu Request/Limit	PVC Size - Recommendation
Prometheus Server	2	4Gi/4Gi	2/2	8Gi
Alert Manager	2	64Mi/64Mi	20m/20m	NA
Fluentd	1 per Worker Node	512Mi/1Gi	100m/200m	NA
Prom-node-exporter	1 per Worker node	512Mi/512Mi	800m/800m	NA
Metal LB speaker	1 per Worker node	100Mi/100Mi	100m/100m	NA
ES Data	3	16Gi/16Gi	1/1	10Gi
ES Master	3	2Gi/2Gi	1/1	30Gi
ES Curator	1	128Mi/128Mi	100m/100m	NA
ES-exporter	1	128Mi/128Mi	100m/100m	NA
Grafana	1	128Mi/128Mi	100m/100m	NA
Kibana	1	500Mi/1Gi	100m/1	NA
kube-state-metrics	1	32Mi/100Mi	20m/20m	NA
jaeger-agent	12	128Mi/512Mi	256m/500m	NA
jaeger-collector	1	512Mi/1Gi	500m/1250m	NA
jaeger-query	1	128Mi/512Mi	256m/500m	NA

2.2.2 cnDBTier Resources

The following table describes resources required by cnDBTier 24.2.0 pods to perform NEF benchmark tests.

Table 2-4 cnDBTier Resources

DB Tier Pods	Replica	vCPU		RAM		Storage PVC
		Request	Limit	Request	Limit	
ndbappmysqld-n	2	8	8	10Gi	10Gi	20Gi
MGMT (ndbmgmd-n) StatefulSet	2	4	4	8Gi	10Gi	15Gi
DB (ndbmtd-n) StatefulSet	4	10	10	16Gi	18Gi	60Gi
SQL (ndbmysqld-n) StatefulSet	2	8	8	10Gi	10Gi	256Gi
nef-db-cluster-db-backup-manager-svc	1	100m	100m	128Mi	128Mi	NA
nef-db-cluster-db-monitor-svc	1	1	2	500Mi	500Mi	NA
mysql-cluster-site1-site2-replication-svc	1	2	2	12G	12G	NA

2.2.3 NEF Resources

The following table provides information about resource requirements to perform NEF benchmark tests.

Table 2-5 NEF Resources

Microservice Name	CPU Request and limit per POD (A)	Memory Request and limit per POD (B)	Scaling Criteria CPU Usage %	Default POD Count (C)	Maximum POD count with scaling (D)	Maximum CPU Total for Default PODs (A*C)	Minimum CPU Total for Default PODs (B*C)	Maximum CPU Total for all PODs (with full scaling and surge) (A*D)	Maximum Memory Total for all PODs (with full scaling and surge) (B*D)
5GC Agent	4	4Gi	60	5	5	20	20Gi	20	20Gi
5GC Egress Gateway	4	4Gi	60	5	5	20	20Gi	20	20Gi
5GC Ingress Gateway	4	4Gi	60	5	5	20	20Gi	20	20Gi
Common Config Hook	1	1Gi	70	1	1	1	1	1	1
APD Manager	4	4Gi	80	5	5	20	20Gi	20	20Gi

Table 2-5 (Cont.) NEF Resources

Microservice Name	CPU Request and limit per POD (A)	Memory Request and limit per POD (B)	Scaling Criteria CPU Usage %	Default POD Count (C)	Maximum POD count with scaling (D)	Maximum CPU Total for Default PODs (A*C)	Minimum CPU Total for Default PODs (B*C)	Maximum CPU Total for all PODs (with full scaling and surge) (A*D)	Maximum Memory Total for all PODs (with full scaling and surge) (B*D)
API Router	4	4Gi	80	5	5	20	20Gi	20	20Gi
App-Info	1	1Gi	70	5	5	5	5Gi	5	5Gi
CCF Client	2	2Gi	60	5	5	10	10Gi	10	10Gi
Config-Server	1	1Gi	70	5	5	5	5Gi	5	5Gi
Diameter Gateway	4	4Gi	No HPA support	5	5	20	20Gi	20	20Gi
Expiry Auditor	4	4Gi	60	5	5	20	20Gi	20	20Gi
External Egress Gateway	4	4Gi	80	5	5	20	20Gi	20	20Gi
External Ingress Gateway	4	4Gi	80	5	5	20	20Gi	20	20Gi
ME Service	4	4Gi	70	5	5	20	20Gi	20	20Gi
NRF Client	1	1Gi	70	5	5	5	5Gi	5	5Gi
Perf-Info	1	1Gi	70	5	5	5	5Gi	5	5Gi
QoS Service	4	4Gi	70	5	5	20	20Gi	20	20Gi
Traffic Influence	4	4Gi	70	5	5	20	20Gi	20	20Gi
Device Trigger	1	1Gi	70	1	1	1	1Gi	1	1Gi
Pool Manager	4	4Gi	70	1	1	4	4Gi	4	4Gi
MSISDNless MO SMS	4	4Gi	70	1	12	4	4Gi	4	4Gi
Console Data Service	4	4Gi	70	1	12	4	4Gi	4	4Gi

Note

Horizontal Pod Autoscaling (HPA) is not supported for Diameter Gateway, here the number of pods should be configured at the time of install or upgrade.

3 Benchmarking Model B

This section describes Model B test topologies and test scenarios for benchmarking NEF. The following features are considered to perform the benchmarking tests:

- Monitoring Event (ME)
- Quality of Service (QOS)
- Traffic Influence (TI)
- Device Trigger (DT)

Using the Model B deployment, the consumer NF performs NF discovery by querying NRF. The consumer NF discovers services available in the network based on service name and target NF type. The consumer NF invokes `Nnrf_NFDiscovery_Request` from an appropriate configured NRF. Based on the discovery result, the consumer NF selects the target producer NF and then sends the service request to that producer NF.

Model B supports the following functionalities that NF and NF services can use to interact with each other based on 3GPP TS 23.501:

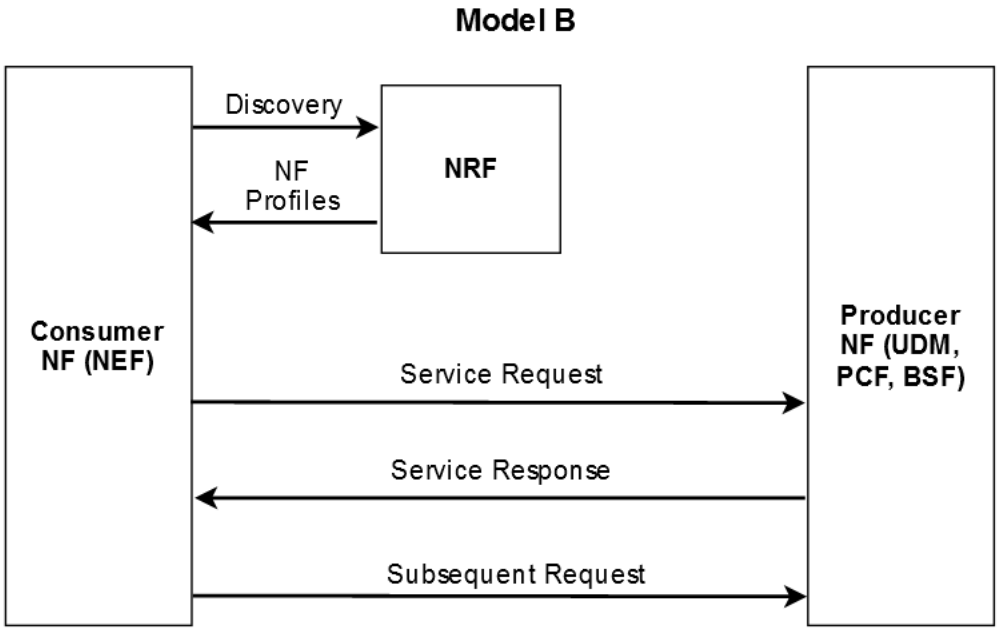
- Service request address
- Discovery using NRF services, no SCP, and direct routing

3.1 Test Topology for Model B Benchmarking

The following image represents a high level topology of Model B consisting of the following network functions:

- Consumer NF (NEF)
- NRF
- Producer NFs

Figure 3-1 Model B test topology



The aforementioned image represents the Model B test topology. In Model B direct 5G SBI communication mode, the consumer NF (NEF) sends NF discovery service requests to NRF. After receiving the discovery response with NF profiles, the consumer NF performs the following tasks:

- NEF sends a service request to producer NF such as UDM, PCF, BSF, and so on.
- The service request contains the address of the selected service producer NF.
- Producer NF responds to NEF with the service response.

> **Note**
>
> - Performance number to be calculated with pre-loaded database with 1 million subscription records for Monitoring (ME), Quality Of Service (QOS), and Traffic Influence (TI).
> - The logging levels in all the involved microservices and stub are set to ERROR. This reduces some execution time.

3.2 Model B Testcases

This section provides information about Model B testcases performed with Converged SCEF+NEF and 5G Callflow methodology.

> **Note**
>
> The setup on which performance was run for Model B testcases is Bulkhead (10.75.150.180).

3.2.1 5G Callflow Testcases

This section captures the testcase objective, input parameters, and results for the following test scenarios:

- Monitoring Event (ME)
- Quality of Service (QOS)
- Traffic Influence (TI)
- Device Trigger (DT)
- MSISDN-Less MO SMS

> **Note**
>
> For the above listed features in 5G callflow, the disk size for sql pod is 256Gi in testcase 3 and 100Gi in testcase 1 and 2.

3.2.1.1 Monitoring Event Testcases

In the network deployments, operators must track the current or the last known location of User Equipment (UE) to provide customized and enhanced network services. Any change in the UE location is considered as an event and NEF facilitates third-party applications or internal Application Functions (AFs) to monitor and get the report about such event.

The Monitoring Event service feature enables NEF to monitor and report the following events to the requested parties:

- Location reporting
- PDU session status
- UE reachability
- Loss of connectivity

The purpose of the ME service feature is to provide the following information to AFs:

- Current location of UE
- Last known location of UE with time stamp
- PDU Session status
- Loss of Connectivity Event
- UE Reachability Event

This functionality is achieved by using the 3GPP defined monitoring event based on the monitoring type. The event is detected based on the event reporting parameters received in the monitoring event subscription request as follows:

- One-time reporting

- Maximum number of reports
- Maximum duration of reporting
- Monitoring type

The following subsections provide information about monitoring event test scenarios considered to perform benchmark tests.

3.2.1.1.1 Single Site

This section provides information about configuration, parameters, and Monitoring Event Model B feature test scenarios of single site.

Deployment Configuration

The following table describes deployment configurations required for NEF microservices while performing benchmarking tests for the Monitoring Event Model B feature.

Table 3-1 Deployment Configuration

Microservice	Replica	CPU	Memory
NEF Ingress Gateway	4	4	4Gi
NEF Egress Gateway	4	4	4Gi
API Router	4	4	4Gi
ME	2	4	4Gi
TI	2	4	4Gi
QOS	2	4	4Gi
Expiry Auditor	2	4	4Gi
FiveGC Agent	4	4	4Gi
FiveGC Egress	4	4	4Gi
FiveGC Ingress	4	4	4Gi
Apd Manager	2	4	4Gi
nrf-client-nfmanagement	1	1	1Gi
nrf-client-nfdiscovery	1	1	1Gi
performance	2	200m	1Gi
Config-server	1	1	1Gi
app-info	1	200m	1Gi

> **Note**
>
> The VAR value for microservice in each scenario is based on the corresponding Replica value.

The following table describes resources required by Stub parameters to perform NEF benchmark tests.

Table 3-2 Stub Parameters

Stub Parameter	Replica	CPU	Memory
GMLC	3	1	1Gi

Table 3-2 (Cont.) Stub Parameters

Stub Parameter	Replica	CPU	Memory
UDM	3	1	1Gi
AF	3	1	1Gi
PCF	3	1	1Gi
BSF	3	1	1Gi
NRF	1	1	1Gi
UDR	3	1	1Gi

3.2.1.1.1.1 Monitoring Event Test Scenario 1

This test scenario describes transactions per second when CPU average utilization reaches around 70%.

Objective

To determine average TPS and average latency at 70% CPU utilization that Subscription, Unsubscription, and Notification endpoints can handle with two replicas of applicable services for both UDM and GMLC request types.

The following table describes the testcase parameters and their values.

Table 3-3 Monitoring Event Testcase Parameters

Input Parameter Details	Configuration Values
Cluster	For more information, see CNE Cluster Details.
Topology	For more information, see Test Topology for Model B Benchmarking.
Deployment Configuration	For more information, see Test Topology for Model B Benchmarking.
NEF Version Tag	24.2.0
Target CPU	70%
Number of NEF Pods for ME	2
ME Pod Profile	4 vCPU and 4 Gi Memory
Traffic Distribution	10% Subscription, 10% Unsubscription, and 80% Notifications

Result and Observation

The following parameters were observed after performing the Monitoring Event test:

- The server used is bulk head, which is a shared Bare metal server.
- The `global.cache.evict.time` is set to a higher value to avoid frequent calls from apd-agent in Fivegcagent microservice to apd-manager microservices. Therefore, `global.cache.evict.time` is set to 90 hrs.
- apdManager was not tuned as it gets hit only when the apd-agent cache evicts at fivegc and that is set to a higher value now.
- Tuning nrf-client related to pods is not required as it is contacted separately through apdmanager and that does not come into our real-time sub/unsub/notification flow.

- Scaling is not required for pods related to nrf-client.

The following table provides observation data for the performance test that can be used for benchmark testing to increase the traffic rate.

Table 3-4 Result and Observation

Parameter	Values
Test Duration	90 hrs
Replica	2

Table 3-5 NEF Microservices and their Utilization

Resources	Subscription	Unsubscription	Notification	Total
Average CPU	NA	NA	NA	~60%
Total TPS	250	250	2000	2500
Average Latency	99.2 ms	106.2 ms	40.6 ms	NA
Average Memory	NA	NA	NA	1.7 GB
Max Memory	NA	NA	NA	2.0 GB

3.2.1.1.1.2 Monitoring Event Test Scenario 2

This test scenario describes transactions per second when CPU average utilization reaches around 70%.

Objective

To determine average TPS and average latency at 70% CPU utilization that Subscription, Unsubscription, and Notification endpoints can handle with two replicas of applicable services for both UDM and GMLC request types.

The following table describes the testcase parameters and their values.

Table 3-6 Monitoring Event Testcase Parameters

Input Parameter Details	Configuration Values
Cluster	For more information, see CNE Cluster Details.
Topology	For more information, see Test Topology for Model B Benchmarking.
Deployment Configuration	For more information, see Test Topology for Model B Benchmarking.
NEF Version Tag	24.2.0
Target CPU	70%
Number of NEF Pods for ME	2
ME Pod Profile	4 vCPU and 4 Gi Memory
Traffic Distribution	10% Subscription, 10% Unsubscription, and 80% Notifications

Result and Observation

The following parameters were observed after performing the Monitoring Event test:

- The server used is bulk head, which is a shared Bare metal server.

- The `global.cache.evict.time` is set to a higher value to avoid frequent calls from apd-agent in Fivegcagent microservice to apd-manager microservices. Therefore, `global.cache.evict.time` is set to 19 hrs.
- apdManager was not tuned as it gets hit only when the apd-agent cache evicts at fivegc and that is set to a higher value now.
- Tuning nrf-client related to pods is not required as it is contacted separately through apdmanager and that does not come into our real-time sub/unsub/notification flow.
- Scaling is not required for pods related to nrf-client.

The following table provides observation data for the performance test that can be used for benchmark testing to increase the traffic rate.

Table 3-7 Result and Observation

Parameter	Values
Test Duration	19 hrs
Replica	2

Table 3-8 NEF Microservices and their Utilization

Resources	Subscription	Unsubscription	Notification	Total
Average CPU	NA	NA	NA	~65%
Total TPS	250	250	2000	2500
Average Latency	100.5 ms	1532 ms	97.3 ms	NA
Average Memory	NA	NA	NA	1.7 GB
Max Memory	NA	NA	NA	1.9 GB

3.2.1.2 Quality of Service Testcases

In network deployments, operators have the requirement to offer services of a certain quality. The quality of service depends on different parameters, such as the availability of a link, the number of bit errors, network latency, and jitter. There are scenarios when operators need to provide different quality of services to different types of subscribers or UEs.

NEF enables the operators to manage the QoS using a set of parameters related to the traffic performance on networks. It also provides the capability to set up different QoS standards for different UE sessions based on the service requirements and other specifications. To perform this functionality, the NEF QoS service communicates with Policy Control Function (PCF) to set up, modify, and revoke an AF session with the required QoS.

The AF session with the QoS service feature allows AF to request a data session for a UE with a specific QoS.

The AF Session with QoS Service functionality enables NEF to perform the following functionality:

- Set up an AF session with the required QoS
- Get the QoS session details for an AF
- Delete an AF session with QoS
- Receive the QoS notifications, such as QoS Monitoring or Usage reports from the PCF and forward them to the subscribed AF

To set up AF sessions with QoS, OCNEF invokes the PCF that is responsible to control the privacy checking of the target subscriber. The invoked PCF authorizes the subscription or notification request, performs the required operation, and sends responses to NEF. NEF exposes the information as and when received from PCF to AF.

The following subsections provide information about quality of service test scenarios considered to perform benchmark tests.

3.2.1.2.1 Single Site

This section provides information about configuration, parameters, and Quality of Service Model B feature test scenarios of single site.

Deployment Configuration

The following table describes deployment configurations required for NEF microservices while performing benchmarking tests for the Quality of Service Model B feature.

Table 3-9 Deployment Configuration

Microservice	Replica	CPU	Memory
NEF Ingress Gateway	4	4	4Gi
NEF Egress Gateway	4	4	4Gi
API Router	4	4	4Gi
ME	2	4	4Gi
TI	2	4	4Gi
QOS	2	4	4Gi
Expiry Auditor	2	4	4Gi
FiveGC Agent	4	4	4Gi
FiveGC Egress	4	4	4Gi
FiveGC Ingress	4	4	4Gi
Apd Manager	2	4	4Gi
nrf-client-nfmanagement	1	1	1Gi
nrf-client-nfdiscovery	1	1	1Gi
performance	2	200m	1Gi
Config-server	1	1	1Gi
app-info	1	200m	1Gi

> **Note**
>
> The VAR value for microservice in each scenario is based on the corresponding Replica value.

The following table describes resources required by Stub parameters to perform NEF benchmark tests.

Table 3-10 Stub Parameters

Stub Parameter	Replica	CPU	Memory
GMLC	3	1	1Gi

Table 3-10 (Cont.) Stub Parameters

Stub Parameter	Replica	CPU	Memory
UDM	3	1	1Gi
AF	3	1	1Gi
PCF	3	1	1Gi
BSF	3	1	1Gi
NRF	1	1	1Gi
UDR	3	1	1Gi

3.2.1.2.1.1 Quality of Service Test Scenario 1

This test scenario describes transactions per second when CPU average utilization reaches around 70%.

Objective

To determine average TPS and average latency at 70% CPU utilization that Subscription, Unsubscription, and Notification endpoints can handle with two replicas of applicable services for PCF requests with BSF enabled.

The following table describes the testcase parameters and their values.

Table 3-11 Quality of Service Testcase Parameters

Input Parameter Details	Configuration Values
Cluster	For more information, see CNE Cluster Details.
Topology	For more information, see Test Topology for Model B Benchmarking.
Deployment Configuration	For more information, see Test Topology for Model B Benchmarking.
NEF Version Tag	24.2.0
Target CPU	70%
Number of NEF Pods for QoS	2
QoS Pod Profile	4 vCPU and 4 Gi Memory
Traffic Distribution	25% Subscription, 25% Unsubscription, and 50% Notification

Result and Observation

The following parameters were observed after performing the Quality of Service test:

- The server used is bulk head, which is a shared Bare metal server.
- The `global.cache.evict.time` is set to a higher value to avoid frequent calls from apd-agent in Fivegcagent microservice to apd-manager microservices. Therefore, `global.cache.evict.time` is set to 90 hrs.
- apdManager was not tuned as it gets hit only when the apd-agent cache evicts at fivegc and that is set to a higher value now.
- Tuning nrf-client related to pods is not required as it is contacted separately through apdmanager and that does not come into our real-time sub/unsub/notification flow.
- Scaling is not required for pods related to nrf-client.

The following table provides observation data for the performance test that can be used for benchmark testing to increase the traffic rate.

Table 3-12 Result and Observation

Parameter	Values
Test Duration	90 hrs
Replica	2

Table 3-13 NEF Microservices and their Utilization

Resources	Subscription	Unsubscription	Notification	Total
Average CPU	NA	NA	NA	~65%
Total TPS	625	625	1250	2500
Average Latency	454.1 ms	353.6 ms	31.6 ms	NA
Average Memory	NA	NA	NA	1.7 GB
Max Memory	NA	NA	NA	1.9 GB

3.2.1.2.1.2 Quality of Service Test Scenario 2

This test scenario describes transactions per second when CPU average utilization reaches around 70%.

Objective

To determine average TPS and average latency at 70% CPU utilization that Subscription, Unsubscription, and Notification endpoints can handle with two replicas of applicable services for PCF requests with BSF enabled.

The following table describes the testcase parameters and their values.

Table 3-14 Quality of Service Testcase Parameters

Input Parameter Details	Configuration Values
Cluster	For more information, see CNE Cluster Details.
Topology	For more information, see Test Topology for Model B Benchmarking.
Deployment Configuration	For more information, see Test Topology for Model B Benchmarking.
NEF Version Tag	24.2.0
Target CPU	70%
Number of NEF Pods for QoS	2
QoS Pod Profile	4 vCPU and 4 Gi Memory
Traffic Distribution	25% Subscription, 25% Unsubscription, and 50% Notification

Result and Observation

The following parameters were observed after performing the Quality of Service test:

- The server used is bulk head, which is a shared Bare metal server.

- The `global.cache.evict.time` is set to a higher value to avoid frequent calls from apd-agent in Fivegcagent microservice to apd-manager microservices. Therefore, `global.cache.evict.time` is set to 19 hrs.
- apdManager was not tuned as it gets hit only when the apd-agent cache evicts at fivegc and that is set to a higher value now.
- Tuning nrf-client related to pods is not required as it is contacted separately through apdmanager and that does not come into our real-time sub/unsub/notification flow.
- Scaling is not required for pods related to nrf-client.

The following table provides observation data for the performance test that can be used for benchmark testing to increase the traffic rate.

Table 3-15 Result and Observation

Parameter	Values
Test Duration	19 hrs
Replica	2

Table 3-16 NEF Microservices and their Utilization

Resources	Subscription	Unsubscription	Notification	Total
Average CPU	NA	NA	NA	~61%
Total TPS	625	625	1250	2500
Average Latency	451.4 ms	360.4 ms	46.8 ms	NA
Average Memory	NA	NA	NA	1.65 GB
Max Memory	NA	NA	NA	2 GB

3.2.1.3 Traffic Influence Testcases

Application Function (AF) influence on traffic routing in NEF is supported through the Traffic Influence service. It allows an AF to send requests to NEF through the Traffic Influence APIs. Based on the availability of UE information in the AF request, NEF determines call flow towards 5G core NFs.

Traffic Influence is used by AFs to influence the routing decisions on the user plane traffic. It allows AF to decide on the routing profile and the route for data plane from UE to the network in a particular PDU session.

The following subsections provide information about traffic influence test scenarios considered to perform benchmark tests.

3.2.1.3.1 Single Site

This section provides information about configuration, parameters, and Traffic Influence Model B feature test scenarios of single site.

Deployment Configuration

The following table describes deployment configurations required for NEF microservices while performing benchmarking tests for the Traffic Influence Model B feature.

Table 3-17 Deployment Configuration

Microservice	Replica	CPU	Memory
NEF Ingress Gateway	4	4	4Gi
NEF Egress Gateway	4	4	4Gi
API Router	4	4	4Gi
ME	2	4	4Gi
TI	2	4	4Gi
QOS	2	4	4Gi
Expiry Auditor	2	4	4Gi
FiveGC Agent	4	4	4Gi
FiveGC Egress	4	4	4Gi
FiveGC Ingress	4	4	4Gi
Apd Manager	2	4	4Gi
nrf-client-nfmanagement	1	1	1Gi
nrf-client-nfdiscovery	1	1	1Gi
performance	2	200m	1Gi
Config-server	1	1	1Gi
app-info	1	200m	1Gi

> **Note**
>
> The VAR value for microservice in each scenario is based on the corresponding Replica value.

The following table describes resources required by Stub parameters to perform NEF benchmark tests.

Table 3-18 Stub Parameters

Stub Parameter	Replica	CPU	Memory
GMLC	3	1	1Gi
UDM	3	1	1Gi
AF	3	1	1Gi
PCF	3	1	1Gi
BSF	3	1	1Gi
NRF	1	1	1Gi
UDR	3	1	1Gi

3.2.1.3.1.1 Traffic Influence Test Scenario 1

This test scenario describes transactions per second when CPU average utilization reaches around 70%.

Objective

To determine average TPS and average latency at 70% CPU utilization that PCF or UDR Subscription, Unsubscription, and SMF Notification endpoints can handle with two replicas of applicable services for PCF requests with BSF enabled.

The following table describes the testcase parameters and their values.

Table 3-19 Traffic Influence Testcase Parameters

Input Parameter Details	Configuration Values
Cluster	For more information, see CNE Cluster Details.
Topology	For more information, see Test Topology for Model B Benchmarking.
Deployment Configuration	For more information, see Test Topology for Model B Benchmarking.
NEF Version Tag	24.2.0
Target CPU	70%
Number of NEF Pods for TI	2
TI Pod Profile	4 vCPU and 4 Gi Memory
Traffic Distribution	25% Subscription, 25% Unsubscription, and 50% Notifications

Result and Observation

The following parameters were observed after performing the Traffic Influence test:

- The server used is bulk head, which is a shared Bare metal server.
- The `global.cache.evict.time` is set to a higher value to avoid frequent calls from apd-agent in Fivegcagent microservice to apd-manager microservices. Therefore, `global.cache.evict.time` is set to 90 hrs.
- apdManager was not tuned as it gets hit only when the apd-agent cache evicts at fivegc and that is set to a higher value now.
- Tuning nrf-client related to pods is not required as it is contacted separately through apdmanager and that does not come into our real-time sub/unsub/notification flow.
- Scaling is not required for pods related to nrf-client.

The following table provides observation data for the performance test that can be used for benchmark testing to increase the traffic rate.

Table 3-20 Result and Observation

Parameter	Values
Test Duration	90 hrs
Replica	2

Table 3-21 NEF Microservices and their Utilization

Resources	Subscription	Unsubscription	Notification	Total
Average CPU	NA	NA	NA	~69%

Table 3-21 (Cont.) NEF Microservices and their Utilization

Resources	Subscription	Unsubscription	Notification	Total
Total TPS	650	650	1250	2550
Average Latency	108.1 ms	107.9 ms	30.0 ms	NA
Average Memory	NA	NA	NA	1.8 GB
Max Memory	NA	NA	NA	1.9 GB

3.2.1.3.1.2 Traffic Influence Test Scenario 2

This test scenario describes transactions per second when CPU average utilization reaches around 70%.

Objective

To determine average TPS and average latency at 70% CPU utilization that PCF or UDR Subscription, Unsubscription, and SMF Notification endpoints can handle with two replicas of applicable services for PCF requests with BSF enabled.

The following table describes the testcase parameters and their values.

Table 3-22 Traffic Influence Testcase Parameters

Input Parameter Details	Configuration Values
Cluster	For more information, see CNE Cluster Details.
Topology	For more information, see Test Topology for Model B Benchmarking.
Deployment Configuration	For more information, see Test Topology for Model B Benchmarking.
NEF Version Tag	24.2.0
Target CPU	70%
Number of NEF Pods for TI	2
TI Pod Profile	4 vCPU and 4 Gi Memory
Traffic Distribution	25% Subscription, 25% Unsubscription, and 50% Notifications

Result and Observation

The following parameters were observed after performing the Traffic Influence test:

- The server used is bulk head, which is a shared Bare metal server.
- The `global.cache.evict.time` is set to a higher value to avoid frequent calls from apd-agent in Fivegcagent microservice to apd-manager microservices. Therefore, `global.cache.evict.time` is set to 19 hrs.
- apdManager was not tuned as it gets hit only when the apd-agent cache evicts at fivegc and that is set to a higher value now.
- Tuning nrf-client related to pods is not required as it is contacted separately through apdmanager and that does not come into our real-time sub/unsub/notification flow.
- Scaling is not required for pods related to nrf-client.

The following table provides observation data for the performance test that can be used for benchmark testing to increase the traffic rate.

Table 3-23 Result and Observation

Parameter	Values
Test Duration	19 hrs
Replica	2

Table 3-24 NEF Microservices and their Utilization

Resources	Subscription	Unsubscription	Notification	Total
Average CPU	NA	NA	NA	~64%
Total TPS	650	650	1250	2500
Average Latency	84.5 ms	560 ms	90.6 ms	NA
Average Memory	NA	NA	NA	1.78 GB
Max Memory	NA	NA	NA	1.9 GB

3.2.1.4 Device Trigger Testcases

Using Device Triggering, applications notify User Equipment (UE) to perform application-specific tasks such as establishing communication with applications, changing device settings, and so on. Device triggering is required when an IP address for the UE is unavailable or unreachable by an application.

This mechanism enhances NEF functionality by introducing new APIs that allow AFs to remotely trigger specific actions on devices within a 5G network.

The Device Trigger feature enables an Application Function (AF) to notify a particular User Equipment (UE) by sending a device trigger request through 5G core (5GC) to perform application-specific tasks such as initiating communication with AF. This is required when the AF does not hold information of IP address for the UE or if the UE is not reachable. Device trigger request includes information required for an AF to send a message to the correct UE and to route the message to the required application. This information including the message to the application and the UE information is known as Trigger payload.

The following subsections provide information about device trigger test scenarios considered to perform benchmark tests.

3.2.1.4.1 Single Site

This section provides information about configuration, parameters, and Device Trigger Model B feature test scenarios of single site.

Deployment Configuration

The following table describes deployment configurations required for NEF microservices while performing benchmarking tests for the Device Trigger Model B feature.

Table 3-25 Deployment Configuration

Microservice	Replica	CPU	Memory
NEF Ingress Gateway	4	4	4Gi

Table 3-25 (Cont.) Deployment Configuration

Microservice	Replica	CPU	Memory
NEF Egress Gateway	4	4	4Gi
Diameter Gateway	4	4	4Gi
API Router	4	4	4Gi
ME	2	4	4Gi
DT	2	4	4Gi
QOS	2	4	4Gi
MSISDNLess Mo SMS	2	4	4Gi
Expiry Auditor	2	4	4Gi
FiveGC Agent	4	4	4Gi
FiveGC Egress	4	4	4Gi
FiveGC Ingress	1	4	4Gi
Apd Manager	2	4	4Gi
nrf-client-nfmanagement	1	1	1Gi
nrf-client-nfdiscovery	1	1	1Gi
performance	2	200m	1Gi
Config-server	1	1	1Gi
app-info	1	200m	1Gi

> **Note**
>
> The VAR value for microservice in each scenario is based on the corresponding Replica value.

The following table describes resources required by Stub parameters to perform NEF benchmark tests.

Table 3-26 Stub Parameters

Stub Parameter	Replica	CPU	Memory
GMLC	3	1	1Gi
UDM	3	1	1Gi
AF	3	1	1Gi
PCF	3	1	1Gi
BSF	3	1	1Gi
NRF	1	1	1Gi
UDR	3	1	1Gi
DIAM	3	1	1Gi

3.2.1.4.1.1 Device Trigger Test Scenario 1

This test scenario describes transactions per second when CPU average utilization reaches around 70%.

Objective

To determine average TPS and average latency at 70% CPU utilization that Subscription, Unsubscription, and Notification endpoints can handle with two replicas of applicable services for PCF requests with BSF enabled.

The following table describes the testcase parameters and their values.

Table 3-27 Device Trigger Testcase Parameters

Input Parameter Details	Configuration Values
Cluster	For more information, see CNE Cluster Details.
Topology	For more information, see Test Topology for Model B Benchmarking.
Deployment Configuration	For more information, see Test Topology for Model B Benchmarking.
NEF Version Tag	24.2.0
Target CPU	70%
Number of NEF Pods for DT	2
DT Pod Profile	4 vCPU and 4 Gi Memory
Traffic Distribution	25% Subscription, 25% Unsubscription, and 50% Notification

Result and Observation

The following parameters were observed after performing the Device Trigger test:

- The server used is bulk head, which is a shared VMware server.
- The `global.cache.evict.time` is set to a higher value to avoid frequent calls from apd-agent in Fivegcagent microservice to apd-manager microservices. Therefore, `global.cache.evict.time` is set to 90 hrs.
- apdManager was not tuned as it gets hit only when the apd-agent cache evicts at fivegc and that is set to a higher value now.
- Tuning nrf-client related to pods is not required as it is contacted separately through apdmanager and that does not come into our real-time sub/unsub/notification flow.
- Scaling is not required for pods related to nrf-client.

The following table provides observation data for the performance test that can be used for benchmark testing to increase the traffic rate.

Table 3-28 Result and Observation

Parameter	Values
Test Duration	90 hrs
Replica	2

Table 3-29 NEF Microservices and their Utilization

Resources	Subscription	Unsubscription	Notification	Total
Average CPU	NA	NA	NA	~68%
Total TPS	650	650	1300	2600

Table 3-29 (Cont.) NEF Microservices and their Utilization

Resources	Subscription	Unsubscription	Notification	Total
Average Latency	78.8 ms	79.8 ms	23.6 ms	NA
Average Memory	NA	NA	NA	1.6 GB
Max Memory	NA	NA	NA	1.9 GB

3.2.1.4.1.2 Device Trigger Test Scenario 2

This test scenario describes transactions per second when CPU average utilization reaches around 70%.

Objective

To determine average TPS and average latency at 70% CPU utilization that Subscription, Unsubscription, and Notification endpoints can handle with two replicas of applicable services for PCF requests with BSF enabled.

The following table describes the testcase parameters and their values.

Table 3-30 Device Trigger Testcase Parameters

Input Parameter Details	Configuration Values
Cluster	For more information, see CNE Cluster Details.
Topology	For more information, see Test Topology for Model B Benchmarking.
Deployment Configuration	For more information, see Test Topology for Model B Benchmarking.
NEF Version Tag	24.2.0
Target CPU	70%
Number of NEF Pods for DT	2
DT Pod Profile	4 vCPU and 4 Gi Memory
Traffic Distribution	25% Subscription, 25% Unsubscription, and 50% Notification

Result and Observation

The following parameters were observed after performing the Device Trigger test:

- The server used is bulk head, which is a shared VMware server.
- The `global.cache.evict.time` is set to a higher value to avoid frequent calls from apd-agent in Fivegcagent microservice to apd-manager microservices. Therefore, `global.cache.evict.time` is set to 19 hrs.
- apdManager was not tuned as it gets hit only when the apd-agent cache evicts at fivegc and that is set to a higher value now.
- Tuning nrf-client related to pods is not required as it is contacted separately through apdmanager and that does not come into our real-time sub/unsub/notification flow.
- Scaling is not required for pods related to nrf-client.

The following table provides observation data for the performance test that can be used for benchmark testing to increase the traffic rate.

Table 3-31 Result and Observation

Parameter	Values
Test Duration	19 hrs
Replica	2

Table 3-32 NEF Microservices and their Utilization

Resources	Subscription	Unsubscription	Notification	Total
Average CPU	NA	NA	NA	~67%
Total TPS	650	650	1300	2600
Average Latency	215.7 ms	210.3 ms	24.8 ms	NA
Average Memory	NA	NA	NA	1.65 GB
Max Memory	NA	NA	NA	1.8 GB

3.2.1.5 MSISDN-Less MO SMS Testcases

Support for MSISDN-Less-MO-SMS feature enables NEF to deliver the MSISDN-less MO-SMS notification message from Short Message Service - Service Center (SMSSC) to Application Function (AF).

With this feature, user equipment (UE) can send messages to AF without using Mobile Station International Subscriber Directory Number (MSISDN) through T4 interface, which is an interface between SMS-SC and NEF. NEF uses the Nnef_MSISDN-Less-MO-SMS API to send UE messages to AF.

The following subsections provide information about MSISDN-Less MO SMS test scenarios considered to perform benchmark tests.

3.2.1.5.1 Single Site

This section provides information about configuration, parameters, and MSISDN-Less MO SMS Model B feature test scenarios of single site.

Deployment Configuration

The following table describes deployment configurations required for NEF microservices while performing benchmarking tests for the MSISDN-Less MO SMS Model B feature.

Table 3-33 Deployment Configuration

Microservice	Replica	CPU	Memory
NEF Ingress Gateway	4	4	4Gi
NEF Egress Gateway	4	4	4Gi
Diameter Gateway	4	4	4Gi
API Router	4	4	4Gi
ME	2	4	4Gi
DT	2	4	4Gi
QOS	2	4	4Gi
MSISDNLess Mo SMS	2	4	4Gi

Table 3-33 (Cont.) Deployment Configuration

Microservice	Replica	CPU	Memory
Expiry Auditor	2	4	4Gi
FiveGC Agent	4	4	4Gi
FiveGC Egress	4	4	4Gi
FiveGC Ingress	1	4	4Gi
Apd Manager	2	4	4Gi
nrf-client-nfmanagement	1	1	1Gi
nrf-client-nfdiscovery	1	1	1Gi
performance	2	200m	1Gi
Config-server	1	1	1Gi
app-info	1	200m	1Gi

> **Note**
>
> The VAR value for microservice in each scenario is based on the corresponding Replica value.

The following table describes resources required by Stub parameters to perform NEF benchmark tests.

Table 3-34 Stub Parameters

Stub Parameter	Replica	CPU	Memory
GMLC	3	1	1Gi
UDM	3	1	1Gi
AF	3	1	1Gi
PCF	3	1	1Gi
BSF	3	1	1Gi
NRF	1	1	1Gi
UDR	3	1	1Gi
DIAM	3	1	1Gi

3.2.1.5.1.1 MSISDN-Less-MO-SMS Test Scenario 1

This test scenario describes transactions per second when CPU average utilization reaches around 70%.

Objective

To determine average TPS and average latency at 70% CPU utilizes Notification endpoints.

The following table describes the testcase parameters and their values.

Table 3-35 MSISDN-Less-MO-SMS Testcase Parameters

Input Parameter Details	Configuration Values
Cluster	For more information, see CNE Cluster Details.
Topology	For more information, see Test Topology for Model B Benchmarking.
Deployment Configuration	For more information, see Test Topology for Model B Benchmarking.
NEF Version Tag	24.2.0
Target CPU	70%
Number of NEF Pods for DT	2
MSISDN-Less-MO-SMS Pod Profile	4 vCPU and 4 Gi Memory
Traffic Distribution	100% Notification

Result and Observation

The following parameters were observed after performing the MSISDN-Less-MO-SMS test:

- The server used is bulk head, which is a shared VMware server.
- The `global.cache.evict.time` is set to a higher value to avoid frequent calls from apd-agent in Fivegcagent microservice to apd-manager microservices. Therefore, `global.cache.evict.time` is set to 90 hrs.
- apdManager was not tuned as it gets hit only when the apd-agent cache evicts at fivegc and that is set to a higher value now.
- Tuning nrf-client related to pods is not required as it is contacted separately through apdmanager and that does not come into our real-time notification flow.
- Scaling is not required for pods related to nrf-client.

The following table provides observation data for the performance test that can be used for benchmark testing to increase the traffic rate.

Table 3-36 Result and Observation

Parameter	Values
Test Duration	90 hrs
Replica	2

Table 3-37 NEF Microservices and their Utilization

Resources	Notification	Total
Average CPU	NA	~55%
Total TPS	6000	6000
Average Latency	10 ms	9.4 ms
Average Memory	NA	1.5 GB
Max Memory	NA	1.75 GB

3.2.1.5.1.2 MSISDN-Less-MO-SMS Test Scenario 2

This test scenario describes transactions per second when CPU average utilization reaches around 70%.

Objective

To determine average TPS and average latency at 70% CPU utilizes Notification endpoints.

The following table describes the testcase parameters and their values.

Table 3-38 MSISDN-Less-MO-SMS Testcase Parameters

Input Parameter Details	Configuration Values
Cluster	For more information, see CNE Cluster Details.
Topology	For more information, see Test Topology for Model B Benchmarking.
Deployment Configuration	For more information, see Test Topology for Model B Benchmarking.
NEF Version Tag	24.2.0
Target CPU	70%
Number of NEF Pods for DT	2
MSISDN-Less-MO-SMS Pod Profile	4 vCPU and 4 Gi Memory
Traffic Distribution	100% Notification

Result and Observation

The following parameters were observed after performing the MSISDN-Less-MO-SMS test:

- The server used is bulk head, which is a shared VMware server.
- The `global.cache.evict.time` is set to a higher value to avoid frequent calls from apd-agent in Fivegcagent microservice to apd-manager microservices. Therefore, `global.cache.evict.time` is set to 19 hrs.
- apdManager was not tuned as it gets hit only when the apd-agent cache evicts at fivegc and that is set to a higher value now.
- Tuning nrf-client related to pods is not required as it is contacted separately through apdmanager and that does not come into our real-time notification flow.
- Scaling is not required for pods related to nrf-client.

The following table provides observation data for the performance test that can be used for benchmark testing to increase the traffic rate.

Table 3-39 Result and Observation

Parameter	Values
Test Duration	19 hrs
Replica	2

Table 3-40 NEF Microservices and their Utilization

Resources	Notification	Total
Average CPU	NA	~68%
Total TPS	6000	6000
Average Latency	9.1 ms	12.9 ms
Average Memory	NA	1.64 GB
Max Memory	NA	1.66 GB