

High Availability

- Feature Summary and Revision History, on page 1
- Feature Description, on page 2
- Feature Description, on page 2
- Configuring NRF High Availability Feature, on page 3

Feature Summary and Revision History

Summary Data

Table 1: Summary Data

Applicable Product(s) or Functional Area	5G-NRF
Applicable Platform(s)	SMI
Feature Default Setting	Enabled – Always-on
Related Changes in this Release	Not Applicable
Related Documentation	Not Applicable

Revision History

Table 2: Revision History

Revision Details	Release
First introduced.	2026.01

Feature Description

The NRF High Availability feature supports an active-active model to increase the availability of NRF service operations toward NFs during NRF and its component failures.

The NRF supports High Availability feature at the following multiple levels:

- NRF pods
- NRF server or worker node
- Site or data center

Feature Description

The NRF High Availability feature supports an active-active load sharing model. The following sections describe the feature at multiple levels.

NRF Pods

1. Provide high availability between NRF REST endpoint and service pods.
2. There must be minimum of two replicas for NRF Rest endpoint and service pods.
3. If one NRF Rest endpoint pod is unavailable, the transactions handled by this endpoint is terminated and the client NF resends the messages.
4. If one NRF service pod is unavailable, the transactions handled by this pod is terminated and the NRF rest endpoint resends the messages to another active service pod.

   Note:

   - If any pod is unavailable, the Kubernetes layer restarts the pod and makes it active.

NRF Server or Worker Node

1. The minimum number of nodes for a type of NRF pod is two to support high availability between Proto VMs (VM with NRF Rest endpoint pods) or service VMs (VM with NRF service pods).
2. A single virtual IP address (VIP) is used as an NRF endpoint for external interface when the REST endpoint pods are deployed on Proto VMs. The VIP is exposed on top of NRF REST endpoint Kubernetes service Cluster IP.
3. If one Proto VM is unavailable, the transactions handled by the endpoints in this VM is terminated and the client NF resends the messages.
4. If one service VM is unavailable, the transactions handled by the service pods in this VM is terminated and the NRF rest endpoint resends the messages to another active service pod.

Note:

- Keepalived running on Proto VMs, which is handled by SMI is used to manage the VIP.

NRF Site or Data Center

The NRF High Availability feature provides high availability between sites or data centers. It supports an active-active model where both the local and remote NRFs serve the NFs corresponding to their sites. During the failure of a local NRF, the remote NRF gains priority to serve the NFs from a different site until the local NRF of the corresponding site is restored and active.

NRF REST Endpoint

If any NRF service pod is unavailable, the transactions handled by this pod is terminated and the NRF rest endpoint resends the messages to another active service pod.

NRF Service Engine

If any NRF REST endpoint pod is unavailable, the transactions handled by this endpoint is terminated and the NRF service pod resends the messages to another active NRF endpoint.

Configuring NRF High Availability Feature

The following sections describe the configuration for the NRF High Availability feature at multiple levels.

Configuring NRF Ops Center

To configure the deployment of pods replicas across nodes, use the following sample configuration.

```
config
instance instance-id instance_id
  endpoint { sbi | service }
    nodes nodes_number
    replicas replicas_number
    end
```

NOTES:

- **endpoint { sbi | service }**: Specify the SBI or service endpoint.
- **nodes** *nodes_number*: Specify the number of nodes for resiliency.
- **replicas***replicas_number*: Specify the number of replicas to be created per node. Default value: 1.
- This configuration deploys Y number of pods on each worker node for X number of worker nodes.
- The total number of NRF endpoints or service pods is X multiplied by Y.

Configuration Example

The following is an example configuration.

```
config
 endpoint sbi
  nodes 3
  replicas 2
  exit
 endpoint service
  nodes 3
  replicas 2
  exit
```

Configuring NRF REST Endpoint Pods and Service Pods

To configure replicas of the NRF REST endpoint pods and service pods at the NRF pod level, use the following sample configuration:

```
config
   instance instance-id instance_id
      endpoint { sbi | service }
      replicas replicas_number
      exit
```

NOTES:

- **replicas** *replicas_number*: Specify the number of replicas per node. Default value: 1.
- All the replicas are deployed on a single worker node with affinity policy.

Configuration Example

The following is an example configuration.

```
config
 instance instance-id 1
  endpoint sbi
  replicas 2
  exit
```

Configuring NRF REST Endpoint and Service Pods at the Server Level

To configure the nodes for the NRF REST endpoint pods and service pods at the server or worker level, use the following sample configuration:

```
config
   instance instance-id instance_id
      endpoint service
         nodes nodes_number
         exit
```

NOTES:

- **nodes** *nodes_number*: Specify the number of node replicas for resiliency.
- All the pods are deployed on multiple worker nodes with anti-affinity policy.

Configuration Example

The following is an example configuration.

```
config
 endpoint service
 nodes 2
 exit
```

Configuring VIP for ProtoVMs

To configure the Keepalived VIP for ProtoVMs , use the sample following configuration:

```
config
virtual-ips sbi
vrrp-interface bond1.AA
vrrp-router-id 5
ipv4-addresses 209.165.200.225
mask 24
broadcast 209.165.200.255
device bond1.AA
exit
hosts nrf01-proto1
priority 2
exit
hosts nrf01-proto2
priority 1
exit
exit
```



Note The VRRP-Router-ID must be unique across the VM-based environment.