Troubleshoot CNDP Cluster Manager HA Issues and Recover

Contents

Introduction

This document describes the procedure to troubleshoot High Availability (HA) issues in the Cloud Native Deployment Platform (CNDP) cluster manager.

Prerequisites

Requirements

Cisco recommends that you have knowledge of these topics:

- Cisco Subscriber Microservices Infrastructure (SMI)
- 5G CNDP or SMI-Bare-metal (BM) architecture
- Distributed Replicated Block Device (DRBD)

Components Used

The information in this document is based on these software and hardware versions:

- SMI 2020.02.2.35
- Kubernetes v1.21.0
- DRBD 8.9.10

The information in this document was created from the devices in a specific lab environment. All of the devices used in this document started with a cleared (default) configuration. If your network is live, ensure that you understand the potential impact of any command.

Background Information

What is SMI?

Cisco SMI is a layered stack of cloud technologies and standards that enable microservices-based applications from Cisco Mobility, Cable, and Broadband Network Gateway (BNG) business units – all of which have similar subscriber management functions and similar datastore requirements.

Attributes:

- Layer Cloud Stack (technologies and standards) to provide top to bottom deployments and also accommodate current user cloud infrastructures.
- Common Execution Environment shared by all applications for non-application functions (data storage, deployment, configure, telemetry, alarm). This provides consistent interaction and experience for all user touchpoints and integration points.
- Applications and Common Execution Environments are deployed in microservice containers and connected with an Intelligent Service Mesh.
- Exposed API for deployment, configuration, and management, in order to enable automation.

What is SMI-BM or CNDP?

Cisco SMI–BM or CNDP is a curated bare-metal platform that provides the infrastructure to deploy Virtual Network Functions (VNF) and Cloud-Native Functions (CNFs), that enable Cisco Mobility, Cable, and BNG business units.

Attributes:

- BM that eliminates the Virtualized Infrastructure Manager (VIM)-related overhead.
- Improved performance: More Cores for applicationFaster application execution
- Automated deployment workflow, integrated with Network Services Orchestrator (NSO) Core Function Packet (CFP)
- Curated Stack to deploy Cisco 5G Network Functions (NFs)
- Simplified order and deployment guide

What is an SMI Cluster Manager?

A cluster manager is a 2-node **keepalived** cluster used as the initial point for both control plane and user plane cluster deployment. It runs a single-node Kubernetes cluster and a set of PODs which are responsible for the entire cluster setup. Only the primary cluster manager is active and the secondary takes over only in case of a failure or brought down manually for maintenance.

What is DRBD?

DRBD is used to increase the availability of data. It is a Linux-based open-source software component that facilitates the replacement of shared storage systems by a networked mirror. In short, you can say this is a "Network-based Raid 1 mirror for the data".

Problem

The cluster manager is hosted in a 2-node cluster with Distributed Replicated Block Device (DRBD) and **keepalived**. The HA can break and it can get into a split-brain state also. This procedure

helps to recover the broken cluster. The desired state of cluster manager HA is that **cluster manager1 (CM1)** is primary and **cluster manager2 (CM2)** is secondary. Here, CM1 is the split-brain victim.

Procedure for the Maintenance

Login to the cluster manager and check the status of the DRBD cluster.

```
cloud-user@pod-name-cm-1:~$ drbd-overview
 0:data/0  WFConnection Secondary/Unknown UpToDate/DUnknown
```

```
cloud-user@pod-name-cm-2:~$ drbd-overview
 0:data/0  StandAlone Primary/Unknown UpToDate/DUnknown /mnt/stateful_partition ext4 568G 147G
392G 28%
```

In this scenario, CM2 is primary and the cluster is in standalone mode. CM1 is currently secondary and in **wait for connection** state.

Here is the correct state of the cluster:

```
cloud-user@pod-name-deployer-cm-1:~$ drbd-overview
0:data/0 Connected Primary/Secondary UpToDate/UpToDate /mnt/stateful_partition ext4 568G 364G
176G 68%
cloud-user@pod-name-deployer-cm-2:~$ drbd-overview
0:data/0 Connected Secondary/Primary UpToDate/UpToDate Move the CM VIP to CM-1 from CM-2 and
make CM-1 as primary –
```

Move the CM VIP to CM-1 from CM-2 and make CM-1 as primary (Cluster manager VIP is **10.x.x.65**

```
On CM-2 issue below command
cloud-user@pod-name-cm-2:~$sudo systemctl restart keepalived
```

```
On CM-1 issue below command (Make sure that the VIP is now switched over to CM-1)
cloud-user@pod-name-cm-1:~$ip a s | grep 10.x.x
inet 10.x.x.70/26 brd 10.x.x.127 scope global vlan1xx   ----> here is the server IP
inet 10.x.x.65/26 scope global secondary vlan1xx.              ----> here is the VIP
```

Identify the DRBD resource (shared over the network):

```
cloud-user@pod-name-deployer-cm-1:/$ cat /etc/fstab
#/mnt/stateful_partition/data /data none defaults,bind 0 0   ---> /data is the resource
#/mnt/stateful_partition/home /home none defaults,bind 0 0
```

```
cloud-user@pod-name-deployer-cm-1:/$ df -h
Filesystem Size Used Avail Use% Mounted on
udev 189G 0 189G 0% /dev
tmpfs 38G 22M 38G 1% /run
/dev/sda1 9.8G 3.5G 5.9G 37% /
tmpfs 189G 0 189G 0% /dev/shm
tmpfs 5.0M 0 5.0M 0% /run/lock
tmpfs 189G 0 189G 0% /sys/fs/cgroup
/dev/sda4 71G 1.5G 66G 3% /tmp
/dev/sda3 71G 11G 57G 16% /var/log
/dev/drbd0 568G 365G 175G 68% /mnt/stateful_partition      -->/dev/drbd0 is the device name
tmpfs 38G 0 38G 0% /run/user/1000
```

```
cloud-user@pod-name-deployer-cm-1:/$ lsblk
NAME MAJ:MIN RM SIZE RO TYPE MOUNTPOINT
```

```
sda 8:0 0 744.1G 0 disk
sda1 8:1 0 10G 0 part /
sda2 8:2 0 10G 0 part
sda3 8:3 0 72.2G 0 part /var/log
sda4 8:4 0 72.2G 0 part /tmp
sda5 8:5 0 577.6G 0 part
 drbd0 147:0 0 577.5G 0 disk /mnt/stateful_partition      ---> /dev/sda5 is used to create drbd0
```

Check the DRBD configuration file for the resource details:

```
cloud-user@pod-name-deployer-cm-1:/$ cat /etc/drbd.d/data.res
resource data {
protocol C;  --->Synchronous replication protocol. Local write operations on the primary node
are considered completed only after both the local and the remote disk write have been
confirmed. As a result, loss of a single node is guaranteed not to lead to any data loss


....
....

device /dev/drbd0;
disk /dev/disk/by-partlabel/smi-state; --> This translates to /dev/sda5
meta-disk internal;
floating 10.192.1.2:7789;
floating 10.192.1.3:7789;
```

Now, perform DRBD recovery:

```
On CM-2

cloud-user@pod-name-cm-2:~$ sudo systemctl stop keepalived ---> stop to avoid VRRP VIP
switchover
cloud-user@pod-name-cm-2:~$ sudo drbdadm disconnect data               ---> data is the cluster
resource
cloud-user@pod-name-cm-2:~$ sudo drbdadm secondary data ---> Make it secondary manually
cloud-user@pod-name-cm-2:~$ sudo drbdadm connect --discard-my-data data ---> Force discard of
all modifications on the split brain victim
cloud-user@pod-name-cm-2:~$ drbd-overview status

On CM-1:
cloud-user@pod-name-cm-1:~$ sudo systemctl stop keepalived  ---> stop to avoid VRRP VIP
switchover
cloud-user@pod-name-cm-1:~$ sudo drbdadm connect data ---> Data will be connected as primary
cloud-user@pod-name-cm-1:~$ drbd-overview status
```

Start **keepalived** process on both CMs. VRRP with the help of **keepalived** selects CM-1 as primary,
based on the connected primary resource **/data**:

```
cloud-user@pod-name-cm-1:~$ sudo systemctl start keepalived
cloud-user@pod-name-cm-1:~$ sudo systemctl status keepalived

cloud-user@pod-name-cm-2:~$ sudo systemctl start keepalived
cloud-user@pod-name-cm-2:~$ sudo systemctl status keepalived
```

Check the DRBD status on CM-1 and CM-2. It must be transformed to the correct cluster state by
now.

```
cloud-user@pod-name-deployer-cm-1:~$ drbd-overview
0:data/0 Connected Primary/Secondary UpToDate/UpToDate /mnt/stateful_partition ext4 568G 364G
```

```
176G 68%
cloud-user@pod-name-deployer-cm-2:~$ drbd-overview
0:data/0 Connected Secondary/Primary UpToDate/UpToDate Move the CM VIP to CM-1 from CM-2 and
make CM-1 as primary
```

/data is mounted only on the primary node.

```
cloud-user@pod-name-deployer-cm-1:~$ df -h
Filesystem Size Used Avail Use% Mounted on
udev 189G 0 189G 0% /dev
tmpfs 38G 22M 38G 1% /run
/dev/sda1 9.8G 3.5G 5.9G 37% /
tmpfs 189G 0 189G 0% /dev/shm
tmpfs 5.0M 0 5.0M 0% /run/lock
tmpfs 189G 0 189G 0% /sys/fs/cgroup
/dev/sda4 71G 1.5G 66G 3% /tmp
/dev/sda3 71G 11G 57G 16% /var/log
/dev/drbd0 568G 364G 175G 68% /mnt/stateful_partition
tmpfs 38G 0 38G 0% /run/user/1000


cloud-user@pod-name-deployer-cm-secondary:~$ df -h
Filesystem Size Used Avail Use% Mounted on
udev 189G 0 189G 0% /dev
tmpfs 38G 2.3M 38G 1% /run
/dev/sda1 9.8G 2.0G 7.3G 22% /
tmpfs 189G 0 189G 0% /dev/shm
tmpfs 5.0M 0 5.0M 0% /run/lock
tmpfs 189G 0 189G 0% /sys/fs/cgroup
/dev/sda3 71G 9.3G 58G 14% /var/log
/dev/sda4 71G 53M 67G 1% /tmp
tmpfs 38G 0 38G 0% /run/user/1000
```