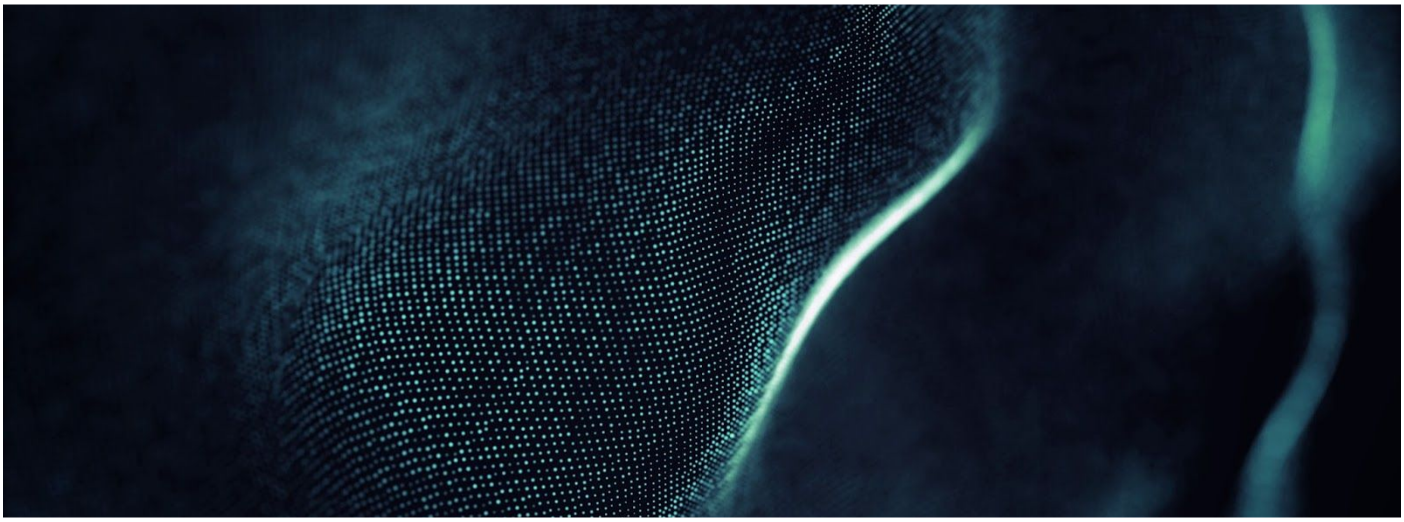
NYRIAD™



U90 Primary/Secondary HA (Active / Passive)

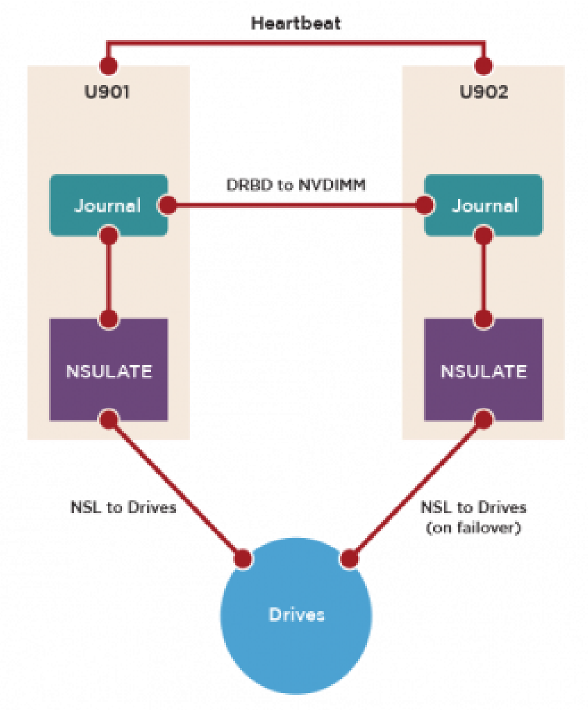
Software Setup Guide

Overview

High Availability Configuration with NSULATE

High Availability can be used to help you minimize downtime experienced by your users. NSULATE can be configured to run on multiple systems in a high availability configuration.

NSULATE is configured on two nodes (u901, u902), where the journal device of each NSULATE array is replicated with DRBD (Distributed Replicated Block Device). The first node is active while the second node is passive, on standby. The passive server acts as a failover node that's ready to take over operation across the same drives, as soon as the first node fails. This configuration is described in the following diagram.



It's important that the two nodes have the same settings. If changes are made on the active node, those changes must be replicated on the passive, failover, node. This ensures that clients won't be able to tell the difference when the failover node takes over.

High Availability Configuration Setup

Hardware Required

Inventec U90G3 series are 4U height ultra-dense storage servers, supporting up to seventy-bay 3.5" large-form-factor hard disk drives and dual server nodes of two-socket mainstream Intel® Xeon® processor E5 v3/v4 family. U90G3 storage servers feature 12G SAS interface and dual domain, supporting two HDD control configurations - single node accessing all 70 drives for best price per drive catering to cold storage usage, or two nodes accessing all 70 drives for those who need failover.

U90G3 - 4U 70 bay Intel® E5 2600 v3/v4 Ultra Dense Storage System

Software Required

NSULATE, the block device.

Heartbeat a subsystem that allows a primary and a back-up Linux server to determine if the other is 'alive' and if the primary isn't, fail over resources to the backup.

DRBD is a kernel block-level synchronous replication facility which serves as an imported shared-nothing cluster building block.

The NFS kernel server is the in-kernel Linux NFS daemon. It serves locally mounted file systems out to clients via the NFS network protocol.

System Specifications

		Comments
Chassis	Inventec U90	Dual Motherboard, 4U, http://ebg.inventec.com/product/info?id=U90G3
Operating System	Ubuntu Server 16.04.5	4.15 Kernel
CPU	Intel Xeon CPU E5-2620 v4 @ 2.10GHz x 2 per node	
GPU	Nvidia Tesla P4 x 1 per node	
RAM	8GB 2667Mhz x 8 per node	
PSU	1400W (220V) Platinum (2+2 redundancy)	Two Slots are active, and two are for fail over purposes
Hard drives	70 x 2 TB HDDs, 2 x 250GB SSDs per node	The 70 hard drives are shared between each node, while the SSDs can be configured in RAID1 for the operating system

Software Setup

NOTE: commands will need to be executed on both machines, unless otherwise specified.

First install Ubuntu Server 16.04.5 separately onto both of the U90 compute servers. This installation does not require any additional configuration.

Then install the following dependencies:

Cuda 9.2

First, add the Nvidia repository from the Nvidia website, and install CUDA 9.2 :

```
wget
https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1604/x86_64/cuda-r
epo-ubuntu1604_9.2.88-1_amd64.deb

dpkg -i cuda-repo-ubuntu1604_9.2.88-1_amd64.deb

apt-key adv --fetch-keys
http://developer.download.nvidia.com/compute/cuda/repos/ubuntu1604/x86_64/7fa2af8
0.pub

apt-get update

apt-get install cuda
```

Once CUDA is installed, run the following commands to extend the PATH environment variable to contain CUDA-9.2.

```
export PATH=/usr/local/cuda-9.2/bin${PATH:+:${PATH}}

echo "PATH=/usr/local/cuda-9.2/bin${PATH:+:${PATH}}" >>  ~/.profile
```

NSULATE

To install NSULATE, first download and install the latest package from the Nyriad Partner Centre. More information can be found in the NSULATE User Manual.

```
apt install -y NSULATE-1.2.*-beta.deb
```

DRBD

```
apt-get install drbd8-utils
```

Heartbeat

```
apt-get install heartbeat
```

Setting up DRBD

First set up a DRBD resource file. This file holds information such as which disks to use, and server names / IP addresses.

The resource file will be saved in :

```
/etc/drbd.d/
```

Here is an example DRBD file created at `/etc/drbd.d/u90.res` for the two U90 servers we set up:

```
# This is the name of the resource file. This will be used for DRDB functions &
# commands eg. drbdadm cstate u90
resource u90 {
    volume 0 {
        # This is the DRDB device that will be used to access the files
        device /dev/drbd0;
    }
    # Node 1's hostname
    on prod-u901 {
        # The IP address and DRDB port of node 2
        address 10.15.21.151:7789;

        # The drive to replicate, In this case we want to use the Journal devices
        # we provide to NSULATE, such as /dev/pmem0 when using NVDIMMs
        disk /dev/disk/by-id/ata-OCZ-TRION100_X5BB52G2KMCX-part1;

        # DRDB journal drive, where we want DRDB to store its metadata
        meta-disk /dev/disk/by-id/ata-OCZ-TRION100_X5BB52G2KMCX-part5;
    }
    # Node 2's hostname
    on prod-u902 {
        # The IP address and DRDB port of Node 2
        address 10.15.21.92:7789;

        # The drive data is replicated to, /dev/pmem0 when using NVDIMMs
        disk /dev/disk/by-id/ata-OCZ-TRION100_X5BB52KQKMCX-part1;

        # DRDB journal drive, where we want DRDB to store its metadata
        meta-disk /dev/disk/by-id/ata-OCZ-TRION100_X5BB52KQKMCX-part5;
    }
}
```

To create the metadata for DRBD use the following command :

```
# Where 'u90' is the resource name created in the previous step
drbdadm create-md u90
```

Bring DRDB up using the resource name :

```
drbdadm up u90
```

Before we set a node to primary we need to set up NSULATE as we will be using the DRBD block device in our NSULATE array.

Setting up NSULATE

As we will be putting the DRBD block device (/dev/drbd0) into the array, we need to set up NSULATE after we have set up DRBD.

First, we will want to get all the hard drives that will be going into our array and put the drive IDs into a text document so we can pass this through to NSULATE when we create the array.

```
#"ST2000NP0011" is the ID of the hard drive used, replace this to suit your needs
ls /dev/disk/by-id/* | grep -e "ST2000NP0011" > ~/devices.conf
```

The basic command to create an NSULATE array is :

```
NSULATE create -n <array_name> -k <data_drive_count> -m <parity_drive_count> -c
<compute_platform> -j <journal_device> <devices>
```

Since we are using DRBD to replicate the journal device, we want to use the DRBD block device as the journal device inside the NSULATE array. Before we can use DRBD in the array we need to set one node to primary so we can write to the DRBD device.

Create the array on the secondary node, and make the secondary node primary by using:

```
drbdadm primary u90
```

This will cause DRBD to start syncing the drives. Since there should be no data, it should not take a long time. It needs to complete before moving on to the next step.

To see the status of the DRBD sync type the following.

```
drbd-overview
```

Once the node is primary and the files are synced, create the array.

```
# The $(cat ~/devices.conf) gets the file we created in the last step, this is a
list of hard drives to use as the data drives
NSULATE create -n narray1 -k 60 -m 10 -c gpu -j /dev/drbd0 $(cat ~/devices.conf)
```

Stop the NSULATE service, and make the node secondary.

```
service nsulated stop

drbdadm secondary u90
```

On the node that is to actually be primary type:

```
drbdadm primary u90
```

Then create the NSULATE array:

```
# $(cat~/devices.conf) is a list of hard drives to use as the data drives
NSULATE create -n narray1 -k 1 -m 1 -c gpu -j /dev/drbd0 $(cat~/devices.conf)
```

Setting up Heartbeat

Heartbeat used to detect if a node is available or not and to pass the resources between nodes, such as IP address and DRBD control.

Heartbeat requires a configuration file named ha.cf.

```
nano /etc/heartbeat/ha.cf
```

In the ha.cf file put:

```
logfacility local0


keepalive 2

# This is the time, in seconds, for how long the node must be offline before it
switches to the second node
deadtime 10

# This is the address of the Ethernet card being using
bcast ens3f0

# These are the hostnames of the nodes
node prod-u901 prod-u902
```

For NSULATE to pick up the array during a failover, heartbeat needs a script that will start NSULATE back up on the secondary node.

On both nodes create a file called startNSULATE.

```
nano /etc/init.d/startNSULATE
```

Inside this file put the following:

```
#!/bin/bash

service nsulated start
exit 0
```

Make the file executable with :

```
chmod u+x /etc/init.d/startNSULATE
```

Create the haresources file, which controls the floating IP address, the DRBD take over, and the start of our startNSULATE script

```
nano /etc/heartbeat/haresources
```

In this file we put the following:

```
prod-u901 IPaddr::10.15.21.233/24/ens3f0 drbddisk::u90 Filesystem::/dev/drbd0
startNSULATE
```

Set up the heartbeat authkeys file so heartbeat is able to authenticate each node:

```
nano /etc/heartbeat/authkeys
```

In this file we choose a random string that will be used as the authentication string:

```
auth 3
3 md5 somerandomstring
```

Change the access of the authkeys file so only root can view it by running:

```
chmod 600 /etc/heartbeat/authkeys
```

After everything is set up, use the following commands to start the services (they may or may not already be running).

```
/etc/init.d/drbd start

/etc/init.d/heartbeat start
```

Failover recovery

To start the failed machine up again, make sure NSULATE is not running by running:

```
service nsulated stop
```

Make sure DRBD is in secondary:

```
drbdadm secondary u90
```

Make sure heartbeat is running by typing:

```
/etc/init.d/heartbeat status
```

If heartbeat is not running type:

```
/etc/init.d/heartbeat restart
```

Summary

This guide will help you set up the Inventec u90 so it is a primary/secondary high availability failover machine sharing data drives, with NSULATE.

Helpful Commands

NSULATE

```
# This displays the NSULATE help
NSULATE help

# This displays details of the current hardware/drives to use
NSULATE detail platform

# Displays information about arrays/information on a specific array
NSULATE display <array_name_optional>

# Displays if the NSULATE daemon is running
NSULATE status

# Shows the current GPU/CPU utilization
NSULATE stat
```

DRBD

```
# Shows the current connection state of drbd
drbdadm cstate <name>

# Shows what role the current node is on
drbdadm role <name>

# Shows information on connection,roles,data sent/received, syncing and time
remaining for syncing
cat /proc/drbd

# Shows information on connect,roles, sent/received and syncing
drbd-overview

# Will start/restart/stop the drbd service
/etc/init.d/drbd start/restart/stop
```

Heartbeat

```
# Starts/restarts/stops the heartbeat service
/etc/init.d/heartbeat start/restart/stop

# Shows the status of the heartbeat service
/etc/init.d/heartbeat status
```