Highly Available NFS Storage with DRBD and Pacemaker

SUSE Linux Enterprise High Availability Extension 12 SP4

Tanja Roth and Thomas Schraitle

This document describes how to set up highly available NFS storage in a two-node cluster, using the following components of SUSE Linux Enterprise High Availability Extension 12 SP4: DRBD* (Distributed Replicated Block Device), LVM (Logical Volume Manager), and Pacemaker, the cluster resource management framework.

Publication Date: November 13, 2020

Contents

1 Usage Scenario 2
2 Installing a Basic Two-Node Cluster 2
3 Creating an LVM Device 2
4 Creating a DRBD Device 3
5 Adjusting Pacemaker's Configuration 6
6 Creating Cluster Resources 7
7 Using the NFS Service 10
8 GNU Free Documentation License 10
1 Usage Scenario

This document will help you set up a highly available NFS server. The cluster used to for the highly available NFS storage has the following properties:

- Two nodes: alice (IP: 192.168.1.1) and bob (IP: 192.168.1.2), connected to each other via network.
- Two floating, virtual IP addresses (192.168.1.10 and 192.168.2.1), allowing clients to connect to the service no matter which physical node it is running on. One IP address is used for cluster administration with Hawk2, the other IP address is used exclusively for the NFS exports.
- A shared storage device, used as an SBD fencing mechanism. This avoids split brain scenarios.
- Failover of resources from one node to the other if the active host breaks down (active/passive setup).
- Local storage on each host. The data is synchronized between the hosts using DRBD on top of LVM.
- A file system exported through NFS.

After installing and setting up the basic two-node cluster, and extending it with storage and cluster resources for NFS, you will have a highly available NFS storage server.

2 Installing a Basic Two-Node Cluster

Before you proceed, install and set up a basic two-node cluster. This task is described in Installation and Setup Quick Start. The Installation and Setup Quick Start describes how to use the ha-cluster-bootstrap package to set up a cluster with minimal effort.

3 Creating an LVM Device

LVM (Logical Volume Manager) enables flexible distribution of hard disk space over several file systems.
To prepare your disks for LVM, do the following:

1. Create an LVM volume group and replace `/dev/sdbX` with your corresponding device for LVM:
   
   ```
   root # pvcreate /dev/sdbX
   ```

2. Create an LVM Volume Group `nfs` that includes this physical volume:
   
   ```
   root # vgcreate nfs /dev/sdbX
   ```

3. Create one or more logical volumes in the volume group `nfs`. This example assumes a 20 gigabyte volume, named `work`:
   
   ```
   root # lvcreate -n work -L 20G nfs
   ```

4. Activate the volume group:
   
   ```
   root # vgchange -ay nfs
   ```

After you have successfully executed the above steps, your system will make visible the following device: `/dev/VOLGROUP/LOGICAL_VOLUME`. In this case it will be `/dev/nfs/work`.

### 4 Creating a DRBD Device

This section describes how to set up a DRBD device on top of LVM. The configuration of LVM as a back-end of DRBD has some benefits:

- Easier setup than with LVM on top of DRBD.
- Easier administration in case the LVM disks need to be resized or more disks are added to the volume group.

As the LVM volume group is named `nfs`, the DRBD resource uses the same name.
4.1  Creating DRBD Configuration

For consistency reasons, it is highly recommended to follow this advice:

- Use the directory `/etc/drbd.d/` for your configuration.
- Name the file according to the purpose of the resource.
- Put your resource configuration in a file with a `.res` extension. In the following examples, the file `/etc/drbd.d/nfs.res` is used.

Proceed as follows:

PROCEDURE 1: CREATING A DRBD CONFIGURATION

1. Create the file `/etc/drbd.d/nfs.res` with the following contents:

   ```
   resource nfs {
   device /dev/drbd0;
   disk   /dev/nfs/work;
   meta-disk internal;

   net {
   protocol C;
   }

   connection-mesh {
   hosts   alice bob;
   }

   on alice {
   address 192.168.1.1:7790;
   node-id 0;
   }

   on bob {
   address 192.168.1.2:7790;
   node-id 1;
   }
   }
   ```

1. The DRBD device that applications are supposed to access.
2. The lower-level block device used by DRBD to store the actual data. This is the LVM device that was created in Section 3, "Creating an LVM Device".
3. Where the metadata format is stored. Using `internal`, the metadata is stored together with the user data on the same device. See the man page for further information.
The specified protocol to be used for this connection. For protocol C, a write is considered to be complete when it has reached all disks, be they local or remote.

Defines all nodes of a mesh. The `hosts` parameter contains all host names that share the same DRBD setup.

Contains the IP address and a unique identifier for each node.

2. Open `/etc/csync2/csnc2.cfg` and check whether the following two lines exist:

```
include /etc/drbd.conf;
include /etc/drbd.d/*.res;
```

If not, add them to the file.

3. Copy the file to the other nodes:

```
root # csync2 -xv
```

For information about Csync2, refer to Book “Administration Guide”, Chapter 4 “Using the YaST Cluster Module”, Section 4.5 “Transferring the Configuration to All Nodes”.

4.2 Activating the DRBD Device

After you have prepared your DRBD configuration, proceed as follows:

1. If you use a firewall in your cluster, open port 7790 in your firewall configuration.

2. The first time you do this, execute the following commands on both nodes (in our example, `alice` and `bob`):

```
root # drbdadm create-md nfs
root # drbdadm up nfs
```

   This initializes the metadata storage and creates the `/dev/drbd0` device.

3. If the DRBD devices on all nodes have the same data, skip the initial resynchronization. Use the following command:

```
root # drbdadm new-current-uuid --clear-bitmap nfs/0
```

4. Make `alice` primary:

```
root # drbdadm primary --force nfs
```
5. Check the DRBD status:

```
root # drbdadm status nfs
```

This returns the following message:

```
nfs role:Primary
disk:UpToDate
alice role:Secondary
peer-disk:UpToDate
```

After the synchronization is complete, you can access the DRBD resource on the block device `/dev/drbd0`. Use this device for creating your file system. Find more information about DRBD in Book "Administration Guide", Chapter 20 "DRBD".

4.3 Creating the File System

After you have finished Section 4.2, “Activating the DRBD Device”, you should see a DRBD device on `/dev/drbd0`:

```
root # mkfs.ext3 /dev/drbd0
```

5 Adjusting Pacemaker's Configuration

A resource might fail back to its original node when that node is back online and in the cluster. To prevent a resource from failing back to the node that it was running on, or to specify a different node for the resource to fail back to, change its resource stickiness value. You can either specify resource stickiness when you are creating a resource or afterward.

To adjust the option, open the crm shell as `root` (or any non-`root` user that is part of the `haclient` group) and run the following commands:

```
root # crm configure
crm(live)configure# rsc_defaults resource-stickiness="200"
crm(live)configure# commit
```

For more information about global cluster options, refer to Book "Administration Guide", Chapter 6 “Configuration and Administration Basics", Section 6.2 “Quorum Determination".
6 Creating Cluster Resources

The following sections cover the configuration of the required resources for a highly available NFS cluster. The configuration steps use the crm shell. The following list shows the necessary cluster resources:

OVERVIEW OF CLUSTER RESOURCES

DRBD Primitive and Multi-state Resource
These resources are used to replicate data. The multi-state resource is switched from and to the Primary and Secondary roles as deemed necessary by the cluster resource manager.

NFS Kernel Server Resource
With this resource, Pacemaker ensures that the NFS server daemons are always available.

NFS Exports
One or more NFS exports, typically corresponding to the file system.

EXAMPLE NFS SCENARIO

- The following configuration examples assume that 192.168.2.1 is the virtual IP address to use for an NFS server which serves clients in the 192.168.2.x/24 subnet.

- The service exports data served from /srv/nfs/work.

- Into this export directory, the cluster will mount ext3 file systems from the DRBD device /dev/drbd0. This DRBD device sits on top of an LVM logical volume with the name nfs.

6.1 DRBD Primitive and Multi-state Resource

To configure these resources, run the following commands from the crm shell:

```bash
crm(live)# configure
crm(live)configure# primitive drbd_nfs \
    ocf:linbit:drbd \ 
    params drbd_resource="nfs" \ 
    op monitor interval="15" role="Master" \ 
    op monitor interval="30" role="Slave"

crm(live)configure# ms ms-drbd_nfs drbd_nfs \ 
    meta master-max="1" master-node-max="1" clone-max="2" \ 
    clone-node-max="1" notify="true"

crm(live)configure# commit
```
This will create a Pacemaker multi-state resource corresponding to the DRBD resource `nfs`. Pacemaker should now activate your DRBD resource on both nodes and promote it to the master role on one of them.

Check the state of the cluster with the `crm status` command, or run `drbdadm status`.

### 6.2 NFS Kernel Server Resource

In the `crm` shell, the resource for the NFS server daemons must be configured as a `clone` of a `systemd` resource type.

```bash
crm(live)configure# primitive nfsserver \
    systemd:nfs-server \
    op monitor interval="30s"
crm(live)configure# clone cl-nfsserver nfsserver \
    meta interleave=true
crm(live)configure# commit
```

After you have committed this configuration, Pacemaker should start the NFS Kernel server processes on both nodes.

### 6.3 File System Resource

1. Configure the file system type resource as follows (but do not commit this configuration yet):

   ```bash
   crm(live)configure# primitive fs_work \
   ocf:heartbeat:Filesystem \
   params device=/dev/drbd0 \
   directory=/srv/nfs/work \
   fstype=ext3 \
   op monitor interval="10s"
   ```

2. Combine these resources into a Pacemaker resource `group`:

   ```bash
   crm(live)configure# group g-nfs fs_work
   ```

3. Add the following constraints to make sure that the group is started on the same node on which the DRBD multi-state resource is in the master role:

   ```bash
   crm(live)configure# order o-drbd_before_nfs inf: \
   ```
4. Commit this configuration:

```
crm(live)configure# commit
```

After these changes have been committed, Pacemaker mounts the DRBD device to `/srv/nfs/work` on the same node. Confirm this with `mount` (or by looking at `/proc/mounts`).

6.4 NFS Export Resources

When your DRBD, LVM, and file system resources are working properly, continue with the resources managing your NFS exports. To create highly available NFS export resources, use the `exportfs` resource type.

To export the `/srv/nfs/work` directory to clients, use the following primitive:

1. Create NFS exports with the following commands:

```
crm(live)configure# primitive exportfs_work \
ocf:heartbeat:exportfs \
    params directory="/srv/nfs/work" \ 
    options="rw,mountpoint" \ 
    clientspec="192.168.2.0/24" \ 
    wait_for_leasetime_on_stop=true \ 
    fsid=100 \ 
    op monitor interval="30s"
```

2. After you have created these resources, append them to the existing `g-nfs` resource group:

```
crm(live)configure# modgroup g-nfs add exportfs_work
```

3. Commit this configuration:

```
crm(live)configure# commit
```

Pacemaker will export the NFS virtual file system root and the two other exports.

4. Confirm that the NFS exports are set up properly:

```
root # exportfs -v 
/srv/nfs/work  IP_ADDRESS_OF_CLIENT(OPTIONS)
```
6.5 Virtual IP Address for NFS Exports

The initial installation creates an administrative virtual IP address for Hawk2. Although you could use this IP address for your NFS exports too, create another one exclusively for NFS exports. This makes it easier to apply security restrictions later. Use the following commands in the crm shell:

```bash
crm(live)configure# primitive vip_nfs IPaddr2 
    params ip=192.168.2.1 cidr_netmask=24 
    op monitor interval=10 timeout=20
crm(live)configure# modgroup g-nfs add vip_nfs
crm(live)configure# commit
```

7 Using the NFS Service

This section outlines how to use the highly available NFS service from an NFS client.

To connect to the NFS service, make sure to use the **virtual IP address** to connect to the cluster rather than a physical IP configured on one of the cluster nodes' network interfaces. For compatibility reasons, use the **full** path of the NFS export on the server.

In its simplest form, the command to mount the NFS export looks like this:

```bash
root # mount -t nfs 192.168.2.1:/srv/nfs/work /home/work
```

To configure a specific transport protocol (**proto**) and maximum read and write request sizes (**rsize** and **wsize**), use:

```bash
root # mount -o rsize=32768,wsize=32768 
    192.168.2.1:/srv/nfs/work /home/work
```

In case you need to be compatible with NFS version 3, include the value **vers=3** after the **-o** option.

For further NFS mount options, consult the **nfs** man page.
We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language. A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Invariant Section then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are, listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, TeXinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and GIF. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History.".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Covers Texts, you must envelope the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy those conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages. If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use as title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more Persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and all Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.
9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

Copyright (c) YEAR YOUR NAME.
Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.