Frequently Asked Questions about DRBD

Contents

General Issues

Please have a look at some of the publications[1] and documentation.

What is DRBD, to begin with?

DRBD[2], developed by PhilippReisner[3] and LarsEllenberg[4], is a
Distributed
Replicated
Block
Device
for the Linux operating system. It allows to have a realtime mirror of your local block devices on a remote machine. In conjunction with heartbeat it allows to create HA (high availability) Linux clusters.

Which license conditions apply to DRBD?

DRBD is released under the GNU GENERAL PUBLIC LICENSE Version 2, Juni 1991 (GPL. Thus, within the conditions of this license it can be freely distributed and modified.)

Where do I get Support for DRBD

At LinBit[5].

Where can I download DRBD?

The latest version of DRBD is available for download from LinBit[6] resp. from drbd.org[7]. DRBD is also included in many Linux distributions, like Debian, SuSE, RedHat and others.

Compiling Issues

Compiling fails with "structure has no member named `nice' " in drbd_syncer.c?

If you mean something similar to this:

```
drbd_syncer.c: In function `drbd_syncer':
drbd_syncer.c:409: structure has no member named `nice'
drbd_syncer.c:439: structure has no member named `nice'
drbd_syncer.c:452: structure has no member named `nice'
make[1]: *** [drbd_syncer.o] Error 1
make[1]: Leaving directory `/usr/local/src/drbd-0.6.3/drbd'
make: *** [all] Error 2
```

have a look into `drbd_config.h` and modify the line starting with `//#define SIGHAND_HACK` so that it reads `#define SIGHAND_HACK`.

Setup and Installation

Can I encrypt/compress the exchanged data?

Of course. But this is no option within DRBD. You 'just' need to setup some VPN, then the network stack will take care of that. For a lightweight solution, have a look at the CIPE project[8]. Of course, IPSEC or OpenVVPN will do, too.

Can I mount the secondary at least readonly?

Short answer: ***No!*** DRBD would not care, but most likely your filesystem will be confused because it will not be aware about changes in the underlying device. This in general means that it cannot work, not with ext2, ext3, reiserFS, JFS or XFS. If you need not just a mirrored, but a shared filesystem, use OCFS2 or GFS or OpenGFS[9] for example. But these are much slower, and typically expect write access on all nodes in question. If we have more than one node concurrently modifying distributed devices, we have some "interessting" problems to decide which part of the device is up-to-date on which node, and what blocks need to be resynchronized in which direction. These are the most important reasons why DRBD does not allow mounting the secondary. Thus, if you want to mount the secondary, set the secondary as the primary first. Both devices mounted at the same time does not work. Actually, DRBD v8 **does** support two Primaries, see the next answer. If you need access to the data from both nodes, and and arbitrary number of other clients, consider using HaNFS[10].

Why does DRBD not allow concurrent access from all nodes? I'd like to use it with GFS/OCFS2...

Actually, DRBD version 8.0.x and later support this. You need to `net { allow-two-primaries; }` ... Also have a look at the DRBD Changelog[11]. An other option would be to have only one node active, export that device via iSCSI, then run OCFS2 on iSCSI.

Can DRBD use two devices of different size?

Generally yes, but there are some issues to consider:
Locally DRBD uses the configured disk-size, which has to be <= physical, and if not given its is set to the physical size. On connect the device size will be set to the minimum of both nodes. And here you could run into problems, if you do things without common sense: if you first use drbd on one node only, without disk-size configured properly, and later connect a node with smaller device size, then the drbd device size shrinks at runtime. you should find a message about *Your size hint is bogus, please change to <some value>* in the syslog in that case. This will confuse the file system on top of your device. Thus, if your device sizes differ, set the size to be used for DRBD explicitely. **DRBD-0.7** stores information about the peers device size in its local meta data, therefore usage of disk-size is deprecated (and is disallowed in the configuration file).

Can XFS be used with DRBD?

XFS uses dynamic block size, thus DRBD 0.7 or later is needed.

When I try to load the drbd module, I am gettin the following error: compiled

for kernel version "some version" while this kernel is "some other version"

The settings for your actual kernel and the `.config` for the kernel source against which drbd was build do not match. On SuSE Linux you can get the right config with the following commands:
`cd /usr/src/linux/ && make cloneconfig && make dep` Ususally, you do not have to recompile your kernel, just drbd. But read INSTALL in the drbd tgz, to learn how to do it the proper way.

Can I use DRBD with LVM?

Yes. With LVM2, snapshots are writeable. So you can replay the journal on the snapshot. But see also A Summary of LVM snapshots with DRBD[12] posted on 2004-04-08 in drbd-user.

I use DRBD and Linux-vServer, and I cannot umount anymore: "file system in use"

Maybe http://linux-vserver.org/advanced+DRBD+mount+issues[13] helps.

What about Xen, DRBD and iSCSI?

Always interesting discussions on http://lists.xensource.com/archives/html/xen-users/[14]

Can I use DRBD with OpenVZ?

See http://wiki.openvz.org/HA_cluster_with_DRBD_and_Heartbeat[15]

Operation Issues

Why is Synchronization (SyncingAll) so slow?

Outdated, applies to drbd versions prior drbd-0.6.4 only For historical reasons replicate used to work backwards. Most physical devices do have a pretty slow thoughput when writing data backwards.

How can I speed up the Synchronization performance?

- double check the value of `sync-max` in the `net {}` section (drbd-0.6) resp. `rate` in the `syncer {}` section (drbd-0.7). Keep in mind that the default value is very low, and the default unit is kByte/sec!
- if you run on top of some local RAID, make sure it is not reconstructing at the same time
- check whether DMA is enabled

How can I speed up write throughput?

First you need to find the bottleneck. This can be your local disk, the network, the remote disk, latency caused by excessive seeks, or the summed up latency of those components.
You may want to play with the values of `protocol` and `sndbuf-size`. If your NIC supports it, you may want to enable "jumbo frames" (up the value of the MTU). If nothing helps, ask the list for known good and performant setups...

Why is my "load average" that high?

Load average is defined as average number of processes in the runqueue during a given interval. A process is in the run queue, if it is

- not waiting for external events (e.g. select on some fd)
- not waiting on its own (not called "wait" explicitly)
- not stopped

Note that all processes **waiting for disk io** are counted as runable! Therefore, if a lot of processes wait for disk io, the "load average" goes straight up, though the system actually may be almost idle cpu-wise ... E.g. crash your nfs server, and start 100 ls /path/to/non-cached/dir/on/nfs/mount-point on a client... you get a "load average" of 100+ for as long as the nfs timeout, which might be weeks ... though the cpu does nothing. Verify your system load by other means, e.g. vmstat, sysstat/sar. This will give you an idea of the bottleneck of your system. Some ideas are using multiple disks (not just partitions!) or even a RAID with 10.000rpm SCSI disks and probably even a Gigabit Ethernet. Even on a Fast Ethernet device you will rarely see more then 6 MByte per second. (100 MBit/s is at most 12.5 MByte/s minus protocol overhead and latency etc.).

What is warning: "Return code 255 from /etc/ha.d/resource.d/datadisk" telling me when using the datadisk script with heartbeat?

DRBD-0.6 only
Exit code 255 is most likely from a script generated `die`, which include a verbose error message. Capture the output of that script. this is the debugfile directive in your ha.cf, iirc. If that does not help, do it by hand, and see what error message it gives. datadisk says something like *cannot promote to primary, sychronization running* or *fsck failed* or ...

When the node goes from secondary to primary the drbd device will not be mounted on the primary. Manually mounting works.

Feature ...
DRBD does not automaticaly mount the partition. The script datadisk (or drbddisk in 0.7) is made for that purpose. It is intended to be called by hartbeat.

When using large devices, the drbd script (0.6.X) fails with "usage: multiply $N by $F"

The exact error output looks like:

```
drbd: pre-parsed needs update, parsing /etc/drbd.conf
.............................
drbd: usage: multiply $N by $F
drbd:   stopped at /etc/init.d/drbd (multiply) line 1026.
drbd:
drbd:   CVSID Id: drbd,v 1.55 2004/03/04 07:38:18 lars Exp
drbd:   BASH_VERSINFO (2 05b 0 1 release i386-pc-linux-gnu)
drbd:
drbd:   If you feel this script is buggy, please report to
drbd:   Lars Ellenberg <l.g.e@web.de>, drbd-user@lists.linbit.com
drbd:
```

The problem is an integer overflow in the `blockdev` tool. Going to search.gmane.org, and searching for *stopped drbd multiply* in group *drbd*, sorted by relevance gives this[16], and then you'll soon find this thread[17] I updated the script to no longer use blockdev, but fdisk -s, which is what I should have done from the begining; sometimes one just does not see the obvious. This and several other small changes meanwhile have been released as drbd-0.6.13[18], if for whatever reason you don't want to update it all, you can of course download the script only[19]. You then need to install/copy it to the right location, of course, i.e. probably `/etc/init.d/drbd`, and maybe copy or relink it to `datadisk` in case datadisk is not a symlink on your box.

When heartbeat is started on node1 and tries to run the datadisk script in order to become the primary node, it fails because it is still synchronizing.

DRBD 0.6 only
Enable `nice_failback on` in `ha.cf`. Though typically you really want to have the drbd init script block until it is fully synchronized, at least with DRBD-0.6, see the example `drbd.conf`.

What is warning: "out of vmalloc space"

For each device, drbd will (try to) allocate X MB of bitmap, plus some constant amount (<1MB). X = storage_size_in_GB/32, so 1 TB storage -> 32 MB bitmap.

By default Linux allocates 128MB to Vmalloc. For systems using more than 4TB, this may cause an issue.

If you get the following error message in /var/log/messages, Try a Linux 2.6 hugemem kernel.

```
kernel: allocation failed: out of vmalloc space - use vmalloc=<size> to increase size.
```

What do the fields like st, ns, nr, dw, dr etc. in /proc/drdb mean?

cs
connection state

Unconfigured	Device waits for configuration.
StandAlone	Not trying to connect to peer, IO requests are only passed on locally.
Unconnected	Transitory state, while bind() blocks.
WFConnection	Device waits for configuration of other side.
WFReportParams	Transitory state, while waiting for first packet on a new TCP connection.
Connected	Everything is fine.
Timeout, BrokenPipe, NetworkFailure	Transitory states when connection was lost.
DRBD-0.6 specific	
SyncingAll	All blocks of the primary node are being copied to the secondary node.

SyncingQuick	The secondary is updated, by copying the blocks which were updated since the now secondary node has left the cluster.
SyncPaused	Sync of this device has paused while higher priority (lower `sync-group` value) device is resyncing.

DRBD-0.7 / DRBD-8; trailing *S* or *T* indicates this node is SyncSource or SyncTarget, respectively.

WFBitMap{S,T}	Transitory state when synchronization starts; "dirty"-bits are exchanged.
SyncSource	Synchronization in progress, this node has the good data.
SyncTarget	Synchronization in progress, this node has inconsistent data.
PausedSync{S,T}	see SyncPaused.
SkippedSync{S,T}	you should never see this. "*Developers only*"

st:Local/Remote
state, the respective node's role for this device.

Primary	the active node; may access the device.
Secondary	the passive node; **must not access the device**; expects mirrored writes from the other node.
Unconfigured	this is not a role, obviously.

ld
local data consistentency (DRBD-0.7, DRBD 8)
ns,nr,dw,dr,...
statistic counters in number of blocks (1KB) respectively number of requests

- **ns** network send
- **nr** network receive
- **dw** disk write
- **dr** disk read
- **al** activity log updates (0.7)
- **bm** bitmap updates (0.7)
- **lo** reference count on local device
- **pe** pending (waiting for ack)
- **ua** unack'd (still need to send ack)
- **ap** application requests expecting io-completion

References

[1] http://www.drbd.org/publications.html
[2] http://www.linux-ha.org/DRBD
[3] http://www.linux-ha.org/PhilippReisner
[4] http://www.linux-ha.org/LarsEllenberg
[5] http://www.linbit.com/en/drbd/drbd/support/
[6] http://www.linbit.com/support/drbd-current/
[7] http://www.drbd.org/download.html
[8] http://sites.inka.de/~bigred/devel/cipe-faq.html
[9] http://opengfs.sourceforge.net/
[10] http://www.linux-ha.org/HaNFS
[11] http://svn.drbd.org/drbd/trunk/ChangeLog
[12] http://thread.gmane.org/gmane.comp.linux.drbd/6175
[13] http://linux-vserver.org/advanced+DRBD+mount+issues
[14] http://lists.xensource.com/archives/html/xen-users/
[15] http://wiki.openvz.org/HA_cluster_with_DRBD_and_Heartbeat
[16] http://search.gmane.org/search.php?query=stopped+drbd+multiply&email=&group=drbd&sort=relevance
[17] http://thread.gmane.org/gmane.linux.network.drbd/5148
[18] http://svn.drbd.org/drbd/tags/drbd-0.6.13/
[19] http://svn.drbd.org/drbd/tags/drbd-0.6.13/scripts/drbd
