Subject: VZ+DRBD+HA and load balancing
Posted by luisdev on Wed, 27 Aug 2008 12:06:16 GMT
View Forum Message <> Reply to Message

Hi there!

This link explains the combination of VZ+DRBD+HA:

http://wiki.openvz.org/HA_cluster_with_DRBD_and_Heartbeat

However, as some other people have pointed in this forum, it would be nice if VEs could be shared between both machines so when both are operative the load would be balanced between them but when one of them is off-line the other took over its VEs.

How would one go about this issue? The biggest problem I see is how to merge both VZ configurations. Symbolic links are a possibility, or a combination filesystem like unionfs or aufs, provided one doesn't use the same VEIDs in both machines.

Some musings:

Consider HNs named cluster11 and cluster12.

There would be two DRBD resources, /vz11 and /vz12 and HA makes them primary in their respective HN and secondary in the other. If one of the HN fails, both will be primary in the remaining node.

/vz11 and /vz12 are configured as in the aforementioned wiki page. /vz is a symbolic link to either /vz11 or /vz12 respectively.

If, e.g., cluster11 fails, at cluster12:
- For every VE in /vz11/cluster/etc/sysconfig/vz-scripts/ that doesn't exist in /vz12/cluster/etc/sysconfig/vz-scripts/:
  - Link its conf, quota, private and root to /vz
  - If onboot=yes, start it.

When cluster11 comes back to life the procedure would be:
- Stop cluster11's VEs.
- Undo the links.
- Keep /vz11 as primary in cluster12 until it is synced at cluster11.
- Give back /vz11 as primary to cluster11.
- Start cluster11's VEs.

This could probably be automated with some HA mojo, I'd have to look at it.

Obvious drawbacks of this approach are:
- Load balancing is manual.
- Must keep close tabs on VEs so there are no collisions (e.g., name them 11xx and 12xx respectively, separate IP address range, etc.)

- Downtime of the VEs during failback (could cook up some live migration scheme, the key point is waiting until /vz11 is back in sync.)

My questions for the forum are:
- Is the merging of VEs via symbolic links a sound approach? It seems to work well here, but perhaps there is some potential pitfall lurking in the darkness.
- I am rather new to VZ+DRBD+HA. Is there a much better solution for this issue that I have overlooked? There is this LBVM project lbvm.sourceforge.net but I haven't heard of anyone using it, so I haven't investigated it yet.

Cheers,

L

Subject: Re: VZ+DRBD+HA and load balancing
Posted by golly on Wed, 27 Aug 2008 12:58:42 GMT
View Forum Message <> Reply to Message

I rolled my own Ubuntu Hardy cluster that I manually load-balance (for now). Once I get approval for it, it'll be going into production. Anyway, the three biggest pains I had was:

1. coming up with the VZ directory layout,
2. learning heartbeat's CRM XML, and
3. getting Heartbeat's ManageVE OCF running correctly.

For the first (VZ dir layout), the biggest issue was that the VE's /private and /root directories *have* to be on the same block device. When VZ starts a VE, it hardlinks /root to /private (along with other processing - but we won't go there now) and I don't know of a way to hardlink across block devices. (The OpenVZ Gurus probably know all about that, though.)

Anyway, once I settled on how to lay out my directories, I encoded that information in the default OpenVZ config file.

Second, learning heartbeat's CRM - read, read, read! Oh, and when you come across the CRM examples that have the DRBD <master_slave> configuration - you really do need it! I tried configuring DRBD in heartbeat without it at first (to see if I really, really needed it) and either nothing happened or only one DRBD side came up (i.e. it never sync'ed to the other node).

Third, on Ubuntu Hardy, Canonical symlinks /bin/sh to DASH instead of BASH. The ManageVE script starts with #!/bin/sh, but it uses a lot of BASH-isms, so it fails to run under heartbeat on Hardy (but runs fine with heartbeat's testing scripts). It took me three or four days to figure out that heartbeat was running ManageVE under DASH instead of BASH, but my cmdline was BASH so ManageVE would work fine.

Anyway, if you would like more details about how I did it, maybe I can put up a wiki page on it (if there is sufficient interest). Let me know.

Regards

Subject: Re: VZ+DRBD+HA and load balancing
Posted by luisdev on Wed, 27 Aug 2008 13:21:53 GMT
View Forum Message <> Reply to Message

golly wrote on Wed, 27 August 2008 08:58
I don't know of a way to hardlink across block devices.

This is simply not possible, since hard links are directory entries that refer to an already used inode. Since the inode table is unique for each block device, one can only link to inodes in the same device.

golly wrote on Wed, 27 August 2008 08:58
Anyway, once I settled on how to lay out my directories

How?

golly wrote on Wed, 27 August 2008 08:58
Second, learning heartbeat's CRM - read, read, read! Oh, and when you come across the CRM examples that have the DRBD <master_slave> configuration - you really do need it! I tried configuring DRBD in heartbeat without it at first (to see if I really, really needed it) and either nothing happened or only one DRBD side came up (i.e. it never sync'ed to the other node).

Ugh! So nothing of the simpler Heartbeat v.1 config, then?

golly wrote on Wed, 27 August 2008 08:58
Third, on Ubuntu Hardy, Canonical symlinks /bin/sh to DASH instead of BASH. The ManageVE script starts with #!/bin/sh, but it uses a lot of BASH-isms,

Nasty one.

golly wrote on Wed, 27 August 2008 08:58
Anyway, if you would like more details about how I did it, maybe I can put up a wiki page on it (if there is sufficient interest). Let me know.

By all means, yes, this issue has come up several times in this forum since it is a very useful setup. But if you don't have the time just attaching your config files would probably be enough to get started.

Subject: Re: VZ+DRBD+HA and load balancing
Posted by golly on Wed, 27 Aug 2008 14:36:24 GMT
View Forum Message <> Reply to Message

This is really quick and rough to give you an idea of what I did. I'll try to follow it up once I have all my notes arranged. I really *JUST* got done setting this up (as in I got it fully running *yesterday*...).

In laying out my directories, I wanted to set the nodes up so that I could recreate the nodes as quickly as possible in the event of a hardware failure. Therefore, I split files into those on a default install of Ubuntu server and those that I modified or that needed to be replicated between nodes.

So, once I had an install of Ubuntu Hardy server running, I set about using the /srv directory as the place where all of my "replication files" went. I mounted a 1G block device at /srv and will be setting up file replication with rsync for files *on that block device only* soon. I didn't want to use DRBD for the /srv block device because it has drbd.conf and heartbeat config files.

Anyway, my current /srv filesystem looks like this (not including files copied over from /etc or /var of the default install of drbd/heartbeat/openvz):

```
/srv
|-/heartbeat
| |-authkeys
| |-ha.cf
|
|-/openvz
| |-/conf
| | |-1008.conf -> ../ve/1008/openvz-ve.conf
| | |-1009.conf -> ../ve/1009/openvz-ve.conf
| |
| |-/cron
| |-/dists
| |-/dump
| |-/lock
| |-/names
| |
| |-/templates
| | |-/cache
| | | |-ubuntu-8.04-i386-minimal.tar.gz
| |
| |-/ve
| | |-/1008
| | |-/1009
|
|-drbd.conf
|-sensors3.conf
```

NOTES:
1. /srv/openvz/ve/{1008,1009,...} are the mount points for the individual virtual machine block

devices.

2. /etc/vz is the only *directory* symlink (symlinked to /srv/openvz), the rest are file symlinks (i.e. drbd.conf, hearbeat's ha.cf and authkeys, etc.)

The virtual machine block devices have this directory layout:
/dump
/private
/root
openvz-ve.conf

NOTE: I will eventually use the VE's /dump to handle live-migration, but its not there yet....

To create a new virtual machine using this layout, follow this outline:

1. create a new DRBD device on two nodes and start sync'ing them.
2. make one side primary and create filesys.
3. create VE mount point (i.e. /srv/openvz/ve/$VEID).
4. mount primary DRBD device on VE mount point.
5. vzctl create....
You can set up the vm here or wait until its running under heartbeat##
6. copy /etc/vz/conf/$VEID.conf to VE mount point and symlink it back to /etc/vz/conf/$VEID.conf.
7. unmount primary DRBD device.
8. wait for DRBD to finish sync'ing and then do "drbdadm down" on both nodes.
9. search 'n replace the constraints and resources XML with your info.
10. sudo cibadmin -C -x vm-constraints.xml -o constraints
11. sudo cibadmin -C -x vm-resources.xml -o resources
12. sudo crm_resource -r ms-##DRBDDEVICE## --meta -p target-role -v "#default"
13. sudo crm_resource -r gp-##VENAME## --meta -p target-role -v "#default"

I've attached all of the config files and heartbeat "template" XML files that I think are central to my setup. If I missed anything, let me know....

Finally, about heartbeat V1 - sorry, I wanted V2 as I don't know if the ManageVE ocf script works with V1. And, while we are talking about the ManageVE script - do yourself a favor and change the #!/bin/sh to #!/bin/bash at the top of /usr/lib/ocf/resource.d/heartbeat/ManageVE.

I can have more detailed instructions for node setup and vm setup on Ubuntu Hardy if there is enough interest.

Enjoy!

File Attachments

1) drbd.conf, downloaded 659 times
2) vz.conf, downloaded 637 times
3) openvz-ve.conf, downloaded 670 times
4) vm-constraints.xml, downloaded 594 times

5) vm-resources.xml, downloaded 590 times

Subject: Re: VZ+DRBD+HA and load balancing
Posted by luisdev on Sun, 31 Aug 2008 23:14:02 GMT
View Forum Message <> Reply to Message

Hi, golly.

I have been looking at your solution. Looks fine to me, thanks for sharing!

Regarding the replication of the configuration, I am not sure a separate partition is the way to go, you still have to create several symlinks in every clone, I'd rather put together a list of the config files involved and rsync --include-from it.

I don't know how this really works, but I am wondering if this setup can somehow make problematic the bookkeeping of the vzquota system, since it is hardcoded to /var/vzquota, so it is difficult to share among two servers since you don't have an option to configure it per VE, as you have for root or private.

However, in your system every VE has its own allocated partition, so I guess you don't need any quota system and you could just disable DISKQUOTA in vz.conf.

I am cooking up some scripts to automate some of the steps you mention. I'll keep posting here but since we are working in similar solutions if you PM your email or your IM perhaps we could discuss and share as we trudge along.

Subject: Re: VZ+DRBD+HA and load balancing
Posted by golly on Mon, 01 Sep 2008 03:18:02 GMT
View Forum Message <> Reply to Message

You're quite welcome for the sharing. (BTW, I've changed a few more things with the config. I'll upload the changes on Tuesday as I'm not at work right now.)

Now about the symlink - the reason I went with the symlink was so that everything needed for the virtual machine was all on the same block device (thinking of off-site back-up). However, rsync'ing the config file would be just as good since it doesn't really change once it is set up and tweaked. However you want to do it - its your cluster .

Oh, and yea, I messed with the disk quota and found that it really didn't do anything for me. However, if I was building an active/passive cluster where all of my VEs were on one DRBD device, then, yea, I'd want it. The only problem I can see is if later I want the per-user disk quota then I'd have to find another mechanism as I would have the per-VE quota turned off. At this point I don't really need disk quotas.

However, I did do research and found out that VZ's disk quota mechanism is quite robust. It will regenerate its information at VE startup if it was out-of-date (and at failover, too, as long as it was a full start and not a migrate - I didn't try the migrate scenario). If you need disk quotas even though the VEs are all on separate block devices, VZ's quota system should be just fine.

Thanks for the heads-up about your scripts! I wanted to give back to the community for all the work they have done by providing such excellent software. You know - the "standing on the shoulders of giants" thing....

Subject: Re: VZ+DRBD+HA and load balancing
Posted by luisdev on Wed, 10 Sep 2008 13:49:38 GMT
View Forum Message <> Reply to Message

Hi, folks!

I investigated Mark's method and it works like a charm, with a couple of modifications:

In the post where he describes it, where it says target-role it should say target_role (underscore instead of hyphen.)
You need to add a rsc_location constraint to express your preference about what node should run every VE.

So I took everything and cooked up a script to automate the process, which you can find attached to this message. A wiki page will follow when we are sure it works as intended, but you are welcome to try it and give us your feedback here.

Off the top of my head, before trying the script be sure to:

Create a LVM volume group to hold your VE volumes.
Set your drbd and heartbeat services to start on boot, but not vz.
Enable crm at your ha.cf
Adjust VE_ROOT and VE_PRIVATE at /etc/vz.conf
You can consider to turn off quotas here, since every VE will have its own partition.
Set ONBOOT=no at your default VPS.conf file
Configure your slave HN (the one where you don't run the script) to accept ssh commands without prompting you for a password (not really necessary and a minor security risk, but very convenient.)

To synchronize the configuration files between both HN I use rsync:

rsync -av --include-from=clusterconffiles / <slave_node>:/

And clusterconffiles contains:

+ /etc/
+ /etc/sysconfig/
+ /etc/sysconfig/vz-scripts
+ /etc/sysconfig/vz-scripts/ve-ambiser.conf-sample
+ /etc/drbd.conf
+ /etc/vz/
+ /etc/vz/vz.conf
+ /etc/ha.d/
+ /etc/ha.d/authkeys
+ /etc/ha.d/ha.cf
- *

(ve-ambiser.conf-sample is my custom default VPS.conf file.)

Did I forget anything?

A bit cryptic, but a full tutorial will follow up soon.

File Attachments
1) mbhave, downloaded 635 times

Subject: Re: VZ+DRBD+HA and load balancing
Posted by tsndcb on Thu, 05 May 2011 19:47:24 GMT
View Forum Message <> Reply to Message

Hello luisdev,

any news about the full tutorial will follow up soon?

Thanks in advanced for your return.

Best Regards

Subject: Re: VZ+DRBD+HA and load balancing
Posted by golly on Thu, 05 May 2011 21:45:13 GMT
View Forum Message <> Reply to Message

Hello, tsndcb!

With all of the changes that have occurred in the high-availability software, the details of our approach are quite outdated. Heartbeat has pretty much been replaced by a stack of software: corosync, pacemaker, openais, etc.

Also, I really lost interest when I found out that Ubuntu didn't have an OpenVZ kernel in lucid (10.04). However, since then I found Linux containers (lxc), so I may try to replicate this using the latest tools (corosync/pacemaker instead of heartbeat and lxc instead of OpenVZ). Since there is an ocf agent that can start and stop OpenVZ virtual machines, you should be able to get everything working just fine in OpenVZ.

Regards!

Subject: Re: VZ+DRBD+HA and load balancing
Posted by tsndcb on Fri, 06 May 2011 08:50:19 GMT
View Forum Message <> Reply to Message

Hello golly,

Thanks for your return, actually I tried to done a fail-over VE (LAMP).

I've two RHEL 5.5 servers (HN)with openvz (ovzkernel-2.6.18-238.9.1.el5.028stab089.1.x86_64.rpm)

I've follow this howto HAclusterwithDRBDandHeartbeat on the wiki with drbd 8.3.10,pacemaker-1.0.11 and heartbeat-3.0.3, corosync 1.2.7 (not running).

"HA" configuration seems OK

I've a VE11 with LAMP (Apache + php + mysql)

Actually, I've done two drbd "FS" /VE11 and /data, on /VE11 there are OS + binary for (apache, php, mysql ...) on /data there are data for apache and mysql database.
In fact I've done a FS mount point /data on the VE11 so /data is a FS for the VE11

My question is how the "primary" VE11 must to migrate on the secondary HN2 when the first HN1 failed ?

Must I need to use drbd only for /data and need to setup an other VE on the secondary HN2 it use only data ? Or must must I need to used drbd for /VE11 and /data ? I wanted to done a fail-over configuration.

so

HN1
VE11 with @IP1
/VE11 and /data with drbd

HN2

VE11 with @IP1
/VE11 and /data with drbd

or

HN1
VE11 with @IP1
only /data with drbd

HN2
VE21 with @IP2
only /data with drbd

thanks for your answer