Subject: *SOLVED* Private disk partitions and disk games
Posted by rollinw on Tue, 27 Jun 2006 17:23:48 GMT

View Forum Message <> Reply to Message

Here's another issue I have come up against.

My hardware node is a CentOS4.x system installed on an IDE disk. I also have a large SATA drive that I would like to put databases on. In particular, I would like to have a partition of the SATA drive dedicated to each VE that will be running mysql.

On CentOS, fdisk treats the partitions as SCSI, with names sda1, sda2, etc. on /dev/sda. Interestingly, SUSE Enterprise 9.3 treats the drive as a hard drive, with names like hdi1, hdi2, etc., on /dev/hdi.

Well, after I created partitions and filesystems on /dev/sda using the hardware node, I tried to mount one of them inside the VPSs. The CentOS VPSs didn't even know about /dev/sda, because there was no such device in /dev. The SUSE VPS knows about the device, but refuses to mount the partitions, saying the device is busy.

I went back to the hardware node to see what I could find. There I discovered that the device-mapper has apparently taken over /dev/sda (device and maps it to /dev/dm-0, /dev/dm-1, etc. (as device 253).

My questions are:

1) Is there a way to mount mapped devices from within a
   VPS? If so, what is needed to set it up?
2) If necessary, I can probably get rid of the mapped
   device using dmsetup. Will /dev/sda from the hardware
   node then become available from within the VPSs for
   dedicated partition mounts using /etc/fstab (granted,
   I would need to create the /dev entries for the VPSs)?

Thanks to anyone who has answers to this problem.

rollinw

Subject: Re: Private disk partitions and disk games
Posted by Vasily Tarasov on Wed, 28 Jun 2006 07:55:36 GMT

View Forum Message <> Reply to Message

To use some device inside VPS directly you should do
vzctl set VEID --devnodes hdX:rw --save

But if you only need to mount some partion inside VEs, you can create one mount point in VE0 (Hardware Node) and then create bind (--bind) mountpoints to /vz/private/<VE ID>/mnt/point.

Subject: Re: Private disk partitions and disk games
Posted by rollinw on Wed, 28 Jun 2006 14:09:49 GMT

View Forum Message <> Reply to Message

I am most grateful for your response. It sounds like the creation of a mount point and the mount bind is what I need to do.

I am concerned about the SUSE VE, because simfs does not exist in /proc/filesystems when it is started. Instead, it grabs another real disk partition. Creating a link from /proc/mounts does not solve anything, because the VE doesn't know about the simfs filesystem.

Based on your post, I will try to fix my other problem.

rollinw

Subject: Private disk partitions and disk games
Posted by rollinw on Thu, 29 Jun 2006 15:58:49 GMT

View Forum Message <> Reply to Message

My thanks to vass for pointing me in the right direction. I will give my solution and then discuss documentation.

The basic problem was that the hardware node (node0) treats SATA drives as SCSI disks. The same problem no doubt exists with real SCSIs.

After formating the sda drive and creating filesystems from node0, my attempts mount sda1, etc. from ANY of the nodes failed, with the message, "device busy". By this time I had already given access by vzctl set 150 --devices 8:0:rw 8:1:rw, etc. to the sda drive. I found the answer in /proc/partitions on node0, which showed that the drive partitions had been mapped to /dev/dm-0, /dev/dm-1, etc. Apparently device-mapper runs when node1 boots.

Now I have no interest in handling the SATA drive as a logical volume, so my first step was to get rid of the mapping. I discovered dmsetup and, after making sure sda was the only mapped device, I ran it:

```
/sbin/dmsetup remove_all
```

Then I was able to mount sda1 on node1. Following the advice from sass, I used vzctl set 151 --devnodes sda rw --save.
Going to node 150 I discovered I still had a problem; the mount command gave an error:

```
mount: wrong fs type, bad option, bad superblock on /dev/sda1,
       or too many mounted file systems
```

I believe this error happened because I originally ran mkfs from node0. I could probably have corrected it on node 150 by running mkfs on the partition there. Instead, I went back to node0 and created a mount path there, and then did a mount --bind to a directory in node 150's private area.

Once this was done, I could mount sda partitions in node 150.

Note that some virtual nodes do not have many devices in /dev, and they may have to be created using mknod.

Bottom line: there are two solutions to mounting an unshared partition on a VPS:

1. Get authorization from node0 to partition the drive directly from the VPS and create its filesystems there.
2. Mount the partition on node0 and use mount --bind to link it to a mount directory on the VPS.

Briefly about documentation.
Usually I try to search for technical info before asking help from a forum. In this case I missed the fact that --devnodes is an option under vzctl. The User's Guide does not discuss this subject at all, though it does mention -devices briefly. Also none of the FAQs deal with non-quota disk management in VPSs.

Sys admins who need to create specialized VPSs have to have more information about setting up VPSs in non-standard ways. This situation can be expected in new technology. Hopefully, after I gain more experience with openvz, I can help out by writing some new FAQs or HOWTOs.

Thanks again,
rollinw

Subject: Re: Private disk partitions and disk games
Posted by rollinw on Thu, 29 Jun 2006 16:29:20 GMT
View Forum Message <> Reply to Message

Sorry, there were some typos in my previous post.

Please read node0 for places I have written node1 and node 150 where I have written node 151. Otherwise, there may be confusion!

Thanks
rollinw

Subject: Re: Private disk partitions and disk games
Posted by Vasily Tarasov on Fri, 30 Jun 2006 06:55:03 GMT
View Forum Message <> Reply to Message

It'll be really cool, if you write a small article at http://wiki.openvz.org/ about expirience you gain!

Subject: Re: Private disk partitions and disk games
Posted by luismi on Mon, 03 Jul 2006 00:27:10 GMT
View Forum Message <> Reply to Message

Vass,

When I created a new partition out of the VE.
Then I mount it and I edit the /etc/fstab of the VE.

The I did a reboot of the VE and it reported a problem with the Quota. I fixed that disable the changes I did into the /etc/fstab and also unmounting the partition.

I think that the problem can b fixed using the config file for the VE, but I don't have any idea.

I will try to recreate the problem and I will put the error message here, but until that thing, can anyone with me any idea about how to mount an external partition into a VE and then after a reboot don't have problems with the VE quota?

Thanks!

Subject: Re: Private disk partitions and disk games
Posted by rollinw on Mon, 03 Jul 2006 16:03:02 GMT
View Forum Message <> Reply to Message

My assumption is that disk quota is allocated from partitions that BELONG to the hardware node (node0), based on the fact that this was the disk space defined when node0 was installed. Perhaps I am wrong; i,e., that node0 claims any disk partitions (both local and foreign) it can see and tries to assign quotas to all of them.

Although my new installation on node0 did try to take over swap spaces from 2 other linux installations on my system, I got rid of the extra 2 by editing /etc/fstab. It is true that all partitions with linux-compatible filesystems are visible in /proc/partitions. However, I do not see how node0 could allocate disk quotas to partitions node0 does not have mounted. Let's assume IT DOES NOT (though I could be wrong).

In a clean, theoretical virtualization concept, node0 owns all resources within its own execution environment and shares these resources with its VEs. Since it controls all the hardware, it could also mount any other disk partitions (i.e., partitions on "foreigh disks" outside its own installation) and share these as well. Having node0 control all resources is probably the purest and most secure use of OpenVZ.

Not all uses of OpenVZ need to be or want to be that theoretically pure. There are special cases a VE MUST have direct access to some of the hardware resources. OpenVZ provides a way for node0 to allocate hardware to a VE. This is through the utility vzctl. There are command modes in vzctl that activate specific kinds of hardware access in a VE. The results of some of these commands are stored in a VE's VEID.conf file. Examples of vzctl commands are:

--netdev_add (allows the VE direct access to a net device)
--devnodes   (gives the VE direct access: r,w,rw, none--to a device)

Note: I have made this work with "foreign" disks. In the near future I may need to try it with a CDROM and/or a floppy drive.

--devices (gives VE ability to control devices; e.g., to partition disks and create filesystems on them)

Note: I couldn't find this one in the vzctl man pages, but it is described in the Advanced Tasks section of the User Guide.

Besides these hardware control options, there is also the vzctl command,
--capability
that gives a VE access to many of its internal system options.

This is longer than I intended, so I will stop.

rollinw