Subject: ploop and trim/ discard support
Posted by Corin Langosch on Wed, 12 Sep 2012 19:13:22 GMT
View Forum Message <> Reply to Message

Hi,

it seems that when creating & deleting files inside a ploop backed container the backing image file doesn't shrink. So the image file is still consuming around 400 GB, while in the container df shows only 100 GB in use.

ll on host: -rw------- 1 root root 384G Sep 12 21:06 root.hdd
df on container: /dev/ploop1 493G 103G 390G 21% /

I assume this is because ploop doesn't get informed by the ext4 fs when blocks are no longer in use. Should it help to to mount the ext4 on the ploop device with the "discard" option? I didn't test if it helps yet, but at least I was able to pass the mount option (thanks to kir's recent enhancements) and it started the container without any problems.

/dev/ploop1 on / type ext4 (rw,relatime,barrier=0,data=ordered,discard)

To free all "free" existing blocks I also tried to do a batch cleanup/ trim from inside the container using "fstrim /" but it doesn't work:

root@test:~# fstrim /
fstrim: /: FITRIM ioctl failed: Operation not permitted

Should this be working or is it not supported yet? Am I on the right path at all? Any better solutions? :)

Corin

Subject: Re: ploop and trim/ discard support
Posted by kir on Wed, 12 Sep 2012 19:34:46 GMT
View Forum Message <> Reply to Message

On Sep 12, 2012 11:20 PM, "Corin Langosch" <info@corinlangosch.com> wrote:
>
> Hi,
>
> it seems that when creating & deleting files inside a ploop backed container the backing image file doesn't shrink. So the image file is still consuming around 400 GB, while in the container df shows only 100 GB in use.
>
> ll on host: -rw------- 1 root root 384G Sep 12 21:06 root.hdd
> df on container: /dev/ploop1 493G 103G 390G 21% /

We have online shrink. It's "ploop balloon discard" or just "vzctl compact". You'd better have ploop and vzctl from git (both are really close to be released).

>
> I assume this is because ploop doesn't get informed by the ext4 fs when blocks are no longer in use. Should it help to to mount the ext4 on the ploop device with the "discard" option? I didn't test if it helps yet, but at least I was able to pass the mount option (thanks to kir's recent enhancements) and it started the container without any problems.
>
> /dev/ploop1 on / type ext4 (rw,relatime,barrier=0,data=ordered,discard)
>
> To free all "free" existing blocks I also tried to do a batch cleanup/ trim from inside the container using "fstrim /" but it doesn't work:
>
> root@test:~# fstrim /
> fstrim: /: FITRIM ioctl failed: Operation not permitted
>
> Should this be working or is it not supported yet? Am I on the right path at all? Any better solutions? :)
>
> Corin
>

Subject: Re: ploop and trim/ discard support
Posted by Corin Langosch on Wed, 12 Sep 2012 21:38:59 GMT
View Forum Message <> Reply to Message

On 12.09.2012 at 21:34 +0200, Kir Kolyshkin <kir@openvz.org> wrote:
>
> On Sep 12, 2012 11:20 PM, "Corin Langosch" <info@corinlangosch.com
> <mailto:info@corinlangosch.com>> wrote:
> >
> > Hi,
> >
> > it seems that when creating & deleting files inside a ploop backed
> container the backing image file doesn't shrink. So the image file is
> still consuming around 400 GB, while in the container df shows only
> 100 GB in use.
> >
> > ll on host: -rw------- 1 root root 384G Sep 12 21:06 root.hdd
> > df on container: /dev/ploop1 493G 103G 390G 21% /
>
> We have online shrink. It's "ploop balloon discard" or just "vzctl
> compact". You'd better have ploop and vzctl from git (both are really

> close to be released).
>

The system's still working heavily but the image is already reduced by
around 60 GB so it seems to work well. :)

Would you suggest putting some script into crontab to compact all ploop
images ex. once a week? Or probably do something like ctid%7 and compact
a few of them every night?

BTW: the man of vzctl contains the "compact" command. But the usage
output of vzctl doesn't. Shall I file a bug report for it?

BTW 2: I just read about the ploop ballooning technique in the wiki. Am
I correct in that it works quite similar to sdelete on windows (
http://pastie.org/4710364 or
 http://technet.microsoft.com/en-us/sysinternals/bb897443.asp x)?

Thanks!

Subject: Re: ploop and trim/ discard support
Posted by kir on Wed, 12 Sep 2012 22:36:22 GMT
View Forum Message <> Reply to Message

On 09/13/2012 01:38 AM, Corin Langosch wrote:
>
> On 12.09.2012 at 21:34 +0200, Kir Kolyshkin <kir@openvz.org> wrote:
>>
>> On Sep 12, 2012 11:20 PM, "Corin Langosch" <info@corinlangosch.com
>> <mailto:info@corinlangosch.com>> wrote:
>> >
>> > Hi,
>> >
>> > it seems that when creating & deleting files inside a ploop backed
>> container the backing image file doesn't shrink. So the image file is
>> still consuming around 400 GB, while in the container df shows only
>> 100 GB in use.
>> >
>> > ll on host: -rw------- 1 root root 384G Sep 12 21:06 root.hdd
>> > df on container: /dev/ploop1   493G 103G 390G 21% /
>>
>> We have online shrink. It's "ploop balloon discard" or just "vzctl
>> compact". You'd better have ploop and vzctl from git (both are really
>> close to be released).
>>
>
> The system's still working heavily but the image is already reduced by

> around 60 GB so it seems to work well. :)
>
> Would you suggest putting some script into crontab to compact all
> ploop images ex. once a week? Or probably do something like ctid%7 and
> compact a few of them every night?

You can use ploop balloon discard --stat to find out how much space
could be compacted, and decide if you need it.

Any contribution in that area is welcome :)

>
> BTW: the man of vzctl contains the "compact" command. But the usage
> output of vzctl doesn't. Shall I file a bug report for it?

NP, I have just fixed it
http://git.openvz.org/?p=vzctl;a=commitdiff;h=60e4b4

But the big problem here is vzctl --help output is too heavy.
I would like to have something that is done for ploop -- you get basic
help and if you specify a command you get a command-specific help.

Good patches for that are welcome :)

Subject: Re: ploop and trim/ discard support
Posted by Kirill Korotaev on Thu, 13 Sep 2012 07:22:27 GMT
View Forum Message <> Reply to Message

On Sep 13, 2012, at 01:38 , Corin Langosch wrote:

>
> On 12.09.2012 at 21:34 +0200, Kir Kolyshkin <kir@openvz.org> wrote:
>> On Sep 12, 2012 11:20 PM, "Corin Langosch" <info@corinlangosch.com> wrote:
>> >
>> > Hi,
>> >
>> > it seems that when creating & deleting files inside a ploop backed container the backing image file doesn't shrink. So the image file is still consuming around 400 GB, while in the container df shows only 100 GB in use.
>> >
>> > ll on host: -rw------- 1 root root 384G Sep 12 21:06 root.hdd
>> > df on container: /dev/ploop1 493G 103G 390G 21% /
>>
>> We have online shrink. It's "ploop balloon discard" or just "vzctl compact". You'd better have ploop and vzctl from git (both are really close to be released).
>>
>

> The system's still working heavily but the image is already reduced by around 60 GB so it seems to work well. :)
>
> Would you suggest putting some script into crontab to compact all ploop images ex. once a week? Or probably do something like ctid%7 and compact a few of them every night?
>
> BTW: the man of vzctl contains the "compact" command. But the usage output of vzctl doesn't. Shall I file a bug report for it?
>
> BTW 2: I just read about the ploop ballooning technique in the wiki. Am I correct in that it works quite similar to sdelete on windows (http://pastie.org/4710364 or http://technet.microsoft.com/en-us/sysinternals/bb897443.asp x)?

No, AFAIR we should use TRIM on ext4 and it simply reports unused space.
Balloon is used for "resize" via allocating some space and hiding it from user, but for compacting it's a bit bad since can cause ENOSPC while it's really not...

Kirill

File Attachments

1) smime.p7s, downloaded 1797 times

Subject: Re: ploop and trim/ discard support
Posted by Corin Langosch on Mon, 17 Sep 2012 15:53:11 GMT
View Forum Message <> Reply to Message

On 13.09.2012 at 09:22 +0200, Kirill Korotaev <dev@parallels.com> wrote:
>
> No, AFAIR we should use TRIM on ext4 and it simply reports unused space.
> Balloon is used for "resize" via allocating some space and hiding it
> from user, but for compacting it's a bit bad since can cause ENOSPC
> while it's really not...
>

So this whole ballooning is only a work around as trim/ discard support
for ext4 is only available in kernel >= 2.6.33? Once openvz is rebased
to a newer kernel (3.2.x?) it can/ will be dropped? :)

Corin

Subject: Re: ploop and trim/ discard support
Posted by Kirill Korotaev on Mon, 17 Sep 2012 16:46:51 GMT
View Forum Message <> Reply to Message

On Sep 17, 2012, at 19:53 , Corin Langosch <corin.langosch@netskin.com> wrote:

>
> On 13.09.2012 at 09:22 +0200, Kirill Korotaev <dev@parallels.com> wrote:
>>
>> No, AFAIR we should use TRIM on ext4 and it simply reports unused space.
>> Balloon is used for "resize" via allocating some space and hiding it
>> from user, but for compacting it's a bit bad since can cause ENOSPC
>> while it's really not...
>>
>
> So this whole ballooning is only a work around as trim/ discard support
> for ext4 is only available in kernel >= 2.6.33? Once openvz is rebased
> to a newer kernel (3.2.x?) it can/ will be dropped? :)
>

You've mixed 2 different scenarios:
1. vzctl set --diskpace
When you resize CT to smaller sizes we do not want to resize live file system and move data around causing I/O.
So we use balloon to reserve some space in CT and "pretend" that CT was made smaller.
TRIM has nothing to do with this scenario, cause it wouldn't prevent file system from allocating its free space.

2. compacting
When CT has used some space and then files were removed image requires "compaction" to free this space back to host.
This is where both ballooning and TRIMing can help. But ballooning reserves disk space, so it can lead to ENOSPC inside CT and thus is better to avoid (remember, it reserves space!).
TRIM on the other hand is a standard way to cause file system to report it's unused space and this is what we use.

TRIM support present in our RHEL6 kernels, so switching to >= 2.6.33 is not required and won't result in any benefits in this area.

Thanks,
Kirill

File Attachments

1) smime.p7s, downloaded 1791 times