Subject: Using a layered filesystem as private dir?
Posted by Rick van Rein on Thu, 05 Jan 2012 11:32:58 GMT
View Forum Message <> Reply to Message

Hello,

I've just started using OpenVZ, and it feels more natural than the alternatives I've seen -- my compliments!

I can get a host runnig from a ZFS volume like /tank/vzdemo, which then also gets shown at /var/lib/vz/vz-$VEID. But what I really want to do is use a layered FS (like aufs) as the private directory for the container. But trying to do that leads to an error:

bash# mount -t aufs -o br:/tank/vzdemo=rw:/tank/squeeze=ro none /mnt
bash# grep VE_ /etc/vz/conf/777.conf
VE_PRIVATE=/mnt
bash# vzctl create 777
Private area already exists in /mnt
Creation of container private area failed

What is this trying to say? Is there a way to do what I am trying to do? Did I understand well that the private area is a directory, not a device?

Thanks,
-Rick

P.S. To capture any "why" questions :- I am trying to share as many resources as possible. Containers beat Xen/KVM/VMware in that respect, and when I can share the base OS and only have a thin layer on top, it should mean that even the buffer cache is shared between containers. It also means that upgrades can be reduced to a minimum of repetition.

Subject: Re: Using a layered filesystem as private dir?
Posted by dowdle on Thu, 05 Jan 2012 17:08:53 GMT
View Forum Message <> Reply to Message

Greeting,

----- Original Message -----
> bash# mount -t aufs -o br:/tank/vzdemo=rw:/tank/squeeze=ro none /mnt
> bash# grep VE_ /etc/vz/conf/777.conf
> VE_PRIVATE=/mnt

> bash# vzctl create 777
> Private area already exists in /mnt
> Creation of container private area failed

/mnt isn't your private directory... it should be /mnt/777/private and of course you need a root directory too.

So far as sharing files between containers in a CoW situation, Virtuozzo and Linux-VServer offer those features, but OpenVZ does not. It appears you are trying to engineer your own solution. I don't know if what you want to do will work or not because I haven't tried it. Nor have I tried ZFS in Linux. I suspect it won't work though... but I do wish you luck.

TYL,
--
Scott Dowdle
704 Church Street
Belgrade, MT 59714
(406)388-0827 [home]
(406)994-3931 [work]

Subject: Re: Using a layered filesystem as private dir?
Posted by Kirill Korotaev on Thu, 05 Jan 2012 18:52:06 GMT
View Forum Message <> Reply to Message

As Scott mentioned we have VZFS in commercial version of Parallels Containers.
It helps to save a lot of IOPS by sharing files between containers and is fully POSIX compliant.

Thanks,
Kirill

On Jan 5, 2012, at 15:32 , Rick van Rein wrote:

> Hello,
>
> I've just started using OpenVZ, and it feels more natural than the
> alternatives I've seen -- my compliments!
>
> I can get a host runnig from a ZFS volume like /tank/vzdemo, which then
> also gets shown at /var/lib/vz/vz-$VEID. But what I really want to
> do is use a layered FS (like aufs) as the private directory for the
> container. But trying to do that leads to an error:
>
> bash# mount -t aufs -o br:/tank/vzdemo=rw:/tank/squeeze=ro none /mnt
> bash# grep VE_ /etc/vz/conf/777.conf
> VE_PRIVATE=/mnt
> bash# vzctl create 777

> Private area already exists in /mnt
> Creation of container private area failed
>
> What is this trying to say? Is there a way to do what I am trying
> to do? Did I understand well that the private area is a directory,
> not a device?
>
>
> Thanks,
> -Rick
>
>
> P.S. To capture any "why" questions :- I am trying to share as many
> resources as possible. Containers beat Xen/KVM/VMware in that
> respect, and when I can share the base OS and only have a thin
> layer on top, it should mean that even the buffer cache is
> shared between containers. It also means that upgrades can be
> reduced to a minimum of repetition.
>

Subject: Re: Using a layered filesystem as private dir?
Posted by jjs - mainphrame on Thu, 05 Jan 2012 19:07:46 GMT
View Forum Message <> Reply to Message

I have postfix servers running on openvz and in general give it high marks,
but there's little point in trying to make it something it's not. If you
have the budget, the extra features of virtuozzo are well worth the money.
There are good reasons why virtuozzo is more expensive, but if you can only
afford openvz, my advice would be to let it do what it does best, and don't
obsess over disk space which is a rather affordable commodity these days.

Joe

On Thu, Jan 5, 2012 at 10:52 AM, Kirill Korotaev <dev@parallels.com> wrote:

> As Scott mentioned we have VZFS in commercial version of Parallels
> Containers.
> It helps to save a lot of IOPS by sharing files between containers and is
> fully POSIX compliant.
>
> Thanks,
> Kirill
>
>
> On Jan 5, 2012, at 15:32 , Rick van Rein wrote:
>
> > Hello,

> >
> > I've just started using OpenVZ, and it feels more natural than the
> > alternatives I've seen -- my compliments!
> >
> > I can get a host runnig from a ZFS volume like /tank/vzdemo, which then
> > also gets shown at /var/lib/vz/vz-$VEID. But what I really want to
> > do is use a layered FS (like aufs) as the private directory for the
> > container. But trying to do that leads to an error:
> >
> > bash# mount -t aufs -o br:/tank/vzdemo=rw:/tank/squeeze=ro none /mnt
> > bash# grep VE_ /etc/vz/conf/777.conf
> > VE_PRIVATE=/mnt
> > bash# vzctl create 777
> > Private area already exists in /mnt
> > Creation of container private area failed
> >
> > What is this trying to say? Is there a way to do what I am trying
> > to do? Did I understand well that the private area is a directory,
> > not a device?
> >
> >
> > Thanks,
> > -Rick
> >
> >
> > P.S. To capture any "why" questions :- I am trying to share as many
> >   resources as possible. Containers beat Xen/KVM/VMware in that
> >   respect, and when I can share the base OS and only have a thin
> >   layer on top, it should mean that even the buffer cache is
> >   shared between containers. It also means that upgrades can be
> >   reduced to a minimum of repetition.
> >

Subject: Re: Using a layered filesystem as private dir?
Posted by Rick van Rein on Thu, 05 Jan 2012 19:12:27 GMT
View Forum Message <> Reply to Message

Hello Scott / others,

Thanks for responding. I have good news for OpenVZ :-)

> /mnt isn't your private directory... it should be /mnt/777/private and of course you need a root directory too.

I will admit that the description in the manual pages
about root and private has me confused... is it not true
that private is the already-mounted directory that will

be used as the root directory, and that this root directory
is the place where VZ will mount its own local work-copy?

But I found I also did something else wrong, namely to run
"vzctl create" instead of "vzctl start". The latter works
fine, I found. Silly me...

> So far as sharing files between containers in a CoW situation, Virtuozzo and Linux-VServer offer those features, but OpenVZ does not. It appears you are trying to engineer your own solution.

Hmm, I've seen what VServer does, which is collecting parts
that are the same and then use CoW. This is a de-duplication
service (that ZFS could also provide) but it is after-the-fact
re-assembly; I would rather build on a designed invariant that
certain parts share. As explained I also hope to avoid running
repetitious upgrades. The savings in disk space are very good,
with 900 MB for a good Debian Squeeze and 35 MB added for Apache
and PHP. The buffer cache should share in the disk savings. I
don't expect that the buffer cache would be shared between VMs
that share blocks with CoW (which isn't really sharing), but I'm
not sure.

> I don't know if what you want to do will work or not because I haven't tried it. Nor have I tried ZFS in Linux. I suspect it won't work though... but I do wish you luck.

I have some encouraging figures to show that it does work!

Test run 1, just after reboot, based on empty1+squeeze and empty2+squeeze:

```
: root# time vzctl exec 777 find / > /dev/null 2> /dev/null
: real    0m16.013s
: user    0m0.264s
: sys     0m1.508s
:
: root# time vzctl exec 777 find / > /dev/null 2> /dev/null
: real    0m8.083s
: user    0m0.120s
: sys     0m0.540s
:
: root# time vzctl exec 777 find / > /dev/null 2> /dev/null
: real    0m0.764s
: user    0m0.056s
: sys     0m0.252s
:
: root# time vzctl exec 778 find / > /dev/null 2> /dev/null
: real    0m1.722s
: user    0m0.076s
: sys     0m0.596s
```

:
: root# time vzctl exec 778 find / > /dev/null 2> /dev/null
: real   0m0.602s
: user   0m0.052s
: sys   0m0.260s
:
: root# time vzctl exec 777 find / > /dev/null 2> /dev/null
: real   0m1.159s
: user   0m0.060s
: sys   0m0.300s
:
: root# time vzctl exec 778 find / > /dev/null 2> /dev/null
: real   0m1.152s
: user   0m0.072s
: sys   0m0.296s

The values over 1s also show up after waiting a while, so it appears
to be caused by background processes that empty/reuse some of the
buffer cache, but this effect is the same as on a single machine:
the first find takes long, the next ones are a lot faster. I am
not sure what to think of the second with its half-way timing though.

I think these figures warrant the statement that the buffer cache is
shared between VMs if the disk blocks are. And that is a great saving
and a big advantage compared to the Qemu model used by KVM and Xen!
I suppose the maximum amount of OpenVZ VMs on a hardware machine just
got multiplied by ten or more?

Cheers,
 -Rick

Subject: Re: Using a layered filesystem as private dir?
Posted by pavel@pronskiy.ru on Thu, 05 Jan 2012 19:32:54 GMT
View Forum Message <> Reply to Message

hi!
my configuration: openvz + squashfs4.0 + aufs
2.1-standalone.tree-32-20110228

/vz/squashfs/gentoo/ - mounted squashfs gentoo image or other distr
/vz/private/123/ - mount squashed gentoo dir and aufs layer from storage
dir /vz/storage/123/
/vz/private/124/ - mount squashed gentoo dir and aufs layer from storage
dir /vz/storage/124/
etc..

basic example:
hac_ostpl_sq_file = /vz/template/cache/gentoo.sq
hac_storage_private_dir = /vz/storage/vpsid/
hac_ostpl_squashfs_dir = /vz/squashfs/gentoo/
hac_private_dir = /vz/private/vpsid/

start ve
mount -o loop $hac_ostpl_sq_file $hac_ostpl_squashfs_dir/
mount -t aufs -o
 br=${hac_storage_private_dir}=rw:${hac_ostpl_squashfs_dir}=r o none
${hac_private_dir}
vzctl start vpsid

stop ve
vzctl stop vpsid
umount /vz/private/vpsid

Making new ve:
1. generate /etc/vz/conf/vpsid.conf
2. mount aufs
3. vzctl start vpsid
profit.

-=>> mount
/vz/template/cache/gentoo.sq on /vz/squashfs/gentoo type squashfs
(ro,relatime)
none on /vz/private/203 type aufs (rw,relatime,si=4b64c719)
/vz/private/203 on /vz/root/203 type simfs (rw,relatime)
proc on /vz/root/203/proc type proc (rw,relatime)
sysfs on /vz/root/203/sys type sysfs (rw,relatime)
rc-svcdir on /vz/root/203/lib/rc/init.d type tmpfs
 (rw,nosuid,nodev,noexec,relatime,size=1024k,nr_inodes=131072 ,mode=755)
devpts on /vz/root/203/dev/pts type devpts
(rw,nosuid,noexec,relatime,gid=5,mode=620)
shm on /vz/root/203/dev/shm type tmpfs
 (rw,nosuid,nodev,noexec,relatime,size=524288k,nr_inodes=1310 72)
...

work perfect on 2.6.32-rhel6 (patched by me and J.R. Okajima author aufs)
If you need support: pavel@pronskiy.ru

> As Scott mentioned we have VZFS in commercial version of Parallels Containers.
> It helps to save a lot of IOPS by sharing files between containers and is fully POSIX compliant.
>
> Thanks,

> Kirill
>
>
> On Jan 5, 2012, at 15:32 , Rick van Rein wrote:
>
>> Hello,
>>
>> I've just started using OpenVZ, and it feels more natural than the
>> alternatives I've seen -- my compliments!
>>
>> I can get a host runnig from a ZFS volume like /tank/vzdemo, which then
>> also gets shown at /var/lib/vz/vz-$VEID. But what I really want to
>> do is use a layered FS (like aufs) as the private directory for the
>> container. But trying to do that leads to an error:
>>
>> bash# mount -t aufs -o br:/tank/vzdemo=rw:/tank/squeeze=ro none /mnt
>> bash# grep VE_ /etc/vz/conf/777.conf
>> VE_PRIVATE=/mnt
>> bash# vzctl create 777
>> Private area already exists in /mnt
>> Creation of container private area failed
>>
>> What is this trying to say? Is there a way to do what I am trying
>> to do? Did I understand well that the private area is a directory,
>> not a device?
>>
>>
>> Thanks,
>> -Rick
>>
>>
>> P.S. To capture any "why" questions :- I am trying to share as many
>>   resources as possible. Containers beat Xen/KVM/VMware in that
>>   respect, and when I can share the base OS and only have a thin
>>   layer on top, it should mean that even the buffer cache is
>>   shared between containers. It also means that upgrades can be
>>   reduced to a minimum of repetition.
>>