Linux kernel client - Bug #55090

mounting subvolume shows size/used bytes for entire fs, not subvolume

03/28/2022 02:51 PM - Dan van der Ster

Status:	New	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:	Luis Henriques		
Category:			
Target version:			
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Crash signature (v1):**	
Severity:	3 - minor	**Crash signature (v2):**	

Description

When mounting a subvolume at the base dir of the subvolume, the kernel client correctly shows the size/usage of a subvol:

```
Filesystem
    Size  Used Avail Use% Mounted on
xxx:6789:/volumes/_nogroup/4db8f9a6-926b-4306-8a6d-0e1b897c1d2f/d2ef5fea-040c-4ec1-b1bb-66073f9fc8
ac  8.8T     0  8.8T   0% /cephfs
```

However if the client mounts a subdir of the subvolume, they see the size/usage of the entire cephfs:

```
Filesystem
              Size  Used Avail Use% Mounted on
xxx:6789:/volumes/_nogroup/4db8f9a6-926b-4306-8a6d-0e1b897c1d2f/d2ef5fea-040c-4ec1-b1bb-66073f9fc8
ac/my/subdir  1.3P  430T  860T  34% /var/lib/service
```

`ceph-fuse` does not have this behaviour -- mounting at a subdir below the subvolume shows the "correct" subvolume size and usage.

Related issues:

Copied to CephFS - Bug #56414: mounting subvolume shows size/used bytes for e...	**Resolved**

History

#1 - 03/28/2022 02:56 PM - Dan van der Ster

kernel is 4.18.0-365.el8.x86_64

#2 - 04/19/2022 02:34 PM - Ramana Raja

Looks like this issue was fixed in kernel 5.2 https://tracker.ceph.com/issues/38482 ?

#3 - 04/19/2022 04:06 PM - Luis Henriques

Ramana Raja wrote:

> Looks like this issue was fixed in kernel 5.2 https://tracker.ceph.com/issues/38482 ?

Yeah, this does indeed look like a duplicate of this tracker. It may be worth trying to test a more recent kernel to see if that's fixed.

#4 - 04/19/2022 04:10 PM - Ramana Raja

Dan van der Ster wrote:

> kernel is 4.18.0-365.el8.x86_64

Dan, can you please check with kernel >= 5.2 as pointed out by https://lists.ceph.io/hyperkitty/list/ceph-users@ceph.io/message/PQPBGGFGQ7CAKYEFVUGEU2OIK64V6FXQ/ ? Thanks!

#5 - 04/19/2022 05:42 PM - Jeff Layton

I don't think a new kernel will probably help. The -365.el8 kernel is up to date with upstream as on ~December 2021. I think the trick here is to "ceph fs subvolume create ..." with a limited "size" and with --namespace-isolated option. That should hopefully replicate what Dan is doing here.

#6 - 04/20/2022 10:05 AM - Luis Henriques

Jeff Layton wrote:

> I don't think a new kernel will probably help. The -365.el8 kernel is up to date with upstream as on ~December 2021. I think the trick here is to "ceph fs subvolume create ..." with a limited "size" and with --namespace-isolated option. That should hopefully replicate what Dan is doing here.

I'm not able to reproduce this issue in a recent kernel, even when using subvolumes. AFAIUI when we create a subvolume with --size we are simply setting the max_bytes quota on a directory. When we mount a subdir of the volume, we should see exactly the same values with 'df' as they are both on the same quota realms. **Unless** the client hasn't the permissions to access the subvolume root where the quotas are set.

Is it possible that the credentials your using for the 2 mount commands are different? Are you using different users/keys or are they exactly the same?

#7 - 04/26/2022 04:22 PM - Luis Henriques

Ok, I've managed to reproduce **a** bug with the same symptoms, but I suspect it's a different bug. Here's how I've done it:

```
# mount .../<subvolume> /mnt
# mkdir /mnt/subdir
# setfattr -n ceph.quota.max_files -v 100000 /mnt/subdir
# umount /mnt
# mount .../<subvolume>/subdir /mnt
```

Running a 'df' at this point will show the filesystem total and not the value in the subvolume quotas.

I think this is not the same issue because the 'max_files' is not mentioned anywhere and I can reproduce it using the fuse client too. Anyway, I've sent out an initial fix for this bug [1], it would be nice if you could give it a try.

Also, would it be possible to get details on the subvolume (ceph fs subvolume info <vol> <subvol>) and the cephx credentials?

[1] https://lore.kernel.org/r/20220426161204.17896-1-lhenriques%40suse.de

#8 - 04/29/2022 10:12 AM - Luis Henriques

I think I managed to understand what's going on. It's a mix of kernel security features (LSMs) and cephfs authentication capabilities configuration.

Inodes have a field ->i_security that is used to store security-related data for certain Linux Security Modules (LSMs). When the ceph kernel client is doing an inode lookup, a request is sent to the MDS with a mask containing the CEPH_CAP_XATTR_SHARED bit set if the ->i_security field isn't NULL. Then, when the MDS is doing the inode lookup, it will require a extra lock (xattrlock) if the CEPH_CAP_XATTR_SHARED was set. This will however result in -EACCESS error if the authentication capabilities configured don't explicitly include the read access to the parent directory that has the quotas set (in this case, to '/volumes/_nogroup/4db8f9a6-926b-4306-8a6d-0e1b897c1d2f').

I do not discard a bug in the MDS in this case, as the code handling this has the following comment:

```
void Server::handle_client_lookup_ino(MDRequestRef& mdr,
                                      bool want_parent, bool want_dentry)
{
...
    // FIXME
    // permission bits, ACL/security xattrs
    if ((mask & CEPH_CAP_AUTH_SHARED) && (issued & CEPH_CAP_AUTH_EXCL) == 0)
      lov.add_rdlock(&in->authlock);
    if ((mask & CEPH_CAP_XATTR_SHARED) && (issued & CEPH_CAP_XATTR_EXCL) == 0)
      lov.add_rdlock(&in->xattrlock);
...
```

But I don't really know this code (or the MDS in general), so I can't help a lot here.

Anyway, there are two options to work around this. The first one is to add the directory to the auth configuration. I assume the client is using something created by:

```
ceph fs authorize cephfs_a client.foo /volumes/_nogroup/4db8f9a6-926b-4306-8a6d-0e1b897c1d2f/d2ef5fea-040c-4ec
1-b1bb-66073f9fc8ac rw
```

Replacing that by:

```
ceph fs authorize cephfs_a client.foo /volumes/_nogroup/4db8f9a6-926b-4306-8a6d-0e1b897c1d2f r /volumes/_nogro
up/4db8f9a6-926b-4306-8a6d-0e1b897c1d2f/d2ef5fea-040c-4ec1-b1bb-66073f9fc8ac rw
```

show do the trick.

The other option is to disable the LSM responsible for setting the inode ->i_security field. In my case, I managed to reproduce the issue when the modules were "capability,yama,bpf,landlock" (you can see the modules being used with 'cat /sys/kernel/security/lsm'). By removing the 'landlock' LSM from this list did the trick: you just need to add an extra kernel parameter (in grub). I've added "lsm=capability,yama,bpf" (removed the 'landlock') and rebooted.

#9 - 05/16/2022 07:50 PM - R Taylor

Hi Luís,

With the cephfs filesystem where I reproduced this, I have read/write access to everything in /volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b4226c3cb2, not just a subdirectory. I am not sure how the cephx keys are generated (it is done by Manila) but I think it would correspond to something like this?

ceph fs authorize cephfs_a client.foo /volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b4226c3cb2 rw

On the Fedora 35 test VM, I disabled the landlock LSM but the problem still happens:

```
$ cat /sys/kernel/security/lsm
lockdown,capability,yama,selinux,bpf,landlock
$ sudo grubby --update-kernel=ALL --args="lsm=lockdown,capability,yama,selinux,bpf"
$ reboot

$ cat /sys/kernel/security/lsm
lockdown,capability,yama,selinux,bpf
$ sudo mount -t ceph 10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9
e2a-81b4226c3cb2 /mnt/ceph1 -o name=eos_rw,secret=AQ...
$ sudo mount -t ceph 10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9
e2a-81b4226c3cb2/testsubdir /mnt/ceph2 -o name=eos_rw,secret=AQ...
# Then the problem is still reproduced:
$ df -h | grep ceph
10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b4226c3cb2
        4.9T  278G  4.7T   6% /mnt/ceph1
10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b4226c3cb2/test
subdir  3.8P  319T  3.5P   9% /mnt/ceph2

# seems like the lockdown LSM can't be removed
$ sudo grubby --update-kernel=ALL --args="lsm=capability,yama,bpf"
$ sudo reboot
$ cat /sys/kernel/security/lsm
lockdown,capability,yama,bpf
# still same result
$ df -h | grep ceph
10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b4226c3cb2
        4.9T  278G  4.7T   6% /mnt/ceph1
10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b4226c3cb2/test
subdir  3.8P  319T  3.5P   9% /mnt/ceph2
```

We are also having this problem on AlmaLinux 8; and I think Dan reported it on a EL8/CentOS8 system. Almalinux 8 does not appear to have landlock by default:

```
$ cat /sys/kernel/security/lsm
capability,yama,selinux,bpf
```

#10 - 05/16/2022 08:14 PM - R Taylor

I checked with our ceph admin and looked up the details of the key I am using with 'ceph auth ls':

```
client.eos_rw
        key: AQC....
        caps: [mds] allow rw path=/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b4226c3cb2
        caps: [mon] allow r
        caps: [osd] allow rw pool=cephfs_data namespace=fsvolumens_55e46a89-31ff-4878-9e2a-81b4226c3cb2
```

#11 - 05/17/2022 04:24 PM - Luis Henriques

I've done a quick grep in the kernel code and, apparently, the following security modules will initialize the inode->i_security field: bpf, landlock, selinux, smack. So, as long as any of these is enabled the kernel client will result in the MDS returning -EACCESS as described above.

So, you can either remove all of these LSMs or simply fix the auth configuration. I **think** it the 'mds' caps line will need to be changed to something like:

```
client.eos_rw
        key: AQC....
        caps: [mds] allow r path=/volumes/_nogroup, allow rw path=/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81
b4226c3cb2
        caps: [mon] allow r
        caps: [osd] allow rw pool=cephfs_data namespace=fsvolumens_55e46a89-31ff-4878-9e2a-81b4226c3cb2
```

#12 - 05/17/2022 04:54 PM - Luis Henriques

- Assignee set to Luis Henriques

#13 - 06/07/2022 10:22 PM - R Taylor

Hi Luis,

Sorry I did not see your reply - apparently posting in an issue does not subscribe you to receive updates on an issue!

It might not be feasible to use a RHEL8-based OS in a standard/secure way with those LSMs disabled; as SELinux and BPF are needed for various things (esp. in kubernetes).
I made a new test VM (Fedora 35) with the default LSM settings.

I made a new keyring with the capabilities you described:

```
 cat ceph.client.eos_test.keyring
[client.eos_test]
    key = AQ...
        caps mds = allow r path=/volumes/_nogroup, allow rw path=/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b
4226c3cb2
        caps mon = allow r
        caps osd = allow rw pool=cephfs_data namespace=fsvolumens_55e46a89-31ff-4878-9e2a-81b4226c3cb2
```

and imported it, shows up as:

```
$ sudo ceph auth ls | grep -A 4 eos_test
client.eos_test
    key: AQ....
    caps: [mds] allow r path=/volumes/_nogroup, allow rw path=/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b422
6c3cb2
    caps: [mon] allow r
    caps: [osd] allow rw pool=cephfs_data namespace=fsvolumens_55e46a89-31ff-4878-9e2a-81b4226c3cb2
```

Mounting again:

```
$ sudo mount -t ceph 10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9
e2a-81b4226c3cb2       /mnt/ceph1 -o name=eos_test,secret=...
$ sudo mount -t ceph 10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9
e2a-81b4226c3cb2/test1 /mnt/ceph3 -o name=eos_test,secret=...
$ df -h | grep ceph
10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b4226c3cb2
   4.9T  278G  4.7T   6% /mnt/ceph1
10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b4226c3cb2/test
1  4.9T  278G  4.7T   6% /mnt/ceph3
```

it fixed the original problem! The reported filesystem size is correct for the subdir mount.

Now I am checking how the reported filesystem usage is updated when I write into the base mount vs the subdir mount; seems like there might be some irregularities. I'll see if I can find anything consistent.

Thanks for your help!

#14 - 06/09/2022 12:55 AM - R Taylor

Hi Luis,
I found two things:
1. With the workaround that adds caps: [mds] allow r path=/volumes/_nogroup, I can actually view the shares of all users on the cloud read-only. Personally for my specific use case on our on-prem cloud that's okay, but I don't think this works as a general solution for all users (e.g. for Manila and Openstack in general).

2. When I have both of these mounts:

```
10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b4226c3cb2
   5120000M  804M  5119196M   1% /mnt/ceph1
10.30.201.3:6789,10.30.202.3:6789,10.30.203.3:6789:/volumes/_nogroup/55e46a89-31ff-4878-9e2a-81b4226c3cb2/test
1  5120000M  604M  5119396M   1% /mnt/ceph3
```

and I write data into CephFS (no matter whether I write to the base share or the subdir) , the filesystem usage reported by df never increases for the subdir mount.
As shown above , there were 604 MB stored, then I added 200 MB, but the reported usage of /mnt/ceph3 never increased.

Only if I unmount and remount the subdir mount on /mnt/ceph3, the reported usage then updates to the correct value. Same thing with file deletions, it is updated correctly for only the mount of the base share not the subdir, unless I remount it.

Thanks!

#15 - 06/29/2022 09:18 AM - Xiubo Li

- Copied to Bug #56414: mounting subvolume shows size/used bytes for entire fs, not subvolume added