Ceph - Bug #1432

libvirt: fix definition for rbd params/sources/etc

08/22/2011 12:07 PM - Sage Weil

Status:	Resolved	**Start date:**	08/22/2011
Priority:	Normal	**Due date:**	
Assignee:	Josh Durgin	**% Done:**	0%
Category:	librbd	**Estimated time:**	0.00 hour
Target version:	v0.39	**Spent time:**	2.00 hours
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

History

#1 - 08/22/2011 12:08 PM - Sage Weil

- translation missing: en.field_position set to 16

#2 - 08/22/2011 12:08 PM - Sage Weil

- translation missing: en.field_position deleted (16)

- translation missing: en.field_position set to 14

#3 - 08/22/2011 01:52 PM - Sage Weil

The virtual disk was added to libvirt in 036ad5052b43fe9f0d197e89fd16715950408e1d.

It only lets you specify server hosts, nothing else. Librados wants

- client name/id
- conf path (optional)
- list of conf key/value pairs

These can be smooshed into one config string if need be (this is what qemu takes, actually). For example, "id=foo:conf=/path/to/conf:otheroption=that:foo=bar".

In the end, the <host> items translate into an option like "mon_host=host1,host2,host3".

#4 - 08/22/2011 02:01 PM - Wido den Hollander

Idea from gregaf was to "abuse" the name attribute to do so, have to look into this.

We should check this with the libvirt guys and see what we can do.

Their idea was to have a general format which can be used for Sheepdog, Ceph and NBD and other future network storage which would come up. Instead for writing a implementation for every project.

Some docs/discussing about the current format:

- http://libvirt.org/formatdomain.html
- http://permalink.gmane.org/gmane.comp.file-systems.ceph.devel/1323

Main thing is that we want to push some key=values down to Qemu so that can do down to librbd without manipulating or interpreting it too much.

Some old hack in libvirt allowed virtual disks like:

```
    <disk type='virtual' device='disk'>
      <driver name='qemu' type='rbd' cache='writeback'/>
      <source path='rbd:rbd/beta:conf=/etc/ceph/ceph.conf'/>
      <target dev='vda' bus='virtio'/>
    </disk>
```

Works like a charm, but not what libvirt wants :)

#5 - 08/27/2011 04:41 AM - Wido den Hollander

You can abuse the current libvirt implementation though.

```
    <disk type='network' device='disk'>
      <driver name='qemu' type='raw' cache='writeback'/>
      <source protocol='rbd' name='rbd/beta:conf=/etc/ceph/ceph.conf:id=admin'>
        <host name='monitor.ceph.widodh.nl' port='6789'/>
        <host name='monitor-sec.ceph.widodh.nl' port='6789'/>
        <host name='monitor-third.ceph.widodh.nl' port='6789'/>
      </source>
      <target dev='vda' bus='virtio'/>
    </disk>
```

That works for me and results in a string to Qemu:

```
CEPH_ARGS=-m monitor.ceph.widodh.nl:6789,monitor-sec.ceph.widodh.nl:6789,monitor-third.ceph.widodh.nl:6789 ...
.. -drive file=rbd:rbd/beta:conf=/etc/ceph/ceph.conf:id=admin,if=none,id=drive-virtio-disk0,boot=on,format=raw
,cache=writeback
```

The benefit of passing down a custom string through libvirt is that we never have to wait for them to implement a new scheme.

When RBD starts to become mainstream people will stick to the libvirt version supplied by their distro. If that means they'll have to wait for 2 years to get a new libvirt on their machines, that's a long time.

Right now you can simply put everything you need in the "name" attribute, but we'll have to check with the libvirt guys how they think about that.

#6 - 08/27/2011 12:15 PM - Sage Weil

Hmm yeah, it's at least doable. I just sent an email to libvir-list asking about a more generic syntax for options... that'll let us eventually support using kernel rbd driver as well (in a reasonably non-hacky way). Seem ok? (Can probably move discussion to the list..)

#7 - 08/27/2011 02:27 PM - Wido den Hollander

Saw the e-mail, let's keep the discussion there to make sure everyone involved is up to date.

We could use this issue for some internal work delegation.

#8 - 08/31/2011 09:43 PM - Sage Weil

I wonder if we can harass someone on IRC to get some partial 'sure, whatever' before we go write patches..

Or Wido, maybe you can follow up on the list and see if someone takes notice?

#9 - 09/06/2011 09:49 PM - Sage Weil

- Target version changed from v0.35 to v0.36

#10 - 09/07/2011 10:58 AM - Sage Weil

- translation missing: en.field_story_points set to 5

- translation missing: en.field_position deleted (37)

- translation missing: en.field_position set to 36

#11 - 09/07/2011 11:49 AM - Sage Weil

- translation missing: en.field_position deleted (39)

- translation missing: en.field_position set to 19

#12 - 09/07/2011 05:11 PM - Sage Weil

- Assignee set to Sage Weil

#13 - 09/25/2011 02:14 PM - Sage Weil

- Target version changed from v0.36 to v0.37

#14 - 10/09/2011 08:39 PM - Sage Weil

- Target version changed from v0.37 to v0.38

#15 - 10/13/2011 10:12 AM - Sage Weil

- Assignee changed from Sage Weil to Josh Durgin

#16 - 10/21/2011 10:51 AM - Sage Weil

- Status changed from New to In Progress

#17 - 10/31/2011 10:40 AM - Sage Weil

- Target version changed from v0.38 to v0.39

#18 - 10/31/2011 11:03 AM - Sage Weil

- translation missing: en.field_position deleted (69)

- translation missing: en.field_position set to 1

#19 - 11/15/2011 04:55 PM - Josh Durgin

- Status changed from In Progress to Resolved

Merged upstream.