Smart Proxy - Bug #17554

ISC dhcpd provider leases monitor doesn't keep up with lease file updates in a large network

12/01/2016 12:07 PM - Anonymous

Status:	Closed		
Priority:	Normal		
Assignee:			
Category:	DHCP		
Target version:	1.14.1		
Difficulty:		**Fixed in Releases:**	
Triaged:		**Found in Releases:**	
Bugzilla link:		**Red Hat JIRA:**	
Pull request:	https://github.com/theforeman/smart-proxy/pull/479		

Description

Related issues:

Related to Smart Proxy - Bug #17373: ISC dhcp provider is unable to handle ve...	**New**

Associated revisions

Revision 9445cae3 - 12/12/2016 10:04 AM - Dmitri Dolguikh

Fixes #17554 - inotify events can't overflow the observer anymore

History

#1 - 12/01/2016 12:12 PM - The Foreman Bot

- Status changed from New to Ready For Testing

- Pull request https://github.com/theforeman/smart-proxy/pull/479 added

#2 - 12/01/2016 12:14 PM - Anonymous

- Related to Bug #17373: ISC dhcp provider is unable to handle very big networks added

#3 - 12/01/2016 12:23 PM - Anonymous

Konstantin, what ruby version are you using? I think the PR will resolve the problem you are seeing, but I couldn't verify it, as I cannot get inotify queue to overflow no matter how hard I tried.

#4 - 12/01/2016 02:17 PM - Konstantin Orekhov

I run ruby 2.0 included with CentOS 7:

ruby 2.0.0p598 (2014-11-13) [x86_64-linux]

I'll test your PR and let you know.

Thanks for staying on top of this!

#5 - 12/01/2016 05:59 PM - Konstantin Orekhov

Alas, I don't see any improvements in large deployment - as soon as inotify events start happening (i'm rsync'ing dhcpd.leases from a prod DHCP server into Foreman smart-proxy VM using syncd[1]).

Please let me know if you want to see a debug log or something...

[1] - https://github.com/drunomics/syncd

#6 - 12/01/2016 06:25 PM - Konstantin Orekhov

Well, to follow up on what I mentioned above - there's a difference in inotify events on a system that actually runs ISC DHCP (small environment):

D, [2016-12-01T16:11:38.475486 #1227] DEBUG -- : caught **:modify** event on /dhcp/conf/dhcpd.leases.

D, [2016-12-01T16:12:36.409427 #1227] DEBUG -- : caught :modify event on /dhcp/conf/dhcpd.leases.
D, [2016-12-01T16:13:05.835620 #1227] DEBUG -- : caught :modify event on /dhcp/conf/dhcpd.leases.
D, [2016-12-01T16:13:17.342263 #1227] DEBUG -- : caught :modify event on /dhcp/conf/dhcpd.leases.
D, [2016-12-01T16:13:45.384206 #1227] DEBUG -- : caught :modify event on /dhcp/conf/dhcpd.leases.
D, [2016-12-01T16:13:48.839135 #1227] DEBUG -- : caught :modify event on /dhcp/conf/dhcpd.leases.

And the systems I replicate the change to with syncd (aka rsync triggered by inotify events):

D, [2016-12-01T16:12:39.147068 #30776] DEBUG -- : caught **:moved_to** event on /dhcp/conf/dhcpd.leases.
D, [2016-12-01T16:13:08.279437 #30776] DEBUG -- : caught :moved_to event on /dhcp/conf/dhcpd.leases.
D, [2016-12-01T16:13:20.252106 #30776] DEBUG -- : caught :moved_to event on /dhcp/conf/dhcpd.leases.
D, [2016-12-01T16:13:47.816966 #30776] DEBUG -- : caught :moved_to event on /dhcp/conf/dhcpd.leases.
D, [2016-12-01T16:13:50.247491 #30776] DEBUG -- : caught :moved_to event on /dhcp/conf/dhcpd.leases.

And what I noticed is that in a first case only the changes are loaded, but in second case the whole file is reloaded, which takes ~30 second each time. So this is definitely contributes to the issue I'm observing.
I'll try to figure out the way to test your PR on actual prod DHCP server and see how the performance over there.

Meanwhile, I have a question for you, Dmitri - is it possible to handle :moved_to event the same way as :modify and load only the differences?

#7 - 12/02/2016 04:22 AM - Anonymous

> Meanwhile, I have a question for you, Dmitri - is it possible to handle :moved_to event the same way as :modify and load only the differences?

moved_to is meant to catch dhcpd recreating leases file, which happens every hour. Incremental updates won't work here -- pre-recreate event file descriptor we are holding isn't valid anymore (is not pointing to the correct leases file anymore to be more specific).

#8 - 12/02/2016 04:41 AM - Anonymous

You might be able to get syncd to work with smart-proxy, but this will require changes to syncd: on modify event it will need to use "--inplace --append" rsync options. moved_to event handler doesn't appear to need any changes.

#9 - 12/08/2016 05:23 PM - Konstantin Orekhov

Dmitri Dolguikh wrote:

> You might be able to get syncd to work with smart-proxy, but this will require changes to syncd: on modify event it will need to use "--inplace --append" rsync options. moved_to event handler doesn't appear to need any changes.

Yes, that trick worked - thanks a lot! However, syncd/rsync seem not be stable enough in large environment (dozens of changes per second) - I noticed that after several hours of rsync'ing dhcpd.leases to 2nd and 3rd Foreman nodes, something gets a "lock" (for the lack of better term) on dhcpd.leases on a destination host so updates stop coming in. And I can't figure out what locks that file - lsof does not show anything, but I know to get out of that condition I have to remove a file so updates start coming in. But that's not all - have to stop/start smart-proxy (restart does not help) as smart-proxy log would show that :modify events detected, but it would not load changes. If you have more ideas on this - I'm all ears.

The other thought I'm having is to go back to shared FS (NFS or GlusterFS maybe) to avoid having to rsync that often. The concern you had with NFS (or any shared FS I assume) is that changes to dhcpd.leases are only detected by inotify only on a machine that actually makes that change (the node that runs dhcpd daemon). So, if that's the only concern, I wonder if there's any other way to poke smart-proxy on 2nd/3rd node to reload dhcpd.leases just as if it detected a change through inotify?
If that's a possibility, syncd maybe used to run that command on 2nd and 3rd nodes instead of starting an rsync to them. Thoughts?

#10 - 12/09/2016 09:41 AM - Anonymous

> I noticed that after several hours of rsync'ing dhcpd.leases to 2nd and 3rd Foreman nodes, something gets a "lock" (for the lack of better term) on dhcpd.leases on a destination host so updates stop coming in. And I can't figure out what locks that file - lsof does not show anything, but I know to get out of that condition I have to remove a file so updates start coming in.

This isn't enough information to diagnose the problem. Are incremental updates being used for move_to events to? If so, this might be the problem: dhcpd recreates leases file once every hour and appends changes to the leases file at all other times. It's imperative that the script performing rsync uses incremental updates on modify events (otherwise the whole leases file is parsed on every update) and completely overwrites the file on move_to events (otherwise issues with read offset in the already open file).

> So, if that's the only concern, I wonder if there's any other way to poke smart-proxy on 2nd/3rd node to reload dhcpd.leases just as if it detected a change through inotify?

If you can tolerate ~30sec? delay (that's roughly the time it takes smart-proxy to load in your network?) during failover, you could keep the leases file synchronized by w/e means you deem suitable, and start/restart smart-proxy when the main node fails (the leases file must be on the local filesystem

before smart proxy is launched though).

#11 - 12/12/2016 11:01 AM - Anonymous

- Status changed from Ready For Testing to Closed

- % Done changed from 0 to 100

Applied in changeset 9445cae3964cf836c28074f700c8f8ba4792bf39.

#12 - 12/13/2016 04:35 AM - Dominic Cleal

- translation missing: en.field_release set to 210

#13 - 12/13/2016 05:04 PM - Konstantin Orekhov

- translation missing: en.field_release deleted (210)

> This isn't enough information to diagnose the problem. Are incremental updates being used for move_to events to? If so, this might be the problem: dhcpd recreates leases file once every hour and appends changes to the leases file at all other times. It's imperative that the script performing rsync uses incremental updates on modify events (otherwise the whole leases file is parsed on every update) and completely overwrites the file on move_to events (otherwise issues with read offset in the already open file).

As per your advice, I've modified rsync options to append, which results in :modify events on a receiving side. Let me separate rsync calls and use different options depending on the event type detected on originating node.

> If you can tolerate ~30sec? delay (that's roughly the time it takes smart-proxy to load in your network?) during failover, you could keep the leases file synchronized by w/e means you deem suitable, and start/restart smart-proxy when the main node fails (the leases file must be on the local filesystem before smart proxy is launched though).

Well, each of my Foreman installations is an active/active 3-node "cluster" behind an LB. Only one node runs dhcpd, but all 3 serve as Foreman and as TFTP, BMC, puppet, etc. smart-proxy. I've been trying to do the same for DHCP smart-proxy. Both DHCP and TFTP data is on NFS-mounted FS at this point. With your latest performance-related changes, things improved a lot for me in this setup, but losing NFS support is a problem. Looks like 3-way syncd for TFTP is working rather good, so now I just need to stabilize DHCP portion. The way I'm trying to make it work is the fact that only active dhcpd node would push the changes to other 2 nodes. When/if failover happens and other node picks up dhcpd, it'll become the active one and start shipping changes to other 2 nodes. Hopefully even w/o restarts of smart-proxy processes. But the latter is not so important as it could be done as part of failover procedure by pacemaker.

Let me continue experiments here. Also, I'd like to say one more time - I very much appreciate your attention to this matter. Most of people do not see these kind of problems because of the scale of the environment. I do and since I like Foreman product, I'd like to make it better and more scalable and thus consumable by large audiences.

Thanks!

#14 - 12/14/2016 03:09 AM - Dominic Cleal

- translation missing: en.field_release set to 210