pitrtools - Bug #5310

Use hard links for unsent WAL files

10/29/2014 09:31 AM - Alexander Shulgin

Status:	Resolved	**Start date:**	10/29/2014
Priority:	Normal	**Due date:**	
Assignee:	Alexander Shulgin	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:			
Resolution:			

Description

The cmd_archiver script will keep a copy of WAL file it was unable to deliver to a slave host, thus potentially using a lot of extra space.

We should use hard links instead.

Make sure that archive directory is on the same partition where pg_xlog is. Make this an option (default: on).

History

#1 - 10/29/2014 03:34 PM - Joshua Drake

On 10/29/2014 09:31 AM, pitrtools-tickets@lists.commandprompt.com wrote:

> The cmd_archiver script will keep a copy of WAL file it was unable to deliver to a slave host, thus potentially using a lot of extra space.
>
> We should use hard links instead.
>
> Make sure that archive directory is on the same partition where pg_xlog is. Make this an option (default: on).

Oh, now that is an interesting limitation. Can we soft link instead?

JD

--
Command Prompt, Inc. - http://www.commandprompt.com/ 503-667-4564
PostgreSQL Support, Training, Professional Services and Development
High Availability, Oracle Conversion, @cmdpromptinc
"If we send our children to Caesar for their education, we should
not be surprised when they come back as Romans."

#2 - 10/30/2014 06:29 AM - Alexander Shulgin

pitrtools-tickets@lists.commandprompt.com writes:

> > The cmd_archiver script will keep a copy of WAL file it was unable to deliver to a slave host, thus potentially using a lot of extra space.
> >
> > We should use hard links instead.
> >
> > Make sure that archive directory is on the same partition where pg_xlog is. Make this an option (default: on).
>
> Oh, now that is an interesting limitation. Can we soft link instead?

No, because postgres will remove the original file afterwards. Or we
can try to detect the situation and only make a copy for the first
failed slave, then hardlink to that copy.

However, I believe the same partition limitation is totally worth the
effect compared to potential explosive disk space usage with the current
aproach.

Another option is to ditch the per-slave queue completely and report
failure from archiver if **any** of the slaves has failed. This will make

postgres keep the WAL file in the same manner that we would do by making
a hard-link. With regard to disk usage on pg_xlog both approaches are
the same.

The potential hazard here is that pg_xlog might fill up due to **one**
slave not responding for a while. I don't really think this is a big
problem since any decent monitoring should take care of that. And if
you only have a single slave, you're in a trouble already.

--
Alex

#3 - 10/30/2014 08:47 AM - Joshua Drake

On 10/30/2014 06:29 AM, pitrtools-tickets@lists.commandprompt.com wrote:

> > Oh, now that is an interesting limitation. Can we soft link instead?
>
> No, because postgres will remove the original file afterwards. Or we
> can try to detect the situation and only make a copy for the first
> failed slave, then hardlink to that copy.

Ahh! Now I understand the problem better and it is not actually a
problem. We can still use hard links.

cmd_archiver archives to a location designated in the cmd_archiver.ini .
It does not care about pg_xlog and in fact the only time that PostgreSQL
cares is if cmd_archiver sends a non-success for the archive. This is
the process:

PostgreSQL hands log to cmd_archiver
cmd_archiver archives it in the slave specific queue
if success: tell postgresql the log has shipped the archiver takes
over management
if fail: act like postgresql always does which is to hold the log
file until it reaches a success state

Therefore the limitation doesn't exist because the hardlink would be
from the l_archivedir to the machine queue which would be a hard link.
So in the new version:

postgresql hands log to cmd_archiver
cmd_archiver archives it in l_archivedir and then hardlinks it to each
of the subscriber/slave queues.
if success: tell postgresql the log has shipped the archiver takes
over management
if fail: act like postgresql always does which is to hold the log
file until it reaches a success state

> However, I believe the same partition limitation is totally worth the
> effect compared to potential explosive disk space usage with the current
> aproach.

No. This is a management issue. PITRtools has the ability to alert on
failure of a shipment. If people don't install monitoring that is their
problem not ours.

> Another option is to ditch the per-slave queue completely and report
> failure from archiver if **any** of the slaves has failed. This will make
> postgres keep the WAL file in the same manner that we would do by making
> a hard-link. With regard to disk usage on pg_xlog both approaches are
> the same.

This is how it used to be and that was considered a bug because we want
to be able to take slaves offline without affecting other slaves and
when the offline slave comes up, it can catch up (similar to
wal_keep_segments without the immediate danger).

> The potential hazard here is that pg_xlog might fill up due to **one**
> slave not responding for a while. I don't really think this is a big
> problem since any decent monitoring should take care of that. And if

> you only have a single slave, you're in a trouble already.

Your argument here justifies exactly what I said above, :P

JD

--
Command Prompt, Inc. - http://www.commandprompt.com/ 503-667-4564
PostgreSQL Support, Training, Professional Services and Development
High Availability, Oracle Conversion, @cmdpromptinc
"If we send our children to Caesar for their education, we should
not be surprised when they come back as Romans."

#4 - 10/30/2014 09:20 AM - Alexander Shulgin

pitrtools-tickets@lists.commandprompt.com writes:

> This is how it used to be and that was considered a bug because we want
> to be able to take slaves offline without affecting other slaves and
> when the offline slave comes up, it can catch up (similar to
> wal_keep_segments without the immediate danger).

OK, I get the idea.

To start work on this I need to know the status of Zam's changes: are
they WIP, done, should be merged, etc.

--
Alex

#5 - 11/05/2014 04:22 AM - Alexander Shulgin

- Status changed from New to In Progress

Note to self: we need to make at least one full copy of the WAL file, since the server might recycle it, not unlink. This also leverages the limitation of same filesystem.

#6 - 11/05/2014 09:08 AM - Joshua Drake

On 11/05/2014 04:22 AM, pitrtools-tickets@lists.commandprompt.com wrote:

> Issue #5310 has been updated by Alexander Shulgin.
>
> Status changed from New to In Progress
>
> Note to self: we need to make at least one full copy of the WAL file, since the server might recycle it, not unlink. This also leverages the limitation of same filesystem.

Correct but the archiver already does that? Remember postgresql hands
the wal log to the archiver, after that postgresql no longer cares, the
archiver and queue manager take care of the rest.

JD

--
Command Prompt, Inc. - http://www.commandprompt.com/ 503-667-4564
PostgreSQL Support, Training, Professional Services and Development
High Availability, Oracle Conversion, @cmdpromptinc
"If we send our children to Caesar for their education, we should
not be surprised when they come back as Romans."

#7 - 11/06/2014 06:02 AM - Alexander Shulgin

pitrtools-tickets@lists.commandprompt.com writes:

> Correct but the archiver already does that? Remember postgresql hands
> the wal log to the archiver, after that postgresql no longer cares, the
> archiver and queue manager take care of the rest.

Yes, it's just that I don't fall the victim of thinking that we can
avoid making any real copy altogether in case l_archive is on the same

partition as pg_xlog.

--
Alex

#8 - 11/12/2014 07:49 AM - Alexander Shulgin

- Status changed from In Progress to Resolved

This is handled in #5306.