Infrastructure - Task #4246

Determine why cn-stage-ucsb-1 LDAP syncrepl is failing

2014-01-28 20:15 - Chris Jones

Status:	In Progress	**Start date:**	2014-02-03
Priority:	Normal	**Due date:**	
Assignee:	Michael Campfield	**% Done:**	30%
Category:	Environment.Stage1	**Estimated time:**	0.00 hour
Target version:	2014.26-Block.4.1		
Milestone:	CCI-1.2	**Story Points:**	
Product Version:	1.2.7		

Description

The three CNs in the stage environment should be sync'ing LDAP entries via multi-master replication (syncrepl). All three CNs respond to port 389 requests from each other, so firewalls don't seem to be the issue. I've restarted slapd, and the updates make it to UNM from ORC, but not to UCSB. Because of this issue, I've removed UCSB from the round robin until we fix this issue. Please look deeper into the LDAP logs in /var/log/messages, and figure out why syncrepl communication is not working to UCSB. I haven't tested in the other direction (update an attribute on UCSB, and look on the others), so that test may be telling.

Subtasks:

Task # 4253: Build slapd stage cn log event logging into Splunk	**Closed**
Task # 4276: Correct cn-stage-orc-1 network configs	**Closed**
Task # 4277: Build slapd event log from sandbox CNs into Splunk	**Closed**

History

#1 - 2014-01-30 16:19 - Robert Waltz

Matt Jones indicated to me that we are invoking a 'push' mechanism when replicating in LDAP. He mentioned that the 'pull' mechanism might be more stable since openldap appears to stop responding to those 'push' commands after a while.

We have noted in the past that openldap has a tendency to get into a state where we need to restart ldap. Which is why i built the mechanism into the code to handle restarts.

the problem with the pull mechanism is that it is a polling method and so the immediate consistency is broken. But since we are inconsistent due to the network problems wreaking havoc on the push mechanism, we should investigate the change.

#2 - 2014-01-30 16:49 - Bruce Wilson

- File slapd connections.xlsx added

- File slapd connections.xlsx added

Interesting differences in connections from nodes to other nodes, summarized in attached Excel file (slapd connections.xlsx). Raw data search in Splunk:

```
index=4246 slapd  earliest="01/28/2014:13:00:50" latest="01/28/2014:15:30:00" ACCEPT | rex "from IP=(?P<client
ip>\d[^:]+):" |
rex "from IP=(?P<clientip>\[[0-9:]+\]):\d\d+" | rename host as server|  stats count by server, clientip
```

#3 - 2014-02-10 20:08 - David Doyle

- File slapd connection lost 020714.xlsx added

- File slapd connection lost 020714.xlsx added

We're seeing some instances where connections will be initiated and then instantly lost. These are happening roughly equally across the stage CNs.

Nine total connections are typically lost per minute - three lost on each node. Connections lost in groups of three every twenty seconds, with one lost per node.

See below for one example. A connection from cn-stage-ucsb is made to the three stage CNs then immediately terminated. Twenty seconds later, this occurs on stage-unm, then twenty seconds later, stage-orc.

2/7/14
1:00:14.000 PM

Feb 7 18:00:14 cn-stage-unm-1 slapd[18566]: conn=208961 fd=45 ACCEPT from IP=128.111.54.76:54410 (IP=0.0.0.0:389)
Feb 7 18:00:14 cn-stage-unm-1 slapd[18566]: conn=208961 fd=45 closed (connection lost)
host=cn-stage-unm-1 sourcetype=syslog source=syslog_stage conn=208961

2/7/14
1:00:14.000 PM

Feb 7 18:00:14 cn-stage-ucsb-1 slapd[15198]: conn=40151 fd=22 ACCEPT from IP=128.111.54.76:41210 (IP=0.0.0.0:389)
Feb 7 18:00:14 cn-stage-ucsb-1 slapd[15198]: conn=40151 fd=22 closed (connection lost)
host=cn-stage-ucsb-1 sourcetype=syslog source=syslog_stage conn=40151

2/7/14
1:00:14.000 PM

Feb 7 18:00:14 cn-stage-orc-1 slapd[30236]: conn=200722 fd=47 ACCEPT from IP=128.111.54.76:52862 (IP=0.0.0.0:389)
Feb 7 18:00:14 cn-stage-orc-1 slapd[30236]: conn=200722 fd=47 closed (connection lost)
host=cn-stage-orc-1 sourcetype=syslog source=syslog_stage conn=200722

Summarized in attached Excel file (slapd connection lost 02072014.xslx). Splunk searches used below:

index="stage-cn" earliest="02/07/2014:13:00:00" latest="02/07/2014:15:30:00" | rex "from IP=(?P[[0-9:]+]):\d\d+" |
rex "from IP=(?P\d[:]+):" | transaction conn maxevents=2 endswith=(connection lost) | stats count by host, clientip

index="stage-cn" earliest="02/07/2014:13:00:00" latest="02/07/2014:18:00:00" | rex "from IP=(?P[[0-9:]+]):\d\d+" |
rex "from IP=(?P\d[:]+):" | transaction conn maxevents=2 endswith=(connection lost) | stats count by host, clientip

#4 - 2014-02-10 21:18 - David Doyle

Additionally, we're seeing several instances of log entries matching the following:

Feb 7 20:27:20 cn-stage-orc-1 slapd[30236]: connection_read(17): no connection!

This message shows up on orc and unm stage, but not ucsb. No regular pattern pops up, and early analysis doesn't connect these messages to longer transactions. (Piping a "no connection" search string to "transaction conn" in Splunk returns no results. See http://docs.splunk.com/Documentation/Splunk/5.0.4/SearchReference/Transaction for more on transaction search function.)

In the six hour period defined in slapd connection lost 020714.xslx (2014-02-07, 13:00 - 18:00 EST), 49 of these show up on orc stage, and 59 show up on unm. Early analysis shows no regularity in occurrence of these events.

openldap.org bug reports appear to indicate that occurrences of this error are linked to an issue with the libldap client library provided by older versions of OpenLDAP. See http://www.openldap.org/its/index.cgi/Software%20Bugs?id=6548 :

Many "connection_read(): no connection!" warnings are written to /var/log/debug
and /var/log/syslog by slapd. As stated at
 http://www.openldap.org/lists/openldap-software/200811/msg00079.html , this is
apparently not a problem with slapd, but a client that is disconnecting without
first unbinding.

This appears to be an issue with the libldap client library provided by OpenLDAP
itself (2.4.21), and not the slapd daemon.

This would be the version we appear to be using, which we might be stuck with until Ubuntu is upgraded over 10.04.

Other potential events found, more to follow. Will look into changing syncrepl entries in slapd.conf to see if that clears anything up this evening or tomorrow.

(Thanks to rwaltz for leads on most of this.)

#5 - 2014-02-10 21:38 - David Doyle

We're seeing several of these warnings on the stage nodes.

Feb 7 22:57:10 cn-stage-unm-1 slapd[18566]: <= bdb_equality_candidates: (d1NodeServiceId) not indexed

Feb 7 22:57:10 cn-stage-unm-1 slapd[18566]: <= bdb_equality_candidates: (d1NodeId) not indexed

Using the timeframe described previously, d1NodeServiceId shows 2700 entries across the stage nodes and d1NodeId shows 3366.

Events show up on ucsb more often than on the other two nodes.

Noting now for work later. Will analyze further tomorrow.

Early Splunk searches:

index="stage-cn" earliest="02/07/2014:15:00:00" latest="02/07/2014:18:00:00" bdb_equality_candidates (d1NodeServiceId)
index="stage-cn" earliest="02/07/2014:15:00:00" latest="02/07/2014:18:00:00" bdb_equality_candidates (d1NodeId)

#6 - 2014-02-12 02:25 - David Doyle

Googling the warning above (bdb_equality_candidates (uid) not indexed) turns up this thread about N-way multimaster replication:
http://ubuntuforums.org/showthread.php?t=1020472&page=2

This thread mentions the bdb_equality_candidates warning as well as others we've seen (do_syncrepl: retrying being one of them). From post 14:

/etc/hosts - Make sure that you have your FQDN in /etc/hosts as 127.0.0.1. I skipped over this initially and it was a big problem.

Looking over /etc/hosts on our stage boxes:

cn-orc

127.0.0.1 localhost
127.0.1.1 cn-stage-orc-1.test.dataone.org cn-stage-orc-1

cn-ucsb

127.0.0.1 localhost
128.111.54.76 cn-stage-ucsb-1.test.dataone.org cn-stage-ucsb-1

cn-unm

127.0.0.1 localhost
64.106.40.8 cn-stage-unm-1.test.dataone.org cn-stage-unm-1

Not an exact match of the problem described, but after looking through /etc/hosts on our other environments, cn-stage-orc-1 appears to be the only box that doesn't reference its own IP address, instead referencing 127.0.1.1. The other ORC boxes (where I assume syncrepl is working as expected) are configured with 127.0.0.1 and their own IP addresses as expected.

Noting this for further discussion, and reading up further in the meantime. Would changing/adding stage-orc's IP address to /etc/hosts change the way syncrepl behaves? Would changing/removing the 127.0.1.1 entry break something else somewhere?

#7 - 2014-02-12 17:55 - David Doyle

Looking into changing over that /etc/hosts entry led me to find that /etc/network/interfaces on cn-stage-orc is set up to use dhcp for eth0. Not sure if that's part of the problem as well, but A) I don't see dhcp being used on any of the other nodes, and B) I'm pretty sure I'll have to change that anyway if I want to change /etc/hosts accordingly.

#8 - 2014-02-14 00:42 - David Doyle

Unfortunately, fixes to cn-stage-orc-1 networking settings changed nothing with multi-master replication behavior on stage.

After testing post-networking changes, entries changed on orc populate to unm and vice versa, but ucsb will not accept changes, and changes made on ucsb do not make it to the other nodes.

Additionally, still seeing warnings and errors listed previously. Will need to check further to compare numbers to see if they're showing up in the same quantities/proportions across the nodes.

Will look into trying an increase in concurrent connections allowed next.

#9 - 2014-02-14 16:02 - David Doyle

One other thing to note: We appear to be getting good slapd connections to and from ucsb stage during the times in question.

Connection from unm to ucsb:

Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 fd=24 ACCEPT from IP=64.106.40.8:47595 (IP=0.0.0.0:389)
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 op=0 EXT oid=1.3.6.1.4.1.1466.20037
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 op=0 STARTTLS
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 op=0 RESULT oid= err=0 text=
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 fd=24 TLS established tls_ssf=128 ssf=128
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 op=1 BIND dn="cn=admin,dc=dataone,dc=org" method=128
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 op=1 BIND dn="cn=admin,dc=dataone,dc=org" mech=SIMPLE ssf=0
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 op=1 RESULT tag=97 err=0 text=
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 op=2 SRCH base="dc=org" scope=2 deref=0 filter="(objectClass=*)"
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 op=2 SRCH attr= +
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 op=2 SEARCH RESULT tag=101 err=0 nentries=0 text=
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 op=3 UNBIND
Feb 14 00:12:21 cn-stage-ucsb-1 slapd[15198]: conn=267698 fd=24 closed

Also, I went through /etc/ldap/slapd.conf line by line on orc and ucsb stage to compare for possible simple typos or differences between them and compared both to the same file on cn-orc-1. Found nothing.

#10 - 2014-02-14 18:29 - David Doyle

- Status changed from New to In Progress

Robert and Matt brought up the possibility of ldap database corruption on stage ucsb being an issue, especially given the errors we're seeing and the fact that some connections to/from ucsb appear to be working fine, but some aren't. Need to read up on slapcat, run it on ucsb and a couple of comparison boxes, and compare the databases to check for issues.

#11 - 2014-02-16 01:22 - David Doyle

Some updates:

- tcpdump analysis via Wireshark shows several ldap connections taking place between stage orc and ucsb, handshakes and tls starts taking place, and then those connections returning encryption errors and closing.
- /var/log/slapd.log on cn-stage-ucsb-1 shows the following lines:

hdb_db_open: database "dc=org": unclean shutdown detected; attempting recovery.
hdb_db_open: database "dc=org": recovery skipped in read-only mode. Run manual recovery if errors are encountered.

No timestamps in place on that log, unfortunately, but this lends credence to the database corruption theory. Will slapcat stage dbs next.

#12 - 2014-02-17 04:36 - David Doyle

- File ucsbslapcat.ldif added

- File orcucsbslapdat.png added

- File unmslapcat.ldif added

- File orcslapcat.ldif added

- File orcucsbslapdat.png added

- File unmslapcat.ldif added

- File ucsbslapcat.ldif added

- File orcslapcat.ldif added

Uploading file - orcucsbslapdat.png

This is a screengrab from a diff ran against two ldif files generated by slapdat on stage ORC (left column) and stage UCSB (right column). -Note the first line in each column - evidence of database corruption?- *Working as intended per cjones, DD 2/26*

Running diff against slapdat ldifs from the stage CNs appears to show that each of the stage CNs is running on a significantly different ldap database from the other two CNs.

#13 - 2014-02-17 04:36 - David Doyle

- File deleted (unmslapcat.ldif)

#14 - 2014-02-17 04:37 - David Doyle

- File deleted (ucsbslapcat.ldif)

#15 - 2014-02-17 04:37 - David Doyle

- File deleted (orcslapcat.ldif)

#16 - 2014-02-26 05:02 - David Doyle

Found what I thought was an issue with ldap replication traffic encryption between orc and ucsb stage, but further analysis with Robert showed that to be working as intended. Robert thinks that database corruption might still be at issue due to what he sees in the tcpdumps we looked at.

Created some more slapcat ldifs from the stage CNs for further comparison. The differences I'm seeing between the databases so far fall into two categories:

- several dns that are otherwise equal show differing entries under entryCSN and modifyTimeStamp
- several dns appear to be either missing altogether from one or more ldap database, or the databases have entries that are in different order from one another

#17 - 2014-03-05 03:22 - David Doyle

After running another round of ldap sync replication tests, it now appears that sync replication is broken across all three stage CNs. I made changes on all three stage CNs, and those changes were not picked up by any other stage CN.

#18 - 2014-03-05 03:39 - David Doyle

Restarting slapd across the stage environment seems to have restored the level of functionality we previously had (ucsb neither replicating nor accepting replications, orc and unm replicating and receiving to/from one another).

Found a possible firewall configuration issue at ucsb. Ufw wasn't configured to accept connections on port 389 from its own IP address, which was the case at orc and unm. I added a rule at ucsb to allow 389 connections from itself before restarting slapd on stage, but this doesn't appear to have helped.

#19 - 2014-07-01 04:51 - Robert Waltz

*- Product Version changed from * to 1.2.7*

- Target version set to 2014.26-Block.4.1

- Assignee changed from David Doyle to Robert Waltz

- Milestone changed from None to CCI-1.2

#20 - 2014-07-21 00:18 - David Doyle

Began post-12.04 testing Friday, 7/18/2014.

Initial tests showed a lack of replication to/from cn-stage-unm-1, but replication to/from cn-stage-ucsb-1 was behaving correctly. Restarted slapd on unm and replication began behaving correctly across all three stage-1 CNs.

Further tests today confirm replication happening as normal across stage-1 CNs.

Will report on Monday and see if further tests are required.

#21 - 2014-07-21 17:05 - David Doyle

- Assignee changed from Robert Waltz to David Doyle

Per Chris at standup today, I will need to induce an outage on one of the stage CNs, make a change to ldap, and then induce a reconnection and see if the stage environment recovers and propagates normally.

#22 - 2014-07-24 04:30 - David Doyle

Spent the evening performing further testing by bringing slapd offline on individual stage CNs, making changes to the directory on other stage CNs, restarting slapd on the tested CNs, and checking the directory on all CNs for replication All other elements being equal, stage CNs appear to be able

to recover from outages.

Testing was hindered somewhat by what appear to be connection issues w/cn-stage-unm-1. Relatively frequent but intermittent cases would come up where replication to cn-stage-unm-1 would not occur naturally and would require a slapd restart. Replication from cn-stage-unm-1 does not appear to exhibit this problem. Based on previous conversations with coredev, this might be due to intermittent connection issues over the wire causing replication attempts to timeout.

Will look over previous notes re: various windows for connection attempts (duration, number of attempts) before timeout.

#23 - 2014-10-31 17:42 - David Doyle

- Assignee changed from David Doyle to Michael Campfield

Moving this over to Michael for future work. Explained the issue, the proposed fix, etc. to him. This is ready to be moved forward on as soon as coredev has time to implement in non-prod, test, then implement in prod, then move changes over to whatever dataone install/postinst scripts they need to be added to.

Files

slapd connections.xlsx	34.7 KB	2014-01-30	Bruce Wilson
slapd connection lost 020714.xlsx	4.93 KB	2014-02-10	David Doyle
orcucsbslapdat.png	35.3 KB	2014-02-17	David Doyle