**November 2016 - MCollective Deployment Issues**

**2016/11/02 Meeting**

Topic: Deployment Issues

Attendees: John, Kel, Russell, Sean, Adrien, Charlie

Plan is to do 2 deployments, first to a small subset of stores, and next to a larger subset of stores.

Due to wrong timing, we missed the smaller deployment, so only caught the second larger deployment.

Just as the deployment to ~820 nodes was about to begin, the pe-activemq service died on one of the hubs. They couldn’t get the hub back up and running before the end of the run, and the run failed on almost all deployments, only about 50 succeeded.

Decide to do a run against ~200 stores with one hub disabled and leave metrics enabled. Listening on e6b, not listening on e5f

Noticed on e5f that there were outgoing connections to one hub, but also to the CA/MoM. Disabled puppet and activemq on lp0e69. Looked at logs on e5f, noticed it was still trying to talk to e69. Charlie asked them to look at /etc/puppetlabs/activemq/activemq.xml and noticed the jetty.xml configuration was missing.

A scheduled test was run against 10 stores (~40 servers). On that run, 5 servers failed, 4 were at one store.

Now we’re going to try and do the run on 200 nodes. All nodes should be running through 6b, since 69 and 64 are both shut down. Noticed that pe-activemq service had just been restarted. Did further investigation and noticed that daemonized puppet runs were turning web console off, and puppet agent -t was turning back on. Decided to puppet agent -disable on 6b and 5f to eliminate that issue for this test.

We did the test, and it only passed on 34/200 nodes

Adrien asked about the network quality, and due to stores being all over the world, the network quality really varies. Although, they haven’t noticed a correlation between issues with the network and the mco issues.

Then ran just mco rpc status against 50 nodes with a 45 s timeout. All 50 came back. Bumped timeout down to 30s, then 5s and we still got all 50 back.

Looking at their broker code, saw that there were --batch 300 and --batch-sleep args in script, so adding those to test. Still worked. Upped hte number of target nodes , all worked at 50, 200, 820 only returned 65,1 and 650 only returned 529. Decided might have something to with r10k because they pass filters when doing r10k runs. Kept doing status tests:

200/200

400/263

300/186

200/200

250/241

225/200

201/201

210/210

220/220

224/208

223/203

221/201

Identified tipping point was 220, Charlie noticed that 220 is divisible by 4, so there may be something weird happening when they hit store 56. Randomized list, ran against 200/199. Randomized again, 200/197. Tipping point stopped being reliable.

Found list of nodes that failed, going to look at /var/log/puppetlabs/mcollective.log - that will show if they’re losing connection to broker or something is happening there. Looked at one successful node, one that failed, and noticed that one received message and one didn’t, both coming from same broker.

Now going to look at the 5a broker. Saw address already in use message on broker

Russell is asking if they are ever going to be able to figure this out because it’s too complex? At the end of the day, they have to be able to send to 2000 stores, 8000 devices. How do they reduce complexity?

Going to keep 5f spoke live, for all other spokes, disable puppet agent, stop the service. Reset log settings on 5e.

On 5f, went into activemq.xml, changed transport 61613 line to add +nio+ssl. Restarted service, pid file said it was running, but service pe-activemq status said not running. Mco pings were responding and servers were responding

Need to get together later today: when we can test, what the debug plan looks like. Next round is Sunday night / Monday morning. Send us /var/log/activemq directory.

\*all spokes were still configured to talk to 69 still, don’t know why

\*if it goes through more than one broker, we still get the same result

**2016/11/3 Meeting**

Topic: MCollective Deployment Issues Update

11/3

Call with John

Blackout period starts Friday 11/18 at 6am. Change window is from Sunday at midnight to Friday at 6am.

Deployment window is 6 hours. Today with opsware they have changes that are rolling out on a regular basis. For tonight, they have changes in 20 minute sections from midnight to 5:30. 3 change sessions per hour per facility. 6 sessions over 5 hours - 30 jobs. In an average night, we have between 20-40 jobs. No hard rules to say deployments have to take less than 20 minutes. Concern is one night I have changes going, one are on single digits, rest is on 4500 devices. They have changes where they put a restriction on how many changes can be batched together. That’s in the 400 range. There’s a lot of changes in that window. If the change took 20 minutes for the whole deployment. That’s ok.

Most changes are being sent to ISP (in store processor), ISPA is pri

STUX01 -

ISPA - in store processor application

PSP - point of sale processor

There’s one tool rental application that was a pilot and it only went to 2 stores. They created a validation package to test deployments. 6 week rollout schedule to stores. They need a VP to say yes

Pilot - 5 stores, 10, 50, ramp up to 200 started last week.

Get confirmation from John on tolerable success rate.

Follow up with Bill Weisse, Mike Stahnke, and Deepak

End of day updates.

Every action that submits a large number to individually addressed nodes has potential to submit a message to the queue that can’t be consumed because the node is unresponsive. That message with persist up to 1min 30 seconds. If in that period additional requests are submitted, that can add more undeliverable requests. If that number exceeds 200, then all further delivery will halt until the undeliverable messages in the queue drop below 200.

This could impact home depot because they’re getting a list of nodes to connect to from an external source, rather than discovery. Use activemq admin command to look at the size of that queue. The solution would be to make the modification to the policy entry for the queue=> and increase the maxPageSize

Found a lightweight way to query metrics that doesn’t require enabling the jmx console. This is a win, since enabling the jmx console can cause pe-activemq to restart. Now support can run a command and give us a result, rather than spending 3 days to enable the JMX console.

Largest number of nodes they address at a time \* number of times they will do that per 90 seconds

Summary of current Home Depot state.

The Home Depot purchased Puppet to replace HPSA. Their primary use case is using Puppet to deploy applications. To do that, they use a series of mcollective commands. They have been doing this successfully in their internal datacenter environment for more than a year.

Recently, they have extended this workflow to their retail store environments. The deadline for developers to deploy their store applications through Puppet rather than HPSA was November 1st. As a result, the number of endpoints mcollective was hitting increased rapidly over the course of the last 2 weeks. The maximum number of endpoints is over 8k.

Last week Thursday 10/27, Home Depot opened a support ticket due to machines dropping connections during deployments. When mcollective sent messages to a large number of geodiverse endpoints, mcollective was not consistently hearing back from the entire set.

Due to the complex nature of their mcollective topology, we have not been able to isolate where the connections are being lost. Furthermore, the component that is driving these connections is ActiveMQ, which mcollective is built on top of, and is not our tooling. Charlie, Adrien Thebo and Elizabeth spent several hours trying to diagnose this last night, and were able to rule out network latency as the issue, but could not determine the source of the lost connections due to lack of activemq expertise.

Most of our larger customers use Mcollective to drive their orchestration. Mcollective is included and supported in the LTS version, and our new orchestrator is also built on top of ActiveMQ. We will continue to experience these issues, and need the expertise to troubleshoot complex ActiveMQ topologies at scale with geo-diverse endpoints.

**11/04/2016**

Did mco ping, came back with 8681, metrics showed 8678

- Running mco status against 300 stores with 30 timeout - 207 out of 300 messages got sent.

- removed, json flag, tried again - got 263 responses, reply line - this suggests it could be caused by the publish\_timeout limit

- edited status script to change puslish\_timeout to 10 s

- got 300/300

- increasing node target to 820 - 707/820

- 13:04 - increased status script to change publish\_timeout to 15

- ran again status again 820, 30

- 820/820 responses

- 13:05:33

- req enque 1777, time 218, qs 0 , dq, 1777

- res 263, 284, 0, 263

Decreased batch size to 50

* Ran status 820, 30
* 13:07
* 803/820 response at 13:11

10:16 Pacific - turned log output off

How long activemq broker has been running on 5f - 13:08 - at 13:08 puppet agent triggered refresh of pe-activemq service. This is due to setting jmx console to false

Rerun at 13:19 - status 820, 30

* 13:22:03
* 818/820
* Noticed 1 or 2 messages sitting in the queue that hadn’t been dispatched.
* Took about 3 minutes

Decreased publish\_timeout to 10s

* Run at 13:27:23
* 820,30
* queue size - 2, 3 (13:28),
* 816/820

Enabling metrics on 6b and 64 - if there’s messages missing, they may have ended up on 69

13:37:16 run 820,30

* Restart pe-activemq part way through
* Most messages going through 64
* 819/820

Shown tha tpublush timeout and batch size have an impact. WE can shut webservice off that will stop those restarts.

Plan for Sunday:

* Batch size and publish timeout parameters - hopefully based on what we’ve seen today that will give us a higher success rate
* Scheduled change for Sunday night starts at Midnight Eastern - start call at 11:30 eastern
* Get amq status script on all of the spokes. Charlie will send a script htat has a throttle on the amount the metrics script output
* Send us the client log -
* Put brokers into debug logging on Sunday