Main page
Recent changes
Server admin log (Prod)
Server admin log (RelEng)
Deployments
SRE/Operations Help
Incident status

Cloud VPS & Toolforge
Cloud VPS documentation
Toolforge documentation
Request Cloud VPS project
Server admin log (Cloud VPS)

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Incident documentation/20150812-LabsOutage

< Incident documentation

Contents [hide]

Summary

Several openstack nova services began flapping around 22:30 on 2015-08-12. Symptoms included intermittent puppet failures, ssh access failures, and general network downtime for labs instances.

The ultimate cause is unclear, but it was probably provoked by "salt 'labvirt*' cmd.run 'puppet agent -tv" or some sort of order-dependency in the restarting of nova services.

The problem was ultimately resolved via a manual restart of rabbitmq-server on labcontrol1001. All services were restored to normal by 23:30.

Timeline

- [22:00] After lots of code-reading and discussion, Chase and Andrew Bogott conclude that the nova setting network_host is obsolete and unused. Andrew merges https://gerrit.wikimedia.org/r/#/c/231177/. Andrew forces a puppet run on labnet1001 and labvirt1009 to verify that nothing breaks -- and (seemingly) nothing does.
- [22:05] Andrew forces a puppet run (via salt) on all labvirt nodes, apparently without ill effect.
- [22:30] Lots of shinken alerts about labs puppet staleness arrive. Andrew investigates and sees that some instances are failing to load their facts, and that many instances are unable to reach the nova metadata service. Around this time icinga throws alerts about nova-compute services being down; Andrew restarts them.
- [22:55] Andrew reverts the previous patch with https://gerrit.wikimedia.org/r/#/c/231189/ -- there's still no theory for why this would have caused the problem, but it seems the safe thing to do.
- [23:00] At this point the most severe user-facing symptom appears: ssh to the labs bastion fails. First, ssh notifies us that the hostkey has changed. If the new hostkey is accepted, subsequent ssh attempts fail due to a rejected key.
- [23:10] Icinga starts throwing alerts about nova-compute services again. Andrew restarts them, again. The logs are full of things like this:

```
2015-08-12 22:44:42.125 55787 TRACE nova.openstack.common.periodic_task   File
"/usr/lib/python2.7/dist-packages/oslo/messaging/_drivers/amqpdriver.py", line
412, in send
2015-08-12 22:44:42.125 55787 TRACE nova.openstack.common.periodic_task
return self._send(target, ctxt, message, wait_for_reply, timeout)
2015-08-12 22:44:42.125 55787 TRACE nova.openstack.common.periodic_task   File
"/usr/lib/python2.7/dist-packages/oslo/messaging/_drivers/amqpdriver.py", line
403, in _send
2015-08-12 22:44:42.125 55787 TRACE nova.openstack.common.periodic_task
result = self._waiter.wait(msg_id, timeout)
2015-08-12 22:44:42.125 55787 TRACE nova.openstack.common.periodic_task   File
"/usr/lib/python2.7/dist-packages/oslo/messaging/_drivers/amqpdriver.py", line
267, in wait
2015-08-12 22:44:42.125 55787 TRACE nova.openstack.common.periodic_task
reply, ending = self._poll_connection(msg_id, timeout)
2015-08-12 22:44:42.125 55787 TRACE nova.openstack.common.periodic_task   File
"/usr/lib/python2.7/dist-packages/oslo/messaging/_drivers/amqpdriver.py", line
217, in _poll_connection
2015-08-12 22:44:42.125 55787 TRACE nova.openstack.common.periodic_task     %
msg_id)
2015-08-12 22:44:42.125 55787 TRACE nova.openstack.common.periodic_task
MessagingTimeout: Timed out waiting for a reply to message ID
b7e615f4909144e480f34f78d1dce791
```

There are also occasional complaints about contacting nova-conductor. Andrew restarts nova-conductor.

- [23:15] When services restart, they appear to be working, but then fail again after a few minutes. Failed services include nova-network on labnet1001.
- [23:23] Andrew explicitly restarts rabbitmq-server on labcontrol1001, then restarts nova services yet again, elsewhere. This time the services come up and stay up.
- [23:25] Normal access is restored. Labs bastions are now returning the original (pre-incident) hostkey, ssh attempts are accepted, all is well.

Conclusions

All of the above symptoms can be explained by a failure of the nova-network service:

- Routing to the metadata service came and went as nova-network came and went. The correlation wasn't completely obvious since it takes nova-network a few minutes to set things up and, of course, it was crashing and timing out a lot.
- The host-key thing can be explained by a failure in routing. If nova-network is down, what happens to labs floating ips? Presumably they are routed directly to labnet1001. And, indeed, I've confirmed that the hostkey offered up during the outage is the host key on labnet1001.
- Similarly with the ssh keypair mismatch -- users were providing a labs key and it was failing to match the production account/key on a production box, labnet1001.

So, symptoms were probably a result of nova-network failure. Andrew's current theory is that it was not due to the specific content of the applied patch (since the problem persisted after reverting) but instead was a result 1) a race or flood caused by restarting too many services at the same time that caused rabbitmq to misbehave, or 2) an unrelated, coincidental failure of rabbitmq.

Actionables

Upgrading all of our openstack infrastructure is probably a good idea.

Vigilance during service restarts is always warranted. Apparently it's also a good idea to roll things out gradually and skip the 'just force puppet to update everything' step.

Our monitoring served us fairly well during this incident, but it would be nice to have something that explicitly monitors rabbitmq.

Chase and Andrew are committed to re-deploying https://gerrit.wikimedia.org/r/#/c/231177/. That should at least eliminate that patch as a culprit.

> UPDATE: That patch has now been merged without incident. andrew (talk) 15:16, 14 August 2015 (UTC)

Category: Incident documentation