Ceph and NTP clock skew

Blog Post created by Ben England on Oct 5, 2017

Like • 1 Comment • 0

some folks in Perf & Scale team have complained lately about Ceph health warning that monitors cannot synchronize due to clock skew. So I started out trying NTP on my laptop to see what was going on.

```
$ ntpq -p
bash: ntpq: command not found...
Install package 'ntp' to provide command 'ntpq'? [N/y] y
...
```

I have fedora 26 on my laptop, I just installed NTP and the default NTP server is 2.fedora.pool.ntp.org, when I ping it I get:

```
$ ping 2.fedora.pool.ntp.org
PING 2.fedora.pool.ntp.org(helium.constant.com (2001:19f0:200:144b::2000)) 56 data bytes
From 2620:52:0:1250::1fd (2620:52:0:1250::1fd) icmp_seq=1 Destination unreachable: Address unreachable
...
```

That's because IPV6 is not allowed in Red Hat's internal network (yes I complained about that). A mojo search found that the corporate NTP server was available at clock.redhat.com, so I tried inserting that into /etc/ntp.conf, restarting NTP and re-running.

```
$ ntpq -p
     remote           refid      st t when poll reach   delay   offset  jitter
==============================================================================
2604:a880:2:d0: .INIT.          16 u    -   64    0    0.000    0.000   0.000
clock01.util.ph .CDMA.           1 u    9   64    1   77.691    0.162   0.000
```

which means clock.redhat.com is giving me 77 msec delay, I think. Why so large? Look at what happens when I traceroute clock.redhat.com. The first time, it resolves to some local time server:

```
$ traceroute clock.redhat.com
traceroute to clock.redhat.com (10.16.255.1), 30 hops max, 60 byte packets
1  gateway (10.18.81.254)  1.366 ms  1.281 ms  1.258 ms
2  10.18.255.66 (10.18.255.66)  0.438 ms 10.18.255.64 (10.18.255.64)  0.420 ms 10.18.255.66 (10.18.255.66)  0.374 ms
3  10.16.253.49 (10.16.253.49)  3.928 ms 10.16.253.55 (10.16.253.55)  1.009 ms  1.529 ms
4  clock.bos.redhat.com (10.16.255.1)  2.173 ms  2.442 ms  2.637 ms
```

The second time, it went intergalactic!

```
$ traceroute clock.redhat.com
traceroute to clock.redhat.com (10.5.27.10), 30 hops max, 60 byte packets
1  gateway (10.18.81.254)  1.486 ms  1.443 ms  1.398 ms
2  10.18.255.66 (10.18.255.66)  11.305 ms  11.375 ms  11.419 ms
3  10.16.253.39 (10.16.253.39)  101.215 ms  101.246 ms  101.223 ms
4  unused (10.4.253.6)  86.131 ms  86.161 ms unused (10.4.253.4)  84.103 ms
5  * * *
...
30  * * *
```

Why different places for same hostname?

```
$ dig clock.redhat.com

; <<>> DiG 9.11.1-P3-RedHat-9.11.1-2.P3.fc26 <<>> clock.redhat.com
...
;; QUESTION SECTION:
;clock.redhat.com.              IN      A

;; ANSWER SECTION:
clock.redhat.com.       37      IN      CNAME   clock.corp.redhat.com.
```

```
clock.corp.redhat.com.  300     IN      A       10.5.26.10
clock.corp.redhat.com.  300     IN      A       10.11.160.238
clock.corp.redhat.com.  300     IN      A       10.16.255.1
clock.corp.redhat.com.  300     IN      A       10.5.27.10
...
```

So NTP daemon will try to communicate with one of these 4 entries, which may or may not be local. However, Ceph only allows 50 millisec of clock skew in monitors by default. So how do I find a local server? This command uses DNS backtranslation to show the hostname of every IP address which appears under above command to resolve hostname clock.corp.redhat.com:

```
$ for ip in `dig clock.redhat.com | awk '/^clock.corp.redhat.com/{print $5}'` ; do \
   dig -x $ip ; done \
  | grep redhat.com
238.160.11.10.in-addr.arpa. 65  IN      PTR     clock1.rdu2.redhat.com.
1.255.16.10.in-addr.arpa. 28800 IN      PTR     clock.bos.redhat.com.
10.27.5.10.in-addr.arpa. 81759  IN      PTR     clock02.util.phx2.redhat.com.
10.26.5.10.in-addr.arpa. 2925   IN      PTR     clock01.util.phx2.redhat.com.
```

so we see that there is a NTP server in the Westford, MA office at 10.16.255.1 (yes you have to read the IP address backwards, ughh) and one in the Raleigh, NC USA RDU2 lab at 10.11.160.238. So I switched to this in ntp.conf:

```
server 2.fedora.pool.ntp.org iburst
server clock.bos.redhat.com
```

Note that the first entry is unreachable - I kept it for when I'm using my laptop outside redhat and I want to use IPV6.

```
$ ntpq -p
     remote           refid      st t when poll reach   delay   offset  jitter
==============================================================================
kcolford.com    .INIT.          16 u    -   64    0    0.000    0.000   0.000
clock.bos.redha .CDMA.           1 u   10   64    1    0.864    0.666   0.000
```

and now my delay is much better. I would recommend a local NTP server for any Ceph cluster. It may take some time after time skew is eliminated before ceph monitors clear the health warning.

Visibility: Ben England's Blog • 83 Views

Last modified on Nov 9, 2017 4:18 PM

Tags: ceph Edit tags

Global Reach
0%

Impact 31
Sentiment **Neutral** 3

0 Comments

Related Content

- TLS Everywhere on OSP12
- Defect DE6795: Time is out of sync on AWS instances
- Daily Work Record-Mar23
- Adding a DNS record to PowerDNS in the Ceph Octo lab
- (RESOLVED) Cannot access bigmachine from RHT

Recommended Content

- OSP15 pre-beta with Ceph
- Is your root password redhat?
- New Password Standards at Red Hat
- How to set up a Yubikey with Red Hat Two Factor Authentication Services
- Red Hat Storage New Hire Guide - Bangalore