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/20130617-Email

< Incident documentation

Copy/paste from Faidon's email to Ops

TL;DR: we've had a mail outage over the past few hours. The underlying cause was spammers using From: $random@wikimedia.org to send spam, which resulted in bounces (backscatter) coming to our servers at a high rate, cascading and producing multiple systems failure. The root causes were fixed/workarounded and mails are flowing again, with delays in mail delivery to be expected over the next few hours.

The outage started at about 01:00 UTC and initially noticed by Ariel at 13:00 UTC which subsequently debugged this as resulting from a sanger LDAP failure. Due to a pre-existing replication misconfiguration/miscommunication, our LDAP was trying to replicate from the non-existent sfo-intranet.corp.wikimedia.org which resulted in Ariel misdiagnosing this as an OIT problem and escalating to Chip, via Chris Johnson. This misconfiguration was a preexisting error and did not contribute in today's outage.

I noticed the IRC discussion at 13:45 UTC and started troubleshooting. I initially fixed the replication issue by placing sfo-aaa1.corp.wikimedia.org's address to a sfo-intranet.corp.wikimedia.org /etc/hosts entry, which quickly worked around the replication issues.

The next problem was that OpenDJ was getting too many LDAP queries and responded too slow, so it reached ulimit for too many open fds (1024) and as soon as it got such an errno shut down the LDAP listening socket altogether (OpenDJ bug). This was worked around by adjusting the ulimit.

The underlying cause for this was that we had too many delivering mail attempts for $random@wikimedia.org which pushed too many queries to the LDAP server which was then unable to handle them. At about 400 searches/sec, OpenDJ took hundreds of milliseconds to respond, only piling up the number of connections there and aggravating the issue. I woke up yan at 14:30 UTC to help with OpenDJ.

While mchenry was unable to process mail due to the LDAP failure (either with the listening socket was down, or with LDAP being slow), mail relays were connecting to our secondary MX, sodium, which happily queued them and attempted to subsequently deliver them to mchenry. I therefore stopped sodium's mail listener and just left exim queue runners to process the queue. mchenry's load went through the roof (~300), which resulted in exim queueing mails (because of queue_only_load = 100.0), which then increased the number of queued mail and runners. I manually set queue_run_max = 50 to make the load a bit more manageable and restore mchenry to be responsive.

Meanwhile, Ryan worked through the OpenDJ issue, initially adding configuration options to get statistics and then diagnosing the issue as twofold: older OpenDJ version plus missing indexes (apologies if I'm oversimplifying :). He added indexes, both to attributes as well as specific VLV (query) indexes, finishing this work at 15:35 UTC.

This made this:

```
 Search: 40939  Avg: 125.223 ms Max: 2521 ms >100ms: 19544 (47%) >1000ms: 84
(0%)
```

into this:

```
 Search: 26429  Avg: 0.198 ms   Max: 252 ms  >100ms: 1 (0%)  >1000ms: 0 (0%)
```

This immediately starting having an effect, increasing mchenry's mail processing speed. I ran exim4 -qff on sodium at 15:35 UTC to force the queue to run overriding cached retry options; its queue had ~180K mails at the time. This resulted into queued mails being delivered to mchenry which subsequently delivered them to their recipients on rejecting them (99% of those being bounced spam which per spec doesn't generate bounces).

Mail queues are on the way to be emptied (as of 17:00 UTC, ~7k mails in the queue) lists was restored moments

ago (so a bit more for lists traffic to be fully restored). Mail that was queued on the sender's side when we were down is going to be coming over the next few hours.

Actions to discuss/take

- Fix OIT replication source to sfo-aaa1 and remove /etc/hosts entry,
- Upgrade sanger's OpenDJ to 2.5,
- Set up a secondary LDAP with ou=corp for usage by the mail infrastructure,
- Make ou=corp LDAP failures paging failures (and set up alternative relay, e.g. neon's for mails to SMS gateway?),
- Make the secondary MX not blindly queue mails but do LDAP queries instead,
- Split lists & secondary MX to be separate mail systems so that we can stop incoming MX traffic without affecting lists,

All these action items have been captured in https://rt.wikimedia.org/Ticket/Display.html?id=6795

Category: Incident documentation