Incident Report: 2018-03-14 04:25 GMT

STATUS:

Fixed

Reporter: Reported by Matthew Scholefield
Incident report opened by Kathy Reid
Incident owned by Steve Penrod

Symptom:
All Mycroft Devices, when rebooted, and which have been paired previously, request to pair again and will issue a Pairing code.

At a technical level, the backend microservices which the Device GETs data from was returning HTTP error code 401 unauthorized, instead of the expected 200 OK.

Status:

01:00 GMT reports from Mycroft users that devices on reboot are requesting pairing codes, even when the device has previously been paired
04:00 GMT Mycroft dev team aware of issue
04:17 GMT Restarted microservices, problem persisted
04:29 GMT Performing a curl request to the https://api.mycroft.ai/v1/device endpoint was returning a 503.

```
mycroft@mark_1:/home/pi$ curl -i https://api.mycroft.ai/v1/device/4f39dda3-XXXX-XXXX-XXXX...
HTTP/1.1 503 Service Unavailable
Server: akka-http/10.0.9
Date: Wed, 14 Mar 2018 04:35:20 GMT
Connection: close
Content-Type: text/plain; charset=UTF-8
Content-Length: 105
Access-Control-Allow-Origin: https://home.mycroft.ai
Access-Control-Allow-Headers: Authorization, Content-Type
Access-Control-Allow-Methods: GET, POST, PUT, PATCH, DELETE, OPTIONS
```

04:40 GMT Performing another curl request to the https://api.mycroft.ai/v1/device endpoint, suddenly returning 200

```
mycroft@mark_1:/home/pi$ curl -i https://api.mycroft.ai/v1/device/4f39dda3-XXXX-XXXX-XXXX...
HTTP/1.1 200 OK
Server: akka-http/10.0.9
Date: Wed, 14 Mar 2018 04:40:29 GMT
Connection: close
Content-Type: application/json
Content-Length: 2979
Access-Control-Allow-Origin: https://home.mycroft.ai
Access-Control-Allow-Headers: Authorization, Content-Type
```

Access-Control-Allow-Methods: GET, POST, PUT, PATCH, DELETE, OPTIONS

04:50 GMT Back to getting 503s. Then a 200.

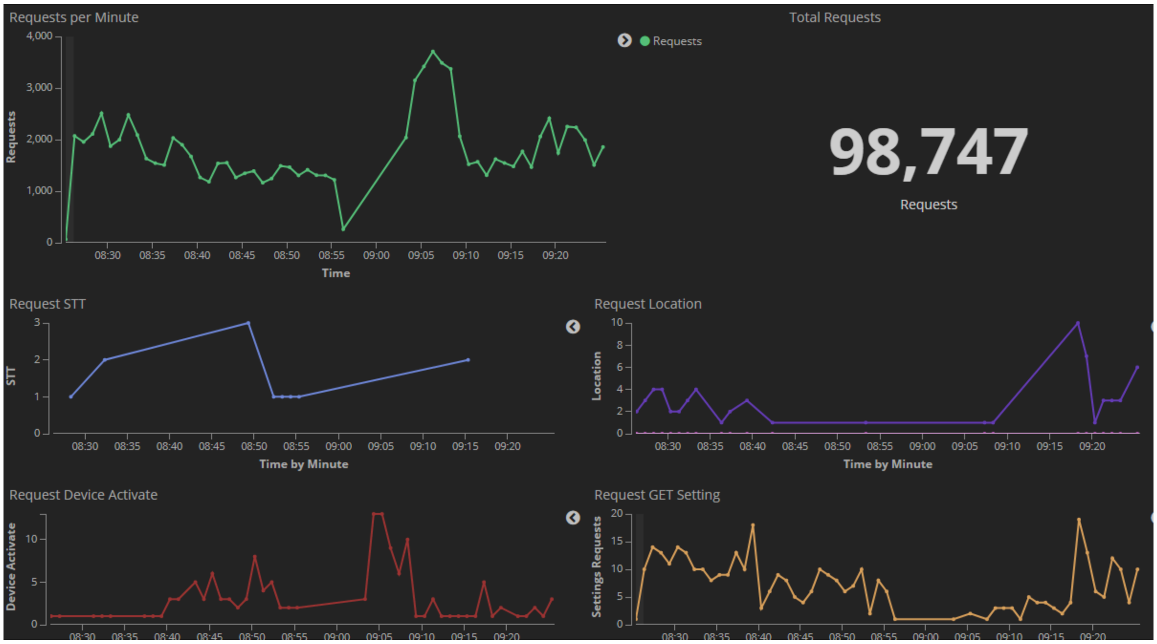
06:08 GMT Noticed that sometimes getting 504s as well as 503s

06:10 GMT Performing another restart of all microservices

07:31 GMT Gateway process consuming 963% of CPU

07:53 GMT Appears that Gateway process is slowly consuming memory. Up to 1.3 GB, growing approx 0.09 GB/min

09:10 GMT Noticed that as Gateway approaches 2.6 GB, it is returning a 401 error. This triggers a request for repairing on devices, which results in a flood of the server. requests jumped from 1400/min to 3400/min. Notice the green peak below



10:33 GMT Identified the root cause. Internal mechanism used to build system metrics was generating a lot of hits in the database used by device micro service. Mechanism Disabled.

14:19 GMT Getting 503s again. Request traffic has spiked since coming back online, approx 3x normal.

16:58 GMT Discovered that the Spotify skill is flooding the server with OAUTH requests -- 100,000+ per hour. Pushed a skill change that disables Spotify. Waiting for it to Propagate

17:23 GMT OAUTH traffic is dropping, down to approx 40k per hour. Expect another 30 minutes for changes to fully propagate.

18:35 GMT Expected trailing is not occurring. Speculating that the skill reload didn’t actually release all of the Spotify skill resource references (Python being a ref-counted language). Going to push a mycroft-core build that disables Spotify. This will cause most mycroft-core instances to restart in the process, cleaning up any dangling references.

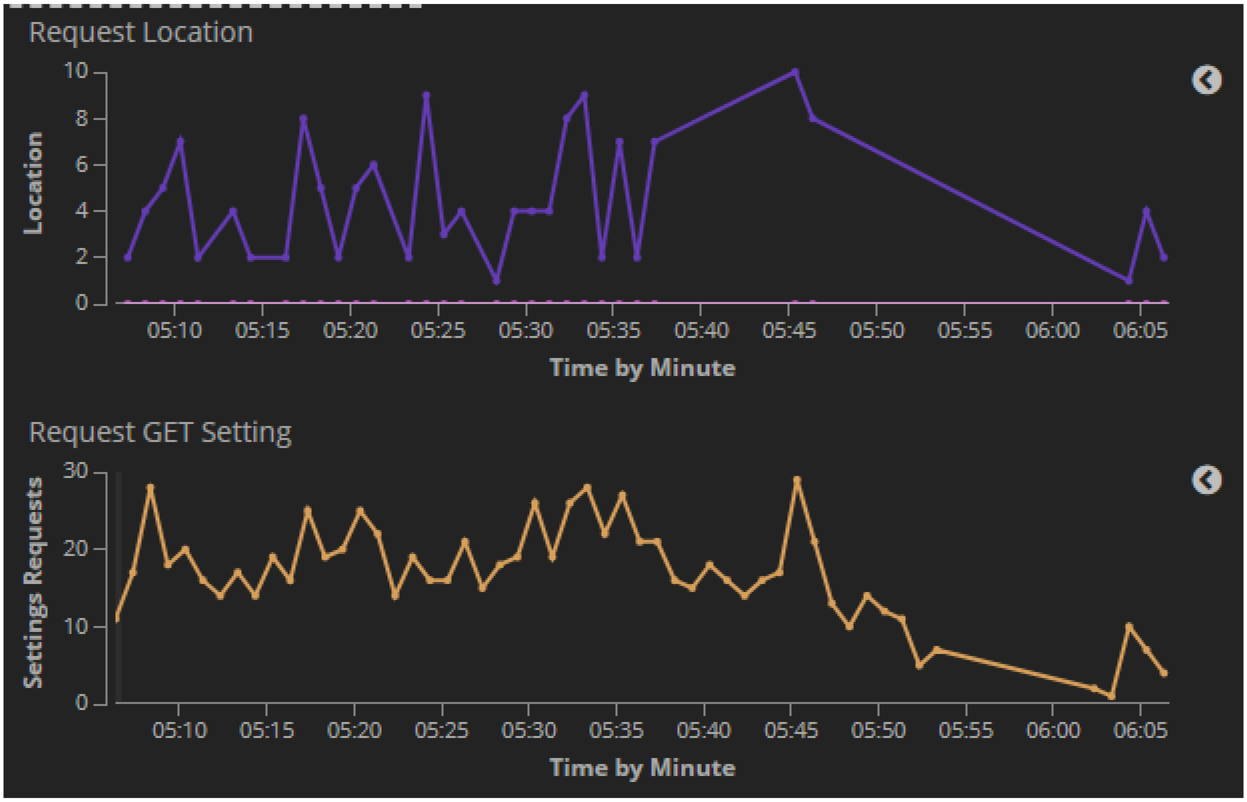
Incident elapsed time 20:25 Current Time: 00:49 GMT

23:43 GMT Systems seem to be back to functioning. However, many devices that had an access token refresh aren’t coming back and require re-pairing. Additionally, some are experiencing a STT lock-out. Clearing that now.

00:49 GMT Cleared the STT locks on 36 users. All should be able to function again. Monitoring.

Notes:

Seeing a dead-time in the Kibana graphs. Notice the lack of data-points on the Location service during 5:45-6:05, even though the Settings requests continued for a while before they went quiet for approx 10 minutes, also.



Cabot is showing “Up” except for the Gateway failing on the “Experimental Pairing” test

System back online:

2018-03-15 00:49 GMT elapsed downtime: 20 hours

Ultimate resolution:

We discovered that a programming error in the Spotify skill was reregistering a callback on each callback. This caused exponential growth of the call for OAUTH in certain situation when starting up. As a result, the backend was hit by nearly 1M requests over the course of the day -- about 100k / hour. This flooded the gateway service and caused timeouts for other types of requests.

Once the culprit was identified, we needed to force the running Mycroft devices in the world to restart the mycroft-skill service to end the cycle (a skill update was inadequate, because Spotify kept the auth object alive until the service was restarted). That was achieved by pushing a new build of mycroft-core out.

During the shutdown, an anti-abuse mechanism for the STT service left the STT request counter at a high value. The restart didn’t clear this, leaving a handful of users without STT (approx 36). We manually cleared this mechanism for today.

Finally, we believe the internal networking timeouts were causing some device pairings to issue a new token/refresh token pair but not actually deliver the pairing to the device. Thus the device was left with an invalid token and an old refresh key, losing their pairing. We are investigating this further.

Postmortem:
<Review incident within 48 hours,. identify what can be learned to detect, prevent, resolve better or quicker>

NOTE: Kathy has cancelled the standing Incident review at 0900hrs Friday 16th because it is a poor time to do a review, she will reschedule at a time that is better for everyone.

- Track down the 401 response instead of 503 when in this state (REDIS connection failure?)
  There is no reason a 401 should have been generated for devices that were paired. Task created to look for the code that incorrectly interprets the failure and responds with 401.
- Missing Jenkins jobs:
  REDIS - need a restart on Jenkins - task
  Kafka - need a restart on Jenkins - task
  Oauth - need a restart on Jenkins - done
  Metrics - need a restart on Jenkins - done
  Zookeeper - need a restart on Jenkins - task
  Neo4j - need a restart on Jenkins - not needed
- The home-web Jenkins job threw an error during the rebuild, investigate
  TODO...
- Source of the "too many requests" issue (not resetting the throttle count)
  Upon restart, the STT service should just clear all these counters. They are all invalid after the startup process completes, anyway.
- Need to evaluate the auth token system to deal with failed refresh token deliveries

  ```
  Today:  auth / refresh
              when auth expires:
                      * Generate new auth / refresh in database
                      * Send new auth / refresh
  Issue is when they weren't received, but DB already changed

  Proposal:auth / refresh
              when auth expires:
                      * Generate new auth2 / refresh2 in database
                      * Send new auth2 / refresh2
              when auth2 is received from client:
                      move auth2 / refresh2 into auth/refresh and clear auth2/refresh2
  ```

  Task created

- Consider https://github.com/Netflix/Hystrix and https://github.com/Netflix-Skunkworks/hystrix-dashboard circuit breakers
  Investigation task created

- The source of the callback bomb was rescheduling a recurring event within the the callback. Adding code to clear duplicate named events.
  PR pending
- In the Kibana graphs, we could see the spike, but no warning or critical thresholds for what constitutes “normal” performance were defined. Do we need to define thresholds for metrics like OAUTH requests and have Cabot issue alerts when thresholds are breached? Our principle should be “we know about an outage before a user reports it”.

Kibana dashboard only: https://www.elastic.co/blog/kibana-dashboard-only-mode Elastic 6.0 only

Using this form:

- Do not mark as “Back online” until completely certain, e.g. one hour has passed without a repeat of the issue.
- Perform postmortem
  - Kathy schedules and runs
  - The backend team (and any others, as appropriate) should review together
  - Fill out the Postmortem, assign Flow tasks
  - Archive this Incident Report by moving it into the **Incidents** subdirectory: File > Move To > Incidents

To change the status, place your cursor in a cell below then hit Shift+Home, Shift+Left Arrow, Ctrl+C (Copy). Then you can move to the top of the sheet, place your cursor in the Status cell and simple Ctrl+V (Paste)

Under investigation
Potentially fixed
Fixed