Search Wikitech

Toolforge webservices are in the final stages of migrating to the toolforge.org domain .
Please help us clean up older documentation referring to tools.wmflabs.org!

Incident documentation/20160623-etherpad

< Incident documentation

Etherpad had a catastrophic and unrecoverable downtime. Recovering from it in a timely manner meant restoring a day old snapshot of the database. Exact cause: Still Unknown'

Contents [hide]

Summary

On 14:27 UTC it became apparent that etherpad was offline. Efforts to restore the service through restarts, software debugging were inconclusive. End result was we took the decision to restore the database from a 1 day old snapshot to restore the service, which means some pad connect may have been irrevocably lost. We are working on that yet

Timeline

- 14:27: It becomes apparent from wikimedia-operations that etherpad.wikimedia.org seems down
- 14:29: Jaime restarts etherpad. Service is not restored
- 14:30: Jaime is debugging the service.
- 14:30: First icinga check noting etherpad is down
- 14:32: <jynus> restarting etherpad-lite.service
- 14:33: Alex joins the debugging.
- 14:34: Icinga reports etherpad down.
- 14:38: <akosiaris> stopping etherpad-lite on etherpad1001, disabling puppet
- 14:38: Jaime after consulting with Alex starts restoring the database so that we can have it handy in case debugging goes south.
- 14:45: <akosiaris> debugging etherpad. Started the service with a blank db, looks like it's working
- 14:47: <akosiaris> change the default message in etherpad to indicate problems
- 15:17: Discovering and commenting on https://github.com/ether/etherpad-lite/issues/2946 . Seems the closest we got to a problem, no response or more help. Stack trace is exactly the same, suspicion the database is corrupted is reinforced.
- 15:20: Restoring the corrupted db and trying to debug the problem. Alex is restricting access to etherpad via ferm to him only (via ssh tunnel). Visits various pads, no reproduction. Starts visiting pads in an effort to find a corrupted pad, delete it and fix the problem. Logs really close to crashes the are the source of pad names. Efforts are in vain.
- 15:26: <akosiaris> stop etherpad-lite, etherpad is down
- 15:27-16:10: investigation continues with no result
- 16:10: Jaime manages to have a restore from backups of the database
- 16:11: ferm restriction is lifted. Service is restored.
- Deliberations between Jaime, Yuvi, Alex as to what to do about the 1 day of lost pads. Proposal to have the service running with the 1 day old data due to wikimania wins, will make efforts to replay the logs and get a clone of the service with the database frozen in time to allow users to access lost pads.
- 21:40 etherpad-restore.wikimedia.org created to allow users to access restored versions of their pads and in a self-serve fashion migrate copies of them in the production instance. Unfortunately it's impossible to do that automatically

Conclusions

- Somehow the etherpad database has been corrupted in a way that caused etherpad's ueberDB component to emit a stack trace and gracefully terminate, leading to etherpad's termination as well. Due to that, the service was down from 14:27 to 16:11 despite efforts to find and fix the problem. Pad content that has not made it to the production instance, may be found on http://etherpad-restore.wikimedia.org &
- It is not clear why etherpad database got corrupted, but it is likely that it was due to some of these reasons (or all of them combined):
  - There was underlying file corruption, made apparent only after one of the 2 following actions happened. This is not backed up by evidence, but it is a possibility given the 31-GB single-table monolitic, highly dynamic structure of etherpad and the old servers it is served from.
  - A failover to this db server (db1016) was done the day before- either that was improperly done or it created certain activity that lead to corruption. It also complicated the recovery as m1 was at the time in reduced redundancy and the previous master being reimaged.
  - Slave (db2010) was under maintenance while the issue happened, and had slave drift on the etherpad database. Correcting/checking this could have caused row corruption in its master
- Backups took too much time to recover because they are too inflexible: we need per table CSVs/TSVs and physical backups. The reason why this is not done now is the lack of proper space

Actionables

- Still investigating what happened and trying to contact upstream at https://github.com/ether/etherpad-lite/issues/2946 &.
- https://phabricator.wikimedia.org/T138516 & - prodiving a host for users to access non restored in etherpad.wikimedia.org pads ✔ **Done** https://etherpad-restore.wikimedia.org &
- Correct m1 backups, that were being done supposing m1-master was still a slave Task T138559 ✔ **Done**
- Improve backup strategy- while the current one allows in general not to lose data, the TTR can be too long Task T138562 🔄 **In progress**

Category:	Incident documentation