



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/20190923-s3 primary db master crash

[< Incident documentation](#)

document status: final

Contents [\[hide\]](#)

- 1 [Summary](#)
 - 1.1 [Impact](#)
 - 1.2 [Detection](#)
- 2 [Timeline](#)
- 3 [Conclusions](#)
 - 3.1 [What went well?](#)
 - 3.2 [What went poorly?](#)
 - 3.3 [Where did we get lucky?](#)
 - 3.4 [How many people were involved in the remediation?](#)
- 4 [Links to relevant documentation](#)
- 5 [Actionables](#)

Summary

s3 primary database master had a [RAID backup battery](#) failure which cause the host to completely crash. It had to be power cycle from the idrac.

Impact

All the s3 wikis (<https://raw.githubusercontent.com/wikimedia/operations-mediawiki-config/master/dblists/s3.dblist>) went read-only as the master wasn't available for writes. Reads were not affected, all the replicas were available.

Detection

- The problem was clear when we saw that db1075 reported HOST DOWN - however, that only sends an IRC alert, not a page. Masters should probably page for HOST DOWN.
- Alerts were sent to IRC and pages.
- Users reporting issues on #wikimedia-operations

Timeline

All times in UTC.

- 18:40 **OUTAGE BEGINS**

```
Sep 22 18:40:38 db1115 mysqld[1630]: OpenTable: (2003) Can't connect to MySQL server on 'db1075.eqiad.wmnet' (110 "Connection timed out")
```

- IRC logs from #wikimedia-operations:

```
18:42:45 <icinga-wm> PROBLEM - Host db1075 is DOWN: PING CRITICAL - Packet loss = 100%
```

```
18:47:03 <AntiComposite> I'm getting a warning on otrs-wiki about a high replay database lock
```

[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](#)

18:47:40 <AntiComposite> It's also slower than usual

18:48:52 <+icinga-wm> PROBLEM - MariaDB Slave IO: s3 on db2105 is CRITICAL: CRITICAL slave_io_state Slave_IO_Running: No, Errno: 2003, Errmsg: error reconnecting to master repl@db1075.eqiad.wmnet:3306 - retry-time: 60 maximum-retries: 86400 message: Cant connect to MySQL server on db1075.eqiad.wmnet (110 Connection timed out)
https://wikitech.wikimedia.org/wiki/MariaDB/troubleshooting%23Depooling_a_slave

18:49:22 <+icinga-wm> PROBLEM - MariaDB Slave IO: s3 on dbstore1004 is CRITICAL: CRITICAL slave_io_state Slave_IO_Running: No, Errno: 2003, Errmsg: error reconnecting to master repl@db1075.eqiad.wmnet:3306 - retry-time: 60 maximum-retries: 86400 message: Cant connect to MySQL server on db1075.eqiad.wmnet (110 Connection timed out)
https://wikitech.wikimedia.org/wiki/MariaDB/troubleshooting%23Depooling_a_slave

18:49:58 <+icinga-wm> PROBLEM - MariaDB Slave IO: s3 on db1095 is CRITICAL: CRITICAL slave_io_state Slave_IO_Running: No, Errno: 2003, Errmsg: error reconnecting to master repl@db1075.eqiad.wmnet:3306 - retry-time: 60 maximum-retries: 86400 message: Cant connect to MySQL server on db1075.eqiad.wmnet (110 Connection timed out)
https://wikitech.wikimedia.org/wiki/MariaDB/troubleshooting%23Depooling_a_slave

<More alerts arriving - not pasting them all here>

18:55:51 <marostegui> I'm connecting

- 18:56 First SMS page sent
- 19:02 Host rebooted from the idrac after seeing it is a BBU issue and the host is not responsive
- 19:03 MySQL started and starts InnoDB recovery
- 19:04 Manual puppet run ran (but failed and went unnoticed)
- 19:06 MySQL finishes recovery
- 19:07 Manually removed read_only=ON from db1075
- 19:08 Lag still being reported
- 19:16 Manual puppet ran (success and pt-heartbeat starts)
- 19:16 **OUTAGE ENDS**

Conclusions

The master lost its BBU and that resulted on a completely host crash, which is something that has been seen before with HP hosts <https://phabricator.wikimedia.org/T231638> <https://phabricator.wikimedia.org/T225391>

The master being unavailable means that writes cannot happen:

<https://grafana.wikimedia.org/d/000000278/mysql-aggregated?orgId=1&var-dc=eqiad%20prometheus%2Fops&var-group=core&var-shard=s3&var-role=master&from=1569174586143&to=1569181181947>

This is part of a batch of 6 servers, and 3 of them have already had BBU issues: <https://phabricator.wikimedia.org/T233569> so we'd need to evaluate if what to do with then next. Definitely replacing the current master and promoting another one which is not part of that batch is what is happening next: <https://phabricator.wikimedia.org/T230783>

What went well?

- Alerts worked fine



s3 master queries, that reflect the impact of the crash

- Rebooting the host from the idrac was successful
- MySQL came back clean

What went poorly?

- Race condition between pt-heartbeat being ran via puppet but mysql wasn't still fully up failed (and went unnoticed) resulted on lag being reported while everything was up, resulting in 8 minutes more of an outage until the second manual puppet run was done.
- A BBU failure shouldnd't result on a completely host crash (but we haven see that before with HP hosts)

Where did we get lucky?

- The master was able to come back after the hardware issue. We had to restart it via idrac but it came back clean, otherwise, we'd have needed to do a fully master failover manually to promote a new replica to master.
- Volunteers and staff noticed the failure even before the alerts caught them

How many people were involved in the remediation?

- 1 DBA, 2 SREs, 1 WMCS, 1 Dev, 1 Volunteer

Links to relevant documentation

- [MariaDB/troubleshooting#Emergency failover](#) (scenario 1)

Actionables

- Implement (or refactor) a script to move replicas when the master is not available (this wasn't needed yesterday, but could be needed in future issues): <https://phabricator.wikimedia.org/T196366>
- Fix mediawiki heartbeat model, change pt-heartbeat model to not use super-user, avoid SPOF and switch automatically to the real master without puppet dependency: <https://phabricator.wikimedia.org/T172497>
- Decide what to do with the same batch of hosts that have already had BBU issues: <https://phabricator.wikimedia.org/T233569>
- Buy a new BBU for db1075 <https://phabricator.wikimedia.org/T233567>
- Remove db1075 from being a master <https://phabricator.wikimedia.org/T230783>
- Address mediawiki spam during readonly/master unavailable <https://phabricator.wikimedia.org/T233623>
- Make sure primary database masters page on HOST DOWN <https://phabricator.wikimedia.org/T233684>
- Better tracking of hardware errors in Netbox <https://phabricator.wikimedia.org/T233774>

Category: [Incident documentation](#)

This page was last edited on 7 April 2020, at 06:17.

Text is available under the [Creative Commons Attribution-ShareAlike License](#); additional terms may apply. See [Terms of Use](#) for details.

[Privacy policy](#) [About](#)

[Disclaimers](#) [Code of Conduct](#) [Developers](#) [Statistics](#) [Cookie statement](#) [Mobile view](#)

Wikitech

