Post mortem for 1,000 validator downtime on Ethereum mainnet on 2022-12-21, node operator CryptoManufaktur

Timeline

00:50 UTC validators start missing attestations

01:09 UTC first Discord warning

01:25 UTC reachout by Lido team on Telegram

01:29 UTC investigation starts

01:49 UTC issue resolved

Root cause

This outage was ultimately caused by insufficient change control procedures.

We run three Ethereum nodes per Lido environment of 1,000 keys, with 2 Vouch (one hot, one cold standby) and 5 Dirk in a 3/5 threshold signing setup.

All three Ethereum nodes were unavailable for submitter duties.

Node A was in planned maintenance for client diversification, moving from Geth to Erigon.

Node B had been removed from submitter duties to protect against a theoretical dDoS vector in Ethereum.

Node C received maintenance to improve its metrics reporting. Documented maintenance procedures were not followed and it failed.

Resolution

After identifying root cause, we brought Node C back up and brought Node B back into submitter rotation.

What worked well

- Alerting worked as designed.
- We responded quickly.
- The separation of Lido environments with a maximum of 1,000 keys per environment worked.

What didn’t work well

- Protecting against a theoretical dDoS attack vector, we DoS’d ourselves by reducing resilience.
- Change control failed to flag that one node was already under maintenance.
- Maintenance procedures were not followed for the remaining node.

Changes made

- All three Ethereum nodes are now in rotation for submitter duties. If the theoretical dDoS vector ever becomes practical, we can deal with that issue then by using a fourth node for “query-only” duties.
- Strengthened procedures around change management, so that environments that are already under maintenance do not receive a second maintenance event.
- Additional training around safe maintenance procedures on an Ethereum node.

Additional planned improvements

- While our time to resolution was good, there is room here for improvement to have a faster initial response, by maybe another 10 to 15 minutes. We will review our escalation procedures.