



P1-critical incident post-mortem

2025-06-10 08:14 CET - Scaleway - Network outage

Customer impact: High, the Keycloak nodes were unreachable

Severity: High, the authentication process is impacted

Time to detection: 5 minutes

Time to identification: 6 minutes

Time to resolution: about 1 hour

Incident management leader: Mathieu

Incident detection source: Cloud-IAM monitoring

Timeline

Disclaimer: this is a general post-mortem regarding the network outage of Scaleway. Please note that your deployment might have been impacted variously by the incident.

The timeline below is in CET timezone

08:10: the automated monitoring probe starts detecting slowness on a deployment
08:14: the automated monitoring probe starts receiving 502 errors on additional deployments
08:15: the on-call team is triggered and starts analysing the issue
08:20: internal diagnostic tools report a network issue with DNS and public gateway communication
08:20: the connection via ssh on the remote servers is impossible
08:22: the team tries to reboot the instances via the Scaleway console
08:27: the monitoring probes of additional customer are failing, they all run on Scaleway fr-par
08:27: the team starts a thread of what looks like a general outage on the Scaleway community slack and opens a case to the support.
08:35: the restart does not seem to recover the network
08:42: the team suspects an issue with the hypervisor. Then the team decides to shutdown and start again the VMs for another customer. This procedure forces the VM to change its hypervisor
08:47: the workaround seems to work, the team applies the hard stop / start to customer VMs, prioritising the production environment before the staging.**08:59:** other customers are complaining on the Scaleway Community slack
09:00: the team is restarting about 60 impacted instances
09:09: Scaleway opens a status on the incident:
<https://status.scaleway.com/incidents/h13h8zphrmxw> (wrongly referring to their K8S infrastructure)
09:27: Scaleway apply a fix on their infrastructure
09:35: all the customer deployments are back to normal from Cloud-IAM point of view
09:37: the team fixes an outage on Cloud-IAM log infrastructure that was also related to the Scaleway issue. The log ingestion and analysis was delayed by a few minutes
09:40: the internal alert manager detects several deployments running with a split brain configuration
09:40: the team applies the procedure to solve such issue
09:47: end of all the incident for the Cloud-IAM customer
10:19: Scaleway opens a new incident regarding what is probably the root cause of the incident: <https://status.scaleway.com/incidents/n0fbmptdn5cc>

Analysis

Abstract

At 08:10 CET, the Cloud-IAM on-call team began investigating a network issue affecting a customer deployment. Shortly afterward, alerts from additional customer environments, also hosted on Scaleway fr-par, started appearing. Recognizing a potential broader infrastructure problem, the team initiated communication with Scaleway via their community Slack and opened a support ticket.

The team attempted several recovery actions:

- Restarting the server's local network manager
- Rebooting the affected servers
- Performing a stop/start operation on the servers to force a hypervisor change

The final workaround proved effective and was systematically applied across all impacted servers, with production environments prioritized over staging.

After the workaround was applied on Cloud-IAM, Scaleway opened an issue on their status page and applied a fix on their infrastructure. This fixed an issue with the Cloud-IAM log ingestion.

Following recovery, the team identified a “split-brain” condition in other deployments. The remediation procedure was applied successfully, resolving the issue promptly.

Actions

- Technical correspondence with Scaleway's support and infrastructure teams to understand the root cause of the incident ⌚
- Find out how we could prevent such outage from the client perspective ⌚