CERNTS post-mortem

OTG0153169
on Saturday, Nov 16th morning

Full report

Mario Rey Regúlez (IT-CD)

The origin of the failure

- ESET Server Security Detection Engine issue
  - It bricked Windows Server 2016 nodes during controlled test-patching (Wed – Fri). No user impact
  - Affected nodes were rebuilt with incomplete configuration
  - In parallel ESET was removed from ~160 Windows servers CERN-wide as a preventive measure (OTG0153165)
- To ensure antivirus protection on Windows Terminal Infrastructure, nodes required a reboot
  - Scheduled in waves through the night (OTG0153166) to minimise impact
  - Half of CERNTS capacity was scheduled to reboot on Saturday 3am
- Rebuilt nodes were not fully configured because of a human error
  - The scheduled reboot closed idle sessions
  - When users connected back in the morning, they were sent to unconfigured nodes, who rejected them

The impact

- Since Friday afternoon new connections to CERNTS could be failing
  - No general problems that we know of before Saturday 10h30
  - Remaining dedicated 40 clusters were not affected
- Based on historic data ~40 users were affected (out of 2500 CERNTS users)
  - Possible performance degradation and connection issues
  - Severe impact on experiments (ALICE)

The timeline (summary)

- [Friday afternoon] The 8 lost nodes (out of 30) during test-patching were recreated and added back to production
- [Saturday 03:15] Half of CERNTS nodes were rebooted as scheduled
- [10:43] ALICE reports the issue on Mattermost
- [10:50] Service manager sees Mattermost. Work starts. Opens OTG, replies to Mattermost
- [11:10] cernts-homeless becomes available as a workaround
- [11:30] Unconfigured nodes are removed from production and stuck ones rebooted. CERNTS becomes operational at lower capacity
- [11:50] All nodes are correctly configured, validated and put back in production. CERNTS recovers normal capacity

The analysis

What went well

- Pre-prepared BC cluster allowed users to regain access sooner
- Knowledge of the deployment and well-documented procedures made recovery quick
- Admin picked up Mattermost messages quickly

What went wrong

- Timing: ESET issue required an urgent operation during a critical patching session
- Taking preventative measures to avoid larger issue with ESET made us overlook incomplete configuration
- Blind spot in our monitoring: it was checking if there was an RDP certificate, not if it was the correct one

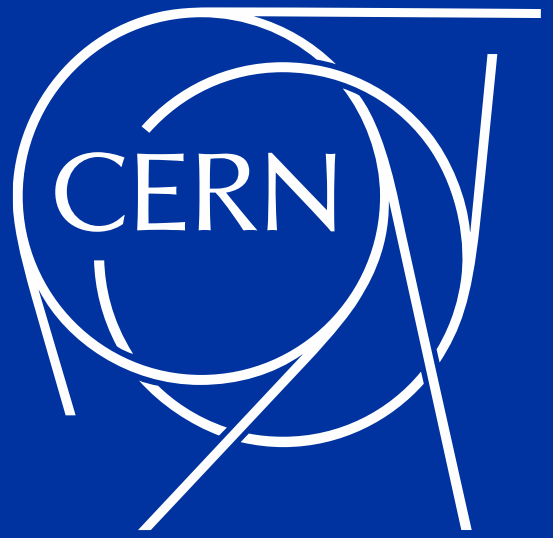
The analysis

Where we got lucky

- ESET problem only affected devices after installing monthly patches and reboot
- It only affected Windows Server 2016 with Remote Desktop role, Internet connectivity and OpenStack virtualization

Follow up Action Items

- Fix monitoring blind spot
- Improve trigger for load balancer expulsion
- Follow up tickets with Microsoft and ESET to understand the root cause
- Update procedure to highlight testing newly added nodes
- Improve BC cluster visibility and isolation


CERN