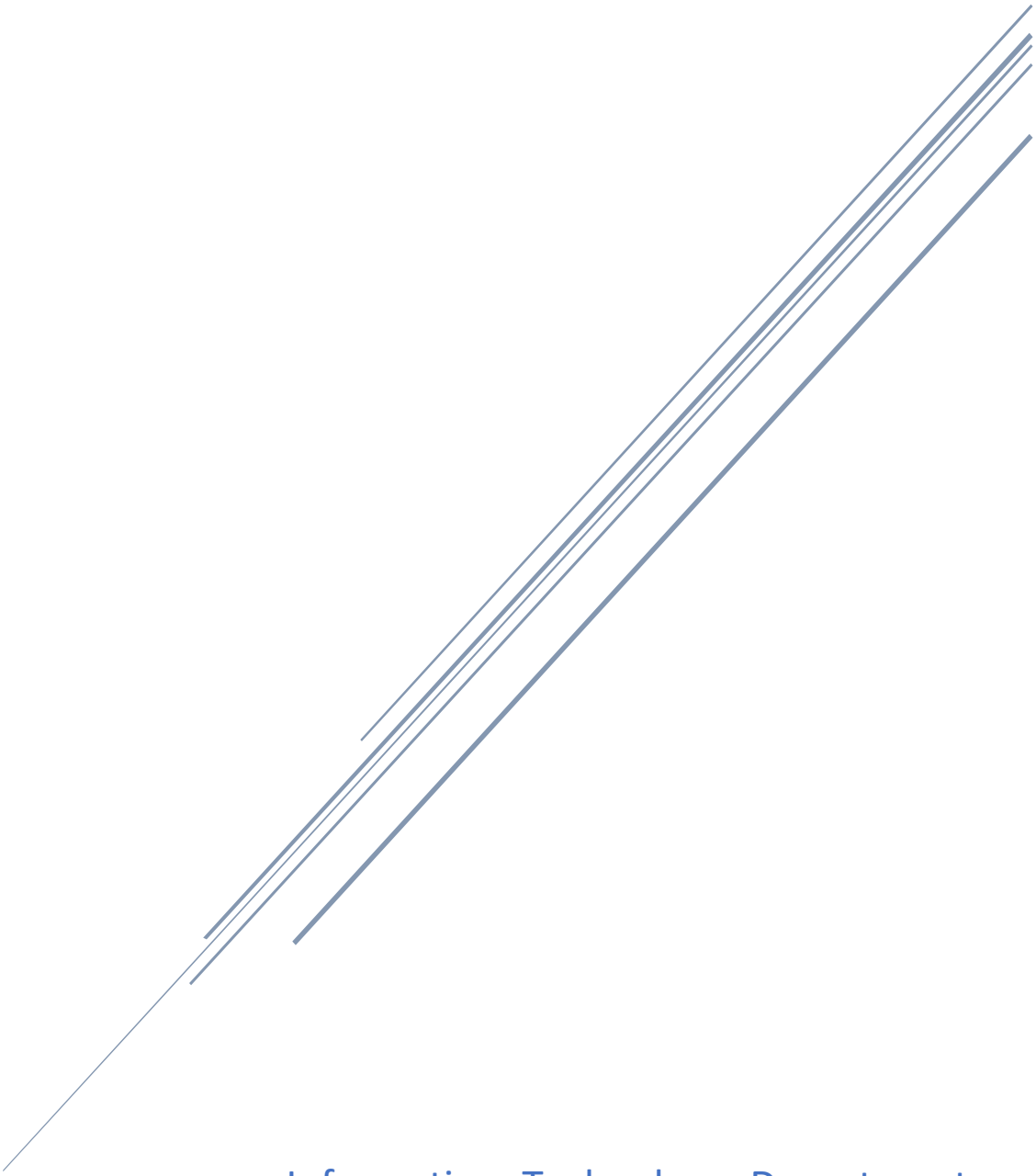
Incident report for DMZ migration and Panel system failure



Information Technology Department

Hyper Group Network Ltd.

Related Parties:

1. Oscar Fung - Database Administrator
2. Jia Jun - Security Architect
3. Anson Tsang - Network Architect
4. Ivan Cheung – Cloud and Infrastructure Administrator
5. Alex Liang – Chairman of Emergency Response Group
6. Dicky Fung – Chief Finance Officer
7. Mary Hun – RCN Infrastructure Engineer
8. Tsang Kwok Wo - Services Engineer
9. Nick Hung – E&O Manager
10. Kevin Li – DC Support Engineer

Related Information:

Incident Location:

- DMZ: Edge Firewall (dmz vdom) & DMZ firewall
- Panel & DB: VMS010 (Production Cluster)

Date: 13 Aug 2024 - 22 Aug 2024

Summarized Symptoms

DMZ not connected, panel not operational, panel database not synced, VMS10 CPU issue

Incident Description:

- 13 Aug 2024

A planned job is issued by the IT department related to the routing changes in DMZ (including DMZ-1, DMZ-2, and CFT networks).

After the change was completed, the CS team reported that some users were not able to access the website including the panel, HNICS, and other services related to web services. RCN Executive Committee member Mary Hui also reported that RCN users cannot access the client's website. Other staff reported all internal systems are unreachable.

At that moment, the network team started to troubleshoot the problem. We have no way to fallback the configuration since this change affects the physical topology of wiring and the logical topology within the internal edge network. The network team verified all BGP sessions are established but the package cannot pass through the firewall(s). The network team raised an emergency support request to the security team to help with the firewall troubleshooting.

At around the end of the day, we conclude this incident is caused by some bugs in the firewall operating system and it auto resumes after rebooting the firewall.

The network team has validated that the network status is good and the service team validated the web is accessible in the production environment. However, they stated the panel system is lagging and sometimes 500 error prompts. We announced services are partially resumed.

- 14 Aug 2024

The Service team observed panel status flapping and the web application is not stable. CS team also reported users are experiencing difficulty to access the service. The Infra team checked the cluster status as healthy and VM were operating as usual. The network team discovered asymmetric routing occurs between the external firewall and the DMZ firewall. The network team has disabled one of the connections in DMZ and the services status resumed in stable condition.

- 15 Aug 2024

The user has reported some unstable conditions for accessing the HNFS/RCN panel. The service team checked if the system was healthy and found that a 500 Error was prompted at the panel interface. The service team reported this issue to the Network team and validated if any issue with network connectivity. The network team report network is working and proposed this issue is caused by the database. The database team checked the database and found that the database is not syntonized and access is extremely slow. After a health check by the database team, the system is unhealthy and “InnoDB” is corrupted due to a segmentation fault (core dumped). Database rebuild the database and flap the database to MASTER-02

- 16 Aug 2024

Our Chief Information and Operations Officer (COIO) issued the 1st class state of emergency and approval by the Chairman of the Emergency Response Group (ERG) as the database is out-of-sync and total service interruption may happen. The database team try to isolate the problematic database node and see if it works.

User starts to report their server are not showing on the panel and the service team starts to validate and confirm the data in HNFS/RCN is not the latest. The database team validated and found that the HA link is not working and the data has been fully out-of-sync since May 2024. After further investigation, it is confirmed the issue is caused by partial hardware failure.

A working group for this issue is formed and coordinated with different parties to work on this issue. The request for a new server is raised and approved by the Chief Finance Officer.

Created a temporary database server and loaded all data within it. Limited the number of access to prevent server overload and kernel panic.

- 17 Aug 2024 & 18 Aug 2024

No technical update. Database team, service team and cloud & infra team perform standard recovery in case the issue happens again.

- 19 Aug 2024

The replacement server has arrived data center and replacement has been arranged by the DC team. Hardware replacement is carried out at night. Database, network, cloud & infra teams are working with new hardware and performing emergency services recovery.

The operating system installation contains some unexpected errors during installation. After investigation, some incompatibility between the storage media and the server is discovered. This is resolved by the Cloud and Infra Team at 23:00.

Reinstall at the OS level was completed at 23:30 and the Network team resumed the network connection at around 23:59.

Application recovery is scheduled for 20 August 2024.

- 20 Aug 2024

The database team has rebuilt the MASTER-01 and MASTER-02 databases and resumed the internal, RCN and HNFS panel system at around 10:00. The HNPS panel are under recovery due to data corruption.

In the afternoon, the paid panel was also fully recovered in a controlled manner after reloading the data infrastructure and rebuilding the SLAVE-1.

Service is resumed and under monitoring by the service team.

- 21 Aug 2024

Services under monitoring. The dev team is performing a fully health check on HNICS and other co-related services.

- 22 Aug 2024

The service team validated the application level of the panel service is resumed. UAT for RCN & HNFS panel is passed.

The Chief Information and Operation Officer requested a close in emergency and was approved by the Chairman of ERG.

- 23 Aug 2024 – 26 Aug 2024

The problem changed from incident to follow-up.

- 27 Aug 2024

Services reported to Dev team and notified user cannot successfully create or modify server via HNICS panel.

The dev team validated that the issue was caused by database schema and worked with the database team.

The database team has confirmed this issue is caused by the auto-increment feature is not correctly configured in the HNFS panel database. After double validation, this is a fault due to the database services brushing too many times during our services recovery period. The database team patched the database and resumed at around 17:00.

UAT is performed by the service team and internal users. The result is positive.

- 28 Aug 2024

Incident case is closed.

Incident Caused:

- Service cutoff is not well planned and fallback cannot be performed since physical wiring is different between 2 major changes.
- A hardware failure on VMS-010 and no immediate replacement can be made since no spare hardware.
- Database not sync was not discovered immediately and caused partial database corruption on MASTER-01/02

Follow up recommendation:

- For all critical services cut off, a fully operational fallback plan should be provided and the DR procedure when infrastructure changes
- The spare server should deploy in datacenter for emergency failover if some hardware failure happens
- Database should be backed up more frequently. Recommended hourly backup towards Production databases.

Report by: Ivan Cheung, Anson Tsang, Oscar Fung
Verified by: Alex Liang
Additional verification: Dicky Fung, Mary Hui, Tsang Kwok Wo, Nick Huang