NETWORK INCIDENT BIT-2C REPORT – 17 02 2021

Summary

A misbehaving customer switch connected to BIT's network caused resource starvation of the leaf-sw1.bit-2c and leaf-sw2.bit-2c switches. As a result uplinks on these switches were not usable. Specific circumstances surrounding this failure caused the shared storage services (with the exception of iSCSI), virtualization services and shared hosting services to be unavailable.

Details

13.46 A customer switch, connected via a port-channel to BIT's network, creates high numbers of MAC address entries on a pair of MC-LAG leaf switches in BIT-2C. The forwarding plane of the BIT switches no longer functions, so no more traffic is forwarded through the leaf switches. Of the four leaf switches in BIT-2C, two have become unusable.

13.47 The network problem causes unavailability of the Ceph cluster from BIT. Services that use this cluster, namely the virtualization services, all shared services of BIT and shared storage (with the exception of iSCSI), become unavailable. The malfunction is picked up by BIT engineers and they start their investigation into the cause.

13.51 The clients using CephFS have been dropped and blocked by the storage nodes because the clients are no longer communicating with the CephFS MetaDataServer (MDS).

~13.52 It became clear the malfunction is happening on one pair of leaf switches in BIT-2C, the investigation is focusing on the cause.

13.56 Because www.bit.nl is also unavailable, the incident is reported on www.bit.org.

14.01 The cause of the malfunctioning forwarding plane on the switches is found in overloaded hardware resources for the forwarding plane. The mechanism used to protect against loops at customers tries to allocate more hardware resources than are available. One port of the port channel for that customer's switch is shut.

14.05 Shutting the port did not solve the problem. Therefore, the first switch of the leaf pair is rebooted.

14.06 Various customers refer in their communication to the incident report on www.bit.org, due to the peak in the number of visitors this site becomes inaccessible.

14.07 The second switch (leaf-sw2.bit-2c) of the leaf pair is rebooted. Immediately after the reboot command, the Ceph cluster becomes available again. Servers that use NFS and servers without mounts become available again. These include the MX, IMAP, SMTP servers and the web servers for, among others, www.bit.nl, portal.bit.nl and webmail.bit.nl. Customer virtual machines without CephFS mounts are also available again.

14.12 The first switch has started successfully after the reboot. Uplinks on this switch are on-line again.

14.12 After a reboot of the www.bit.org server, that site becomes available again.

14.14 The second switch has successfully started after the reboot. The uplinks on this switch are also back on-line.

~14.15 Due to the blocking of the CephFS clients, clients with a CephFS mount require remounts/reboots. All servers with these mounts are checked and remounted or rebooted where necessary.

14.29 The incident is closed on www.bit.org.

14.30 The shared hosting services are back on-line.

14.43 The incident is reported on www.bit.nl with a reference to the incident report on www.bit.org.

~15.00 The overloaded hardware resources were only possible in combination with a specific configuration on uplinks for customers. All switches are checked for the presence of such a configuration. It turns out this configuration is used only for the aforementioned customer. The configuration for this customer is changed everywhere, where necessary.

16.15 All virtual servers with CephFS mounts have been checked and mounts are made available again where necessary.

Conclusion

Network

A configuration on a pair of MC-LAG leaf switches causes those switches to be vulnerable to hardware resource starvation. Those resources are used by the forwarding plane for, among others, port security and storm control. The forwarding plane of the switch stops functioning at said starvation, while the control plane of the switch continues to work. As a result, specific MAC and IP addresses seem to still be reachable via this switch, while the traffic is subsequently not delivered to the correct switch port. The packets and frames not being forwarded while continuing to announce that everything connected to the switch is reachable is extremely undesirable behavior.

By default, a counter is kept for each port on the switches in the resources for the number of MAC addresses that are on that port. As soon as that counter hits a limit, the port in question is shut. For one customer that configuration has been changed so that no counter is kept, but a limited access list of MAC addresses was put in place instead. As soon as the limit is reached, the port is not closed, but no new MAC addresses are learned anymore. This deviating configuration has been set because in the past this specific customer regularly triggered port security due to many MAC address changes, resulting in port shuts for this customer. BIT was not aware that this configuration introduced a risk of resource starvation. The hardware resources are monitored so that timely action can be taken in case of high load on the resources. Due to the large numbers of MAC addresses that were rapidly learned on the ports of this specific customer, the monitoring systems were unable to timely warn for the (inordinately) high load.

The resource starvation caused the switches, leaf-sw1.bit-2c and leaf-sw2.bit-2c, to stop forwarding traffic. Customers with uplinks on those switches could no longer exchange traffic with the switches until the problematic configuration was removed and the switches rebooted.

Storage

BIT's shared storage platform is based on Ceph and is fully redundant with nodes in the data centers BIT-1, BIT-2A and BIT-2C. In the event of a data center failure, all data remains available on the Ceph platform. However, the specific circumstances of the network incident made the Ceph cluster inaccessible.

The BIT switches use EVPN, an extension to the BGP protocol. Traffic is routed across the network instead of being switched. Using Equal Cost Multi Path (ECMP) routing, all storage nodes in the platform are used, if available. A full explanation of these techniques can be found on our site[1].

Because the control planes of the problematic leaf (access) switches continued to work, they also continued to advertise routes for the storage nodes behind these switches. The storage nodes are linked to each other via a separate (storage) network, which was online as normal. Because routes to the storage nodes in BIT-2C remained advertised and the storage network, also in BIT-2C, itself was available, the nodes appeared to be available in BIT-2C. The clients using the storage platform therefore received routes for the access network to the storage nodes in BIT-2C. As a result of ECMP, approximately one third of the storage traffic was sent to BIT-2C, also by clients in other data centers. Because all traffic was not handled correctly, all shared storage for the clients was effectively useless.

The shared storage platform of BIT is used for shared storage services, all virtual machines on the virtualization platform and all shared services of BIT, including email and web. As soon as the problematic switches were functioning properly again, all services that do not use CephFS mounts became available again. Clients with CephFS mounts, on the other hand, had the storage nodes automatically blocked after 300 seconds of inaccessibility. These clients required remounts or reboots before the mount became available again.

[1] https://www.bit.nl/news/2773/88/Deep-Dive-in-het-colocatienetwerk-van-BIT

www.bit.org

Because www.bit.nl was not available, the incident was reported on www.bit.org. This notification is automatically texted to customers who have indicated that they wish to receive SMS text messages about incidents and the notification is also placed in an RSS feed. The large amount of visitors who viewed the incident report on www.bit.org caused an overload on the web server for www.bit.org.

After an earlier incident in which www.bit.nl could not be reached and www.bit.org could not handle the load, measures were taken to allow that site to process more visitors. During this incident it appeared that these measures had not been sufficiently effective. The high load on www.bit.org is partly caused by customers who mention the incident report on www.bit.org in communication with their respective customers. This however should not pose a problem for the accessibility of www.bit.org.

After a reboot of the www.bit.org server, the site was accessible again.

Points of improvement

To reduce the chance and impact of a disruption like this recurring, a number of measures have been taken already and others will be taken in the near future:

- All switches have been checked for the configuration that presents risks. Other than the configurations for the specific customer, no other ports/port channels were provided with this configuration.
- The switch supplier is engaged to jointly evaluate the incident. If this evaluation leads to desired changes, these will be implemented.
- The switch supplier has been asked to investigate what options are available to ensure that resource starvation has less or no impact on performance.
- www.bit.org will be hosted in an adapted form that will limit the risk of inaccessibility during peaks in visitor numbers.

Update 01-03-2021

The above report was published on February 18, 2021, one day after the incident. At the time, there was some uncertainty about some aspects of this incident. There is now certainty about these aspects and the incident report has therefore been updated.

- Some responses to the published report involved questions about the risk of the same incident occurring again. That risk is absent. The configuration that (partly) gave rise to the incident is no longer present in the network. On Wednesday, February 17, the configuration in question was removed everywhere. Incidentally, this configuration was deployed for only one customer.
- Re-deploying the specific configuration has been made impossible in the scripts with which configurations are placed on the switches.
- A joint analysis of the incident with the switch supplier has revealed that the switch's forwarding plane should have continued to function despite the resource starvation. The switch vendor has therefore identified this issue as a bug and will work on fixing this bug. To be clear, if the bug is absent in future releases of the firmware, this will not be an incentive for BIT to re-deploy the configuration in question.
- The switch supplier has checked whether the bug can also be triggered by other types of exhaustion of the hardware resources. The supplier has informed us that this is not the case.
- Closer examination of the server hosting www.bit.org revealed that the website did not fail under load. The server had to contend with IRQ errors that were triggered by load on the server. The server has been updated and has not logged any errors since then, nor under load tests. Nevertheless, the server will be replaced. In addition to the monitoring that was already active on the server, regular load tests have been set up for this server.

Contact

If you have any questions regarding this report, please contact our Customer Care department on 0318 648 688 or support@bit.nl.