restena
réseau · sécurité · .lu

Report
Crisis on 19 May 2023

Recorded by: Stefan Winter, CTO

Version: 1.0
Dissemination: PUB

TABLE OF CONTENTS

1 Executive Summary

Restena operates multiple POPs across the country. The two most important ones are POP B ("main POP") and POP L ("Disaster Recovery POP"). Technical precautions are taken such that redundant devices and services can switch between both data centers without any disruption ; particularly for the firewalls that protect the two POPs from the internet. This is achieved by pairs of devices that share and synchronise the common state of the network configuration.

It is important that such pairs of devices mutually agree which of the two is actively forwarding internet traffic at any point in time. A direct connection between the devices is responsible for determining the active and the failover unit.

The root cause for an initial short and controlled outage of the outside connectivity was a failure in this direct connection on the main firewall units. The issue was handled immediately after detection, and a spare part for the faulty module was made ready.

The root cause for the subsequent long-lasting full outage and thus crisis was a software problem in the pair of devices triggered by the replacement of the module in question.

The nature of this software problem created very bizarre issues on the networks connected to the pair of firewalls that made conclusive diagnosis and problem-fixing very difficult. As the sequence of events unfolded, the team discovered many aspects of the existing crisis management plans that were positive and helped in overcoming the situation; but also identified some blind spots that need to be improved for the future. These findings are highlighted in Green and Red in the report.

The actual duration of the outage was approx. 2 h 47 min.

2 Context and Technical Setup

Restena operates multiple POPs across the country. The two most important ones are POP B ("main POP") and POP L ("Disaster Recovery POP").

According to the BCP/DRP Plan and setup, L2 connections are spread between the two, and L3 connections exist to both POPs from the backbone. L3 connectivity is gated with an active/passive failover pair of firewalls, one on each POP.

Routing is dynamically activated between the firewall pair and the backbone by using a private iBGP protocol deployment between the two backbone routers in both POPs with the two main perimeter firewalls in both POPs.

Given the active/passive failover setup, logically, only one of the firewalls is active at any time and assumes the L3 addresses for routing, and maintains the BGP sessions. In this setup, it is important that the two members of the pair negotiate exactly one active node; failure to do so will lead to the gateway L3 addresses being advertised twice by different devices at different places in the network ("dual primaryship", "split-brain") and has significant to catastrophic impact on routing.

The negotiation of primaryship happens via two dedicated dark fibre links (channels of a DWDM spectrum on two distinct fibres) interconnecting the two nodes. The physical device layout allows only exactly one "HA Control" link which is the only way to negotiate primaryship [1]. The second link is "HA Data" and is used mainly to exchange real-time state data between the nodes, so that in case of a failover event, all connections from the original device can be picked up without any disruption from the failover device.

The HA Control link is critical to the functionality of the failover process. Operational experience has shown that it can fail, and the associated protocol often infers correctly that the active unit is still functioning, and the passive unit will not false assume primaryship just because of an HA Control link failure. HA Data is used as a secondary criterion - if state data is incoming via HA Data, but HA Control indicates a failure, the secondary unit will permanently disable itself due to unknown state of the peer.

Once the passive unit has entered disabled state, only a device reboot will bring it back to normal working order. During the reboot phase, the secondary criterion HA Data is NOT being considered, and HA Control being available is then the only way of preventing dual primaryship.

Unfortunately, operational experience has shown that the question whether HA Control is functioning normal or not cannot be determined without doubts: the link itself contains many DWDM channels, and if other channels are functioning, then the HA Control channel itself must also be. On SFP+ level, the electronic components can be queried from the CLI; and failures of the SFP+ module as a whole can be observed. However, there is no way to query the optics layer of the SFP+ module from the device CLI.

3 Prelude

The on-duty engineer alerts the firewall admin that a secondary monitoring system keeps sending notices about some SNMP values that can't be read from the firewall cluster.

The firewall admins checks the device status: the secondary, passive unit is in "Disabled" state due to an HA Control link failure. This could have been due to a short fibre outage in the past, or an ongoing HA Control link problem.

The DWDM channels are transmitting payload traffic, so there is no fibre cut at this point in time. The device reports the interfaces of HA Control "up". In all probability, this indicates it is safe to reboot the secondary unit in the disaster recovery POP to resume to secondary state.

Knowing that there was a failed SFP+ module in HA Control a few years back, which resulted in dual primaryship at the time, the firewall administrators goes to the POP Data Center to be prepared to physically remove all cabling in case the device reboots into dual primaryship.

A remote reboot is initiated.

The devices enter the dual primaryship state.

Within seconds, the firewall administrator unplugs all payload cables, keeping the HA Control, HA Data and MGMT ports connected to allow further investigation. In this situation, both devices still are primary, but the unplugged payload interfaces prevent the local unit from communicating its IP addresses to the outside world, a "lock-in" situation. This allows to connect to the device via out-of-band management for further investigation and makes the DR POP unit handle all L3 communication. This is a known procedure that was executed before.

The few seconds of interruption during the dual primaryship situation are expected, monitoring alerts raise and recover quickly. There is one previously unobserved discrepancy: the BGP session for IPv4 outbound comes up as expected, while the session for IPv6 does not. All POP-internal IPv6 communication works however, so there is an issue with BGP. This in itself is not a crisis, it is logged as a Priority 2 incident by the firewall administrator because only one of 2 IP stacks is affected, and all services continue to be reachable.

Connectivity checks with ping towards the BGP peers reveal an unusual situation: IPv4 pings succeed, IPv6 pings fail with "can't assign requested address". The configuration is unchanged from before, and the device complains that it cannot assign its own address to itself as the source address for the ping. This is an explanation for the BGP session failing - the device is not sending any packets - but there is no good explanation for the address assignment phenomenon in itself.

The issue at hand was fixed by unconfiguring and re-configuring the exact same address on the outside interface. No further change was required; BGP for IPv6 came back up and the situation returned to normal. The fact that this fix worked suggests that the problem may have been due to Duplicate Address Detection in IPv6, which led to permanent disabling of the address.

Subsequent investigations with the SFP+ module of HA Control indicated that the SFP+ module was indeed faulty: after re-seating the module, its activity LED flashed for an extended amount of time (>1 min) but stopped at some time.

Naturally, for the SFP+ module in question, a spare was in stock. As long as HA Control did not work, the failover capacity was not working, and a long weekend with official holidays was ahead. The SFP+ module should be replaced soon.

4 Crisis

The firewall administrator unplugs the defunct SFP+, and plugs in the replacement SFP+ module, as has been done on earlier occasions.

The entire network in both the main POP and DR POP becomes unresponsive. [first alerts at 15:43 local time]

> *Observation (positive): This incident corresponds to an identified crisis scenario in Restena's BCP/DRP plans.*
> *It was useful that all BCP/DRP policy documents were available as an offline copy, because file server functions were not accessible.*

A crisis is formally declared if this particular type of incident lasts longer (or is expected to) than 2 hours. Nevertheless, all on-site personnel is alerted immediately and the crisis plan is invoked.

Notably, the connection to the firewall cluster's out-of-band management ports is not reachable any more; a connection which does not traverse the data plane of the firewalls in question at all.

The connection from a third-party internet connection via the company VPN can be established, but there is no name resolution available inside the VPN.

> *Observation (negative): Restena workstations and servers are by default configured to contact two distinct DNS resolvers for name resolution; one of which is an anycast system that exists both inside and outside the data center in question. However, when logically inside the DC, the anycast routing in all cases needs to traverse the pair of firewalls in question; which is not possible in this scenario. Lesson learnt: the backup resolver also has a dedicated unicast address in the DRP site which continues to operate. In the future, servers and workstations will be configured to use that unicast address as a fallback for DNS resolution.*

While all data cables in the primary unit were and remain disconnected, and it is proven not to cause issues that the "lock-in" dual mastership situation exists, the only useful explanation for an observed connectivity loss at this stage is that, somehow, the primary unit does process traffic and the two units are in a dual primaryship situation. This cannot be confirmed without doubts since no out-of-band connectivity to the units can be established. To resolve the situation, the firewall administrator decides to fully turn off the primary unit, so that there is no possibility for a dual primaryship, and that traffic will resume to normal.

The primary firewall unit gets turned off. There is no change to the sorry state of network connectivity.

At this point, crisis communication to external parties should have commenced. This was delayed due to access issues to the foreseen communications channels:

> *Observation (negative): Passwords for access to platforms like Twitter were stored in a server-based password manager; for a successful access, DNS resolution for the communications team on their workstations, and the connections to the webserver frontend as well as database backend are needed. As a result, the communications team had significant difficulties reaching out to the external parties.*

Further investigations show: name resolution from outside the VPN from the internet works, with significant delay, jitter and packet loss. From this it follows that the secondary firewall in the DRP POP is in principle working, and routing traffic. Delay, jitter and packet loss are almost always signs of a layer 2 forwarding issue; like a loop which forwards packets indefinitely and saturates all available bandwidth.

With this information, the firewall admin informs the network team that the issue must include a L2 component. The network team begins an investigation on all the L2 network equipment across the POPs. At this point, work is progressing very slowly due to the lack of DNS resolution: it is not possible to connect to management UIs to check link saturation (requires access to web servers, with SNI and host name resolution); it is also required to resolve host names of the switches from outside the engineering/VPN network, and then connect directly with those slowly retrieved IP addresses to each switch, one at a time.

> *Observation (positive): a temporary loss of DNS resolution is already anticipated in the Business Continuity Plan, and as a precaution, some key names and IP addresses are stored statically in a Knowledge Base article.*

> *Observation (positive): the loss of accessibility to the Knowledge Base (hosted on a web server after all) is already anticipated in the Business Continuity Plan, and the contents of the Knowledge Base are stored in an office copy with the CTO (who happens to be the firewall admin himself).*

With enough IP addresses resolved, the network team logs in and attempts to close possible forwarding loops in the switch configuration by turning off select network ports. In the course of that, the L2 connectivity between the main POP and DRP POP inadvertedly gets disrupted entirely.

This leads to the unavailability of the network-side out-of-band management network, and the L2 connectivity loss cannot be un-configured remotely.

The network team sets out to re-configure the needed network ports directly at the chassis, using RS232 serial cables for direct management access.

As L2 connectivity between the POPs gets restored, it is the subjective experience of the author of this document that the jitter/loss/delay issue got significantly better, and the remainder of the problems were visible more deterministically. It is probable that the L2 cut-off was maybe overdoing it, but it may very well have contributed to the resolution of the issue by removing parts of the symptoms.

Access to the firewall's own out-of-band management ports became possible by utilising a layer 2 cable pre-configured for cases of disaster in the CTO's office, which was directly connected to the VLAN of the server out-of-band network. Connectivity to the firewall cluster became more easily possible, and it was possible to confirm the configuration and failover state via the out-of-band ports. The firewall reported a normal operational state.

> *Observation (positive): the availability of this physical link to the out-of-band network was helpful, even if not codified in the BCP/DRP plans.*

The DNS team provided information about a DNS resolver residing in a dedicated DRP network which could be accessed without traversing the impacted firewalls. As staff manually reconfigured their DNS resolvers to this auxiliary address, debugging became a lot simpler (see observation above).

As more IP addresses got resolved, L2 connectivity restored, and more reachability tests with ICMP Echo Req/Reply could be carried out, it surfaced that many individual links were fully functional, but routing was not in order. One example, which is still not fully explained even in hindsight, is a link (engineer workstation) -> (internal firewall) -> (external firewall) -> (DNS resolver). Each and every direct hop worked flawlessly, and all routes in both directions were correctly configured, but still, it was not possible to reach the target routed via these hops. Investigating where the packet got dropped (tcpdump...) showed a bizarre situation: the internal firewall accepted the incoming packet, but did not even attempt to forward the packet on its outgoing interface towards the firewall cluster. This explained well why the DNS resolution was deterministally broken from inside the engineer/VPN network; but continued to work somewhat from the outside internet. To date, no viable explanation for this behaviour of the internal firewall was found. This apparent misbehaviour of the internal firewall shadowed the actual problem.

At some point, the main POP firewall was turned on again, for it to take over traffic in case the secondary unit had a yet unobserved issue. Failover was initiated and the main POP firewall took over. There was no change at all to the connectivity situation.

The firewall admin and network team members discussed possible further causes; there were none. As a straw, the team considered the possibility that the runtime state between both firewalls might be corrupt, and that any failovers between the two instances would carry over the corrupt state. It was decided to eliminate this possibility by turning off both instances of the firewall cluster simultaneously. This would forcibly lose all runtime state and cause the firewall to restart from pure configuration.

This measure was executed by permanently powering off the remote unit in DRP POP, and rebooting the local unit in POP B. Prior to the reboot, configuration was dumped to compare it with a known-good state later on.

Network connectivity was restored and the situation returned to normal. [recovery alerts came in around 18:30]. The immediate issue was thus resolved after approx. 2:47.

5 Aftermath

After waiting some time for the network and monitoring system to stabilise, it was decided that the secondary unit in the recovery site should be turned back on to restore redundancy.

This involved travel to the secondary site to push the device’s power button. When the secondary device came back up [19:42 production link up] it correctly assumed its secondary failover state and is working normally since. The crisis was declared overcome at this point.

A check was performed on whether the firewall configuration was still the same as the last known-good state before the incident. The comparison with the configuration in Git compared to the cluster’s current config reveals that the cluster is working with a ruleset that is four days old. It appears that configurations were not backed up for some time, hinting to the existence of a bad cluster state prior to the immediate crisis at hand. It was not possible to find out in hindsight why no backups were completed in these four days preceding the incident.

The configuration was brought back up to date, and the firewall cluster is behaving normally since.