Inspire Net Incident Report

What	Core Juniper Routing Platform at Inspire
Time from	27 March 2019 0900
Time to	27 March 2019 1830

Background

Inspire Net's main core routing network runs on 2 * Juniper MX480 routers, with full resilient power and routing engines. These run the latest long term stable releases of Junos, with routing engines updated using in service upgrades when required. They had an uptime of 560 days in some cases.

The main routing platform in Palmerston North starting exhibiting non-standard behaviour at approximately 9am on 27 March, after various upstream routing changes that caused increased load on the routing control plane.

Once the control plane became overwhelmed, the router started intermittently and randomly dropping services.

Outage Cause

NOC diary entries at Inspire for the event

1am : eBGP to Spark started flapping – non catastrophic, failed over to other upstream providers

9am : Logged fault with Spark

9.10am : Realised that other eBGP sessions are flapping. Logged JTAC ticket.
***At this stage with the TAC we did a graceful RE changeover that was very ungraceful. This caused an MPC / PFE reboot. This likely triggered a RPD bug - potentially also related to Juniper PR 1092009 .**

At this stage we're also likely being affected by Juniper PR 1311224

10.15am : Did first graceful RE switchover on pmr-inspire-cr-1. This wasn't graceful and our MPC (line card) in slot 0 rebooted.
*** This then caused LACP instability which caused OSPF instability from about 10.15am to mid-afternoon.**

Diagnosing the issue was hindered by this bug – Juniper PR 1220061

10.30am : Problem changed to be a OSPF flapping issue. Update JTAC ticket.

11am : Turn off OSPF BFD.

11.30am : Realise that OSPF is flapping due to AE / LACP issues. This is because the control plane cannot process LACP packets. JTAC (wrongly) blame our lack of loopback filter reject rule.

1pm : We have a loopback filter rule in place. OSPF flapping is still occurring. Carry on updating JTAC ticket.

2pm : next session with JTAC (phone call). We do a non graceful RE switchover. We fix the OSPF/LACP issue at this stage. Everything seems stable.
*** Mid afternoon we did another RE reboot and graceful switchover. This then triggered the main bug of Juniper PR 1312308**

2.30pm : We (the NOC) realise that now we have interface issues with VPLS instances. Instances won't come up as they have no active interfaces. Other VPLS and interface issues occur.

3pm : Ticket is escalated to ATAC (advanced JTAC).

3.30pm : We realise that interface changes are just being queued and not handed by the various daemons.

4pm : Call starts with ATAC. During this call diagnostics information is collected on the issue.
***Fixes are applied to REs due to PR1312308. Then around 6pm need to do a non-graceful restart of REs again. Around 6.30pm call ends and everything is stable.**

6.30pm : Start working on remaining issues, moving traffic around, fixing snmp and nagios.

7.30pm : Everything seems good.

To briefly summarise
Junos, the operating system on these routers, runs in a container in a virtual environment on the Juniper hardware. There are some processes in the lower level virtual machine that caused the routing plane to become less responsive over time, which in turn caused intermittent processing issues on the router. This should have been resolved by a graceful failover to the other routing plane, but this was in turn catastrophic due to a fault introduced by the first bug.

Diagnosing all of the issues during the day was hindered by us being affected by 3 or 4 separate problem reports. An upgrade to fix 2 of these issues should have occurred on Jan 27 2019 but unfortunately failed due to one of the other issues. A decision was made to migrate certain customers away from this core router before attempting the upgrade again and unfortunately there were problems scheduling outage windows with particular customers. However, this previously planned upgrade on Jan 27 would not have prevented the main virtualisation bug that occurred on March 27.

Restoration

A temporary full restoration of services point was made at 6:30pm to a stable and working routing environment. A decision was made to stop work and let technicians rest and recover and plan a full router reboot to address issues raised by JTAC / ATAC

Full mitigation of the bug in the router will be made on Friday 29 March 2019, which will require up to 3 * up to 45 minute outages between 2am and 6am during an emergency outage window.

During this outage we will be performing patches on the hardware / virtual environment on the router to prevent future occurrences of this fault

Actions arising / Comments

This fault effected approx. 80% of our customers on an intermittent basis. Even though we have a fully resilient routing platform, failover of a lot of services did not occur due to the intermittent nature of the fault, with residential customers only failing over on a total outage basis, and several larger customers not using BGP or other resilient options to both our core routers for their DR plans to either us or to another provider. One of our larger customers quoted to us "you are our DR plan, you never have outages", which while it would be nice if it was true, outages are unavoidable with some of the things that are beyond our control.

This is the second longest fault we have had in 20 years, with the last major fault requiring Vendor equipment to be shipped in from Ireland, which is why we installed a fully resilient routing solution.

Our Vendor has confirmed that the issues outlined above resulted in the fault we experienced and that the mitigations put in place during the emergency outage window address the root causes of the outage.